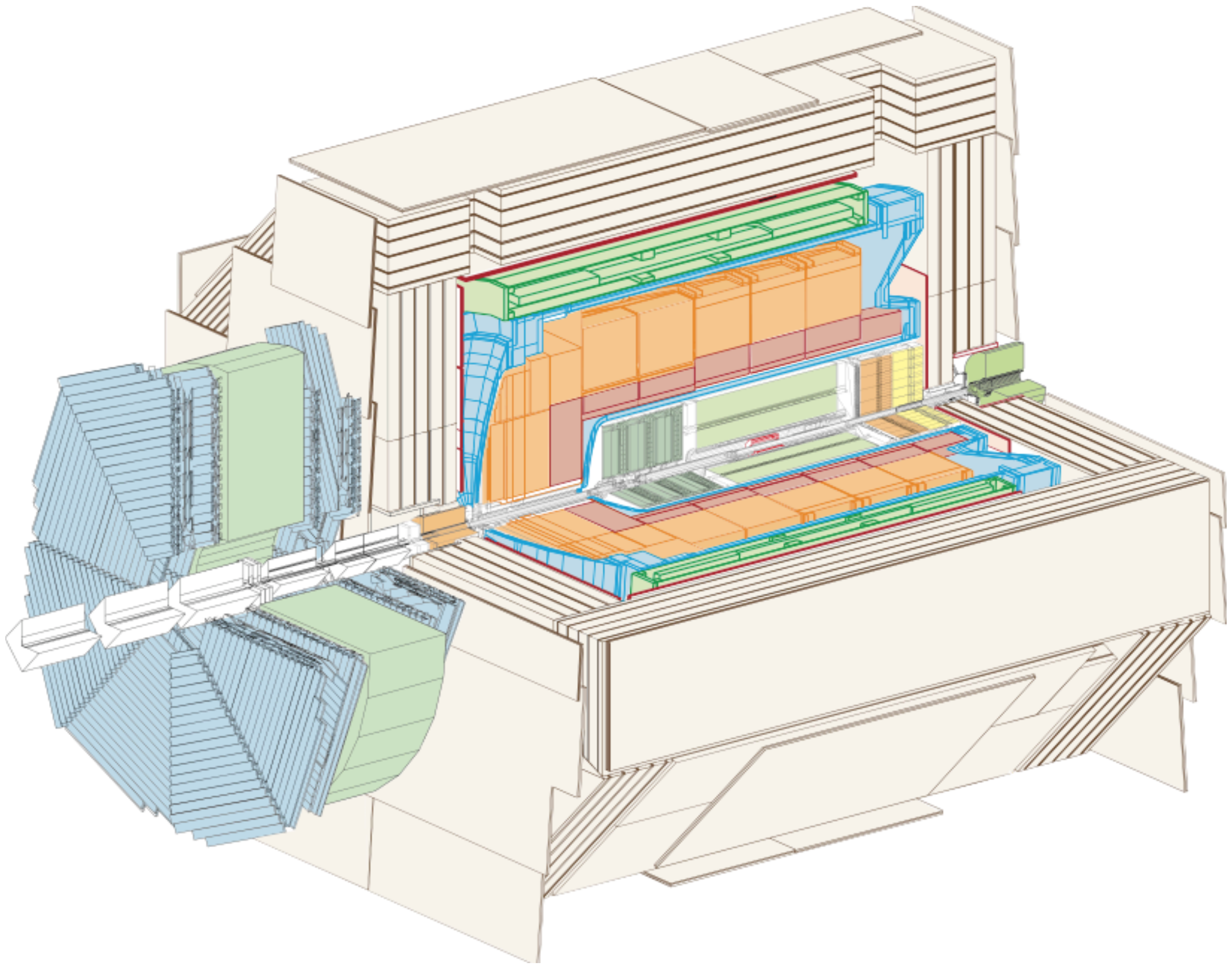
Unfolding Measurements at H1 using Machine Learning

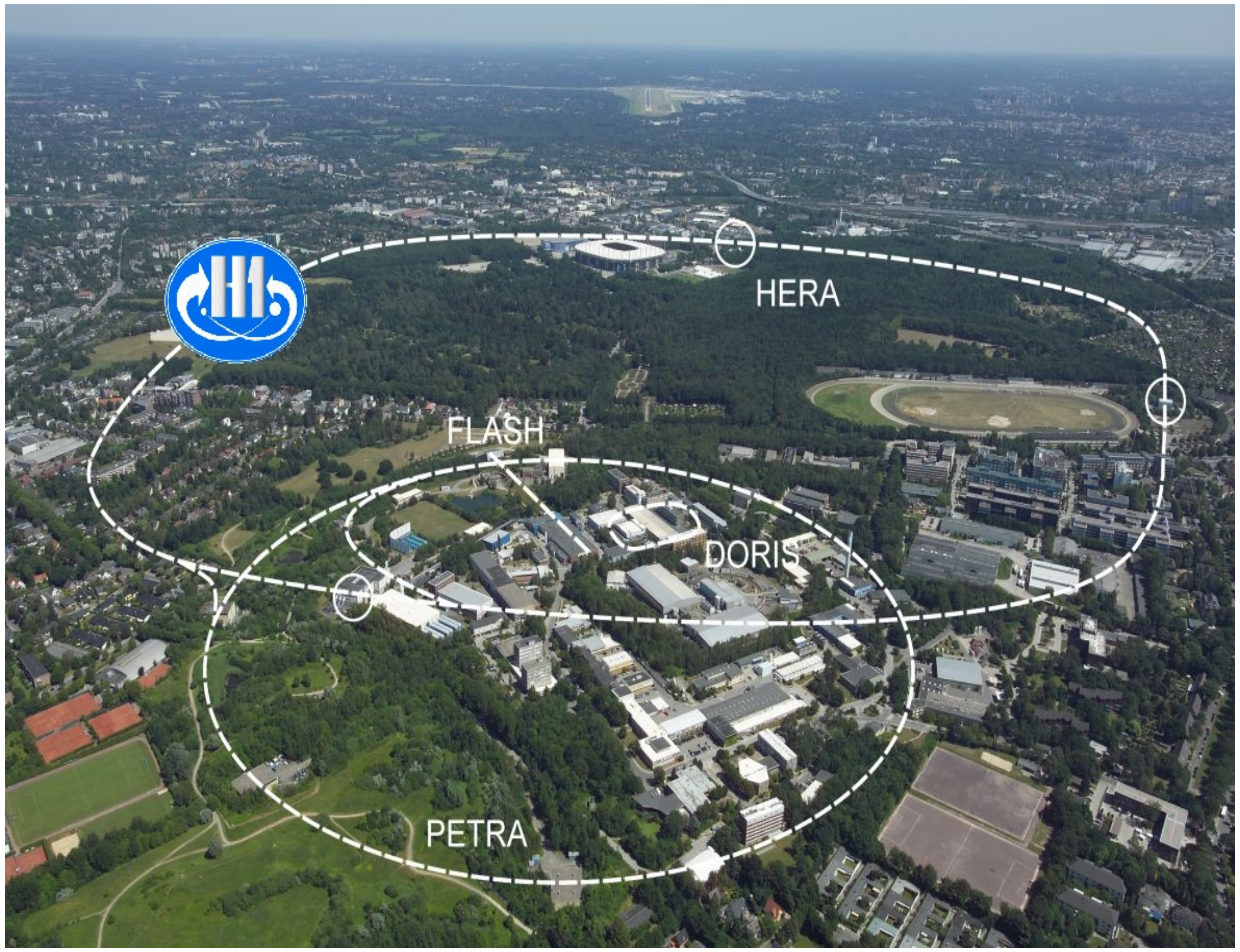
Fernando Torales Acosta

Quick Overview

- Unfolding + OmniFold
- First OmniFold Measurement
- Previously inaccessible observable
  - Made possible with OmniFold
- OmniFold for Jet Substructure

H1 at HERA





- **H1 Detector at the positron-proton collider, HERA. Hosted in Hamburg Germany**
- **Major goal was to study internal structure of the proton through deep inelastic scattering**

$$e(k) + q(p_1) \rightarrow e'(k_\ell) + jet(k_J) + X$$

HERA publication overview

- HERA operated from 1992- 2007
- Both ZEUS and H1 are still active
  - Data *AND* simulation are available to members for analysis
- HERA data used to study PDFs and perturbative QCD, low-x and diffraction, transition from soft to hard QCD

H1+ZEUS combined	8 publication
H1	223 publications
ZEUS	250 publications

Top-ten cited (excluding detector papers)

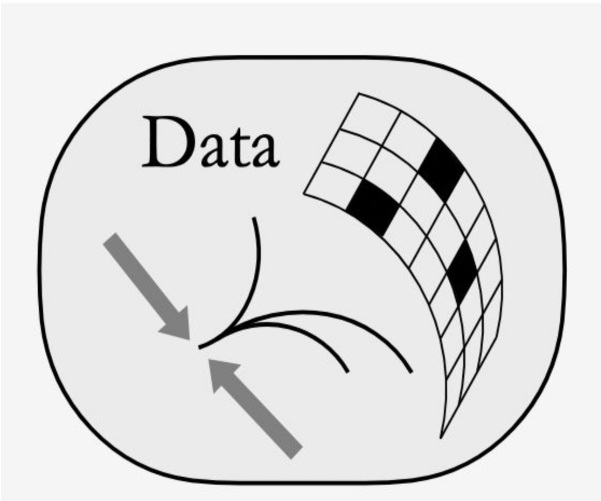
JHEP 1001 (2010) 109	H1+ZEUS	1000+ Data combination, PDF
Eur.Phys.J. C21 (2001) 33	H1	700+ Low-x, PDF, alpha_s
Nucl.Phys. B470 (1996) 3	H1	500+ Low-x, PDF
Eur.Phys.J. C21 (2001) 443	ZEUS	500+ Low-x, PDF
Phys.Lett. B315 (1993) 481	ZEUS	500+ Observation of diffraction
Nucl.Phys. B407 (1993) 515	H1	400+ Rise of F2 at low-x
Eur.Phys.J. C75 (2015) 580	H1+ZEUS	400+ Data combination, Low-x, PDF
Phys.Lett. B316 (1993) 412	ZEUS	400+ Rise of F2 at low-x
Z.Phys. C76 (1997) 613	H1	400+ Difffractive PDF
Z.Phys. C74 (1997) 207	ZEUS	400+ High Q^2 DIS

Really great example of maintaining ‘legacy’ datasets as our analysis methods improve

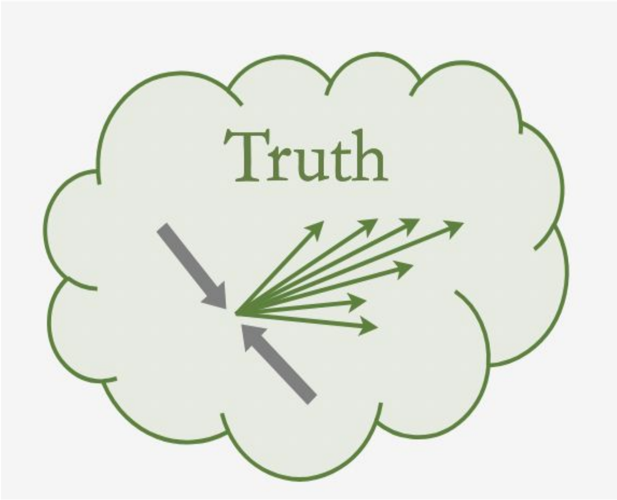
Unfolding

- Essentially: We want to remove unwanted detector effects from our experimental data
  - correct a whole dataset on a statistical level
  - combine data from multiple sources
- Un-binned?
  - Re-bin option for future analysis
  - Modify phase space in the future
  - New Observables that are function of previous unfolding

Detector-level

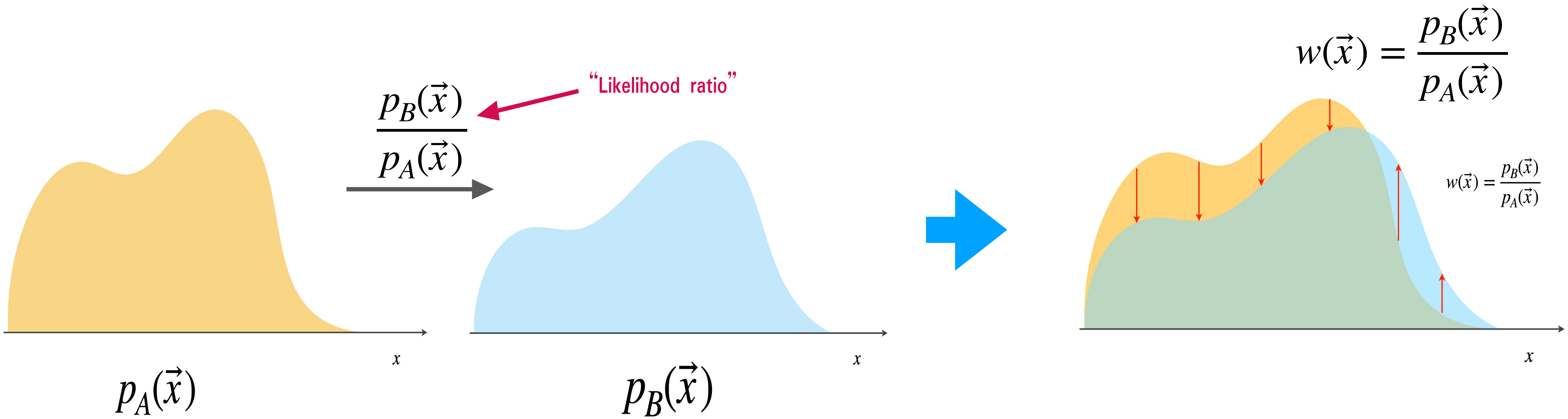


Particle-level



Motivating ML + Liklehood Ratios

How can we adjust one distribution to look like another?



- In practice, directly learning the individual densities, $p_A(\overrightarrow{x})$ and $p_B(\overrightarrow{x})$ is difficult

- Machine learning (classifiers) can directly approximate the ***ratio*** of the liklehoods

Credit: Mariel Pattee

Classifier functions can be re-used to directly approximate a likelihood ratio.

A vanilla NN classifying between two classes could be trained using **binary cross-entropy loss:**

$$L_{\mathrm{BCE}}[f] = -\int \mathrm{d}x \,\big(p_A(x)\,\log(f(x)) + p_B(x)\,\log(1-f(x))\big)$$

where $f(x)$ is the output of a NN classifier, and our datasets are sampled from these two probability distributions $p_A(x)$ and $p_B(x)$.

Credit: Mariel Pattee

Classifier functions can be re-used to directly approximate a likelihood ratio.

A vanilla NN classifying between two classes could be trained using **binary cross-entropy loss:**

$$L_{\mathrm{BCE}}[f] = -\int \mathrm{d}x \, \big(p_A(x)\,\log(f(x)) + p_B(x)\,\log(1 - f(x))\big)$$

To find where this is minimized, we need to find the extremum, i.e. differentiate with respect to $f(x)$ and set equal to 0:

$$\begin{aligned}\frac{\partial L}{\partial f} &= -\frac{\partial}{\partial f}\big(p_A(x)\,\log(f(x)) + p_B(x)\,\log(1 - f(x))\big) \\ &= -\frac{p_A(x)}{f(x)} + \frac{p_B(x)}{1 - f(x)}\end{aligned}$$

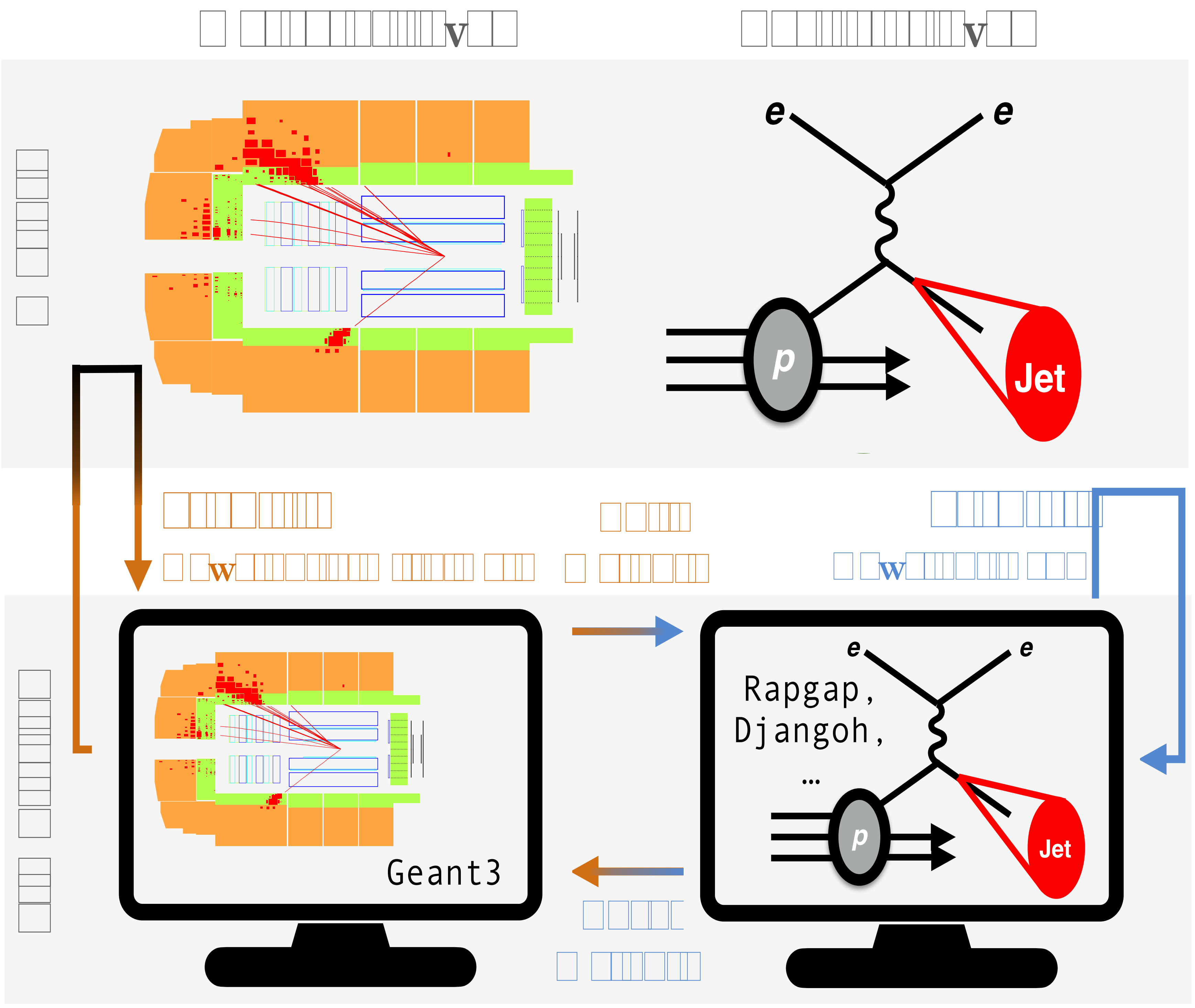
$$\frac{\partial L}{\partial f} = 0 \Rightarrow \frac{f(x)}{1 - f(x)} = \frac{p_A(x)}{p_B(x)}$$

Rescaling of classifier output

Likelihood ratio

Credit: Mariel Pattee

OmniFold



2 step iterative approach

1. Events from detector level sim. are reweighted to match the data
2. Create a “new simulation” by transforming weights to a proper function of the generated events

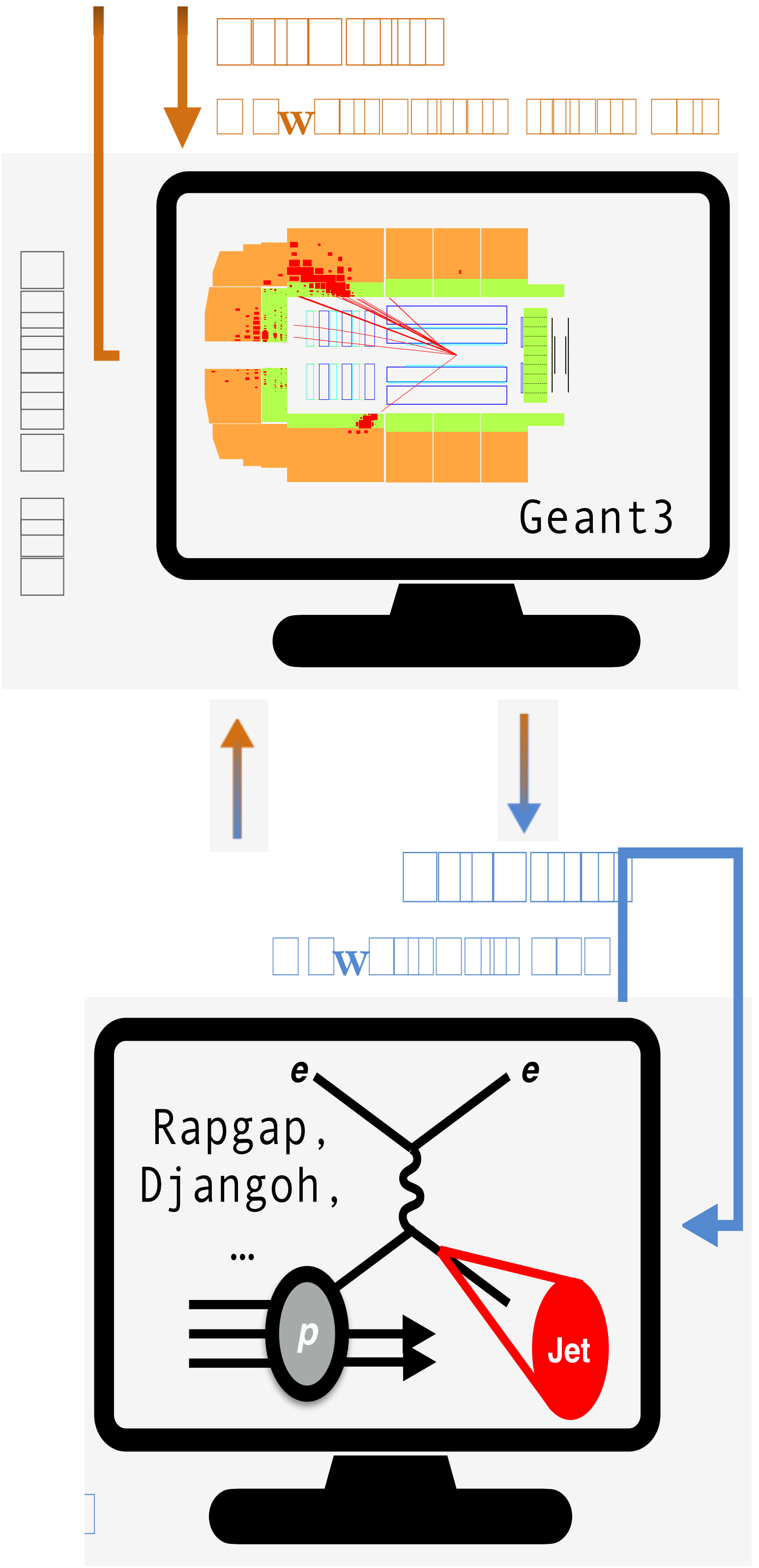
Classifiers used to approximate **2** likelihood functions:

1. reco MC to Data reweighting
2. **Previous** and **new Gen** reweighting

Some pretty consistent numbers: 4-5 Iterations, for single ensemble, ~5 ensembles

OmniFold Terms

- Dimensionality:
  - 1 Observable: UniFold
  - Many: MultiFold
    - (Often used interchangeably with OmniFold)
  - All: OmniFold
- Steps
  - Step 1: Detector sim to Data
  - Step 2: Old Particle-level to new Particle Level
- Iterations: Loops of OmniFold
- Ensembles: Repetition of the unfolding
  - To mitigate randomness from the model



Experimental Setup

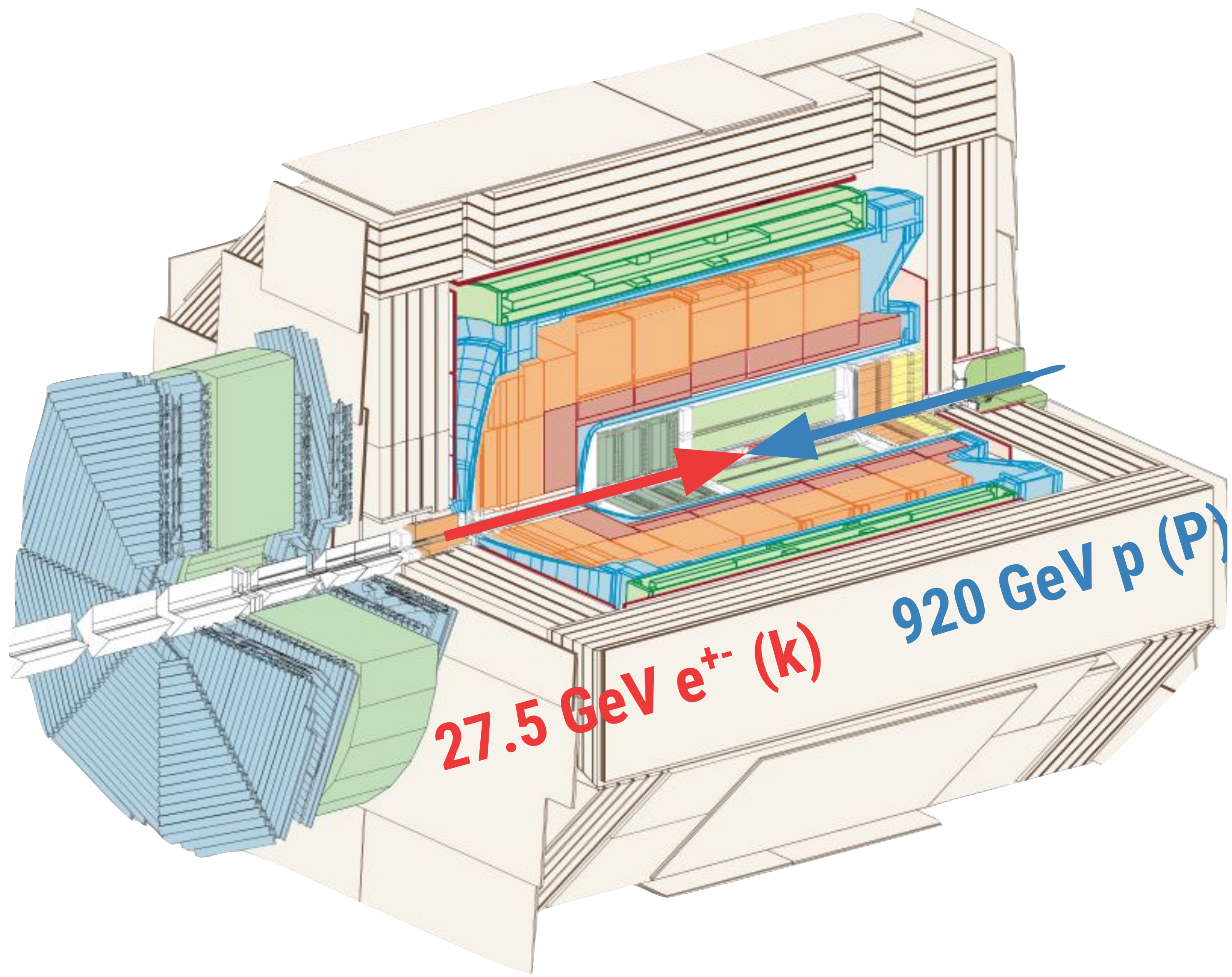
- H1 Data from 2006 and 2007 periods at 228 $\mathrm{pb}^{-}1$(130 pb^{-1})
  - Positron-proton and electron proton collisions
  - $\sqrt{s} = 318$ GeV

- Fiducial Cuts:
  - $0.2 < y < 0.7$
  - $Q^2 > 150\ \mathrm{GeV}^2$
  - $p_T^{\mathrm{jet}} > 10\ \mathrm{GeV}$
  - $-1 < \eta_{\mathrm{lab}} < 2.5$
  - $\mathrm{k_T}, R = 1.0$
  - $q_\perp / Q < 0.25$
  - $q_\perp / p_{\mathrm{T},jet} < 0.3$

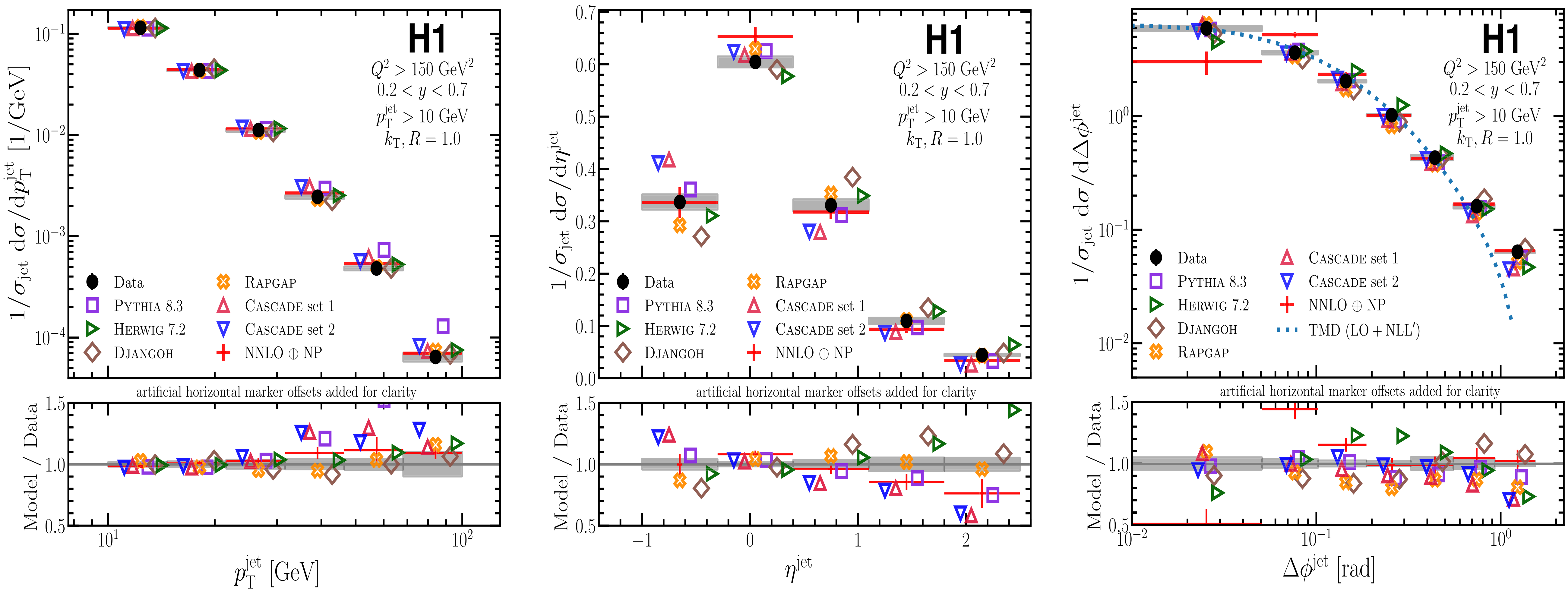
$Q^2 = - q^2$

$y = Pq / pk$

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

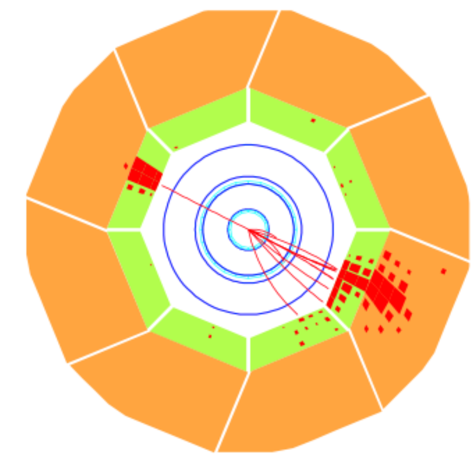


H1 Differential Cross Sections (Lepton-Jet correlations)



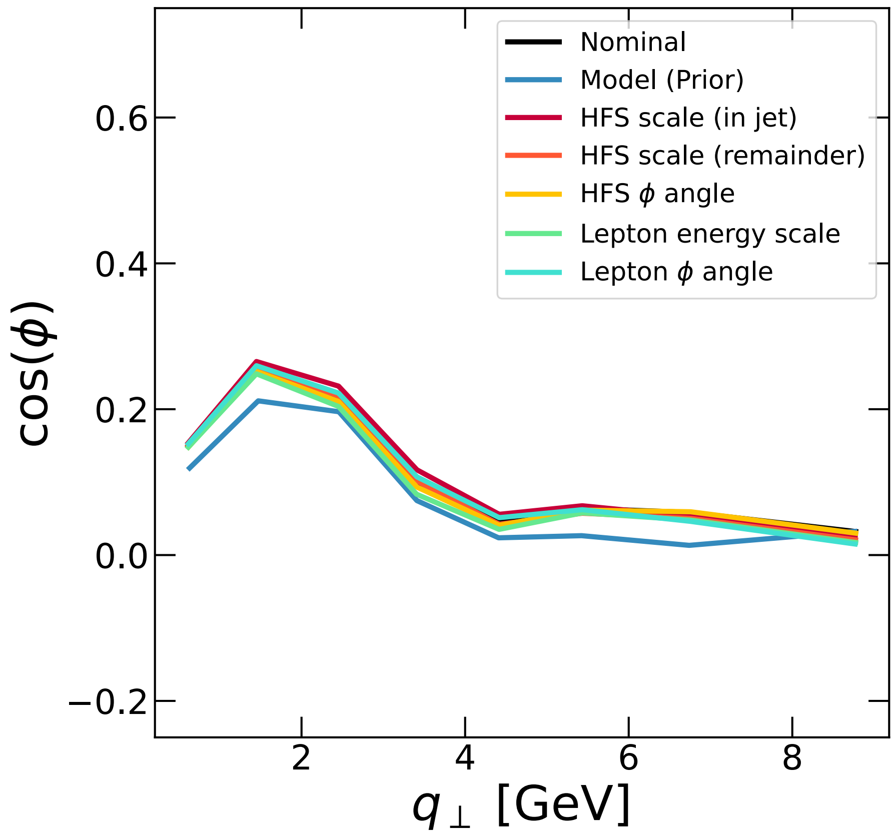
First multidimensional un-binned unfolding using OmniFold

- Production cross sections as a function of p_T^{jet}, η^{jet}, $\Delta\phi^{jet}$
- Simultaneous unfolding of eight observables:
  - $p_x^e,\ p_y^e,\ p_z^e,\ p_T^{jet},\ \eta^{jet},\ \phi^{jet},\ \Delta\phi^{jet},\ q_T^{jet}/Q$



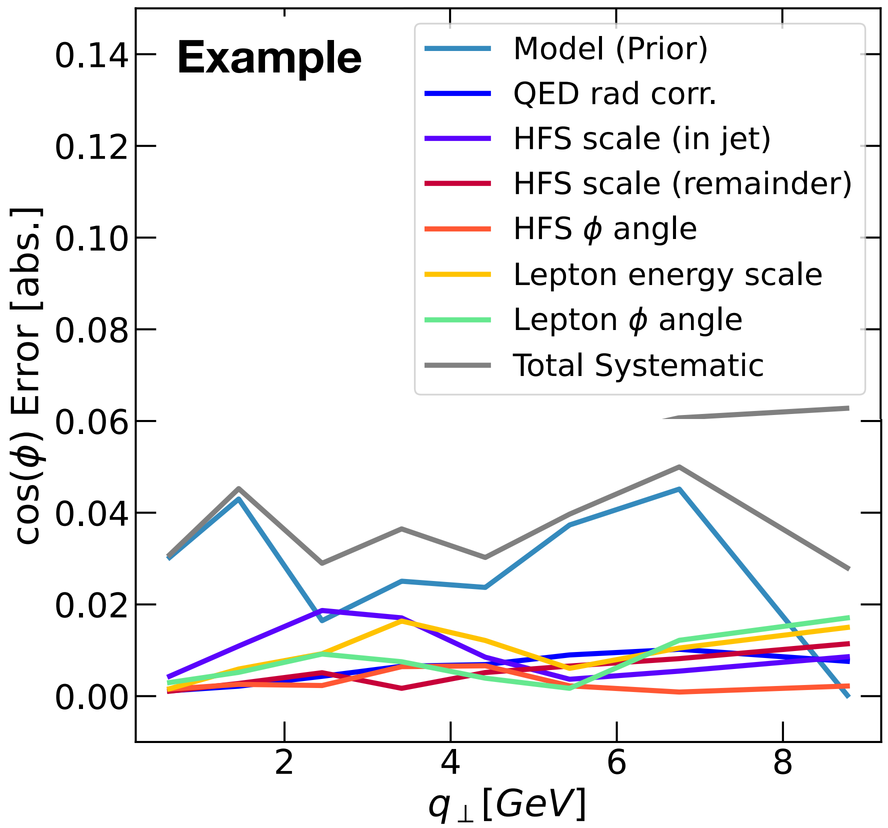
$$e(k) + q(p_1) \to e'(k_\ell) + jet(k_J) + X$$

Systematic Uncertainties



General Procedure

- Systematically vary MonteCarlo
- Both detector level and generator level sim.
- Re-do entire analysis, **including unfolding**
- Take full difference of systematic variations as uncertainty

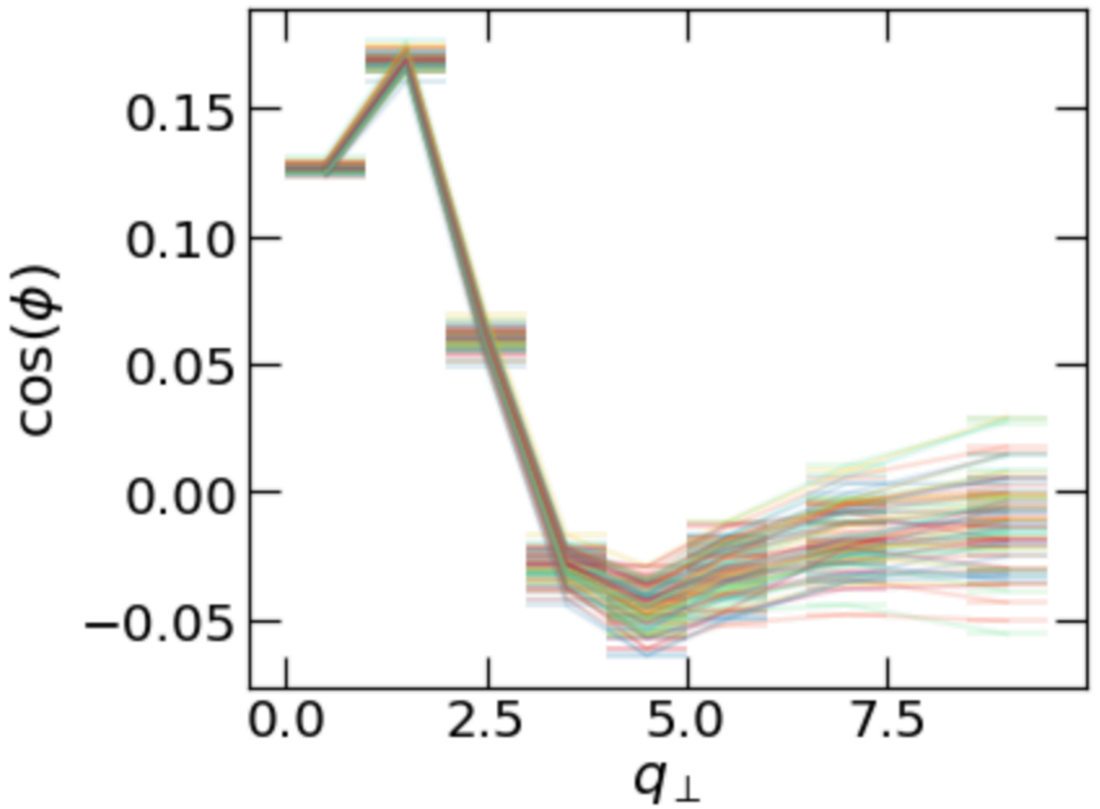


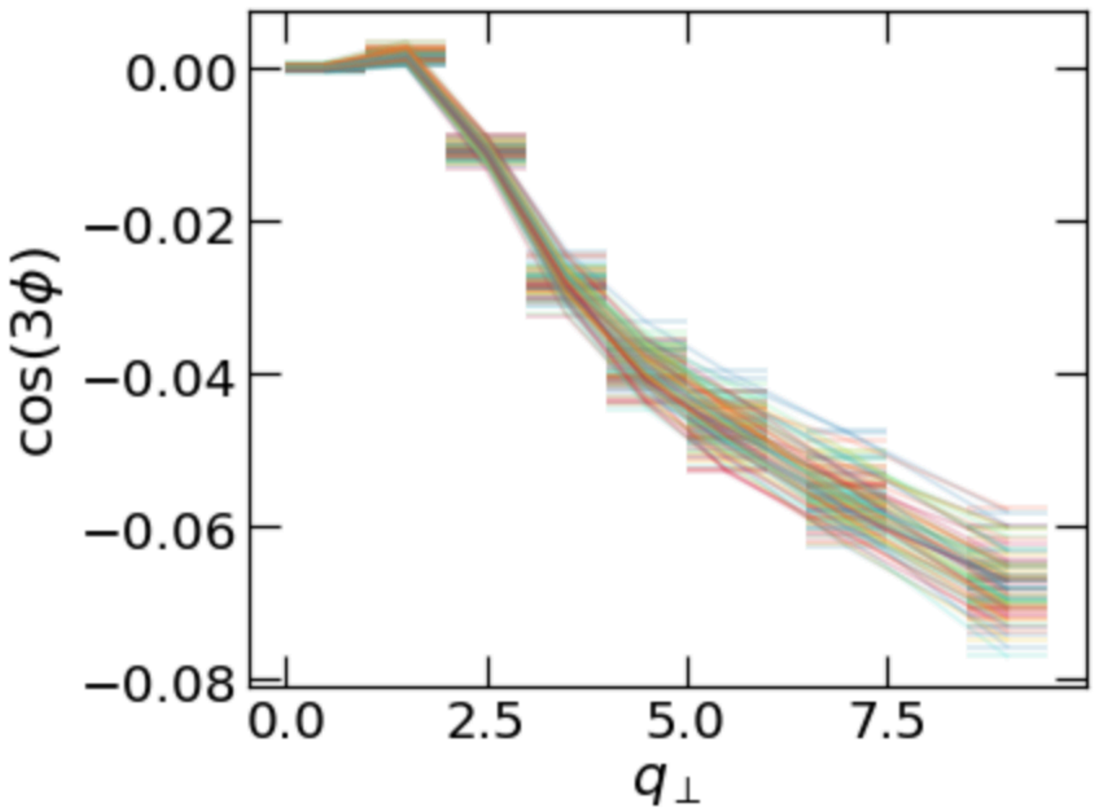
Systematic uncertainties considered

- **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
- **QED uncertainty**: Use the variation of measured quantities when radiation is turned off in the simulation

Bootstrapping Uncertainty

- Simulate different ensembles of data
  - Each event is given an initial weight according to a poisson distribution with $\mu = 1$
  - Simulates ~100 “pseudo datasets”
  - Estimates statistical uncertainty of dataset
- Repeat entire unfolding process with different ensembles
  - Save final NN weights of OmniFold Procedure
  - Take the standard deviation of the spread in the unfolded results as the statistical uncertainty



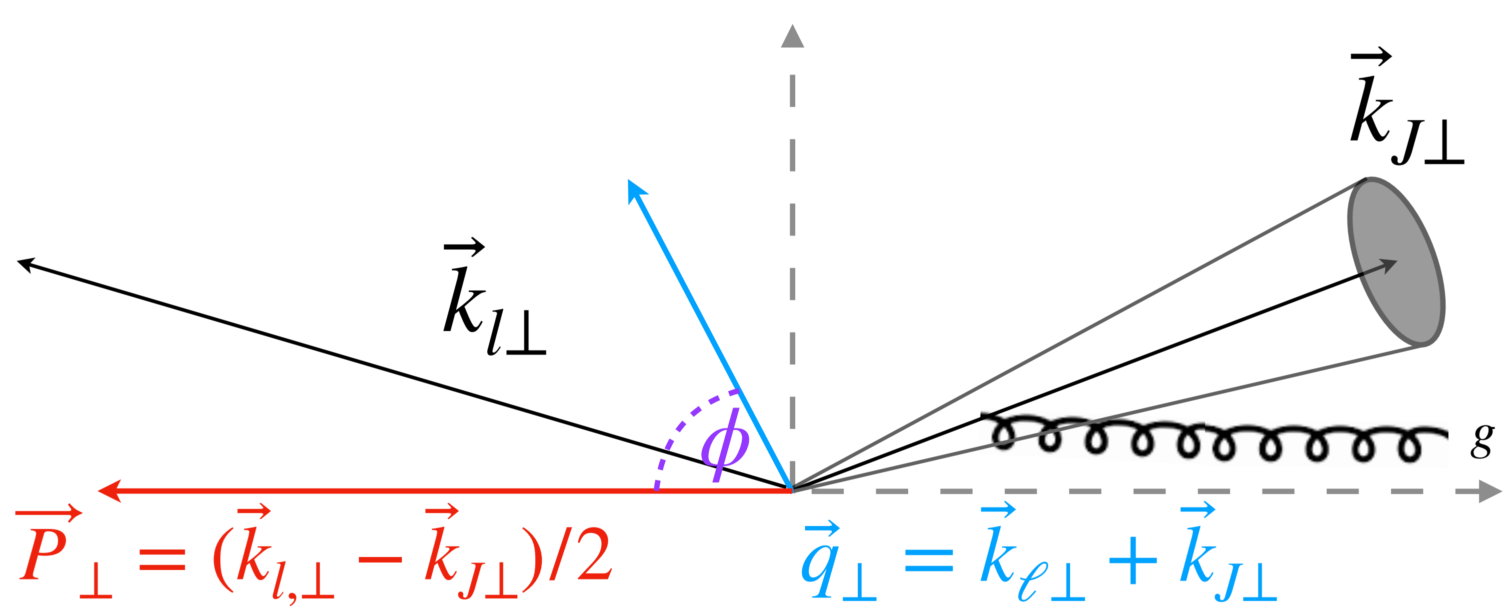


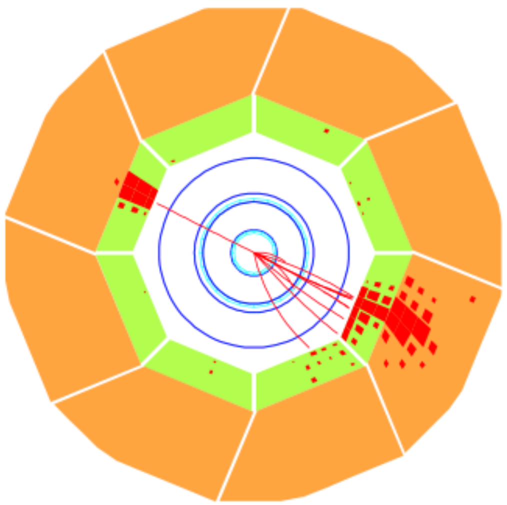
Lepton Jet Asymmetry

Observable that was previously impossible to unfold!

- Total transverse momentum of the outgoing system $\vec{q}_\perp = \vec{k}_{\ell\perp} + \vec{k}_{J\perp}$, is typically *small* but *nonzero*
- Imbalance can come from perturbative initial and final state radiation
  - e.g. Emission of soft gluon with momentum $k_{\perp g}$
  - unrelated to TMDs or intrinsic transverse momentum of target gluons
- Depending on kinematics, soft gluon radiation can dominate
  - Radiative corrections enhanced approximately as $(\alpha_s \ln^2 P_\perp^2/q_\perp^2)^n$

$$P_\perp \gg q_\perp$$





$e(k) + q(p_1) \to e'(k_\ell) + jet(k_J) + X$

Lepton Jet Asymmetry

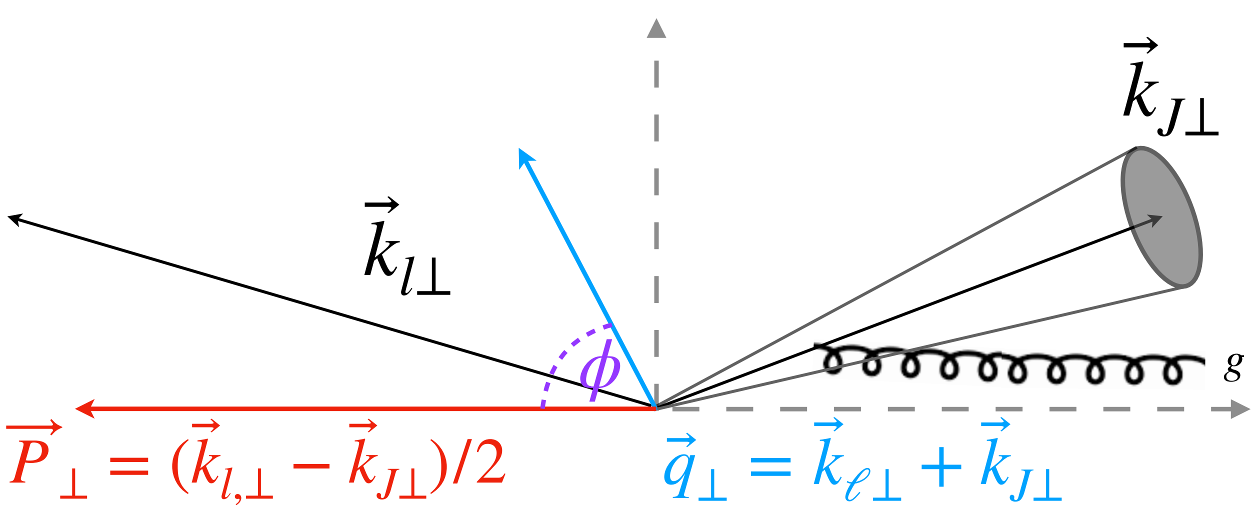
Key Ingredients:

- $q_\perp$ = ***Total* transverse momentum**
- $P_\perp$ = **Transverse momentum *difference***
- $\boxed{\phi = \text{Angle between } q_\perp \text{ and } P_\perp}$

$$\vec{q}_\perp = \vec{k}_{\ell\perp} + \vec{k}_{J\perp}$$

$$\overrightarrow{P_\perp} = (\vec{k}_{\ell\perp} - \vec{k}_{J\perp}) \;/\; 2$$

$$\phi = \text{acos}[(\vec{q}_\perp \cdot \overrightarrow{P_\perp}) \;/\; \vec{q}_\perp \quad \overrightarrow{P_\perp} \;]$$



Momentum conservation:
$\vec{q}_\perp = -\sum_i^{soft} \vec{k}_{i\perp}$

Final Observable:
$\langle \cos(n\phi) \rangle$ **for n = 1, 2, 3**

Multifold used to unfold:
$p_x^e,\ p_y^e,\ p_z^e,\ p_\mathrm{T}^\mathrm{jet},\ \eta^\mathrm{jet},\ \phi^\mathrm{jet},\ \Delta\phi^\mathrm{jet},\ q_\mathrm{T}^\mathrm{jet}/Q$

Asymmetry Motivation

1. Probes soft gluon radiation $S(g)$
   - Soft gluon radiation can be the primary contribution to asymmetry
   - 10.1103/PhysRevD.104.054037

2. Asymmetry is perturbative
   - Opportunity to compare to unfolded H1 data

3. May represent a vital reference for other signals, in particular TMD PDF measurements
   - Factorize contributions TMD PDFs and Soft gluon radiation

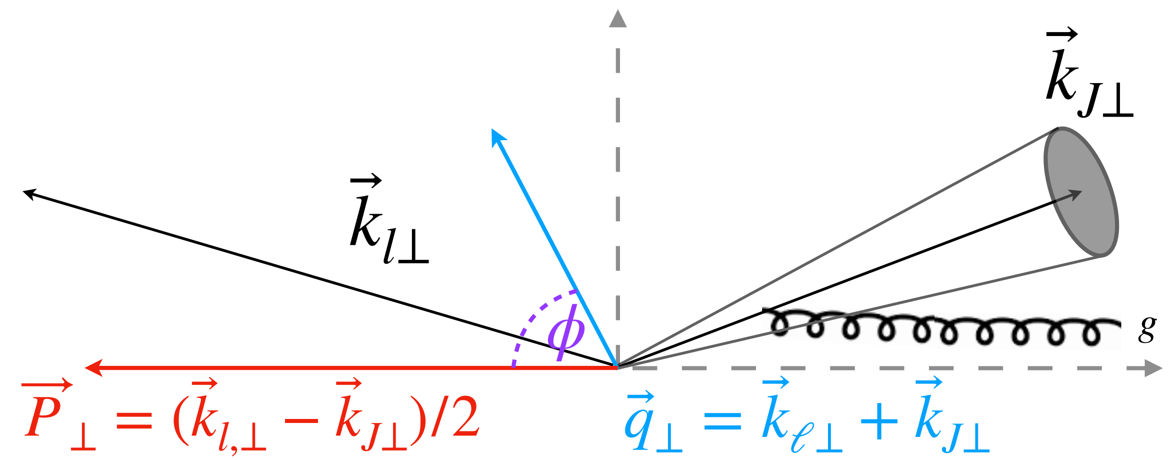
4. Observable is sensitive to gluon saturation phenomena, possibly measurable at the EIC
   - 10.1103/PhysRevLett.130.151902

Putting it Together*

$$\phi = \mathrm{acos}[(\vec{q}_\perp \cdot \overrightarrow{P_\perp}) \ / \ \ \vec{q}_\perp \ \ \overrightarrow{P_\perp} \ ]$$

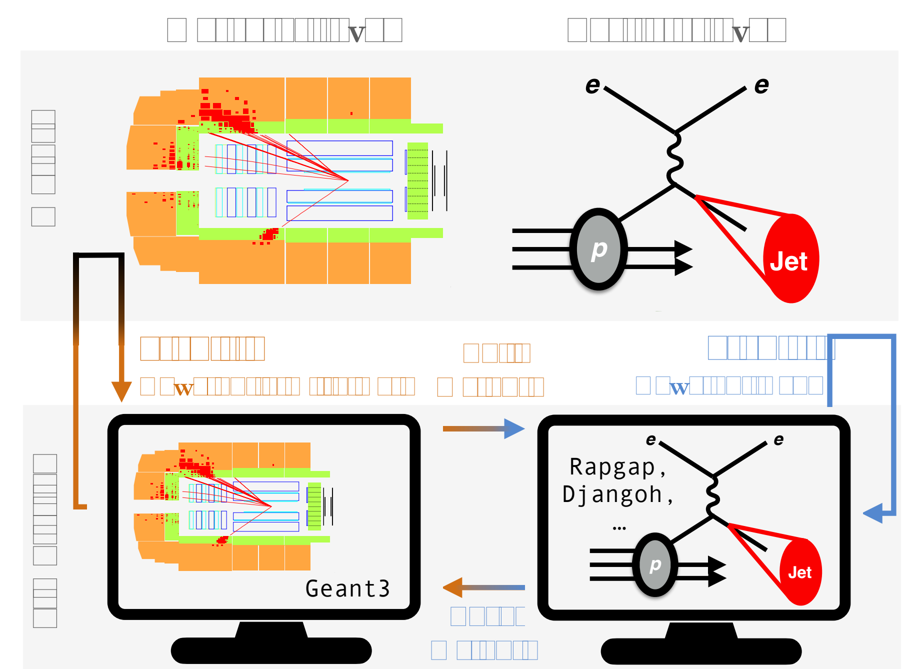
1. Obtain the azimuthal asymmetry angle, ϕ, in each event

2. Obtain unfolding event weight from MultiFold Step 2, ω_i, for each event, i



$$\frac{\sum_i \omega_i \cos(n\phi_i)}{\sum_i \omega_i} \text{ for } n = 1,\ 2,\ 3$$

Done in bins of $\vec{q}_\perp$ GeV/c

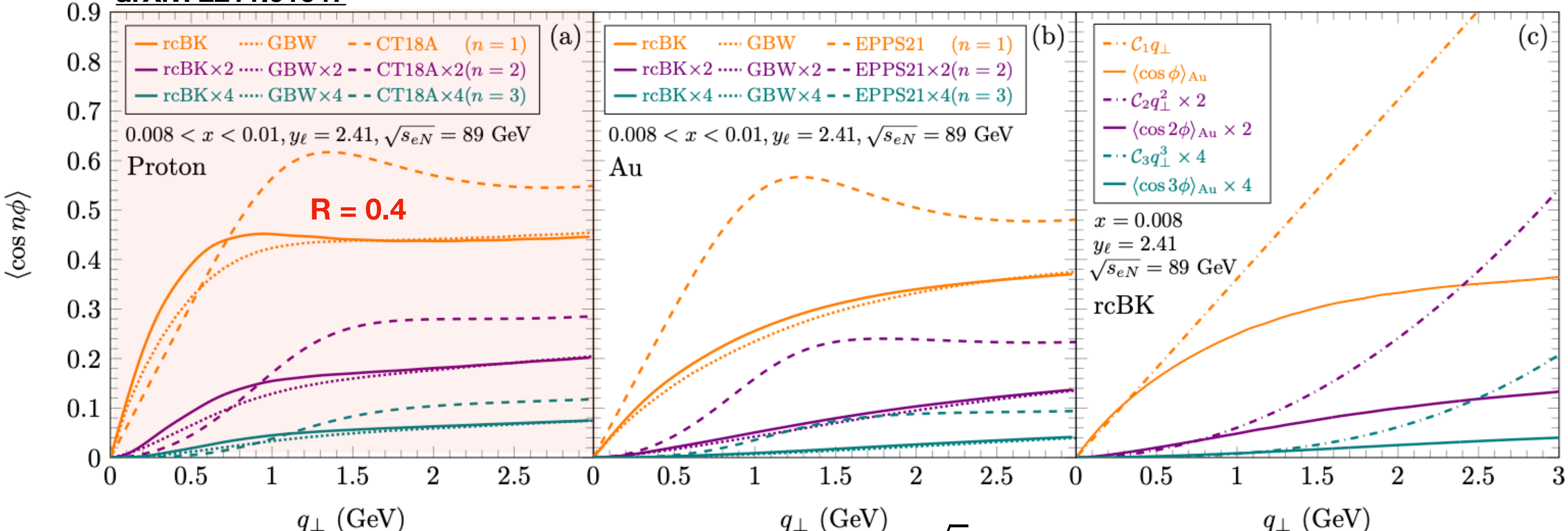


Multifold already used to unfold:

$p_x^e,\ p_y^e,\ p_z^e,\ p_\mathrm{T}^\mathrm{jet},\ \eta^\mathrm{jet},\ \phi^\mathrm{jet},\ \Delta\phi^\mathrm{jet},\ q_\mathrm{T}^\mathrm{jet}/Q$

EIC Calculation @ HERA kinematics

<u>**arXiv: 2211.01647**</u>



$$\vec{q}_\perp = \vec{k}_{\ell\perp} + \vec{k}_{J\perp}$$

$$\overrightarrow{P_\perp} = (\vec{k}_{\ell\perp} - \vec{k}_{J\perp})\ /\ 2$$

$\sqrt{s}$ **=140 GeV,** $P_\perp$ **= 20 GeV,**

y_l **= 1.5,** Q **= 25 GeV**

Radiative corrections

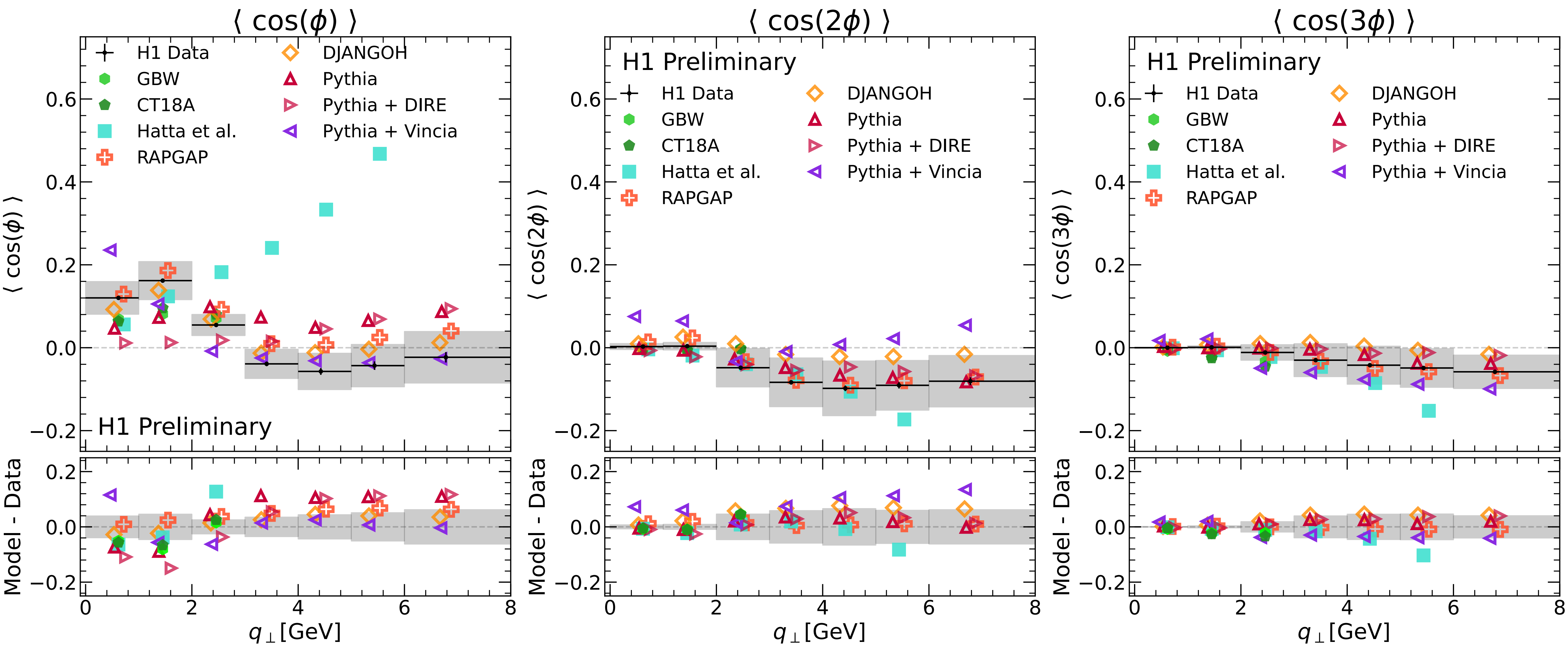
enhanced $\propto (\alpha_s \ln^2 P_\perp^2/q_\perp^2)^n$

Plots above are for R = 0.4. Calculation done for this measurement w/ R = 1.0,

Very good example of observable from 'legacy' dataset influencing future colliders

Harmonics of saturation with the inputs <u>GBW</u> model and a TMD calculation CT18A PDF

Moments of Asymmetry Results

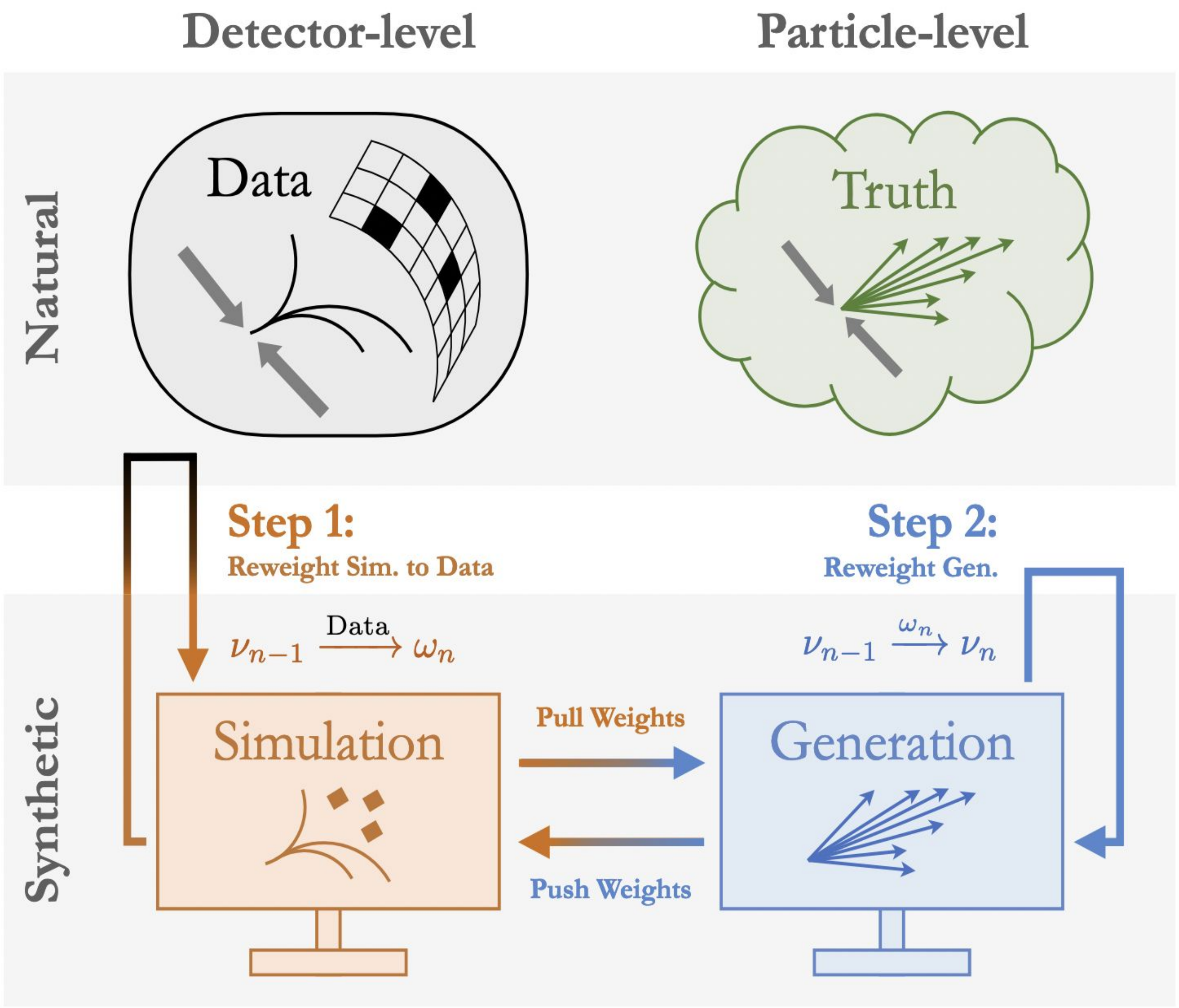


- **Three harmonics of the azimuthal angular asymmetry between the lepton and leading jet as a function of $q_\perp$.**
- **Predictions from multiple simulations as well as a pQCD calculation are shown for comparison.**

Taking OmniFold one step ***Further***

- Neural networks are well suited for handling high dimensional inputs
- We no longer *bin* for unfolding, but still use the same typical physics objects as inputs
  - Ex: Scattered lepton and Jet properties
- Why not expand what we use as inputs for the unfolding?

Quick OmniFold Recap



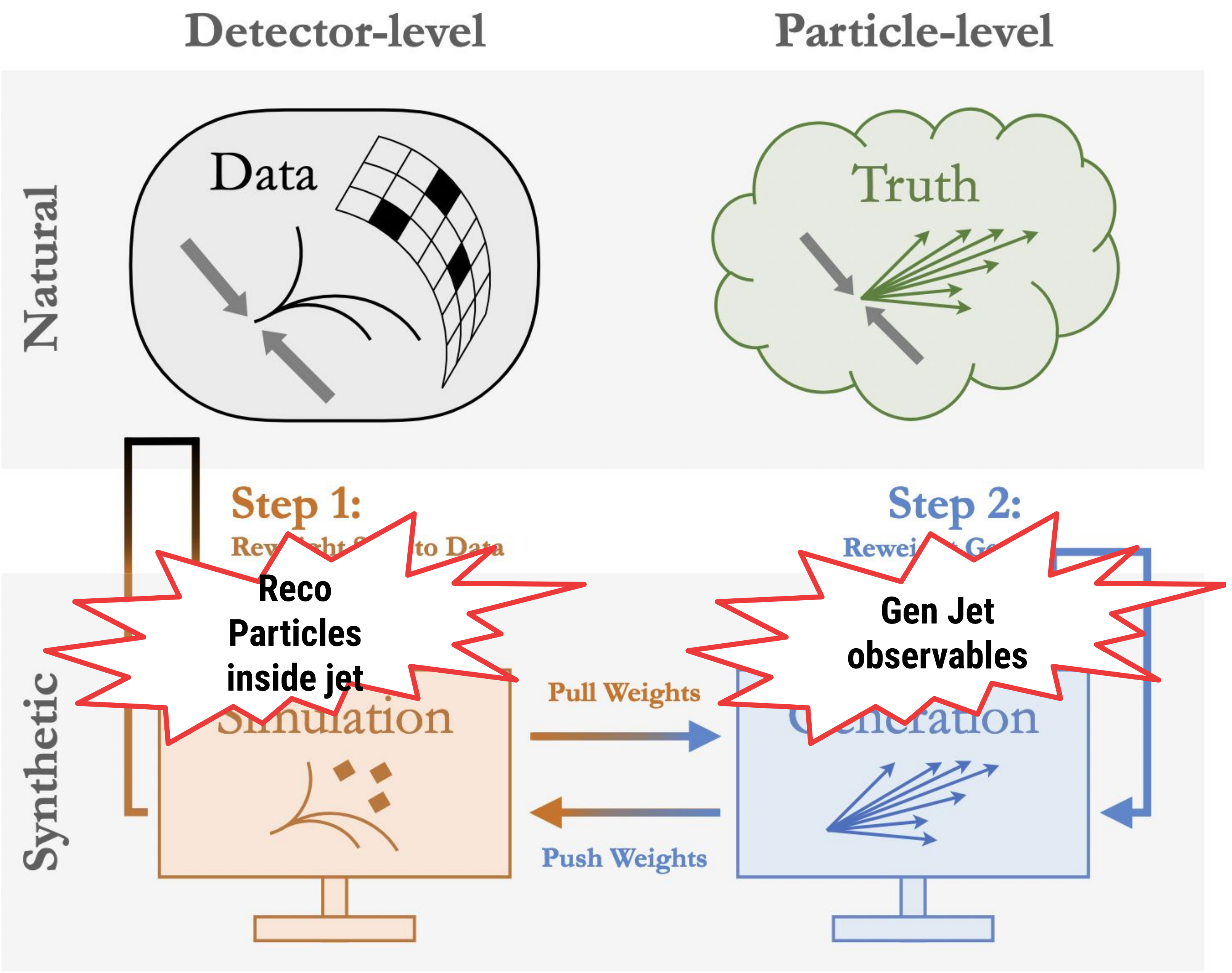
2 step iterative approach

1. Events from detector level sim. are reweighted to match the data
2. Create a "new simulation" by transforming weights to a proper function of the generated events

Classifiers used to approximate **2** likelihood functions:

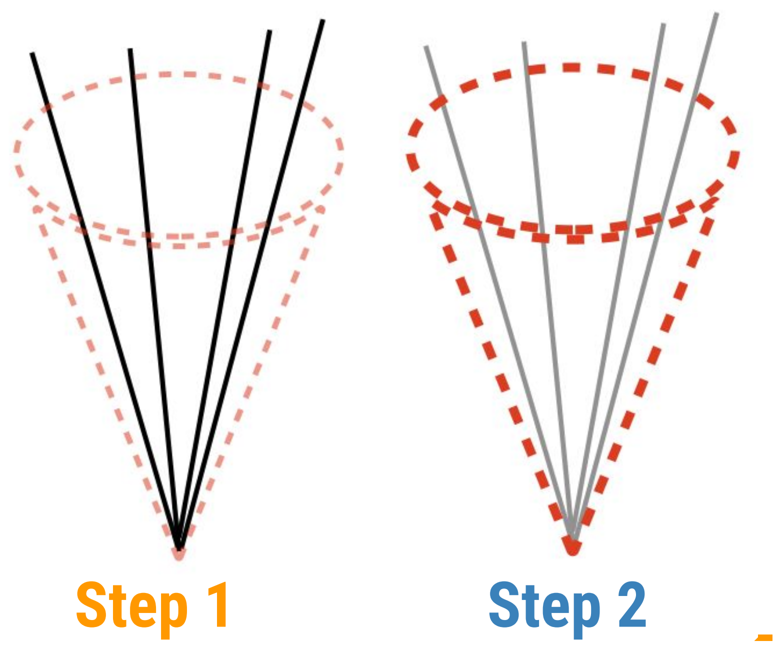
1. reco MC to Data reweighting
2. **Previous** and **new Gen** reweighting

Different OmniFold input



Different input levels for each step

1. Particles are as inputs
2. Set of gen obs. Planned to unfold

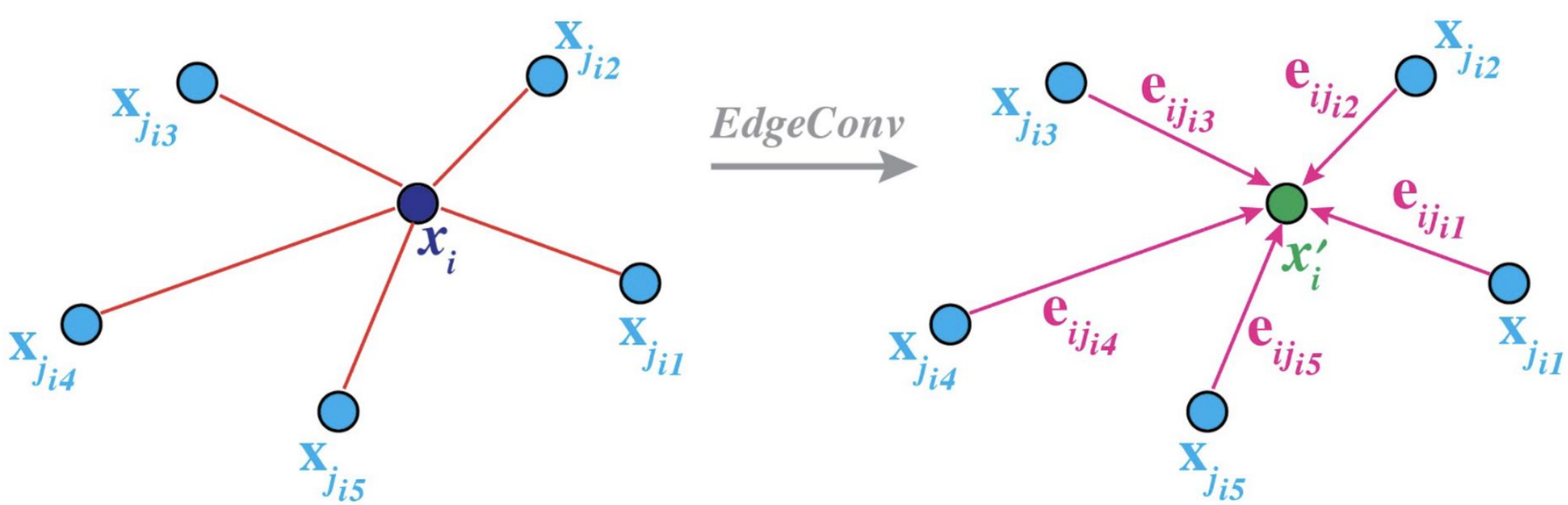


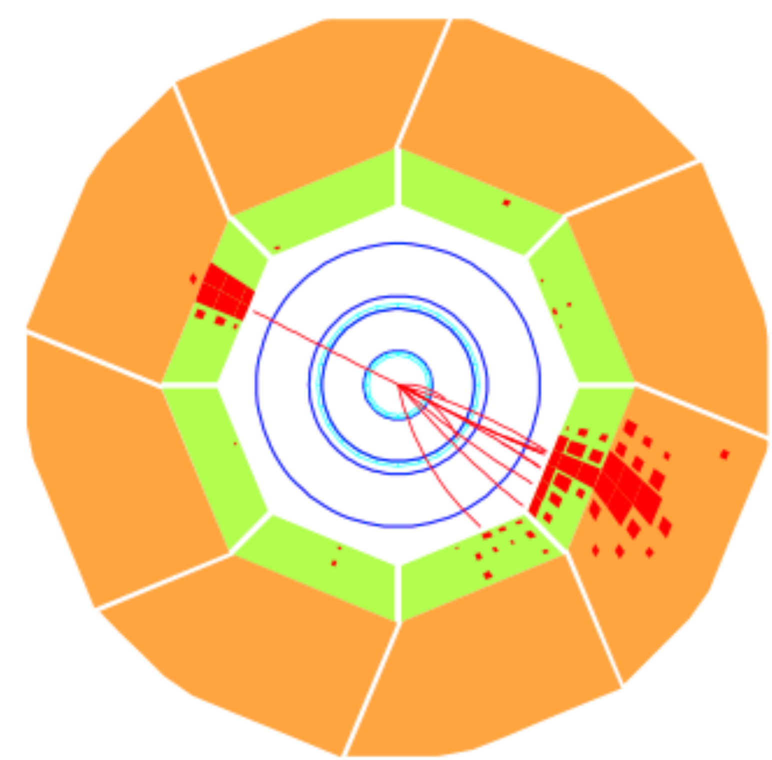
Reminder: The output of each step is an event weight, $w(\overrightarrow{x})$

Point Cloud Input

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

$$e(k) + q(p_1) \rightarrow e'(k_\ell) + jet(k_J) + X$$





Summary: The model is given much richer data in step 1, accounting for all possible covariates of the detector response

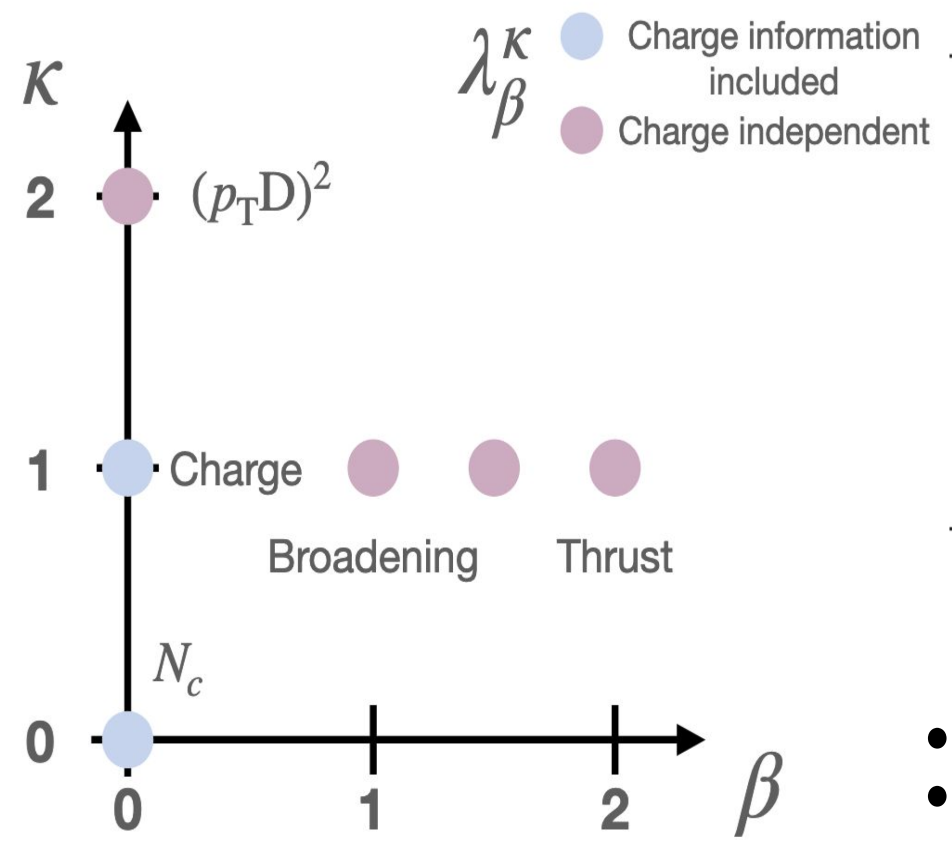
Credit: Vinicius Mikuni

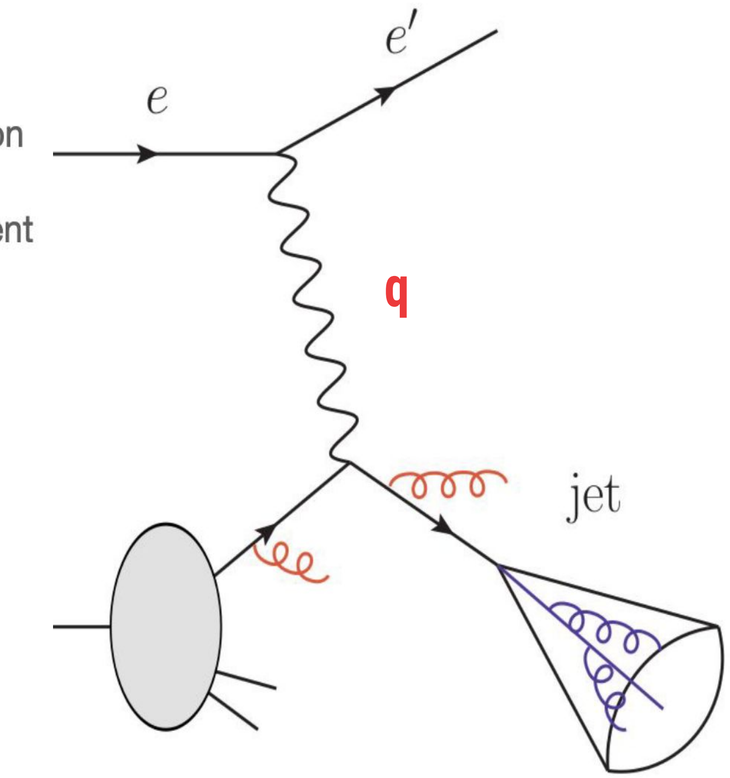
* V. Mikuni and F. Canelli 2021 *Mach. Learn.: Sci. Technol.* **2** 035027

Simultaneously Measuring 6 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 $\mathbf{p_T D}$ angularities
- Charge dependent observables: $\mathbf{Q_j}$ and $\mathbf{N_c}$
- Study the evolution of the observables with energy scale $\mathbf{Q^2 = -q^2}$





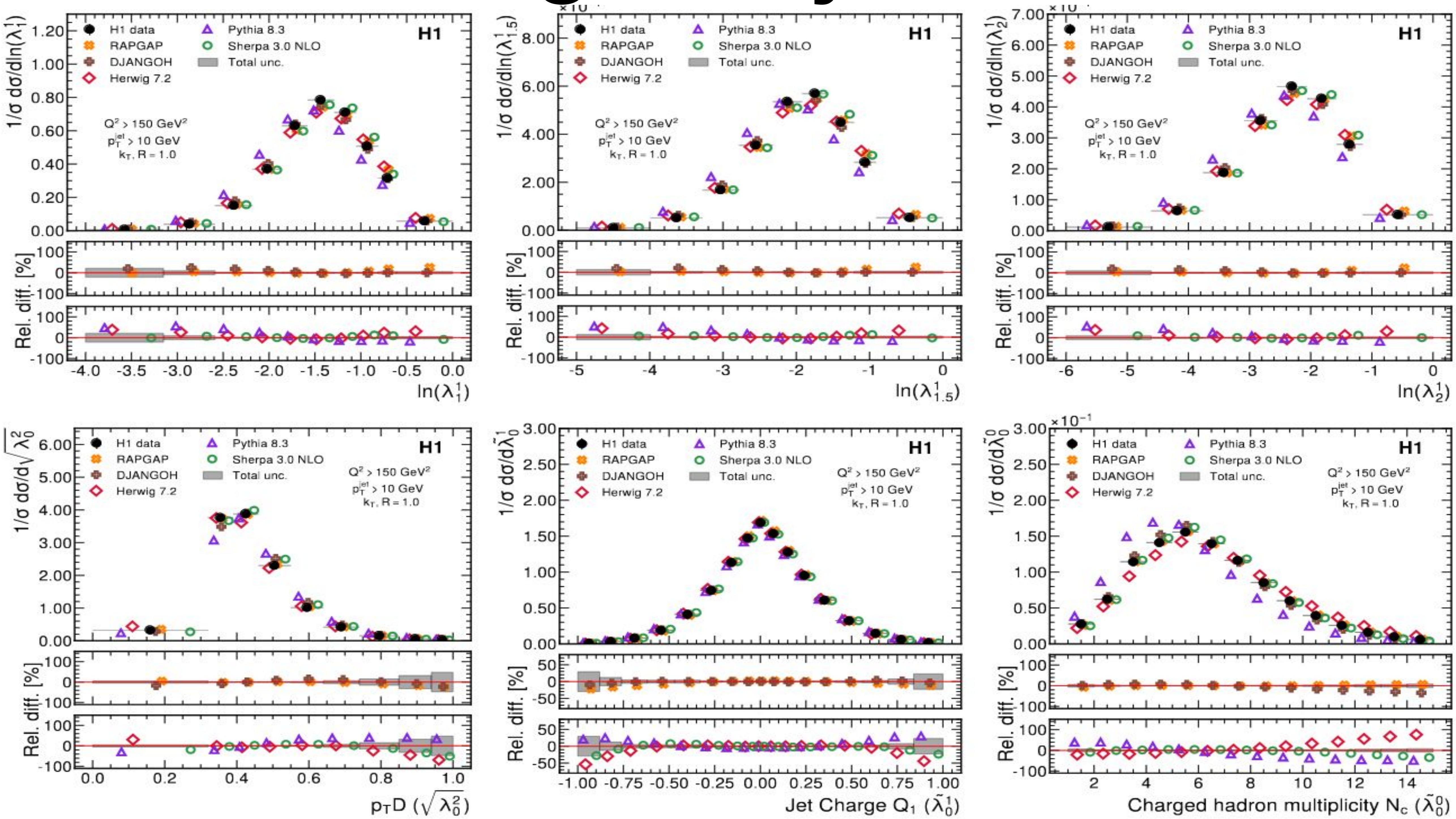
- z_i: longitudinal momentum fraction
- q_i: charge
- R_i distance from jet axis in (eta,phi)

$$\lambda^{\kappa}_{\beta} = \sum_{i \in \text{jet}} z_i^{\kappa} \left(\frac{R_i}{R_0} \right)^{\beta}$$

$$\tilde{\lambda}^{\kappa}_{0} = Q_{\kappa} = \sum_{i \in \text{jet}} q_i \times z_i^{\kappa}.$$

Credit: Vinicius Mikuni

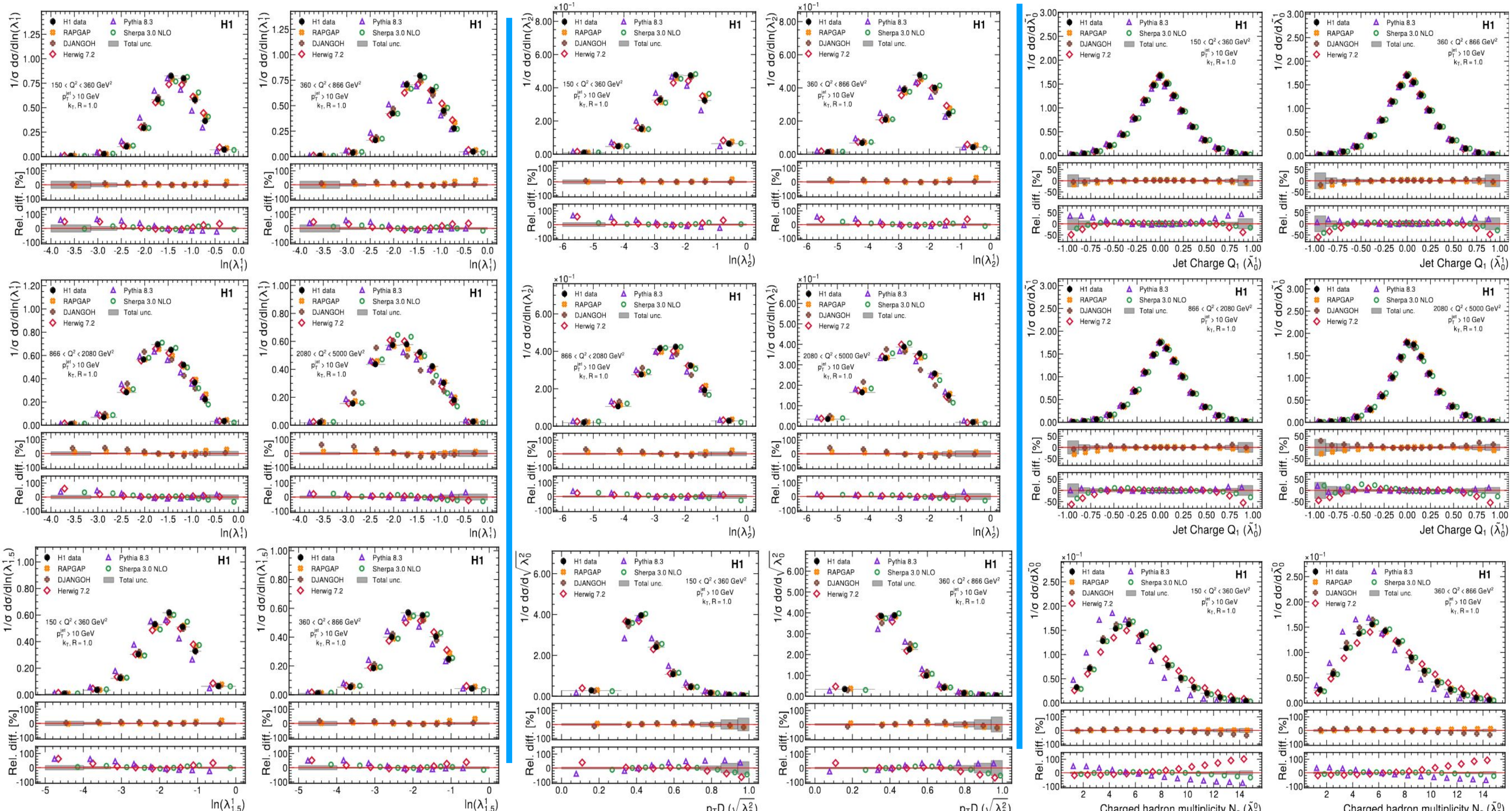
Jet Angularity Results



- **Dedicated DIS generators do good job (Rapgap and Djangoh)**
- **Herwig, Pythia, and Sherpa do a decent job**

Credit: Vinicius Mikuni

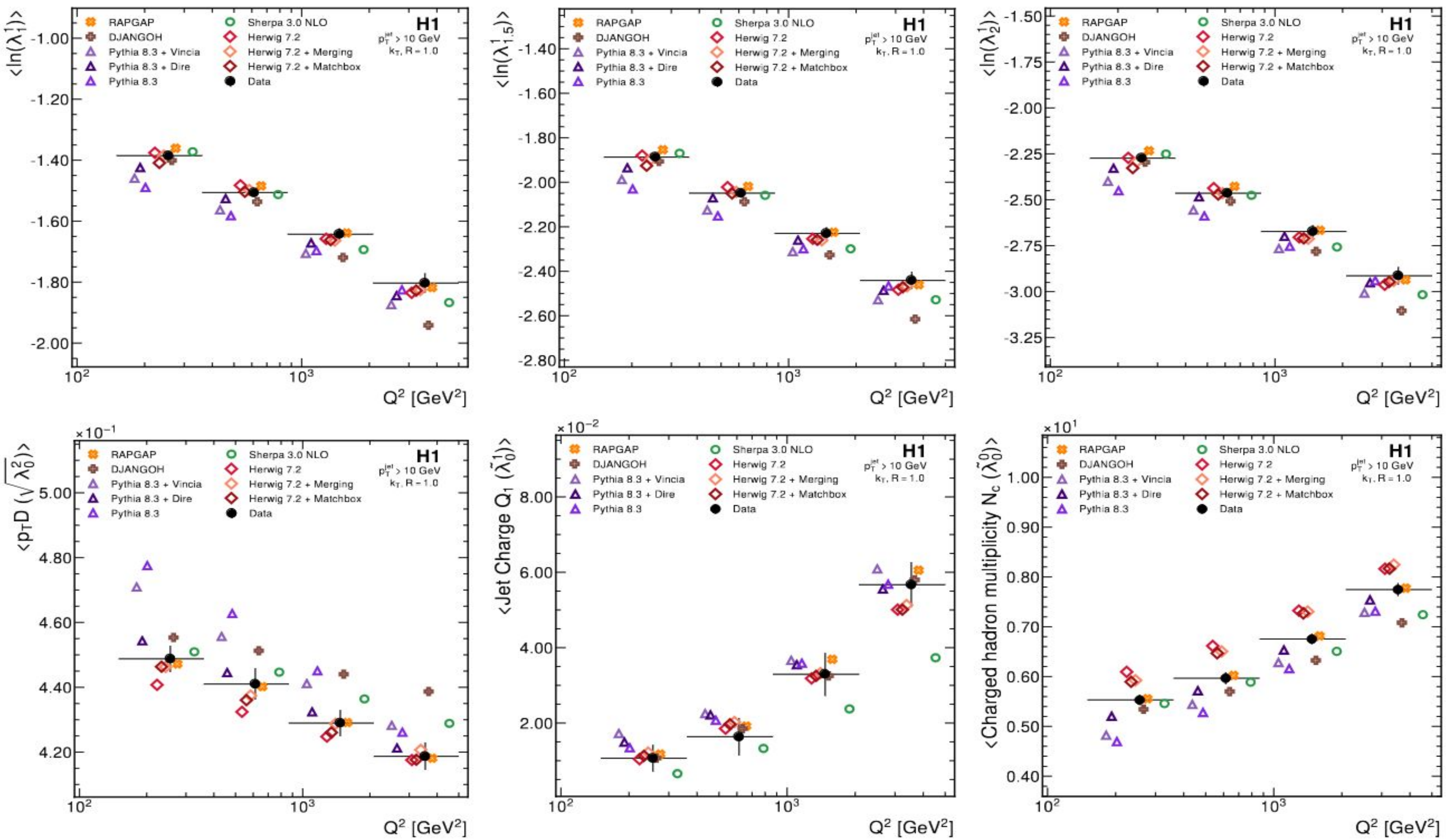
Multi-Differential Results



- Q^2 **distribution simultaneously unfolded, showing the energy scale dependence**

Credit: Vinicius Mikuni

Mean Value, for free



- **More quark-like at higher energies: mean charge increases**
- **Better agreement at higher Q^2**

Credit: Vinicius Mikuni

Conclusions

- First Multidimensional un-binned unfolding using OmniFold and real data
- Promising measurement to probe soft gluon radiation, with importance for EIC
- Simultaneous unfolding for Jet Substructure
- MultiFold
  - This work presents a measurement of *moments*, requiring the *un-binned unfolding!*
  - Re-usability (cross sections + asymmetry measurement)
  - LHC measurement! https://arxiv.org/pdf/2405.20041
- H1 is a great example of exciting measurements using legacy datasets

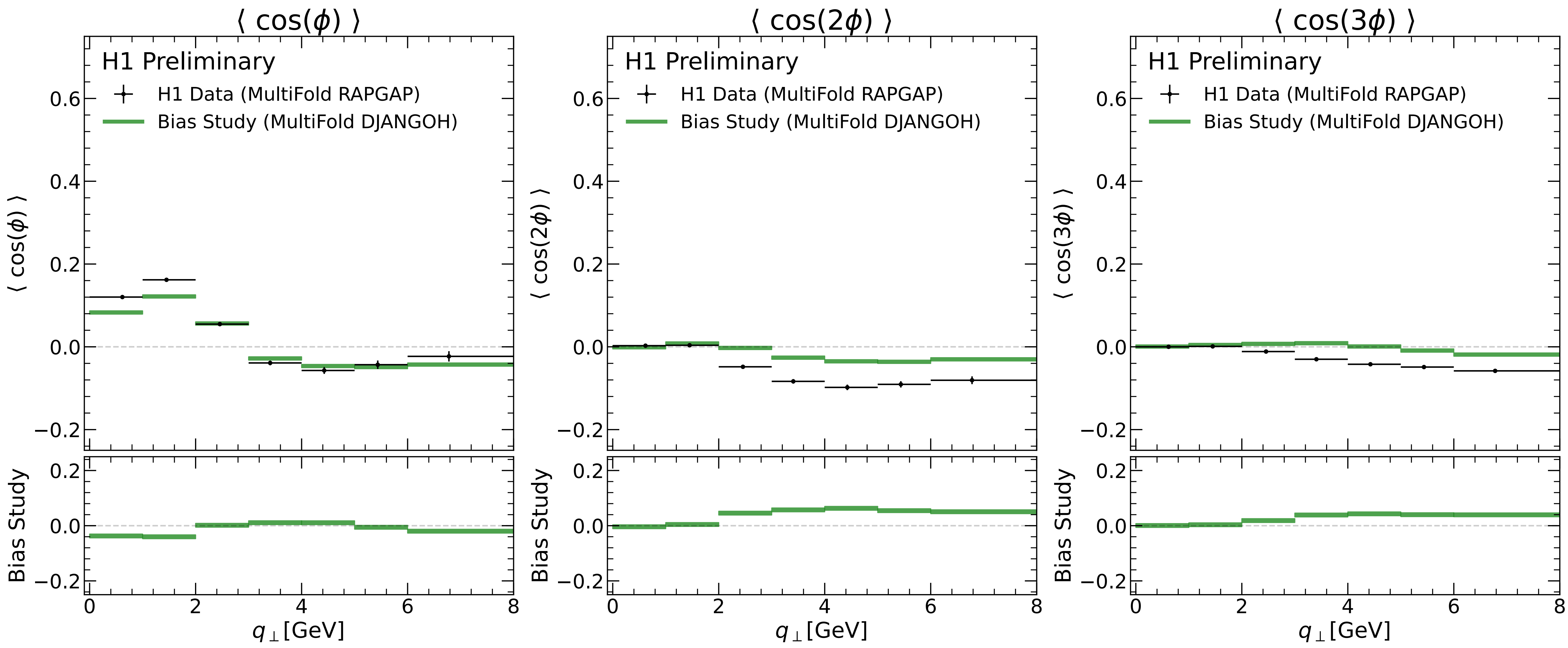
PhysRevLett.128.132002

https://doi.org/10.1016/j.physletb.2023.138101

END

Backup

Investigation of Model Bias vs. $q_\perp$ [GeV]



- Leading uncertainty is model bias in the unfolding for $\cos(2\phi)$ and $\cos(3\phi)$
- Difference in the result when unfolding using RAPGAP and DJANGO
- Reporting Abs. Errors; <u>central values are very close to 0.0</u>
- The Total Uncertainty is quite stable between harmonics

Systematic Uncertainties

- Model Dependance:
  - The bias of the unfolding procedure is determined by taking the difference in the result when unfolding using RAPGAP and DJANGO
  - The two generators have different underlying physics, thus providing a realistic evaluation of the procedure bias

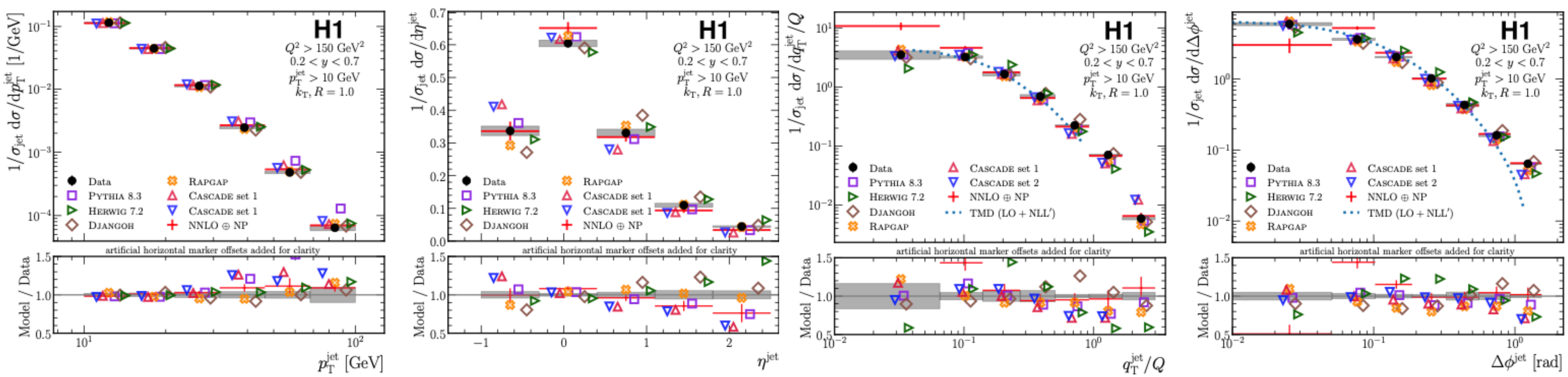
- QED Radiation Corrections
  - Difference of correction between RAPGAP and DJANGO
  - Take RAPGAP with and without QED corrections
  - Take DJANGO with and without QED corrections

- Systematic uncertainties are determined by varying an aspect of the simulation and repeating the unfolding
  - These values detail the magnitude of variation:
  - HFS-object energy scale: $\pm 1\ \%$
  - HFS-object azimuthal angle: ± 20 mrad
  - Scattered lepton azimuthal: ± 1 mrad
  - Scattered lepton energy: $\pm 0.5 - 1.0\ \%$

Further Background

- Machine learning (OmniFold) is used to perform an 8-dimensional, unbinned unfolding. Present four, binned results:

- Use the 8-dimensional result to explore the Q^2 dependence and any other observables that can be computed from the electron-jet kinematics



Extracted from the same phase-space as Yao's analysis, but reporting a different observable

OmniFold

1. $\omega_n(m) = \nu_{n-1}^{\text{push}}(m) L[(1,\text{Data}), (\nu_{n-1}^{\text{push}}, \text{Sim.})](m)$

$$\omega_n^{\text{pull}}(t) = \omega_n(m)$$

- Detector level simulation is weighted to match the data
- $L[(1,\text{Data}), (\nu_{n-1}^{\text{push}}, \text{Sim.})](m)$ approximated by classifier trained to distinguish the *Data* and *Sim*.

2. $\nu_n(t) = \nu_0(t) L[(\omega_n^{\text{pull}}, \text{Gen.}), (\nu_0, \text{Gen.})](t)$

- Transform weights to a proper function of the generated events to create a new simulation
- $L[(\omega_n^{\text{pull}}, \text{Gen.}), (\nu_{n-1}, \text{Gen.})](t)$ approximated by classifier trained to distinguish Gen. with *pulled* weights from Gen. using $\text{weights}_{\text{old}}$ / $\text{weights}_{\text{new}}$

Each iteration of step 2 learns the correction from the original ν_0 weights

Advantage: Easier implementation, no need to store previous ν_n model

Disadvantage: Learning correction from ν_0 is more computationally expensive

IBU Generalization

$$t_j^{(n)} = \sum_i \Pr_{n-1}(\text{truth is } j | \text{measure } i) \Pr(\text{measure } i)$$

$$= \sum_i \frac{R_{ij} t_j^{(n-1)}}{\sum_k R_{ik} t_k^{(n-1)}} \times m_i,$$

$$L[(w, X), (w', X')](x) = \frac{p_{(w,X)}(x)}{p_{(w',X')}(x)},$$

Differential Cross Section

- Back-to-back electron-jet production from *ep* collision,

$$e(l) + p(P) \to e(l') + J_q(p_J) + X$$

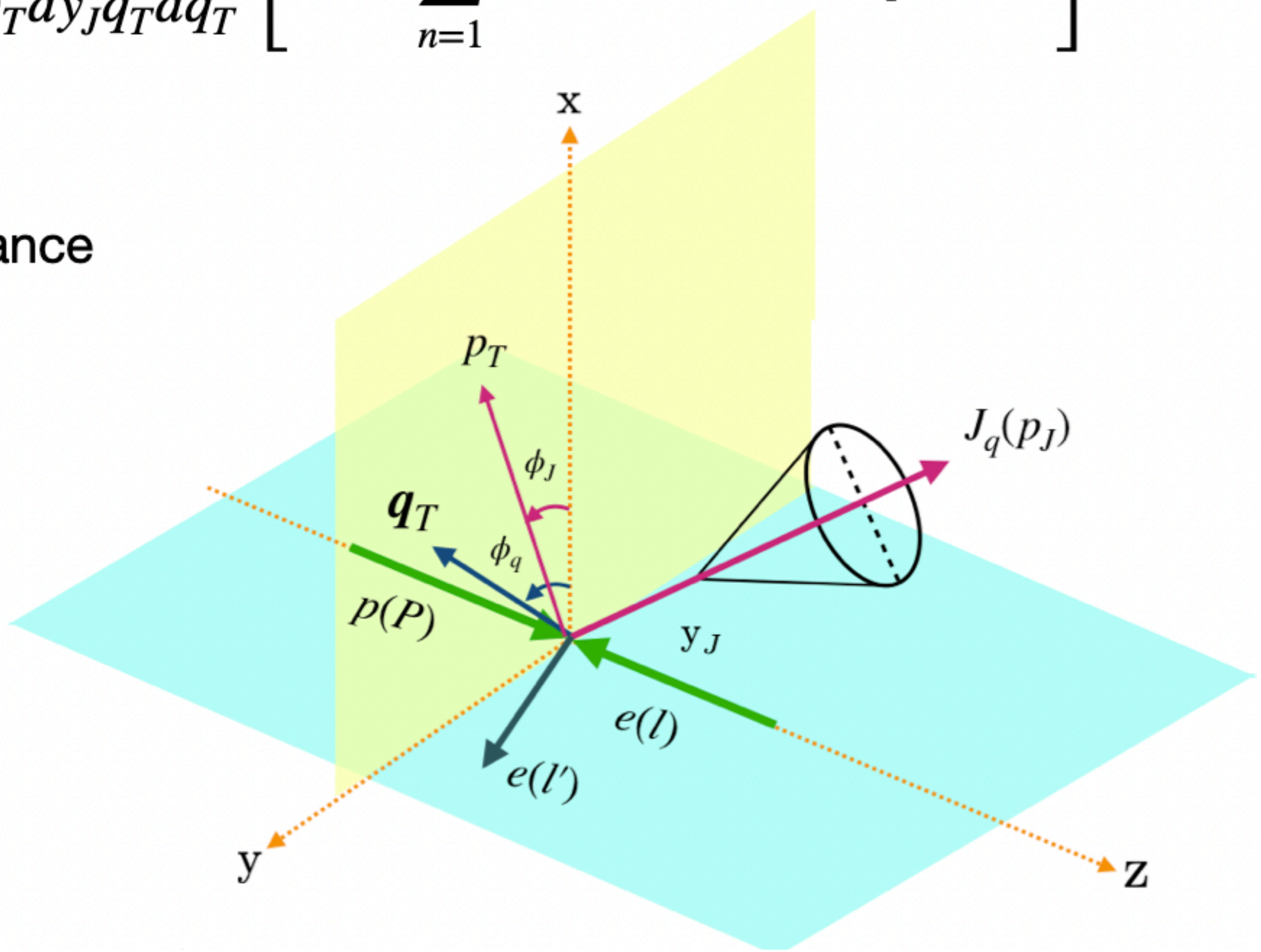
$$\frac{d\sigma}{d^2\boldsymbol{p}_T dy_J d\phi_J d^2\boldsymbol{q}_T} = \frac{d\sigma}{2\pi d^2\boldsymbol{p}_T dy_J q_T dq_T}\left[1 + 2\sum_{n=1}^{\infty} v_n(p_T, y_T)\cos(n(\phi_q - \phi_J))\right]$$

q_T : transverse momentum imbalance

$$\boldsymbol{q}_T = \boldsymbol{l}'_T + \boldsymbol{p}_{JT}$$

p_T : jet transverse momentum

y_J : jet rapidity



Note: slightly different angle definition, but background still applies]

Credit: Fanyi Zhao