University of Zurich[UZH]

Machine-learning (ML) techniques for hadronic reconstruction and calibration, and machine learning in analyses with jets

Weijie Jin

Machine-learning-based unfolding analysis

Measurement of Event Shapes in Minimum Bias Events at $\sqrt{s}$ = 13 TeV (CMS)

NEW

CMS-PAS-SMP-23-008

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

NEW

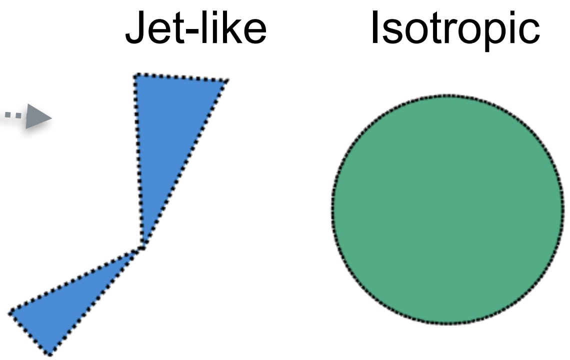
arxiv:2405.20041

Machine-learning-based unfolding measurement of event shapes

Event shape observables:

Variables describing the “**shapes** ” of the events

→ Functions of the momentum of the final state particles

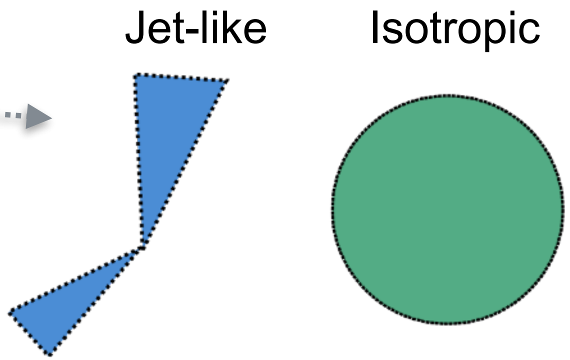


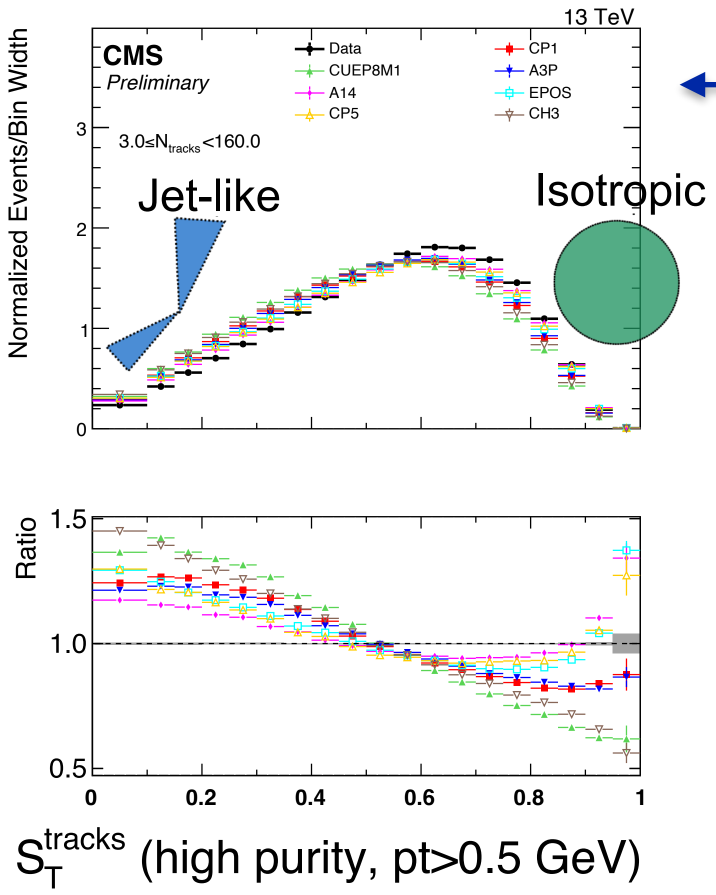
Machine-learning-based unfolding measurement of event shapes

Event shape observables:

Variables describing the "**shapes** " of the events

→ Functions of the momentum of the final state particles

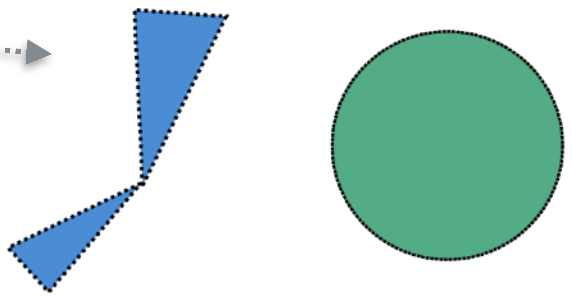




An example: transverse sphericity

others: (transverse) thrust, broadening, isotropy etc.

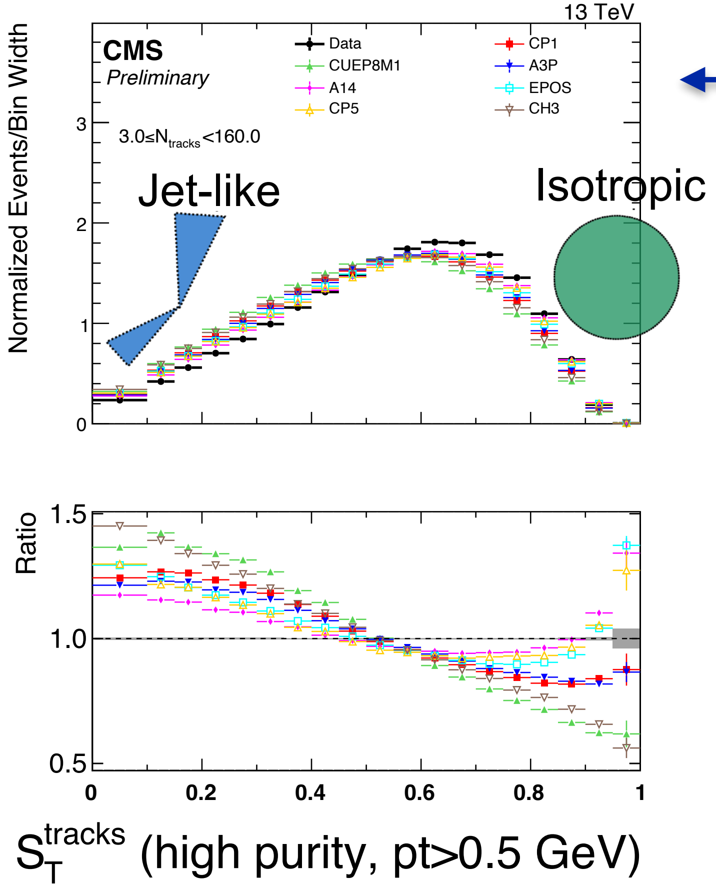
Machine-learning-based unfolding measurement of event shapes



Event shape observables:

Variables describing the “**shapes** ” of the events

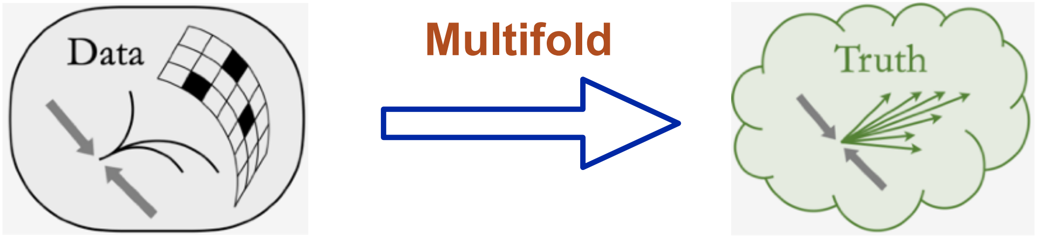
→ Functions of the momentum of the final state particles



An example: transverse sphericity

others: (transverse) thrust, broadening, isotropy etc.

Unfold with a **machine-learning-based** algorithm: **Multifold***



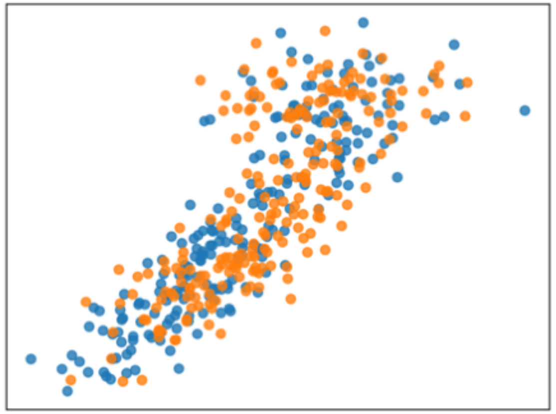
Event shapes of detector-level objects

Event shapes of particles

→ theoretical interpretation, generator tuning …

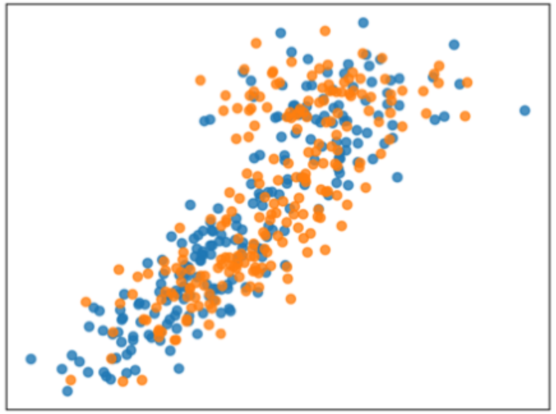
* https://arxiv.org/abs/1911.09107, https://arxiv.org/abs/2105.04448

Unbinned multi-dimensional unfolding and uncertainty estimation



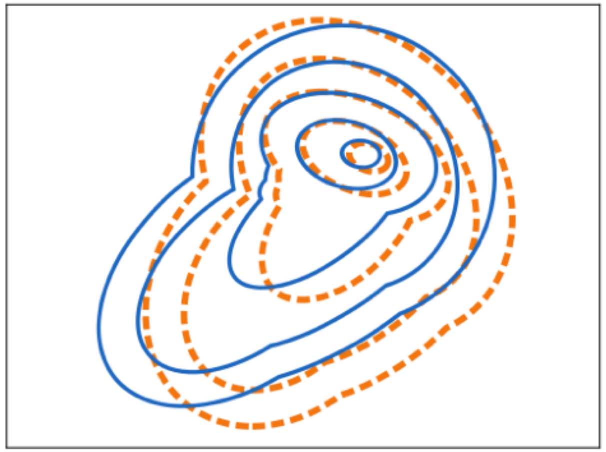
← A typical binary classifier to distinguish two sets

Unbinned multi-dimensional unfolding and uncertainty estimation

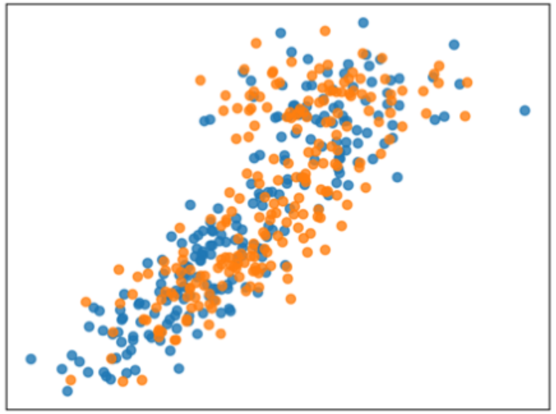


← A typical binary classifier to distinguish two sets

What it actually did: learn the differences in the distributions →

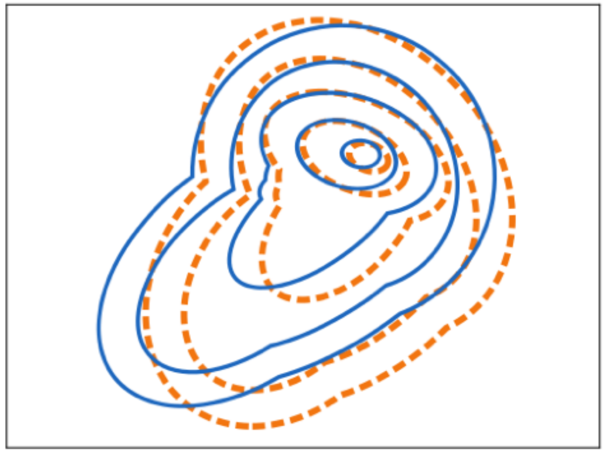


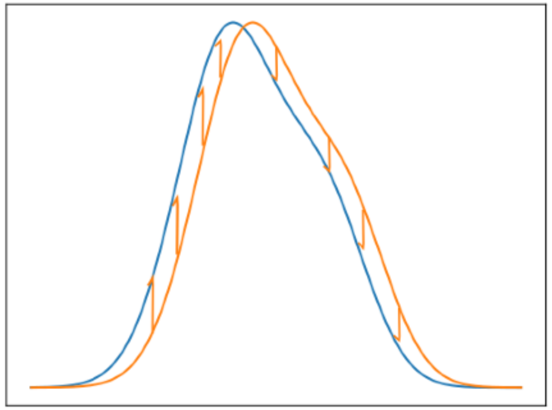
Unbinned multi-dimensional unfolding and uncertainty estimation



← A typical binary classifier to distinguish two sets

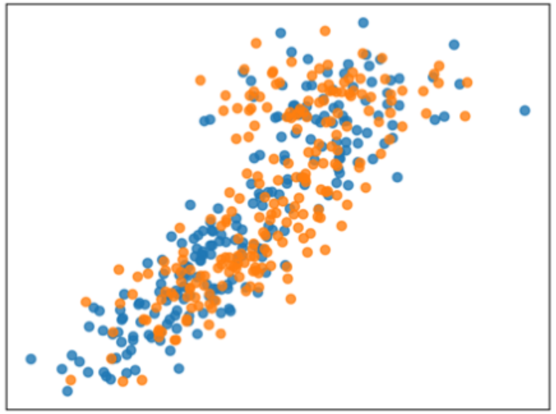
What it actually did: learn the differences in the distributions →





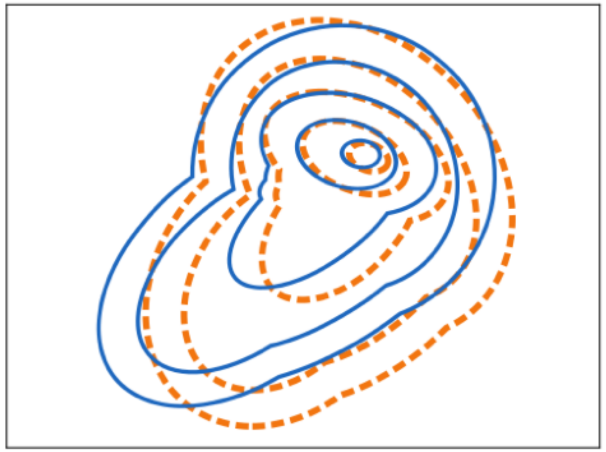
← We can use the classification scores to weight **MC** to **data**, and **nominal sample** to **systematic variations**

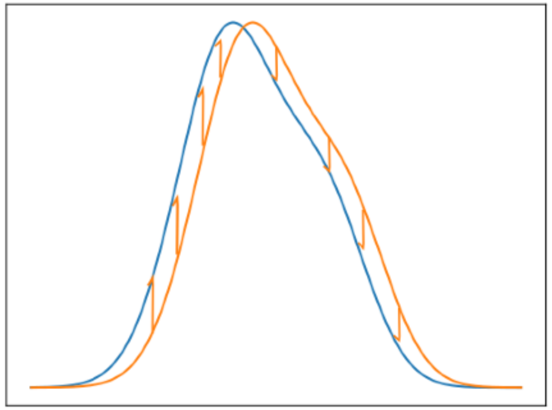
Unbinned multi-dimensional unfolding and uncertainty estimation



← A typical binary classifier to distinguish two sets

What it actually did: learn the differences in the distributions →



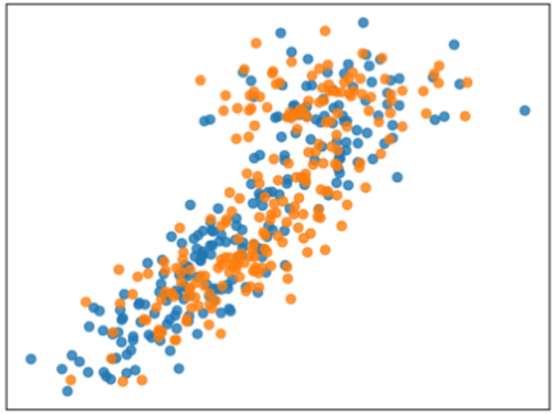


← We can use the classification scores to weight **MC** to **data**, and **nominal sample** to **systematic variations**

Event-wise unfolding → the result independent of binning

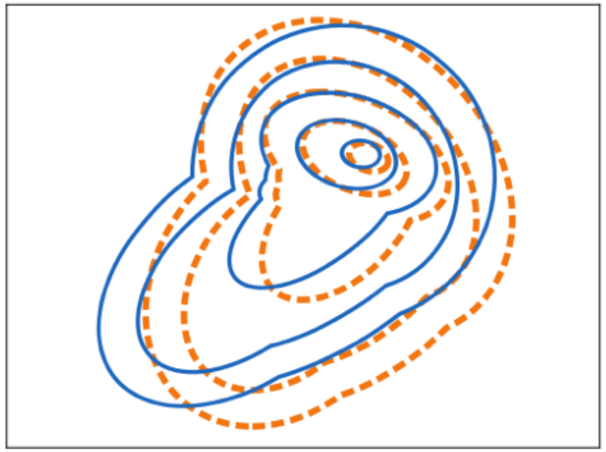
The actual unfolding in iterations:
- Step 1: weight **MC** to **data**, at detector level
- Step 2: pull back the weights to particle(truth) level

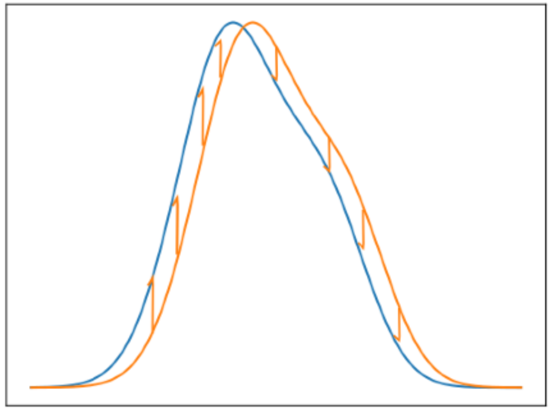
Unbinned multi-dimensional unfolding and uncertainty estimation



← A typical binary classifier to distinguish two sets

What it actually did: learn the differences in the distributions →





← We can use the classification scores to weight **MC** to **data**, and **nominal sample** to **systematic variations**

Event-wise unfolding → the result independent of binning

The actual unfolding in iterations:

- Step 1: weight **MC** to **data**, at detector level
- Step 2: pull back the weights to particle(truth) level

Event-wise uncertainty template → unbinned unfolding uncertainty & covariance

Unfolding results

Simultaneously unfold all the variables for ML-based weighting

Add a variable to the unfolding:

Methods based on **binned** histograms:

Add **another dimension** in binning

→ require **higher statistics**

→ more **computation** in simulation and unfolding

This method:

Add **a feature** in the ML training and evaluation

→ much easier to scale up the dimensions

Unfolding results

Simultaneously unfold all the variables for ML-based weighting

Add a variable to the unfolding:

Methods based on **binned** histograms:

Add **another dimension** in binning

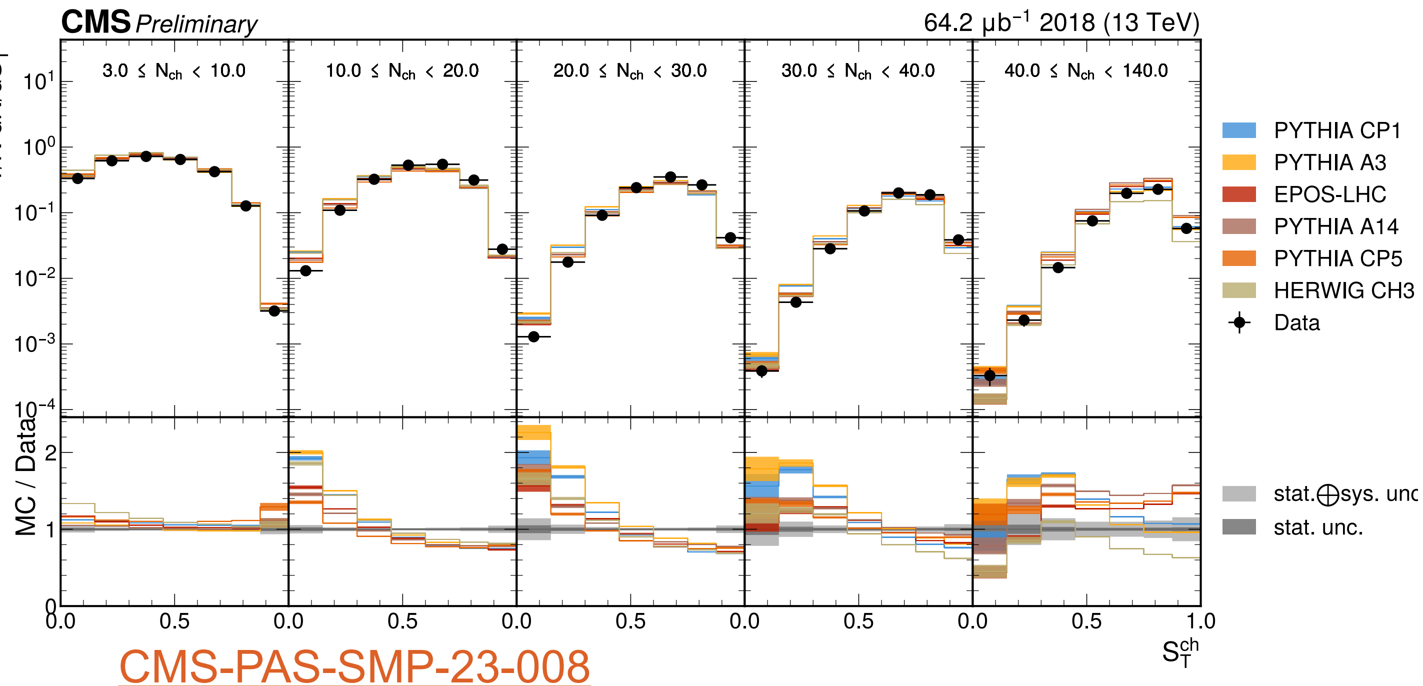
→ require **higher statistics**

→ more **computation** in simulation and unfolding

This method:

Add **a feature** in the ML training and evaluation

→ much easier to scale up the dimensions



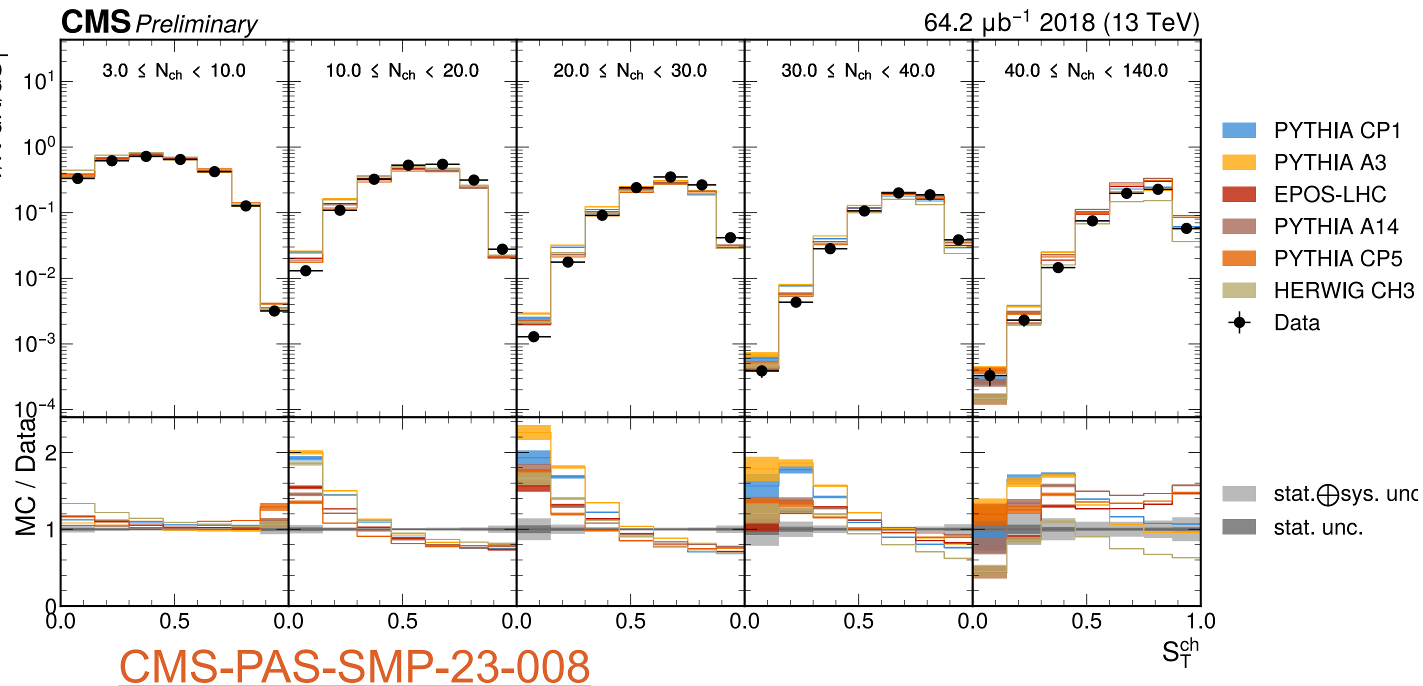
CMS-PAS-SMP-23-008

Unfolding results as **weighted MC events**

← 2D visualisation of transverse sphericity in charged particle multiplicity slices

Customise binning and **variable choices** are supported with the **event-wise unfolded data**

Unfolding results



CMS-PAS-SMP-23-008

More isotropic data than MC:
multi-parton-interaction model? collective effects? instantons?
→ We provide the unfolded results for theoretical interpretation

Simultaneously unfold all the variables for ML-based weighting

Add a variable to the unfolding:

Methods based on **binned** histograms:
- Add **another dimension** in binning
- → require **higher statistics**
- → more **computation** in simulation and unfolding

This method:
- Add **a feature** in the ML training and evaluation
- → much easier to scale up the dimensions

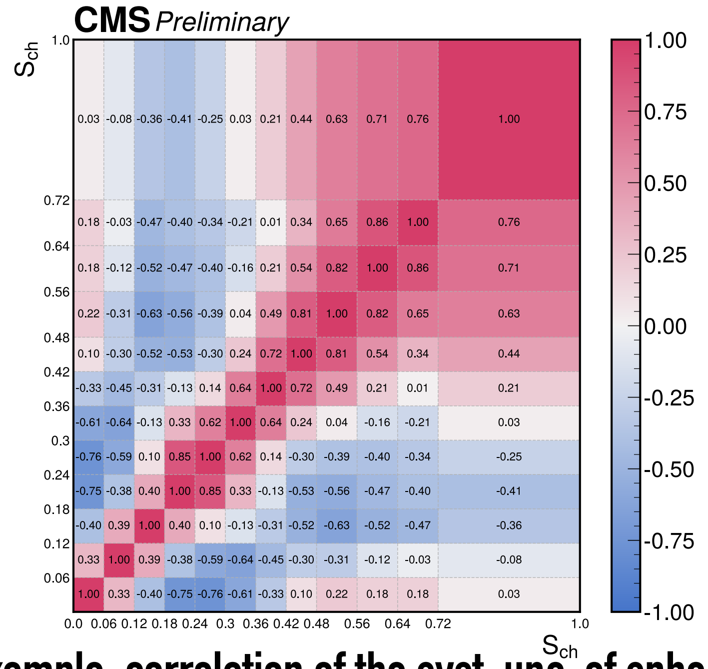
Unfolding results as **weighted MC events**
← 2D visualisation of transverse sphericity in charged particle multiplicity slices

Customise binning and **variable choices** are supported with the **event-wise unfolded data**

Unbinned uncertainty estimation

ML-based reweighting → **Uncertainty templates** as sets of **weights on nominal MC**
→ **Continuous nuisance** parameters can be assigned to the **event-weights**
→ Uncertainty **covariance** can be estimated from **toy experiments**

- Unfold with **“bootstraps” of MC** with **variations of nuisance parameters** → **Syst. Unc + Covariance**
- Unfold with “**bootstraps” of resampled data** → **Stat. Unc. + Covariance**



Example: correlation of the syst. unc. of sphericity

CMS-PAS-SMP-23-008

Customise binning and **variable choices** are supported with the **event-wise unfolded data**

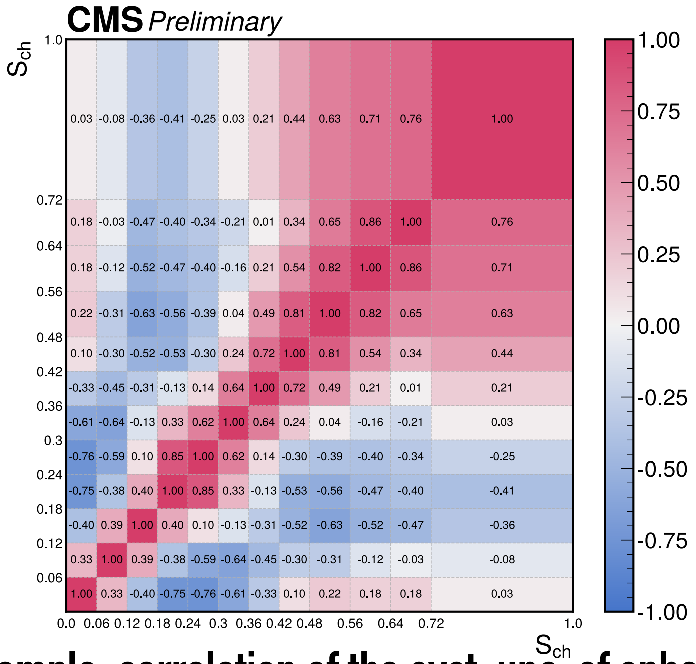
+

Uncertainties+Covariance on the results

Unbinned uncertainty estimation

ML-based reweighting → **Uncertainty templates** as sets of **weights on nominal MC**
→ **Continuous nuisance** parameters can be assigned to the **event-weights**
→ Uncertainty **covariance** can be estimated from **toy experiments**

- Unfold with **"bootstraps" of MC** with **variations of nuisance parameters → Syst. Unc + Covariance**
- Unfold with "**bootstraps" of resampled data → Stat. Unc. + Covariance**



CMS-PAS-SMP-23-008

Example: correlation of the syst. unc. of sphericity

Customise binning and **variable choices** are supported with the **event-wise unfolded data**

+

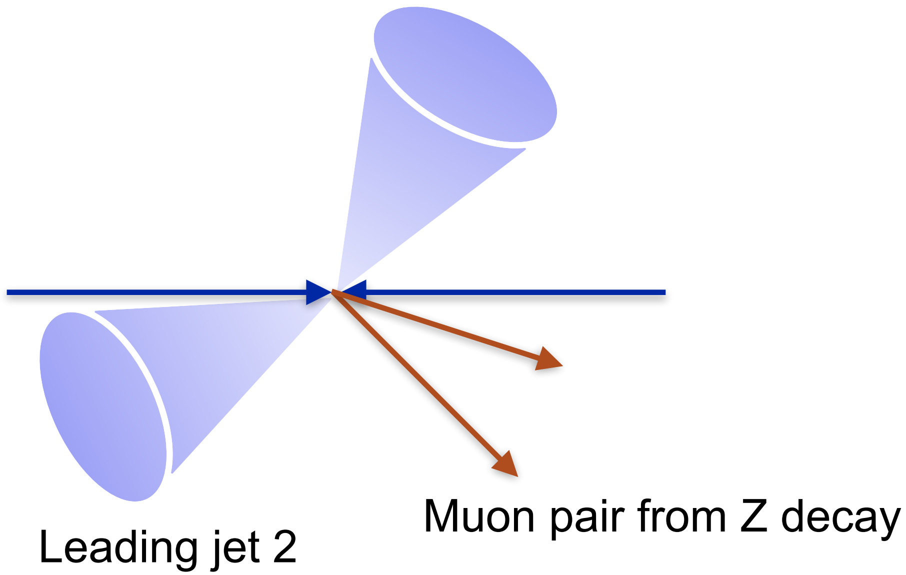
Uncertainties+Covariance on the results

The way to improve the usability of **unfolded results**
- Publish the **unbinned** results on **event-level**
- Publish the **weight sets** from **toy experiments**
  → Unc. + Covariance

⇨

Unbinned fit for **theoretical interpretation**
Unbinned generator tuning
(Or **any binning** chosen by the user)

Machine-learning-based unfolding of Z+jet kinematic observables



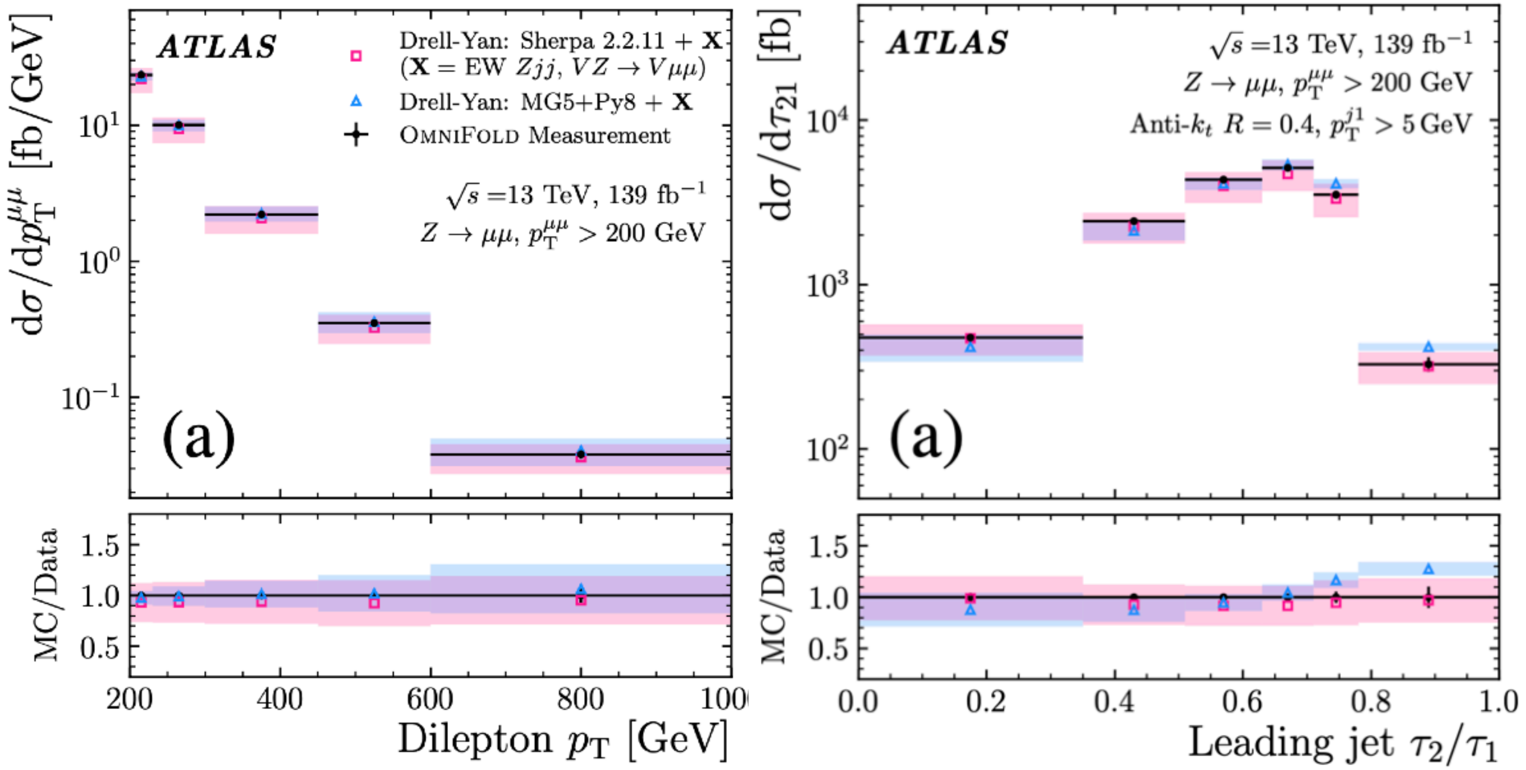
Observables to be measured

- Kinematics of the **di-muon system from Z decay**
  $p_T^{\mu\mu}, y_{\mu\mu}$
  → probe **Z boson production** kinematics
- Kinematics of the **two muons**
  $p_T^{\mu 1}, p_T^{\mu 2}, \eta_{\mu 1}, \eta_{\mu 2}, \phi_{\mu 1}, \phi_{\mu 2}$
  → probe **Z boson decay** kinematics
- Kinematics of **two leading charged particle jets**
  $p_T^{j1}, p_T^{j2}, y_{j1}, y_{j2}, \phi_{j1}, \phi_{j2}$
- **Substructure** of the two leading charged particle jets
  mass: (m_{j1}, m_{j2}), charged particle multiplicity: $(n_{ch}^{j1}, n_{ch}^{j2})$,
  N-subjettiness: $\tau_1^{j1}, \tau_1^{j2}, \tau_2^{j1}, \tau_2^{j2}, \tau_3^{j1}, \tau_3^{j2}$

Also **unfolded** with **Multifold*** → **Simultaneous** unfolding of **24 variables**

* https://arxiv.org/abs/1911.09107, https://arxiv.org/abs/2105.04448

Unfolding results

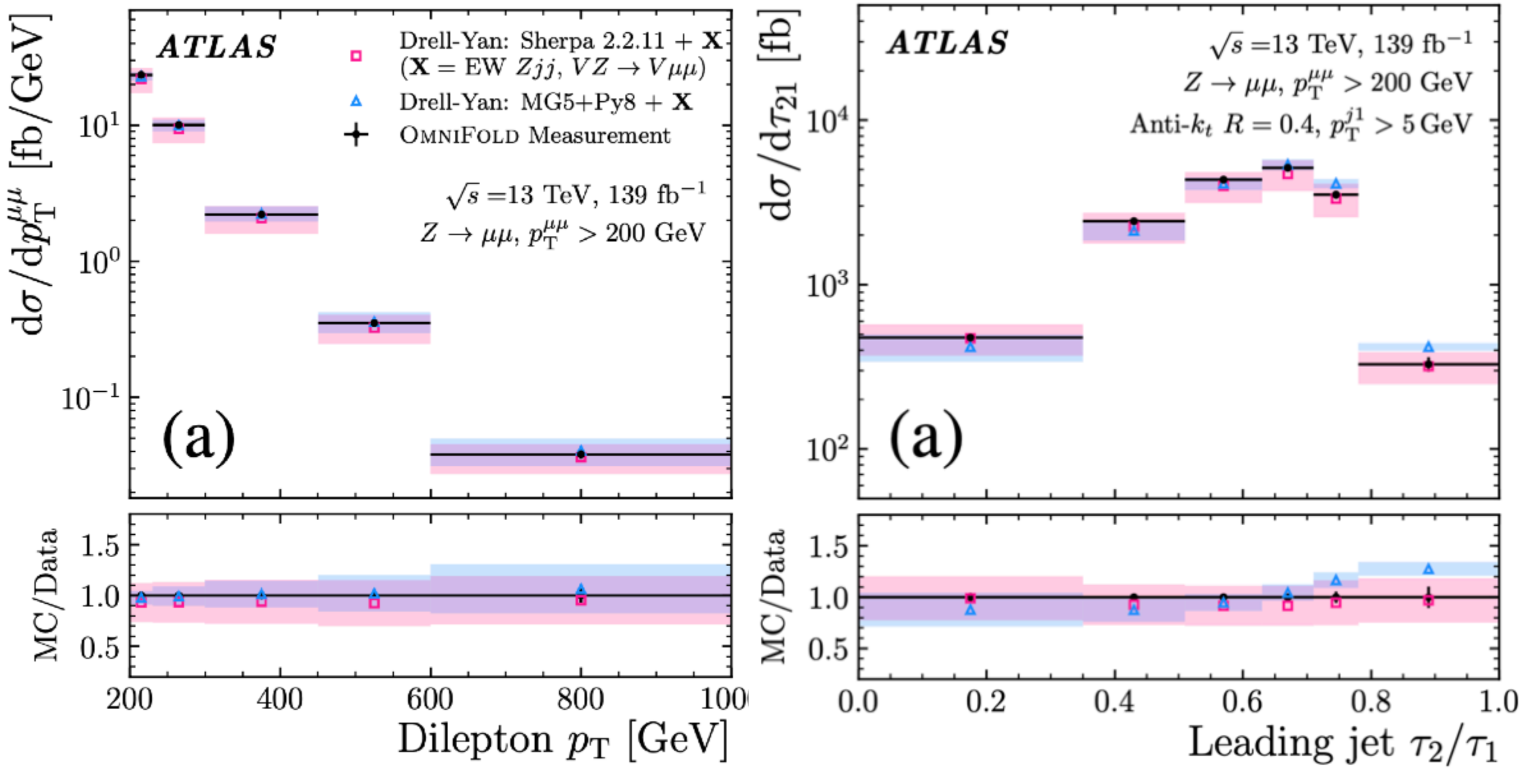


Unfolded results are **event-wise** (weighted MC events)

← 1D visualisation of **dilepton p_T** and **leading jet 2-subjettiness(τ_2) / 1-subjettiness(τ_1)**
Unfolded data versus **Sherpa** and **MadGraph+Pythia** predictions

arxiv:2405.20041

Unfolding results



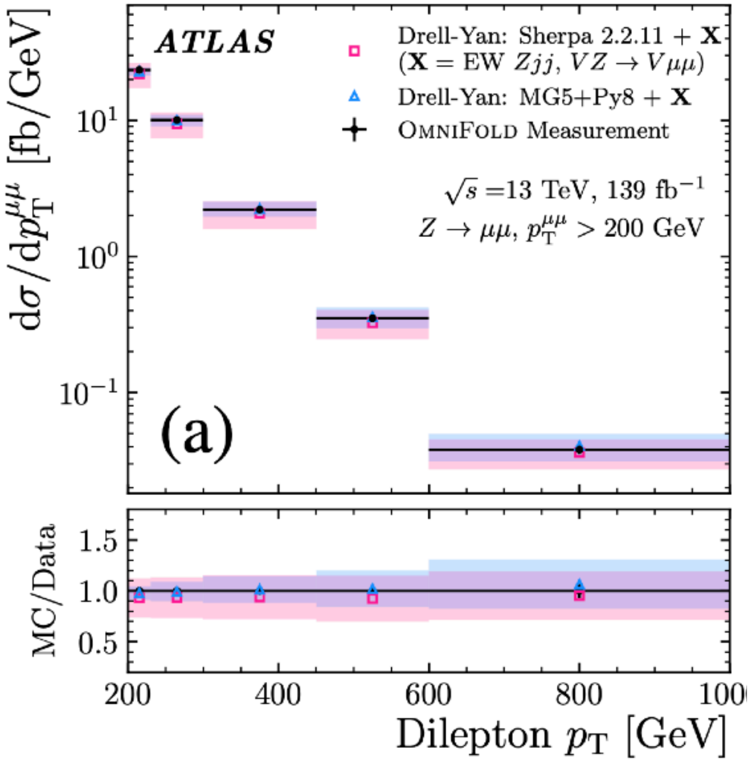
arxiv:2405.20041

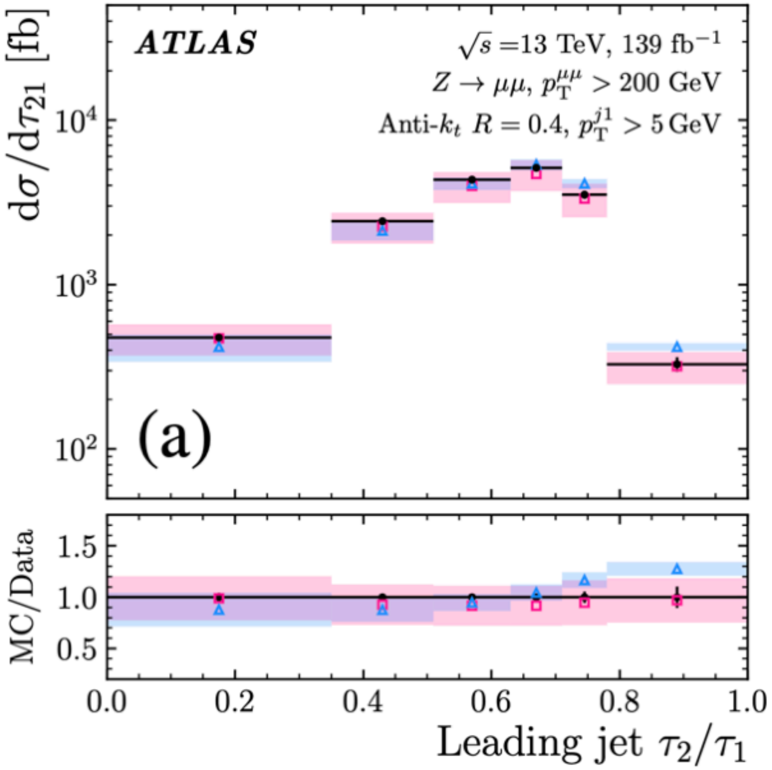
Unfolded results are **event-wise** (weighted MC events)

← 1D visualisation of **dilepton p_T** and **leading jet 2-subjettiness(τ_2) / 1-subjettiness(τ_1)**
Unfolded data versus **Sherpa** and **MadGraph+Pythia** predictions

τ_2, τ_1 are unfolded,
but τ_2/τ_1 **is not directly unfolded**
→ The unfolding **preserves the relation** among variables

Unfolding results





Unfolded results are **event-wise** (weighted MC events)

← 1D visualisation of **dilepton p_T** and **leading jet 2-subjettiness(τ_2) / 1-subjettiness(τ_1)**
Unfolded data versus **Sherpa** and **MadGraph+Pythia** predictions

τ_2, τ_1 are unfolded,
but τ_2/τ_1 **is not directly unfolded**
→ The unfolding **preserves the relation** among variables

arxiv:2405.20041

Event-level unbinned unfolding results
(weighted nominal MC)

Perturbations on the input samples according to **uncertainties**
→ **Unfold** with these **alternative samples**
→ Unfolding **uncertainty** as **alternative weights**

Unfolded results with **customised bins**
+ **uncertainties**

Machine-learning for jet calibration and tagging

Measurement of the radius dependence of charged-particle jet suppression in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 5.02 TeV (ALICE)

Phys. Lett. B 849 (2024) 138412

Performance of new jet techniques based on machine-learning for $H \to b\bar{b}$ and $H \to c\bar{c}$ searches (LHCb)

LHCB-FIGURE-2023-029

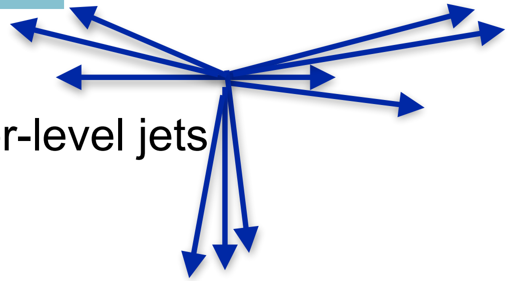
Simultaneous energy and mass calibration of large-radius jets with the ATLAS detector using a deep neural work (ATLAS)

arxiv:2311.08885

Machine-learning-based jet p_T reconstruction under large background

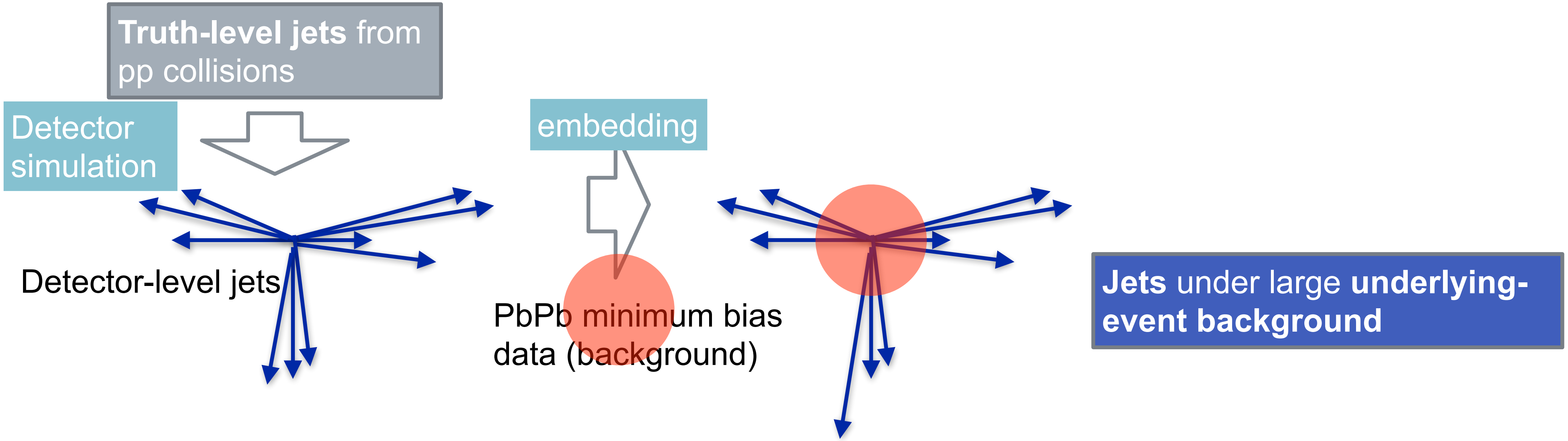
Truth-level jets from pp collisions

Detector simulation

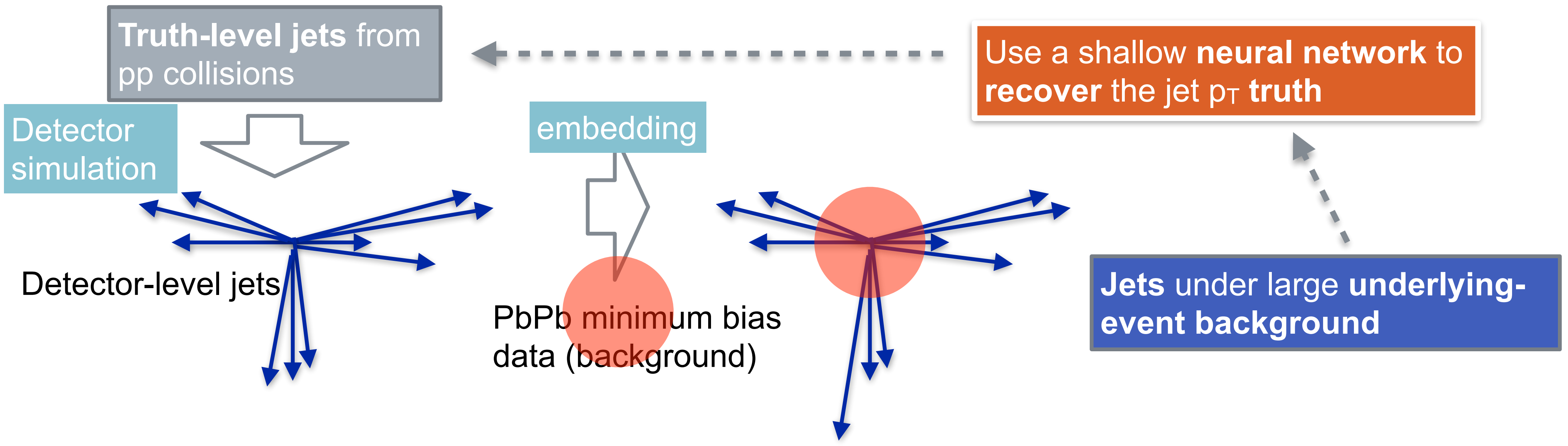


Detector-level jets

Machine-learning-based jet p_T reconstruction under large background

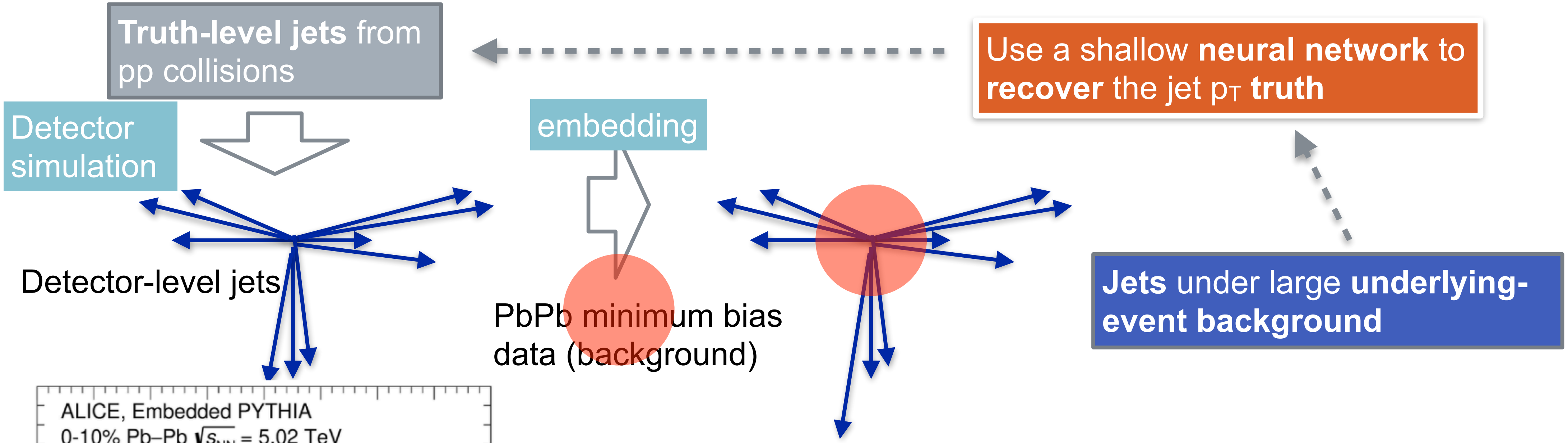


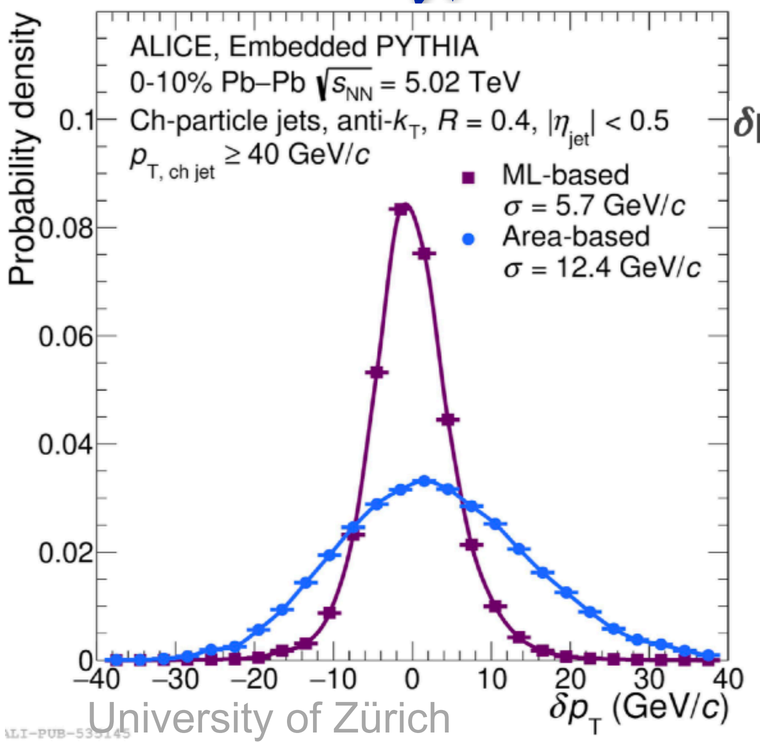
Machine-learning-based jet p_T reconstruction under large background



Training input for the NN: **jet** and **constituent** (p_T of leading tracks) properties

Machine-learning-based jet p_T reconstruction under large background





$\delta p_T = p_{T,\,rec} - p_{T,\,true}$

Training input for the NN: **jet** and **constituent** (p_T of leading tracks) properties

Large improvement of jet p_T reconstruction w.r.t **standard area-base approach**!
narrower δp_T → reduced background

Improves the measurement of **jet-quenching** in **Pb-Pb** collisions especially for jets with **large radius** and **low p_T**

Phys. Lett. B 849 (2024) 138412

Regression technique for Higgs mass reconstruction ($H \to b\bar{b}, H \to c\bar{c}$)

$H \to b\bar{b}$ and $H \to c\bar{c}$ search is based on a fit to **invariant mass**
$\to$ **sensitivity** relies on **precise dijet mass reconstruction**

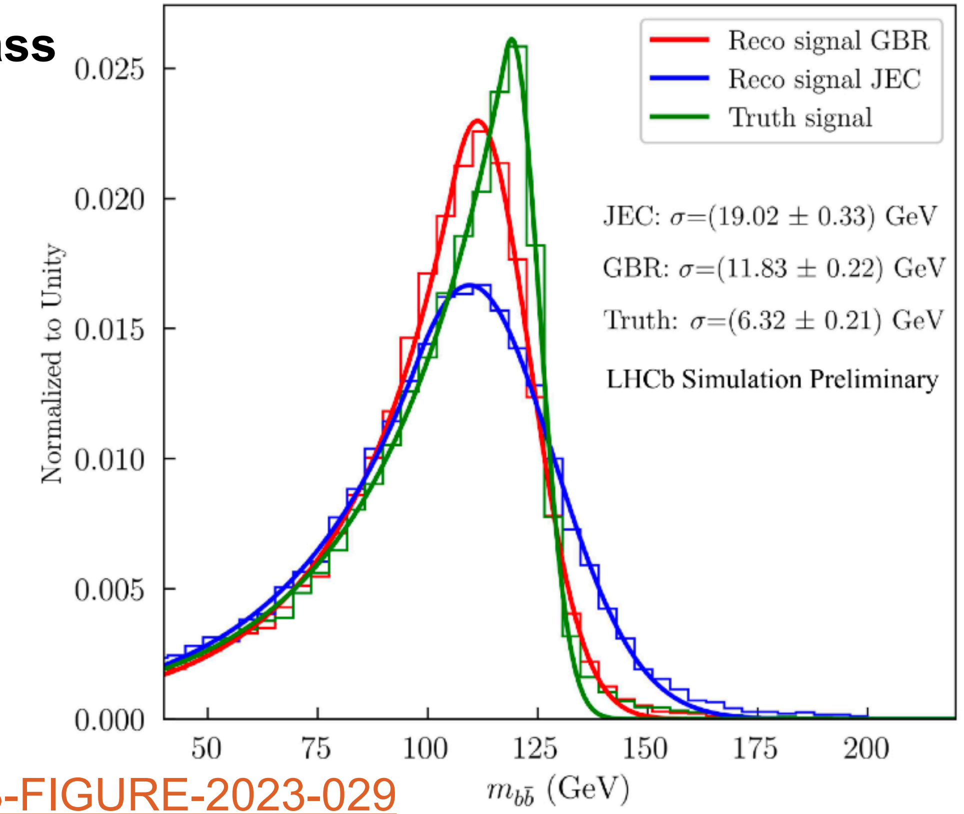
Jet kinematics and substructures

$\Downarrow$ **Input**

Gradient Boosted Regressor (GBR)

$\Downarrow$ **Fit**

Dijet invariant mass



LHCB-FIGURE-2023-029

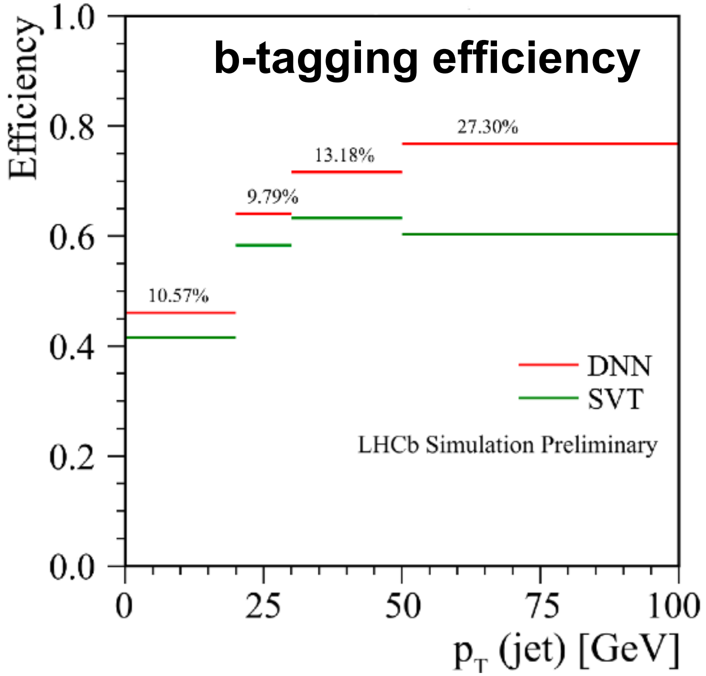
The **reconstructed mass** from **GBR** has a **narrower peak** than that from **standard Jet Energy Correction** (JEC) tools
$\to$ 50% improvement on Higgs mass reconstruction!

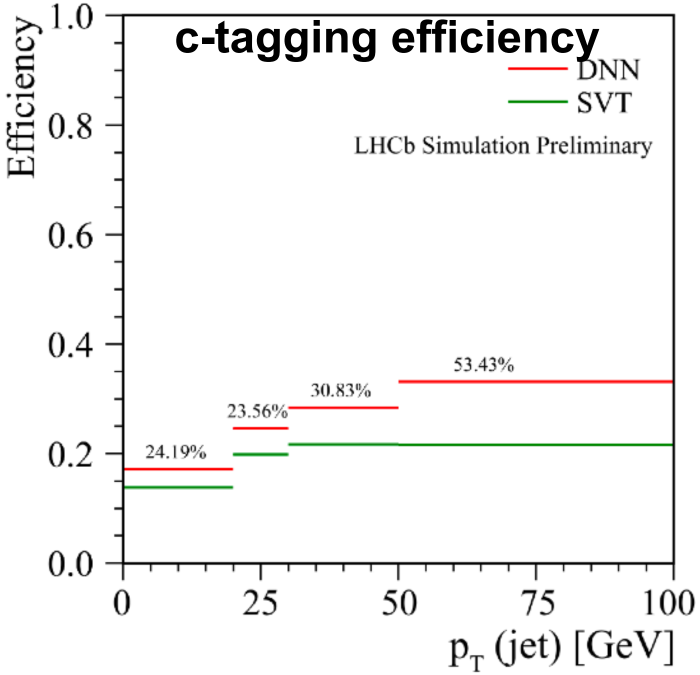
b-, c- and light-flavor- jet tagging for $H \to b\bar{b}$, $H \to c\bar{c}$

Standard **secondary-vertex-tagging (SVT)** relies heavily on secondary vertex (SV) identification
→ **limited** by the **SV reconstruction efficiency**

The **Deep Neural Network (DNN)** approach uses jet observables instead
- **Inputs**: features from **individual constituents** + jet **substructures** and **global** features
- **3 outputs**: probabilities to be **b-, c-** or **light** jets

→ includes **more information** into tagging





LHCB-FIGURE-2023-029

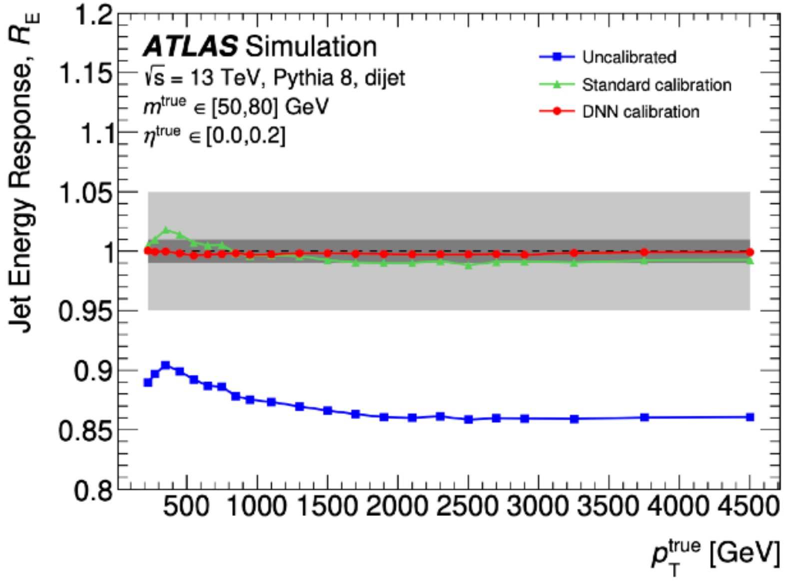
Higher tagging efficiency is achieved by **DNN** than **SVT** !

Simultaneous energy and mass calibration for large-radius jets

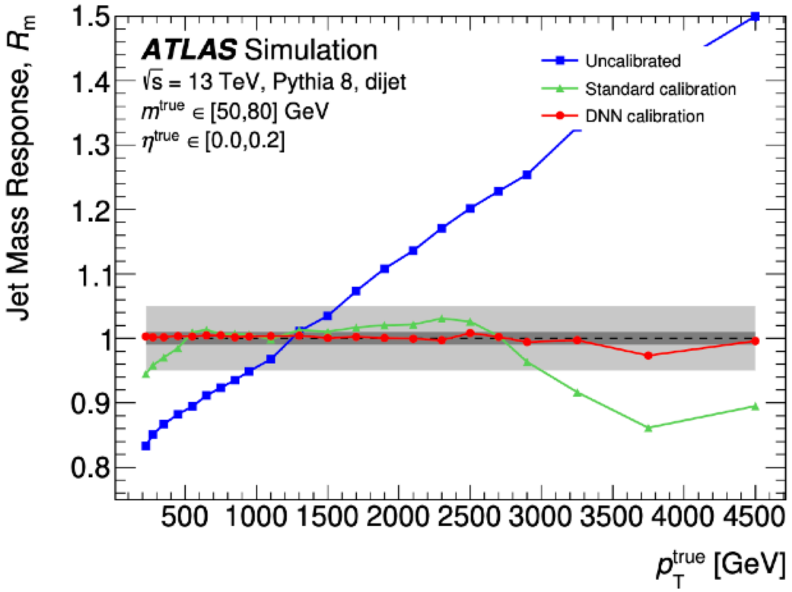
Special deep neural network regression
- Train on jet variables
- Aim to calibrate the energy & mass as close as possible to truth

Efforts to **improve** the performance
- Encoding of **jet position** w.r.t. detector
- Special **loss** to learn the response mode
- **Architecture** & **training** designs

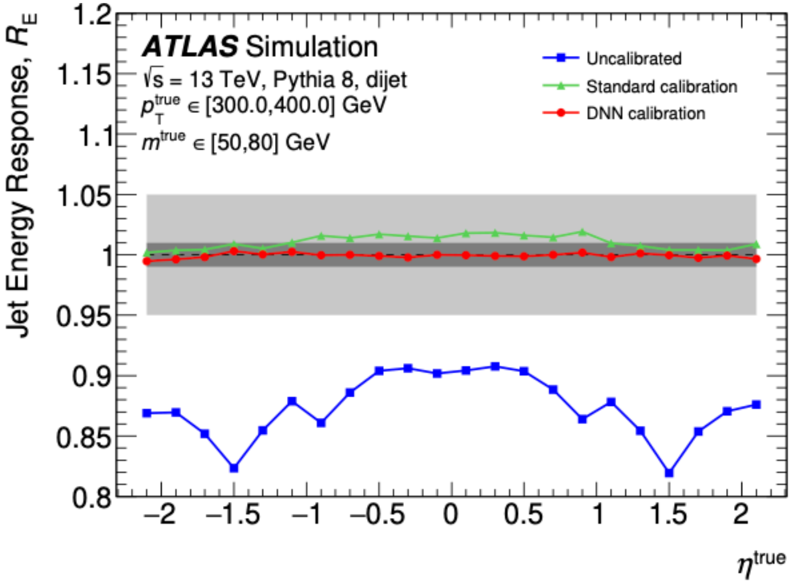
calibrated/truth energy w.r.t. p_T



calibrated/truth mass w.r.t. p_T



calibrated/truth mass w.r.t. η



The **DNN calibration** is **superior** to the **standard calibration**

arxiv:2311.08885

The calibration to **large-radius jets** is important for **heavy-particle search**

Summary

Machine-learning (ML) in analysis with jets
→ Both based on Multifold, event-wise, multi-dimensional

- **ML-based unfolding** with **event shapes** of minimum bias events (CMS)
  - **Unbinned unfolding** and **uncertainty estimation** with **ML-based weighting**
  - **Simultaneous** unfolding of **multiple** variables with **full covariance**
- **ML-based unfolding** of 24 kinematic variables of **Z+jets** (ATLAS)
  - **Unbinned unfolding results**
  - **Different uncertainty estimation** strategy + **background** treatment
- ML-based data-driven dijet anomaly search is covered by Amandeep's talk, Dag's talk

ML techniques for jet calibration and tagging

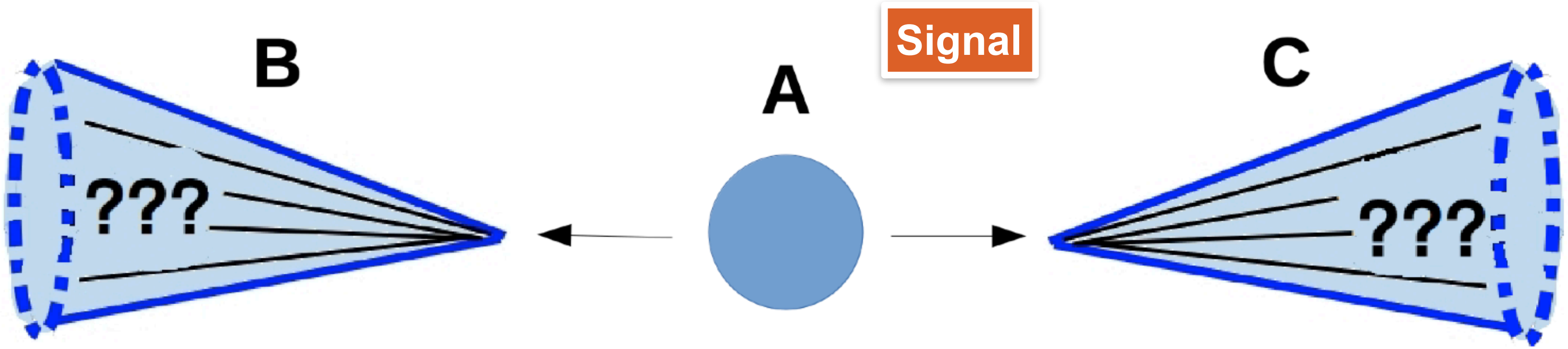
- ML-based jet calibration
  - **Jet p_T** reconstruction under **large background** of underlying event in PbPb collisions (ALICE)
  - **Dijet mass reconstruction** in $H \to b\bar{b}$ and $H \to c\bar{c}$ search (LHCb)
  - **Simultaneous energy and mass** calibration in large radius jets (ATLAS)
- ML-based jet tagging
  - **b and c** tagging against **light-flavor** jets in $H \to b\bar{b}$ and $H \to c\bar{c}$ search (LHCb)
  - More jet tagging results are covered by Andrea's talk on June 7

Backup

Machine-learning-based search

Search for Dijet Resonances with Anomalous Substructure (CMS)

Machine-learning based data-driven dijet anomaly search



heavy particle **A** → much **lighter** daughters **B, C** → boosted decay products as **jets**

Anomalous jet **substructure** from B, C decay → Be used to **distinguish signal** & **QCD background** → **improve the search sensitivity in bump-hunt**

But we prefer **not** to **rely** on specific **models of B & C decay**

Let the **data** tell the **anomalies: Anomaly detector trained directly on data**

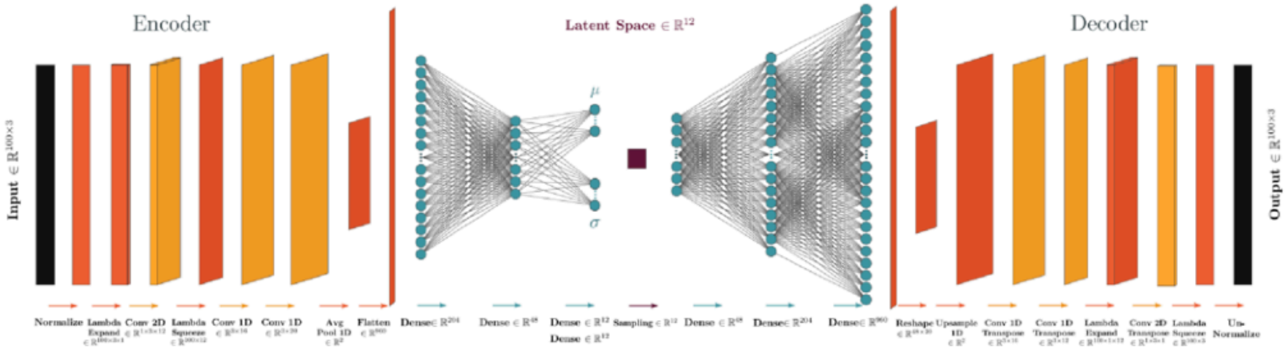
- Outlier detection (VAE-QR)
- Weak supervision (CWoLa Hunting, TNT, CATHODE)
- Multi-signal priors

Entirely data-driven with no MC input

Train on background and mixture of signals

Data-driven anomaly detection: Outlier detection

Variational autoencoder (VAE)



Jet variables from data control region → Compressed → Recovered variables

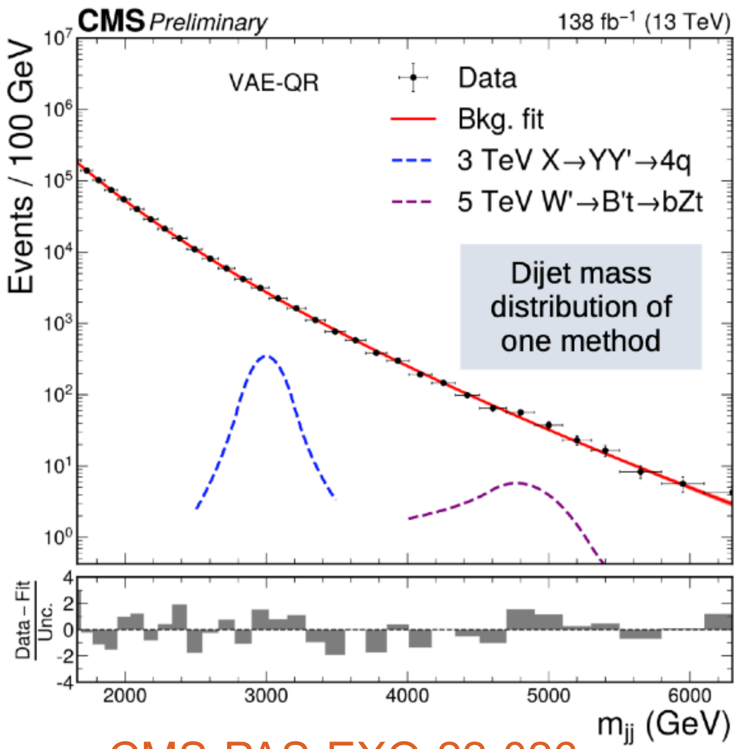
Anomalous score defined as **differences** between the two sets

The Network learned to **compress and decompress** the QCD **background**
But doesn't know how to do this for **anomalous** jets
→ **Lower** anomalous scores for **background**
→ **Higher** scores for **signal**

Cut on the scores for background removal
Additional 'quantile regression' to **decouple** the **cut** with **dijet mass**

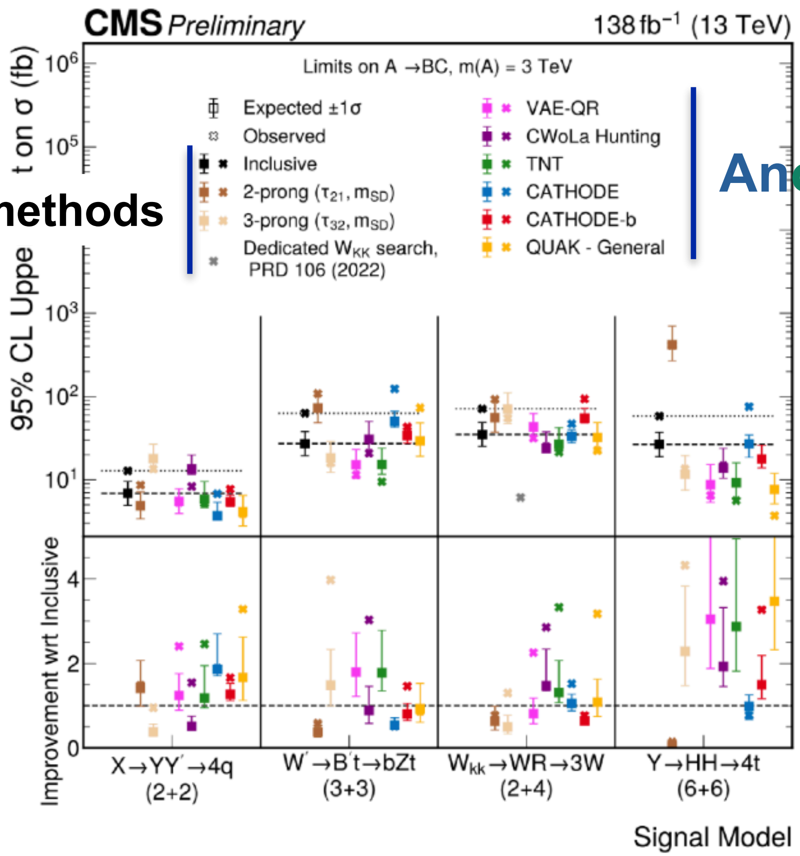
Data with **reduced background** for dijet-mass **bump-hunt**

Sensitivity improvement by the anomaly detectors



CMS-PAS-EXO-22-026

Standard methods



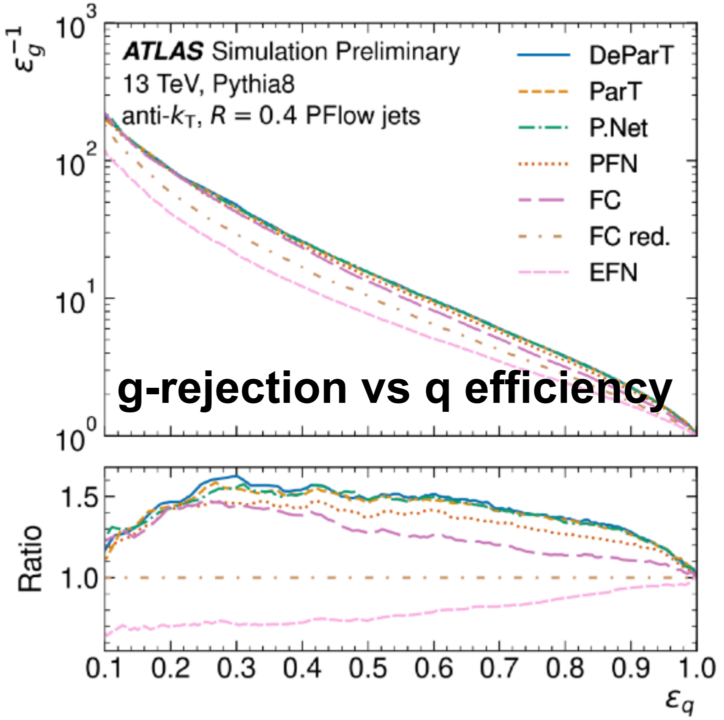
Anomaly detection

No significant excesses from any methods

Test the limits on **several benchmark signals** with varying jet substructures
- Anomaly detection **improves** the **sensitivity** compared to inclusive search
- **More generalisable** than searches for specific substructure
- First usage of anomalous detection in CMS!

Quark vs gluon and W tagging with advanced techniques

Various classifiers are explored for **q/g tagging**
- Particle Flow Network (PFN), Energy Flow Network (EFN), ParticleNet (P.Net), Particle Transformer (ParT), Dynamically-enhanced Particle Transformer (DeParT)
- Reference: Fully Connected (FC), FC reduced

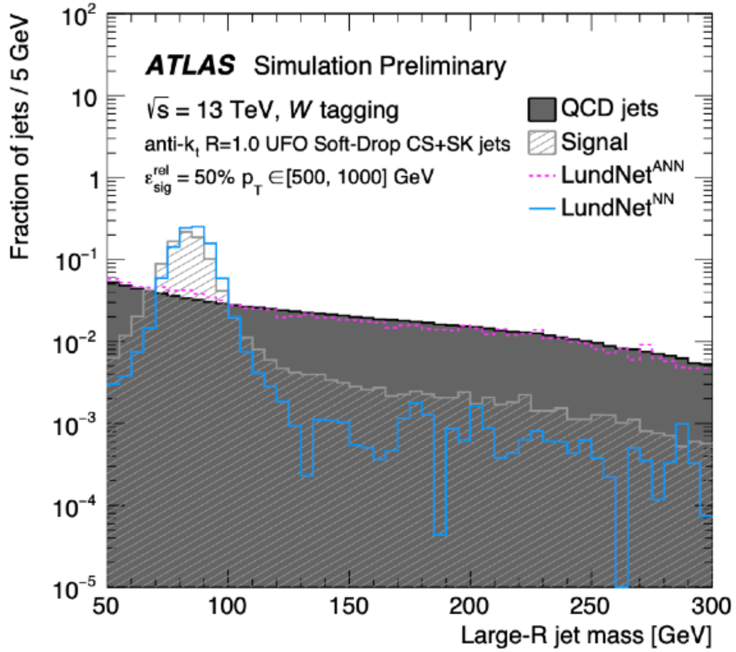


Most advance techniques outperform the reference taggers (except EFN)

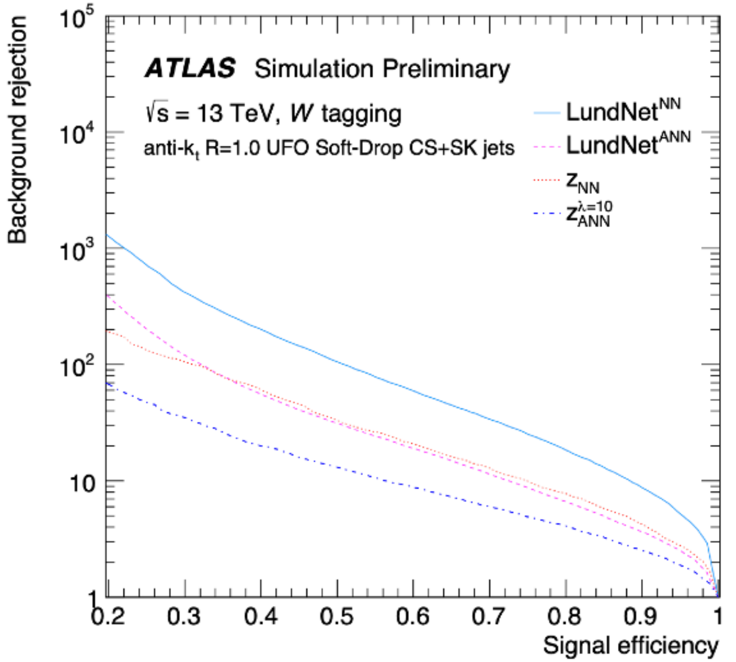
ATL-PHYS-PUB-2023-032

W-tagging with **Lund-plan** tagger
- Use history of jet shower
- Graphical Neural Network (GNN) to learn the "graphs" of the jet

The tagger **changes the background jet mass**
→ Use **Adversarial NN (ANN)** to **decorrelate** mass & tagger

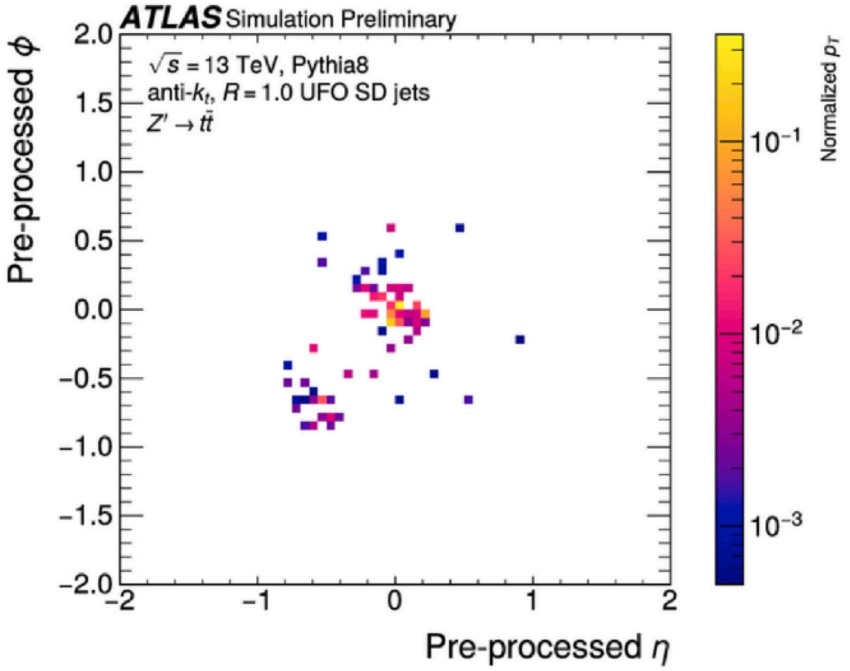


ATL-PHYS-PUB-2023-017



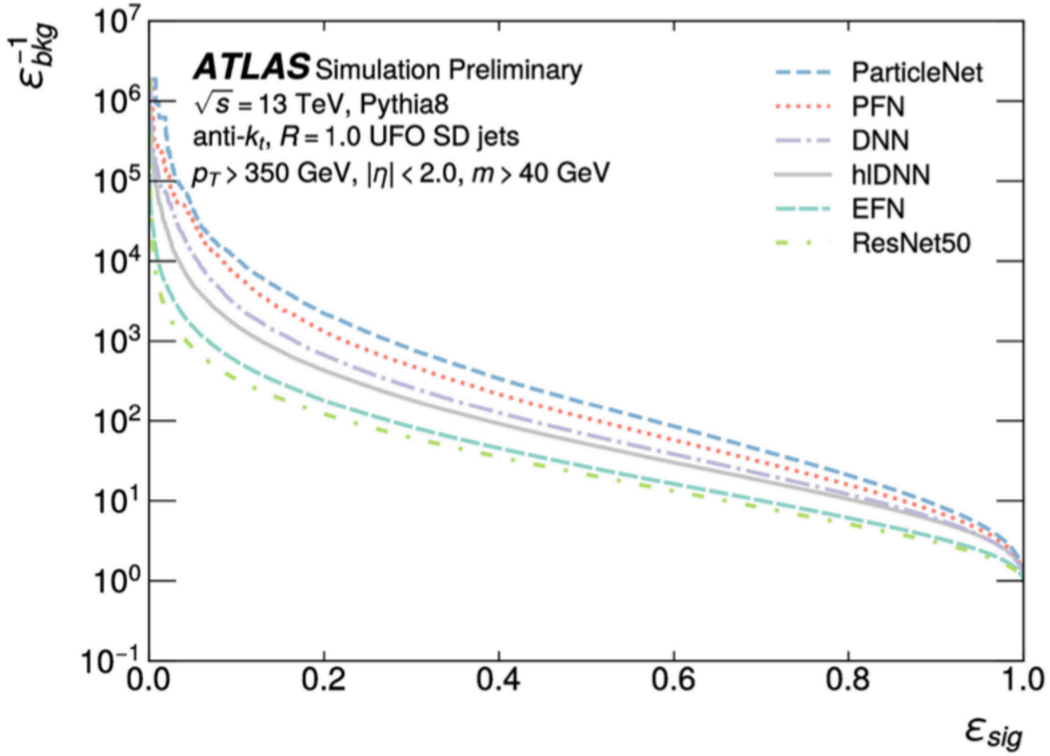
- **LundNet outperforms** the baselines
- mass & tagger decorrelation (ANN) worsen the performance

Top tagging with advanced techniques



ATL-PHYS-PUB-2022-039

Large-scale convolutional neural network for **image classification** is tested (ResNet50)
→ train directly on 2D "jet images"



ParticleNet, PFN, DNN surpass the ResNet50 performance

Systematic uncertainty estimation based on unbinned reweighting

1. **MC statistics**

   Derive the templates by **weighting** the **nominal MC** with **Poisson(1)**
   (similar to the treatment of data stat.)

2. **Track reconstruction efficiency** uncertainty
   - Step1: Randomly drop 2.1%(1%) tracks with pT<20 GeV (>20 GeV) in nominal MC*
   - Step2: weight the nominal MC to Step1 output at particle- and detector-level

* The uncertainty of track reco. eff. is given by D* analysis: https://cds.cern.ch/record/2810814/

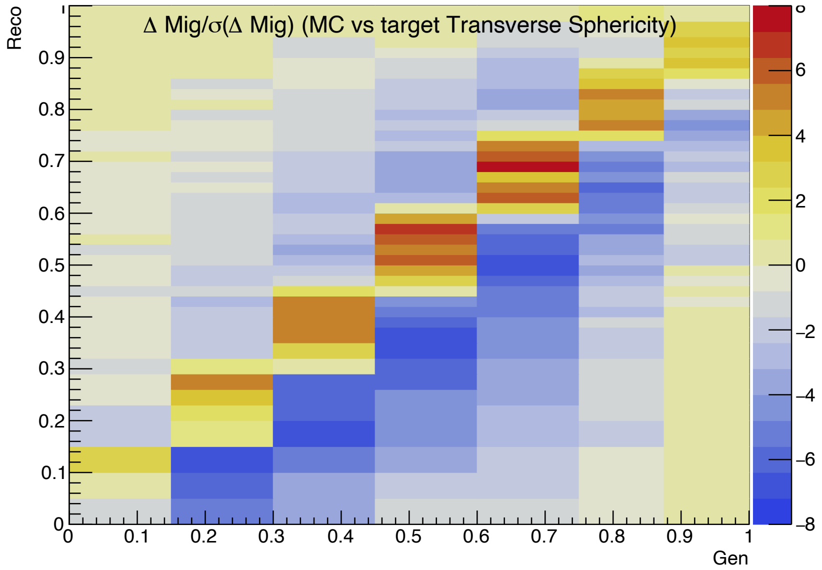
Systematic uncertainty estimation based on unbinned reweighting

1. **MC statistics**
   Derive the templates by **weighting** the **nominal MC** with **Poisson(1)**
   (similar to the treatment of data stat.)

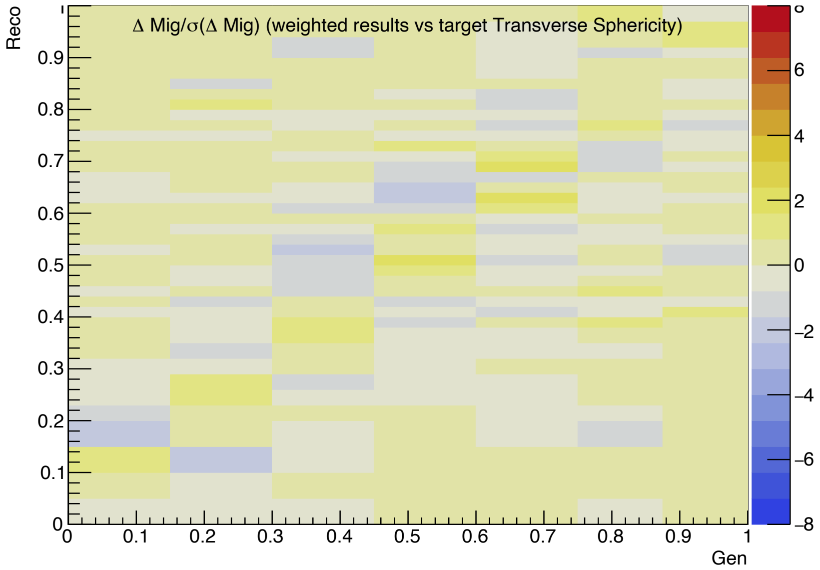
2. **Track reconstruction efficiency** uncertainty
   - Step1: Randomly drop 2.1%(1%) tracks with pT<20 GeV (>20 GeV) in nominal MC*
   - Step2: weight the nominal MC to Step1 output at particle- and detector-level

Example:
Gen → reco migration
of transverse sphericity

Difference between nominal **MC** and **target**
before weighting



After weighting



* The uncertainty of track reco. eff. is given by D* analysis: https://cds.cern.ch/record/2810814/

Systematic uncertainty estimation based on unbinned reweighting

3. **Mismodelling of observables used directly in unfolding**

   Derive the templates by **weighting** **nominal MC** to **alternative MC** at the **particle-level**

   → **ML-based** unbinned weighting

   → output: **weighted nominal MC events**

   - same **particle-level distribution** as **alternative MC**
   - keeps the **gen. → reco. migration** of the **nominal MC**

Systematic uncertainty estimation based on unbinned reweighting

3. **Mismodelling of observables used directly in unfolding**

Derive the templates by **weighting nominal MC** to **alternative MC** at the **particle-level**

→ **ML-based** unbinned weighting

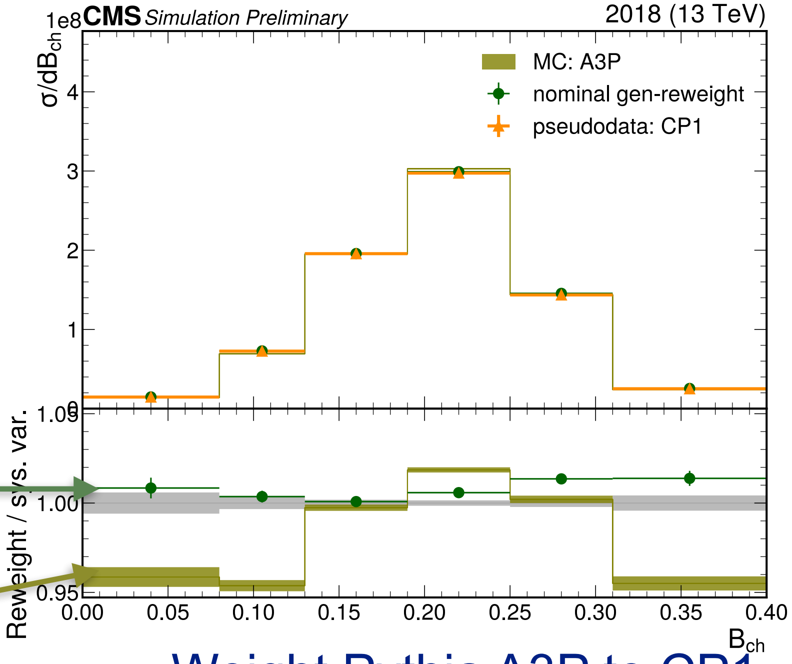
→ output: **weighted nominal MC events**

- same **particle-level distribution** as **alternative MC**
- keeps the **gen. → reco. migration** of the **nominal MC**

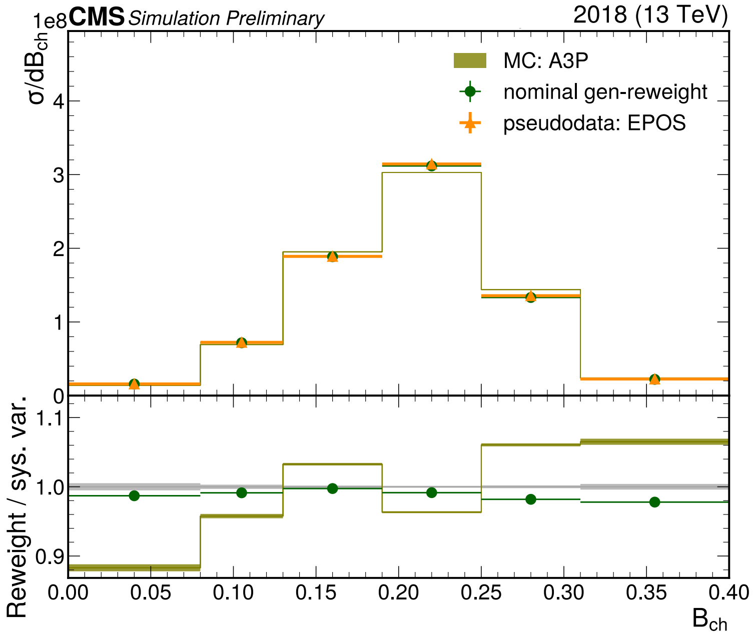
Example:
particle-level broadening
before & after weighting

After reweighting at the gen-level

Nominal MC



Weight Pythia A3P to CP1



Weight Pythia A3P to EPOS

Systematic uncertainty estimation based on unbinned reweighting

4. **Mismodelling of other observables which may change detector response**
   Derive the templates with two-step weighting

Systematic uncertainty estimation based on unbinned reweighting

4. **Mismodelling of other observables which may change detector response**
   Derive the templates with two-step weighting

- **Step 1**: weight the **alternative MC** to **nominal MC** at the **particle-level**
  → output: **weighted alternative MC**
    - with migration function of alternative MC
    - particle-level distributions of nominal MC

Systematic uncertainty estimation based on unbinned reweighting

4. **Mismodelling of other observables which may change detector response**
Derive the templates with two-step weighting

- **Step 1**: weight the **alternative MC** to **nominal MC** at the **particle-level**
  → output: **weighted alternative MC**
    - with migration function of alternative MC
    - particle-level distributions of nominal MC
- **Step 2**: weight the **nominal MC** to the **Step 1 output** at **particle- and detector-level**
  → output: **weighted nominal MC**
    - with migration function of alternative MC
    - particle-level distributions of nominal MC

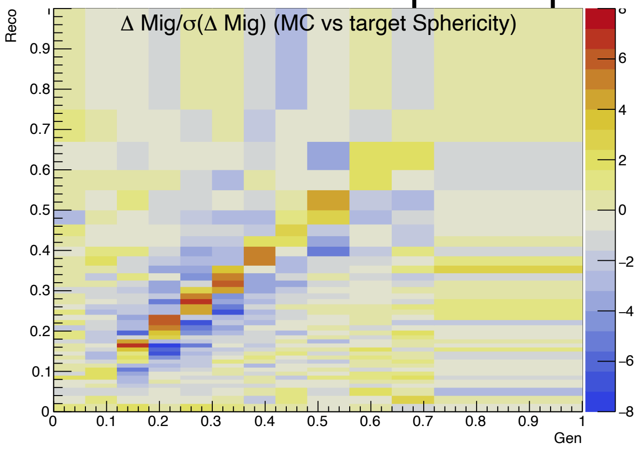
Systematic uncertainty estimation based on unbinned reweighting

4. **Mismodelling of other observables which may change detector response**
Derive the templates with two-step weighting

- **Step 1**: weight the **alternative MC** to **nominal MC** at the **particle-level**
  → output: **weighted alternative MC**
    - with migration function of alternative MC
    - particle-level distributions of nominal MC
- **Step 2**: weight the **nominal MC** to the **Step 1 output** at **particle- and detector-level**
  → output: **weighted nominal MC**
    - with migration function of alternative MC
    - particle-level distributions of nominal MC

Example:
Gen → reco migration of spherocity

Before step 2 weighting:
nominal MC - step 1 output



After step 2 weighting:
weighted result - step 1 output