

Early Science Pseudo-Data Production

Update on the SIDIS binning definition

Salvatore Franzese

September 08, 2026

In this work

Goal: Generate Early Science pseudo-data for unpolarized TMD impact studies.

The observable: The four-dimensional SIDIS cross section

$$\frac{d^4\sigma}{dx_B dQ^2 dz dP_{hT}},$$

Output: For each identified hadron ($\pi^\pm$, $K^\pm$), we produce tables containing the multidimensional bin, the extracted cross section and its uncertainty.

Purpose of this update: Define a binning that preserves the relevant kinematic information without introducing intervals that are too narrow compared with the experimental resolution.

Data sample and analysis selection

The study is performed on ep collisions from the **July 2026 ePIC simulation campaign**, generated with **PYTHIA 6-eic** for the 9×275 GeV and 9×130 GeV beam configurations.

Event selection

- DIS events are selected through the reconstructed scattered-electron kinematics.
- The reconstructed electron is truth-matched to the generated electron.
- $|\eta_e| \leq 3.5$, $0.01 \leq y \leq 0.95$, and $p_e > 1$ GeV.

Hadron selection

- $\pi^\pm$ and $K^\pm$ tracks are selected within the DIS sample and truth-matched to the generated hadrons.
- $|\eta_h| < 3.5$; reconstructed and generated particle species are required to agree.

The present study concerns the kinematic binning. The event selection and reconstruction procedure are unchanged with respect to the previous update.

Reconsidering the binning

We start from the binning previously used in the SIDIS analysis. The aim of this study is to assess whether the existing bin widths are compatible with the detector resolution, and to identify the regions where a **larger binning is needed**.

Bins that are too narrow compared to the detector resolution are strongly affected by migrations.

We therefore use the reconstructed resolution to identify the regions where the original binning should be **merged into wider bins**, and validate the resulting choice through migration matrices afterwards.

Results shown: π^+ at 9×275 GeV. The binning was also validated independently for π^- , $K^\pm$, and at 9×130 GeV. See [here](#).

Resolution variables

The resolution is studied separately for each kinematic variable, using truth–reconstructed associated pairs and integrating over the other three variables.

For Q , x_B , and z , we use the relative residual

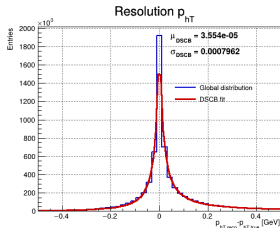
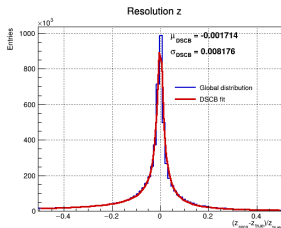
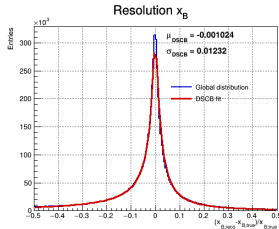
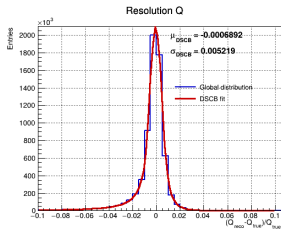
$$\delta_v = \frac{v_{\text{reco}} - v_{\text{true}}}{v_{\text{true}}}, \quad v = Q, x_B, z.$$

For P_{hT} , we use the absolute residual

$$\delta_{P_{hT}} = P_{hT,\text{reco}} - P_{hT,\text{true}}.$$

For P_{hT} , which extends down to zero, the relative residual becomes ill-defined near the origin, so the absolute residual is used instead.

Resolution distributions



Global resolution

The reconstructed distributions contain a gaussian core component together with non-Gaussian tails. This is particularly relevant for z and P_{hT} :

- the RMS is very sensitive to the extreme events;
- the width of the Gaussian or DSCB core alone would instead give too little weight to the populated tails.

Definition of the resolution range

For each candidate bin, we consider its residual distribution and select the interval containing the central 95% of the events. The mean μ_i and the RMS are then evaluated within this interval.

The mean quantifies the **reconstruction bias** in that bin. The RMS corresponds to the **resolution scale**.

We retain most of each non-Gaussian distribution, while limiting the influence of the most extreme 5% of events.

Note that events in the outer 5% are not removed from the analysis and are fully included in the subsequent migration study.

Binning requirements

For each candidate bin i , its width is compared with the corresponding local resolution scale:

$$R_i = \frac{\Delta_i^{(\text{res})}}{\text{RMS}_i^{95}}$$

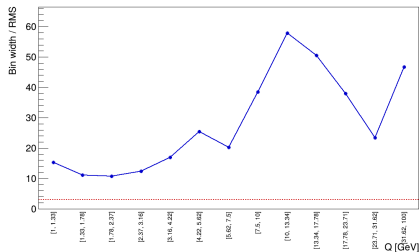
This quantity is used to construct the binning and the final performance is evaluated independently through migration matrices.

Variable	Requirement
Q	$R_i \geq 3$
x_B	$R_i \geq 3$
z	$R_i \geq 1$
P_{hT}	$R_i \geq 1$

Q and x_B allow the more conservative requirement $R_i \geq 2$ while retaining a detailed partition.

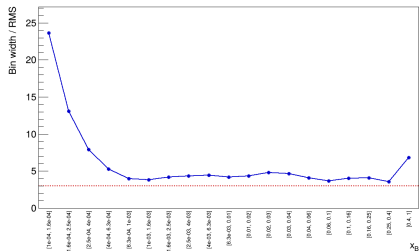
For z and P_{hT} , we use $R_i \geq 1$ to preserve the granularity relevant for fragmentation and TMD studies.

Reconstruction bias and resolution: Q , x_B



For both inclusive variables, the populated intervals satisfy the resolution requirement $R_i \geq 2$.

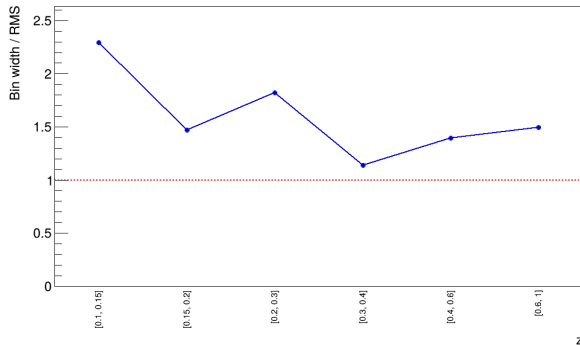
The original bin edges for Q are retained without modifications.



For x_B we remove the intervals below 10^{-4} (empty for the considered beam energies, while at high x_B , the edge at 0.63 is removed, merging the intervals $[0.40, 0.63]$ and $[0.63, 1.0]$).

Binning result: z , p_{hT}

Some original intervals are narrower than the local resolution scale and therefore do not satisfy $R_i \geq 1$. Adjacent intervals are merged until every final bin fulfills the requirement.



Original z binning

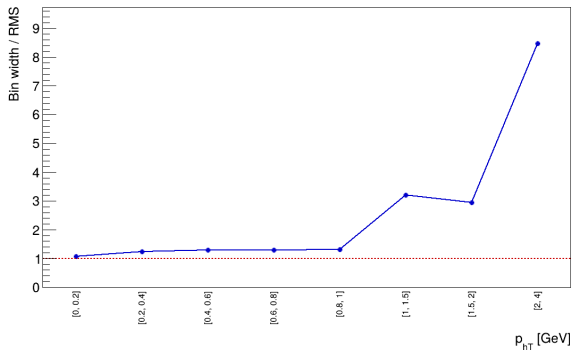
[0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30,
0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00]

Final z binning

[0.10, 0.15, 0.20, 0.30, 0.40, 0.60, 1.00]

Although well reconstructed, the region $z < 0.1$ is excluded since it is outside the region targeted by the SIDIS phenomenological analysis.

Binning result: P_{hT}



Original P_{hT} binning [GeV]

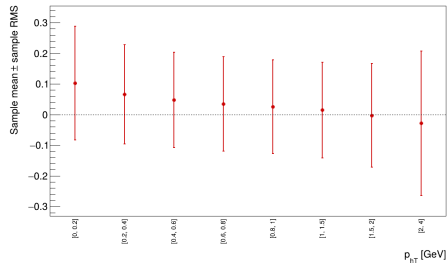
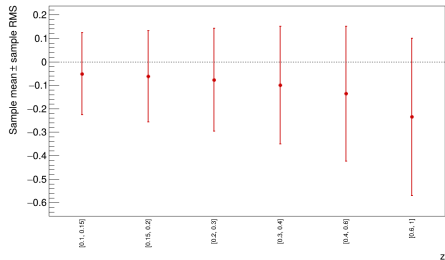
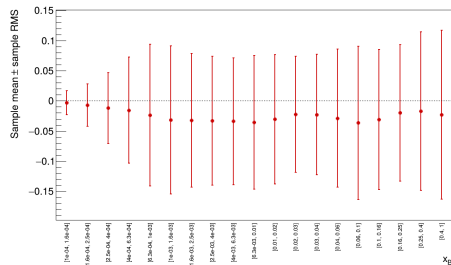
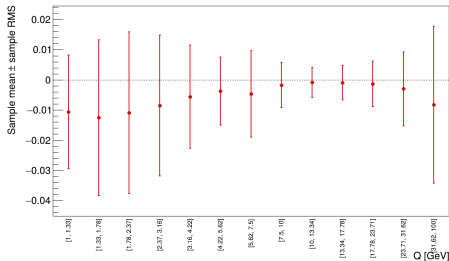
[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
0.8, 0.9, 1.0, 1.5, 2.0, 4.0]

Final P_{hT} binning [GeV]

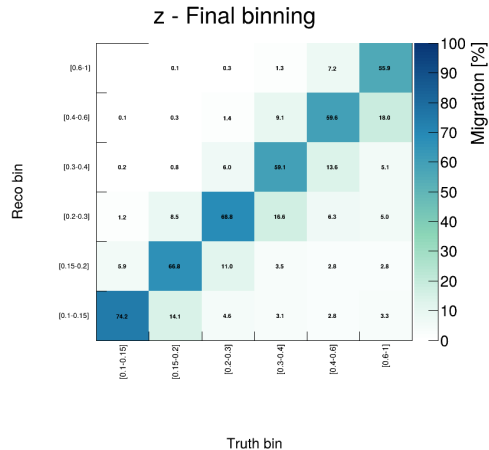
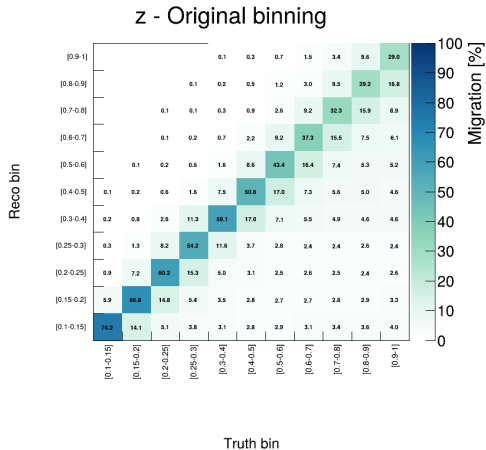
[0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 4.0]

Fine binning is retained below 1 GeV, where the cross section is most sensitive to the non-perturbative transverse-momentum dependence. At higher P_{hT} , wider bins are sufficient for the intended TMD fits.

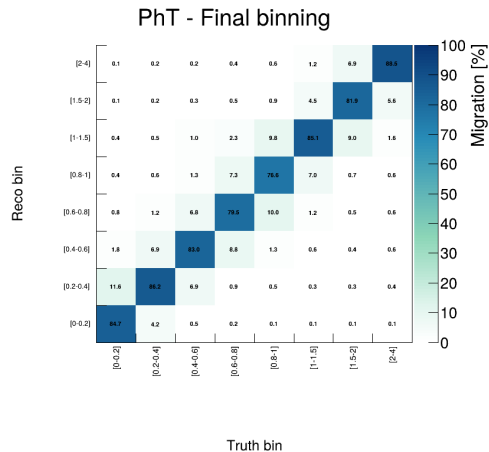
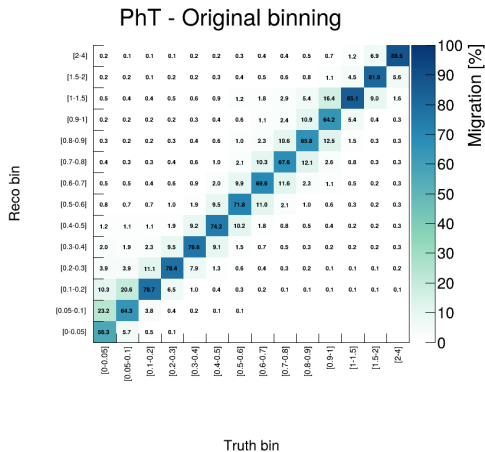
Reconstruction bias and resolution



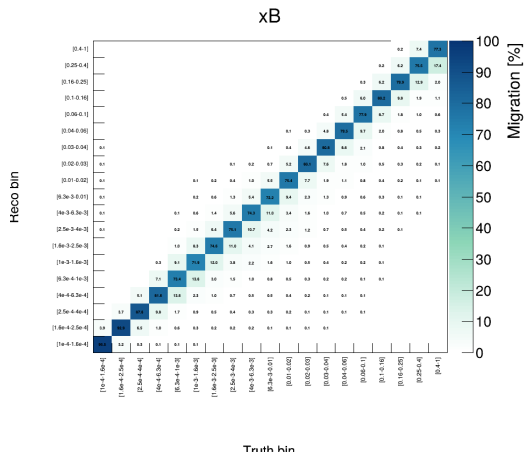
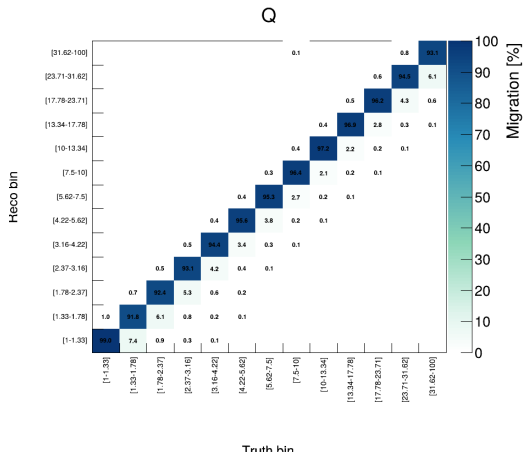
Migration matrix in z: previous vs final binning



Migration matrix in P_{hT} : previous vs final binning



Migration matrices in Q and x_B



Summary and outlook

Binning definition

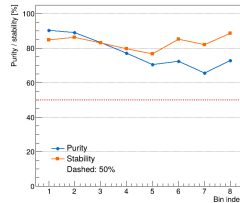
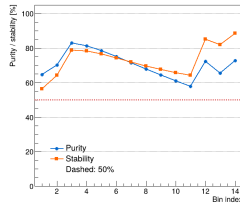
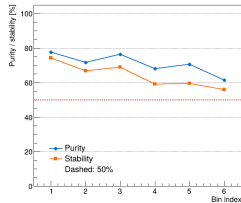
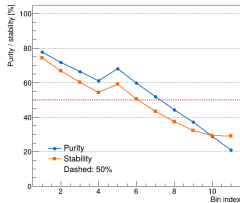
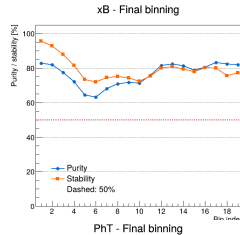
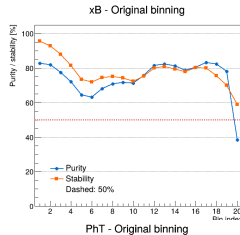
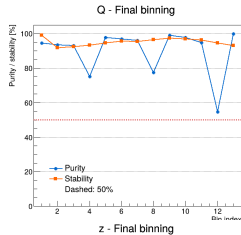
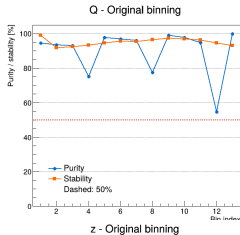
- The binning is constructed from a marginal resolution study, independently for Q , x_B , z and P_{hT} .
- A global central-95% residual interval makes the local RMS less sensitive to extreme non-Gaussian events without reducing it to the Gaussian core.

Validation

- Migration matrices are evaluated only after the edges are fixed.
- The final partition substantially improves the weakest bins, most clearly at high z .

Backup

Purity and stability



Complete final binning

Q bin edges:

1, 1.33417, 1.77764, 2.37065, 3.16228, 4.21900, 5.62139,
7.49667, 10, 13.3417, 17.7764, 23.7065, 31.6228, 100

x_B bin edges:

10^{-4} , 1.59×10^{-4} , 2.51×10^{-4} , 3.98×10^{-4} , 6.31×10^{-4} ,
 10^{-3} , 1.59×10^{-3} , 2.51×10^{-3} , 3.98×10^{-3} , 6.31×10^{-3} ,
 10^{-2} , 1.59×10^{-2} , 2.51×10^{-2} , 3.98×10^{-2} , 6.31×10^{-2} ,
0.1, 0.159, 0.251, 0.398, 1

z bin edges:

0.10, 0.15, 0.20, 0.30, 0.40, 0.60, 1.00

P_{hT} bin edges [GeV]:

0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 4.0

Why not use the full RMS or the DSCB core width?

Full sample RMS

Includes all tails, but a small number of very distant events can dominate it through their squared residuals and force excessive merging.

Gaussian or DSCB σ

Describes the central component. The DSCB tail parameters characterize the non-Gaussian behavior, but its σ remains primarily the core width and would be too optimistic as the only binning scale.

Adopted solution

The sample RMS in the global central-95% interval preserves most of the tails while limiting the influence of the most extreme events.

Why use different R requirements?

For Q and x_B , the reconstructed resolution allows $R \geq 2$ while maintaining many intervals over the full kinematic range. There is therefore no reason to adopt a less conservative requirement.

For z and P_{hT} , the same condition would remove much of the structure that the pseudo-data are intended to constrain:

- the evolution in z is directly related to fragmentation;
- the detailed dependence below $P_{hT} = 1$ GeV is particularly relevant for TMD studies.

We therefore require $R \geq 1$ and then test this choice against the actual migrations. The final minimum purity and stability remain above 50% for both variables.

Validation with migration matrices

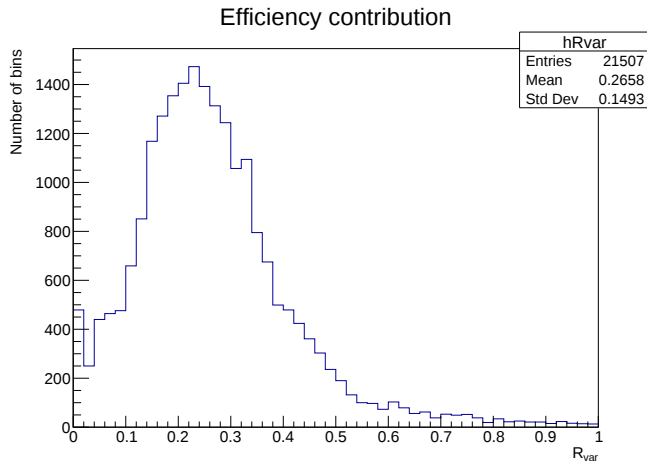
After fixing the binning through the resolution criterion, truth–reconstructed **migration matrices** are produced as validation.

To summarize the diagonal content, we also compute purity $P_i = N_{ii}/N_i^{\text{reco}}$ and stability $S_i = N_{ii}/N_i^{\text{truth}}$.

$P_i \longrightarrow$ contamination entering the bin

$S_i \longrightarrow$ migration out of the bin

Efficiency uncertainty



The relative contribution of the efficiency uncertainty to the nominal statistical variance is quantified as

$$R_{\text{var}} = \frac{(\delta\epsilon/\epsilon)^2}{1/N_{\text{reco}} + (\delta\epsilon/\epsilon)^2}.$$

The distribution has a mean value

$$\langle R_{\text{var}} \rangle \simeq 0.27.$$