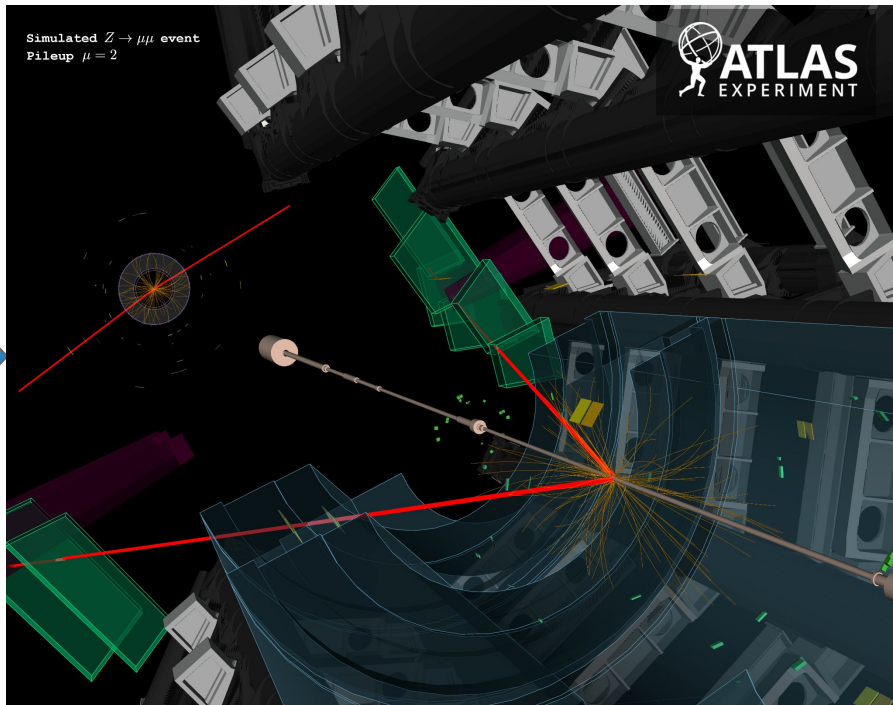




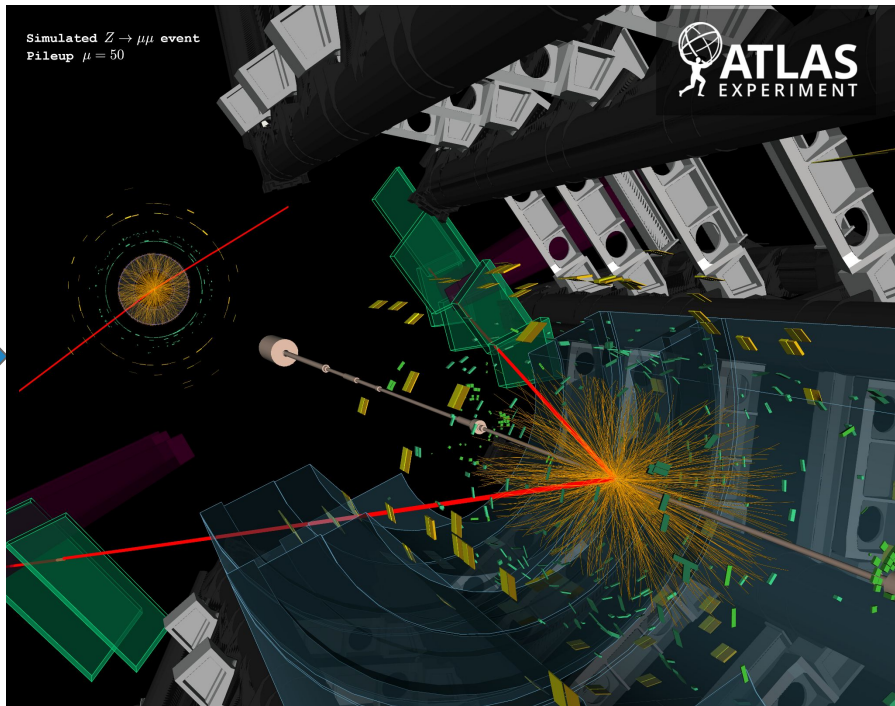
Weakly Supervised Training for Optimal Transport Pileup Mitigation Strategies at Hadron Colliders

Nathan Suri, Vinicius Mikuni
ID Graphs/Transformers Meeting
2/13/24

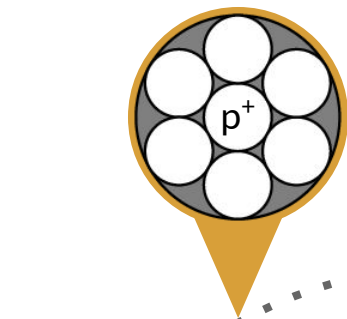




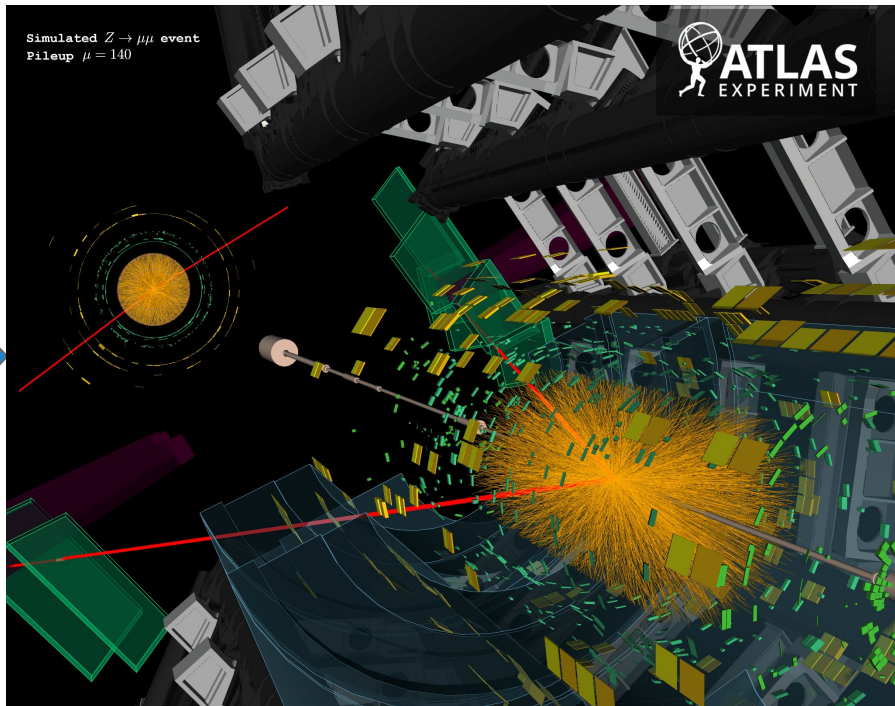
Charged + neutral pileup



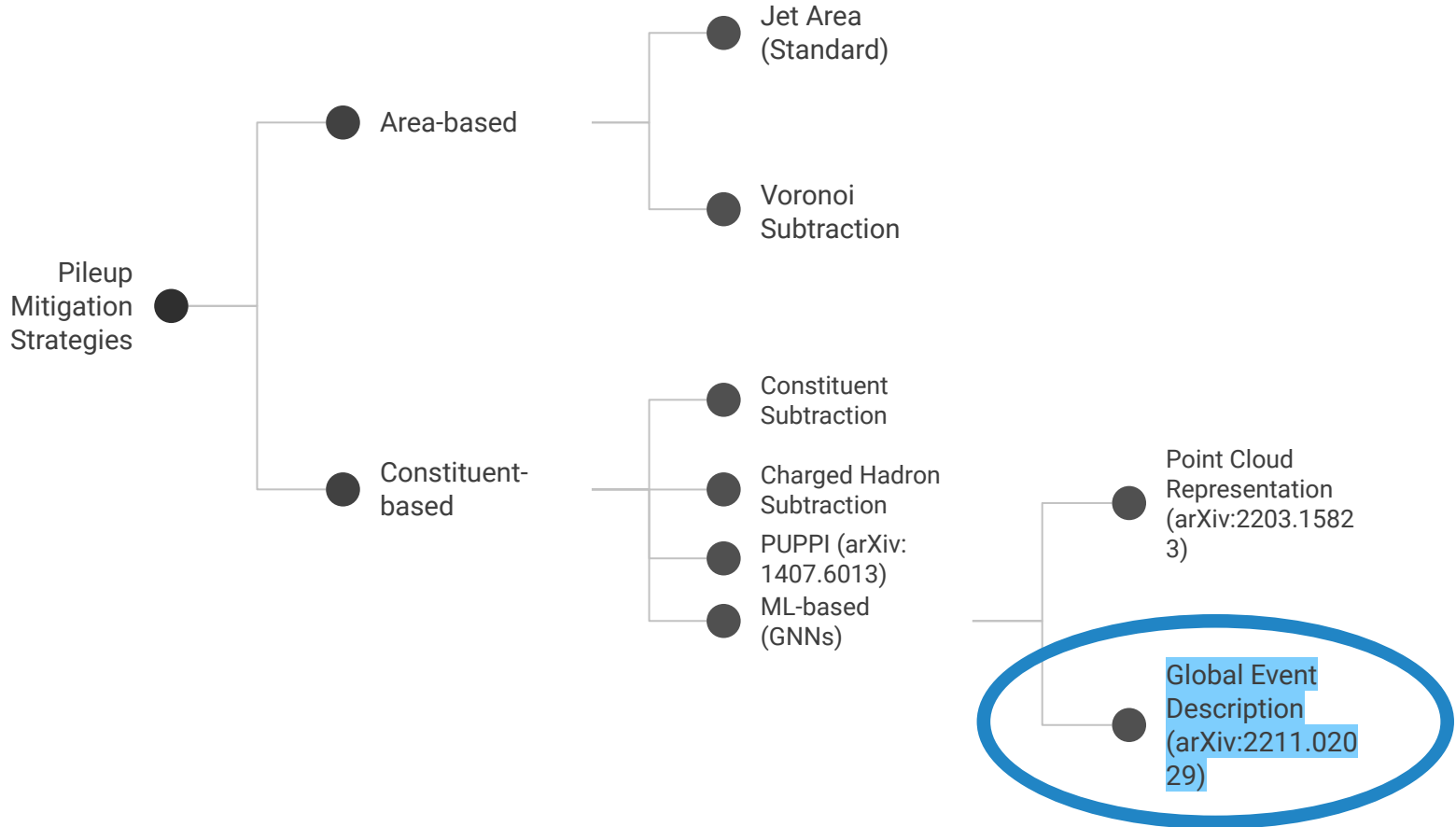
Simulated $Z \rightarrow \mu\mu$ event
Pileup $\mu = 50$



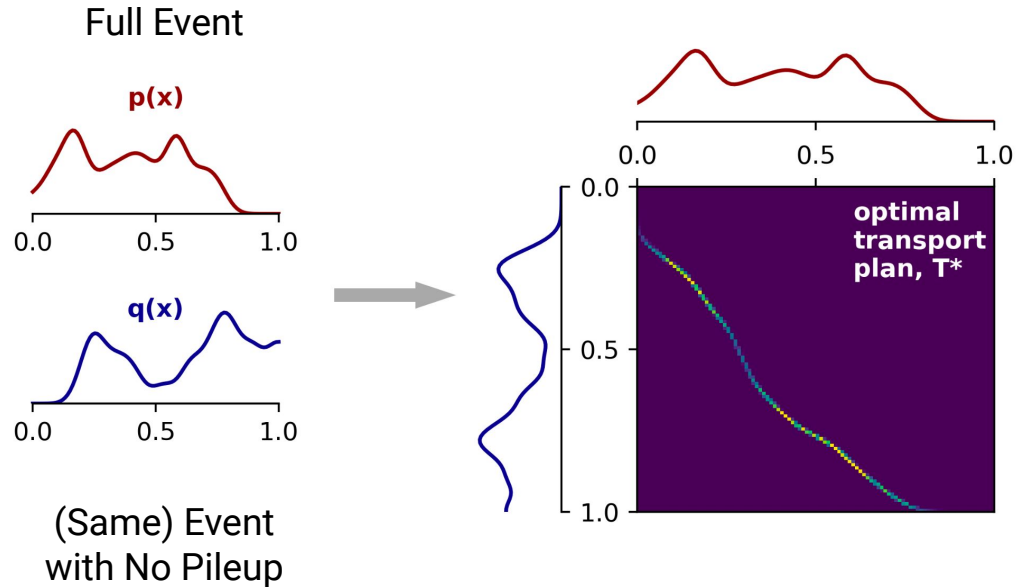
Charged + neutral pileup



Charged + neutral pileup

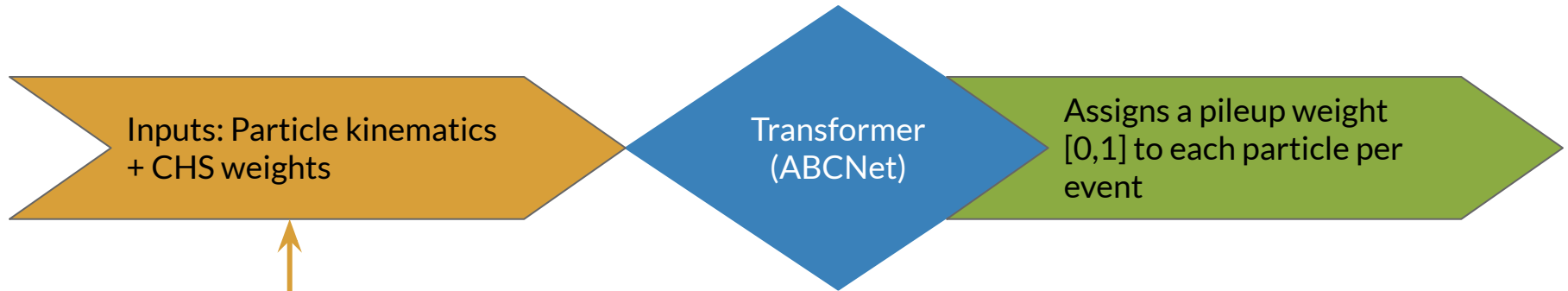


TOTAL: Training Optimal Transport with Attention Learning



$$\mathcal{L} = \text{SWD}(x'_p, x_{np}) + \lambda \text{MSE}(p_T^{\text{miss}}(x'_p), p_T^{\text{miss}}(x_{np}))$$

TOTAL: Training Optimal Transport with Attention Learning



1. Point cloud is represented as a graph with vertices represented by the points themselves
2. Edges are constructed by connecting the points to their k-nearest neighbours, while the edge features are taken as the difference between features of each point and its k-neighbours

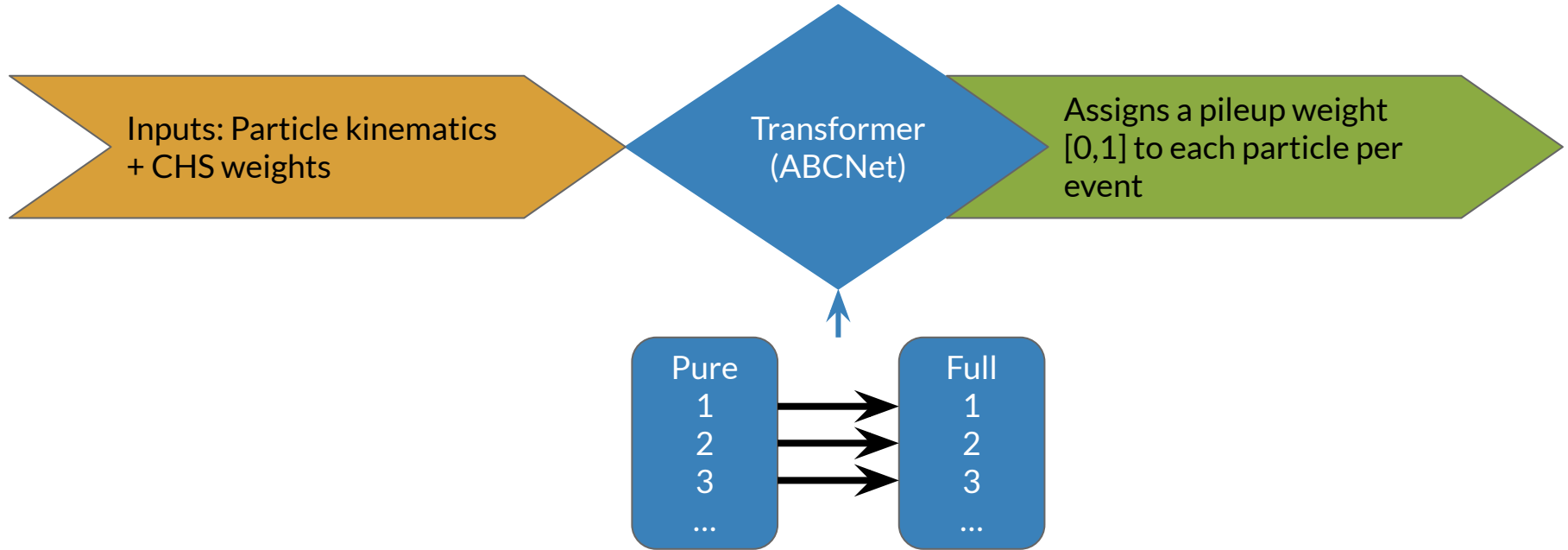
TOTAL: Training Optimal Transport with Attention Learning



1. Calculate optimal transport cost function per event
2. Take average of all squared differences of costs, which yields the optimal transport mapping between paired events

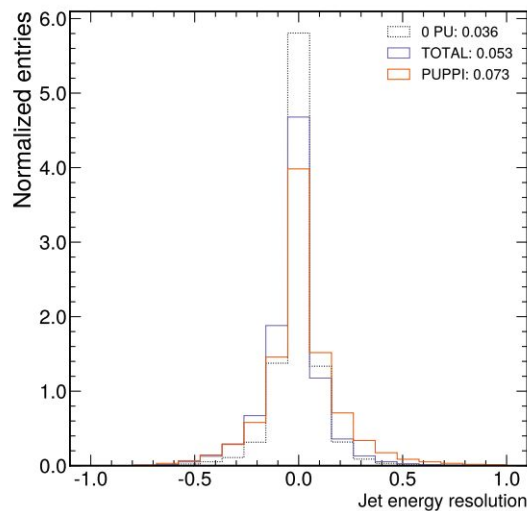
(arXiv:2001.05311)

TOTAL: Training Optimal Transport with Attention Learning

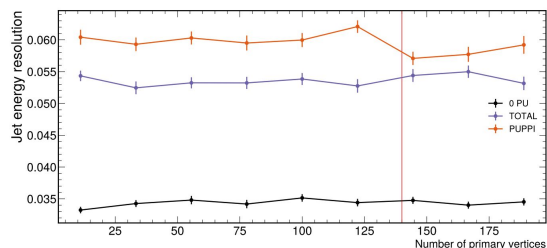


TOTAL: Training Optimal Transport with Attention Learning

- + Outperforms traditional and ML-based alternatives
- + Relies on global event descriptions
- + Robustly learns pileup characteristics as a transport function



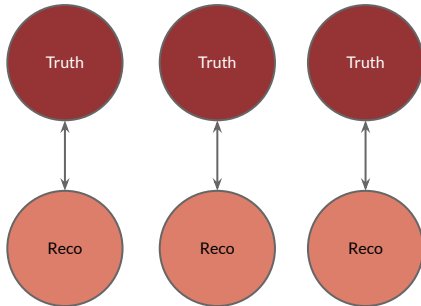
- Requires direct matching of events
- Overall limited due to supervision



TOTAL: Training Optimal Transport with Attention Learning

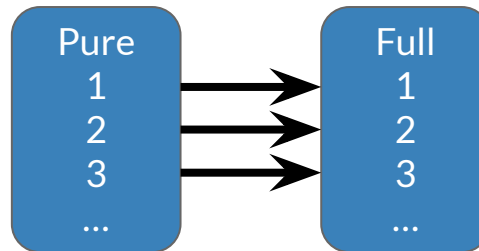
ML Competitors

- Matching between truth and reco at particle level (MC correction)



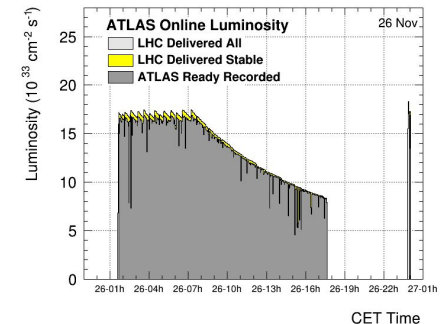
TOTAL

- Matching between pileup events and same event without pileup vertices (data-driven*)



Δ TOTAL

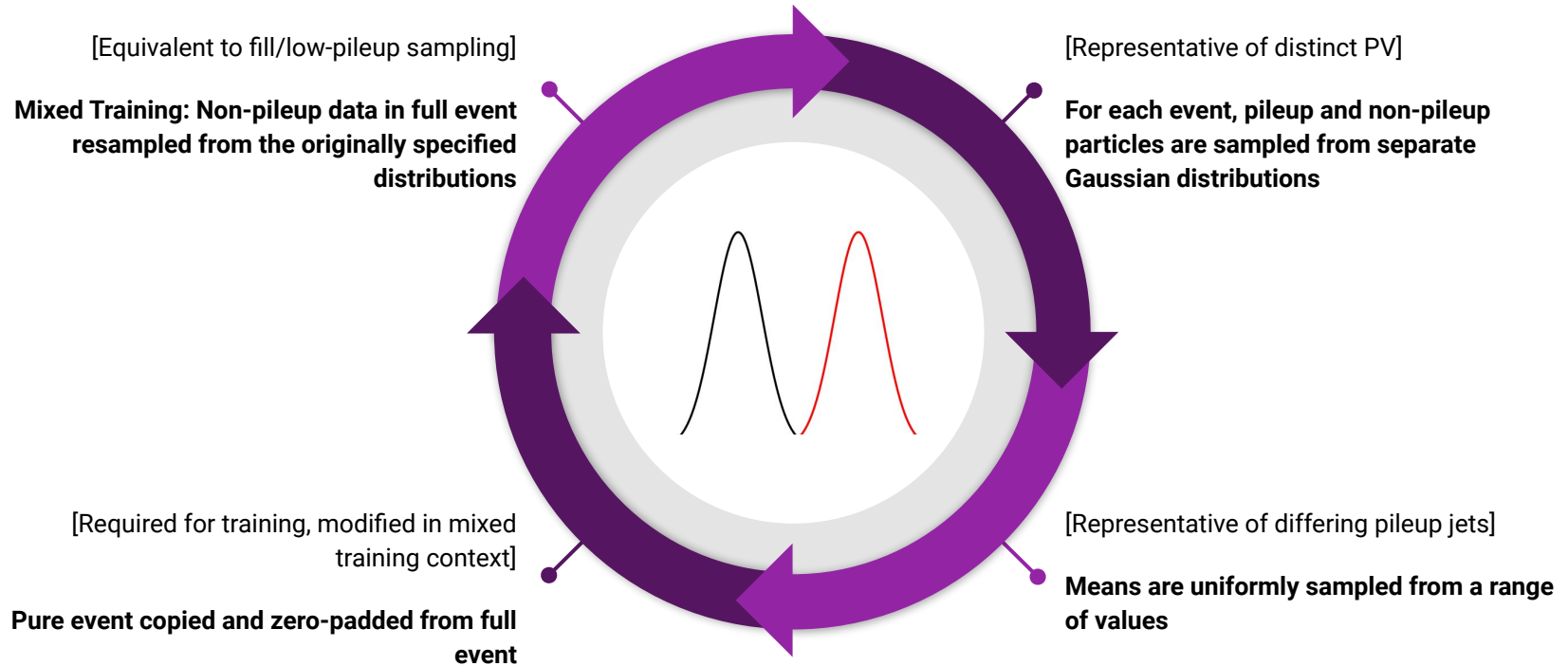
- Matching between ensembles of events with different relative pileup densities (fully data-driven)



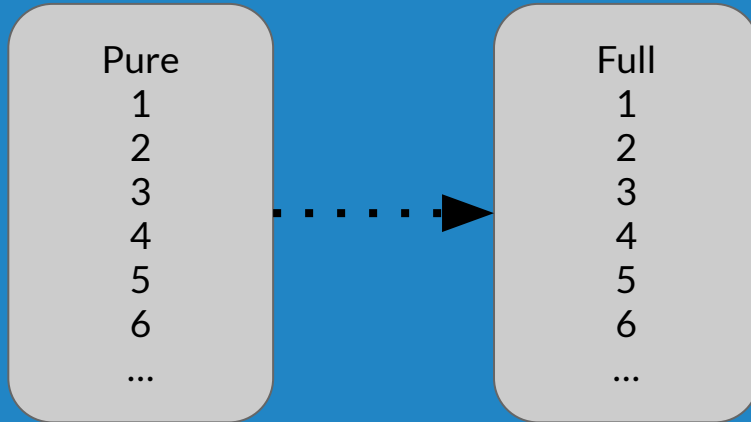


*What happens if we do not require
direct matching?*

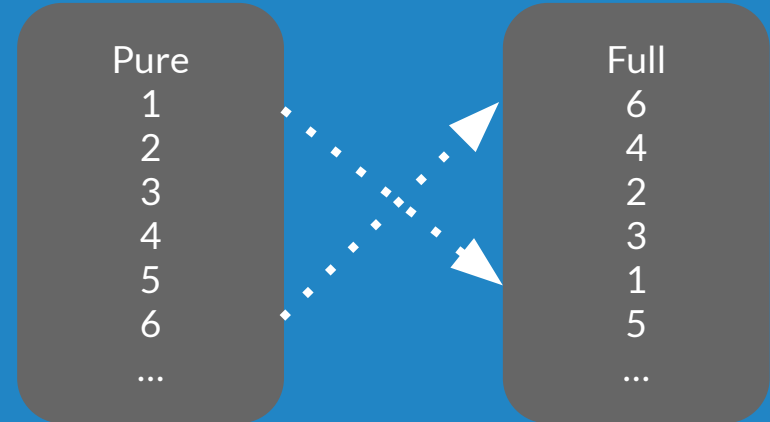
Toy Example Generation



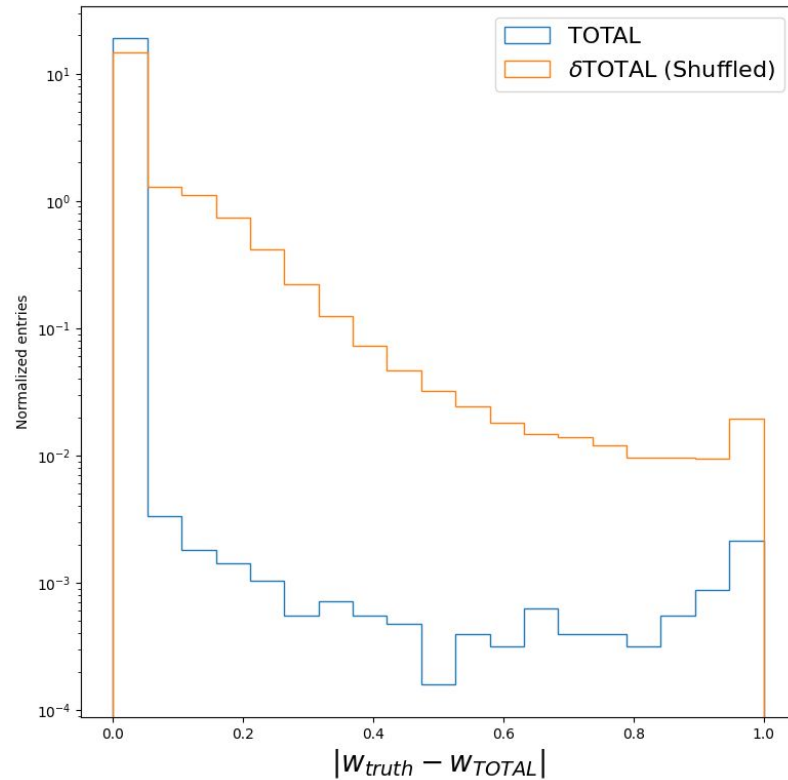
Original



Shuffling



Toy Example

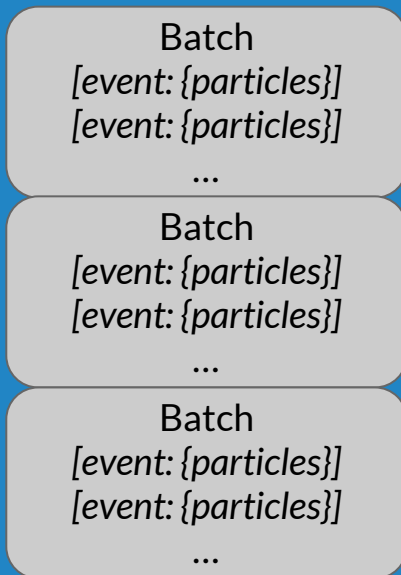




*How can we mitigate the
information loss of not matching
events?*

Original

$[n_{\text{batch}} \times n_{\text{particles}} \times n_{\text{features}}]$

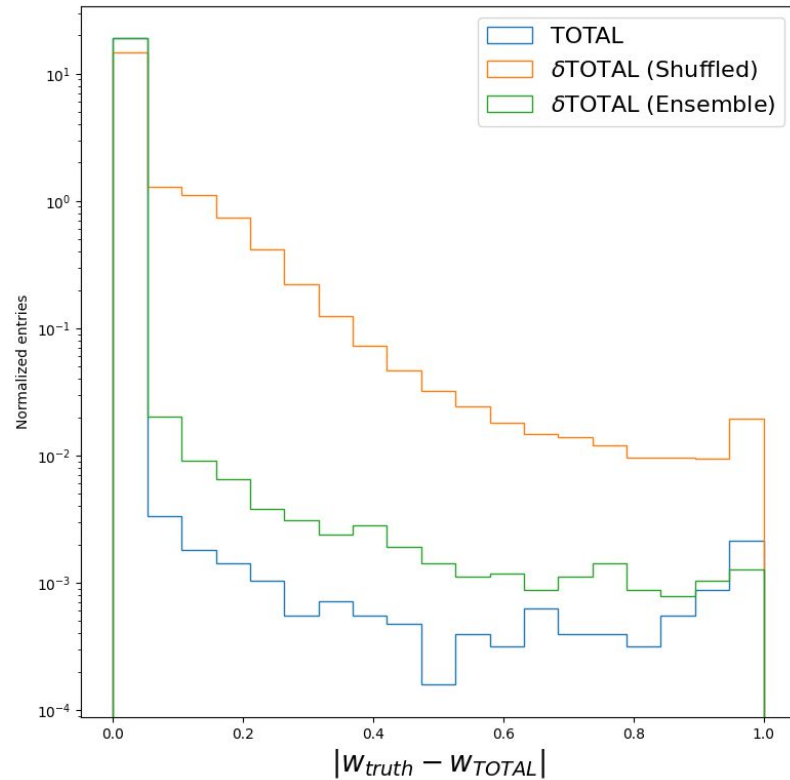


Enhanced

$[(n_{\text{batch}} \times n_{\text{particles}}) \times n_{\text{features}}]$



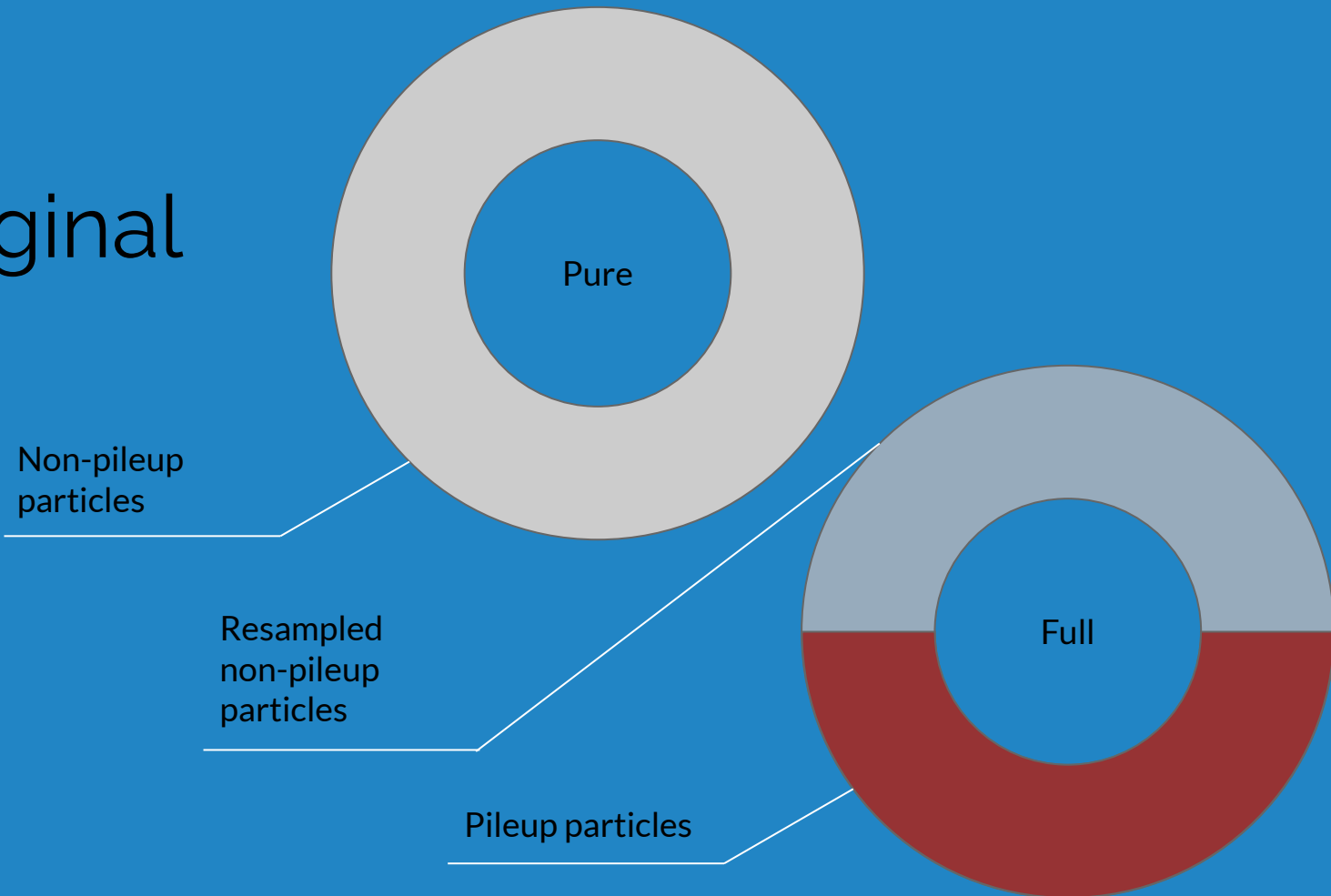
Toy Example



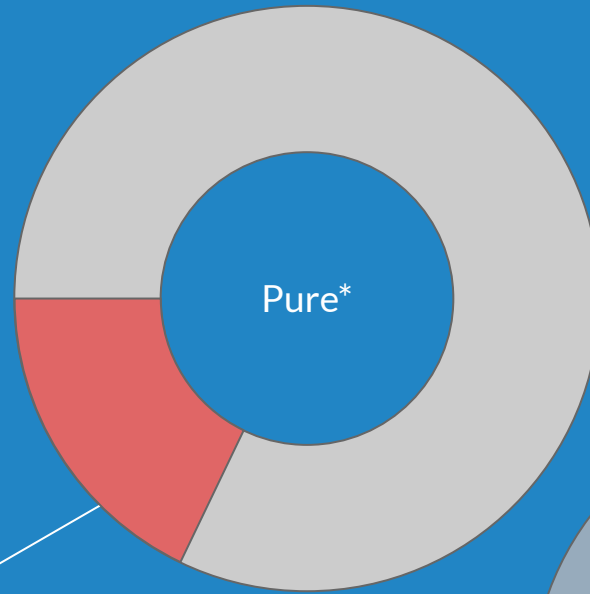


What happens if we decrease the purity of the non-pileup sample?

Original

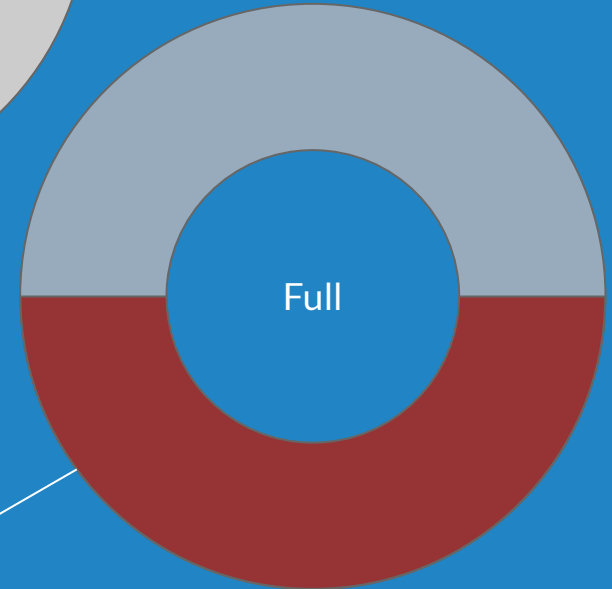


Mixed



Injected pileup particles (c%)

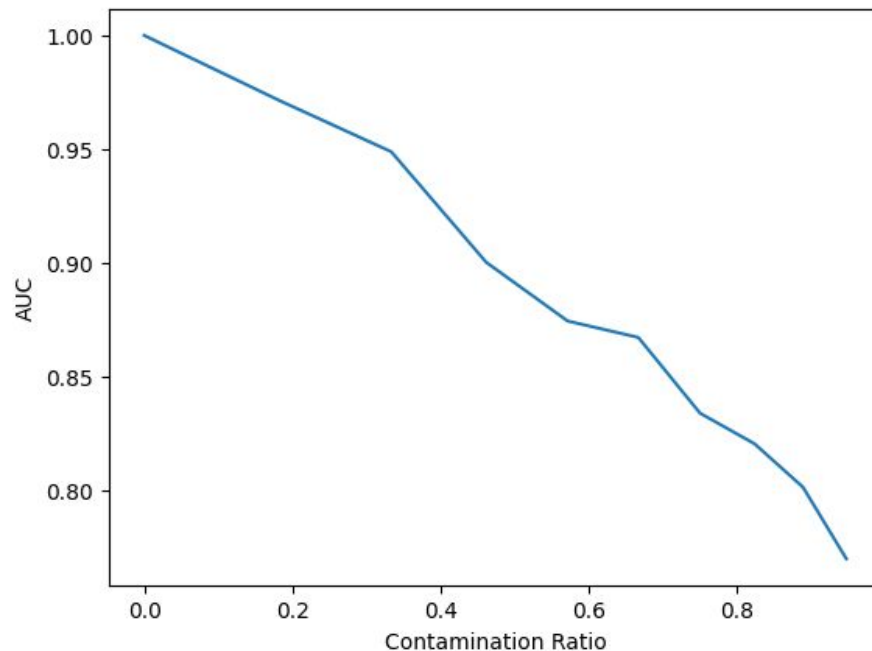
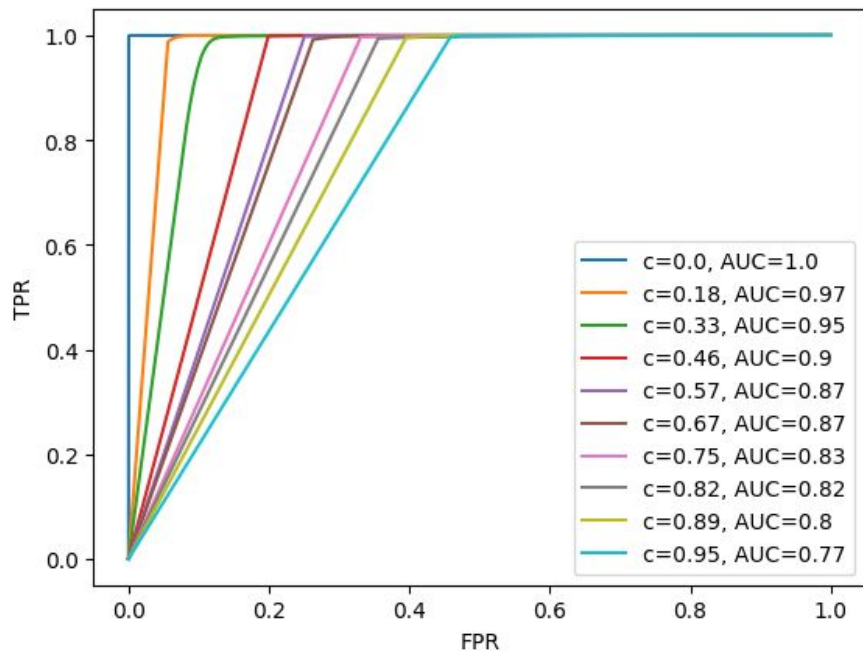
$$c = \frac{\frac{n_{p,pure}}{(n_{p,pure} + n_{np,pure})}}{\frac{n_{p,full}}{(n_{p,full} + n_{np,full})}}$$



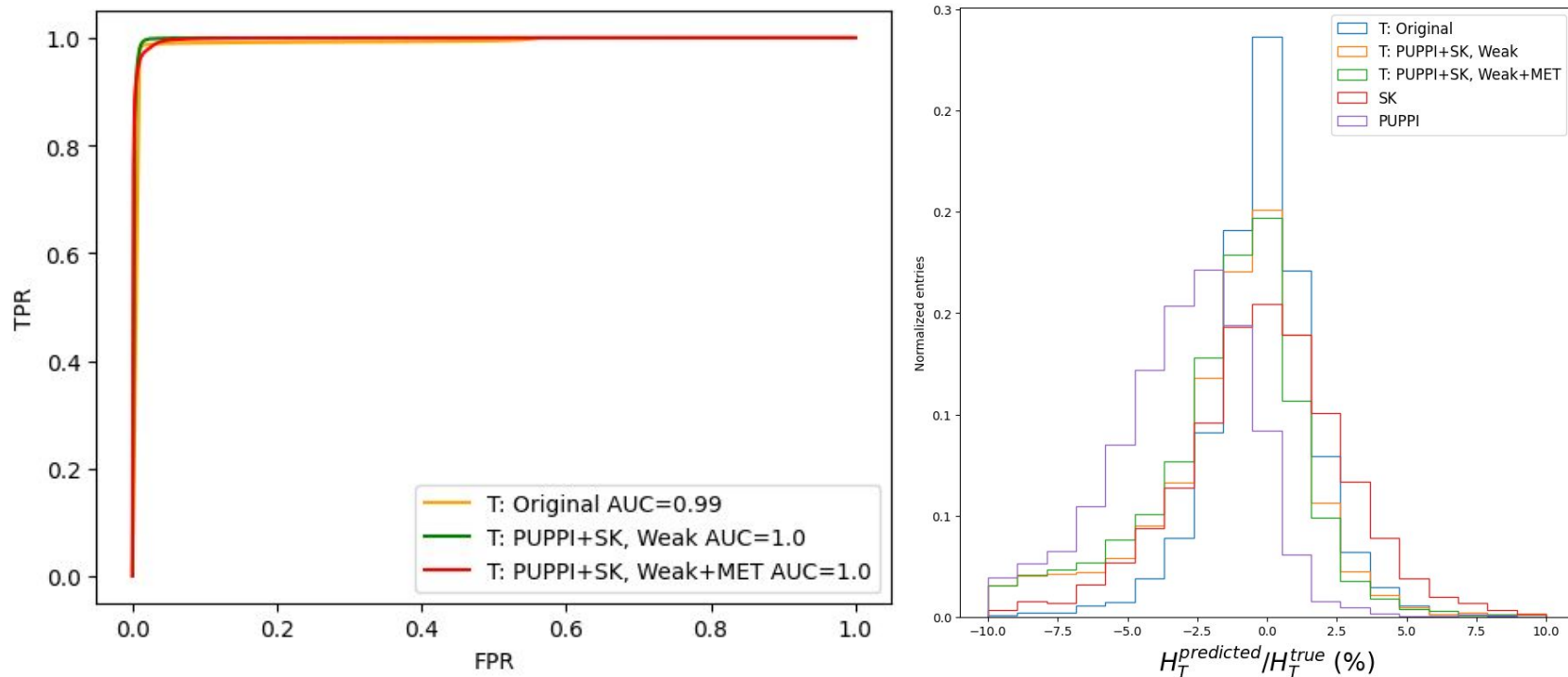
Resampled pileup particles

Toy Example

$$c = \frac{\frac{n_{p,pure}}{(n_{p,pure} + n_{np,pure})}}{\frac{n_{p,full}}{(n_{p,full} + n_{np,full})}}$$



Physics Example: High p_T Jets



$$\mathcal{L} = \text{SWD}(x'_p, x_{np}) + \lambda \text{MSE}(E_T(x'_p), E_T(x_{np}))$$

Key Takeaways

Δ TOTAL represents efforts to increase the flexibility without sacrificing performance

Having removed the supervision of event matching, Δ TOTAL only requires matching ensembles of events.



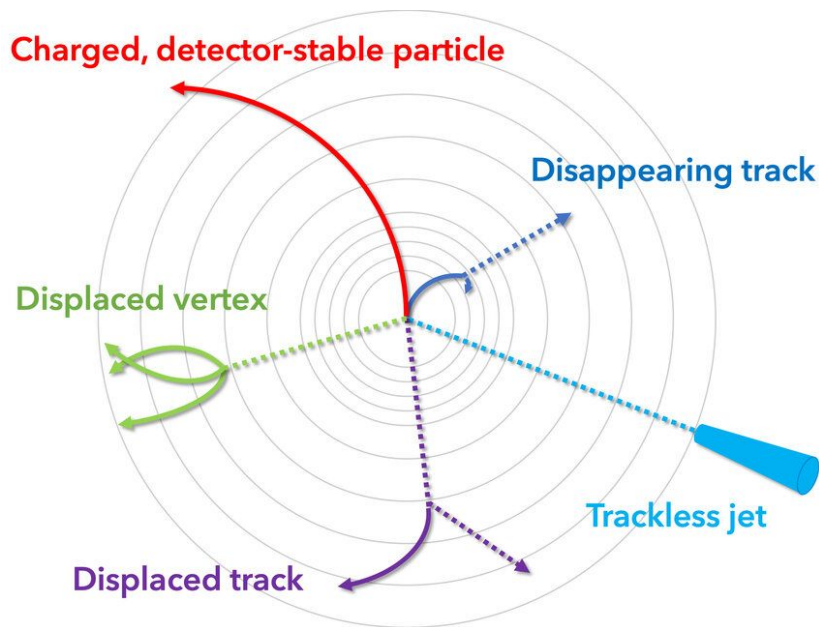
Initial results show promise in pursuing weak supervision

Toy studies show the potential power of Δ TOTAL in moving towards a weakly supervised context. Current physics results show equal performance to leading conventional strategies, but more gains are to be expected.

TOTAL is a completely data-driven pileup mitigation technique

While competing ML methods require particle-level truth and reco matching, TOTAL only requires a match between events with and without pileup.

Applications to LLP (DV) Searches



- ▷ Global Event Description
- ▷ OT vs. BCE Approach
 - Aforementioned advantage is filtered down to constituent level
 - Handles particle-particle correlations better
- ▷ Topology-aware methods?

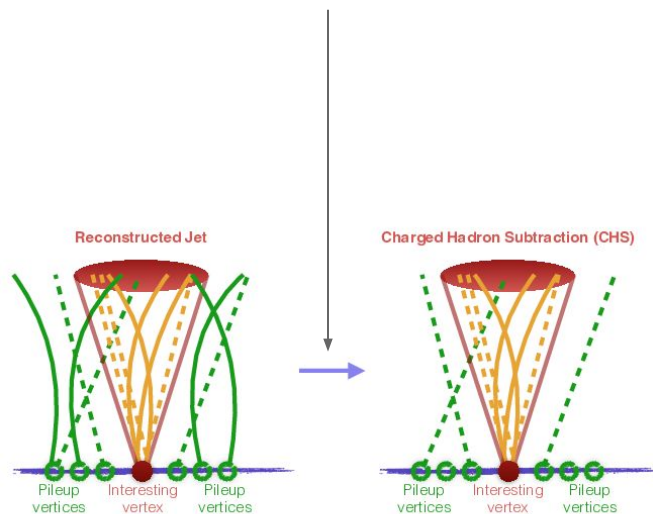


Backup Slides

Charged Hadron Subtraction

$$\text{CVF} = \frac{\sum_{\text{tracks,HS}} p_{\text{T}}^{\text{track}}}{\sum_{\text{tracks,HS}} p_{\text{T}}^{\text{track}} + \sum_{\text{tracks,PU}} p_{\text{T}}^{\text{track}}}$$

- ▷ Benefits
 - Very effective at removing charged pileup due to track information
- ▷ Drawbacks
 - Inapplicable to neutral pileup
- ▷ (arXiv:2012.06271)



TOTAL: Training Optimal Transport with Attention Learning

$$\mathcal{L} = \text{SWD}(x'_p, x_{np}) + \lambda \text{MSE}(p_T^{\text{miss}}(x'_p), p_T^{\text{miss}}(x_{np}))$$

TOTAL: Training Optimal Transport with Attention Learning

$$\mathcal{L} = \text{SWD}(x'_p, x_{np})$$

- Wasserstein distance (WD): Finds the transport function that keeps hard scattering particles and removes those from simultaneous vertices
- Sliced WD to compensate for poor scaling of computational costs of calculating WD at high dimensions

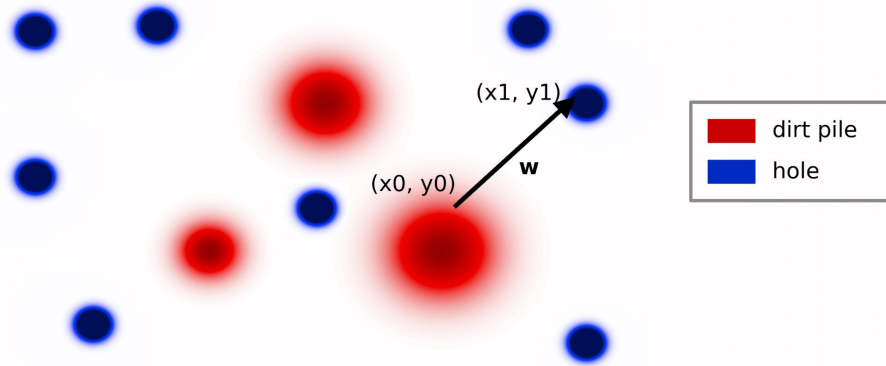
TOTAL: Training Optimal Transport with Attention Learning

$\mathcal{L} =$

- Mean Square Error of missing p_T
- Lambda denotes the importance of the MET regularization

$$+ \lambda \text{MSE}(p_T^{\text{miss}}(x'_p), p_T^{\text{miss}}(x_{np}))$$

Wasserstein Metric



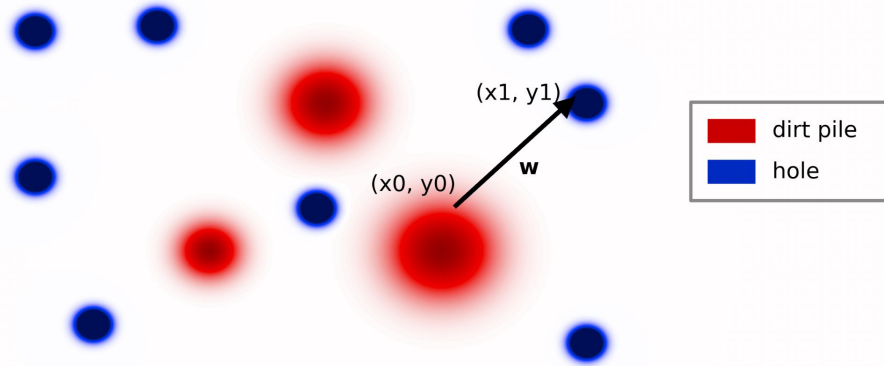
- ▷ Assumption: Total volume of the holes = total volume of the dirt piles
- ▷ Piles as the probability density function of P and holes as the probability density function of Q
- ▷ Per unit transportation cost:

$$C(x_0, y_0, x_1, y_1) = (x_0 - x_1)^2 + (y_0 - y_1)^2$$

- ▷ Transportation Plan:

$$T(x_0, y_0, x_1, y_1) = w$$

Wasserstein Metric



$$\int \int T(x_0, y_0, x, y) dx dy = p(x_0, y_0)$$

$$\int \int T(x, y, x_1, y_1) dx dy = q(x_1, y_1)$$

$$\text{Total Cost} = \int \int \int \int C(x_0, y_0, x_1, y_1) \cdot T(x_0, y_0, x_1, y_1) dx_0 dy_0 dx_1 dy_1$$

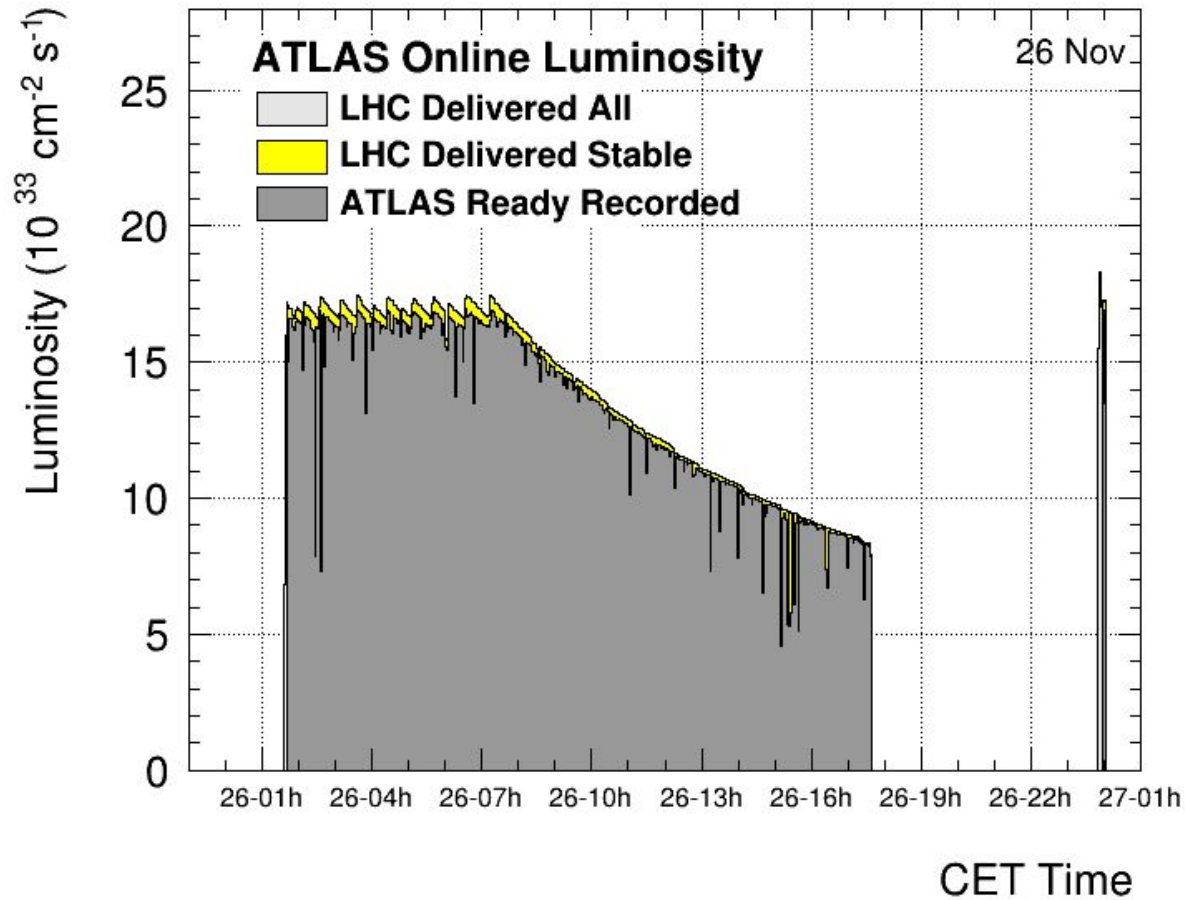
TOTAL: Training Optimal Transport with Attention Learning

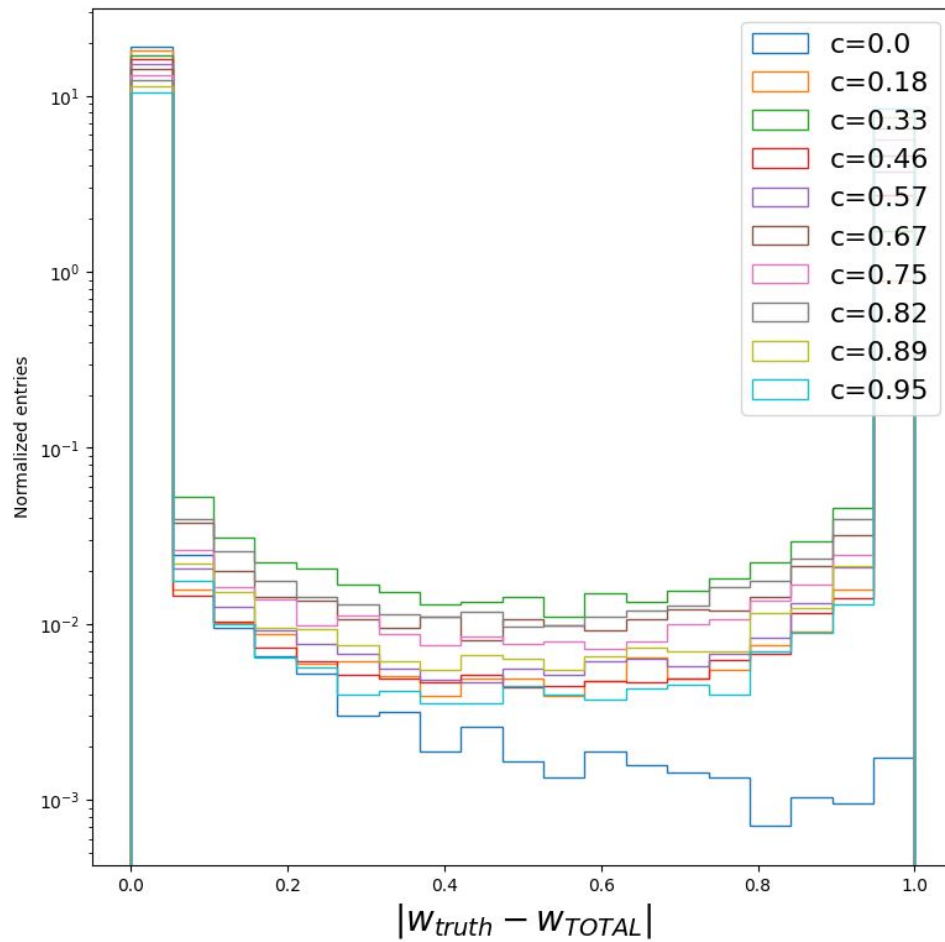
$$\mathcal{L} = \text{SWD}(x'_p, x_{np}) + \lambda \text{MSE}(p_T^{\text{miss}}(x'_p), p_T^{\text{miss}}(x_{np}))$$



Modification for jet-based dataset (PUMML)

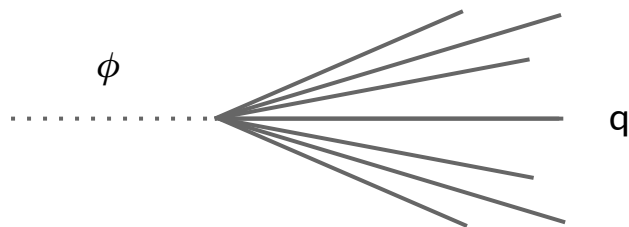
$$\mathcal{L} = \text{SWD}(x'_p, x_{np}) + \lambda \text{MSE}(E_T(x'_p), E_T(x_{np}))$$



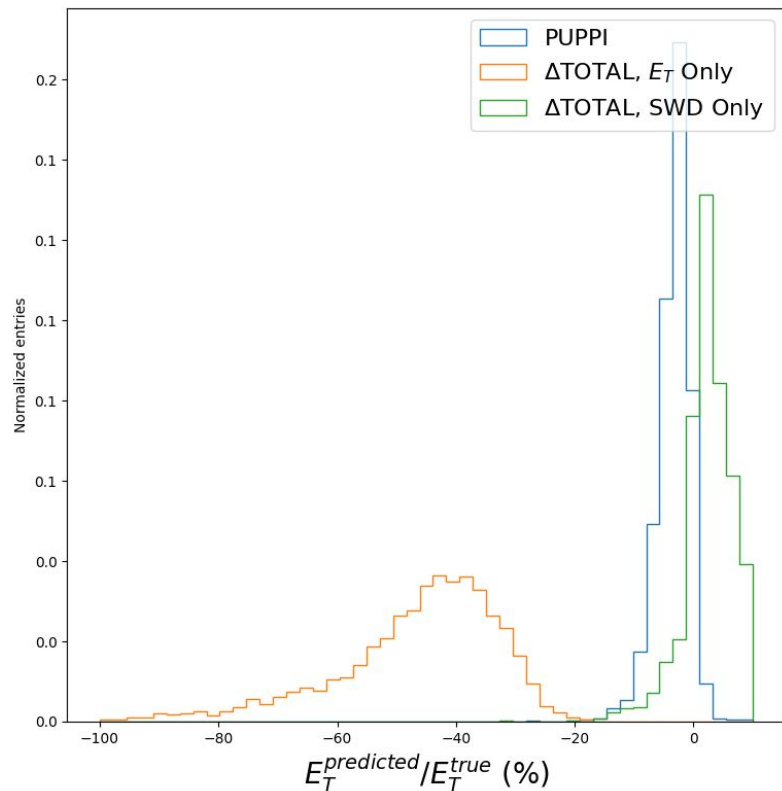
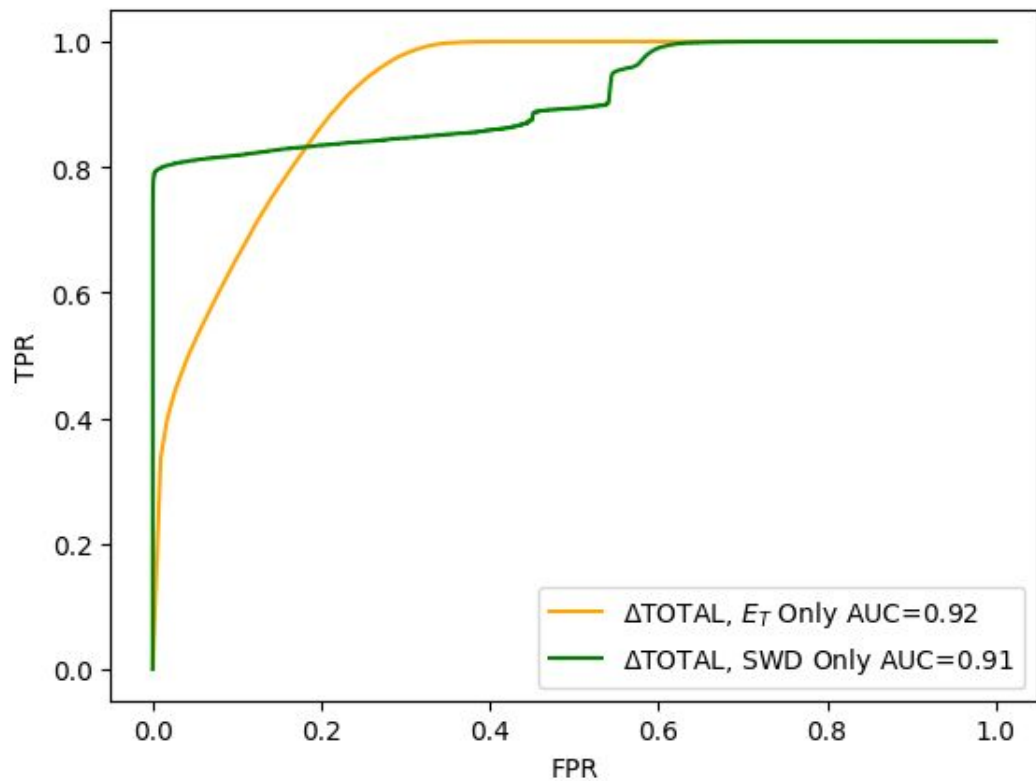


Toy Example: Contamination Ratio Test

Physics Example: High p_T Jets



- ▷ PUMML Dataset:
<https://zenodo.org/records/2652034>
- ▷ Datasets
 - *mH_Mu140: Set PV count, varied scalar mass*
 - **Mu_mH500: Varied PV count, set scalar mass**
 - PV: 130-141



Result with SWD turned off (only E_T)