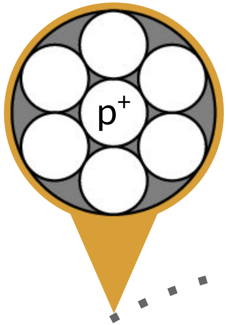
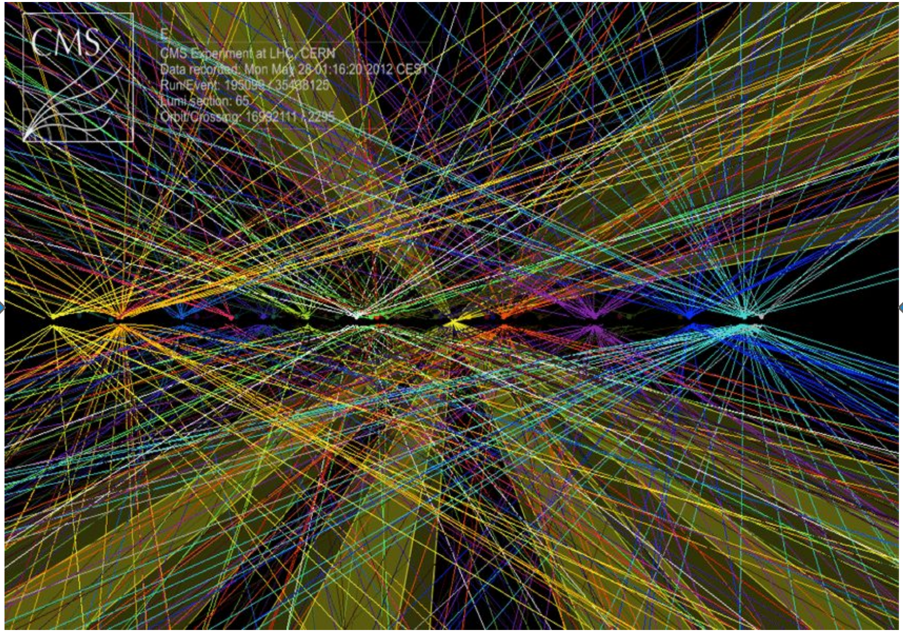
# Weakly Supervised Training for Optimal Transport Pileup Mitigation Strategies at Hadron Colliders

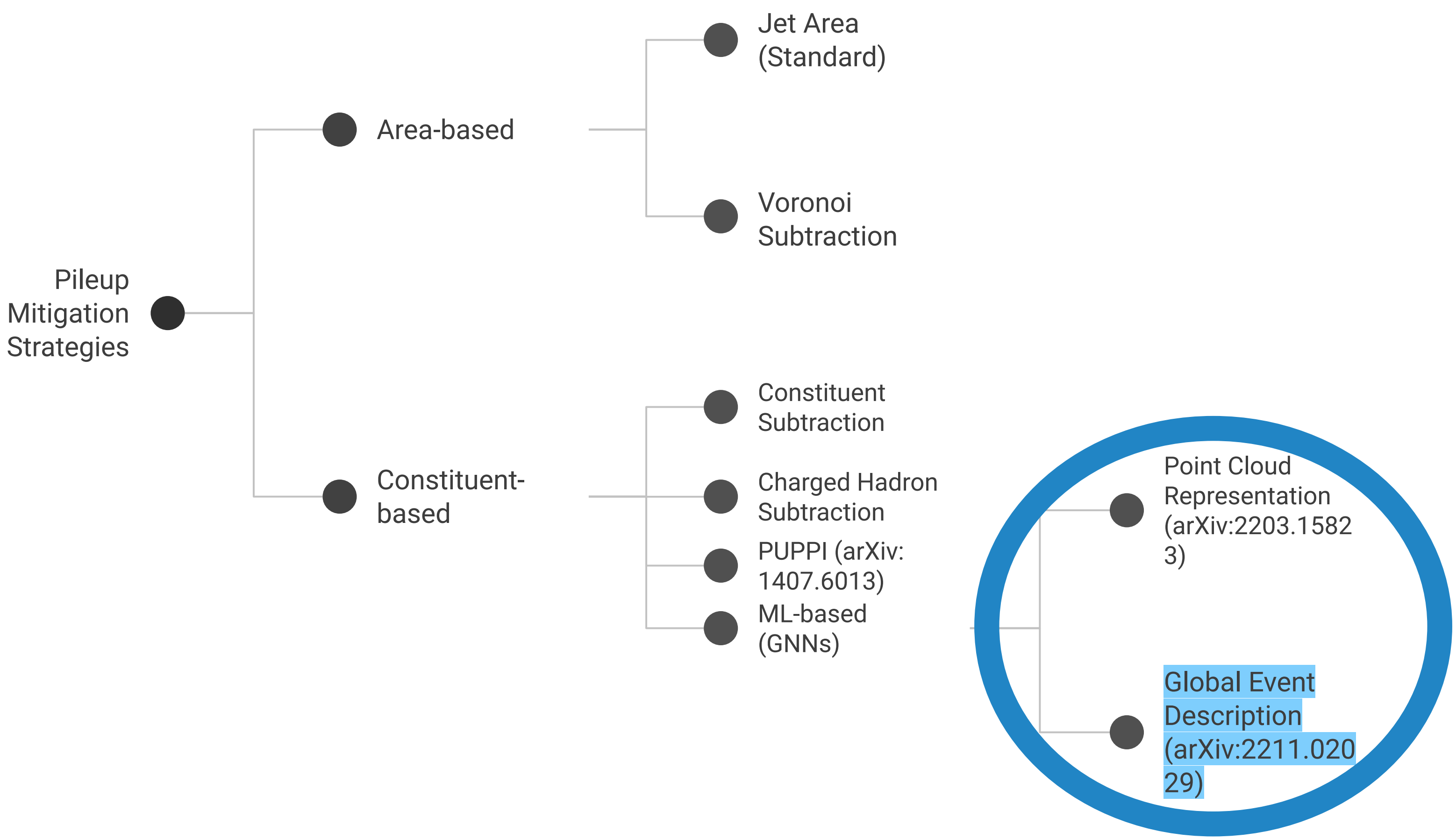
**Nathan Suri**, Vinicius Mikuni

ML4Jets 2023

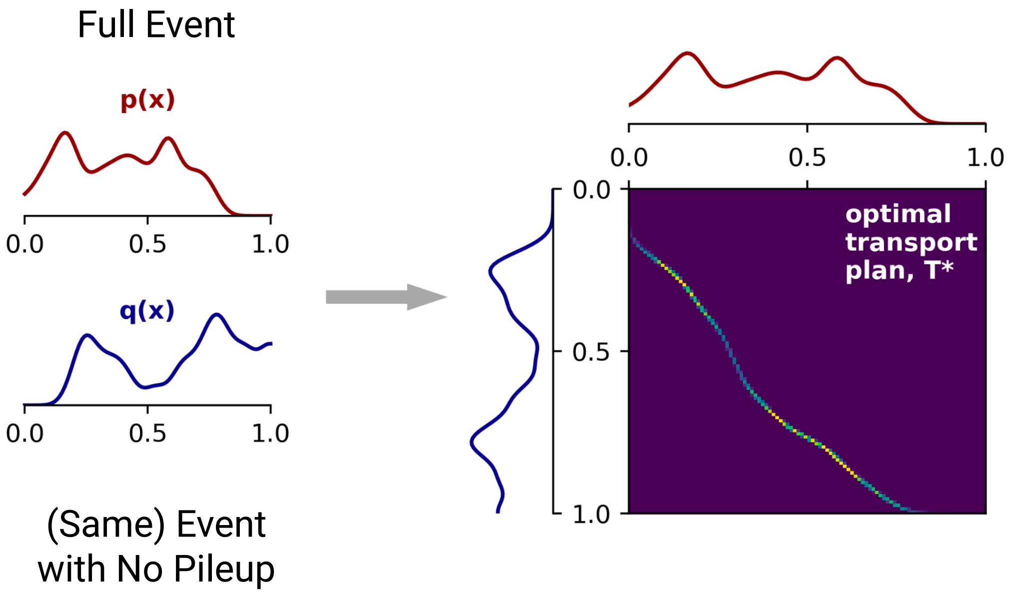



Charged + neutral pileup


Pileup Mitigation Strategies
Area-based
Jet Area (Standard)
Voronoi Subtraction
Constituent-based
Constituent Subtraction
Charged Hadron Subtraction
PUPPI (arXiv: 1407.6013)
ML-based (GNNs)
Point Cloud Representation (arXiv:2203.15823)
Global Event Description (arXiv:2211.02029)

# TOTAL: Training Optimal Transport with Attention Learning




$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\,\mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

# TOTAL: Training Optimal Transport with Attention Learning

$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\,\mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

# TOTAL: Training Optimal Transport with Attention Learning

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

- 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

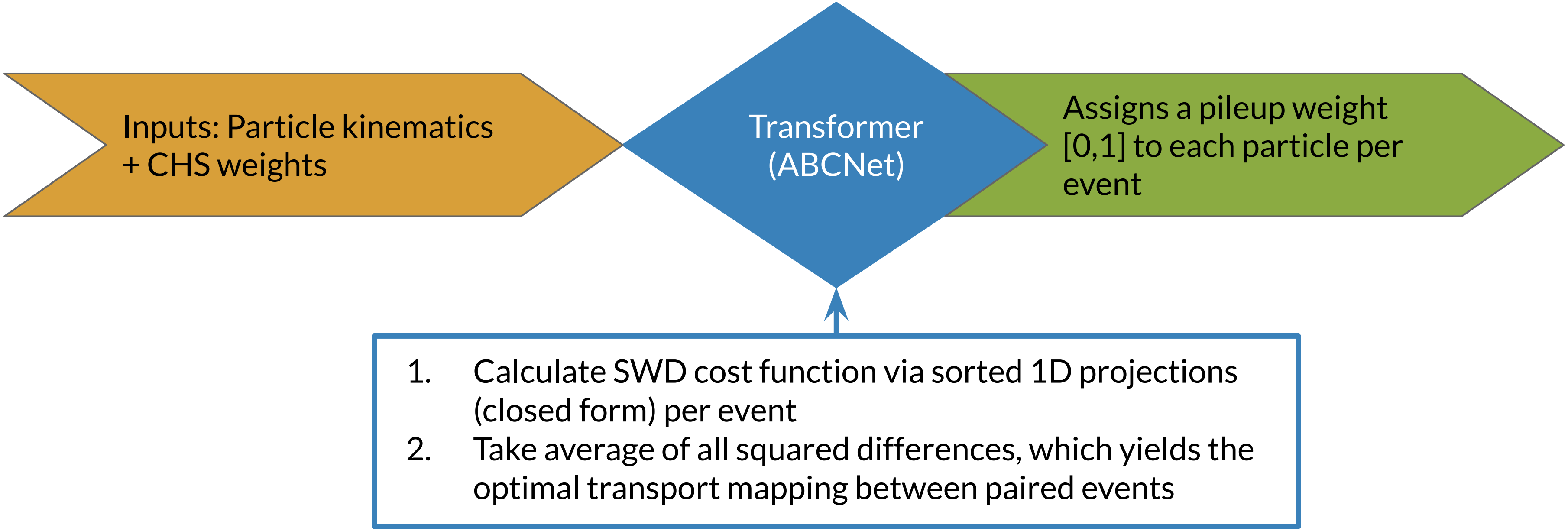
# TOTAL: Training Optimal Transport with Attention Learning

$$\mathcal{L} =$$

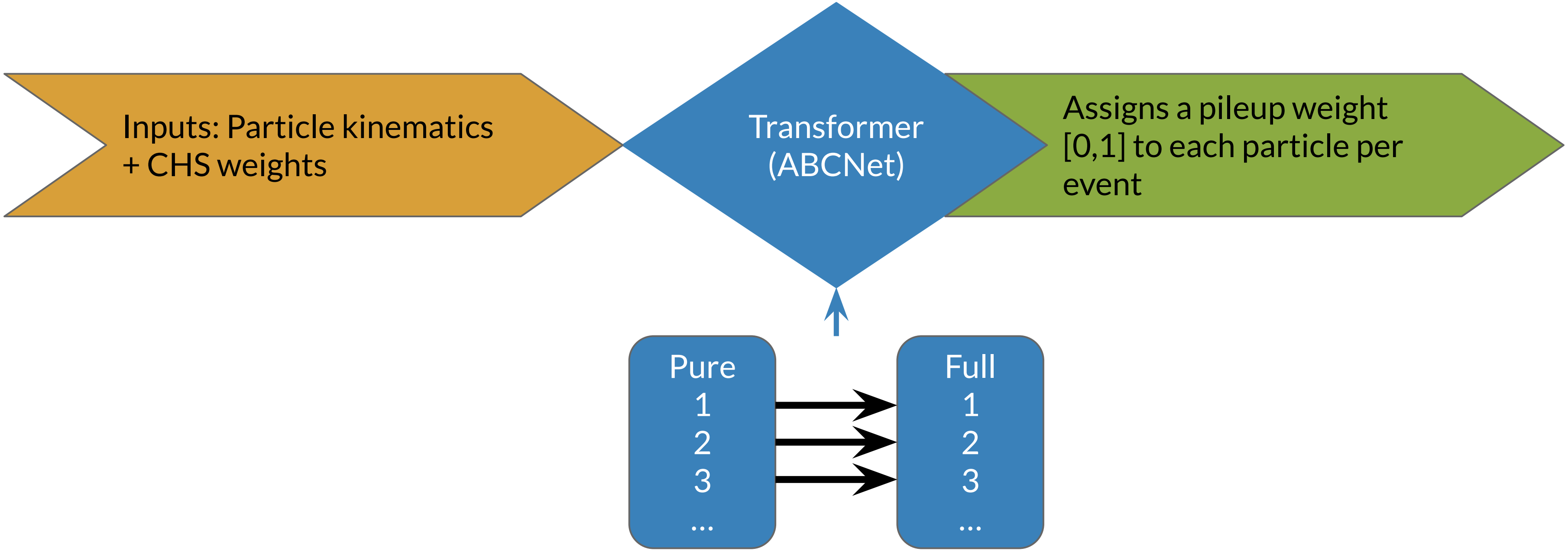
- Mean Square Error of missing $p_T$
- Lambda denotes the importance of the MET regularization

$$+ \lambda \, \mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

# TOTAL: Training Optimal Transport with Attention Learning

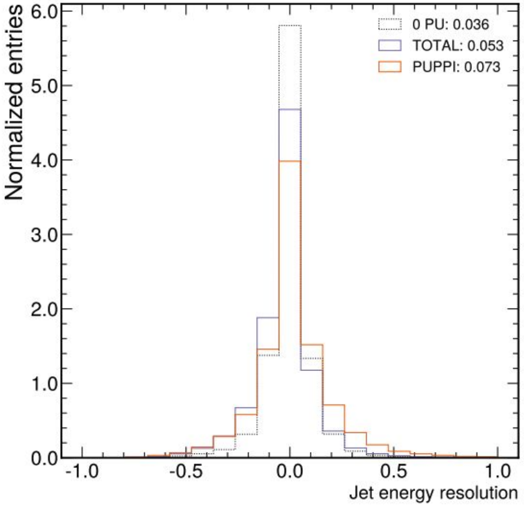


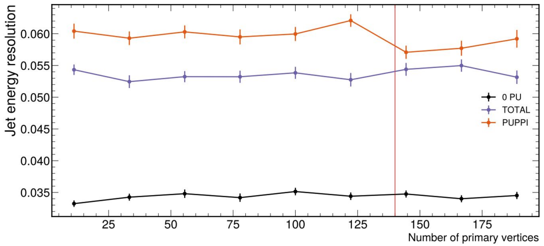
# TOTAL: Training Optimal Transport with Attention Learning



# TOTAL: Training Optimal Transport with Attention Learning

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



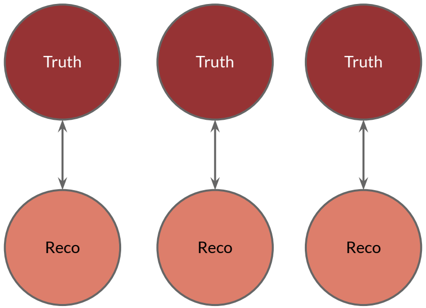



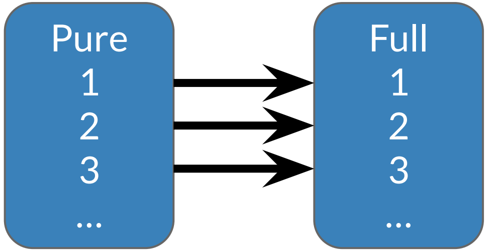

(arXiv:2211.02029)

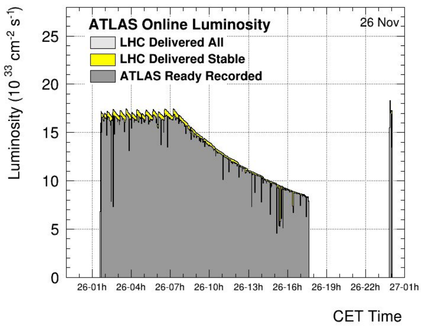
- Requires direct matching of events
- Overall limited due to supervision

# TOTAL: Training Optimal Transport with Attention Learning

| ML Competitors | TOTAL | ΔTOTAL |
| --- | --- | --- |
| Matching between truth and reco at particle level (MC correction) | Matching between pileup events and same event without pileup vertices (data-driven*) | Matching between ensembles of events with different relative pileup densities **(fully data-driven)** |



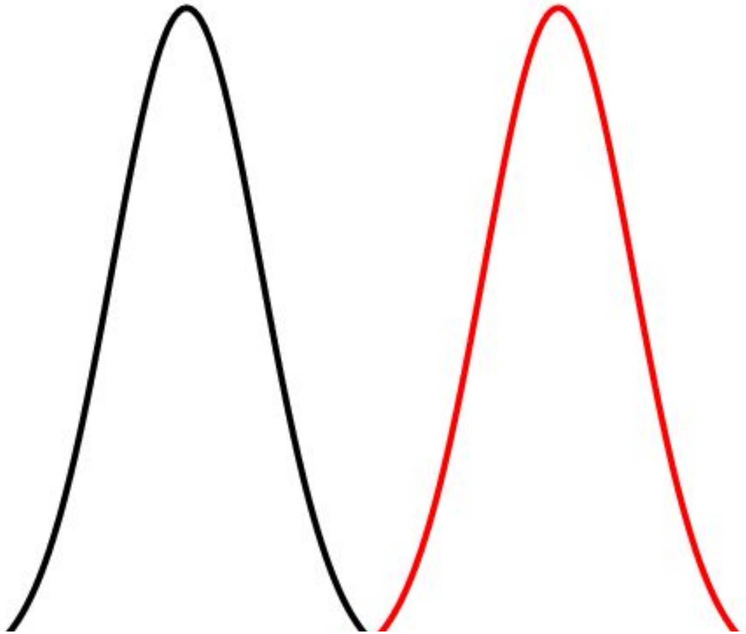
“

*How can we improve TOTAL’s flexibility?*

“

*Reduce supervision!*

# Toy Example

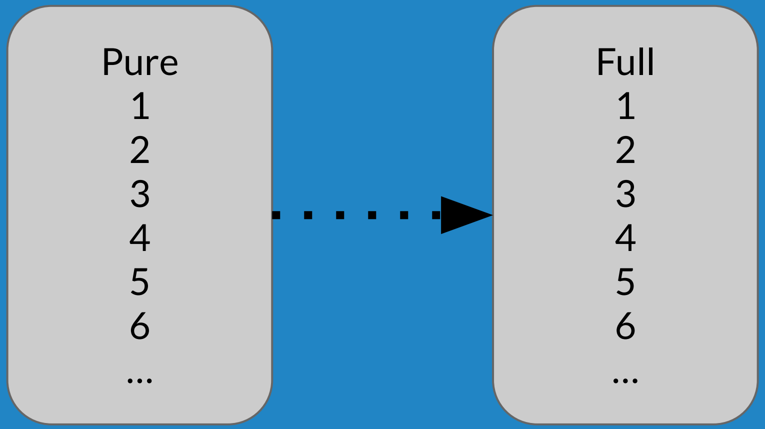


1. For each event, pileup and non-pileup particles are sampled from separate Gaussian distributions
   a. Means are uniformly sampled from a range of values
2. Pure event copied and zero-padded from full event
3. *Mixed Training: Non-pileup data in full event resampled from the originally specified distributions*

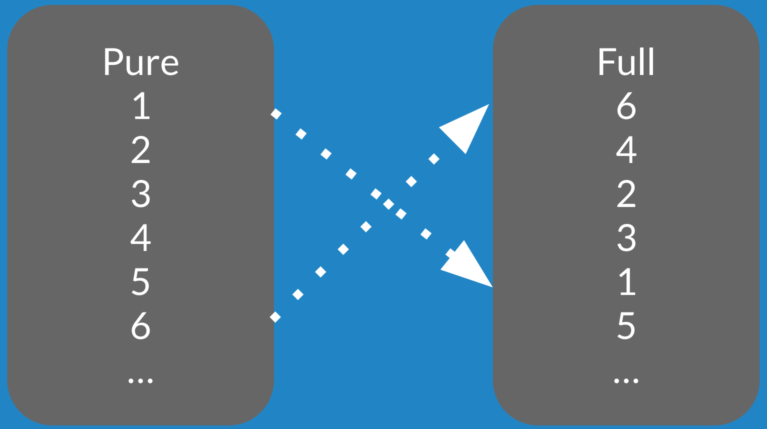
“

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

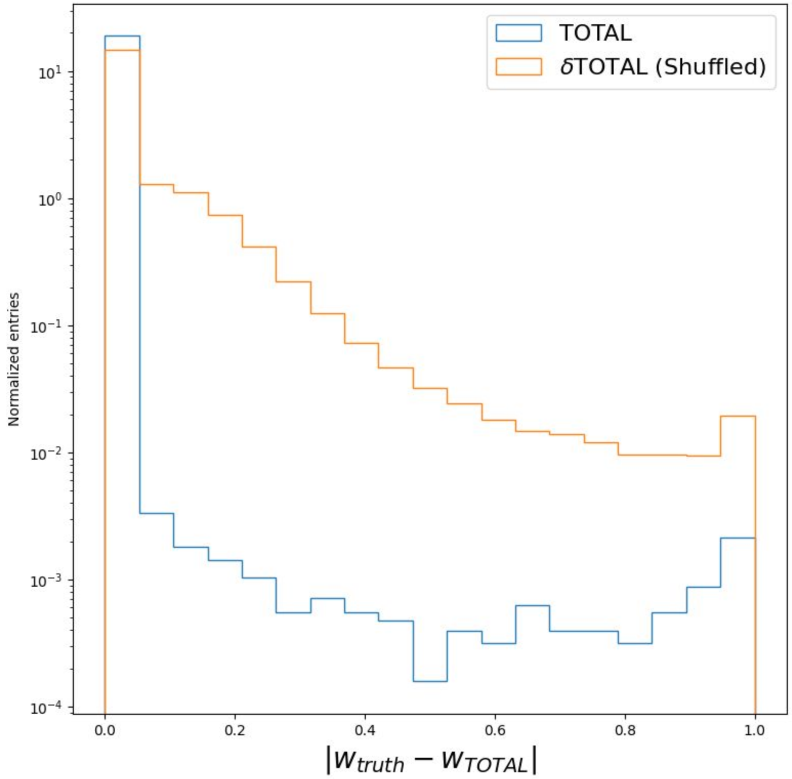
# Original




# Shuffling



# Toy Example



TOTAL
δTOTAL (Shuffled)
Normalized entries
$10^1$
$10^0$
$10^{-1}$
$10^{-2}$
$10^{-3}$
$10^{-4}$
0.0
0.2
0.4
0.6
0.8
1.0
$|w_{truth} - w_{TOTAL}|$

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

# Original

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

Batch
*[event: {particles}]*
*[event: {particles}]*
…

Batch
*[event: {particles}]*
*[event: {particles}]*
…

Batch
*[event: {particles}]*
*[event: {particles}]*
…

# Enhanced

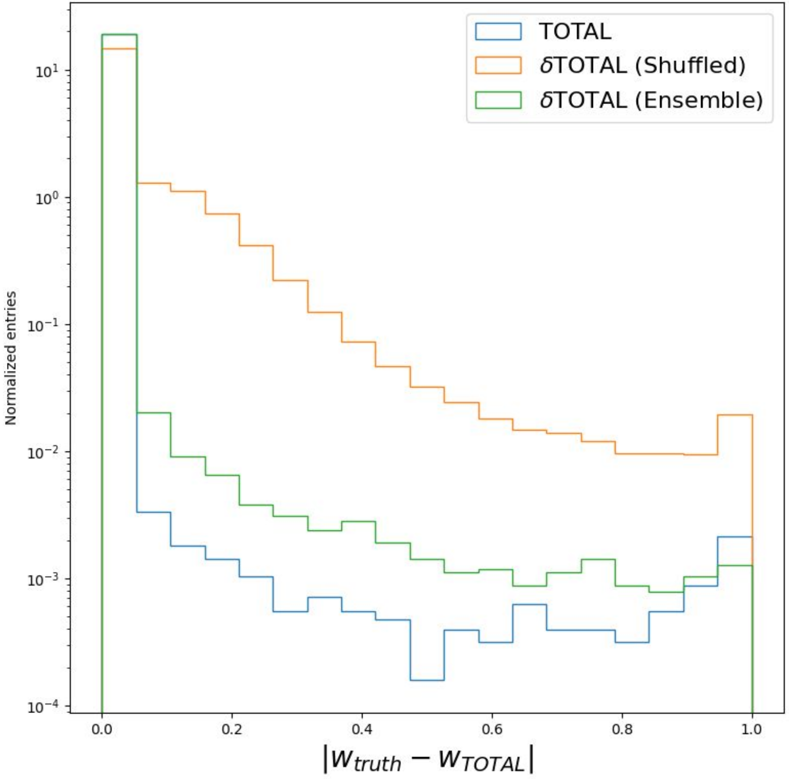
$[(n_{batch} \times n_{particles}) \times n_{features}]$

Batch
*[event ensemble: {particles}]*

Batch
*[event ensemble: {particles}]*

Batch
*[event ensemble: {particles}]*

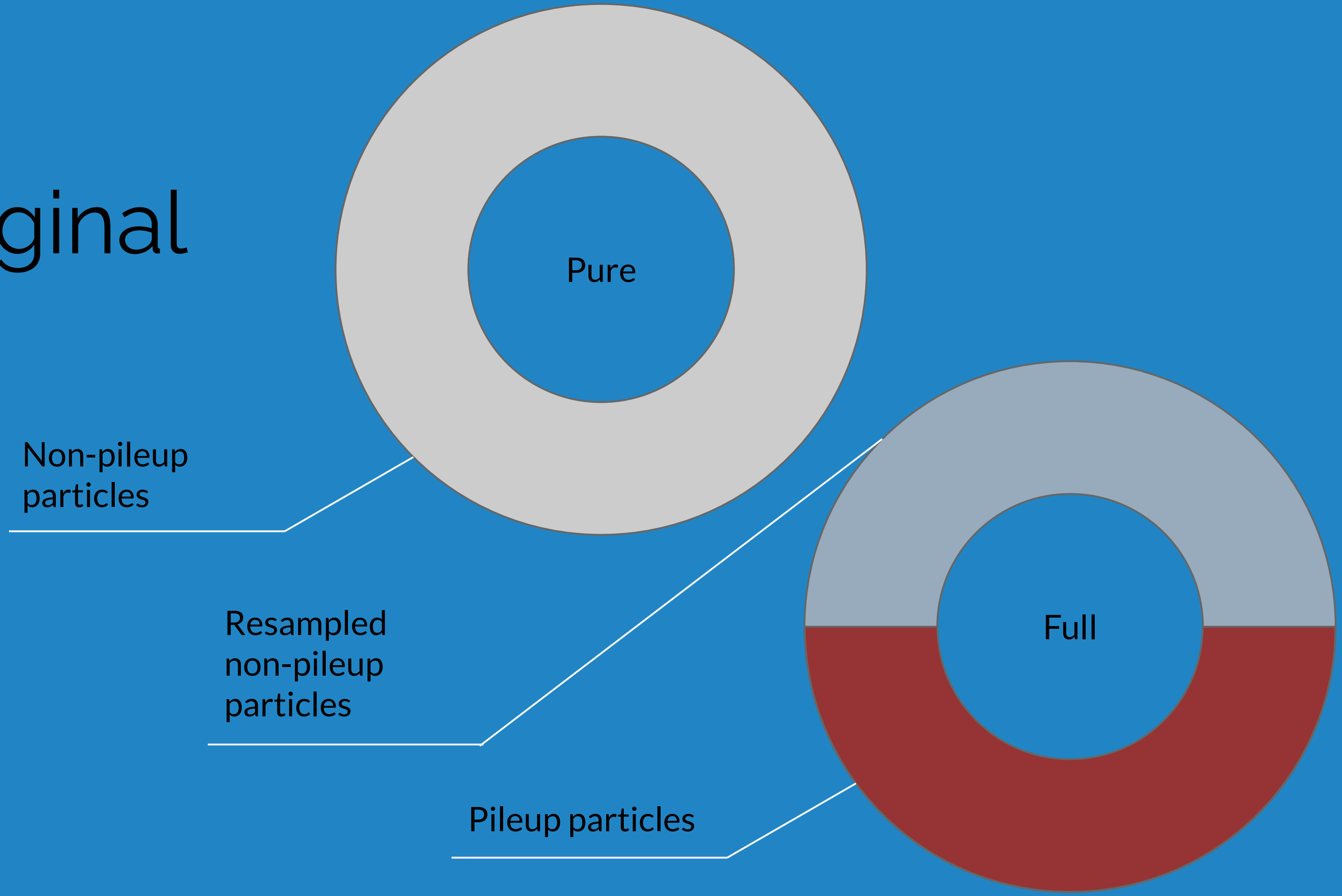
# Toy Example


TOTAL
δTOTAL (Shuffled)
δTOTAL (Ensemble)
Normalized entries
$10^{1}$
$10^{0}$
$10^{-1}$
$10^{-2}$
$10^{-3}$
$10^{-4}$
0.0
0.2
0.4
0.6
0.8
1.0
$|w_{truth} - w_{TOTAL}|$

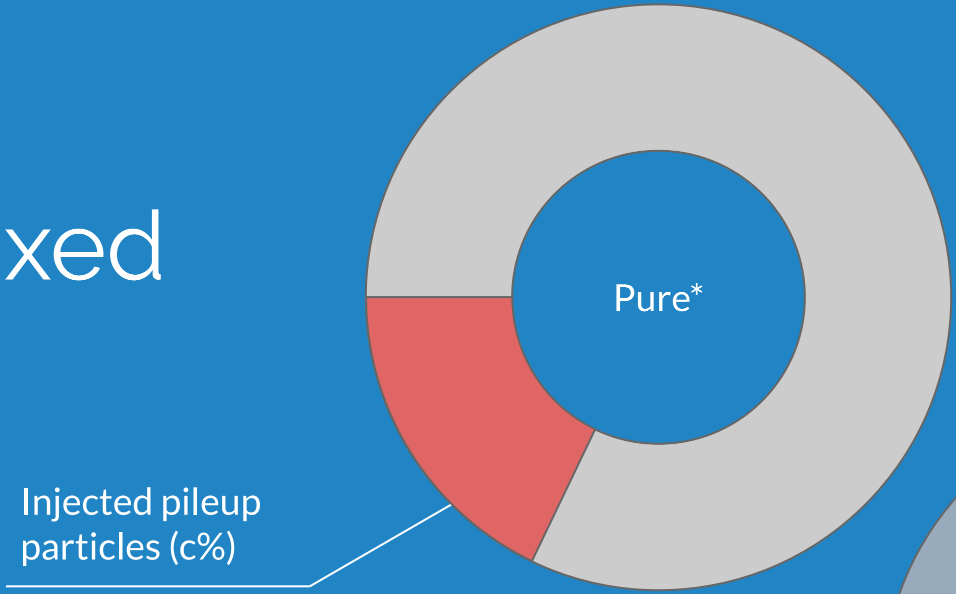
“

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

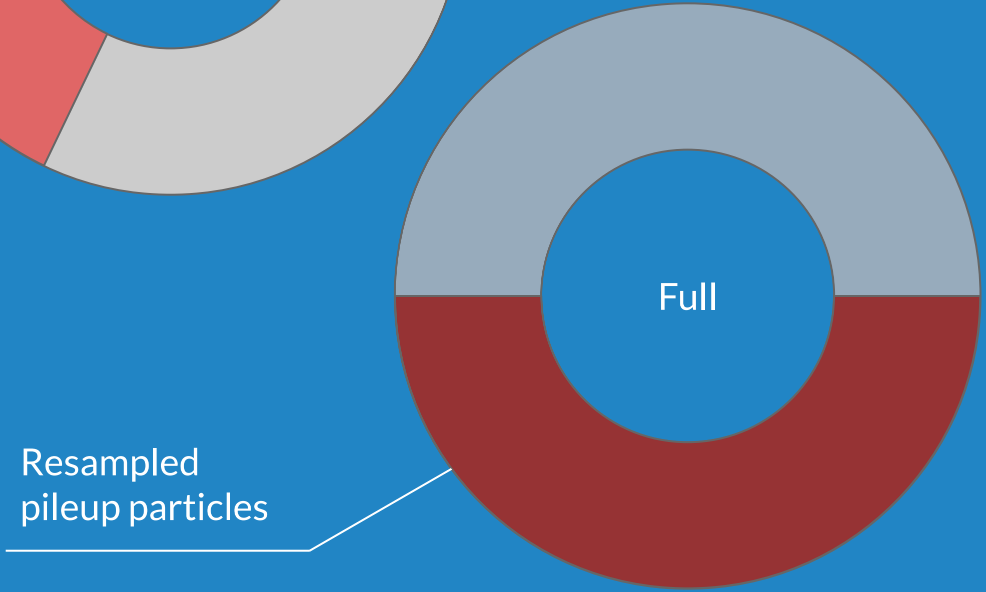
# Original


Pure
Non-pileup particles
Resampled non-pileup particles
Full
Pileup particles

# Mixed

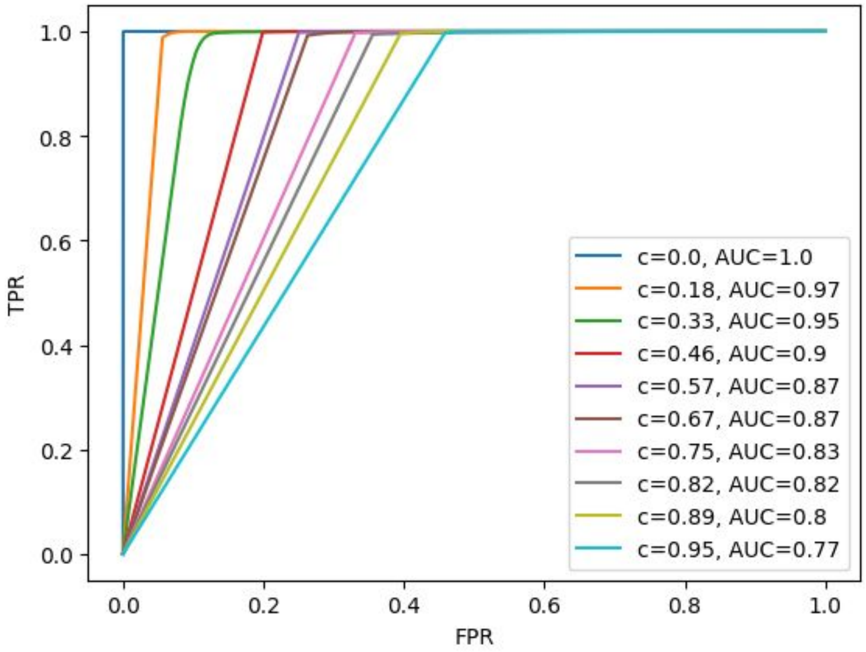


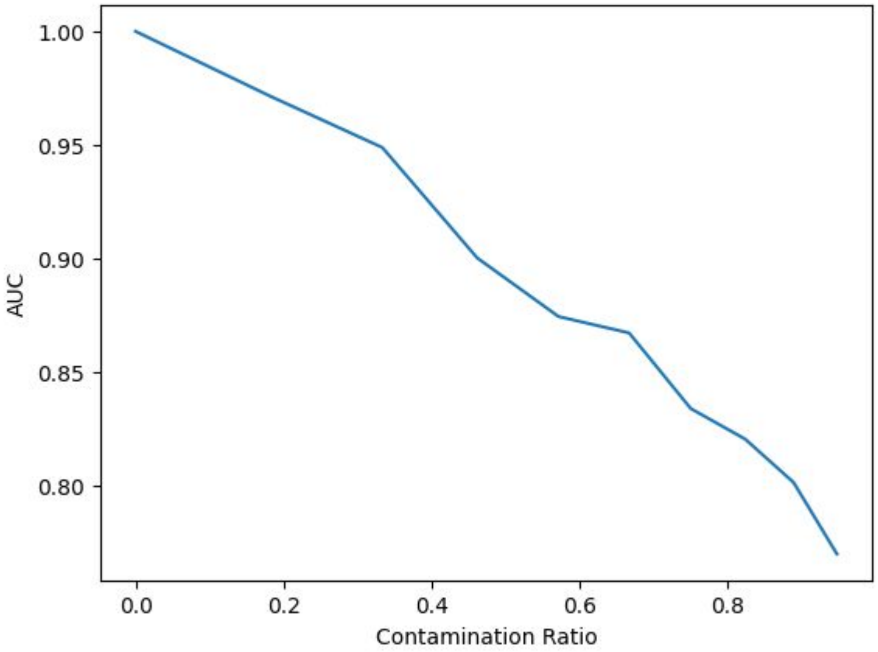

$$c = \frac{\frac{n_{p,pure}}{(n_{p,pure}+n_{np,pure})}}{\frac{n_{p,full}}{(n_{p,full}+n_{np,full})}}$$



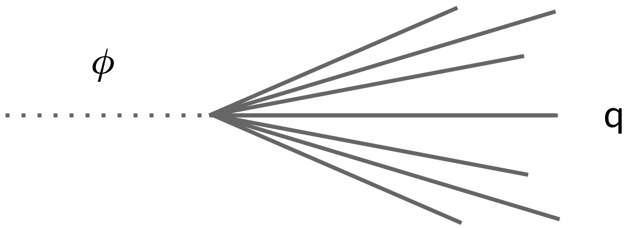
# Toy Example

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



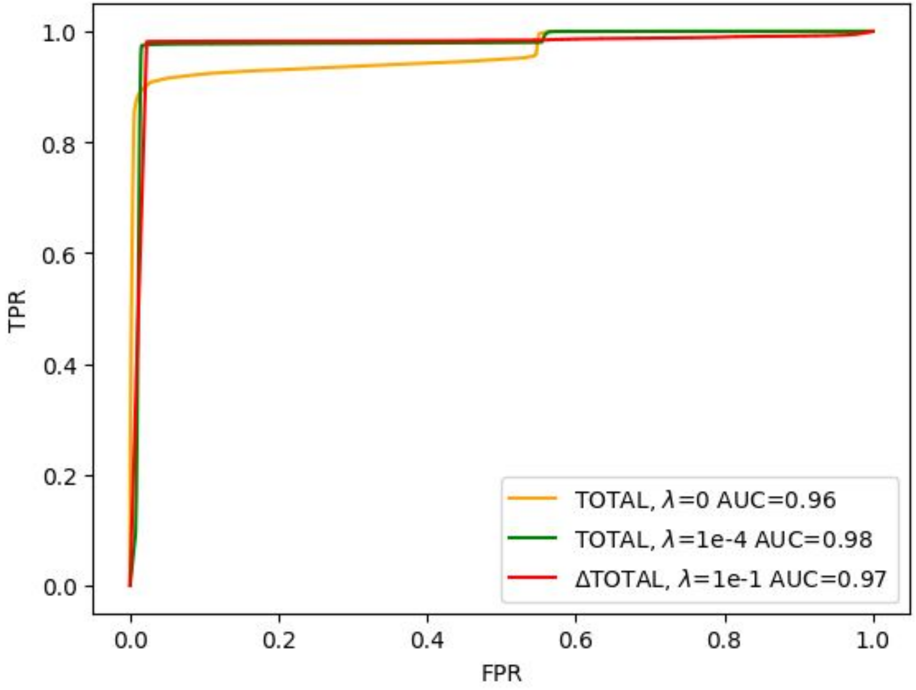



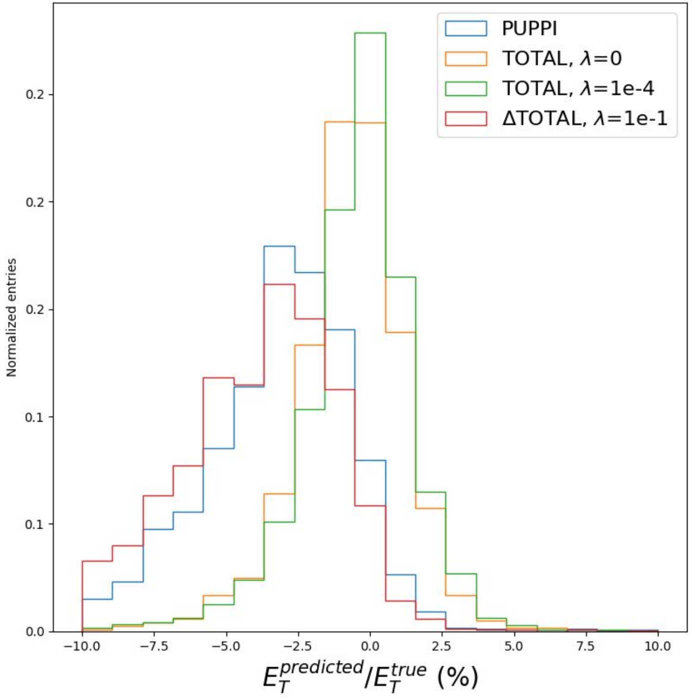
# Physics Example: High $p_T$ Jets




- 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

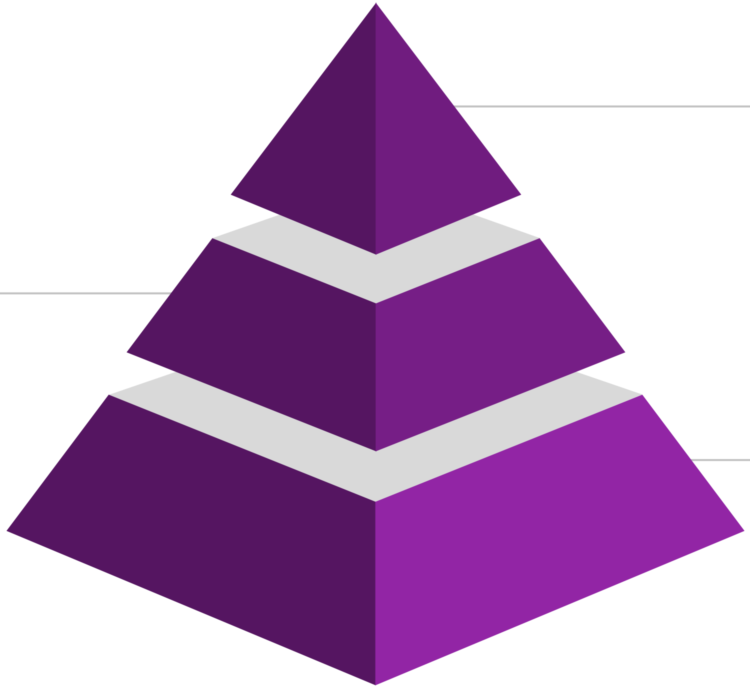
# Physics Example: High $p_T$ Jets




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

# Key Takeaways



1

## 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.

2

## ΔTOTAL represents efforts to increase the flexibility without sacrificing performance

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

3

## Initial results show promise in pursuing weak supervision

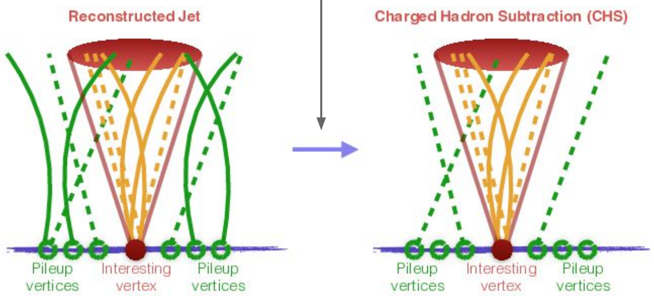
Toy studies show the potential power of ΔTOTAL in moving towards a weakly supervised context. Current physics results show equal performance to leading conventional strategies, but more gains are to be expected.

# Backup Slides

# Charged Hadron Subtraction

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

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

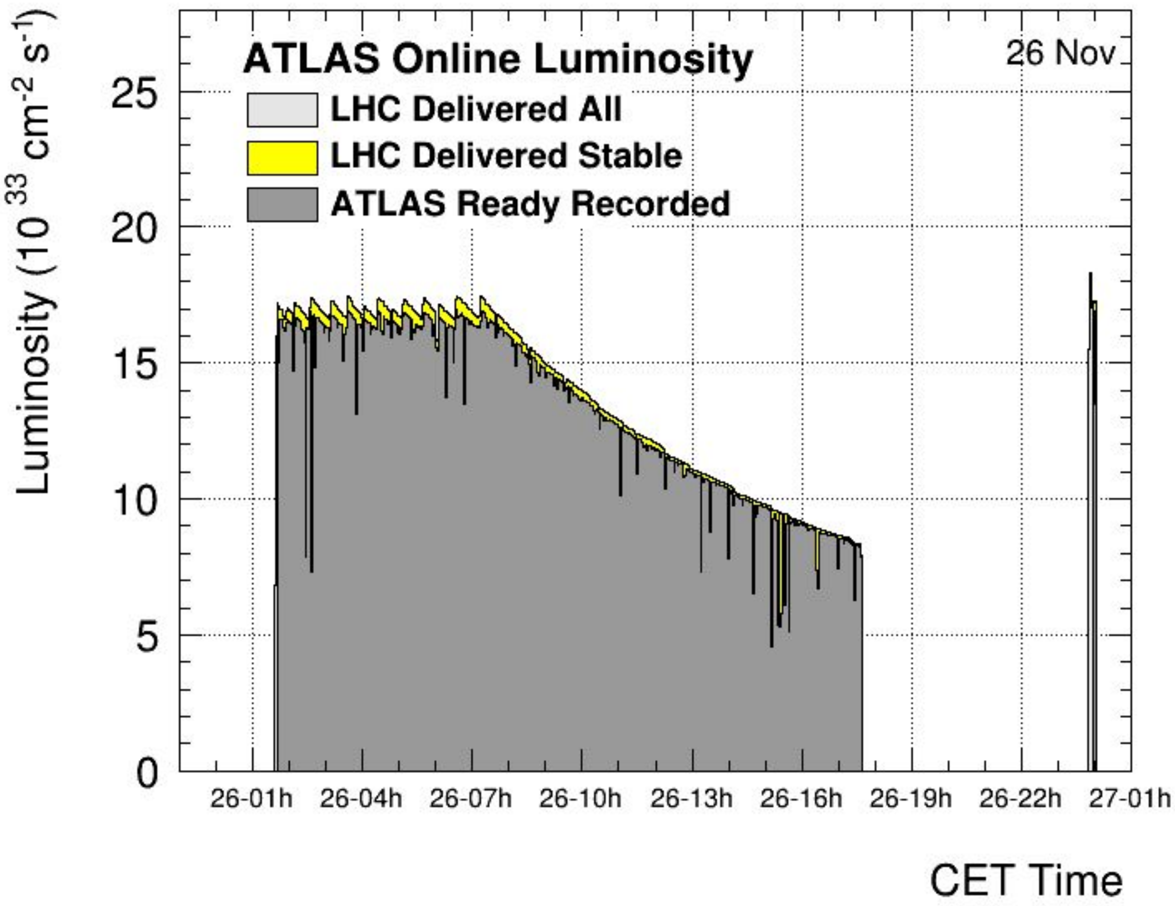


# TOTAL: Training Optimal Transport with Attention Learning

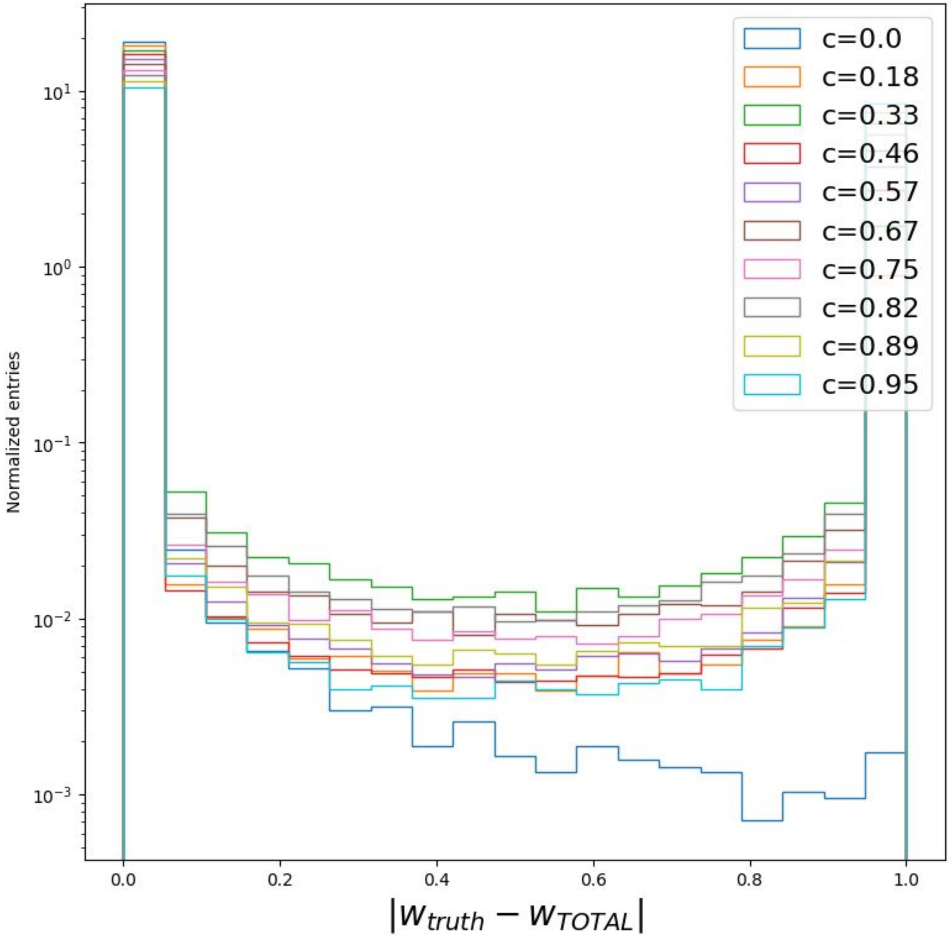
$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\,\mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

↓ Modification for jet-based dataset (PUMML)

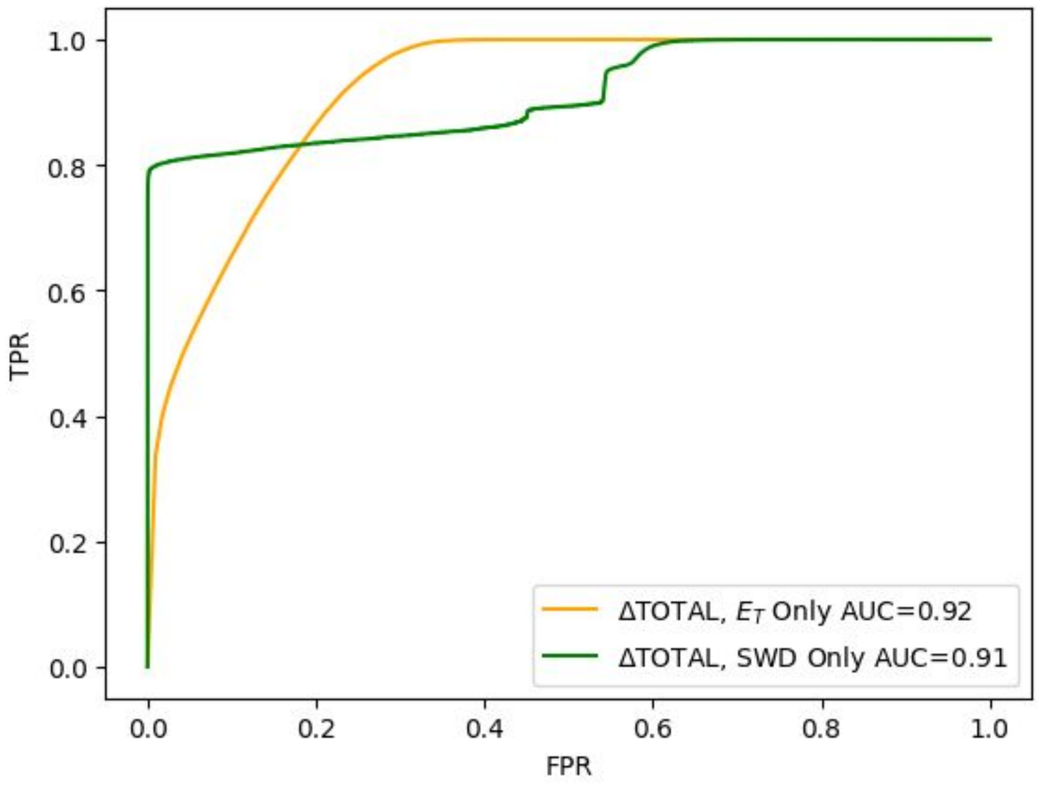
$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda \mathrm{MSE}(E_T(x'_p), E_T(x_{np}))$$

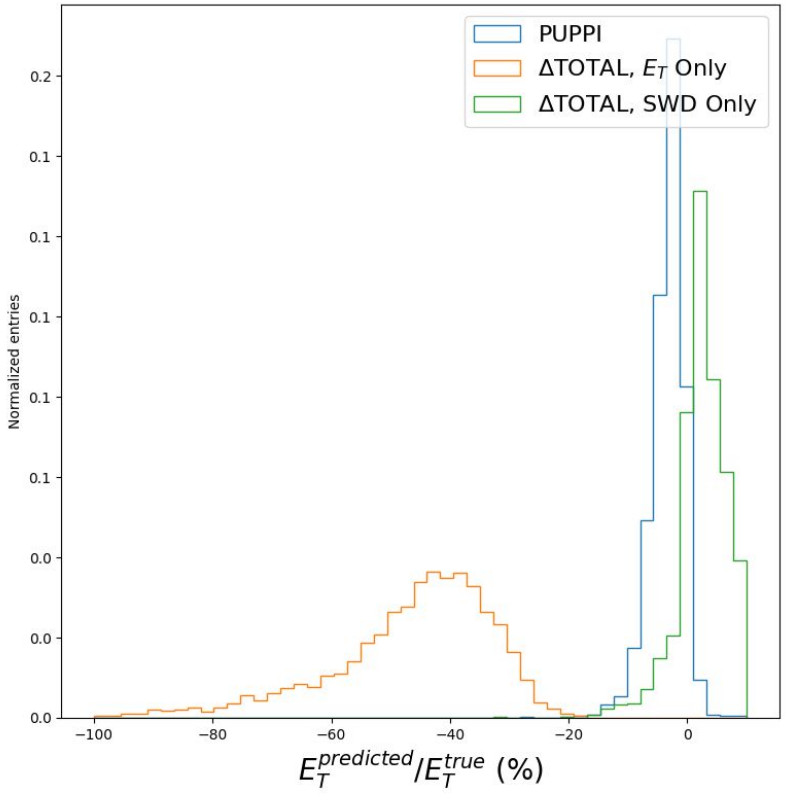



ATLAS Fill Luminosity




Toy Example: Contamination Ratio Test




Result with SWD turned off (only $E_T$)