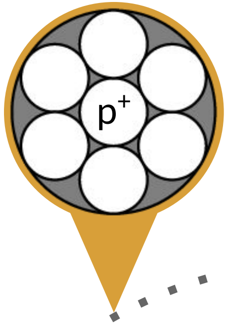
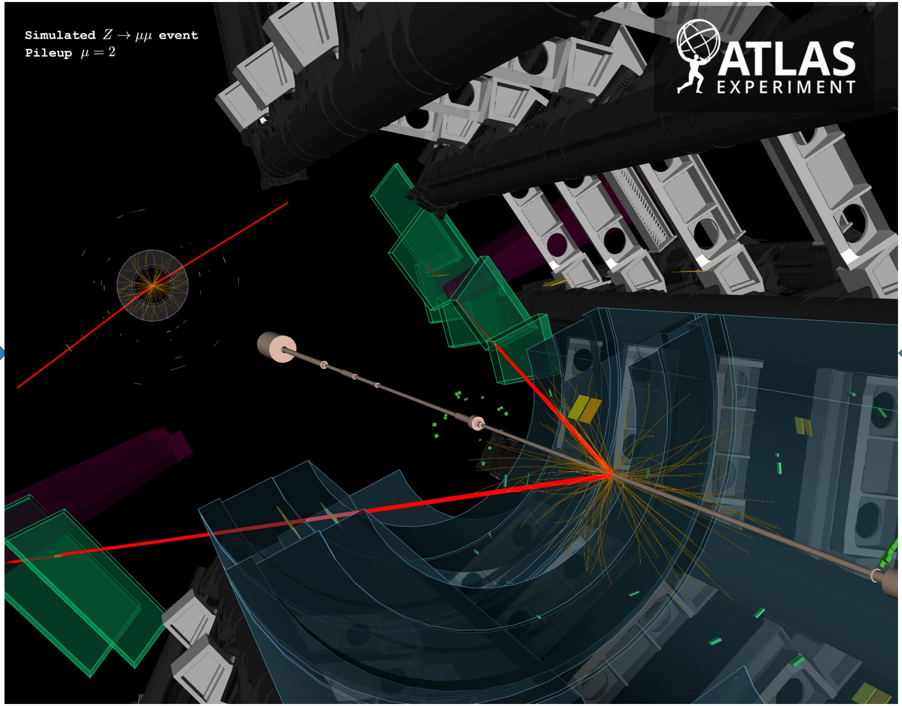
Weakly Supervised Training for Optimal Transport Pileup Mitigation Strategies at Hadron Colliders

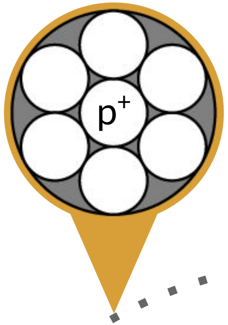
Nathan Suri, Vinicius Mikuni

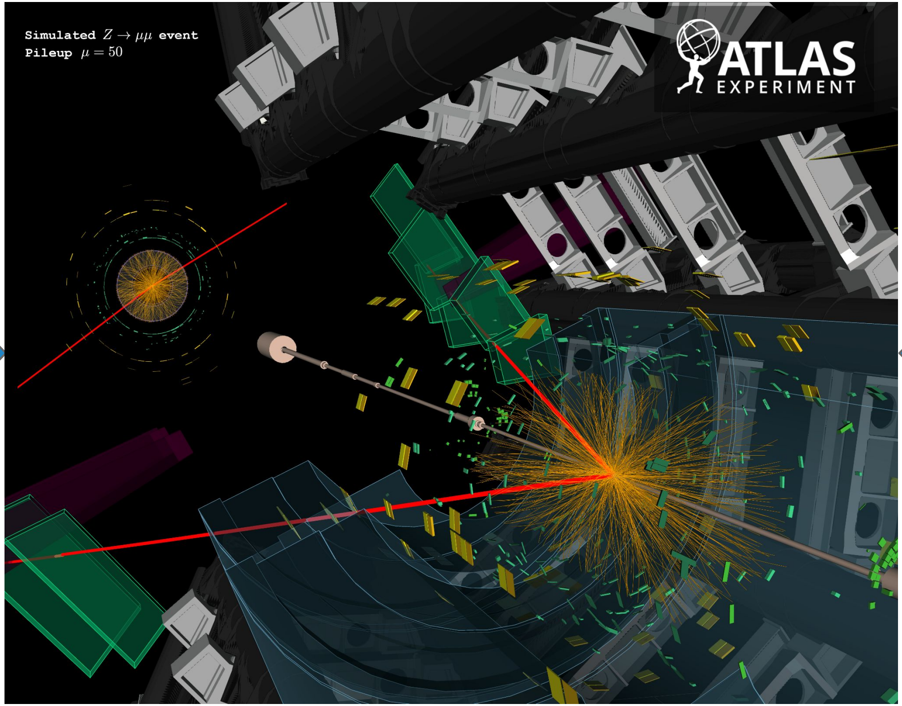
US LUA Meeting 2023



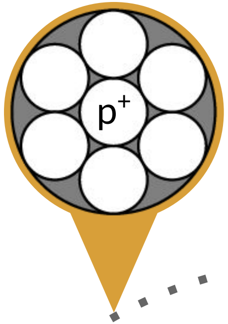


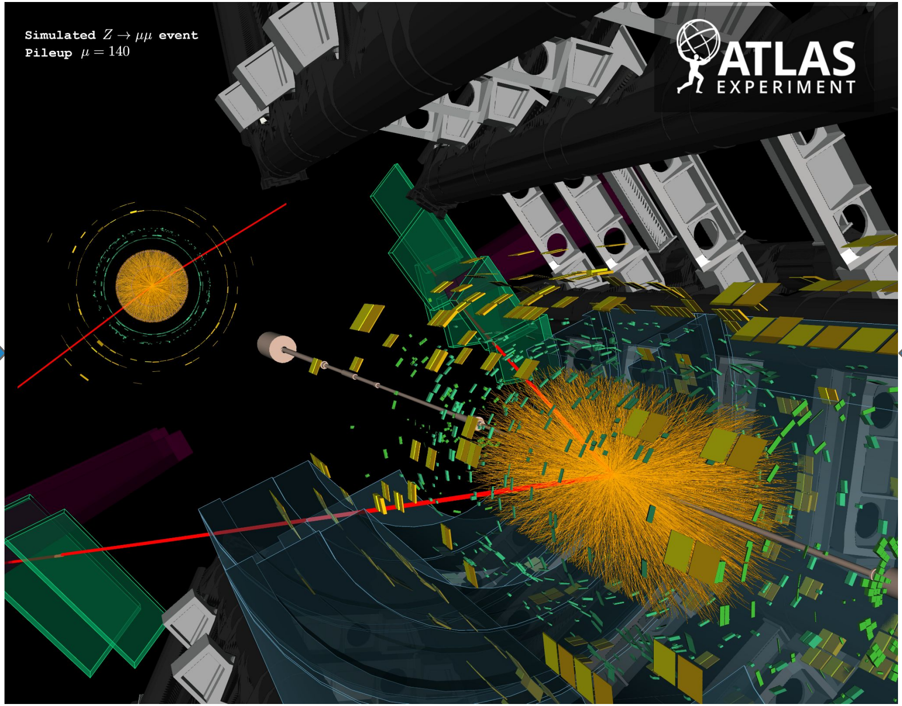
Charged + neutral pileup



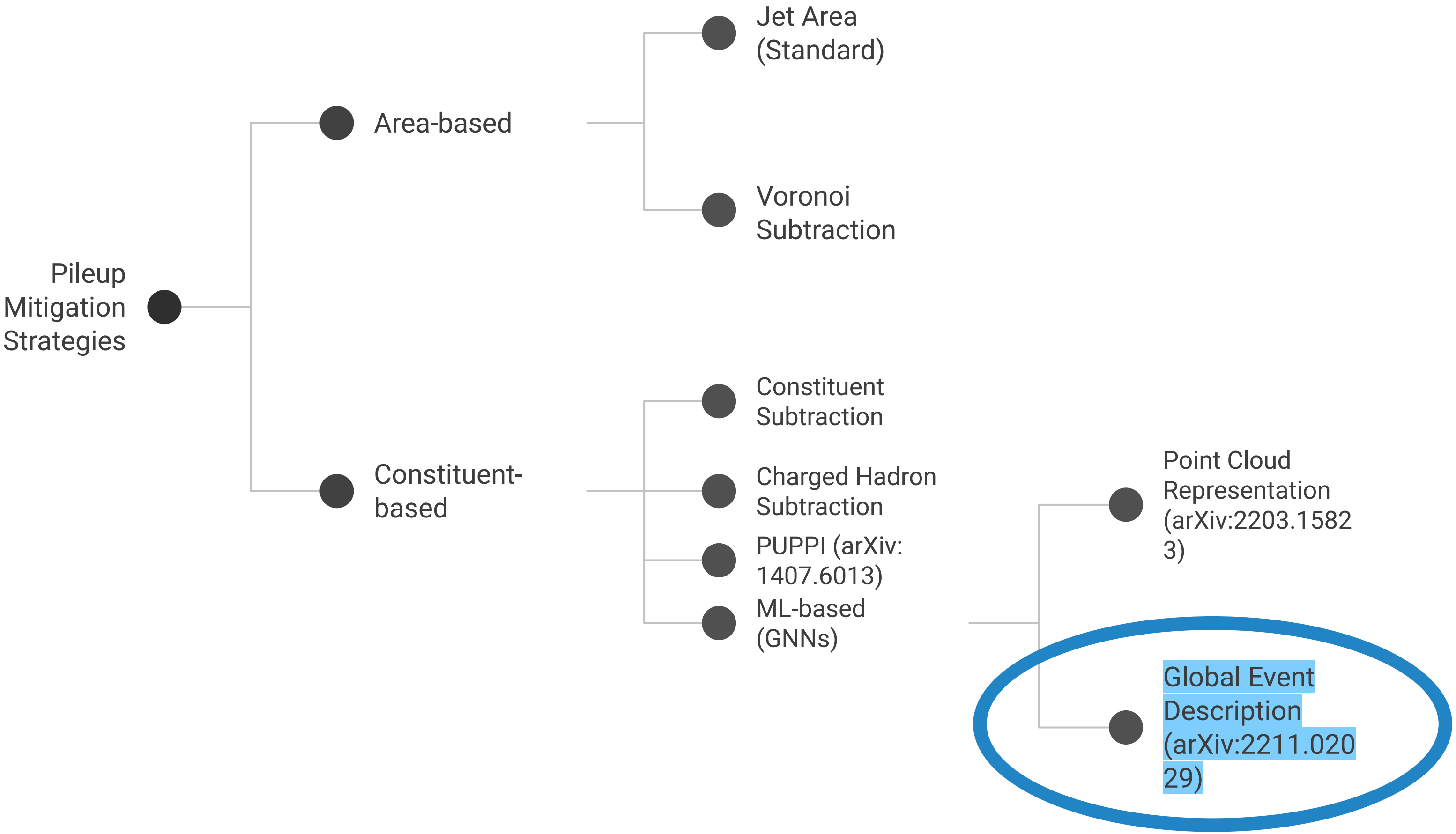


Charged + neutral pileup

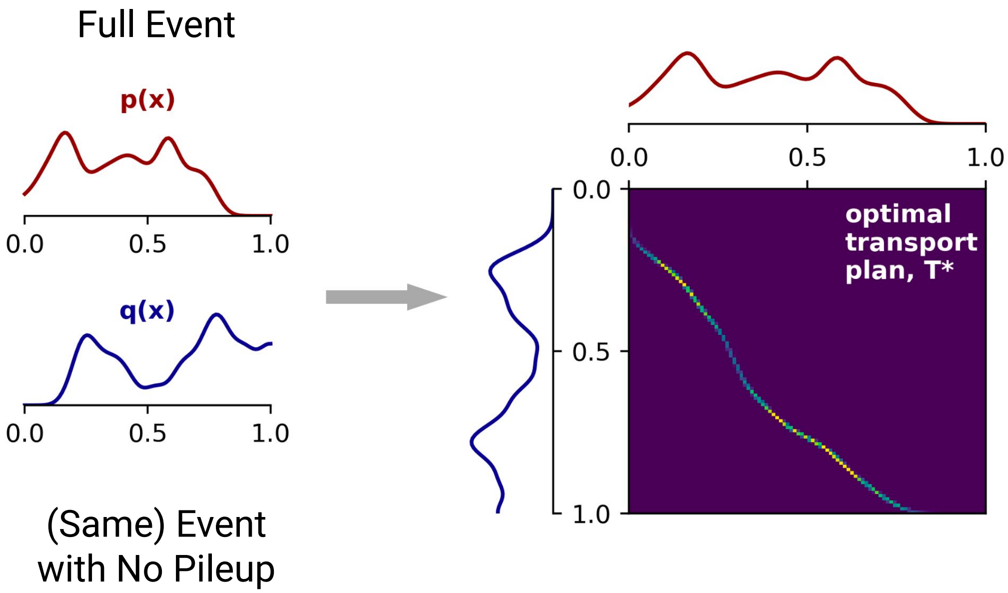




Charged + neutral pileup

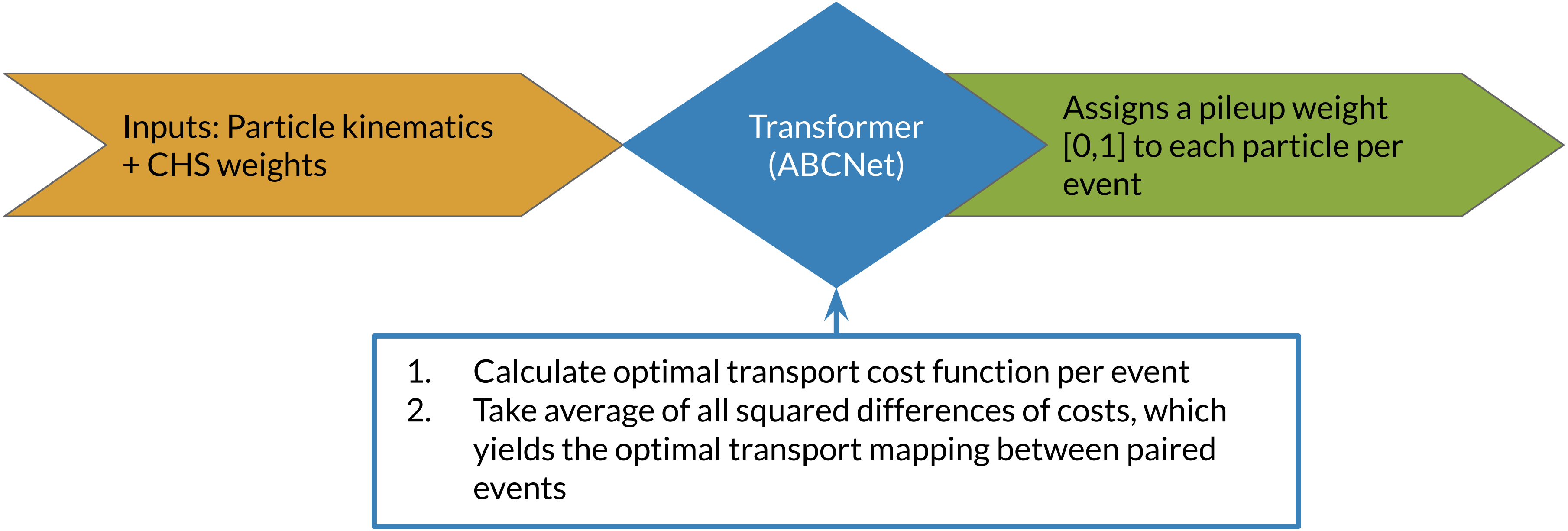

Pileup Mitigation Strategies
Area-based
Jet Area (Standard)
Voronoi Subtraction
Constituent-based
Constituent Subtraction
Charged Hadron Subtraction
PUPPI (arXiv: 1407.6013)
ML-based (GNNs)
Point Cloud Representation (arXiv:2203.15823)
Global Event Description (arXiv:2211.02029)

TOTAL: Training Optimal Transport with Attention Learning

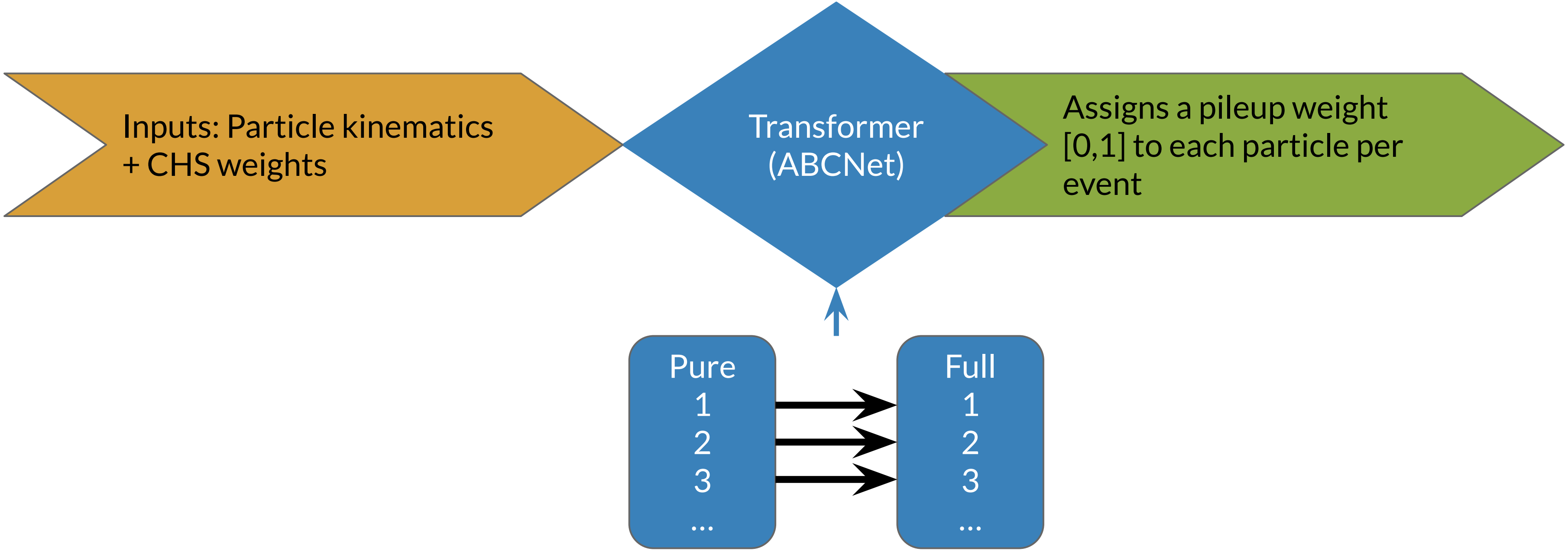


$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\, \mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

TOTAL: Training Optimal Transport with Attention Learning

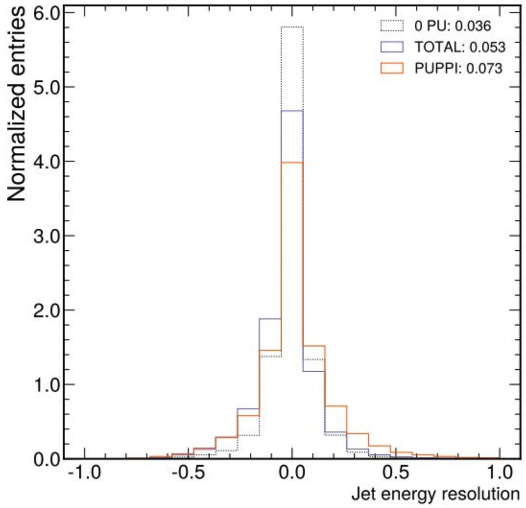


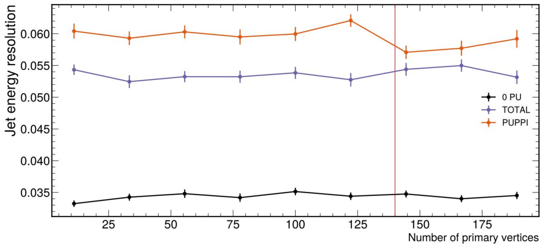
TOTAL: Training Optimal Transport with Attention Learning



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





(arXiv:2211.02029)

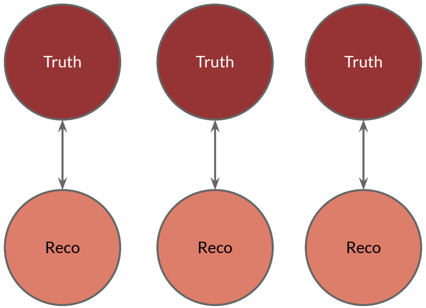
- Requires direct matching of events
- Overall limited due to supervision

TOTAL: Training Optimal Transport with Attention Learning

ML Competitors → TOTAL → ΔTOTAL

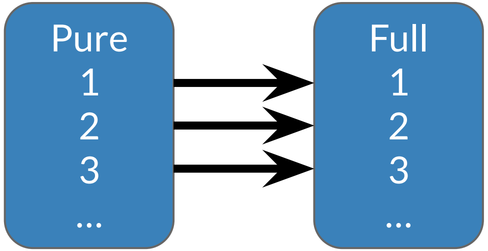
ML Competitors

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



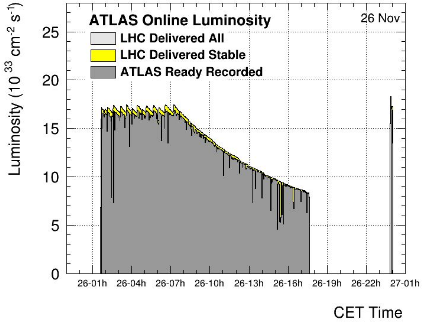
TOTAL

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



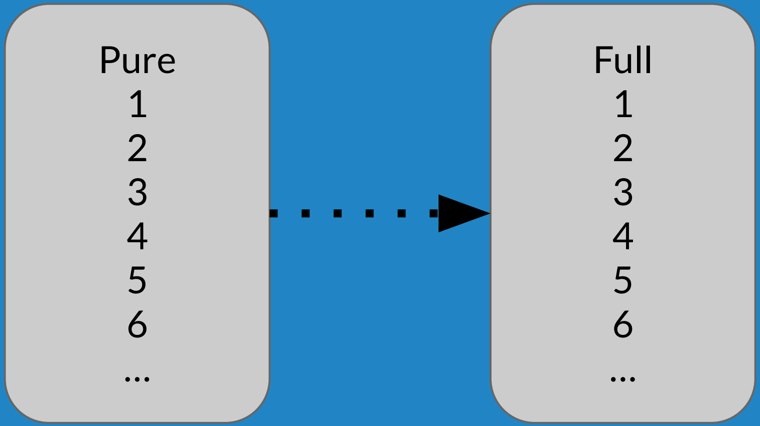
ΔTOTAL

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

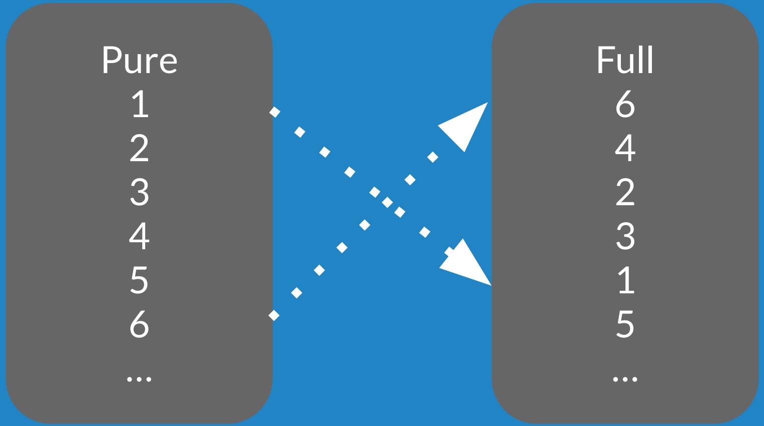


What happens if we do not require direct matching?

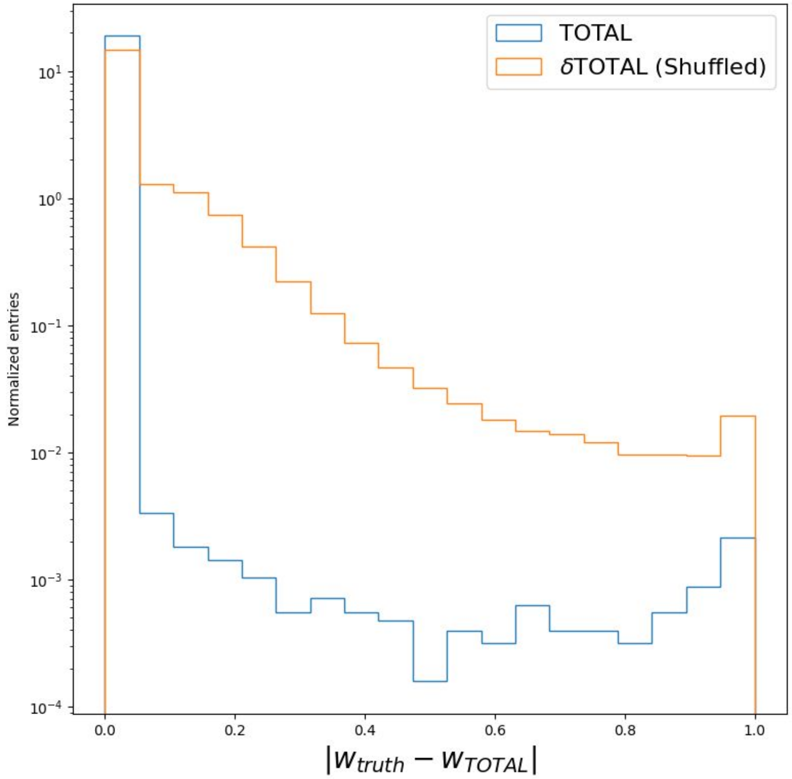
Original



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

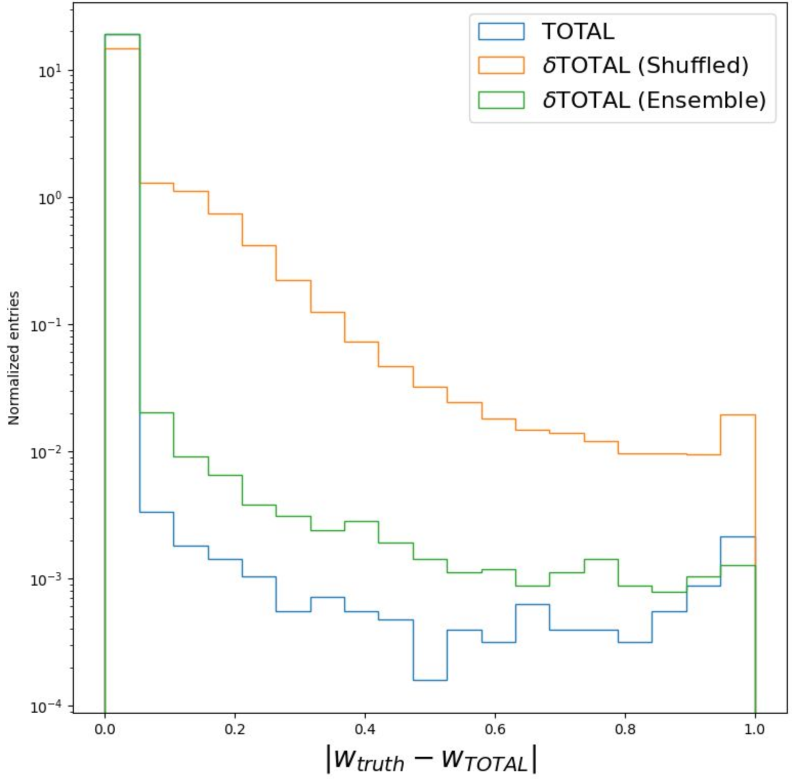
$[(n_{batch} \times n_{particles}) \times n_{features}]$

Batch
[event ensemble: {particles}]

Batch
[event ensemble: {particles}]

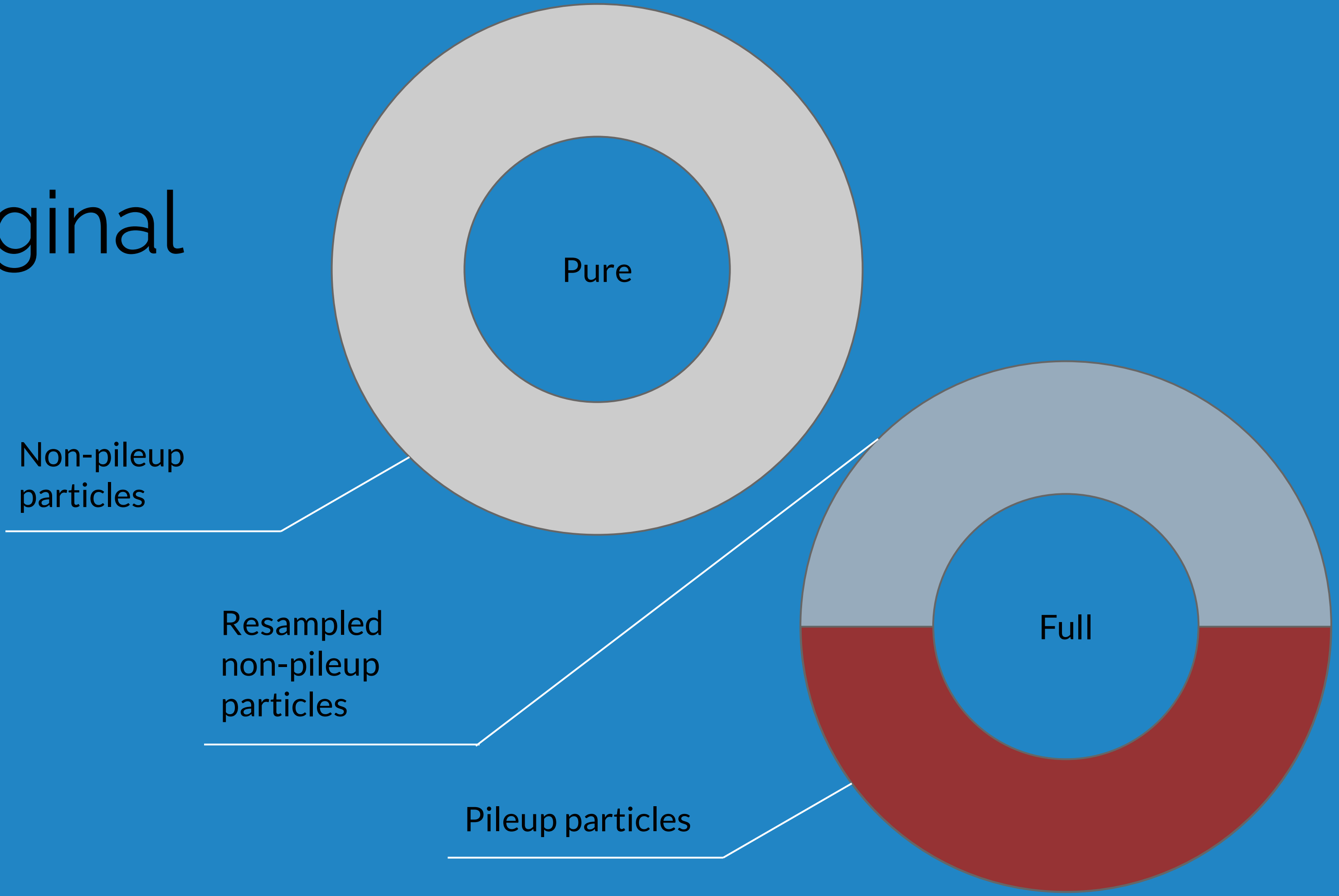
Batch
[event ensemble: {particles}]

Toy Example

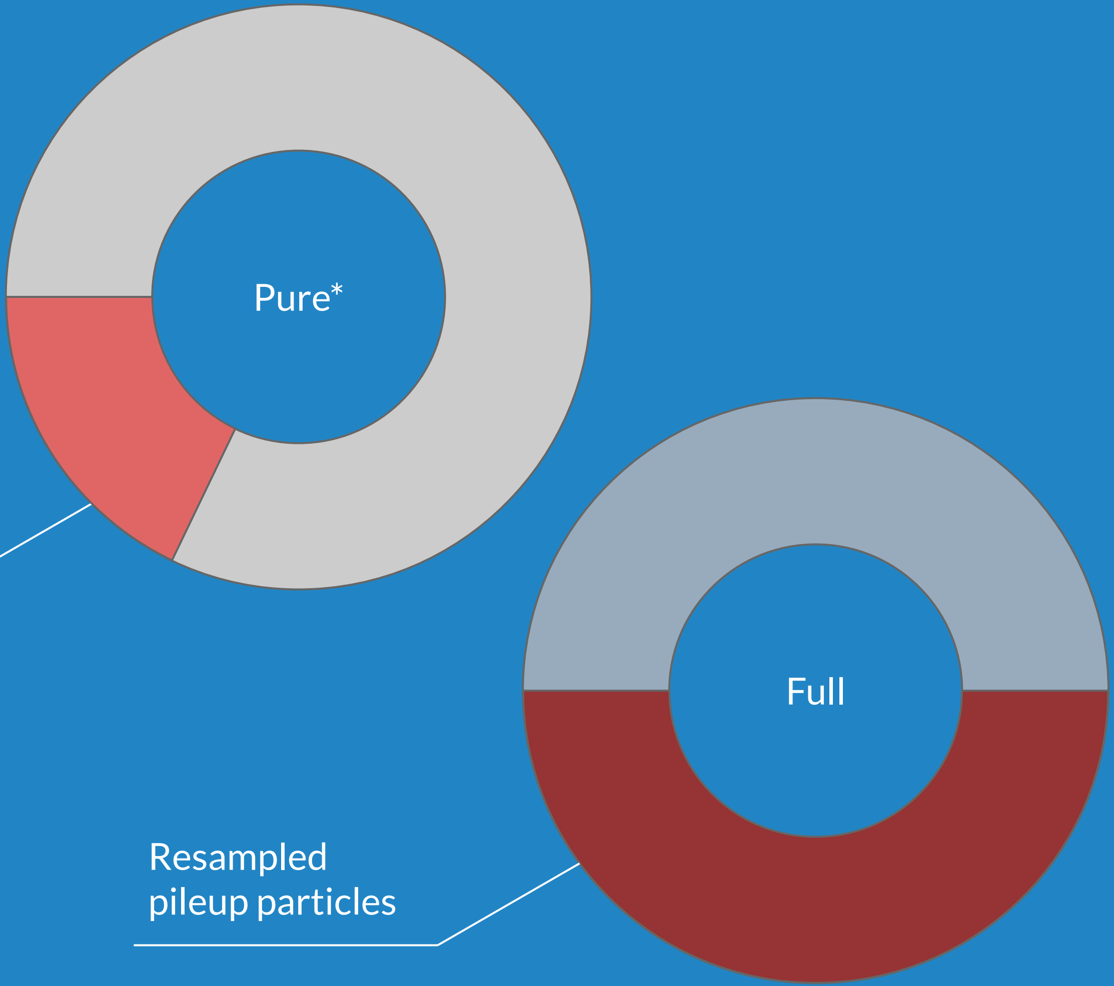

TOTAL
δTOTAL (Shuffled)
δTOTAL (Ensemble)
Normalized entries
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0.0
0.2
0.4
0.6
0.8
1.0
$|w_{truth} - w_{TOTAL}|$

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

Original


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

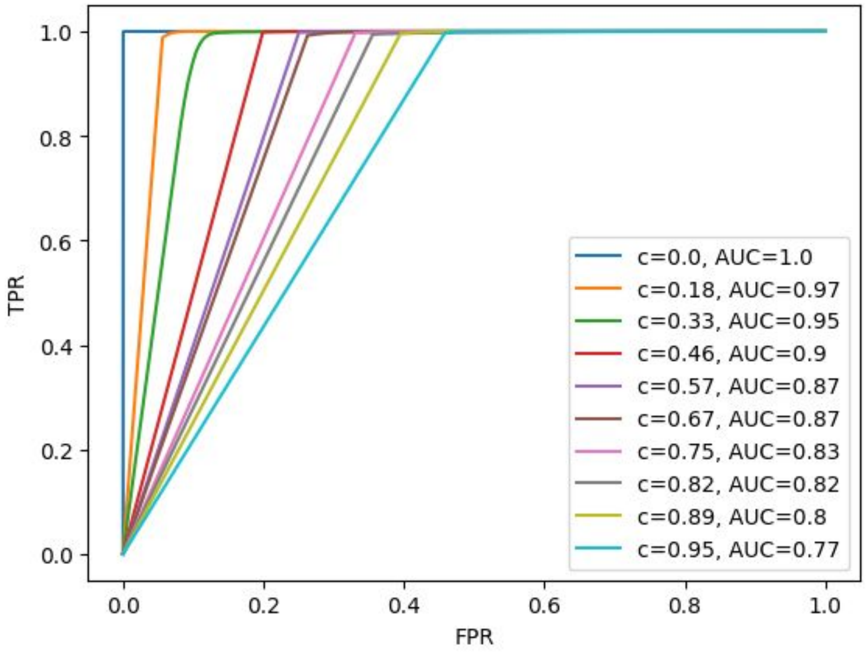
Mixed

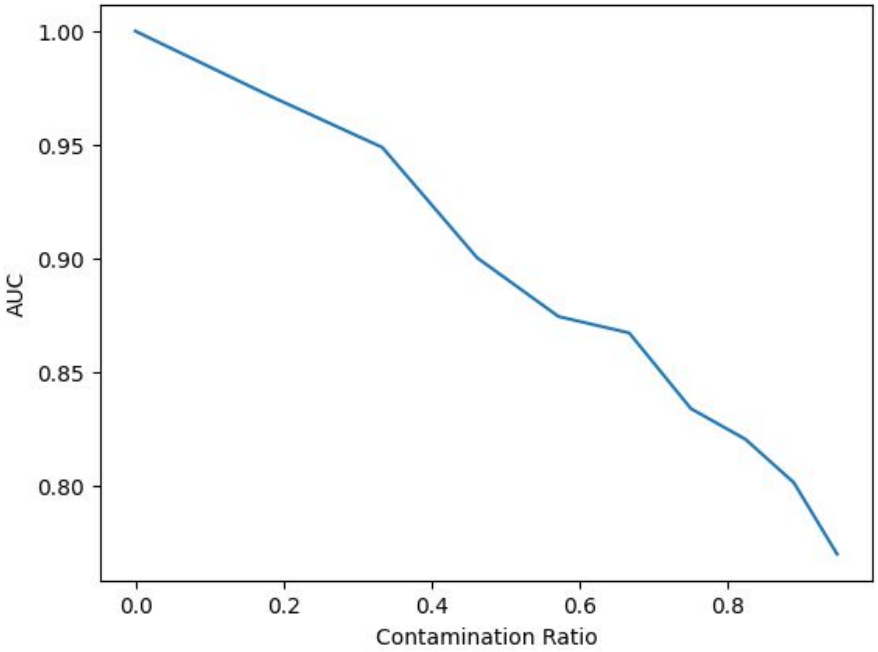


$$c = \frac{\frac{n_{p,pure}}{(n_{p,pure}+n_{np,pure})}}{\frac{n_{p,full}}{(n_{p,full}+n_{np,full})}}$$

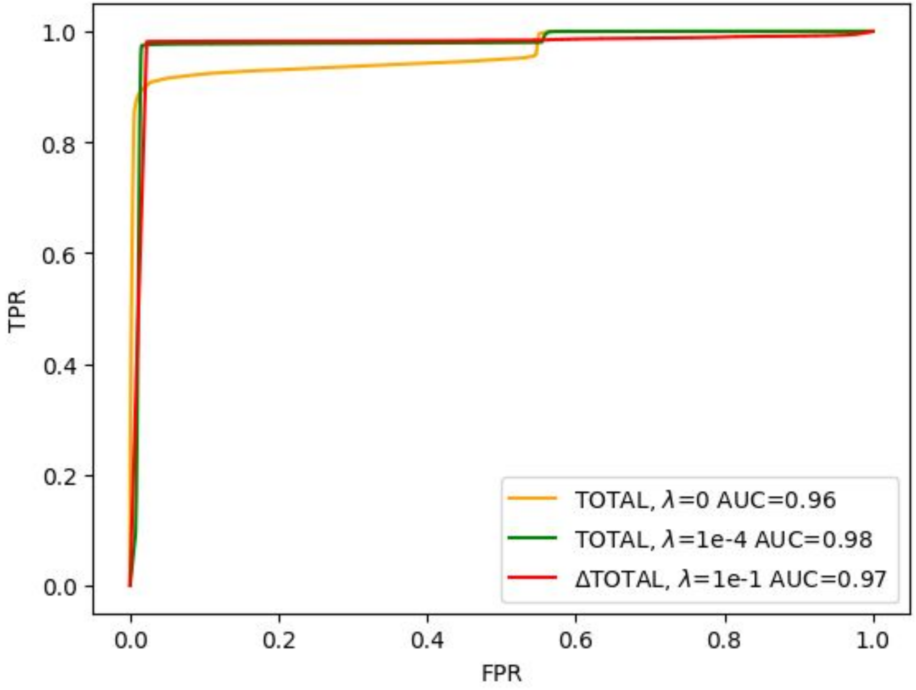
Toy Example

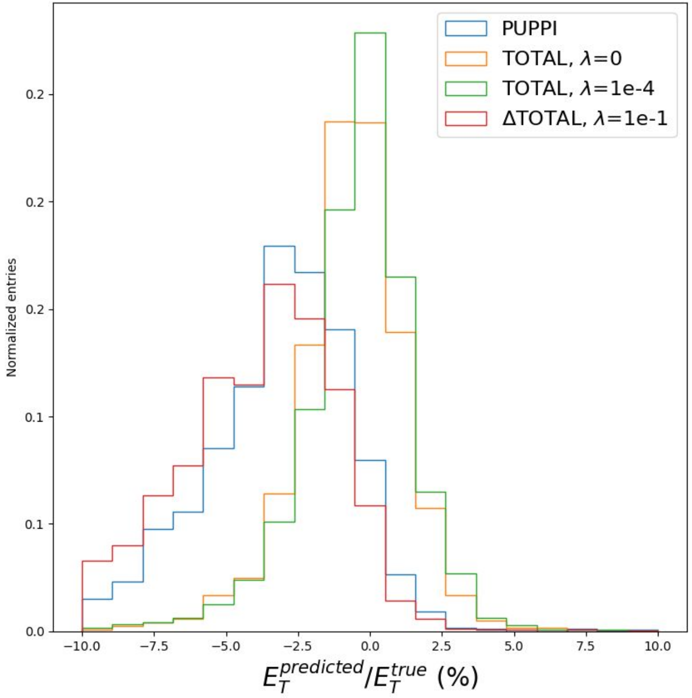
$$c = \frac{\frac{n_{p,pure}}{(n_{p,pure}+n_{np,pure})}}{\frac{n_{p,full}}{(n_{p,full}+n_{np,full})}}$$





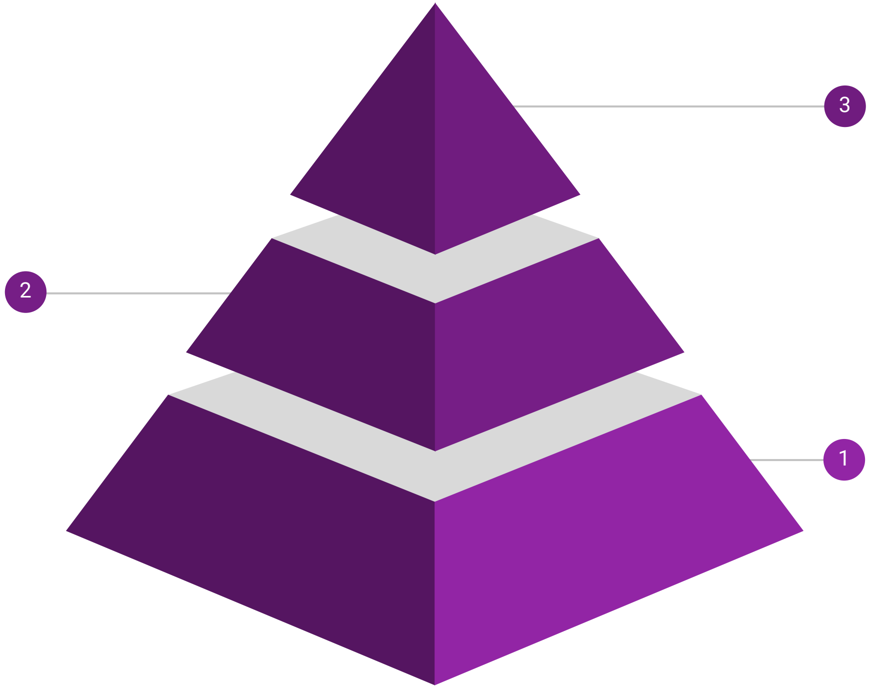
Physics Example: High p_T Jets





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

Key Takeaways



ΔTOTAL represents efforts to increase the flexibility without sacrificing performance

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

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.

TOTAL is a completely data-driven pileup mitigation technique

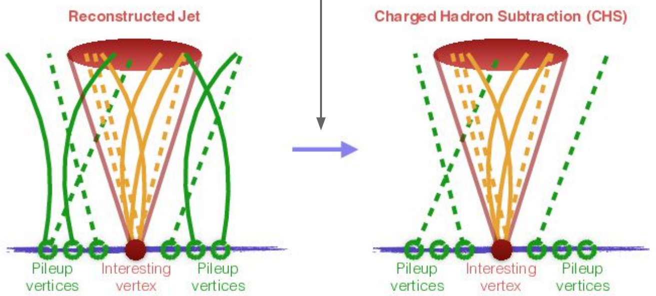
While competing ML methods require particle-level truth and reco matching, TOTAL only requires a match between events with and without pileup.

Backup Slides

Charged Hadron Subtraction

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

- 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

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

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

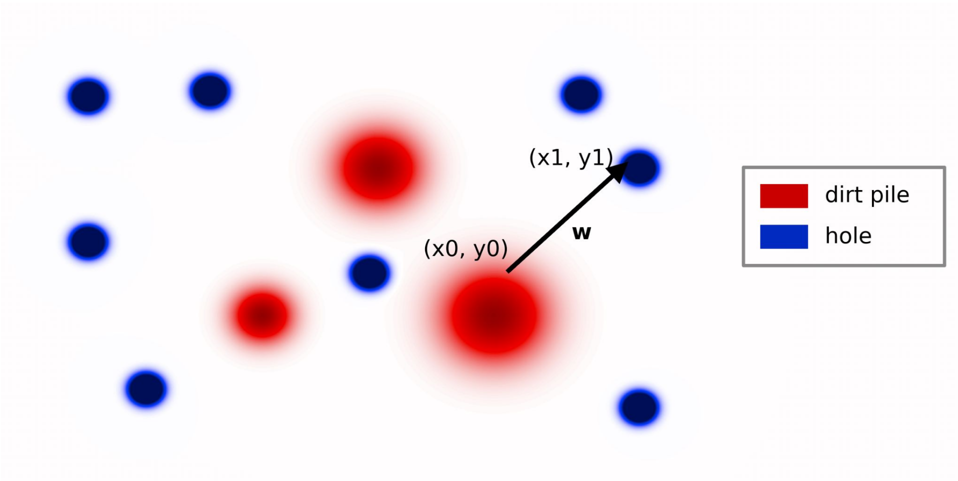
TOTAL: Training Optimal Transport with Attention Learning

$$\mathcal{L} =$$

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

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

Wasserstein Metric



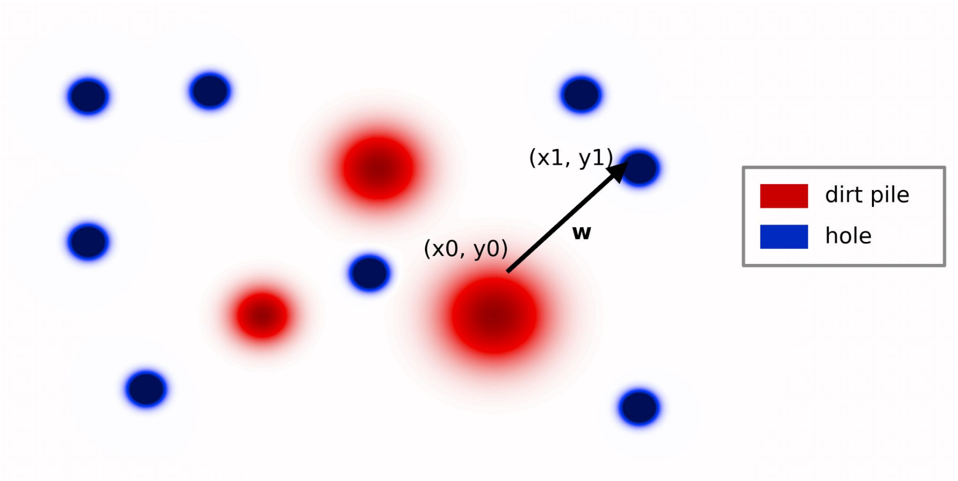
- Assumption: Total volume of the holes = total volume of the dirt piles
- Piles as the probability density function of P and holes as the probability density function of Q
- Per unit transportation cost:

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

- 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

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

↓ Modification for jet-based dataset (PUMML)

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

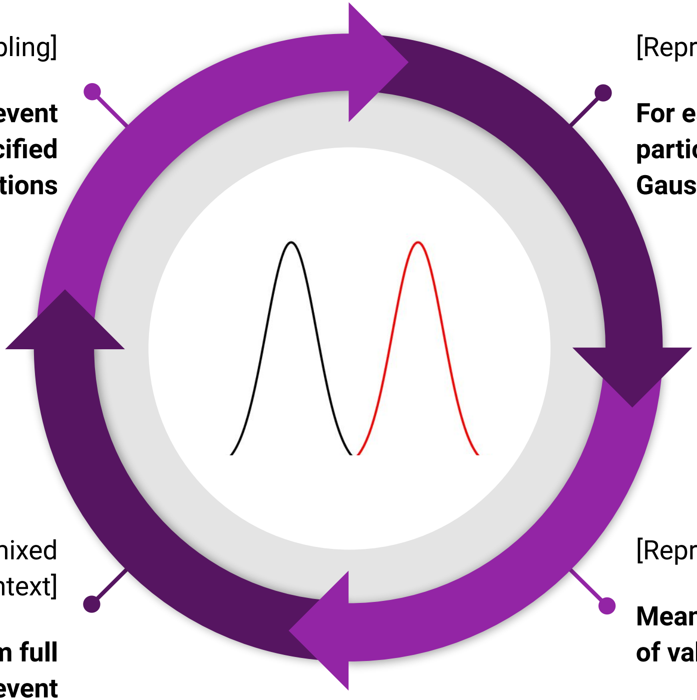
Toy Example Generation

[Equivalent to fill/low-pileup sampling]

Mixed Training: Non-pileup data in full event resampled from the originally specified distributions

[Representative of distinct PV]

For each event, pileup and non-pileup particles are sampled from separate Gaussian distributions

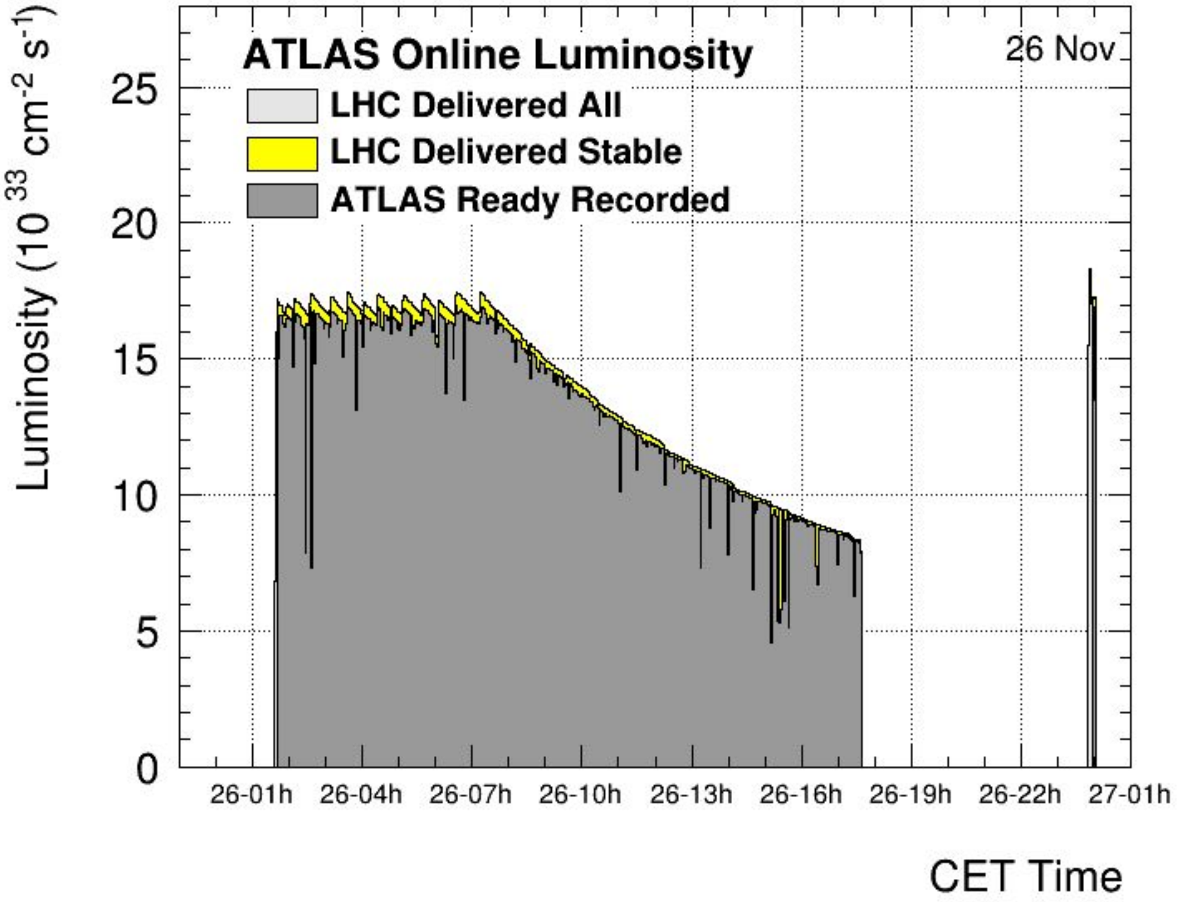


[Required for training, modified in mixed training context]

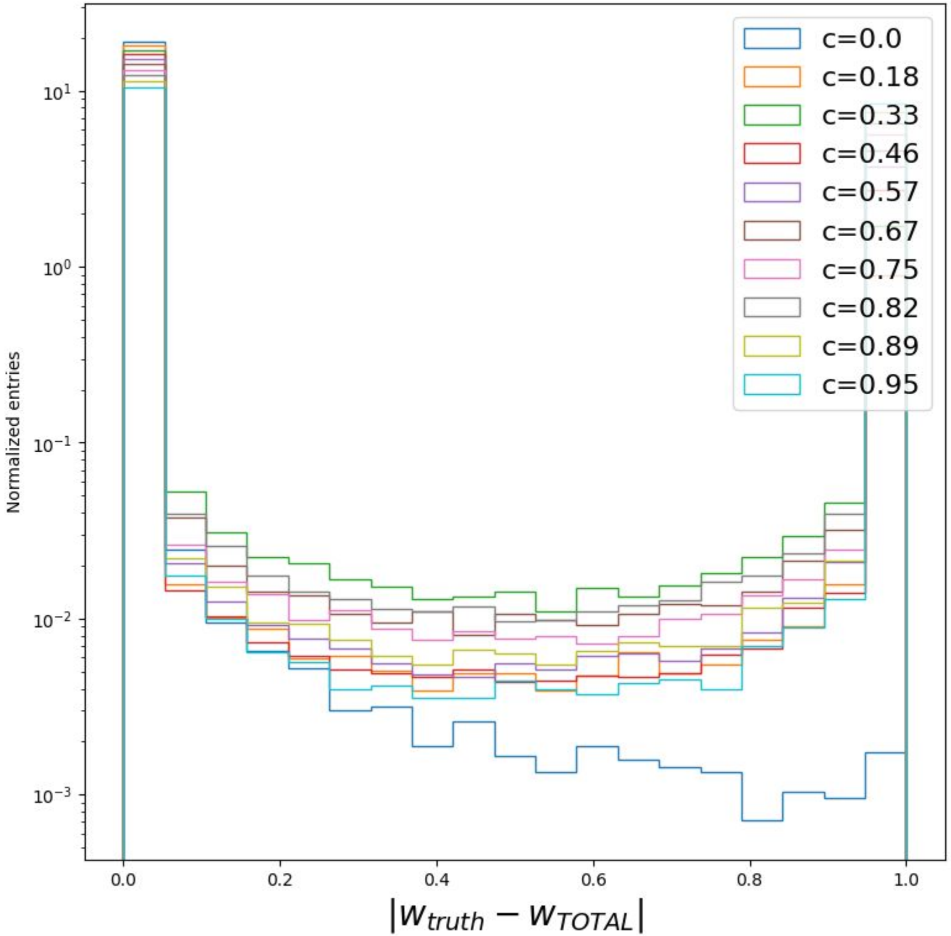
Pure event copied and zero-padded from full event

[Representative of differing pileup jets]

Means are uniformly sampled from a range of values

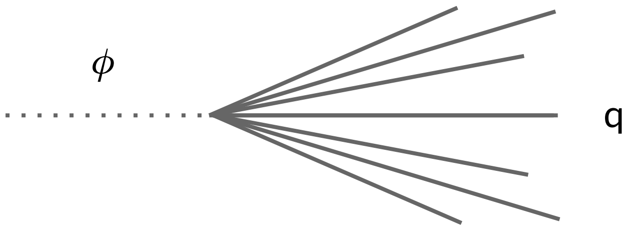


ATLAS Fill Luminosity



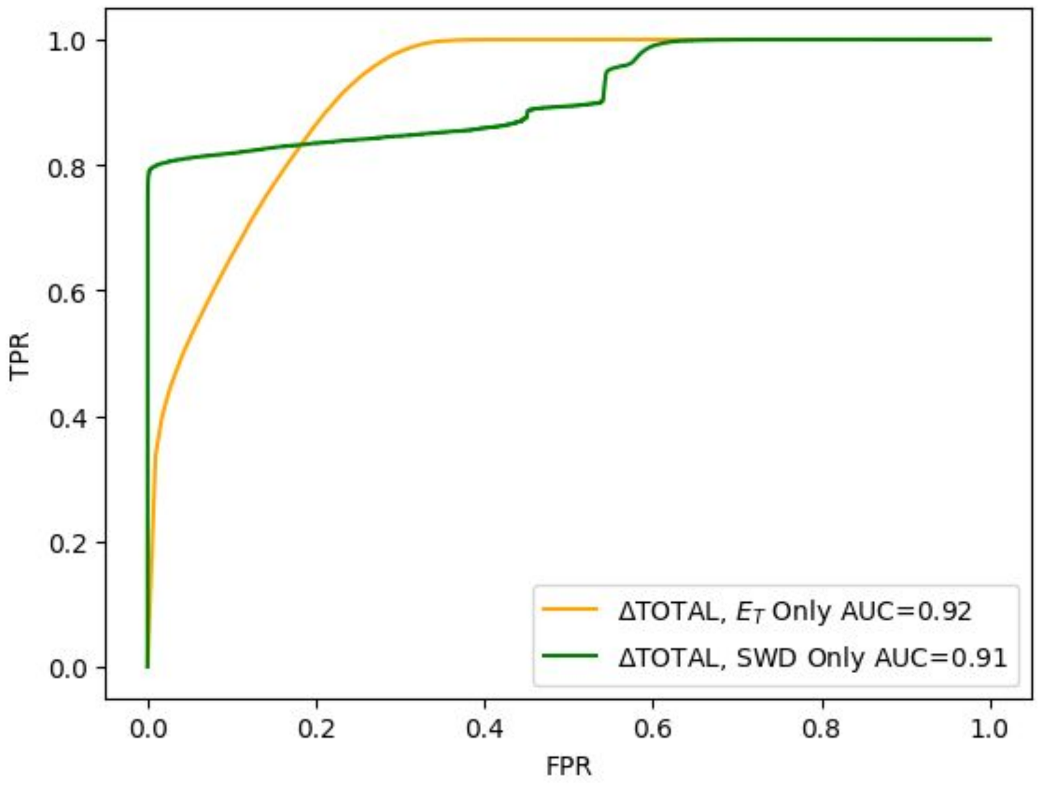
Toy Example: Contamination Ratio Test

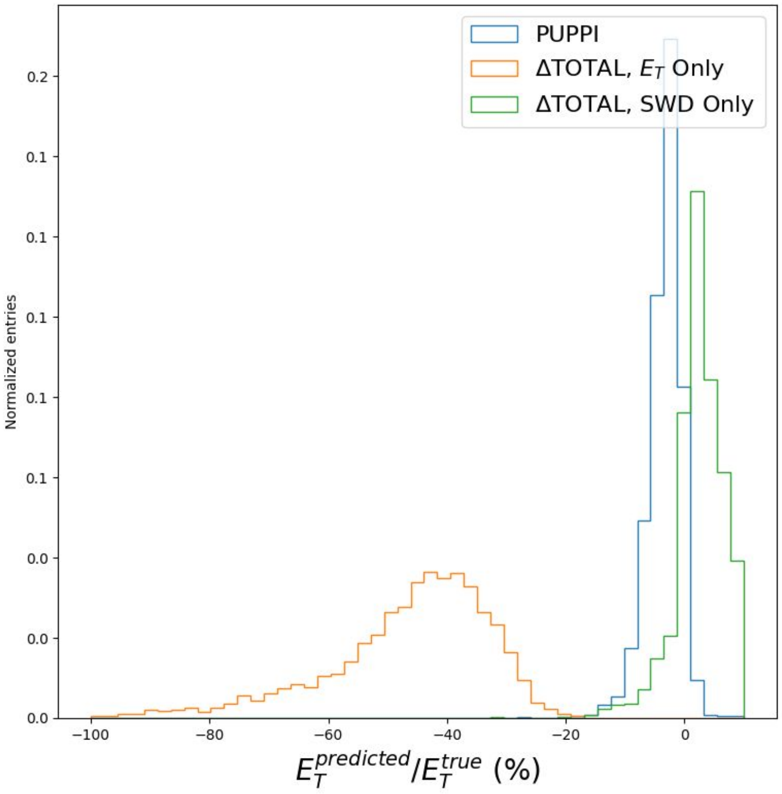
Physics Example: High p_T Jets



- 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)