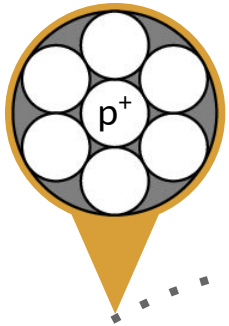
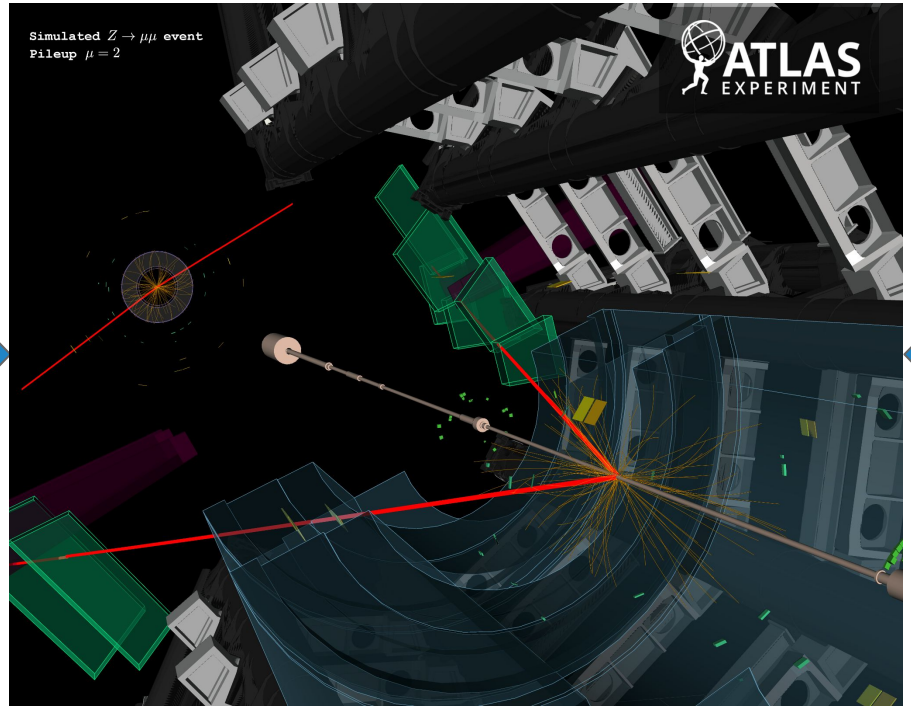
WOTAN: Weakly-supervised Optimal Transport Attention-based Noise Mitigation

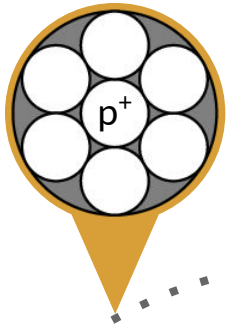
Nathan Suri, Vinicius Mikuni, Benjamin Nachman

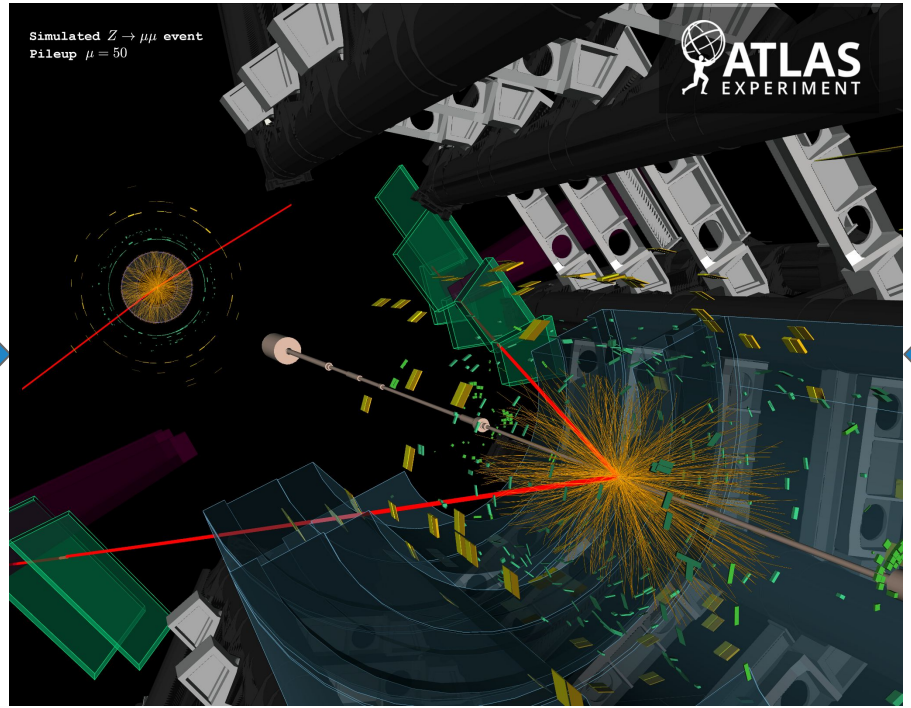
ML4Jets 2024



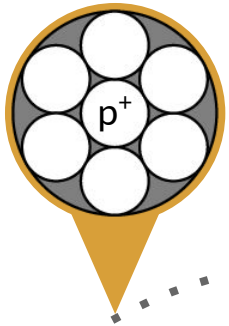


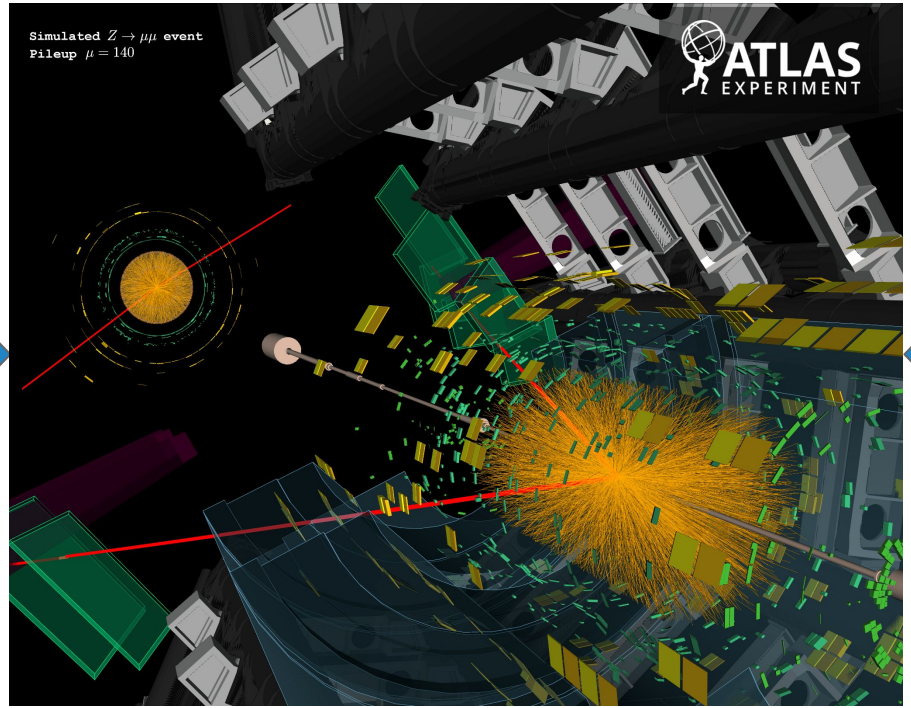
Charged + neutral pileup
In-time + out-of-time pileup



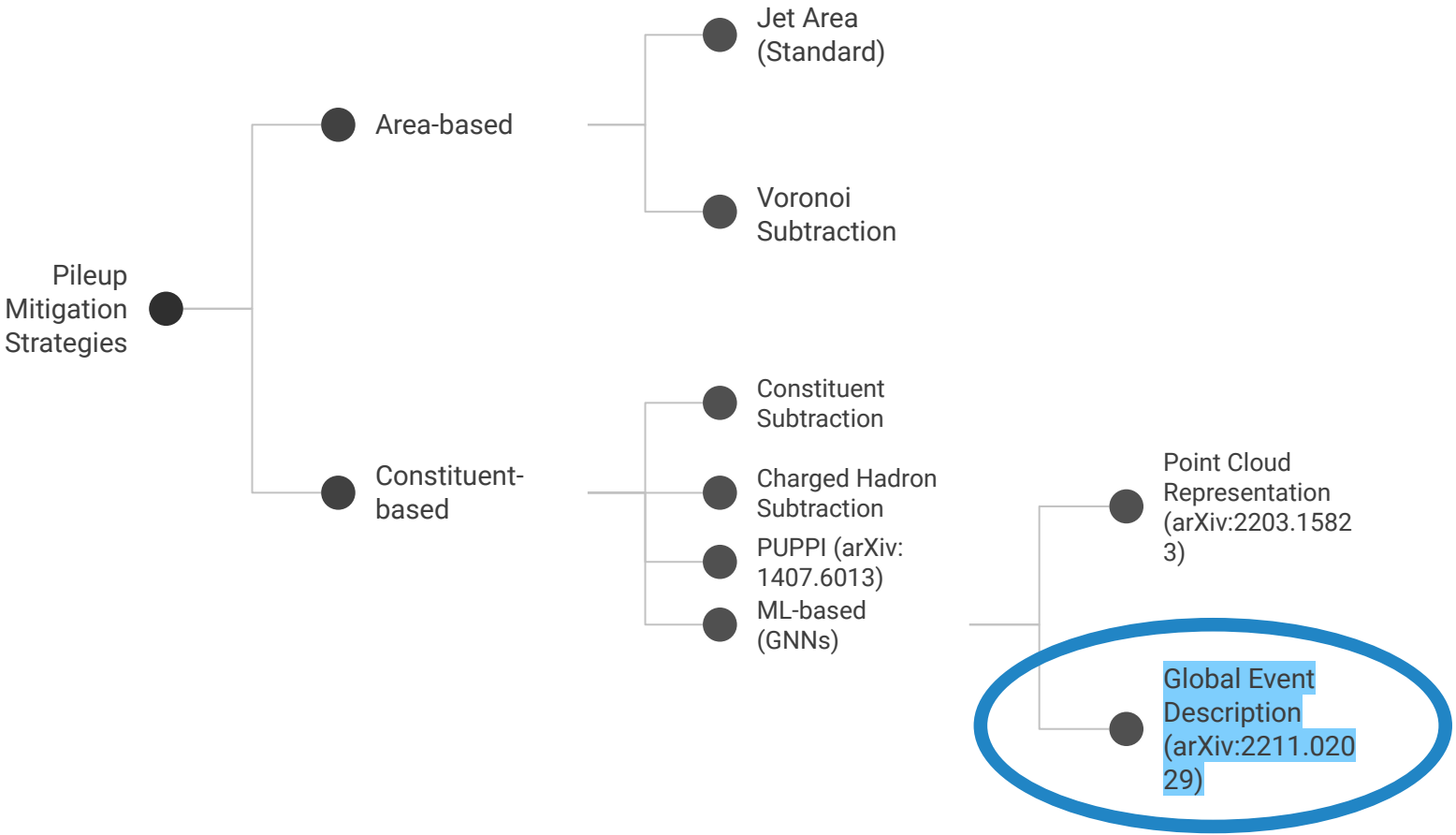


Charged + neutral pileup
In-time + out-of-time pileup

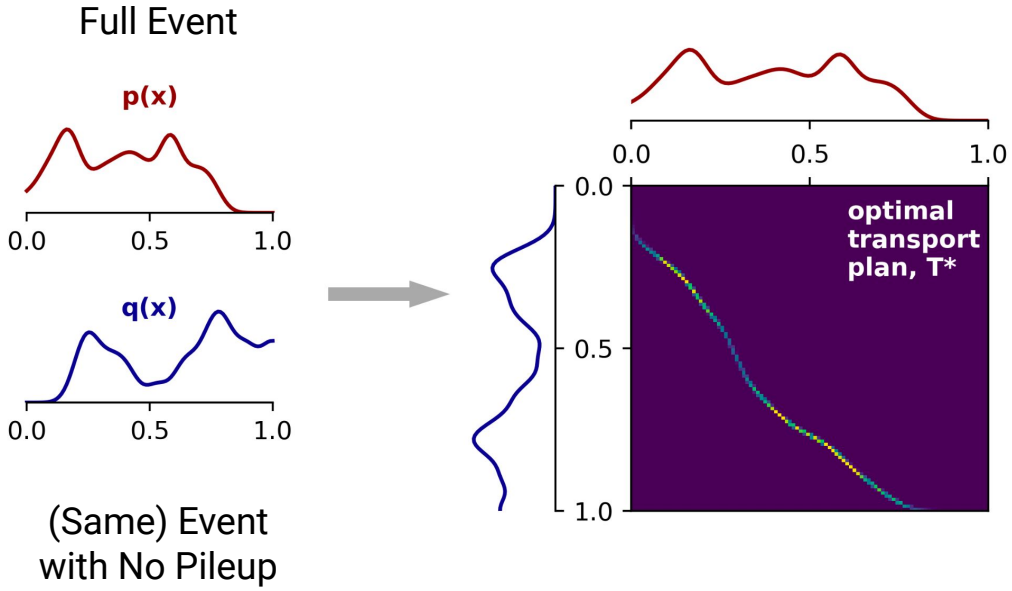




Charged + neutral pileup
In-time + out-of-time pileup


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)

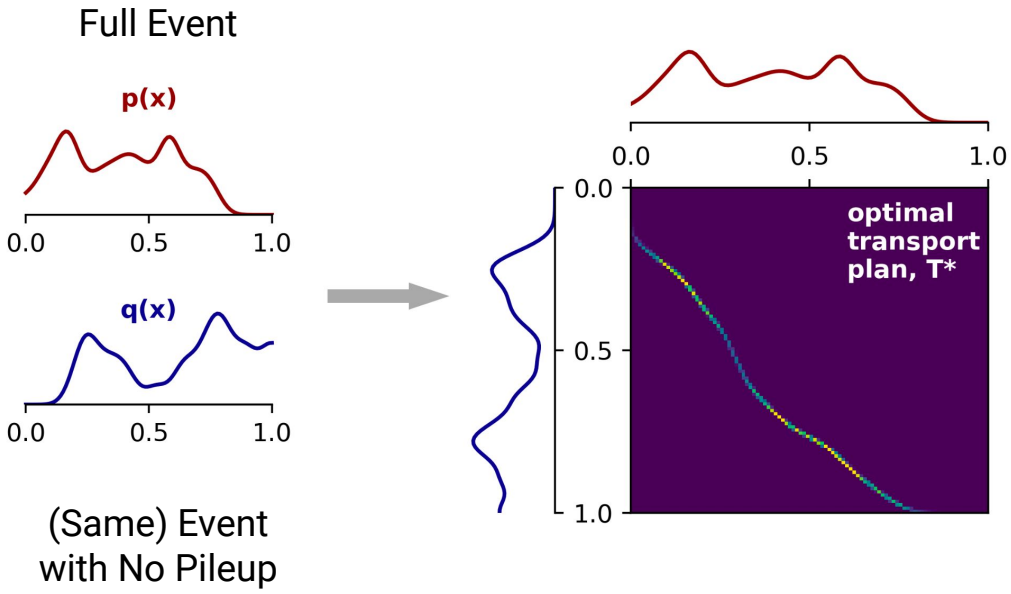
TOTAL: Training Optimal Transport with Attention Learning



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

TOTAL: Training Optimal Transport with Attention Learning

- The probability density is intractable, but we can approximate the density
- Realizations of the density are accessible
- Optimal transport over the space of inputs allows for approximation



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

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

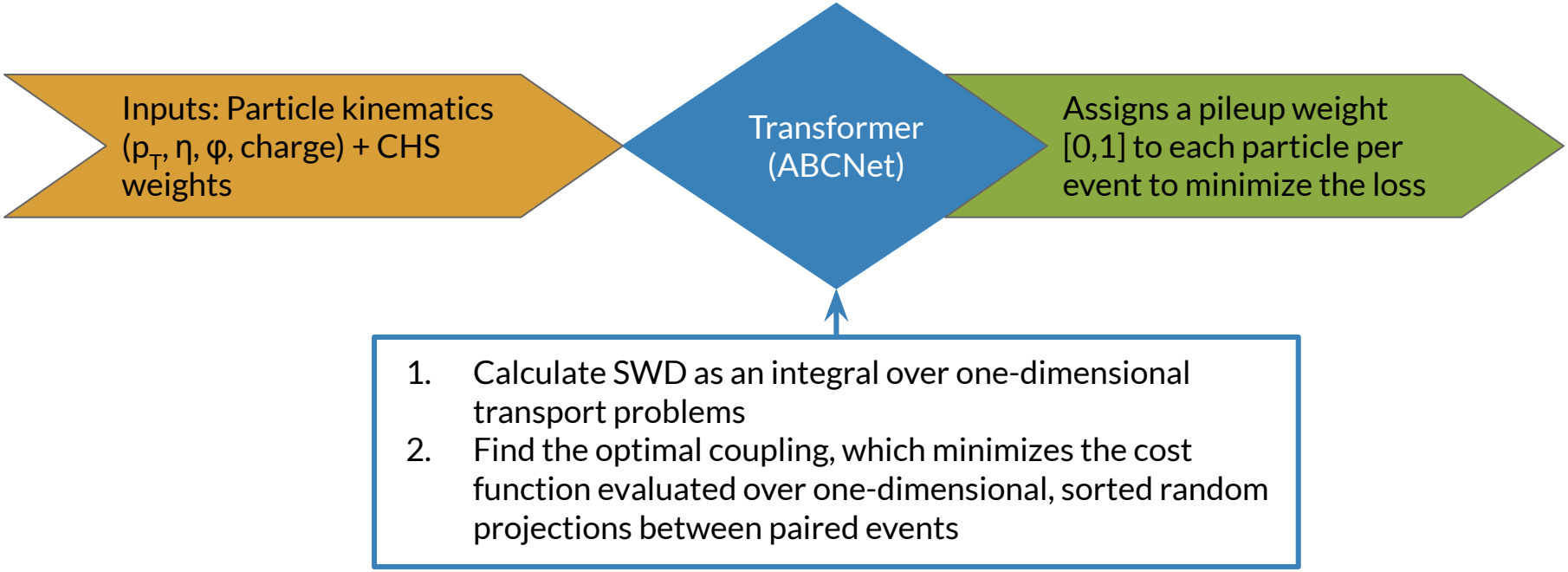
$$x'_p = \omega x_p$$

TOTAL: Training Optimal Transport with Attention Learning

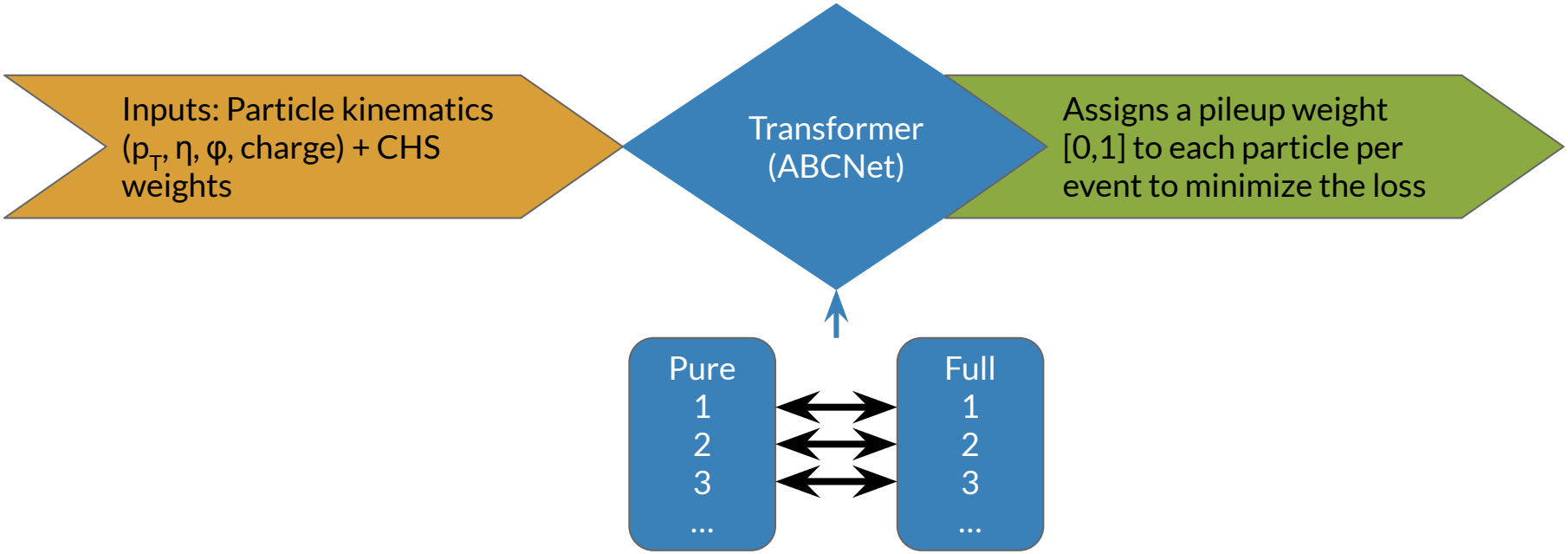
$$\mathcal{L} = + \lambda \, \mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

- Scaled Mean Square Error of missing p_T
- Forces energy conservation between the pure and full samples

TOTAL: Training Optimal Transport with Attention Learning

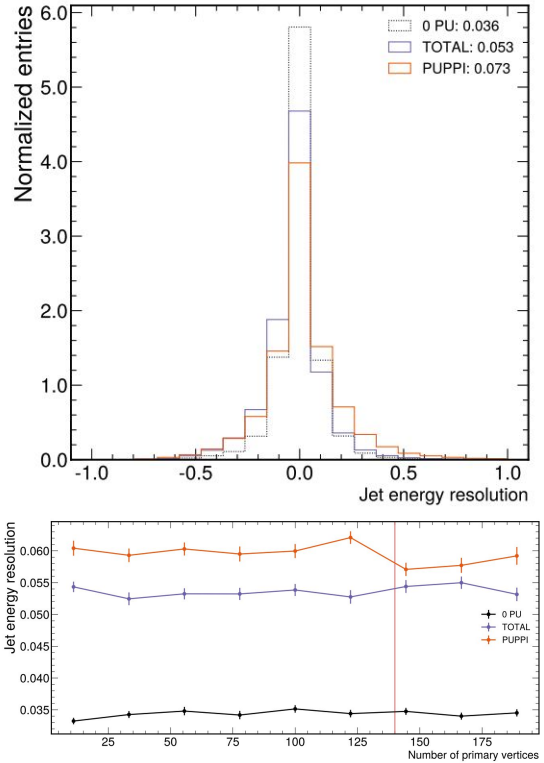


TOTAL: Training Optimal Transport with Attention Learning



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



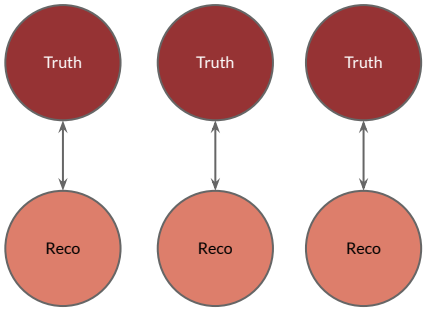
- Requires direct matching of events
- Overall limited due to supervision

(arXiv:2211.02029)

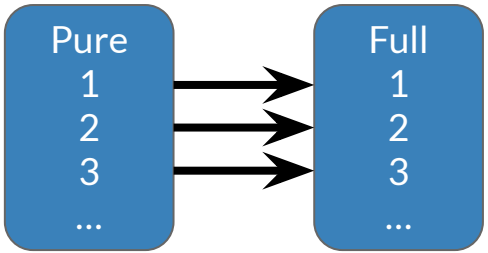
TOTAL: Training Optimal Transport with Attention Learning

ML Competitors | TOTAL | WOTAN

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

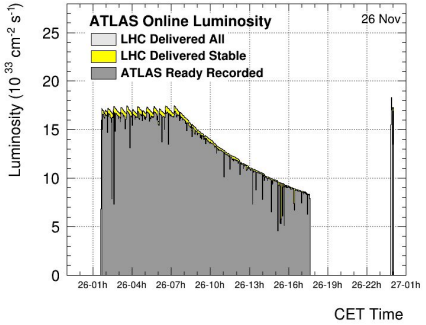


- Matching between pileup events and same event without pileup vertices (data-driven*)



*Such supervision is not physically realizable

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



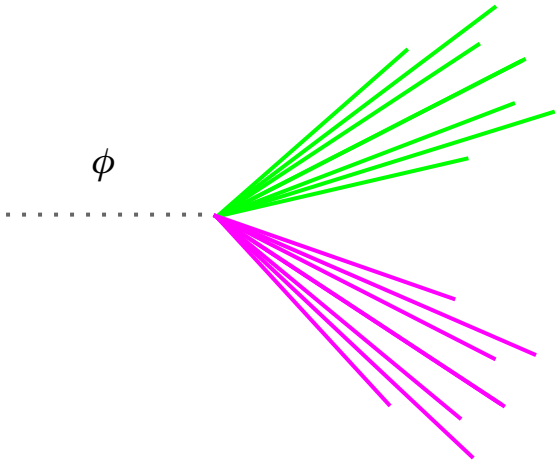
How can we improve TOTAL's flexibility?

Nathan Suri, 2024

Reduce supervision!

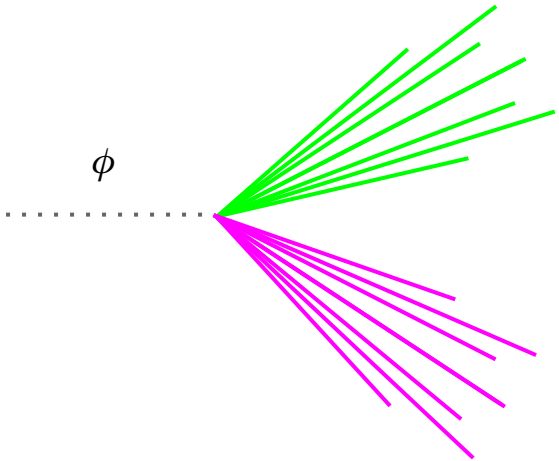
Nathan Suri, 2024

Physics Example: High p_T Jets



- PUMML Dataset: https://zenodo.org/records/2652034
- Process: q-qbar
  light-quark-initiated jets from the
  from the decay of a Higgs-like scalar
  - Pileup was generated by overlaying soft QCD on top of signal

Physics Example: High p_T Jets

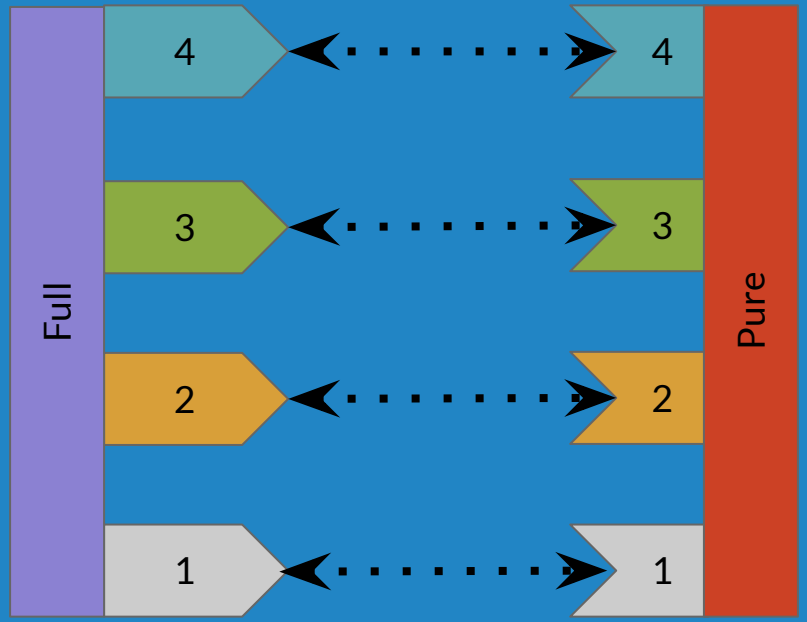


- PUMML Dataset: https://zenodo.org/records/2652034
- Datasets
  - μ = 140, Δm: Set pileup vertex count, varied scalar mass
  - Δμ, m = 500 GeV: Varied pileup vertex (PV) count, set scalar mass
    - PV: 130-141

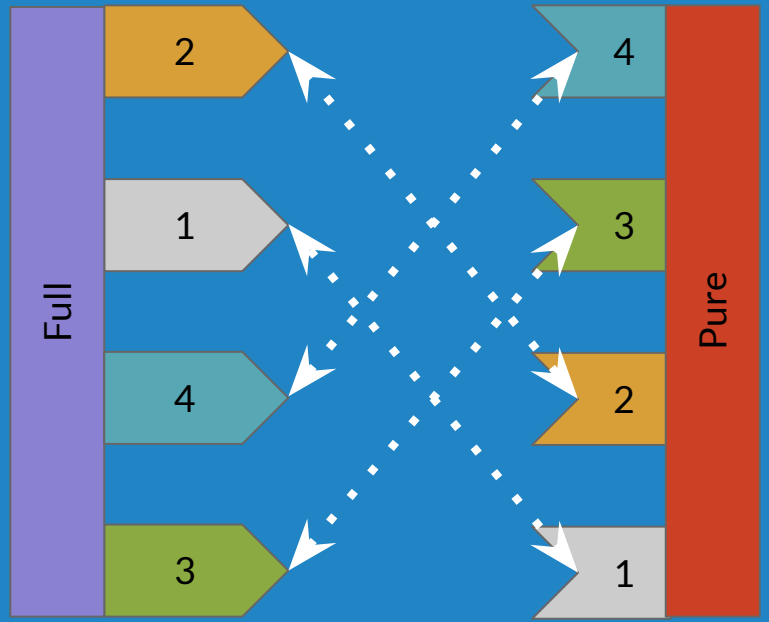
What happens if we do not require direct matching?

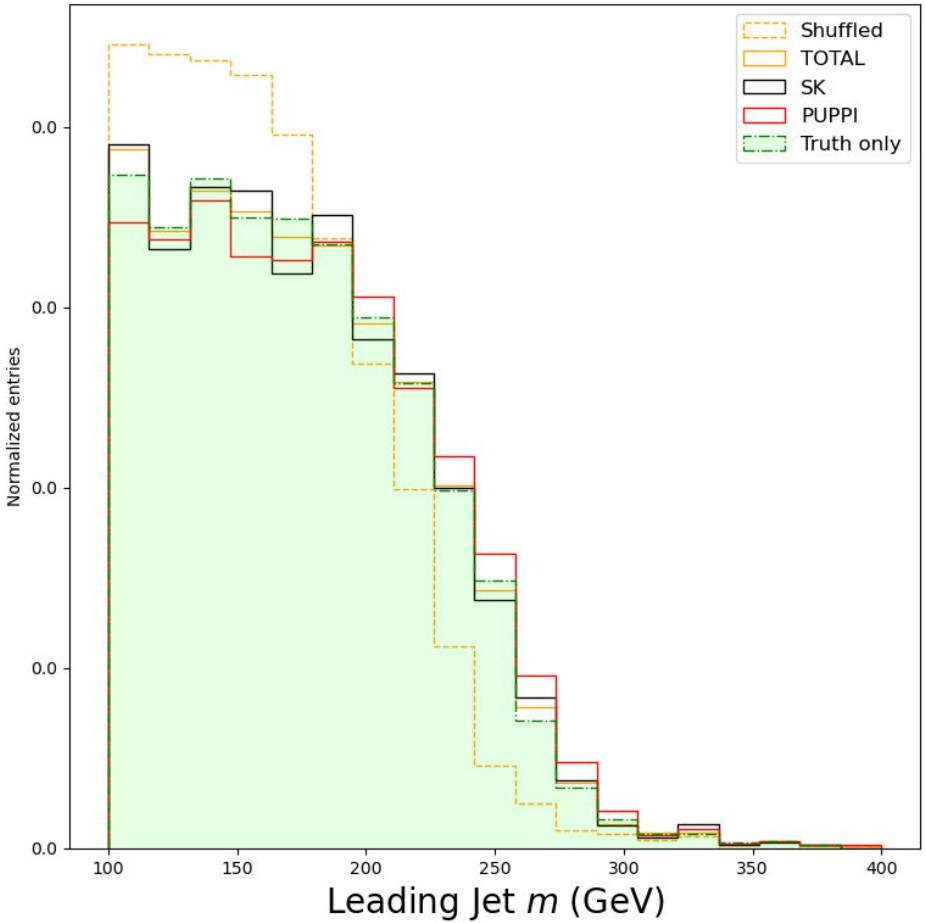
Nathan Suri, 2024

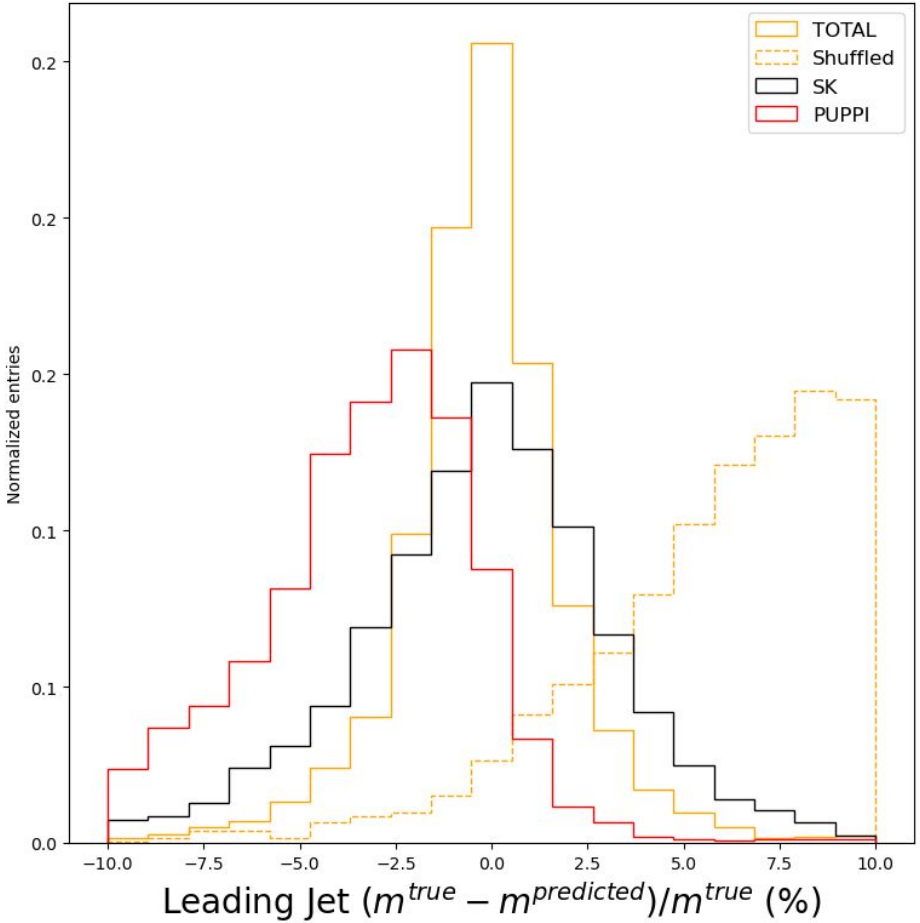
TOTAL



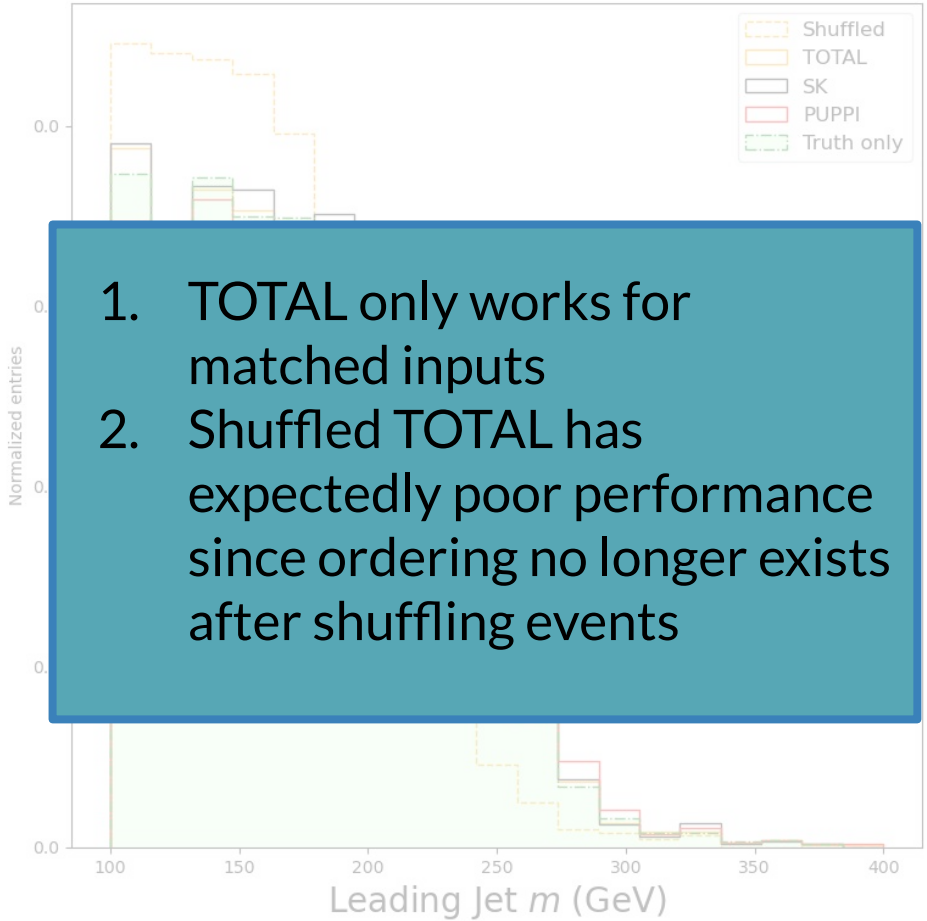
Shuffled



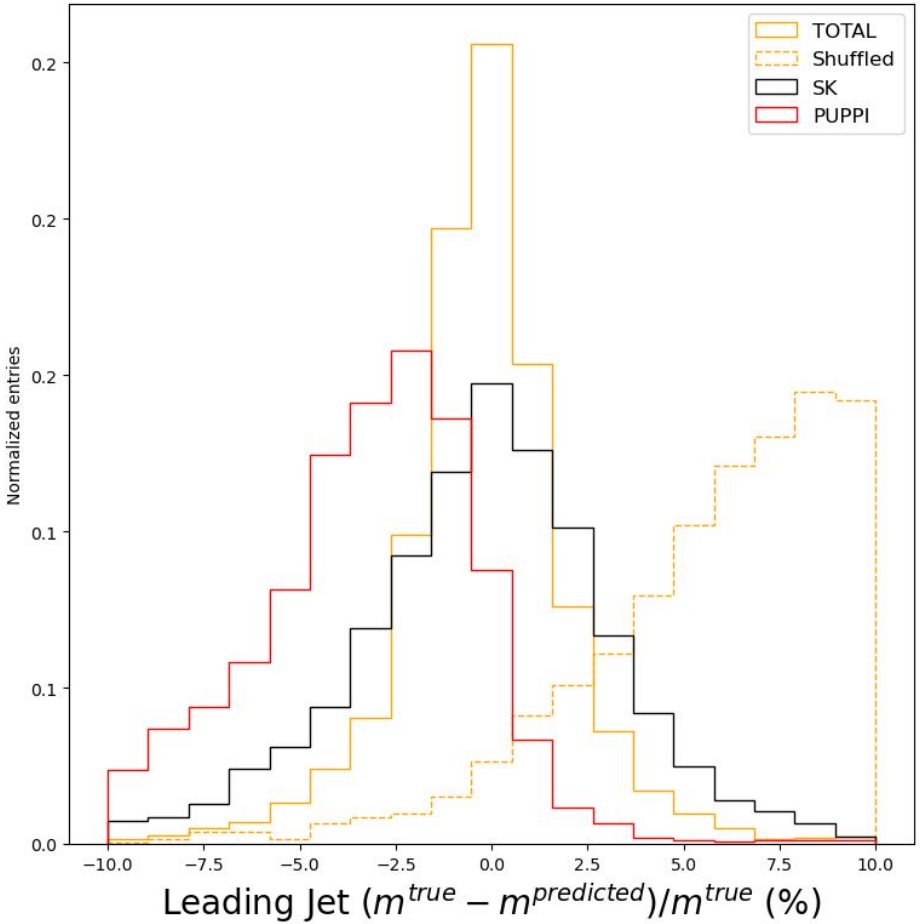




**Each result averaged over 5 trainings*



1. TOTAL only works for matched inputs
2. Shuffled TOTAL has expectedly poor performance since ordering no longer exists after shuffling events



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

TOTAL

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

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

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

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

WOTAN

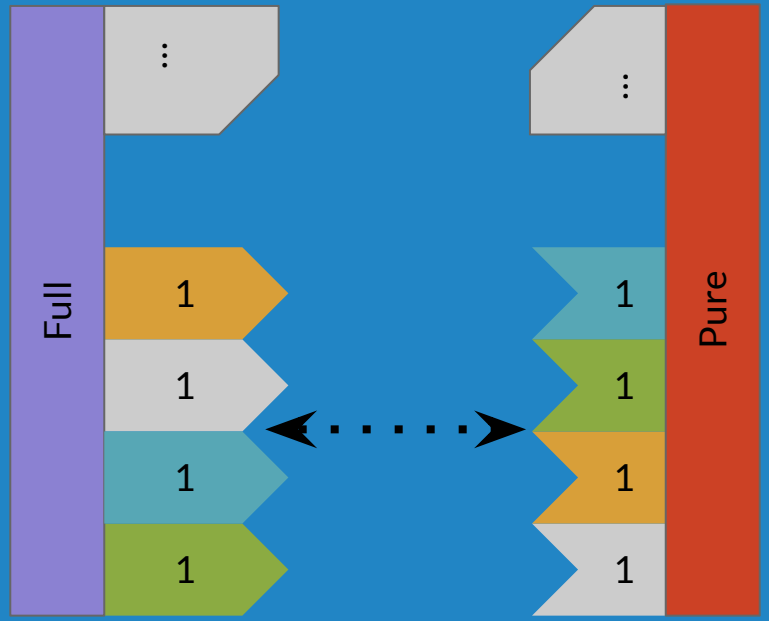
$[(n_{batch} \times n_{particles}) \times n_{features}]$

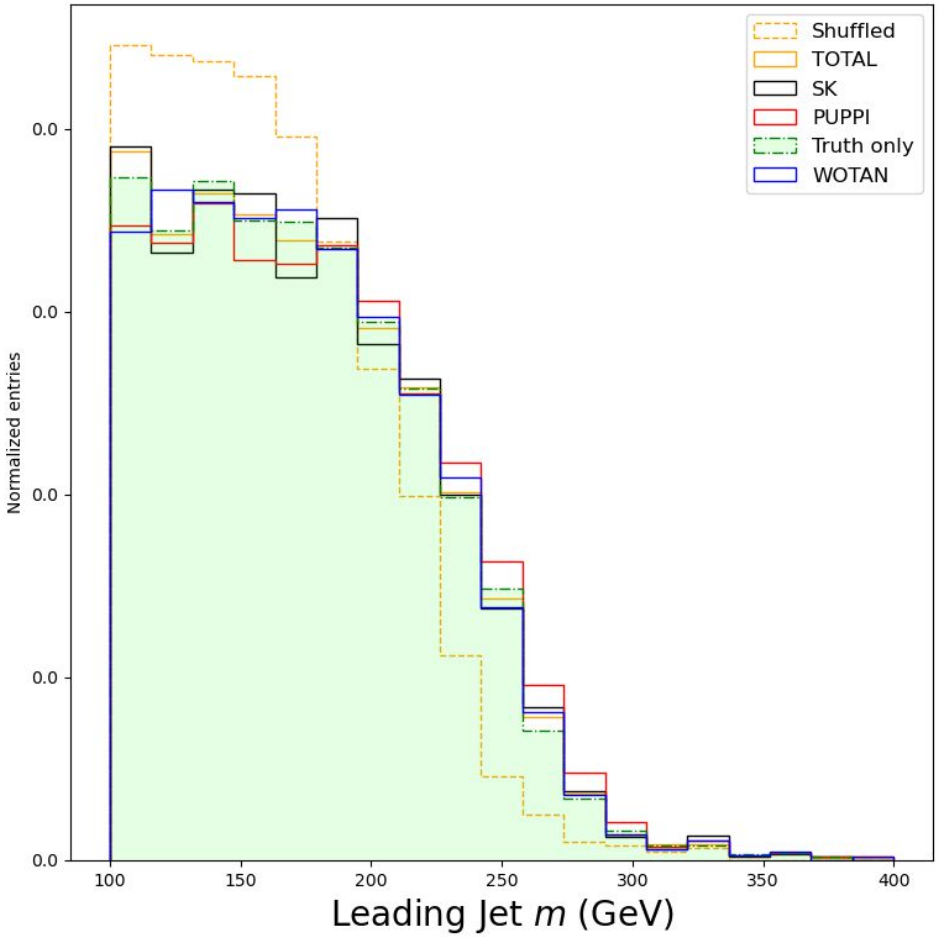
Batch
[event ensemble: {particles}]

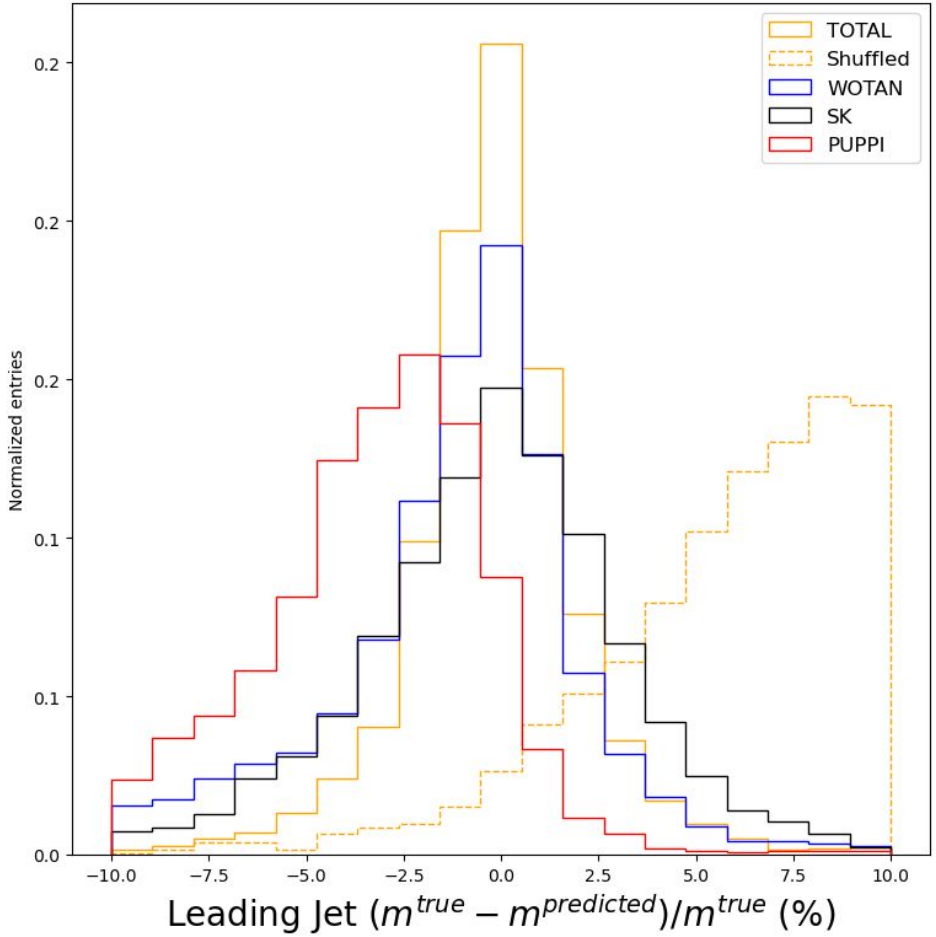
Batch
[event ensemble: {particles}]

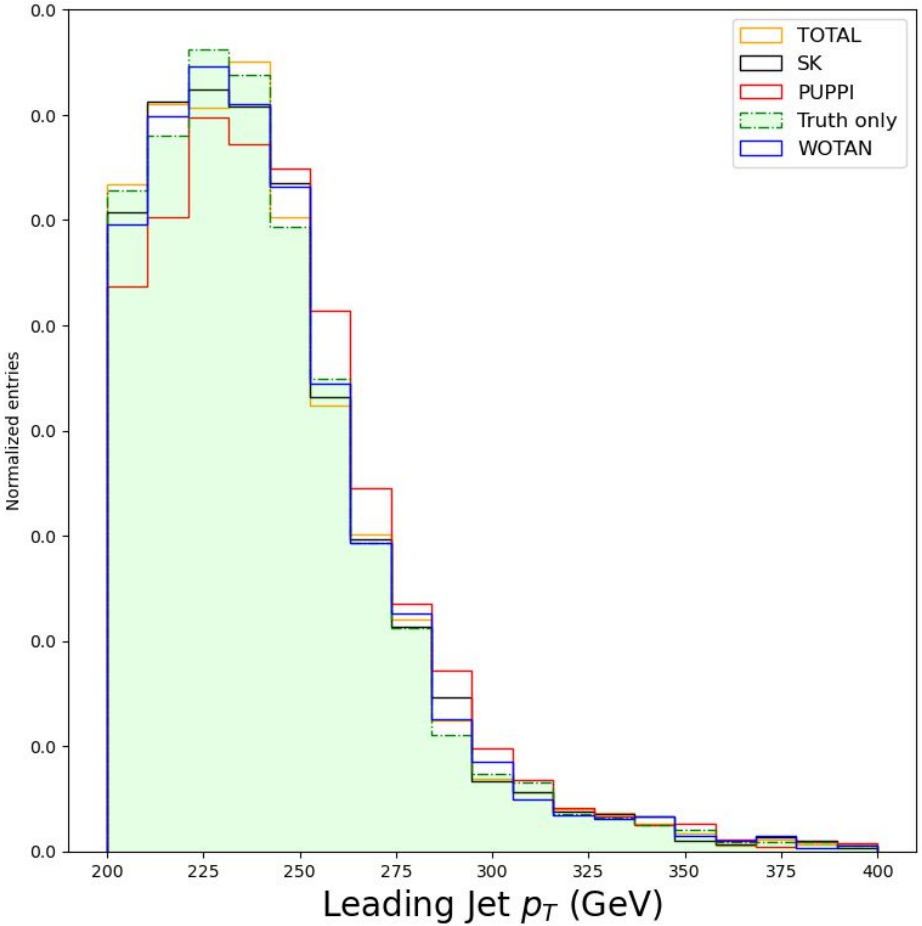
Batch
[event ensemble: {particles}]

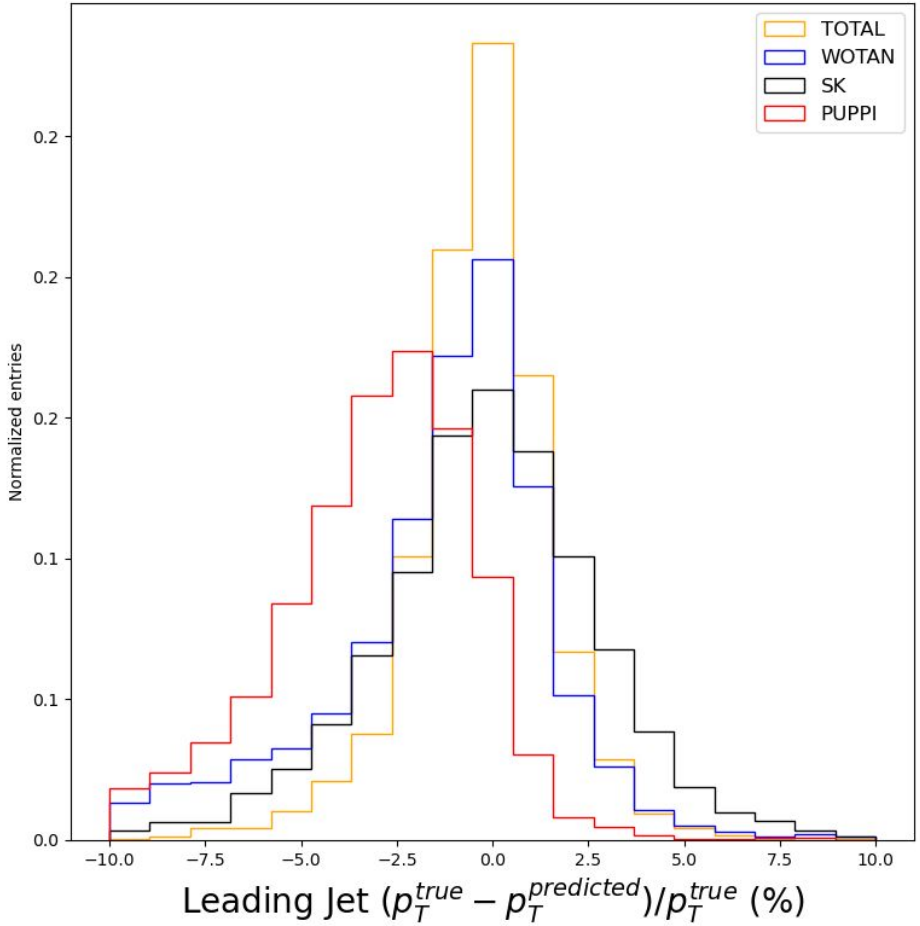
WOTAN

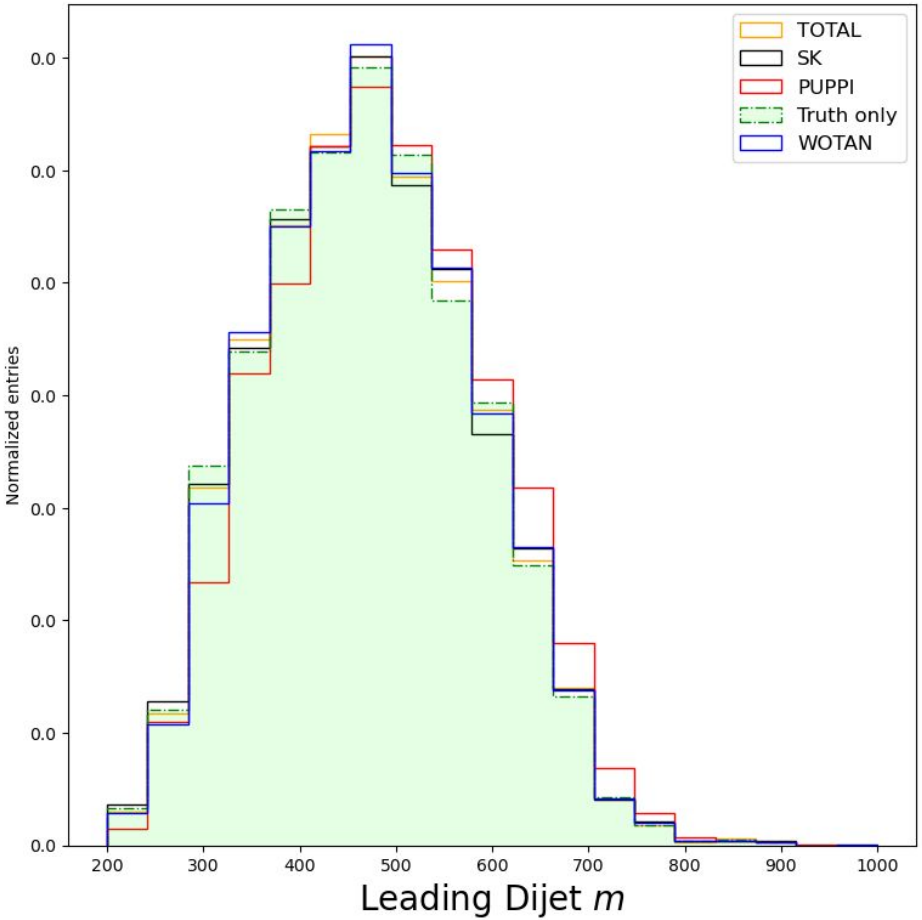


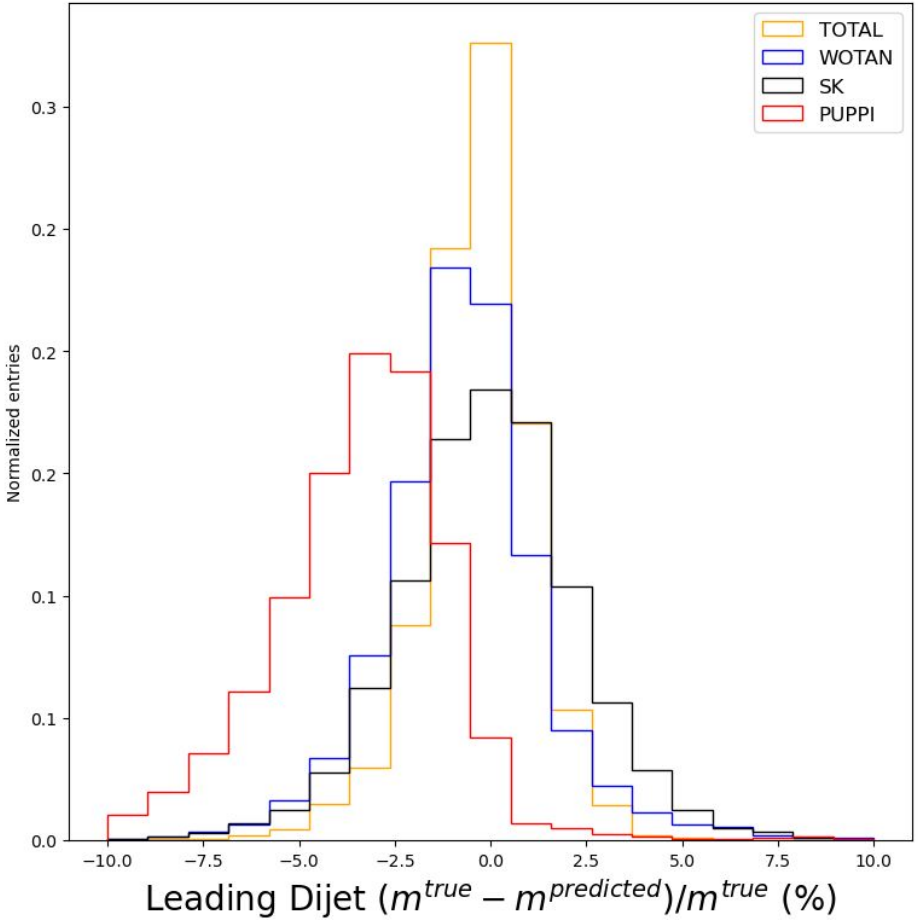

Shuffled
TOTAL
SK
PUPPI
Truth only
WOTAN
Normalized entries
0.0
0.0
0.0
0.0
100
150
200
250
300
350
400
Leading Jet m (GeV)


TOTAL
Shuffled
WOTAN
SK
PUPPI
Normalized entries
0.2
0.2
0.2
0.1
0.1
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Leading Jet $(m^{true} - m^{predicted})/m^{true}$ (%)

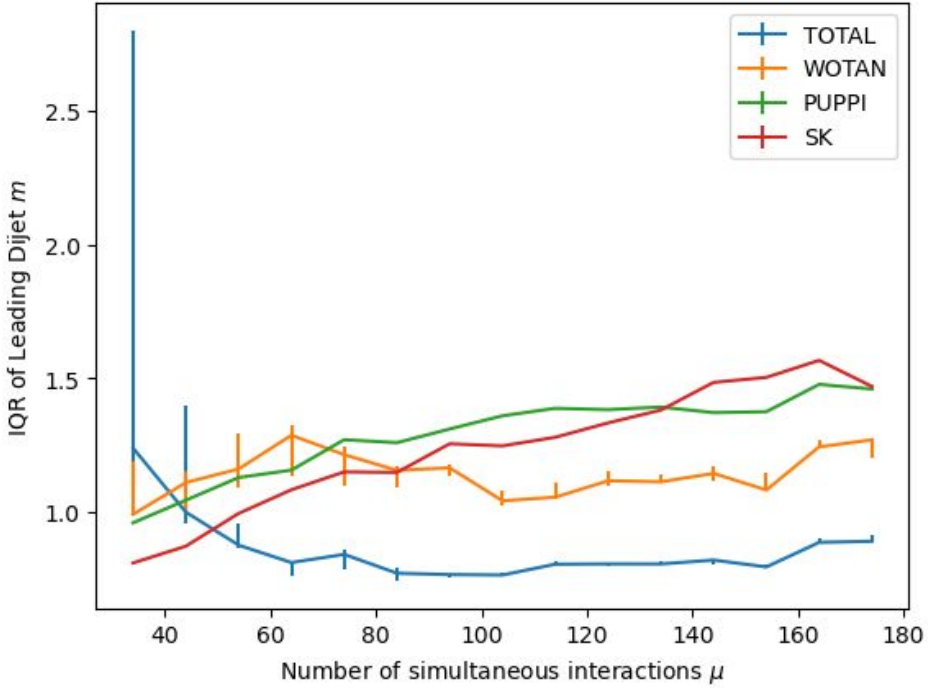

TOTAL
SK
PUPPI
Truth only
WOTAN
Normalized entries
Leading Jet p_T (GeV)

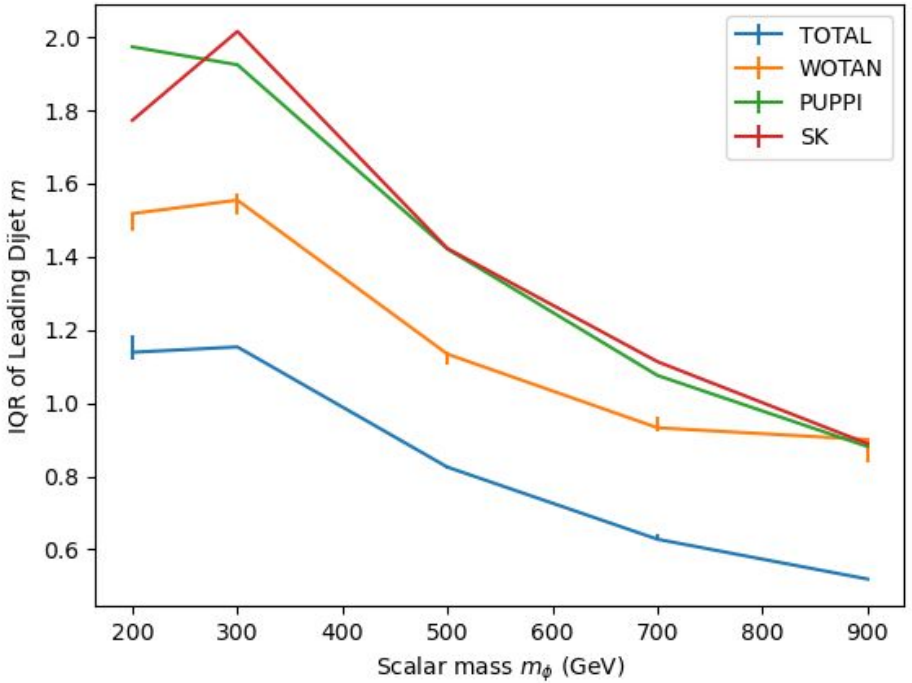

TOTAL
WOTAN
SK
PUPPI
Normalized entries
Leading Jet $(p_T^{true} - p_T^{predicted})/p_T^{true}$ (%)


TOTAL
SK
PUPPI
Truth only
WOTAN
Normalized entries
0.0
0.0
0.0
0.0
0.0
0.0
200
300
400
500
600
700
800
900
1000
Leading Dijet m


TOTAL
WOTAN
SK
PUPPI
Normalized entries
0.3
0.2
0.2
0.2
0.1
0.1
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Leading Dijet $(m^{true} - m^{predicted})/m^{true}$ (%)

IQR	TOTAL	WOTAN	SoftKiller	PUPPI
Leading Jet p_T Resolution	0.967	1.704	1.749	1.714
Leading Jet m Resolution	1.117	1.922	2.112	2.143
Leading Dijet m Resolution	0.793	1.117	1.422	1.421


TOTAL
WOTAN
PUPPI
SK
IQR of Leading Dijet m
2.5
2.0
1.5
1.0
40
60
80
100
120
140
160
180
Number of simultaneous interactions μ


TOTAL
WOTAN
PUPPI
SK
IQR of Leading Dijet m
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
200
300
400
500
600
700
800
900
Scalar mass m_ϕ (GeV)

Key Takeaways

01 WOTAN is a completely data-driven pileup mitigation technique

02 WOTAN outperforms conventional pileup mitigation strategies without requiring unphysical supervision

03 WOTAN is generalizable to any denoising problem that matches the outline discussed in this talk

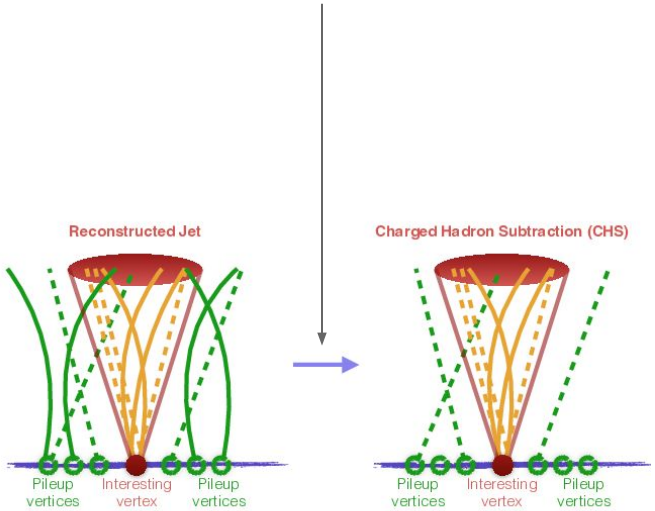
Stay tuned!
Final results to be released soon
arXiv 2411.XXXXX

Backup Slides

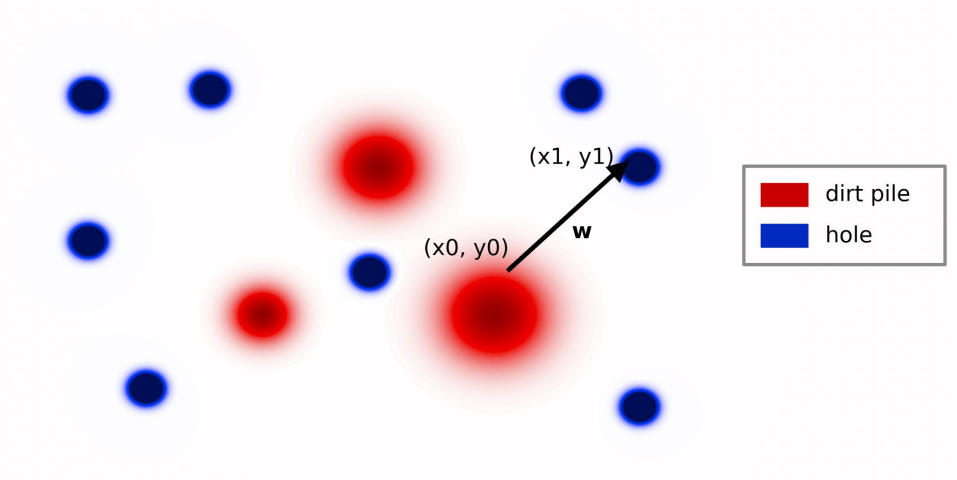
Charged Hadron Subtraction

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

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



Earth Mover's Distance = W_1



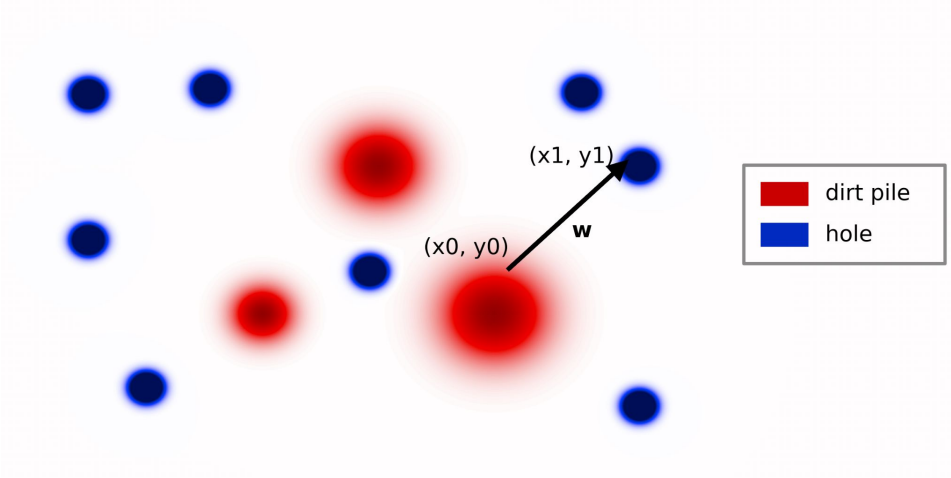
- Assumption: Total volume of the holes = total volume of the dirt piles
- Piles as the probability density function of P and holes as the probability density function of Q
- Per unit transportation cost:

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

- Transportation Plan:

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

Earth Mover's Distance = W_1



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

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

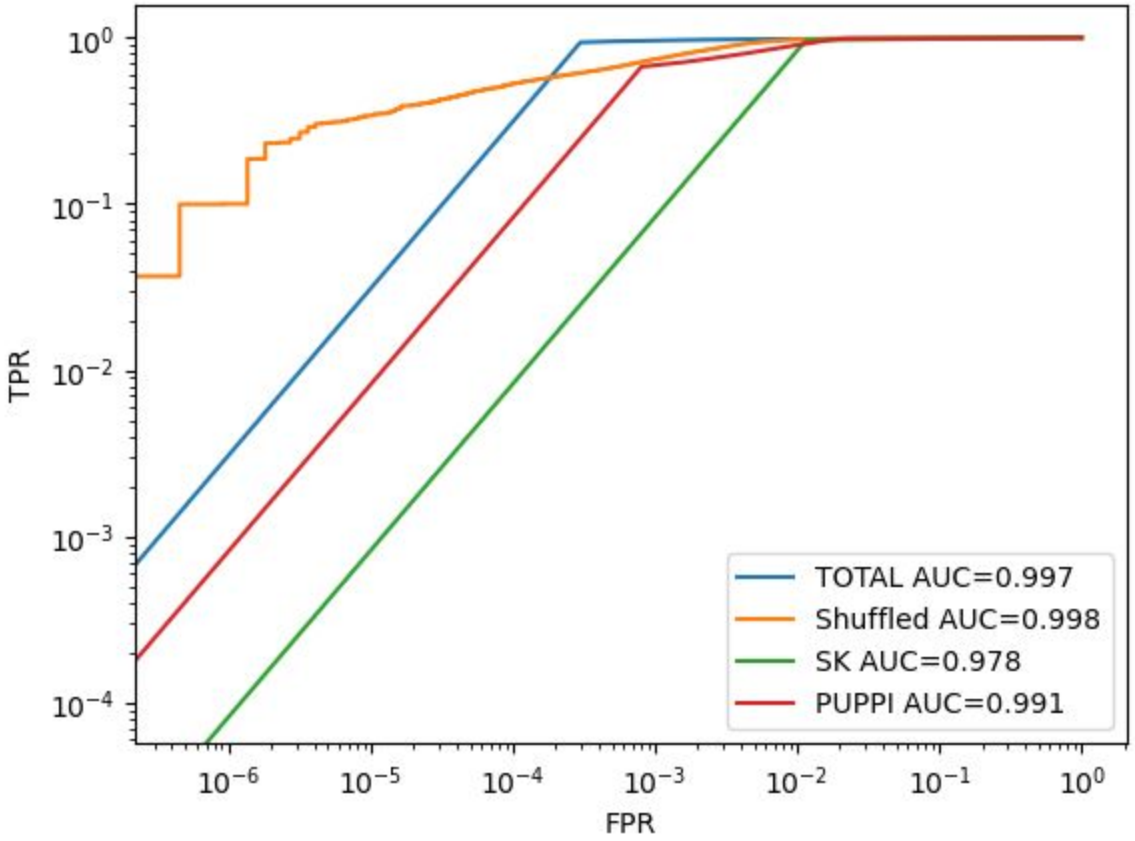
$$\text{Total Cost} = \int\int\int\int C(x_0, y_0, x_1, y_1) \cdot T(x_0, y_0, x_1, y_1)dx_0dy_0dx_1dy_1$$

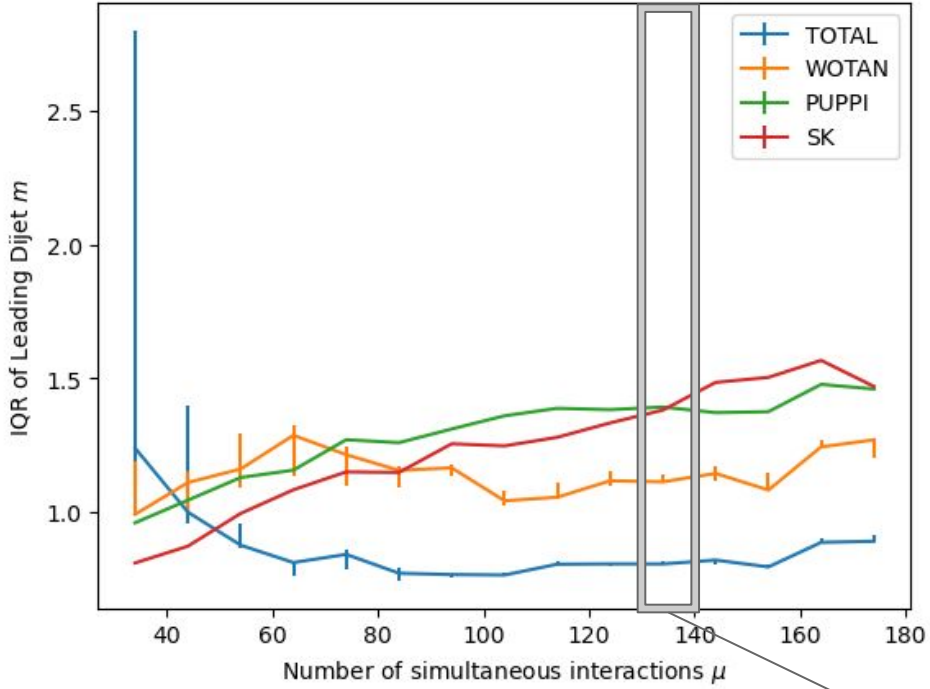
TOTAL: Training Optimal Transport with Attention Learning

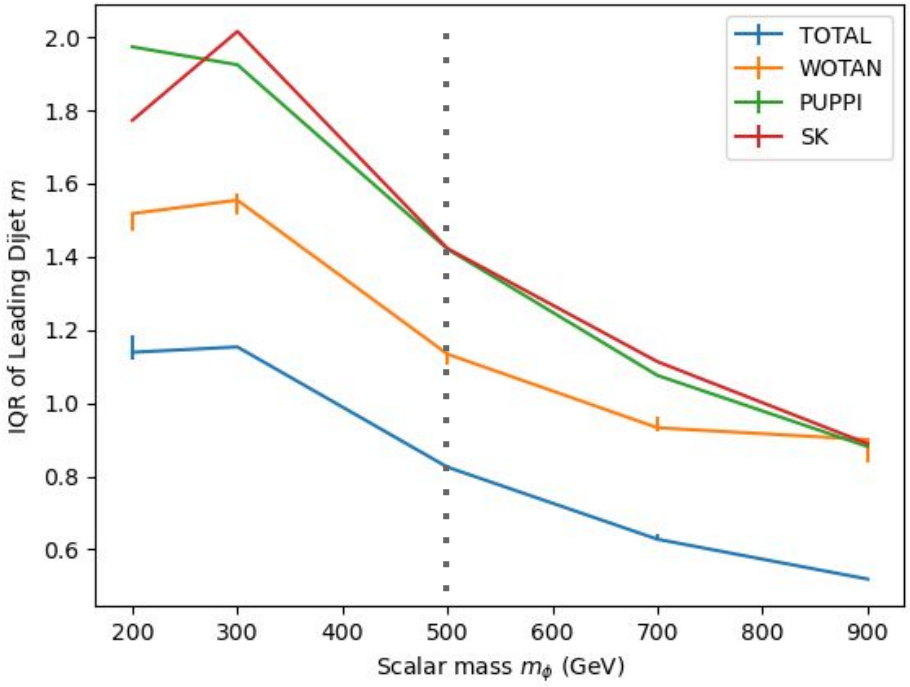
$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\,\mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

Modification for jet-based dataset (PUMML)

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


TPR
FPR
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
TOTAL AUC=0.997
Shuffled AUC=0.998
SK AUC=0.978
PUPPI AUC=0.991





**Initial training ranges*