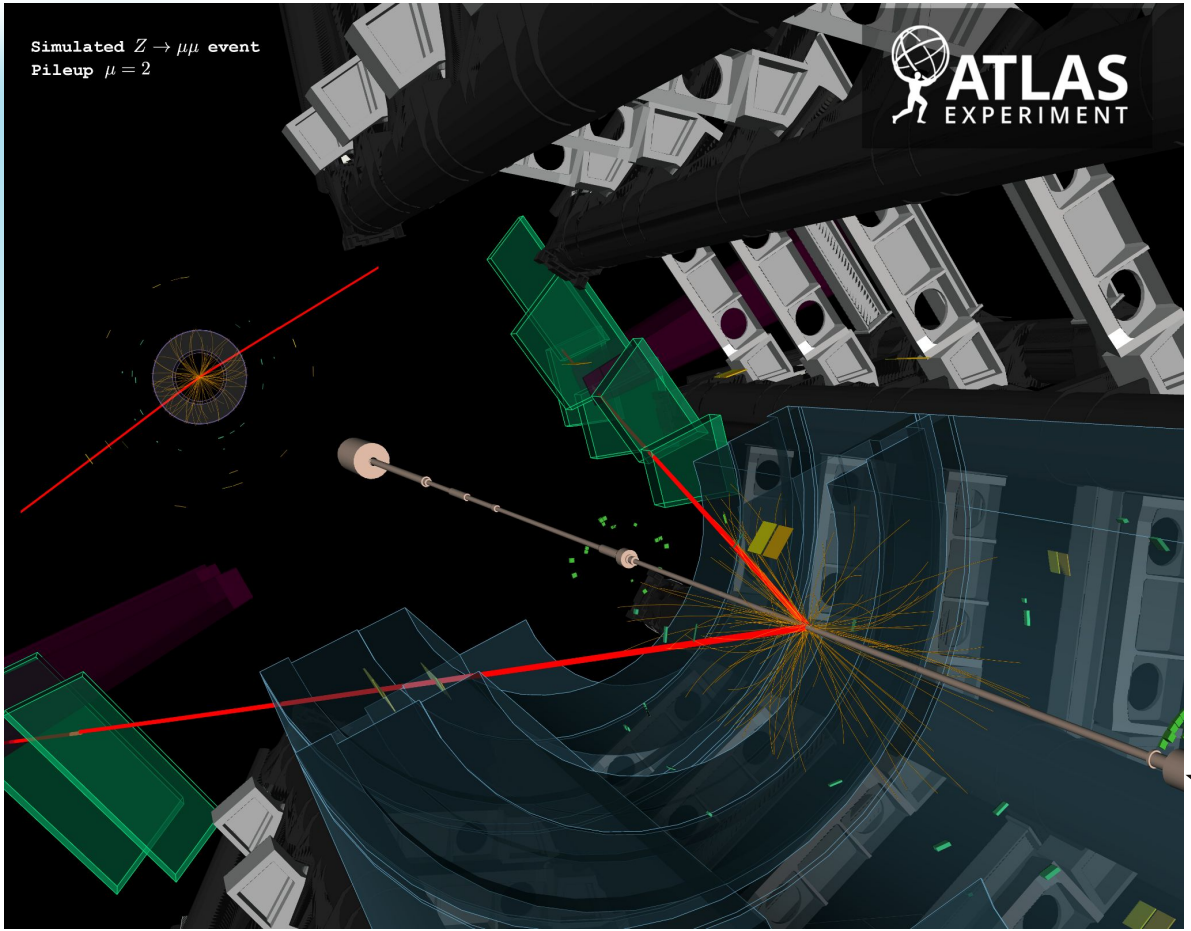
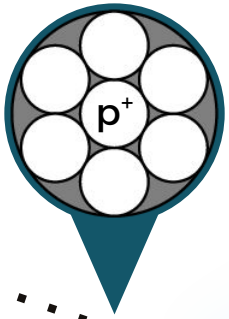
Nathan Suri, Vinicius Mikuni, Ben Nachman
ML4Jets 2026

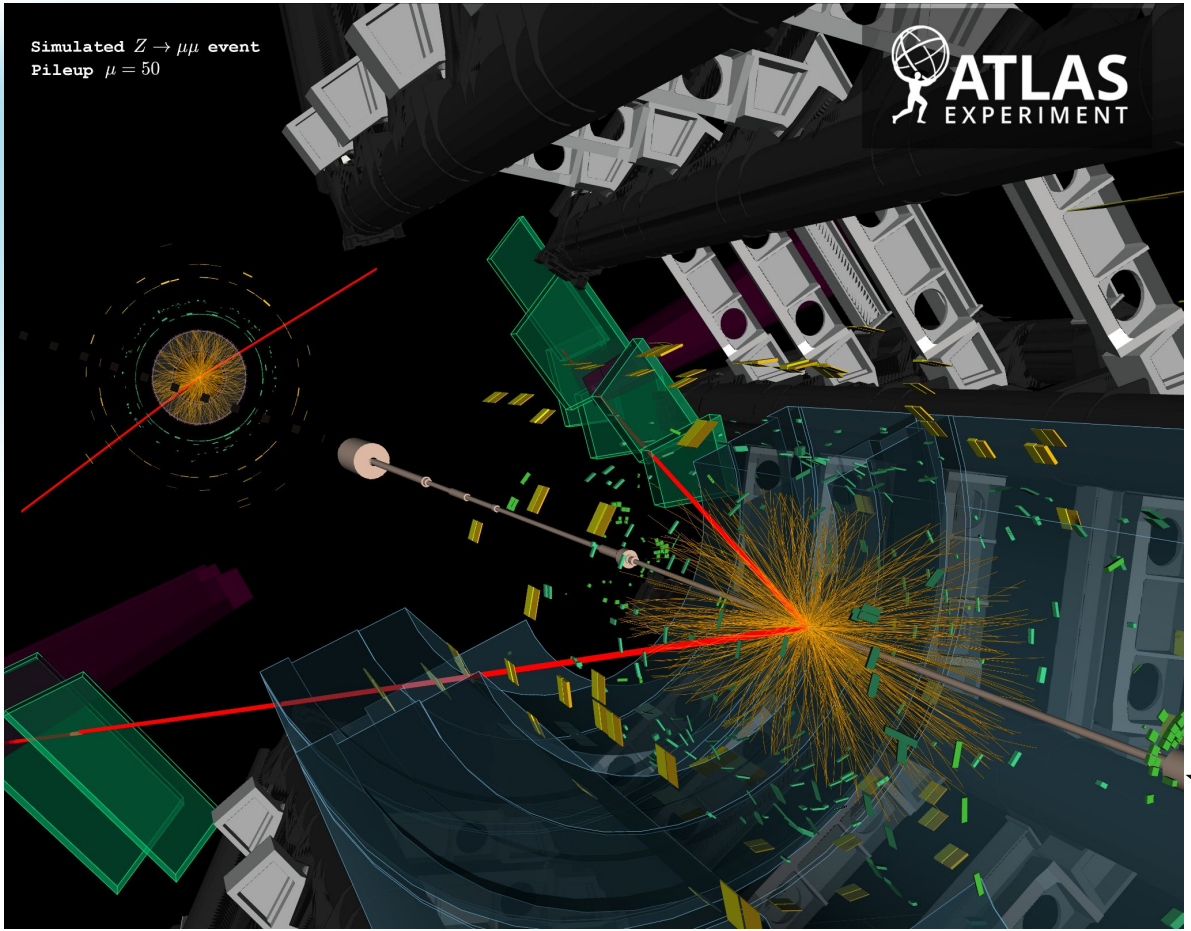
Optimizing Optimal Transport-based Pileup Mitigation



Pileup Descriptors

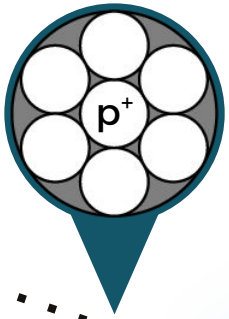
1. Charged versus neutral
2. In-time versus out-of-time

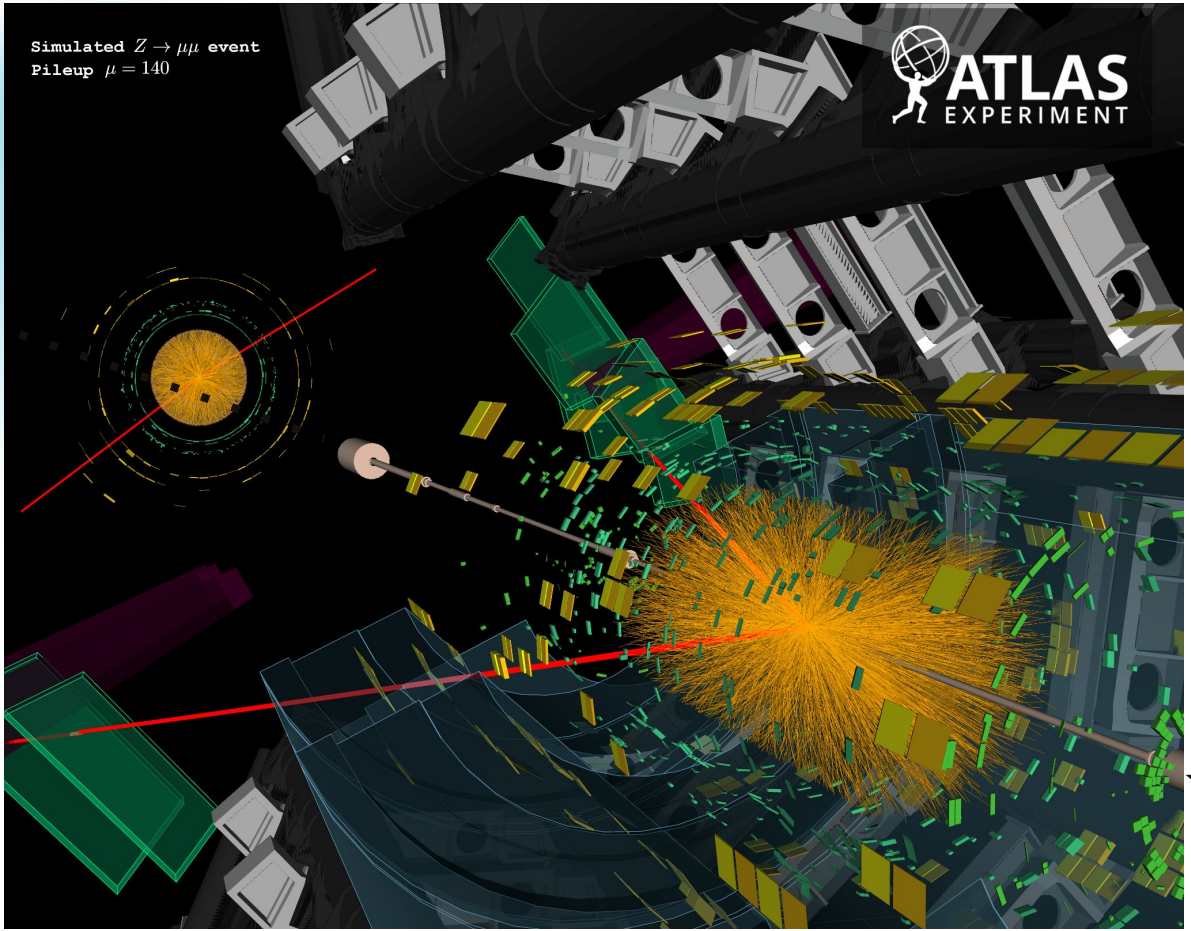




Pileup Descriptors

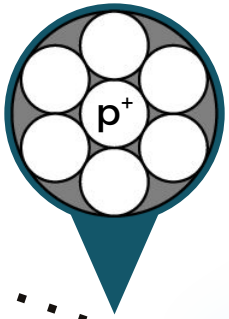
1. Charged versus neutral
2. In-time versus out-of-time

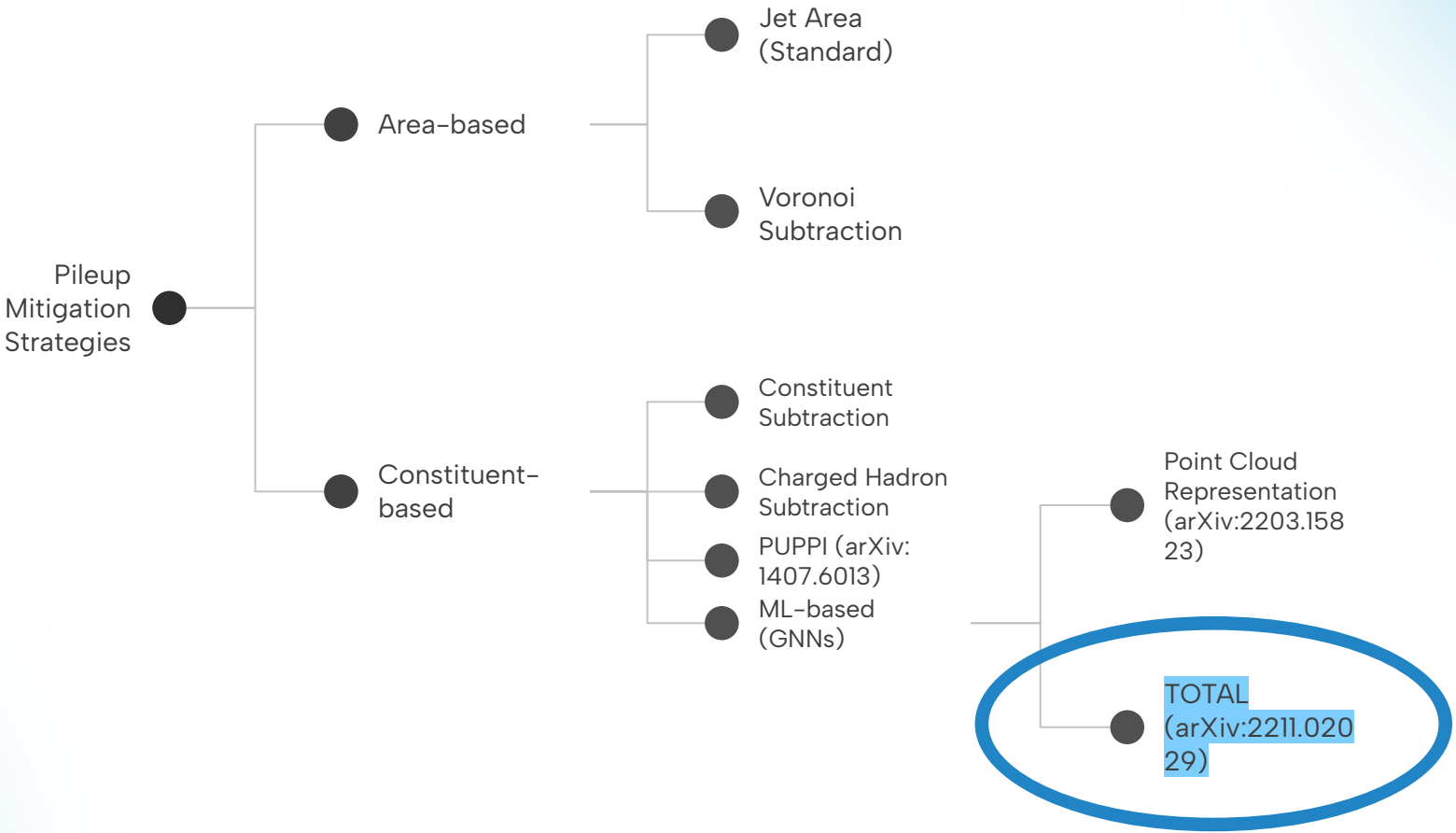




Pileup Descriptors

1. Charged versus neutral
2. In-time versus out-of-time

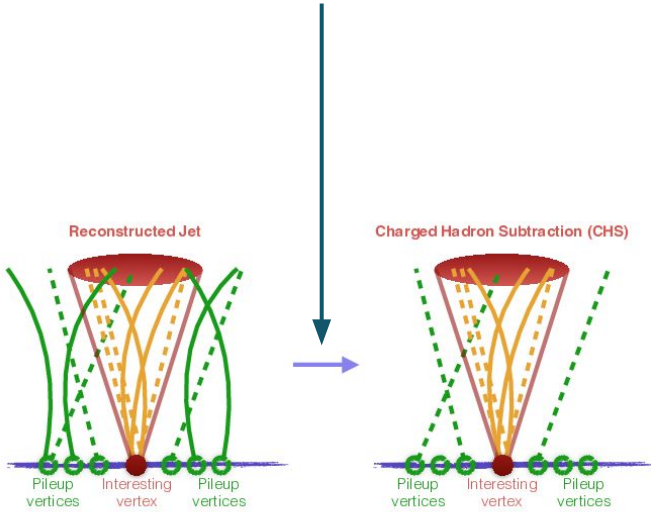



Pileup Mitigation Strategies
Area-based
Jet Area (Standard)
Voronoi Subtraction
Constituent-based
Constituent Subtraction
Charged Hadron Subtraction
PUPPI (arXiv: 1407.6013)
ML-based (GNNs)
Point Cloud Representation (arXiv:2203.15823)
TOTAL (arXiv:2211.02029)

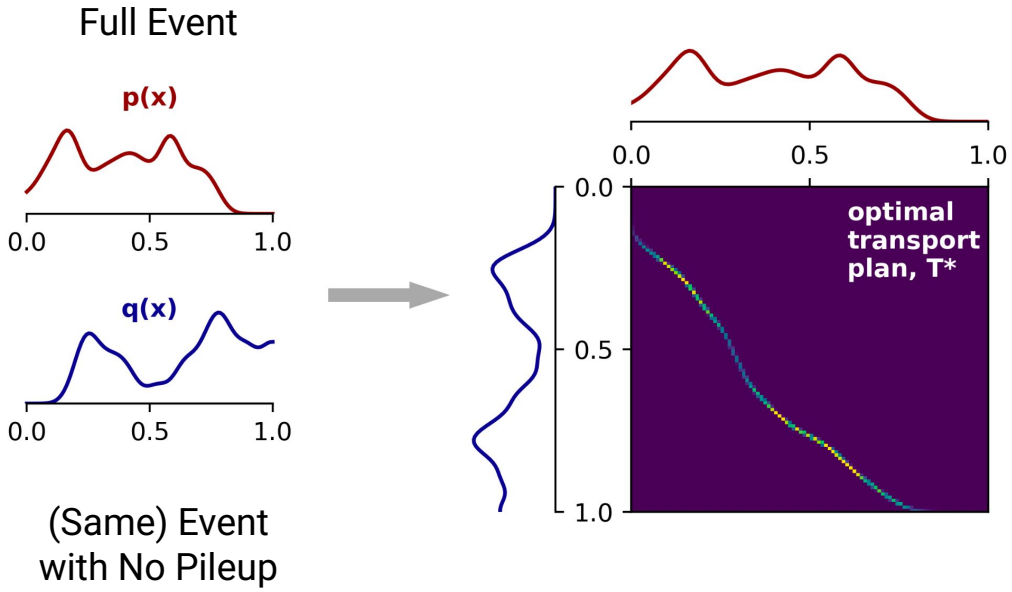
Charged Hadron Subtraction

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

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



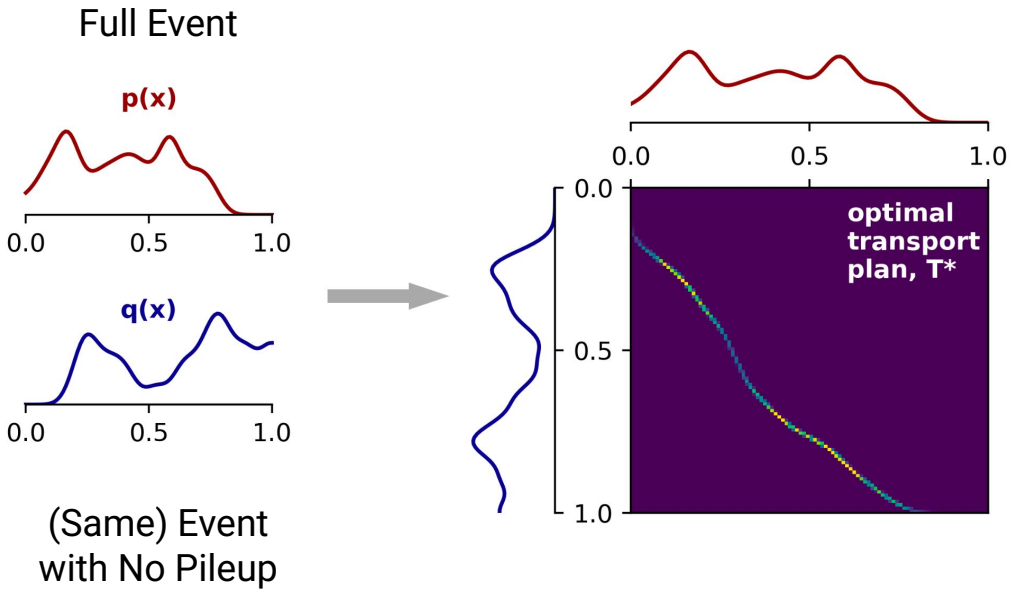
TOTAL: Training Optimal Transport with Attention Learning



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

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

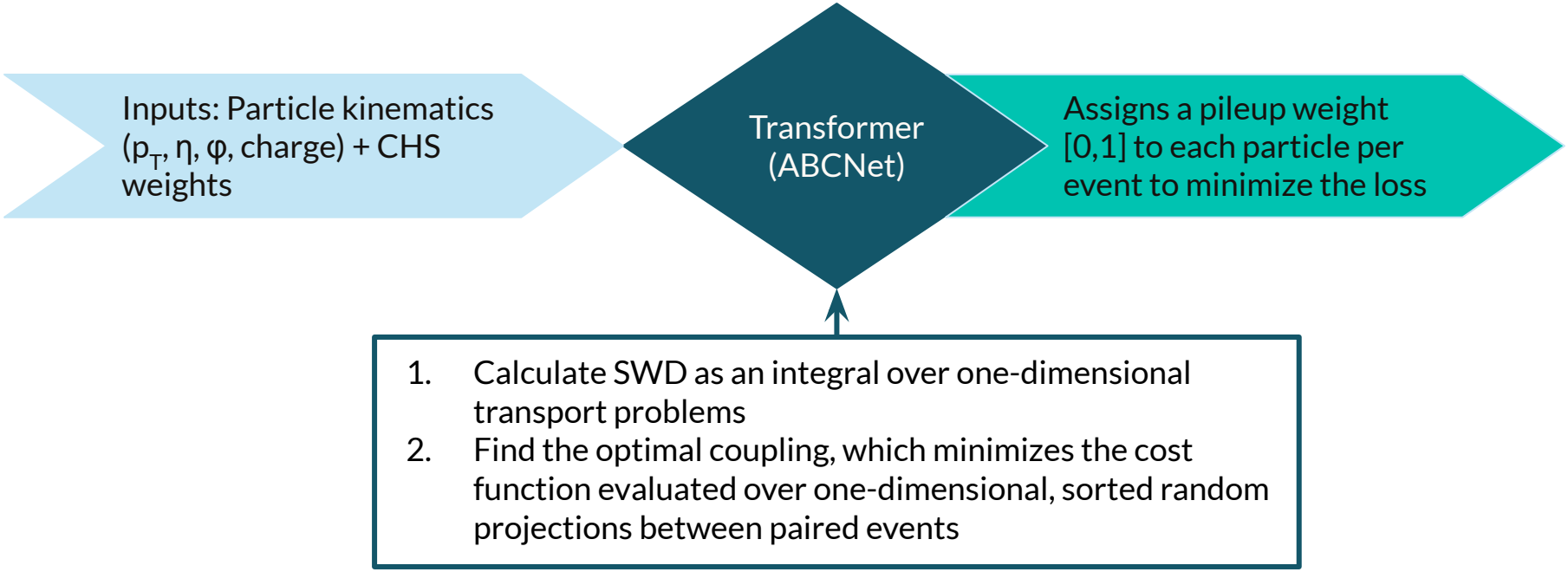
$$x'_p = \omega x_p$$

TOTAL: Training Optimal Transport with Attention Learning

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

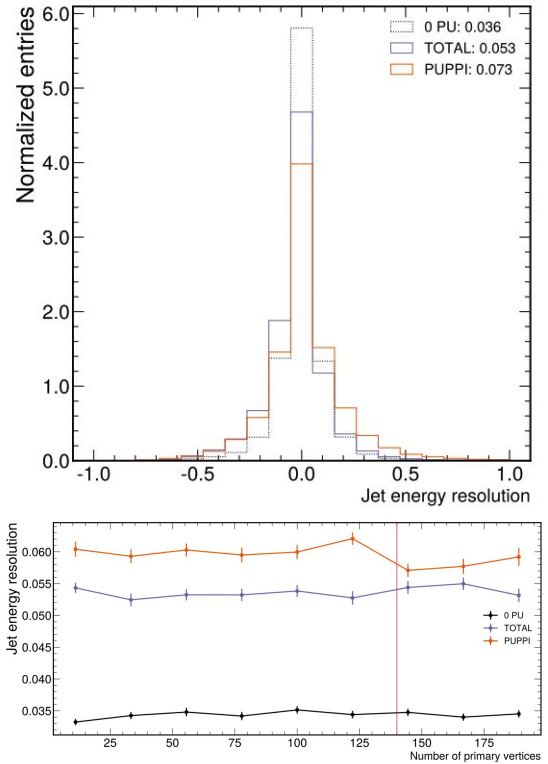
- Scaled Mean Square Error of missing p_T
- Forces energy conservation between the pure and full samples

TOTAL: Training Optimal Transport with Attention Learning



TOTAL: Training Optimal Transport with Attention Learning

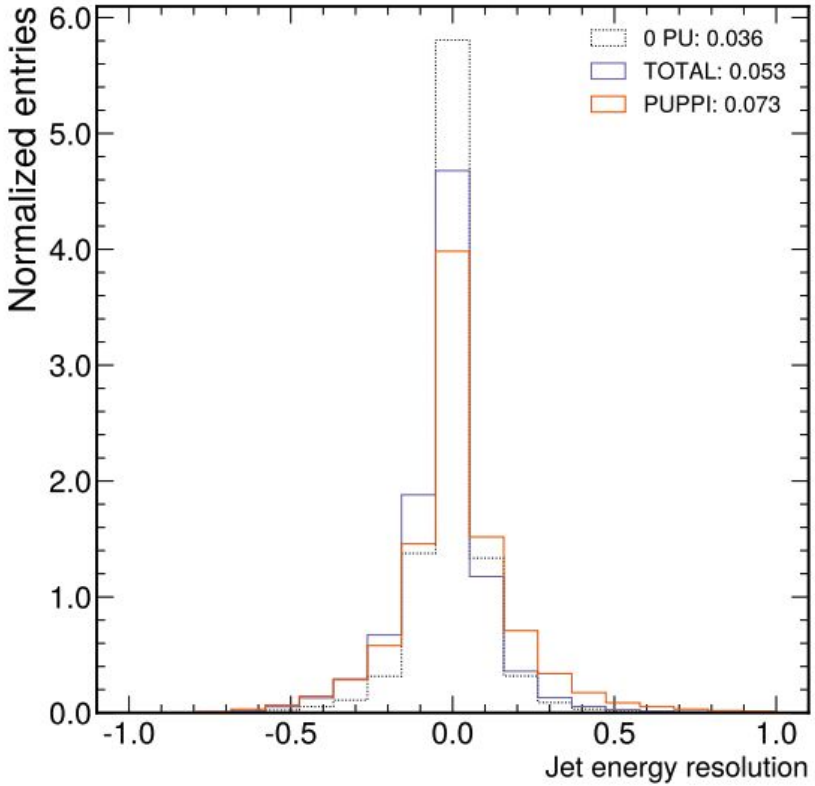
+ Outperforms traditional and ML-based alternatives
+ Relies on global event descriptions without MC corrections
+ Robustly learns pileup characteristics as a transport function



- Requires direct matching of events
- Limited due to supervision

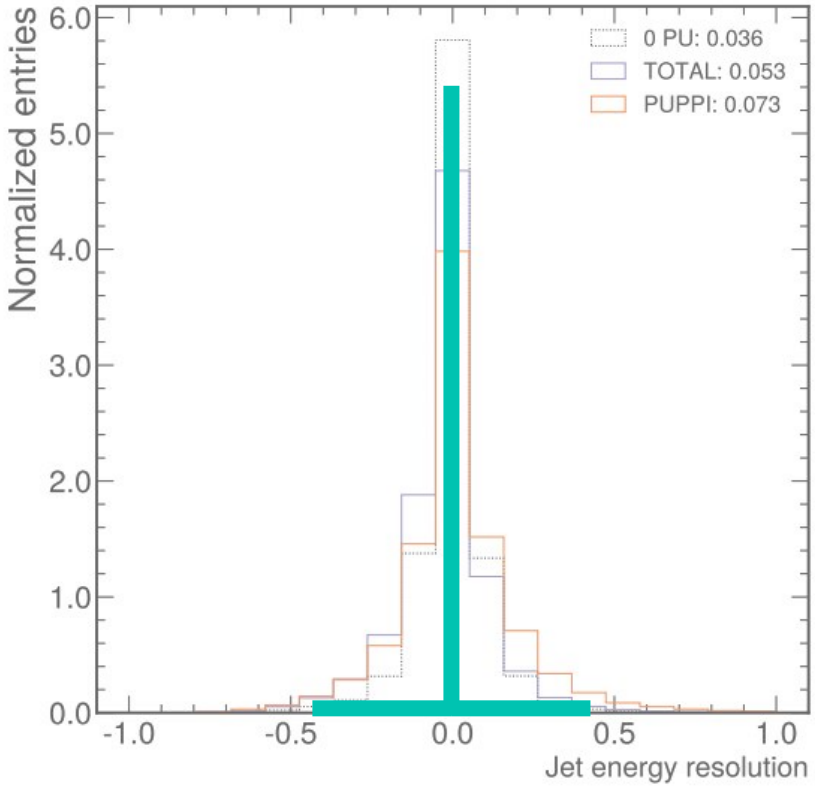
(arXiv:2211.02029)

TOTAL: Training Optimal Transport with Attention Learning



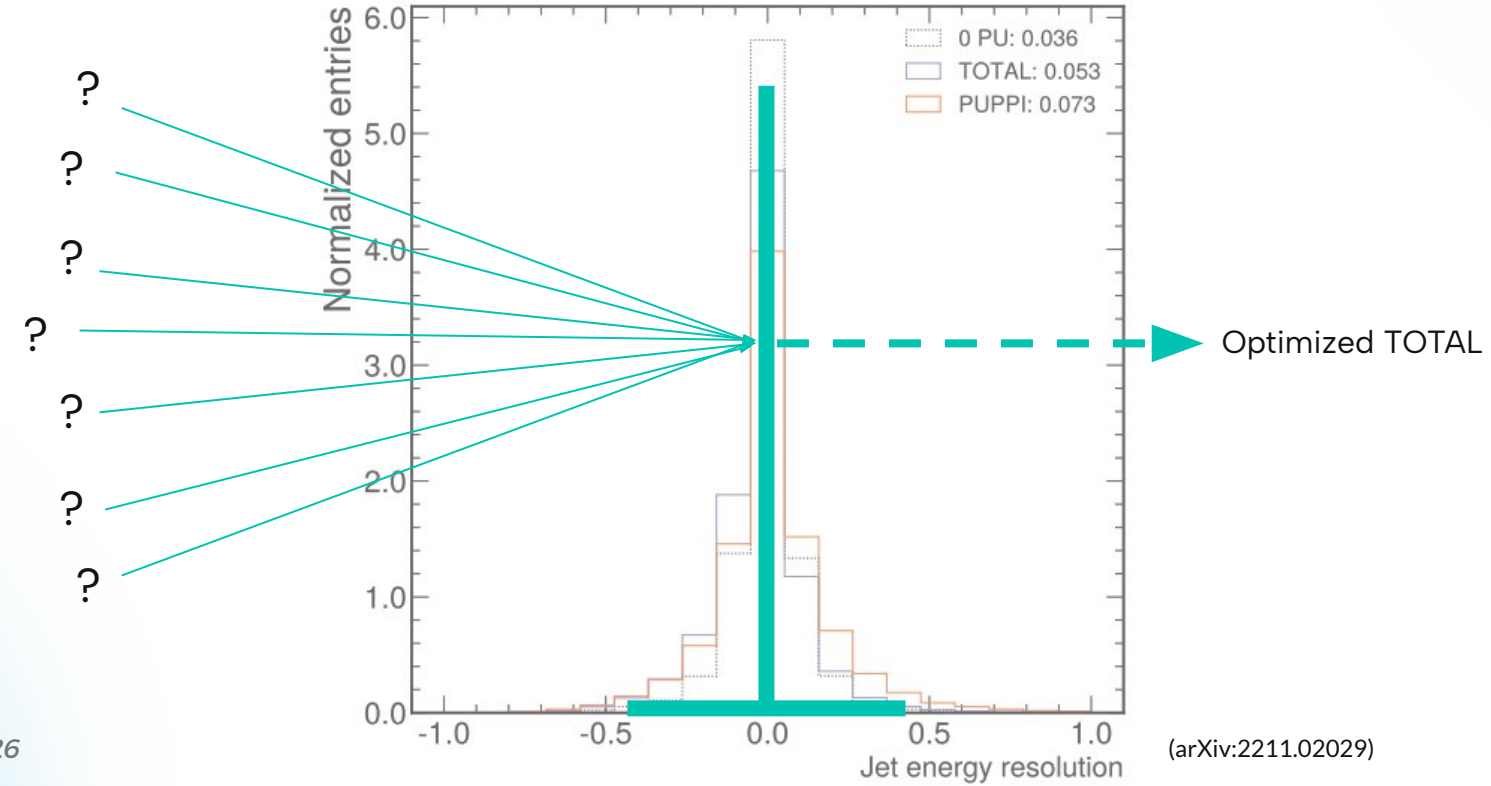
(arXiv:2211.02029)

TOTAL: Training Optimal Transport with Attention Learning



(arXiv:2211.02029)

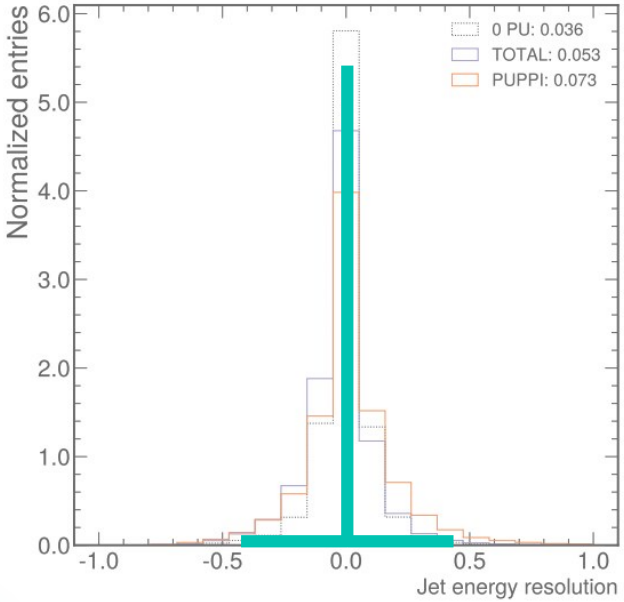
TOTAL: Training Optimal Transport with Attention Learning



(arXiv:2211.02029)

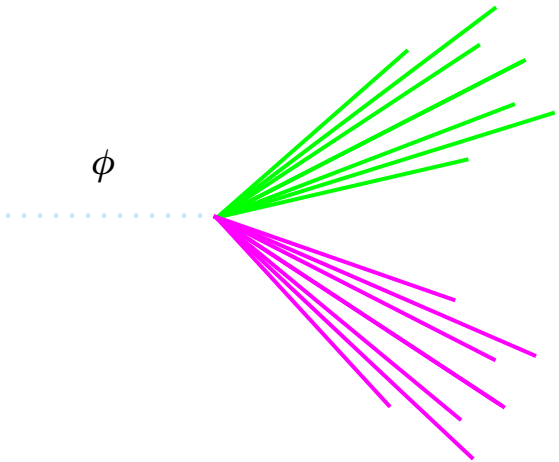
TOTAL Upgrades

1. Physics-based constraints
2. Alternative loss functions



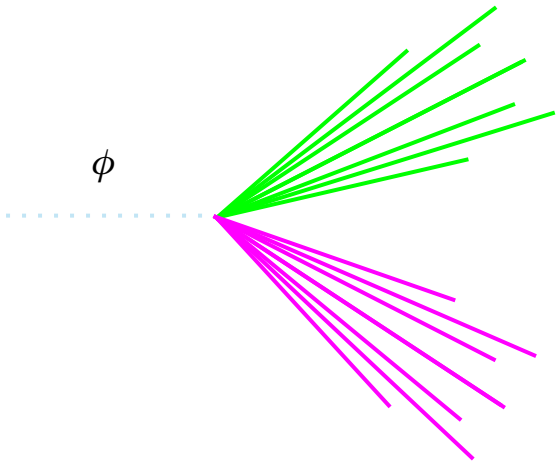
(arXiv:2211.02029)

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

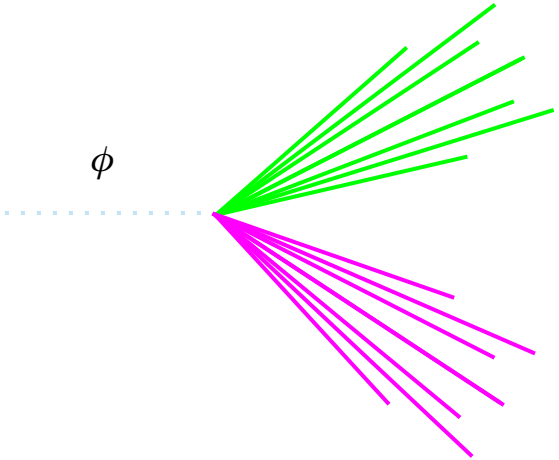
- Datasets
  - μ = 140, Δm: Set pileup vertex count, varied scalar mass
  - Δμ, m = 500 GeV: Varied pileup vertex (PV) count, set scalar mass

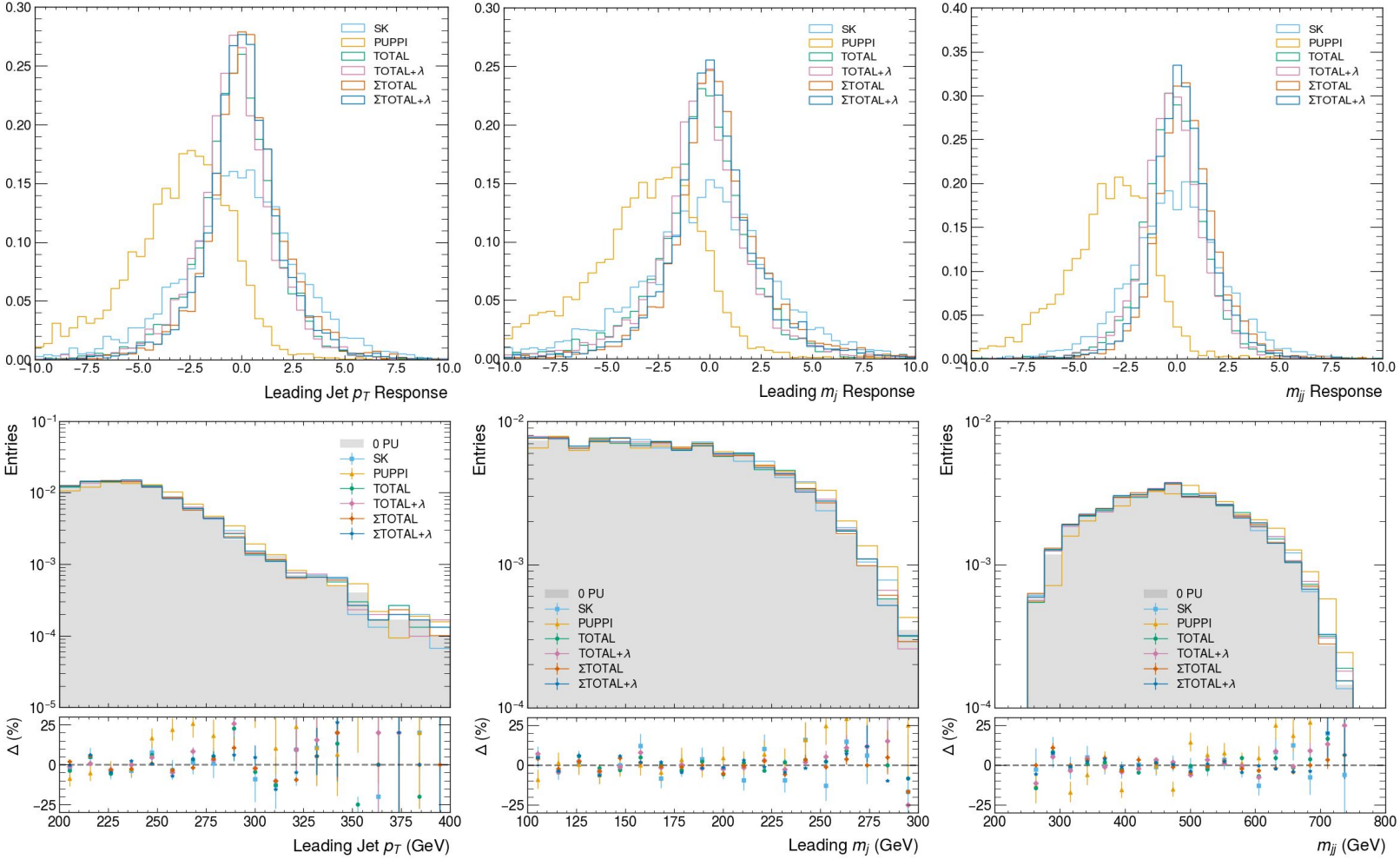
Physics-based Constraints

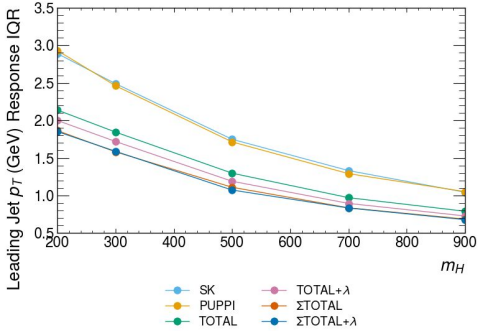
1. **Event-level constraints (λ):** Add MET and H_T constraints to the loss to control and regularize event-level energy scales

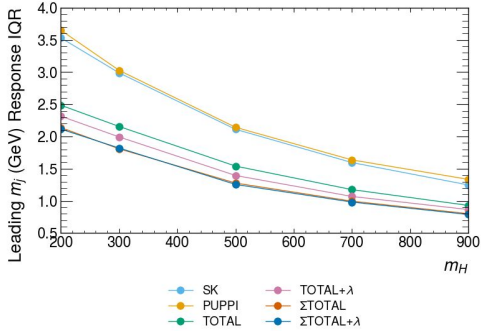
2. **Constituent splitting (Σ):** Sum two separate losses for charged and neutral constituents

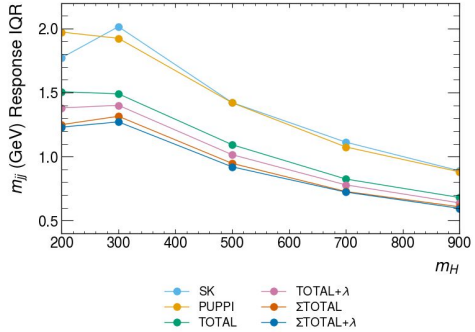
3. **Activation function*:** Adjust from using a hard sigmoid to a sigmoid activation function

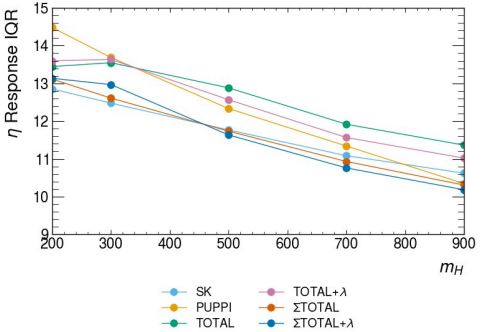


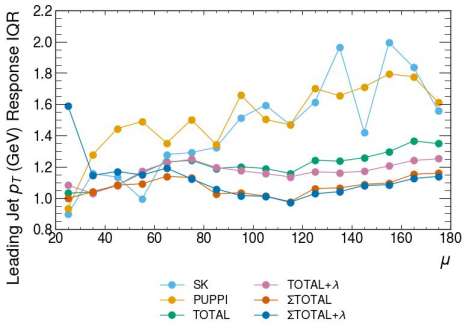

SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ
Leading Jet p_T Response
Leading m_j Response
m_{jj} Response
Entries
0 PU
Δ (%)
Leading Jet p_T (GeV)
Leading m_j (GeV)
m_{jj} (GeV)

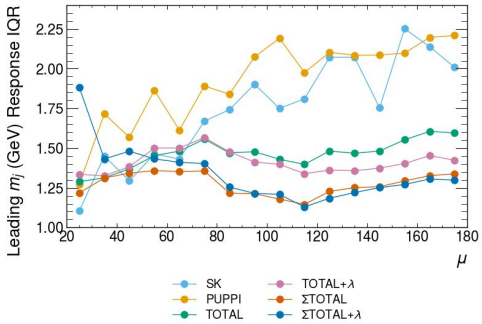

Leading Jet p_T (GeV) Response IQR
m_H
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ

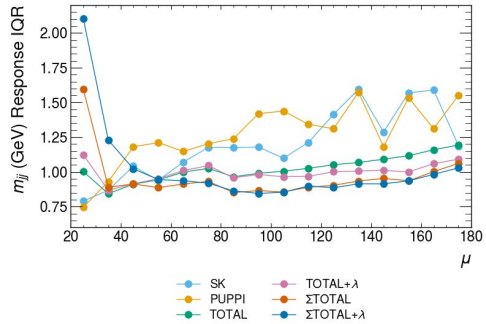

Leading m_j (GeV) Response IQR
m_H
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ

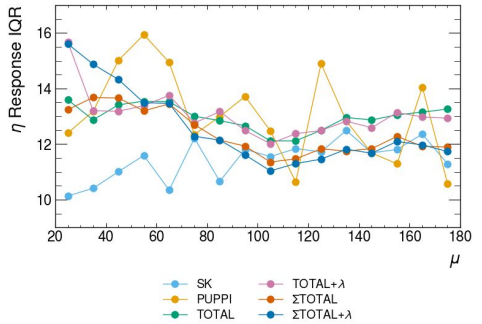

m_{jj} (GeV) Response IQR
m_H
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ


η Response IQR
m_H
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ


Leading Jet p_T (GeV) Response IQR
μ
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ

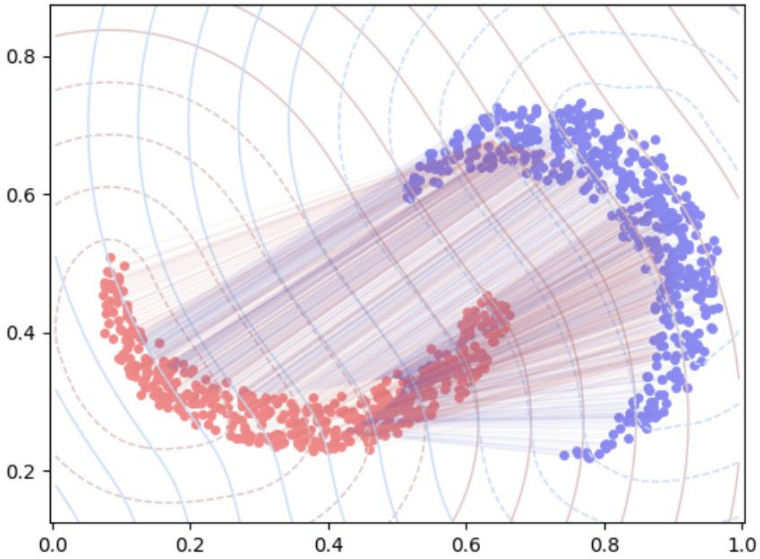

Leading m_j (GeV) Response IQR
μ
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ

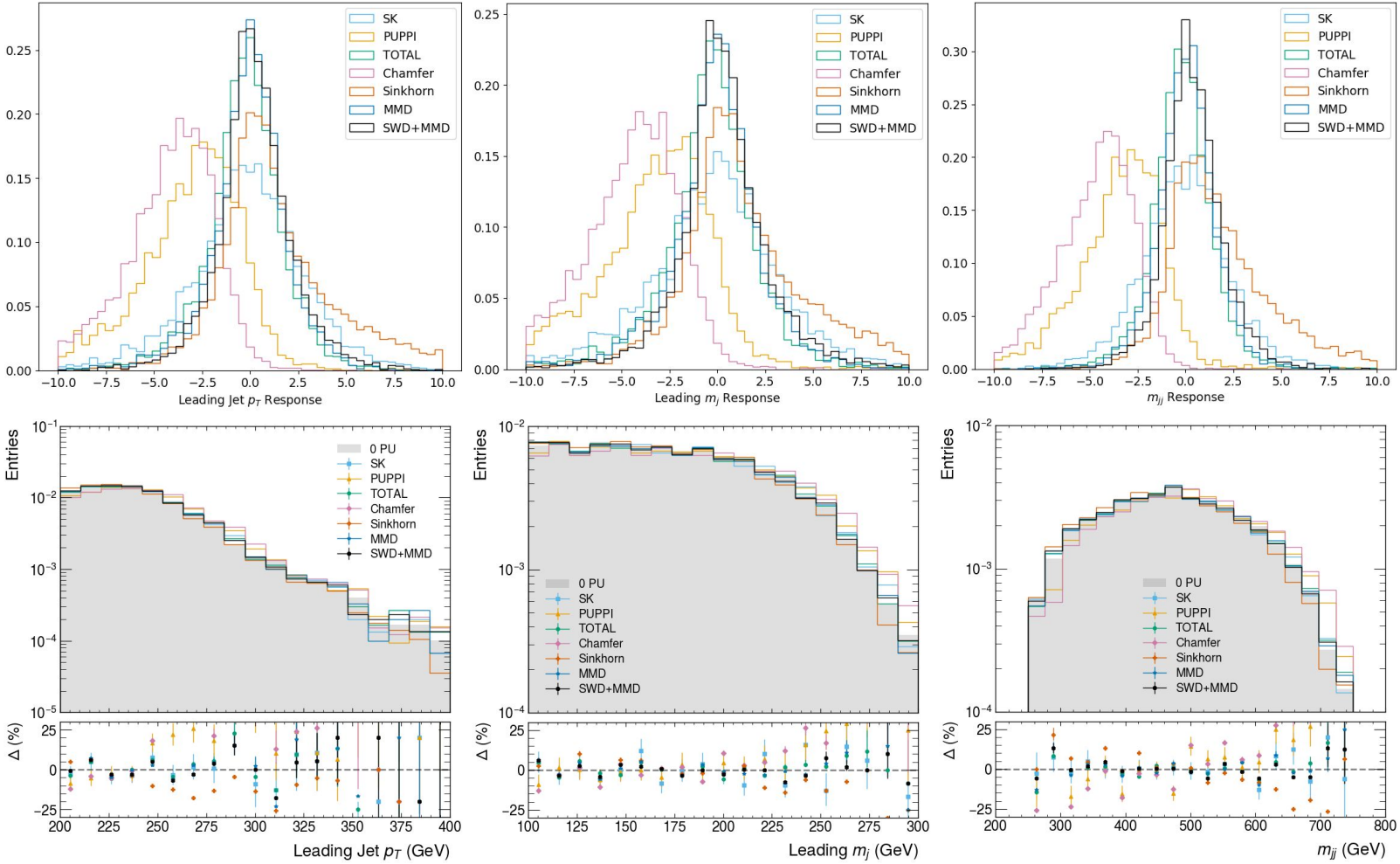

m_{jj} (GeV) Response IQR
μ
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ

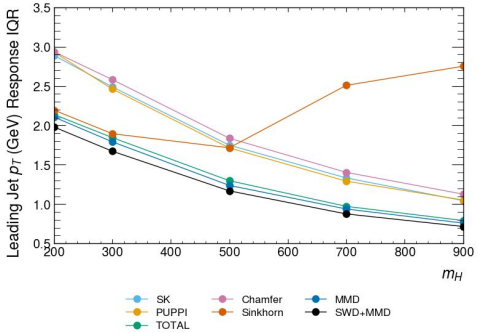

η Response IQR
μ
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ

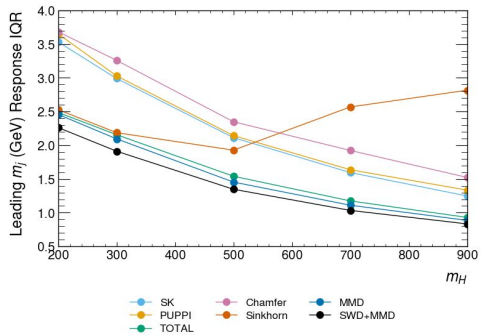
(Alternative) Loss Functions

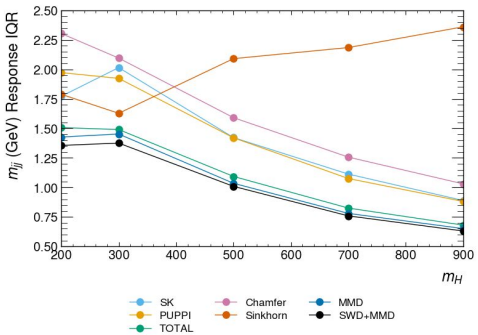
1. **Sliced Wasserstein Distance (SWD)**
2. Chamfer distance
3. Sinkhorn divergence
4. Maximum Mean Discrepancy (MMD)
    a. MMD + SWD

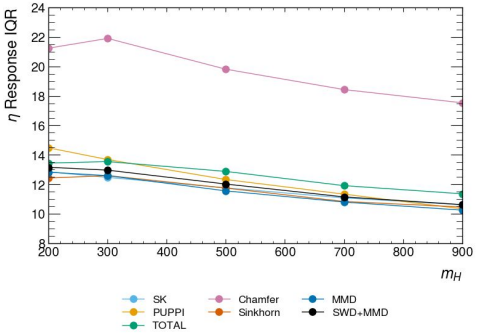


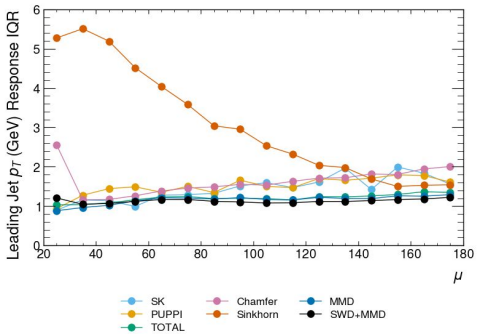

SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD
Leading Jet p_T Response
Leading m_j Response
m_{jj} Response
Entries
0 PU
Δ (%)
Leading Jet p_T (GeV)
Leading m_j (GeV)
m_{jj} (GeV)

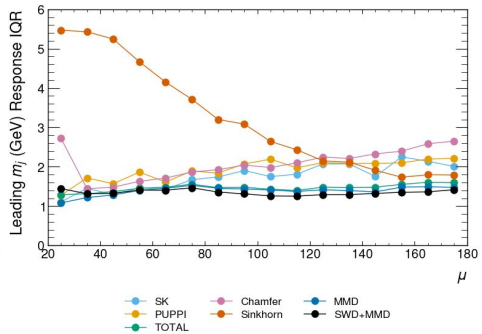

Leading Jet p_T (GeV) Response IQR
m_H
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD

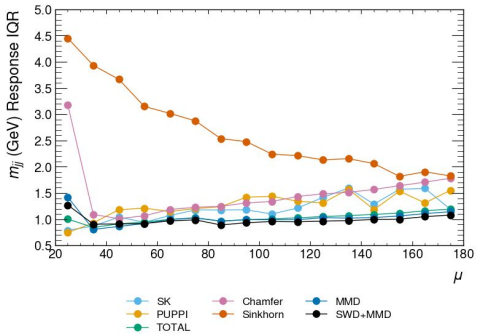

Leading m_j (GeV) Response IQR
m_H
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD

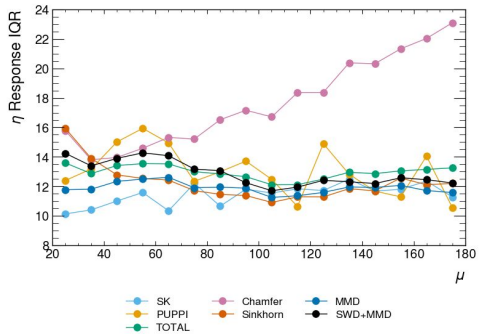

m_{jj} (GeV) Response IQR
m_H
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD


η Response IQR
m_H
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD

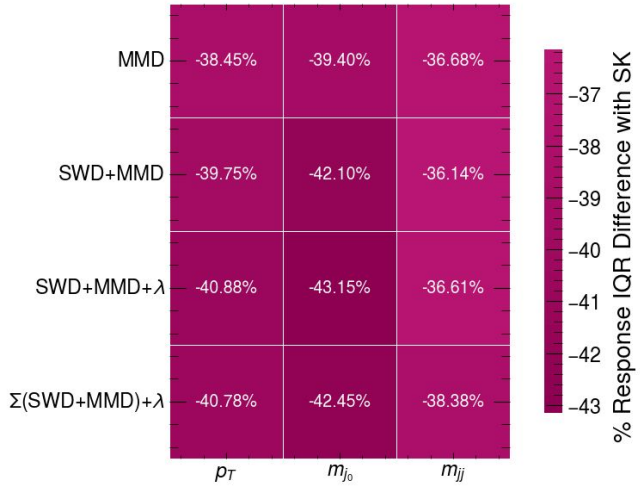

Leading Jet p_T (GeV) Response IQR
μ
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD

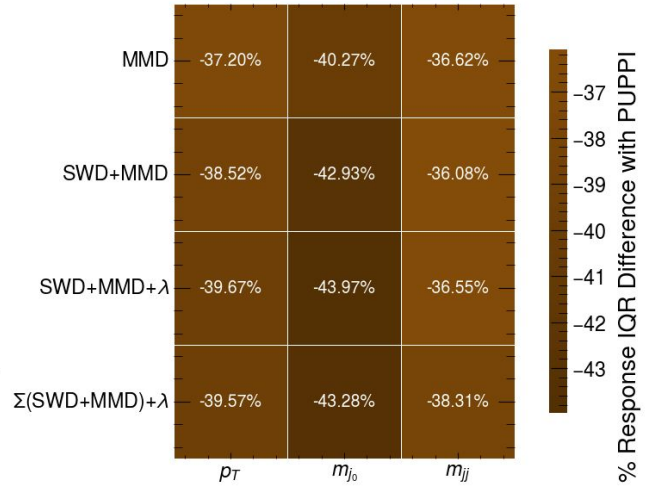

Leading m_j (GeV) Response IQR
μ
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD

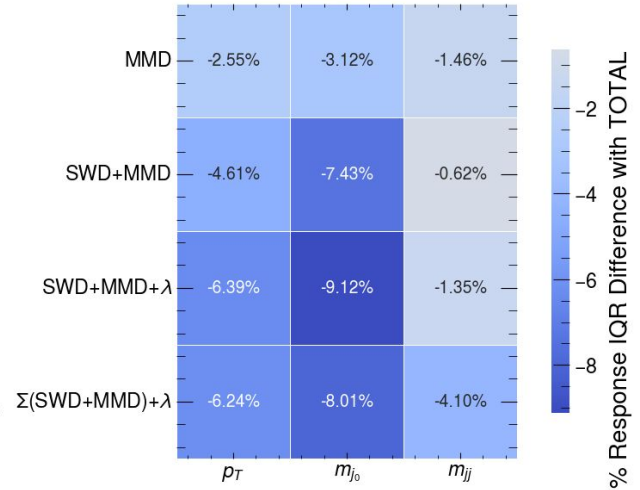

m_{jj} (GeV) Response IQR
μ
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD


η Response IQR
μ
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD

So what is the *total* TOTAL?


MMD
SWD+MMD
SWD+MMD+λ
Σ(SWD+MMD)+λ
-38.45%
-39.40%
-36.68%
-39.75%
-42.10%
-36.14%
-40.88%
-43.15%
-36.61%
-40.78%
-42.45%
-38.38%
p_T
m_{j_0}
m_{jj}
-37
-38
-39
-40
-41
-42
-43
% Response IQR Difference with SK


MMD
SWD+MMD
SWD+MMD+λ
Σ(SWD+MMD)+λ
-37.20%
-40.27%
-36.62%
-38.52%
-42.93%
-36.08%
-39.67%
-43.97%
-36.55%
-39.57%
-43.28%
-38.31%
p_T
m_{j_0}
m_{jj}
-37
-38
-39
-40
-41
-42
-43
% Response IQR Difference with PUPPI


MMD
SWD+MMD
SWD+MMD+λ
Σ(SWD+MMD)+λ
-2.55%
-3.12%
-1.46%
-4.61%
-7.43%
-0.62%
-6.39%
-9.12%
-1.35%
-6.24%
-8.01%
-4.10%
p_T
m_{j_0}
m_{jj}
-2
-4
-6
-8
% Response IQR Difference with TOTAL


MMD
SWD+MMD
SWD+MMD+λ
Σ(SWD+MMD)+λ
-38.45%
-39.40%
-36.68%
-39.75%
-42.10%
-36.14%
-40.88%
-43.15%
-36.61%
-40.78%
-42.45%
-38.38%
p_T
m_{j_0}
m_{jj}
-37
-38
-39
-40
-41
-42
-43
% Response IQR Difference with SK
MMD
SWD+MMD
SWD+MMD+λ
Σ(SWD+MMD)+λ
-37.20%
-40.27%
-36.62%
-38.52%
-42.93%
-36.08%
-39.67%
-43.97%
-36.55%
-39.57%
-43.28%
-38.31%
p_T
m_{j_0}
m_{jj}
-37
-38
-39
-40
-41
-42
-43
% Response IQR Difference with PUPPI
MMD
SWD+MMD
SWD+MMD+λ
Σ(SWD+MMD)+λ
-2.55%
-3.12%
-1.46%
-4.61%
-7.43%
-0.62%
-6.39%
-9.12%
-1.35%
-6.24%
-8.01%
-4.10%
p_T
m_{j_0}
m_{jj}
-2
-4
-6
-8
% Response IQR Difference with TOTAL

Key Takeaways

Stay tuned!
Final results to be released soon
arXiv 2607.XXXXX

01 TOTAL is a data-driven, optimal transport-based pileup mitigation technique

02 The addition of physics-based constraints and a new loss function boosts TOTAL's performance

03 Optimized TOTAL performs ~7% better than TOTAL and ~40% better than PUPPI and SK

Appendices

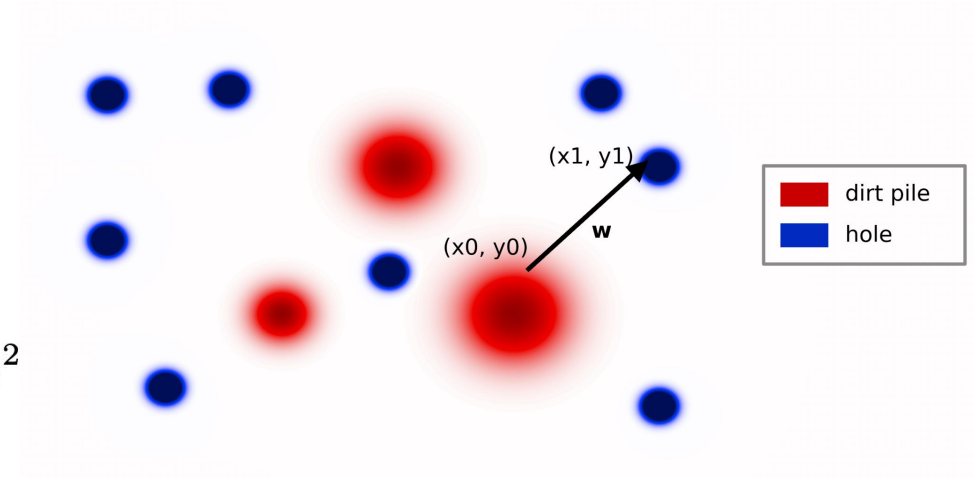
Earth Mover's Distance = W_1

- 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:

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



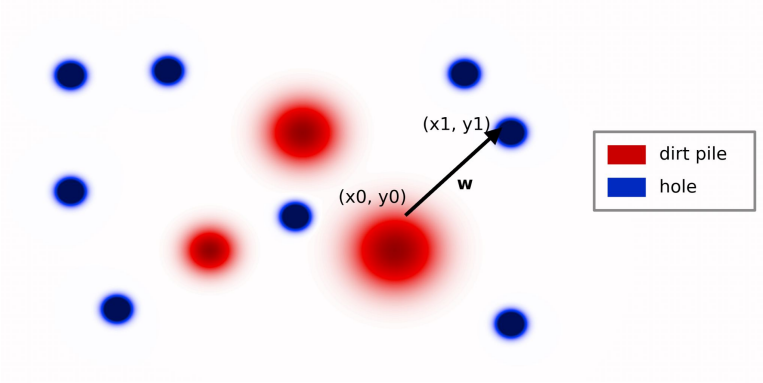
Earth Mover's Distance = W_1

- 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:

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

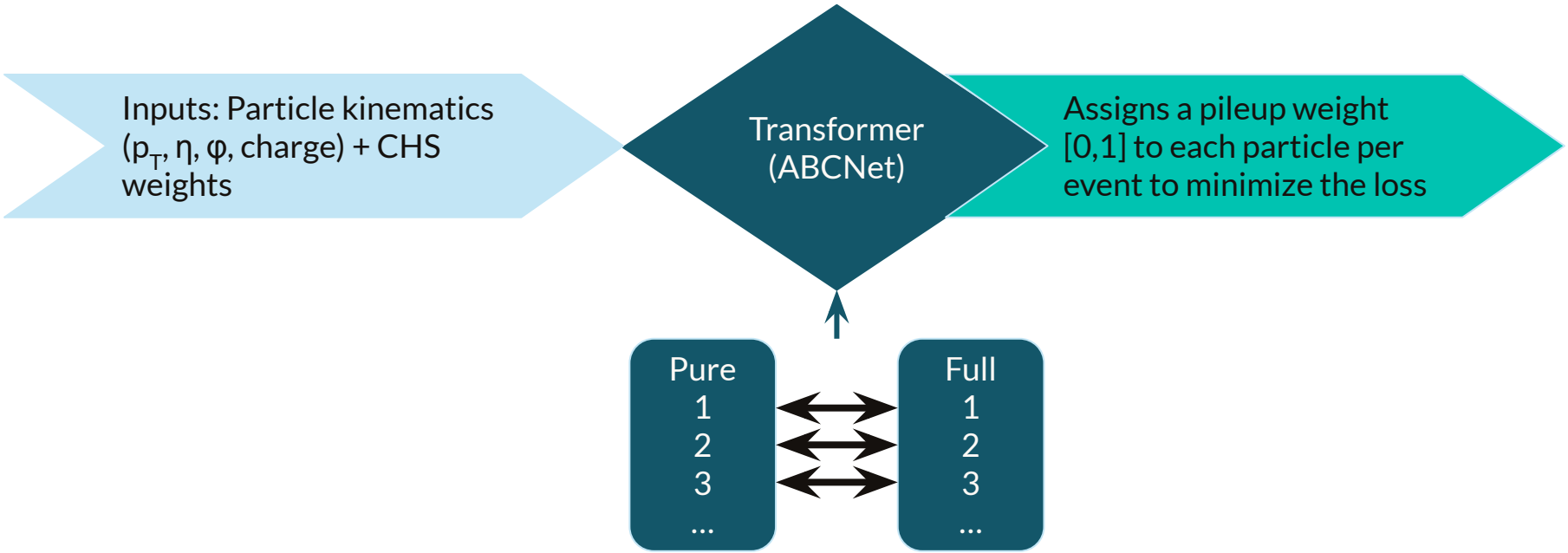


$$\int \int T(x_0, y_0, x, y) dx dy = p(x_0, y_0)$$

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

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

TOTAL: Training Optimal Transport with Attention Learning



TOTAL: Training Optimal Transport with Attention Learning

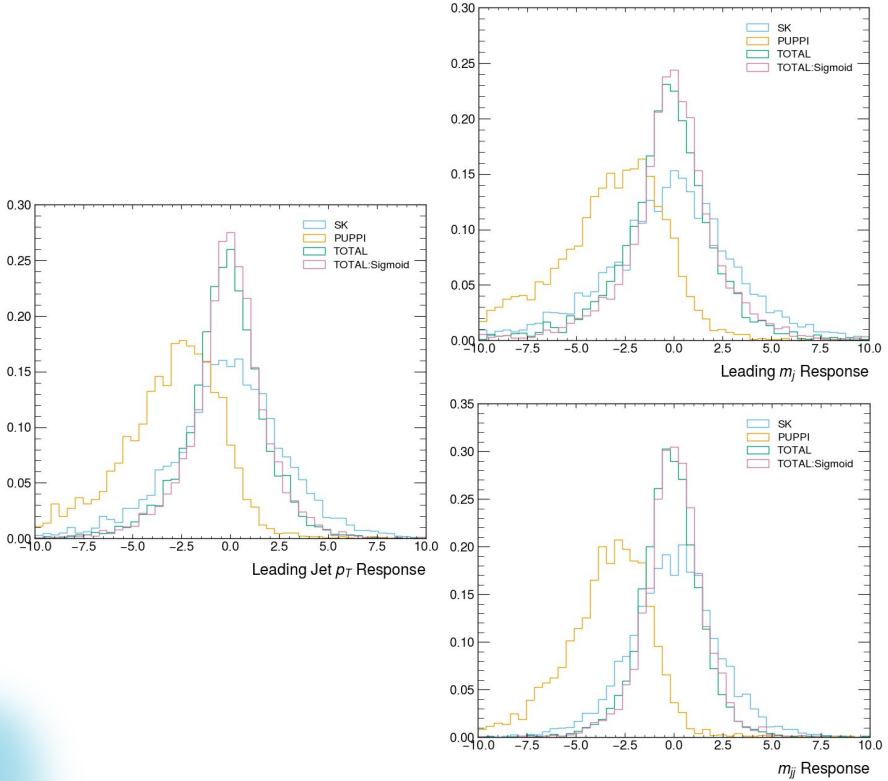
$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\,\mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np}))$$

Modification for jet-based dataset (PUMML)

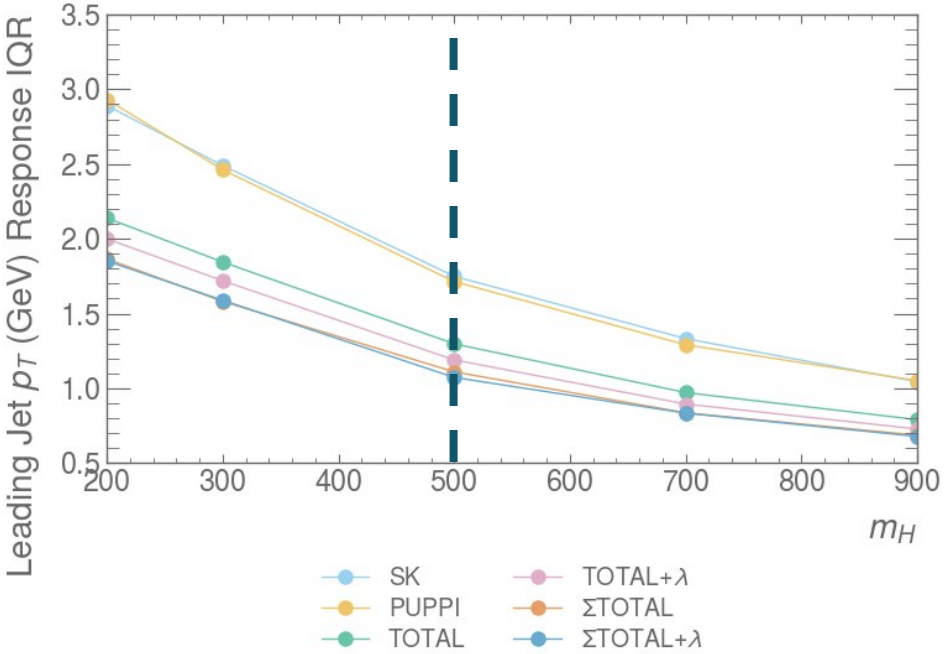
$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda\mathrm{MSE}(E_T(x'_p), E_T(x_{np}))$$

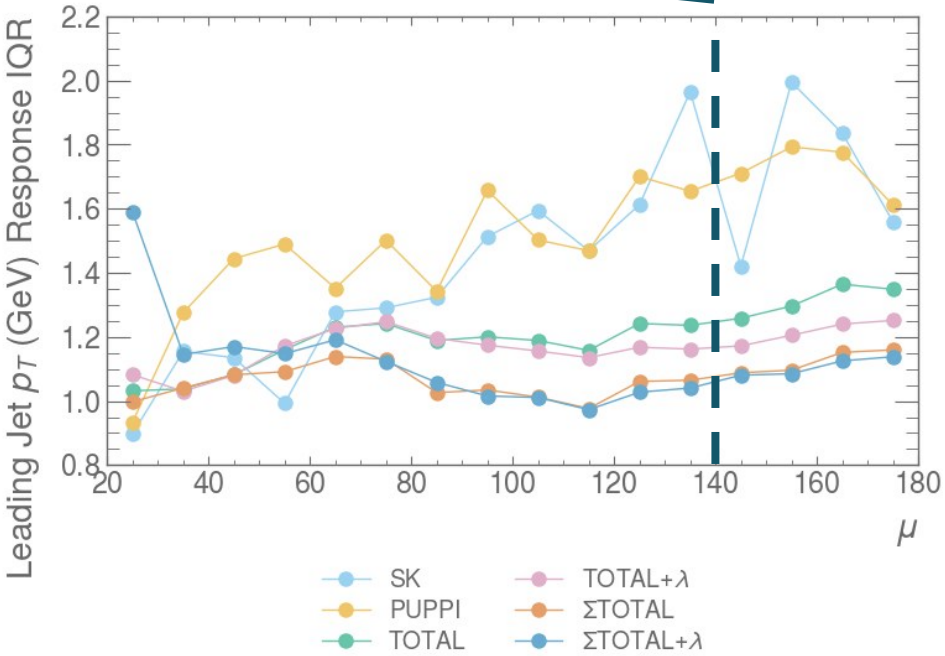
Physics-based Constraints: Activation Function

- Hard Sigmoid
  - Fast as the function is piecewise-linear and quantizes easily for low-precision inference
  - Forces a bimodal output distribution for the classifier labels
  - Events with differing pileup distributions can appear more similar due to the reduced resolution



**Initial training value*



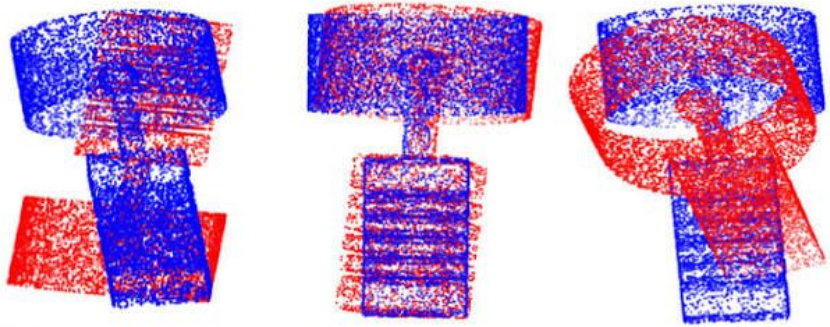


Alternative Loss Functions

1. **Chamfer distance**
   a. Nearest-neighbor matching function
   b. For every point in the first set, it finds and measures the distance to the closest point in the second set and vice versa before summing these distances together
   c. Simpler and faster than SWD

$$c(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} d(a, b),$$

$$d_C(A, B) = c(A, B) + c(B, A).$$



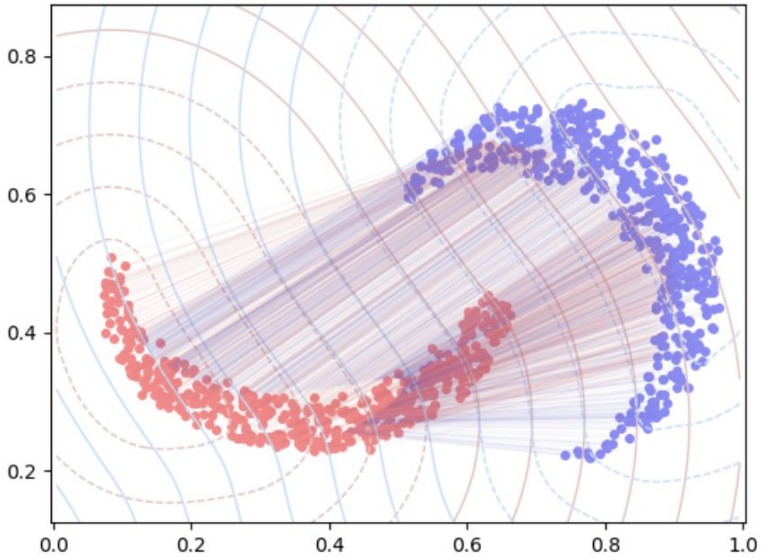
Rotation angle 6.58
Chamfer distance 0.12

Rotation angle 2.98
Chamfer distance 0.22

Rotation angle 4.59
Chamfer distance 0.12

Alternative Loss Functions

1. Chamfer distance
2. **Sinkhorn divergence**
    a. Computes a regularized transport coupling, allowing the metric to respect density in a superior fashion to the others including the projections of SWD.
    b. Well-conditioned, differentiable gradients
    c. Interpolates between true OT and MMD by ε



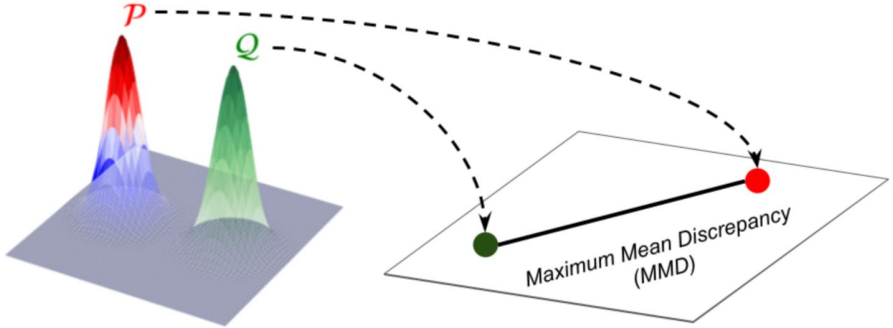
$$\mathrm{OT}_\varepsilon(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times X} c(x,y)\, d\pi(x,y) + \varepsilon\, \mathrm{KL}(\pi\|\mu\otimes\nu)$$

$$\mathrm{KL}(\pi\|\rho) = \int \frac{d\pi}{d\rho} \log \frac{d\pi}{d\rho} - 1\, d\rho$$

$$S_\varepsilon(\mu,\nu) := \mathrm{OT}_\varepsilon(\mu,\nu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu)$$

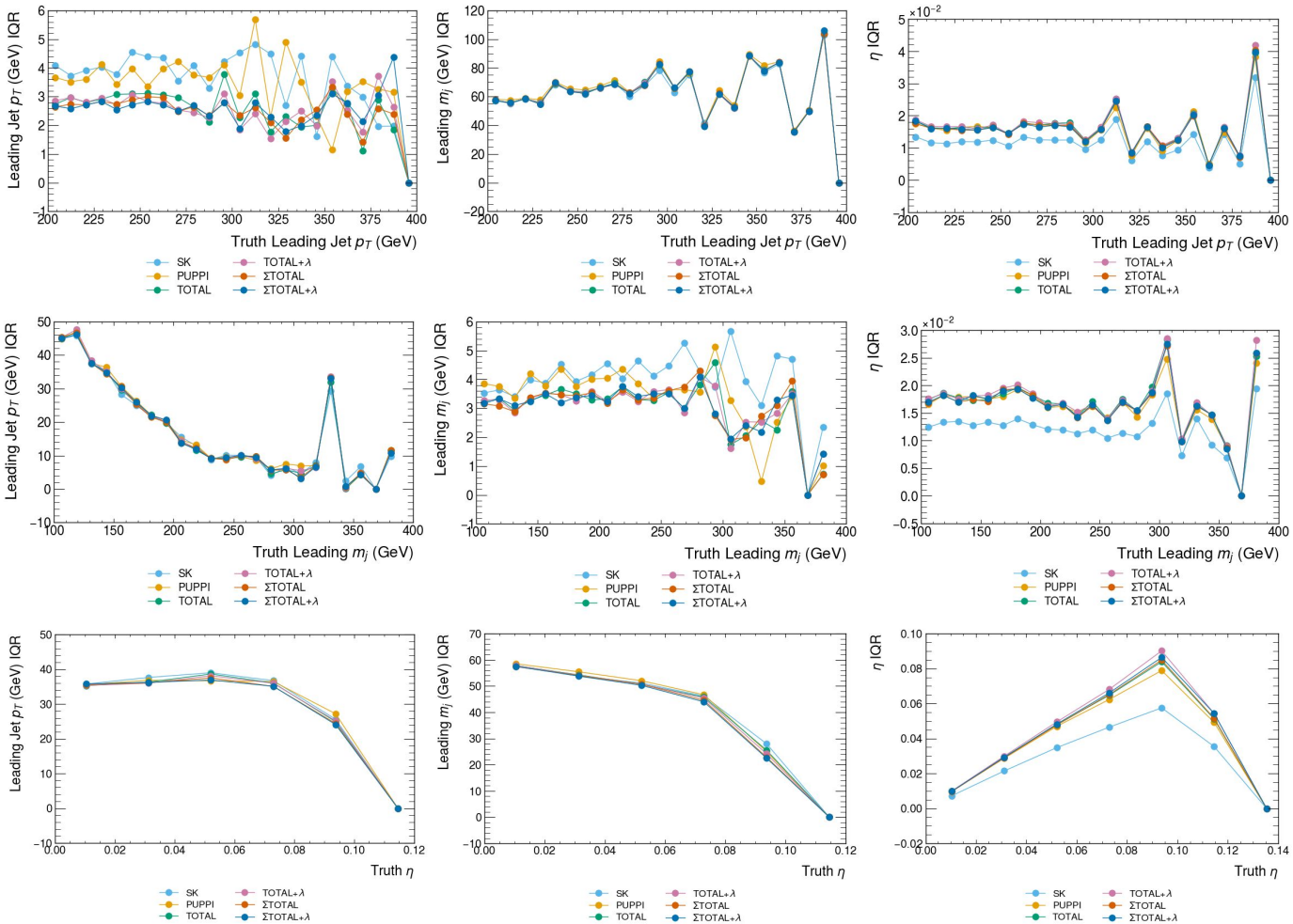
Alternative Loss Functions

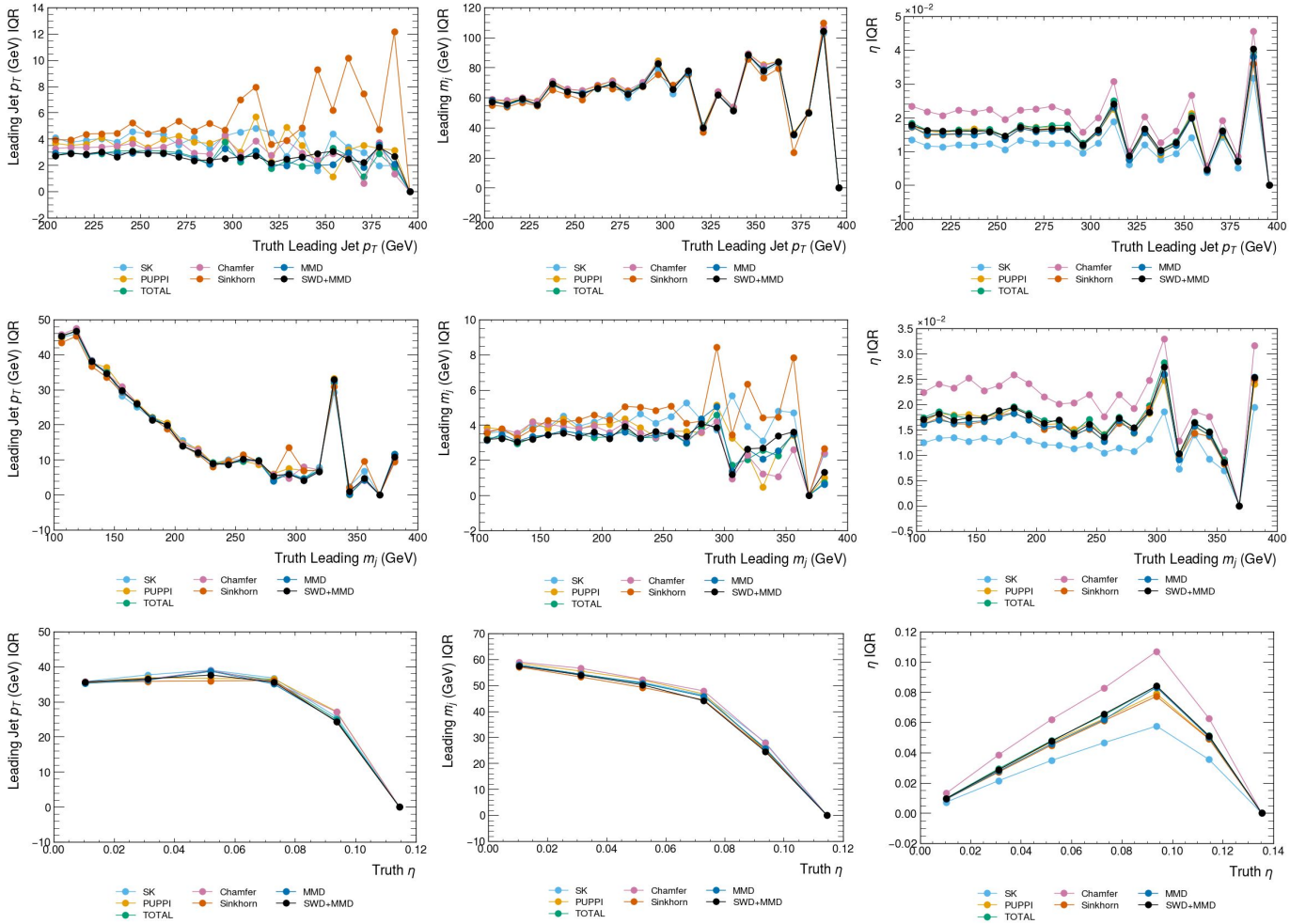
1. Chamfer distance
2. Sinkhorn divergence
3. **Maximum Mean Discrepancy (MMD)**
   a. A kernel-based integral probability metric that maps data points from two distinct sets into a high-dimensional space known as a reproducing kernel Hilbert space (RKHS) using a kernel function
   b. A closed-form estimator that does not rely on random projections
   c. Highly sensitive to tuning of the kernel bandwidth



$$\mathcal{L}_{\mathrm{D}}(\mathbf{X}^s, \mathbf{X}^t) = \mathrm{MMD}(\mathbf{X}^s, \mathbf{X}^t) = \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \phi(\mathbf{x}_i^s) - \frac{1}{n^t} \sum_{j=1}^{n^t} \phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2$$

ϕ: mapping function
$\boldsymbol{X}^s$: feature matrix in source domain
$\boldsymbol{X}^t$: feature matrix in target domain


Leading Jet p_T (GeV) IQR
Truth Leading Jet p_T (GeV)
Leading m_j (GeV) IQR
η IQR
Truth Leading m_j (GeV)
Truth η
SK
PUPPI
TOTAL
TOTAL+λ
ΣTOTAL
ΣTOTAL+λ


Leading Jet p_T (GeV) IQR
Truth Leading Jet p_T (GeV)
Leading m_j (GeV) IQR
η IQR
Truth Leading m_j (GeV)
Truth η
SK
PUPPI
TOTAL
Chamfer
Sinkhorn
MMD
SWD+MMD