# Exploring Optimal Transport for Event-Level Anomaly Detection at the LHC

**Based on recent work guided by:**

Nathaniel Craig[1,2] and Jessica Howard[1]

1 Kavli Institute for Theoretical Physics, UC Santa Barbara
2 Department of Physics, University of California, Santa Barbara

https://arxiv.org/abs/2401.15542

**Hancheng Li**

Undergraduate Student at UC Santa Barbara
hanchengli@ucsb.edu

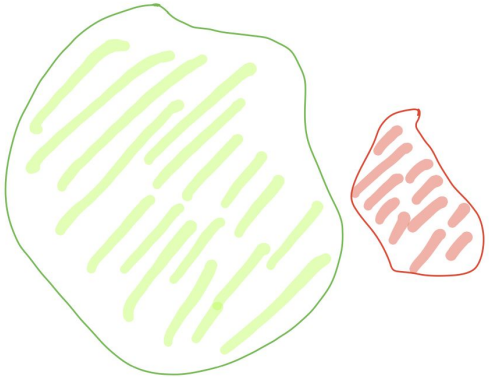
# Goal of Anomaly Detection in Collider Physics

★ Trying to find evidence of BSM events(aka. anomalies) in collisions

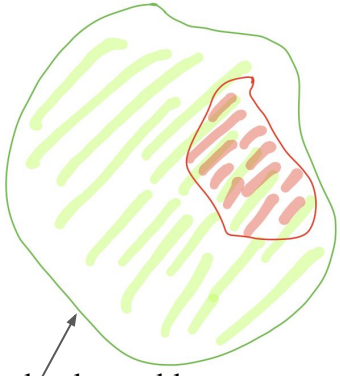
# Goal of Anomaly Detection in Collider Physics

★ Trying to find evidence of BSM events(aka. anomalies) in collisions

**Outliers**: Trying to identify events that emerge in some unexpected region.



**Over-densities**: Trying to identify region of events that are abnormally dense.

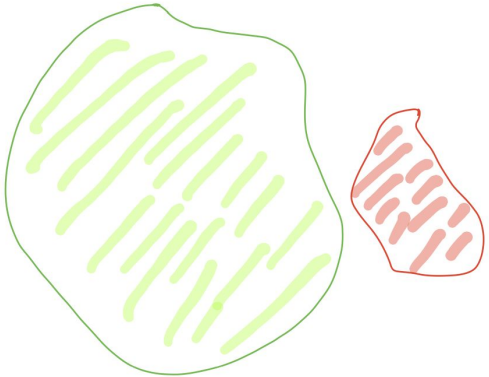


Generally a harder problem

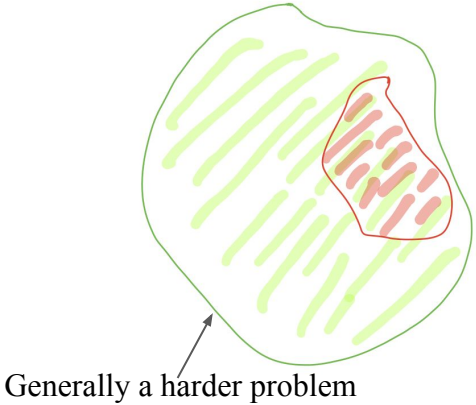
# Goal of Anomaly Detection in Collider Physics

★ Trying to find evidence of BSM events(aka. anomalies) in collisions

**Outliers**: Trying to identify events that emerge in some unexpected region.



**Over-densities**: Trying to identify region of events that are abnormally dense.




- State-of-the-art machine learning algorithm: **Autoencoders**
  - Learn to minimize reconstruction loss on background data
  - Anomalies are expected to have higher reconstruction losses while testing
- Fast to evaluate but not very interpretable
- Can cost a lot of computational resources for large network
- Lots of interesting work has gone into improving them[1][2]

[1] E. Govorkova, et al. “Autoencoders on FPGAs for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider” (2021)
[2] Javier Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics" (2018)

# Possible alternative: Optimal Transport

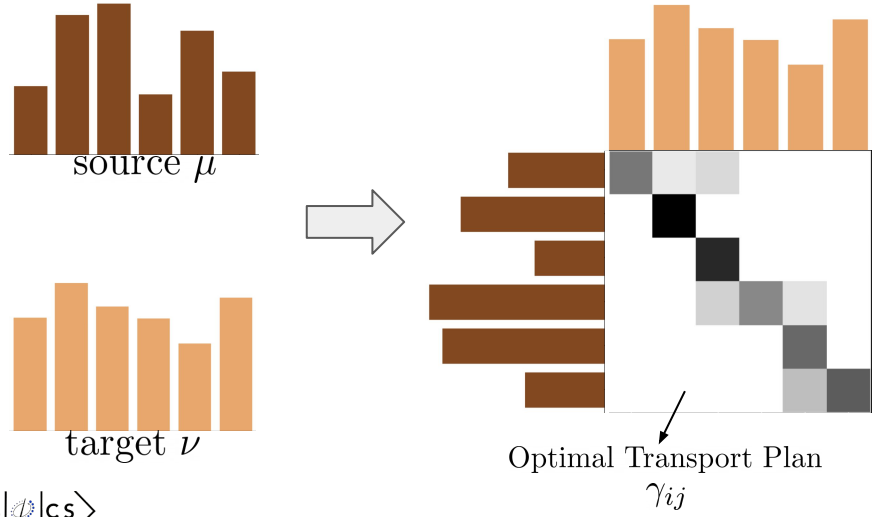
- What is Optimal Transport?

# Possible alternative: Optimal Transport

- What is Optimal Transport?

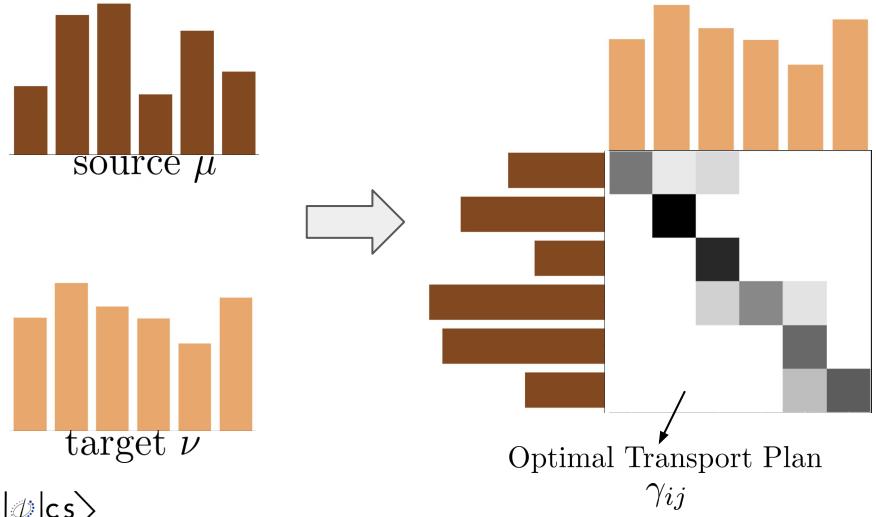
# Possible alternative: Optimal Transport

➢ What is Optimal Transport?

# Possible alternative: Optimal Transport

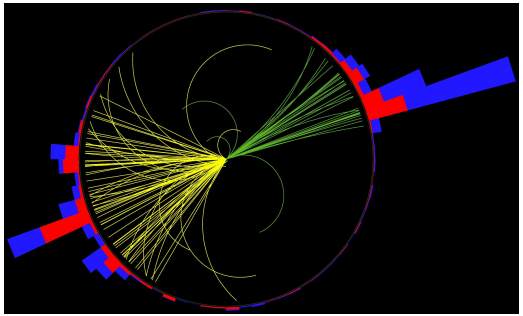
➢ What is Optimal Transport?

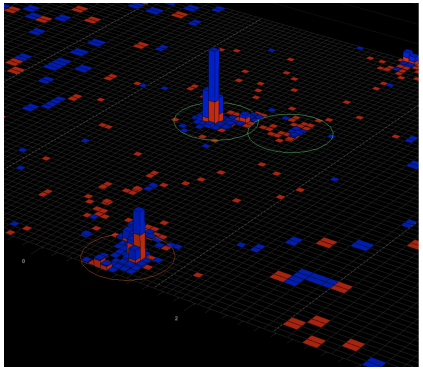




$$W_2(\mu,\nu) = \min_{\gamma_{ij}\in\Gamma(\mu,\nu)} \left( \sum_{ij} \gamma_{ij} \|x_i - \tilde{x}_j\|^2 \right)^{\frac{1}{2}}$$

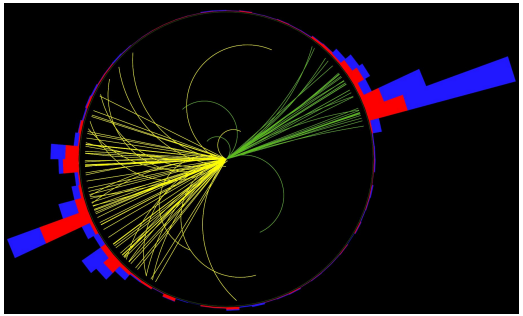
$$\Gamma(\mu,\nu) = \{\gamma_{ij} : \gamma_{ij} \geq 0, \sum_i \gamma_{ij} = u_i, \sum \gamma_{ij} = v_j\}$$

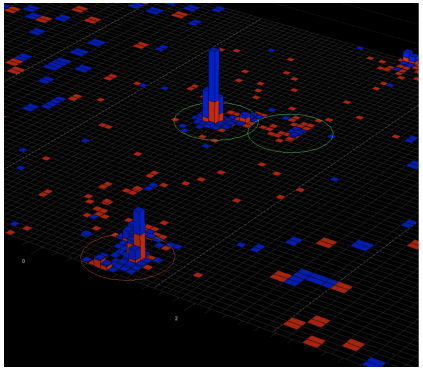
# Optimal Transport in particle physics





# Optimal Transport in particle physics





- OT provides a natural and rigorous definition of distance between events
- First introduced by Ref[1]
- Used in LHC classification tasks[2][3]
  - Typically on jet constituent data
  - Either on its own or combined with machine learning
- Also has been used for anomaly detection task[4][5](also, mostly on jet constituent data)
  - Relatively under-explored

[1] Patrick T. Komiske, Eric M. Metodiev, Jesse Thaler, "The Metric Space of Collider Events", Phys. Rev. Lett. 123, 041801, (2019)
[2] P. T. Komiske, E. M. Metodiev and J. Thaler, “The Hidden Geometry of Particle Collisions.” JHEP 07, 006 (2020)
[3] T. Cai, J. Cheng, K. Craig and N. Craig. “Which metric on the space of collider events?” Phys. Rev. D 105(7), 076003 (2022)
[4] K. Fraser, S. Homiller, R. K. Mishra, B. Ostdiek and M. D. Schwartz. “Challenges for unsupervised anomaly detection in particle physics.” JHEP 03, 066 (2022)
[5] S. E. Park, P. Harris and B. Ostdiek. “Neural embedding: learning the embedding of the manifold of physics data.” JHEP 07, 108 (2023)

# Dataset

Using the ADC 2021 dataset[1]

**SM Background: 4,000,000 events**

| | |
|---|---|
| (59.2%) | $pp \to W^{\pm} + \text{jets} \to l^{\pm}\nu_l + \text{jets}$ |
| (6.7%) | $pp \to Z + \text{jets} \to l^{+}l^{-} + \text{jets}$ |
| (0.3%) | $pp \to t\bar{t} + \text{jets}$ |
| (33.8%) | $pp \to \text{jets}$ |

**4 BSM Signal cases**

**Neutral scalar boson $A$: 55,969 events**

$m_A = 50$ GeV $\quad pp \to A + X \to Z^*Z^* + X,\ Z^* \to l^{+}l^{-}$

**Scalar boson $h^0$: 691,283 events**

$m_{h^0} = 60$ GeV $\quad pp \to h^0 + X \to \tau^{+}\tau^{-} + X$

**Charged scalar $h^{\pm}$: 760,272 events**

$m_{h^{\pm}} = 60$ GeV $\quad pp \to h^{\pm} + X \to \tau\nu + X$

**Leptoquark ($LQ$): 340,544 events**

$m_{LQ} = 80$ GeV $\quad pp \to LQ \to \tau b$

Jessica Howard

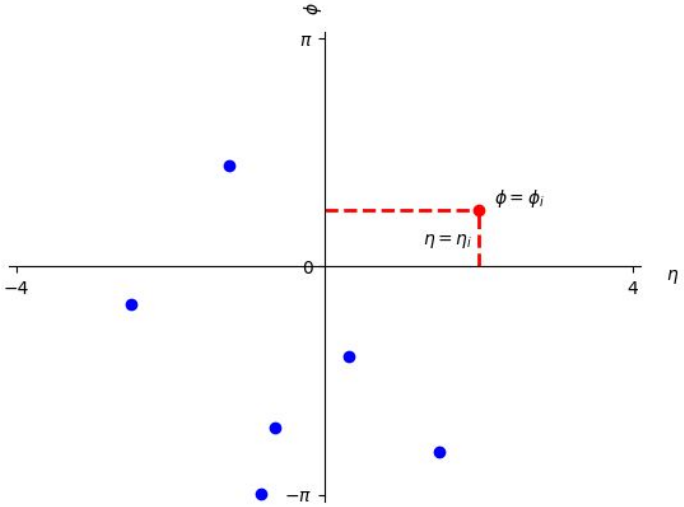
**Event = (19,3); zero-padded**

| | $p_T$ | $\eta$ | $\phi$ |
|---|---|---|---|
| MET | | | |
| 4 $e/\gamma$ | | | |
| 4 $\mu$ | | | |
| 10 jet | | | |

- Coarse-grained L1 trigger-level data
- Goal:
  - Train on SM background data to learn the distribution
  - Test on BSM signal data and (hopefully) see good performances on all signal cases

[1] E. Govorkova et al., LHC physics dataset for unsupervised New Physics detection at 40 MHz. Sci. Data 9, 118 (2022), doi: 10.1038/s41597-022-01187-8, arXiv:2107.02157.

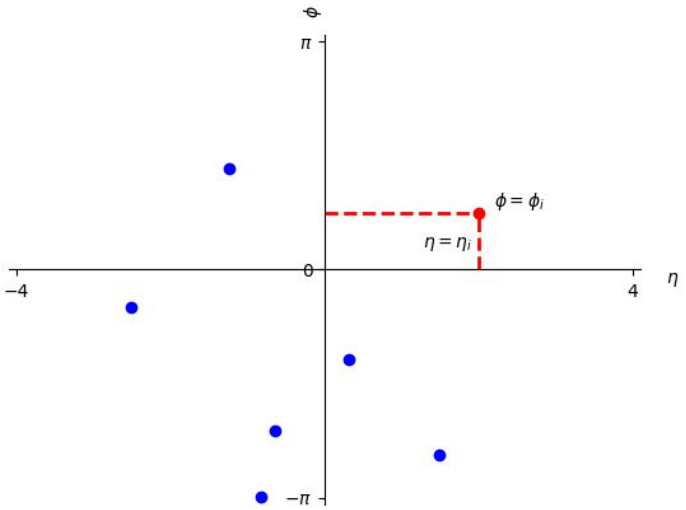
# Choice of Ground Space




2D Ground Space: $(\eta, \phi)$

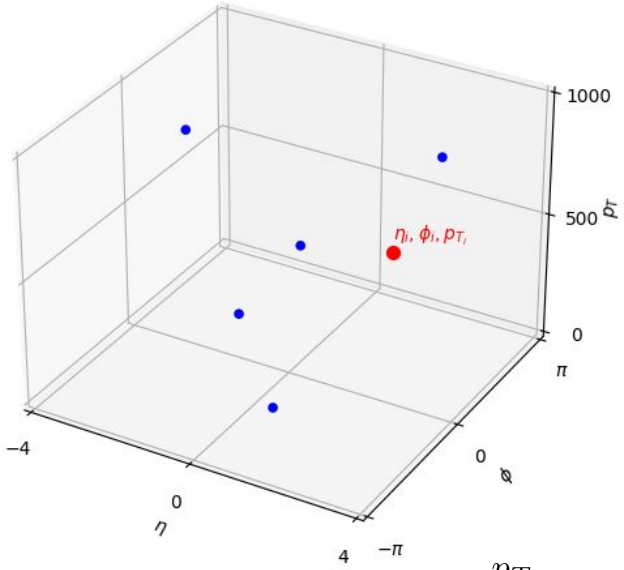
$$m_i = \frac{p_{T_i}}{\sum_0^{18} p_{T_i}}$$

# Choice of Ground Space




2D Ground Space: $(\eta, \phi)$

$$m_i = \frac{p_{T_i}}{\sum_0^{18} p_{T_i}}$$




3D Ground Space: $(\eta, \phi, \frac{p_T}{GeV})$

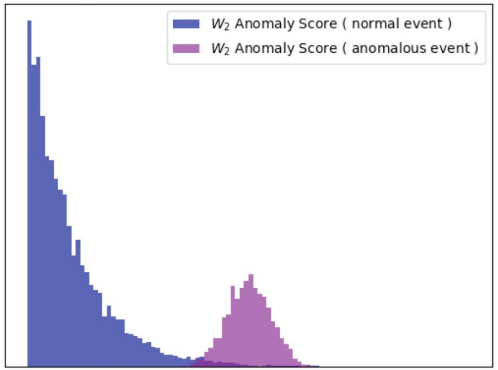
$$m_i = \frac{1}{19}$$

# OT as anomaly score

Procedure:

1. Randomly choose a set of background events as reference events $\longrightarrow \boldsymbol{BKG_{ref}}$

2. For each test event $\boldsymbol{T}$, compute the OT distance to $\boldsymbol{BKG_{ref}}$ and get the minimum $\longrightarrow$ Anomaly Score of $\boldsymbol{T}$



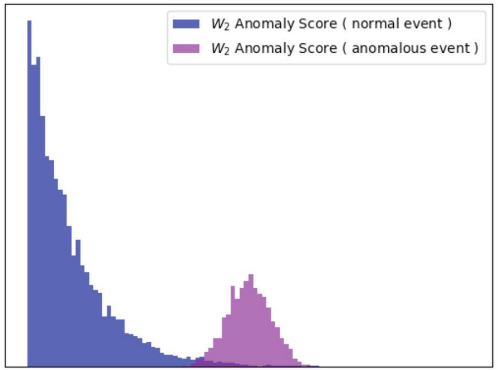

Jessica Howard

# OT as anomaly score

Procedure:

1. Randomly choose a set of background events as reference events $\longrightarrow \boldsymbol{BKG_{ref}}$

2. For each test event $\boldsymbol{T}$, compute the OT distance to $\boldsymbol{BKG_{ref}}$ and get the minimum $\longrightarrow$ Anomaly Score of $\boldsymbol{T}$




Jessica Howard

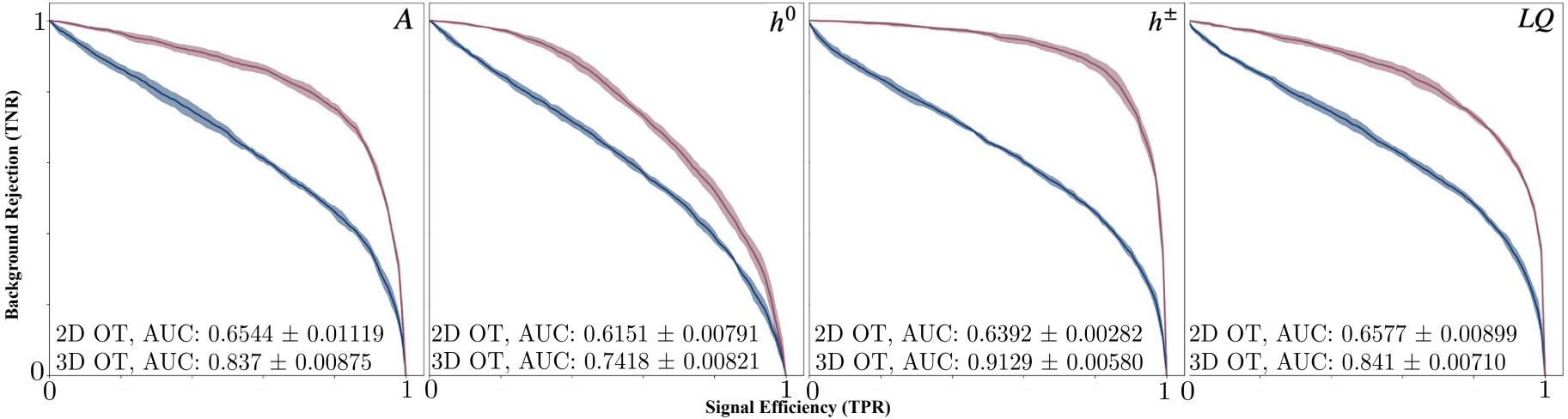
Advantage:

- Fully interpretable
- No training needed, purely based on the distinguishability of OT

Potential disadvantage:

- Some SM events may be further apart in OT distance than SM and BSM events
- Didn't take into account of symmetries of the problem

# Results

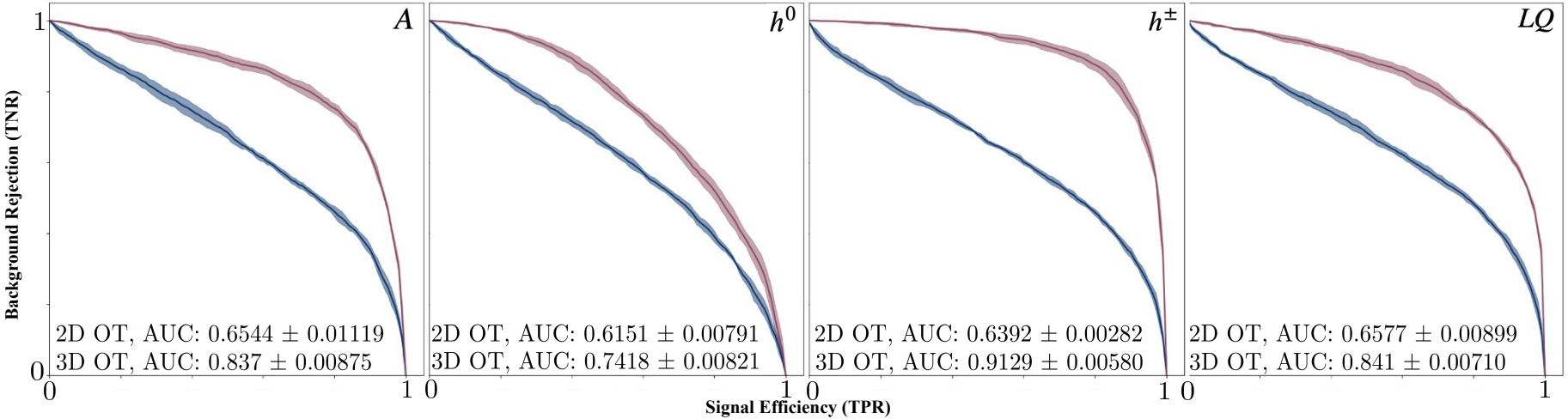



Raw Autoencoder Result from [1]

| | | |
|---|---|---|
| AUC | $A$ | $0.885 \pm 0.002$ |
| | $h^0$ | $0.755 \pm 0.002$ |
| | $h^\pm$ | $0.900 \pm 0.004$ |
| | $LQ$ | $0.856 \pm 0.002$ |

[1] B.M. Dillon, L. Favaro, F. Feiden, T. Modak, T. Plehn. "Anomalies, Representations, and Self-Supervision." arXiv: 2301.04660

# Results




Raw Autoencoder Result from [1]

| | | |
|---|---|---|
| AUC | $A$ | $0.885 \pm 0.002$ |
| | $h^0$ | $0.755 \pm 0.002$ |
| | $h^\pm$ | $0.900 \pm 0.004$ |
| | $LQ$ | $0.856 \pm 0.002$ |

- Choice of ground space is essential
- In the 3D case, OT is approaching the performance of raw autoencoders

[1] B.M. Dillon, L. Favaro, F. Feiden, T. Modak, T. Plehn. "Anomalies, Representations, and Self-Supervision." arXiv: 2301.04660

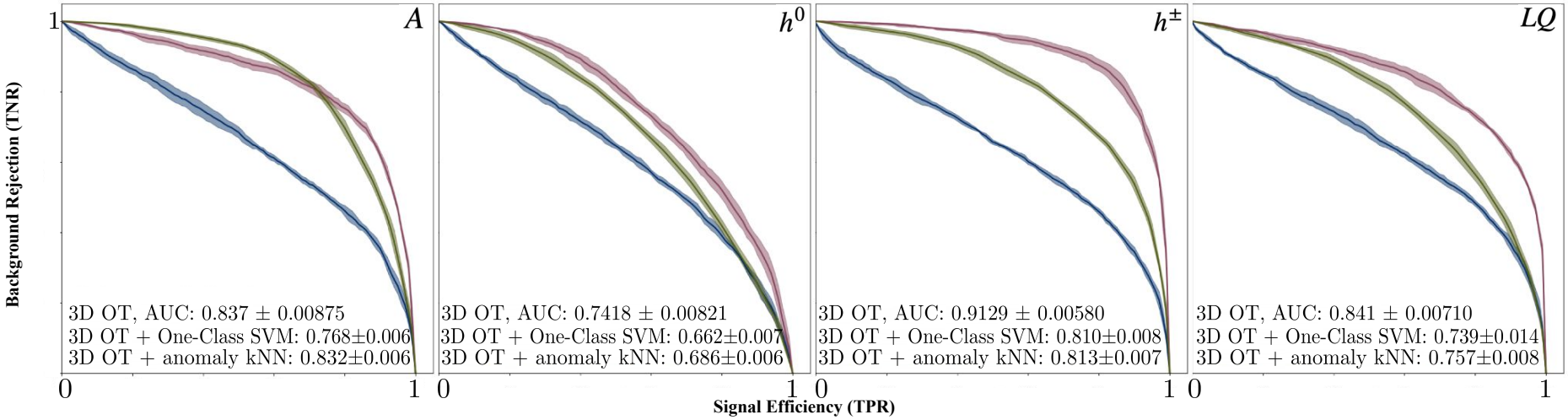
# OT + Machine Learning

- anomaly-kNN is kNN trained on background data + anomaly augmented background data[1] (to imitate the distribution of signal dataset)
- One-class SVM is an unsupervised learning algorithm that trains on background data only to learn its distribution

[1] B.M. Dillon, L. Favaro, F. Feiden, T. Modak, T. Plehn. "Anomalies, Representations, and Self-Supervision." arXiv: 2301.04660

# OT + Machine Learning

- anomaly-kNN is kNN trained on background data + anomaly augmented background data[1] (to imitate the distribution of signal dataset)
- One-class SVM is an unsupervised learning algorithm that trains on background data only to learn its distribution




- 3D OT generally does better than OT + simple ML algorithms!

[1] B.M. Dillon, L. Favaro, F. Feiden, T. Modak, T. Plehn. "Anomalies, Representations, and Self-Supervision." arXiv: 2301.04660

# Conclusion and future work

- OT is a promising method for anomaly detection, both as a plain metric and combined with machine learning algorithms

# Conclusion and future work

- OT is a promising method for anomaly detection, both as a plain metric and combined with machine learning algorithms
  - Choice of ground space is essential

# Conclusion and future work

- OT is a promising method for anomaly detection, both as a plain metric and combined with machine learning algorithms
  - Choice of ground space is essential
  - Already give decent results and still has plenty of room for improvement:

# Conclusion and future work

- OT is a promising method for anomaly detection, both as a plain metric and combined with machine learning algorithms
  - Choice of ground space is essential
  - Already give decent results and still has plenty of room for improvement:
    - Symmetries in particles collisions ⟶ **Gromov-Wasserstein distance**[1]

[1] Planned follow-up work with N. Craig and J. Howard

# Conclusion and future work

- OT is a promising method for anomaly detection, both as a plain metric and combined with machine learning algorithms
  - Choice of ground space is essential
  - Already give decent results and still has plenty of room for improvement:
    - Symmetries in particles collisions ⟶ **Gromov-Wasserstein distance**[1]
    - There is more information collected by the detector that we can added to our ground space(e.g charge and particle species) ⟶ **Multi-species OT**[2]

[1] Planned follow-up work with N. Craig and J. Howard
[2] Ongoing work with T. Cai, K. Craig, N. Craig, and J. Howard

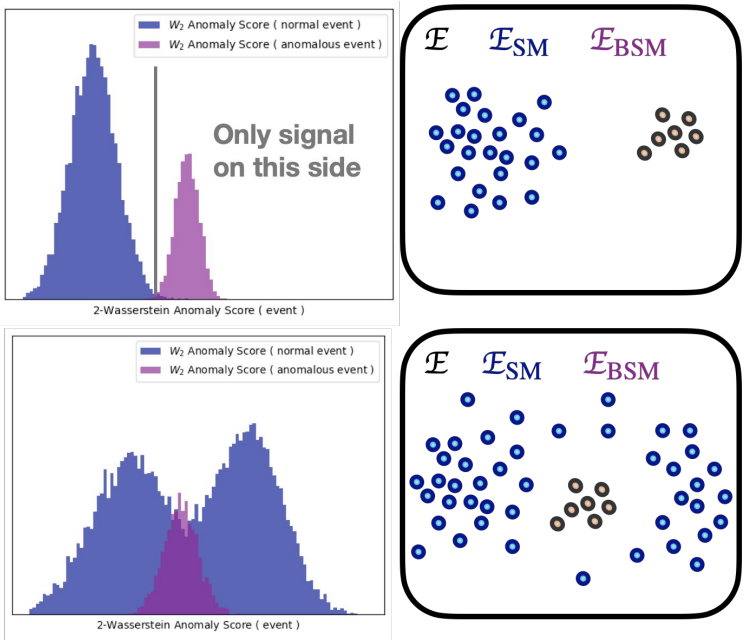
# Conclusion and future work

- OT is a promising method for anomaly detection, both as a plain metric and combined with machine learning algorithms
  - Choice of ground space is essential
  - Already give decent results and still has plenty of room for improvement:
    - Symmetries in particles collisions ⟶ **Gromov-Wasserstein distance**[1]
    - There is more information collected by the detector that we can added to our ground space(e.g charge and particle species) ⟶ **Multi-species OT**[2]
    - Computational Feasibility ⟶ **Linearized Optimal Transport**[3]

[1] Planned follow-up work with N. Craig and J. Howard
[2] Ongoing work with T. Cai, K. Craig, N. Craig, and J. Howard
[3] T. Cai, J. Cheng, K. Craig, N. Craig. "Linearized Optimal Transport for Collider Events". Phys. Rev. D 102, 116019 (2020). arXiv: 2008.08604

# Backup Slides

# OT + Machine Learning (Classification)




Jessica Howard

AUC

| | 2D | 2D planed total $p_T$ [GeV] | | 3D |
|---|---|---|---|---|
| | | $p_T \in (50, 100)$ | $p_T \in (500, 1000)$ | |
| $A$ | $0.6948 \pm 0.01795$ | $0.7974 \pm 0.008994$ | $0.7989 \pm 0.02231$ | **$0.9021 \pm 0.01426$** |
| $h^0$ | $0.6698 \pm 0.01178$ | $0.6761 \pm 0.03569$ | $0.5829 \pm 0.02124$ | **$0.7713 \pm 0.01814$** |
| $h^{\pm}$ | $0.8103 \pm 0.01673$ | $0.6071 \pm 0.02694$ | $0.6203 \pm 0.03418$ | **$0.9198 \pm 0.006547$** |
| $LQ$ | $0.7906 \pm 0.02531$ | $0.7469 \pm 0.02149$ | $0.5285 \pm 0.01536$ | **$0.8766 \pm 0.01415$** |

| 3D OT |
|---|
| $0.8370 \pm 0.008752$ |
| $0.7418 \pm 0.008213$ |
| $0.9129 \pm 0.005798$ |
| $0.8410 \pm 0.007098$ |

- Possible reasons why OT + Simple ML algorithms for anomaly detection doesn't do very well:
  - Anomaly-augmented dataset cannot imitate the distribution of signal dataset perfectly
  - One-class SVM is known to have issues with complex datasets where anomalies can form dense clusters

# Anomaly augmented background data

This is the background anomaly augmentation scheme used in paper, defined by [1]

1. **Multiplicity increase:** This transformation randomly adds some number of $e/\gamma$, $\mu$, and jets to each event in the range $(n_{e/\gamma}, 4-n_{e/\gamma})$, $(n_{\mu}, 4-n_{\mu})$, $(n_{\text{jets}}, 10-n_{\text{jets}})$, respectively. The $p_T$ of each new particle is a randomly chosen fraction of the highest $p_T$ of any object in that event added to the base required by that object's selection criteria. The $\eta, \phi$ of each object are uniformly chosen within the allowed limits corresponding to that object's type. After all objects are added, the MET of the event is recalculated.

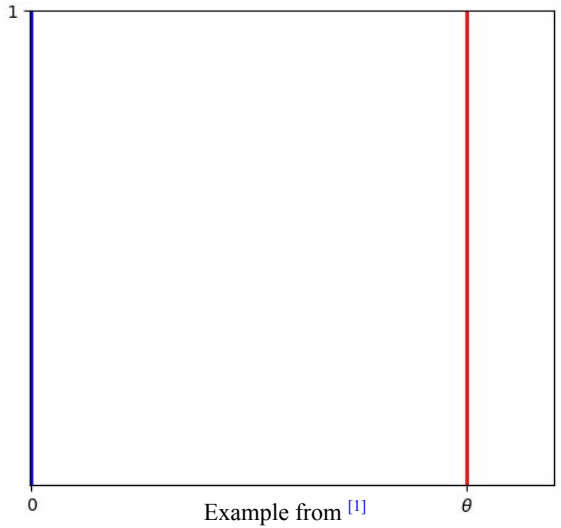
2. **Multiplicity increase, with constant** MET **and** $p_{\text{T}}$**:** This transformation keeps the total $p_{\text{T}}$ and MET constant while still increasing the overall object multiplicity. This is achieved by splitting an existing object to create two new objects. The combined $p_{\text{T}}$ of the two new objects is equal to the original. The $\eta, \phi$ of each new object are then randomly smeared with Gaussian noise.

3. MET **and** $p_{\text{T}}$ **shift:** This transformation randomly shifts the a) MET, b) reconstructed object $p_{\text{T}}$, or c) both by a constant multiplicative factor. Transformations (a),(b),(c) are chosen with equal probability. To satisfy selection criteria, the multiplicative factor for shifting $p_{\text{T}}$ is chosen uniformly in the range $[1, 5)$ (i.e. the $p_{\text{T}}$ of objects will never be down-shifted such that it might violate an object's minimum $p_{\text{T}}$ criteria). No such restriction exists for MET, so its multiplicative factor is chosen uniformly in the range $[0.5, 5)$.

In general, we apply augmentations (1), (2), and (3) with equal probability. However, there are certain events for which augmentation (2) could not transform the event without causing the event selection criteria to be violated (approximately 8%). For these events, we instead apply augmentations (1) and (3) with equal probability. Taking this into account, approximately 37.3%, 29.3%, 37.3% of events are transformed with augmentations (1), (2), and (3), respectively.

[1] B.M. Dillon, L. Favaro, F. Feiden, T. Modak, T. Plehn. "Anomalies, Representations, and Self-Supervision." arXiv: 2301.04660

# Advantage of OT




Example from [1]

$$KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

$$JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

$$W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$$

[1] Arjovsky et al., ICML 2017. “Wasserstein Generative Adversarial Network”