# Linearized Optimal Transport for Collider Events


Student Name: Junyi Cheng

Authors: Tianji Cai[1], Junyi Cheng[1], Katy Craig[2], Nathaniel Craig[1]

[1]Department of Physics, University of California, Santa Barbara

[2]Department of Mathematics, University of California, Santa Barbara


## Introduction

- How to analysis collider data?
- Combine optimal transport with particle physics: define a metric space & calculate the distances between collider events.
- Results of same interaction processes have small distances in metric space; results of different processes are far away in metric space.
- Use linear embedding to speed up calculation.

## Linearized Optimal Transport (LOT)

Optimal transport (OT): Most efficient way to move between 2 probability distributions, solutions are p-Wasserstein distances

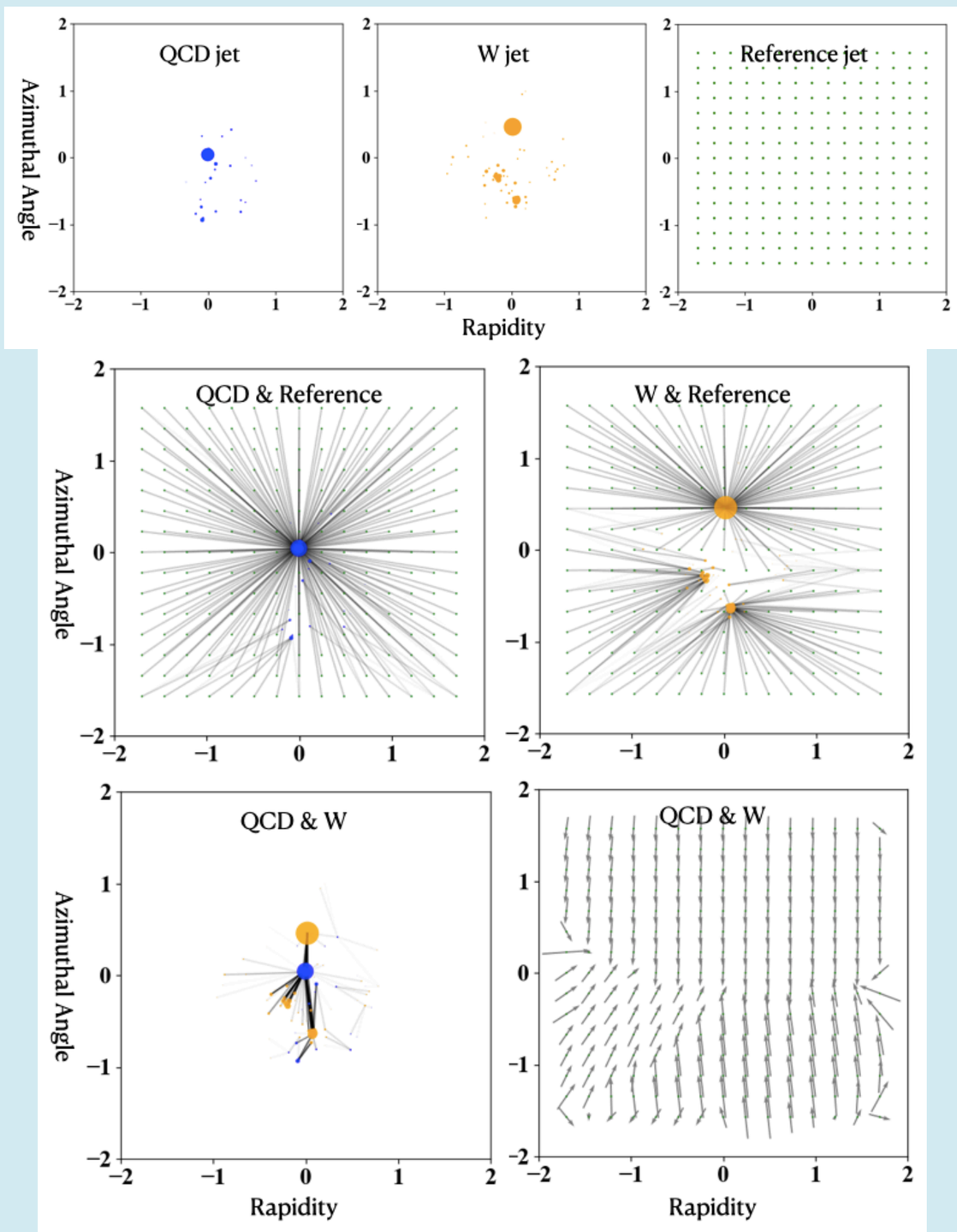
$$W_p(\mathcal{E},\tilde{\mathcal{E}}) = \min_{g_{ij}\in\Gamma(\mathcal{E},\tilde{\mathcal{E}})} \left(\sum_{ij} g_{ij}\|x_i-\tilde{x}_j\|^p\right)^{1/p}$$

$$\Gamma(\mathcal{E},\tilde{\mathcal{E}}) = \left\{g_{ij} : g_{ij}\geq 0,\ \sum_j g_{ij} = E_i,\ \sum_i g_{ij} = \tilde{E}_j\right\}$$

LOT: Linear approximation to 2-Wasserstein distance, much faster than OT, saves storage space, while almost as good as OT.

Calculate bary centers -> Euclidean distances between bary centers

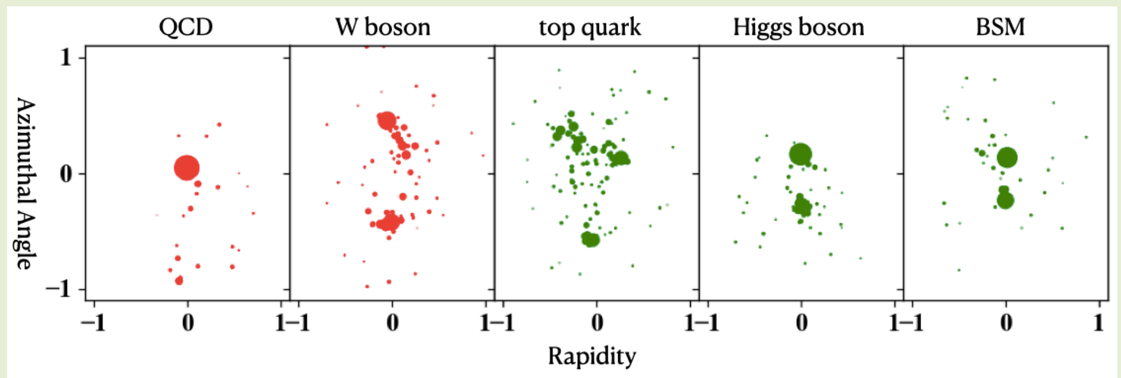
$$z_i := \frac{1}{R_i}\sum_j r_{ij}x_j \qquad LOT_{r,\tilde{r}}(\mathcal{E},\tilde{\mathcal{E}}) = \left(\sum_i R_i\|z_i-\tilde{z}_i\|^2\right)^{1/2}$$




## Object Classification with LOT

Collider simulation: collision -> hadronization -> jet classification

Jet samples:




Jet tagging tasks: QCD vs. W, t vs. QCD, t vs. W, Higgs vs. QCD, Higgs vs. W, BSM vs. QCD, BSM vs. W

## Machine Learning Analysis

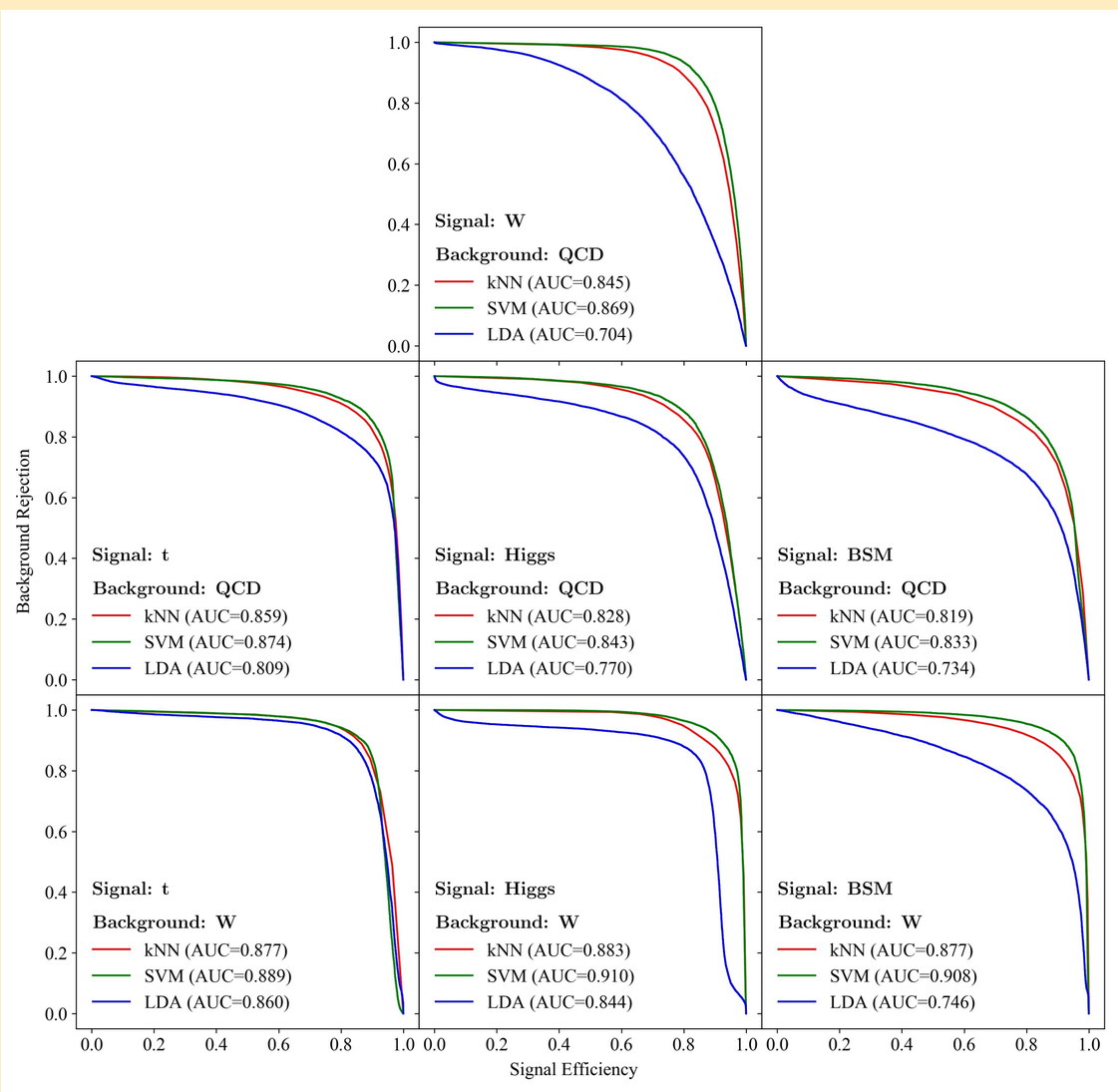
Models: [1]supervised & [2]unsupervised

①Linear Discriminant Analysis (LDA)[1] + visualization; ②k Nearest Neighbors (kNN)[1]; ③Support Vector Machine (SVM)[1]; ④k-medoids[2]

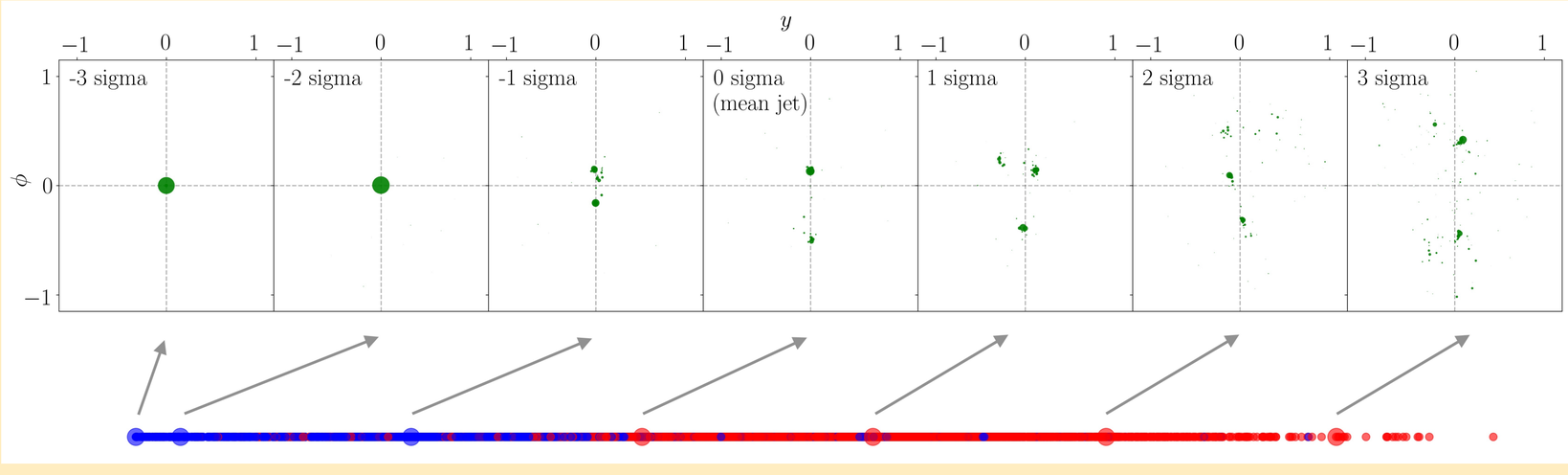
**General conclusions:**

- LOT framework classifies better than traditional observables hand picked by physicists, and is comparable to exact OT metrics (but much faster) and DNNs (but gives more human interpretability).
- Only nearby jets are significant for classification.
- LDA is good for quick first check; SVM performs the best; k-medoids encourages future classifications beyond supervised ML.
- In LOT framework, it's easier to reject W than QCD as background for top, Higgs and BSM signals.

**ROC curves & AUC values:**




**LDA visualization for W-top tagging:**




## Future Directions

Unbalanced LOT and event-level LOT


## Acknowledgements

My work was funded by Create Fund Summer Undergraduate Fellowship, thanks to the generosity of CCS donors.