

## Which metric on the space of collider events?

Tianji Cai  and Junyi Cheng 

*Department of Physics, University of California, Santa Barbara, California 93106, USA*

Katy Craig 

*Department of Mathematics, University of California, Santa Barbara, California 93106, USA*

Nathaniel Craig

*Department of Physics, University of California, Santa Barbara, California 93106, USA,  
Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA,  
and Berkeley Center for Theoretical Physics, University of California, Berkeley, California 94720, USA*



(Received 14 December 2021; accepted 10 March 2022; published 7 April 2022)

Which is the best metric for the space of collider events? Motivated by the success of the energy mover's distance in characterizing collider events, we explore the larger space of unbalanced optimal transport distances, of which the energy mover's distance is a particular case. Geometric and computational considerations favor an unbalanced optimal transport distance known as the Hellinger-Kantorovich distance, which possesses a Riemannian structure that lends itself to efficient linearization. We develop the particle linearized unbalanced optimal transport framework for collider events based on the linearized Hellinger-Kantorovich distance and demonstrate its efficacy in boosted jet tagging. This provides a flexible and computationally efficient optimal transport framework ideally suited for collider physics applications.

DOI: [10.1103/PhysRevD.105.076003](https://doi.org/10.1103/PhysRevD.105.076003)

### I. INTRODUCTION

The distance between collider events is a promising tool for analyzing particle physics data. But what is the best way to define such a distance? Recently, Komiske *et al.* [1] pioneered a new approach to the problem using the tools of optimal transport: The *energy mover's distance* (EMD) treats collider events as angular distributions of energy (or closely related analogs such as transverse momentum distributions in the rapidity-azimuth plane at hadron colliders) and computes a modified form of the canonical Earth mover's distance between these distributions. Intuitively, this quantifies the “work” required to rearrange one collider event to look like the other or, in other words, how to “optimally transport” one collider event into another. *Optimal transport* distances are naturally suited to the space of collider events. They allow for the comparison of raw event data (that is, events composed of a finite number of energy measurements at prescribed locations) without smoothing or binning, and they respect the underlying notion of distance in the reference frame of

the detector. Furthermore, they are robust to high-frequency noise, ensuring infrared and collinear safety [1,2].

Application of the EMD to collider events has led to new bounds on the modification of infrared- and collider-safe observables by hadronization [1]; jet classification using interpretable, distance-based machine learning algorithms [1,3]; visualization of the metric space of jets using CMS open data [4]; and definition of new shape variables [5,6]. In Ref. [2], the EMD formed the foundation of a unified approach to collider observables based on the distance between an event and a manifold on the space of events, recasting decades of collider physics in the language of optimal transport. Optimal transport also underlies a number of recently developed machine learning frameworks for model discrimination, anomaly detection, and event generation at the LHC [7–14].

The advantages of applying optimal transport to collider events are increasingly clear. But which optimal transport distance is best? The Earth mover's distance, on which the EMD was based, is only one example of a family of balanced optimal transport distances defined between equally normalized distributions, known as *p*-Wasserstein distances. Although the various *p*-Wasserstein distances are qualitatively similar, they differ in key respects. For instance, only the 2-Wasserstein distance has a Riemannian structure that admits a computationally efficient linearized approximation [15,16]. Additional choices arise when comparing events with unequal total energy. The EMD is obtained by extending

---

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP<sup>3</sup>.*

the Earth mover’s distance with an additional term to account for differences in total energy or  $p_T$  among collider events, but this extension is far from unique. Rather, there are many possible approaches to the *unbalanced* optimal transport problem, one of which enjoys a Riemannian structure akin to the 2-Wasserstein distance. Of course, the “best” optimal transport distance for collider events depends on the relevant criteria. Practical considerations include simplicity, robustness, and computational speed, while theoretical considerations (such as the conceptual advantages of a geometric language for collider physics) favor geometric interpretability.

In this paper, we develop and demonstrate an optimal transport framework for collider events that builds on the EMD with an eye toward these goals. We first situate the EMD in the space of unbalanced optimal transport distances, showing that the EMD is an example of the *partial transport distance* [17–21]. The dynamic formulation of the partial optimal transport distance, in turn, suggests a general framework for unbalanced optimal transport that may be applied to collider events. In particular, optimizing for both computational speed and geometric interpretability favors an unbalanced optimal transport distance known as the *Hellinger-Kantorovich distance* [22–26]. Loosely speaking, this distance generalizes the 2-Wasserstein distance to unbalanced distributions, preserving a Riemannian structure that lends itself to both geometric interpretation and computational efficiency via linearization [26,27].

The latter advantage is particularly salient. Constructing the geometry of a collider event sample is computationally expensive, requiring the determination of  $\mathcal{O}(N_{\text{evt}}^2)$  optimal transport distances. However, the computational cost may be significantly decreased by leveraging optimal transport distances admitting a Riemannian structure, such as the 2-Wasserstein distance (for balanced distributions) and the Hellinger-Kantorovich distance (for unbalanced ones). This allows the calculation of optimal transport distances to be linearized by projecting onto the tangent plane at a chosen reference event and computing simpler  $\ell^2$  distances on this plane.

For the balanced case, the linearized optimal transport (LOT) approximation [28] to the 2-Wasserstein distance is competitive with EMD in boosted jet classification at a fraction of the computational cost [3]. Furthermore, as the numerical discretization of the reference event is refined, the linearized metric converges to a true metric on the space of events (see Corollary 1 in Ref. [3]). In the case where the reference measure is chosen to be an approximation of the uniform measure, the LOT embedding is closely related to comparing power diagrams [2,29,30]. Recent work by Delalande and Mériqot [31] has quantified the relationship between the original 2-Wasserstein distance and its linearization, implying, in particular, that LOT preserves benefits of infrared and collinear safety [1,2].

For the unbalanced case, the corresponding linearization of the Hellinger-Kantorovich distance was recently developed in Ref. [27]. In analogy with LOT, we will refer to the linearization of the Hellinger-Kantorovich metric for discrete measures as particle linearized unbalanced optimal transport (PLUOT). As with LOT, when the reference measure is an approximation of the uniform measure, the PLUOT embedding is closely related to comparing generalized Laguerre diagrams [2,29]. Here, we apply PLUOT to the classification of boosted  $W$  and QCD jets, finding that simple machine learning algorithms based on PLUOT perform comparably to the same algorithms based on the EMD in a fraction of the time. This provides a computationally efficient linearized optimal transport framework suitable for comparing collider events with different total energy or  $p_T$ .

On one hand, we are optimistic that the computational advantages of our approach will enable entirely new applications of optimal transport methods in particle physics. On the other hand, as we will see, the advantages of applying the Hellinger-Kantorovich distance to collider events extend beyond the computational speedup arising from linearization. The Hellinger-Kantorovich distance and its linearization depend on a length scale  $\kappa$  which determines the relative cost of transporting vs creating or destroying energy. When applied to the classification of boosted  $W$  and QCD jets, simple machine learning algorithms using PLUOT with an optimized choice of  $\kappa$  outperform the same algorithms using the EMD when the jets are drawn from a sufficiently large  $p_T$  range. We also find that boosted jet classification based on the Hellinger-Kantorovich distance is more robust against pileup contamination than traditional approaches such as  $N$ -subjettiness.

This paper is organized as follows: In Sec. II, we review the classical Earth mover’s distance and other  $p$ -Wasserstein distances of balanced optimal transport. We then turn to the energy mover’s distance, showing that it is a special case of the partial transport distance. We use a dynamic formulation of the partial transport distance to introduce a more general family of unbalanced optimal transport distances. Within this family, we focus on the Hellinger-Kantorovich distance, reviewing the linearization introduced in Ref. [27] and introducing the PLUOT framework for collider events. In Sec. III, we consider the classification of boosted  $W$  and QCD jets, using the PLUOT distance as input to a number of simple machine learning algorithms. We study the performance of these algorithms as a function of the choice of reference measure, Hellinger-Kantorovich scale parameter, and  $p_T$  range, finding performance comparable or superior to the same algorithms using EMD distances with much lower computation and storage costs. Finally, in Sec. IV, we study the effects of pileup on PLUOT-based classification, finding a surprising level of robustness compared to, e.g., classification based on  $N$ -subjettiness ratios. We conclude in Sec. V, reserving tables of numerical results for the Appendix.

## II. UNBALANCED OPTIMAL TRANSPORT

Consider two discrete measures  $\mathcal{E}$  and  $\mathcal{E}'$ , which assign positive masses  $\{E_i\}_{i \in I}$  and  $\{E'_j\}_{j \in J}$  to particles at locations  $\{x_i\}_{i \in I}$  and  $\{x'_j\}_{j \in J}$ , respectively, in a domain  $\Omega \subseteq \mathbb{R}^d$ . For example, in the collider physics context, such measures are often used to represent jet events, where the mass assigned at a point represents the energy measured by the calorimeters.<sup>1</sup> When the measures have the same total mass,  $\sum_i E_i = \sum_j E'_j$ , the theory of (balanced) optimal transport provides a natural notion of distance between the two measures. For example, if  $d_{ij} = \|x_i - x'_j\|$  represents the distance between particles in each discrete measure, the classical Earth mover's distance is given by

$$W_1(\mathcal{E}, \mathcal{E}') := \min_{\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}}} \sum_{ij} d_{ij} \gamma_{ij},$$

$$\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}} := \left\{ \gamma_{ij} : \gamma_{ij} \geq 0, \sum_j \gamma_{ij} = E_i, \sum_i \gamma_{ij} = E'_j \right\}.$$

This can be interpreted as finding a way to rearrange the distribution of mass in  $\mathcal{E}$  to match  $\mathcal{E}'$ , using the least amount of effort:  $\gamma_{ij}$  represents the amount of mass moved from particle  $i$  to particle  $j$ , and  $\sum_{ij} \gamma_{ij} d_{ij}$  represents the required effort. By making a mild modification to the notion of effort, one may also consider the  $p$ -Wasserstein distance between two discrete measures of equal total energy, for  $p \geq 1$ :

$$W_p(\mathcal{E}, \mathcal{E}') := \min_{\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}} \left( \sum_{ij} \gamma_{ij} d_{ij}^p \right)^{1/p}. \quad (1)$$

In recent years, there has been substantial interest in generalizing the above metric to measures with unequal total energy, known as the *unbalanced* optimal transport problem [18–26, 32, 33]. We begin by recalling the notion of *partial optimal transport*, as already used in collider physics applications by Komiske *et al.* [1]. Next, we describe how partial optimal transport distances relate to the Hellinger-Kantorovich distance [22–26], which will be the key tool in the present work. As we will describe, the Hellinger-Kantorovich distance is unique among related distances due to the fact that it has a Riemannian structure [26, 27]. This allows us to linearize the metric and vastly improve its computational performance on classification tasks.

### A. Partial optimal transport

The first unbalanced optimal transport metric considered in collider physics was the energy mover's distance studied

<sup>1</sup>In what follows, we will largely use the lexicon of optimal transport, emphasizing that energy or  $p_T$  (rather than jet or particle masses) will play the role of “mass” in collider physics applications.

by Komiske *et al.* [1]. In this case, for fixed  $R \geq \max_{ij} d_{ij}/2$ , the distance between discrete measures  $\mathcal{E}, \mathcal{E}'$  is

$$\text{EMD}^R(\mathcal{E}, \mathcal{E}') = \min_{\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}}} \frac{1}{R} \sum_{ij} d_{ij} \gamma_{ij} + \left| \sum_i E_i - \sum_j E'_j \right|,$$

where a transport plan  $\gamma_{ij}$  belongs to the set  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}}$  in case it satisfies the following four criteria:

- (a)  $\gamma_{ij} \geq 0$ ,
- (b)  $\sum_j \gamma_{ij} \leq E_i$ ,
- (c)  $\sum_i \gamma_{ij} \leq E'_j$ , and
- (d)  $\sum_{ij} \gamma_{ij} = \min(\sum_i E_i, \sum_j E'_j)$ .

These criteria ensure that (i) the amount of mass moved between any two particles is always non-negative, (ii) the maximum amount of mass that can be moved from location  $i$  in  $\mathcal{E}$  to any location in  $\mathcal{E}'$  is  $E_i$ , (iii) the maximum amount of mass that can be moved to location  $j$  in  $\mathcal{E}'$  from any location in  $\mathcal{E}$  is  $E'_j$ , and (iv) the total mass that is moved equals the total mass of whichever event has smaller mass. If strict inequality holds in (b), we will say  $E_i - \sum_j \gamma_{ij}$  mass has been *destroyed* at  $x_i$ , and if strict inequality holds in (c), we will say  $E'_j - \sum_i \gamma_{ij}$  mass has been *created* at  $x'_j$ . Note that, when the measures  $\mathcal{E}$  and  $\mathcal{E}'$  have equal total mass,  $\text{EMD}^R(\mathcal{E}, \mathcal{E}') = \frac{1}{R} W_1(\mathcal{E}, \mathcal{E}')$ .

In fact, the energy mover's distance is a special case of the partial transport distance studied by Georgiou *et al.* [17], Caffarelli and McCann [18], Figalli [19], and Piccoli and Rossi [20, 21]: For  $\kappa > 0$ ,  $p \geq 1$ , define

$$T_p^\kappa(\mathcal{E}, \mathcal{E}') = \min_{\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}} (\sum_{ij} d_{ij}^p \gamma_{ij})^{1/p} + \frac{\kappa}{2} (|\sum_i E_i - \sum_j \gamma_{ij}| + |\sum_j E'_j - \sum_i \gamma_{ij}|), \quad (2)$$

where a transport plan  $\gamma_{ij}$  belongs to the set  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}$  in case it satisfies criteria (a)–(c) above. The two main differences between  $\text{EMD}^R$  and  $T_p^\kappa$  are that, first, the partial transport distances allow  $p \geq 1$  and, second, they permit the amount of mass that is rearranged from  $\mathcal{E}$  to  $\mathcal{E}'$  to differ from the total mass of whichever event has smaller mass. To see this, assume without loss of generality that  $\mathcal{E}$  has smaller total mass,  $\sum_i E_i \leq \sum_j E'_j$ . The distance  $\text{EMD}^R$  requires that all of the mass in  $\mathcal{E}$  be rearranged: Exactly  $\sum_j E'_j - \sum_i E_i$  mass is created in  $\mathcal{E}'$ , and no mass is destroyed. On the other hand,  $T_p^\kappa$  allows for  $\sum_{ij} \gamma_{ij} \in (0, \min(\sum_i E_i, \sum_j E'_j))$  mass to be rearranged:  $\sum_i E_i - \sum_{ij} \gamma_{ij}$  mass is destroyed in  $\mathcal{E}$ , and  $\sum_j E'_j - \sum_{ij} \gamma_{ij}$  mass is created in  $\mathcal{E}'$ .

Now we show why, for  $\kappa = 2R \geq \max_{ij} d_{ij}$ ,  $\text{EMD}^R$  coincides with (a constant multiple of)  $T_1^\kappa$ . First, note that the EMD constraint set is a subset of the Piccoli-Rossi constraint set,  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}} \subseteq \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{PR}}$ . Furthermore, if  $\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}}$ , then the values of the objective function in each minimization problem coincide, up to a factor of

$\kappa = 2R$ . Thus, if we can show that  $\kappa = 2R \geq \max_{ij} d_{ij}$  ensures that the optimizer  $\gamma_{ij}^*$  of  $T_1^\kappa$  belongs to the stricter constraint set  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}^{\text{EMD}}$ , we can conclude that

$$T_1^\kappa(\mathcal{E}, \mathcal{E}') = T_1^{2R}(\mathcal{E}, \mathcal{E}') = R \text{EMD}^R(\mathcal{E}, \mathcal{E}'). \quad (3)$$

Observe that, using properties (b) and (c) of the constraint set  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}$ , we may remove the absolute value signs in the definition of  $T_1^\kappa$  and express it equivalently as

$$T_1^\kappa(\mathcal{E}, \mathcal{E}') = \min_{\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \mathcal{E}')}} \sum_{ij} (d_{ij} - \kappa) \gamma_{ij} + \frac{\kappa}{2} (\Sigma_i E_i + \Sigma_j E'_j). \quad (4)$$

Thus, if  $\kappa \geq \max_{ij} d_{ij}$ , the coefficient on  $\gamma_{ij}$  is always negative, so the optimal  $\gamma_{ij}^*$  for the  $T_1^\kappa$  distance will be as large as possible, subject to the constraints (b) and (c). In particular, the optimal  $\gamma_{ij}^*$  will satisfy constraint (d) and belong to  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}$ .

The above argument not only establishes the equivalence between  $T_1^\kappa$  and  $\text{EMD}^R$  for  $\kappa = 2R \geq \max_{ij} d_{ij}$ , but also sheds light on the role of the parameter  $\kappa > 0$ . From Eq. (2), we observe that smaller  $\kappa$  makes creation and destruction cheaper and transport comparatively more expensive. In fact, using Eq. (4), we can make this quantitative: If  $\gamma_{ij}^*$  is the optimizer, then, for any  $i, j$  such that  $d_{ij} > \kappa$ , we must have  $\gamma_{ij}^* = 0$ . (If not, we could find a strictly better choice of  $\gamma$  in  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}$  by setting  $\gamma_{ij} = 0$ , contradicting that  $\gamma_{ij}^*$  was the optimizer.) In other words, energy will never be transported over a distance greater than  $\kappa$ .

## B. From partial optimal transport to the Hellinger-Kantorovich distance

One of the key contributions of Piccoli and Rossi's work on the partial optimal transport distance  $T_p^\kappa$  is a *dynamic* formulation of the distance [21]. This dynamic perspective is most clear when  $T_p^\kappa$  is stated in full generality, as a distance on the space of finite Borel measures  $\mathcal{M}(\Omega)$ : For  $\mu, \mu' \in \mathcal{M}(\Omega)$ ,  $\kappa > 0$ , and  $p \geq 1$ ,

$$T_p^\kappa(\mu, \mu') = \inf_{\gamma \in \Gamma_{\leq(\mu, \mu')}} \left( \iint |x - x'|^p d\gamma(x, x') \right)^{1/p} + \frac{\kappa}{2} \left( \left| \int \mu - \iint \gamma \right| + \left| \int \mu' - \iint \gamma \right| \right), \quad (5)$$

where we say  $\gamma \in \Gamma_{\leq(\mu, \mu')}$  in case  $\gamma \in \mathcal{M}(\Omega \times \Omega)$  satisfies  $\gamma(B \times \Omega) \leq \mu(B)$  and  $\gamma(\Omega \times B) \leq \mu'(B)$  for any Borel set  $B$ . Note that Eq. (5) reduces to Eq. (2) when  $\mu = \sum_{i \in I} \delta_{x_i} E_i$  and  $\mu' = \sum_{j \in J} \delta_{x'_j} E'_j$ .

Piccoli and Rossi [21, 24] show that  $T^\kappa$  has the following equivalent dynamic formulation:

$$\begin{aligned} T_p^\kappa(\mu, \mu') &= \inf_{\rho, v, \psi \in \mathcal{C}(\mu, \mu')} (A_p^\kappa[\rho, v, \psi])^{1/p}, \\ A_p^\kappa[\rho, v, \psi] &= \int_0^1 \int_\Omega (|v(x, t)|^p + (\kappa/2)|\psi(x, t)|) \rho(x, t) dx dt, \\ \mathcal{C}(\mu, \mu') &= \{ \rho \in C([0, 1], \mathcal{M}(\Omega)), v \in L^2(d\rho_t dt), \\ &\quad \psi \in L^1(d\rho_t dt) : \partial_t \rho + \nabla \cdot (\rho v) = \psi \rho, \rho(\cdot, 0) \\ &\quad = \mu, \rho(\cdot, 1) = \mu' \}. \end{aligned}$$

In other words, one can find the  $T_p^\kappa$  distance from  $\mu$  to  $\mu'$  by considering all curves  $\rho$  connecting  $\mu$  to  $\mu'$  with velocity  $v$  and reaction rate  $\psi$  and finding the curve with least action  $A_p^\kappa[\rho, v, \psi]$ .

This dynamic perspective reveals a general framework for unbalanced optimal transport problems, in terms of minimizing different notions of action. In particular, as observed in Ref. [24], for any  $\kappa > 0$ ,  $p \geq 1$ , and  $q \geq 1$ , one may consider

$$\begin{aligned} A_{p,q}^\kappa[\rho, v, \psi] &= \int_0^1 \int_\Omega (|v(x, t)|^p \\ &\quad + (\kappa/2)^q |\psi(x, t)|^q) \rho(x, t) dx dt. \end{aligned}$$

As before, large values of  $\kappa > 0$  penalize creation and destruction. In particular, sending  $\kappa \rightarrow +\infty$  (see Theorem 7.24 in Ref. [23]),

$$\begin{aligned} \lim_{\kappa \rightarrow +\infty} \inf_{\rho, v, \psi \in \mathcal{C}(\mu, \mu')} (A_{p,q}^\kappa[\rho, v, \psi])^{1/p} \\ = \begin{cases} W_p(\mu, \mu') & \text{if } \int \mu = \int \mu', \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

While minimizing the action  $A_{p,q}^\kappa[\rho, v, \psi]$  with  $q = 1$  yields the partial transport distance  $T_p^\kappa$  described in the previous section, minimizing it for  $p = q = 2$  yields the Hellinger-Kantorovich distance

$$\text{HK}^\kappa(\mu, \mu') = \inf_{\rho, v, \psi \in \mathcal{C}(\mu, \mu')} (A_{2,2}^\kappa[\rho, v, \psi])^{1/2}. \quad (6)$$

This case is distinguished among all  $p, q \geq 1$ , since it is the only choice that directly gives rise to an infinite-dimensional Riemannian manifold [24, 26]. Furthermore, not only does the Hellinger-Kantorovich metric have a well-defined limit as  $\kappa \rightarrow +\infty$  whenever  $\mu$  and  $\mu'$  have equal mass,  $\lim_{\kappa \rightarrow +\infty} \text{HK}^\kappa(\mu, \mu') = W_2(\mu, \mu')$ , the  $\kappa \rightarrow 0$  limit is also well defined for arbitrary  $\mu, \mu'$ :

$$\lim_{\kappa \rightarrow 0} \frac{1}{\kappa} \text{HK}^\kappa(\mu, \mu') = \left( \int \left| \sqrt{\frac{d\mu}{dx}} - \sqrt{\frac{d\mu'}{dx}} \right|^2 dx \right)^{1/2}, \quad (7)$$

which is known as the Hellinger distance [22–24].

Like the original definition of the partial optimal transport distances  $T_p^\kappa$ , it can also be expressed in terms of a

static minimization problem, which for simplicity, we state in the case of fully discrete measures  $\mathcal{E}, \mathcal{E}'$ :

$$\begin{aligned} \text{HK}^\kappa(\mathcal{E}, \mathcal{E}') &= \min_{\gamma_{ij} \geq 0} \sum_{ij} (\ell^\kappa(d_{ij})\gamma_{ij} + \kappa^2 \text{KL}(\mathcal{G}, \mathcal{E})) \\ &\quad + \kappa^2 \text{KL}(\mathcal{G}', \mathcal{E}')^{1/2}, \end{aligned} \quad (8)$$

where  $\mathcal{G}$  and  $\mathcal{G}'$  are auxiliary discrete measures, with  $\mathcal{G}$  assigning mass  $G_i = \sum_j \gamma_{ij}$  to location  $x_i$  and  $\mathcal{G}'$  assigning mass  $G'_j = \sum_i \gamma_{ij}$  to  $x_j$ , and

$$\begin{aligned} \ell^\kappa(s) &= \begin{cases} -2\kappa^2 \log(\cos^2(s/\kappa)) & \text{if } s < \frac{\pi}{2}\kappa, \\ +\infty, & \text{otherwise,} \end{cases} \\ \text{KL}(\mathcal{G}, \mathcal{E}) &= \sum_i E_i f\left(\frac{G_i}{E_i}\right), \quad f(s) = s \log(s) - s + 1. \end{aligned} \quad (9)$$

The equivalence between Eqs. (6) and (8) is a significant mathematical result, due to Liero, Mielke, and Savaré, based on a surprising connection with cone geometry [22,23].

The optimizer  $\gamma_{ij}$  of Eq. (8) represents how much mass is transported from  $x_i$  in  $\mathcal{G}$  to  $x'_j$  in  $\mathcal{G}'$ ; that is,  $\gamma_{ij}$  is the optimal transport plan from  $\mathcal{G}$  to  $\mathcal{G}'$ . In general,  $G_i \neq E_i$  and  $G'_j \neq E'_j$ , and the energy that is not transported can be thought of as having been created or destroyed. In particular,

- (i) if  $G_i > E_i$ , we say energy was created at  $x_i$ ;
- (ii) if  $G_i < E_i$ , we say energy was destroyed at  $x_i$ ;
- (iii) if  $G_j > E'_j$ , we say energy was destroyed at  $x_j$ ;
- (iv) if  $G_j < E'_j$ , we say energy was created at  $x_j$ .

[Note that the first and third options did not arise for the  $T_p^\kappa$  distance, due to requirements (b) and (c) for the set of transport plans  $\Gamma_{\leq(\mathcal{E}, \mathcal{E}')}$ .] While until now we have always assumed that our discrete measures have strictly positive energy at every location,  $E_i, E'_j > 0$ , observe that it is possible for  $G_i$  or  $G'_j$  to be zero.

The first term in the minimization problem in Eq. (8) penalizes transporting energy over long distances. As with  $T_p^\kappa$ , small values of  $\kappa$  penalize transport: Energy will never be transported over distance greater than  $\frac{\kappa\pi}{2}$ . The second two terms penalize the difference between  $\mathcal{G}$  and  $\mathcal{E}$  and between  $\mathcal{G}'$  and  $\mathcal{E}'$ , in terms of the Kullback-Liebler divergence.

A major difference between the Hellinger-Kantorovich metric and the 2-Wasserstein metric considered in the authors' previous work [3] is that the Hellinger-Kantorovich metric allows for the comparison of events with unequal total energy. However, even when the total energy of events  $\mathcal{E}$  and  $\mathcal{E}'$  coincide,  $\text{HK}^\kappa(\mathcal{E}, \mathcal{E}')$  is, in general, not equal to  $W_2(\mathcal{E}, \mathcal{E}')$ . This can be seen, for example, from Eqs. (8) and (9): Mass will never be transported more than distance  $\frac{\kappa\pi}{2}$ . Interestingly, the

converse is also true: If mass is not transported from  $x_i$  to  $x'_j$ , that is, if  $\gamma_{ij} = 0$ , then we must have  $d_{ij} = \|x_i - x_j\| \geq \frac{\kappa\pi}{2}$  (see Lemma 3.13 in Ref. [27]).

As already observed in the collider physics context for the special case  $\kappa = +\infty$ ,  $p = q = 1$  [1], Chizat *et al.* identified that minimizing the action  $A_{p,q}^\kappa$  has a dual characterization in terms of the following maximization problem:

$$\begin{aligned} \inf_{\rho, v, \psi \in \mathcal{C}(\mu, \mu')} A_{p,q}^\kappa[\rho, v, \psi] &= \sup_{\varphi \in \text{HJ}_{p,q}^\kappa} \int_\Omega \varphi(x, 1) d\mu'(x) \\ &\quad - \int_\Omega \varphi(x, 0) d\mu(x), \\ \text{HJ}_{p,q}^\kappa &= \left\{ \varphi \in C^1([0, 1] \times \Omega) : \right. \\ &\quad \left. \partial_t \varphi + \frac{|\nabla \varphi|^{p'}}{p'} + (\kappa/2)^{-q'} \frac{|\varphi|^{q'}}{q'} \leq 0 \right\}, \end{aligned}$$

where  $p' = \frac{p}{p-1}$  and  $q' = \frac{q}{q-1}$  [24]; this has been rigorously justified in the cases  $p = q = 2$  [23] and  $p = q = 1$  [21]. In the above equations, we use the following convention if either  $p = 1$  or  $q = 1$ : As  $p \rightarrow 1$  or  $q \rightarrow 1$ , the second and third terms in the sum would diverge to  $+\infty$  unless  $|\nabla \varphi| \leq 1$  or  $|\varphi| \leq \kappa/2$ , in which case each term would converge to zero. Consequently, if either  $p = 1$  or  $q = 1$ , we drop the respective term from the inequality and add the constraint  $|\nabla \varphi| \leq 1$  or  $|\varphi| \leq \kappa/2$ . Furthermore, if we first send  $\kappa \rightarrow +\infty$ , to reduce to the  $p$ -Wasserstein, since  $q' \geq 1$ , the second term in the inequality constraint will vanish. In this way, for the case  $\kappa = +\infty$ ,  $p = q = 1$  considered by Komiske *et al.* [1], we see that we may replace  $\text{HJ}_{1,1}^{+\infty}$  by  $\{\varphi(x, t) : \varphi(x, t) = \Phi(x) \in C^1(\mathbb{R}^d), \sup_x |\nabla \Phi(x)| \leq 1\}$ , the space of 1-Lipschitz functions. Indeed, it is the comparatively simple formulation of the dual problem in the case  $p = q = 1$  for all  $\kappa > 0$  that makes such metrics popular in practice. However, we choose to work with  $p = q = 2$ , due to its superior geometric properties, which allow us to linearize the metric, vastly improving computational efficiency of our method, while preserving key features of the optimal transport distance.

### C. Linearized Hellinger-Kantorovich metric

We now describe the linearization of the Hellinger-Kantorovich metric, as introduced in Ref. [27], that we use in the present work. We begin by explaining how to construct an embedding of events  $\mathcal{E}$  into Euclidean space  $\mathbb{R}^{dn} \times \mathbb{R}^n$ , using the optimal transport plan  $\gamma_{ij}$  in the Hellinger-Kantorovich metric; see Eq. (8). We will then describe how computing the distance between the embeddings provides a linearization of the Hellinger-Kantorovich distance.

Let  $\mathcal{R}$  denote a discrete reference measure, consisting of particles at locations  $\{x_i\}_{i \in I}$  with positive masses  $\{R_i\}_{i \in I}$ .

For any discrete measure  $\mathcal{E}'$ , let  $\gamma_{ij}$  denote an optimizer of Eq. (8), which represents an optimal transport plan from the auxiliary measures  $\mathcal{G}$  to  $\mathcal{G}'$ . (Note that there may be more than one optimizer.) In general, the transport plan  $\gamma_{ij}$  may send mass from  $x_i$  in  $\mathcal{G}$  to many different locations in  $\mathcal{G}'$ . In order to linearize the Hellinger-Kantorovich metric, we first consider the average of these locations, weighted by how much mass is sent to each place and normalized by the amount of mass starting at  $x_i$  in  $\mathcal{G}$ :

$$z_i = \begin{cases} \frac{1}{G_i} \sum_j \gamma_{ij} x'_j & \text{if } G_i > 0, \\ x_i & \text{if } G_i = 0. \end{cases} \quad (10)$$

Next, we consider the average amount that mass starting at location  $x_i$  needs to be rescaled, via creation or destruction, in order for  $\mathcal{R}$  to become  $\mathcal{E}'$ : For each  $x'_j$ , consider the ratio  $E'_j/G'_j$ , between the amount of mass that must end up at location  $x'_j$  and the amount of mass transported by  $\gamma_{ij}$  to  $x'_j$ . If  $E'_j/G'_j > 1$ , mass needs to be created at  $x_j$ , and if  $E'_j/G'_j < 1$ , mass needs to be destroyed at  $x_j$ . Note that this quantity is well defined only for  $G'_j = \sum_i \gamma_{ij} > 0$ . In fact, this is a necessary assumption for the Hellinger-Kantorovich metric to be linearized in a manner that admits a Euclidean embedding (see p. 18 in Ref. [27]). Recall from the previous section that a sufficient condition for  $\gamma_{ij} > 0$  is  $d_{ij} = \|x_i - x'_j\| < \frac{\kappa\pi}{2}$ . Consequently, in what follows, we will suppose that  $\kappa$  is sufficiently large so that, for each  $x'_j$ , there exists  $x_i$  so that

$$\|x_i - x'_j\| < \frac{\kappa\pi}{2}. \quad (11)$$

This will ensure  $G'_j > 0$  for all  $j$ .

With this assumption in hand, we now consider, for each fixed  $x_i$ , the weighted average of this ratio, representing how much mass needs to be created or destroyed at  $x_j$ , with respect to how much mass  $\gamma_{ij}$  transports to each  $x'_j$ , normalized by the amount of mass  $G_i$  originally starting at  $x_i$ :

$$u_i = \begin{cases} \frac{1}{G_i} \sum_j \left( \frac{E'_j}{G'_j} \right) \gamma_{ij} & \text{if } G_i > 0, \\ 0 & \text{if } G_i = 0. \end{cases} \quad (12)$$

While the coordinate  $z_i$ , defined in Eq. (10), represents the average location that mass starting at  $x_i$  is transported to in  $\mathcal{E}'$ , the coordinate  $u_i$  represents the average amount of creation or destruction that will happen to mass that started at  $x_i$ , after it is transported.

With these quantities in hand, we may now state the formula for the linearized Hellinger-Kantorovich metric. In analogy with the linearized 2-Wasserstein metric, which is known as LOT [3,28], we will refer to the linearization of the Hellinger-Kantorovich metric for discrete measures as

PLUOT, emphasizing that it is a discrete particle approximation of the continuum linearization of the Hellinger-Kantorovich metric studied in previous work by the first two authors [27]:

$$\begin{aligned} \text{PLUOT}^\kappa(\mathcal{E}', \tilde{\mathcal{E}}') &= \left( \sum_i R_i \|v_i - \tilde{v}_i\|^2 + \frac{\kappa^2}{4} R_i |\alpha_i - \tilde{\alpha}_i|^2 \right)^{1/2}, \\ v_i &= \kappa \text{sgn}(z_i - x_i) \sqrt{u_i G_i / R_i} \sin(\|z_i - x_i\| / \kappa), \\ \alpha_i &= 2(\sqrt{u_i G_i / R_i} \cos(\|z_i - x_i\| / \kappa) - 1). \end{aligned} \quad (13)$$

Note that this approximation depends on the choice of the optimal transport plans  $\gamma_{ij}$ ,  $\tilde{\gamma}_{ij}$  via their dependence on  $x_i$ ,  $z_i$ ,  $\tilde{x}_i$ ,  $\tilde{z}_i$ ; see Eqs. (10) and (12).

As in the definition of  $\text{HK}^\kappa$ , the unusual expressions for  $v_i$  and  $\alpha_i$  in PLUOT derive from the surprising connection to cone geometry [22–24]. In particular, when comparing the locations and masses of particles  $(x_i, E_i)$ , the cone structure is used to identify all points with mass zero as the same point. For example, in one spatial dimension [and under assumption (11)],  $(x_i, E_i)$  corresponds to the point  $(E_i \cos(x_i/\kappa), E_i \sin(x_i/\kappa))$  in the plane.

To see the connection with Eq. (13), consider the original location and mass of the  $i$ th particle,  $(x_i, R_i)$ , along with the average location to which its mass is sent and the average mass at that location after creation or destruction,  $(z_i, u_i G_i)$ . The *constant speed geodesic* in the *cone metric* between these two points represents how the location  $x_i$  is optimally transported to  $z_i$ , while simultaneously mass is created and destroyed to convert  $R_i$  into  $u_i G_i$  [23]. (In one spatial dimension, this is just the line connecting the two points in the plane.) The coordinate  $v_i$  represents the velocity of the spatial trajectory at time zero, while  $R_i \alpha_i$  represents the rate of change of the mass at time zero. From this perspective, PLUOT measures the difference between two events  $\mathcal{E}'$  and  $\tilde{\mathcal{E}}'$  in terms of how a reference event  $\mathcal{R}$  deforms into  $\mathcal{E}'$  and  $\tilde{\mathcal{E}}'$ , by comparing the velocities by which particles in the reference event move and the rates at which their masses change.

In analogy with LOT for balanced optimal transport, a key benefit of the PLUOT approximation of the Hellinger-Kantorovich metric is that it provides a natural embedding

$$\mathcal{E}' \mapsto (v_i, \alpha_i)_{i \in I} \in \mathbb{R}^{dn} \times \mathbb{R}^n, \quad (14)$$

where  $d$  is the dimension of the underlying domain  $\Omega$  in which particles are located and  $n$  is the number of particles in the discrete reference measure,  $n = |I|$ . This vector may be interpreted geometrically as an approximation of the tangent vector from  $\mathcal{R}$  to  $\mathcal{E}$  with respect to the Hellinger-Kantorovich geometry, an interpretation that may be made precise when  $\mathcal{R}$  is a finite Borel measure that is absolutely continuous with respect to Lebesgue measure

(see Definition 4.5 in Ref. [27]). In this way, it is natural to compare two discrete measures  $\mathcal{E}'$  and  $\tilde{\mathcal{E}}'$  by computing the distance between the vectors  $(v_i, \alpha_i)$  and  $(\tilde{v}_i, \tilde{\alpha}_i)$  as elements of the tangent space at  $\mathcal{R}$ , as in Eq. (13) above.

The above Euclidean embedding [Eq. (14)] is useful from the perspective of classification algorithms for two reasons. First, while computing the embedding  $\mathcal{E}' \mapsto (v_i, \alpha_i)$  for each event  $\mathcal{E}'$  in a sample requires  $O(N_{\text{evt}})$  computations of the Hellinger-Kantorovich metric in Eq. (8), computing the linearized distance between all pairs of events  $\mathcal{E}'$  and  $\tilde{\mathcal{E}}'$  using PLUOT requires only  $O(N_{\text{evt}}^2)$  computations of the weighted Euclidean metric; see Eq. (13). Given that computing a weighted Euclidean metric is several orders of magnitude faster than computing the Hellinger-Kantorovich metric, our approach using PLUOT offers a substantial computational advantage compared to computing the exact Hellinger-Kantorovich distance between all pairs of events.

The second reason that the Euclidean embedding in PLUOT is useful in classification tasks is that it allows us to apply a wider range of classification algorithms directly to the vectors  $(v_i, \alpha_i)$  and  $(\tilde{v}_i, \tilde{\alpha}_i)$  representing the discrete measures  $(\mathcal{E}', \tilde{\mathcal{E}}')$ , including classification algorithms that require a Euclidean structure. In particular, we are often able to delegate the computation of the entire pairwise distance matrix to efficient downstream methods, leading to a large storage advantage over other methods.

In our analysis of the linearized Hellinger-Kantorovich metric as a tool to classify jets, we will investigate the effects of creation and destruction in the Hellinger-Kantorovich metric separately from the fact that it allows for the comparison of events with unequal total energy. We do this by separately analyzing the classification performance of the linearization of  $\text{PLUOT}^\kappa(\mathcal{E}, \mathcal{E}')$  with the classification performance of  $\text{PLUOT}^\kappa(\mathcal{E}/\sum_i E_i, \mathcal{E}'/\sum_j E'_j)$ , where  $\mathcal{E}/(\sum_i E_i)$  denotes the normalized measure, in which the mass  $E_i$  of each particle in  $\mathcal{E}$  is replaced by  $E_i/(\sum_i E_i)$ . The Hellinger-Kantorovich metric exhibits a simple scaling under this transformation (see Theorem 3.3 in Ref. [34]): Denoting  $m = \sum_i E_i$ ,

$$\begin{aligned} (E'_j)^{\text{norm}} &= m^{-1/2} E'_j, & \gamma_{ij}^{\text{norm}} &= m^{-1/2} \gamma_{ij}, \\ G_i^{\text{norm}} &= m^{-1/2} G_i, & (G'_j)^{\text{norm}} &= m^{-1/2} G'_j, \\ z_i^{\text{norm}} &= z_i, & u_i^{\text{norm}} &= m^{-1/2} u_i, \\ v_i^{\text{norm}} &= m^{-1/2} v_i, & \alpha_i^{\text{norm}} &= m^{-1/2} \alpha_i + 2(m^{-1/2} - 1). \end{aligned} \quad (15)$$

### III. JET CLASSIFICATION WITH UNBALANCED OPTIMAL TRANSPORT

We now demonstrate the practical relevance of linearized, unbalanced optimal transport to collider physics by

applying the PLUOT framework to the task of boosted jet tagging. Discrimination between boosted jets stemming from the decay of heavy particles and QCD backgrounds composed of quark and gluon jets is a key component of many analyses at the LHC. Here, we focus exclusively on distinguishing boosted  $W$  boson jets from QCD backgrounds for ready comparison to previous studies applying optimal transport techniques to boosted jet tagging [1,3], though the same analysis can be easily extended to other pairwise tagging tasks.

Compared to previous transport-only methods [3], unbalanced optimal transport now allows for the creation and destruction of energy in comparing distributions. For the Hellinger-Kantorovich distance, the relative importance of mass creation and destruction over mass transportation is controlled by the intrinsic length scale parameter  $\kappa$ , as explained in Sec. II and more extensively in Ref. [27]. Insofar as collider physics applications typically feature one or more additional length scales (such as the jet clustering radius  $R$ ), a key question is how HK-based classification depends on the value of  $\kappa$  relative to other length scales. To this end, we consider values of  $\kappa$  ranging over several orders of magnitude, i.e.,  $\kappa \in [0.01, 100]$ .

To develop intuition for the behavior of the linearized Hellinger-Kantorovich as a function of  $\kappa$ , in Fig. 1 we show the optimal transport (OT) plans corresponding to either the  $W_2$  or  $\text{HK}^\kappa$  metrics for various values of  $\kappa$ ; see Eqs. (1) and (8). With an eye toward linearization, the OT plans are calculated between sample jets and uniform reference measures. In the first row, OT plans are constructed between an artificial jet composed of a single particle and a reference measure consisting of an  $8 \times 8$  grid of particles. In the second row, OT plans are constructed between a simulated boosted  $W$  jet and a  $15 \times 15$ -particle uniform measure.

For the single-particle artificial jet, both the mass transportation and mass creation and destruction are symmetric with respect to the origin. Less and less mass is created or destroyed as  $\kappa$  is increased, and the HK distance approaches the value of the  $W_2$  distance; see the first three columns in Fig. 1. On the other hand, a small  $\kappa$  essentially reduces the HK distance to the Hellinger distance, which corresponds to the ordinary Euclidean difference between rescaled images, where no mass is being transported; see Eq. (7). The regime of intermediate  $\kappa$  is perhaps most interesting, as contributions from both mass transportation and creation and destruction are important.

To study the performance of the linearized Hellinger-Kantorovich distance in boosted jet classification, we consider simulated data consisting of 200 000 boosted  $W$  jets and QCD jets, generated as in Ref. [3]. Proton-proton collision events at  $\sqrt{s} = 14$  TeV are simulated in MADGRAPH 2.9.2 [35] with  $W$  bosons being pair produced, gluons generated via  $q\bar{q} \rightarrow Z \rightarrow \nu\bar{\nu}g$ , and quarks via  $qg \rightarrow Z \rightarrow \nu\bar{\nu}q$ . The particles are then hadronized and decayed in PYTHIA 8.302 [36], where default tuning and

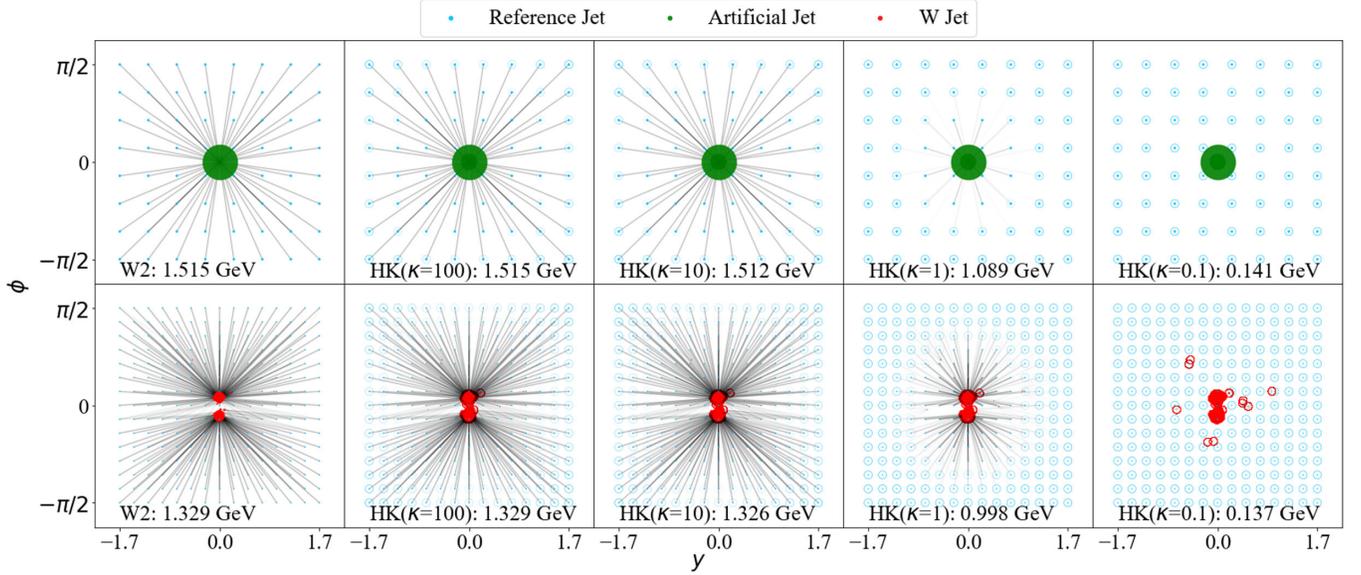


FIG. 1. Optimal transport plans from uniform reference measures to jets. In the top row, the reference measure (blue) consists of  $8 \times 8$  particles and is transported to an artificial jet (green), composed of a single particle at the origin. In the second row, the uniform reference measure (blue) consists of  $15 \times 15$  particles and is transported to a simulated  $W$  jet (red). The columns correspond to different choices of OT metric:  $W_2$  and  $\text{HK}^\kappa$  with  $\kappa = 100, 10, 1, 0.1$  (from left to right). The size of the filled dots indicates the amount of  $p_T$  at that point. The darkness of the lines indicates how much  $p_T$  is moved from one particle to another. For  $\text{HK}^\kappa$ , the thickness of the circles around the points represents how much  $p_T$  is destroyed for that particular particle. Also shown at the bottom of the plots are the total OT distances between the jets, which are similar for  $\kappa = +\infty, 100, 10$ , the transport regime.

showering parameters are used. Afterward, we cluster the events into jets using `FastJet 3.3.4` [37] with an anti- $k_T$  algorithm (jet radius  $R = 1$ ) where at most two jets are kept with  $|y| \leq 1.7$  and  $|\phi| \leq \frac{\pi}{2}$ . Before calculating their linearized  $W_2$  or  $\text{HK}$  embedding, we boost and rotate the jets to center the jet 4-momentum and vertically align the principal component of the constituent  $p_T$  flow in the  $y - \phi$  plane using the `EnergyFlow` package [1,4,38–40]. The pre-processing procedure removes artificial differences in the energy flows of the jets.

Once the Euclidean embedding for each jet is acquired via PLUOT [see Eq. (14)], we employ simple machine learning algorithms such as  $k$ -nearest neighbors (kNN) and support vector machine (SVM) to classify the jets; see Sec. IV in Ref. [3] for a detailed discussion of the models. Given that our main interest here is the optimal transport distance itself, we limit ourselves mostly to kNN and do not consider other potentially more powerful machine learning models or deep neural networks, though their use can be easily implemented and incorporated into the present framework.

We consider classification of boosted jets in datasets consisting of either 10 000 or 200 000  $W$  and QCD jets. For datasets with 10 000 jets, we use 5000 jets to train kNN and SVM, 2500 for validation in order to pick the best model hyperparameter(s), and the remaining 2500 jets as the test dataset to obtain the model performance. We try  $k \in [10, 100]$  with an increment of 10 for kNN and

$C, \gamma \in [10^{-2}, 10^5]$  for SVM where only powers of 10 are considered. These ranges are chosen by experience so as to ensure the coverage of the optimal value of the hyperparameter(s) but as little of anything else as possible. When dealing with the full 200 000 dataset, we use 150 000 jets to train the models and 50 000 to evaluate the performance, where the model hyperparameter(s) are already picked by smaller runs with the 10 000 datasets.

We compare the tagging performance of kNN based on PLUOT to that of  $N$ -subjettiness  $\tau_N$ , a popular jet substructure observable designed to capture the prongedness of a jet [41,42]. Since  $W$  jets typically have two prongs and QCD jets are more diffuse and single pronged, the  $N$ -subjettiness ratio  $\tau_{21} = \tau_2/\tau_1$  is particularly well suited to the task at hand. Here  $\tau_N$  is determined using the `Nsubjettiness` plug-in package in `FastJet` [41,42].

Another benchmark for gauging the performance of PLUOT is the pairwise EMD distance matrix [1] coupled with the same machine learning models. We test the EMD both on normalized jets, where the jets are first rescaled to have  $p_T = 1$ , as well as on unnormalized jets. The ability to compare both normalized and unnormalized jets is implemented by a built-in function in the `EnergyFlow` package [1,4,38–40], with the parameters  $R, \beta = 1$  and the normalization parameter norm set, respectively, to true or false.

In a similar manner, the PLUOT framework also presents us with two options to calculate the Euclidean embedding. One way is to compute the unbalanced OT

distances directly between jets with different total  $p_T$ , as in Eq. (8). Alternatively, we can again normalize the jets so that each has  $p_T = 1$  and then compute the unbalanced  $\text{HK}^\kappa$  distance between the normalized jets. We emphasize that, even when two jets have equal total  $p_T$ , as in the case of balanced OT, the HK distance still allows for local mass to be created and destroyed. In fact, the normalized and unnormalized approaches are related by simple scaling transformations; see Eq. (15). For this reason, in practice we begin by computing the Euclidean embedding of normalized jets and then invert Eq. (15) to recover the embedding of the unnormalized jets. Hereafter, we abbreviate the distances calculated on normalized jets with a subscript of  $N$  and those obtained for unnormalized jets with subscript  $unN$ .

Previous studies of jet classification based on OT have been relatively insensitive to differences in total  $p_T$  among different jets in the sample, typically considering events drawn from narrow (50 GeV)  $p_T$  bins. Indeed, in Ref. [3], it was observed that classification based on balanced optimal transport distances between normalized jets drawn from a 50 GeV  $p_T$  bin modestly outperformed unbalanced optimal transport distances using the EMD. To better assess the effects of unbalanced samples, we explore events drawn from a broader range of total  $p_T$ , extending from [500, 550] GeV in our previous study [3] to [500, 1500] GeV. This is achieved by stacking 20 datasets, each containing 10 000 jets with a  $p_T$  bin of 50 GeV, i.e.,  $p_T \in [500, 550]$  GeV for the first dataset,  $p_T \in [550, 600]$  GeV for the second, and so forth. In this way, in addition to the 20 datasets each with 50 GeV  $p_T$  bin width, we have a combined dataset of 200 000 jets in which the total jet  $p_T$  is approximately uniformly distributed between 500 and 1500 GeV. We examine the PLUOT framework on the classification of jets with widely different total  $p_T$  in Sec. III A.

When working with linearized OT, a reference measure should be chosen in advance against which the OT distance of simulated jets are computed. Loosely speaking, the reference measure is the point on the manifold of events that defines the tangent plane for linearized OT. Our default is a uniform jet with a total  $p_T = 750$  GeV and  $15 \times 15 = 225$  constituent particles, whose  $p_T$  is distributed uniformly on the  $y - \phi$  rectangle  $[-1.7, 1.7] \times [-\pi/2, \pi/2]$ . Such a reference measure has about the same number of particles as in a typical  $W$  or QCD jet in our sample of simulated events. Other uniform references considered in this study have  $4 \times 4 = 16$ ,  $8 \times 8 = 64$ ,  $30 \times 30 = 900$ , and  $60 \times 60 = 3600$  constituent particles. The interparticle spacing  $l$  of these reference measures differs widely, ranging from roughly 0.05 to 0.85. This defines yet another length scale in addition to the HK scale parameter  $\kappa$ , the jet clustering radius  $R$ , and the characteristic angular separation of the partonic decay products of a boosted particle of mass  $m$ ,  $\propto m/p_T$ . We study the effect of the reference spacing  $l$  on

the tagging performance in Sec. III B and summarize the interplay of the various scales in Sec. III C.

### A. $p_T$ range

In this subsection, we study the performance of classification based on unbalanced optimal transport as a function of the  $p_T$  range of the simulated events, comparing the tagging performance of  $W$  vs QCD jets whose total  $p_T \in [500, 550]$  or  $[500, 1500]$  GeV. The three OT distances examined are (i) the EMD distance on normalized jets ( $\text{EMD}_N$ ) and unnormalized jets ( $\text{EMD}_{unN}$ ); (ii) the balanced  $W_2$  distance on normalized jets; and (iii) the  $\text{HK}^\kappa$  distance on both normalized and unnormalized jets (denoted as  $\text{HK}_N$  or  $\text{HK}_{unN}$ , respectively). The N-subjettiness ratio  $\tau_{21}$  is also computed for each jet where classification using  $\tau_{21}$  serves as a benchmark.

For the  $\text{HK}^\kappa$  distance, we consider the  $\kappa$  values  $+\infty, 100, 10, 1, 0.5, 0.4, 0.3, 0.2, 0.1, 0.07, 0.05, 0.03$ , and  $0.01$ , with  $\kappa = +\infty$  denoting the  $W_2$  distance. Here, the reference measure is taken to be a uniform jet with a total  $p_T = 750$  GeV and  $15 \times 15 = 225$  particles. Since it is impossible to calculate and store the entire distance matrix for 200 000 jets using the EMD approach with reasonable computational resources, we compute EMD distances on only the 10 000 datasets, whereas the linear  $W_2$  and  $\text{HK}^\kappa$  embedding can be calculated efficiently for the full 200 000 datasets.

Figure 2 shows the tagging performance in terms of the area under the curve (AUC) score, a number in  $[0, 1]$  where 1 indicates a perfect classifier and 0.5 corresponds to random guessing. A more detailed table including the true positive rate (TPR), the false positive rate (FPR), and the optimal hyperparameters is presented in Appendix, where results from other tasks in Sec. III B are also included. A discussion of the general trends of the tagging performance not specific to the present task is deferred to Sec. III C.

As can be seen, for jets drawn from a 50-GeV-wide  $p_T$  bin (column 1), classification performance on either normalized or unnormalized jets is almost indistinguishable for linearized  $\text{HK}^\kappa$  distances with small  $\kappa$  values ( $\kappa \leq 1$ ). The EMD approach also produces similar AUC scores regardless of whether or not the jets are normalized, with kNN slightly preferring the normalized approach and SVM favoring the unnormalized version. The percentage differences in the AUC are within 1.5%, consistent with statistical fluctuations. Such behavior is to be expected, since normalization should not make a big difference when the total  $p_T$  difference among jets is small. Additionally, the tagging performance of the LOT approximation, including  $W_2$  and  $\text{HK}^\kappa$  (with the exception of  $\text{HK}_{unN}^\kappa$  for large  $\kappa$ ) approaches the same (or better, in the case of SVM) level of accuracy of the EMD method, with far less computational expense.

The effect of normalization becomes significant when the  $p_T$  bin width is broadened. For jets with

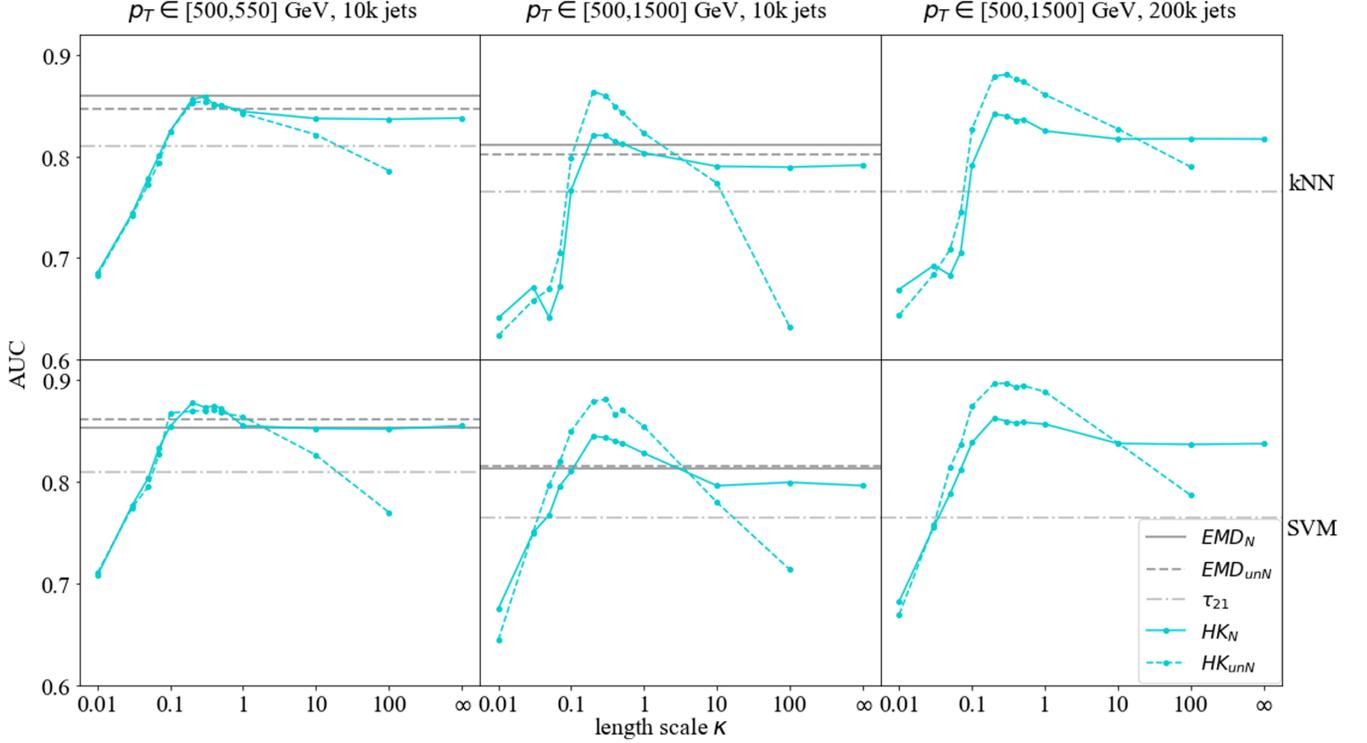


FIG. 2. AUC scores for classifying  $W$  vs QCD jets using kNN and SVM models coupled to linear  $W_2$  and  $HK^\kappa$  embedding with  $\kappa \in [0.01, +\infty]$ . Jet  $p_T$  ranges from 500 to 550 (1500) GeV in the first (second, third) column. The datasets for column 1 and column 2 (column 3) have 10 000 (200 000) jets. Solid (dashed) blue lines show the results calculated on normalized (unnormalized) jets; horizontal gray solid (dashed) lines use the EMD metrics on normalized (unnormalized) jets; and gray dash-dotted lines give the performance using  $\tau_{21}$  as the discriminator.

$p_T \in [500, 1500]$  GeV (10 000 for column 2 and 200 000 for column 3), the HK distance with  $\kappa$  in its optimal range calculated directly on the unnormalized jets (dashed blue lines) gives superior performance to the normalized jets (solid blue lines), whether we use kNN or SVM as the coupled model. The increase in AUC reaches about 5% at its peak when  $\kappa \sim 0.2$ . There the AUC from the HK distance, whether normalized or not, is noticeably higher than when using the EMD distance.

Interestingly, such performance gain is not observed in the EMD approach. Here, it makes no notable difference whether we use  $EMD_N$  (solid gray line) or  $EMD_{unN}$  (dashed gray line). This implies that, though the difference in total jet  $p_T$  has potential discriminating power, not all approaches to unbalanced optimal transport take advantage of it. A simple difference term like  $|p_T(\text{jet } 1) - p_T(\text{jet } 2)|$ , as included in the original EMD formulation, does not lead to improved discrimination for samples drawn from a larger  $p_T$  range. In contrast, unbalanced  $HK^\kappa$ , especially  $HK_{unN}$ , appears to take better advantage of this information by allowing local mass to be created and destroyed in addition to being transported.

Note that, while the original formulation of the EMD in the particle physics literature considered a fixed scale parameter  $R = \frac{\kappa}{2} \geq \max_{ij} d_{ij}/2$  coinciding with the jet

clustering radius, one could perform a similar analysis by using the more general partial transport distance, investigating how different choices of  $R = \frac{\kappa}{2}$  lead to different amounts of creation and destruction and, potentially, improved AUC in certain regimes. However, due to the fact that such metrics lack a Riemannian structure amenable to linearization, the analysis of finding the optimal parameter  $R = 2\kappa$  would be extremely computationally intensive.

## B. Reference measures

In the PLUOT framework, we are, in principle, free to pick any reasonable measure as our reference jet. Ideally, the choice of a reference measure should not exert too large an impact on the calculated linear  $W_2$  and HK embedding and the downstream tagging performance. Since the reference measure is associated with its own scale, the interparticle spacing  $l$ , it is natural to consider the interplay between  $l$  and the HK scale parameter  $\kappa$  in determining what constitutes a reasonable measure in practice.

To this end, we examine five uniform reference jets consisting of  $4 \times 4$ ,  $8 \times 8$ ,  $15 \times 15$  (the default),  $30 \times 30$ , and  $60 \times 60$  particles, respectively, denoted as uniref4, uniref8, uniref15, uniref30, and uniref60. In Fig. 3, we show the distribution of the Euclidean norms of the LOT

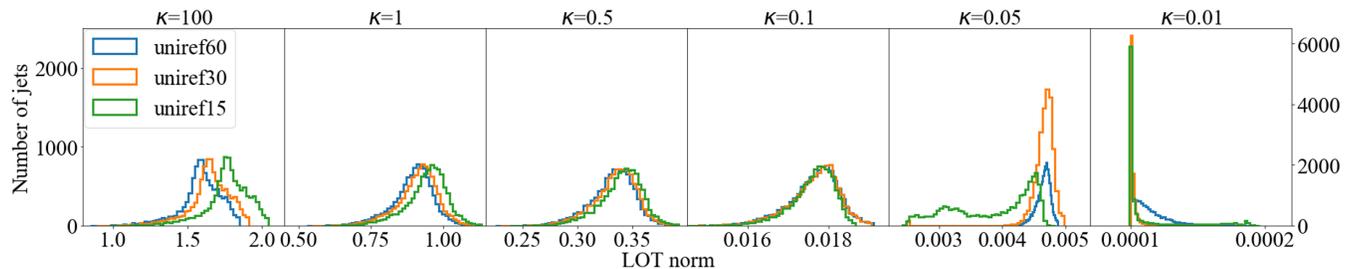


FIG. 3. Distribution of the LOT norm, i.e., the distance from each jet’s LOT coordinate to the origin, for the HK distance with  $\kappa = 100, 1, 0.5, 0.1, 0.05, 0.01$ . The uniform reference measures used include uniref15 with  $15 \times 15$  particles, uniref30, and uniref60. The y coordinate of the rightmost plot follows the scale on the right, while the other plots follow the scale on the left.

coordinates of 10 000 jets ( $p_T \in [500, 550]$  GeV) with  $\text{HK}^\kappa$  using uniref15, uniref30, and uniref60.<sup>2</sup> As  $\kappa$  is decreased from large values  $\kappa \sim 100$ , the distribution of the norms using the  $\text{HK}^\kappa$  distance becomes more and more similar for different reference measures. The closest agreement occurs for  $\kappa \sim 0.1$ , which we will see later is the  $\kappa$  value that gives the optimal tagging performance. As  $\kappa$  is decreased below  $\kappa \sim 0.1$  and we enter a scaled Euclidean image difference regime, the discrepancy of the norms using different reference measures becomes noticeable. We will see that this instability with respect to the chosen reference measure translates to deterioration of the tagging performance for small  $\kappa$  values.

Figure 4 shows the tagging performance on 10 000 jets with total  $p_T \in [500, 550]$  (first row) and  $[500, 1500]$  GeV (second row) using  $\text{EMD}_N$ ,  $\text{EMD}_{unN}$ ;  $\text{HK}_N$ ,  $\text{HK}_{unN}$ ; and the N-subjettiness ratio  $\tau_{21}$ . Tagging performance is plotted in terms of AUC as a function of  $\kappa$  for the HK distances. Apart from similar behaviors already discussed in Sec. III A, we observe here that the peak tagging performance is roughly the same for all reference measures except uniref4, which does not attain tagging performance comparable to any EMD distance for any value of  $\kappa$ . Although uniref8 yields tagging performance comparable to  $\text{EMD}_{unN}$  for jets with total  $p_T \in [500, 550]$  GeV, it does not reach the tagging performance of  $\text{EMD}_N$ . In contrast, the tagging performance of PLUOT using uniref15, uniref30, and uniref60 meets or exceeds the tagging performance of the EMD distances for optimized values of  $\kappa$ . This suggests that the classification performance of the linear  $W_2$  and HK distance is rather robust to the choice of the reference for uniref15 and finer measures. Considering that the finest reference measure under consideration (uniref60) incurs a relatively high computational cost without significant improvement in tagging performance, in what follows we largely favor the default  $15 \times 15$  reference jet, reserving some comparisons with uniref30 for the Appendix.

<sup>2</sup>As we will see, the uniref4 and uniref8 reference measures are too coarse to capture the relevant structure of the jets for any value of  $\kappa$ , and the distribution of Euclidean norms for these measures are correspondingly omitted from Fig. 3.

Table I lists the  $\kappa$  value that produces the best AUC score for each task using  $\text{HK}_N$  and  $\text{HK}_{unN}$  metrics. Ignoring uniref4, the optimal value  $\kappa_{\text{best}}$  lies between 0.2 and 0.5 for all others, regardless of the interparticle spacing  $l$ . No obvious relationship is observed between  $l$  and  $\kappa_{\text{best}}$ .

### C. Discussion

The tagging performance of the HK-based metrics in Figs. 2 and 4 exhibits three distinct regimes as a function of  $\kappa$ . In the regime where mass creation or destruction dominates ( $\kappa \lesssim 0.1$ ), the AUC scores for both  $\text{HK}_N$  and  $\text{HK}_{unN}$  are comparable and decrease with decreasing  $\kappa$ . From Eq. (9), we know that no mass is allowed to be moved a distance more than  $\frac{\pi}{2}\kappa$ . When  $\kappa$  becomes so small such that  $\frac{\pi}{2}\kappa < l$  (where  $l$  is the interparticle spacing of the reference jet), mass transportation is largely forbidden when computing the distance between a jet and the reference measure. Furthermore, in this regime the assumption that, for each particle  $x'_j$  in the jet, there exists a particle  $x_i$  in the reference measure so that  $\|x_i - x'_j\| < \frac{\kappa}{2}$  is often violated, causing the linearization to break down; see Sec. II C. While this breakdown could be avoided in a continuum formulation of the linearization [27], one would still have to contend with the fact that, as  $\kappa \rightarrow 0$ , the rescaled Hellinger-Kantorovich metric converges to the Hellinger metric, in which all information on the spatial distribution of the jets is discarded and their distance is based purely on the difference between their energies at each location; see Eq. (7). We observe this breakdown at the level of the AUCs in Fig. 3, considering the value  $\kappa = 0.01$ . At the other end, at large  $\kappa$ , the tagging performance using  $\text{HK}_N$  stabilizes for  $\kappa \gtrsim 1$ , whereas the AUC score deteriorates significantly using  $\text{HK}_{unN}$ . As  $\kappa$  grows sufficiently large, it becomes increasingly expensive to create or destroy mass. Once we enter this transport-only regime,  $\kappa$  no longer plays any role for  $\text{HK}_N$ . On the other hand, whenever the total energies of the events are unequal,  $\text{HK}_{unN}$  diverges to  $+\infty$  as  $\kappa \rightarrow +\infty$ .

In between these two extremes,  $0.1 \lesssim \kappa \lesssim 1$ , the tagging performance of both  $\text{HK}_N$  and  $\text{HK}_{unN}$  is optimized, matching or exceeding the EMD approach. In this regime

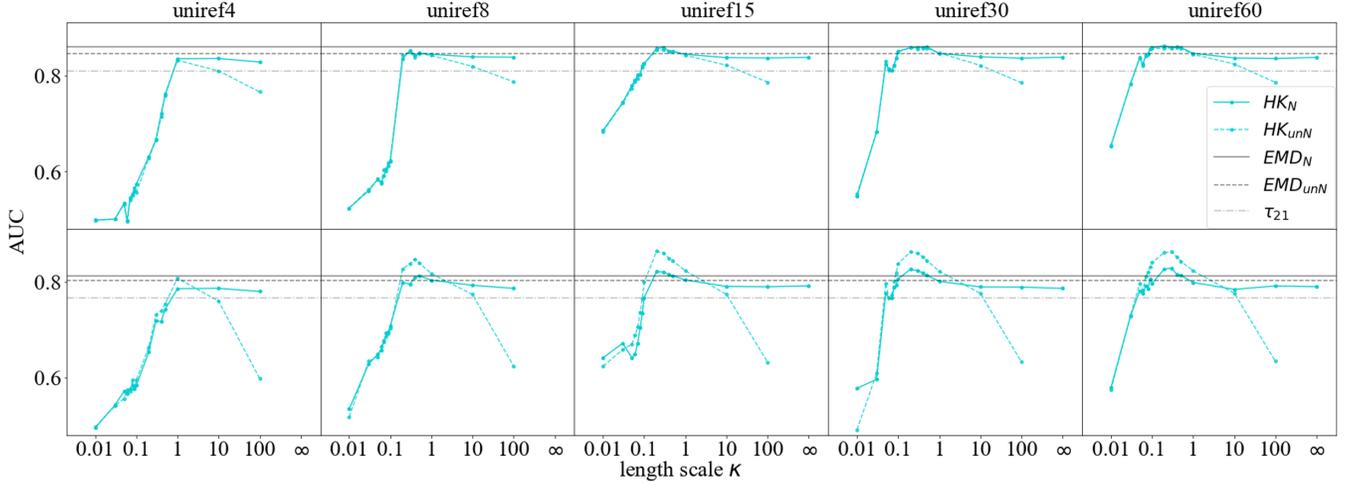


FIG. 4. AUC scores for classifying 10 000  $W$  vs QCD jets using different reference measures, with uniref4, uniref8, uniref15, uniref30, and uniref60 (from left to right). The machine learning model used here is kNN. Jet  $p_T$  is in between 500 and 550 (1500) GeV in the first (second) row. Solid (dashed) blue lines show the results calculated on normalized (unnormalized) jets for  $W_2$  and HK distance; horizontal gray solid (dashed) lines use the EMD metrics on normalized (unnormalized) jets; and gray dash-dotted lines give the tagging performance of  $\tau_{21}$ .

both mass transportation and creation or destruction are relevant. We have not observed any strong correlations between the optimal value  $\kappa_{\text{best}}$ , reference spacing  $l$ , the jet clustering radius  $R$ , and the typical angular separation of boosted partonic decay products  $\propto m/p_T$ , and no definite conclusion can be drawn at this stage regarding the dependence of  $\kappa_{\text{best}}$  on various jet length scales. We leave this question to future studies.

A major advantage of the PLUOT framework over EMD is that PLUOT significantly speeds up computation and requires much less storage space. On average, only about an hour is needed to calculate the LOT distance for 10 000 jets on a laptop, with further speedup in the  $\text{HK}^\kappa$  embedding for smaller  $\kappa$  values. In contrast, computing EMD for 10 000 jets takes approximately 15 h for jets drawn from a

50-GeV-wide  $p_T$  bin and 30–40 h for jets drawn from a 1-TeV-wide  $p_T$  bin.

To make matters worse, the EMD outputs a huge distance matrix, which for large datasets is impossible to store on a local computer or even on a moderate cluster. The matrix is then directly fed into downstream machine learning (ML) models, and, since there are fewer models that can directly handle the distance matrix rather than the coordinate of each jet, the EMD approach also limits the practical choices of ML models. Of course, the model training step is now more time efficient than PLUOT, since it now needs only to read in the pairwise distance from the matrix instead of computing the Euclidean distance on the fly. But the difference in time is not significant: For kNN, EMD takes seconds while PLUOT takes minutes; and for SVM, EMD needs minutes whereas PLUOT requires about 3 h. This gain in the latter ML step is not big enough to offset the huge time gap in OT computation.

TABLE I. Optimal  $\kappa$  values and their corresponding AUC scores for kNN classification of  $W$  vs QCD jets using different reference measures.

Jet $p_T$ (GeV)		[500, 550]		[500, 1500]	
Reference		$\text{HK}_N$	$\text{HK}_{unN}$	$\text{HK}_N$	$\text{HK}_{unN}$
uniref4 ( $4 \times 4$ )	$\kappa_{\text{best}}$	10	1	10	1
	AUC	0.835	0.832	0.786	0.807
uniref8	$\kappa_{\text{best}}$	0.3	0.3	0.5	0.4
	AUC	0.852	0.849	0.813	0.847
uniref15	$\kappa_{\text{best}}$	0.3	0.3	0.2	0.2
	AUC	0.859	0.854	0.821	0.863
uniref30	$\kappa_{\text{best}}$	0.5	0.2	0.2	0.2
	AUC	0.860	0.859	0.826	0.862
uniref60	$\kappa_{\text{best}}$	0.2	0.4	0.3	0.3
	AUC	0.862	0.858	0.828	0.863

#### IV. PILEUP ROBUSTNESS OF OPTIMAL TRANSPORT-BASED CLASSIFICATION

Pileup contamination reduces the efficacy of many commonly used jet physics observables [43–45] such as jet mass and dijet mass, where in Ref. [46] the impact of different levels of pileup on dijet mass is studied. This, in turn, motivates the invention of various pileup mitigation techniques [47–52]. Pileup mitigation has recently been recast in the language of optimal transport [2], but the robustness of OT-based approaches to jet classification has yet to be studied. Here, we carry out a preliminary study of the effects of pileup on  $W$  and QCD jet classification.

Again, we use the same  $W$  and QCD dataset with  $p_T \in [500, 550]$  GeV and jet radius  $R = 1$  as described in

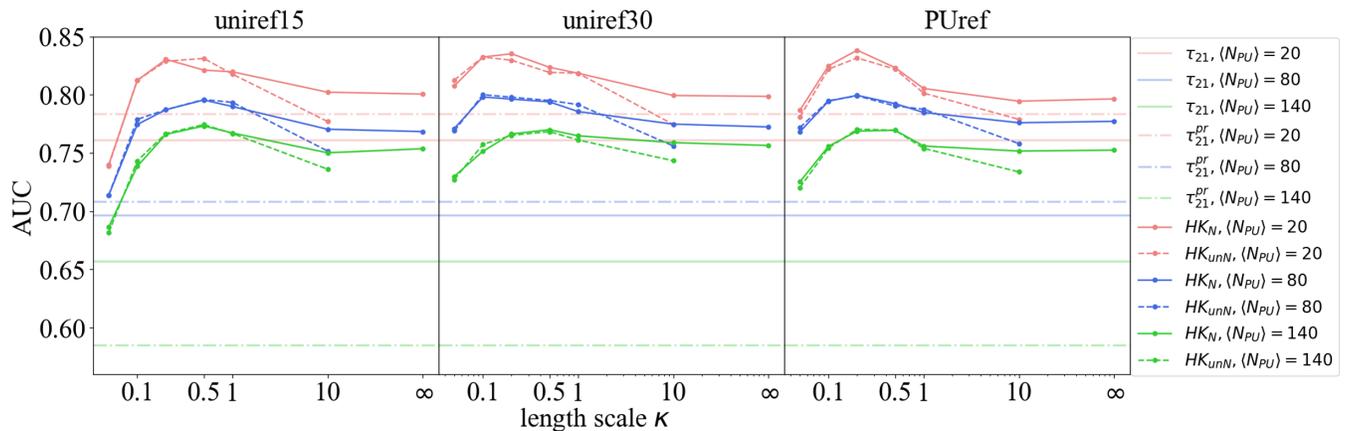


FIG. 5. AUC scores for using kNN to classify 10 000 W vs QCD jets with different amounts of pileup where the average numbers of pileup particles in each event are  $\langle N_{PU} \rangle = 20$  (red), 80 (blue), and 140 (green). From left to right, the reference measures used are the  $15 \times 15$  uniform reference, the  $30 \times 30$  uniform reference, and a jet drawn from the pileup template corresponding to each  $N_{PU}$ . As usual, solid (dashed) lines show the AUC scores using the linear HK and  $W_2$  distances on normalized (unnormalized) jets, and solid horizontal lines give the tagging performance of  $\tau_{21}$  on unpruned jets, whereas the dash-dotted lines are the results using  $\tau_{21}$  on pruned jets (denoted by  $\tau_{21}^{pr}$ ).

Sec. III. Background contamination is generated in PYTHIA where the actual number of pileup events per bunch crossing follows a Poisson distribution around  $\langle N_{PU} \rangle$ . We consider three different pileup benchmarks with  $\langle N_{PU} \rangle = 20, 80, 140$ . These pileup templates are then added to each event and FastJet is used to group the pileup-contaminated events into jets. We then follow the same procedure as before, applying the PLUOT framework to the pileup-contaminated jets.

Here, three reference measures are included: the default  $15 \times 15$  uniform reference; the  $30 \times 30$  uniform reference; and a “pileup” reference jet picked from one of the pileup templates for each value of  $\langle N_{PU} \rangle$ . For example, when examining jets contaminated by pileup with  $\langle N_{PU} \rangle = 80$ , the reference measure is taken to be another Poisson distribution with  $\langle N_{PU} \rangle = 80$ . The motivation behind the choice of uniref30 is that, since the number of particles in the reference is close to that of the jets contaminated by pileup with  $\langle N_{PU} \rangle = 80, 140$ , uniref30 should better capture the true underlying differences between W vs QCD jets not obscured by the superficial pileup addition.

Again, the N-subjettiness ratio  $\tau_{21}$  serves as a benchmark, where  $\tau_{21}$  is computed on the datasets both with and without pileup. The one without pileup is generated by pruning the contaminated datasets, accomplished in FastJet by a pruner that reclusters the jets with a Cambridge-Aachen algorithm and removes constituent particles that are soft or at large angles with other particles [53,54]. Figure 5 displays the resulting AUC vs  $\kappa$  curves, where we use kNN + linear  $W_2$  and HK distances with  $\kappa = +\infty, 10, 1, 0.5, 0.2, 0.1$ , and 0.05 on both normalized and unnormalized jets.

It is clear from the figure that, comparing to  $\tau_{21}$  (horizontal lines), the tagging performance of PLUOT

behaves rather well and does not decay significantly as pileup increases. Especially for high pileup scenarios, the AUC scores of kNN +  $W_2$  and HK distances on unpruned jet samples using any of the three references are far better than the corresponding AUCs of kNN +  $\tau_{21}$ , where, for  $\langle N_{PU} \rangle = 140$ ,  $\tau_{21}$  on pruned jets behaves much worse than that on unpruned jets, corroborating the observation in Ref. [55] that N-subjettiness on groomed jets is less discriminant than being computed on ungroomed jets. More studies need to be performed in order to examine in detail the influence of background contamination such as pileup on OT-based metrics, but its potential advantage over traditional methods is already clear.

## V. CONCLUSION

The energy mover’s distance [1] illustrates the many advantages to be gained by equipping the space of collider events with a metric, from unifying the panoply of collider observables to enabling the use of interpretable distance-based machine learning algorithms. These successes invite further exploration of the larger space of optimal transport distances with an eye toward collider physics. In this paper, we have generalized the EMD by situating it within a multiparameter family of unbalanced optimal transport metrics. Among the other metrics in this family, the Hellinger-Kantorovich distance stands out as particularly suited to collider applications insofar as it (uniquely) preserves a Riemannian structure. The resulting manifold of collider events inherits various satisfying properties, including a well-defined tangent space at each point on the manifold. This enables the computationally efficient linearization of unbalanced optimal transport distances [27] in close analogy with linearized optimal transport in the balanced case [3,28].

Motivated by these advantages, we developed the particle linearized unbalanced optimal transport framework for computing linearized Hellinger-Kantorovich distances. To illustrate the efficacy of PLUOT for collider physics, we used it as input to simple distance-based machine learning algorithms for boosted  $W$  and QCD jet classification and studied the performance as a function of various scale parameters. For optimized parameter choices, we found that PLUOT matched or exceeded the same algorithms using EMD distances in a fraction of the computing time. Although the effects of pileup on optimal transport distances have yet to be extensively studied, we found that boosted jet classification based on PLUOT also exhibited an encouraging degree of robustness against pileup contamination compared to the  $N$ -subjettiness shape observable.

There are numerous avenues for further exploration. Within the PLUOT framework itself, there is considerable room to explore the interplay between the Hellinger-Kantorovich length scale parameter, the jet clustering radius, and the scale(s) associated with the choice of reference measure. We have focused on boosted jet classification as an initial application to collider physics, but PLUOT should be generally well suited to the same array of applications as the energy mover's distance. More broadly, linearization is but one of the many potential advantages of applying the Hellinger-Kantorovich distance to collider physics. The Riemannian event manifold obtained with the Hellinger-Kantorovich distance is likely

to have interesting properties and may reveal further hidden structure in the space of collider events.

## ACKNOWLEDGMENTS

We thank B. Schmitzer and M. Thorpe for many helpful conversations and for sharing their code to compute the Hellinger-Kantorovich distance. The work of T. C. and N. C. is supported by the U.S. Department of Energy under Grant No. DE-SC0011702. N. C. thanks Lawrence Berkeley National Laboratory (LBNL) and the Berkeley Center for Theoretical Physics (BCTP) for hospitality during the completion of this work. T. C. and J. C. are grateful for a Worster Fellowship from the UCSB Department of Physics. The work of K. C. is supported by National Science Foundation (NSF) Division of Mathematical Sciences (DMS) Grant No. 1811012 and a Hellman Faculty Fellowship. K. C. gratefully acknowledges the support from the Simons Center for Theory of Computing, at which this work was completed.

## APPENDIX: RESULT TABLES

Here, we include two tables with numerical results for the classification tasks mentioned in Secs. III and IV. Both tables consist of AUC scores for  $W$  vs QCD jet classification using the kNN model. In Table II, the dataset has default jet radius  $R = 1$ , and the reference measures used are  $15 \times 15$  uniform measure (uniref15) and  $30 \times 30$  uniform measure (uniref30). Jet  $p_T$  ranges are [500, 550] GeV

TABLE II. AUC scores for kNN classification of  $W$  vs QCD jets.

Ref jet	$\kappa$	Normalization	$p_T$		
			[500, 550] GeV, 10 000 jets	[500, 1500] GeV, 10 000 jets	[500, 1500] GeV, 200 000 jets
uniref15	$+\infty$	N	0.838	0.791	0.817
	100	N	0.836	0.789	0.817
	10	unN	0.786	0.632	0.790
		N	0.837	0.790	0.817
	1	unN	0.821	0.774	0.827
		N	0.844	0.803	0.825
	0.5	unN	0.842	0.823	0.861
		N	0.850	0.812	0.836
	0.2	unN	0.850	0.843	0.874
		N	0.856	0.821	0.842
	0.1	unN	0.853	0.863	0.879
		N	0.825	0.767	0.791
	0.05	unN	0.825	0.799	0.827
		N	0.779	0.642	0.683
	0.01	unN	0.773	0.669	0.708
		N	0.685	0.641	0.669
		unN	0.683	0.624	0.644

(Table continued)

TABLE II. (*Continued*)

Ref jet	$\kappa$	Normalization	$p_T$		
			[500, 550] GeV, 10 000 jets	[500, 1500] GeV, 10 000 jets	[500, 1500] GeV, 200 000 jets
uniref30	$+\infty$	N	0.838	0.786	0.815
	100	N	0.836	0.789	0.815
		unN	0.785	0.633	0.791
	10	N	0.839	0.789	0.815
		unN	0.821	0.776	0.827
	1	N	0.846	0.801	0.827
		unN	0.847	0.822	0.860
	0.5	N	0.860	0.813	0.839
		unN	0.856	0.844	0.874
	0.2	N	0.857	0.826	0.842
		unN	0.859	0.862	0.880
	0.1	N	0.851	0.806	0.827
		unN	0.849	0.837	0.861
	0.05	N	0.823	0.775	0.802
		unN	0.830	0.797	0.825
	0.01	N	0.549	0.577	0.566
unN		0.552	0.492	0.567	
EMD	N	0.859	0.812	...	
	unN	0.846	0.802		
	$\tau_{21}$		0.810	0.766	0.765

TABLE III. AUC scores for kNN classification of  $W$  vs QCD jets with different levels of pileup.

Ref jet	$\kappa$	Normalization	$\langle N_{PU} \rangle = 20$	$\langle N_{PU} \rangle = 80$	$\langle N_{PU} \rangle = 140$	
uniref15	$+\infty$	N	0.801	0.768	0.754	
	10	N	0.802	0.770	0.750	
		unN	0.777	0.752	0.736	
	1	N	0.820	0.790	0.767	
		unN	0.818	0.794	0.767	
	0.5	N	0.821	0.796	0.773	
		unN	0.831	0.796	0.774	
	0.2	N	0.830	0.787	0.766	
		unN	0.829	0.787	0.767	
	0.1	N	0.812	0.775	0.739	
		unN	0.812	0.779	0.743	
	0.05	N	0.740	0.714	0.686	
		unN	0.739	0.714	0.682	
	uniref30	$+\infty$	N	0.799	0.772	0.757
		10	N	0.799	0.775	0.759
			unN	0.775	0.756	0.743
1		N	0.819	0.786	0.765	
		unN	0.819	0.792	0.761	
0.5		N	0.824	0.794	0.770	
		unN	0.819	0.795	0.768	
0.2		N	0.835	0.796	0.766	
		unN	0.830	0.798	0.765	
0.1		N	0.832	0.798	0.752	
		unN	0.832	0.800	0.757	
0.05		N	0.808	0.771	0.730	
		unN	0.813	0.769	0.727	

*(Table continued)*

TABLE III. (Continued)

Ref jet	$\kappa$	Normalization	$\langle N_{PU} \rangle = 20$	$\langle N_{PU} \rangle = 80$	$\langle N_{PU} \rangle = 140$	
PUref	$+\infty$	N	0.797	0.777	0.753	
	10	N	0.795	0.776	0.752	
		unN	0.779	0.758	0.734	
	1	N	0.805	0.785	0.756	
		unN	0.801	0.788	0.754	
	0.5	N	0.823	0.792	0.770	
		unN	0.822	0.790	0.770	
	0.2	N	0.838	0.800	0.769	
		unN	0.832	0.799	0.770	
	0.1	N	0.825	0.795	0.756	
		unN	0.822	0.795	0.754	
	0.05	N	0.787	0.768	0.725	
		unN	0.781	0.772	0.720	
		$\tau_{21}$		0.761	0.697	0.657
		Pruned $\tau_{21}$		0.784	0.708	0.585

with 10 000 jets, [500, 1500] GeV with 10 000 jets, and [500, 1500] GeV with 200 000 jets. Also included are tagging results using EMD metrics and  $\tau_{21}$ .

In Table III, the pileup contaminated dataset has the default jet radius  $R = 1$ , and their  $p_T$  range is [500,

550] GeV, with pileup levels being  $\langle N_{PU} \rangle = 20, 80, 140$ . The reference measures used are uniref15, uniref30, and a jet drawn from the pileup template corresponding to each  $N_{PU}$ . The tagging results using  $\tau_{21}$  on jets both with and without pruning are included as well.

- 
- [1] P. T. Komiske, E. M. Metodiev, and J. Thaler, 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, *J. High Energy Phys.* **07** (2020) 006.
- [3] T. Cai, J. Cheng, N. Craig, and K. Craig, Linearized optimal transport for collider events, *Phys. Rev. D* **102**, 116019 (2020).
- [4] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, Exploring the space of jets with CMS open data, *Phys. Rev. D* **101**, 034009 (2020).
- [5] C. Cesarotti and J. Thaler, A robust measure of event isotropy at colliders, *J. High Energy Phys.* **08** (2020) 084.
- [6] C. Cesarotti, M. Reece, and M. J. Strassler, The efficacy of event isotropy as an event shape observable, *J. High Energy Phys.* **07** (2021) 215.
- [7] A. David and G. Passarino, Use and reuse of SMEFT, [arXiv:2009.00127](https://arxiv.org/abs/2009.00127).
- [8] G. Stein, U. Seljak, and B. Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, in *34th Conference on Neural Information Processing Systems* (2020).
- [9] G. Kasieczka *et al.*, The LHC olympics 2020: A community challenge for anomaly detection in high energy physics, *Rep. Prog. Phys.* **84**, 124201 (2021).
- [10] J. N. Howard, S. Mandt, D. Whiteson, and Y. Yang, Foundations of a fast, data-driven, machine-learned simulator, [arXiv:2101.08944](https://arxiv.org/abs/2101.08944).
- [11] G. Di Guglielmo *et al.*, A reconfigurable neural network ASIC for detector front-end data compression at the HL-LHC, *IEEE Trans. Nucl. Sci.* **68**, 2179 (2021).
- [12] R. Kansal, J. Duarte, H. Su, B. Orzari, T. Tomei, M. Pierini, M. Touranakou, J.-R. Vlimant, and D. Gunopulos, Particle cloud generation with message passing generative adversarial networks, [arXiv:2106.11535](https://arxiv.org/abs/2106.11535).
- [13] J. H. Collins, An exploration of learnt representations of W jets, [arXiv:2109.10919](https://arxiv.org/abs/2109.10919).
- [14] K. Fraser, S. Homiller, R. K. Mishra, B. Ostdiek, and M. D. Schwartz, Challenges for unsupervised anomaly detection in particle physics, [arXiv:2110.06948](https://arxiv.org/abs/2110.06948).
- [15] J. Lott, Some geometric calculations on Wasserstein space, [arXiv:math/0612562](https://arxiv.org/abs/math/0612562).
- [16] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Springer Science and Business Media, New York, 2008).
- [17] T. T. Georgiou, J. Karlsson, and M. S. Takyar, Metrics for power spectra: An axiomatic approach, *IEEE Trans. Signal Process.* **57**, 859 (2008).
- [18] L. Caffarelli and R. J. McCann, Free boundaries in optimal transport and monge-ampere obstacle problems, *Ann. Math.* **171**, 673 (2010).

- [19] A. Figalli, The optimal partial transport problem, *Arch. Ration. Mech. Anal.* **195**, 533 (2010).
- [20] B. Piccoli and F. Rossi, Generalized Wasserstein distance and its application to transport equations with source, *Arch. Ration. Mech. Anal.* **211**, 335 (2014).
- [21] B. Piccoli and F. Rossi, On properties of the generalized Wasserstein distance, *Arch. Ration. Mech. Anal.* **222**, 1339 (2016).
- [22] M. Liero, A. Mielke, and G. Savaré, Optimal transport in competition with reaction: The hellinger–kantorovich distance and geodesic curves, *SIAM J. Math. Anal.* **48**, 2869 (2016).
- [23] M. Liero, A. Mielke, and G. Savaré, Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures, *Inventiones Mathematicae* **211**, 969 (2018).
- [24] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, An interpolating distance between optimal transport and fisher–rao metrics, *Found. Comput. Math.* **18**, 1 (2018).
- [25] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, Scaling algorithms for unbalanced optimal transport problems, *Math. Comput.* **87**, 2563 (2018).
- [26] S. Kondratyev, L. Monsaingeon, D. Vorotnikov *et al.*, A new optimal transport distance on the space of finite radon measures, *Adv. Differ. Equat.* **21**, 1117 (2016).
- [27] T. Cai, J. Cheng, B. Schmitzer, and M. Thorpe, The linearized Hellinger–Kantorovich distance, [arXiv:2102.08807](https://arxiv.org/abs/2102.08807).
- [28] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, A linear optimal transportation framework for quantifying and visualizing variations in sets of images, *Int. J. Comput. Vis.* **101**, 254 (2013).
- [29] D. P. Bourne, B. Schmitzer, and B. Wirth, Semi-discrete unbalanced optimal transport and quantization, [arXiv:1808.01962](https://arxiv.org/abs/1808.01962).
- [30] Q. Mérigot, A multiscale approach to optimal transport, in *Computer Graphics Forum* (Wiley Online Library, New York, 2011), Vol. 30, pp. 1583–1592.
- [31] A. Delalande and Q. Merigot, Quantitative stability of optimal transport maps under variations of the target measure, [arXiv:2103.05934](https://arxiv.org/abs/2103.05934).
- [32] W. Gangbo, W. Li, S. Osher, and M. Puthawala, Unnormalized optimal transport, *J. Comput. Phys.* **399**, 108940 (2019).
- [33] E. K. Ryu, W. Li, P. Yin, and S. Osher, Unbalanced and partial  $L_1$  /  $L_1$  monge–kantorovich problem: A scalable parallel first-order method, *J. Sci. Comput.* **75**, 1596 (2018).
- [34] V. Laschos and A. Mielke, Geometric properties of cones with applications on the hellinger–kantorovich space, and a new distance on the space of probability measures, *J. Funct. Anal.* **276**, 3529 (2019).
- [35] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, *J. High Energy Phys.* **07** (2014) 079.
- [36] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to pythia8.2, *Comput. Phys. Commun.* **191**, 159177 (2015).
- [37] M. Cacciari, G. P. Salam, and G. Soyez, Fastjet user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [38] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, *J. High Energy Phys.* **04** (2018) 013.
- [39] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, *J. High Energy Phys.* **01** (2019) 121.
- [40] P. T. Komiske, E. M. Metodiev, and J. Thaler, Cutting multiparticle correlators down to size, *Phys. Rev. D* **101**, 036019 (2020).
- [41] J. Thaler and K. Van Tilburg, Identifying boosted objects with n-subjettiness, *J. High Energy Phys.* **03** (2011) 015.
- [42] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing n-subjettiness, *J. High Energy Phys.* **02** (2012) 093.
- [43] Pileup jet identification, No. CMS-PAS-JME-13-005, 2013.
- [44] Performance of the ATLAS inner detector track and vertex rReconstruction in the high pile-up LHC environment, No. ATLAS-CONF-2012-042, 2012.
- [45] G. Aad *et al.* (ATLAS Collaboration), Performance of pile-up mitigation techniques for jets in  $pp$  collisions at  $\sqrt{s} = 8$  TeV using the ATLAS detector, *Eur. Phys. J. C* **76**, 581 (2016).
- [46] G. P. Salam, Towards jetography, *Eur. Phys. J. C* **67**, 637 (2010).
- [47] G. Soyez, Pileup mitigation at the LHC: A theorist’s view, *Phys. Rep.* **803**, 1 (2019).
- [48] P. Berta, M. Spousta, D. W. Miller, and R. Leitner, Particle-level pileup subtraction for jets and jet shapes, *J. High Energy Phys.* **06** (2014) 092.
- [49] D. Krohn, M. D. Schwartz, M. Low, and L.-T. Wang, Jet cleansing: Pileup removal at high luminosity, *Phys. Rev. D* **90**, 065020 (2014).
- [50] D. Bertolini, P. Harris, M. Low, and N. Tran, Pileup per particle identification, *J. High Energy Phys.* **10** (2014) 059.
- [51] M. Cacciari, G. P. Salam, and G. Soyez, SoftKiller, a particle-level pileup removal method, *Eur. Phys. J. C* **75**, 59 (2015).
- [52] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, Pileup mitigation with machine learning (PUMML), *J. High Energy Phys.* **12** (2017) 051.
- [53] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Techniques for improved heavy particle searches with jet substructure, *Phys. Rev. D* **80**, 051501 (2009).
- [54] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches, *Phys. Rev. D* **81**, 094023 (2010).
- [55] Identifying hadronically decaying vector bosons merged into a single jet, No. CMS-PAS-JME-13-006, 2013.