Editors' Suggestion | Featured in Physics

Metric Space of Collider Events

Patrick T. Komiske,[*] Eric M. Metodiev,[†] and Jesse Thaler[‡]
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
and Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

(Received 15 February 2019; published 26 July 2019)

When are two collider events similar? Despite the simplicity and generality of this question, there is no established notion of the distance between two events. To address this question, we develop a metric for the space of collider events based on the earth mover's distance: the "work" required to rearrange the radiation pattern of one event into another. We expose interesting connections between this metric and the structure of infrared- and collinear-safe observables, providing a novel technique to quantify event modifications due to hadronization, pileup, and detector effects. We showcase how this metrization unlocks powerful new tools for analyzing and visualizing collider data without relying upon a choice of observables. More broadly, this framework paves the way for data-driven collider phenomenology without specialized observables or machine learning models.

DOI: 10.1103/PhysRevLett.123.041801

High-energy particle collisions produce a tremendous number of intricately correlated particles, especially when energetic quarks and gluons are involved. Behind this apparent complexity, however, the overall flow of energy in an event is a robust memory of its simpler partonic origins [1–8]. Surprisingly, no definition of the similarity between events presently exists that sharply captures this correspondence. In the absence of a metric, efforts typically fall back upon *ad hoc* methods such as comparing specific observables [9–13] or matching the pixels of calorimeter images [13–17]. These approaches suffer from significant pathologies: disparate event topologies can give rise to identical observable values, while pixels lack stability under small perturbations. A theoretically and experimentally robust definition of the "distance" between events would profoundly expand our ability to explore the structure of collider data and unlock entirely new ways to probe events.

In this Letter, we advocate for the earth (or energy) mover's distance (EMD) [18–22] as a metric for the space of collider events. We propose a variant of the EMD, inspired by Refs. [21,22], that allows events with different total energies to be sensibly compared. The EMD is the minimum "work" required to rearrange one event $\mathcal{E}$ into the other $\mathcal{E}'$ by movements of energy f_{ij} from particle i in one event to particle j in the other:

$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\geq 0\}} \sum_{ij} f_{ij}\frac{\theta_{ij}}{R} + \left|\sum_i E_i - \sum_j E'_j\right|,$$
$$\sum_j f_{ij} \leq E_i, \qquad \sum_i f_{ij} \leq E'_j, \qquad \sum_{ij} f_{ij} = E_{\mathrm{min}}, \tag{1}$$

where i and j index particles in events $\mathcal{E}$ and $\mathcal{E}'$, respectively, E_i is the particle energy, θ_{ij} is an angular distance

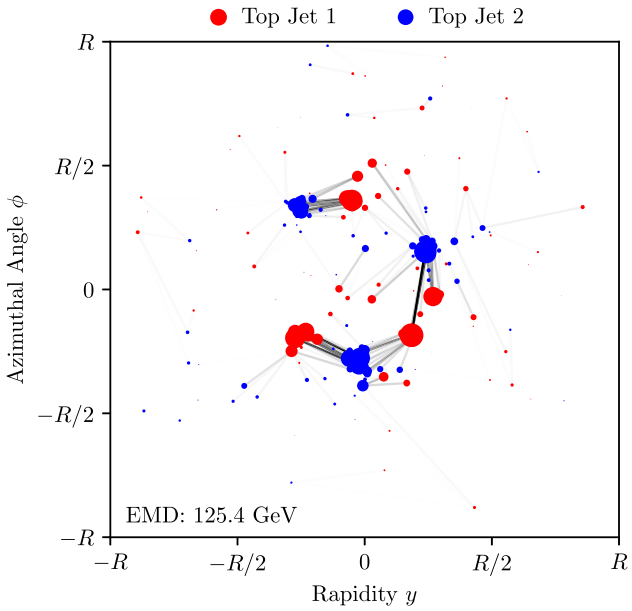

FIG. 1. The optimal movement to rearrange one top jet (red) into another (blue). Particles are shown as points in the rapidity-azimuth plane with areas proportional to their transverse momenta. Darker lines indicate more transverse momentum movement. The energy mover's distance in Eq. (1) is the total "work" required to perform this rearrangement.

between particles, and $E_{\text{min}} = \min(\sum_i E_i, \sum_j E'_j)$ is the smaller of the two total energies. R is a parameter that controls the relative importance of the two terms. While energies and angles are used here for clarity, we will use transverse momenta p_T and rapidity-azimuth (y, ϕ) distances for our applications relevant for the Large Hadron Collider (LHC).

The EMD that we propose in Eq. (1) has dimensions of energy, where the first term quantifies the difference between the two radiation patterns and the second term accounts for the creation or destruction of energy. It is a true metric (satisfying the triangle inequality) as long as θ_{ij} is a metric and $R \geq \frac{1}{2}\theta_{\text{max}}$, where θ_{max} is the maximum attainable angular distance between particles. For instance, R must be at least the jet radius for conical jets. Formally, the EMD metrizes the energy flow as it treats events differing only by soft particles or collinear splittings identically. This hints at a deep connection to infrared and collinear (IRC) safety of observables [23–26], which we explore further below.

A metric for comparing events is particularly relevant for probing the substructure of jets [27–37], collimated sprays of particles resulting from the fragmentation and hadronization of high-energy quarks, and gluons via quantum chromodynamics (QCD). Here, we will consider three classes of jets that have different intrinsic topologies: three-pronged boosted top quark jets, two-pronged boosted W boson jets, and single-pronged QCD (quark or gluon) jets. We generate proton-proton collision events at the LHC with PYTHIA 8.235 [38] at $\sqrt{s} = 14$ TeV including hadronization and multiple particle interactions. Anti-k_T jets [39] with a jet radius of 1.0 are clustered using FASTJET 3.3.1 [40], and up to two jets with $p_T \in [500, 550]$ GeV and $|y| < 1.7$ are kept. This p_T selection is representative of an intermediate energy range for jets at the LHC and allows for sensitivity to the effects of both terms in Eq. (1). Jets are longitudinally boosted and rotated to center the jet four-momentum at $(y, \phi) = 0$ as well as to vertically align the principal component of the constituent transverse momentum flow in the rapidity-azimuth plane; this removes the dependence of the EMD on these jet isometries.

We record the final-state hadrons, as well as the partons (before hadronization) and the hard W/top decay products, that are within a jet radius of the jet four momentum. We use the Python Optimal Transport [41] library to compute EMDs with the minimal choice of $R = 1.0$, the jet radius. The energy difference penalty in Eq. (1) is implemented using a fictitious particle at a distance R from all other particles. Figure 1 shows the optimal energy movement between two example top jets.

We begin by highlighting a remarkable mathematical property of the EMD, which provides a quantitative understanding of an observable's sensitivity to the radiation pattern. Specifically, we relate the EMD to additive IRC-safe observables via the Kantorovich-Rubinstein [42]

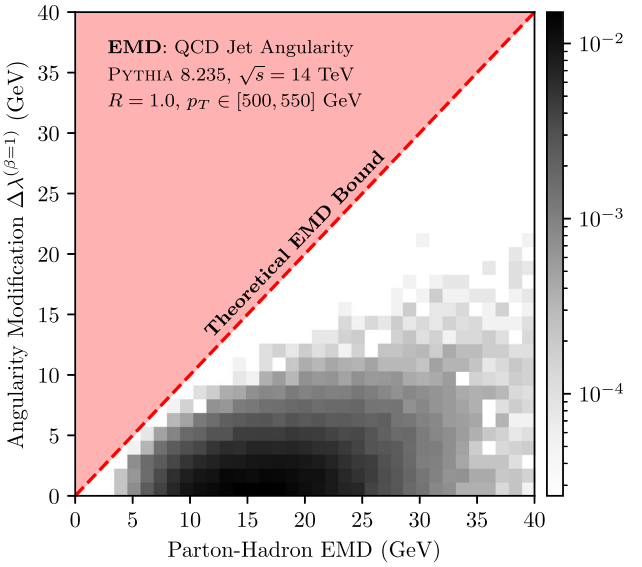

FIG. 2. Two-dimensional histogram of the EMD between 30 000 QCD jets before and after hadronization versus the corresponding $\beta = 1$ angularity modification. The red region is excluded based on the bound in Eq. (3), shown as a dashed red line. The bound is clearly satisfied and is nearly saturated for EMD $\lesssim 10$ GeV.

duality theorem. Applying this theorem to our variant of the EMD, we derive the following mathematical bound between two events $\mathcal{E}$ and $\mathcal{E}'$:

$$\frac{1}{RL}\left|\sum_i E_i \Phi(\hat{p}_i) - \sum_j E'_j \Phi(\hat{p}_j{}')\right| \leq \text{EMD}(\mathcal{E}, \mathcal{E}'), \quad (2)$$

where i, j index $\mathcal{E}, \mathcal{E}'$, respectively, $\hat{p}_i$ is the particle angular position, and Φ is any L-Lipschitz function (essentially, with gradient size bounded by L) which vanishes at the center of the space (e.g., the jet axis). The implications of Eq. (2) are simple yet profound: the similarity of events according to the EMD metric guarantees the closeness of their $\mathcal{O} = \sum_{i=1}^{M} E_i \Phi(\hat{p}_i)$ observable values in a precise way that depends on Φ. By formulating IRC-safe observables in the language of additive energy-weighted structures [43,44], Eq. (2) can be applied to provide a robust bound.

As a concrete example, we demonstrate how the EMD bounds hadronization modifications of jet angularities [45] (see also Refs. [46–49]), $\lambda^{(\beta)} = \sum_i p_{T,i} \theta_i^{\beta}$ where θ_i is the rapidity-azimuth distance to the jet axis. These angularities are evidently of the form in Eq. (2) with $\Phi(y_i, \phi_i) = (y_i^2 + \phi_i^2)^{\beta/2}$, which for $\beta \geq 1$ is a β-Lipschitz function over our $R = 1.0$ jet cone; hence:

$$\Delta\lambda^{(\beta)} = |\lambda^{(\beta)}(\mathcal{E}) - \lambda^{(\beta)}(\mathcal{E}')| \leq \beta\ \text{EMD}(\mathcal{E}, \mathcal{E}'). \quad (3)$$

The EMD between two events yields a robust upper bound of the difference in their $\beta \geq 1$ angularity values. This bound is borne out in Fig. 2, where the angularity

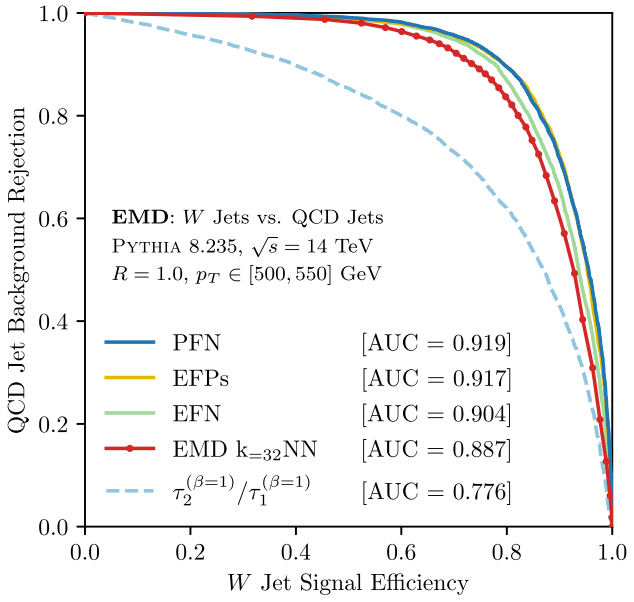

FIG. 3. ROC curves showing the boosted W classification performance of a $k = 32$ nearest-neighbor EMD classifier, which requires no choice of observables or parametrized machine learning architectures. The EMD classifier is competitive with machine learning techniques known to be good multiprong classifiers, such as PFNs, EFNs, and EFPs.

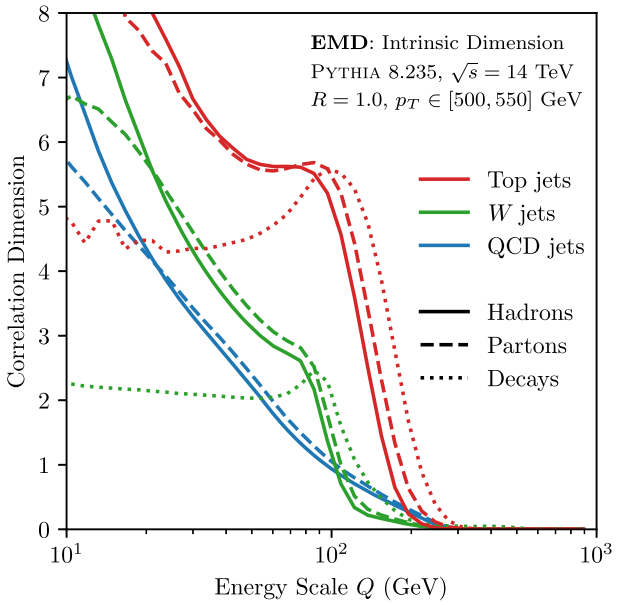

FIG. 4. The correlation dimension of top, W, and QCD jets as a function of the energy scale Q using hadrons (solid), partons (dashed), and hard decay products (dotted). Generally, QCD jets are the lowest dimensional and top jets are the highest dimensional. By comparing partons and hadrons, one sees that hadronization affects the structure of the space at scales below about 30 GeV. Similarly, the hard decay structure of top and W jets governs their dimension at high scales. Below about 10 GeV, the data become very high dimensional and sparse, making dimension estimation difficult.

differences and EMDs are computed for the same QCD jets before and after hadronization. For this jet p_T range, hadronization modifies events by EMD $\lesssim 30$ GeV and correspondingly modifies $\lambda^{(\beta=1)}$ by no more than this amount. The intuitive picture of parton-hadron duality [5], that the energy flow in an event is robust to nonperturbative effects, is quantified by considering the EMD that these nonperturbative effects can induce.

A metric space is also useful for classification without requiring specially designed observables or parametrized machine learning algorithms. One of the simplest examples of a nonparametric classifier is the k-nearest-neighbor (kNN) algorithm [50], whereby a given event's closest k neighbors in a reference set are used to determine class membership. We build a kNN classifier applied to the problem of discriminating W jets from QCD jets using a balanced training sample of 100 000 total jets. The classifier output is the number of W jets among the $k = 32$ nearest neighbors by EMD. This method should approach the optimal IRC-safe classifier with a sufficiently large dataset. The performance of the resulting EMD kNN classifier is shown in Fig. 3 as a receiver operating characteristic (ROC) curve, with the area under the ROC curve (AUC) also shown. For comparison, we include an energy flow network (EFN) and a particle flow network (PFN) [44] as well as a linear classifier trained on energy flow polynomials (EFPs) [43]. All classifiers are trained on a 100 000 training sample and evaluated on a 20 000 test sample, with the neural networks using 20% of the training sample for validation and a batch size of 125 (see Ref. [44] for additional details). The kNN approaches the performance of these state-of-the-art classifiers and significantly outperforms a ratio $\tau_2^{(\beta=1)}/\tau_1^{(\beta=1)}$ of N-subjettiness observables [51,52] designed to identify two-prong substructure. It is expected that the performance of the kNN method would improve with more sophisticated kernel density estimation techniques.

It is worth noting that while searching through a large reference set of events to find neighbors naïvely requires every possible pairwise comparison, in a metric space the triangle inequality can provide a great deal of simplification. Specialized data structures known as metric trees [53–56] have been developed to achieve query times that are approximately logarithmic in the size of the dataset. While we use direct searches throughout this Letter, this is not a fundamental limitation and we leave metric tree query optimizations to future work.

Once a space has been equipped with a metric, it is natural to ask about the structure of the induced manifold. The most basic aspect of the manifold underlying the data is its dimension, and several notions of its intrinsic dimension exist [57]. The correlation dimension [58,59], a type of fractal dimension, is suitable for our purposes and is defined using only pairwise distances:

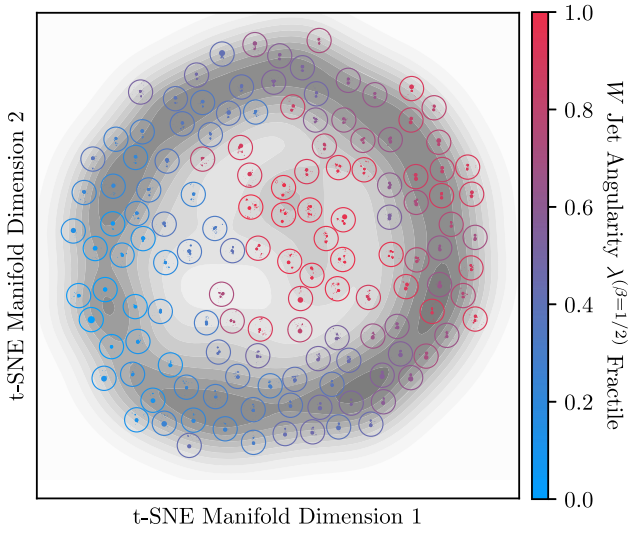

FIG. 5. The embedding of W jets into a two-dimensional space with t-SNE. The gray contours represent the density of embedded jets. Examples of W jets are shown throughout the space. The color of each jet corresponds to its angularity $\lambda^{(\beta=1/2)}$ fractile to quantify the energy sharing of the two prongs. An annulus emerges with jets in the lower (upper) region of the manifold having a more energetic lower (upper) subjet. More complex topologies with the largest angularity values populate the center of the manifold.

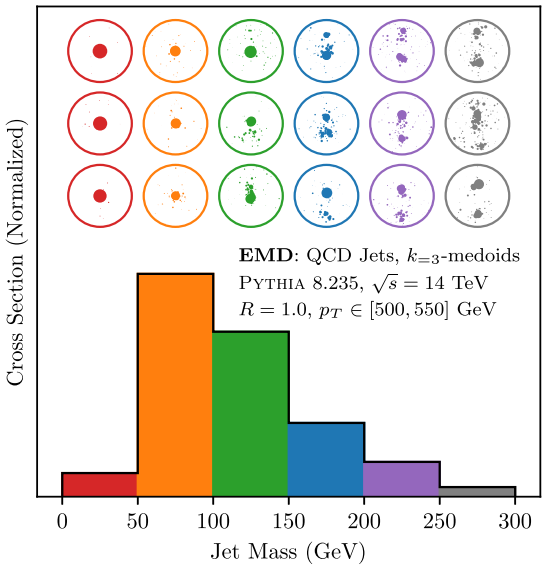

FIG. 6. The jet mass distribution for QCD jets, with $k = 3$ medoids shown above each bin. This visualization highlights that simple one-prong topologies dominate low jet masses and complex two-prong topologies exist at high jet masses.

$$\dim(Q) = Q\frac{\partial}{\partial Q}\ln \sum_{1\le k<\ell\le N} \Theta[\mathrm{EMD}(\mathcal{E}_k, \mathcal{E}_\ell) < Q], \quad (4)$$

where N is the total number of events and the summand indicates whether event k is within EMD Q of event ℓ.

The correlation dimension is an intrinsically scale-dependent quantity, which is particularly useful as we anticipate different physical effects to dominate jets at different scales. Shown in Fig. 4 is the intrinsic dimension of our top, W, and QCD samples over energy scales Q ranging from 10 to 1000 GeV obtained from Eq. (4) with 25 000 jets. At high energy scales Q, the EMD is governed by the hard decay kinematics, resulting in a relatively simple manifold with low intrinsic dimension. At energy scales Q approaching the fragmentation and hadronization scales, the structure of the events becomes increasingly complex and the dimension correspondingly increases. It is satisfying that the dimension is relatively low for a wide range of relevant energies, which is critical for a variety of metric-based techniques such as classification and low-dimensional visualization to work effectively with a realistic amount of data.

Beyond probing its dimension, the entire space of jets can be visualized using techniques such as t-distributed stochastic neighbor embedding (t-SNE) [60–63], which finds a low-dimensional embedding of the data that attempts to respect the distances between points. Figure 5 shows a t-SNE embedding of 5 000 W jets with $p_T \in [500, 510]$ GeV into a two-dimensional manifold using SCIKIT-LEARN [64]. The narrower p_T range focuses the EMD on the jet substructure and was found to yield sharper visualizations, with other choices also yielding sensible results. The W jets populate a circular subspace roughly corresponding to the energy sharing of the two prongs. As the W jet originates from a resonant decay, the two decay quarks (after rotation) are solely described by their energy sharing, which satisfyingly emerges from the manifold of W jets. Moreover, the center of the ring, distant from the annulus, tends to contain the most complex jet topologies, resulting in a type of automatic anomaly detection.

Finally, we illustrate the use of EMD for a new kind of visualization strategy that clusters events to better understand observable distributions. To describe a given set of events, such as those in a histogram bin, we find the k events (called medoids) which best describe the set in that the sum of distances of each event to its closest medoid is minimized. This procedure works for any observable and provides an immediate glimpse of the types of event topologies that correspond to a given observable value. We use an iterative approximation of k medoids from the PYCLUSTERING Python package [65]. As an illustration, Fig. 6 shows the jet mass for QCD jets with $k = 3$ medoids per bin, providing a snapshot of the different event topologies at different masses.

In conclusion, we have equipped the space of events with a metric, thereby allowing a powerful suite of new tools and techniques to be directly applied to collider physics. There are many potential applications of the EMD at colliders beyond those presented here. Pileup mitigation or detector reconstruction could use the EMD to

benchmark performance and thus benefit from the quantitative bounds on IRC-safe observable modifications. Further, machine learning models could be trained to optimize the EMD, related to recent efforts in generative modeling [66–69]. By counting neighbors, one could also perform density estimation in the space of events [70]. While we have focused on jet substructure, analogous studies could be carried out at the event level, which may require working with composite objects such as jets for realistic computation times. It would be interesting to explore an EMD strategy for unfolding by matching detector-level and simulated events. One might consider alternatives to the EMD, such as symmetry-projected metrics [22] or p-Wasserstein metrics [71,72] beyond our $p = 1$ case, though our conclusions should hold for any physically sensible metric. Further, using the EMD for model-independent anomaly detection [73–79] by finding isolated or clustered event topologies could empower searches for physics beyond the standard model at the LHC.

We would like to thank Felice Frankel, Marat Freytsis, Paul Ginsparg, Aram Harrow, Gregor Kasieczka, Andrew Larkoski, Katherine Liu, Benjamin Nachman, Miruna Oprescu, Katherine Quinn, and Jonathan Walsh for helpful discussions. We benefited from the hospitality of the Harvard Center for the Fundamental Laws of Nature, the Fermilab Distinguished Scholars program, and the Aspen Center for Physics. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under Grant No. DE-SC0011090 and the DOE Office of High Energy Physics under Grants No. DE-SC0012567 and No. DE-SC0019128. J. T. is supported by the Simons Foundation through a Simons Fellowship in Theoretical Physics. Cloud computing resources were provided through a Microsoft Azure for Research award and through a Google Cloud allotment from the MIT Quest for Intelligence.

*pkomiske@mit.edu
†metodiev@mit.edu
‡jthaler@mit.edu

[1] G. F. Sterman and S. Weinberg, Jets from Quantum Chromodynamics, Phys. Rev. Lett. **39**, 1436 (1977).
[2] H. Georgi and M. Machacek, A Simple QCD Prediction of Jet Structure in e^+e^- Annihilation, Phys. Rev. Lett. **39**, 1237 (1977).
[3] J. F. Donoghue, F. E. Low, and S.-Y. Pi, Tensor Analysis of Hadronic Jets in Quantum Chromodynamics, Phys. Rev. D **20**, 2759 (1979).
[4] G. Altarelli, Partons in Quantum Chromodynamics, Phys. Rep. **81**, 1 (1982).
[5] Y. L. Dokshitzer, V. A. Khoze, and S. I. Troian, On the concept of local parton hadron duality, J. Phys. G **17**, 1585 (1991).
[6] F. V. Tkachov, Measuring multi-jet structure of hadronic energy flow or What is a jet, Int. J. Mod. Phys. A **12**, 5411 (1997).
[7] N. A. Sveshnikov and F. V. Tkachov, Jets and quantum field theory, Phys. Lett. B **382**, 403 (1996).
[8] D. M. Hofman and J. Maldacena, Conformal collider physics: Energy and charge correlations, J. High Energy Phys. 05 (2008) 012.
[9] M. Cacciari and G. P. Salam, Pileup subtraction using jet areas, Phys. Lett. B **659**, 119 (2008).
[10] M. Cacciari, G. P. Salam, and G. Soyez, SoftKiller, a particle-level pileup removal method, Eur. Phys. J. C **75**, 59 (2015).
[11] D. Bertolini, P. Harris, M. Low, and N. Tran, Pileup Per Particle Identification, J. High Energy Phys. 10 (2014) 059.
[12] J. Arjona Martínez, O. Cerri, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Pileup mitigation at the large hadron collider with graph neural networks, arXiv:1810.07988.
[13] 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.
[14] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, Jet-images: Computer vision inspired techniques for jet tagging, J. High Energy Phys. 02 (2015) 118.
[15] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images—deep learning edition, J. High Energy Phys. 07 (2016) 069.
[16] M. Paganini, L. de Oliveira, and B. Nachman, Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters, Phys. Rev. Lett. **120**, 042003 (2018).
[17] M. Paganini, L. de Oliveira, and B. Nachman, CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks, Phys. Rev. D **97**, 014021 (2018).
[18] S. Peleg, M. Werman, and H. Rom, A unified approach to the change of resolution: Space and gray-level, IEEE Trans. Pattern Anal. Mach. Intell. **11**, 739 (1989).
[19] Y. Rubner, C. Tomasi, and L. J. Guibas, A metric for distributions with applications to image databases, in *Proceedings of the Sixth International Conference on Computer Vision, ICCV '98, Bombay, India, 1998* (IEEE Computer Society, Washington, DC, 1998), pp. 59–66.
[20] Y. Rubner, C. Tomasi, and L. J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. **40**, 99 (2000).
[21] O. Pele and M. Werman, A linear time histogram metric for improved SIFT matching, in *Computer Vision—ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, 2008, Proceedings, Part III* (Springer-Verlag, Berlin, Heidelberg, 2008), pp. 495–508.
[22] O. Pele and B. Taskar, The tangent earth mover's distance, in *Geometric Science of Information—First International Conference, GSI 2013, Paris, France, 2013. Proceedings* (Kluwer Academic Publishers, Norwell, MA, 2013), pp. 397–404.
[23] T. Kinoshita, Mass singularities of Feynman amplitudes, J. Math. Phys. (N.Y.) **3**, 650 (1962).
[24] T. D. Lee and M. Nauenberg, Degenerate systems and mass singularities, Phys. Rev. **133**, B1549 (1964).

[25] R. Brock *et al.* (CTEQ Collaboration), Handbook of perturbative QCD: Version 1.0, Rev. Mod. Phys. **67**, 157 (1995).
[26] S. Weinberg, *The Quantum Theory of Fields. Vol. 1: Foundations* (Cambridge University Press, Cambridge, England, 2005).
[27] M. H. Seymour, Tagging a heavy Higgs boson, in *ECFA Large Hadron Collider Workshop, Aachen, Germany, 1990: Proceedings. 2.* (CERN, Geneva, 1991), pp. 557–569.
[28] M. H. Seymour, Searches for new particles using cone and cluster jet algorithms: A Comparative study, Z. Phys. C **62**, 127 (1994).
[29] J. M. Butterworth, B. E. Cox, and Jeffrey R. Forshaw, *WW* scattering at the CERN LHC, Phys. Rev. D **65**, 096014 (2002).
[30] J. M. Butterworth, J. R. Ellis, and A. R. Raklev, Reconstructing sparticle mass spectra using hadronic decays, J. High Energy Phys. 05 (**2007**) 033.
[31] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Jet Substructure as a New Higgs Search Channel at the LHC, Phys. Rev. Lett. **100**, 242001 (2008).
[32] A. Abdesselam *et al.*, Boosted objects: A probe of beyond the Standard Model physics, Eur. Phys. J. C **71**, 1661 (2011).
[33] A. Altheimer *et al.*, Jet substructure at the tevatron and LHC: New results, new tools, new benchmarks, J. Phys. G **39**, 063001 (2012).
[34] A. Altheimer *et al.*, Boosted objects and jet substructure at the LHC, Eur. Phys. J. C **74**, 2792 (2014).
[35] D. Adams *et al.*, Towards an understanding of the correlations in jet substructure, Eur. Phys. J. C **75**, 409 (2015).
[36] A. J. Larkoski, I. Moult, and B. Nachman, Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning, arXiv:1709.04464.
[37] L. Asquith *et al.*, Jet substructure at the large hadron collider: Experimental review, arXiv:1803.06991.
[38] 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 PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).
[39] M. Cacciari, G. P. Salam, and G. Soyez, The Anti-k(t) jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[40] M. Cacciari, G. P. Salam, and G. Soyez, FastJet User Manual, Eur. Phys. J. C **72**, 1896 (2012).
[41] R. Flamary and N. Courty, Pot python optimal transport library, https://pot.readthedocs.io/en/stable/, 2017.
[42] L. V. Kantorovich and G. S. Rubinstein, On a space of completely additive functions, Vestn. Leningr. Univ. Fiz. Khim. **13**, 52 (1958).
[43] 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.
[44] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.
[45] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, Gaining (mutual) information about quark/gluon discrimination, J. High Energy Phys. 11 (**2014**) 129.
[46] C. F. Berger, T. Kucs, and G. F. Sterman, Event shape/energy flow correlations, Phys. Rev. D **68**, 014012 (2003).
[47] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, and J. Virzi, Substructure of high-p_T Jets at the LHC, Phys. Rev. D **79**, 074017 (2009).
[48] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, Jet shapes and jet algorithms in SCET, J. High Energy Phys. 11 (2010) 101.
[49] A. J. Larkoski, D. Neill, and J. Thaler, Jet shapes with the broadening axis, J. High Energy Phys. 04 (2014) 017.
[50] T. M. Cover and P. E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory **13**, 21 (1967).
[51] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.
[52] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.
[53] J. K. Uhlmann, Satisfying general proximity/similarity queries with metric trees, Inf. Proc. Lett. **40**, 175 (1991).
[54] P. N. Yianilos, Data structures and algorithms for nearest neighbor search in general metric spaces, in *Proceedings of the Fourth Annual ACM/SIGACT-SIAM Symposium on Discrete Algorithms, 1993, Austin, Texas, USA* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1993), pp. 311–321.
[55] S. Brin, Near neighbor search in large metric spaces, in *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, edited by U. Dayal, P. M. D. Gray, and S. Nishio (Morgan Kaufmann, San Francisco, CA, 1995), pp. 574–584.
[56] T. Bozkaya and Z. Meral Özsoyoglu, Indexing large metric spaces for similarity search queries, ACM Transactions on Database Systems **24**, 361 (1999).
[57] F. Camastra, Data dimensionality estimation methods: A survey, Pattern Recognit. **36**, 2945 (2003).
[58] P. Grassberger and I. Procaccia, Characterization of Strange Attractors, Phys. Rev. Lett. **50**, 346 (1983).
[59] B. Kégl, Intrinsic dimension estimation using packing numbers, in *Advances in Neural Information Processing Systems 15, Neural Information Processing Systems, NIPS 2002, 2002, Vancouver, British Columbia, Canada* (MIT Press, Cambridge, MA, 2002), pp. 681–688.
[60] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. **9**, 2579 (2008).
[61] L. van der Maaten, Learning a parametric embedding by preserving local structure, in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, 2009* (PMLR, Clearwater Beach, Florida, 2009), pp. 384–391.
[62] L. van der Maaten and G. E. Hinton, Visualizing non-metric similarities in multiple maps, Mach. Learn. **87**, 33 (2012).
[63] L. van der Maaten, Accelerating t-sne using tree-based algorithms, J. Mach. Learn. Res. **15**, 3221 (2014).
[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn:

Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[65] A. Novikov, PyClustering: Data mining library, J. Open Source Software, **4**, 1230 (2019).
[66] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein Generative Adversarial Networks, Proceedings of the 34th International Conference on Machine Learning, PMLR 70:214-223, 2017.
[67] M. Erdmann, L. Geiger, J. Glombitza, and D. Schmidt, Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks, Comput. Softw. Big Sci. **2**, 4 (2018).
[68] M. Erdmann, J. Glombitza, and T. Quast, Precise simulation of electromagnetic calorimeter showers using a Wasserstein generative adversarial network, Comput. Softw. Big Sci. **3**, 4 (2019).
[69] V. Chekalina, E. Orlova, F. Ratnikov, D. Ulyanov, A. Ustyuzhanin, and E. Zakharov, Generative models for fast calorimeter simulation, in *23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018) Sofia, Bulgaria, 2018* (EDP Sciences, Les Ulis, 2018).
[70] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, JUNIPR: A framework for unsupervised machine learning in particle physics, Eur. Phys. J. C **79**, 102 (2019).
[71] L. N. Wasserstein, Markov processes over denumerable products of spaces describing large systems of automata, Probl. Inf. Transm. **5**, 47 (1969).
[72] R. L. Dobrushin, Prescribing a system of random variables by conditional distributions, Theory Probab. Its Appl. **15**, 458 (1970).
[73] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018).
[74] A. De Simone and T. Jacques, Guiding new physics searches with unsupervised learning, Eur. Phys. J. C **79**, 289 (2019).
[75] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics, arXiv:1807.10261.
[76] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or What?, SciPost Phys. **6**, 030 (2019).
[77] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, arXiv:1808.08992.
[78] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Variational Autoencoders for New Physics Mining at the Large Hadron Collider, J. High Energy Phys. 05 (2019) 036.
[79] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99**, 014038 (2019).