Metric Methods with Open Collider Data

Eric M. Metodiev[*] **Patrick T. Komiske, Radha Mastandrea, Preksha Naik, Jesse Thaler**
Center for Theoretical Physics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
*metodiev@mit.edu

Abstract

We introduce a metric for collider data, synthesizing ideas from optimal transport and perturbative quantum field theory. The metric is the "work" required to rearrange one collider event into another, based on the earth mover's distance. Endowing collider data with a metric allows for distance-based unsupervised learning techniques to be used and can provide a simple alternative to sophisticated machine learning approaches. We use this metric to identify the most representative or anomalous events, to visualize the space of events, and to quantify the dimensionality of the dataset. We apply the metric to jets, sprays of particles from high-energy quarks and gluons, using public collider data from the CMS experiment at the Large Hadron Collider. We also make the processed jet dataset publicly available to empower future jet studies with open collider data.

1 Introduction

High energy proton collisions at the Large Hadron Collider (LHC) give rise to thousands of outgoing particles. There has been growing interest in applying sophisticated machine learning methods to collider data directly at the level of particles for tasks such as classifying the initiating particles in an event [1, 2, 3] or mitigating the effects of pileup [4]. Despite these developments in machine learning for collider physics, there remains little middle ground between traditional collider observables and sophisticated machine learning models. For instance, a notion of similarity or distance between collider events has yet to be developed to allow for simple metric-based unsupervised techniques to be applied. A robust notion of the "distance" between collider events would unlock new ways to probe events and significantly expand our ability to explore collider data.

Here, we present a metric for the space of collider data [5] based on ideas from optimal transport, particularly the earth mover's distance [6, 7, 8]. We apply this metric to public collider data from the CMS experiment [9]. In particular, we focus on applying various unsupervised learning techniques to jets. Jets are collimated sprays of particles that arise from the fragmentation and hadronization of outgoing high energy quarks and gluons. We use this metric to identify the most representative jets using the k-medoids algorithm, to identify the most anomalous jets in a dataset, to visualize the entire space of jets at once, and to quantify the fractal dimensionality of the dataset. Finally, we can relate this metric with origins in optimal transport to rich ideas from perturbative quantum field theory. We make our dataset of jets processed from CMS Open Data publicly available for reproducibility and to enable future jet studies with public collider data.

2 Dataset

CMS Open Data The CMS experiment [10] at the LHC has taken the unprecedented step of releasing public research-grade collider data [11], beginning in November 2014. These data releases

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

have enabled new exploratory studies of jets and jet substructure [12, 13], new physics searches [14], and machine learning studies on simulated events [15, 16, 17].

We make use of the CMS 2011A Jet primary dataset [18] and associated Monte Carlo files [19, 20, 21, 22, 23, 24, 25], which contain generated events before and after GEANT4 detector simulation. The jets considered in this study are "AK5" anti-k_T jets [26] with radius parameter $R = 0.5$, operationally defined by clustering the particles in an event with a hierarchical agglomerative clustering algorithm. Jets with transverse momenta $p_T \geq 375$ GeV are kept based on the firing of the `Jet300` trigger. We established that essentially all such jets seen by the CMS detector during this period of data taking are kept in the dataset.

Jet Dataset Our complete jet dataset processed from the CMS Open Data is available on the Zenodo platform, both for jets in data [27] as well as the associated simulated datasets [28, 29, 30, 31, 32, 33, 34, 35]. The dataset contains a total of 1,785,625 jets recorded by CMS, which yields about 40,000 jets after the additional kinematic jet selections described below.

Each jet is stored as a list of particle candidates reported by CMS, with transverse momentum p_T, rapidity y, azimuthal angle ϕ, mass m, particle identification code, and vertex information stored for each particle. From this jet dataset, we further restrict to jets with transverse momenta $p_T \in [399, 401]$ GeV, of "medium" quality, and with pseudorapidity $|\eta| < 1.9$ to be in the tracking region of the detector, where charged particles can be accurately vertexed. Charged particles from pileup collisions, namely those other than the collision of interest, are identified with vertex information and removed using the charged hadron subtraction procedure [36]. Further, we restrict to charged particles (tracks) with $p_T > 1$ GeV to minimize the impact of detector resolution and neutral pileup. Jets are centered and rotated to vertically align their principal component in the rapidity-azimuth plane. The jets are finally rescaled to have their constituent transverse momenta sum to 400 GeV to highlight the jet substructure.

Two example jets from the dataset are shown in Figs. 1a and 1b.

3 Method

We characterize a jet $\mathcal{J}$ by its distribution of energy flowing into the detector. Specifically, we focus on the distribution ρ of transverse momentum in the rapidity-azimuth (y, ϕ) plane:

$$\rho(y,\phi) = \sum_{j \in J} p_{T,j}\,\delta(y - y_j)\,\delta(\phi - \phi_j), \tag{1}$$

where y and ϕ are coordinates which parameterize the detector cylinder.

The earth (or energy) mover's distance (EMD) between two jets $\mathcal{I}$ and $\mathcal{J}$ is then:

$$\text{EMD}(\mathcal{I}, \mathcal{J}) = \min_{\{f_{ij}\}} \sum_{i \in I} \sum_{j \in J} f_{ij}\theta_{ij}, \tag{2}$$

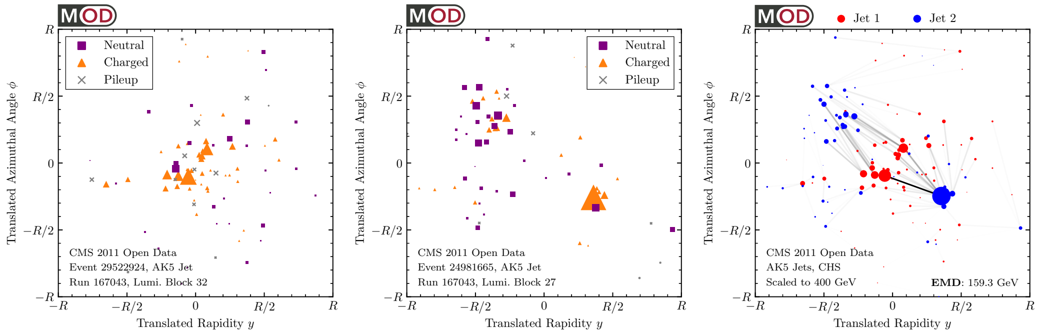
where $\theta_{ij}^2 = ((y_i - y_j)^2 + (\phi_i - \phi_j)^2)/R^2$ is a rapidiy-azimuth distance between particles, with $R = 0.5$ being the jet radius in our case. Here, f_{ij} is the amount of energy moved from particle i in jet $\mathcal{I}$ to particle j in jet $\mathcal{J}$, with the natural constraints:

$$f_{ij} \geq 0, \quad \sum_{i \in \mathcal{I}} f_{ij} = p_{T,j}, \quad \sum_{j \in \mathcal{J}} f_{ij} = p_{T,i}, \quad \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} f_{ij} = \sum_{i \in \mathcal{I}} p_{T,i} = \sum_{j \in \mathcal{J}} p_{T,j}. \tag{3}$$

The EMD is a true metric in that it is symmetric, non-negative, and satisfies the triangle inequality. Jets with different total transverse momenta can be compared by slightly modifying this definition, which we avoid here by rescaling them to 400 GeV to focus on the jet substructure. Finding the minimum f_{ij} in Eq. 2 subject to the constraints in Eq. 3 is an optimal transport problem which can be solved via the network simplex algorithm, where we use the python optimal transport library [37].

The optimal transportation plan between two example jets in the dataset is shown in Fig. 1c.

It is worth remarking that this metric, which has been used extensively for image retrieval and point cloud comparisons, is also deeply connected to core ideas in perturbative quantum field theory. Infrared and collinear (IRC) safety is a key concept that guarantees that an observable is perturbatively



(a) (b) (c)

Figure 1: (a,b) Two jets from the CMS Open Data, with the size of each symbol indicating the particle transverse momentum and the style indicating the charge. Charged pileup particles are indicated by gray crosses and removed from the jet. (c) The two jets represented as energy distributions, along with the optimal transportation plan to rearrange one jet into the other, with the intensity of each line corresponding to the transported energy between the particles. The two jets are 159.3 GeV apart.

finite and calculable. An observable is IRC safe if it is insensitive to the addition of a zero-energy particle and the collinear splitting of one particle into two. The jet energy flow in Eq. 1 is manifestly IRC safe, and jets which are close in the EMD are guaranteed to be close in any IRC safe observable.

4 Results

Representative Jets Endowing the space of collider events with a metric unlocks a number of unsupervised learning techniques. We begin with finding the k-medoids, namely the most representative jets in the dataset. Namely, we want to find k jets $\{\mathcal{K}_1, \ldots, \mathcal{K}_k\}$ minimizing the distance of the dataset to those jets:

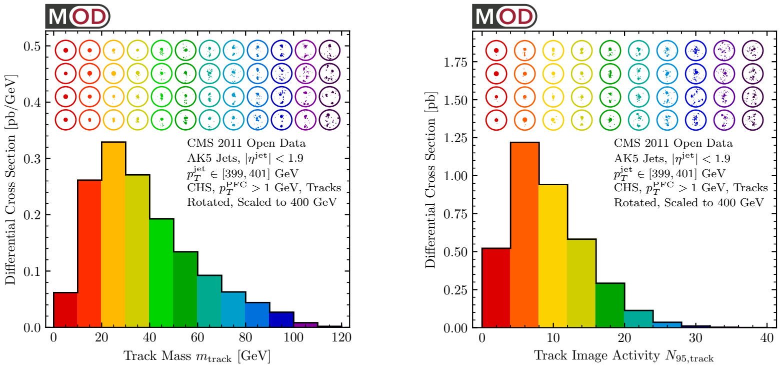
$$\mathcal{V}_k = \frac{1}{N}\sum_{i=1}^{N} \min\{\text{EMD}(\mathcal{J}_i, \mathcal{K}_1), \ldots, \text{EMD}(\mathcal{J}_i, \mathcal{K}_k)\}. \tag{4}$$

We use the approximation algorithm in the pyclustering python package [38] to find the k medoids in this work. In Fig. 2, we visualize the physics behind distributions of jet substructure observables by finding the 4-medoids in each histogram bin. We focus on the invariant mass of the jet and the image activity: the number of pixels in a 33×33 image that account for 95% of the transverse momentum. This visualization highlights for instance that the mass of a jet is dominantly generated by the formation of an additional hard prong. We also find and display the 25-medoids of the entire dataset in Fig. 4b, which we discuss more below. More broadly, this notion of representative events may be useful for "triggering" or robust compression of collider datasets.

Anomaly Detection Model-agnostic and data-driven anomaly detection techniques have been of recent interest in collider physics, motivated in part due to the lack of new physics discoveries at the LHC using targeted methods. We can use the EMD to determine which jets are most anomalous, by finding the ones farthest from the rest of the dataset. We quantify this by computing the mean distance $\overline{Q}$ of each jet $\mathcal{I}$ to all the jets in the dataset:

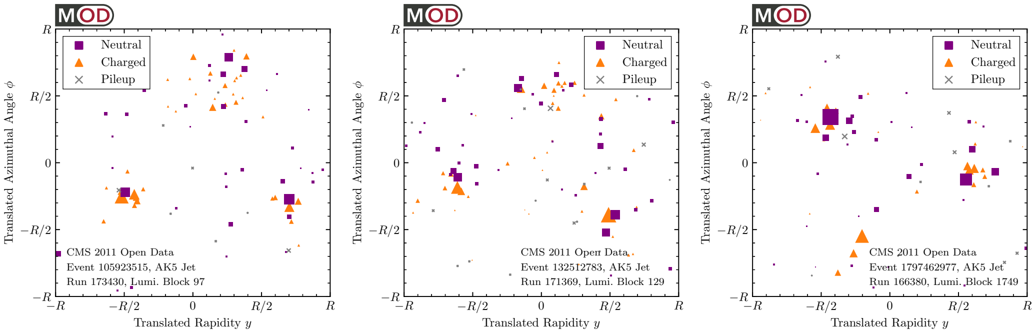
$$\overline{Q}(\mathcal{I}) = \frac{1}{N}\sum_{j=1}^{N} \text{EMD}\left(\mathcal{I}, \mathcal{J}_j\right). \tag{5}$$

Small values of $\overline{Q}$ correspond to jets with typical substructure, and large values of $\overline{Q}$ correspond to anomalous jets far from most jets in the dataset. In Fig. 3, we show the three most anomalous jets in the dataset according to Eq. 5. We can see that indeed they have uncommon three-pronged topologies, indicating that this measure of distance indeed captures aspects of uncommon substructure.



(a) (b)

Figure 2: Distributions of two jet substructure observables (a) the jet mass and (b) the jet image activity in the CMS Open Data, showing the 4-medoids in each histogram bin.



(a) (b) (c)

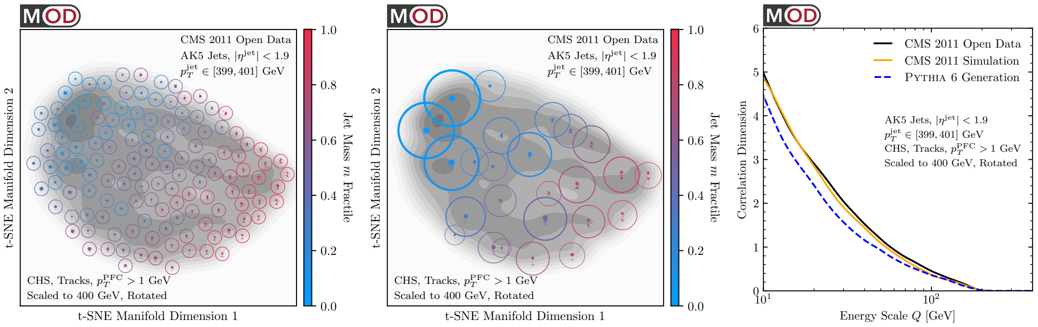
Figure 3: Event displays of the three most anomalous jets in our CMS Open Data sample based on their mean distance to the dataset. Jets with more exotic, three-pronged substructure emerge as the most anomalous.

The Space of Jets We can also visualize the entire space of jets from the CMS Open Data using t-distributed Stochastic Neighbor Embedding (t-SNE) [39], which seeks to find a low-dimensional embedding that respects the distance between points. Shown in Fig. 4a is the t-SNE embedding of the dataset using the implementation in scikit-learn [40], with example jets distributed throughout the embedding. The two-dimensional embedding includes regions of one- and two-pronged jets, with varying energy sharing fractions between the prongs. Also shown in Fig. 4b is the embedding together with the 25-medoids of the dataset, which can be seen to "tile" the space.

Correlation Dimension The dimensionality of the space of jets can be probed solely using pairwise distances between points alone. To that end, we use the notion of fractal dimension, specifically the correlation dimension [41, 39]:

$$\dim(Q) = Q\frac{\partial}{\partial Q}\ln\sum_{1\le k<\ell\le N}\Theta[\mathrm{EMD}(\mathcal{J}_k,\mathcal{J}_\ell) < Q], \tag{6}$$

where Θ is a step function indicating whether jet k is within EMD Q of jet ℓ. The correlation dimension is a scale-dependent quantity, with different physics dominating at different energy scales Q. Fig. 4c shows the correlation dimension in the CMS Open Data, compared to Monte Carlo samples before and after detector simulation. We see an increase of dimensionality with decreasing energy scale. This behavior can also be understood and computed in perturbative quantum chromodynamics, which we will develop and showcase in future work.



(a)

(b)

(c)

Figure 4: (a,b) A two-dimensional t-SNE embedding of jets from the CMS Open Data, with the gray contours indicating the density of jets in the space. Overlaid are (a) example jets are shown uniformly throughout the space, colored according to their jet mass fractile in the dataset, and (b) the 25-medoid jets of the dataset, sized with area proportional to the number of jets nearest to them. (c) The correlation dimension of the CMS Open Data, together with Monte Carlo samples before and after detector simulation.

5 Conclusion

Data-driven methods that circumvent a reliance on simulated truth information or specific new physics models are of growing interest for collider physics. To enable the use of many unsupervised learning techniques at the LHC, we have established a new metric for collider data. We showcased that metric on jets from public collider data from the CMS experiment. Applications beyond those discussed here can be built using this metric, such as k-nearest neighbors jet classifiers or clustering of collider events. We have released our full processed jet dataset to facilitate such future explorations.

Acknowledgments

We thank CERN, the CMS collaboration, and the CMS Data Preservation and Open Access (DPOA) team for making research-grade collider data available to the public. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under Grant No. DESC0011090 and the DOE Office of High Energy Physics under Grant Nos. DE-SC0012567 and DE-SC0019128. The authors benefited from the hospitality of the Harvard Center for the Fundamental Laws of Nature. Cloud computing resources were provided through a Google Cloud allotment from the MIT Quest for Intelligence.

References

[1] Gilles Louppe, Kyunghyun Cho, Cyril Becot, and Kyle Cranmer. QCD-Aware Recursive Neural Networks for Jet Physics. *JHEP*, 01:057, 2019.

[2] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy Flow Networks: Deep Sets for Particle Jets. *JHEP*, 01:121, 2019.

[3] Huilin Qu and Loukas Gouskos. ParticleNet: Jet Tagging via Particle Clouds. 2019.

[4] J. Arjona Martínez, Olmo Cerri, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Pileup mitigation at the Large Hadron Collider with graph neural networks. *Eur. Phys. J. Plus*, 134(7):333, 2019.

[5] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Metric Space of Collider Events. *Phys. Rev. Lett.*, 123(4):041801, 2019.

[6] Shmuel Peleg, Michael Werman, and Hillel Rom. A unified approach to the change of resolution: Space and gray-level. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):739–742, 1989.

[7] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 59–66, Washington, DC, USA, 1998. IEEE Computer Society.

[8] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, November 2000.

[9] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. Exploring the Space of Jets with CMS Open Data. 2019.

[10] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[11] CERN Open Data Portal. `http://opendata.cern.ch`.

[12] Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, and Jesse Thaler. Jet Substructure Studies with CMS Open Data. *Phys. Rev.*, D96(7):074003, 2017.

[13] Andrew Larkoski, Simone Marzani, Jesse Thaler, Aashish Tripathee, and Wei Xue. Exposing the QCD Splitting Function with CMS Open Data. *Phys. Rev. Lett.*, 119(13):132003, 2017.

[14] Cari Cesarotti, Yotam Soreq, Matthew J. Strassler, Jesse Thaler, and Wei Xue. Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum. *Phys. Rev.*, D100(1):015021, 2019.

[15] Celia Fernández Madrazo, Ignacio Heredia Cacha, Lara Lloret Iglesias, and Jesús Marco de Lucas. Application of a Convolutional Neural Network for image classification to the analysis of collisions in High Energy Physics. 2017.

[16] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos. End-to-End Physics Event Classification with the CMS Open Data: Applying Image-based Deep Learning on Detector Data to Directly Classify Collision Events at the LHC. 2018.

[17] Michael Andrews, John Alison, Sitong An, Patrick Bryant, Bjorn Burkle, Sergei Gleyzer, Meenakshi Narain, Manfred Paulini, Barnabas Poczos, and Emanuele Usai. End-to-End Jet Classification of Quarks and Gluons with the CMS Open Data. 2019.

[18] CMS Collaboration. Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD). *CERN Open Data Portal*, 2016.

[19] CMS Collaboration. Simulated dataset QCD_Pt-170to300_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[20] CMS Collaboration. Simulated dataset QCD_Pt-300to470_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[21] CMS Collaboration. Simulated dataset QCD_Pt-470to600_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[22] CMS Collaboration. Simulated dataset QCD_Pt-600to800_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[23] CMS Collaboration. Simulated dataset QCD_Pt-800to1000_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[24] CMS Collaboration. Simulated dataset QCD_Pt-1400to1800_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[25] CMS Collaboration. Simulated dataset QCD_Pt-1800_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive). *CERN Open Data Portal*, 2016.

[26] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008.

[27] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Open Data | Jet Primary Dataset | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[28] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 170-300 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[29] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 300-470 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[30] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 470-600 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[31] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 600-800 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[32] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 800-1000 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[33] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 1000-1400 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[34] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 1400-1800 | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[35] Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, and Jesse Thaler. CMS 2011A Simulation | Pythia 6 QCD 1800-inf | pT $>$ 375 GeV | MOD HDF5 Format. *Zenodo*, 2019.

[36] CMS Collaboration. Pileup Removal Algorithms. Technical Report CMS-PAS-JME-14-001, 2014.

[37] R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017.

[38] Andrei Novikov. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230, apr 2019.

[39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[40] 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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[41] Peter Grassberger and Itamar Procaccia. Characterization of Strange Attractors. *Phys. Rev. Lett.*, 50:346–349, 1983.