Interaction networks for the identification of Higgs boson decays to bottom quark-antiquark pairs

Eric A. Moreno, Olmo Cerri, Thong Q. Nguyen, Jean-Roch Vlimant, Harvey B. Newman, Avikar Periwal, Aidana Serikova, Maria Spiropulu
California Institute of Technology

Javier M. Duarte
University of California San Diego
Fermi National Accelerator Laboratory (FNAL)

Maurizio Pierini
European Center for Nuclear Research (CERN)

Abstract

We investigate the performance of a jet identification algorithm based on an interaction network to identify high-momentum Higgs bosons decaying to bottom quark-antiquark pairs and distinguish them from ordinary jets originating from the hadronization of quarks and gluons. The algorithm's inputs are features of the reconstructed charged particles in a jet and the secondary vertices associated to them. Describing the jet shower as a combination of particle-to-particle and particle-to-vertex interactions, an interaction network is trained to learn a jet representation with which to classify the jet. The algorithm is trained on simulated samples of accurate LHC collisions, released by the CMS collaboration on the CERN Open Data Portal. The interaction network achieves a drastic improvement in the identification performance with respect to state-of-the-art algorithms.

1 Introduction

Jets are collimated cascades of particles produced at particle accelerators. Quarks and gluons originating from hadron collisions, such as the proton-proton collisions at the CERN Large Hadron Collider (LHC), generate a cascade of other particles (mainly other quarks or gluons) that then arrange themselves into hadrons. The stable and unstable hadrons' decay products are observed by large particle detectors, reconstructed by algorithms that combine the information from different detector components, and then clustered into jets, using physics-motivated sequential recombination algorithms such as those described in Ref. [1–3]. Jet identification, or *tagging*, algorithms are designed to identify the nature of the particle that initiated a given cascade, inferring it from the collective features of the particles generated in the cascade. Recently, several approaches based on deep learning have been proposed to optimize jet tagging algorithms using convolutional neural networks [4–9], physics-inspired dense neural networks [10–12], or recurrent and recursive neural networks [13–15]. For instance, the LHC collaborations and other researchers have investigated the optimal way to combine substructure, tracking, and vertexing information to enhance the tagging efficiency for high transverse momentum (p_{T}) $\mathrm{H} \to \mathrm{b\bar{b}}$ decays [16–20]. This is an important task in particle physics because measurements of high-p_{T} $\mathrm{H} \to \mathrm{b\bar{b}}$ decays may help resolve the loop induced and tree-level contributions to the gluon fusion process [21, 22] and provide an alternative approach to study the top quark Yukawa coupling in addition to the $\mathrm{t\bar{t}H}$ process [23, 24].

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In this work, we propose to accomplish this task with an interaction network (IN). In Ref. [25], INs were introduced to describe complex physical systems and predict their evolution after a certain amount of time. Particle jets, distributed sensor networks, and power grids are examples of systems that involve multiple entities with complex interactions. Graphs provide a natural representation for encoding such relational information without requiring an arbitrary ordering of elements, a fixed-Euclidean-grid assumption, special handling of sparse inputs, or substantial pre-processing.

Geometric deep learning, including graph convolution networks [26–31] and graph generative models [32, 33], leverages a graph representation to learn directly from structured data. In particle physics, graph neural networks have recently been used for jet tagging, matching the performance of other deep learning approaches [34–36], for event classification [37, 38], charged particle tracking in a silicon detector [39], pileup subtraction at the LHC [40], and particle reconstruction in irregular calorimeters [41] as well as in the IceCube experiment [38]. In this paper, we extend the use of graph neural networks to the case of $\mathrm{H} \to \mathrm{b\bar{b}}$ tagging. In particular, we investigate the use of INs to learn a collective representation of the tracking, vertexing, and substructure properties of the jet and employ this optimized representation to enhance the tagging efficiency. By placing charged particles and secondary vertices on a graph, the network can learn a representation of each particle-to-particle and particle-to-vertex *interaction*, and exploit this information to categorize a given jet as signal ($\mathrm{H} \to \mathrm{b\bar{b}}$) or background (QCD).

The study is carried out using a sample of fully-simulated LHC collision events, released by the CMS collaboration on the CERN Open Data Portal [42]. The open data simulation allows for a more in depth and realistic study of the efficacy of machine learning methods on high-energy physics experiments. We compare the performance to a different algorithm that we trained with open data for $\mathrm{H} \to \mathrm{b\bar{b}}$ tagging based on the architecture of the deep double-b (DDB) tagger created by the CMS collaboration [17]. The DDB tagger is a convolutional and recurrent neural network model based on the 27 high-level features used in Ref. [16], as well as 8 charged particle features, and 2 properties of secondary vertices associated with the jet.

2 Data Samples

The dataset is obtained from the samples provided on CERN Open Data Portal, which consist of CMS simulated data from $\mathrm{H} \to \mathrm{b\bar{b}}$ and QCD processes corresponding to the 2016 data-taking conditions [43]. Jets are clustered from the reconstructed particles using the anti-k_{T} algorithm [3, 44] with a jet-size parameter $R = 0.8$ (AK8 jets) and subsequently their energy is corrected. In order to remove soft, wide-angle radiation from the jet, the soft-drop (SD) algorithm [45, 21] is applied, with angular exponent $\beta = 0$, soft cutoff threshold $z_{\mathrm{cut}} < 0.1$, and characteristic radius $R_0 = 0.8$ [46]. The soft-drop mass (m_{SD}) is then computed from the four-momenta of the remaining constituents. The dataset is reduced by requiring the jets to have $300 < p_{\mathrm{T}} < 2400\,\mathrm{GeV}$ and $40 < m_{\mathrm{SD}} < 200\,\mathrm{GeV}$. The IN uses 30 features related to charged particles and 14 features related to the secondary vertices described in Ref. [43].

In this study, we use samples consisting of 3.9 million $\mathrm{H} \to \mathrm{b\bar{b}}$ jets and 1.9 million inclusive QCD jets, split into training, validation, and test samples with proportions of 80%, 10%, and 10%, respectively.

3 Model architecture

The IN is based on two input collections comprising N_p particles, each represented by a feature vector of length P, and N_v vertices, each represented by a feature vector of length S. The input consists of an ensemble of X and Y matrices, with sizes $P \times N_p$ and $P \times N_v$, respectively.

A particle graph $\mathcal{G}_p$ is constructed by connecting each particle to each other particle through $N_{pp} = N_p(N_p - 1)$ directed edges. Similarly, a particle-vertex graph $\mathcal{G}_{pv}$ is constructed by connecting each particle to each vertex through $N_{pv} = N_p N_v$ undirected edges. For the graph $\mathcal{G}_p$, a receiving matrix (R_R) and a sending matrix (R_S) are defined, both of size $N_p \times N_{pp}$. For the second graph, the corresponding adjacency matrices R_K (of size $N_p \times N_{vp}$) and R_V (of size $N_v \times N_{vp}$) are defined.

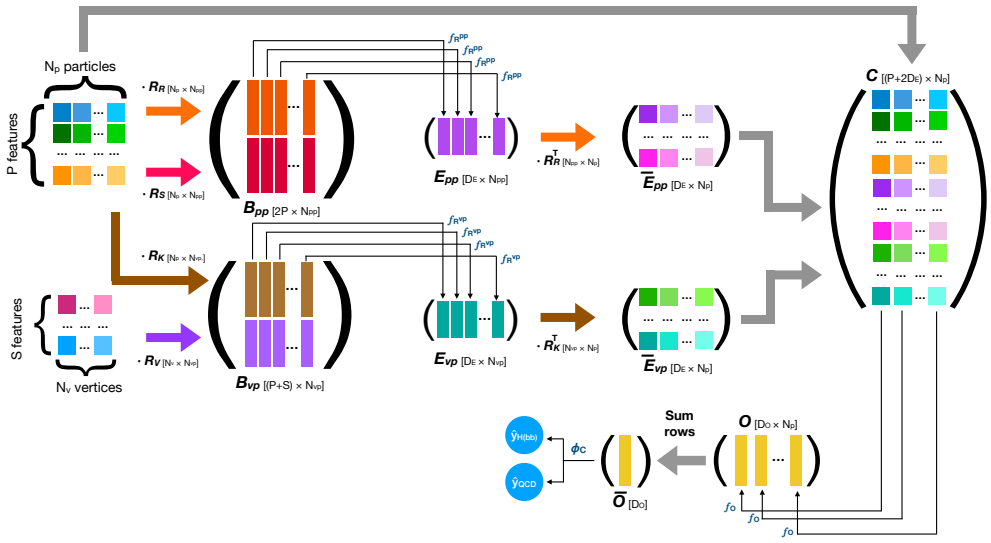

Figure 1: Illustration of the IN architecture.

The input processing begins by creating the $2P \times N_{pp}$ particle-particle interaction matrix B_{pp} and the $(P + S) \times N_{vp}$ particle-vertex interaction matrix B_{vp} defined as:

$$B_{pp} = \begin{pmatrix} X \cdot R_R \\ X \cdot R_S \end{pmatrix}, \tag{1}$$

$$B_{vp} = \begin{pmatrix} X \cdot R_K \\ Y \cdot R_V \end{pmatrix}, \tag{2}$$

where $\cdot$ indicates the ordinary matrix product. Processing each column of B_{pp}, one builds an internal representation of the particle-particle interaction with a neural network $f_R^{pp} : \mathbb{R}^{2P} \mapsto \mathbb{R}^{D_E}$, where D_E is the size of the internal representation. This results in an *effect matrix*: E_{pp} with dimensions $D_E \times N_{pp}$. We similarly build the E_{vp} matrix, with dimensions $D_E \times N_{vp}$, using a neural network $f_R^{vp} : \mathbb{R}^{P+S} \mapsto \mathbb{R}^{D_E}$.

We then propagate the particle-particle interactions back to the particles receiving them, by building $\overline{E}_{pp} = E_{pp} R_R^\top$ with dimension $D_E \times N_p$. We also build $\overline{E}_{vp} = E_{vp} R_V^\top$ with dimension $D_E \times N_p$, which collects the information of the particle-vertex interactions for each particle and across all of the vertices.

The next step consists of building the C matrix, with dimensions $(P + 2D_E) \times N_O$, by combining the input information for each particle (X) with the learned representation of the particle-particle ($\overline{E}_{pp}$) and particle-vertex ($\overline{E}_{vp}$) interactions:

$$C = \begin{pmatrix} X \\ \overline{E}_{pp} \\ \overline{E}_{vp} \end{pmatrix}. \tag{3}$$

The final aggregator combines the input and interaction information to build the post-interaction representation of the graph, summarized by the matrix O, with dimensions $D_O \times N_p$. The aggregator consists of a function $f_O : \mathbb{R}^{P+2D_E} \mapsto \mathbb{R}^{D_O}$, which computes the elements of the O matrix. The elements of the O matrix are computed by a function $f_O : \mathbb{R}^{P+2D_E} \mapsto \mathbb{R}^{D_O}$, which returns the post-interaction representation for each of the input particles.

The learned representation of the post-interaction graph, represented by the elements of the O matrix, can be used to solve the specific task at hand. We sum O along each row to produce a feature vector $\overline{O}$ with length D_O for the jet as a whole. This is passed to a final function $\phi_C : \mathbb{R}^{D_O} \mapsto \mathbb{R}^N$, which produces the output of the classifier. The overall architecture of the IN is illustrated in Fig. 1. We

stress the fact that the chosen architecture makes the outcome of the IN tagger independent of the order used to label the N_p input particles and N_v input vertices.

We train the model by minimizing the categorical cross-entropy loss using the Adam optimizer [47] with an initial learning rate of 10^{-4} and a minibatch size of 128 for up to 100 epochs, enforcing early stopping [48] on the validation loss with a patience of 5 epochs.

4 Decorrelation with the jet mass

Many possible applications of this algorithm would require that the tagger score be uncorrelated from the jet mass so that a selection based on the tagger score does not change the jet mass distribution, particularly for the background events. This is especially crucial for the applicability of modeling the background jet mass distribution with an analytic function in the final physics analysis.

We test 3 different techniques to minimize the IN tagger's effects on the jet mass distribution of the selected events: (i) performing an adversarial training [49–51], where the classifier is penalized if its output score can be used to reconstruct the jet mass; (ii) removing or reweighting events [52] such that the background jet mass distribution is indistinguishable from the signal jet mass distribution then training the network; and (iii) defining mass-dependent thresholds as in the "designing decorrelated taggers" (DDT) procedure [53], where each selection threshold corresponds to k-percentile of the background's output score distribution given a mass value, with $(1-k)$ being the overall false-positive rate of the classifier's working point. A mass-decorrelation technique is also applied to the DDB tagger using the Kullback–Leibler divergence as a penalty in the loss function.

5 Performance

As shown in Fig. 2 (left), the IN provides an improved performance with respect to our recreation of the DDB tagger [17], which was trained on the same CMS open data simulation as the IN. At a 1% FPR, the IN tagger outperforms the DDB tagger by 36% in true positive rate (TPR). Likewise, at an 80% TPR, the IN tagger yields a factor of 4 smaller false positive rate (FPR) than the the DDB tagger.

Fig. 2 (right) shows an illustration of how the background jet mass distribution changes after applying a threshold on the tagger score for different decorrelation algorithms. The DDT procedure provides the best decorrelation of the IN tagger followed by the reweighted training and the domain-adversarial training, respectively. After applying the mass decorrelation techniques, the performance worsens slightly but still significantly outperforms the DDB taggers. At a 1% FPR, the DDT-decorrelated IN tagger has a TPR of 76% compared to the decorrelated DDB tagger with a 51% TPR, corresponding to an improvement of 49%.

6 Conclusions

We demonstrate a novel jet-tagging technique using graph representations of the jet's constituents and secondary vertices based on an interaction network. The IN can operate on a variable number of jet constituents and secondary vertices and does not depend on the ordering schemes of these objects. Different mass decorrelation techniques are explored to minimize the selection bias of the classifier output toward any values of the jet mass. The IN significantly outperforms an alternative tagger based on one created by the CMS collaboration and also allows for more flexible representations of jet data. These results motivate further exploration of using interaction networks for different object tagging in experimental high energy physics.

References

[1] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, *Better jet clustering algorithms*, *JHEP* **08** (1997) 001, [hep-ph/9707323].

[2] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl. Phys. B* **406** (1993) 187–224.

[3] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063, [arXiv:0802.1189].

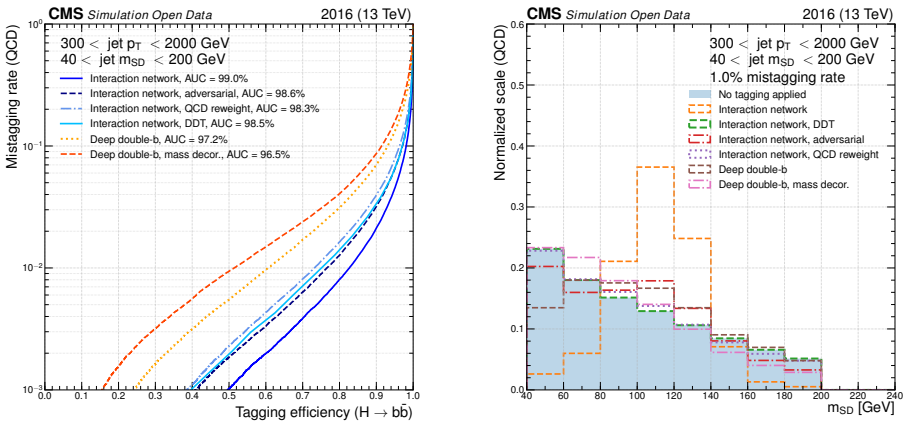

Figure 2: (Left) Performance of different algorithms quantified as a receiver operator characteristic (ROC) curve where the y-axis is the mistagging rate or false positive rate and the x-axis is the tagging efficiency or true positive rate. The interaction network's performance is shown before and after applying methods to decrease dependence of the tagger score on the jet mass. These results are compared with the ROC curves for the deep double-b tagger trained on CMS open data simulation without and with a mass decorrelation technique using the Kullback–Leibler divergence as a penalty in the loss function. (Right) An illustration of the "sculpting" of the background jet mass distribution after applying a threshold on the tagger score for several different algorithms. The unmodified interaction network is highly correlated with the jet mass, but after applying the methods described in the text, the correlation is reduced.

[4] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, *Jet-images – deep learning edition*, *JHEP* **07** (2016) 069, [arXiv:1511.05190].

[5] D. Guest, J. Collado, P. Baldi, et al., *Jet flavor classification in high-energy physics with deep neural networks*, *Phys. Rev.* **D94** (2016), no. 11 112002, [arXiv:1607.08633].

[6] S. Macaluso and D. Shih, *Pulling out all the tops with computer vision and deep learning*, *JHEP* **10** (2018) 121, [arXiv:1803.00107].

[7] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, *Deep-learning top taggers or the end of QCD?*, *JHEP* **05** (2017) 006, [arXiv:1701.08784].

[8] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110, [arXiv:1612.01551].

[9] A. Schwartzman, M. Kagan, L. Mackey, B. Nachman, and L. De Oliveira, *Image Processing, Computer Vision, and Deep Learning: new approaches to the analysis and physics interpretation of LHC events*, *J. Phys. Conf. Ser.* **762** (2016), no. 1 012035.

[10] K. Datta and A. J. Larkoski, *Novel jet observables from machine learning*, *JHEP* **03** (2018) 086, [arXiv:1710.01305].

[11] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, *Deep-learned top tagging with a Lorentz layer*, *SciPost Phys.* **5** (2018), no. 3 028, [arXiv:1707.08966].

[12] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, *JHEP* **04** (2018) 013, [arXiv:1712.07124].

[13] G. Louppe, K. Cho, C. Becot, and K. Cranmer, *QCD-aware recursive neural networks for jet physics*, arXiv:1702.00748.

[14] S. Egan, W. Fedorko, A. Lister, J. Pearkes, and C. Gay, *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC*, arXiv:1711.09059.

[15] T. Cheng, *Recursive Neural Networks in Quark/Gluon Tagging*, *Comput. Softw. Big Sci.* **2** (2018), no. 1 3, [arXiv:1711.02633].

[16] **CMS** Collaboration, A. M. Sirunyan et al., *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, *JINST* **13** (2018), no. 05 P05011, [arXiv:1712.07158].

[17] **CMS** Collaboration, *Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector*, Jul, 2018.

[18] **CMS** Collaboration, *Boosted jet identification using particle candidates and deep neural networks*, Nov, 2017.

[19] **ATLAS** Collaboration, G. Aad et al., *Identification of boosted Higgs bosons decaying into b-quark pairs with the ATLAS detector at 13 TeV*, arXiv:1906.11005.

[20] J. Lin, M. Freytsis, I. Moult, and B. Nachman, *Boosting $H \to b\bar{b}$ with Machine Learning*, *JHEP* **10** (2018) 101, [arXiv:1807.10768].

[21] 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** (2008) 242001, [arXiv:0802.2470].

[22] K. Becker, F. Caola, A. Massironi, et al., *Recommended predictions for the boosted-Higgs cross section*, Tech. Rep. LHCHXSWG-2019-002, CERN, Geneva, Mar, 2019.

[23] **CMS** Collaboration, A. M. Sirunyan et al., *Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair*, *Phys. Rev. Lett.* **120** (2018), no. 7 071802, [arXiv:1709.05543].

[24] **ATLAS** Collaboration, *Search for boosted resonances decaying to two b-quarks and produced in association with a jet at $\sqrt{s} = 13$ TeV with the ATLAS detector*, 2018.

[25] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, and K. Kavukcuoglu, *Interaction Networks for Learning about Objects, Relations and Physics*, *ArXiv e-prints* (Dec., 2016) [arXiv:1612.00222].

[26] M. Niepert, M. Ahmed, and K. Kutzkov, *Learning convolutional neural networks for graphs*, in *International conference on machine learning*, pp. 2014–2023, 2016.

[27] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, *arXiv preprint arXiv:1609.02907* (2016).

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, *Pointnet: Deep learning on point sets for 3d classification and segmentation*, *CoRR* **abs/1612.00593** (2016) [arXiv:1612.00593].

[29] Y. Wang, Y. Sun, Z. Liu, et al., *Dynamic graph CNN for learning on point clouds*, *CoRR* **abs/1801.07829** (2018) [arXiv:1801.07829].

[30] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, *Geometric deep learning: Going beyond euclidean data*, *IEEE Signal Processing Magazine* **34** (July, 2017) 18–42.

[31] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, *Gated graph sequence neural networks*, *arXiv e-prints* (Nov., 2015) arXiv:1511.05493.

[32] A. Grover, A. Zweig, and S. Ermon, *Graphite: Iterative generative modeling of graphs*, *arXiv preprint arXiv:1803.10459* (2018).

[33] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec, *Graphrnn: Generating realistic graphs with deep auto-regressive models*, *arXiv preprint arXiv:1802.08773* (2018).

[34] I. Henrion et al., *Neural message passing for jet physics*, in *Proceedings of the Deep Learning for Physical Sciences Workshop at NIPS*, 2017.

[35] H. Qu and L. Gouskos, *ParticleNet: Jet tagging via particle clouds*, 2019.

[36] E. A. Moreno, O. Cerri, J. M. Duarte, et al., *JEDI-net: a jet identification algorithm based on interaction networks*, 2019.

[37] M. Abdughani, J. Ren, L. Wu, and J. M. Yang, *Probing stop with graph neural network at the LHC*, arXiv:1807.09088.

[38] N. Choma et al., *Graph neural networks for IceCube signal classification*, *CoRR* **abs/1809.06166** (2018).

[39] S. Farrell et al., *Novel deep learning methods for track reconstruction*, in *4th International Workshop Connecting The Dots 2018 (CTD2018) Seattle, Washington, USA, March 20-22, 2018*, 2018. arXiv:1810.06111.

[40] 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.

[41] S. R. Qasim, J. Kieseler, Y. Iiyama, and M. Pierini, *Learning representations of irregular particle-detector geometry with distance-weighted graph networks*, arXiv:1902.07987.

[42] "CERN Open Data Portal." http://opendata.cern.ch.

[43] **CMS** Collaboration, J. Duarte, "Sample with jet, track and secondary vertex properties for hbb tagging ml studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC." CERN Open Data Portal.

[44] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet user manual*, *Eur. Phys. J. C* **72** (2012) 1896, [arXiv:1111.6097].

[45] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, *Towards an understanding of jet substructure*, *JHEP* **09** (2013) 029, [arXiv:1307.0007].

[46] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146, [arXiv:1402.2657].

[47] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, *CoRR* **abs/1412.6980** (2014) [arXiv:1412.6980].

[48] Y. Yao, L. Rosasco, and A. Caponnetto, *On early stopping in gradient descent learning*, *Constructive Approximation* **26** (2007), no. 2 289–315.

[49] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, *Domain-Adversarial Neural Networks*, *arXiv e-prints* (Dec, 2014) arXiv:1412.4446, [arXiv:1412.4446].

[50] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, 2016.

[51] **ATLAS** Collaboration, *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, Tech. Rep. ATL-PHYS-PUB-2018-014, CERN, Geneva, Jul, 2018.

[52] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, *Mass Agnostic Jet Taggers*, arXiv:1908.08959.

[53] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, *Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure*, *JHEP* **05** (2016) 156, [arXiv:1603.00027].