Optimal transport for a novel event description at hadron colliders

L. Gouskos,[1,6] F. Iemmi,[2,*] S. Liechti,[3] B. Maier,[1,4,†] V. Mikuni,[5] and H. Qu[1]

[1]*European Organization for Nuclear Research (CERN), Meyrin 1211, Switzerland*
[2]*Institute of High Energy Physics (IHEP), Shijingshan District, Beijing 100049, China*
[3]*University of Zurich (UZH), Zurich 8057, Switzerland*
[4]*Karlsruhe Institute of Technology (KIT), POB 6980, 76049 Karlsruhe, Germany*
[5]*National Energy Research Scientific Computing Center, Berkeley Lab, Berkeley, California 94720, USA*
[6]*Department of Physics, Brown University, Providence, Rhode Island 02912, USA*

(Received 17 February 2023; revised 6 June 2023; accepted 22 September 2023; published 1 November 2023)

We propose a novel strategy for disentangling proton collisions at hadron colliders such as the LHC that considerably improves over the current state of the art. Employing a metric inspired by optimal transport problems as the cost function of a graph neural network, our algorithm is able to compare two particle collections with different noise levels and learns to flag particles originating from the main interaction amidst products from up to 200 simultaneous pileup collisions. We thereby sidestep the critical task of obtaining a ground truth by labeling particles and avoid arduous human annotation in favor of labels derived *in situ* through a self-supervised process. We demonstrate how our approach—which, unlike competing algorithms, is trivial to implement—improves the resolution in key objects used in precision measurements and searches alike and present large sensitivity gains in searching for exotic Higgs boson decays at the High-Luminosity LHC.

DOI: 10.1103/PhysRevD.108.096003

I. INTRODUCTION

At the High-Luminosity LHC (HL-LHC), up to 200 proton collisions will take place simultaneously. This poses an unprecedented challenge to the reconstruction algorithms of experiments such as ATLAS and CMS, hindering their ability to search for new physics with the highest sensitivity if collision products are not disentangled properly. A performant rejection of particles from subordinate proton collisions (pileup) is therefore paramount to the success of the LHC physics program. Exploiting the excellent position resolution of tracking systems, charged particles from pileup collisions can be effectively mitigated [1,2] by discarding particles not associated with the primary vertex. Contributions from neutral particles (photons and neutral hadrons), however, can only be reconstructed using the calorimeter systems with comparably poor spatial resolution. To this end, dedicated algorithms based on a physics-motivated, rule-based selection were developed to obtain a set of per-particle probabilities indicating whether neutral particles originate from the leading primary vertex or not. For instance, the PUPPI algorithm is the state of the art at the CMS experiment [3,4].

The complexity of the pileup mitigation task motivated the proposition of machine learning (ML)-based algorithms: the PUMML algorithm [5] relies on image recognition techniques to identify particles stemming from pileup vertices; other algorithms—e.g., in Refs. [6,7]—exploit graph neural networks (GNNs) or transformers. In all cases, the ML-based approaches yield sizable improvements compared to rule-based algorithms.

One of the main limitations of these ML algorithms is that they require a sound definition of a ground truth—i.e., the assignment of a label to each particle, indicating if it originates from the leading primary vertex or not—to train a fully supervised ML model. Such labels are available in simplified detector simulations, such as the ones implemented in DELPHES [8]. However, due to merged energy deposits and the much more complex event reconstruction, an unambiguous ground truth definition for neutral particles is intractable in data and in the full-scale simulations based on Geant4 [9] that are used by experiments such as ATLAS and CMS. Therefore, per-particle target labels based on human annotation are very hard to obtain and exhibit insufficient sharpness, rendering fully supervised strategies suboptimal. Thus, none of the previously described ML-based solutions can be implemented in a straightforward manner by the experiments at the LHC. Recently, it has been proposed in Ref. [10] to train a

*fabio.iemmi@cern.ch
†benedikt.maier@cern.ch

semisupervised network using charged particles, whose provenance can be obtained under the assumption of perfect tracking, and to extrapolate the results to neutral particles. While this approach would allow us to train on full-scale simulations and on data, it only works in a central detector region where tracking is available, whereas pileup is predominant in the forward regions of the detectors. Therefore, our primary motivation to develop the algorithm for training optimal transport with attention learning (TOTAL) in this article is to address these bottlenecks. We achieve this by employing metrics inspired by optimal transport (OT) [11] problems as the cost function of a self-supervised, attention-based GNN, whose architecture closely follows the development in [12]. This network is used to assimilate two simulated samples, one containing only the particles arising from the primary interaction and the other containing also contributions from pileup. By training in a self-supervised fashion—i.e., not relying on per-particle truth labels from human annotation—our approach can be realistically implemented in full-scale simulations and does not rely on any kind of extrapolation. Following this strategy, we are able to derive an event description that is straightforward to implement and that exhibits greatly improved precision, yielding a global sensitivity enhancement for SM measurements and searches alike, explained as follows.

II. SLICED WASSERSTEIN DISTANCE

Instead of relying on truth labels for reconstructed particles, we design an alternative objective that transforms the entire set of particles from multiple simultaneous interactions into the same collision event, containing only the primary interaction. Borrowing from concepts of OT, we are interested in finding the transport function that leaves particles with similar features between sets unchanged while removing contributions from additional interactions. The Wasserstein distance [13] can leverage geometric information in the probability space to estimate the distance between probability measures. Given two probability measures $\alpha \in P(X)$ and $\beta \in P(Y)$, both defined in the set of probability measures $P(\Omega)$, we identify the q-Wasserstein distance problem as the determination of the transportation plan $\gamma \in \Gamma(\alpha, \beta)$ that satisfies

$$\mathrm{OT}_q(\alpha,\beta) = \left(\inf_{\gamma\in\Gamma(\alpha,\beta)} \int_{\Omega\times\Omega} c(\mathbf{x},\mathbf{y})^q d\gamma(\mathbf{x},\mathbf{y})\right)^{1/q}, \quad (1)$$

where c is the cost function evaluated over observations $\mathbf{x}$, $\mathbf{y}$ drawn from the sets X, Y, and the order q can be chosen to give the qth root of the total cost incurred. In high-dimensional spaces, solving the OT problem becomes computationally expensive, since only the one-dimensional case allows a closed-form solution [14]. This observation motivates the formulation proposed in [14,15] to introduce the sliced Wasserstein distance (SWD) as an integral over one-dimensional transport problems:

$$\mathrm{SWD}(\alpha,\beta) = \int_{\mathcal{S}^{d-1}} \mathrm{OT}_q(\mathcal{R}_\theta(\alpha), \mathcal{R}_\theta(\beta)) d\theta. \quad (2)$$

The term $\mathcal{R}_\theta$ represents the linear operation that projects (or "slices") the probability measures over the one-dimensional space and is integrated over a uniform measure θ in the unit sphere $\mathcal{S}^{d-1} \in \mathbb{R}^d$. The dimensionality d corresponds to the number of particle features considered. Since the integral is intractable, we can instead use a Monte Carlo approach to replace it with multiple random projections. In this formulation, the optimal coupling is the one that minimizes the cost function evaluated over one-dimensional sorted projections, thus replacing the expensive OT problem with several (one for each projection) simple sorting problems.

Accordingly, we denote the set of particles in the sample with pileup as $x_p \in \mathbb{R}^{N\times d}$, while the same particle collision in the set without pileup is denoted as $x_{np} \in \mathbb{R}^{N\times d}$. Since the set of particles without pileup is considerably smaller than the one with, we keep the overall number of particles N fixed by zero-padding as necessary. Given M projections with permutations μ and ν that sort N particles of the sets $\mathcal{R}_{\theta_m}(x_p)$ and $\mathcal{R}_{\theta_m}(x_{np})$, the SWD is calculated as

$$\mathrm{SWD}(x_p, x_{np}) = \frac{1}{M}\sum_j^M \sum_i^N c\left(\mathcal{R}_{\theta_j}(x_{p,\mu(i)}), \mathcal{R}_{\theta_j}(x_{np,\nu(i)})\right), \quad (3)$$

where the cost function $c(\mathbf{x},\mathbf{y}) \equiv |\mathbf{x}-\mathbf{y}|^2$ is used in the following studies.

Our method aims to train a neural network to output a set of weights $\omega \in [0,1]$, for each particle, that removes particles from pileup collisions. While the network is trained using a larger set of features, the SWD calculation in Eq. (3) is carried out using only the four-vector (p_x, p_y, p_z, E) for each particle. The weights ω are learned by minimizing $\mathrm{SWD}(x'_p, x_{np})$, with the weighted set $x'_p = \omega x_p$. By multiplying the four-vector with a single weight, we are able to rescale the magnitude of the momentum vector while preserving the direction in the calculation.

As a result, particles created from pileup collisions increase $\mathrm{SWD}(x'_p, x_{np})$ and should populate ω values closer to zero, whereas particles from the primary collision are assigned $\omega \approx 1$.

An extra term can be added to Eq. (3) to control the energy scale of the events in the sample containing pileup. To that end, a constraint on the missing transverse momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$) is introduced to the loss function as the mean square error (MSE) of the $p_{\mathrm{T}}^{\mathrm{miss}}$ values between samples:

$$\mathcal{L} = \mathrm{SWD}(x'_p, x_{np}) + \lambda \mathrm{MSE}(p_{\mathrm{T}}^{\mathrm{miss}}(x'_p), p_{\mathrm{T}}^{\mathrm{miss}}(x_{np})). \quad (4)$$

The parameter λ controls the strength of the regularization. In this paper, we present results for $\lambda = 0$

(no regularization), and for $\lambda = 10^{-3}$, resulting in the $p_{\mathrm{T}}^{\mathrm{miss}}$ constraint having the same order of magnitude as the SWD term.

We simulate proton collisions at a center-of-mass energy $\sqrt{s} = 14$ TeV, employing the PYTHIA v8.244 event generator [16,17]. The DELPHES v3.4.3pre01 detector simulation [8] is used to obtain reconstructed particles with a detector layout resembling the Phase-II upgrade of the CMS detector. The simulated physics processes include jets produced via quantum chromodynamics (QCD); the production of a top quark-antiquark pair ($\mathrm{t\bar{t}}$), where both W bosons decay leptonically; the vector boson fusion (VBF) production of a Higgs boson decaying into undetectable dark matter particles; the production of a heavy resonance (Z') decaying to $\mathrm{t\bar{t}}$, where both W bosons decay hadronically; and the production of a W boson in association with at least one jet ($\mathrm{W} + \mathrm{jets}$). The former three processes are used for training and inference, while the Z' and $\mathrm{W} + \mathrm{jets}$ processes are used to assess the performance on jet substructure and the robustness of the algorithm, respectively.

Starting from the same simulated hard interaction, we generate two event samples. The first one consists solely of particles produced during the hard interaction process. In the second sample, we add contributions from pileup before reconstructing the event. The number of pileup interactions follows a Poisson distribution with a mean of 140 to match HL-LHC conditions.

The TOTAL algorithm accomplishes the task of pileup mitigation by taking the available properties of each particle as input features. These are the four-vector $(p_{\mathrm{T}}, \eta, \phi, E)$, the impact parameters in the transverse plane and along the beam axis, the particle ID, the electric charge, and the vertex identification for charged particles. The inputs are preprocessed so that the order of magnitude of all features is 1. We use the output weights $\omega \in [0, 1]$ to rescale the four-vector of each particle in the slicing process. Per event, we consider the first 9000 particles (sorted by descending p_{T} and including zero-padding) and gather the 20 nearest neighbors in the η-ϕ plane for each particle when building the graph. The SWD in Eq. (4) is computed taking 128 random projections per collision event. Different choices of numbers of projections were tested, and changes in the results were found to be negligible when using more projections.

III. RESULTS

The network trained with $\lambda = 0$ [cf. Eq. (4)] is evaluated on samples statistically independent from the ones used during training. The per-particle weights obtained from the evaluation process are used to rescale the four-momenta of the particles in the event. We then use the FastJet software [18] to cluster two TOTAL jet collections using the anti-k_t algorithm [19] with small- and large-radius parameters of $R = 0.4$ and $R = 0.8$, respectively. The set of rescaled particles is also used to compute $p_{\mathrm{T}}^{\mathrm{miss}}$. We evaluate the network on three benchmark processes: QCD multijet production, which is ubiquitous at hadron colliders and constitutes an important background to many measurements and searches; dileptonic $\mathrm{t\bar{t}}$, which often enters analyses either as a signal or as a background process; and Z', which is enriched in events with hadronically decaying top quarks merged in large-radius jets and is well suited to study jet substructure. These processes cover different physics scenarios, namely cases where sizable $p_{\mathrm{T}}^{\mathrm{miss}}$ comes from detector and reconstruction inefficiencies, and events where genuine $p_{\mathrm{T}}^{\mathrm{miss}}$ is caused by neutrinos.

To compare the performance of different pileup mitigation algorithms, a matching in $\Delta\mathrm{R} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$ is performed between reconstructed jets and generator-level jets. The latter are obtained by clustering only the particles coming from the leading primary vertex, before any detector or reconstruction effects. The following studies are based on generator-matched jets—namely, reconstructed jets for which a generator-level jet is found within $\Delta\mathrm{R} < 0.3$.

We define the response for a given observable x as the difference between the reconstructed and generator-level values, divided by the generator-level value, $(x_{\mathrm{reco}} - x_{\mathrm{gen}})/x_{\mathrm{gen}}$. In Fig. 1, we show the distributions for different kinematic variables and the corresponding response functions in the $\mathrm{t\bar{t}}$ and Z' samples. The spread of the response is indicative of the experimental resolution of the reconstruction algorithms. We define the resolution in an observable as the spread in its response, given by

$$\frac{q_{75\%} - q_{25\%}}{2}, \tag{5}$$

where $q_{n\%}$ represents the nth percentile of the response distribution. Lower resolutions result in a better reconstruction and increased power of kinematic variables such as invariant masses or $p_{\mathrm{T}}^{\mathrm{miss}}$.

Figure 2 shows the jet resolution in the transverse momentum p_{T} (jet energy resolution, JER) of generator-matched jets, computed as functions of p_{T} and η of the matched generator-level jet, for $\mathrm{t\bar{t}}$ and QCD multijet events, respectively, and the large-radius jet resolution in τ_3/τ_2, which is a variable that provides a handle on the substructure of the jet [20], as a function of the p_{T} of the matched generator-level jet for Z' events. While truth labels for pileup particles are not always available, we investigate the differences in performance of the TOTAL algorithm with respect to a fully supervised network trained using the same strategy as the one used in [7]. Results are presented in the Appendix.

Across the entire p_{T} spectrum considered, TOTAL jets are found to have a better JER than PUPPI jets, with improvements of ∼20% at low p_{T}. The PUPPI parameters used for the comparison have been taken from [21]. We observe the convergence of all algorithms in the high-p_{T} regime, as expected due to the reduced effect of soft pileup particles in

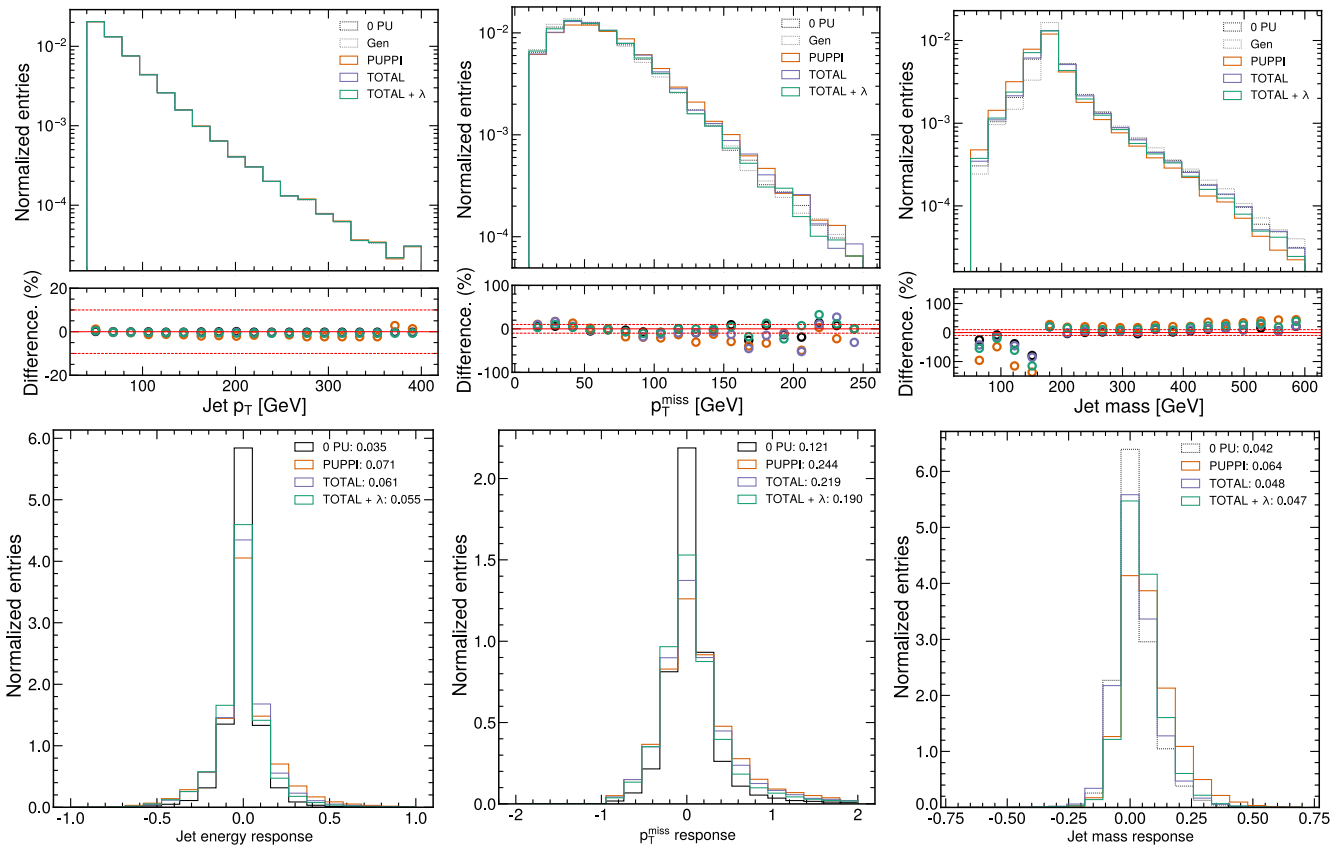

FIG. 1. Distributions of the jet p_T and missing transverse momentum p_T^{miss} for $t\bar{t}$ events, and of the jet mass for Z' events (top row); and the respective response functions (bottom row). Numbers shown in the response function represent the resolution of the observable. We compare the TOTAL algorithm with (green) and without (violet) additional energy conservation, and the PUPPI algorithm (orange), with the ideal scenario of zero pileup interactions (black), and the distribution at generator level (gray). The ratio panels display the percentage difference in resolution of different algorithms compared to generator-level events.

highly energetic jets. Similarly, we observe a better JER across all the considered η spectrum, with improvements of the order of 15% in the forward region of the detector ($|\eta| > 2.5$). The JER degrades in the forward region for all algorithms, due to a worse intrinsic performance of the instrumentation and higher levels of pileup. Finally, we observe a better performance in the τ_3/τ_2 variable with an improvement up to roughly 10%. The degradation of the

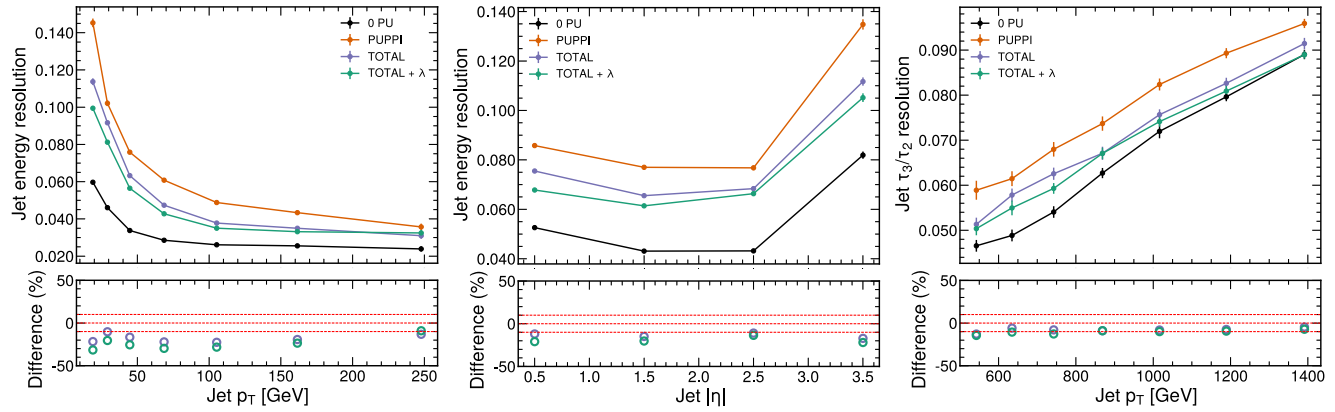

FIG. 2. JER as a function of the generator-level jet p_T for $t\bar{t}$ events (left); JER as a function of the generator-level jet η for QCD events (center); and large-radius jet τ_3/τ_2 resolution as a function of the generator-level jet p_T for $Z' \to t\bar{t}$ events (right). We compare the TOTAL algorithm with (green) and without (violet) additional energy conservation, and the PUPPI algorithm (orange) with the ideal scenario of zero pileup interactions (black). The ratio panels display the percentage difference in the resolution of TOTAL compared to PUPPI.

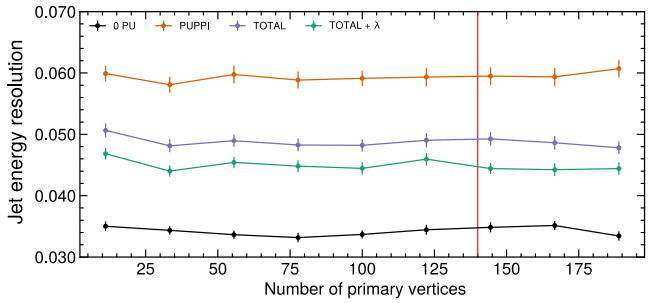

FIG. 3. Jet energy resolution as a function of the number of pileup interactions for W + jets events. We compare the TOTAL algorithm with (green) and without (violet) additional energy conservation, and the PUPPI algorithm (orange) with the ideal scenario of zero pileup interactions (black). The vertical line indicates the average number of PU collisions in the training.

τ_3/τ_2 resolution with increasing p_T is interpreted as the effect of particles becoming more collimated, resulting in a more challenging definition of the centers of the energy prongs. Finally, we observe that enforcing the energy conservation further improves the performance of the TOTAL algorithm in all observables by up to 10%.

A robust pileup mitigation algorithm is required to show stable performance with respect to various physics processes and to the number of pileup interactions in a wide range of values, since pileup conditions can change during data-taking. To check the stability of our algorithm, we evaluate our model on W + jets events generated with a uniform distribution in the number of primary vertices, NPV, ranging from 0 to 200, and compute the JER as a function of the number of pileup interactions. In Fig. 3, the performance of TOTAL is found to be stable and consistently better than PUPPI across the entire NPV spectrum. This testifies to the ability of the network to adapt to pileup scenarios and processes different from the ones experienced during training.

Finally, to illustrate the benefit TOTAL brings to searches for new physics, we study its impact in a search for invisible Higgs boson decays, which is one of the essential search channels at the LHC. Such decays are highly suppressed in the Standard Model, rendering any observation of an enhanced decay rate an unambiguous sign for new physics. Here, the Higgs boson is assumed to be produced via the fusion of two vector bosons in association with two jets close to the beam axis. A predominant background in this search is the production (mediated by the strong interaction) of a Z boson decaying into neutrinos. In Fig. 4, we demonstrate the improvement in significance $S/\sqrt{B}$, with S being the signal yield and B the background yield, as a function of a selection on a linear classifier constructed from p_T^{miss} and the dijet mass. We find the improvement from TOTAL to be of the order of 15% compared to PUPPI, consistent with the improvements in resolution for jets and p_T^{miss}. This would lead to a better sensitivity in the search for such decays and significantly improve the expected upper limit on the branching ratio of

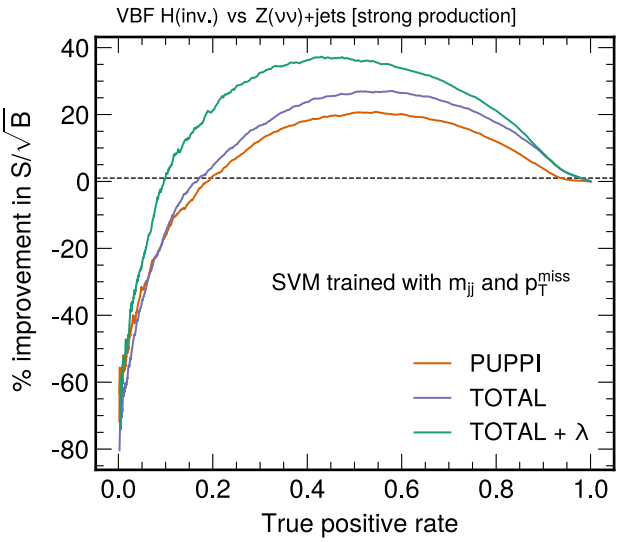

FIG. 4. A linear classifier (support vector machine, SVM) trained with p_T^{miss} and the dijet mass results in a sizable improvement in $S/\sqrt{B}$ for the TOTAL algorithms over PUPPI.

Higgs bosons to invisible particles quoted by the ATLAS and CMS Collaborations [22,23].

IV. CONCLUSIONS

We have presented a novel method to identify particles from the primary interaction and reject pileup particles in proton collisions at hadron colliders. Without relying on training labels, our self-supervised algorithm is able to morph an inclusive particle collection into a collection containing just the products of the primary interaction. Our algorithm provides a perspective for mitigating the impact of pileup at the High-Luminosity LHC, yielding an improvement over the current state of the art of up to 25% in the resolution of key observables used in searches for new physics and in precision measurements. To the best of our knowledge, this is the first application of optimal transport concepts embedded into graph neural networks to solve a highly relevant problem of current and future particle collider experiments. Not relying on per-particle truth labels, our approach can be implemented in full-scale simulations of detectors such as ATLAS and CMS. Thus, we encourage our colleagues to test this strategy. Additionally, our approach can be applied to other problems inside and outside the field of high-energy physics, where denoising is of great importance and where a realistic simulation of noise exists even if the noise distribution is intractable. For instance, it could be used for shower reconstruction in highly granular calorimeters, or for mitigating noise in time-series-based astronomical data. Future studies should focus on the expansion of the method to compare statistically independent samples with different noise levels. This would allow for a data-driven, global determination of pileup at the LHC—e.g., by comparing

distributions obtained at the beginning of a given proton fill and towards the end, when less pileup is present in the data.

ACKNOWLEDGMENTS

F. I. is supported by the Ministry of Science and Technology of China, Project No. 2018YFA0403901 and National Natural Science Foundation of China, Projects No. 12188102, and No. 12061141003. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC Award No. HEP-ERCAP0021099. This research was supported in part by the Swiss National Science Foundation (SNF) under Project No. 200020_204975/1. This project was supported by funding from the Alexander von Humboldt Foundation.

APPENDIX: COMPARISON WITH A FULLY SUPERVISED NETWORK

The TOTAL algorithm is trained using a self-supervised setting to avoid the need of labels from human annotation that cannot perfectly be obtained in full simulation or in data. However, to compare our results with a fully supervised setting using perfect labels, which are available in the simplified reconstruction of the DELPHES framework, we retrain the backbone architecture of TOTAL, ABCNet, using the same strategy presented in [7], with a regression objective that aims to learn the energy fraction of the primary collision carried by each individual particle and resulting in 0 labels for pileup interactions. Similarly, we also compare with the PUMA algorithm to check the differences in performance obtained by the use of different backbone architectures. Results for the jet energy resolution versus the jet transverse momentum for QCD events are shown in Fig. 5. We observe the resolution obtained by TOTAL to be similar to the one obtained by the supervised training of both ABCNet and PUMA, with differences in performance below 10% in all generator-level p_{T} intervals. Moreover, Fig. 5 demonstrates the performance of a truth definition obtained by matching (in angular separation and momentum difference) particles between the samples with and without pileup contributions to identify particles coming from the hard interaction. For the matching to be successful, particles have to be matched within $\Delta R = 0.04$ and have transverse momenta compatible within 10%. A successful matching of particles could then be used as truth labels for a fully supervised algorithm *à la* PUMA. However, as is visible from a comparison with even the classical benchmark PUPPI, the performance in $p_{\mathrm{T}}^{\mathrm{miss}}$ resolution is severely degraded for such a truth definition even after cumbersome manual tuning of the matching parameters, rendering the application of fully supervised algorithms an arduous if not impossible task if one wants to achieve results competitive with our self-supervised approach.

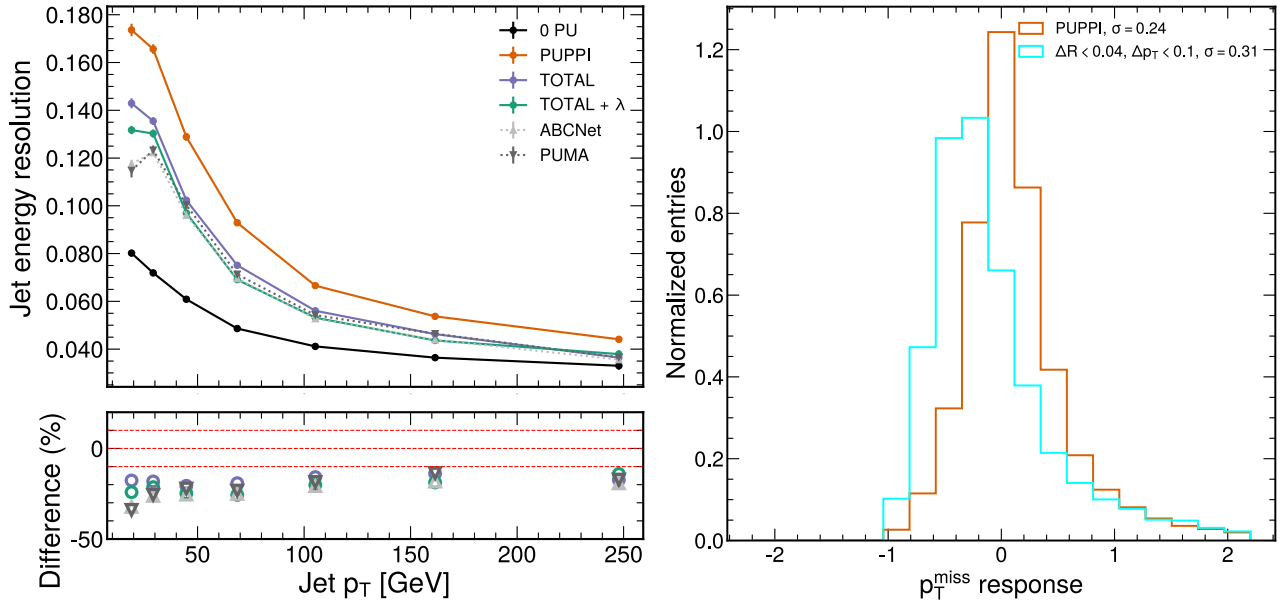

FIG. 5. Left: jet energy resolution for QCD events as a function of the generator level p_{T}. We compare the TOTAL algorithm with (green) and without (violet) additional energy conservation, and the PUPPI algorithm (orange) with the ideal scenario of zero pileup interactions (black). Additional supervised results are shown for ABCNet (light gray) and PUMA (dark gray), both using unrealistic, perfect labels. Right: comparison of $p_{\mathrm{T}}^{\mathrm{miss}}$ resolutions for PUPPI and for a human-annotation-based per-particle truth, where particles are labeled as coming from the primary interaction if they can be matched to a particle in the same event without pileup overlaid, and from pileup otherwise. The poor resolution when using truth definition renders fully supervised algorithms relying on these targets inferior to our self-supervised strategy.

[1] 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).
[2] A. M. Sirunyan *et al.* (CMS Collaboration), Particle-flow reconstruction and global event description with the CMS detector, J. Instrum. **12**, P10003 (2017).
[3] D. Bertolini, P. Harris, M. Low, and N. Tran, Pileup per particle identification, J. High Energy Phys. 10 (2014) 059.
[4] A. M. Sirunyan *et al.* (CMS Collaboration), Pileup mitigation at CMS in 13 TeV data, J. Instrum. **15**, P09018 (2020).
[5] 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.
[6] J. Arjona Martínez, O. Cerri, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Pileup mitigation at the Large Hadron Collider with graph neural networks, Eur. Phys. J. Plus **134**, 333 (2019).
[7] B. Maier, S. M. Narayanan, G. de Castro, M. Goncharov, C. Paus, and M. Schott, Pile-up mitigation using attention, Mach. Learn. Sci. Tech. **3**, 025012 (2022).
[8] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, and M. Selvaggi, DELPHES 3: A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[9] S. Agostinelli *et al.*, Geant4: A simulation toolkit, Nucl. Instrum. Methods Phys. Res., Sect. A **506**, 250 (2003).
[10] T. Li, S. Liu, Y. Feng, G. Paspalaki, N. V. Tran, M. Liu, and P. Li, Semi-supervised graph neural networks for pileup noise removal, Eur. Phys. J. C **83**, 99 (2023).
[11] G. Monge, *Mémoire sur la théorie des déblais et des remblais* (De l'Imprimerie Royale, Paris, 1781).
[12] V. Mikuni and F. Canelli, ABCNet: An attention-based method for particle tagging, Eur. Phys. J. Plus **135**, 463 (2020).
[13] C. Villani, *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften (Springer, Berlin, Heidelberg, 2008).
[14] J. Rabin, G. Peyré, J. Delon, and M. Bernot, Wasserstein barycenter and its application to texture mixing, in *Scale Space and Variational Methods in Computer Vision*, edited by A. M. Bruckstein, B. M. ter Haar Romeny, A. M. Bronstein, and M. M. Bronstein (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012), pp. 435–446.
[15] C. Lee, T. Batra, M. H. Baig, and D. Ulbricht, Sliced Wasserstein discrepancy for unsupervised domain adaptation, arXiv:1903.04064.
[16] 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).
[17] R. Corke and T. Sjostrand, Interleaved parton showers and tuning prospects, J. High Energy Phys. 03 (2011) 032.
[18] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[19] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[20] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.
[21] CMS Collaboration, Pileup removal algorithms, Technical Report No. CMS-PAS-JME-14-001, CERN, Geneva, 2014.
[22] G. Aad *et al.*, Search for invisible Higgs-boson decays in events with vector-boson fusion signatures using 139 fb^{-1} of proton-proton data recorded by the ATLAS experiment, J. High Energy Phys. 08 (2022) 104.
[23] A. Tumasyan *et al.*, Search for invisible decays of the Higgs boson produced via vector boson fusion in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. D **105**, 092007 (2022).