Holistic approach to predicting top quark kinematic properties with the covariant particle transformer

Shikai Qiu,[1,*] Shuo Han,[2,†] Xiangyang Ju,[2,‡] Benjamin Nachman,[2,3,§] and Haichen Wang[2,1,‖]

[1]*Department of Physics, University of California, Berkeley, Berkeley, California 94720, USA*
[2]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*

(Received 19 May 2022; accepted 11 May 2023; published 22 June 2023)

Precise reconstruction of top quark properties is a challenging task at the Large Hadron Collider due to combinatorial backgrounds and missing information. We introduce a physics-informed neural network architecture called the covariant particle transformer (*CPT*) for directly predicting the top quark kinematic properties from reconstructed final state objects. This approach is permutation invariant and partially Lorentz covariant and can account for a variable number of input objects. In contrast to previous machine learning-based reconstruction methods, *CPT* is able to predict top quark four-momenta regardless of the jet multiplicity in the event. Using simulations, we show that the *CPT* performs favorably compared with other machine learning top quark reconstruction approaches.

DOI: 10.1103/PhysRevD.107.114029

I. INTRODUCTION

For the Large Hadron Collider (LHC) experiments, the kinematic reconstruction of top quarks is critical to many precision tests of the Standard Model (SM) as well as direct searches for physics beyond the SM. Once produced, the top quark decays to a bottom quark (b-quark) and a W boson, with a branching ratio close to 100% [1]. Subsequently, the W boson decays into a lepton or quark pair. In the final state, quarks originating from top quark decays and other colored partons hadronize, resulting in collimated sprays of hadrons, known as jets. Conventional top quark methods assume that a hadronically decaying top quark produces three jets in the final state. Therefore, these methods are tuned to identify triplets of jets, which are considered as proxies for the three quarks originating directly from the top quark and W boson decays. The estimated top quark four-momentum is computed from the sum of measured four-momenta over the triplet of jets. Essentially, top quark reconstruction is treated as a combinatorial problem of sorting jets, and most methods use jet kinematic and flavor tagging information to construct likelihood-based [2] or machine learning-based [3–10] metrics to identify triplets of jets as proxies to top quarks and similar particles.

While the conventional top quark reconstruction approaches have been implemented in a variety of forms and extensively used at hadron collider experiments, they have fundamental flaws and shortcomings. The one-to-one correspondence between a parton (quark or gluon) and a jet, assumed by the conventional approaches, is only an approximation. Partons carry color charges but jets only consist of colorless hadrons. The formation of a jet, by construction, has to be contributed to by multiple partons. On the other hand, a single parton may contribute to the formation of multiple jets, particularly when the parton is highly energetic. In addition, triplet-based top quark reconstruction requires the presence of a certain number of jets in the final state. This jet multiplicity requirement can be inefficient because of kinematic thresholds, limited detector coverage, or the merging of highly collimated parton showers.

In this paper, we propose a new machine learning-enabled approach to determine the top quark properties through a holistic processing of the event final state. Our goal is to predict top quark four-momenta in a collision event with a given number of top quarks. The number of top quarks can itself be learned from the final state or it can be posited for a given hypothesis. As discussed earlier, the kinematic information of a top quark is not localized in a triplet of jets, rather, it is possessed by all particles in the event collectively. This motivates the

*calvin_qiu@berkeley.edu
†shuohan@lbl.gov
‡xju@lbl.gov
§bpnachman@lbl.gov
‖haichenwang@berkeley.edu

use of particle identification (ID) and kinematic information from all detectable particles in the final state as input to the determination of the top quark four-momenta. Specifically, the four-momenta and ID of all detectable final state particles are input to a deep neural networks regression model, which is constructed and trained to predict the four-momenta of a given number of top quarks. This approach offers three major advantages compared to conventional approaches. First, we no longer deal with the conceptually ill-defined jet-triplet identification process. Second, we can account for noisy or missing observations due to limited acceptance, detector inefficiency and resolution, as the regression model can learn such effects from Monte Carlo (MC) simulations. Third, the holistic processing of the event final state offers a unified approach to determining the top quark properties for both the hadronic and semileptonic top quark decays, which may simplify analysis workflows. Finally, our approach has a runtime polynomial in the number of final state objects as opposed to superexponential for standard reconstruction-based approaches which need to consider all possible permutations, making ours the first tractable method for processes with high multiplicity final state such as $t\bar{t}t\bar{t}$.

To realize the holistic approach of top quark property determination, we propose a physics-informed transformer [11] architecture termed covariant particle transformer (*CPT*).[1] *CPT* takes as input properties of the final state objects in a collision event and outputs predictions for the top quark kinematic properties. Like other recent top reconstruction proposals [7–9], *CPT* is permutation invariant under exchange of the inputs. A novel attention mechanism [11,12], referred to as covariant attention, is designed to learn the predicted kinematic properties as a function of the set of final state objects as a whole, and guarantees that the predictions transform covariantly under rotation and/or boosts of the event along the beamline. While not fully Lorentz covariant like Ref. [13], our approach captures the most important covariances relevant to hadron collider physics with minimal computational overhead and enjoys a much simpler implementation, which allows it to be easily adopted for a broad range of tasks in collider physics.

This paper is organized as follows. Section II introduces the construction and properties of *CPT*. Synthetic datasets used for demonstrating the performance of *CPT* are introduced in Sec. III. Numerical results illustrating the performance of *CPT* are presented in Sec. IV. In Sec. V, we explore what aspects of *CPT* give rise to the excellent performance. The paper ends with conclusions and outlook in Sec. VI.

II. COVARIANT PARTICLE TRANSFORMER

A. Symmetries and covariance

At the LHC, the beamline determines a special direction and reduces the relevant symmetry group of collision events from the proper orthochronous Lorentz group $\mathrm{SO}^+(1,3)$ to $\mathrm{SO}(2) \times \mathrm{SO}^+(1,1)$, which contains products of azimuthal rotations and longitudinal boosts along the beamline. The covariant particle transformer extends the original transformer architecture to properly account for these symmetry transformations, by ensuring that if the four-momenta of all final state objects undergo such a transformation, the resulting prediction of the top quark four-momenta will undergo the same transformation. At its core, this is achieved through the novel covariant attention mechanism, which modifies the standard attention mechanism to ensure that all intermediate learned features have well-defined transformation properties.

Covariance[2] under rotations and boosts [13,14] and input permutations [15] have been studied in a variety of recent high energy physics (HEP) papers. A number of additional studies have explored permutation invariant architectures [16–20] (see also other graph network approaches [21,22]). Compared to prior works in this direction, we make the following important contributions:

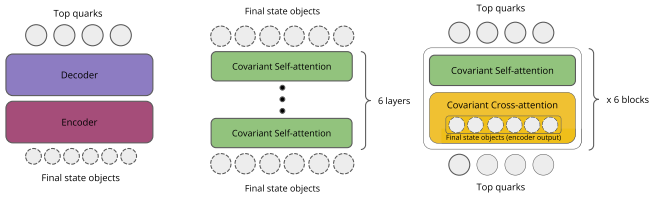
(i) We develop the first transformer architecture that enforces Lorentz covariance. Transformers are a powerful class of neural networks that have revolutionized many areas of machine learning applications, such as natural language processing [11,23], computer vision [24], and recently protein folding [25]. By integrating the transformer architecture with Lorentz covariance, *CPT* combines the current state-of-the-art of machine learning with physics-specific knowledge to become a powerful tool for applications in collider physics, as we will illustrate in this work.

(ii) We develop a simple, efficient, and effective way of achieving partial Lorentz covariance. While previous works have developed Lorentz covariant neural networks using customized architectures, they incur significant computational overhead compared to a standard neural network due to computations of continuous group convolutions [14] or irreducible representations of the Lorentz group [13]. By contrast, *CPT* only requires a simple modification to the standard attention mechanism with minimal computational overhead.

(iii) We are the first to demonstrate the benefit of using a Lorentz covariant architecture for regression problems where the targets are four-momenta of the particles. Previous works on Lorentz covariant

[1]We make our code available at https://github.com/shikaiqiu/Covariant-Particle-Transformer.

[2]This term is referred to as *equivariance* in the domain of machine learning.



(a) CPT (b) Encoder (c) Decoder

FIG. 1. An illustration of the covariant particle transformer (*CPT*) architecture. The encoder consists of six covariant self-attention layers, while the decoder consists of six covariant cross-attention layers and six covariant self-attention layers interleaved.

neural networks only evaluate on classification problems such as jet tagging where the Lorentz group acts trivially (i.e. as an identity) on the targets. There Lorentz symmetry plays a less significant role since the neural network only needs to be Lorentz invariant but not covariant.

B. Architecture

The covariant particle transformer consists of an encoder and a decoder. To ensure permutation invariance, we remove the positional encoding [11] in the original transformer encoder. The encoder produces learned features of the final state objects, which include jets, photons, electrons, muons, and missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$).[3]

Each object is represented by its transverse momentum p_{T}, rapidity y, azimuthal angle ϕ expressed as a unit vector $(\cos(\phi), \sin(\phi))$ to avoid mod π calculations, mass m, and particle identification ID. The encoder uses six covariant self-attention layers to update the feature vectors of the final state objects. The decoder uses 12 covariant attention layers to produce learned features of the top quarks. Six of these layers use self-attention, which updates the feature vector of each top quark as a function of itself and the feature vectors of other top quarks, and the other six layers use cross attention, which updates the feature vector of each top quark as a function of itself and the feature vectors of the final state objects. Finally, the feature vectors of top quarks are converted to predicted physics variables, which are the top quark four-momenta expressed in transverse momentum p_{T}, rapidity y, azimuthal angle unit vector, and mass m. Figure 1 illustrates the architecture of the covariant particle transformer. Detailed descriptions of input featurization, *CPT* architecture, and the covariant attention mechanism are provided in Appendix A.

C. Loss function

The model is trained to minimize a supervised learning objective that measures the distance between the true and predicted values of the target variables.[4] Auxiliary losses are included to stabilize training the model. We provide a detailed description of the loss function in Appendix A 6.

III. DATASETS

We use MadGraph@NLO (v2.3.7) [27] to generate pp collision events at next-to-leading order (NLO) in QCD. The decays of top quarks and W bosons are performed by MadSpin [28]. We generate 9.2×10^6 $t\bar{t}H$ events, 5.4×10^6 $t\bar{t}t\bar{t}$ events, 1.3×10^6 $t\bar{t}$ events, 1.3×10^6 $t\bar{t}W$ events, and 1×10^6 $t\bar{t}H$ events with a CP-odd top-Yukawa coupling ($t\bar{t}H_{CP\text{-odd}}$). In our generation, Higgs bosons decay through the diphoton channel for simplicity and all other objects such as top quarks and W bosons decay inclusively. The Higgs characterization model [29] is used to generate the $t\bar{t}H_{CP\text{-odd}}$ events. The generated events are interfaced with the PYTHIA 8.235 [30] for parton shower. We do not emulate detector effects as the salient features of the problem already present from the parton shower and hadronization. The generated hadrons are used to construct anti-k_t [31] $R = 0.4$ jets using FastJet 3.3.2 [32,33].

Jets are required to have $|y| \leq 2.5$ and $p_T \geq 25$ GeV, while leptons are required to have $|y| \leq 2.5$ and $p_T \geq 10$ GeV. A jet is removed if its distance[5] in ΔR with a photon or a lepton is less than 0.4. Jets that are ΔR matched to b quarks at the parton level are labeled as b jets; this label is removed randomly for 30% of the b jets, to mimic the inefficiency of a realistic b tagging [34,35]. We further apply a preselection on the testing set of $N_{\mathrm{bjet}} > 0$, and ($N_{\mathrm{jet}} \geq 3$ and $N_{\mathrm{lepton}} = 0$) or $N_{\mathrm{lepton}} > 0$, to mimic realistic data analysis requirements. The $t\bar{t}H$ and $t\bar{t}t\bar{t}$ samples are each divided to training, validation, and testing sets, corresponding to a split of 75%:12.5%:12.5%. The other samples ($t\bar{t}$, $t\bar{t}W$, and $t\bar{t}H_{CP\text{-odd}}$) are used only for testing. While a single model can be trained to learn from a mixture of processes such as $t\bar{t}H$ and $t\bar{t}t\bar{t}$ for greater generality, we leave this exciting direction to future work.

As we compare the performance of *CPT* to that of a conventional approach, we refer to top quarks that can be matched to a triplet of jets as "truth matched" and those that cannot as "unmatched." Specifically, a top quark is considered as truth matched if it decays hadronically and each of the three quarks originating from its decay is matched ($\Delta R < 0.4$) to exactly one jet. According to this definition, semileptonically decaying tops are always unmatched, which is motivated by the fact that we cannot physically detect its neutrino (at best we can estimate its kinematics such as p_{T}). The vast majority (e.g., 76% for $t\bar{t}H$) of tops

[3] $E_{\mathrm{T}}^{\mathrm{miss}}$ is implemented as a massless particle with zero longitudinal momentum component.

[4] Note that learning the true value from reconstructed quantities introduces a prior dependence [26]. This is true for nearly all regression approaches in HEP.

[5] ΔR is defined as $\sqrt{\Delta y^2 + \Delta\phi^2}$, where Δy is the difference of two particles in pseudorapidity and $\Delta\phi$ is the difference in azimuthal angle.

are unmatched, and therefore cannot be fully reconstructed due to incomplete information about their decay products. For events passing the preselection, the fraction of hadronically decaying top quarks that can be truth matched is 36% for $t\bar{t}H$, 37% for $t\bar{t}$, 38% for $t\bar{t}W$, and 38% for $t\bar{t}t\bar{t}$.

IV. PERFORMANCE

We study three different performance aspects of *CPT*. First, we evaluate the resolution of the predictions of individual top quark kinematic variables. Second, we compare the correlation between the predicted variables to the correlations between the true top quark properties. Finally, we assess the model dependence of *CPT* by applying the model trained on $t\bar{t}H$ events to alternative processes. We study these metrics inclusively for events passing the preselection, and we also break down the performance for top quarks where a matching triplet of jets can be identified using truth information and for top quarks where no matching triplet of jets can be identified. For the former case, we also compare *CPT* prediction with the calculation from a triplet-based reconstruction method. The latter scenario corresponds to the case where the conventional triplet-based reconstruction method does not apply.

A. Resolution

Figure 2 shows the predicted and truth variable distributions for p_{T}, y, ϕ of the top quarks in the $t\bar{t}H$ sample. To quantify the resolution, we calculate the width of $\Delta p_{\mathrm{T}}/p_{T,\mathrm{truth}}$, Δy and $\Delta\phi$, the model's prediction error for the three variables (relative error for p_{T}). The width is quantified using half of the 68% interquantile range, which corresponds to 1 standard deviation in the Gaussian case. The top quark mass is part of the four-momentum prediction, but we do not show it here as it is nearly a delta function. Since the model predicts the four-momenta of two top quarks, the predicted top quarks are matched to truth top quarks during the resolution calculation to minimize the sum of ΔR between all matched pairs. Table I summarizes the prediction resolutions for all top quarks in the predicted $t\bar{t}H$ events, separated into truth matched top quarks and unmatched top quarks. As expected, *CPT*'s performance is worse for unmatched tops due to incomplete information. Over all tops (truth matched and unmatched) in the test set, the median values of $\Delta p_{\mathrm{T}}/p_{\mathrm{T}}$, Δy, and $\Delta\phi$, are -0.02, 0.002 and -0.002, showing that there is no significant statistical bias in *CPT*'s prediction.

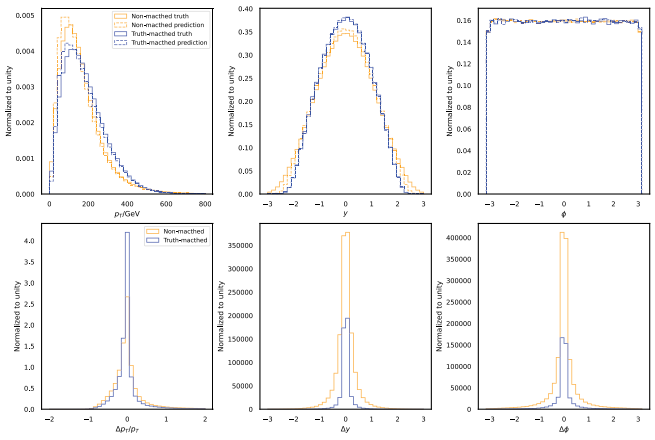


FIG. 2. Top row: distributions of truth and predicted top quark four-momentum components, p_{T}, y, and ϕ from the $t\bar{t}H$ sample. Bottom row: the distributions of dimensionless errors $\Delta p_{\mathrm{T}}/p_{\mathrm{T}}$, Δy, and $\Delta\phi$, where Δ means prediction minus truth. The area under each histogram is normalized to unity. As expected, *CPT*'s performance is worse for unmatched tops due to incomplete information. Over all tops (truth matched and unmatched) in the test set, the median values of $\Delta p_{\mathrm{T}}/p_{\mathrm{T}}$, Δy, and $\Delta\phi$, are -0.02, 0.002 and -0.002, showing that there is no significant bias in *CPT*'s prediction.

B. Relative performance

The model prediction resolutions are compared to the intrinsic resolutions of reconstructing top quarks using jet triplets. The intrinsic resolutions are calculated from truth matched triplets of jets, where the four-momenta of the truth matched jet triplet are considered as the predictions. In this case, the resolution arises from the effects of quark hadronization and jet reconstruction. For truth matched top quarks, the ratio of the prediction resolution from *CPT* to the intrinsic resolution is 1.5 for p_{T}, 2.3 for the rapidity y, and 2.0 for the azimuthal angle ϕ.

To compare *CPT* with a strong baseline, we also evaluate a triplet-based reconstruction method, where a neural network is trained to identify the triplet associated with each top quark. The baseline resolutions have prediction-to-intrinsic ratios of 2.2 for p_{T}, 2.8 for y, and 3.1 for ϕ. Therefore, even when evaluated on truth matched top quarks, *CPT* achieves significantly better resolution than the triplet-based method. The comparison is visualized in Fig. 3. Details on the baseline implementation is available in Appendix B.

TABLE I. Summary of resolutions of top quark four-momentum components in various scenarios for $t\bar{t}H$, $t\bar{t}$ and $t\bar{t}W$ processes.

		$\sigma_{p_{\mathrm{T}}}$	σ_y	σ_ϕ
$t\bar{t}H$	Intrinsic	0.10	0.04	0.07
	Truth matched	0.15	0.09	0.14
	Unmatched	0.27	0.25	0.26
$t\bar{t}$	Intrinsic	0.11	0.04	0.09
	Truth matched	0.19	0.11	0.20
	Unmatched	0.31	0.32	0.37
$t\bar{t}W$	Intrinsic	0.12	0.04	0.08
	Truth matched	0.27	0.15	0.28
	Unmatched	0.45	0.36	0.50

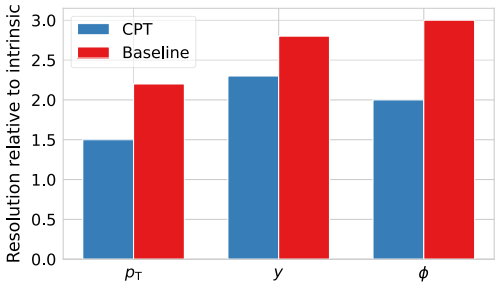


FIG. 3. Resolution (smaller means better) achieved by *CPT* and the triplet-based reconstruction (baseline) normalized by the intrinsic resolution arising from effects of quark hadronization and jet reconstruction, evaluated on truth matched tops in $t\bar{t}H$ events. *CPT* achieves significantly better resolution than the reconstruction-based approach.

In the preselected $t\bar{t}H$ events, 76% of the top quarks are unmatched. Specifically, 43% out of the total 67% of tops that decay hadronically do not have a matching triplet and 33% of all tops decay semileptonically. For these unmatched top quarks, *CPT* achieves a prediction-to-intrinsic resolution ratio of 2.5 for p_T, 6.5 for y, and 3.6 for ϕ. Because of incomplete information about the tops' decay products, *CPT*'s performance degrades as expected for unmatched top quarks, though the absolute resolutions remain below 30%. Note these top quarks cannot otherwise be fully reconstructed using reconstruction-based alternatives due to incomplete information about their decay products. While there exist procedures to approximately recover some of the missing information, such as the neutrino kinematics, combining these additional estimators with a reconstruction-based method to handle unmatched tops introduces additional complexity and sources of error and it is highly unlikely that the resulting approach will outperform a regression model.

C. Correlation

Between the six variables of interest, only three pairs of variables have a linear correlation beyond 5% in the truth sample. These correlations are 74% for $(p_{T,1}, p_{T,2})$, 50% for (y_1, y_2), and −31% for (ϕ_1, ϕ_2). The corresponding correlations observed in the covariant particle transformer prediction are 75% for $(p_{T,1}, p_{T,2})$, 43% (y_1, y_2), and −34% for (ϕ_1, ϕ_2). The correlation between top quarks is well reproduced in *CPT*'s predictions.

D. Process dependence

We assess the process dependence of *CPT* by applying the model trained with $t\bar{t}H$ to $t\bar{t}W$, $t\bar{t}$ and $t\bar{t}H_{CP\text{-odd}}$ events, respectively. Table I compares the intrinsic and prediction resolutions between $t\bar{t}H$, $t\bar{t}W$, and $t\bar{t}$ processes. *CPT* trained exclusively on the $t\bar{t}H$ sample can be applied without any retraining to yield a similar level of performance for $t\bar{t}$ events. This level of generalization is not

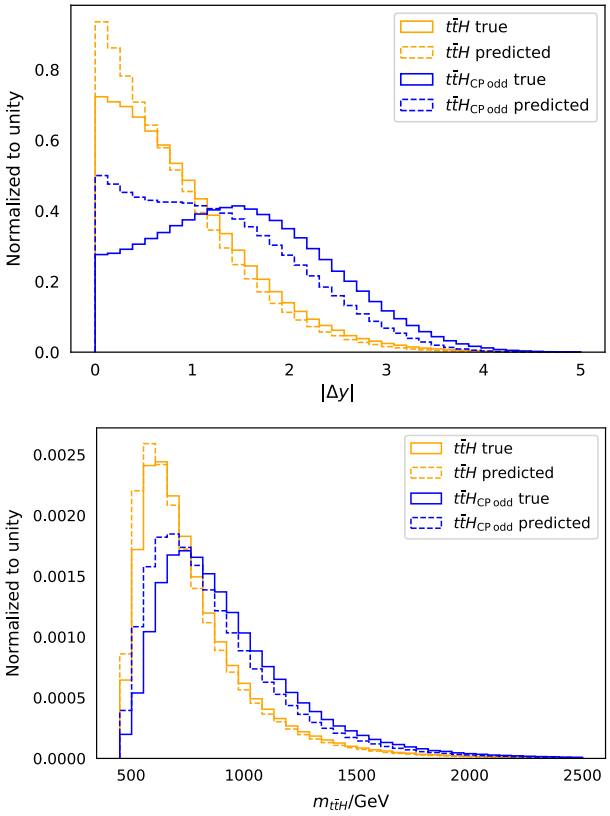


FIG. 4. Predicted and truth distributions for system-level observables $|\Delta y|$ (top) and $m_{t\bar{t}H}$ (bottom) in the $t\bar{t}H$ sample (orange) and $t\bar{t}H_{CP\text{-odd}}$ sample (blue). $|\Delta y|$ is the absolute difference between the rapidities of two tops, and $m_{t\bar{t}H}$ is the invariant mass of the $t\bar{t}H$ system, where the Higgs four-momentum is taken to be its ground-truth value. The area under each histogram is normalized to unity. As *CPT* is not trained on the $t\bar{t}H_{CP\text{-odd}}$ sample, its prediction for $t\bar{t}H_{CP\text{-odd}}$ events is worse as expected.

trivial since these two processes induce different statistics in the final state objects and top quarks. The $t\bar{t}W$ events constitute a much more challenging test set since additional jets, leptons, and neutrinos are produced from the W decay which introduces more complex correlations among the objects that are not present in *CPT*'s training set. Consequently, *CPT* yields a larger resolution on the $t\bar{t}W$ test set. The process dependence can be mitigated by a number of strategies, such as training *CPT* with a more representative sample or possibly active decorrelation strategies [36–49], which we defer to future studies. Figure 4 shows distributions of the system-level observables constructed from individual top quark four-momenta for $t\bar{t}H$ and $t\bar{t}H_{CP\text{-odd}}$ samples. A reasonable agreement between the predictions and ground-truth properties is observed for these observables, indicating *CPT* captures the subtle difference in the kinematics between the two processes and reproduces correlation in the four-momentum between the two top quarks. The agreement can be improved by applying preselection such as the

TABLE II. Summary of resolutions of top quark four-momentum components in various scenarios in the $t\bar{t}t\bar{t}$ sample.

	σ_{p_T}	σ_y	σ_ϕ
Intrinsic	0.19	0.05	0.09
Truth matched	0.29	0.16	0.24
Unmatched	0.42	0.32	0.36

requirement of at least one truth matched top. Importantly, although the model prediction is not perfect, the separation between $t\bar{t}H$ and $t\bar{t}H_{CP\text{-odd}}$ events is preserved by *CPT* predictions, showing the promise of applying *CPT* to produce discriminating kinematic variables.

E. High multiplicity final state

CPT can predict the four-momenta of an arbitrary (fixed) number of top quarks in a collision event. We test the prediction ability of *CPT* in the extreme case at the LHC where four top quarks are produced in the same event. We configure *CPT* to predict the four-momenta of four top quarks and train it with the $t\bar{t}t\bar{t}$ sample described in Sec. III. Table II shows the intrinsic and prediction resolutions from this test. Compared to the prediction for the $t\bar{t}H$ sample, the prediction for $t\bar{t}t\bar{t}$ is worse. However, the intrinsic resolution in the $t\bar{t}t\bar{t}$ sample is also worse than that in the $t\bar{t}H$ sample, suggesting that the top quarks in $t\bar{t}t\bar{t}$ events are inherently more complex and challenging to reconstruct. We expect the gap between the intrinsic and *CPT*'s resolution can be reduced by further architectural improvements and more training data. We stress that the exploding combinatorics in $t\bar{t}t\bar{t}$ events render reconstruction-based methods prohibitively expensive to be successfully applied in this setting, whereas we can easily apply *CPT* without any modification. To predict top quarks' kinematics from N jets, a standard reconstruction-based method has a super-exponential computational complexity of $O(N!)$, the number of all possible permutations within N objects, while *CPT* only has a polynomial complexity of $O(N^2)$ since the attention mechanism only involves pairwise interactions among the objects.

V. ABLATION STUDIES

We demonstrate the effects of removing important components of *CPT* to show how they contribute to the final performance. All comparisons are done on the $t\bar{t}H$ dataset. Resolutions are reported on all top quarks passing the preselection, regardless of truth-matching status.

A. Attention mechanism

The attention mechanism is an important part of the model as it allows the model to selectively focus on a subset of the final state objects in determining the four-momentum of each top quark. We demonstrate its benefit by training an otherwise identical model except with all attention weights set to a constant $\frac{1}{N_{\text{in}}}$, where N_{in} is the number of final state objects in the event. Comparisons between the resolution achieved by this model and the nominal model is shown in Table III. We observe the model with uniform attention achieves worse resolutions, which demonstrates the benefit of the attention mechanism.

TABLE III. Comparison of resolutions of top quark four-momentum components in the $t\bar{t}H$ sample achieved by *CPT* and its variant applying uniform attention for each final state object.

	σ_{p_T}	σ_y	σ_ϕ
CPT	0.24	0.21	0.23
CPT (uniform attention)	0.27	0.23	0.28

B. Covariant attention

CPT employs a covariant attention mechanism to exploit the symmetries in collision data. When the covariant attention is replaced by a regular attention mechanism which does not guarantee covariance, we observe increasing degradation in performance as the size of the training sample becomes smaller. Figure 5 compares the resolutions achieved by *CPT* and its variant using a regular attention mechanism, as a function of the number of training events. For example, the increase in p_T resolution can be as large as 16% when only 0.1% of the events in the nominal training sample is used. This shows that the covariant attention enables *CPT* to be more data efficient and provide more accurate predictions in the low-data regime compared to noncovariant models.

C. Alternative architectures

Finally, we compare with two alternative permutation-invariant architectures, graph convolutional networks [50]

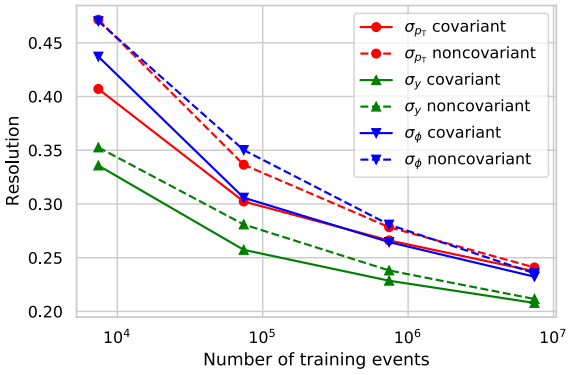


FIG. 5. Resolution on in the $t\bar{t}H$ sample achieved by using the covariant attention and noncovariant attention. The covariant attention offers clear benefit particularly in the low-data regime.

TABLE IV. Comparison of resolutions of top quark four-momentum components in the $t\bar{t}H$ sample achieved by *CPT*, GCN, and DeepSets.

	σ_{p_T}	σ_y	σ_ϕ
CPT	0.24	0.21	0.23
GCN	0.38	0.35	0.42
DeepSets	0.36	0.32	0.36

and DeepSets [51]. Applied to this task, graph convolutional networks (GCNs) use graph convolutions to process information in the final state objects represented as a complete graph, while DeepSets uses a fully connected neural network encoder to learn the feature vector of each final state object individually. In both cases, the feature vectors of all final state objects are then summed and fed into a fully connected neural network to predict the top quark four-momenta. The covariant particle transformer mainly differs from these two architectures by utilizing an attention mechanism, implementing partial Lorentz covariance, and using a decoder module. We use six graph convolutional layers and six encoder layers for the GCN and the DeepSet models, and a feature dimension of 128 for both. A comparison of resolutions between the models is shown in Table IV. *CPT* significantly outperforms the other two methods, showing its outstanding effectiveness on this task. We did not perform extensive hyperparameter optimizations for any of the three architectures. However, we hypothesize that the performance ordering would persist after such an optimization given the magnitude of the observed differences. We defer this study to future work.

VI. CONCLUSION

In this paper, we propose a new machine learning-enabled approach to determining top quark kinematic properties by processing the full event information holistically. Our approach offers three major advantages compared to conventional approaches. First, we no longer deal with the conceptually ill-defined jet-triplet identification process. Second, we can account for noisy or missing observations due to limited detector acceptance, inefficiency, and resolution, as the regression model can learn such effects from simulations. Third, the holistic processing of the event final state offers a unified approach to determine the top quark properties for both the hadronic and semileptonic top quark decays, which simplifies the analysis workflow. Finally, our approach has a run-time polynomial in the number of final state objects as opposed to superexponential for reconstruction-based approaches which need to consider all possible permutations, making ours the first tractable method for processes with high multiplicity final state such as $t\bar{t}t\bar{t}$.

To realize this holistic approach to predicting top quark kinematic properties, we propose the covariant particle transformer (*CPT*). *CPT* takes as input properties of the final state objects in a collision event and outputs predictions for the top quark kinematic properties. Using a novel covariant attention mechanism, *CPT* prediction is invariant under permutation of the inputs and covariant under rotation and/or boosts of the event along the beamline. *CPT* can recover 76% (75%) of the top quarks produced in the $t\bar{t}H$ ($t\bar{t}t\bar{t}$) events that cannot be truth matched to a jet triplet and thus not fully reconstructable by conventional methods. For $t\bar{t}H$ events, *CPT* achieves a resolution close to the intrinsic resolution of jet triplet and outperforms a carefully tuned triplet-based top reconstruction method on top quarks that can be matched to a jet triplet. In addition, we demonstrate that *CPT* can generalize to top production processes not seen during training, though its performance degrades as the test process becomes more complex and distinct from the training process. Finally, we demonstrate that by building Lorentz covariance into *CPT*, it achieves higher data efficiency and outperforms the noncovariant alternative when the training set is small.

In the future, it may be possible to improve and extend *CPT*. *CPT* training uses simulation to learn to invert parton shower and hadronization (and in the future, detector effects). Training strategies that rely less on parton shower and hadronization simulations like those in Ref. [52] may be able to improve the robustness of *CPT*. Furthermore, as a direct regression approach, *CPT* is prior dependent. A variety of domain adaptation and other strategies may be able to further improve the resilience of *CPT*. It may also be possible to include lower-level, higher-dimensional inputs directly into *CPT* instead of first clustering jets.

As it uses a generic representation for collision events as sets of particles, *CPT* can be directly applied to predict kinematic properties of other heavy decaying particles, such as the W, Z, and Higgs boson, and potential heavy particles beyond the SM. The predicted kinematics of these heavy decaying particles can be used to construct discriminating variables for searches or observables for differential cross-section measurements. The ability to predict properties of heavy decaying particles through a holistic analysis of the collision event can enable measurements that otherwise suffer extreme inefficiencies using conventional reconstruction methods.

ACKNOWLEDGMENTS

This work is supported by the U.S. Department of Energy, Office of Science under Contract No. DE-AC02-05CH11231. The work of H. W. is partly supported by the U.S. National Science Foundation under the Award No. 2046280.

Note added.—Recently, we became aware of Ref. [53], which proposes another Lorentz equivariant architecture.

In contrast to that paper, we integrate Lorentz covariance with the transformer, a state-of-the-art neural network architecture that revolutionized many areas of machine learning applications such as natural language processing, computer vision, and protein folding. We have also considered a completely different application: namely regression instead of classification, where the Lorentz group acts nontrivially (not an identity) on the target variables.

APPENDIX A: CPT IMPLEMENTATION

1. Attention mechanism

Attention mechanisms are a way to update a vector of n features $\{x_i\}_{i=1}^n$, given a context $\{c_j\}_{j=1}^m$. Learnable query, key, and value matrices $\{W_Q, W_K, W_V\}$ are used to generate d-dimensional query, key, and value vectors $\{q_i\}_{i=1}^n$, $\{k_j\}_{j=1}^m$, and $\{v_j\}_{j=1}^m$, via

$$q_i = W_Q x_i \tag{A1}$$

$$k_j = W_K c_j, \tag{A2}$$

$$v_j = W_V c_j. \tag{A3}$$

The inner product between $q_i^\top$ and k_j is used to compute the attention weights α_{ij} through

$$\alpha_{ij} = \frac{\exp(q_i^\top k_j/\sqrt{d})}{\sum_j \exp(q_i^\top k_j/\sqrt{d})}, \tag{A4}$$

where the $\sqrt{d}$ is a normalization factor. A weighted sum of the value vectors are then used to compute update vectors $\{m_i\}_{i=1}^n$,

$$m_i = \sum_j \alpha_{ij} v_j, \tag{A5}$$

which is then used to update x_i by, for example, addition $x_i' = x_i + m_i$. Intuitively, the attention weights α_{ij} represent how important the information contained in c_j is to x_i. When the context $\{c_j\}$ is simply $\{x_i\}$, this is termed as self-attention, otherwise cross attention. It is common to use a slight extension of the method above, called multiheaded attention, where H different query, key, and value matrices $\{(W_Q^h, W_K^h, W_V^h)\}_{h=1}^H$ are learned. Each head follows the above procedure to independently produce attention weights $\{a_{ij}^h\}_{ijh}$ and then update vectors $\{m_i^h\}_{i=1,h=1}^{n,H}$. The H update vectors $\{m_i^h\}_{h=1}^H$ received by each x_i are concatenated to produce a final update vector,

$$m_i = \bigoplus_{h=1}^H m_i^h, \tag{A6}$$

which is then used to update x_i as before.

2. Particle representation

We represent each particle with a feature vector h_i, and $h_i = (x_i, \omega_i)$ consists of an invariant feature vector x_i, and a covariant feature vector ω_i. x_i is an invariant quantity under a rotation and boost along the beamline, while $\omega_i = (y_i, \cos(\phi_i), \sin(\phi_i))$ represents the flight direction of the object and is a covariant quantity. As input to the covariant particle transformer, $x_i = (p_{\mathrm{T},i}, m_i, \mathrm{id})$ where id is a one-hot vector indicating particle identity. The model learns to update these feature vectors while maintaining their invariance/covariance property through the covariant attention.

3. Covariant attention

To update the learned feature vectors of each object in the event, we use covariant attention, an extension of the regular attention mechanism to process kinematics information and guarantee covariance properties of the predictions. In general, covariant attention updates feature vectors $\{h_i\}$ of a subset of the objects in the event using feature vectors $\{h_j\}$ of a (potentially different) subset as context. First, it computes the flight direction of each context object as viewed in i's frame: $\omega_{ij} = (y_j - y_i, \cos(\phi_j - \phi_i), \sin(\phi_j - \phi_i)$, which is invariant under longitudinal boosts and azimuthal rotations. Then it computes the d-dimensional query, keys, and value vectors as follows:

$$\hat{x}_i = \mathrm{LayerNorm}(x_i), \tag{A7}$$

$$v_{ij} = W_V(\hat{x}_j + \mathrm{MLP}(\omega_{ij})), \tag{A8}$$

$$k_{ij} = W_K(\hat{x}_j + \mathrm{MLP}(\omega_{ij})), \tag{A9}$$

$$q_i = W_Q \hat{x}_i, \tag{A10}$$

where W_V, W_K, W_Q are learned matrices and MLP is a multilayer perceptron. The inner products between q_i and k_{ij} are then sent through a softmax operator so as to weight the value vectors. The weighted sum produces an aggregated message vector m_i^x which is added to x_i:

$$\alpha_{ij} = \frac{\exp(q_i^\top k_{ij}/\sqrt{d})}{\sum_j \exp(q_i^\top k_{ij}/\sqrt{d})}, \tag{A11}$$

$$\tilde{m}_i^x = \sum_j \alpha_{ij} v_{ij}, \tag{A12}$$

$$m_i^x = \sigma(\mathrm{Linear}(x_i, \tilde{m}_i^x)) \odot \tilde{m}_i^x, \tag{A13}$$

$$x_i' = x_i + m_i^x, \tag{A14}$$

where σ is the sigmoid function and $\odot$ denotes elementwise (Hadamard) product. Gating is applied to the attention weights following the gated attention network [54]. A multiheaded version of covariant attention can be constructed in the same way as in regular attention, and is omitted here. x_i' is then passed through a feed-forward network as done in the original transformer. When it is desirable to also update the covariant feature ω_i, we produce another update vector m_i^ω from m_i^x via

$$\tilde{m}_i^\omega = \mathrm{MLP}(m_i^x) \tag{A15}$$

$$m_i^\omega = \sigma(\mathrm{Linear}(x_i, \tilde{m}_i^\omega)) \odot \tilde{m}_i^\omega, \tag{A16}$$

where m_i^ω is a three-dimensional vector. Its first component is used as a boost with rapidity δy_i, while its last two components v_i converted to a rotation matrix $R(v_i)$, which is used to rotate the azimuthal angle ϕ_i:

$$y_i' = y_i + \delta y_i \tag{A17}$$

$$\begin{pmatrix} \cos(\phi_i') \\ \sin(\phi_i') \end{pmatrix} = R(v_i) \begin{pmatrix} \cos(\phi_i) \\ \sin(\phi_i) \end{pmatrix} \tag{A18}$$

$$\omega_i' = (y_i', \cos(\phi_i'), \sin(\phi_i')), \tag{A19}$$

where $R(v_i)$ is obtained as follows:

$$u_i = v_i + (1, 0), \tag{A20}$$

$$w_i = \frac{u_i}{\|u_i\|} = (\cos(\theta_i), \sin(\theta_i)), \tag{A21}$$

$$R(v_i) = \begin{pmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{pmatrix}, \tag{A22}$$

where we added $(1, 0)$ to v_i to bias the rotation matrix to an identity for stability. The covariance of $\{\omega_i'\}$ follows from the fact that only invariant information is used to construct its update, and prior to the update, $\{\omega_i\}$ are themselves covariant. An inductive argument establishes the end-to-end covariance of compositions of covariant attention updates. We denote the above covariant attention update as $h_i \leftarrow \mathcal{A}_{x\omega}^{x\omega}(h_i, \{h_j\})$ where the subscript indicates that it makes use of both the invariant and covariant feature vector, and the superscript indicates that it updates both the invariant and covariant feature vector. The following variants are used to build the full model:

(i) $x_i \leftarrow \mathcal{A}_{x\omega}^{x}(h_i, \{h_j\})$: the covariant feature vector is not updated.

(ii) $x_i \leftarrow \mathcal{A}_{x}^{x}(x_i, \{x_j\})$: the covariant feature vector is not updated nor used to construct the key and value vectors. This reduces to the regular attention mechanism.

4. Encoder

The encoder uses six layers of covariant attention to update the input invariant features $x_i^{\mathrm{in}} \leftarrow \mathcal{A}_{x\omega}^{x}(h_i^{\mathrm{in}}, \{h_j^{\mathrm{in}}\})$. The covariant features associated with the input objects $\{\omega_i^{\mathrm{in}}\}$ are not updated.

5. Decoder

a. Initialization

The decoder first initializes the invariant feature vectors associated with the top quarks using the Set2Set module [55], which takes in the set $\{x_i^{\mathrm{in}}\}$ and outputs $\{x_i^{\mathrm{out}}\}$, the initial invariant feature vectors of the output objects. The decoder then updates $\{x_i^{\mathrm{out}}\}$ by having each output attend to the input objects, using invariant features only, $x_i^{\mathrm{out}} \leftarrow \mathcal{A}_{x}^{x}(x_i^{\mathrm{out}}, \{x_j^{\mathrm{in}}\})$. The attention weights α_{ij} computed in the previous attention update are used to initialize the output covariant feature vectors:

$$y_i^{\mathrm{out}} = \sum_j \alpha_{ij} y_j^{\mathrm{in}}, \tag{A23}$$

$$\begin{pmatrix} \cos(\phi_i^{\mathrm{out}}) \\ \sin(\phi_i^{\mathrm{out}}) \end{pmatrix} = \frac{\sum_j \alpha_{ij} \begin{pmatrix} \cos(\phi_j^{\mathrm{in}}) \\ \sin(\phi_j^{\mathrm{in}}) \end{pmatrix}}{\left\| \sum_j \alpha_{ij} \begin{pmatrix} \cos(\phi_j^{\mathrm{in}}) \\ \sin(\phi_j^{\mathrm{in}}) \end{pmatrix} \right\|}. \tag{A24}$$

The covariance of y_i^{out} follows from the fact that $\sum_j \alpha_{ij} = 1$, and $\{y_j^{\mathrm{in}}\}$ transforms by an overall additive constant under a boost. The covariance of ϕ_i^{out} follows from the fact that its unit vector representation is a linear combination of $\{\binom{\cos(\phi_j^{\mathrm{in}})}{\sin(\phi_j^{\mathrm{in}})}\}_j$, each of which transform linearly by a rotation.

b. Interleaved covariant cross attention and self-attention

After initialization, the decoder consists of $L_{\mathrm{out}} = 6$ decoder blocks. In each block, the output invariant and covariant feature vectors are updated using two covariant attention layers:

$$h_i^{\mathrm{out}} \leftarrow \mathcal{A}_{x\omega}^{x\omega}(h_i^{\mathrm{out}}\{h_j^{\mathrm{out}}\}) \quad \forall\ i, \tag{A25}$$

$$h_i^{\mathrm{out}} \leftarrow \mathcal{A}_{x\omega}^{x\omega}(h_i^{\mathrm{out}}, \{h_j^{\mathrm{in}}\}) \quad \forall\ i. \tag{A26}$$

After each decoder block, indexed by $\ell \in \{1, \ldots, L_{\mathrm{out}}\}$, an intermediate set of predictions $\{p_i^\ell\}_i$ for the top quark four momenta is constructed as follows:

$$\begin{aligned} &(p_{T_i}^\ell/\mathrm{GeV}, y_i^\ell, \phi_i^\ell, m_i^\ell/\mathrm{GeV}) \\ &\quad = (100(x_i^\ell)_0, y_i, \phi_i, 5(x_i^\ell)_1 + 173), \end{aligned} \tag{A27}$$

where $(x_i^\ell)_0, (x_i^\ell)_1$ denotes the first and second entry of the invariant feature vector associated with each top at the ℓth block. The shift and scaling is to keep the feature vectors small and centered to facilitate training.

6. Loss function and optimization details

For each event, the main component of loss function is the L_2 norm of the difference between the model prediction and ground truth for the top quark four-momenta in $(p_x/100\text{ GeV}, p_y/100\text{ GeV}, y, m/5\text{ GeV})$ coordinates, averaged over the N top quarks present in the event:

$$\mathcal{L}_{\text{final}} = \frac{1}{N}\sum_{i=1}^{N} \|p_i - p_i^*\|, \tag{A28}$$

where $\{p_i\}$ are the model predictions at the final decoding block and $\{p_i^*\}$ are the ground truths. We chose this set of coordinates so that each component of the four-momenta has standard deviation of $O(1)$, encouraging the model to pay equal attention to each of them. The N predictions from the model are matched to the N ground truths through a permutation π^* that minimizes the average ΔR between each matched pair:

$$\pi^* = \underset{\pi:\text{permutations}}{\text{argmin}} \frac{1}{N}\sum_{i=1}^{N} \sqrt{(y_i - y_i^*)^2 + (\phi_i - \phi_i^*)^2}. \tag{A29}$$

We add two auxiliary losses $\mathcal{L}_{\text{intermediate}}$ and $\mathcal{L}_{\text{unit-norm}}$ to stabilize training models with many layers. The intermediate loss $\mathcal{L}_{\text{intermediate}}$ measures the intermediate prediction errors at earlier decoder blocks,

$$\mathcal{L}_{\text{intermediate}} = \frac{1}{L_{\text{out}} - 1}\sum_{\ell=1}^{L_{\text{out}}-1}\left(\frac{1}{N}\sum_{i=1}^{N}\|p_i^\ell - p_i^*\|\right), \tag{A30}$$

where $\{p_i^\ell\}_{i=1}^N$ are intermediate predictions at the ℓth decoder. The unit-norm loss $\mathcal{L}_{\text{unit-norm}}$ encourages the vectors u_i to have unit-norm before being normalized and converted to rotation matrices in Eq. (A20) in each decoder block:

$$\mathcal{L}_{\text{unit-norm}} = \frac{1}{L_{\text{out}}}\sum_{\ell=1}^{L_{\text{out}}}\left(\frac{1}{N}\sum_{i=1}^{N}|\|u_i^\ell\| - 1|\right). \tag{A31}$$

The two auxiliary losses are inspired by similar auxiliary losses in AlphaFold2 [25]. The total loss is a weighted combination of the above three terms,

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{final}} + \lambda_2\mathcal{L}_{\text{intermediate}} + \lambda_3\mathcal{L}_{\text{unit-norm}}. \tag{A32}$$

We use $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 0.02$. All models used to report our results are trained using the Lamb optimizer [56] with a batch size of 256 and a learning rate of 10^{-4} for 30 epochs and 10^{-5} for another ten epochs. A weight decay [57] of 0.01 is applied. The model from the epoch achieving minimum validation loss is used for final evaluation. This training protocol is sufficient to saturate validation performance for all variants of the model and datasets of various processes and sizes used to present our results.

APPENDIX B: BASELINE

We train a neural network to identify triplets of jets that originate from top decays. This task can be formulated as a link prediction problem on a graph, where the nodes are detected jets in an event and any two jets that belong to a triplet are connected by a link. Specifically, every event is represented by a fully connected graph using the four-momenta and particle types as node features and a graph neural network (GNN) predicts a probability $p_{ij} \in (0, 1)$ that a link exists between jet i and jet j for every pair of jets in the event. The particular architecture we use is the interaction network [58], followed by an MLP applied per edge to output the per-edge probabilities. The GNN is trained to minimize the cross-entropy loss so that p_{ij} is encouraged to be 1 if the jets belong to the same triplet and 0 otherwise. It uses the same training, validation, and test set as used by *CPT*. We tune the hyperparameters to maximize validation accuracy and settled on four interaction network blocks, two layers and 128 hidden units for all MLPs, Adam optimizer, and a learning rate of 0.001. At test time, we sort all possible links (i, j) by decreasing order in p_{ij} and sequentially form one or two predicted triplets depending on the number of available jets in the event. Each predicted top four-vector is the system four-vector of the predicted triplet. The predicted tops are ΔR matched to the true tops following the same procedure in *CPT* defined in Eq. (A29). We note that this method provides a strong baseline as it uses a neural network architecture that has demonstrated state-of-the-art performance on reasoning about object and relations in a wide range of complex problems such as N-body dynamics and estimating physical quantities [58].

[1] Particle Data Group, Review of particle physics, Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).

[2] J. Erdmann, S. Guindon, K. Kroeninger, B. Lemmer, O. Nackenhorst, A. Quadt, and P. Stolte, A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework, Nucl. Instrum. Methods Phys. Res., Sect. A **748**, 18 (2014).

[3] M. Aaboud *et al.* (ATLAS Collaboration), Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Rev. D **97**, 072016 (2018).

[4] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the $t\bar{t}b\bar{b}$ production cross section in the all-jet final state in pp collisions at $\sqrt{s} = 13$ TeV, Phys. Lett. B **803**, 135285 (2020).

[5] J. Erdmann, T. Kallage, K. Kröninger, and O. Nackenhorst, From the bottom to the top—Reconstruction of $t\bar{t}$ events with deep learning, J. Instrum. **14**, P11015 (2019).

[6] G. Aad *et al.* (ATLAS Collaboration), *CP* Properties of Higgs Boson Interactions with Top Quarks in the $t\bar{t}H$ and tH Processes Using $H \to \gamma\gamma$ with the ATLAS Detector, Phys. Rev. Lett. **125**, 061802 (2020).

[7] M. J. Fenton, A. Shmakov, T.-W. Ho, S.-C. Hsu, D. Whiteson, and P. Baldi, Permutationless many-jet event reconstruction with symmetry preserving attention networks, Phys. Rev. D **105**, 112008 (2022).

[8] J. S. H. Lee, I. Park, I. J. Watson, and S. Yang, Zero-permutation jet-parton assignment using a self-attention network, arXiv:2012.03542.

[9] A. Shmakov, M. J. Fenton, T.-W. Ho, S.-C. Hsu, D. Whiteson, and P. Baldi, SPANet: Generalized permutationless set assignment for particle physics using symmetry preserving attention, SciPost Phys. **12**, 178 (2022).

[10] A. Badea, W. J. Fawcett, J. Huth, T. J. Khoo, R. Poggi, and L. Lee, Solving combinatorial problems at particle colliders using machine learning, Phys. Rev. D **106**, 016001 (2022).

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems* (2017), p. 5998, arXiv:1706.03762.

[12] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473.

[13] A. Bogatskiy, B. Anderson, J. T. Offermann, M. Roussi, D. W. Miller, and R. Kondor, Lorentz group equivariant neural network for particle physics, arXiv:2006.04780.

[14] C. Shimmin, Particle convolution for high energy physics, arXiv:2107.02908.

[15] M. J. Dolan and A. Ore, Equivariant energy flow networks for jet tagging, Phys. Rev. D **103**, 074022 (2021).

[16] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.

[17] H. Qu and L. Gouskos, ParticleNet: Jet tagging via particle clouds, Phys. Rev. D **101**, 056019 (2020).

[18] E. A. Moreno, O. Cerri, J. M. Duarte, H. B. Newman, T. Q. Nguyen, A. Periwal, M. Pierini, A. Serikova, M. Spiropulu, and J.-R. Vlimant, JEDI-net: A jet identification algorithm based on interaction networks, Eur. Phys. J. C **80**, 58 (2020).

[19] V. Mikuni and F. Canelli, ABCNet: An attention-based method for particle tagging, Eur. Phys. J. Plus **135**, 463 (2020).

[20] V. Mikuni and F. Canelli, Point cloud transformers applied to collider physics, Mach. Learn. Sci. Tech. **2**, 035027 (2021).

[21] J. Shlomi, P. Battaglia, and J.-R. Vlimant, Graph neural networks in particle physics, Mach. Learn. **2**, 021001 (2020).

[22] M. Feickert and B. Nachman, A living review of machine learning for particle physics, arXiv:2102.02770.

[23] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, Language models are few-shot learners, Adv. Neural Inf. Process. Syst. **33**, 1877 (2020), https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, An image is worth 16×16 words: Transformers for image recognition at scale, arXiv:2010.11929.

[25] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with AlphaFold, Nature (London) **596**, 583 (2021).

[26] ATLAS Collaboration, Generalized numerical inversion: A neural network approach to jet calibration, Technical Report No. ATL-PHYS-PUB-2018-013, 2018.

[27] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.

[28] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations, J. High Energy Phys. 03 (2013) 015.

[29] P. Artoisenet *et al.*, A framework for Higgs characterisation, J. High Energy Phys. 11 (2013) 043.

[30] 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).

[31] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

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

[33] M. Cacciari and G. P. Salam, Dispelling the N^3 myth for the k_t jet-finder, Phys. Lett. B **641**, 57 (2006).

[34] G. Aad *et al.* (ATLAS Collaboration), ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV, Eur. Phys. J. C **79**, 970 (2019).

[35] A. M. Sirunyan *et al.* (CMS Collaboration), Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV, J. Instrum. **13**, P05011 (2018).

[36] G. Louppe, M. Kagan, and K. Cranmer, Learning to pivot with adversarial networks, Adv. Neural Inf. Process. Syst. **30**, 981 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/hash/48ab2f9b45957ab574cf005eb8a76760-Abstract.html.

[37] J. Stevens and M. Williams, uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers, J. Instrum. **8**, P12013 (2013).

[38] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Søgaard, Decorrelated jet substructure tagging using adversarial neural networks, Phys. Rev. D **96**, 074034 (2017).

[39] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, Mass agnostic jet taggers, SciPost Phys. **8**, 011 (2020).

[40] Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, Report No. ATL-PHYS-PUB-2018-014, 2018.

[41] G. Kasieczka and D. Shih, DisCo Fever: Robust Networks Through Distance Correlation, Phys. Rev. Lett. **125**, 122001 (2020).

[42] L.-G. Xia, QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics, Nucl. Instrum. Methods Phys. Res., Sect. A **930**, 15 (2019).

[43] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Machine learning uncertainties with adversarial neural networks, Eur. Phys. J. C **79**, 4 (2019).

[44] S. Wunsch, S. Jórger, R. Wolf, and G. Quast, Reducing the dependence of the neural network function to systematic uncertainties in the input space, Comput. Software Big Sci. **4**, 5 (2019).

[45] A. Rogozhnikov, A. Bukva, V. V. Gligorov, A. Ustyuzhanin, and M. Williams, New approaches for boosting to uniformity, J. Instrum. **10**, T03002 (2014).

[46] CMS Collaboration, A deep neural network to search for new long-lived particles decaying to jets, Mach. Learn. **1**, 035012 (2020).

[47] J. M. Clavijo, P. Glaysher, and J. M. Katzy, Adversarial domain adaptation to reduce sample bias of a high energy physics classifier, Mach. Learn. Sci. Tech. **3**, 015014 (2022).

[48] O. Kitouni, B. Nachman, C. Weisser, and M. Williams, Enhancing searches for resonances with machine learning and moment decomposition, J. High Energy Phys. 04 (2021) 070.

[49] M. J. Dolan and A. Ore, Meta-learning and data augmentation for mass-generalised jet taggers, Phys. Rev. D **105**, 094030 (2022).

[50] T. N. Kipf and M. Welling, in Semi-supervised classification with graph convolutional networks, *Proceedings of the 5th International Conference on Learning Representations* (2016), arXiv:1609.02907.

[51] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola, in Deep sets, *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), pp. 3394–3404, arXiv:1703.06114.

[52] J. N. Howard, S. Mandt, D. Whiteson, and Y. Yang, Foundations of a fast, data-driven, machine-learned simulator, Sci. Rep. **12**, 7567 (2022).

[53] S. Gong, Q. Meng, J. Zhang, H. Qu, C. Li, S. Qian, W. Du, Z.-M. Ma, and T.-Y. Liu, An efficient Lorentz equivariant graph neural network for jet tagging, J. High Energy Phys. 07 (2022) 030.

[54] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung, in Gaan: Gated attention networks for learning on large and spatiotemporal graphs, *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (2018), arXiv:1803.07294.

[55] O. Vinyals, S. Bengio, and M. Kudlur, Order matters: Sequence to sequence for sets, arXiv:1511.06391.

[56] Y. You, J. Li, S. J. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, in Large batch optimization for deep learning: Training bert in 76 minutes, *Proceedings of ICLR* (2020), arXiv:1904.00962.

[57] I. Loshchilov and F. Hutter, in Decoupled weight decay regularization, *Proceedings of ICLR* (2019), arXiv:1711.05101.

[58] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, and K. Kavukcuoglu, Interaction networks for learning about objects, relations and physics, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), https://proceedings.neurips.cc/paper/2016/hash/3147da8ab4a0437c15ef51a5cc7f2dc4-Abstract.html.