A 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, §] and Haichen Wang[2, 1, ¶]

[1] *Department of Physics, University of California, Berkeley, Berkeley, CA 94720, USA*

[2] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

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.

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 to a lepton or quark pair. In the final state, quarks originating from top quark decays and other colored partons hadronize and result in collimated sprays of hadrons, known as jets. The conventional top quark methods target top quark hadronic decays and assume that the hadronic top quark decay would result in three jets in the final state. Therefore, those conventional methods are tuned to identify triplets of jets, which are considered as a proxy to the top quark decay products, out of a number of jets in the final state. The top quark four momentum is computed from the measured four-momenta of the triplet of jets. Essentially, the top quark reconstruction is treated as a process of sorting combinatorics, and most methods use jet kinematic and flavor tagging information to construct likelihood-based [2] or machine learning-based [3–9] metrics to identify triplets of jets as proxies to top quarks.

While the conventional top quark reconstruction approach has been implemented in a variety of forms and extensively used at hadron collider experiments, it has fundamental flaws and shortcomings. The one-to-one correspondence between a parton (quark or gluon) and a jet, assumed by the conventional approach, 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 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, and the merging of highly collimated jets.

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, which can be determined upstream in the analysis or an analysis assumption. 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 use of the particle identification (ID) and kinematic information from all detectable particles in the event final state as input to the determination of top quark four-momenta. Technically, the kinematics and ID of all detectable final state particles can be input to a deep neural networks based regression model, which is constructed and trained to predict the four momenta of a given number of top quarks. This new approach offers three major advantages compared to the conventional approach. First, we no longer deal with the conceptually ill-defined jet-triplet identification process. Second, we can recover information lost 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 determine the top quark properties for both the hadronic and semi-leptonic top quark decays, which simplifies the analysis workflow.

To realize the holistic approach of top quark property determination, we propose a physics-informed Transformer [10] architecture termed Covariant Particle Transformer (CPT). The CPT takes as input properties of the final state objects in a collision event and outputs predictions for the top quark kinematics. Like other recent top

[*] calvin_qiu@berkeley.edu
[†] shuohan@lbl.gov
[‡] xju@lbl.gov
[§] bpnachman@lbl.gov
[¶] haichenwang@berkeley.edu

reconstruction proposals [6, 7, 9], the CPT is permutation invariant under exchange of the inputs. A novel attention mechanism [10, 11], referred to as covariant attention, is used to learn the predicted kinematics 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. [12], our approach captures important covariances relevant for hadron collider physics.

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

II. COVARIANT PARTICLE TRANSFORMER

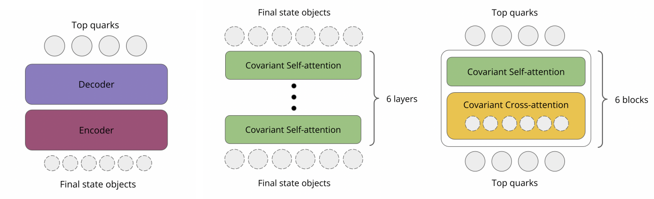
A. Symmetries and covariance

In 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 covariant attention mechanism, which modifies the regular attention mechanism to ensure that all intermediate learned features have well-defined transformation properties.

While the first integration with a Transformer architecture, covariance under rotations/boosts [12, 13] and input permutations [14] have been studied in a variety of recent High Energy Physics (HEP) papers. A number of additional studies have explored permutation invariant architectures [15–19] (see also other graph network approaches [20]).

B. Architecture

Covariant Particle Transformer consists of an encoder and a decoder. To ensure permutation invariance, we remove the positional encoding [10] 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 ($E_{\mathrm{T}}^{\mathrm{miss}}$) [1]. Each object is represented by its transverse momentum p_{T}, rapidity y, azimuthal angle ϕ expressed as a unit vector $(\cos(\phi), \sin(\phi))$, 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 attentionlayers to produce learned features of the top quarks. 6 of such 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 6 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 ϕ as a unit vector, and mass m. Figure 1 illustrates the architecture of the Covariant Particle Transformer. Detailed descriptions of the architecture and the covariant attention mechanism is provided in the Appendix.



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

FIG. 1: The Covariant Particle Transformer architecture illustration. The encoder consists of 6 covariant self-attention layers, while the decoder consists of 6 covariant cross-attention layers and 6 covariant self-attention layers interleaved.

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[2]. Auxiliary losses are included to stabilize training the model. We provide detailed descriptions of the loss function in the Appendix.

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

[2] Note that learning the true value from reconstructed quantities introduces a prior dependence [21]. This is true for nearly all regression approaches in HEP.

III. DATASETS

We use Madgraph@NLO (v2.3.7) [22] to generate collision events at NLO in QCD. The decays of top quarks and W bosons are implemented by MadSpin [23]. 9.2 million $t\bar{t}H$ events, 10 million $t\bar{t}$ events, 10 million $t\bar{t}W$ events, and 5.4 million $t\bar{t}t\bar{t}$ events are generated. The generated events are passed through a parton showering process modeled by Pythia8.X [24]. We do not introduce any detector effects to the collision events. The truth hadrons are used to construct anti-k_t [25] $R = 0.4$ jets using FastJet 3.X [26, 27].

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, ΔR [3], with a photon or a lepton is within 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 Afterwards, each dataset is split at random into 75% training, 12.5% validation, and 12.5% testing. We further apply a preselection on the testing set of $N_{\text{bjet}} \geq 0$, and ($N_{\text{jet}} \geq 3$ and $N_{\text{lepton}} = 0$), or $N_{\text{lepton}} \geq 0$.

As we compare the performance of Covariant Particle Transformer to that of the conventional approach, we separate top quarks that can be matched to a triplet of jets and top quarks that cannot. A top quark is considered as "truth-matched" if each of the three quarks originating from its decay can be uniquely matched to a jet using a criterion of their distance is less than 0.4. 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 the 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 that between the truth values of target variables. Finally, we assess the model dependence of the CPT by applying the model trained on the $t\bar{t}H$ process 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 the CPT prediction with the calculation from the triplet-based reconstruction method. The latter scenario corresponds to the case where the conventional triplet-based reconstruction method does not apply.

Resolution: Figure IV shows the predicted and truth variable distributions for p_{T}, y, ϕ of the top quark in the $t\bar{t}H$ sample. To quantify the resolution, we calculate the width of the relative $p_{\text{T}}$and absolute y and ϕ response. The width is quantified using half of the 68% inter-quantile range, which corresponds to one standard deviation in the case of a Gaussian resolution. The top quark mass is part of the 4-vector prediction, but we do not show it here as it has nearly no spread. This is because there is little to no information in the final state objects about the off-shell top quark mass and therefore the network predicts the average value.

Table I summarizes the prediction resolutions for all top quarks in the predicted $t\bar{t}H$ events, "truth-matched" top quarks, and "not-matched" top quarks. The result is also compared to that from a triplet-based top reconstruction method, following a procedure similar to that used in Ref. [8].

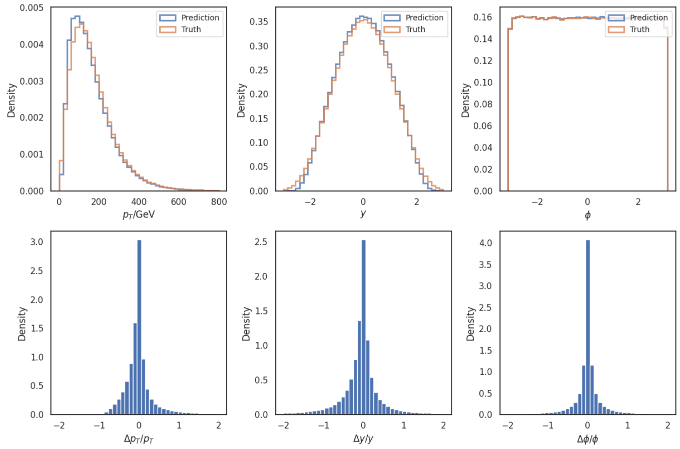


FIG. 2: Distributions of truth and predicted top quark four-momentum components, p_{T}, y, and ϕ (top row) from the $t\bar{t}H$ sample. The distributions of the pull, defined as $(x_{\text{prediction}} - x_{\text{truth}})/x_{\text{truth}}$ with x being the variable of interest, are also shown (bottom row). The area under each histogram is normalized to unity. The mean values of the pull distributions are consistent with zero, indicating that there is no significant bias in the prediction.

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

	$\sigma_{p_{\text{T}}}$	σ_y	σ_ϕ
Intrinsic	0.10	0.04	0.07
Triplet-based reconstruction	0.22	0.11	0.22
Truth-matched	0.15	0.09	0.14
Non-matched	0.27	0.25	0.26

[3] 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.

We calculate the resolution for variables based on triplets of jets that are truth-matched to top quarks. The model prediction resolution is compared to the intrinsic resolution of reconstructing top quarks using jet-triplets. The intrinsic resolutions are calculated from truth-matched triplets of jets, where the four momentum of the truth-matched jet-triplet is considered as the prediction. In this case, the resolution arises from the effects of quark hadronization and jet reconstruction. For the same sample of truth-matched top quarks, the ratio of the prediction resolution from the 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. By comparison, we also present the resolution of top quark property predicted from the triplet-based reconstruction method. The prediction-to-intrinsic resolution ratio is 2.2 for p_T, 2.8 for y, and 3.1 for ϕ. Therefore, for truth-matched top quarks, the CPT achieves significantly better resolution than the triplet-based method.

In the preselected $t\bar{t}H$ events, 76% of the top quarks cannot be matched to any triplet of jets. which include those that decay through the semi-leptonic channel. For these non-matched top quarks, the CPT achieves a prediction-to-intrinsic resolution ratio of 2.5 for p_T, 6.5 for y, and 3.6 for ϕ. While the performance degrades in the case of non-matched top quarks, these top quarks otherwise cannot be reconstructed using the triplet-based top reconstruction method.

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 ($p_{\mathrm{T},1}$,$p_{\mathrm{T},2}$), 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 ($p_{\mathrm{T},1}$,$p_{\mathrm{T},2}$), 43% (y_1,y_2), and −34% for (ϕ_1,ϕ_2). The correlation between top quarks is well reproduced in the Covariant Particle Transformer predictions.

Process dependence: We assess the process dependence of the Covariant Particle Transformer by applying the model on $t\bar{t}H$ to $t\bar{t}W$ and $t\bar{t}$ events. Table II compares the intrinsic and prediction resolutions between the $t\bar{t}H$, $t\bar{t}W$, and $t\bar{t}$ samples. As expected, a moderate level of process dependence is observed. The process-dependence can be mitigated by a number of strategies, such as training the Covariant Particle Transformer with a more representative sample, which is beyond the scope of this study. Figure 3 shows distributions of $t\bar{t}$-system observables constructed from individual top quark four-momenta for $t\bar{t}H$ and $t\bar{t}$ samples. A reasonable agreement between the prediction and truth is observed for these $t\bar{t}$-system observables, indicating the correlation in the four-momentum between the two top quarks is well reproduced by the Covariant Particle Transformer. The level of agreement between the prediction and truth is also similar for $t\bar{t}H$ and $t\bar{t}$ samples, and the separation between $t\bar{t}H$ and $t\bar{t}$ events is preserved in the prediction.

TABLE II: Summary of resolutions of top quark four-momentum components in various scenarios.

		σ_{p_T}	σ_y	σ_ϕ
$t\bar{t}H$	Intrinsic	0.10	0.04	0.07
	Truth-matched	0.15	0.09	0.14
	Non-matched	0.27	0.25	0.26
$t\bar{t}W$	Intrinsic	0.12	0.04	0.08
	Truth-matched	0.27	0.15	0.28
	Non-matched	0.45	0.36	0.50
$t\bar{t}$	Intrinsic	0.11	0.04	0.09
	Truth-matched	0.19	0.11	0.20
	Non-matched	0.31	0.32	0.37

High multiplicity final state: The Covariant Particle Transformer can predict the four-momenta of an arbitrary number of top quarks in a collision event. We test the prediction ability of the CPT in the extreme case at the LHC where four top quarks are produced in the same event. We configure the CPT to predict the four-momenta of four top quarks and train it with the $t\bar{t}t\bar{t}$ sample described in Section III. Table III 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 slightly worse.

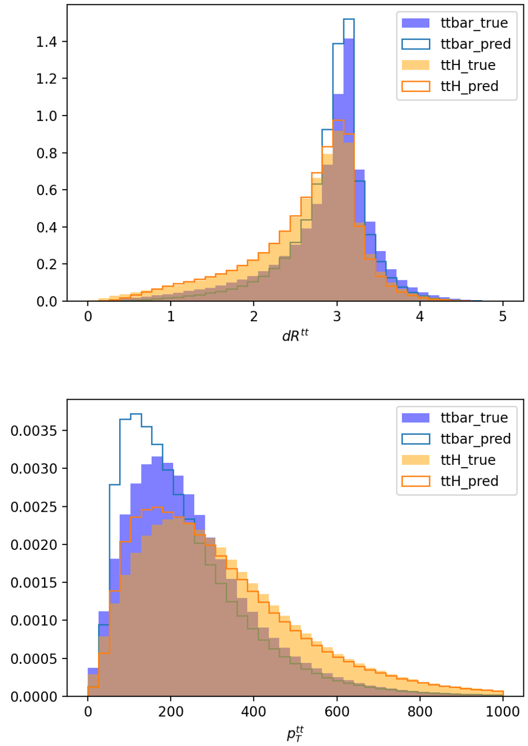


FIG. 3: Predicted and truth distributions for $t\bar{t}$-system observables $\Delta R_{t\bar{t}}$ (top) and $p_{\mathrm{T},t\bar{t}}$ (bottom) in the $t\bar{t}H$ sample (orange) and $t\bar{t}$ sample (blue). The area under each histogram is normalized to unity.

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 $t\bar{t}t\bar{t}$ events are more complex and challenging for the CPT to make predictions.

TABLE III: 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
Not truth-matched	0.42	0.32	0.36

V. ABLATION STUDIES

We demonstrate the effects of removing important parts of our model to show how individual components 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.

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 $\frac{1}{N_\mathrm{in}}$, 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 IV. We observe the model with uniform attention achieves worse resolutions, which demonstrates the benefit of the attention mechanism.

TABLE IV: 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

Covariant attention: The 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 degradation in performance that increases as the size of the training sample becomes smaller. Figure 4 compares the resolutions achieved by the 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 the CPT to be more data-efficient and provide more accurate predictions in the low-data regime compared to non-covariant models.

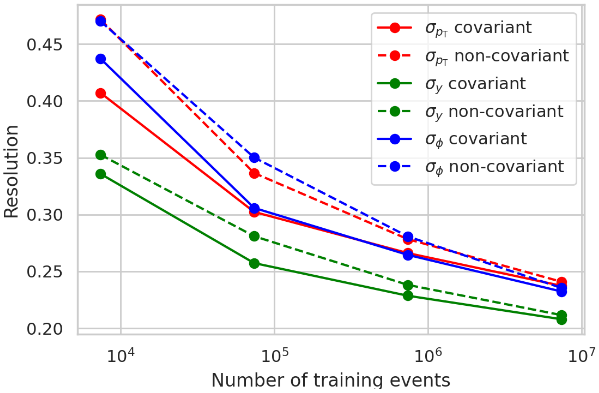


FIG. 4: Resolution on in $t\bar{t}H$ sample achieved by using the covariant attention and non-covariant attention. The covariant attention offers clear benefit particularly in the low-data regime.

Alternative architectures Finally, we compare with two alternative permutation-invariant architectures, Graph Convolutional Networks[28] and DeepSets [29]. 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 use an MLP 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 an MLP to predict the top quark four-momenta. The Covariant Particle Transformer mainly differ from these two architectures by utilizing an attention mechanism, implementing partial Lorentz covariance, and using a decoder module. We use 6 graph convolutional layers and 6 MLP encoder layers for the GCN and the DeepSet model, and a feature dimension of 128 for both. Comparison between the resolutions achieved by these models is shown in Table V. The 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 as we expect the same conclusion would continue to hold given the size of the gap.

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

	$\sigma_{p_{\mathrm{T}}}$	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

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 the conventional approach. First, we no longer deal with the conceptually ill-defined jet-triplet identification process. Second, we can recover information lost due to limited acceptance, detector 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 semi-leptonic top quark decays, which simplifies the analysis workflow.

To realize this holistic approach to predicting top quark kinematic properties, we propose the Covariant Particle Transformer (CPT). The CPT takes as input properties of the final state objects in a collision event and outputs predictions for the top quark kinematics. Using a novel covariant attention mechanism, its prediction is invariant under permutation of the inputs and covariant under rotation and/or boosts of the event along the beamline. The 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 discarded by the conventional triplet-based reconstruction method. For top quarks that can be matched to a jet-triplet, the CPT achieves a resolution close to the intrinsic resolution of jet-triplet and outperforms a carefully tuned triplet-based top reconstruction method. In addition, we demonstrate that the CPT only has a modest process dependence and achieves high data efficiency by utilizing symmetries of the collision data.

As it uses a generic representation for collision events as sets of particles, the Covariant Particle Transformer 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 the conventional reconstruction method.

ACKNOWLEDGEMENTS

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

Appendix A

1. Attention mechanism

The attention mechanism is a way to update a set of feature vectors $\{x_i\}_{i=1}^n$, 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 pairwise inner product between q_i and k_j used to compute the attention weights α_{ij} through

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

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 self-attention, otherwise cross-attention. It is common to use a slight extension of the method above, called Multi-headed attention, where H different query, key, and value matrices $\{(W_Q^h, W_K^h, W_V^h)\}_{h=1}^H$ are learnt. Each head follow 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 featrue vector ω_i. x_i is an invariant quantity under a rotation and boost along the beamline, while $\omega_i = (y_i, \cos(\phi_i), \sin(\phi_i))$ represnets 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 extenstion of the regular attention mechanism to process kinematics information and gaurantee 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 G-invariant. 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 an MLP. The exponentiated inner products between q_i and k_{ij} are then used to weight the value vectors. The weighted sum produces an aggregated message vetor m_i^x which is added to x_i:

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

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

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

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

where σ is the sigmoid function and $\odot$ denotes element-wise product. Gating is applied to the attention weights following the Gated Attention Networks [30]. A multi-headed 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's desirable to also update the covariant feature ω_i, we produce another update vector m_i^ω from m_i^x via

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

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

$$\tag{A17}$$

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 :

$$y_i' = y_i + \delta y_i \tag{A18}$$

$$\begin{pmatrix} \cos(\phi_i') \\ \sin(\phi_i') \end{pmatrix} = R(v_i) \begin{pmatrix} \cos(\phi_i) \\ \sin(\phi_i) \end{pmatrix} \tag{A19}$$

$$\omega_i' = (y_i', \cos(\phi_i'), \sin(\phi_i')). \tag{A20}$$

where $R(v_i)$ is obtained as follows

$$u_i = v_i + (1, 0), \tag{A21}$$

$$w_i = \frac{u_i}{\|u_i\|} = (\cos(\theta_i), \sin(\theta_i)), \tag{A22}$$

$$R(v_i) = \begin{pmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{pmatrix}, \tag{A23}$$

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:

- $x_i \leftarrow \mathcal{A}_{x\omega}^{x}(h_i, \{h_j\})$: the covariant feature vector is not updated
- $x_i \leftarrow \mathcal{A}_{x}^{x}(x_i, \{x_j\})$: 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 6 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 [31], 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 attends to the input objects, using invariant features only, $x_i^{\text{out}} \leftarrow \mathcal{A}_x^x(x_i^{\text{out}}, \{x_j^{\text{in}}\})$. The attention weights α_{ij} computed in the previous attention update is used to intialize the output covariant feature vectors:

$$y_i^{\text{out}} = \sum_j \alpha_{ij} y_j^{\text{in}}, \tag{A24}$$

$$\begin{pmatrix}\cos(\phi_i^{\text{out}})\\ \sin(\phi_i^{\text{out}})\end{pmatrix} = \frac{\sum_j \alpha_{ij}\begin{pmatrix}\cos(\phi_j^{\text{in}})\\ \sin(\phi_j^{\text{in}})\end{pmatrix}}{\left\|\sum_j \alpha_{ij}\begin{pmatrix}\cos(\phi_j^{\text{in}})\\ \sin(\phi_j^{\text{in}})\end{pmatrix}\right\|} \tag{A25}$$

The covariance of y_i^{out} follows from the fact that $\sum_j \alpha_{ij} = 1$, and $\{y_j^{\text{in}}\}$ 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 $\left\{\begin{pmatrix}\cos(\phi_j^{\text{in}})\\ \sin(\phi_j^{\text{in}})\end{pmatrix}\right\}_j$, each of which transform linearly by a rotation.

b. *Interleaved covariant cross- and self-attention*

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

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

$$h_i^{\text{out}} \leftarrow \mathcal{A}_{x\omega}^{x\omega}(h_i^{\text{out}}, \{h_j^{\text{in}}\}) \quad \forall i. \tag{A27}$$

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

$$(p_{T_i}^\ell/\text{GeV}, y_i^\ell, \phi_i^\ell, m_i^\ell/\text{GeV}) = (100(x_i^\ell)_0, y_i, \phi_i, 5(x_i^\ell)_1 + 173), \tag{A28}$$

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 ℓ-th block. The shift and scaling is to keep the feature vectors small to stabilize 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.

$$\mathcal{L}_{\text{final}} = \frac{1}{N}\sum_{i=1}^{N} \|p_i - p_i^*\|, \tag{A29}$$

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.

$$\pi^* = \underset{\pi:\text{permutations}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}\sqrt{(y_i - y_i^*)^2 + (\phi_i - \phi_i^*)^2}. \tag{A30}$$

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,

$$\mathcal{L}_{\text{intermediate}} = \frac{1}{L_{\text{out}} - 1}\sum_{\ell=1}^{L_{\text{out}}-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|p_i^\ell - p_i^*\right\|\right), \tag{A31}$$

, where $\{p_i^\ell\}_{i=1}^N$ are intermediate predictions at the $\ell-$th 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 Equation A21 in each output decoding block.

$$\mathcal{L}_{\text{unit-norm}} = \frac{1}{L_{\text{out}}}\sum_{\ell=1}^{L_{\text{out}}}\left(\frac{1}{N}\sum_{i=1}^{N}\left|\left\|u_i^\ell\right\| - 1\right|\right), \tag{A32}$$

The two auxiliary losses are inspired by similar auxiliary losses in AlphaFold2 [32]. The final loss is a weighted combination of the above three terms,

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{final}} + \lambda_2\mathcal{L}_{\text{intermediate}} + \lambda_3\mathcal{L}_{\text{unit-norm}}. \tag{A33}$$

We use $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 0.02$. All models used to report our results are trained using Lamb optimizer with a batch size of 256, learning rate of 10^{-4} for 30 epochs and 10^{-5} for 10 epochs. A weight decay of 0.01 is applied. 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.

[1] Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), 08 2020. 083C01.

[2] Johannes Erdmann, Stefan Guindon, Kevin Kroeninger, Boris Lemmer, Olaf Nackenhorst, Arnulf Quadt, and Philipp Stolte. A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework. *Nucl. Instrum. Meth. A*, 748:18–25, 2014.

[3] Morad Aaboud et al. 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(7):072016, 2018.

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

[5] Johannes Erdmann, Tim Kallage, Kevin Kröninger, and Olaf Nackenhorst. From the bottom to the top—reconstruction of $t\bar{t}$ events with deep learning. *JINST*, 14(11):P11015, 2019.

[6] Michael James Fenton, Alexander Shmakov, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, and Pierre Baldi. Permutationless Many-Jet Event Reconstruction with Symmetry Preserving Attention Networks. 10 2020.

[7] Jason Sang Hun Lee, Inkyu Park, Ian James Watson, and Seungjin Yang. Zero-Permutation Jet-Parton Assignment using a Self-Attention Network. 12 2020.

[8] Georges Aad et al. 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(6):061802, 2020.

[9] Alexander Shmakov, Michael James Fenton, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, and Pierre Baldi. SPANet: Generalized Permutationless Set Assignment for Particle Physics using Symmetry Preserving Attention. 6 2021.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

[12] Alexander Bogatskiy, Brandon Anderson, Jan T. Offermann, Marwah Roussi, David W. Miller, and Risi Kondor. Lorentz Group Equivariant Neural Network for Particle Physics. 6 2020.

[13] Chase Shimmin. Particle Convolution for High Energy Physics. 7 2021.

[14] Matthew J. Dolan and Ayodele Ore. Equivariant Energy Flow Networks for Jet Tagging. *Phys. Rev. D*, 103(7):074022, 2021.

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

[16] Huilin Qu and Loukas Gouskos. ParticleNet: Jet Tagging via Particle Clouds. *Phys. Rev. D*, 101(5):056019, 2020.

[17] Eric A. Moreno, Olmo Cerri, Javier M. Duarte, Harvey B. Newman, Thong Q. Nguyen, Avikar Periwal, Maurizio Pierini, Aidana Serikova, Maria Spiropulu, and Jean-Roch Vlimant. JEDI-net: a jet identification algorithm based on interaction networks. *Eur. Phys. J. C*, 80(1):58, 2020.

[18] Vinicius Mikuni and Florencia Canelli. ABCNet: An attention-based method for particle tagging. *Eur. Phys. J. Plus*, 135(6):463, 2020.

[19] Vinicius Mikuni and Florencia Canelli. Point cloud transformers applied to collider physics. *Mach. Learn. Sci. Tech.*, 2(3):035027, 2021.

[20] Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. Graph Neural Networks in Particle Physics. 7 2020.

[21] Generalized Numerical Inversion: A Neural Network Approach to Jet Calibration. Technical report, CERN, Geneva, Jul 2018. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/A PHYS-PUB-2018-013.

[22] 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. *JHEP*, 07:079, 2014.

[23] Pierre Artoisenet, Rikkert Frederix, Olivier Mattelaer, and Robbert Rietkerk. Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations. *JHEP*, 03:015, 2013.

[24] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

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

[26] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J. C*, 72:1896, 2012.

[27] Matteo Cacciari and Gavin P. Salam. Dispelling the N^3 myth for the k_t jet-finder. *Phys. Lett. B*, 641:57–61, 2006.

[28] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations*, 2016.

[29] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. 2017. cite arxiv:1703.06114Comment: NIPS 2017.

[30] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. *Conference on Uncertainty in Artificial Intelligence*, 2018.

[31] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets, 2016.

[32] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.