# A Holistic Approach to Predicting Top Quark Kinematic Properties with the Covariant Particle Transformer

Shuo Han, Xiangyang Ju, Benjamin Nachman, **Shikai Qiu**, and Haichen Wang

University of California, Berkeley

Nov 1, 2022

Berkeley
UNIVERSITY OF CALIFORNIA

# Acknowledgements

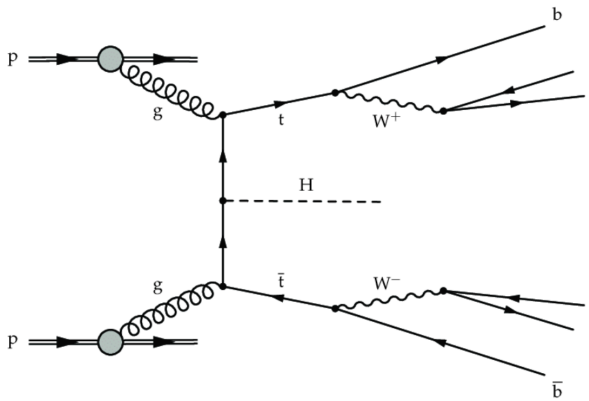

- Thanks to my collaborators in this work: Shuo Han, Xiangyang Ju, Benjamin Nachman, and Haichen Wang, for their support and insightful discussions.
- This work was done during my time at UC Berkeley. Currently, I'm a Ph.D. student in Computer Science at NYU Courant.

# Top quark kinematics prediction

- **Motivation:**
  1. Top Yukawa coupling is a uniquely important Standard Model parameter as it is the largest among all elementary fermions.
  2. Identifying and reconstructing top quarks using final state particles is critical to the success of such measurements because one can use kinematics of reconstructed top quarks as discriminating variables and therefore improve measurement sensitivity.
- **Challenge:** top Yukawa coupling induced processes (*tttt*, *tH*, *ttH*) involve cascade decays of top quarks and their final states have a high multiplicity of particles.

# The conventional approach: reconstruction

- Reconstruct the top by solving the combinatorics problem when possible, e.g:
  for a hadronically decaying top, identify a triplet of jets that originate from its decay, out of a high number of jets that are in the final state.
- ML-based methods include BDT and recently GNN, mostly dealing only with hadronic decay.



# Challenges with the reconstruction-based approach

- At the analysis level, a jet may be missing due to acceptance (too soft or too forward).
- In practice, parton-to-jet correspondence may not be one-to-one. A jet may consist of hadrons originating from more than one partons, e.g: when a top is highly boosted.
- $\rightarrow$ In *ttH*, Less than $40\%$ the hydronically decaying tops are reconstructable (have a triplet of jets $\Delta R-$matched to the top).
- In principle, parton-to-jet correspondence is ill-defined. Partons are colored, but jets consist of color-neutral hadrons. This means that a jet always receives contributions from more than one partons. So even if one can $\Delta R-$match jets to separate partons, there is a limitation in jet serving as a proxy of parton.

# Our approach

- Forget about combinatorics, just do regression:

$$\text{NeuralNetwork} : \{\text{final state objects } p_i^\mu\}_{i=1}^{N_{\text{final state obj}}} \mapsto \{\text{top quarks } p^\mu\}_{i=1}^{N_{\text{tops}}}$$

- Several advantages:
  1. Simultaneously examine all inputs (final state objects): access to event-level information
  2. Simultaneously predict all outputs (top quarks): capture and utilize correlation in the top quarks's kinematics variables
  3. Can make statistically informed predictions when faced with missing jets or leptonic decay, where a top is fundamentally not reconstructable via traditional methods.

# Exploiting the problem structure, i.e., physics

There exist non-trivial structures in both the inputs and outputs which we can use to build models with more calibrated inductive biases:

- Permutation invariance: no canonical ordering between objects in an event.
  a) Predictions should be invariant to a permutation of the input final state objects
  b) Loss function should be invariant to a permutation of the predicted tops
- Lorentz covariance:

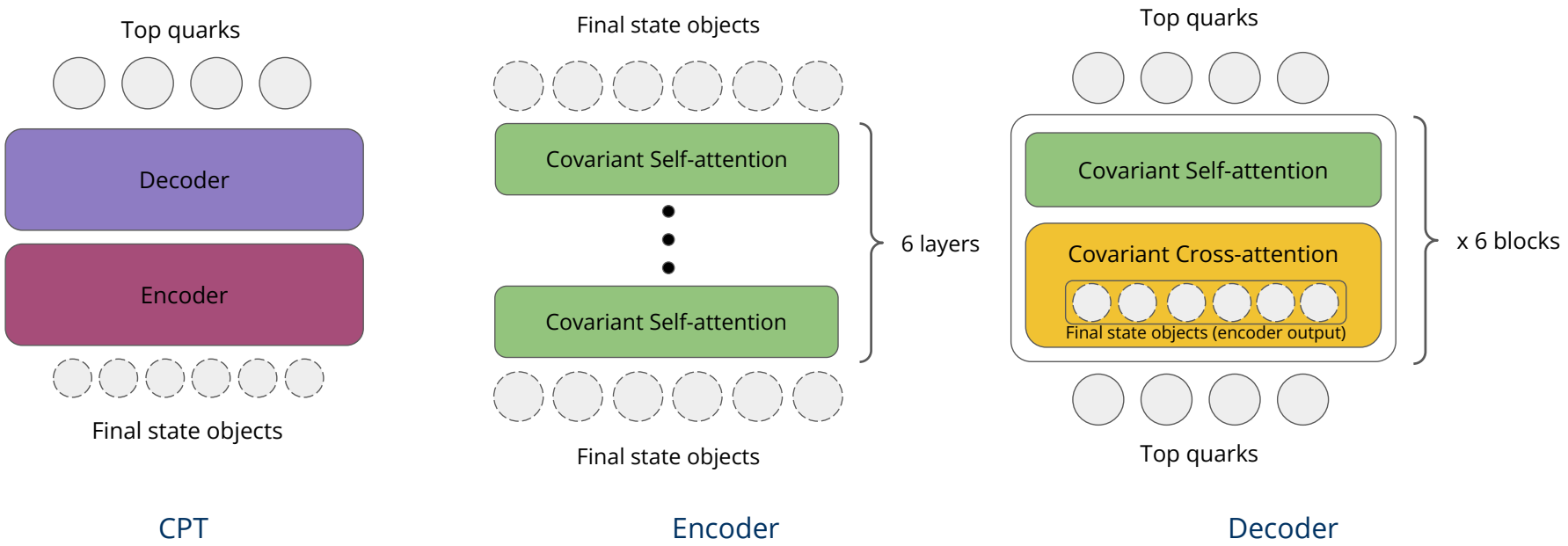
$$\text{NeuralNetwork} : \Lambda\{\text{final state objects } p_i^\mu\}_{i=1}^{N_{\text{final state obj}}} \mapsto \Lambda\{\text{top quarks } p^\mu\}_{i=1}^{N_{\text{tops}}}$$

  At the LHC, the beamline determines a special direction and reduces the relevant symmetry group of collision events from the proper orthochronous Lorentz group $SO^+(1,3)$ to $SO(2) \times SO^+(1,1)$, which contains products of azimuthal rotations and longitudinal boosts along the beamline. Our model only needs to be covariant w.r.t. this subset of transformations.

# Architecture: Covariant Particle Transformer

- We introduce the **Covariant Particle Transformer (CPT)**, a general architecture for high energy physics applications performing set-to-set predictions, satisfying Lorentz covariance.
- Based on the Transformer architecture, the model has two components:
  1. Encoder: process information in the final state objects
  2. Decoder: make predictions for top 4-vectors and iteratively improve the predictions
- We achieve Lorentz covariance through the covariant attention mechanism, which modifies the regular attention mechanism to ensure that all intermediate learned features have well-defined transformation properties (they are either Lorentz scalars or 4-vectors).

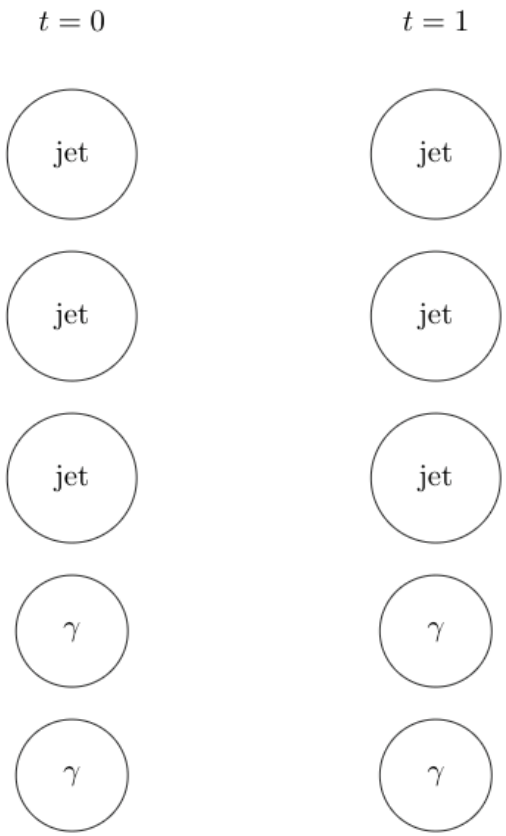
# Diagrammatic representation of CPT




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.

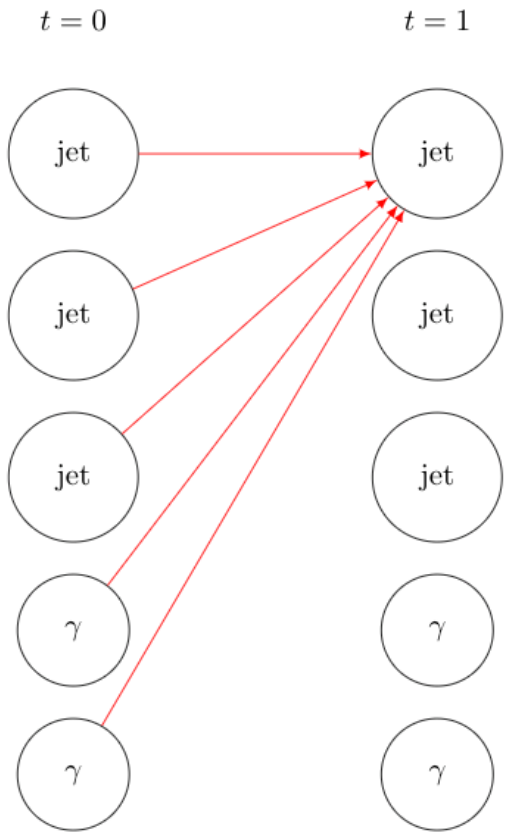
# Encoder

Stacked covariant self-attention layers, producing learned features of each object.
$t = 0$ : features are the 4-vectors of the objects and their identity (jet, photon, ...)

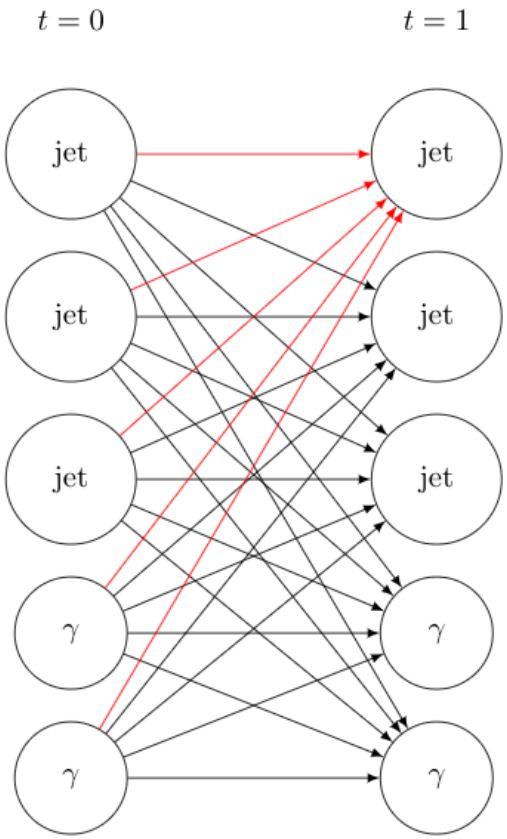


# Encoder

In each layer, update each final state object's learned feature vector with other objects as input.

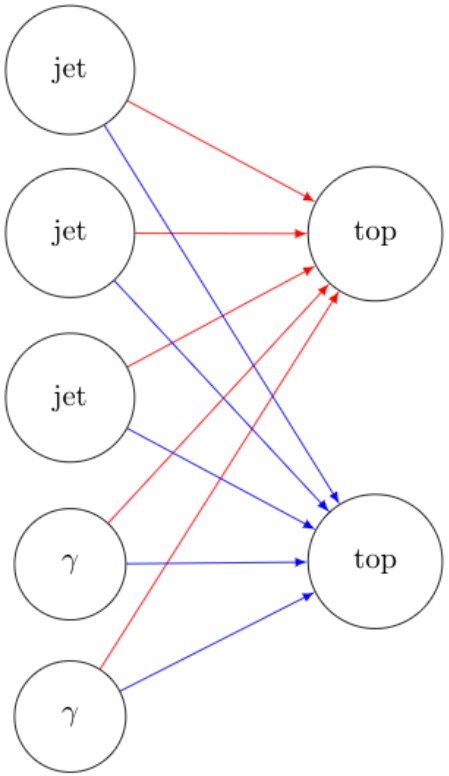


# Encoder



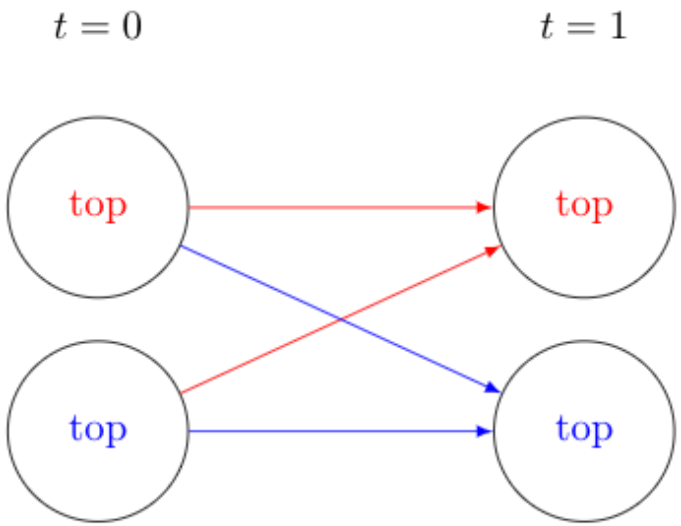
# Decoder

Interleaved covariant cross/self-attention layers.
Covariant cross-attention layer: update top quarks' feature vectors (including their 4-vectors) with final state objects as input.



# Decoder

Covariant self-attention layer: update top quark feature vectors with top quarks themselves as input.



# Covariant attention layer

- Every object $i$ is represented by a feature vector $f_i = (z_i, \omega_i)$ that consists of $z_i$ a list of Lorentz scalars (pid, $p_T$, $m$, ...) and $\omega_i = (\eta_i, \phi_i)$ a pair of non-scalar components of its 4-vector.
- The set of all Lorentz scalars $\{z_i\}$ and $\{\omega_i - \omega_j\}_{ij}$ are used produce update vectors $\{\Delta_i\}$ for each object through the standard attention mechanism, which are too Lorentz scalars.
- Each update vector is structured as $\Delta_i = (\delta z_i, \delta\eta_i, \delta\phi_i)$ and is used to update $f_i = (z_i, \omega_i)$ to $f_i' = (z_i', \omega_i')$ given by

$$z_i' = z_i + \delta z_i, \omega_i' = (\eta_i + \delta\eta_i, \phi_i + \delta\phi_i).$$

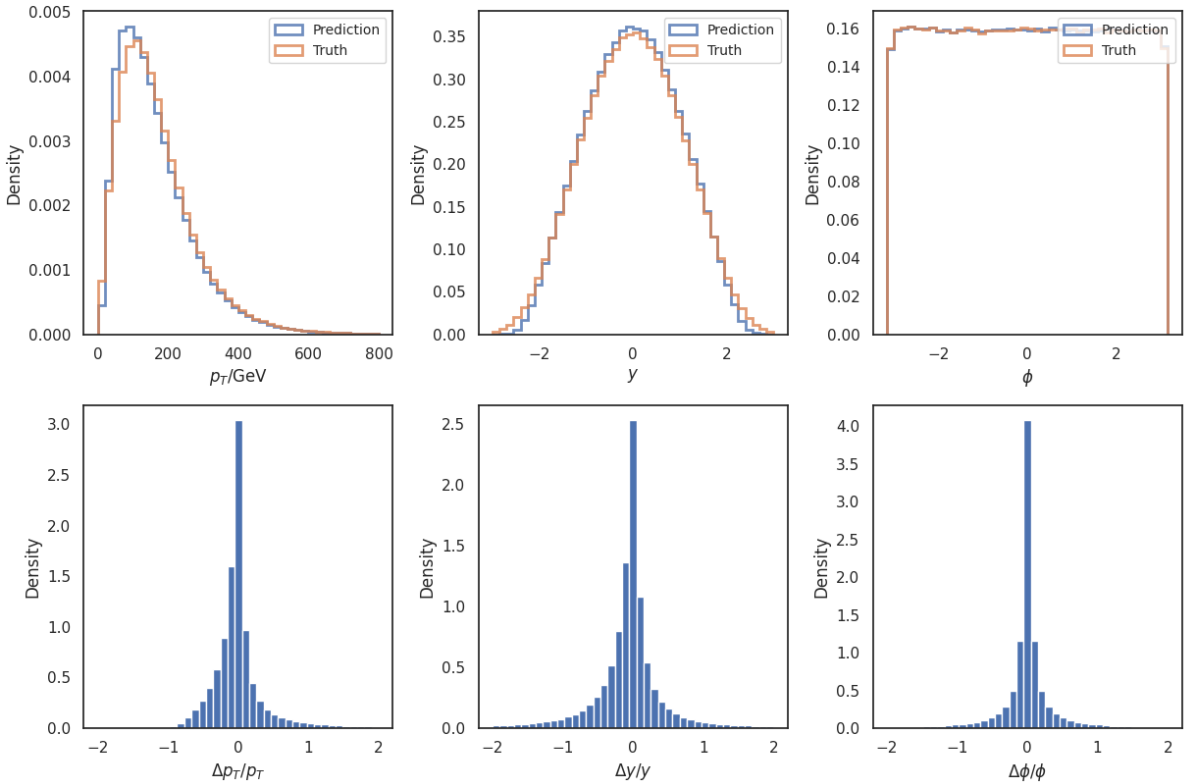
- It is straightforward to check the transformation property (invariance/covariance) of each quantity is preserved after each layer.
- This approach is much simpler to implement and much less computationally intensive compared to alternatives based on irreducible representations of the Lorentz group.

# Experimental setup

- We use MC events generated at NLO in QCD with parton showering. We use truth-level information and do not model detector response.
- Jets are required to have $|\eta| \leq 2.5$ and $p_T \geq 25$GeV, while leptons are required to have $|\eta| \leq 2.5$ and $p_T \geq 10$GeV. A jet is removed if it has $\Delta R \leq 0.4$ with a photon or a lepton.
- Further apply a cut on the testing set of $N_{\text{bjet}} \geq 0$, and ($N_{\text{jet}} \geq 3$ and $N_{\text{lepton}} = 0$) or $N_{\text{lepton}} \geq 0$.
- We show results for two models, one trained on 7M $ttH(H \to \gamma\gamma)$ events, the other trained on 4M *tttt* events. All results are reported using the test set.

# Results on $ttH(H \to \gamma\gamma)$

- Metric: percentage resolution = median value of the normalized prediction error (e.g $\frac{\Delta p_T}{p_T}$) over the test set




| $p_T$ | $\eta$ | $\phi$ |
|---|---|---|
| 12.1% | 16.9% | 8.4% |

# Resolution given preselections

The first column shows the preselection applied, and the second column shows the fraction out of all events that pass the preselection.

| Preselection | Fraction | $p_T$ | $\eta$ | $\phi$ |
|---|---|---|---|---|
| Inclusive | 100.0% | 12.1% | 16.9% | 8.4% |
| Reconstructable | 23.9% | 7.2% | 7.0% | 4.8% |
| Not reconstructable | 76.1% | 13.9% | 20.9% | 9.8% |

Reconstructable = having a matched triplet of jets

- Prediction is more accurate for reconstructable tops (having a matched triplet of jets), but they only constitute $\leq 25\%$ of all the tops.
- Resolution for rot reconstructable tops is worse due to loss of information, but the model still provides reasonable predictions.
- Without this new approach, “not reconstructible“ tops are basically lost in analysis

# Comparison with alternatives

Reconstruction-based methods deal with triplet identification, so we only compare on reconstructable tops that have a matched triplet.

| Model | $p_T$ | $\eta$ | $\phi$ | $m$ |
|---|---|---|---|---|
| Truth-triplets | 5.6% | 3.2% | 2.7% | 7.4% |
| GNN Top Reco [1] | 9.1% | 6.7% | 5.4% | 7.9% |
| Our method | 7.0% | 7.0% | 4.7% | 0.4% |

"Truth-triplets" directly use the ground truth four-vector of the matched triplets systems

- Even restricted to reconstructable tops, our method offers better resolution (expect for $\eta$) than the best current approach trained specifically to identify triplets (this happens because the identified triplet can be incorrect).
- Resolution of our method is approaching that of the truth-triplets, which is the optimal resolution achievable by any method based on reconstruction.



[1] A method developed by Allison Xu, Shuo Han, Ryan Roberts, Haichen Wang et al. link

# Generalization across top production process

- Machine learning models often perform poorly on inputs that look different from training data.
- To test if the performance of our model generalizes to other *tt* production processes, we apply the model trained only on $ttH(H \to \gamma\gamma)$ directly to *tt* and $ttW(W \to \text{inclusive})$.

| Process | $p_T$ | $\eta$ | $\phi$ |
|---|---|---|---|
| *ttH* | 12.1% | 16.9% | 8.4% |
| *tt* | 15.7% | 18.0% | 12.0% |
| *ttW* | 19.6% | 20.4% | 15.3% |

- Resolution gets worse but is still within $20\%$, even for *ttW* which produces more complicated final state objects.

# Applicability to four-top

- No reconstruction-based method has been successfully applied to 4-top due to extreme combinatorics.
- Our method can be applied to 4-top with zero modification except for the switching the training set.

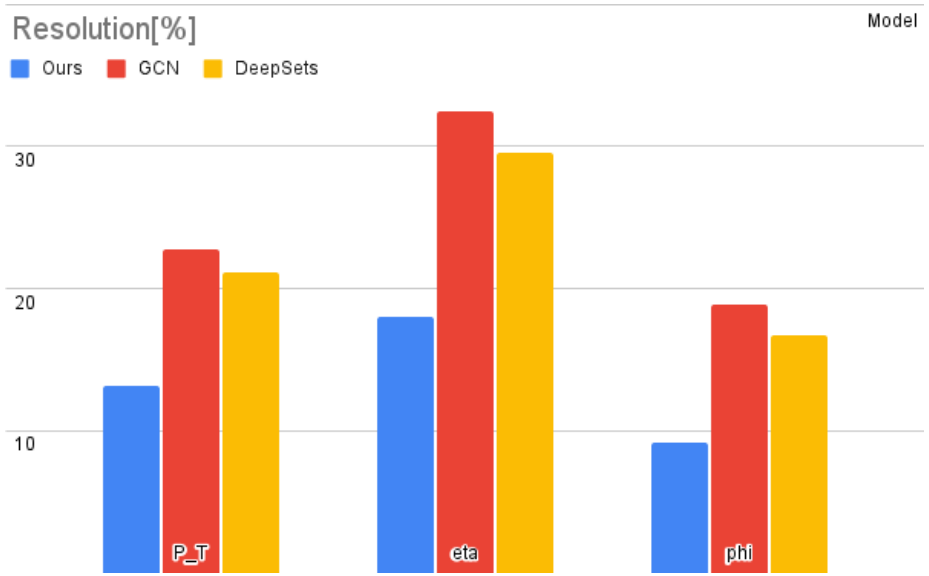
| Preselection | Fraction | $p_T$ | $\eta$ | $\phi$ |
|---|---|---|---|---|
| Inclusive | 100.0% | 20.6% | 22.4% | 12.6% |
| Reconstructable | 25.3% | 16.0% | 14.2% | 9.0% |
| Not reconstructable | 74.7% | 22.3% | 25.6% | 14.0% |

Resolution of *tttt*, trained on *tttt*.

- Decent resolution $\sim 20\%$ given the complexity of final state objects.

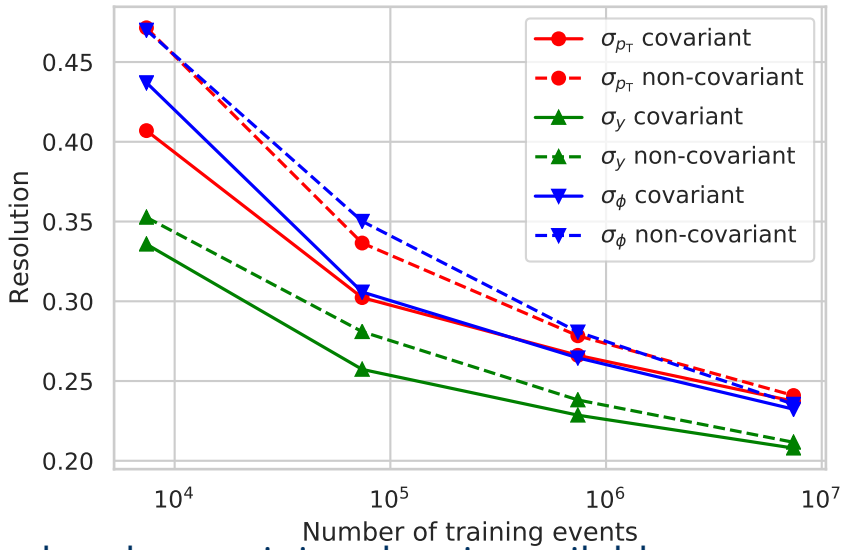
# Importance of the architecture

On $ttH(H \to \gamma\gamma)$, our method significantly outperforms popular baseline methods used to process graph or set structured data: 1. Graph Convolutional Network (GCN), 2. DeepSet.



# Benefit of enforcing Lorentz covariance

We compares the resolution achieved by a covariant and a non-covariant model as a function of number of training events on the $ttH(H \to \gamma\gamma)$ dataset.




- Significant advantage when less training data is available.
- Covariance enables the model to provably generalize across a family of events related by a Lorentz transformation $\to$ more data-efficient.

# Conclusion

- We introduce the **Covariant Particle Transformer**, a general architecture for high energy physics applications performing set-to-set predictions, satisfying Lorentz covariance.
- Compared to alternative methods, we achieve Lorentz covariance with a much simpler implementation and a much smaller computational overhead.
- We formulate top kinematics prediction as a regression as opposed to combinatorics problem and demonstrate that CPT enables state-of-the-art predictive accuracy across a wide range of processes $(ttH, tt, ttW, tttt)$, and shows non-trivial generalization across processes.
- Our method has the following advantages over reconstruction-based approach:
  1. It works for $\sim 75\%$ of the top quarks that are not reconstructable.
  2. It provides equally accurate predictions on reconstructable ones.
  3. It works out-of-the-box on processes with complicated final state objects such as $tttt$, where most reconstruction-based methods are infeasible due to extreme combinatorics.
- We hope the methodology and novel architecture developed in this work can be applied to solve other traditionally challenging problems in HEP by leveraging the flexibility of modern deep learning models as well as our ability to bake in prior knowledge of physics.