A Methodology for Developing Foundational Transformer Models in Collider Physics Analysis

E. Abasov,[1] L. Dudko,[1] E. Iudin,[1] A. Markina,[1]
P. Volkov,[1] M. Perfilov,[1] and A. Zaborenko[1,*]

[1] *Skobeltsyn Institute of Nuclear Physics of Lomonosov Moscow State University (SINP MSU), 1(2), Leninskie gory, GSP-1, Moscow 119991, Russian Federation*

We present a methodology for training foundational transformer models capable of processing collider data with diverse kinematic signatures. Our universal foundation model is designed for simultaneous analysis of all processes involving from one to four top-quarks production with their corresponding background processes. The approach employs multi-task pre-training on combined datasets of simulated events, enabling the model to capture the full spectrum of interaction physics while extracting universal patterns across different final states prior to task-specific fine-tuning. This unified architecture eliminates the need for separate analysis frameworks for different final signatures and specific tasks.

The transformer-based pre-training strategy explicitly preserves unique interaction patterns through adaptive attention mechanisms while establishing cross-process correlations. We plan to demonstrate how this architecture maintains sensitivity to rare high-multiplicity topologies (3t and 4t) without compromising performance on conventional channels ($t\bar{t}$, tX, $t\bar{t}H$), effectively bridging the gap between disparate analysis paradigms in collider physics.

I. FOUNDATIONAL MODELS

In the fields of computer vision [1] and natural language processing [2], foundational models pre-trained on large and diverse datasets are widely used. Such training allows models to uncover data patterns useful for numerous downstream tasks.

In High Energy Physics (HEP), this approach shows particular promise for event classifi-

* zaborenko.ad18@physics.msu.ru

cation. A foundational model capturing general physical patterns (e.g., energy-momentum correlations, jet substructure relationships) could enhance the accuracy of specialized analyses through transfer learning.

The study investigates:

- Architectural adaptations for collider data (input space formulation, numerical embeddings)
- Comparative evaluation of pre-training techniques (masked reconstruction vs contrastive learning)
- Transfer learning capabilities through downstream regression and classification tasks

For analyzing complex 3- and 4-top-quark production processes, we developed a large neural network (NN) architecture (Fig. 1), pre-trained on kinematically distinct production processes sharing underlying Standard Model (SM) physics. This model bridges the gap between specialized networks (trained on specific kinematic regimes) and broad foundational models capable of pattern recognition across entire HEP domains.

II. UNIFIED INPUT VARIABLE SPACE

For testing these approaches, a dataset of simulated collider events was used, comprising five process groups categorized by top-quark multiplicity (Fig. 2). Event generation utilized MadGraph5 [4] and CompHEP [5, 6] packages with the number of events for each process scaled according to available computational resources, while detector response simulation employed in DELPHES [7] for the CMS detector (13 TeV) [8, 9].

The dataset is categorized into five classes based on the number of top quarks in the final state:

- **Zero-top class:**
  - W^+W^- production (diboson)
  - W+jets production (inclusive W boson with jets)
- **Single-top class:**

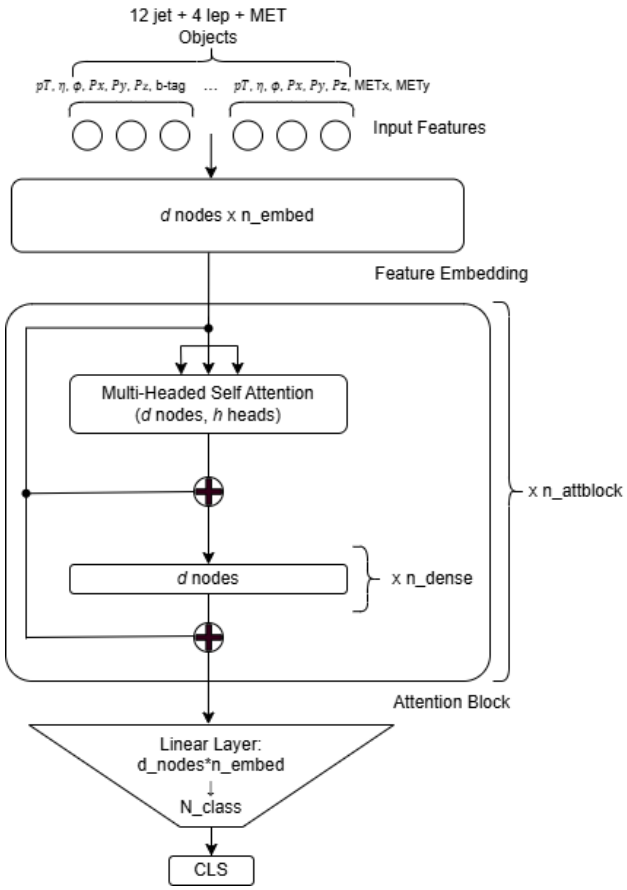


Figure 1: Graphical representation of the Transformer model. Input features are embedded through a LinearEmbedding [3] layer, then passed through several Attention Blocks employing Multi-Headed Self-Attention. The output of the Transformer is flattened and passed through the final Linear Layer to convert its dimension to N_{class} for classification. This approach can be used to convert the output of the model into any number of dimensions for regression or classification tasks.

  - Single-top production via t-channel (anti-top and top quarks)
  - Single-top production in association with a W boson (tW-channel for anti-top and top quarks)

- **Two-top class:**

  - $t\bar{t}$ production (top quark pair) in dileptonic and semileptonic decay channels

- **Three-top class:**

  - Standard Model (SM) processes: $TTTJ$ (triple-top + jet) and $TTTW$ (triple-top + W boson)

- **Four-top class:**

  - SM four-top-quark production ($TTTT$)

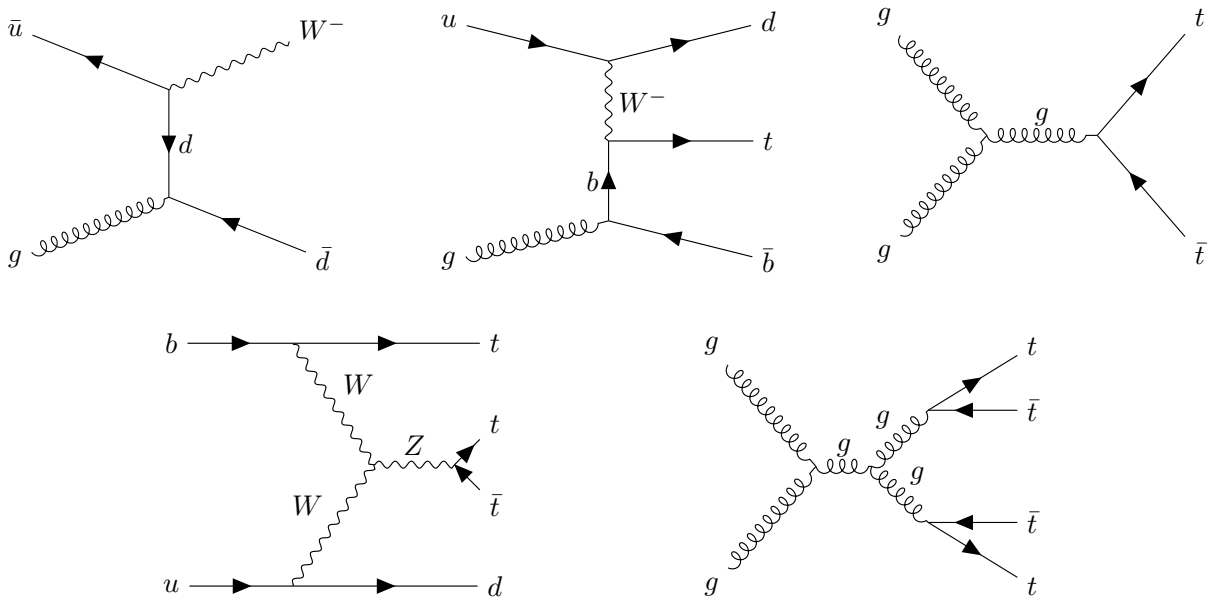


Figure 2: Example Feynman diagrams for process groups with different top-quark counts.

Class	Count	Percentage (%)	Class Weight
Zero-top	1,787,876	21.81	0.917
Single-top	2,174,999	26.53	0.754
Two-top	3,329,795	40.62	0.492
Three-top	31,403	0.38	52.214
Four-top	874,355	10.66	1.875
TOTAL	8,198,428	–	–

Table I: Class distribution and balanced class weights applied to each sample in the corresponding class.

To mitigate class imbalance (Table I), each event was assigned a sample weight inversely

proportional to its class frequency, such that all classes contribute equally to the training. To unify the input variable space, the following minimized approach was proposed:

- An event is represented as a collection of kinematic variables for all possible objects. For each object, we use:
  - p_T: Transverse momentum
  - η: Pseudorapidity
  - ϕ: Azimuthal angle
  - P_x, P_y, P_z: Cartesian momentum components
- Event-wide variables include:
  - N_{jets}: Number of jets
  - $N_{b\text{-jets}}$: Number of b-tagged jets
  - N_ν: Number of neutrinos
  - N_l: Number of leptons
  - MET: Missing Transverse Energy
- Maximum object counts: 12 jets (4 b-jets) + 4 leptons.
- Objects within each group are sorted by energy.
- Missing objects are represented with zero-valued variables, handled separately.

III. EVENT REPRESENTATION LEARNING

A potential step for analyzing processes with different kinematic features involves creating a unified representation space. Typically, this is the output of an intermediate network layer trained on a combined dataset. This space is expected to better reveal data patterns and cluster definitions. Common implementations use dimensionality reduction (autoencoders) or masked variable modeling [10].

For our task, masked variable reconstruction was chosen (Fig. 3). Compared to autoencoders, this method may improve precision by masking all variables during training, forcing

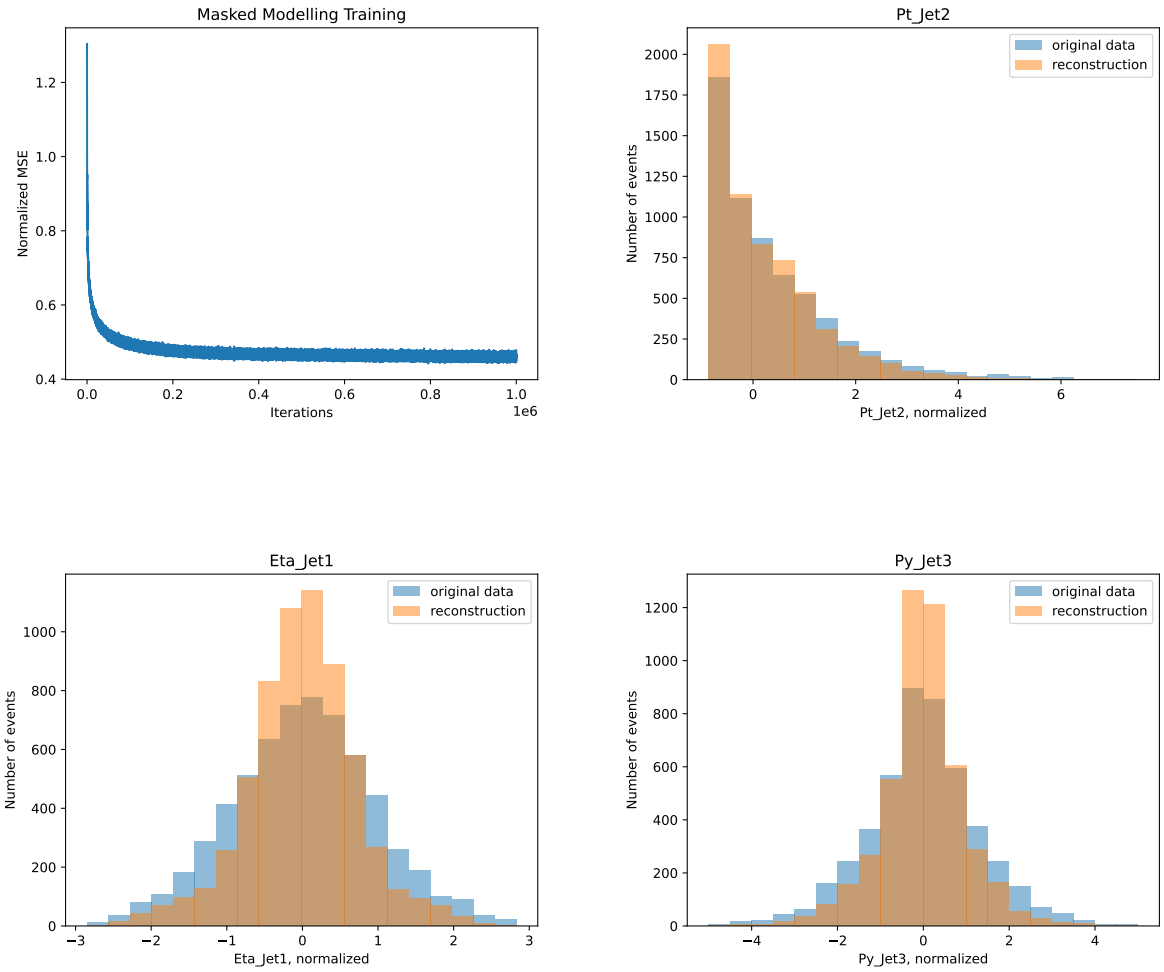
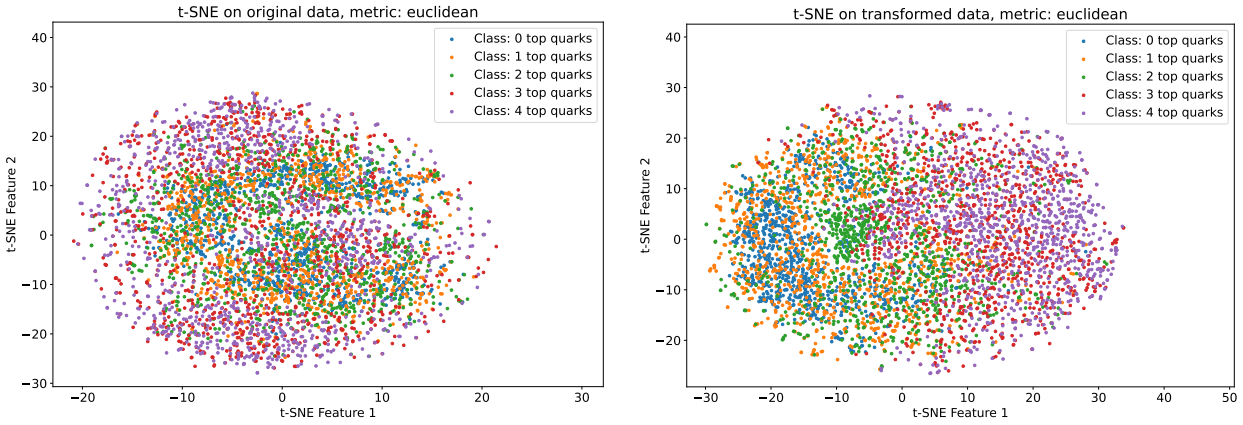


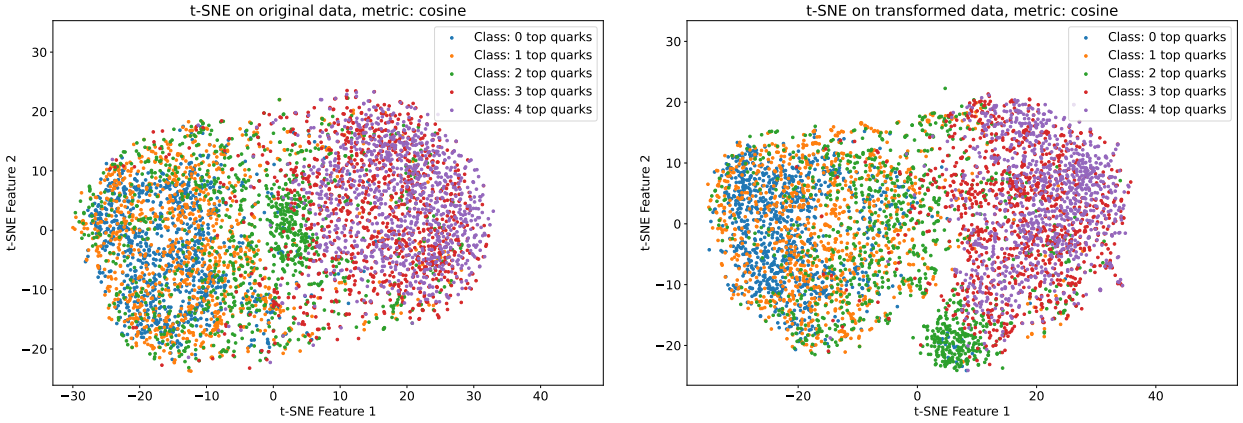
Figure 3: Examples of pre-training results for masked reconstruction (average reconstruction error for the test split and event distributions for some of the reconstructed variables). Masking probability: 30%, standardized (i.e. subtracted the mean and divided by the standard deviation) variables.

the model to learn inter-variable relationships rather than relying on average reconstruction quality. The visualization of the data manifold before and after applying the masked modelling approach are presented in (Fig. 4). The clusters containing events with the same class are more pronounced after applying the model.

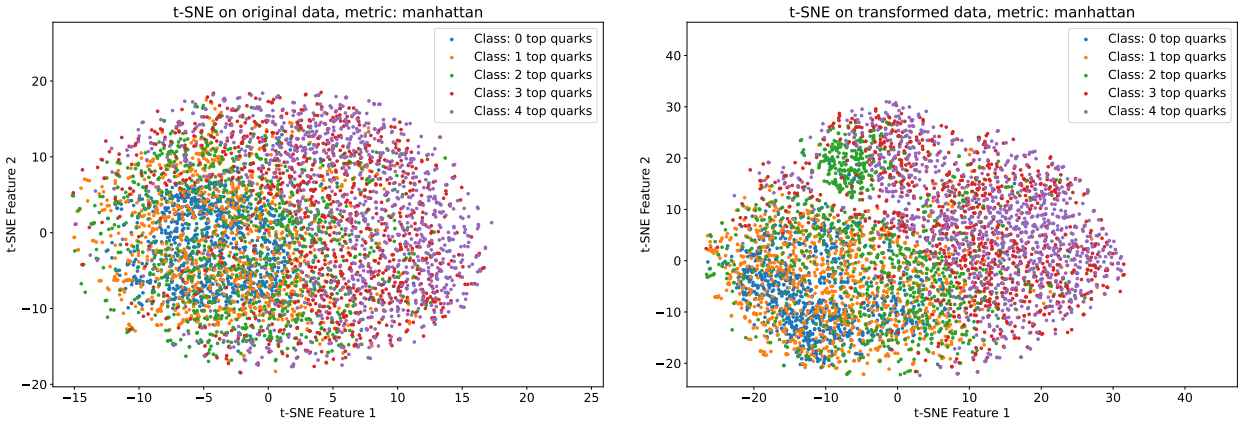
Potential pre-training directions include using generator-level information: neutrino component regression and combinatorial challenges, with unfolding to parton-level information.



(a) Euclidean distance: original input (left) vs. transformed space (right).



(b) Cosine distance: original input (left) vs. transformed space (right).



(c) Manhattan distance: original input (left) vs. transformed space (right).

Figure 4: t-SNE [11] visualization of the effect of DNN transformation across distance metrics. Each row compares original input space (left) with transformed space (right).

IV. TRANSFORMER TRAINING

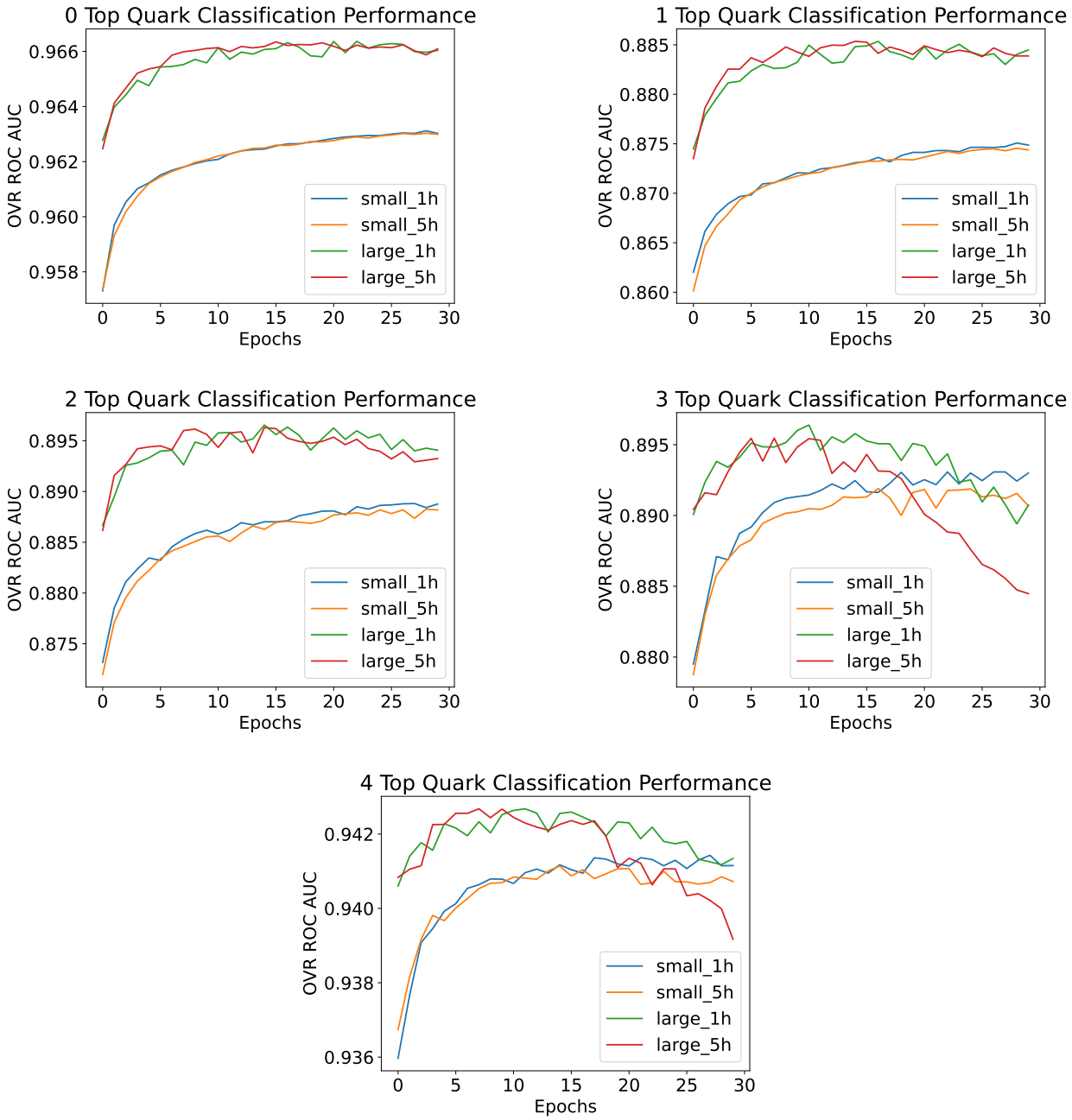


Figure 5: One-vs-rest ROC AUC metrics for all classes and reported architectures.

To study architectural dependencies, four Transformer network variants [12] were prepared (Table II). Preliminary comparisons (Fig. 5) revealed:

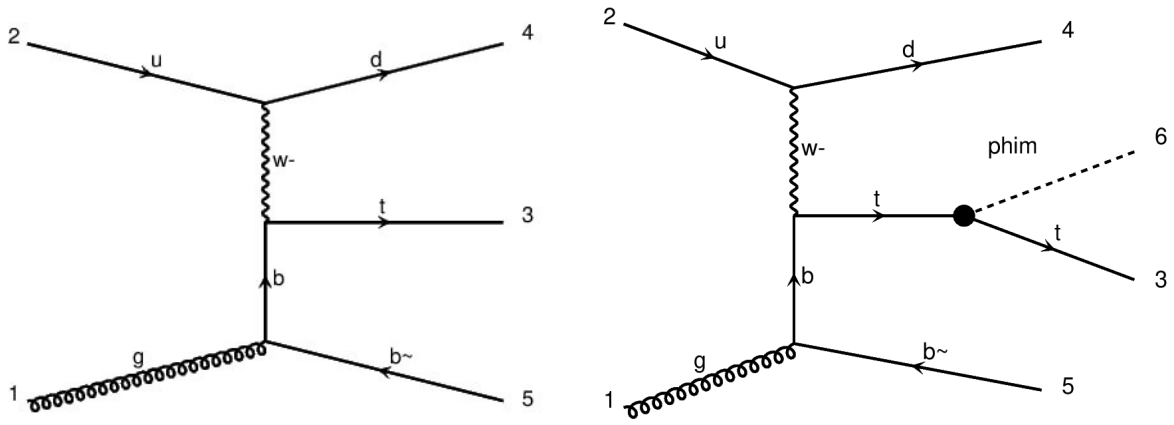
- Model size increases improve classification metrics but require more training resources
- Adding attention heads degrades performance when overfitting occurs

Table II: Transformer architecture variants and training hyperparameters. All models use AdamW optimizer with learning rate 0.0001 and no weight decay. The model metrics are calculated as the maximum mean one-vs-rest Receiver Operating Characteristic Area Under the Curve (ROC AUC) on the test set.

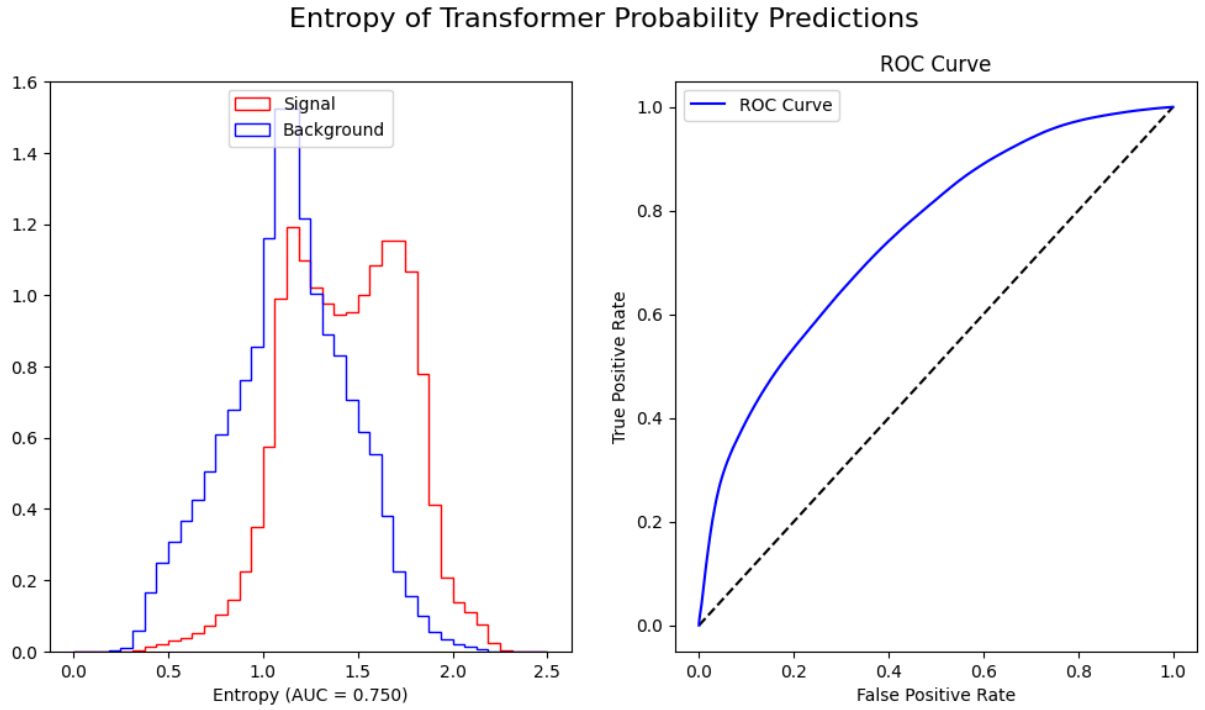
Parameter	Small_1h	Small_5h	Large_1h	Large_5h
Input size	125	125	125	125
Output size	5	5	5	5
Feedforward dim	128	128	512	512
Hidden layers	5	5	5	5
Embedding dim	16	20	125	125
Attention heads	1	5	1	5
Dropout	0.1	0.1	0.1	0.1
Max of Mean ROC AUC	0.9128	0.9121	0.9172	0.9169

V. EXAMPLE OF DOWNSTREAM APPLICATION

As an application example, the entropy of the Transformer's probability distribution (without fine-tuning) was used to identify events with dark matter (DM) mediator in single top-quark production (Fig. 6). The specifics of the DM model are covered in [13, 14]. It is important to emphasize that the Transformer was trained exclusively on the Standard Model events and was not exposed to DM events during training. The entropy function $H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$ was applied to predicted probabilities for each event, comparing t-channel single top-quark production in Standard Model (SM) and in association with scalar dark matter mediator in Simplified Model scenario (DM).



(a) Examples of Feynman diagrams for background (SM, left) and signal (DM, right) processes.



(b) Separating power of entropy distributions for background vs. signal events.

Figure 6: Feynman diagrams and entropy-based separation of background and signal.

VI. PROSPECTS

Future work will focus on building a universal model for different processes with top-quarks. One of the most promising applications is the study of 3- and 4-top-quarks production processes for dark matter searches [15, 16]. To enhance sensitivity to these rare processes, foundational model pre-training will be implemented. Training efficient models on diverse datasets could reveal complex patterns potentially boosting specialized analysis sensitivity in all possible processes with top-quark in different theories. Such foundational NN models already show promise in HEP [17, 18], demonstrating potential for unifying analyses of disparate processes.

Some additional studies could include incorporating various methods implemented in state-of-the-art neural networks for particle reconstructions and unfolding. One approach is utilizing the tensor attention mechanism [19], which includes symmetry between decay products in both the network architecture and loss function. Since the data has azimuthal symmetry, the covariant attention mechanism [20, 21], which uses Lorentz symmetry to account for this fact, also looks promising; however, studies conducted in [20–22] show that while this approach improves network performance if trained on small datasets ($< 10^5$ events), its effect becomes negligible as the size of the dataset increases. Even in the current state of our research, we use approximately $5 * 10^6$ events for training, so the traditional attention mechanism will provide sufficient performance.

At this stage of the work, only basic kinematic variables are used as inputs for the network. This standardized set is used in almost all state-of-the-art event reconstruction methods, such as SPA-NET [22] and HyPER [23] and has proven to be sufficient for highly efficient models. However, it was demonstrated that additional variables responsible for the pairwise and cluster interaction between particles, such as scalar products, can improve performance by decreasing the possible order of nonlinearity in the functional dependence of the input space and desired output [24]. We believe that the inclusion of these variables can make a positive impact on our network and this will be studied in future research.

[1] Sonain Jamil, Md. Jalil Piran, and Oh-Jin Kwon. A comprehensive survey of transformers for computer vision. *Drones*, 7(5), 2023.

[2] Narendra Patwardhan, Stefano Marrone, and Carlo Sansone. Transformers in the real world: A survey on nlp applications. *Information*, 14(4), 2023.

[3] Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. In *NeurIPS*, 2022.

[4] 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.

[5] E. Boos, V. Bunichev, M. Dubinin, L. Dudko, V. Ilyin, A. Kryukov, V. Edneral, V. Savrin, A. Semenov, and A. Sherstnev. CompHEP 4.4: Automatic computations from Lagrangians to events. *Nucl. Instrum. Meth. A*, 534:250–259, 2004.

[6] A. Pukhov, E. Boos, M. Dubinin, V. Edneral, V. Ilyin, D. Kovalenko, A. Kryukov, V. Savrin, S. Shichanin, and A. Semenov. CompHEP: A Package for evaluation of Feynman diagrams and integration over multiparticle phase space. 1999. `arXiv:hep-ph/9908288`.

[7] Michele Selvaggi. DELPHES 3: A modular framework for fast-simulation of generic collider experiments. *J. Phys. Conf. Ser.*, 523:012033, 2014.

[8] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[9] DELPHES CMS Card. https://github.com/delphes/delphes/blob/master/cards/delphes_card_CMS.tcl. Accessed: 2025-10-02.

[10] Tobias Golling, Lukas Heinrich, Michael Kagan, Samuel Klein, Matthew Leigh, Margarita Osadchy, and John Andrew Raine. Masked particle modeling on sets: towards self-supervised high energy physics foundation models. *Mach. Learn. Sci. Tech.*, 5(3):035074, 2024.

[11] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] E. Abasov, L. Dudko, E. Iudin, A. Markina, P. Volkov, G. Vorotnikov, M. Perfilov, and A. Zaborenko. Reconstruction of angular correlations in the associated top quark and the dark matter mediator production. *ArXiv*, 4 2025. `arXiv:2504.14303 [hep-ph]`.

[14] E. Boos, V. Bunichev, and L. Dudko. Dark Matter Mediators in Processes of Single Top Quark Production. *Physics of Particles and Nuclei*, 56(2):429–433, 2025.

[15] Nathaniel Craig, Jan Hajer, Ying-Ying Li, Tao Liu, and Hao Zhang. Heavy Higgs bosons at low $\tan\beta$: from the LHC to 100 TeV. *JHEP*, 01:018, 2017.

[16] Georges Aad et al. Search for $t\bar{t}H/A \to t\bar{t}t\bar{t}$ production in the multilepton final state in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, 07:203, 2023.

[17] Georges Aad et al. Measurement of the associated production of a top-antitop-quark pair and a Higgs boson decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC. *Eur. Phys. J. C*, 85(2):210, 2025.

[18] Huilin Qu, Congqiao Li, and Sitian Qian. Particle Transformer for jet tagging. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18281–18292, 2022. `arXiv:2202.03772 [hep-ph]`.

[19] 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. *SciPost Phys.*, 12(5):178, 2022.

[20] Congqiao Li, Huilin Qu, Sitian Qian, Qi Meng, Shiqi Gong, Jue Zhang, Tie-Yan Liu, and Qiang Li. Does Lorentz-symmetric design boost network performance in jet physics? *Phys. Rev. D*, 109(5):056003, 2024.

[21] Shikai Qiu, Shuo Han, Xiangyang Ju, Benjamin Nachman, and Haichen Wang. Holistic approach to predicting top quark kinematic properties with the covariant particle transformer. *Phys. Rev. D*, 107(11):114029, 2023.

[22] Michael James Fenton, Alexander Shmakov, Hideki Okawa, Yuji Li, Ko-Yang Hsiao, Shih-Chieh Hsu, Daniel Whiteson, and Pierre Baldi. Reconstruction of unstable heavy particles using deep symmetry-preserving attention networks. *Commun. Phys.*, 7(1):139, 2024.

[23] Callum Birch-Sykes, Brian Le, Yvonne Peters, Ethan Simpson, and Zihan Zhang. Reconstructing short-lived particles using hypergraph representation learning. *Phys. Rev. D*, 111(3):032004, 2025.

[24] Lev Dudko, Georgi Vorotnikov, Petr Volkov, Dmitri Ovchinnikov, Maxim Perfilov, Artem Shporin, and Andrei Chernoded. General recipe to form input space for deep learning analysis of HEP scattering processes. *Int. J. Mod. Phys. A*, 35(21):2050119, 2020.