Bumblebee: Foundation Model for Particle Physics Discovery

Andrew J. Wildridge
Department of Physics and Astronomy
Purdue University
West Lafayette, IN 47907
awildrid@purdue.edu

Jack P. Rodgers
Department of Physics and Astronomy
Purdue University
West Lafayette, IN 47907
jprodger@purdue.edu

Ethan M. Colbert
Department of Physics and Astronomy
Purdue University
West Lafayette, IN 47907
colberte@purdue.edu

Yao Yao
Department of Physics and Astronomy
Purdue University
West Lafayette, IN 47907
yao317@purdue.edu

Andreas W. Jung
Department of Physics and Astronomy
Purdue University
West Lafayette, IN 47907
anjung@purdue.edu

Miaoyuan Liu
Department of Physics and Astronomy
Purdue University
West Lafayette, IN 47907
liu3173@purdue.edu

Abstract

Bumblebee is a foundation model for particle physics discovery, inspired by BERT. By removing positional encodings and embedding particle 4-vectors, Bumblebee captures both generator- and reconstruction-level information while ensuring sequence-order invariance. Pre-trained on a masked task, it improves dileptonic top quark reconstruction resolution by 10-20% and excels in downstream tasks, including toponium discrimination (AUROC 0.877) and initial state classification (AUROC 0.625). The flexibility of Bumblebee makes it suitable for a wide range of particle physics applications, especially the discovery of new particles.

1 Introduction

The intersection of machine learning (ML) and particle physics offers immense potential to improve our understanding of fundamental particles and their interactions [1]. Foundation models like BERT [2] excel in capturing complex relationships and achieving state-of-the-art results, but their inherent sequence sensitivity presents challenges in particle physics, where events, represented by particle 4-vectors, are naturally invariant to order. Additionally, developing an effective pre-training objective that aids in particle discovery remains nontrivial.

We propose **Bumblebee**, a foundation model inspired by BERT but tailored for particle physics. By removing positional encoding, we ensure sequence-order invariance and modify the embedding to capture particle 4-vectors instead of words, allowing Bumblebee to learn both generator-level (truth) and reconstruction-level (observed) information. Our pre-training objective, akin to BERT's *Cloze* task [3], trains Bumblebee to learn the kinematics within the event topology and the transformation between the generator and reconstruction levels. This intricate knowledge of the interactions and

event topology required to predict any decay product four momenta enables Bumblebee to perform state-of-the-art regression and classification tasks.

Bumblebee significantly outperforms state-of-the-art methods in dileptonic top quark reconstruction, a challenging task due to the presence of two neutrinos, which only manifest as missing transverse energy (MET) in Large Hadron Collider (LHC) detectors. Moreover, the pre-trained model can be fine-tuned to search for a potential top quark-antiquark ($t\bar{t}$) bound state (toponium) and to enhance the degree of quantum entanglement in pair-produced top quarks at the LHC [4]. This opens new pathways for precision measurements in quantum information science at the highest energies yet achieved.

2 Related Work

Foundation models have been successfully created for data domains such as images [5, 6, 7], text [2, 8, 9], and speech [10, 11]. So far, applications of foundation models in particle physics have focused on the reconstruction of particle "objects" such as jets [12, 13] and tracks [14] and fine-tuning for downstream tasks. Beyond foundation models, the transformer architecture has been successful in particle physics from generating events [15, 16] to achieving state-of-the-art performance in the jet-parton assignment problem [17]. This strongly motivates building a foundation model with the transformer architecture for particle physics discovery.

3 Bumblebee

The Bumblebee model is a transformer-based model [18] designed for particle physics. Similar to BERT [2] and other foundation models, our framework consists of a pre-training and fine-tuning step. We will be using the event topology of the dileptonic $t\bar{t}$ as a case study. Dileptonic $t\bar{t}$ means that we have a lepton (antilepton), a b antiquark (quark), and an antineutrino (neutrino) in the final state for each top antiquark (quark).

3.1 Model architecture

Bumblebee is a multilayer bidirectional transformer encoder based on the original implementation detailed in Ref. [18]. Unlike traditional transformer encoders, which rely on positional encodings to handle sequence order, Bumblebee eliminates positional encodings to ensure that they remain invariant to the order in which particles are processed. This architectural decision reflects the physical reality that particles in an event are not ordered in any meaningful way, and thus preserving this permutation invariance is critical for accurately modeling particle physics collisions.

We use L to denote the number of layers, the hidden size as d_{model}, and the number of heads of self-attention as A. Specifically, the Bumblebee model as reported here is (L=8, d_{model}=768, A=16, Total Parameters=57M).

3.2 Input representation

Bumblebee takes as input the 4-vectors of particles at both the generator and reconstruction levels, allowing it to learn correlations between partonic truth and reconstructed observables, improving downstream prediction quality. Each 4-vector consists of the transverse momentum p_{T}, pseudorapidity η, azimuthal angle ϕ, and mass m of the particle. Due to color confinement, generator-level quarks form jets [19]. Bumblebee receives b-tag scores to indicate the likelihood of a jet originating from a b quark. Non-jet particles and generator-level b quarks are assigned b-tag scores of 0 and 1, respectively. Neutrinos manifest as MET because of the conservation of momentum, with assigned pseudorapidity and mass of zero, because their z-momentum and energy are unmeasured. These five-dimensional vectors are linearly embedded in the d_{model}-dimensional space.

Additionally, Bumblebee uses three learned embedding tables: (1) to differentiate between reconstruction-level (isReco) and generator-level (isNotReco) particles, (2) to distinguish particle types using a modified PDG ID scheme where b-tagged jets are assigned a PDG ID of 5, non b-tagged jets 41, and MET 40, with all IDs shifted by $+50$ to map to positive indices, and (3) to indicate whether particles are masked (isMasked) or not (isNotMasked). This masking is essential

for Bumblebee's pre-training task (Section 3.3). The final input to Bumblebee is the unweighted sum of the four d_{model}-dimensional embedded vectors, as illustrated in Fig. 1 for a dileptonic $t\bar{t}$ event.

Figure 1: The embedding procedure for Bumblebee for a $t\bar{t}$ dileptonic decay. The final embedded input given to Bumblebee is the unweighted sum of the particle vector embedding, PDG ID embedding, level type embedding, and mask status embedding.

We apply object and event selection criteria typical of a dileptonic $t\bar{t}$ analysis at the LHC [20, 21]. Bumblebee is pre-trained and fine-tuned on events and objects after these selection criteria.

3.3 Pre-training Bumblebee

We pre-train Bumblebee using the *Cloze* task [3], where particles are randomly masked with a $(1/n_{particles})\%$ probability for half of the training. The other half involves masking all 4-vectors at the generator or reconstruction level. Only the particle vector embedding is masked in both scenarios, shown as $\vec{0}$ in Fig. 1. The model minimizes the batch-average mean squared error (MSE) on the predicted masked 4-vectors. Validation and testing of Bumblebee focus on predicting generator-level 4-vectors from reconstruction-level information, framed as dileptonic top quark reconstruction. During pre-training, we use a linearly decaying learning schedule with a warm-up of 9,000 iterations, peaking at a learning rate of $\sim 10^{-4}$. The Adam optimizer [22] is used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay for regularization. Epsilon, dropout, batch size and weight decay are set to 10^{-8}, 0.05, 16, and 10^{-3}, respectively. Training is carried out on 2 V100 GPUs for 10 epochs, and the model with the lowest validation loss is used for test predictions.

3.4 Fine-tuning

After pre-training Bumblebee on the $t\bar{t}$ system, we fine-tune the model for downstream tasks. For classification, a masked vector $(1, 0, 0, 0, 0)$ is added for signal and $(0, 0, 0, 0, 0)$ for background, with a "PDG ID" of 50, using the reconstruction-level (isReco) and masked (isMasked) embeddings. Generator-level information is omitted during fine-tuning, and the model is trained to predict this vector, similar to the CLS token in BERT. To avoid catastrophic forgetting [23], the learning rate is reduced by an order of magnitude. Hyperparameters such as weight decay, dropout, batch size, and ϵ are optimized during this phase and trained for 4 epochs.

4 Datasets

To test the performance of Bumblebee, we generate a 7M $t\bar{t}$ Monte Carlo sample at next-to-leading order using POWHEGV2 [24, 25, 26, 27]. Additionally, we use a toy model [28] to generate a 1M Monte Carlo sample of the ground state, η_t, of the $t\bar{t}$ bound state toponium at leading order with MADGRAPH5_AMC@NLO [29]. Both processes perform parton showering and hadronization with PYTHIA [30]. Detector simulation is performed with DELPHES [31] using the default card for the

CMS detector card. We use a 70/15/15 training/validation/test split for each dataset and all results are shown on the withheld test set.

5 Experiments

We conducted several experiments to evaluate the ability of Bumblebee to learn foundational $t\bar{t}$ physics and fine-tune on downstream tasks.

Dileptonic top quark reconstruction   Our pre-training task doubles as a dileptonic top quark reconstruction challenge. The reconstructed top quark is the sum of predicted generator-level daughter four-vectors. A 10-20% improvement in the resolution of the $t\bar{t}$ system's invariant mass ($m(t\bar{t})$) is achieved compared to a supervised transformer, as shown in Fig. 2C. The improved $m(t\bar{t})$ reconstruction resolution at high invariant mass is of great importance for heavy resonance searches of physics beyond the Standard Model [32].

Toponium discrimination   We benchmark the ability of Bumblebee to discover new particles using the ground state of toponium (η_t), a hypothetical particle predicted by the Standard Model [28, 33, 34, 35, 36]. Due to the low resolution in $m(t\bar{t})$ relative to η_t's width (3 GeV) [37], observation of η_t is challenging and an appropriate benchmark for Bumblebee. We fine-tune Bumblebee on weighted binary cross-entropy. With early stopping and a class imbalance of 10:1, the model achieves an AUC of 0.877, as seen in Fig. 2A.

Initial state classification   Top quark pairs at the LHC originate from gluon-gluon or quark-antiquark interactions with a rich dependence on this origination [4, 38]. Initial-state discrimination enhances the search for new physics [39]. Bumblebee achieves a 0.625 AUC in this task, marking the first attempt at initial-state classification and outperforming supervised machine learning models, as shown in Fig. 2B.

6 Limitations

The primary limitation of this work is its focus on the dileptonic decays of top quark pairs at the LHC, driven by the computational cost of generating Monte Carlo for each process. Although photons are absent in dileptonic decays, there is nothing in our embedding procedure that inherently restricts Bumblebee from handling photons at the reconstruction or generator level. Another limitation is the event topology: in dileptonic decays, the main challenge is not jet-parton assignment but reconstructing the missing neutrino 3-vectors from MET. More complex topologies, such as $t\bar{t}H$ [40, 41] and $t\bar{t}t\bar{t}$ [42, 43], feature numerous jets and multiple neutrinos, offering further avenues for exploration.

7 Ablation study

We present an ablation study where we remove embeddings in the input representation and measure the performance regarding the MSE loss on the validation set. In Fig. 3, it is clear that the most important embedding is the PDG ID embedding as this results in the largest increase of MSE loss. This is expected as this defines the event topology at the generator and reconstruction levels. The next most important embedding is the level type embedding which is obvious when considering scenarios where a particle is masked both at the reconstruction and generator level.

8 Conclusion

The Bumblebee model has demonstrated its ability to outperform state-of-the-art methods in dileptonic top quark reconstruction, while also generalizing this knowledge to various downstream tasks. Its flexibility may extend beyond the $t\bar{t}$ process, as any particle with kinematics can be embedded, making it highly versatile for a wide range of applications. Although we focused on specific discrimination and pre-training tasks, Bumblebee can be fine-tuned for other uses. With its strong performance and adaptability, Bumblebee shows evidence of being a possible path to constructing foundation models applicable to many particle physics processes.
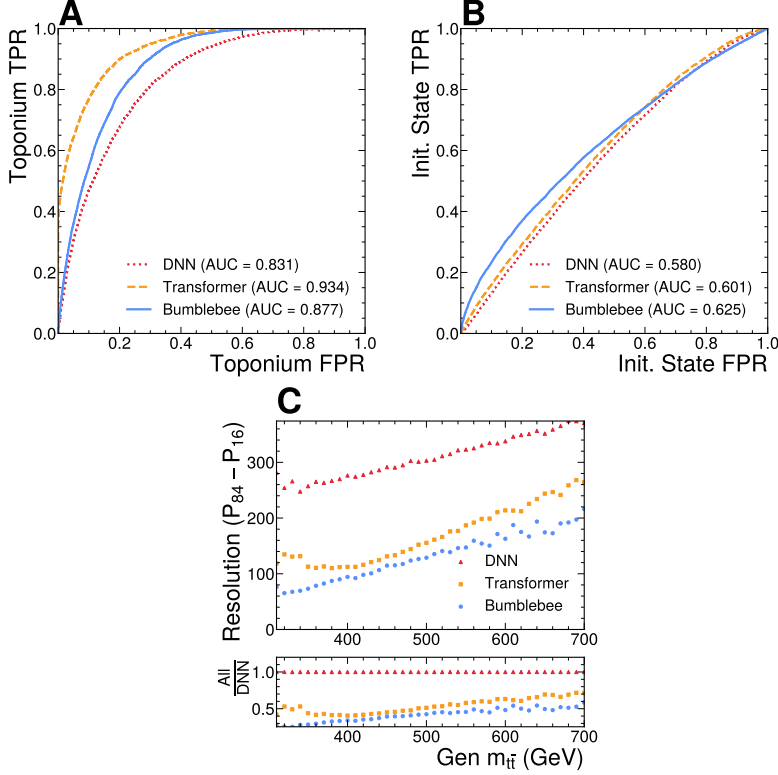
Figure 2: A) The receiver operating characteristic (ROC) curve for Bumblebee fine-tuned on discriminating toponium against $t\bar{t}$. Two supervised models, DNN and Transformer, are shown for comparison. B) The ROC curve for Bumblebee fine-tuned on discriminating the initial state of $t\bar{t}$. The positive class is the gluon-gluon initial state. Two supervised models, DNN and Transformer, are shown for comparison. C) The resolution of $m(t\bar{t})$ given as the difference between the 16th and 84th percentiles of the $m(t\bar{t})$ residuals ($P_{84} - P_{16}$) as a function of the true $m(t\bar{t})$.
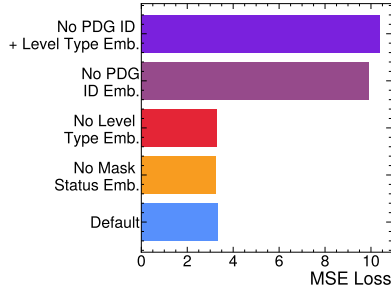

Figure 3: The comparison of the best validation loss obtained on the pre-training objective when removing individual embeddings present in the input representation. The embeddings removed are labeled on the y-axis.

Acknowledgments and Disclosure of Funding

References

[1] Kim Albertsson et al. *Machine Learning in High Energy Physics Community White Paper*. 2019. arXiv: 1807.02876 [physics.comp-ph].

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* (2018). arXiv: 1810.04805.

[3] Timo Schick and Hinrich Schütze. "Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference". In: *CoRR* (2020). arXiv: 2001.07676.

[4] Yoav Afik and Juan Ramón Muñoz de Nova. "Entanglement and quantum tomography with top quarks at the LHC". In: *The European Physical Journal Plus* 136.9 (2021). DOI: 10.1140/epjp/s13360-021-01902-1.

[5] Zelong Liu et al. *VISION-MAE: A Foundation Model for Medical Image Segmentation and Classification*. 2024. arXiv: 2402.01034 [eess.IV].

[6] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV].

[7] Kaiming He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. 2020. arXiv: 1911.05722 [cs.CV].

[8] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[9] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[10] Steffen Schneider et al. *wav2vec: Unsupervised Pre-training for Speech Recognition*. 2019. arXiv: 1904.05862 [cs.CL].

[11] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS].

[12] Tobias Golling et al. *Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models*. 2024. arXiv: 2401.13537 [hep-ph].

[13] Joschka Birk, Anna Hallin, and Gregor Kasieczka. *OmniJet-α: The first cross-task foundation model for particle physics*. 2024. arXiv: 2403.05618 [hep-ph].

[14] Andris Huang et al. *A Language Model for Particle Tracking*. 2024. arXiv: 2402.10239 [hep-ph].

[15] Thorben Finke et al. "Learning the language of QCD jets with transformers". In: *Journal of High Energy Physics* 2023.6 (2023). DOI: 10.1007/jhep06(2023)184.

[16] Anja Butter et al. *Jet Diffusion versus JetGPT – Modern Networks for the LHC*. 2023. arXiv: 2305.10475 [hep-ph].

[17] Alexander Shmakov et al. "SPANet: Generalized permutationless set assignment for particle physics using symmetry preserving attention". In: *SciPost Physics* 12.5 (2022). DOI: 10.21468/scipostphys.12.5.178.

[18] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* (2017). arXiv: 1706.03762.

[19] Andrew J. Larkoski. *QCD Masterclass Lectures on Jet Physics and Machine Learning*. 2024. arXiv: 2407.04897 [hep-ph].

[20] CMS Collaboration. *Observation of quantum entanglement in top quark pair production in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. 2024. arXiv: 2406.03976 [hep-ex].

[21] CMS Collaboration. "Measurement of the top quark polarization and $t\bar{t}$ spin correlations using dilepton final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: *Physical Review D* 100.7 (2019). DOI: 10.1103/physrevd.100.072002.

[22] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[23] Yun Luo et al. *An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning*. 2024. arXiv: 2308.08747 [cs.CL].

[24] S. Frixione, G. Ridolfi, and P. Nason. "A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction". In: *JHEP* 09 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. arXiv: 0707.3088 [hep-ph].

[25] S. Frixione, P. Nason, and C. Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

[26] P. Nason. "A new method for combining NLO QCD with shower Monte Carlo algorithms". In: *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146.

[27] S. Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph].

[28] Benjamin Fuks et al. "Signatures of toponium formation in LHC run 2 data". In: *Physical Review D* 104.3 (2021). DOI: 10.1103/physrevd.104.034023.

[29] Johan Alwall et al. "MadGraph 5: going beyond". In: *Journal of High Energy Physics* 2011.6 (2011). DOI: 10.1007/jhep06(2011)128.

[30] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024.

[31] J. de Favereau et al. "DELPHES 3: a modular framework for fast simulation of a generic collider experiment". In: *Journal of High Energy Physics* 2014.2 (2014). DOI: 10.1007/jhep02(2014)057.

[32] CMS Collaboration. "Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2019.4 (2019), p. 31. DOI: 10.1007/JHEP04(2019)031.

[33] Y. Kiyo et al. "Top-quark pair production near threshold at LHC". In: *The European Physical Journal C* 60.3 (2009), p. 375. DOI: 10.1140/epjc/s10052-009-0892-7.

[34] Wan-Li Ju et al. "Top quark pair production near threshold: single/double distributions and mass determination". In: *Journal of High Energy Physics* 2020.6 (2020). DOI: 10.1007/jhep06(2020)158.

[35] Yukinari Sumino and Hiroshi Yokoya. "Bound-state effects on kinematical distributions of top quarks at hadron colliders". In: *Journal of High Energy Physics* 2010.9 (2010). DOI: 10.1007/jhep09(2010)034.

[36] V. Fadin and V Khoze. "On the threshold behaviour of heavy top production". In: *Zeitschrift für Physik C Particles and Fields* 48.4 (1990). DOI: 10.1007/BF01614696.

[37] J. A. Aguilar-Saavedra. *Toponium Hunter's Guide*. 2024. arXiv: 2407.20330 [hep-ph].

[38] CMS Collaboration. "Measurement of the $t\bar{t}$ charge asymmetry in events with highly Lorentz-boosted top quarks in pp collisions at $\sqrt{s} = 13TeV$". In: *Physics Letters B* 846 (2023). DOI: 10.1016/j.physletb.2023.137703.

[39] J. A. Aguilar-Saavedra. "Overview of models for the $t\bar{t} asymmetry$". In: *Il Nuovo Cimento C* 35.3 (2012), p. 167. DOI: 10.1393/ncc/i2012-11240-7.

[40] CMS Collaboration. "Observation of $t\bar{t}H Production$". In: *Physical Review Letters* 120.23 (2018). DOI: 10.1103/physrevlett.120.231801.

[41] ATLAS Collaboration. "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector". In: *Physics Letters B* 784 (2018), p. 173. DOI: 10.1016/j.physletb.2018.07.035.

[42] CMS Collaboration. "Observation of four top quark production in proton-proton collisions at $\sqrt{s} 13TeV$". In: *Physics Letters B* 847 (2023). DOI: 10.1016/j.physletb.2023.138290.

[43] ATLAS Collaboration. "Observation of four-top-quark production in the multilepton final state with the ATLAS detector". In: *The European Physical Journal C* 83.6 (2023). DOI: 10.1140/epjc/s10052-023-11573-0.