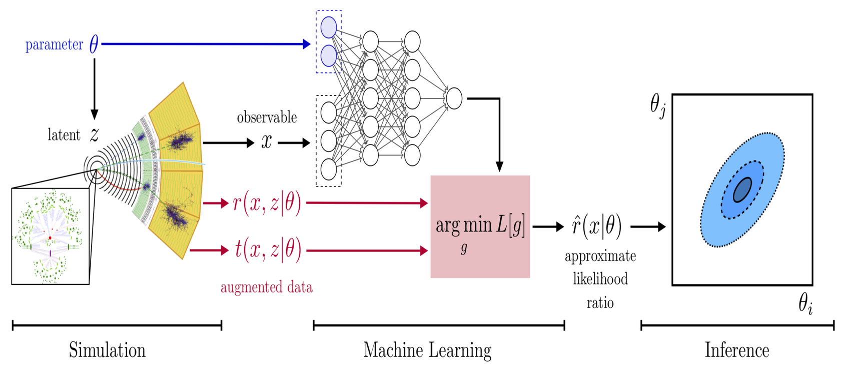
Bumblebee: Foundational Model for Particle Physics Discovery

A.J. Wildridge, **Jack Rodgers**, Ethan Colbert, Miaoyuan Liu, Andreas Jung

Introduction

With high energy physics becoming much more data-driven in the past few years, the introduction of machine learning models to perform inference on certain observables has become much more prevalent.



The most common uses in this domain are:

- **Anomaly Detection**
- **Classification**
- **Regression**

But with generative models / LLMs coming into the fold, there is an opportunity to create models to perform much more complex tasks such as:

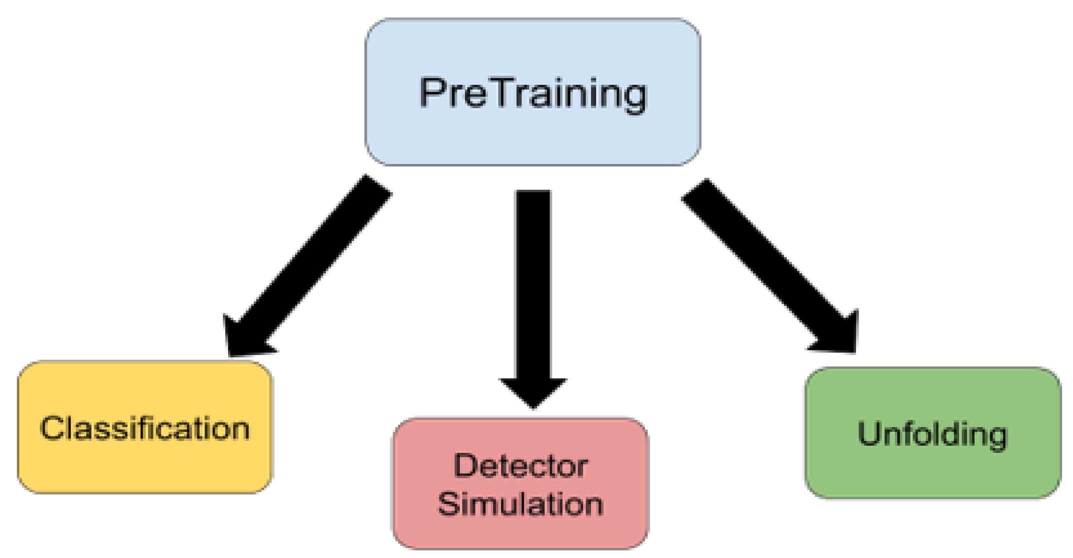
- **Unfolding** to rectify poor detector performance on some observables
- **Detector Simulation** to simulate events at a much quicker rate than traditional analytic methods.

Problem Formulation

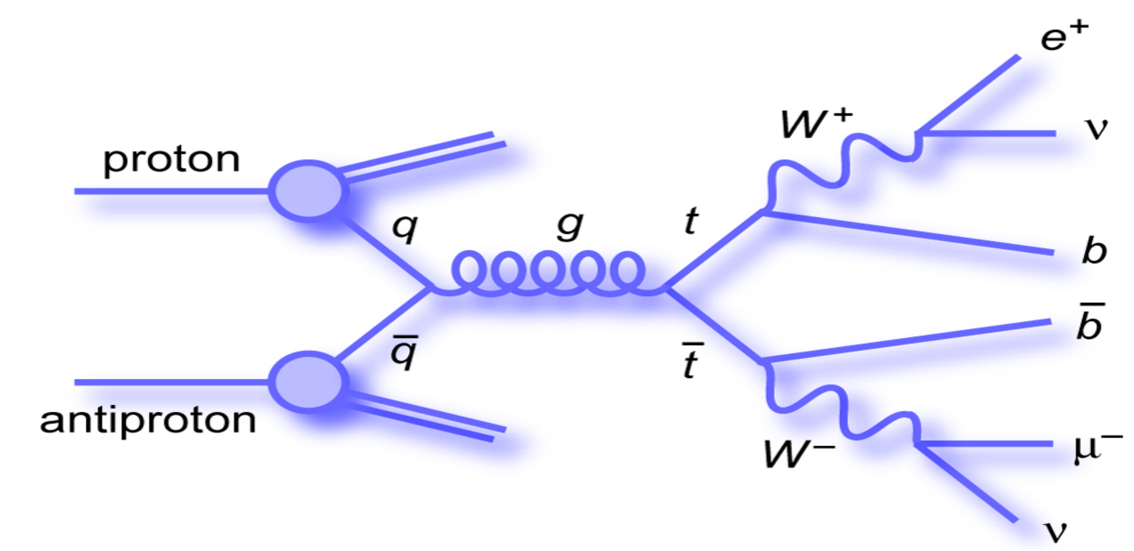
Approach: Create a model inspired by the success of LLMs such as RoBERTa[1] that can be generally pretrained on data and finetuned to perform inference on many different tasks that come up in high energy physics. As a use case, we use the dileptonic $t\bar{t}$ decay.

Motivation:

- LLMs and in particular, the multi headed attention mechanism, has demonstrated very good performance in areas of contextual learning and modeling sequential data.
- Rather than use a variety of models for different tasks in high energy physics, we introduce **one model** that can do the work of many different models and **achieve high performance** standards.



Methodology



- **Step a) PreTrain:** Using masked language modeling, we mask a particle's kinematics and ask the model to predict those kinematics using the surrounding information.
- **Step b) Test model:** Using the predicted lepton, neutrino, and b quark information, we can calculate the predicted kinematics for our $t\bar{t}$ system and compare to the m_{lb} method.
- **Step c) Finetune:** The model configuration that performed the best on our pretraining task is **stored and reused** on downstream tasks such as:
  - **Toponium Classification**
  - **Initial State Classification**
  - **Unfolding/Reconstruction**

Performance Plots

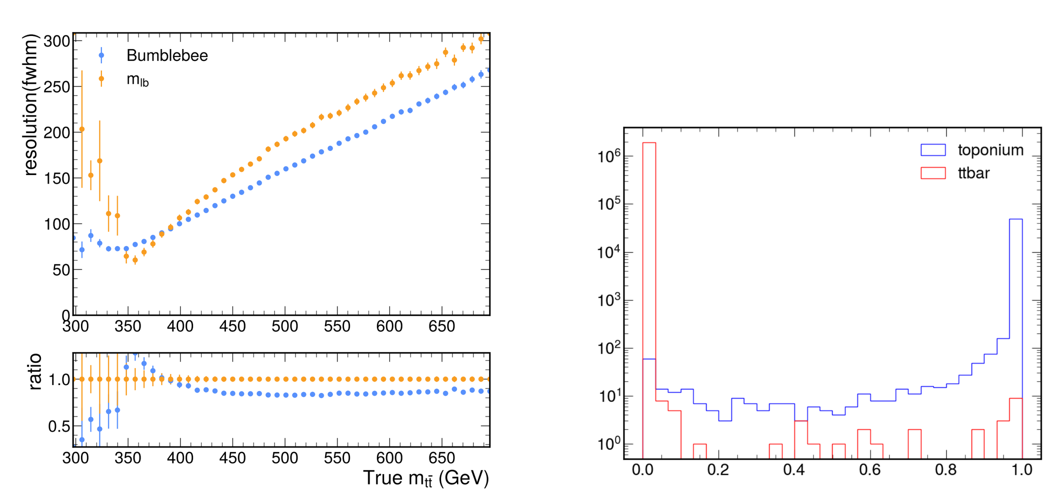


Figure 1: (Left) Resolution of $t\bar{t}$ mass compared to the analytic method of m_{lb} weighting.

Figure 2: (Right) Score distribution of toponium vs. $t\bar{t}$ classification (AUC: 0.9

Model Architecture

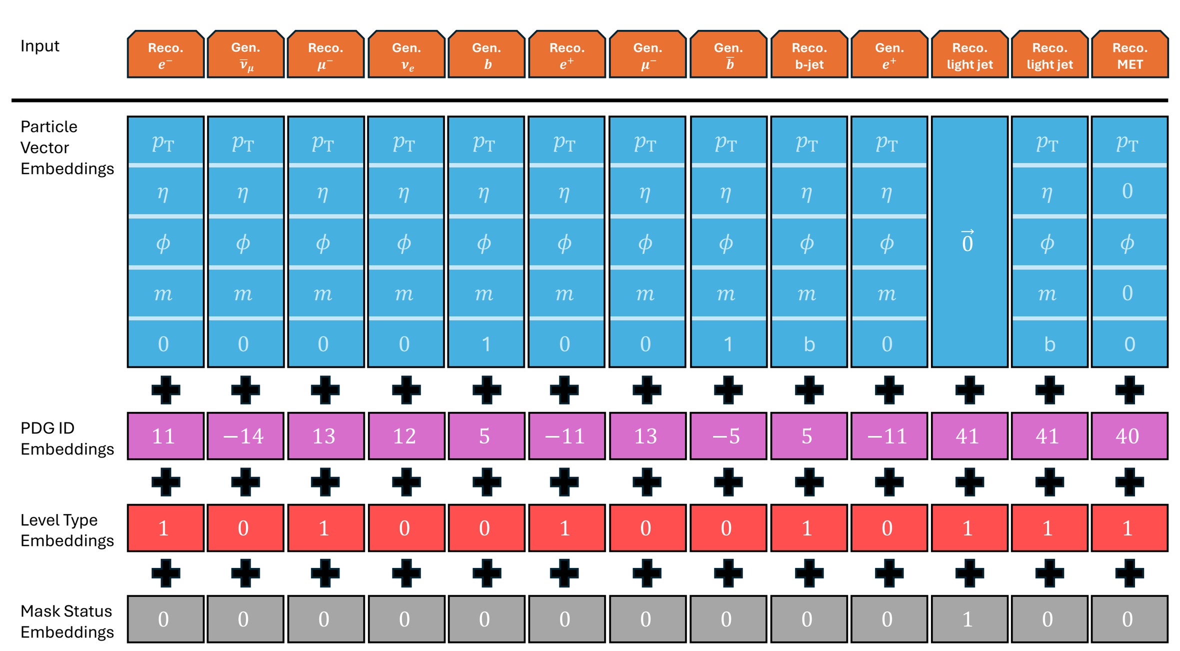


Figure 3: Bumblebee Embedding Structure

Conclusion

Future Directions:

- **Add generative tasks** such as detector simulation to the capabilities of Bumblebee.
- Search for more model-agnostic ways of **detecting Toponium** and understand why its predictions are so good.
- **Add mixed states** to Initial State Classification in an attempt to get a better performance on $q\bar{q}$ signal region.

Overall, the model has been shown to perform very well due to its ability to **gather inherent knowledge** of the system through pretraining via masked language modeling. Utilizing multi headed attention, the model can **generalize quite well** to different tasks such as classification and unfolding tasks, saving quite a lot of effort in inference tasks to make a specialized model for each new task that comes up.

Main References

[1] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, 2017.

[4] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2024.