Complex Data for High Energy Physics at CERN: Two Projects in Cluster Computing and Machine Learning

Vera Staub[1] | Dmitri Tsybychev[2]

[1]Wellesley College [2]Stony Brook University

Introduction

In the field of high energy physics, the most pressing, consistent endeavor is the processing and analysis of large quantities of data coming from CERN (Conseil Européen pour la Recherche Nucléaire/European Council for Nuclear Research) near Geneva, Switzerland. The Large Hadron Collider (LHC) at CERN is the world's largest particle accelerator. In this ring accelerator, subatomic particles collide at close to the speed of light. These events provide information about how the particles interact and give insight into fundamental natural laws. The Higgs boson was discovered in 2012 by the ATLAS and CMS collaborations with a mass of m_H = 125 GeV and the subsequent research of its properties has become of vital importance since then. The portion of Higgs bosons decaying into pairs of b-quarks is the largest among all possible decays. Therefore, the decay $H \rightarrow bb$ is important to shed light on the properties of the Higgs boson. The two projects detailed here represent some of the technologies in use for the task of analyzing data from Hbb decays.

Cluster Computing

Background and Methodology

One consequence of the high volume and complexity of the data produced by CERN is the difficulty of manipulating the data on one computer. Cloud storage, which has very widespread usage, employs machines in an interconnected system to store small amounts of data. In a similar way, cloud/cluster computing outsources computing work to a large cluster of computers. Stony Brook University has its own computer cluster, known as the SeaWulf cluster. We worked to utilize this cluster and others to run code analyzing data from high energy physics experiments. In the ATLAS experiment, information about the Hbb decay is taken in a format called xAOD. Then the processed datatsets are prepared by running the CxAOD framework. Based on CxAODs, various analyses can be done, e.g: data-vs-Monte Carlo comparisons, multivariate analysis. Final results can be extracted from histograms that are produced from CxAODs. It can be further processed using the CxAODReader package. Alternatively, a tuple can be produced in parallel using the CxAODMaker. The CxAOD framework can be run as a batch job or interactively. A batch job allows us to run other code on a local machine while the cluster handles the code separately, but an interactive job shows the progress of the job and can make errors or failures easier to understand. Moving code from one environment to another also often requires a container, which preserves a certain setting with all its characteristics and allows a user to essentially recreate that environment in another place.

Results and Conclusions

We were first able to successfully run the CxAOD framework on the sbaint cluster, which is specifically designated for high energy physics work at Stony Brook University. Next, we ran the framework on lxplus, CERN's computer cluster. This represents a step in adapting the framework to a new environment. We plan to move the framework to the SeaWulf cluster and build it there, which requires containers. Migrating our work from this external cluster to Stony Brook's own cluster will consolidate and facilitate future work and progress.

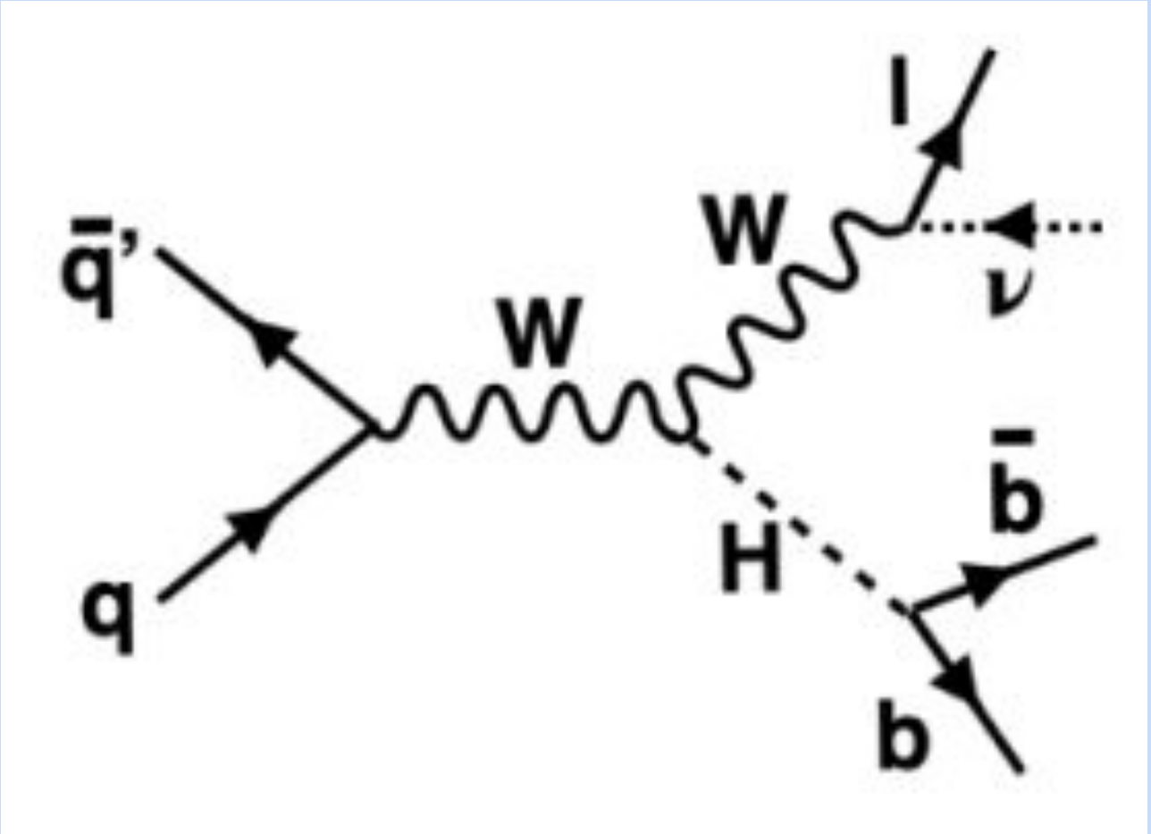

Fig. 2: diagram of VHbb process, where a vector boson is produced alongside the Higgs boson

Machine Learning

Background and Methodology

Much of the work done at CERN produces collision data that are cumbersome and often confusing or difficult to wrangle. The Deep Sets Neural Network (DSNN) is a tool to estimate modeling uncertainties for analysis of the Hbb process. Analysis of this process can be time-consuming, but the use of a neural network can help simplify. A neural network is a computer structure modeled on the human brain, intended to perform analytical or predictive tasks. In this case, the DSNN uses TensorFlow software for machine learning. Given unordered sets of particles' properties as event inputs and categorical labels as event outputs, the neural network can learn to simulate the events of interest and predict the behavior of particles to estimate shape uncertainties. Running the DSNN on the SeaWulf computer cluster would further streamline the analysis. The script is sent to run as a batch job on the SeaWulf cluster using a job scheduling software called Slurm, which takes in information such as job runtime, memory usage, or number of nodes required. Slurm then assigns the job a place in the correct queue to be executed when space is available. Running the initial script produces arrays for three data samples. These arrays are then combined and scaled. The outputs are split in half, with one one half going to training and the other half going to testing. It is also possible to plot data using ROOT, a software designed specifically for analysis of data in high energy physics.

Results and Conclusions

We hope to be able to utilize the SeaWulf cluster to run the DSNN training and plotting scripts. In past work, the DSNN has been run on other clusters, including the cluster at Brookhaven National Laboratory. The SeaWulf cluster is a very promising new option and resource for this neural network to handle the copious amounts of data coming from CERN's VHbb analysis. The cluster at BNL can execute jobs submitted through Slurm very quickly after they are submitted, so the Stony Brook cluster will be evaluated on the ability to execute jobs without delays and in comparison to the BNL cluster.

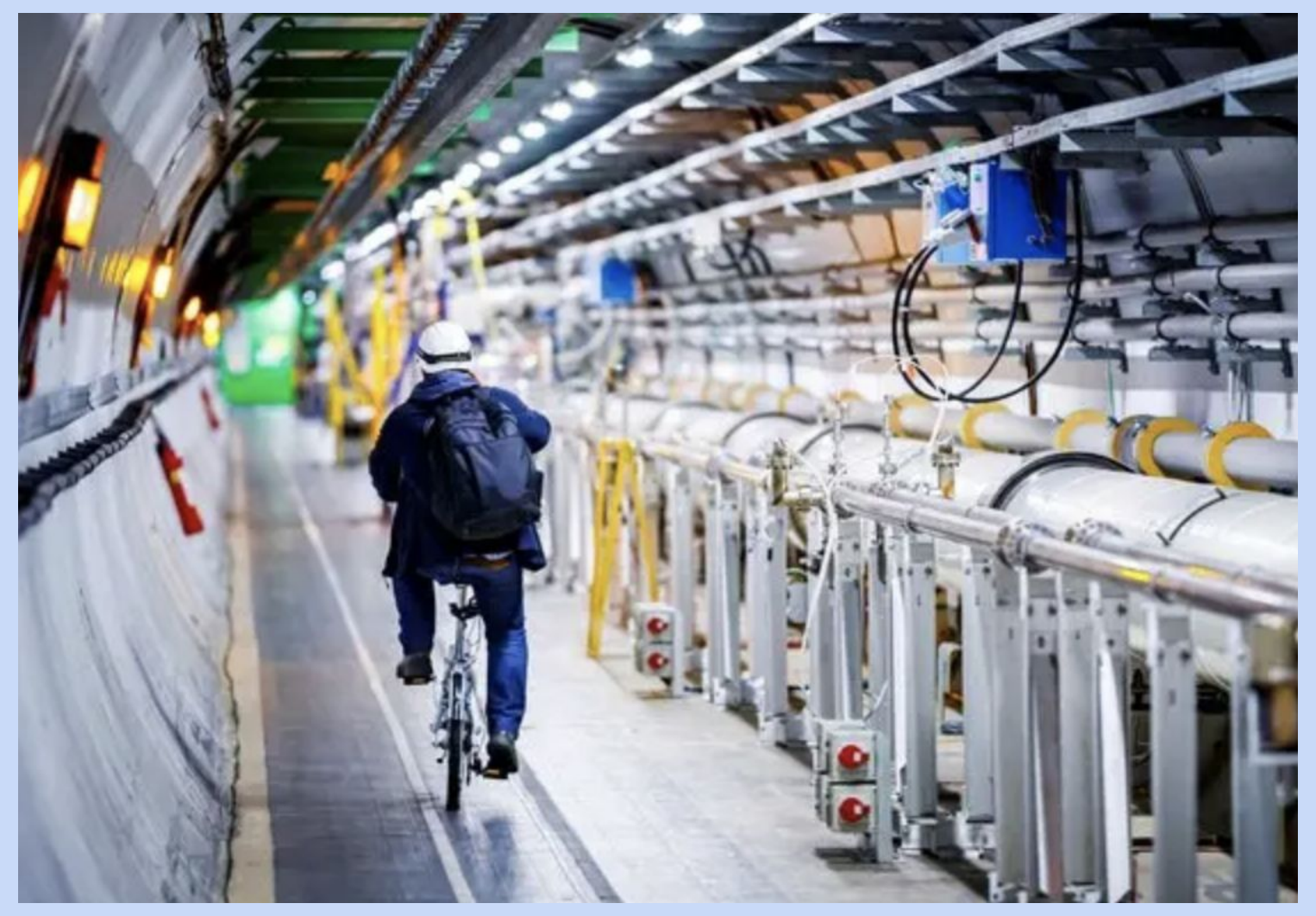

Fig. 1: a biker inside the LHC

Acknowledgements and References

Many, many thanks to Prof. Dmitri Tsybychev for his guidance and mentorship and to Yan Ke and Fang-Ying Tsai for their indispensable help. Thanks as well go to Prof. Navid Vafaei-Najafabadi for his leadership as the organizer of the REU program. This work was funded by National Science Fooundation award no. PHY-2243856.

Y. Chou et al. "Application of the Deep Sets neural network for evaluation of systematic uncertainties: $H \rightarrow bb$ ¯decays in association with a vector boson at ATLAS," March 16, 2023.
Han, T. "Collider Phenomenology: Basic Knowledge and Techniques," August 9, 2005, https://doi.org/48550/arXiv.hep-ph/0508097.