

Mitigating the Impact of Simulation Mis-Modeling on DNN Training: Building Robust DNNs in the Presence of Detector Mis-Modeling

Applicant: Andrii Len

Mentors: Dmytro Kovalskyi

Project duration: 12 Weeks

Proposed start date: 2 June 2025

Introduction and Project Description

The quality of physics analyses at the Large Hadron Collider critically depends on how well the Monte-Carlo (MC) simulation [1] reproduces detector response and event kinematics. Any significant *mis-modeling* between data and simulation can introduce biases that degrade the performance of multivariate classifiers and the reliability of physics measurements. Deep Neural Networks (DNNs), while powerful, are particularly vulnerable: when trained on imperfect simulation they may learn to discriminate *Data vs. MC* instead of *Signal vs. Background*, leading to sub-optimal sensitivity.

The rare decay $K_S^0 \rightarrow \mu^+ \mu^-$ [2] which we will focus on is an excellent laboratory to study such effects. In our previous research a DNN trained on simulation demonstrated promising background rejection, but further studies revealed non-negligible mis-modeling which should be taken into account in the form of huge corrections which impair the sensitivity of the analysis.

There are multiple ways to address this issue, such as training solely on data samples (means not using MC at all) or modifying the loss function to include penalty terms for mis-modeled features[3][4] to prevent model from learning what it is not supposed to learn. The main drawback of the first approach is that it requires a signal-like control region with sufficient statistics, which is not always available (in our case we, fortunately, have such a signal-like control region), whereas the MC-vs-data strategy can be applied to any relevant process, making it much more versatile and if done correctly, we think, it can potentially be more powerful. In this project, we will compare these two methods to assess their relative performance and identify common trends.

Software Deliverables

All developments will be implemented in Python using mostly TensorFlow/Keras, PyTorch, and scikit-learn packages for ML and other auxiliary ones for data preparation.

Preliminary Timeline

Week 1-2 Using 2022-2024 CMS[5] data to perform Data/MC comparison studies of existing variables. Study the materials on how to apply efficient penalty to model's training.

Weeks 3-4 Training an optimal Data-vs-Data model. Understanding how it utilize different features.

Weeks 5-10 Develop an advanced MC-vs-Data architecture with modified loss function to include penalty terms for mis-modeled features.

Weeks 10-11 Comprehensive performance assessment. Comparison of two methods.

Week 12 Prepare final report and presentation.

References

- [1] S. Agostinelli *et al.*, “GEANT4 — a simulation toolkit,” *Nucl. Instrum. Meth. A* **506** (2003) 250–303.
- [2] A. Dery, M. Ghosh, “ $K_S^0 \rightarrow \mu^+ \mu^-$ as a clean probe of short-distance physics” (2021), arXiv:2104.06427.
- [3] Walter, David. (2018, September 3). Domain Adaptation Studies in Deep Neural Networks for Heavy-Flavor Jet Identification Algorithms with the CMS Experiment
- [4] E. M. Metodiev, B. Nachman, “Classification without labels: Learning from mixed samples in high energy physics” (2017), arXiv:1708.02949.
- [5] CMS Collaboration, “The CMS experiment at the CERN LHC,” *JINST* **3** (2008) S08004.