CMS ML KNOWLEDGE GROUP NEWSLETTER

November 2024, v2

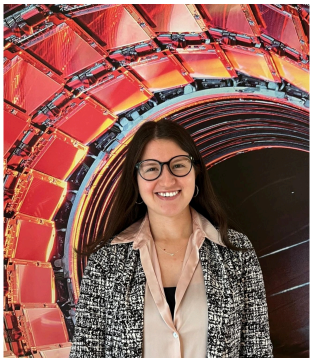
Welcome to the second edition of the CMS ML Knowledge Group Newsletter

Quick Peak

- Scholar Spotlight: Meet Valentina Guglielmi from DESY
- ML Corner: Meet Florian Eble from ETH Zurich
- Events on our Radar
- Help Wanted

SCHOLAR SPOTLIGHT

Meet Valentina Guglielmi, a Ph.D student at DESY



Could you tell us a little more about your background and how you got into particle physics?

I completed my bachelor's and master's degrees at the University of Milano-Bicocca, and I'm currently pursuing my PhD at DESY in Hamburg. I've been fascinated by science since I was young, but I decided to follow particle physics specifically in 2012. It was July, and I was on a trip to Geneva. During that trip, I had the chance to take a guided tour of CERN. When we entered the main auditorium—where just some weeks before the discovery of the Higgs boson had been announced—I decided at that moment that I wanted to be a particle physicist and contribute to a CERN experiment.

Could you tell us a little more about your research?

In my PhD, I am working on several topics, including QCD, Top quark physics, and Machine Learning (ML). A key focus of my research is applying the Deep neural networks using Classification for Tuning and Reweighting (DCTR) method developed by A. Andreassen and B. Nachman to reweight Monte Carlo (MC) samples of top quark production in the CMS experiment. Accurate simulations of particle collisions and detectors are critical in particle physics, but they come with significant computational costs, especially for detector simulations. My work uses ML to reweight these samples for different model parameters or entirely different models, significantly reducing the need for multiple simulations and cutting computational costs. This ML-based approach is already used in CMS but it will become more and more crucial for the High-Luminosity LHC.

A lot of us use centrally provided ML taggers and frameworks. But, you work in designing something brand new. What is a challenge you have faced with your research and how have you overcome it?

It was very fascinating to try to do something new. One of the main challenges I faced was learning how to navigate this unfamiliar territory. In particular, the novelty of the project lies in adapting an existing ML method to a concrete and demanding context, such as a physics analysis at the LHC. Achieving the level of precision required for analysis in a chaotic environment like the LHC was particularly challenging. This required a significant amount of time to fine-tune the method, as well as to find clever tricks to reach the high level of accuracy needed (within a few percent).

Do you have any advice for researchers who may want to get involved with creating a new ML approach like you have?

I especially encourage young scientists to explore machine learning, as it will be crucial for the future. My advice is to focus as much as possible on understanding the underlying concepts. At first, machine learning may seem like a black box, but over time, you'll learn valuable tips and tricks that can help you improve the knowledge and accuracy of your results.

Where are you in your academic journey? What are you interested in doing in the future?

I am currently in the final year of my PhD. So I am facing the most dreaded moment for any PhD student: writing my thesis :) After my PhD, I want to continue my academic career. My dream is to work at CERN, since it is a very international environment.

Is there anything else you would like to share?

In my free time, I love hanging out with friends and dancing. I think it’s really important to have a hobby. Doing something different can actually spark new ideas for our work!

ML CORNER

Meet Florian Eble, a PhD student at ETH Zurich

How did you start working on normalized autoencoders?

It all started at a workshop on semivisible jets in July 2022 at ETH, where Barry Dillon presented the work he and his collaborators did on normalized autoencoders [1]. At this time in our group, we were trying to understand why the autoencoder we used to identify semivisible jets worked well against QCD but not top jets. I got immediately very excited about the normalized autoencoder paradigm! The idea of being able to gain insight into what the autoencoder was actually learning was an obvious step to move forward on the issue we were facing for our search.

Can you tell us a little more about the Wasserstein normalized autoencoder (WNAE) that you designed? What makes this type of autoencoder different from other autoencoders?

The basic idea behind using autoencoders for anomaly detection is to train an autoencoder such that it learns how to reconstruct with low error the examples from the data it is trained on, but fails to do so on different, anomalous, data. Translated to searches for new physics, this means training an autoencoder on standard model events/jets in order to identify anomalous events/jets in an unsupervised fashion. The main shortcoming of standard autoencoders is that they are free to reconstruct with low error examples outside the training phase space! This is called outlier reconstruction. The fact that they can generalize outside the training phase space, usually good for most ML applications, is actually detrimental to using the reconstruction error of the autoencoder as a discriminator to detect anomalies, as anomalies can have as low reconstruction error as the training data.

The concept of normalized autoencoders (NAE) was coined in 2021 by Sangwoong Yoon, Yung-Kyun Noh, and Frank Park [2], who provided a strategy to suppress outlier reconstruction, borrowed from energy-based models. The key idea is to ensure that the higher the probability density of the training data, the lower the reconstruction error of the autoencoder. To achieve this, the probability distribution of the autoencoder is defined as the Boltzmann distribution where the energy is the reconstruction error of the autoencoder. By sampling from this probability distribution via a Markov chain Monte Carlo (MCMC), the so-called negative samples, which follow the probability distribution of the autoencoder to have low reconstruction error, can be obtained! The loss of the NAE is defined as the difference between the reconstruction error of the training samples and the negative samples. This is theoretically equivalent to minimizing the negative log-likelihood of the training data given the probability distribution of the autoencoder! Although this formulation is very sound theoretically, we found that it suffers from a number of failure modes, and that a more robust measure of the distance between the training data probability distribution and the autoencoder probability distribution can be achieved by computing the Wasserstein distance between the training samples and the negative samples. This is how the Wasserstein normalized autoencoder was born!

Could you briefly describe some of the failure modes of a normalized encoder and how your Wasserstein normalized encoder overcomes them?

The first failure mode of normalized autoencoders is that the loss can be negative and diverge! The reconstruction error of the negative samples can be higher than that of the training samples, resulting in negative loss. And as a consequence, the network is incentivized to diverge. This was traditionally addressed by the means of ad hoc regularization, which we found to be unsatisfying: the network will most likely converge to a configuration where the reconstruction error of the negative samples is different from that of the training samples. In this case, the autoencoder probability distribution is necessarily different from that of the training data! In addition, it is possible for different phase space regions to be energy-degenerate and have the same reconstruction error, in which case the NAE fails to suppress outlier reconstruction. These failure modes are naturally overcome by measuring the distance between the training data probability distribution and the autoencoder probability distribution using the Wasserstein distance.

What kinds of computing tools and libraries did you use?

The WNAE code is written in PyTorch. The flexibility of this library makes it a natural choice to implement the MCMC and compute the Wasserstein distance, differentiable with respect to the autoencoder's weights. This is technically quite challenging, as the gradient with respect to the feature space needs to be calculated to compute each MCMC step, but at the same time the dependence on the neural network weights must be kept in order to compute the gradient with respect to the weights for the backpropagation. The Wasserstein distance is computed using the POT library, providing an exact solution to the optimal transport problem.

What type of physics study did you do to test your WNAE?

We developed the WNAE in the context of searching for dark matter, in theories where the dark sector is made of particles that interact via a new confining force. In analogy with the standard model, this new interaction is called dark QCD. The experimental signature of such a dark sector is the production of semivisible jets: jets of visible standard model particles and invisible dark matter states, with different internal structure than standard model jets. Because of the many parameters that a complete dark QCD theory would depend on (e.g. number of dark flavors, colors, dark hadron masses, dark hadronization scale) and the dependence of the specific features of semivisible jets on these parameters, this search is an ideal physics case where anomaly detection offers a complementary approach to traditional supervised strategies. We are looking forward to applications to different physics cases!

Anything else you would like to add?

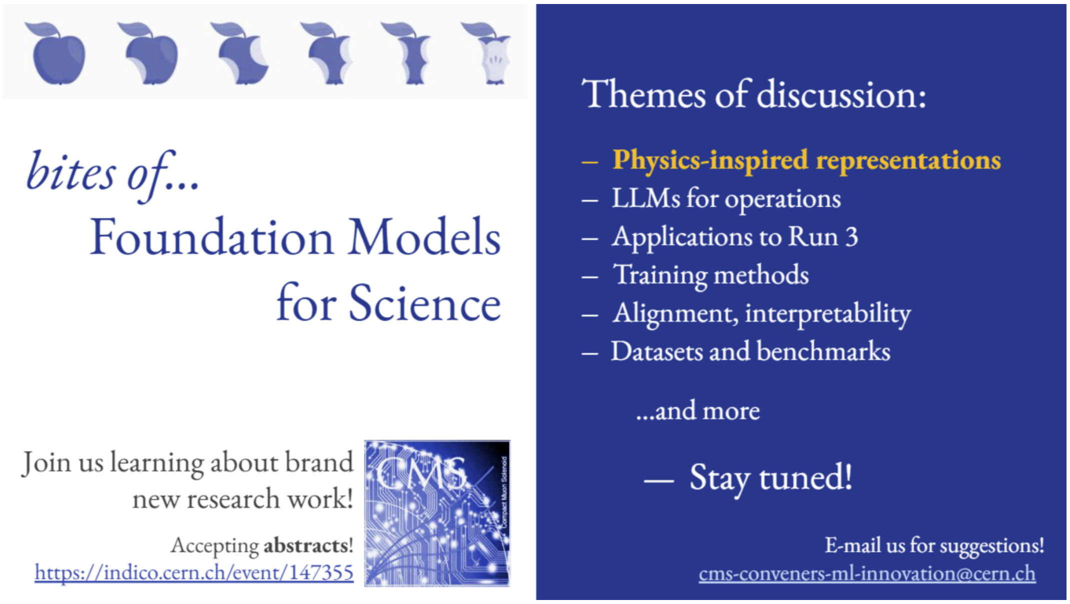
The ongoing developments in anomaly detection in CMS and outside open new doors in HEP! I think we are all very excited to see more usage of anomaly detection for new physics searches as well as for triggers!

References:

[1] B.M. Dillon, L. Favaro, T. Plehn, P. Sorrenson and M. Krämer, A Normalized Autoencoder for LHC Triggers, arXiv:2206.14225

[2] S. Yoon, Y.-K. Noh and F. Park, Autoencoding under normalization constraints, in ICML, pp. 12087 12097, PMLR, 2021, arXiv:2105.05735

EVENTS ON OUR RADAR



bites of...
Foundation Models
for Science

Join us learning about brand
new research work!

Accepting **abstracts**!
https://indico.cern.ch/event/147355

CMS
Compact Muon Solenoid

Themes of discussion:

- **Physics-inspired representations**
- LLMs for operations
- Applications to Run 3
- Training methods
- Alignment, interpretability
- Datasets and benchmarks

...and more

— Stay tuned!

E-mail us for suggestions!
cms-conveners-ml-innovation@cern.ch

Bites of Foundation Models for Science: 2024: Nov. 20

Following up on the Foundation Models for Science mini-workshop, the CMS ML innovation team is excited to announce Bites of Foundation Models for Science, deep diving into specific themes in this area.

The first of these bites will be held remotely on Nov. 20th 2024 (15:00-19:00 CERN time) with the theme of learning representations inspired by and aimed for physics applications. We aim to connect researchers from CMS, ATLAS, and beyond HEP working on building powerful, robust representations for science.

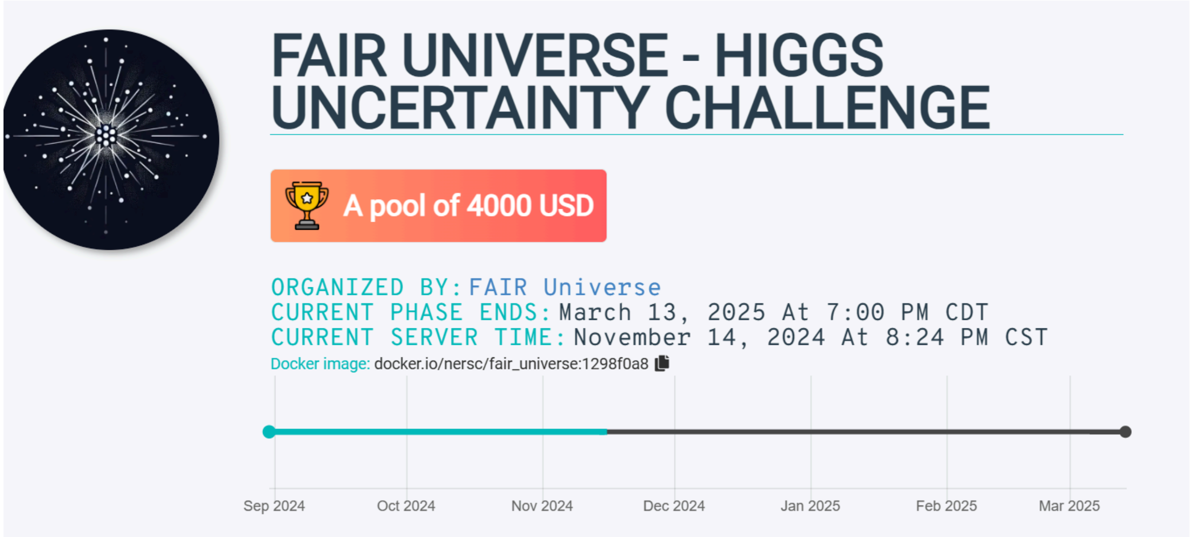
Join at: https://indico.cern.ch/event/1473554/

NeurIPS: 2024: Dec 9-15

HACKATHONS

Join the **CERN FAIR Universe & HDR ML Challenge**: 2024: Dec. 14
This hackathon explores uncertainty-aware AI techniques for HEP, and also features monetary prizes!

To learn more:
https://www.codabench.org/competitions/2977/
https://fair-universe.lbl.gov/Uncertainty-Challenge-Workshop.html
https://indico.cern.ch/event/1477109/



FAIR UNIVERSE - HIGGS UNCERTAINTY CHALLENGE

A pool of 4000 USD

ORGANIZED BY: FAIR Universe
CURRENT PHASE ENDS: March 13, 2025 At 7:00 PM CDT
CURRENT SERVER TIME: November 14, 2024 At 8:24 PM CST
Docker image: docker.io/nersc/fair_universe:1298f0a8

JOIN US

This newsletter was brought to you by the CMS knowledge group. We meet every three weeks, and welcome new members!

If you've enjoyed this newsletter, please let us know. Also, if you have an idea for something you would like to see in this newsletter, would like to nominate someone for an interview or ML corner spotlight, please let us know!

Contact: Melissa Quinnan & Jieun Yoo at: cms-conveners-ml-knowledge@cern.ch

HELP WANTED

Do you need EPR? Join the CMS ML Knowledge Group. We are looking for contributors. See our most recent list of tasks here (note: CMS internal web only)

Do you LLM? CMS is investigating LLMs: See this link: (note: CMS internal web only)