Early Career Research Proposal:
Uncertainty-Robust Machine Learning

Applicant/Institution: Walter Hopkins, Argonne National Laboratory
Postal Address:
Argonne National Laboratory, 9700 S. Cass Avenue, Building 360, Lemont, IL 60439
PI name: Walter Hopkins
Position title for PI: Physicist
PI telephone number, email: (630) 252 7551, whopkins@anl.gov
Administrative Point of Contact name, telephone number, email:
Diane Hart, (630) 252 7677, dhart@anl.gov
FOA Number: DE-FOA-0003176
DOE/SC Program Office: HEP
Topic Area: Experimental Research at the
Energy Frontier in High Energy Physics
DOE/SC Program Office Technical Contact: Abid Patwa
Year Doctorate Awarded: 2013
Eligibility Extension Included in Approved Pre-application: No
Does the PI have tenure: No
Number of Times Previously Applied: 2
PAMS Pre-application Number: PARE-0000036282
PECASE Eligible: Yes
Proposal Contains Data Management Plan in Appendix 4: Yes
Proposal Contains PIER Plan in Appendix 5: Yes

Contents

1 Introduction

Experiments at the Large Hadron Collider (LHC) have confirmed many predictions of the highly successful Standard Model (SM), including the landmark discovery of the Higgs boson [1, 2]. Despite its success, the SM does not account for several observed phenomena, such as dark matter and the matter-antimatter asymmetry, thereby fueling the search for Beyond the Standard Model (BSM) physics. Future LHC upgrades will shift focus from increasing energy to enhancing precision, marked by a tenfold expansion of the Run 2 dataset in the High Luminosity-LHC (HL-LHC) phase. Crucially, enhancing the identification efficiencies of physics objects, such as photons, and minimizing systematic uncertainties are vital for future breakthroughs in collider physics. Enhanced object identification effectively increases the usable data, akin to extending LHC operation times, while reducing uncertainties heightens the sensitivity to subtle deviations from SM predictions.

Machine learning (ML) has increasingly become a critical tool in High-Energy Physics (HEP), offering significant advancements in various tasks such as identifying particles and distinguishing between signal and background processes. ML techniques allow physicists to leverage complex correlations among a wide range of observables, from the trajectories and energies of particles to their interactions within detectors. However, the application of ML in HEP is accompanied by certain challenges that need careful consideration: 1) Sensitivity of ML models to differences between Monte Carlo simulations and recorded data, which can introduce additional uncertainties into ML predictions and affect the overall systematic uncertainties in physics measurements; 2) Sensitivity of ML algorithms to experimental parameters (e.g., the energy resolution of a subdetector) which increase systematic uncertainties in the final result; 3) Tendency of ML models to sometimes create unintended correlations between their outputs (such as estimated energy or particle identification) and other variables critical for calibrations or background estimations.

Recent applications of adversarial [3] and distance correlation [4] techniques have shown promise in reducing uncertainties due to data-simulation differences in a long-lived particle search [5] and in decorrelating jet substructure variables from jet mass [6]. These approaches are part of a group of methods that aim to improve the 'domain adaptation' of ML algorithms, i.e., the ability of an ML model trained with one dataset (e.g., simulations) to be robust enough to yield similar performance on a different dataset with somewhat different features. This robustness can be achieved in several ways, for example, by including the alternative dataset during training and penalizing differences in performance.

domain-adaptation techniques can enhance the resilience of ML models used in HEP against data-simulation discrepancies, varying detector and accelerator conditions (e.g., the number of simultaneous proton-proton interactions, known as pile-up), and changes in the underlying parameters of simulations, such as those used for estimating the properties of physics objects (e.g., photon energy, jet transverse momentum, etc). These techniques also aid in identifying (ID) physics objects (e.g., photons, jets containing b-hadrons, etc). Thus, incorporating domain adaptation into the HEP ML workflow can significantly reduce the overall uncertainties in physics results.

This proposal **outlines the development of a framework designed to facilitate the deployment of various domain-adaptation techniques within existing and future HEP ML workflows. This framework aims to ensure that ML models are significantly more resilient against experimental systematic uncertainties, data-simulation discrepancies, and changes in detector or collider conditions.** Given recent initiatives within HEP to adapt ML models for use with field-programmable gate arrays (FPGAs), the proposed framework will also enhance ML models deployed at the trigger level. Furthermore, the PI's team will introduce a new-to-HEP domain adaptation method, "feature representation transfer", that adds the ability to produce generally applicable features as compared to the already applied-in-HEP methods of distance correlation and adversarial networks which are more application specific.

Precise measurements of SM processes, particularly those involving Higgs bosons, play a crucial role in probing for BSM effects caused by particles too massive to be directly produced at the LHC. Higgs differential cross-section measurements, i.e., as a function of kinematic properties such as Higgs p_{T}, are critical probes for BSM physics because they are sensitive to deviations in the Higgs boson self-coupling and its coupling to the top quark. The latest ATLAS $H(\rightarrow \gamma\gamma)$ differential cross-section measurements, which include the associated production of a Higgs with a top pair ($t\bar{t}H$), are limited by statistical uncertainties in each kinematic bin. Enhancing object ID efficiencies, especially for photons which improves the statistical power of the measurement, and reducing dominant detector-based systematic uncertainties, such as photon ID and energy resolution, are crucial for improving these measurements and thereby maximizing sensitivity to BSM physics.

The proposed framework and techniques will initially be developed and applied to photon ID and energy. Studies within ATLAS have shown a potential improvement of approximately 10-20% in both ID efficiency (equivalent to collecting more data) and energy resolution when all available information for ID and energy estimation is included. However, once calibration—a process that involves evaluating ID efficiencies and properties in simulations and adjusting them to match true properties before reconstruction or to match observed data, the gains in efficiency and enhancement in resolution were lost. For the ID, ML approaches were found to correlate with quantities that must remain independent of the ID for the calibration method, while for energy evaluation, ML methods were sensitive to mis-modeling of shower shapes in the simulation. Therefore, the proposed domain adaptation framework and approaches could enhance all ATLAS measurements involving $H(\rightarrow \gamma\gamma)$, which are vital to the HL-LHC BSM search program.

As collider physics enters the precision era and ML becomes increasingly integral to our methodologies, it is crucial to refine ML models to achieve heightened precision in anticipation of the HL-LHC startup. My extensive background in ML, demonstrated through leadership roles such as convening the ATLAS ML Forum and co-developing a novel feature extraction technique [7], aligns perfectly with the goals of this project. Additionally, my deep experience with calorimetry and simulations, combined with my involvement in prominent searches involving $t\bar{t}$ final states and recent analyses using two-photon final states, equips me uniquely to lead the proposed research.

At Argonne the advanced computing resources that have supported substantial ATLAS group analyses and ML initiatives present a distinct advantage. These facilities will play a critical role in developing and implementing the proposed ML framework. This convergence of timing, expertise, and resources at Argonne makes it the ideal moment and place for this essential research. By advancing ML techniques now, we can ensure that the HL-LHC's full potential for discovery is realized.

1.1 Precision Differential Cross-Section Measurements at the HL-LHC

The HL-LHC will deliver a vast HEP dataset providing opportunities to perform new measurements and with unprecedented precision. Many measurements which were statistically constrained will become systematic uncertainties which will underscores the pivotal role of systematic uncertainties in shaping the landscape of precision measurements within the HEP. There are, however, measurements with potentially significant BSM contributions that will benefit from an even greater increase of the dataset size. Such an increase is only possible by improving the ID efficiency of physics objects that are expected to be present in various final states.

Measurements involving the SM Higgs become increasingly sensitive to high mass BSM contributions as the precision of measurements improves. As the size of the LHC dataset grows, differential cross-section measurements, which can further expose the effects of BSM physics in high-energy regions (see Figure 1 for an illustration of this effect), become more powerful tools in the search for BSM physics. The results of differential cross-section measurements can probe BSM effects by being interpreted in the context of the Standard Model Effective Field Theory (SMEFT) [8–10] and in the case of Higgs-related measurements, results can also be interpreted in terms of coupling strengths within the κ-framework [11].

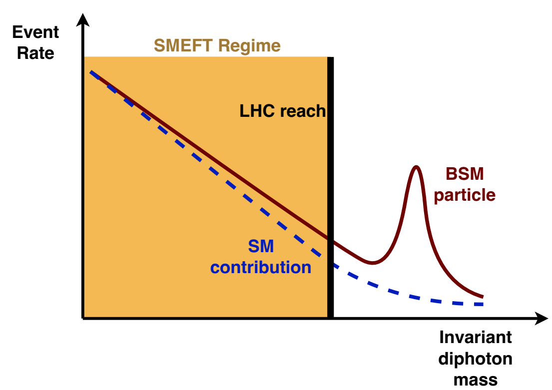

Figure 1: Example of how effects of a heavy BSM particle can leak into high-energy regions of distributions measured at LHC experiments.

A recent differential cross-section measurement involving the decay of the SM Higgs to two photons [12] utilized the Simplified Template Cross Section (STXS) method. This approach measures cross-sections in bins across several kinematic dimensions to constrain BSM effects via the SMEFT interpretation. Although current measurements are statistically limited and suffer from substantial systematic uncertainties—up to approximately 40% in some differential cross-section bins—the HL-LHC's larger dataset will mitigate statistical limitations in most bins, thus amplifying the significance of reducing systematic uncertainties. However, bins in high-energy regions, such as those measuring properties of a Higgs produced in association with a top pair ($t\bar{t}H$), will still face statistical limitations. Improving photon ID efficiency could reduce statistical uncertainties by up to 17% in final states that include two photons. This improvement requires ensuring that photon ID is uncorrelated with track isolation. The methods, such as distance correlation and adversarial techniques, incorporated into the proposed framework can be applied to a new ML-based photon ID algorithm to achieve increased statistical power in topologies with two photons."

1.2 Domain Adaptation for HEP ML Models

Machine learning has been utilized for the identification of various physics objects, such as electrons and jets containing b-hadrons. However, many of these ML models rely on simulations for training and do not automatically account for data-simulation differences during this phase—this issue is typically addressed during calibration. Furthermore, algorithms may need to remain independent from variables that are part of the calibration process, for example, those used to determine the rate of fake objects [14], or variables that may vary during collider or detector operations, such as pile-up.

Table 1: domain-adaptation techniques that will be studied and applied to various tasks detailed in this proposal.

Method	**Maturity in HEP**	**Main Benefit**	**Compute Intensity**
Adversarial	used	complex correlations	high
Distance Correlation	used	non-linear correlations	low
Feature Transfer	not used	features for multiple tasks	medium

Enhancing the robustness of ML models used to identify objects and assess their properties—thus mitigating these effects—will have far-reaching impacts on physics results.

There have been recent advancements in using domain-adaptation techniques to decorrelate ML models from certain quantities using methods such as adversarial discriminants and distance correlation. Adversarial approaches involve two ML models, one that generates high-dimensional distributions and another model that is trained to discriminate between the ML-generated data and the real data. More recently, adversarial approaches have been used to discriminate between recorded and simulated data from the output of another discriminator [5]. This strategy, i.e., to impose penalties on an ML model for producing inconsistent results across various datasets, can be applied to minimize differences between data and simulations or between simulation datasets with different parameter values (e.g., the energy resolution of jets). This is achieved by including two optimization quantities (i.e., cost or loss functions), one for each ML model. Using two opposing ML models to improve domain adaptations allows for complex multi-dimensional correlation to be taken into account. However, this dual-faceted optimization process introduces a significant computational demand and presents challenges in establishing definitive criteria for the cessation of training.

An alternative method that requires less computing resources, but may not be able to take advantage of as much information as adversarial approaches, is using distance correlation which summarizes the dependence of sets of variables. Unlike the more frequently used measure of the Pearson correlation coefficient, which can only detect linear correlations, distance correlation is sensitive to non-linear correlations, which are commonly present in HEP datasets. The distance correlation is zero only and only if there are no correlations between the input variables. This feature allows it to easily be added to typical loss functions which are minimized to optimize ML models. Thus, distance correlation is simpler to implement than adversarial techniques with less hyperparameters (since only a loss term is added rather than an entire NN) and more stable training characteristics. Another, less explored within HEP, domain adaptation technique is feature representation transfer. The idea is to learn a feature representation that is domain-invariant. Techniques such as autoencoders networks (which take inputs and encode them into a lower-dimensional space) can be used to learn such representations where the model minimizes the difference between the source (e.g., simulation) and target (e.g., recorded data) domain features, making the model's predictions less dependent on the domain-specific features. The advantage of feature representation transfer over distance correlation and the adversarial techniques is that the resulting domain-adaptive features can be used for a variety of tasks. These features would be akin to physicist developed variables that are ratios of quantities that behave similar after a change in assumption (e.g., both the numerator and denominator increase at a similar scale if an energy scale is changed).

Domain adaptation allows ML models to maximize the use of all information while avoiding regions of kinematic phase space that are poorly modelled in simulation or to ensure an ML model is independent of a quantity, such as pile-up or isolation. Depending on the particular application, adversarial, autoencoder based feature representation transfer, or distance correlation approaches can be the optimal solution and all should be studied when developing a particular ML model. A summary of the approaches that will be developed and applied are listed in Table 1.

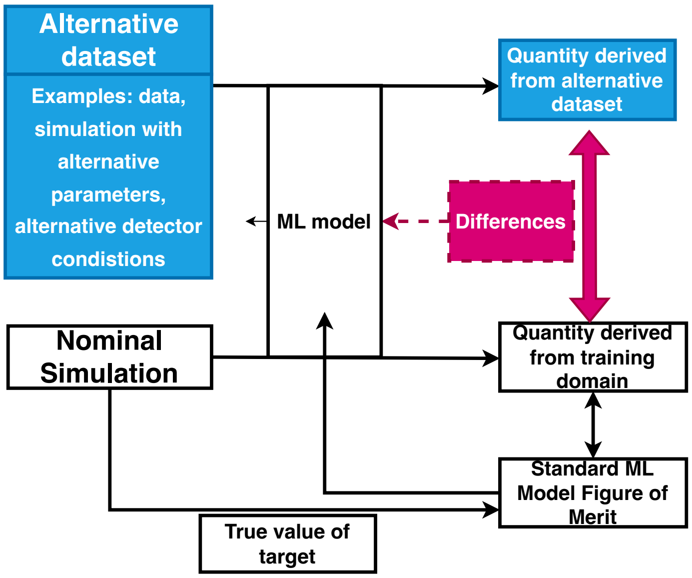

Figure 2: Sketch of the proposed framework. The pink box and arrow highlight the new parts of ML training workflows. The "true value of target" represents known quantities in the simulations such as the event classification (signal vs background), object identification (real photon vs fake), or a quantity such as the true energy of a photon before it is measured by a detector.

Domain adaptation allows ML models to maximize the use of all information while avoiding regions of kinematic phase space that are poorly modelled in simulation or to ensure an ML model is independent of a quantity, such as pile-up or isolation. Depending on the particular application, adversarial, autoencoder based feature representation transfer, or distance correlation approaches can be the optimal solution and all should be studied when developing a particular ML model. A summary of the approaches that will be developed and applied are listed in Table 1.

Recently, ML frameworks that are multi-modal, i.e., they can handle various types of input data (particle trajectories, jets, etc) and that are multi-task, that is they can build complex ML models for various tasks (identifying jets which contain hadrons that are initiated charm or bottom quarks, jet energy corrections, etc). SALT [15], is such a framework developed within ATLAS but not limited to use within ATLAS. These frameworks are powerful tools but lack infrastructure to automatically produce models that are robust against simulation mismodelling, experimental systematic uncertainties, or pile-up (see Fig. 2 for a depiction of how domain adaptation fits into the current training paradigms). Incorporating domain-adaptation techniques within SALT will reduce systematic uncertainties which will propagate through to many physics results.

1.3 Initial Applications of Domain Adaptation Framework: Photon ID and Energy

Machine-learning-based photon ID techniques (a boosted-decision-tree, BDT) have been studied within ATLAS demonstrating a potential gain in signal efficiency of 5-10% (e.g., moving from 88% to 95% efficiency for unconverted photons with $p_{\mathrm{T}}{\sim}60$ GeV) resulting in an increase of statistics of 10-20% (thus reducing the statistical uncertainty by 10-17% for events with two photons) while retaining the same background rejections as the current rectangular-requirement-based ID. An increase in efficiency would impact all $H(\rightarrow \gamma\gamma)$ measurements by improving the statistical power of these measurements with the same amount of HL-LHC data while increasing the background rejection (once the signal efficiency has saturated) will result in an even purer $H(\rightarrow \gamma\gamma)$ signal. Even more efficiency gains and increased background rejection have been observed when using convolutional neural networks (CNNs). However, photon ID calibration procedures and fake-photon background estimation techniques used in many ATLAS analyses necessitates that the photon ID is uncorrelated from track isolation (a measure of the energy associated with the photon candidate vs the energy around the candidate). This requirement of orthogonality to particular quantities has limited the use of ML for photon ID within ATLAS. Additionally, the Phase II upgrade of the ATLAS Inner Tracker (ITk) will improve the track isolation, which yields an opportunity to revisit the ATLAS photon ID and its calibration to maximize performance during the HL-LHC.

The evaluation of photon energies was studied using both advanced ML architectures, i.e., CNNs and graph neural networks (GNNs), and more information than the current BDT-based algorithm used. The resulting calibration showed a $\sim$20% improvement in photon energy resolution but was sensitive to simulation mismodelling in variables that summarized the shape of the energy deposits in the calorimeter. Ensuring that ML models are insensitive to the observed mismodelling in certain corners of the dataset could result in significantly improved photon energy resolution which in turn will reduce the systematic uncertainties for analyses that include photons in their final states (which includes all channels that involve $H(\to \gamma\gamma)$ decays such as the Higgs mass measurement).

2 Project Objectives

The objective of the proposed work is to enhance the discovery potential of the HL-LHC by both improving the statistical power of measurements and by reducing the effect of systematic uncertainties on precision measurements. This is enabled by the incorporation of domain-adaptation techniques into an ML framework. Domain adaptation will first be applied to the ATLAS photon ID and energy evaluation which will in turn be used in the upcoming $H(\to \gamma\gamma)$ differential cross-section measurements. The proposed domain-adaptation techniques will also be used in the optimization of $H(\to \gamma\gamma)$ to improve background rejection. The ML framework resulting from the proposed work will provide streamlined access to domain-adaptation techniques that automatically mitigate the effects of systematic uncertainties and simplify calibration tasks (e.g., adjusting simulations to better match recorded data), for a wide range of applications within HEP.

- Improve the precision of future HEP measurements
    - By incorporating resistance to parameter changes (e.g., resolution in the assessment of systematic uncertainties), data-simulation discrepancies, and changing detector conditions into ML models through the use of domain-adaptation techniques
- Create a framework that inherently integrates domain-adaptation techniques during the development of ML models
- Use ML framework with domain adaptation to improve Run 3 $H(\to \gamma\gamma)$ differential cross-section measurement

3 Proposed Research and Method

The PI's team, consisting of two postdocs at 2.0 FTE, will customize and deploy domain-adaptation techniques within an ATLAS common framework, i.e., "SALT" [15]. These techniques will first be applied to improve photon energy evaluation and photon ID. The more established approaches, distance correlation and adversarial networks, will be studied in parallel with the feature representation transfer. One postdoc will gain expertise in photon energy evaluation while the other postdoc will work on photon ID using the same ML domain-adaptation methods. Both postdocs will contribute to developing feature representation transfer and will work on preparing the next two iterations of the $H(\rightarrow \gamma\gamma)$ differential cross-section measurement.

domain-adaptation techniques will first be applied in conjunction with BDTs and will then be adapted to be used with more advanced networks such as GNNs. The adversarial approaches as well as GNNs require significant computing resources to train and tune. The PI will leverage both computing resources and expertise at Argonne to develop domain adaptive models. The team will assess the effectiveness of domain-adaptation techniques by examining residual correlations and comparing the performance of ML models with and without domain adaptation. Upon successful validation and integration of these techniques into SALT, the PI's team will with members of the Argonne ATLAS group (who will not be funded by this proposal, if awarded), who are working within the ATLAS flavor tagging group, to implement the domain-adaptation approaches for flavor tagging.

In parallel to preparing domain adaptation for use in ATLAS physics object ID and property assessment, the PI's team will contribute to two differential cross-section measurements of $H(\rightarrow \gamma\gamma)$: an intermediate analysis using a subset of Run 3 data, and a comprehensive final result utilizing the entire dataset. The improved photon energy evaluation and ID calibrations will be used with the full Run 3 $H(\rightarrow \gamma\gamma)$ measurement. The PI's team will focus on ensuring that the MVA used for signal-background separation does not sculpt the diphoton mass (an essential ingredient to the background estimation) which will benefit measurements in all the bins of this diifferential cross section measurement. The PI's team will focus on the $t\bar{t}H$ measurement drawing from previous experience with final states that include $t\bar{t}$ within Supersymmetry searches and more recently, final states that include two photons within a charged Higgs search.

3.1 Photon-Energy Evaluation

The evaluation of photon and electron energy involves several steps as detailed in [16]. This proposal aims to test and enhance the simulation-based calibration that utilizes variables derived from energy deposits grouped in the calorimeter. The process aims to adjust for energy that is lost in material before the calorimeter, deposited in adjacent cells, or not captured by the LAr calorimeter. The current algorithm uses a BDT to regress the corrected photon energy using shower-related variables that summarize shower properties (see Figure 3). The BDT is trained with a single particle simulation with a flat transverse momentum (p_{T}) spectrum up to 150 GeV. The performance of the BDT is then evaluated by comparing the energy predicted by the BDT with the true particle energy. Internal ATLAS studies have shown that adding lateral shower shape variables to the BDT could improve the energy resolution after calibration by 10-15%. However, these variables are poorly modelled by the simulation when compared to recorded data, preventing them from being used in the calibration. The application of domain-adaptation techniques—such as integrating real data during training, reweighting training datasets to better reflect actual data distributions, employing adversarial methods, and using distance correlation strategies—could potentially reclaim some benefits of including shower shape variables, despite the discrepancies between simulation and actual data.

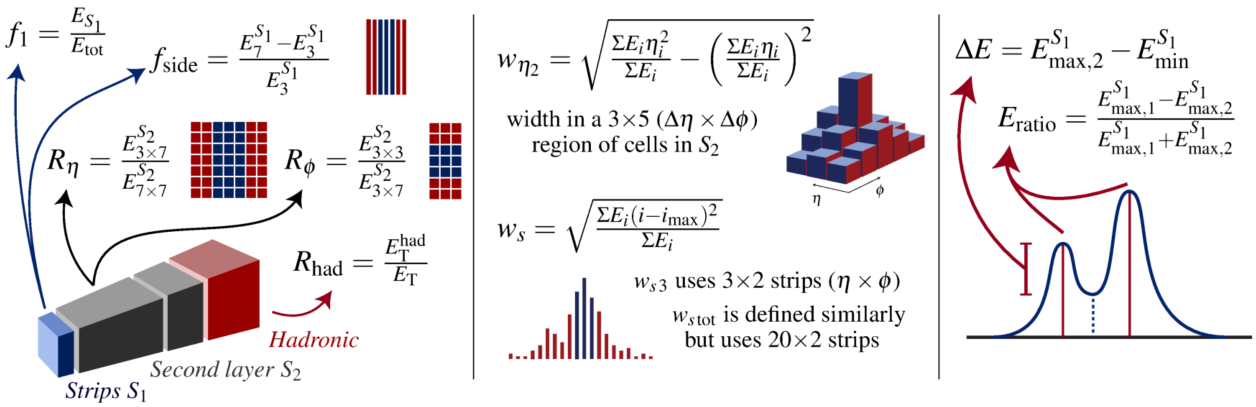

Figure 3

Two domain-adaptation methods, which have already been applied within other contexts within HEP will be studied in parallel, first on the photon-energy calibration using the existing BDT energy regression, and then using a more advanced approach that utilizes GNNs and low-level calorimeter information. The PI's team will apply the adversarial and distance correlation approaches studied by the PI on a simplified toy example using the existing BDT. One postdoc will focus on the adversarial technique while the other will focus on the distance correlation approach, each at 0.5 FTE. The existing BDT has been extensively studied for various improvements and will serve as a benchmark to ensure that the domain-adaptation techniques do not effect the original ML algorithm in a negative way.

The PI's team will first, together with ATLAS e/γ experts, produce a simulation dataset that has been modified to resemble the data. This "fudged" simulation will allow for more control of properties such as photon purity, energy scale, and sample size for initial studies; the team will also study the use of recorded data during training after the domain-adaptation techniques reach a mature state. The team will use the previously developed toy model as a basis to develop domain-adaptation techniques for the energy regression BDT. As a former convener of the ATLAS Full Simulation group, which aims to improve both the physics and computational performance of the ATLAS detector simulation based on GEANT4 [17], the PI will also work with the Full Simulation experts to revisit understanding the source of the mismodelling.

The figure of merit for determining the best domain adaptation approach, of the two more established methods, will be the energy resolution. However, the framework developed to apply these methods will be carried over for use on more complex ML calibrations. The approach that results in the best physics performance, i.e., with the best energy resolution which is quantified by the lowest 50% interquartile range, and with the best data-simulation agreement will be chosen for further validation and integration into the ATLAS e/γ photon calibration procedure. This improved energy calibration could then be used across $H(\rightarrow \gamma\gamma)$ measurements during the LHC Phase-II shutdown using the full Run 3 dataset.

A GNN-based photon-energy calibration is being developed by the University of Edinburgh, University of California at Berkeley, Lawrence Berkeley National Laboratory, and Michigan State University, and has shown promise in improving the photon energy resolution significantly (up to 20%). However, the calibration resulting from the GNN-based approach differs for data and simulation. The PI's team with collaborate with the GNN calibration team to introduce domain-adaptation approaches to the powerful GNN calibration. This will be done after the distance correlation and adversarial techniques have been applied and tested on the simpler BDT-based calibration. The

The PI's team will use the toy data set to develop a domain adaptation approach which will extract features that are robust against simulation mismodelling using the toy fudged-simulation dataset. This will be done using an autoencoder, which reduces the dimensionality of input features, that has an additional domain adaptation figure of merit during training. The domain-adaptation methods studied for the photon energy regression, i.e., distance correlation and adversarial networks, will be used to ensure that the encoded features are insensitive to data-simulation differences. The learned features could then be used for a variety of tasks, e.g., energy determination or object ID.

The deliverables for the photon-energy calibration domain adaptation work as follows:

- An improved BDT-based photon-energy calibration with up to a 15% better resolution and similar performance in recorded and simulated data
- Experience employing domain-adaptation techniques, i.e., distance correlation and adversarial networks, for calibration tasks within ATLAS
- Further improved (up to 20% over the BDT) photon-energy calibration that is robust against data-simulation differences
- A new approach, feature representation transfer, that automatically produces features that are robust against data-simulation differences or other domain changes

3.2 Photon Identification

The ATLAS photon ID utilizes similar quantities to what are used for the photon-energy calibration, of which some are highlighted in Figure 3. Unlike the photon-energy calibration, the ATLAS photon ID does not make use of a multivariate approach. This is due to the calibration procedure for the photon ID, which requires the ID algorithm to be independent of track isolation. This calibration procedure derives correction factors that adjust efficiencies derived from simulation to match efficiencies measured in data. The requirement for the ID algorithm to be independent from isolation comes from the procedure of measuring the efficiency in recorded data. The domain-adaptation techniques, that this proposal aims to prepare for calibration tasks, may be the key to moving the photon ID to a multivariate technique such as a BDT or approaches that use low-level variables such as GNNs. These multivariate approaches have been shown to improve the ID efficiency up to 10% at a similar background rejection to the existing rectangular-requirement-based method. Once the move to multivariate methods is enabled by domain-adaptation techniques, these same approaches can then be extended to minimize the very data-simulation difference that the calibration procedure aims to correct.

The PI's team will work closely with photon reconstruction experts at Northern Illinois University (NIU) who have previously studied the use of BDTs and CNNs for photon identification, members of the ALCF data science group who have studied PointCloud [18] NNs for object ID to develop a multivariate photon ID method, and other ATLAS collaborators who have used ML approaches for photon ID. The domain-adaptation techniques will first be applied to a BDT or an NN that takes high-level quantities as inputs, similar to those currently used in the rectangular-requirements-based approach. As with the energy resolution calibration, both distance correlation and adversarial techniques will be studied in parallel by the two postdocs of the team. Once the output of the multivariate technique has been shown to be independent of the isolation, the team will work to fully calibrate the new ID algorithm together with collaborators at NIU (e.g., working with students who visit Argonne via the ATC or the Office of Science Graduate Student Research Program) using the standard techniques described in [19, 20].

The experience gained from applying the domain-adaptation techniques on a high-level inputs will then be used to apply similar methods to a GNN that uses low-level inputs i.e., energy deposits in the calorimeter cells.

Finally, the photon ID algorithm will be trained with fudged simulation to make it more robust against mismodelling, as was done with the photon energy estimation.

The following deliverables are expected from the proposed photon ID work:

- An improved BDT- or NN-based photon ID algorithm that is independent from isolation and thus can be calibrated using existing techniques

- A GNN-based photon ID algorithm that utilized more information than BDT/NN-based method and that inherently has the same behavior for both simulated and recorded data

3.3 Differential Cross-Section Measurements in $H(\to \gamma\gamma)$

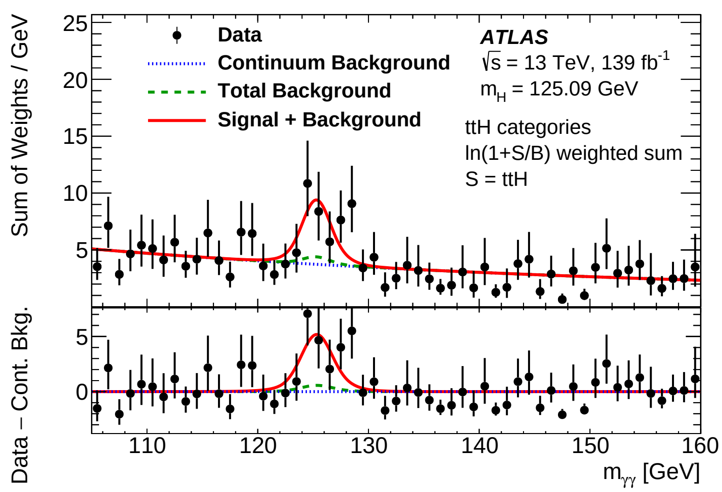

Figure 4: Distribution of the diphoton invariant mass in the $t\bar{t}H$ channel for the latest STXS ATLAS result [12]. The data (dots) are shown together with the sum of the fitted signal plus background (solid line). The blue dotted line represents the sum of the fitted continuum background, while the dashed line combines the contributions of continuum background and other Higgs boson events. The shape of the mass distribution must not be altered by the MVA which is used to discriminate background from signal. The domain-adaptation approaches proposed in this text can improve the discrimation power of MVAs while ensuring the MVA does not sculpt the $m_{\gamma\gamma}$ spectrum.

The most recent ATLAS results that performed many differential cross-section measurements in the $H(\to \gamma\gamma)$ channel made use of BDTs to both discriminate between the differential cross-section bins and to separate background from signal processes. One of the main backgrounds in many channels is the continuum diphoton background which is estimated using a functional fit to the $m_{\gamma\gamma}$ spectrum (see Figure 4 for an example). Due to this background estimation methodology, only variables that were found to have less than a 5% linear correlation with the $m_{\gamma\gamma}$ spectrum were used in the BDT training. The domain-adaptation techniques used to improve the photon energy calibration can also be used to ensure that an ML classifier is independent of the $m_{\gamma\gamma}$ spectrum. This application of domain-adaptation approaches is similar to techniques that have been proposed to decorrelate algorithms to identify substructure within a jet from the mass of the jet, allowing the technique to be used for any mass of a hypothesised BSM particle.

The PI's team will contribute to the "intermediate" $H(\rightarrow \gamma\gamma)$ differential cross-section measurement which was initiated in early 2024 and will use a part of the Run 3 data. The team will focus on re-optimizing the BDTs trained to discriminate between the differential cross-section bins and between signal and the dominant background. This effort will familiarize the team with the existing ML strategy that does not incorporate domain adaptation to ensure that the BDTs are not correlated with the $m_{\gamma\gamma}$ spectrum.

Domain adaptation will be applied to the final Run 3 analysis that will use the full Run 3 dataset and that is expected to finalize near the end of the Phase-II shutdown. Using the framework that was developed for the photon ID MVA and energy calibration, the two postdocs and the PI will work in parallel to determine which domain adaptation approach, distance correlation vs adversarial vs feature representation transfer, is optimal for ensuring that the diphoton invariant mass spectrum remains uncorrelated with the signal-to-background MVA (either a BDT or NN). The method that both ensures that the background shape remains unchanged and results in the highest signal significance, which will now take advantage of more variables, will be used in the $H(\rightarrow \gamma\gamma)$ measurement.

In addition to using domain adaptation to ensure the stability of the $m_{\gamma\gamma}$ spectrum, the improved photon ID will be ready to be used in the final Run 3 $H(\rightarrow \gamma\gamma)$ measurement. The PI's team will focus on optimizing sensitivity to the $t\bar{t}H(\rightarrow \gamma\gamma)$ which is statistically limited and thus will especially benefit from the improved photon efficiency (reducing the statistical uncertainty by as much as 17%).

The deliverables associated with the differential cross-section measurement are:

- A publication describing the intermediate Run 3 $H(\rightarrow \gamma\gamma)$ measurement
- A publication describing the final Run 3 $H(\rightarrow \gamma\gamma)$ measurement with improved photon ID and background rejection enabled by domain adaptation

3.4 Integrating Domain Adaptation into an ML Framework

domain-adaptation techniques will be explored and applied across various ML models (BDTs, NNs, and GNNs) and their applications, including object classification, property regression, and event classification. These models will be developed within the SALT framework, which is frequently used within ATLAS, to facilitate the incorporation of domain adaptation into a widely-used machine learning framework. SALT has already been employed to construct ML models for identifying boosted decays to two b-quarks, jets initiated by heavy flavor quarks, τ-leptons, and more. Integrating features into SALT that allow ML models to be decorrelated from specific variables—such as data-simulation disparities, pile-up, and others - would reduce uncertainties stemming from discrepancies between recorded and simulated data and equip numerous algorithms for the demanding conditions of the HL-LHC.

Given that the photon ID, photon energy regression, and the $H(\rightarrow \gamma\gamma)$ signal-to-background ML models incorporating domain adaptation will be partially built using SALT (which predominantly supports GNN-based models), the PI's team will have already embedded domain-adaptation techniques within specific SALT implementation. This groundwork will ease the inclusion of a comprehensive domain adaptation framework into SALT. This enhancement will encompass the capability to use manipulated simulation data for preliminary tests of domain adaptation strategies and the implementation of various domain-adaptation methods. Finally, validation plots will be produced to demonstrate any residual correlations using different techniques, such as distance correlation when adversarial models are the used and the performance of a discriminator trained to differentiate between both domains when distance correlation is employed.

The effort to include domain adaptation, which can require significant computing resources to include during the training of an ML model, into SALT will be aided by synergistic activities within the Argonne ATLAS group. These efforts are geared towards equipping SALT to utilize HPCs, which are expected to drastically cut down training times (preliminary studies indicate a reduction from several days to just a few hours). Leveraging the extensive computing capabilities of HPCs will ensure that the implementation of domain adaptation is not constrained by computational limitations.

3.5 Timetable of Activities

The timetable of activities is shown in Figure 5.

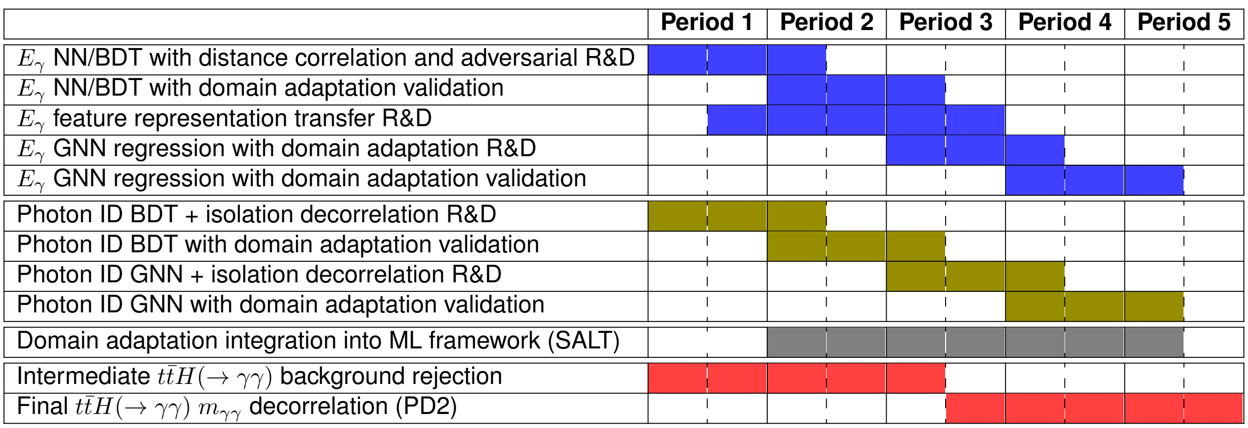

Figure 5: Expected time spent on the work proposed in this text by the PI's team which will consists of two postdocs (at 1.0 FTE each) and about half of the PI's effort. Each period is 12 months long with Period 1 starting on August 1st 2024 and Period 5 ending on July 31st 2029 which is two months after the HL-LHC is slated to start operations (May 2029). The tasks include the application of domain adaptation techniques to photon energy (E_γ) evaluation. Two established techniques will be studied, distance correlation and adversarial networks (AN), as well as a new technique, feature representation transfer. A subset of domain-adaptation techniques which decorrelate the ML model's output from certain features will be used for photon ID and signal-to-background separation in the $t\bar{t}H$ differential cross-section measurement.

Below is a summary of the milestones:

1. Reduction in photon energy resolution uncertainty due to mitigation of data-simulation differences using domain adaptation.

2. Multivariate photon ID that is decorrelated from isolation

3. Updated intermediate $H(\rightarrow\gamma\gamma)$ differential cross-section measurement using part of the Run 3 dataset

4. Improved background rejection in $H(\to \gamma\gamma)$ enabled by domain adaptation being applied to decorrelate signal-to-background discriminating ML from the $m_{\gamma\gamma}$ spectrum

5. Integration of domain adaptation into ML framework (SALT) for broad use within ATLAS

3.6 Personnel and Resources

The PI will lead a team to execute the proposed project, involving an average of 2.0 postdoctoral FTEs and contributing 50% of the PI's effort. These postdocs, fully funded by this award, are slated to begin at the start of the first and second budget periods and will continue through the fifth budget period. Their responsibilities will be divided between implementing domain adaptation strategies and conducting two differential cross-section measurements in the $H(\to \gamma\gamma)$ channel.

Additionally, the PI will engage with graduate and undergraduate students on this research through established connections with local universities, supported by several programs. The PI has a track record of mentoring students who participate in programs such as the DOE Office of Science Graduate Student Research program and the Science Undergraduate Laboratory Internships. Furthermore, the Argonne ATLAS Center, which welcomes graduate students from ATLAS collaboration member universities, offers further collaborative opportunities. The PI will work with students from institutions that have existing ties with the Argonne ATLAS group, including NIU, which boasts expertise in photon and Higgs sectors and runs a computing traineeship aligning with the computational demands (i.e., the infrastructure needed for the framework) of this proposal.

For computing needs, the PI will access resources at the Argonne through the Laboratory Computing Resources Center (LCRC), which boasts a significant setup (825 nodes, each with 128 cores) frequently utilized by the Argonne ATLAS group and ATLAS Analysis Center (ATC) university collaborators for preparing ATLAS measurements for publication. Additionally, the PI will use a specialized system at LCRC, comprising six nodes with eight GPUs each, for ML model development. The Argonne ATLAS group holds a renewable computing allocation at LCRC, which can be expanded as required for the proposed tasks. The Argonne Leadership Computing Facility (ALCF) provides further resources and expertise, including both NVIDIA and Intel GPU capabilities, to support the development and training of complex ML models. The PI's team will leverage existing relationships with ALCF experts to develop and refine domain-adaptation approaches, crucial to computing scientists, using ALCF resources. Requests for resource allocations will be made as needed to support these activities.

4 Summary

Machine learning has become an indispensable tool in HEP, enhancing capabilities in tasks such as particle identification and distinguishing between signal and background processes. However, as we approach the era of the HL-LHC with its unprecedented data scale, ML models must become more robust to effectively mitigate the impact of systematic uncertainties on physics measurements. Domain-adaptation approaches have shown potential in reducing these uncertainties and improving the robustness of models against experimental conditions and data-simulation discrepancies. This proposal outlines the integration of domain-adaptation techniques into an ML framework to make these methods broadly accessible, thus potentially impacting a wide range of HEP measurements. This effort is crucial for future discoveries in collider physics, especially in the precision era of the HL-LHC, where precise measurements are key to probing the effects of BSM physics.

Appendix 1: Biographical sketch

NSF BIOGRAPHICAL SKETCH

NAME: Hopkins, Walter

POSITION TITLE & INSTITUTION: Assistant Physicist, Argonne National Laboratory

(a) PROFESSIONAL PREPARATION -(see PAPPG Chapter II.C.2.f.(a))

INSTITUTION	LOCATION	MAJOR / AREA OF STUDY	DEGREE (if applicable)	YEAR YYYY
Rochester Institute of Technology	Rochester, NY	Physics and Applied Mathematics	BS	2007
Cornell University	Ithaca, NY	High Energy Physics	PHD	2013
University of Oregon	Eugene, OR	Searches for Supersymmetry and the Liquid Argon Calorimeters with the ATLAS experiment	Postdoctoral Fellow	2013 - 2018

(b) APPOINTMENTS -(see PAPPG Chapter II.C.2.f.(b))

2018 - present Assistant Physicist, Argonne National Laboratory, Lemont, IL

(c) PRODUCTS -(see PAPPG Chapter II.C.2.f.(c))

Products Most Closely Related to the Proposed Project

1. ATLAS Collaboration. Search for a scalar partner of the top quark in the all-hadronic $t{\bar{t}}$ plus missing transverse momentum final state at $\sqrt{s}=13$ TeV with the ATLAS detector. Eur. Phys. J. C. 2020; 80(2020):737. Available from: https://arxiv.org/abs/2004.14060 DOI: 10.1140/epjc/s10052-020-8102-8
2. Benjamin D, Chekanov S, Hopkins W, Li Y, Love J. Automated detector simulation and reconstruction parametrization using machine learning. JINST. 2020 May; 15(5):P05025–P05025. Available from: https://arxiv.org/abs/2002.11516 DOI: 10.1088/1748-0221/15/05/p05025
3. ATLAS Collaboration. Searches for third-generation scalar leptoquarks in $\sqrt{s}=13$ TeV pp collisions with the ATLAS detector. JHEP. 2019; 6(2019):144. Available from: http://dx.doi.org/10.1007/JHEP06(2019)144 DOI: 10.1007/jhep06(2019)144
4. ATLAS Collaboration. Summary of searches for dark matter and dark energy using $\sqrt{s}$=13 TeV pp collisions with the ATLAS detector at the LHC. JHEP. 2019; 5(2019):142. Available from: https://arxiv.org/abs/1903.01400 DOI: 10.1007/jhep05(2019)142
5. ATLAS Collaboration. Search for a scalar partner of the top quark in the jets plus missing transverse momentum final state at $\sqrt{s}=13$ TeV with the ATLAS detector. JHEP. 2017; 12(2017):085. Available from: https://arxiv.org/abs/1709.04183 DOI: 10.1007/jhep12(2017)085

Other Significant Products, Whether or Not Related to the Proposed Project

1. ATLAS Collaboration. ATLAS Run 1 searches for direct pair production of third-generation squarks at the Large Hadron Collider. Eur. Phys. J. C. 2015; 75(2015):10. Available from: http://arxiv.org/abs/1506.08616 DOI: 10.1140/epjc/s10052-015-3726-9
2. ATLAS Collaboration. ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design

519

Report. CERN. 2013. Available from: https://cds.cern.ch/record/1602230

3. CDF Collaboration. Search for $B_s^0\to\mu^+\mu^-$ and $B^0\to\mu^+\mu^-$ Decays with CDF II Full Data Set. Phys. Rev. D. 2013; 87(2013):072003. Available from: http://arxiv.org/abs/1301.7048 DOI: 10.1103/physrevd.87.072003
4. CDF Collaboration. Search for $B_s^0\to\mu^+\mu^-$ and $B^0\to\mu^+\mu^-$ Decays with CDF II. Phys. Rev. Lett.. 2011; 107(2011):191801. Available from: https://arxiv.org/abs/1107.2304 DOI: 10.1103/PhysRevLett.107.191801

(d) SYNERGISTIC ACTIVITIES -(see PAPPG Chapter II.C.2.f.(d))

1. April 2020-present: SUSY Strong Production Subgroup convener. Reviewed SUSY Strong analyses for unblinding approval and preparation for publication.
2. August 2020-present: member of Geant4 Optimization Task Force. Studied sources of computational bottlenecks of the ATLAS implementation of Geant4. Also studied possible ML and non-ML based methods to reduce the computational cost of Geant4.
3. 2018-present: PI for the Argonne ATLAS Aurora Early Science Project. Preparing both an ATLAS ML workload, flavor tagging with uncertainty quantification, and standard workload, event generation with MadGraph, for use on the upcoming Aurora supercomputer. Madgraph is being prepared with CERN collaborators to make use of GPU resources which will make up a significant part of future computing resources.
4. April 2020-present: Snowmass topical group co-convener for the Experimental Algorithm Parallelization group. Prepare a summary document to be used as input for the final Snowmass Computational Frontier report. The group is focused on non-simulation and non-ML algorithms used in various experiments and that will need to be adapted for use on future computing resources.
5. 2014-2020: ATLAS SUSY Stop Search Analysis Team contact. Co-lead late Run 1 and all Run 2 searches for the supersymmetric top partner in the all-hadronic final state.

Appendix 2: Bibliography and References

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012) 1, arXiv: 1207.7214 [hep-ex].

[2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716** (2012) 30, arXiv: 1207.7235 [hep-ex].

[3] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, 2017, arXiv: 1611.01046 [stat.ML].

[4] G. Kasieczka and D. Shih, *Robust Jet Classifiers through Distance Correlation*, Phys. Rev. Lett. **125** (12 2020) 122001, URL: https://link.aps.org/doi/10.1103/PhysRevLett.125.122001.

[5] ATLAS Collaboration, *Search for neutral long-lived particles in pp collisions at $\sqrt{s}$ = 13 TeV that decay into displaced hadronic jets in the ATLAS calorimeter*, Journal of High Energy Physics **2022** (2022), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP06(2022)005.

[6] *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, tech. rep., CERN, 2018, URL: https://cds.cern.ch/record/2630973.

[7] ATLAS Collaboration, *Search for New Phenomena in Two-Body Invariant Mass Distributions Using Unsupervised Machine Learning for Anomaly Detection at* $\sqrt{s} = 13$ TeV *with the ATLAS Detector*, Phys. Rev. Lett. **132** (8 2024) 081801, URL: https://link.aps.org/doi/10.1103/PhysRevLett.132.081801.

[8] W. Buchmüller and D. Wyler, *Effective lagrangian analysis of new interactions and flavour conservation*, Nucl. Phys. B **268** (1986) 621.

[9] B. Grzadkowski, M. Iskrzyński, M. Misiak, and J. Rosiek, *Dimension-six terms in the Standard Model Lagrangian*, JHEP **10** (2010) 085, arXiv: 1008.4884 [hep-ph].

[10] I. Brivio, *SMEFTsim 3.0 — a practical guide*, JHEP **04** (2021) 073, arXiv: 2012.11343 [hep-ph].

[11] LHC Higgs Cross Section Working Group, D. de Florian, et al., *Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*, CERN-2017-002-M (2017), arXiv: 1610.07922 [hep-ph].

[12] ATLAS Collaboration, *Measurement of the properties of Higgs boson production at $\sqrt{s}$ = 13 TeV in the $H \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment*, Journal of High Energy Physics **2023** (2023), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP07(2023)088.

[13] ATLAS Collaboration, *Deep Learning for Pion Identification and Energy Calibration with the ATLAS Detector*, tech. rep., All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-018: CERN, 2020, URL: https://cds.cern.ch/record/2724632.

[14] ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016*, The European Physical Journal C **79** (2019), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-019-6650-6.

[15] ATLAS Collaboration, *SALT*, https://gitlab.cern.ch/atlas-flavor-tagging-tools/algorithms/salt, 2024.

[16] A. Collaboration, *Electron and photon energy calibration with the ATLAS detector using LHC Run 2 data*, 2023, arXiv: 2309.05471 [hep-ex].

[17] S. Agostinelli et al., GEANT4—*a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250.

[18] *Physics Object Localization with Point Cloud Segmentation Networks*, tech. rep., CERN, 2021, URL: https://cds.cern.ch/record/2753414.

[19] ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run-1 data*, Eur. Phys. J. C **76** (2016) 666, arXiv: 1606.01813 [hep-ex].

[20] ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016*, Eur. Phys. J. C **79** (2019) 205, arXiv: 1810.05087 [hep-ex].

Appendix 3: Facilities and other resources

Argonne offers various resources, including office space for postdocs and potential students as well as several computing resources. The computing resources expected to be utilized include CPU and GPU resources available through the Laboratory Computing Resource Center at Argonne. These resources are expected to be used to produce input data for ML algorithms as well as training the ML algorithms. The PI also plans to use computing resources at the ALCF resources, such as Polaris (which has NVIDIA GPUs) and Aurora (which contains Intel GPUs), via a discretionary allocation. These resources will be used for large-scale training and optimization of the proposed approach.

An additional important resource is the ATLAS experiment at the LHC at CERN. The PI's team will make use of ATLAS data and simulation for the proposed studies and will occasionally travel to CERN to work with international collaborators.

Appendix 4: Equipment

The requested funding will be used to purchase two laptops for the postdocs that the PI will supervise. The combined cost of the three laptops is expected to be $6,418.

Appendix 5: Data management plan

The majority of the data and code produced will come from the ATLAS experiment. ATLAS has its own data management plan which the PI's team will adhere to; the details of that plan can be found at https://po.usatlas.bnl.gov/programoffice/datamanagementpolicy.php.

Scientific results that used ATLAS data will be published in a scientific journal or an ATLAS publication note, which will be publicly available on https://arxiv.org/ and https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SimulationPublicResults, respectively. Figures and tables will be made available in the same way as other ATLAS results, via HEPData (https://www.hepdata.net/) and ATLAS public pages.

Additional data that are produced outside of ATLAS for initial ML algorithm development will be stored on the Argonne High Energy Physics divisional nodes in the HDF5 format and will consist of energy deposits in a simplified detector. The generated data are expected to occupy a small fraction of the available data storage on these nodes. Code and configurations used for the development of the algorithm and production of data will be kept at the Argonne Computing, Environment and Life Sciences (CELS) gitlab repository: https://xgitlab.cels.anl.gov/. Results from these studies, if published, will be publicly available on https://arxiv.org/.

Appendix 6: Promoting Inclusive and Equitable Research (PIER) Plan

Achieving and maintaining long-term success requires the creation of a safe and welcoming environment that fosters diversity, equity, and inclusion (DEI). An environment that promotes DEI will enhance the proposed research by nurturing creativity in problem-solving and by encouraging constructive criticism from individuals with diverse backgrounds and perspectives. This appendix summarizes the key aspects of how the PI will approach DEI for the proposed work within the ATLAS group and the HEP division at Argonne, as well as how the PI plans to improve DEI.

The PI will ensure an inclusive and equitable work environment. All personnel, including two postdocs hired for this proposal and any students and postdocs from collaborating institutions, will be encouraged to express their ideas in an environment that promotes curiosity and openness. This will be facilitated through both semi-formal and formal settings. Formal idea exchanges will include group meetings with collaborators, where members of the working group are encouraged to ask questions or suggest improvements to the proposed work. Topics will be discussed among all involved personnel to foster leadership on a subtopic by postdocs and students. Additionally, postdocs will be assigned a mentor outside the ATLAS group, in accordance with HEP division policy, who will discuss career goals from a non-ATLAS perspective.

The PI will also meet individually with group members at least once a week to discuss ideas for the proposed work. These meetings are intended to build the confidence of junior postdocs and students, enabling them to express ideas and provide constructive feedback both within and outside the group.

6.1 How This Work Will Enhance DEI

The PI is committed to fostering opportunities and removing obstacles for those on the path to becoming career scientists. Outreach programs are instrumental in enlarging the candidate pool, particularly for underrepresented minorities (URMs). Drawing from previous outreach experience, the PI will illuminate the diverse career paths available to students who pursue advanced degrees in HEP, both within academia and beyond. It is often overlooked that qualifications in HEP can lead to fulfilling careers outside the field, particularly in industries that value expertise in ML and computing. During outreach activities, the PI will emphasize the broad spectrum of career opportunities that ML and computing skills can unlock.

To further support these efforts, the PI will recruit undergraduate students for proposed projects through various summer initiatives. The USATLAS Summer Undergraduate Program for Exceptional Researchers (SUPER), sponsored by USATLAS, provides financial assistance to exceptional undergraduates in the US for summer research in areas like ATLAS physics analysis and computing. Expanded in 2022, SUPER now includes students from non-USATLAS minority-serving institutions (MSIs). Additionally, the PI will continue recruiting through the Science Undergraduate Laboratory Internships (SULI) and participating in the Look@Argonne event, which promotes internship opportunities at Argonne to Louis Stokes Alliances for Minority Participation affiliated institutions, highlighting the ATLAS group's opportunities through panel discussions.

The PI has also been instrumental in leading traineeship programs that mentor early-stage researchers. Alongside USATLAS universities, the PI co-organized the CAMPFIRE events in 2019, 2022, and 2023, which are designed to build a supportive and inclusive community of USATLAS graduate students. These events, featuring educational, technical, and social activities, strive to foster connections among students from historically underrepresented groups, addressing potential feelings of isolation they might encounter at their home institutions or at CERN, and preparing them for future stages of their careers.

In addition to engaging with graduate students through traineeship programs, the PI has hosted graduate students from collaborating universities via the USATLAS ATLAS Center (ATC) and the Office of Science Graduate Student Research (SCGSR) Program. The PI plans to deepen these collaborations, actively involving more graduate students at Argonne in the proposed projects. This integration not only provides graduate students with experience working at a national laboratory on ML projects but also contributes valuable insights and efforts to the research.

In hiring practices, the PI adheres to the best practices of the Argonne HEP division, which include advertising positions directly to MSIs and ensuring that selection committees, often led by a chair who has completed bias training from the Argonne Human Resources department, are aware of implicit biases. The PI also implements a co-developed strategy within the ATLAS group that includes blind reviews of candidates to prevent groupthink and bias, enhancing the fairness and inclusivity of the hiring process.

6.2 Conclusion

The PI is dedicated to cultivating a research environment that not only supports but also thrives on DEI. Central to these efforts is the development of a work culture that welcomes input from all members, nurturing leadership skills among postdocs and students, and facilitating their involvement in high-impact research topics. Moreover, the PI's proactive approach to expanding the diversity of the research pool—through targeted recruitment from minority-serving institutions and outreach initiatives—highlights a strategic commitment to accessibility and representation in science. The PI's leadership in events like the CAMPFIRE traineeship programs further exemplifies a commitment to mentorship and the development of an inclusive community of physicists. Finally, adherence to equitable hiring practices, including bias training and blind review processes, underscores a systematic approach to fostering an inclusive research environment.

Appendix 7: Other attachments

There are no other attachments.