Contents

1 Introduction

Machine learning (ML) has increasingly become a critical tool in High-Energy Physics (HEP), offering significant advancements in various tasks such as simulating calorimeter showers, identifying particles, and distinguishing between signal and background processes. ML techniques allow physicists to delve into 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. One such challenge is the alignment of input variable distributions between Monte Carlo (MC) simulations and experimental data. Differences between MC simulations and recorded data can introduce additional uncertainties into ML predictions, affecting the overall systematic uncertainties in physics measurements. Additionally, ML algorithms may be sensitive to experimental systematic uncertainties which are typically assessed by varying underlying experimental parameters (e.g., the energy resolution of a subdetector). Furthermore, ML models can sometimes create unintended correlations between their outputs (such as estimated energy or particle identification) and other variables critical for calibrations or background estimations.

Recently proposed ML approaches such as adversarial [1] and distance correlation (DisCo) [2] techniques have shown promise in various applications such as reducing uncertainties in a long-lived particle search [3] and decorrelating jet substructure variables from jet mass [4]. 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 set of data to be robust enough to be applied to data that is expected to be different from the training data. Domain adaptation techniques have the potential to be more broadly applied when developing ML models to estimate physics object properties (e.g., photon energy, jet transerve momentum, etc), to identify (ID) physics objects (e.g., photons, jets containing b-hadrons, etc), to reduce the sensitivity data-MC discrepancies, detector and accelerator conditions (e.g., the number of simultaneous proton-proton interactions, pile-up), and to changes in the underlying parameters of simulations. All these applications of domain adaptation approaches could reduce the total uncertainties of physics results within HEP.

This proposal presents **the development of a framework to deploy various domain adaptation techniques to minimize uncertainties when using machine-learning-based physics object ID and property estimation by ensuring that ML models are more resilient against experimental systematic uncertainties, data-MC differences, and changes in detector conditions. Furthermore, the PI's team intends to refine domain adaptation methods that generate durable features, less affected by specific assumptions (such as detector resolution and pile-up), that have not been previously applied in HEP, for implementation in HEP contexts.** The framework has broad applications but will first be used to maximize object ID efficiencies and property estimation precision. The proposed techniques could also improve object ID and property evaluations (potentially improving both the efficiency and resolution) at the trigger level, especially given the recent work on fast inference on field-programmable gate arrays (FPGAs).

Improving both ID efficiencies (which is equivalent to recording more data) and reducing systematic uncertainties will be essential to enabling future discoveries in collider physics. Results from the Large Hadron Collider (LHC) experiments have verified the predictions of the highly successful Standard Model (SM), culminating with the discovery of the Higgs boson [5, 6]. However, the SM lacks an explanation for several observed phenomena (e.g., dark matter, the matter-antimatter asymmetry, etc) motivating the search for Beyond the Standard Model (BSM) physics. Future LHC upgrades will no longer include substantial increases in energy and move HEP into the precision era, with a tenfold increase (High Luminosity-LHC, HL-LHC) of the Run 2 data set.

Precision measurements of SM processes, especially interaction involving Higgs bosons, can probe for BSM effects resulting from particles which may have large masses that prevent them from being directly produced at the LHC. An essential aspect to improving the precision of measurements, which will maximize sensitivity to BSM physics, is the calibration of physics object ID and properties. This calibration involves evaluating ID efficiencies and properties in MC simulations and correcting these quantities to the true property before reconstruction or to what is observed in data.

The Higgs boson self coupling and the coupling to the top quark are especially sensitive to BSM effects [7]. The associative production of a top quark pair with a Higgs boson ($t\bar{t}H$) is sensitive to both of these couplings [8] and thus the $t\bar{t}H$ total and particularly differential cross section measurements, e.g., as function of Higgs p_{T}, can serve as probes for BSM physics. The latest ATLAS $H(\rightarrow \gamma\gamma)$, which includes $t\bar{t}H(\rightarrow \gamma\gamma)$, differential cross section results [9] remain limited by the statistical uncertainty in each kinematic observable bin. However, the photon ID and energy resolution uncertainties are the dominant detector-based systematic uncertainties and are currently on par with the other dominant source of systematic uncertainty, theory uncertainties. Reducing these detector systematic uncertainties will become significantly more important for all $H(\rightarrow \gamma\gamma)$ measurements, as ATLAS drastically increases the size of its data set near the end of HL-LHC data taking. Thus, the unprecedented data volume of the HL-LHC offers an opportunity to improve the precision cross section measurements as a function of kinematic variables in channels with high $t\bar{t}H$ purity, i.e., where the Higgs decays to two photons ($t\bar{t}H(\rightarrow \gamma\gamma)$).

The proposed framework and techniques will first be developed and applied to photon ID and energy resolution because these are key ingredients to maximizing the sensitivity of Higgs precision measurements. Studies have been performed within ATLAS that showed a potential ∼10-20% improvement in both ID efficiency (which is equivalent to collecting 10-20% more data) and energy resolution when including all available information for ID and energy estimation. However, as soon as the calibration of the ID efficiency and energy resolution was taken into account, the gain in efficiency and enhancement in resolution were lost. For the ID, ML approaches were found to be correlated with quantities that needed to be independent of the ID for the calibration while for the energy resolution it was found that ML methods were sensitive MC mismodelling of shower shapes. Thus, **the proposed domain adaptation techniques could improve all ATLAS measurements involving $H \rightarrow \gamma\gamma$ which are essential to the HL-LHC BSM search program.**

The PI's experience in ML, e.g., as one of the ATLAS ML Forum conveners and a co-developer of unsupervised learning techniques for use in ATLAS, as well as his experience with the LAr calorimeter and calorimeter simulations will aid the success of this proposal. The PI will also draw from his work within the ATLAS SUSY group, as a leader of flagship searches [10–13] involving $t\bar{t}$ final states.

1.1 Background and Motivation

Machine learning has already been used for the identification of various physics objects (i.e., electrons, jets containing b-hadrons, etc) and to calibrate properties of objects (e.g., for pions [14]). Many of these ML models, however, use simulations for training and do not automatically take data-MC differences into account during training. Additionally, algorithms may be required to be independent from variables that are either used in part of the calibration procedure, e.g., to determine the rate of fake objects [15], or that might vary during the operations of the collider or detector, e.g., pile-up which will be particularly challenging environment at the HL-LHC.

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_{T}∼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 $H \to \gamma\gamma$ 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 $H \to \gamma\gamma$ 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 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 calibration 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 ∼20% improvement in photon energy resolution but was sensitive to data-MC mismodelling in variables that summarized the shape of the energy deposits in the calorimeter. Ensuring that ML models are invariant with respect to these mismodelled quantities 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).

There have been recent advancements in decorreclating ML models from variables (e.g., jet substructure from jet mass) using methods such as adversarial discriminants and DisCo. Adversarial approaches have been deployed to generate data, including calorimeter showers [16, 17], and 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. This strategy, i.e., to use a discriminator to penalize an ML model for giving differing results, can be applied to minimize differences between data and MC and to remove correlations to certain variables by including two loss (i.e., cost) functions, one for each ML model. The two opposing optimizations can, however, cause the training to be compute intensive and make it difficult to determine when to stop training. Distance correlation summarizes the dependence of sets of variables and is sensitive to non-linear correlations, which are commonly present in HEP datasets, unlike the more frequently used measure of the Pearson correlation coefficient which can only detect linear correlations. The DisCo 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 (or cost) functions which are minimized to optimize ML models. Thus, DisCo is simpler to implement than adversarial techniques with less hyperparameter (since only a loss term is added rather than an entire NN) and more stable training characteristics. These decorrelation methods allow ML techniques to maximize the use of all information while avoiding regions of kinematic phase space that are poorly modelled in simulation and/or ensure an ML model is independent of a quantity, such as pile-up or isolation. Finally, these approaches also avoid regions where there are differences between recorded data and simulations within non-linear correlations. Both adversarial and DisCo approaches can be the optimal solution for a particular application and should be studied to maximize the performance of a particular ML model. The rise of AI/ML-focused High-Performance Computers (HPCs), suchs at Aurora at the ALCF, will facilitate the development of these decorrelation techniques.

Typically, physics object properties are corrected for differences between MC and data in bins of

variables affected by changing detector effects (e.g., the pseudorapidity, η, and transverse momentum, p_{T}, of an object). The dimensionality of these bins is limited by practical considerations with the bin edges being chosen manually. Another set of ML techniques, known as clustering, not only allows for an automated way to group similar events at higher dimensions, it also enables correlations between these variables to be taken into account. Clustering has already been used to choose groupings in a two dimensional space for an analysis [18] and the PI has studied these techniques to group theory models with similar final states. These ML approaches can help not only improve the calibration of physics objects but can also improve the speed of the calibration cycle by automating part of the process.

The HL-LHC, heralds a new frontier in HEP, characterized by unparalleled precision and voluminous datasets that transition many experiments from being statistically constrained to being dominated by systematic uncertainties. This paradigm shift underscores the pivotal role of systematic uncertainties in shaping the landscape of precision measurements within the HEP domain. Measurements, especially those involving the SM Higgs, become increasingly sensitive to high mass BSM contributions as the precision of measurements improves. As the size of the LHC data set 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 sections measurements can probe BSM effects by being interpreted in the context of the Standard Model Effective Field Theory (SMEFT) [19–21] and in the case of Higgs-related measurements, results can also be interpreted in terms of coupling strengths within the κ-framework [22]. An example of a recent differential cross section measurement involving the decay of the SM Higgs to two photons [9] used the simplified cross section template (STXS) method, which performs measurements in bins in several kinematic dimensions (see Figure **??**) to constrain BSM effects. Current measurements statistically limited but have substantial systematic uncertainties (up to ~40% in one of the differential cross section bins) which will become crucial in the HL-LHC era. For many of these measurements photon resolution and ID are some of the largest experimental uncertainty (see Figure 2).

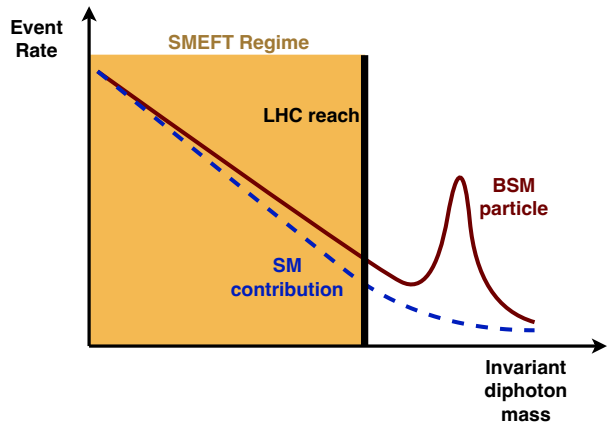
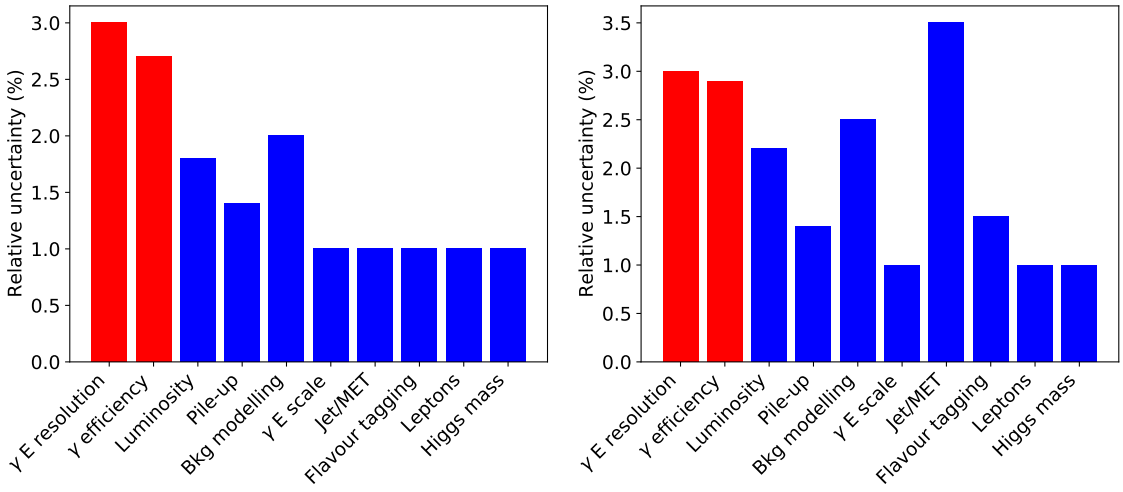


Figure 1: Example of how effects of a heavy BSM particle can leak into high-energy regions of distributions measured at LHC experiments.

2 Project Objectives

The objective of the proposed work is to gain experience with decorrelation techniques for calibration tasks and to build a framework to easily deploy such methods for various calibration tasks. These approaches can help incorporate calibration considerations (e.g., differences between recorded data and simulation) into ML models which will result in a decrease in uncertainties. This will maximize the power of precision measurements by improving physics object ID and energy evaluation. This approach will first be applied to the ATLAS photon ID and energy resolution calibration



(a) ggF (b) $t\bar{t}H$

Figure 2: Expected contributions from the main sources of experimental systematic uncertainty to the total uncertainty in the measurement of the cross-section times $H \to \gamma\gamma$ branching ratio for gluon-gluon fusion Higgs production and $t\bar{t}H$ processes [9]. The uncertainty from each source is shown as a fraction of the total expected cross-section (σ). The magnitude of these uncertainties grows significantly for the differential cross section measurements.

which will in turn be used in the upcoming $H(\to \gamma\gamma)$, focusing on $t\bar{t}H(\to \gamma\gamma)$, differential cross section measurements. Decorrelation techniques will also be used in the optimization of $H(\to \gamma\gamma)$ measurements by ensuring that final multivariate analysis (MVA) does not change the background $\gamma\gamma$ spectrum. The decorrelation framework will then be integrated into existing ATLAS ML frameworks such as "salt" [**salt**].

- Develop method to incorporate calibration considerations into ML physics object ID
  - Using recent decorrelation schemes
- Apply decorrelation techniques to calibrate photon ID efficiency and photon energy
- Study the use of clustering for automated calibration binning
- Perform updated $t\bar{t}H(\to \gamma\gamma)$ differential cross section measurement with improved photon ID and resolution
- Integrate framework into existing ATLAS ML frameworks

3 Proposed Research and Method

The PI's team, consisting of two postdocs at 2.0 FTE, will customize and deploy decorrelation techniques to improve photon energy and photon ID calibration. The two general approaches, decorrelation and adversarial, will be studied in parallel by the two postdocs, with one postdoc gaining expertise in adversarial techniques and the other in decorrelation methods which will have the same application and thus input features, data, and caveats.

The decorrelation techniques will first be applied to ML calibration and ID approaches using BDTs and will then be adapted to be used with more nascent and advanced networks such as GNNs. The adversarial approaches as well as GNNs require significant computing resources to train and tune. The PI will make use of the computing resources at Argonne via the Laboratory Computing Resources Center (LCRC) and the resources and expertise at Argonne Leadership Computing Facility (ALCF). The Argonne ATLAS group already has a renewing computing allocation at LCRC which includes a cluster (six nodes with eight graphical processing units, GPUs, each) that will be used to build models that include decorrelation. Additional resources and expertise is available at the ALCF which includes both NVIDIA and Itel GPU resources. The PI's team will make use of previous connections with ALCF experts to help develop and train the decorrelation approaches on ALCF resources. Allocations for these resources will be requested as they are needed.

The photon energy and ID calibrations will then be used for full Run 3 $H \to \gamma\gamma$ STXS 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 photon final states within a charged Higgs search.

Once the proposed approaches have been developed and validated for photons, the PI's team will work members of the Argonne ATLAS group (who will not funded by this proposal, if awarded), who are involved in flavor tagging, to study the decorrelation techniques for other ID tasks.

3.1 Photon energy calibration

Photon and electron energy is calibrated using several steps which are detailed in [23]. The MC-based calibration, which this proposal seeks to use as a test case and to improve, uses variables derived from groupings of energy deposits in the calorimeter and aims to correct for the energy lost in the material upstream of the calorimeter, for the energy deposited in the cells neighbouring the cluster, and the energy lost beyond 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 MC 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. Including data or simulations that has been weighted to mimic data distributions together with decorrelation techniques could recover some of the improvements resulting from including shower shape variables without being affected by data/MC differences.

The two decorrelation methods 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 methods studied by the PI on a simplified toy example and described in [**<empty citation>**] using the existing BDT. One postdoc will focus on the adversarial technique while the other will focus on the decorrelation approach, each at 0.5 FTE. The current BDT does not suffer from large discrepancies between data and simulations but studies within ATLAS have shown that adding additional variables, specifically quantities that describe the shower shape, which are poorly modelled in simulations could improve the performance of the BDT. The BDT has been extensively studied for various improvements and will serve as a benchmark to ensure that the mismodeling reducing (i.e., decorrelation) techniques do not effect the original ML algorithm in a negative way.

The PI's team will first, together with ATLAS e/γ, produce a simulation data set that has been

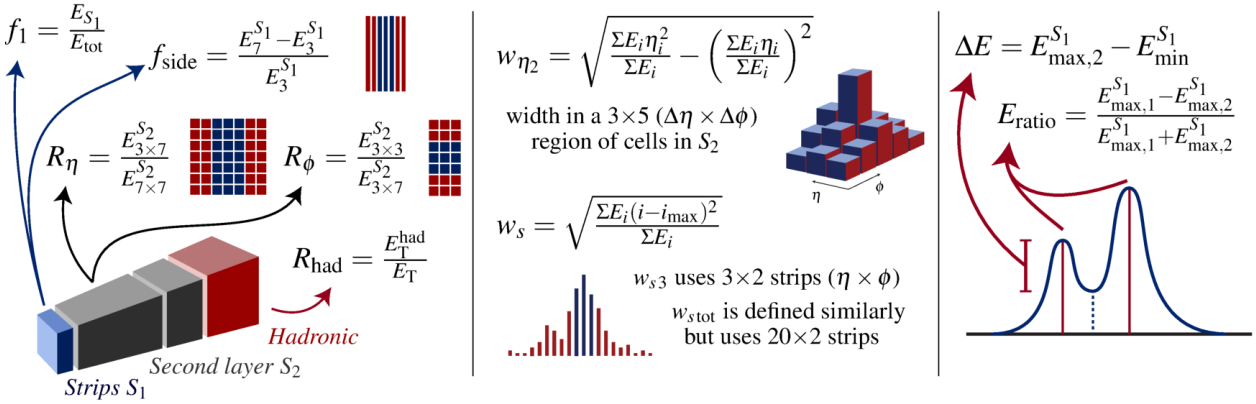


Figure 3

modified to resemble the data. This "fudged" MC 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 data during training after the decorrelation models reach a mature state. The team will use the previously developed toy model as a basis to develop decorrelation methods 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 [24], 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 decorrelation approach will 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 lowest 50% interquartile range, and with the best data/MC agreement will be chosen for further validation and integration into the ATLAS e/γ photon calibration procedure.

A GNN-based 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 MC. The PI's team with collaborate with the GNN calibration team to introduce decorrelation approaches to the powerful GNN calibration.

3.2 Photon identification

The ATLAS photon identification 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 mainly due to the calibration procedure for the photon ID, which requires the ID algorithm to be independent of the isolation. The decorrelation techniques, that this proposal aims to develop, 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. Once the move to multivariate methods is possible, the decorrelation can be extended to minimize data-MC difference as proposed for the photon energy calibration.

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 [25] NNs for object ID to develop a multivariate photon ID method, and other ATLAS collaborators who have used ML approaches for

photon ID. The decorrelation techniques will first be applied to a BDT or NN that takes high-level quantities as inputs, similar to those currently used in the cut-based approach. As with the energy resolution calibration, both the decorrelation 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 using the standard techniques described in [26, 27].

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

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

3.3 $t\bar{t}H(\rightarrow \gamma\gamma)$

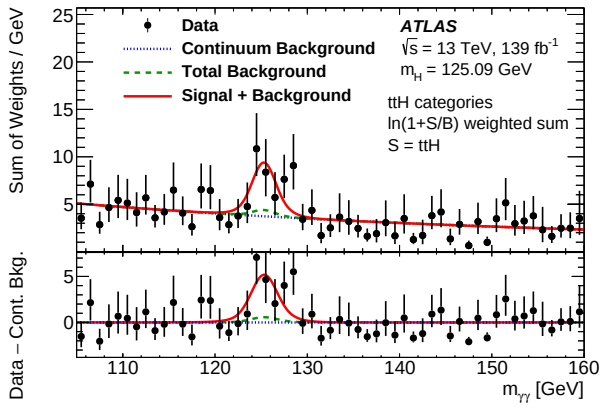


Figure 4: Distribution of the diphoton invariant mass in the $t\bar{t}H$ channel for the latest STXS ATLAS result [9]. The data (dots) is shown together with the sum of the fitted signal plus background (solid line). The blue dotted line represents the sum of the fitted continuum background, while the dashed line combines the contributions of continuum background and other Higgs boson events.

The most recent ATLAS results that performed many differential cross section measurements in the $H(\rightarrow \gamma\gamma)$ 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 background 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 decorrelation techniques to improve the photon calibration can also be used to ensure that an ML classifier is independent of the $m_{\gamma\gamma)}$ spectrum. This application of decorrelation 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.

Using the framework that was developed for the photon ID MVA and energy calibration, the two postdocs will work in parallel to determine which decorrelation approach, decorrelation versus adversarial, 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 first used in the $t\bar{t}H$ channel and then will be extended to other channels.

The PI's team will also contribute to the final interpretations of the $H(\rightarrow \gamma\gamma)$ result. The two interpretations, each of which will be performed by one of the postdocs and the PI, are withing the κ-modifier framework and the SMEFT.

3.4 Timetable of Activities

The timetable of activities is shown in Figure 5.

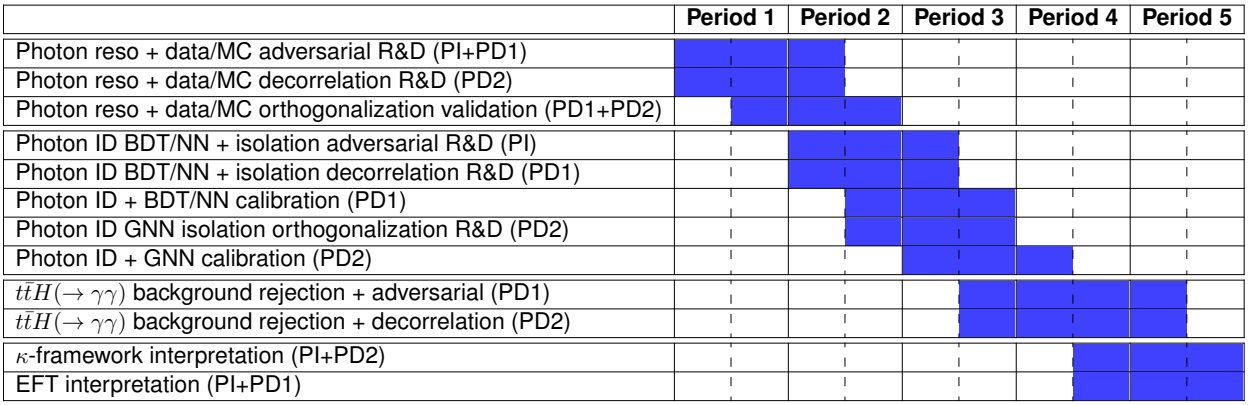


Figure 5: Expected time spent on the work proposed in this text by the PI's team which will consists of two postdocs (at 1.0 FTE each) and about half of the PI's effort. Each period is 12 months long with Period 1 starting on August 1st 2024 and Period 5 ending on July 31st 2029 which is two months after the HL-LHC is slated to start operations (May 2029).

Below is a summary of the milestones:

1. Finalization of framework that can apply both decorrelation as well as adversarial methods to decorrelate multivariate algorithms from certain quantities.
2. Reduction in photon energy resolution uncertainty due to decorrelation of data/MC differences.
3. Multivariate photon ID that is decorrelated from isolation
4. Improved background rejection in $t\bar{t}H$
5. SMEFT and κ interpretation of $t\bar{t}H$ result

3.5 Personnel and Resources

The PI will lead a team to conduct the proposed work, which will consist of an average of 2.0 postdoctoral FTEs and 50% of the PI's effort.

The postdocs, covered 100% by this award, is expected to start work at the beginning of the first Period. Both postdocs will be work on the proposed projects until Period 5. The postdoc's time will be split between developing decorrelation approaches and preparing the new $t\bar{t}H(\rightarrow \gamma\gamma)$ result.

The PI will also collaborate with graduate and undergraduate students on the proposed research, through the DOE Office of Science Graduate Student Research program and the Science Undergraduate Laboratory Internships, respectively. Furthermore, Argonne's ATLAS Center, which hosts graduate students from universities that are members of the ATLAS collaboration, provides other opportunities to collaborate on the proposed work. The PI will work with students from institutions with established collaborations with the Argonne ATLAS group, such as NIU. NIU has both photon and Higgs expertise but also has a computing traineeship which has synergies with this proposal due to the computing required to perform the proposed research.

Computing resources available at the ALCF include Polaris (which is an NVIDIA GPU farm) and Aurora (which has Intel GPUs). The resources are available to the PI in the form of

Director's Discretionary Allocations, which are routinely granted with consideration to clearly defined specifications. An additional computing resource for smaller-scale testing will also be available through the Laboratory Computing Resource Center.

4 Summary

Appendices

A 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

379

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.

B Bibliography and References

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

[2] 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.

[3] 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.

[4] *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, tech. rep., All figures including auxiliary figures are available at https://atlas.web.cern.ch/At CERN, 2018, URL: https://cds.cern.ch/record/2630973.

[5] 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].

[6] 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].

[7] P. Agrawal, D. Saha, L.-X. Xu, J.-H. Yu, and C.-P. Yuan, *Determining the shape of the Higgs potential at future colliders*, Phys. Rev. D **101** (2020) 075023, arXiv: 1907.02078 [hep-ph].

[8] F. Maltoni, D. Pagani, A. Shivaji, and X. Zhao, *Trilinear Higgs coupling determination via single-Higgs differential measurements at the LHC*, The European Physical Journal C **77** (2017).

[9] 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.

[10] ATLAS Collaboration, *Search for direct pair production of the top squark in all-hadronic final states in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, JHEP **09** (2014) 015, arXiv: 1406.1122 [hep-ex].

[11] ATLAS Collaboration, *ATLAS Run 1 searches for direct pair production of third-generation squarks at the Large Hadron Collider*, Eur. Phys. J. C **75** (2015) 510, [Erratum: Eur.Phys.J.C 76, 153 (2016)], arXiv: 1506.08616 [hep-ex].

[12] 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 **12** (2017) 085, arXiv: 1709.04183 [hep-ex].

[13] 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 **80** (2020) 737, arXiv: 2004.14060 [hep-ex].

[14] 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.

[15] 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.

[16] M. Paganini, L. de Oliveira, and B. Nachman, *CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, Phys. Rev. D **97** (2018) 014021, arXiv: 1712.10321 [hep-ex].

[17] A. Collaboration, *AtlFast3: The Next Generation of Fast Simulation in ATLAS*, Comput. Softw. Big Sci. **6** (2022) 7, arXiv: 2109.02551 [hep-ex].

[18] CMS Collaboration, *Evidence for associated production of a Higgs boson with a top quark pair in final states with electrons, muons, and hadronically decaying τ leptons at $\sqrt{s} = 13$ TeV*, Journal of High Energy Physics **2018** (2018), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP08(2018)066.

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

[20] 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].

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

[22] 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].

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

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

[25] *Physics Object Localization with Point Cloud Segmentation Networks*, tech. rep., All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ CERN, 2021, URL: https://cds.cern.ch/record/2753414.

[26] 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].

[27] 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].

C 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.

D Equipment

The requested funding will be used to purchase three laptops for the postdocs that the PI will supervise. The combined cost of the three laptops is expected to be $8,524.

E 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/.

F Promoting Inclusive and Equitable Research (PIER) Plan

G Recruitment and Retention of Students and Early-stage Investigators

H Other attachments

There are no other attachments.