

Incorporating Physical Priors into Weakly Supervised Anomaly Detection

Chi Lung Cheng^{1,2,*}, Gup Singh^{2,†} and Benjamin Nachman^{2,3,4,5,‡}

¹Department of Physics, *University of Wisconsin, Madison, Wisconsin 53706, USA*

²Physics Division, *Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

³Berkeley Institute for Data Science, *University of California, Berkeley, California 94720, USA*

⁴Department of Particle Physics and Astrophysics, *Stanford University, Stanford, California 94305, USA*

⁵Fundamental Physics Directorate, *SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA*



(Received 4 June 2024; revised 14 October 2024; accepted 13 May 2025; published 7 July 2025)

We propose a new machine-learning-based anomaly detection strategy for comparing data with a background-only reference (a form of weak supervision). The sensitivity of previous strategies degrades significantly when the signal is too rare or there are many unhelpful features. Our prior-assisted weak supervision (PAWS) method incorporates information from a class of signal models to significantly enhance the search sensitivity of weakly supervised approaches. As long as the true signal is in the prespecified class, PAWS matches the sensitivity of a dedicated, fully supervised method without specifying the exact parameters ahead of time. On the benchmark LHC Olympics anomaly detection dataset, our mix of semisupervised and weakly supervised learning is able to extend the sensitivity over previous methods by a factor of 10 in cross section. Furthermore, if we add irrelevant (noise) dimensions to the inputs, classical methods degrade by another factor of 10 in cross section while PAWS remains insensitive to noise. This new approach could be applied in a number of scenarios and pushes the frontier of sensitivity between completely model-agnostic approaches and fully model-specific searches.

DOI: [10.1103/8259-wt5p](https://doi.org/10.1103/8259-wt5p)

Introduction—Even if we could enumerate all potential new physics scenarios discoverable at particle physics experiments, it would be impossible to conduct a search for each one. Therefore, physics model-agnostic *anomaly detection* searches are an essential complement to the traditional physics model-specific searches. Modern machine learning has ignited a rapidly growing literature on innovative anomaly detection strategies that are projected to qualitatively extend the sensitivity of the particle physics search program [1–3]. Several of these strategies have already been applied to data [4–11].

One powerful set of anomaly detection strategies is based on classifiers trained to distinguish data from a background-only reference [12–24]. If the background-only reference is statistically identical to the background in the data, then the classifier will be able to pick up on any nonbackground component in the data. This weakly supervised (WS) strategy is asymptotically optimal [17] and various methods are distinguished by how they

construct the background reference, how they construct the classifier, and how they determine the statistical compatibility of the data with the reference.

Until recently, most studies of the background reference strategy used low-dimensional, high-level features. Additional features can degrade the sensitivity of the search, especially when the new features have the same spectrum for signal and background [25,26]. The authors of Ref. [27] have shown how to employ the background reference strategy to high-dimensional, low-level features, but the sensitivity degrades rapidly with low signal cross sections. We hypothesize that existing methods use overly flexible classifiers, which allow them to model unphysical features, thereby reducing sensitivity. Our solution is to incorporate physical priors to reduce the space of functions.

To this end, we introduce a prior-assisted weak supervision (PAWS) strategy for anomaly detection. Our approach starts with specifying a class of signal models, parametrized by $\theta \in \mathbb{R}^N$. These parameters could be masses, couplings, or other physical properties. As a first step, we use simulated signal samples to pretrain neural network classifiers parametrized by θ [28,29] using features $x \in \mathbb{R}^M$. Then, we use the same classifier to distinguish the data from the background-only reference. In the second classification step, we freeze all of the network weights and promote θ to be learnable. Whereas existing anomaly detection methods degrade quickly with the number of features M , the performance of PAWS is largely independent of M , but degrades with N . This is because each

*Contact author: ccheng84@wisc.edu

†Contact author: gupsingh@berkeley.edu

‡Contact author: bpnachman@lbl.gov

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

parameter dimension effectively behaves like an extra feature. Typically, $N \ll M$, so the stability of PAWS with M is beneficial. In other words, the set of PAWS classifiers exist on a lower-dimensional manifold of functions than a typical unconstrained classifier. By only considering these physical classifiers, the sensitivity may be significantly improved compared to more general searches without the pretraining step. Additionally, since θ is optimized like any other neural network parameter via gradient descent, far fewer parameter values are checked compared to a classical search that would scan over all possibilities, which is itself intractable in many dimensions. The trade-off is that we assume the anomaly is in our signal class (and can be reasonably simulated). This is a strong constraint, but this still gives us access to large classes of physics models that cannot be searched for individually.

There have been a number of studies that have investigated related aspects of our approach. References [30,31] use a set of signal models to indirectly improve the performance to out-of-distribution signals. A pair of recent studies showed that less expressive classifiers (boosted decision trees) can help mitigate degradation due to irrelevant features [25,26]. References [32–35] use transfer learning from models trained on similar tasks to inject some physics information through the initialization of the classifier. Instead of pretraining on signal models, Ref. [36] recycles a background prediction to enhance the sensitivity to small signals. Such an approach is complementary to our proposal. Another way of viewing our approach is as a variation of neural simulation-based inference [37], where the signal fraction μ and physics parameters θ are inferred from the data. In the anomaly detection case, we are (mostly) interested in knowing if the classifier learns anything non-trivial, but a by-product of the approach is that there is also an estimate of the physics parameters.

Dataset—As a representative example, we use collisions at the Large Hadron Collider (LHC). We build on the LHC Olympics (LHCO) R&D dataset [1,38] and the extended background dataset from Ref. [19] for our studies. The original dataset consists of generic quark-gluon scattering processes to produce dijet events as the standard model background and $W' \rightarrow XY$ with $X, Y \rightarrow q\bar{q}$ as signal. Both processes were simulated using PYTHIA8.219 [39,40] and DELPHES3.4.1 [41,42]. When $m_{W'} \gg m_X, m_Y$, the X and Y particle decay products will be mostly captured inside a single jet each and their invariant mass will be close to $m_{W'}$. The original dataset had $m_{W'} = 3.5$ TeV, $m_X = 500$ GeV, and $m_Y = 100$ GeV. To train our parametrized classifier and to demonstrate the performance on a variety of signal models, we use the same simulator settings to generate an entire grid of 144 signal models with $m_X, m_Y < 600$ GeV in increments of 50 GeV. The upper mass bound ensures that the quarks are well contained within a single jet. In addition to the two-prong $X, Y \rightarrow qq$ decay, we also include the three-prong $Y \rightarrow qqq$ decay from Ref. [38] to increase the dimensionality of our parameter space.

In total, each decay features 144 signal models, with around 100 thousand events each. The parameters of our signal model set are $\theta = \{m_X, m_Y, \alpha \equiv \text{BR}(Y \rightarrow qqq)\}$, so $N = 3$. When it would be ambiguous, we will denote $\hat{\theta}$ to mean the true values instead of the fitted values. We briefly discuss the sensitivity to signals outside of this set in the End Matter.

The reconstructed particles of each event are clustered into jets using the anti- k_T algorithm [43–45] with $R = 1.0$. All events are required to satisfy $p_T > 1.2$ TeV. The two highest p_T jets J_1 and J_2 are sorted such that $m_{J_1} > m_{J_2}$. Each event is characterized by the same six features ($M = 6$) that have been widely used in LHCO studies: $x = \{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}, \tau_{32}^{J_1}, \tau_{32}^{J_2}\}$, where the n -subjettiness ratios are defined as $\tau_{ij} \equiv \tau_i/\tau_j$ [46,47]. These are sensitive to the two- (τ_{21}) or three- (τ_{32}) prong nature of the jets.

The signal is resonant in m_{JJ} and a full analysis would scan over m_{JJ} to form signal regions and sideband regions. As in many previous studies, for simplicity, we focus on the one signal region $m_{JJ} \in [3.3, 3.7]$ TeV that contains the signal. The sideband $m \notin [3.3, 3.7]$ TeV would be used to both estimate the background-only distribution of features $p_B(x)$ for training classifiers and to estimate the total number of background events after applying a threshold in the classifier. Previous studies have shown that deep learning is very effective for estimating p_B from data (see, e.g., Ref. [19]) and so we assume it is already known; instead, we focus on how to use samples from p_B to achieve signal sensitivity. Future experimental studies would synthesize all of these components into a full analysis. In practice, our reference dataset is statistically identical to and has the same size [48] as the background in our actual “data.”

Methods—Given features x and parameters θ , we start with a fully supervised (FS) parametrized classifier $f_{\text{FS}}(x, \theta)$ trained with the binary cross entropy (BCE) loss:

$$-\sum_{(x,\theta) \in S} \log[f_{\text{FS}}(x, \theta)] - \sum_{(x,\theta) \in B \times \Theta} \log[1 - f_{\text{FS}}(x, \theta)], \quad (1)$$

where S are signal events specified by their features and their parameters, and B are background events specified by their features. The background here is the same as the reference dataset. The signal dataset is generated such that every event uses parameters θ_i in the set Θ ; i.e., $\theta_i \in \Theta$. The notation $B \times \Theta$ in Eq. (1) represents the fact that the θ values used for the background sample are independent from the features and distributed in the same way as for the signal. This means that θ itself is not useful for classification. In practice, there are about 700 thousand background events that are repeated 144 times to form $B \times \Theta$ along with the approximately 10 million signal events to form S .

Asymptotically, the optimal f_{FS} from Eq. (1) will approximate

$$f_{\text{FS}}^* \approx \frac{P_S(x|\theta)p(\theta)}{P_S(x|\theta)p(\theta) + P_B(x)p(\theta)} = \frac{P_S(x|\theta)}{P_S(x|\theta) + P_B(x)}, \quad (2)$$

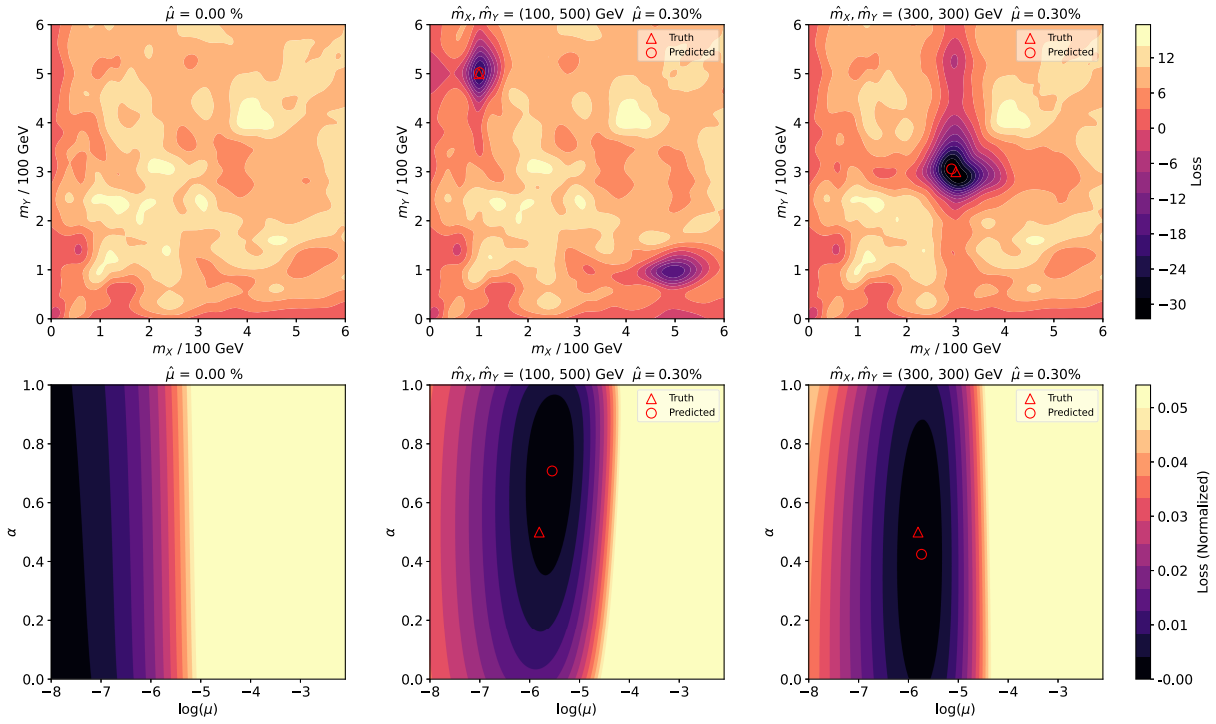


FIG. 1. The weakly supervised loss as a function of m_X, m_Y, μ , and α . The top row varies m_X and m_Y with fixed true values μ and α , while the bottom row scans μ and α with m_X and m_Y fixed to the true values. The left-hand column shows the loss for a pure background sample, while the middle and right-hand columns correspond to a 0.3% signal injection at $(\hat{m}_X, \hat{m}_Y) = (100, 500)$ and $(300, 300)$ GeV, respectively. Loss values in the μ - α scans are min-max normalized per plot for better visual comparison across different signal injections, which would otherwise have vastly different scales due to the behavior of μ .

where P_i is the probability density p_i multiplied by the number of events N_i used in training. Define $\kappa(\theta) \equiv N_B/N_S(\theta)$; the θ dependence is due to acceptance effects that results in slightly uneven number of signal events in training as a function of the particle masses. We can extract the likelihood ratio from the classifier: $p_S(x|\theta)/p_B(x) \approx \kappa(\theta)f_{\text{FS}}^*(x)/[1 - f_{\text{FS}}^*(x)] \equiv \Lambda_{\text{FS}}(x|\theta)$.

Our weakly supervised classifiers are also trained using the BCE loss [12,13,49]:

$$-\sum_{(x) \in D} \log[f_{\text{WS}}(x)] - \sum_{(x) \in R} \log[1 - f_{\text{WS}}(x)], \quad (3)$$

where D is the data with $p_D(x) = \mu p_S(x|\theta) + (1 - \mu)p_B(x)$ and R is the reference with $p_R = p_B$. Asymptotically, the optimal f_{WS} from Eq. (3) will approximate

$$f_{\text{WS}}^*(x) \approx \frac{P_D(x)}{P_D(x) + P_R(x)} = \frac{\mu \frac{p_S(x|\theta)}{p_B(x)} + (1 - \mu)}{\mu \frac{p_S(x|\theta)}{p_B(x)} + 2(1 - \mu)}. \quad (4)$$

Previous approaches parametrized f_{WS} as a generic neural network. In our new approach PAWS, we insert f_{FS}^* by replacing the likelihood ratio in Eq. (4) with Λ_{FS} :

$$\tilde{f}_{\text{WS}}(x|\mu, \theta) = \frac{\mu \Lambda_{\text{FS}}(x|\theta) + (1 - \mu)}{\mu \Lambda_{\text{FS}}(x|\theta) + 2(1 - \mu)}, \quad (5)$$

with the functional form of Λ_{FS} fixed and only θ and μ are optimized. When $p(x|\theta)$ factorizes, it is possible to add additional structure. This allows us to consider both two-prong and three-prong decays of the Y particle. In particular, we train separate supervised classifiers for each decay mode and then mix the likelihood ratios according to the branching fraction α .

Each model f_{FS} and f_{WS} is parametrized as a three-layer fully connected neural network with node configurations of (256,128,64). The model $\tilde{f}_{\text{WS}}$ is parametrized in terms of f_{FS} according to Eq. (5). The networks are implemented in TENSORFLOW [50] and optimized using ADAM [51] with early stopping. The data are divided so that 50% are used for training, 25% are used for validation, and the rest for testing. In weakly supervised training, the background events are split evenly between D and R and the signals are included in D according to μ and α . The datasets used to train f_{FS} and f_{WS} are mutually independent to prevent cross-contamination of training samples. In practice, a k -fold procedure would be used to ensure that all data could participate in both training and testing [4,12,13]. To improve robustness, we train 20 models with the same input data but different random network initializations. In the

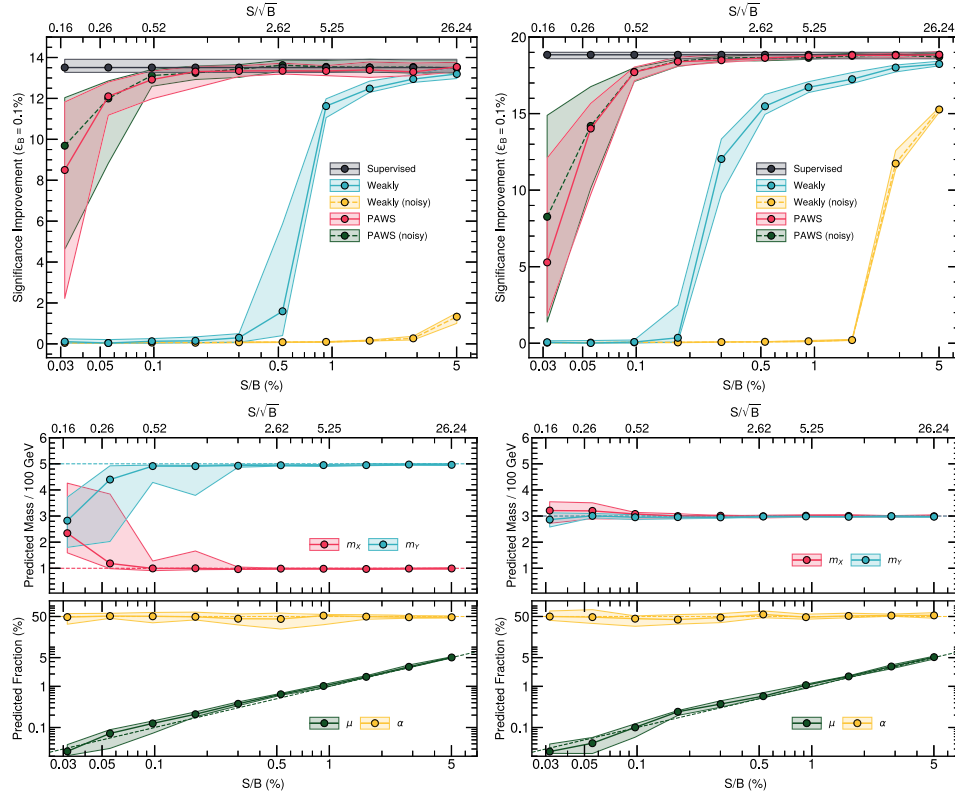


FIG. 2. Significance improvement at a fixed value of $\varepsilon_B = 0.1\%$ (top) and predicted parameter values (bottom) with varying amounts of signal injections for $(m_X, m_Y) = (100, 500)$ GeV (left) and $(300, 300)$ GeV (right). The shaded uncertainty bands quantify the standard deviation (around the median) from retraining the respective models with random realizations of the inputs. The top axis indicates the starting significance of the signal ($S/\sqrt{B}$). The dashed lines from the top plots include scenarios where 10 dimensions of pure noise are added to the inputs. The dashed lines from the bottom plots indicate the truth values of the parameters represented by the inputs.

case of $\tilde{f}_{WS}$, this corresponds to sampling uniformly in the range of masses with fixed values of $\mu = \exp(-4)$ and $\alpha = \exp(-1)$. Per event, the scores of the 10 models with the lowest validation loss are averaged. As there is a local minimum when $m_X \leftrightarrow m_Y$, we always check the symmetric point after training and take the model with the lowest validation loss.

We also found that it was useful to regularize the network with gradient clipping and the loss function with a penalty for θ that grows exponentially with the distance away from physically valid regions. That is, we constrain m_X and m_Y to lie within the range $[50, 600]$ GeV, penalizing values outside the boundary. The $f_{FS}(x|\theta)$ model was validated against a series of models trained with a particular value of θ and excellent agreement was observed. We did not extensively optimize the hyperparameters; the full set of choices can be found in the public software.

Our figure of merit for comparing different approaches is the significance improvement characteristic defined as $SIC = \varepsilon_S / \sqrt{\varepsilon_B}$, where ε_S and ε_B are the signal and background efficiencies, respectively. The SIC represents the multiplicative factor by which the signal significance increases for a selection on the network corresponding to ε_B . In practice, we do not know *a priori* what selection to

make, so we consider the value of the SIC at a fixed value of $\varepsilon_B = 0.1\%$, which we can determine using R . This is also more numerically stable than the maximum of the curve which has often been used in the past. In order to quantify the uncertainty from the finite data statistics, we compute the SIC for 10 random realizations of the training inputs.

Results—We find that PAWS works well across the entire grid of m_X and m_Y values. To illustrate the performance in detail, we focus on two benchmark points: $m_X, m_Y = (300, 300)$ and $(100, 500)$ GeV, showcasing scenarios of highly symmetric and asymmetric masses, respectively. The asymmetric model is also the same as the one in the R&D dataset of the LHCO. The branching ratio for the three-prong decay of the Y particle is set to $\alpha = 0.5$ and we scan over μ to show our sensitivity to low levels of signal.

Unlike traditional weak supervision with fully flexible neural networks, for PAWS, the learned parameters are physical and thus directly interpretable when the anomaly is in the pretrained model class. Figure 1 shows the weakly supervised loss [Eq. (3)] as a function of m_X and m_Y for α and μ fixed to their correct values. The μ values are chosen to approximately correspond to just before and then just about where the classical weakly supervised approach becomes sensitive. All four plots in Fig. 1 show clear

minima at the correct values of m_X and m_Y . In the asymmetric case, there is also a (local) minimum for the incorrect point with $m_X \leftrightarrow m_Y$. The classifier at this point is still very effective because m_{J_1} and m_{J_2} have nearly the same spectra as for the correct mass point.

When the correct mass points are chosen, PAWS matches the performance of the fully supervised model using the same features. This is shown in Fig. 2. The fact that our model matches the dedicated fully supervised approach for $S/B \gtrsim 0.1\%$ also is a validation that the parametrized model is effective. PAWS achieves sensitivity to signals that are more than a factor of 10 weaker than for the classical weakly supervised approach. This sensitivity is reaching the limit—for a signal fraction of 0.03%, the injected significance is only 0.16, and the fully supervised significance improvement is around 15–20. This means that the fully supervised approach would enhance the signal to only about 2–3. Therefore, no method can achieve useful performance below 0.03%.

The lower panels of Fig. 2 also demonstrate that the physics parameters are accurate. Because of the subtle differences between the two- and three-prong models, we are not very sensitive to the exact value of α .

In order to show the robustness of our approach to the dimensionality of the feature space, we added 10 pure noise dimensions to the input [25,26] ($M = 16$). Each of these dimensions is sampled from a standard normal distribution for both the signal and the background. Since our pre-trained models only consider physical information, they effectively learn to ignore these additional features. As a result, PAWS is unaffected by the noise, as shown by the dotted line in Fig. 2. In contrast, the traditional weakly supervised methods are strongly dependent on the noise and their sensitivity degrades by about a factor of 10 in μ .

Conclusions—We have introduced a prior-assisted weak supervision approach to anomaly detection. Like previous weakly supervised methods, PAWS is based on training a classifier to distinguish data from a background-only reference dataset. We impose structure on the function class of classifiers through a pretraining step based on a set of signal models. This trades off breadth for depth of sensitivity—we show that PAWS is able to improve the sensitivity of previous methods by a factor of 10–100 in signal cross section. While we require prespecifying the signal class, it can (and in our case, does) contain a large number of models that would be infeasible to probe one at a time. One could scan, in steps consistent with the experimental resolution, over the parameters of the models. In addition to the computational challenge, this would also incur a large trial factor that PAWS does not pay since it uses pretraining and gradient descent. Furthermore, the sensitivity of PAWS is at the limit of what is possible for any method using the same input features and so there is little to be lost compared to a dedicated approach.

There are a number of extensions of PAWS that would be exciting to explore in future work. One direction would be to

push the dimensionality of the parameter space. In our numerical example, the parameter space was three dimensional (m_X , m_Y , and α) in addition to the signal fraction μ . Adding parameters would simply require adding features to the classifiers, but it may be advisable to not scan over them uniformly when building the parametrized classifier. Another direction would be to push the dimensionality of the feature space. In our numerical example, the feature space was six dimensional. Preliminary studies suggest that the setup works well when using the full phase space [27], but the performance is so good that more data are required for training and testing.

It would also be interesting to explore the sensitivity to signals that are similar to the ones used in the pretraining, but not exactly the same. We found empirically that the model does not extrapolate well to m_X and m_Y outside of the training dataset, but this is not conclusive because the jet substructure also changes significantly (e.g., the decay products are no longer contained within a single jet). While our example focused on resonant anomalies, we expect that the PAWS strategy would work in any scenario where data are compared to a background-only reference. Combining PAWS with other weakly supervised approaches for signal sensitivity and background estimation could lead to further improvements. Ultimately, we expect that a variety of strategies with complementary assumptions and strengths will be required to achieve broad discovery potential at the LHC and beyond.

Acknowledgments—We thank V. Mikuni, R. Mastandrea, and D. Shih for helpful discussions and feedback on the manuscript. C. L. C., G. S., and B. N. are supported by the U.S. Department of Energy (DOE), Office of Science under Contract No. DE-AC02-05CH11231. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships (SULI) program. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC Award No. HEP-ERCAP0021099. C. L. C. is also supported by the U.S. DOE Office of Science under Contract No. DE-SC0017647.

Data availability—The original LHC Olympics datasets are available on Zenodo. Our additional signal data, with the corresponding scripts to generate them, are also on Zenodo. The code we used to train and evaluate our models is available at [52].

-
- [1] G. Kasieczka *et al.*, The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics, *Rep. Prog. Phys.* **84**, 124201 (2021).
 - [2] T. Aarrestad *et al.*, The dark machines anomaly score challenge: Benchmark data and model independent event

- classification for the large hadron collider, *SciPost Phys.* **12**, 043 (2022).
- [3] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, Machine learning in the search for new fundamental physics, *Nat. Rev. Phys.* **4**, 399 (2022).
 - [4] Aad, Georges *et al.* (ATLAS Collaboration), Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector, *Phys. Rev. Lett.* **125**, 131801 (2020).
 - [5] Aad, Georges *et al.* (ATLAS Collaboration), Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $s = 13$ TeV with the ATLAS detector, *Phys. Rev. Lett.* **132**, 081801 (2024).
 - [6] Aad, Georges *et al.* (ATLAS Collaboration), Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector, *Phys. Rev. D* **108**, 052009 (2023).
 - [7] Chekhovsky, Vladimir *et al.* (CMS Collaboration), Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV, Technical Report, CERN, Geneva, 2024.
 - [8] D. Shih, M. R. Buckley, L. Necib, and J. Tamanas, via machinae: Searching for stellar streams using unsupervised machine learning, *Mon. Not. R. Astron. Soc.* **509**, 5992 (2021).
 - [9] D. Shih, M. R. Buckley, and L. Necib, Via Machinae 2.0: Full-sky, model-agnostic search for stellar streams in Gaia DR2, *Mon. Not. R. Astron. Soc.* **529**, 4745 (2024).
 - [10] M. Pettee, S. Thanvantri, B. Nachman, D. Shih, M. R. Buckley, and J. H. Collins, Weakly-supervised anomaly detection in the milky way, *Mon. Not. R. Astron. Soc.* **527**, 8459 (2023).
 - [11] D. Sengupta, S. Mulligan, D. Shih, J. A. Raine, and T. Golling, SkyCURTAINS: Model agnostic search for Stellar Streams with Gaia data, *Mon. Not. R. Astron. Soc.* **536**, 1104 (2024).
 - [12] J. H. Collins, K. Howe, and B. Nachman, Anomaly detection for resonant new physics with machine learning, *Phys. Rev. Lett.* **121**, 241803 (2018).
 - [13] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, *Phys. Rev. D* **99**, 014038 (2019).
 - [14] R. T. D’Agnolo and A. Wulzer, Learning new physics from a machine, *Phys. Rev. D* **99**, 015014 (2019).
 - [15] R. T. D’Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning multivariate new physics, *Eur. Phys. J. C* **81**, 89 (2021).
 - [16] A. Andreassen, B. Nachman, and D. Shih, Simulation assisted likelihood-free anomaly detection, *Phys. Rev. D* **101**, 095004 (2020).
 - [17] B. Nachman and D. Shih, Anomaly detection with density estimation, *Phys. Rev. D* **101**, 075042 (2020).
 - [18] K. Benkendorfer, L. Le Pottier, and B. Nachman, Simulation-assisted decorrelation for resonant anomaly detection, *Phys. Rev. D* **104**, 035003 (2021).
 - [19] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Classifying anomalies through outer density estimation (CATHODE), *Phys. Rev. D* **106**, 055006 (2022).
 - [20] R. T. d’Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning new physics from an imperfect machine, *Eur. Phys. J. C* **82**, 275 (2022).
 - [21] T. Golling, S. Klein, R. Mastandrea, and B. Nachman, Flow-enhanced transportation for anomaly detection, *Phys. Rev. D* **107**, 096025 (2023).
 - [22] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, Resonant anomaly detection without background sculpting, *Phys. Rev. D* **107**, 114012 (2023).
 - [23] T. Golling, G. Kasieczka, C. Krause, R. Mastandrea, B. Nachman, J. A. Raine, D. Sengupta, D. Shih, and M. Sommerhalder, The interplay of machine learning-based resonant anomaly detection methods, *Eur. Phys. J. C* **84**, 241 (2024).
 - [24] D. Sengupta, S. Klein, J. A. Raine, and T. Golling, CURTAINS flows for flows: Constructing unobserved regions with maximum likelihood estimation, *SciPost Phys.* **17**, 046 (2024).
 - [25] T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangechaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, Tree-based algorithms for weakly supervised anomaly detection, *Phys. Rev. D* **109**, 034033 (2024).
 - [26] M. Freytsis, M. Perelstein, and Y. C. San, Anomaly detection in the presence of irrelevant features, *J. High Energy Phys.* **02** (2024) 220.
 - [27] E. Buhmann, C. Ewen, G. Kasieczka, V. Mikuni, B. Nachman, and D. Shih, Full phase space resonant anomaly detection, *Phys. Rev. D* **109**, 055015 (2024).
 - [28] K. Cranmer, J. Pavez, and G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers, [arXiv:1506.02169](https://arxiv.org/abs/1506.02169).
 - [29] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, Parameterized neural networks for high-energy physics, *Eur. Phys. J. C* **76**, 235 (2016).
 - [30] S. E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, and P. Harris, Quasi anomalous knowledge: Searching for new physics with embedded knowledge, *J. High Energy Phys.* **06** (2021) 030.
 - [31] C. K. Khosa and V. Sanz, Anomaly awareness, *SciPost Phys.* **15**, 053 (2023).
 - [32] H. Beauchesne, Z.-E. Chen, and C.-W. Chiang, Improving the performance of weak supervision searches using transfer and meta-learning, *J. High Energy Phys.* **02** (2024) 138.
 - [33] T. Golling, L. Heinrich, M. Kagan, S. Klein, M. Leigh, M. Osadchy, and J. A. Raine, Masked particle modeling on sets: Towards self-supervised high energy physics foundation models, *Mach. Learn. Sci. Tech.* **5**, 035074 (2024).
 - [34] V. Mikuni and B. Nachman, Solving key challenges in collider physics with foundation models, *Phys. Rev. D* **111**, L051504 (2025).
 - [35] V. Mikuni and B. Nachman, Method to simultaneously facilitate all jet physics tasks, *Phys. Rev. D* **111**, 054015 (2025).
 - [36] R. Das, G. Kasieczka, and D. Shih, Residual ANODE, [arXiv:2312.11629](https://arxiv.org/abs/2312.11629).
 - [37] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
 - [38] G. Kasieczka, B. Nachman, and D. Shih, Official datasets for LHC Olympics 2020 anomaly detection challenge (Version v6) [Data set] (2019), [10.5281/zenodo.4536624](https://doi.org/10.5281/zenodo.4536624).

- [39] T. Sjöstrand, S. Mrenna, and P.Z. Skands, PYTHIA6.4 physics and manual, *J. High Energy Phys.* **05** (2006) 026.
- [40] T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C.O. Rasmussen, and P.Z. Skands, An introduction to PYTHIA8.2, *Comput. Phys. Commun.* **191** (2015) 159.
- [41] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, and M. Selvaggi (DELPHES3 Collaboration), DELPHES3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [42] A. Mertens, New features in DELPHES3, *J. Phys. Conf. Ser.* **608**, 012045 (2015).
- [43] M. Cacciari and G.P. Salam, Dispelling the N^3 myth for the k_t jet-finder, *Phys. Lett. B* **641**, 57 (2006).
- [44] M. Cacciari, G.P. Salam, and G. Soyez, FastJet user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [45] M. Cacciari, G.P. Salam, and G. Soyez, The anti- k_t jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [46] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, *J. High Energy Phys.* **03** (2011) 015.
- [47] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, *J. High Energy Phys.* **02** (2012) 093.
- [48] Previous studies have shown that there may be advantages to oversampling the reference [19]—this could be combined with our approach in future studies.
- [49] E.M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, *J. High Energy Phys.* **10** (2017) 174.
- [50] M. Abadi *et al.*, TENSORFLOW: Large-scale machine learning on heterogeneous systems (2015), Software available from tensorflow.org.
- [51] D.P. Kingma and J. Ba, ADAM: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [52] C.L. Cheng, G. Singh, and B. Nachman, PAWS: Prior-Assisted Weak Supervision, v1.1.0, Zenodo, <https://github.com/hep-lbdl/PAWS>.

End Matter

Sensitivity Beyond the Prior—The left-hand column of Fig. 1 shows that in the absence of a signal, no bias is introduced in the selected phase space. The other figures demonstrate state-of-the-art sensitivity to signals within the family of priors used for training. For anomaly detection, this is sufficient: when there is nothing to find, nothing is found; and there exist new signals that can be found. Nonetheless, it is interesting to explore the case where new physics is present, but not exactly in the family of signals parametrized by the prior. While we cannot make completely general statements, we empirically observe that PAWS still retains good sensitivity if the signal is reasonably similar to the pretrained models. For example, a larger natural width (comparable to detector resolution) yields no appreciable degradation relative to Fig. 2. By contrast, Fig. 3 shows a more extreme mismatch: m_X and m_Y are well below the training range. Although the parametrized network must extrapolate, its fitted masses are still near the true values (albeit with greater variance). At low signal cross sections, the prior-based ansatz continues to outperform purely model-agnostic anomaly detection, but with enough signal, the latter eventually catches up. For intermediate masses (e.g., 20–50 GeV), performance lies between these two extremes.

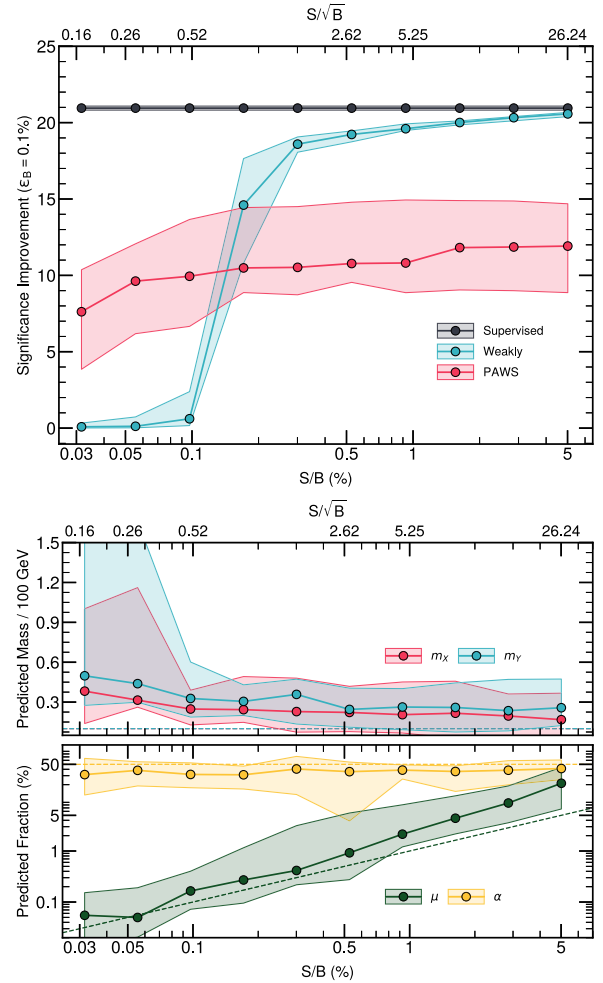


FIG. 3. Same as Fig. 2, but with $(m_X, m_Y) = (10, 10)$ GeV, well below the training range of 50–600 GeV.