

How to pick the best anomaly detector?

Marie Hein^{1,*} Gregor Kasieczka^{2,†} Michael Krämer^{1,‡} Louis Moureaux^{2,§}
Alexander Mück^{1,||} and David Shih^{3,¶}

¹*Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen University,
D-52056 Aachen, Germany*

²*Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*

³*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*



(Received 26 January 2026; accepted 29 June 2026; published 16 July 2026)

Anomaly detection has the potential to discover new physics in unexplored regions of the data. However, choosing the best anomaly detector for a given dataset in a model-agnostic way is an important challenge which has hitherto largely been neglected. In this paper, we introduce the data-driven ARGOS metric, which has a sound theoretical foundation and is empirically shown to robustly select the most sensitive anomaly detection model given the data. Focusing on weakly supervised, classifier-based anomaly detection methods, we show that the ARGOS metric outperforms other model selection metrics previously used in the literature, in particular the binary cross-entropy loss. We explore several realistic applications, including hyperparameter tuning as well as architecture and feature selection. In all cases, we demonstrate that ARGOS is highly sensitive to the small amounts of signal present in anomaly detection analyses while at the same time being robust to background overfitting and background template mismodeling.

DOI: [10.1103/9kcn-slvt](https://doi.org/10.1103/9kcn-slvt)

I. INTRODUCTION

In recent years there has been an explosion of new methods for model-agnostic new physics searches at the Large Hadron Collider (LHC), leveraging powerful modern machine-learning (ML) anomaly detection methods such as autoencoders and weak supervision [1–4]. These methods are able to find anomalies in the data with varying degrees of model independence: either signal-model-independence, in that they make no use of any specific new physics models in their training, and/or background-model-independence in that they make no use of Standard Model (SM) simulations in their training. As such, these new model-agnostic search strategies offer significant new opportunities for discovery in the vast, unexplored phase space of the LHC data, beyond what model-specific search strategies have been able to achieve to date.

However, one important question has been left largely unaddressed to date: how does one pick the best anomaly detector from all the available options? In our case, we consider the “best” anomaly detector to be the *most sensitive to unknown signals*, as measured by the standard significance improvement characteristic (SIC),

$$\text{SIC} = \frac{\epsilon_S}{\sqrt{\epsilon_B}}, \quad (1)$$

where $\epsilon_{S/B}$ is the signal/background efficiency after selecting the most anomalous events with the anomaly detector. The SIC is a multiplicative factor that quantifies the extent to which the anomaly detector enhances the significance of the new physics hiding in the data. Our goal is to select an anomaly detector that can achieve a high SIC sensitivity to any signal that is actually present in the given dataset.

So far, when it comes to different approaches (e.g. autoencoders vs weak supervision), the idea has been to just try them all, as seen in Ref. [5]. Within a given approach, there are myriad smaller but still vital choices to make: which model epoch to pick during training, which model architecture and hyperparameters to use, or even which features to use. Metrics traditionally used to guide these choices, such as the SIC and area under curve (AUC), are based on truth labels and require the use of benchmark signals. This is clearly not satisfactory for anomaly detection as the added signal dependence may, in the worst case, work against finding a signal actually present in the data.

*Contact author: marie.hein@rwth-aachen.de

†Contact author: gregor.kasieczka@uni-hamburg.de

‡Contact author: mkraemer@physik.rwth-aachen.de

§Contact author: louis.moureaux@cern.ch

||Contact author: mueck@physik.rwth-aachen.de

¶Contact author: shih@physics.rutgers.edu

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.

Consequently, none of the existing experimental anomaly detection analyses [5–12] report systematic model optimization. Model architectures and training parameters are generally derived from the original publication introducing the corresponding method, with adjustments to network capacity to deal with the larger experimental datasets. In some cases [5,8,11], parameters have been retuned using a small set of benchmark signals. In Refs. [5,11], anomaly scores have been selected from different classifier runs or epochs in certain approaches—CWoLa [13,14] Hunting and TNT [15] used the number of events selected in the signal region at a fixed background efficiency in the side bands as a metric; and CATHODE [16] used the validation loss, following Ref. [16]. In Ref. [17], combining different anomaly detection models has been studied in the context of the NPLM methodology [18].

In this paper we introduce ARGOS (above random gain of SIC), a new *fully data-driven* metric for selecting the best anomaly detector. The key ingredient (besides the anomaly detector) that is required to calculate ARGOS is what we refer to as a background template (BT), which is a sample of events following the distribution of SM background in the signal region (SR). Then ARGOS is defined to be

$$\text{ARGOS} = \frac{\epsilon_{\text{SR}}}{\sqrt{\epsilon_{\text{BT}}}} - \sqrt{\epsilon_{\text{BT}}}, \quad (2)$$

where $\epsilon_{\text{SR/BT}}$ is the efficiency to select SR/BT events for a given anomaly threshold. For a worse-than-random classifier, ARGOS would become negative, so we would never select models with a preference to select BT events. By definition, ARGOS is equivalent to the standard test statistic for an excess in a cut-and-count experiment with N_{SR} signal region events, where the observed number of events $\epsilon_{\text{SR}}N_{\text{SR}}$ after the anomaly cut is compared to the expected event number $\epsilon_{\text{BT}}N_{\text{SR}}$ for the null hypothesis. This test statistic has also been used for specific anomaly detection methods [19,20]. Here, however, we advocate ARGOS as a metric for optimizing anomaly detection performance. In contrast to SIC, which cannot be calculated for real (unlabeled) data, ARGOS is accessible for any anomaly detector if there is a corresponding background template in addition to the (unlabeled) dataset. In Sec. II, we show that ARGOS has a sound theoretical foundation, being monotonic with the actual signal-to-background SIC for any unknown signal at a fixed background efficiency when the background template is perfect. More generally, ARGOS is proportional to SIC to a very good approximation in any meaningful setting.

For concreteness, we demonstrate the utility of ARGOS within the context of weak supervision (also called classification without labels, CWoLa [21]), where a classifier is trained to distinguish two mixed-label datasets, for resonant anomaly detection. However, nothing in principle limits ARGOS to this context. Using the LHC olympics 2020

(LHCO) dataset [1], introduced in Sec. III, we demonstrate how ARGOS can consistently select the best hyperparameters, architecture (e.g. neural networks vs different boosted decision trees) and model epoch in Sec. IV. As an outlook, we also briefly discuss the potential for ARGOS to select the best features in a fully data-driven way before we conclude in Sec. V. Technical aspects of our experiments are discussed in several Appendixes.

II. THE ARGOS METRIC

In this section, we show that ARGOS, as introduced in Sec. I, is the optimal data-driven metric for evaluating anomaly detectors when using an ideal BT. For this purpose, we rewrite the ARGOS metric as given in Eq. (2) in terms of the background and signal efficiencies using the probability density of the SR,

$$p_{\text{SR}}(x) = f_S \cdot p_S(x) + (1 - f_S) \cdot p_B(x), \quad (3)$$

where f_S is the signal fraction in the signal region and $p_{S/B}(x)$ denote the probability density for the signal/background events. For an ideal BT, we have

$$p_{\text{BT}}^{\text{ideal}}(x) = p_B(x), \quad (4)$$

which we try to approximate in real-world applications as discussed in Sec. III. Using these densities, which imply $\epsilon_{\text{BT}} = \epsilon_B$, we find

$$\text{ARGOS} = \frac{f_S \cdot \epsilon_S + (1 - f_S) \cdot \epsilon_B}{\sqrt{\epsilon_B}} - \sqrt{\epsilon_B} \quad (5)$$

$$= f_S \cdot (\text{SIC} - \sqrt{\epsilon_B}). \quad (6)$$

Thus we see that ARGOS and SIC are closely related, at least in the idealized setting.

In more detail, there are multiple ways one can view this result:

- (i) At fixed ϵ_B , ARGOS is monotonic with SIC, since the signal fraction f_S is a constant. Then the best anomaly detector with respect to the ARGOS metric is also the best anomaly detector with respect to SIC. Additionally, ARGOS and the number of selected signal region events are also monotonic in this case and therefore also result in the same optimal decisions. The latter was used in Refs. [5,11] to select epochs from multiple trainings of the CWoLa Hunting and TNT approaches.
- (ii) For ARGOS, however, there is no need to fix the background efficiency a priori. For a good anomaly detector, $\sqrt{\epsilon_B} \ll \text{SIC}$ also holds, such that the maximum value of ARGOS as a function of score threshold is also close to the maximum value of SIC. Then an added benefit of using the ARGOS metric in this way is that one can also select the optimal working point of the anomaly detector in a

data-driven way. In Ref. [20], instead anomaly detectors with different background efficiencies have been combined using the corresponding test statistic.

In Sec. IV, we take the second approach and employ the maximum value of ARGOS for optimization. We will show that it retains good performance for realistic BTs.

III. SETUP

A. Dataset

In Sec. IV, we perform experiments in a number of different scenarios, which test different metrics for optimizing the setup for anomaly detection. These experiments are performed using the LHC olympics 2020 (LHCO) [1] research and development dataset [22], which is a common benchmark dataset for LHC anomaly detection. It contains 10^6 QCD dijet background events and 10^5 signal events. The signal is a W' resonance at $m_{W'} = 3.5$ TeV with the decay $W' \rightarrow X(\rightarrow qq)Y(\rightarrow qq)$, where X and Y have masses $m_X = 100$ GeV and $m_Y = 500$ GeV, respectively. Additionally, we use 612858 SR background events from Ref. [23], which were generated for Ref. [16].

The simulation pipeline consists of Pythia 8 [24,25] and Delphes 3.4.1 [26]. Using FastJet [27], reconstructed particles are clustered with the anti- k_T algorithm [28] with a distance parameter of $R = 1$. All events must pass a leading-jet trigger using $p_T > 1.2$ TeV. For the classification, we use the features also used in Ref. [16]. We select the two highest p_T jets and use the mass of the lighter jet m_{J_1} , the mass differences of the two jets $\Delta m_J = m_{J_2} - m_{J_1}$ as well as both jets' 21-subjettiness ratios [29,30]. In Sec. IV E, additional feature sets based on Ref. [31] are used. The dijet mass m_{JJ} is used as the resonant feature.

For the anomaly detection scenarios, we use all 1M background events and inject N_{sig} signal events. We use the same signal events in all runs in analogy to the fixed dataset available in an analysis. We define our signal region using the dijet mass by requiring $3.3 \text{ TeV} \leq m_{JJ} \leq 3.7 \text{ TeV}$, i.e., a 0.4 TeV window around the $m_{W'}$ resonance, resulting in approximately 120000 SR background events. The additional SR background events from Ref. [23] constitute the background template for the idealized anomaly detector (IAD, see Sec. III B 1), and also form part of the test set, for which a total of 340 000 SR background and 20 000 SR signal events are used.

B. Weakly supervised methods for resonant anomaly detection

We focus on classifier-based resonant anomaly detection in this work, since these methods are predicated on the existence of an accurate BT that describes the smooth background in the SR. For resonant anomalies, the data-driven construction of the BT is usually based on adjacent sideband (SB) regions that facilitate smooth interpolation into the SR. There are a variety of different methods of

constructing the background template [13,14,16,21,32–41], and we use the following three to illustrate the utility of the ARGOS metric:

1. Idealized anomaly detector

The idealized anomaly detector (IAD) [16] is used as a benchmark which is available for simulated data only. Instead of constructing the BT in a data-driven way, the IAD simply uses background events simulated with the same pipeline that generated the background data in the SR. Hence, by definition, $p_{\text{BT}}^{\text{IAD}}(x) = p_{\text{BT}}^{\text{ideal}}(x) = p_B(x)$. We use approximately 272 000 such events.

2. CWoLa Hunting

CWoLa Hunting [13,14] directly uses data from short sidebands of 0.2 TeV on either side of the signal region as the background template, assuming there is little correlation of the features with the dijet invariant mass. Using short sidebands reduces correlations with the dijet mass. As these short sidebands are in total as wide as the SR, the resulting background template has a similar size to the SR.

3. CATHODE

CATHODE (classifying anomalies through outer density estimation) [16] uses conditional density estimation to interpolate the SB distribution into the SR: as the SB contains only very low amounts of signal events, learning $p(x|m \in \text{SB})$ is largely equivalent to learning the background distribution $p_B(x|m)$, which can then be sampled in the SR to obtain background events in the SR. The density estimator can be used to generate more BT events than there are SR events in order to obtain better background statistics. We use a fixed oversampling factor of 4 as in Ref. [5]. The implementation of the density estimation is discussed in the Appendix A 1.

C. Classifiers

We use a total of three different classifier setups for our test cases:

- (1) A standard multilayer perceptron (MLP) neural network (NN) with three hidden layers, which is trained using the binary cross-entropy (BCE) loss with the Adam optimizer for 100 epochs. We ensemble the best ten epochs according to our different metrics, following the ensembling strategy used in Ref. [16], where the best ten epochs based on the validation loss are used.
- (2) The HistGradientBoostingClassifier, which is a gradient boosted decision tree classifier implementation in scikit-learn [42]. It histograms the classification features before training. We ensemble 50 independent trainings with different 50-50 training-validation splits as was done in Ref. [31].

- (3) The `AdaBoostingClassifier`, which is an adaptive boosted decision tree classifier implementation in `scikit-learn` [42]. We build our implementation in analogy to Ref. [31] and perform an ensembling of ten independently trained classifiers using random 50-50 splits of the training dataset.

A more detailed explanation of these classifiers as well as their hyperparameters can be found in the Appendix A 2.

D. Performance metrics

Throughout this work, we use the following performance metrics:

- (1) The maximum SIC value with a statistics cutoff at a relative statistical error on the background efficiency of 20%. This is a supervised metric for the anomaly detection performance calculated on the large held-out test set defined in Sec. III A. We refer to this metric as “max SIC” in the following.
- (2) The BCE loss calculated on the validation set, which is the standard metric used to perform model optimization or select the best epochs during training (see e.g. Refs. [5,16]).
- (3) The maximum ARGOS calculated on the validation set as defined in Eq. (2). We refer to this metric as “max ARGOS” in the following.

IV. EXPERIMENTS

There are a number of different applications for a metric that can analyze the performance of weakly supervised anomaly detection. Here, we discuss four examples, namely: selecting the best epochs to ensemble from a neural network training; optimizing hyperparameters; selecting an architecture; and selecting features for anomaly detection. However, we first highlight the correlation of ARGOS and SIC.

A. Correlation of ARGOS and SIC

In Fig. 1, we show our three performance metrics (see Sec. III D) tracked throughout a NN training for the IAD setup (see Sec. III B 1).

Comparing the standard BCE validation loss to the max SIC, there is no visible correlation. Instead, the validation loss very quickly shows overtraining, where the classifier tries to tell apart indistinguishable background events on the basis of statistical fluctuations. As the BCE loss takes the average over all events and most events are background, they dominate the metric.

ARGOS instead focuses on the events with the highest anomaly score, since max ARGOS is found for large background rejection $1/\epsilon_B$. Therefore it is not as sensitive to background overtraining. While we expect an almost perfect correlation between max SIC and max ARGOS according to Eq. (5), the validation-set ARGOS is severely

statistics limited by the small number of signal events in the data. Thus the max ARGOS has a slightly noisy correlation with the max SIC but shows the same underlying trend.

In this particular example, the signal is not found at the beginning of the training but instead after about 30 epochs, where the BCE already shows significant overtraining. This is not the case for every network initialization but it is common for the max SIC to show a delayed onset of the signal identification for lower signal injections, which the BCE is unable to identify. Hence, using early stopping could be counter productive.

B. NN epoch selection

As noted in Sec. III C, we ensemble the ten best epochs for the NN to obtain a lower variance and to stabilize the anomaly detection performance. Which epochs are selected strongly depends on the metric used, as can be seen in Fig. 1 for example. The impact of ensembling based on our three tested metrics for the IAD, CWoLa Hunting and CATHODE is shown in Fig. 2. We show the median anomaly detection performance of ten trainings as well as an error band representing the 16% to 84% quantiles.

By definition, selecting the epochs based on our supervised target metric, the max SIC, results in the optimal performance in all cases. Selecting using the max ARGOS results in a slightly lower and somewhat noisier performance but clearly outperforms the selection based on the BCE loss, which especially falls short for low signal injections.

This difference is particularly pronounced for CWoLa Hunting: when using BCE for epoch selection, the max SIC drops below 10 already at $N_{\text{sig}} = 900$, whereas with ARGOS this drop only occurs around $N_{\text{sig}} = 600$. In our setup, CWoLa Hunting generally has the worst quality background template of the three methods considered.¹ Therefore, the NN can learn to distinguish background template and signal region based on the background events only, which results in a lower validation BCE loss but does not transfer onto a better performance on the anomaly detection test set. Instead using the ARGOS metric, and in particular max ARGOS, allows the epoch selection to focus on the events with the highest anomaly score.

An interesting feature can be seen in the performance of the max SIC based epoch selection at $N_{\text{sig}} = 0$, where the performance is slightly better than random. The fluctuations of the classifier weights in the absence of signal still lead to some epochs with nonzero anomaly detection performance, which the max SIC identifies. Selecting the epochs using signal information can therefore enhance the false positive rate. For our data-driven metrics, the BCE loss and ARGOS,

¹The background template of the IAD contains no mismodeling by definition and the flow-matching based density estimate used for CATHODE is highly accurate for the small feature space used here.

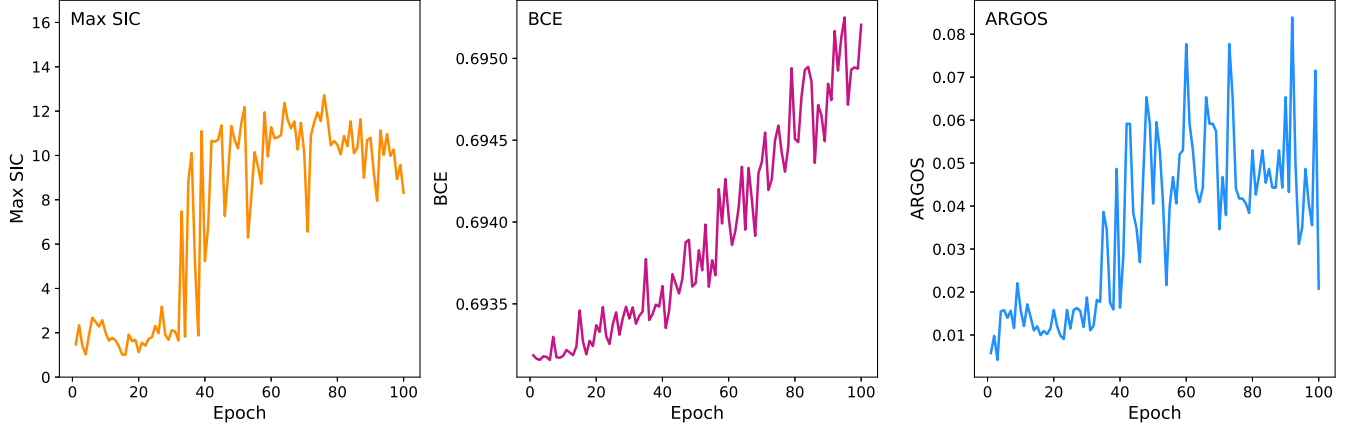


FIG. 1. Example of metrics tracked throughout a NN training with $N_{\text{sig}} = 400$ signal events using the default hyperparameters. We show the supervised max SIC metric (left) evaluated on the test set as well as BCE (middle) and max ARGOS (right), both evaluated on the validation set.

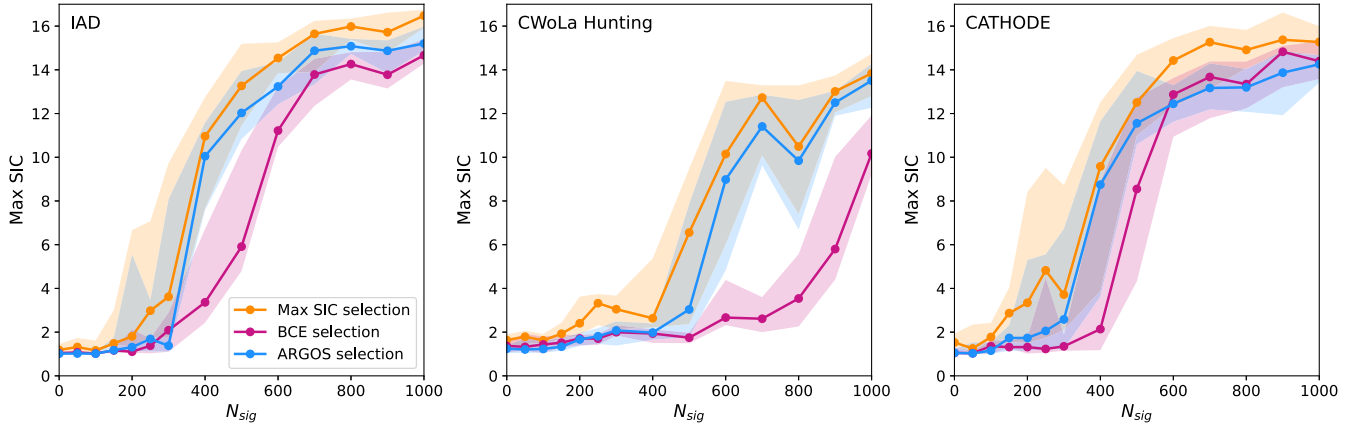


FIG. 2. Anomaly detection performance (max SIC) as a function of the number of signal events N_{sig} after epoch selection. The epoch selection is performed using the supervised benchmark metric max SIC and the two data-driven metrics (max ARGOS and BCE), shown for IAD (left), CWoLa Hunting (middle) and CATHODE (right).

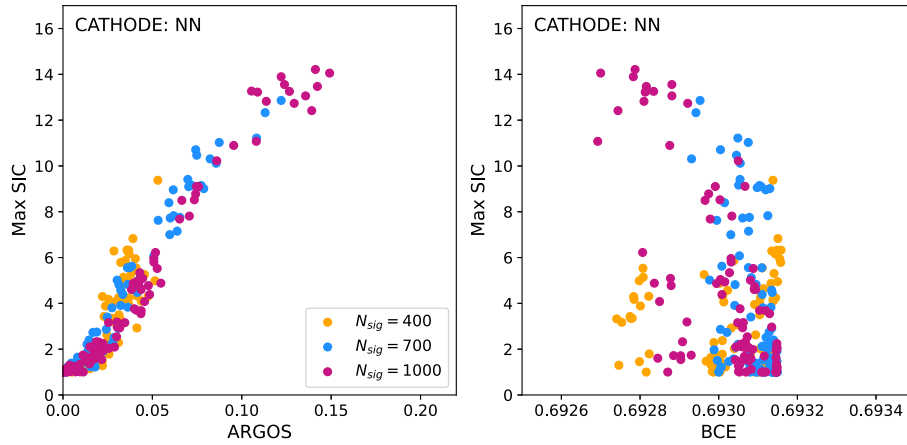


FIG. 3. Correlation between median anomaly detection performance max SIC and two data-driven metrics max ARGOS (left) and BCE (right) for all 100 hyperparameter sets for CATHODE using the NN classifier at three example signal injections.

this effect is not observed and in fact cannot occur. When no signal is present in the signal region, in contrast to maxSIC, the data-driven metrics are only calculated on datasets which also do not contain any signal. Hence, by construction, neither ARGOS nor the BCE loss are able to select fluctuations of the classifier weights which select the signal. We discuss why we do not generally expect an enhanced rate of false discoveries or enhanced background sculpting when selecting models on data in the Appendix D.

In the rest of the paper, we select epochs using the max ARGOS metric, as this is the best-performing data-driven metric for this application.

C. Hyperparameter optimization

An especially important application of a data-driven metric for weak supervision is data-driven hyperparameter optimization. As an illustration, we perform this

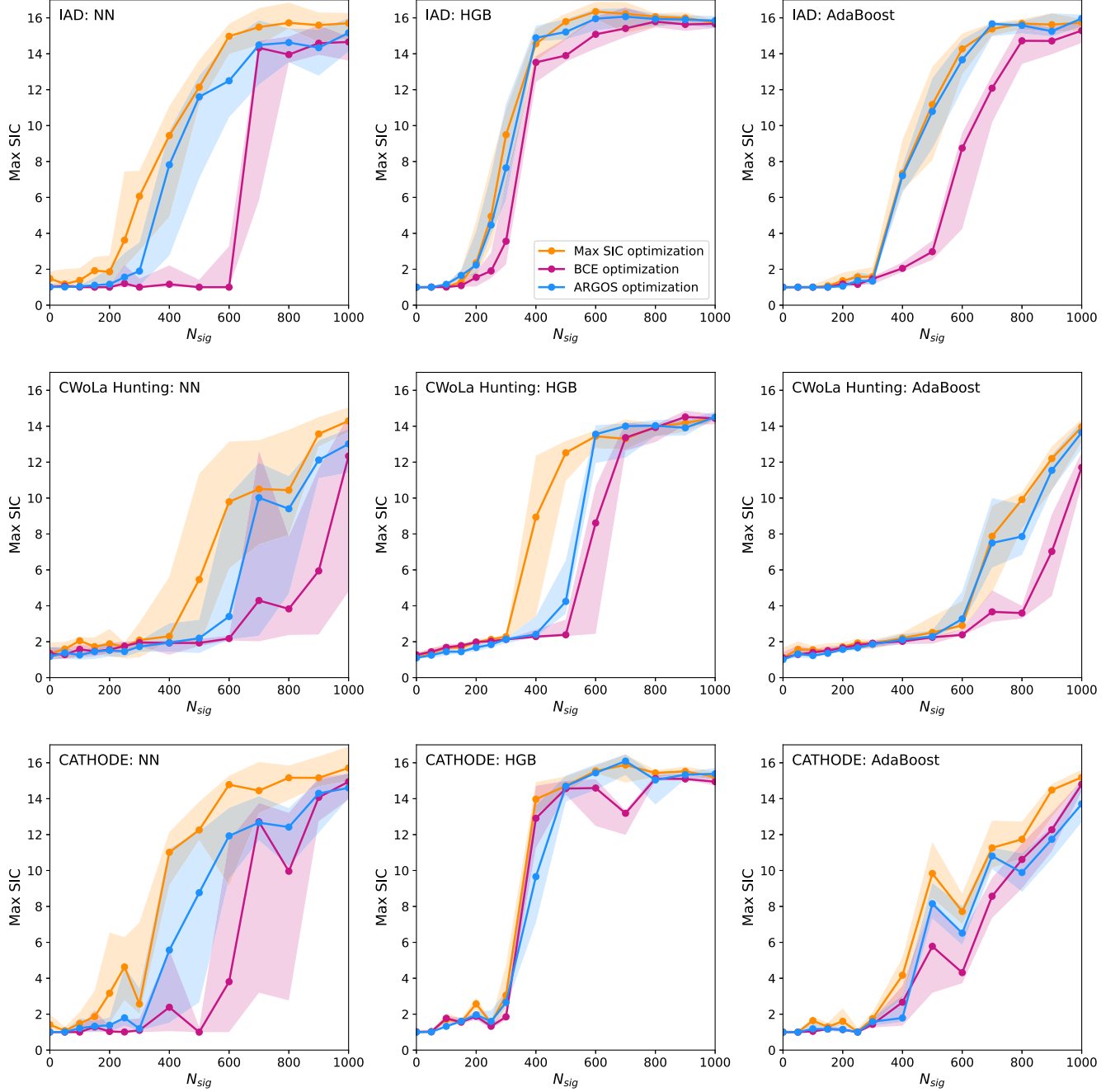


FIG. 4. Anomaly detection performance (max SIC) after hyperparameter optimization for the NN (left), HGB (middle) and AdaBoost (right) classifiers. The optimization is performed using the supervised benchmark metric max SIC and the two data-driven metrics (max ARGOS and BCE), shown for IAD (top), CWoLa Hunting (middle), and CATHODE (bottom).

optimization using a random-search strategy, where we randomly sample different hyperparameters $N_{\text{hp}} = 100$ times from a range of possible values as given in the Appendix A 2. Using each hyperparameter set, the classifier is trained, and all three metrics are evaluated; BCE and ARGOS on the validation set and max SIC on the test set. Then the best hyperparameter set based on each metric is selected. This procedure is performed ten times for each signal injection N_{sig} to obtain a stable median as well as 16% to 84% quantile error bands. A more detailed explanation of the procedure can be found in the Appendix B.

In Fig. 3, we show the correlation between our data-driven metrics and the anomaly detection performance (max SIC) for the NN based CATHODE optimization at three different signal injections. The range of different hyperparameter configurations result in a large spectrum of performance. We observe a much stronger correlation with max SIC for max ARGOS than for BCE, particularly for lower signal injections. For example, we find relatively low BCE values for low max SIC values at $N_{\text{sig}} = 400$. These low BCE values are likely achieved by optimizing on the differences between BT and SR background. We do not see a similar effect for the ARGOS metric. Analogous results are also obtained for the IAD and CWoLa Hunting and the other classifier architectures. Hence, we expect the optimization based on ARGOS to be more stable and to yield better results, especially at lower signal injections.

This is demonstrated in Fig. 4, which shows the significance improvement after hyperparameter optimization with the different metrics. Starting with the IAD results of the hyperparameter optimization in the top row of Fig. 4, we can once more see that an optimization using the supervised max SIC metric of course leads to ideal results. Second best in all cases is optimizing based on ARGOS, which in particular for the NN leads to a slightly noisier but still similar performance compared to optimizing on the

max SIC. Optimizing on the BCE loss results in a worse anomaly detection performance, which is especially significant for the NN and AdaBoost. This effect is less prominent for the HGB, likely because its histogramming of the input data limits overtraining. These trends generally persist for CWoLa Hunting and CATHODE in the bottom two rows of Fig. 4, although the imperfect background templates cause the results to be slightly noisier, especially for CATHODE where the individual density estimation runs cause the BT quality to slightly vary between different signal injections.

D. Architecture selection

In addition to optimizing the hyperparameters of a given architecture, one may also be interested in selecting the architecture itself. We train our three architectures (with hyperparameters optimized using ARGOS) on 50% of our standard training and validation data. On the other half, we calculate BCE and ARGOS once the trainings are complete. We can then select the best architectures based on these two metrics as well as on the max SIC of the half-statistics training. The selected architecture is then trained on the full statistics and evaluated on the test set. We perform a separate architecture selection for each of the ten optimized hyperparameter sets obtained in Sec. IV C at each signal injection. We plot these ten runs as a median and an error band corresponding to the 16% to 84% quantiles. A more detailed description of the architecture-selection procedure can be found in the Appendix C.

Figure 5 shows the results of the architecture selection for IAD, CWoLa Hunting and CATHODE. Again, selecting based on max SIC or ARGOS results in very similar performance, while a worse performance is observed when selecting based on the BCE. Interestingly, the effect is not the same for the IAD as it is for CATHODE or CWoLa Hunting. For the IAD, BCE results in a performance drop at higher signal injections compared to

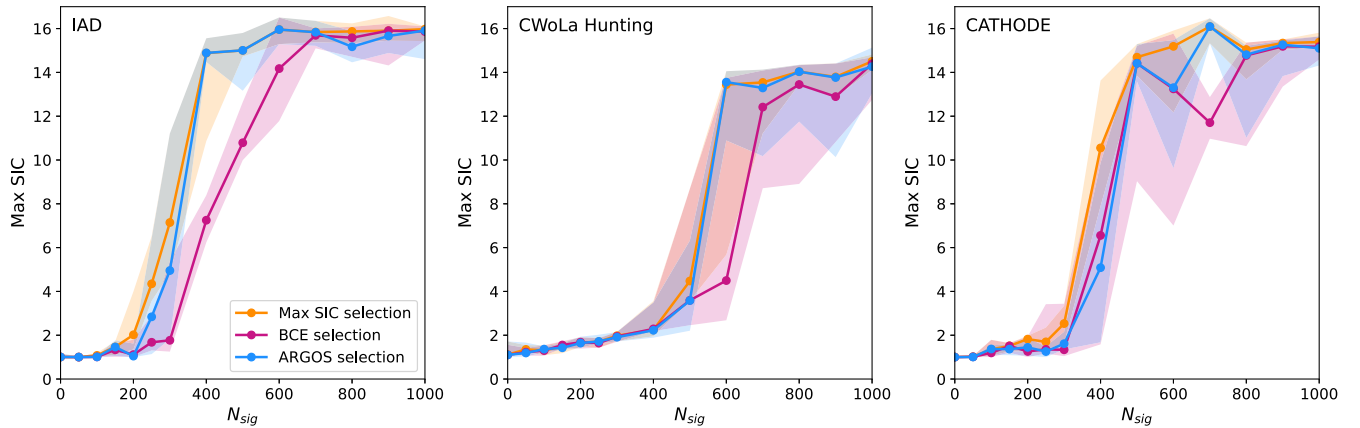


FIG. 5. Anomaly detection performance (max SIC) after architecture selection based on the supervised metric max SIC as a benchmark and the two data-driven metrics max ARGOS and BCE for IAD (left), CWoLa Hunting (middle) and CATHODE (right).

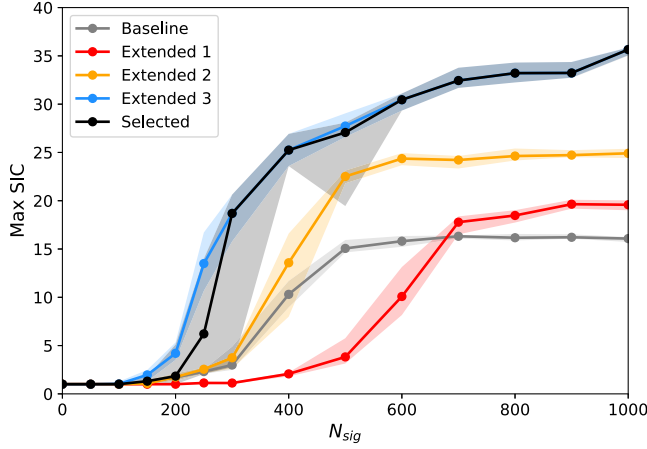


FIG. 6. Anomaly detection performance max SIC for different subjecttiness-based feature sets and for the feature set selection using max ARGOS in black. We use the IAD setup with the HGB classifier. The baseline set is defined in Sec. III. The extended sets are introduced in Ref. [31].

ARGOS and max SIC, whereas for CWoLa Hunting and CATHODE the dominant effect is a larger error band. This large error band originates from the same metric selecting different classifier architectures across the ten runs. For example, in the case of CATHODE, the BCE metric largely chooses the HGB architecture, which leads to the median largely matching the optimal performance; but sometimes (3 times out of 10) it chooses AdaBoost, which leads to the large downward shift of the error band. For CWoLa Hunting the selection of worse models becomes common enough to also lead to a downward shift of the median.

In our case, it was necessary to perform multiple independent architecture selections in order to obtain error bands to help compare the different metrics. In an actual analysis, however, it may be more effective to select architectures based on the mean or median metric obtained from multiple trainings of the same hyperparameter configuration, which could further stabilize the performance and reduce downward fluctuations of the selection.

E. Preliminary studies on feature selection

For anomaly detection to be truly signal model agnostic, feature sets should not be tuned on specific signals. To test the concept of choosing the best feature set based on data-driven metrics, we apply the procedure used for the architecture selection (see Sec. IV D) to the feature sets used in Ref. [31] for the IAD in Fig. 6. The largest feature set, extended set 3, is in this case most sensitive and at high signal injections always chosen by the ARGOS metric. At lower signal injections ARGOS sometimes also selects less performing feature sets as can be seen from the increased error band and the slightly earlier performance dropoff.

This IAD proof-of-concept study shows the potential viability of data-driven feature selection in general.

V. CONCLUSION AND OUTLOOK

In this paper, we have introduced the fully data-driven ARGOS metric for choosing the best anomaly detector. Given an accurate template of background events, we show that the ARGOS metric has a sound theoretical grounding—it is monotonic with the significance improvement characteristic for any unknown signal in the data. This guarantees the effectiveness of the ARGOS metric for comparing arbitrary anomaly detectors.

We demonstrated the power of the ARGOS metric with a number of realistic use cases in the context of weak supervision. For two well-known methods (CWoLa Hunting and CATHODE), applications to epoch selection, hyperparameter optimization, and architecture selection showed that the ARGOS metric is a better proxy for the true anomaly detection performance than the binary cross-entropy loss, which is the default standard currently used in classifier-based weakly supervised anomaly detection.

In this work, we employ the maximum of the ARGOS metric as a function of ϵ_{BT} ; i.e., we optimize the working point ϵ_{BT} and the anomaly detector at the same time. One could just as easily evaluate ARGOS at a specific background template efficiency ϵ_{BT} , if ϵ_{BT} is fixed in a given analysis.

We stress that this is not the only context where the ARGOS metric can be used. For example, it is directly applicable to ANODE [32] and R-ANODE [43], which are density-estimator-based resonant anomaly detection methods. Beyond resonant anomaly detection, different outlier detectors (e.g. autoencoders) could also be compared to one another if a background template is available. In contrast to the binary cross-entropy loss for classifier-based anomaly detection, ARGOS does not assume predictions to represent a class probability. ARGOS is a cut-based metric, which is only affected by the relative ordering of different events and invariant to rescaling or shifting of the entire anomaly score. Therefore, ARGOS can be applied to any anomaly score.

As an illustration of the potential for data-driven feature selection, we carried out a proof-of-concept study using the IAD setup, where ARGOS reliably identifies the best-performing feature sets at high signal injections. However, for a realistic background template, ARGOS will exhibit a bias towards worse background templates. The mitigation of this issue needs to be studied in future work in order for data-driven feature selection to be viable on real data.

In this work, we have only studied how ARGOS can improve the significance for a single, well-studied benchmark signal. Since the default architectures we use in our study have already been developed using the LHCO research and development data signal as a benchmark, room for improvement might actually be particularly

limited for this signal model and ARGOS might perform even better for other signals. To gain even more confidence in the power of ARGOS, investigations of the sensitivity gains for different benchmark models (e.g. the signal models published in Refs. [11,44]) should be studied in future work. The big advantage of data-driven optimization is the potential to become more sensitive to the signal topology which is actually present in the data.

The code for this paper can be found in Ref. [45].

ACKNOWLEDGMENTS

The authors thank Tobias Quadfasel for his contributions in the early stages of this project. M. H. is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 400140256—GRK 2497: The physics of the heaviest particles at the Large Hadron Collider. The research of M. K. and A. M. is supported by the DFG under Grant No. 396021762—TRR 257: Particle Physics Phenomenology after the Higgs Discovery. G. K. and L. M. acknowledge support by the DFG under Germany’s Excellence Strategy No. 390833306—EXC 2121: Quantum Universe. D. S. is supported by DOE Grant No. DOE-SC0010008. Computations were performed with computing resources granted by RWTH Aachen University under Project No. rwth0934.

DATA AVAILABILITY

The data that supports the findings of this article are openly available [43].

APPENDIX A: ARCHITECTURES AND TRAINING

1. Implementation of the CATHODE density estimation

In order to obtain state of the art results, we use a conditional flow matching (CFM) density estimator based on Ref. [39]. In this implementation [46], the conditional vector field needed for CFM is learned using a ResNet-style [47] architecture from `nflows` [48]. The hyperparameters from Ref. [39] are used for both the model and the training. For each signal injection, ten density estimators are trained and ensembled by combining their samples in order to limit fluctuations of the density estimation quality between the different signal injections.

2. Classifier implementations and hyperparameter settings

a. NN

For the NN, we use a standard MLP based on the architecture used in Ref. [16] with an implementation in PyTorch [49] based on Ref. [50]. We choose a fixed number of three hidden layers with 64 nodes each and ReLU activation. We train on the binary cross-entropy loss using

TABLE I. Default NN hyperparameter settings and discrete choices sampled from for hyperparameter optimization.

Hyperparameter	Default	Choices
Learning rate	10^{-3}	[0.01, 0.005, 0.001, 0.0005, 0.0001]
Batch size	128	[64, 128, 256, 512, 1024, 2048, 5096]
Dropout	0	[0, 0.1, 0.2, 0.3, 0.4, 0.5]
Weight decay	0	[0, 10^{-5} , 10^{-4} , 10^{-3} , 10^{-2}]
Momentum	0.9	[0.8, 0.9, 0.95, 0.99]

the ADAM optimizer [51] for 100 epochs. Other hyperparameters such as batch size, dropout, learning rate and other ADAM settings are optimized later on. Their default values can be found in Table I. For the NN, we draw hyperparameters from a set of choices, which are also included in the table.

We choose a 50-50 training and validation split, calculating BCE and max ARGOS on the validation set and max SIC on the test set after every epoch. After training, we ensemble the best ten epochs as decided by the different metrics. This follows the ensembling strategy used in Ref. [16], where the best ten epochs based on the validation loss are used. The impact of choosing different metrics can be seen in Sec. IV B. As a default, we pick the epochs on max ARGOS.

b. HGB

`HistGradientBoostingClassifier` is a gradient boosted decision tree classifier implementation in `scikit-learn` [42], which histograms the classification features before training. It is currently state of the art for weakly supervised anomaly detection [31,52] with high-level features. Following Ref. [31], we perform 50 independent trainings with different 50-50 training-validation splits, which we ensemble by taking the mean over the predictions. Our default hyperparameters are largely equivalent to the default of the `scikit-learn` package and can be found in Table II. For hyperparameters optimization, we sample hyperparameters from the distributions also shown in the table.

TABLE II. Default HGB hyperparameter settings and distributions sampled from for hyperparameter optimization.

Hyperparameter	Default	Sampling	Min	Max
Learning rate	0.1	Log10Uniform	10^{-4}	$10^{-0.1}$
Max. iterations	200	UniformInt	2	200
Max. leaf nodes	31	UniformInt	2	100
Max. depth	None	UniformInt	2	15
Max. bins	255	UniformInt	31	255
L2 regularization	0	Uniform	0	10

TABLE III. Default AdaBoost hyperparameter settings and distributions sampled from for hyperparameter optimization.

Hyperparameter	Default	Sampling	Min	Max
Learning rate	1	Log10Uniform	10^{-4}	$10^{-0.1}$
Number estimators	50	UniformInt	2	100
Max. leaf nodes	31	UniformInt	2	100
Max. depth	None	UniformInt	2	15
Min. samples per leaf	20	UniformInt	1	100

c. AdaBoost

AdaBoostingClassifier is an adaptive boosted decision-tree classifier implementation in `scikit-learn`, which was investigated as an alternative to HGB in [31] but did not achieve the performance or robustness of HGB. We build our implementation in analogy to [31] and perform an ensembling of ten independently trained classifiers using random 50-50 splits of the training dataset. AdaBoost does not make use of a validation set. However, as training is deterministic, the random data split results in variations between different trainings, which can be used to obtain a better performance through ensembling. Our default hyperparameters reflect the default used for AdaBoost in [31] and can be found in Table III. The distributions used to randomly sample hyperparameters for optimization can be found there, as well.

APPENDIX B: HYPERPARAMETER OPTIMIZATION PROCEDURE

As explained in Sec. IV C, the hyperparameter optimization in this work follows a random search strategy with a random sampling of $N_{hp} = 100$ hyperparameter configurations. The distributions these hyperparameters are sampled from can be found in the Appendix A 2. A

classifier is trained with each of these configurations, BCE and max ARGOS are evaluated on the validation set and max SIC on the test set. We use slightly different strategies of stabilizing the metrics and speeding up the computation time for each of our architectures:

- (i) For the NN, each network is trained for 20 epochs, with the epoch number set to 100 after optimization. All three metrics are evaluated after each epoch and the best value for each (minimum for BCE, maximum for max SIC and max ARGOS) is selected.
- (ii) For AdaBoost and HGB, we do not ensemble during the optimization, instead the metrics evaluated after training are averaged over ten runs to obtain more stable results.

Based on the metric values obtained in this way, the best hyperparameter configuration based on each metric is selected. This entire procedure is performed 10 times for each signal injection N_{sig} .

APPENDIX C: DETAILS OF THE ARCHITECTURE SELECTION

In this Appendix, we explain the procedure of the architecture selection in more detail using the example of picking the best architecture for the IAD using ARGOS in Fig. 7. In Fig. 5 of Sec. IV D this is shown as a single line with an error band but in Fig. 7 the entire procedure can be seen.

The first step of our architecture selection is to train our architecture options—NN, HGB and AdaBoost—on 50% of our normal training and validation set. The anomaly detection performance characterized by the max SIC of these runs is shown in the middle panel of Fig. 7. On the other half of the data, we then calculate the ARGOS metric, the result of which is shown in the left panel of Fig. 7. For both of these panels we show the median and 16% to 84% quantiles of ten runs as lines and error bands respectively. These ten runs contain not only ten independent trainings

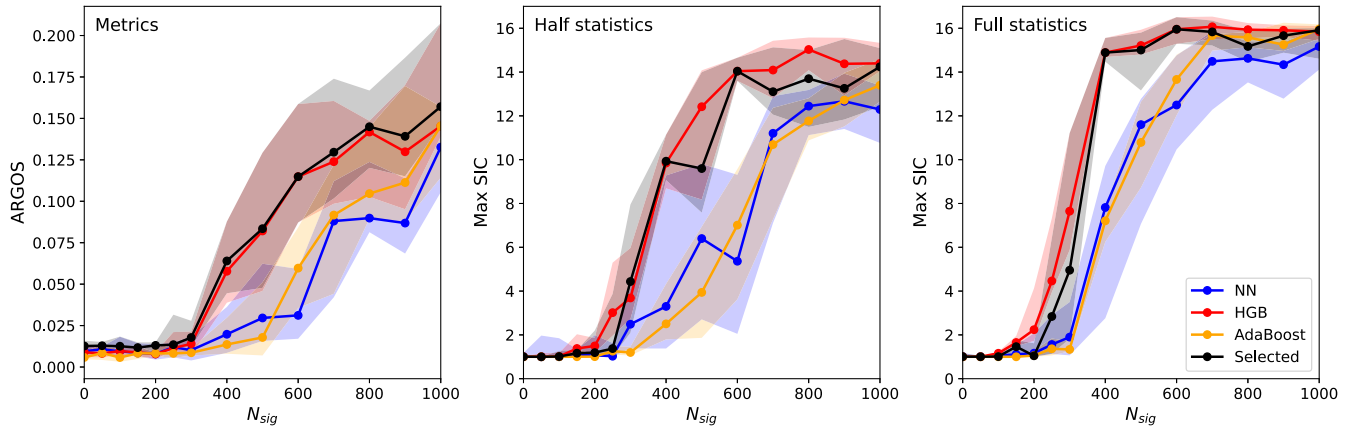


FIG. 7. Architecture selection for IAD using max ARGOS. Left: Max ARGOS calculated on half the data when trained on the other half. Middle: Max SIC of the trainings of half the data. Right: Max SIC of full statistics training. In each case the selected runs are shown in black. Note that the x-axis shows the full statistics signal injection values even for the half statistics runs.

but also ten different hyperparameter configurations per model as we use the results from Sec. IV C. For each of these sets of ten runs, we then pick the best-performing architecture between NN, HGB and AdaBoost based on the obtained max ARGOS values and plot the resulting set of ten again as median and quantile-error band. We perform this 50-50 split of the dataset in order to have the metric values used for architecture selection be calculated on a held-out evaluation set. As all classifiers use the validation set in one form or another, this is not the case when directly selecting on the validation set.

For this set of selected architectures, we then run another training using the full statistics, which leads to the black line and error band in the right panel of Fig. 7. For comparison, we also show the full statistics performance of the different architectures. One can see that the final performance of the selected runs largely matches that of the best performing architecture, namely the HGB, with only some downward fluctuations.

APPENDIX D: FALSE DISCOVERIES AND BACKGROUND SCULPTING

When optimizing on data, it is important to understand the effect that this has on the statistics of an analysis. In our case, we seek to understand whether an increased rate of

false discoveries may follow, which could stem from one of two issues:

- (1) Selecting the best classifier on data could lead to “p-hacking”: The selected classifier may prioritize purely statistical fluctuations in the training data, leading to a look elsewhere effect [53] and mis-calibrated significances.
- (2) The selected classifier may prioritize regions of the data with the most systematic background mismodeling, which could lead to false excesses in both fit-based bump hunts (as performed in Refs. [5,6,9,54]) and in analyses with direct background estimation [19].

The first issue can be completely eliminated by evaluating on an independent test set. This will be the case for the remainder of this appendix.

Regarding the second issue, the classifier most able to distinguish between data and background is fundamentally the better anomaly detector. If this leads to false positives due to background mismodeling, this is not a fault of the anomaly detector *per se*.

Nevertheless, we will investigate this issue in more detail, to see what the effect of improved model selection has on it.

Shown in Fig. 8 is the χ^2 histogram difference for the background m_{JJ} distribution before and after a cut on the

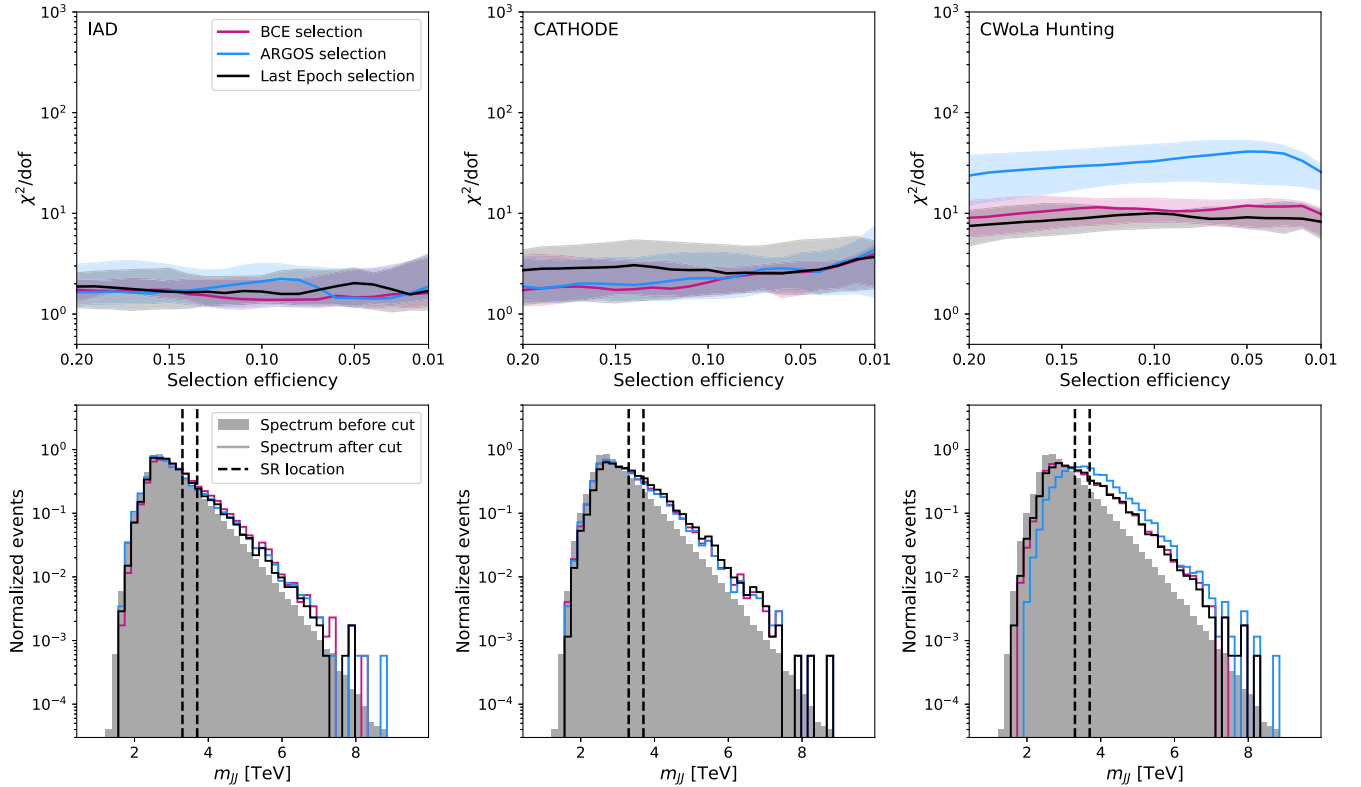


FIG. 8. Background sculpting for IAD (left), CATHODE (middle) and CWoLa Hunting (right) when selecting epochs based on validation BCE loss, max ARGOS or selecting the last epoch characterized by the χ^2 metric from [36] (top) and example spectra obtained when selecting 1% most anomalous events (bottom).

anomaly score, as a function of background selection efficiency. This metric was previously introduced in Ref. [36] and is a way to quantify the effect of background sculpting due to a cut on the anomaly score. We compare the sculpting for an anomaly detector selected using the BCE loss (as used in Ref. [5]), the ARGOS metric and using the last epoch. We see that for the IAD and CATHODE the sculpting does not change significantly based on the different model selection methods.

For CWoLa Hunting, the amount of sculpting is already more severe compared with IAD or CATHODE when model selecting with BCE loss or using the last epoch. This is further exacerbated when the selection is performed with ARGOS. This is entirely expected as the background in the SR is not perfectly modeled by the SB data, and with the ARGOS metric we are picking this up more effectively.

Background sculpting is a general problem for anomaly detection analyses and goes beyond weakly supervised methods. As such, any fit-based bump hunt using anomaly detectors requires careful treatment of this potential sculpting, as detailed e.g. in Refs. [5,11]. Using ARGOS, such a careful treatment is also necessary but we do not see any indication that it is worse than other model selection methods.

If significances are obtained through direct background estimation the magnitude of the systematic shift between BT and SR needs to be estimated in any case for the given classifier setup [19]. Using a better classifier based on ARGOS optimization, the systematic shift might increase. However, estimating the systematic shift properly, an increased rate of false discoveries will be avoided.

-
- [1] G. Kasieczka *et al.*, *Rep. Prog. Phys.* **84**, 124201 (2021).
 - [2] T. Aarrestad *et al.*, *SciPost Phys.* **12**, 043 (2021).
 - [3] V. Belis, P. Odagiu, and T. K. Aarrestad, *Rev. Phys.* **12**, 100091 (2023).
 - [4] HEP ML Community, A living review of machine learning for particle physics, <https://iml-wg.github.io/HEPML-LivingReview/>.
 - [5] CMS Collaboration, *Rep. Prog. Phys.* **88**, 067802 (2024).
 - [6] ATLAS Collaboration, *Phys. Rev. Lett.* **125**, 131801 (2020).
 - [7] ATLAS Collaboration, *Phys. Rev. D* **108**, 052009 (2023).
 - [8] ATLAS Collaboration, *Phys. Rev. Lett.* **132**, 081801 (2024).
 - [9] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. D* **112**, 072009 (2025).
 - [10] ATLAS Collaboration, *Phys. Rev. D* **112**, 012021 (2025).
 - [11] CMS Collaboration, Machine-learning techniques for model-independent searches in dijet final states, Technical Report, CERN, CMS Physics Analysis Summary CMS-PAS-MLG-23-002, 2025.
 - [12] CMS Collaboration, *Eur. Phys. J. C* **86**, 137 (2026).
 - [13] J. H. Collins, K. Howe, and B. Nachman, *Phys. Rev. Lett.* **121**, 241803 (2018).
 - [14] J. H. Collins, K. Howe, and B. Nachman, *Phys. Rev. D* **99**, 014038 (2019).
 - [15] O. Amram and C. M. Suarez, *J. High Energy Phys.* **01** (2021) 153.
 - [16] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, *Phys. Rev. D* **106**, 055006 (2021).
 - [17] G. Grosso and M. Letizia, *Eur. Phys. J. C* **85**, 4 (2024).
 - [18] R. T. D'Agnolo and A. Wulzer, *Phys. Rev. D* **99**, 015014 (2019).
 - [19] R. Das, T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, and D. Shih, *Phys. Rev. D* **111**, 094041 (2025).
 - [20] G. Grosso, D. Sengupta, T. Golling, and P. Harris, *Eur. Phys. J. C* **85**, 1074 (2025).
 - [21] E. M. Metodiev, B. Nachman, and J. Thaler, *J. High Energy Phys.* **10** (2017) 174.
 - [22] G. Kasieczka, B. Nachman, and D. Shih, R&D dataset for lhc olympics 2020 anomaly detection challenge, <https://zenodo.org/record/6466204> (2019).
 - [23] D. Shih, Additional QCD background events for LHCO2020 R&D (signal region only), <https://zenodo.org/record/5759086> (2021).
 - [24] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *J. High Energy Phys.* **05** (2006) 026.
 - [25] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *Comput. Phys. Commun.* **178**, 852 (2008).
 - [26] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), *J. High Energy Phys.* **02** (2014) 057.
 - [27] M. Cacciari, G. P. Salam, and G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012).
 - [28] M. Cacciari and G. P. Salam, *Phys. Lett. B* **641**, 57 (2006).
 - [29] J. Thaler and K. Van Tilburg, *J. High Energy Phys.* **03** (2011) 015.
 - [30] J. Thaler and K. Van Tilburg, *J. High Energy Phys.* **02** (2012) 093.
 - [31] T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, *Phys. Rev. D* **109**, 034033 (2023).
 - [32] B. Nachman and D. Shih, *Phys. Rev. D* **101**, 075042 (2020).
 - [33] A. Andreassen, B. Nachman, and D. Shih, *Phys. Rev. D* **101**, 095004 (2020).
 - [34] K. Benkendorfer, L. L. Pottier, and B. Nachman, *Phys. Rev. D* **104**, 035003 (2020).
 - [35] J. A. Raine, S. Klein, D. Sengupta, and T. Golling, *Front. Big Data* **6**, 899345 (2022).
 - [36] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, *Phys. Rev. D* **107**, 114012 (2022).

- [37] T. Golling, S. Klein, R. Mastandrea, and B. Nachman, *Phys. Rev. D* **107**, 096025 (2023).
- [38] T. Golling, G. Kasieczka, C. Krause, R. Mastandrea, B. Nachman, J. A. Raine, D. Sengupta, D. Shih, and M. Sommerhalder, *Eur. Phys. J. C* **84**, 241 (2023).
- [39] R. Das and D. Shih, *Phys. Rev. D* **112**, 074040 (2025).
- [40] M. Leigh, D. Sengupta, B. Nachman, and T. Golling, *J. High Energy Phys.* **12** (2025) 105.
- [41] I. Oleksiyuk, S. Voloshynovskiy, and T. Golling, *J. High Energy Phys.* **07** (2025) 177.
- [42] F. Pedregosa *et al.*, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [43] R. Das, G. Kasieczka, and D. Shih, [arXiv:2312.11629](https://arxiv.org/abs/2312.11629).
- [44] R. Das, M. Hein, G. Kasieczka, M. Krämer, L. Lang, R. Mastandrea, L. Moureaux, A. Mück, and D. Shih, [arXiv:2604.20965](https://arxiv.org/abs/2604.20965).
- [45] M. Hein, ARGOS metric for anomaly detection, <https://github.com/mariehein/ARGOS-metric-for-anomaly-detection> (2025).
- [46] R. Das, Density Estimation for “Accurate and robust methods for direct background estimation in resonant anomaly detection”, https://github.com/rd804/cut_and_count_FM (2024).
- [47] K. He, X. Zhang, S. Ren, and J. Sun, [arXiv:1512.03385](https://arxiv.org/abs/1512.03385).
- [48] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, nflows: Normalizing flows in PyTorch (2020), [10.5281/zenodo.4296287](https://zenodo.org/record/4296287).
- [49] J. Ansel *et al.*, in *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)* (ACM, New York, NY, USA, 2024), [10.1145/3620665.3640366](https://doi.org/10.1145/3620665.3640366).
- [50] O. Amram, J. Birk, and M. Sommerhalder, sk_cathode, https://github.com/uhh-pd-ml/sk_cathode (2024).
- [51] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [52] M. Freytsis, M. Perelstein, and Y. C. San, *J. High Energy Phys.* **02** (2023) 220.
- [53] M. Hein, B. Nachman, and D. Shih, [arXiv:2512.13787](https://arxiv.org/abs/2512.13787).
- [54] R. Gambhir, R. Mastandrea, B. Nachman, and J. Thaler, *Phys. Rev. Lett.* **135**, 021902 (2025).