Preserving new physics while simultaneously unfolding all observables

Patrick Komiske,[1,2,*] W. Patrick McCormack,[3,4,†] and Benjamin Nachman[4,5,‡]

[1]*Center for Theoretical Physics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA*
[2]*The NSF AI Institute for Artificial Intelligence and Fundamental Interactions,
Cambridge, Massachusetts 02139, USA*
[3]*Department of Physics, University of California, Berkeley, California 94720, USA*
[4]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[5]*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*

(Received 9 June 2021; accepted 30 September 2021; published 27 October 2021)

Direct searches for new particles at colliders have traditionally been factorized into model proposals by theorists and model testing by experimentalists. With the recent advent of machine learning methods that allow for the simultaneous unfolding of all observables in a given phase space region, there is a new opportunity to blur these traditional boundaries by performing searches on unfolded data. This could facilitate a research program where data are explored in their natural high dimensionality with as little model bias as possible. We study how the information about physics beyond the Standard Model is preserved by full phase space unfolding using an important physics target at the Large Hadron Collider (LHC); exotic Higgs boson decays involving hadronic final states. We find that if the signal cross section is high enough, information about the new physics is visible in the unfolded data. We will show that in some cases, quantifiably all of the high-level information about the new physics is encoded in the unfolded data. Finally, we show that there are still many cases when the unfolding does not work fully or precisely, such as when the signal cross section is small. This study will serve as an important benchmark for enhancing unfolding methods for the LHC and beyond.

DOI: 10.1103/PhysRevD.104.076027

I. INTRODUCTION

Analyses at the Large Hadron Collider (LHC) are generally classified as *measurements* or *searches* if their goal is to search for indirect or direct signs of physics beyond the Standard Model (SM), respectively. An important reason for this distinction is that measurements assume that deviations to the SM are small. This is required so that the removal of detector distortions (unfolding) can be based on SM simulations. Traditional unfolding methods [1–7] are based on low-dimensional and binned observables. The detector response may depend on additional unmeasured features and may vary strongly within a given bin. If these properties are significantly different for new particles, then an unfolding derived with SM simulations is likely to be inaccurate.

This feature of current unfolding methods has been studied in [8] and limits the applicability of recasting tools such as CONTUR [9]. Recasting is the task of taking a published result and reinterpreting it in the context of a signal model that was not used in the original analysis. A variety of complementary tools have been developed to fold model predictions with a detector response, including MadAnalysis [10–14], RECAST [15–17], Checkmate [18,19], SModelS [20,21], FASTLIM [22], and XQCAT [23]. In addition to limitations from recasting approximations, these approaches are limited by the minimal (binned) search results that are usually highly optimized for particular signal models.

One possibility is to perform model-agnostic approaches at detector level using one of the growing number of anomaly detection methods [24–65] (for reviews, see Refs. [60,66]). These techniques can achieve broad and deep sensitivity by learning directly from data. However, methods that do not rely on any signal information (unsupervised) are not particularly sensitive [47,65] and methods that use noisy or partial signal information (weakly and semisupervised, respectively) are not recastable after the search is performed [44].

A new solution that has emerged is to perform an unbinned unfolding using all of the available information.

[*] pkomiske@mit.edu
[†] wpmccormack@lbl.gov
[‡] bpnachman@lbl.gov

If the high frequency and high dimensional aspects of the detector response are part of the unfolding procedure, then differences between signal and background will not be a source of bias. Unbinned and high-dimensional unfolding are now possible with advances in machine learning [67–69] (for other machine learning and unbinned proposals, see Refs. [70–74]). Of these, only the OmniFold [67] can currently process the full phase space that includes all observable particles and their properties. Unlike proposals based on generative models, OmniFold is built on neural network (NN) classifiers that are used to iteratively reweight simulations to match the data. Classifiers designed to process variable-length, unordered sets of particles allow this technique to access the full phase space [75,76]. While OmniFold has yet to be applied to collider data in the full phase space, it has recently been deployed in a low-dimensional case with the H1 experiment [77].

In this paper, we investigate the ability of OmniFold to preserve information about new particles present in the data. In particular, we will study the direct production of new particles which have spectra that are not similar to the SM background. Our benchmark example will be the exotic decay of a Higgs bosonlike particle decaying into a Z boson and a light color singlet that decays into hadrons. The dominant background to this process is the SM production of Z bosons and jets. We will see to what extent the information about new particles are preserved in the unfolding. Recently, the authors of Ref. [69] showed that generative model approaches can preserve new physics with a relatively large cross section. Our first example will be motivated by this example and then we will explore how the sensitivity depends on variations in the signal-model parameters and the unfolding setup.

This paper is organized as follows. Section II briefly reviews full phase space unfolding and introduces the benefits and challenges of the existing approach in the context of physics beyond the SM. The simulation samples and machine learning setup are introduced in Sec. III. Section IV explores a case where a model-independent new physics search technique, such as bump hunting, could be applied in unfolded data. Section V then studies an example of exotic Higgs boson decays, where simple bump-hunting would be less fruitful. Implications of model-dependent search program in unfolded data are explored in this section as well. The paper ends with conclusions and outlook in Sec. VI.

II. REVIEW OF OmniFold UNFOLDING

The OmniFold method is represented visually in Fig. 1. There are two inputs; natural data from experiment and synthetic data from simulation. The goal is to remove the detector distortions from the observations ("Data") to infer the underlying particle-level distribution ("Truth"). Synthetic particle-level events ("Generation") provide the initial guess for the Truth and we have an event-by-event match between the Generation and detector-level synthetic data ("Simulation"). As in all unfolding algorithms, we assume that the detector response is well modeled. The $Z+$jets final state is chosen because it can be identified and reconstructed with high efficiency and high purity. Minor backgrounds, and acceptance effects are not considered in this analysis, although they can be accounted for in the OmniFold framework—see Ref. [78] for details.

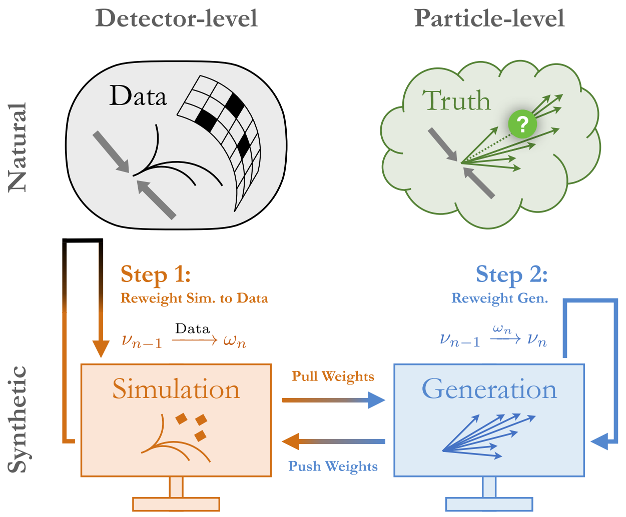

FIG. 1. Visual representation of the OmniFold method. This method relies on experimental "Data", which was caused by some underlying "Truth" distribution, and synthetic datasets at both particle level ("Generation") and detector level ("Simulation"). The Simulation is reweighted to match the Data using a classifer, and these weights, ω_i are applied to Generation. The unweighted Generation is reweighted to match the Generation with ω_i applied, giving particle-level weights, ν_i. This process is repeated iteratively, as the ν_i are applied to the Simulation, and this reweighted Simulation is once again reweighted to Data. Based on the corresponding figure from Ref. [67].

The first step in the OmniFold method is to train a classifier to distinguish the Data from the Simulation. An optimally-trained classifier that minimizes one of the standard loss functions (cross entropy or mean-squared error) will predict the probability that an event x_{det} is drawn from Data instead of Simulation: $\Pr(\mathrm{Data}|x_{\mathrm{det}})$. By applying the weight

$$\omega_1 = \frac{\Pr(\mathrm{Data}|x_{\mathrm{det}})}{1 - \Pr(\mathrm{Data}|x_{\mathrm{det}})} \propto \frac{p(x_{\mathrm{det}}|\mathrm{Data})}{p(x_{\mathrm{det}}|\mathrm{Simulation})}, \qquad (1)$$

to each simulated event, the Simulation will closely resemble the Data. Note that in Eq. (1), "Data" and "Simulation" denote class labels while x_{det} represents the observable features that are the input to the training.

The detector-level weights, ω_1 are then applied to the corresponding particle-level events, resulting in a reweighted distribution of particle-level events. A second classification step is required because these *pulled back* weights are not a proper function of the particle-level phase space; the same phase-space point can be mapped to two different detector-level points with different weights under the stochastic mapping of the detector. The second classifier distinguishes the nominal Generation from the one using the weights from the first step. As with the first step, an optimally-trained classifier will learn

$$\nu_1 = \frac{\Pr(\text{reweighted Gen}|x_{\text{part}})}{1 - \Pr(\text{reweighted Gen}|x_{\text{part}})} \tag{2}$$

$$\propto \frac{p(x_{\text{part}}|\text{reweighted Gen})}{p(x_{\text{part}}|\text{Generation})}. \tag{3}$$

The matching between Generation and Simulation can be used to push the weights to detector level and the entire process can be repeated N times, with each complete pass through being called an "iteration". The final result is the Generation dataset with a set of per-event weights, v_N. For a given unfolding problem, the optimal value of N and the corresponding optimal weights v_N can be chosen by comparing the detector-level distributions after several iterations. The iteration with the best agreement can be selected. Agreement tends to plateau after fewer than ten iterations, but it is possible to train for additional iterations if the agreement is continuing to improve. There is no true notion of "overtraining" here, as the OmniFold weights are not meant to be applied in any context beyond the data and simulation present at the start of the procedure. Within an individual iteration, part of the Data and Simulation sets are held out as validation sets to check for overtraining of the network designated to distinguish between them and to check for overtraining in second classification step. Statistical uncertainty on the weights—and thereby the unfolded distribution—can be determined with bootstrapping techniques.

We will parametrize the classifiers as neural networks. When x is the full phase space, i.e., a complete list of reconstructed or true particles with their observable properties, we need a neural-network architecture that can process variable length, unordered sets. For this purpose, we use Particle Flow Networks (PFN) [75,76]. A reduced alternative approach will use a fixed number of high-level observables, which will use a standard fully connected neural network. To distinguish between these two cases, we will call the full phase-space version OmniFold and the reduced version MultiFold.

A. OmniFold in the presence of new physics

The main benefit of the OmniFold method comes from its use of low-level observables and the freedom from fixed bins. By using information about each reconstructed particle in an event or jet, the full phase space is exploited. Therefore, as long as the interaction of individual particles with the detector is modeled well, beyond the Standard Model (BSM) physics should not negatively affect the ability to unfold. If there is BSM physics present in the data, then the most BSM-like events in the Simulation will be upweighted as appropriate.[1] In traditional unfolding schemes that use regularized-matrix inversion of binned histograms, the presence of BSM physics could affect the detector response matrix in ways that are not accounted for in a SM-only simulation. OmniFold is less affected by this, and is not affected by the possibility of suboptimal binning choices for BSM sensitivity.

However, there is a key assumption in the OmniFold method; the initial Simulation and Data must have overlapping support. For example, if a heavy resonance existed at a mass well beyond the last data point in the simulation, then it would not be possible to upweight events to match the resonance even in a binned histogram case. In the OmniFold case, the initial simulation should span the data in *all* dimensions. Empirically, the simulations often used at the LHC share the same support as the data. In practice, one needs the ratios of probability densities to not be too far from unity because even if the support is overlapping, a very small likelihood ratio will have a large weight and thus poor statistical uncertainty.

III. SIMULATION AND MACHINE LEARNING SETUP

Many models of new physics predict new heavy particles that decay to SM particles. If the invariant mass of the decay particles is computed, a resonant enhancement in the mass spectrum should occur, centered on the new particle's mass. As an initial exploration of the efficacy of OmniFold in the presence of BSM physics, we consider the case of a new heavy scalar particle. Our study is based on proton-proton collisions generated at $\sqrt{s} = 14$ TeV with Tune 26 [79] of PYTHIA 8.243 [80–82]. Signal events are generated as $h \to Za$, $a \to gg$, where m_h has been set to 125 or 250 GeV, and various a masses have been used. This final state (with low m_a and $m_h = 125$ GeV) was recently studied by the ATLAS Collaboration in Ref. [83]. Detector effects are emulated with DELPHES 3.4.2 [84], using the CMS detector card, which uses particle-flow reconstruction. For this study, the Data, Truth, Simulation, and Generation sets consist of 200,000 events. Jets with radius parameter $R = 0.4$ are clustered using either all-particle flow objects (detector-level) or stable non-neutrino truth particles (particle level) with the anti-k_T algorithm [85] implemented in FastJet 3.3.2 [86,87]. We consider leptonic decays of the Z boson, which can be precisely reconstructed. The target final state is then Z-boson production in association with one jet that has nontrivial substructure. Events are selected if

[1] Events may be down weighted in the case of interference effects.

there is at least one truth-level and one detector-level particle within the jet.[2]

For OMNIFOLD, we use all of the particles in the leading jet. Each particle is specified with its p_T, rapidity, y, azimuthal angle, ϕ, and particle identification number. Furthermore, the invariant mass of the $Z +$ jet system, the jet mass, and the jet multiplicity are included as global features. These data are processed using PFNs implemented in the EnergyFlow Python package [88]. The PFN architecture is composed of an encoder followed by a fully connected network. The encoder has two hidden layers of 200 nodes each and outputs a 256-dimensional latent vector. These vectors are summed over all particles and then the subsequent fully connected network is composed of three layers of 100 nodes each.

For MULTIFOLD, we use ten features from each event, based on the Z boson properties and the leading jet. These features include the invariant mass of the $Z +$ jet system, the jet mass, the jet constituent multiplicity, the jet p_T, the Z p_T, the jet Les Houches Angularity [89,90], the jet width [90–93], the groomed jet mass with Soft Drop parameters $z_{\rm cut} = 0.1$ and $\beta = 0$ [94], the groomed jet momentum fraction (same Soft Drop parameters), and the jet image activity, which is the minimum number of pixels in a jet image that contain 95% of the total p_T [95]. The MULTIFOLD neural networks are composed of three hidden layers of 100 nodes each.

For each iteration of OMNIFOLD and MULTIFOLD, the neural network was trained with 120 and 20 epochs, respectively, and included an early stopping condition based on validation loss improvement. The validation sample was constructed from a random 20% of the events. The models are randomly initialized in the first iteration and subsequently warm-started using the model from the previous iteration. All neural networks are implemented using KERAS [96] with the TensorFlow backend [97] and optimized with ADAM [98].

IV. HEAVY SCALAR DECAY STUDY

First, we study the case of $m_h = 250$ GeV where 10% of the data is BSM physics. This composition and signal model relative to the $Z +$ jets background are qualitatively similar to the example presented in Ref. [69], which used generative models.

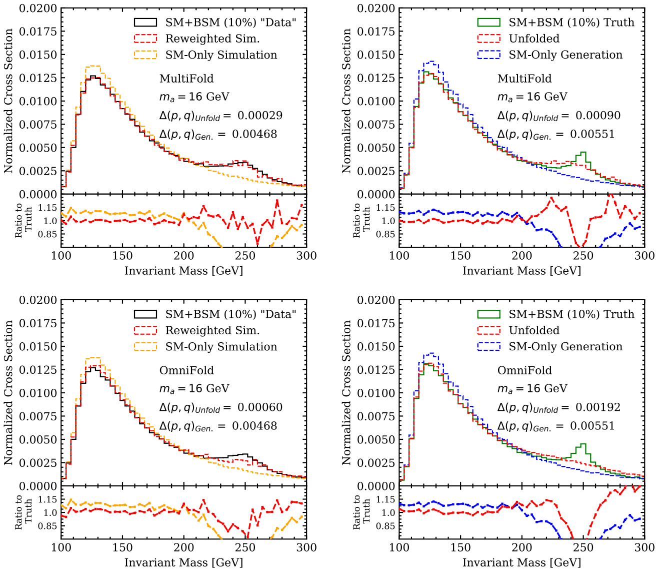

FIG. 2. Distributions of the $Z +$ jet invariant mass spectrum for both MULTIFOLD (top row) and OMNIFOLD (bottom row). Distributions are shown for both detector-level (Data and Simulation) and truth-level (Truth and Generation) values. The Truth and Data distributions are a combination of 180,000 PYTHIA 8$Z +$ jet events and 20,000 $h \to Za$, $a \to gg$, where $m_h = 250$ GeV and $m_a = 16$ GeV. The Generation and Simulation are 200,000 SM-only events. The weights are taken after five iterations of the respective unfolding procedure. The triangular discriminator [99–101] $\Delta(p,q) = \int d\lambda \frac{(p(\lambda)-q(\lambda))^2}{p(\lambda)+q(\lambda)}$ is used to quantify the difference between distributions.

Figure 2 shows the detector-level and truth-level distributions of $Z +$ jet invariant mass before and after unfolding. At detector level, which corresponds to the first step in an iteration of the OMNIFOLD method, the distributions exhibit good agreement after unfolding; the height and width of the mass peak are reproduced accurately, especially in the MULTIFOLD case. At truth level, the peaks are not reproduced as sharply. In the MULTIFOLD case, the height and width of the peak are similar to that seen at detector level, and in the OMNIFOLD case, the peak is considerably broader.

Part of the broadening is an inherent challenge with nontrivial resolutions and limited statistics. The truth-level peak quality can be recovered by modifying the Generation. In particular, the spectrum of synthetic events at particle level before beginning the unfolding procedure [$p(x_{\rm part}|{\rm Generation})$] can be viewed as a “prior” in the sense of an initial guess on the unfolded distribution. We are free to choose whatever Generation we want as OMNIFOLD is a maximum-likelihood estimator that is formally prior independent. However, the closer the prior is to the data, the more accurate the unfolding will be with finite statistics. To test this idea, the same Truth sample was used as above, with 180,000 SM events and 20,000 $h \to Za$, $a \to gg$, where $m_h = 250$ GeV and $m_a = 16$ GeV. However, now the Generation was taken to include 200,000 SM events, 10,000 $h \to Za$, $a \to gg$ events with $m_h = 125$ GeV for each of $m_a = 0.5$, 1, 2, 4, 8, and 16 GeV, and 10,000 $h \to Za$, $a \to gg$ events with $m_h = 250$ GeV for each of the same m_a values, for a total of 320,000 events. The truth-level results of unfolding with the same OMNIFOLD setup discussed above are

[2] We ignore acceptance effects from the jet selection, which can be made arbitrarily small in this case by using only the Z to choose events.

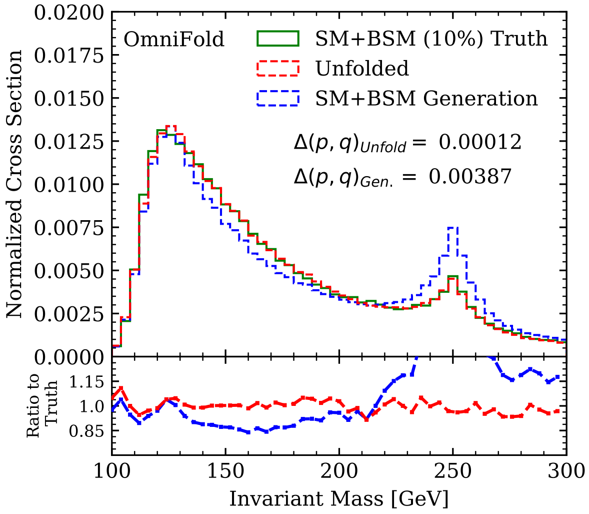

FIG. 3. Truth-level distribution of $Z +$ jet invariant mass for the case that OmniFold is performed with BSM events in Generation. The BSM event included in the Generation were drawn from events with $m_h = 125$ GeV and $m_h = 250$ GeV. The weights are taken after five OmniFold iterations.

shown in Fig. 3. Here, both the height and weight of the truth-level peak are reproduced well by the reweighted sample. The fact that this works well, when an application of OmniFold with SM-only events did not, shows the importance of sufficiently covering the relevant regions of phase space.

Adding BSM physics to the Generation sample begs the question of what the invariant mass distribution would look like after unfolding if the Data does not itself contain BSM physics. To test this, the same Generation sample with 320,000 events from the preceding paragraph is used again, but Data and Truth are taken to be 200,000 SM-only events. The OmniFold method is applied in the same way as above, and the resulting $Z +$ jet invariant mass distributions are shown in Fig. 4. The bump in the unfolded distribution has been eliminated despite the fact that almost 40% of events in the Generation sample were drawn from BSM samples. To optimize the unfolding, care must be taken when choosing the BSM models to include in the Generation sample as well as when choosing the number of BSM events to include—manipulating these parameters effectively corresponds to choosing different priors for what is expected in the Data. It may also be possible to achieve a similar performance without inserting such a localized signal. For this idea to be useful, it will be important to establish a procedure that minimizes the model dependence for picking non-SM contributions.

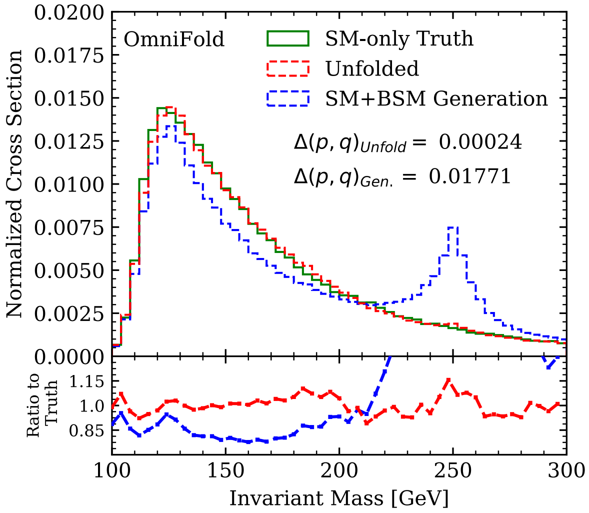

FIG. 4. Truth-level distribution of $Z +$ jet invariant mass for the case that OmniFold is performed with BSM events in Generation despite a lack of BSM events in Truth. The BSM event included in the Generation were drawn from events with $m_h = 125$ GeV and $m_h = 250$ GeV. The weights are taken after 5 OmniFold iterations.

This section has demonstrated that MultiFold, and to some extent OmniFold, can qualitatively preserve a relatively large[3] and prominent resonant signature from the data. A sideband technique could then be used to perform a search with these data. In the next section, we will explore the ability of OmniFold to precisely preserve the phase space so that a multivariate classifier could be used for a search with the unfolded data.

V. EXOTIC HIGGS DECAY

Given the visual prominence of the signal in the data plots of Fig. 2, it is likely that a discovery would be made when performing inclusive cross-section measurements of $Z +$ jets, such as routinely performed by the ATLAS and CMS Collaborations [102,103]. However, not all new physics processes can be searched for with such a simple approach. To explore such a scenario, we consider $m_h = 125$ GeV. In this case, the signal bump is near the background peak and so additional features beyond just the $Z +$ jet invariant mass are required. We explore the possibility of performing a model-dependent search that uses dedicated BSM vs SM discriminating variables. If OmniFold effectively unfolds the full phase space, it should be possible to use any combination of variables in unfolded data.

A. Unfolding with MultiFold

First, we can consider the case of MultiFold with two working points for BSM physics:

(a) 0.1% of Data and Truth events are BSM physics, with $m_a = 16$ GeV.

(b) 10% of Data and Truth events are BSM physics, with $m_a = 16$ GeV.

[3] In fact, the amount of signal is so large, that it would result in a significant detection from the cross section alone, which is well known for $Z +$ jets. We revisit this in Sec. VI.

For each working point, the Data, Truth, Simulation, and Generation sets will be once again be 200,000 events. In each case, the Simulation and Generation sets are drawn from the SM-only PYTHIA 8 sample. The SM events in Data and Truth are also drawn from the PYTHIA 8 sample, but no SM event can be used in both Data and Simulation.

The distributions of $Z+$jet invariant mass, jet mass, and jet multiplicity for Truth, Generation, and unfolded Generation are shown in Fig. 5. The impact of a 0.1% signal is difficult to detect in these one-dimensional histograms and MULTIFOLD has correspondingly left the phase space mostly untouched. For the 10% signal, MULTIFOLD improves the agreement of every variable's distribution with that of the truth-level generation, based on the triangular-discriminator metric. Similar trends hold for alternative m_a values as well (not shown).

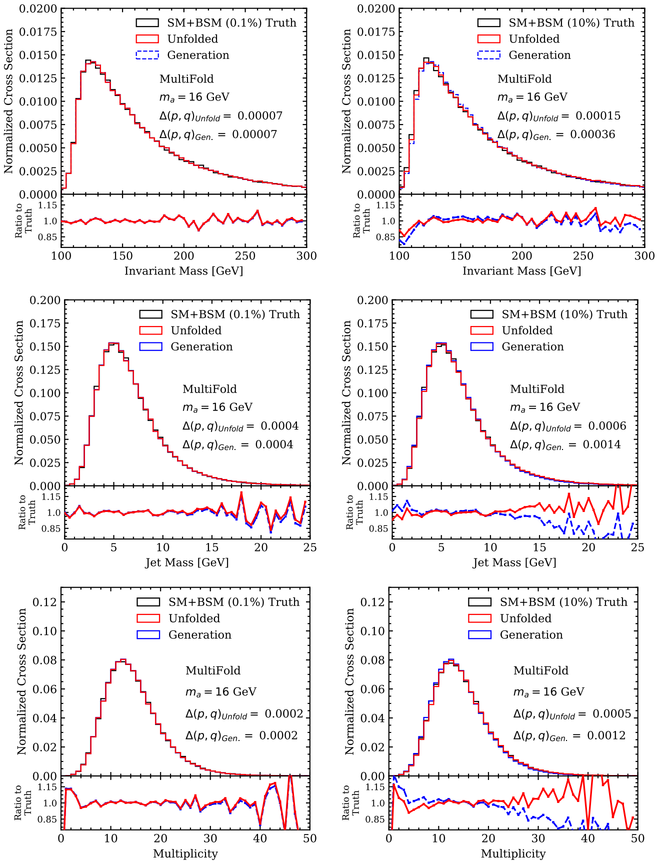

FIG. 5. Truth, Generation, and unfolded Generation distributions for the MULTIFOLD case, where BSM $h \to Za$, $a \to gg$ events have been included in the Truth, but not the Generation. SM events in these samples come from PYTHIA 8$Z+$jets simulation. In the left column, 200 out of 200,000 Truth events come from the BSM sample, and in the right column, 20,000 out of 200,000 Truth events are BSM physics. Distributions are given for the invariant mass of the $Z+$jet, the jet mass, and the jet multiplicity. The ratios of the Generation distributions are given to Truth for each plot. The weights are taken after five iterations of MULTIFOLD.

The ratio panels in Fig. 5 show that the distributions after MULTIFOLD are flatter with respect to truth than the distributions prior to unfolding. To investigate the degree to which BSM physics is encoded in the unfolded data, we emulate a model-dependent search by training a fully supervised classifier to distinguish $Z+$jets events from the $m_a = 16$ GeV signal. A sample of 90,000 SM and 90,000 BSM events was used for training, with 30%

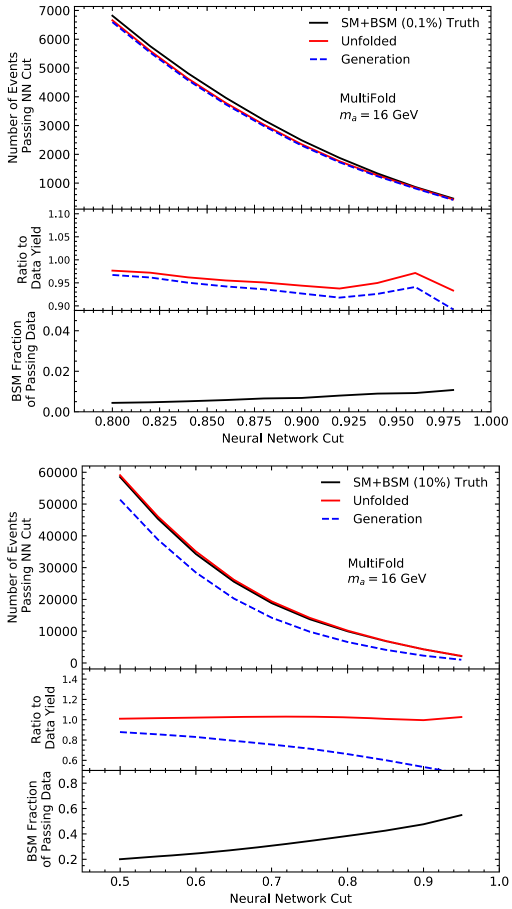

FIG. 6. The number of Truth, Generation, and unfolded Generation events passing a cut on the neural network score, as a function of the cut value, in the case that 0.1% of the data comes from BSM physics (top) and 10% of the data comes from BSM physics (bottom). The unfolding was performed with MULTIFOLD. The neural network was specifically trained to distinguish SM from BSM physics. The middle segment of both plots shows the ratio of the pre and postunfolding Generation yields to the Truth yield. The yield from the preunfolding Generation is not expected to agree well with the Truth, as there are no BSM events. However, if the most BSM-like events get upweighted by the unfolding, then the weighted sum of passing events increases. In a fully accurate unfolding the unfolded yield would match the yield from Truth. The bottom segment shows the fraction of Truth events that are BSM events at truth level; as the cut value approaches 1, the events passing the selection must be more BSM-like, which is why the BSM-purity increases as a function of cut value.

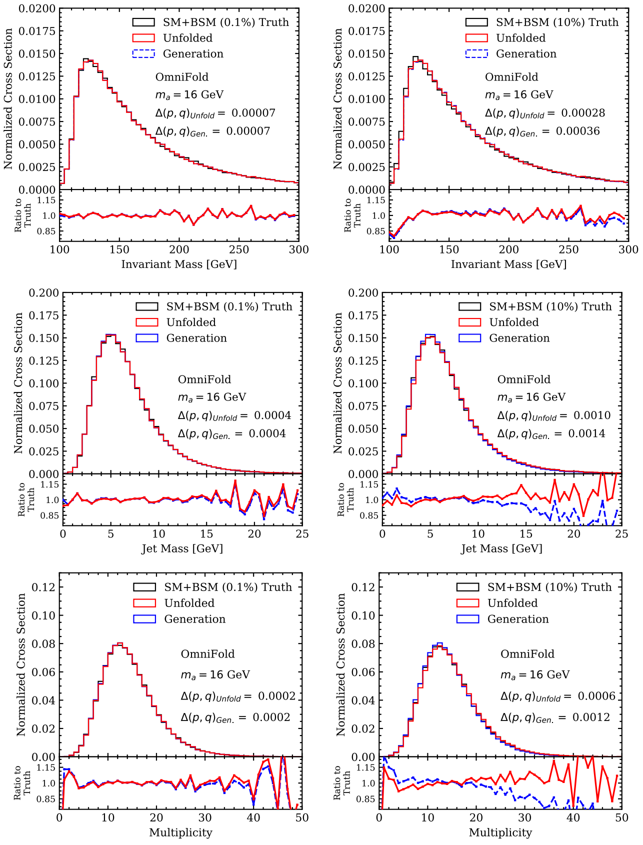

FIG. 7. Truth, Generation, and unfolded Generation distributions for the OmniFold case, where BSM $h \rightarrow Za$, $a \rightarrow gg$ events have been included in the Truth, but not the Generation. SM events in these samples come from PYTHIA 8$Z+$jets simulation. In the left column, 200 out of 200,000 Truth events come from the BSM sample, and in the right column, 20,000 out of 200,000 Truth events are BSM physics. Distributions are given for the invariant mass of the $Z+$jet, the jet mass, and the jet multiplicity. The ratios of the Generation distributions are given to Truth for each plot. The weights are taken after three iterations of OmniFold.

randomly held out as a validation set. The neural network has the same inputs and architecture as the one used for MultiFold.[4] If MultiFold preserves the complete phase space—the event-by-event distribution of all variables in the data sample including BSM physics, then any threshold cut on this classifier should have the same efficiency with the unfolded data as it does with the Truth.

The number of Truth, Generation, and unfolded Generation events passing a cut on the neural network score, as a function of the cut value, is shown in Fig. 6. In both the 0.1% and 10% signal case, the number of events in the reweighted samples more closely matches the Truth than

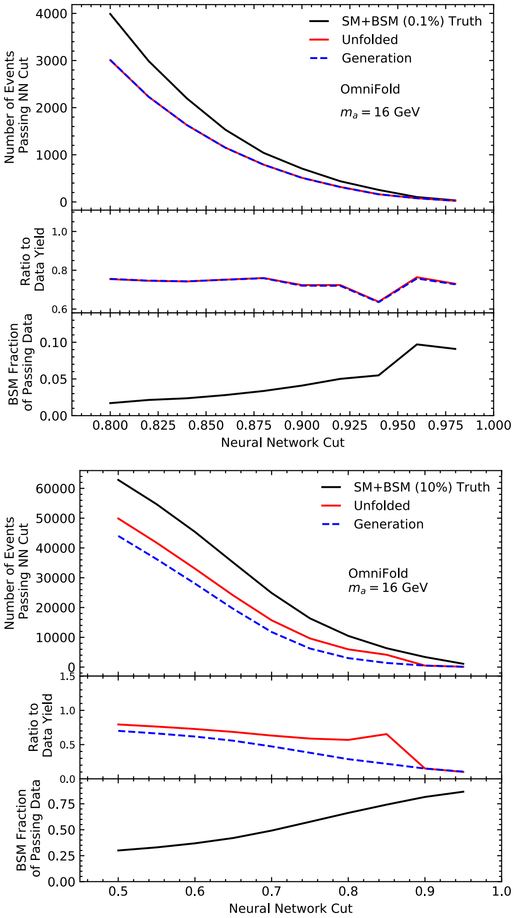

FIG. 8. The number of Truth, Generation, and unfolded Generation events passing a cut on the PFN score, as a function of the cut value, in the case that 0.1% of the data comes from BSM physics (top) and 10% of the data comes from BSM physics (bottom). The unfolding was performed with OmniFold. The PFN was specifically trained to distinguish SM from BSM physics. The middle segment of both plots shows the ratio of the pre- and post-unfolding Generation yields to the Truth yield. The bottom segment shows the fraction of Truth events that are BSM events at truth level.

the raw Generation samples. The agreement between the reweighted sample and Truth in the 10% case is an impressive achievement, as the Truth and Unfolded yields after the application of the NN score cut is stable at one. The NN score is a specialized value that was not used in training, so it is clear that in this case, the most BSM-like events are being upweighted to an appropriate degree. In contrast, the unfolding has not upweighted the BSM events enough in the 0.1% case, highlighting the difficulty of working with such a small signal.

B. Unfolding with OmniFold

An investigation similar to the MultiFold case can be performed with OmniFold. The same m_a and contamination values are investigated.

[4] It is important to note that in the MultiFold case, the neural network is trained to distinguish between Data and Simulation, whereas the discriminator neural network is trained to distinguish truth-level SM events from BSM events.

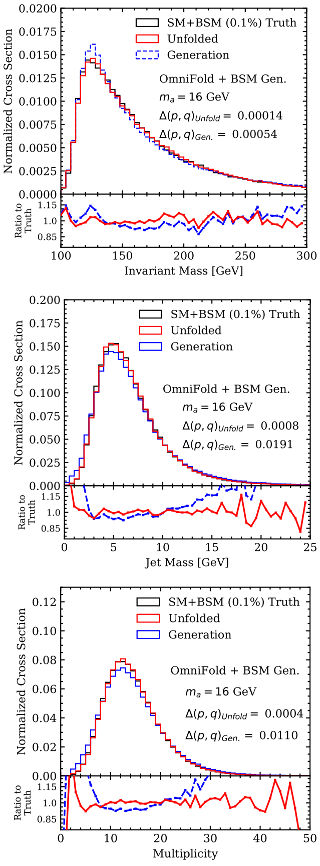

FIG. 9. Truth, Generation, and unfolded Generation distributions for OmniFold, where BSM $h \to Za$, $a \to gg$ events have been included in the Truth and the Generation. SM events in these samples come from PYTHIA 8 $Z+$jets simulation. In the left column, 200 out of 200,000 Truth events come from the BSM sample, and in the right column, 20,000 out of 200,000 Truth events are BSM physics. In all cases, 60,000 Generation events out of 260,000 come from $h \to Za$, $a \to gg$ events with different m_a values. Distributions are given for the invariant mass of the $Z+$jet, the jet mass, and the jet multiplicity. The ratios of the Generation distributions are given to Truth for each plot. The weights are taken after five iterations of OmniFold.

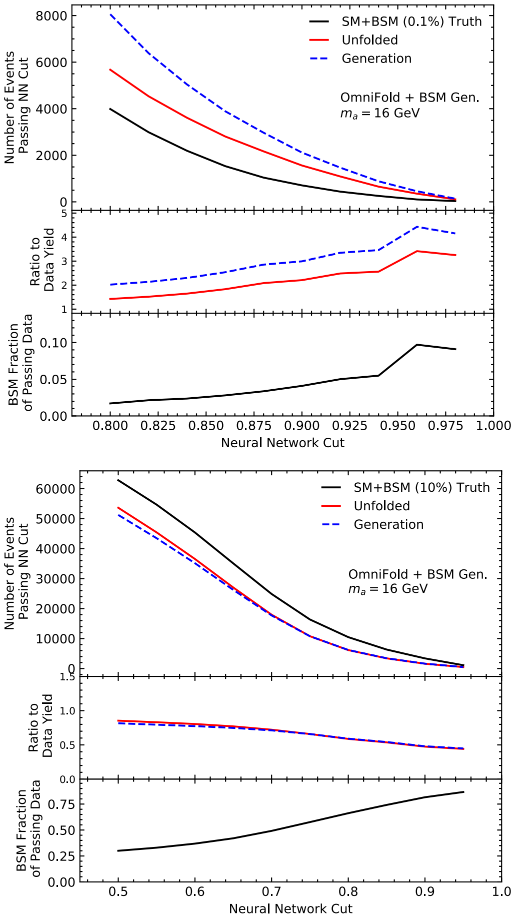

FIG. 10. The number of Truth, Generation, and unfolded Generation events passing a cut on the PFN score, as a function of the cut value, in the case that 0.1% of the data comes from BSM physics (top) and 10% of the data comes from BSM physics (bottom). In both cases, 60,000 Generation events out of 260,000 come from $h \to Za$, $a \to gg$ events with different m_a values. The unfolding was performed with OmniFold. The PFN was specifically trained to distinguish SM from BSM physics. The middle segment of both plots shows the ratio of the pre and postunfolding Generation yields to the Truth yield. The bottom segment shows the fraction of Truth events that are BSM events at truth level.

The distributions of $Z+$jet invariant mass, jet mass, and jet multiplicity for Truth, Generation, and unfolded Generation are shown in Fig. 7. The performance in these plots is similar to, but slightly worse than, that observed in Fig. 5.

It is also possible to train a PFN to distinguish SM from BSM events. This PFN is set up in the same way as the PFN used for OmniFold, but it is trained to discriminate a sample of 90,000 SM from 90,000 BSM events. Here, the area under the ROC curve is 0.94, achieving superior discrimination to the neural network described in Sec. V A (AUC of 0.73). Figure 8 shows the number of Truth, Generation, and unfolded Generation events that pass cuts on the PFN score. If this figure is compared to Fig. 6, it is evident that the post-cut yields in the OmniFold case do not agree with truth as well as in the MultiFold case. This is due to the challenges discussed in Sec. II A: the phase space of the OmniFold case is significantly larger than that of the MultiFold case. Within the full particle-level phase space used by OmniFold, BSM and SM events are more separable than they are in the ten-variable phase space used in our MultiFold example. Therefore, a PFN specifically trained to discriminate SM from BSM events is highly accurate. The Generation sample poorly populates the very BSM-like region of this discriminator especially in the case of 0.1% contamination, where there were only 200 BSM events in the data sample; the unfolded dataset would not enable a discovery of new physics, as the weights do not strongly affect the postcut yields.

TABLE I. Average weights applied to events in Generation based on the event type. Here Generation was 260,000 events, with 200,000 SM $Z+\mathrm{jet}$ events, and 10,000 events each from $h \to Za$, $a \to gg$ samples with different m_a values, as given in the table. Generation was used as an initial distribution for unfolding to a Truth sample of 200,000 events that either had 0.1% or 10% of its events drawn from the $h \to Za$, $a \to gg$ sample with $m_a = 16$ GeV. The average weight is 0.77 to match the normalization of the Truth sample, which has 200,000 events total, rather than 260,000.

Event type	Average weight (0.1% case)	Average weight (10% case)
$m_a = 0.5$ GeV	0.47	0.50
$m_a = 1$ GeV	0.51	0.54
$m_a = 2$ GeV	0.63	0.64
$m_a = 4$ GeV	0.76	0.78
$m_a = 8$ GeV	0.75	0.80
$m_a = 16$ GeV	0.75	0.81
SM	0.81	0.80

C. Including BSM physics in Generation

Similar to the end of Sec. IV, we explore how the performance in the previous section changes if we add in BSM to the Generation. This can effectively populate the regions of phase space that are under-populated by the SM to enable a more precise postunfolding search. For this purpose, we take the same 200,000 SM events in Sec. V B and add 10,000 events each from $h \to Za$, $a \to gg$ samples with $m_a = 0.5$, 1, 2, 4, 8, and 16 GeV, for a total of 260,000 events in Generation. The OmniFold method is carried out in exactly the same way as in Sec. V B. Distributions for the invariant $Z+\mathrm{jet}$ mass, jet mass, and jet multiplicity are shown in Fig. 9. By comparing the triangular-discriminator metric between Figs. 9 and 7, it can be seen that when BSM physics is included in the Generation, the distributions generally agree slightly better after unfolding, particularly in the case of 10% contamination, despite worse initial agreement.

The SM vs BSM discriminator PFN of Sec. V B can applied to this new Generation sample. The results of this application are shown in Fig. 10. The PFN is also able to discriminate events with different m_a values relatively accurately,[5] and it is clear here that the OmniFold reweighted sample does not predict postcut yields well. The average weights found for the different m_a components of Generation are given in Table I. While the $m_a = 16$ GeV events are upweighted relative to the lighter m_a events, it is clear that in the 0.1% contamination case, the $m_a = 16$ GeV events are not adequately downweighted, and in the 10% contamination case, they are not adequately upweighted. Together with Fig. 8, it can be seen that while OmniFold is performed using the full phase space, it has difficulty properly weighting extreme regions of phase space that can be particularly useful to model-dependent searches.

VI. CONCLUSIONS AND OUTLOOK

The OmniFold and MultiFold methods can be used for unbinned, all-variable unfolding in the presence of BSM physics, but there are inherent limitations on its applicability for truth-level searches for new physics.

In general, the distributions of high-level observables are unfolded well, as in Figs. 2, 5, and 7. This would enable model-independent searches or searches that use relatively high-level variables as discriminants, especially if the new physics has a high cross section. However, it is possible to devise strong BSM vs SM discriminating variables that are not necessarily unfolded well, such as the neural network scores shown in Figs. 6 and 8. These discriminants, which probe relatively subtle regions of phase space would most likely be applied in a model-dependent search. While OmniFold uses the full phase space, it has difficulty unfolding such specialized variables. The best performance highlighted above is the 10% BSM contamination case with MultiFold, where the postcut yield in the unfolded sample closely matches that found in Data. In the 0.1% contamination case with MultiFold, there is also an enhancement in the reweighted sample relative to the Generation sample, but the agreement with Truth is not as stable as the 10% case. Together with the lack of agreement in the OmniFold case, this suggests that it would be difficult to make a discovery of BSM physics unless the new physics comprises $> 1\%$ of data events. Such high rates of BSM contamination would likely be discovered through conventional means by experimentalists prior to the release of unfolded datasets.

A significant issue in any attempt to perform a search with unfolded data is the inverse problem highlighted by Fig. 2. Information is lost as particles pass through the detector, as seen in the smearing of the truth-level peak. We have shown how this can be partially recovered by adding BSM events to the Generation. This also helps to populate the most BSM-like regions of phase space. For example, Fig. 3 shows that this can be a powerful means to accurately reproduce an invariant mass peak even at truth level. However, this raises the natural question of how to choose the correct events to include in Generation. The study in Sec. V C highlights the fact that even though high-level distributions can be unfolded well when BSM events are included in Generation, specialized variables may not be unfolded well; in particular, the reweighted distributions in Fig. 9 are significantly different from both the Data *and* from what would be expected in a SM-only case. Because of this, a model-dependent search with the PFN discriminator would be ineffective in the 10% case and return a false positive in the 0.1% case.

[5] Such that few events with $m_a = 4$ GeV will pass the NN cut, for example.

Full phase-space unfolding is a promising direction for postmeasurement searches for resonant new physics; for example, a bump, such as that shown in Fig. 2 can be reproduced in unfolded data even if the prior distribution does not have a bump. However, significant work is required to increase the precision of the unfolding, to understand how to quantify the statistical significance of such an anomaly, and to cope with cases where there are phase space regions with a large-likelihood ratio. It is likely that nonresonant new physics, which may be modeled using effective field theory methods, will be more successful because the likelihood ratio is never too far from unity. This is closer to the previously studied case that investigated the impact of different SM simulations [67]. The resonant examples presented in this paper will serve as an important benchmark for the community as existing methods are extended and new techniques are developed to empower a new class of analyses at the LHC and beyond.

VII. CODE AND DATA

The code for this paper can be found at https://github.com/wpmccormack/OmniFoldBSM.

ACKNOWLEDGMENTS

We thank Eric Metodiev for collaboration in the early stages of this project. B. N. and W. P. M. are supported by the U.S. Department of Energy (DOE), Office of Science under Contract No. DE-AC02-05CH11231. P. K. is supported by the National Science Foundation under Cooperative Agreement No. PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/), and by the U.S. DOE Office of High Energy Physics under Grant No. DE-SC0012567. We would also like to thank Jesse Thaler for useful discussions.

[1] G. Cowan, A survey of unfolding methods for particle physics, Conf. Proc. C **0203181**, 248 (2002), https://inspirehep.net/literature/599644.

[2] V. Blobel, Unfolding methods in particle physics, *Proceedings of PHYSTAT2011* (CERN, Geneva, Switzerland, 2011), p. 240.

[3] V. Blobel, Unfolding, in *Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods* (Wiley-VCH, Weinheim, Germany, 2013), p. 187.

[4] R. Balasubramanian, L. Brenner, C. Burgard, G. Cowan, V. Croft, W. Verkerke, and P. Verschuuren, Statistical method and comparison of different unfolding techniques using RooFit, Int. J. Mod. Phys. A **35**, 2050145 (2020).

[5] G. D'Agostini, A Multidimensional unfolding method based on Bayes' theorem, Nucl. Instrum. Methods Phys. Res., Sect. A **362**, 487 (1995).

[6] A. Hocker and V. Kartvelishvili, SVD approach to data unfolding, Nucl. Instrum. Methods Phys. Res., Sect. A **372**, 469 (1996).

[7] S. Schmitt, TUnfold: An algorithm for correcting migration effects in high energy physics, J. Instrum. **7**, T10003 (2012).

[8] G. Facini, K. Merkotan, M. Schott, and A. Sydorenko, On the model dependence of fiducial cross-section measurements, Mod. Phys. Lett. A **34**, 2050065 (2019).

[9] J. M. Butterworth, D. Grellscheid, M. Krämer, B. Sarrazin, and D. Yallup, Constraining new physics with collider measurements of Standard Model signatures, J. High Energy Phys. 03 (2017) 078.

[10] E. Conte, B. Fuks, and G. Serret, MadAnalysis 5, A user-friendly framework for collider phenomenology, Comput. Phys. Commun. **184**, 222 (2013).

[11] E. Conte, B. Dumont, B. Fuks, and C. Wymant, Designing and recasting LHC analyses with MADANALYSIS 5, Eur. Phys. J. C **74**, 3103 (2014).

[12] E. Conte and B. Fuks, Confronting new physics theories to LHC data with MADANALYSIS 5, Int. J. Mod. Phys. A **33**, 1830027 (2018).

[13] B. Dumont, B. Fuks, S. Kraml, S. Bein, G. Chalons, E. Conte, S. Kulkarni, D. Sengupta, and C. Wymant, Toward a public analysis database for LHC new physics searches using MADANALYSIS 5, Eur. Phys. J. C **75**, 56 (2015).

[14] J. Y. Araz, M. Frank, and B. Fuks, Reinterpreting the results of the LHC with MADANALYSIS 5: Uncertainties and higher-luminosity estimates, Eur. Phys. J. C **80**, 531 (2020).

[15] K. Cranmer and I. Yavin, RECAST: Extending the impact of existing analyses, J. High Energy Phys. 04 (2011) 038.

[16] ATLAS Collaboration, Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework, CERN Report No. ATL-PHYS-PUB-2020-007 (2020), http://cds.cern.ch/record/2714064.

[17] ATLAS Collaboration, RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two *b*-quarks, CERN Report No. ATL-PHYS-PUB-2019-032 (2019), http://cds.cern.ch/record/2686290.

[18] M. Drees, H. Dreiner, D. Schmeier, J. Tattersall, and J. S. Kim, CheckMATE: Confronting your favourite new physics model with LHC data, Comput. Phys. Commun. **187**, 227 (2015).

[19] D. Dercks, N. Desai, J. S. Kim, K. Rolbiecki, J. Tattersall, and T. Weber, CheckMATE 2: From the model to the limit, Comput. Phys. Commun. **221**, 383 (2017).

[20] S. Kraml, S. Kulkarni, U. Laa, A. Lessa, W. Magerl, D. Proschofsky-Spindler, and W. Waltenberger, SModelS: A tool for interpreting simplified-model results from the LHC and its application to supersymmetry, Eur. Phys. J. C **74**, 2868 (2014).

[21] G. Alguero, J. Heisig, C. K. Khosa, S. Kraml, S. Kulkarni, A. Lessa, P. Neuhuber, H. Reyes-González, W. Waltenberger, and A. Wongel, New developments in SModelS, Proc. Sci. TOOLS2020 (**2021**) 022.

[22] M. Papucci, K. Sakurai, A. Weiler, and L. Zeune, Fastlim: A fast LHC limit calculator, Eur. Phys. J. C **74**, 3163 (2014).

[23] D. Barducci, A. Belyaev, M. Buchkremer, G. Cacciapaglia, A. Deandrea, S. De Curtis, J. Marrouche, S. Moretti, and L. Panizzi, Framework for model independent analyses of multiple extra quark scenarios, J. High Energy Phys. 12 (2014) 080.

[24] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018).

[25] R. T. D'Agnolo and A. Wulzer, Learning new physics from a machine, Phys. Rev. D **99**, 015014 (2019).

[26] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99**, 014038 (2019).

[27] R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning multivariate new physics Eur. Phys. J. C **81**, 89 (2021).

[28] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, Phys. Rev. D **101**, 075021 (2020).

[29] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or What?, SciPost Phys. **6**, 030 (2019).

[30] T. S. Roy and A. H. Vijay, A robust anomaly finder based on autoencoder, arXiv:1903.02032.

[31] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Variational autoencoders for new physics mining at the large hadron collider, J. High Energy Phys. 05 (2019) 036.

[32] A. Blance, M. Spannowsky, and P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches, J. High Energy Phys. 10 (2019) 047.

[33] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics, Phys. Rev. D **101**, 076015 (2020).

[34] A. De Simone and T. Jacques, Guiding new physics searches with unsupervised learning, Eur. Phys. J. C **79**, 289 (2019).

[35] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, Does SUSY have friends? A new approach for LHC event analysis, J. High Energy Phys. 02 (2021) 160.

[36] G. M. Alessandro Casa, Nonparametric semisupervised classification for signal detection in high energy physics, arXiv:1809.02977.

[37] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Uncovering latent jet substructure, Phys. Rev. D **100**, 056002 (2019).

[38] A. Andreassen, B. Nachman, and D. Shih, Simulation assisted likelihood-free anomaly detection, Phys. Rev. D **101**, 095004 (2020).

[39] B. Nachman and D. Shih, Anomaly detection with density estimation, Phys. Rev. D **101**, 075042 (2020).

[40] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, J. High Energy Phys. 11 (2017) 163.

[41] M. Romão Crispim, N. Castro, R. Pedro, and T. Vale, Transferability of deep learning models in searches for new physics at colliders, Phys. Rev. D **101**, 035042 (2020).

[42] M. C. Romao, N. Castro, J. Milhano, R. Pedro, and T. Vale, Use of a generalized energy mover's distance in the search for rare phenomena at colliders Eur. Phys. J. C **81**, 192 (2021).

[43] O. Knapp, G. Dissertori, O. Cerri, T. Q. Nguyen, J.-R. Vlimant, and M. Pierini, Adversarially learned anomaly detection on CMS open data: Re-discovering the top quark, Eur. Phys. J. Plus **136**, 236 (2021).

[44] ATLAS Collaboration, Dijet Resonance Search with Weak Supervision using 13 TeV pp Collisions in the ATLAS Detector, Phys. Rev. Lett. **125**, 131801 (2020).

[45] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning the latent structure of collider events, J. High Energy Phys. 10 (2020) 206.

[46] M. C. Romao, N. Castro, and R. Pedro, Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders Eur. Phys. J. C **81**, 27 (2021).

[47] O. Amram and C. M. Suarez, Tag N' train: A technique to train improved classifiers on unlabeled data, J. High Energy Phys. 01 (2021) 153.

[48] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, Variational autoencoders for anomalous jet tagging, arXiv:2007.01850.

[49] C. K. Khosa and V. Sanz, Anomaly awareness, arXiv:2007.14462.

[50] P. Thaprasop, K. Zhou, J. Steinheimer, and C. Herold, Unsupervised outlier detection in heavy-ion collisions, Phys. Scr. **96**, 064003 (2021).

[51] S. Alexander, S. Gleyzer, H. Parul, P. Reddy, M. W. Toomey, E. Usai, and R. Von Klar, Decoding dark matter substructure without supervision, arXiv:2008.12731.

[52] J. A. Aguilar-Saavedra, F. R. Joaquim, and J. F. Seabra, Mass unspecific supervised tagging (MUST) for boosted jets, J. High Energy Phys. 03 (2021) 012.

[53] K. Benkendorfer, L. L. Pottier, and B. Nachman, Simulation-Assisted Decorrelation for Resonant Anomaly Detection, Phys. Rev. D **104**, 035003 (2021).

[54] Adrian Alan Pol, Victor Berger, Gianluca Cerminara, Cecile Germain, and Maurizio Pierini, Anomaly detection with conditional variational autoencoders, arXiv:2010.05531.

[55] V. Mikuni and F. Canelli, Unsupervised clustering for collider physics, Phys. Rev. D **103**, 092007 (2021).

[56] M. van Beekveld, S. Caron, L. Hendriks, P. Jackson, A. Leinweber, S. Otten, R. Patrick, R. Ruiz de Austri, M. Santoni, and M. White, Combining outlier analysis algorithms to identify new physics at the LHC, J. High Energy Phys. 09 (2021) 024.

[57] 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.

[58] D. A. Faroughy, Uncovering hidden patterns in collider events with Bayesian probabilistic models, Proc. Sci. ICHEP2020 (**2021**) 238.

[59] G. Stein, U. Seljak, and B. Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, arXiv:2012.11638.

[60] G. Kasieczka *et al.*, The LHC olympics 2020: A community challenge for anomaly detection in high energy physics, arXiv:2101.08320.

[61] P. Chakravarti, M. Kuusela, J. Lei, and L. Wasserman, Model-independent detection of new physics signals using interpretable semi-supervised classifier tests, arXiv:2102.07679.

[62] J. Batson, C. G. Haaf, Y. Kahn, and D. A. Roberts, Topological obstructions to autoencoding, J. High Energy Phys. 04 (2021) 280.

[63] A. Blance and M. Spannowsky, Unsupervised event classification with graphs on classical and photonic quantum computers, arXiv:2103.03897.

[64] B. Bortolato, B. M. Dillon, J. F. Kamenik, and A. Smolkovič, Bump hunting in latent space, arXiv:2103.06595.

[65] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, Comparing weak- and unsupervised methods for resonant anomaly detection, Eur. Phys. J. C **81**, 617 (2021).

[66] B. Nachman, Anomaly detection for physics analysis and less than supervised learning, arXiv:2010.14554.

[67] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler, OmniFold: A Method to Simultaneously Unfold All Observables, Phys. Rev. Lett. **124**, 182001 (2020).

[68] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, and U. Köthe, Invertible networks or partons to detector and back again, SciPost Phys. **9**, 074 (2020).

[69] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, and R. Winterhalder, How to GAN away detector effects, SciPost Phys. **8**, 070 (2020).

[70] G. Zech and B. Aslan, Binning-free unfolding based on monte carlo migration, eConf **C030908**, TUGT001 (2003), https://inspirehep.net/literature/637561.

[71] L. Lindemann and G. Zech, Unfolding by weighting Monte Carlo events, Nucl. Instrum. Methods Phys. Res., Sect. A **354**, 516 (1995).

[72] N. D. Gagunashvili, Machine learning approach to inverse problem and unfolding procedure, arXiv:1004.2006.

[73] A. Glazov, Machine learning as an instrument for data unfolding, arXiv:1712.01814.

[74] K. Datta, D. Kar, and D. Roy, Unfolding with generative adversarial networks, arXiv:1806.00433.

[75] M. Zaheer, S. Kottur, S. Ravanbhakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola, Deep sets, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Curran Associates Inc., Red Hook, NY, USA, 2017), p. 3394–3404.

[76] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.

[77] H1 Collaboration, Measurement of lepton-jet correlations in high Q^2 neutral-current DIS with the H1 detector at HERA, Report No. H1prelim-21-031, 2021.

[78] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, A. Suresh, and J. Thaler, Scaffolding simulations with deep learning for high-dimensional deconvolution, arXiv:2105.04448.

[79] The ATLAS Collaboration, ATLAS Run 1 Pythia8 tunes, Technical Report No. ATL-PHYS-PUB-2014-021, CERN, Geneva, 2014, https://cds.cern.ch/record/1966419?ln=en.

[80] T. Sjöstrand, S. Mrenna, and P. Z. Skands, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun. **178**, 852 (2008).

[81] T. Sjöstrand, S. Mrenna, and P. Z. Skands, PYTHIA 6.4 physics and manual, J. High Energy Phys. 05 (2006) 026.

[82] 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 PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).

[83] G. Aad *et al.* (ATLAS Collaboration), Search for Higgs Boson Decays into a Z Boson and a Light Hadronically Decaying Resonance using 13 TeV pp Collision Data from the ATLAS Detector, Phys. Rev. Lett. **125**, 221802 (2020).

[84] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.

[85] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

[86] M. Cacciari, G. P. Salam, and G. Soyez, FastJet User Manual, Eur. Phys. J. C **72**, 1896 (2012).

[87] M. Cacciari and G. P. Salam, Dispelling the N^3 myth for the k_t jet-finder, Phys. Lett. B **641**, 57 (2006).

[88] P. T. Komiske, E. M. Metodiev, M. Feickert, and A. Andreassen, Energyflow, https://github.com/pkomiske/EnergyFlow (2018).

[89] J. R. Andersen *et al.*, Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report, arXiv:1605.04692.

[90] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler, Systematics of quark/gluon tagging, J. High Energy Phys. 07 (2017) 091.

[91] S. Catani, G. Turnock, and B. R. Webber, Jet broadening measures in e^+e^- annihilation, Phys. Lett. B **295**, 269 (1992).

[92] P. E. L. Rakow and B. R. Webber, Transverse momentum moments of hadron distributions in QCD jets, Nucl. Phys. **B191**, 63 (1981).

[93] R. K. Ellis and B. R. Webber, QCD jet broadening in hadron hadron collisions, Conf. Proc. C **860623**, 74 (1986), https://inspirehep.net/literature/234764.

[94] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, J. High Energy Phys. 05 (2014) 146.

[95] J. Pumplin, How to tell quark jets from gluon jets, Phys. Rev. D **44**, 2025 (1991).

[96] F. Chollet, Keras, https://github.com/fchollet/keras (2017).

[97] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard*et al.*, Tensorflow: A system for large-scale machine learning, in *OSDI* (TensorFlow, 2016), Vol. 16, pp. 265–283, https://www.tensorflow.org/.

[98] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[99] D. Boutigny *et al.* (*BABAR* Collaboration), *The BABAR Physics Book: Physics at an Asymmetric B Factory* (Stanford Linear Accelerator Center, Menlo Park, CA, 1998).

[100] A. Hocker *et al.*, TMVA—Toolkit for Multivariate Data Analysis (CERN, Geneva, Switzerland, 2007).

[101] F. Topsoe, Some inequalities for information divergence and related measures of discrimination, IEEE Trans. Inf. Theory **46**, 1602 (2000).

[102] G. Aad *et al.*, Measurements of the production cross-section for a Z boson in association with b-jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, J. High Energy Phys. 07 (2020) 044.

[103] A. M. Sirunyan *et al.*, Measurement of differential cross sections for Z boson production in association with jets in proton-proton collisions at $\sqrt{s} = 13$ TeV, Eur. Phys. J. C **78**, 965 (2018).