Using Machine Learning to disentangle LHC signatures of Dark Matter candidates

Charanjit K. Khosa,[1] Veronica Sanz,[1, 2, 3] and Michael Soughton[1]

[1]*Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK*
[2]*Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK*
[3]*Instituto de Física Corpuscular (IFIC), Universidad de Valencia-CSIC, E-46980, Valencia, Spain*

(Dated: September 19, 2020)

We study the prospects of detecting and characterising Dark Matter at colliders using Machine Learning (ML) techniques. We focus on the monojet and missing transverse energy (MET) channel and propose a set of benchmark models for the study: a typical WIMP Dark Matter candidate in the form of a SUSY neutralino, a pseudo-Goldstone impostor in the shape of an Axion-Like Particle, and a light Dark Matter impostor whose interactions are mediated by a heavy particle. All these benchmarks are tensioned against each other and against the main SM background (Z+jets). Our analysis uses both the leading-order kinematic features as well as the information in additional hard jets. We use different representations of the data, from a simple event data sample with values of kinematic variables fed into a Logistic Regression algorithm or a Neural Network, to a transformation of the data into images related to probability distributions, fed to Deep and Convolutional Neural Networks. We also study the robustness of our method against including detector effects, dropping kinematic variables, or changing the number of events per image. In the case of signals with more combinatorial possibilities (events with more than one hard jet), the most crucial data features are selected by performing a Principal Component Analysis. We compare the performance of all these methods, and find that using the 2D images significantly improves the performance.

I. INTRODUCTION

After the Higgs boson discovery at the Large Hadron Collider (LHC), a strong focus has shifted towards Dark Matter (DM) searches. The discovery of DM and its characterisation would have profound consequences in Particle Physics, Cosmology and Astrophysics and the LHC could be the key to it. In spite of having experimental evidence of the presence of DM, we do not know what is its true nature, its mass scale, spin and interactions; or even if DM is a particle or a whole sector of new particles and interactions as in the SM.

The unknown properties of DM open the possibility of many different types of DM candidates being consistent with the DM relic density determination from the Cosmic Microwave Background (CMB), as well as other astrophysical constraints. To further explore DM, searches are being conducted in three main directions: underground experiments aiming to directly detect the interaction of DM with nuclei (*direct detection*), astrophysical observatories searching for an excess of light or charged particles in the sky (*indirect detection*), and collider searches for imprints of DM in collisions of protons and leptons (*collider searches*).

Among the hypothesized DM candidates, the category of Weakly-Interacting Massive Particles (WIMPs) enjoy a privileged position, as a WIMP DM could in turn link to other issues plaguing the Standard Model (SM) of Particle Physics. Indeed, the WIMP *paradigm* is realised in many extensions of the SM, such as Supersymmetry (SUSY), where the WIMP is typically a new stable Majorana fermion with electroweak couplings and mass linked to the breaking of SUSY.

Typical DM collider searches are based on the idea that a stable and neutral particle, if produced at colliders, would leave the detector without resistance. Hence, the collider strategy is to search for traces of the DM presence via the associated production of other particles, namely the identification of singular objects within the detector (*mono searches*), where a single object could be a jet, W or Z boson, top quark, photon or $t\bar{t}$ pair. The motivation for using these channels is that DM candidates (which cannot be directly detected) could be exposed through a momentum-mismatch in the final state, where the detected objects appear to recoil against nothing.

A variety of DM scenarios could also be indirectly probed at the LHC via displaced track and vertices signatures using the long lived particle searches (see e.g. ref. [1] and references therein). Collider-stable particles are particles which do not interact with the detector and do not decay before leaving it, hence manifesting as missing momentum and mimicking DM particles. Note that a collider-stable particle could be completely stable (as is thought to be the case for DM particles) or be unstable but with a lifetime long enough such that it is not likely to decay before it leaves the detector. In our search for New Physics at the LHC, we must then attempt to remove this degeneracy. This task is even more challenging than analysing the signal and background for DM searches because here we are concerned with two or more (unknown) physics models.

This complex analysis would likely benefit from the implementation of new (more sophisticated) techniques beyond conventional search strategies. This task calls for the use of Machine Learning (ML), which is emerging as a powerful tool for New Physics searches at LHC (see e.g. the recent review by Radovic *et al.* [2] and references therein).

More conventional ML methods like boosted decision

trees (BDTs) have already been incorporated into many data analysis packages which made a significant impact in the analysis. Several recent studies have shown exciting applications of the ML methods for various tasks, like constraining Wilson coefficients of higher-dimensional operators in the EFT framework[3–5], top-tagging[6, 7], cosmological phase transitions[8], parameter exclusion in SUSY models[9], quark-gluon tagging[10] etc. ML techniques have also been used recently for non-collider DM searches using substructure probes[11, 12], for cosmological DM[13] and in direct detection experiments[14].

In this work, we will apply ML techniques to explore the possibility of disentangling different scenarios for DM and impostors. We will consider the canonical search for DM: events with one jet recoiling against missing transverse energy (*monojet searches*), and use Machine Learning techniques in DM signal characterization. The main focus is to explore the potential of ML techniques for the characterization of the discovered DM candidate [1]. Therefore our task is to explore the features of the signal and to be able to distinguish one signal from another. We will compare the features of DM signals from different Beyond the Standard Model (BSM) models. Specifically, we will use as benchmarks of comparison three types of models: a heavy WIMP dark matter from SUSY, Axion-Like Particles (ALPs) [15, 16] and a simplified DM model with a spin-1 mediator. These will provide enough variety of characteristics to analyse differences and degeneracies among models.

Note we will base our analysis solely on differential information, not on overall cross-sections. The reason to restrict the analysis on kinematics alone is due to the freedom one has on the values of couplings in each model. For example, the production cross section of SUSY WIMPs depends strongly on its nature, e.g. Bino-like or Higgsino-like would lead to different values on the total number of events, yet differential kinematics (divided by the total number of events) would not sizeably change. By restricting the analysis to differential information, we can draw more robust conclusions about the ability to distinguish different scenarios. Note though that discovery potential does depend on both discrimination power and production cross section, and this work focuses on the aspects of the analysis which can then be translated into different benchmarks.

The paper is organised as follows. In the next section II we describe the models which we use for benchmarking DM scenarios. In Sec. III, we discuss kinematic features in the monojet signal, both at leading-order in the QCD expansion (LO) and next-to-leading-order (NLO) and show differences between the SUSY benchmark and the SM background. In section IV we describe the ML methods and address the WIMP DM characterization us-

ing ML methods considered. In the last section, we discuss our findings and conclude.

II. DESCRIPTION OF BENCHMARK MODELS

In this section we describe our choices of benchmark models. These models span a large enough range of kinematic features to compare with SM processes, as well as to see the strength and limitations of the task of disentangling different DM scenarios.

Model	Mass	Type of coupling
SUSY1	$m_{\tilde{\chi}^0} = 100$ GeV	Bino-like
SUSY2	$m_{\tilde{\chi}^0} = 200$ GeV	Bino-like
SUSY3	$m_{\tilde{\chi}^0} = 300$ GeV	Bino-like
ALP	negligible	gluon-ALP
EFT	negligible	4-fermion

TABLE I. Summary of benchmark models.

We first consider a set of three benchmarks in the WIMP DM scenario based on Bino-like SUSY neutralinos and with masses in the range 100 to 300 GeV, see Table I. We did not consider heavier WIMPs as they would likely be very hard to find at the LHC in monojet events. The cross section of production in monojet final states decreases quickly with the WIMP mass [17].

To contrast against the WIMP we consider two alternative cases. The first is an Axion-Like Particle (or ALP) which is a paradigm for signatures from pseudo-Goldstone bosons. They can be light and collider-stable due to the derivative (suppressed) nature of their couplings. ALPs themselves could be a DM particle or can be a DM mediator (see e.g. [18]). These exotic particles are constrained by Astrophysics as well as colliders in a complementary fashion [15]. ALPs are also searched in axion experiments, which are designed to target their couplings with the photons.

In this work, we do not restrict the ALP to be a DM candidate, and could decay after being produced, just not inside the detector [2]. It escapes detection as it has no charge under $SU(3)_c \times U(1)_Y$, hence its signatures are the same as DM, mono-searches. For the monojet channel, the ALP relevant couplings are to gluons, as couplings to quarks are mass-suppressed:

$$\mathcal{L}_a \supset -\frac{g_{agg}}{2} a \, \text{Tr} \left[G_{\mu\nu} \tilde{G}^{\mu\nu} \right].\tag{1}$$

Note that the ALP-gluon coupling has the following bound [15, 16]

$$g_{agg} \lesssim 1.1 \cdot 10^{-5} \, \text{GeV}^{-1} \quad (90\% \text{ C.L.}) \quad \text{for} \quad m_a \lesssim 60 \, \text{MeV}\tag{2}$$

———

[1] However in the appendix we also perform the classification for the SM background versus signal benchmarks.

———

[2] If the ALP decays inside the detector, its detection would still be difficult due to its lightness. Nevertheless, one can still search for its effects on tails of distributions [19].

(a) Monojet process in linear ALPs

(b) Monojet process in case of spin 1 mediator

(c) Monojet process in MSSM

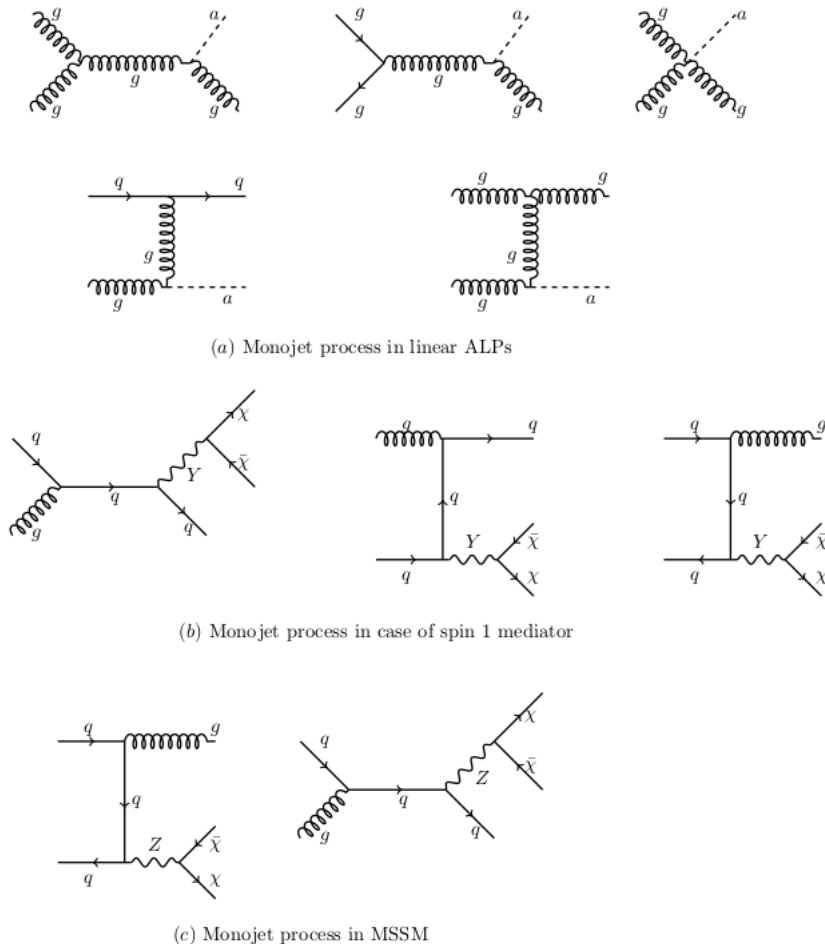



FIG. 1. Feynman diagrams for monojet process in the linear ALP (top), EFT framework with a massive spin-1 mediator (middle) and SUSY WIMP DM cases (bottom).

Note that coupling of ALPs with photons or massive particles would lead to mono-photon, -W, -Z, -top and -Higgs signatures.

We consider a second alternative scenario to WIMP DM based on light DM produced from the decay of the heavy mediator. We label this EFT DM, as the effective interaction between the SM and DM is via a four-fermion higher-order operator. The simplified model Lagrangian which describes the interaction of DM (χ) with the vector mediator (Y) is given by

$$\mathcal{L}_Y = \bar{\chi}\gamma_\mu g_\chi^V \chi Y^\mu. \tag{3}$$

and similarly for the interaction between the mediator and quarks. Note that the kinematic distribution of events is not very sensitive to the dark matter mass (see Fig. A1 in the appendix) in the limit of $m_Y \gg m_\chi$.

Within these models, monojet signatures would result from the processes shown in Fig. 1: a pair of DM particles (or a single ALP) produced in association with one initial state radiation (ISR) gluon or quark.

Finally, we consider the dominant SM background given by Z+jets, where the Z boson decays to neutrinos.

III. KINEMATIC DISTRIBUTIONS

After presenting our benchmark scenarios for New Physics, in this section we describe the kinematic features which the Machine Learning algorithms will be able to optimize over. The aim of this section is to explain some of the features that the ML results will show.

The input for the ML studies is a list of events with their kinematic features simulated using a particular benchmark model. We perform both parton- and detector-level simulations to generate the data samples. More details related to event generation can be found in Appendix A.

The kinematic features of those events are generically multi-dimensional, but one can project into 1D and 2D

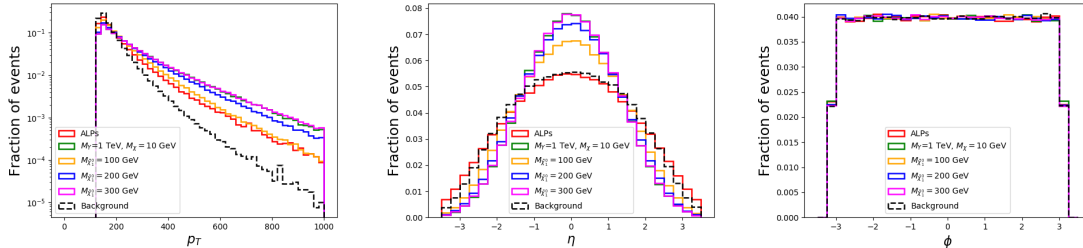



FIG. 2. Monojet parton-level LO 1D histograms for WIMP, EFT and ALP scenarios, as well as the SM background.

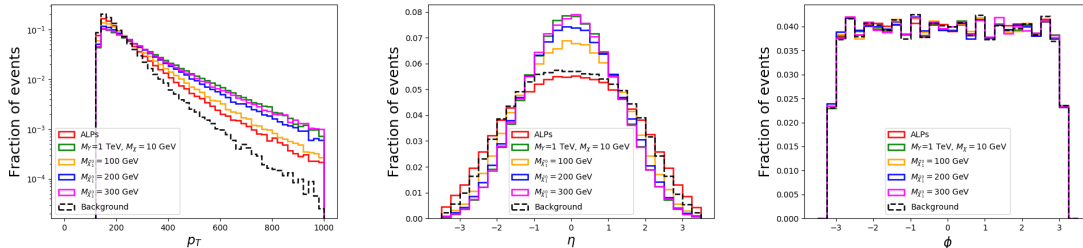



FIG. 3. Monojet detector-level 1D histograms for WIMP, EFT and ALP scenarios, as well as the SM background.

kinematic space. For example, in Fig. 2 we show a set of 1D distributions for this process. Note that the jet transverse momentum in SUSY1 (lightest SUSY WIMP) and ALP is very similar, and also close to the SM background distribution, whereas the other SUSY scenarios and the EFT exhibit a harder spectrum and easier discrimination with respect to the SM. Additionally, the pseudorapidity η distribution of the jet is very similar for the ALP and SM background cases.

At the level of 1D distributions, one cannot distinguish any preferred direction of the azimuthal angle ϕ distribution for both the New Physics signals and SM background processes. Additional information can be obtained when moving from 1D to 2D distributions, as we will show in section IV C and is discussed (in the Bayesian context) in Ref. [20], where we show the event distribution in the 2D plane of p_T^j vs η. Different models are compared to each other in this plane. The top three panels show a clear difference in the high-p_T region of SUSY2, SUSY3 and EFT benchmarks against the SM background. In the lower-leftmost plot we show that this is not the case for ALPs, whose p_T distributions are not so peaked. Nevertheless, the η and p_T information is still useful. Finally, the two other lower plots show ALPs as compared to the EFT benchmark and the light SUSY case. The EFT case leads to harder p_T spectrum and easier to differentiate from ALPs than the light SUSY case. We will see how the ML techniques will exhibit a similar trend. ALPs will be hard to separate from SUSY1, or light DM, whereas EFT and SUSY3 will have the most overlap, as both exhibit similar hard spectra.

The characteristics of these rough features, broadening in pseudorapidity and p_T reach, and what conclusions

one can draw from them, depend on level of accuracy of our simulation. To explore robustness against showering and detector effects, we promoted the simulation to the `Pythia` and `Delphes` level. In Fig. 3 we show the same choices distributions as in Fig. 2 but now at detector-level. The overall behaviour is indeed maintained from parton-level to detector-level.

Finally, we discuss how one would use another source of information from these monojet topologies. Monojet events do often contain an additional hard jet. Splitting or additional ISR emission is not extremely rare, and with an additional object in the final state more information can be extracted of the DM nature. To account for the additional jet, we simulate detector-level events for the monojet (LO) and dijet (NLO) production. For all the cases, we consider data samples of 50K events with one additional jet of $p_T > 25$ GeV. 1D histograms for the constructed kinematic variable are shown in Fig. 4. An additional jet provides a new source of kinematic information, enhancing the discriminating power of our study. For example, notice that the distributions of $\Delta\phi_{j_1 j_2}$ (j_1: leading jet, j_2: sub-leading jet) and $\Delta\phi_{MET}^{j_2}$ are quite characteristic for the signals and background.

IV. DM CHARACTERIZATION USING MACHINE LEARNING

In this section we start our discussion on the use of supervised Machine Learning techniques for Dark Matter signal characterization meaning that we focus on disentangling the different DM signals from each other rather than the discovery of any DM signal from the back-

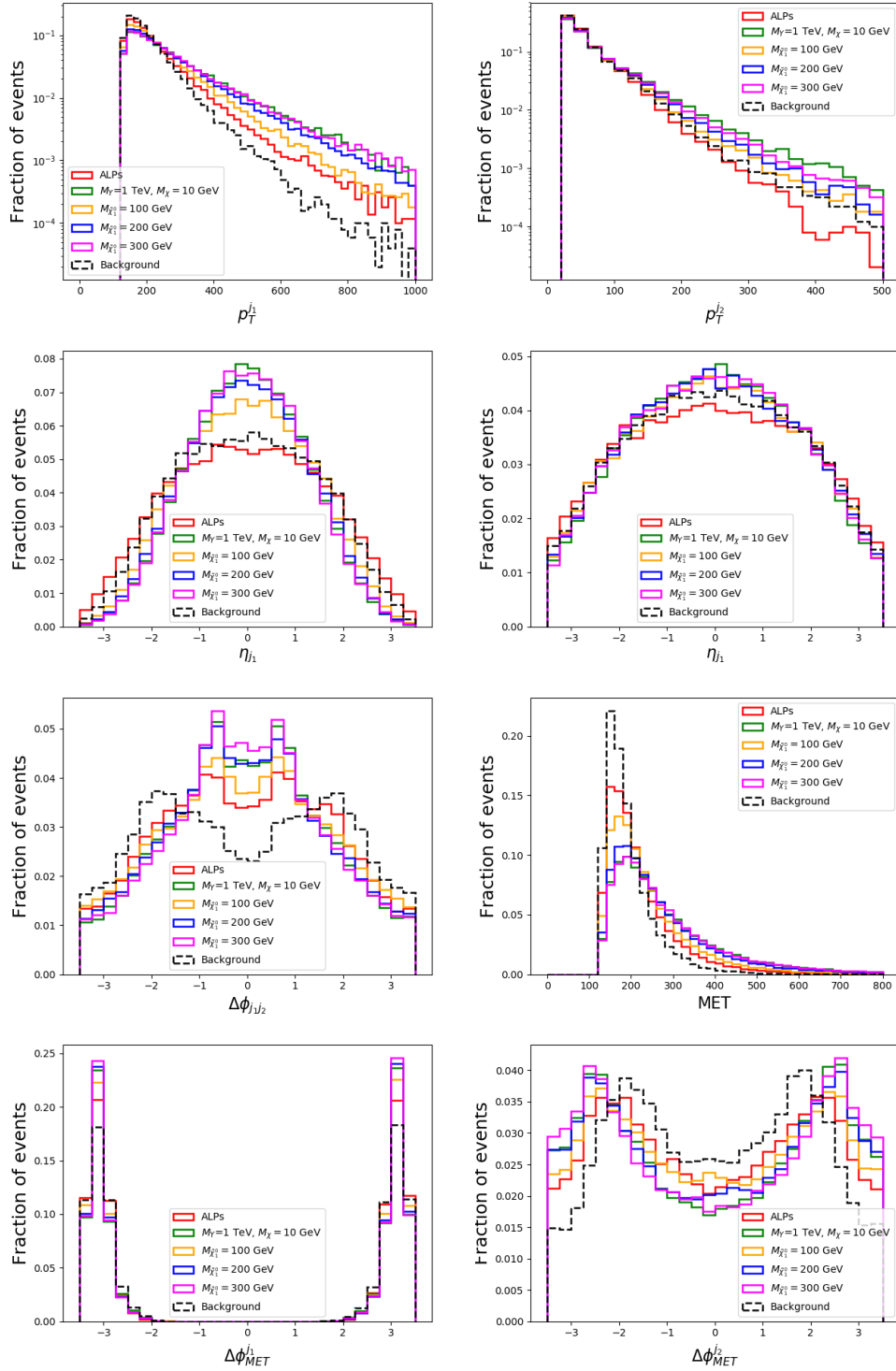



FIG. 4. Detector-level 1D histograms for the dijet process for WIMP, EFT and ALP scenarios, as well as the SM background.

ground. Our analysis is done for the idealised scenario where all signals and the background have the same number of events as mentioned earlier, though we show results for comparison with background in Appendix C. The main question we shall endeavour to answer is then whether it is possible to disentangle an electroweak-scale

WIMP DM signal from an ALP or an EFT signal with very light DM.

The input data we used for the analysis consists of three kinematic features of the monojet i.e. p_T, η_j and ϕ_j. We use this data both as an array input and also in the form of 2D histograms to build a ML algorithm.

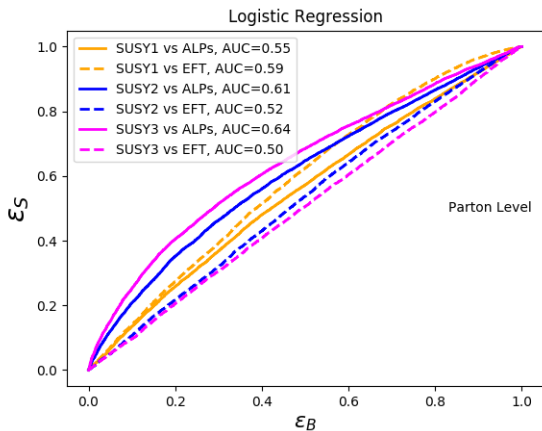



FIG. 5. ROC curves using logistic regression for different WIMP scenarios versus ALP/EFT signals for the LO parton level analysis.

In the following sections, we will compare the performance of different methods, and also parton-level versus detector-level simulations. For this analysis, we will consider the kinematics of the leading jet ($p_T > 130$ GeV) for the LO analysis and in addition to this also of the sub-leading jet for the NLO data sample. For the reasons mentioned in Sec. II, we do not use the cross-section information, so a balanced dataset is considered for all the classes. Data samples are divided in $70\% : 30\%$ proportions for the training and test samples.

A. Logistic Regression

As a first step, we perform logistic regression for the monojet data without any data processing. We used SGDclassifier of SKLEARN python library with a log-loss function. For the parton level events, Receiver operating characteristic (ROC) curves for SUSY signals vs ALP and EFT signals are shown in Fig 5. We consider the three benchmark values for the neutralino mass in the WIMP scenario described in Sec. II. We can see that the AUC[3] value varies from 0.50 to 0.64 for the SUSY3 versus EFT and SUSY3 versus ALP, and the other four cases lie in-between. In other words, one can *easily* separate heavy WIMPs from the ALP monojet, however EFT monojet would not be efficiently classified. The regularization parameter $\lambda = 10^{-5}$ is used for all the cases.

---

[3] AUC (area under the ROC curve) is a measure of the algorithm's performance.

B. Neural Networks-kinematic features

We investigate the classification accuracy using Deep Neural Networks (DNN) with the same input features i.e. p_T, η and ϕ of the jet. We used five fully-connected hidden layers for the network as including more layers does did improve the performance.

For all the layers, except the final one, the number of neurons equal to the number of data features that are considered. For the intermediate layers, a ReLU activation function is used, whereas we use a 'sigmoid' function for the output layer. We considered the binary cross-entropy loss function. The 'dropouts' option is also activated, with 0.2 as the optimized choice. Finally, the 'adadelta' optimizer, and batch-size and epochs are set to 500 and 300, respectively.

Before discussing the performance of DNN for the signal classification, we discuss some aspects of the analysis when using DNNs with the three kinematic inputs: the first aspect relates to the robustness of the analysis with changes in the level of simulation detail. We see from the similarity between figures A3 and A4 in Appendix C for the signal versus background scenario that the DNN only performs slightly worse when using the more noisy detector-level data than the pure parton-level data. We also compared the performance of DNNs with just two variables (by excluding the ϕ variable) and found that there is no degradation in the classification accuracy. This is expected, as the three input variables are redundant once energy-momentum conservation within the event is taken into account. We also explore the benefit of considering NLO dijet processes in addition to mono-jet for the signal versus background scenario. In this case we consider eight features: $p_T^{j_1}, p_T^{j_2}, \eta^{j_1}, \eta^{j_2}$, MET, $\Delta\phi_{j_1 j_2}$, $\Delta\phi_{MET}^{j_1}$, and $\Delta\phi_{MET}^{j_2}$. Using the DNN with the same architecture for this data sample we get indeed an enhancement in the classification accuracy as shown in Fig. A5 in Appendix C.

We use NNs with kinematic features as input for the classification of different signals. As mentioned earlier, classification accuracy is not very different for the detector level simulated data, hence we consider parton-level data in this section. In Fig. 6, we show the ROC curves using the Neural Network with raw kinematic features for the WIMP signal, where the other new physics signals are taken as competitors. The AUC values for WIMP SUSY3 versus EFT, and SUSY1 versus axion signal are 0.50 and 0.58 respectively (same as logistic regression case). Therefore the likelihood to misidentifying light WIMPs and axions is very high, and the same is true for heavy WIMP and EFT DM.

Another way to express this statement is shown in Fig. 7, where we plot the likelihood that a true WIMP signal is misidentified for an ALP or EFT DM scenario. The likelihood is computed as the false positive rate for the optimal point on the ROC curve. Since we are not considering the weight of the cross-section for different signals, we find the optimal point by minimizing the Eu-

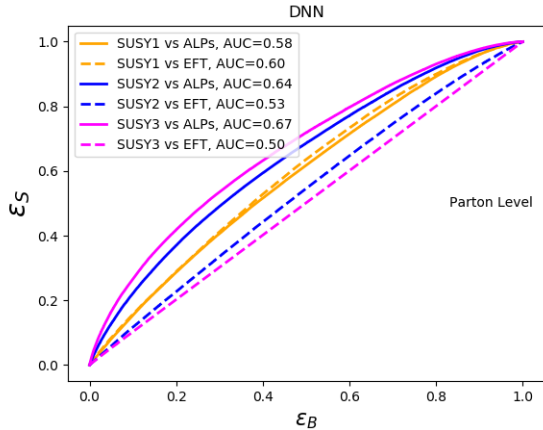



FIG. 6. ROC curves using Neural Network for different WIMP scenarios versus ALP/EFT signals using the raw kinematic features (LO parton-level analysis).

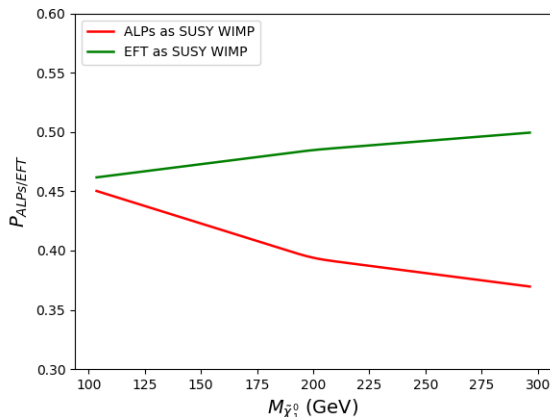



FIG. 7. Probability of misidentifying ALPs or EFT DM scenarios with a true SUSY WIMP as a function of the neutralino mass (parton-level LO data).

clidean distance from the (1=TPR[4],0=FPR) value.

C. DNN with 2D histograms

As mentioned earlier, the information in the monojet event is saturated by choosing two variables, p_T^j and η^j. Therefore, inspired by the use of Convolutional Neural Networks (CNNs) in the classification of images, we construct 'images' from 2D histograms using p_T and η of the jet.

---

[4] TPR is fraction of signal events correctly identified by the algorithm, and FPR is the fraction of background events identified as signal events by the algorithm.

One has choices on how to group the events into images. The simulated dataset contains N_{Tot} total number of events, and is divided into N_{Images} number of images, such that each image contains $r = N_{\mathrm{Tot}}/N_{\mathrm{Images}}$ number of events.

Creating a number of 'images' to train a network is a powerful tool since each image is itself giving an approximation of the joint Probability Density Function (PDF) distribution for both p_T and η. The degree to which each image approximates the PDF depends on the number of events r chosen to be in an image. For a fixed total number of events (determined by the LHC integrated luminosity) there will then be a trade-off between r and N_{Images} that will affect the accuracy of the model.

We will examine this method in both the monojet-LO case and the dijet-NLO case.

1. Monojet-LO

Before analysing the data with a CNN, we first attempt solving the problem with a DNN. This is achieved by decomposing the 2D histogram into a 1D array with values corresponding to the normalised number of events in each bin. Note that whilst the may seem like we are just reconstructing the original distributions, this is not the case since this data now contains correlations from individual distributions. A few illustrative pictures of these plots are shown in Fig. 8.

We consider p_T and η in range [130, 2000] GeV and [-4 to 4], respectively with 29×29 bins. We use the information of event density in this grid as an input for a DNN. The network is well optimised with two fully connected hidden layers, both consisting of twenty neurons with a ReLU activation function and a *softmax* activation function for the output layer. After investigation, we do not find over-training to be an issue, so only include a small number of dropout neurons. The network is trained for 300 epochs with a batch size of 500. We evaluate the network's performance through accuracy (found from finding whether the predicted result (ranging from 0 to 1) round to 0 or to 1). This is done within the training dataset whilst training, then with the validation dataset whilst evaluating how the network is training, and finally with the test dataset.

We find that averaging over more number of events per image improves the classification accuracy. As an illustration of this effect, in Fig. 9, we show the accuracy plot for WIMP SUSY3 versus ALPs data and how it depends on the number of events injected in one picture for the fixed data sample. The different colored curves correspond to varying the total data sample. For a fixed data sample, the number of images a DNN is trained and tested is reduced when we increase the number of events per image. As we know that with more events per image the input data approaches closer to the true likelihood function for the event distributions and this is clearly reflected in the increase in accuracy with number

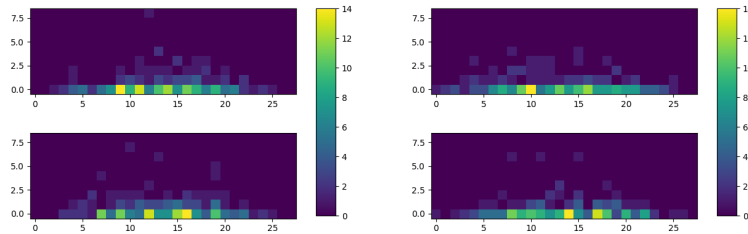



FIG. 8. Illustrative 2D histograms for SUSY1 (left) and ALP case (right) averaging over 200 events.

of events per image.

in Appendix D.

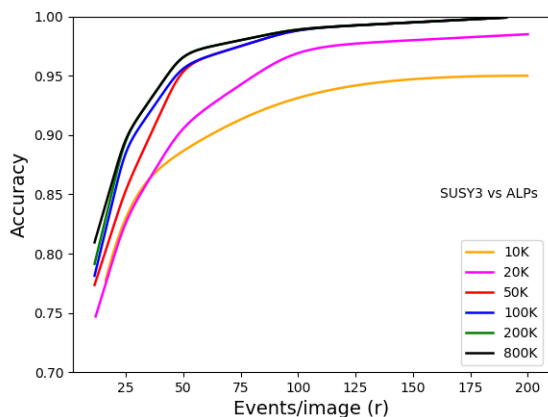
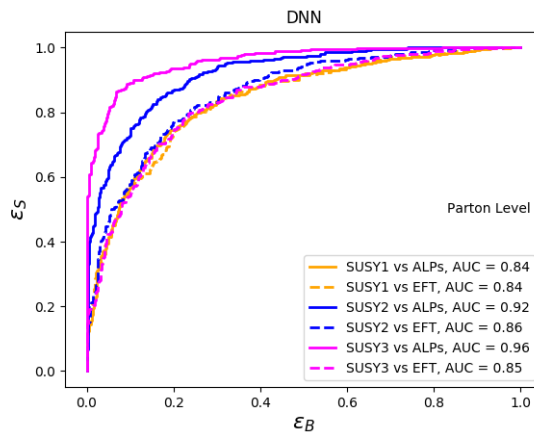



FIG. 9. Accuracy versus number of events per image, with varying total number of events for SUSY3 versus ALP. We use fully-connected NNs with two hidden layers for different data sample size. The parton-level data set is considered for this plot.

FIG. 10. ROC curves (DNN with 2D histograms, $r = 20$) for SUSY WIMP benchmarks versus other signals, for the parton-level with 50K events are used for each signal.

We studied the dependence of the accuracy with the size of the data sample. Once we have enough data to train the DNN, adding more events does not improve accuracy, whereas the accuracy decreases for smaller values of the ratio r. In Fig. 9 central values of accuracy are plotted for different total number of events. Using the total number of 100K data sample the DNN performance is very stable, but it starts becoming unstable for the 50K data sample. We also note that using less than 20k total events leads to large errors in accuracy. ROC curves for the signal-to-signal identification using DNN with 2D histograms are shown in Figs. 10 for the LO data sample.

All of the analysis here has been conducted with the assumption that the number of different type of signal events are the same and that we can fully decompose events into purely signal and background. This has allowed us to investigate disentangling the various DM model presented, however it is not a realistic scenario. We therefore also explore how effectively the classifiers can disentangle DM models with varying levels of signal

2. Dijet-NLO

We also consider the Next-to-Leading-Order (NLO) processes in which two parton-level jets in addition to the 1-jet process are produced. We also simulate detector-level for the NLO processes, and use both the leading and sub-leading jet. More details about how two jet events are selected can be found in the Appendix A.

Since for the NLO process we have eight different features, there are 28 possible variations of 2D histograms that we could produce. If this were to be fully taken into account, then it might prove fruitful to combine the results of those 28 ML algorithms together in some manner. Alternatively, one could produce three (or higher)-dimensional histograms and deconvolve them for the DNN, or use a three (or higher)-dimensional CNN. However, we shall not take this approach here to avoid the risk of over-complicating an analysis which may not benefit from it and obscure the physics. Instead, we perform a Principal Component Analysis (PCA) to the NLO data in order to determine which features are more relevant.

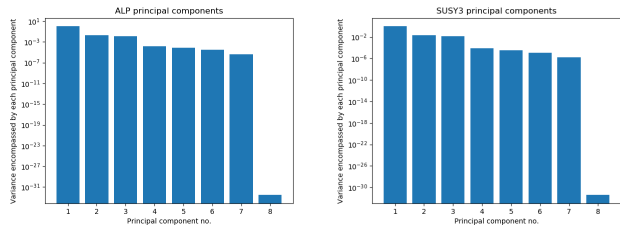



FIG. 11. The explained variance ratio of the new principal component axes (with the number of principal component axes being 8, the same number of original features), highlighting the relative importance of the principal components in terms of capturing variance in the datasets.

Figure 11 (and Fig. A2 in the Appendix) shows that whilst all but one feature columns contribute to the variance of the dataset, a significant amount of the variance within the datasets is captured within the first three principal component axes. By finding the correlation between the original features and the new principal component axes, we can determine which features are most important in terms of holding information. We expect an increase in classification performance if we were to use the principal component axes as new feature columns, however we choose instead to use the original feature columns that correspond most strongly to the most important principal component axes, as we are primarily interested in establishing which features are most relevant for our 2D analysis as a proof-of-concept.

The correlations can be seen in Table II for ALP and the SUSY3 benchmark, and for other cases in Tables B1, B2, B3, and B4 of the Appendix B. One can see from the table that the three most important features are the same for all datasets, namely $p_T^{j_1}$, $\Delta\phi_{j_1 j_2}$, and $\Delta\phi_{\rm MET}^{j_1}$.

We do find an increase in classifier performance when using the $p_T^{j_1}$ and $\Delta\phi_{\rm MET}^{j_1}$ distributions over $p_T^{j_1}$ and η_{j_1}. The ROC curves for a DNN (with the same architecture as before) are shown in Fig. 12.

We note that using dijet data with the 2D histogram approach performs slightly worse than when using the monojet data showing that more information is contained within the p_T^{j} and η_j for a monojet process than the $p_T^{j_1}$ and $\Delta\phi_{\rm MET}^{j_1}$ of one jet in a dijet process and therefore it is still useful to include the other features of a dijet event, even if they do offer diminishing returns on performance after the second. Note that our results are on this are not intended as a comprehensive evaluation but rather as evidence that it is possible and prudent to include dijet events for the 2D histogram method.

D. CNN with 2D histograms

After transforming our events into histograms/images (which are the approximations of PDFs) and processing them using a DNN, we move onto applying CNNs to them. We use a CNN with two convolutional layers, two

ALP PCA correlations							
	$p_T^{j_1}$	$p_T^{j_2}$	η_{j_1}	η_{j_2}	$\Delta\phi_{j_1 j_2}$	MET	$\Delta\phi_{\rm MET}^{j_1}$	$\Delta\phi_{\rm MET}^{j_2}$
PC-1	0.66	0.36	0.01	0.00	-0.01	0.67	0.01	0.01
PC-2	-0.01	0.00	-0.02	-0.02	-0.38	-0.01	0.76	0.52
PC-3	0.01	-0.01	0.05	0.04	-0.76	0.00	0.06	-0.64
PC-4	0.00	0.00	-0.70	-0.71	-0.04	0.00	-0.02	-0.06
PC-5	-0.29	0.93	0.05	-0.04	-0.01	-0.22	-0.01	-0.01
PC-6	0.02	-0.06	0.71	-0.71	0.00	0.01	-0.00	0.00
PC-7	0.70	0.05	0.00	-0.00	0.00	-0.71	-0.00	0.00
PC-8	0.00	-0.00	0.00	0.00	0.52	-0.00	0.64	-0.56

SUSY3, $M_{\tilde{\chi}_1^0} = 300$ GeV PCA correlations							
	$p_T^{j_1}$	$p_T^{j_2}$	η_{j_1}	η_{j_2}	$\Delta\phi_{j_1 j_2}$	MET	$\Delta\phi_{\rm MET}^{j_1}$	$\Delta\phi_{\rm MET}^{j_2}$
PC-1	0.67	0.32	0.00	0.01	0.00	0.67	-0.00	-0.00
PC-2	-0.00	-0.01	0.01	-0.01	0.32	-0.00	-0.76	-0.57
PC-3	0.00	-0.00	-0.07	-0.06	-0.80	0.00	0.11	-0.58
PC-4	-0.01	0.02	0.70	0.70	-0.07	-0.01	0.01	-0.05
PC-5	-0.22	0.95	-0.00	-0.02	0.00	-0.23	-0.01	-0.00
PC-6	-0.00	0.01	-0.71	0.71	0.01	-0.00	-0.01	-0.00
PC-7	0.71	-0.01	-0.00	0.00	-0.00	-0.71	0.00	0.00
PC-8	0.00	0.00	0.00	-0.00	0.50	0.00	0.64	-0.58

TABLE II. PCA original feature to principal component correlations for the ALP and SUSY3 benchmark (rounded to 2 d.p.).

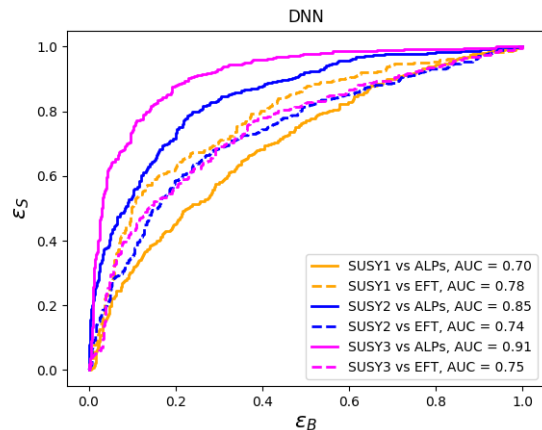



FIG. 12. ROC curves (DNN with 2D histograms, $r = 20$) for SUSY WIMP versus signal for the dijet NLO detector-level simulation. For these plots, 50K events are used for each process.

max-pooling layers and one dense flatten hidden layer, with `ReLu` activation function for all the cases.

ROC curves are shown in Figs. 13 and 14, for LO and NLO processes respectively. As mentioned earlier, CNNs are able to retain the information of spatial correlations and usually perform better than DNNs for the image data. But in our case, because images are constructed from highly processed information (PDFs are already an abstract concept) we find no sizeable improvement respect to the DNN results. The performance for the LO (detector-level) and NLO data sample is the same and is lower than the parton-level LO analysis for the ALPs and SUSY1 case.

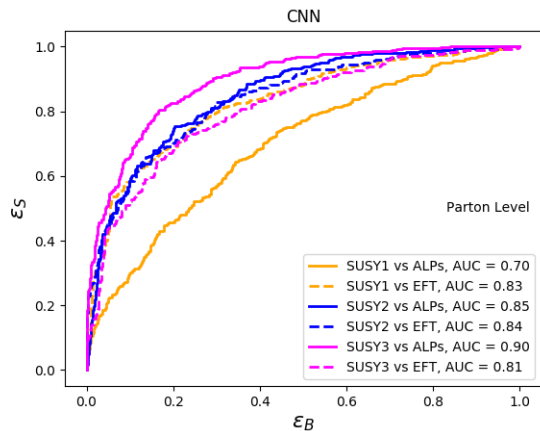



FIG. 13. ROC curves (CNN with 2D histograms, $r = 20$) for SUSY WIMP benchmarks versus other signals, for the parton-level with 50K events are used for each signal.

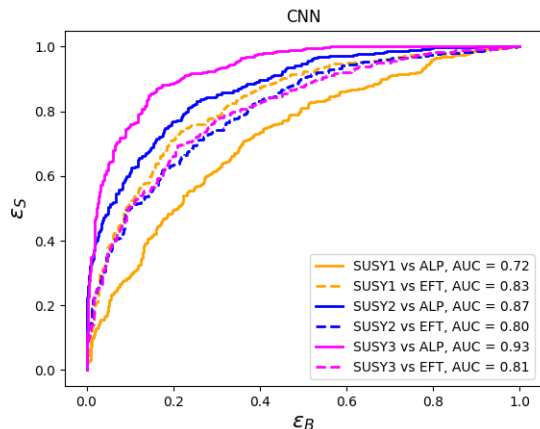



FIG. 14. ROC curves (CNN with 2D histograms, $r = 20$) for signal versus signal for the dijet NLO detector level simulation. For these plots, 50K events are used for each process.

V. CONCLUSIONS AND OUTLOOK

In this paper, we used supervised Machine Learning algorithms for the identification of WIMP dark matter in the monojet and missing transverse energy channel. In addition to considering the prospects of detecting WIMP monojet signal over the SM background, we also address the key question of what is the probability that other signals may mimic WIMP dark matter. We considered LO and NLO parton- detector-level events for the data samples. We found that for the LO signal there is not a significant degradation in the performance when we perform the realistic detector level analysis.

We used different ML approaches to the classification problem: logistic regression, DNNs and CNNs. We found that for the kinematic variables NN performs slightly better than logistic regression.

We then created images made of approximations to the probability density of events in a kinematic 2D plane. We then feed the images to a DNN and a CNN. The processing of these images offers a better classification accuracy than lists of events with values of kinematic observables. We show that this method can produce good results for the characterisation of different DM models. We found a trade-off between the accuracy and number of events averaged for one histogram. In a more realistic situation, one would expect a small sample of DM events, hence our exploration of low values of r.

We also investigated NLO monojet processes and found that the accuracy could increase as compared to the LO monojet, simply because an event with more than one object contains more information. In this case, we performed a PCA analysis to decide the most important combination of features, and then constructed 2D histograms of those although we find that when using the 2D histogram method we obtain slightly worse performance when using dijet events with only two variables however we do demonstrate that it is possible to incorporate the additional information from dijet events into our methods.

The techniques proposed could be used for the actual data incorporating the information of the cross-section by choosing specific model benchmarks. Our study focused instead on features which are independent of these. Nevertheless, one could consider the jet images constructed event-by-event instead of our choice of probability distributions.

Moreover, this analysis could be extended for other channels in collider dark matter searches, e.g. for the Vector Boson Fusion (VBF) topology where we have two additional forward jets. It would be interesting to compare the performance of proposed methods for the dijet and VBF case. A combined analysis of different channels may offer enhanced sensitivity which will be investigated in the near future.

Finally, the promising results of this analysis adds a further motivation to explore the possibility of using unsupervised techniques.

VI. ACKNOWLEDGMENTS

C.K.K. wishes to acknowledge support from the Royal Society-SERB Newton International Fellowship (NF171488). The work of V.S. is supported by the Science Technology and Facilities Council (STFC) under grant number ST/P000819/1. M.S. wishes to acknowledge the support by the Data Intensive Science Center in the South East Physics Network (DISCnet), an extension of the STFC, under grant number ST/P006760/1. We would also like to thank Alexander Belyaev for his comments on the levels of signal and background.

Appendix A: Analysis set-up

1. Leading-order(LO) parton-level simulation

We generate parton-level events for monojet and missing energy signals and centre-of-mass energy $\sqrt{s} = 14$ TeV using MadGraph_aMC@NLO v2.6.3.2 [21]. For SUSY WIMP, we used MSSM-SLHA2 model (all the SUSY spectrum is set to be very heavy except the lightest neutralino). We use the `Feynrules` model file [22, 23] for the linear ALPs [24] and spin-one mediator case (DMsimp_s_spin1 [25] model) in the EFT framework.

Using these models, we generate the following processes

$$pp \to aj \qquad \text{ALPs} \qquad (A1)$$

$$pp \to \chi\bar{\chi}j \qquad \text{EFT, spin-1 mediator} \qquad (A2)$$

$$pp \to \tilde{\chi}_1^0 \tilde{\chi}_1^0 j \qquad \text{SUSY-WIMP} \qquad (A3)$$

For SM monojet background, we consider the following dominant background:

$$pp \to Zj(Z \to \nu\bar{\nu}). \qquad (A4)$$

For all the processes 400K events are generated using a cut of $p_T > 130$ GeV for the jet p_T. The following kinematic variables are constructed;

$$p_T^j(\text{MET}), \eta^j, \phi^j.$$

2. Leading-order (LO) and Next-to-Leading order (NLO) detector-level simulation

We generate both monojet and dijet + MET processes. As earlier, hard processes are generated using Madgraph and PYTHIA [26] is used for hadronization and showering. Detector effects are incorporated using DELPHES [27] with the default ATLAS card. A generation level cut of $p_T^j > 130$ GeV is used for leading jet p_T. For the NLO case, while analysing the root file, in addition to the leading jet p_T cut we also demand $p_T^j > 25$ GeV for the sub-leading jet.

We have used 200K events for all the cases after root file analysis. To avoid the double-counting from showering, a jet merging scheme (MLM) with xqcut 20 GeV is applied. For the monojet, we consider the same three kinematic features as in the parton-level monojet events. For the dijet, we construct the following kinematic variables :

- $p_T^{j_1}$: transverse momentum of the leading jet.

- $p_T^{j_2}$ transverse momentum of the sub-leading jet.

- η^{j_1}: pseudo-rapidity of the leading jet.

- η^{j_2}: pseudo-rapidity of the sub-leading jet.

- MET: missing transverse momentum.

- $\Delta\phi_{j_1 j_2}$: angular separation between the leading and sub-leading jet.

- $\Delta\phi_{\text{MET}}^{j_1}$: angular separation between the MET and leading jet.

- $\Delta\phi_{\text{MET}}^{j_2}$: angular separation between the MET and sub-leading jet.

For the dijet sample, we consider 50K events for all the processes.

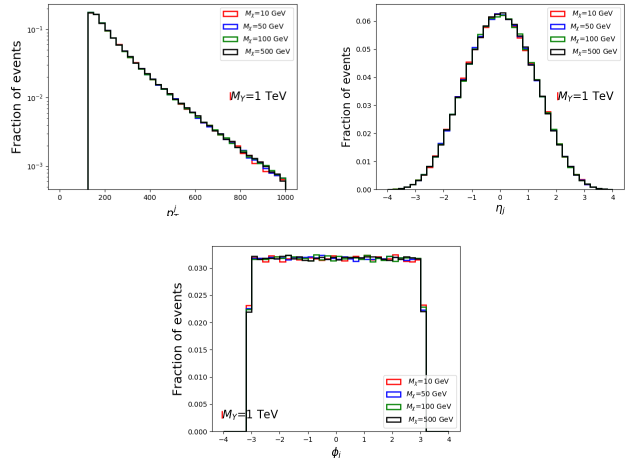



FIG. A1. The p_T, η, and ϕ distributions for the monojet and MET process in case of EFT spin-1 mediator (at parton-level) for fixed mediator mass and various dark matter masses.

Appendix B: PCA correlations

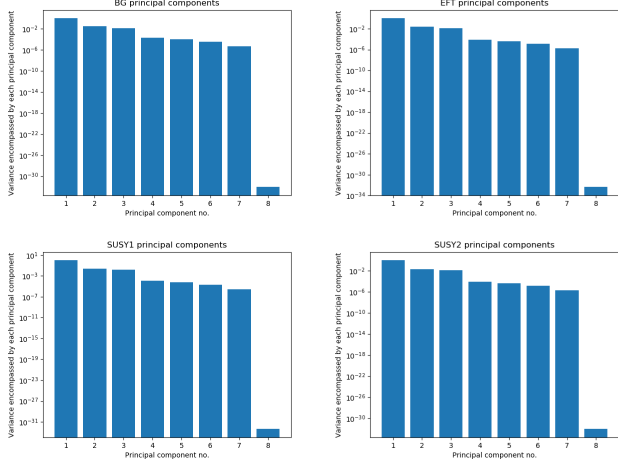



FIG. A2. The variance ratio of the new principal component axes (with the number of principal component axes being 8, the same number of original features), highlighting the relative importance of the principal components in terms of capturing variance in the datasets.

Background PCA correlations							
	$p_T^{j_1}$	$p_T^{j_2}$	η_{j_1}	η_{j_2}	$\Delta\phi_{j_1 j_2}$	MET	$\Delta\phi_{\mathrm{MET}}^{j_1}$	$\Delta\phi_{\mathrm{MET}}^{j_2}$
PC-1	0.67	0.41	-0.01	-0.01	-0.01	0.62	0.00	-0.00
PC-2	0.00	0.00	0.01	0.00	0.45	-0.00	-0.77	-0.45
PC-3	-0.00	-0.00	0.09	0.10	-0.70	-0.00	0.00	-0.70
PC-4	-0.01	0.00	-0.70	-0.70	-0.09	-0.01	-0.01	-0.10
PC-5	-0.12	0.89	-0.01	0.02	-0.00	-0.45	0.00	0.00
PC-6	-0.01	0.01	0.71	-0.71	-0.01	-0.01	0.01	-0.00
PC-7	0.73	-0.23	0.00	0.00	0.00	-0.65	0.00	0.00
PC-8	0.00	0.00	-0.00	-0.00	0.54	-0.00	0.64	-0.54

TABLE B1. PCA original feature to principal component correlations for the SM background (rounded to 2 d.p.).

EFT PCA correlations							
	$p_T^{j_1}$	$p_T^{j_2}$	η_{j_1}	η_{j_2}	$\Delta\phi_{j_1 j_2}$	MET	$\Delta\phi_{\mathrm{MET}}^{j_1}$	$\Delta\phi_{\mathrm{MET}}^{j_2}$
PC-1	0.67	0.34	-0.00	0.00	0.00	0.66	0.01	0.02
PC-2	-0.01	-0.01	-0.00	0.01	-0.33	-0.01	0.76	0.56
PC-3	0.01	-0.00	0.00	0.01	-0.80	0.01	0.10	-0.59
PC-4	-0.00	0.01	0.71	0.71	0.01	-0.00	-0.00	0.00
PC-5	-0.22	0.94	-0.03	0.02	-0.01	-0.27	0.00	-0.00
PC-6	0.01	-0.04	-0.71	0.71	0.01	0.01	-0.01	0.00
PC-7	-0.71	0.04	0.00	-0.00	-0.00	0.70	-0.00	-0.00
PC-8	-0.00	0.00	-0.00	0.00	0.51	0.00	0.64	-0.57

TABLE B2. PCA original features to principal component correlations for the EFT simplified framework (rounded to 2 d.p.).

SUSY1, $M_{\tilde\chi_1^0} = 100$ GeV PCA correlations							
	$p_T^{j_1}$	$p_T^{j_2}$	η_{j_1}	η_{j_2}	$\Delta\phi_{j_1 j_2}$	MET	$\Delta\phi_{\mathrm{MET}}^{j_1}$	$\Delta\phi_{\mathrm{MET}}^{j_2}$
PC-1	0.67	0.35	0.01	0.00	-0.00	0.66	0.01	0.00
PC-2	0.00	0.01	-0.01	-0.01	0.37	0.00	-0.76	-0.53
PC-3	0.00	-0.02	-0.01	-0.01	0.77	0.01	-0.07	0.63
PC-4	-0.00	-0.01	0.71	0.71	0.01	-0.00	-0.01	-0.00
PC-5	-0.21	0.93	-0.03	0.04	0.01	-0.28	0.00	0.01
PC-6	-0.01	0.05	0.71	-0.71	0.00	-0.02	-0.00	-0.00
PC-7	0.71	-0.05	-0.00	-0.00	0.00	-0.70	-0.00	0.00
PC-8	0.00	0.00	0.00	0.00	-0.52	-0.00	-0.65	0.56

TABLE B3. PCA original features to principal component correlations for the SUSY1 case (rounded to 2 d.p.).

SUSY2, $M_{\tilde\chi_1^0} = 200$ GeV PCA correlations							
	$p_T^{j_1}$	$p_T^{j_2}$	η_{j_1}	η_{j_2}	$\Delta\phi_{j_1 j_2}$	MET	$\Delta\phi_{\mathrm{MET}}^{j_1}$	$\Delta\phi_{\mathrm{MET}}^{j_2}$
PC-1	0.67	0.32	0.00	-0.00	-0.01	0.67	0.01	0.01
PC-2	0.01	0.01	0.00	-0.01	0.34	0.01	-0.76	-0.56
PC-3	-0.00	0.01	0.03	0.06	0.79	-0.00	-0.09	0.60
PC-4	0.00	-0.01	0.71	0.70	-0.05	0.00	0.00	-0.04
PC-5	-0.22	0.95	0.01	0.01	-0.01	-0.24	0.00	-0.01
PC-6	0.00	0.00	-0.71	0.71	-0.01	0.00	-0.00	-0.01
PC-7	0.71	-0.01	0.00	-0.00	0.00	-0.71	-0.00	-0.00
PC-8	0.00	0.00	-0.00	0.00	0.51	-0.00	0.65	-0.57

TABLE B4. PCA original features to principal component correlations for the SUSY2 case (rounded to 2 d.p.).

Appendix C: Signal to background analysis

We show here results for signal to background classification. Note that these results are not intended to be used for discovery since we are assuming the same number of signal and background events which would not be the case for real data, but rather to demonstrate the performance of our methods and provide a reference for how well it is possible to distinguish the different signals from background depending on the signal kinematics.

In Fig. A3, we show the ROC curves for various signals at parton-level, whereas the classification accuracy for the detector-level LO and NLO event generation is shown in Fig A4 and A5 respectively. ALPs are more difficult to pick up from the SM background, a task that becomes simpler and simpler as we increase the mass of the DM particle. SUSY3 and EFT, as expected, exhibit very similar performances, and in Fig. A3 the two lines overlap each other. Nevertheless, the SUSY3 benchmark corresponds to a DM particle of 300 GeV, whereas the EFT corresponds to a DM particle with negligible mass.

Appendix D: Classifier performance at different levels of signal

Here we present how the DNN and CNN for 2D histograms perform for different levels of signal. We run the classifiers as in sections IV C and IV D but this time we construct the histograms from a mixture of both signal and background events. We set the total number of

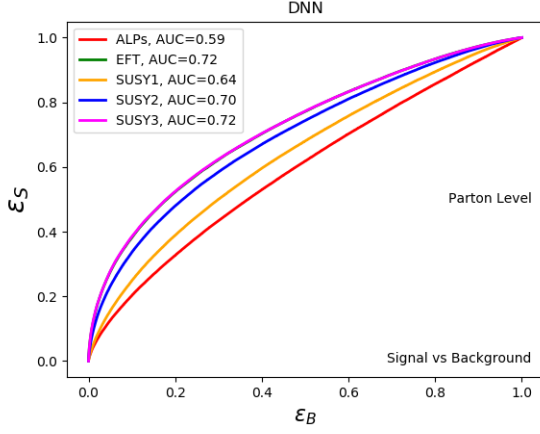



FIG. A3. ROC curves using Neural Networks for different signals versus background for the LO parton-level analysis.

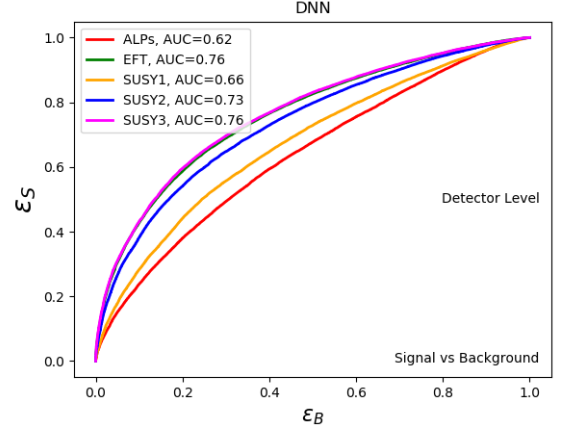



FIG. A5. Neural Network NLO detector-level ROC curves for different signals versus background.

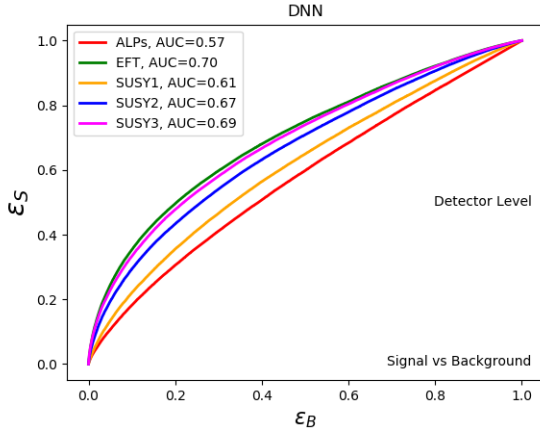



FIG. A4. ROC curves using Neural Networks for different signals versus background for the LO detector-level simulated data.

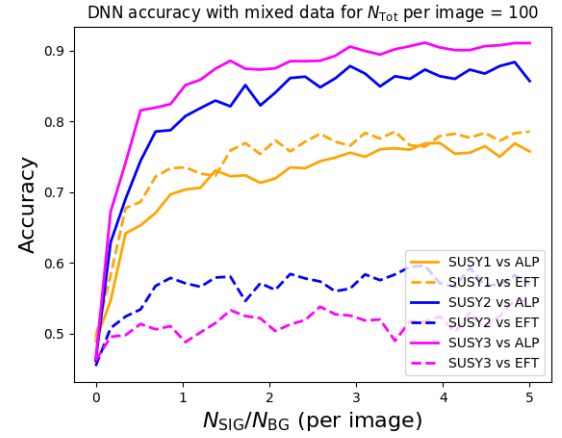



FIG. A6. DNN for 2D histogram accuracy for differing number of signal events for 50 background events trained for 100 epochs

events per histogram to be fixed and change the ratio of number of signal events N_{Sig} to background events N_{BG} and show the model accuracy in Figure A6 for when distinguishing between SUSY WIMP events and ALP and EFT signals for various event ratios. Note that in the limit $N_{\text{Sig}}/N_{\text{BG}} \to \infty$ we reduce to the previous situation where we are comparing only signal with no background. We see that it is harder to distinguish SUSY signals from the EFT than the ALP, especially around the $m_{\tilde{\chi}^0} = 300$ GeV range.

The plot shows us that we require comparable levels of signal to background for good performance, illustrating that this method is not suitable for discovery since discovery would necessitate good performance for $N_{\text{Sig}} \ll N_{\text{BG}}$ (it would be expected that performance increases with total number of events - perhaps with a very large number of events the algorithm would be much more sensitive for lower levels of signal but that is beyond the scope of this paper.

The point of having the mixing between signal and background is to demonstrate how effectively we can disentangle DM candidates beyond the perfect idealised scenario where $N_{\text{Sig}} = N_{\text{BG}}$ - assuming a discovery of signal is made it is good to know benchmarks for performance since we would not be able to fully separate signal from background in a realistic scenario. Knowing these benchmarks is also useful for when constructing unsupervised algorithms where it is useful to know the number

of events required for good performance.

[1] J. M. No, P. Tunney and B. Zaldivar, arXiv:1908.11387 [hep-ph].

[2] A. Radovic *et al.*, Nature **560** (2018) no.7716, 41.

[3] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, Phys. Rev. D **98** (2018) no.5, 052004, [arXiv:1805.00020 [hep-ph]].

[4] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, Phys. Rev. Lett. **121** (2018) no.11, 111801, [arXiv:1805.00013 [hep-ph]].

[5] F. F. Freitas, C. K. Khosa and V. Sanz, Phys. Rev. D **100** (2019) 035040, [arXiv:1902.05803 [hep-ph]].

[6] G. Kasieczka, T. Plehn, M. Russell and T. Schell, JHEP **1705** (2017) 006, [arXiv:1701.08784 [hep-ph]].

[7] G. Kasieczka *et al.*, SciPost Phys. **7** (2019) 014, [arXiv:1902.09914 [hep-ph]].

[8] M. L. Piscopo, M. Spannowsky and P. Waite, Phys. Rev. D **100** (2019) no.1, 016002, [arXiv:1902.05563 [hep-ph]].

[9] S. Caron, J. S. Kim, K. Rolbiecki, R. Ruiz de Austri and B. Stienen, Eur. Phys. J. C **77** (2017) no.4, 257, [arXiv:1605.02797 [hep-ph]].

[10] G. Kasieczka, N. Kiefer, T. Plehn and J. M. Thompson, SciPost Phys. **6** (2019) 069, [arXiv:1812.09223 [hep-ph]].

[11] J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe and K. Cranmer, arXiv:1909.02005 [astro-ph.CO].

[12] S. Alexander, S. Gleyzer, E. McDonough, M. W. Toomey and E. Usai, arXiv:1909.07346 [astro-ph.CO].

[13] C. Escamilla-Rivera, M. A. C. Quintero and S. Capozziello, arXiv:1910.02788 [astro-ph.CO].

[14] U. Simola, B. Pelssers, D. Barge, J. Conrad and J. Corander, JINST **14** (2019) no.03, P03004, [arXiv:1810.09930 [astro-ph.IM]].

[15] K. Mimasu and V. Sanz, JHEP **1506** (2015) 173, [arXiv:1409.4792 [hep-ph]].

[16] I. Brivio, M. B. Gavela, L. Merlo, K. Mimasu, J. M. No, R. del Rey and V. Sanz, Eur. Phys. J. C **77** (2017) no.8, 572, [arXiv:1701.05379 [hep-ph]].

[17] D. Barducci, A. Belyaev, A. K. M. Bharucha, W. Porod and V. Sanz, JHEP **1507** (2015) 066, [arXiv:1504.02472 [hep-ph]].

[18] H. M. Lee, M. Park and V. Sanz, JHEP **1303** (2013) 052, [arXiv:1212.5647 [hep-ph]].

[19] M. B. Gavela, J. M. No, V. Sanz and J. F. de Trocóniz, arXiv:1905.12953 [hep-ph].

[20] F. Ferreira, S. Fichet and V. Sanz, Phys. Lett. B **778**, 35 (2018) doi:10.1016/j.physletb.2018.01.008 [arXiv:1702.05106 [hep-ph]].

[21] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, JHEP **1106** (2011) 128, [arXiv:1106.0522 [hep-ph]].

[22] N. D. Christensen and C. Duhr, Comput. Phys. Commun. **180** (2009) 1614, [arXiv:0806.4194 [hep-ph]].

[23] C. Degrande, C. Duhr, B. Fuks, D. Grellscheid, O. Mattelaer and T. Reiter, Comput. Phys. Commun. **183** (2012) 1201, [arXiv:1108.2040 [hep-ph]].

[24] http://feynrules.irmp.ucl.ac.be/wiki/ALPsEFT

[25] http://feynrules.irmp.ucl.ac.be/wiki/DMsimp

[26] T. Sjostrand, S. Mrenna and P. Z. Skands, Comput. Phys. Commun. **178** (2008) 852, [arXiv:0710.3820 [hep-ph]].

[27] J. de Favereau *et al.* [DELPHES 3 Collaboration], JHEP **1402** (2014) 057, [arXiv:1307.6346 [hep-ex]].