nature astronomy

Supplementary information

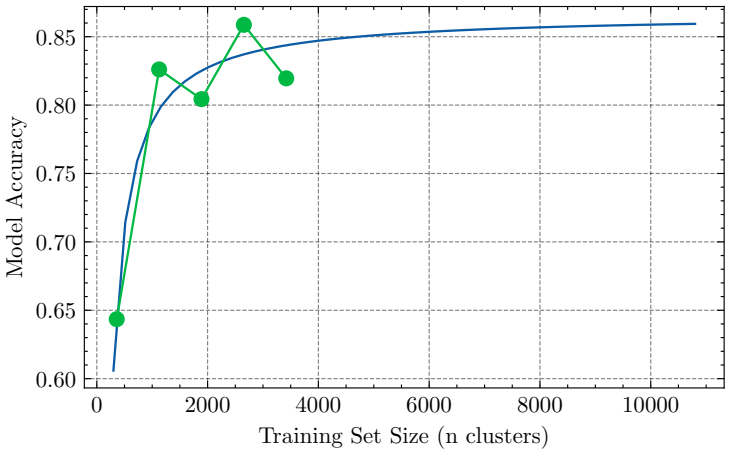
https://doi.org/10.1038/s41550-024-02322-8

A deep-learning algorithm to disentangle self-interacting dark matter and AGN feedback models

In the format provided by the authors and unedited

Supplementary Information 1 Training Set Size

We choose to keep a small training set by only projecting clusters from the simulations to one-dimension (where in principal we could increase the training set by a factor of 3 by projecting to the other two dimensions). We do this to keep the training time to a minimum since we carry out a multitude of tests here. In future work where it is applied to real data, one could reasonably extract all clusters to get the best possible model. Here we estimate the loss of information by keeping the training set small. We take increasingly large training samples and train the Inception model with idealised three channel setup. We then test of simulations generated from a different seed to gauge the accuracy of the model (see the main text for more information regarding this). Figure 1 shows the results. We find that the model becomes relatively insensitive to the training set size above 3000 clusters. We would expect the model to improve by about 2% if we included the full set of clusters.

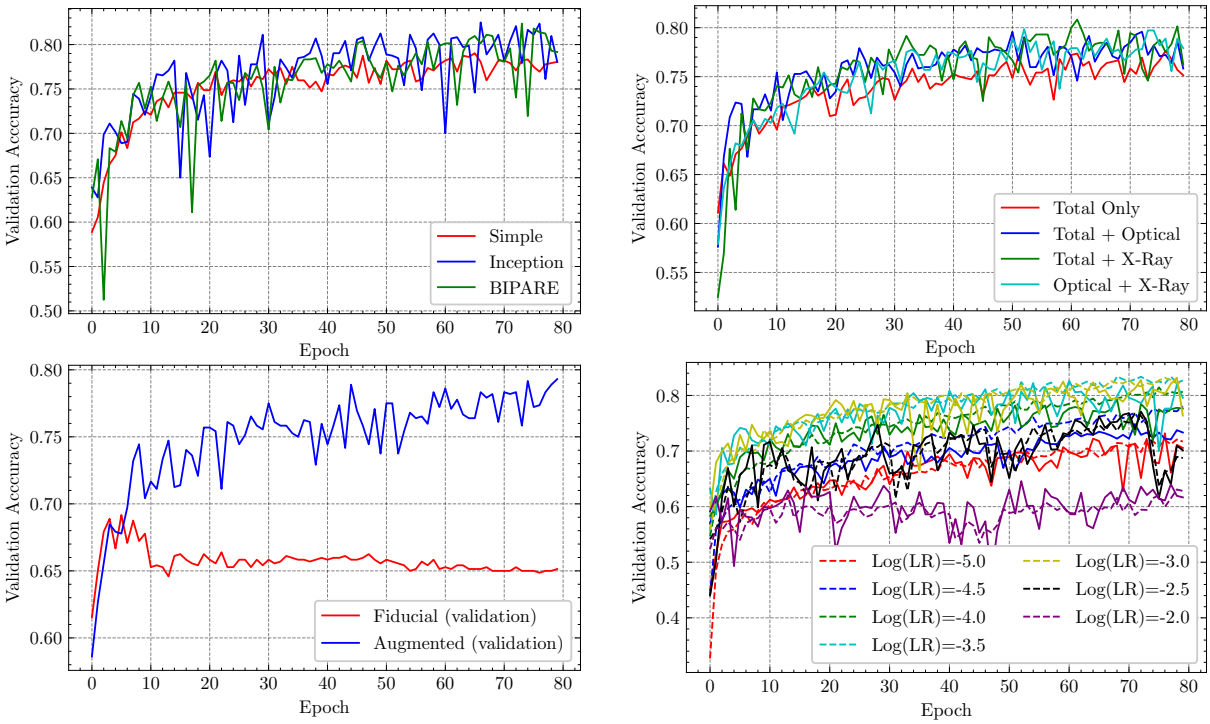


Supplementary Figure 1: We estimate the dependence of the model accuracy on the training set size (i.e. number of simulated clusters). By projecting clusters in to one dimension we only have 3'600 of the 10'800 clusters available. However this dramatically speeds up the training time of our model. Here we show the incremental improvement of the model with increasing training size with a fitted model in pull. We find we would only expect $\sim 2\%$ improvement if we used all available clusters.

Supplementary Information 2 Calibration of hyper-parameters

As in any model there are hyper-parameters that must be tuned to garner the best performance. Here we tune the architecture, the number of channels or colours in the input image, whether to augment the data and the Learning Rate of the model, and present the results of which can be found in Figure 2. For each test we split the available data in to a training sample (80% of the total amount of data) and a validation set (20% of the total amount of data). We initially train on three dark matter models - Fiducial Cold Dark Matter, SIDM0.1 (i.e. $\sigma_{\rm DM}/m = 0.1{\rm cm}^2/{\rm g}$) and SIDM1.0 (i.e. $\sigma_{\rm DM}/m = 1.0{\rm cm}^2/{\rm g}$). We ensure for both the training and test set that there is an equal number of each class represented, and that there are equal number of clusters from each redshift slice, to ensure we do not bias the training sample. We optimise using stochastic gradient descent (ADAM) and since this is a classification problem we use the Sparse Categorical Cross-entropy as the loss function. We monitor the accuracy and validation accuracy as the main metric of success. We carry out four key tests -

1. Figure 2, top left : The dependency of the validation accuracy with the architecture of the CNN. We take three example architectures: a simple CNN consisting of 5 layers and $1,203,747$ free parameters; BIPARE, a model taken from [1], consisting of $4,185,156$ parameters and finally Inception also from [1] consisting of $13,462,756$. We find that there is a mild improvement in the accuracy between the three models. The simple model although doing much worse is significantly quicker to train, whereas the time between BIPARE and Inception is negligible despite Inception consisting of three times as many parameters. We also find that the scatter in the validation accuracy is much lower than for the



Supplementary Figure 2: Four key tests associated with our CNN. In each case we show the validation accuracy as a function of the training epoch *T*op Left, Architecture: We test three models with increasing complexity. We find that the simple model (red) does not perform as well as the more complicated Inception (blue) and BIPARE (red), however the simple model does have a lower scatter and is more stable. *B*ottom Left, Data Augmentation: We show the validation accuracy both with data augmentation (blue, random rotation and random flip) and without (red). We see the significant improvement in the accuracy of the model with data augmentation. *T*op Right, Input Channels: We show the model performance for different combinations of input channels. We find that in this situation, total matter and stellar matter contribute the most information with X-ray containing almost none in deciphering between dark matter models. *B*ottom right, Learning Rate: We optimise the learning-rate (LR) meta-parameter for the Inception model. We find the best fit LR is LR=10^-3.

Inception and BIPARE. This will be due to the low training volume spread out over larger number of free parameters. We therefore do two things. The first is we use the simple model to carry out tests on the data where architecture is not important (for example input channels and data augmentation). This speeds up the tests. Whereas any tests directly on the CNN (for example learning rate), we use the Inception model. We choose this model for one main reason, that is although the improvement is negligible and time to train is comparable to BIPARE, by choosing Inception we retain consistency between this study and [1], which uses the Inception architecture.

2. Figure 2, bottom left : The dependency of the validation accuracy with the data augmentation. Data augmentation is a useful way to increase the size of the available training set. It was used for the classification of galaxies [2] and we adopt it here. This figure shows the improvement in the validation accuracy if we include a random rotation (where we reflect the true values at the border where there exists no data) and a random flip. We find that the accuracy improves by almost 20%. Clearly showing the need for this augmentation. We do also test other forms of augmentation, for example, random cropping, zooming and contrast. However, these all do not improve the network. This is expected, since rotating and flipping is physically invariant, whereas cropping, contrasting and zooming actively alter the cluster, (making it less or more concentrated, changing the ratios between channels). Therefore this acts to remove information from the dataset, not enhance it.
3. Figure 2, top right : The dependency of the validation accuracy with input channel. We test to see how much information is in each channel of the CNN. Often image classification algorithms use the standard three RGB channels of images. In this case we use three classic astrophysics channels : the

distribution of total matter (from weak gravitational lensing), the distribution of stellar matter (from optical imaging), and the X-ray emission maps. We find that total matter alone is able to achieve $> 75\%$ in accuracy with stellar matter improving the performance by $\sim 5\%$. We find adding X-ray does not significantly improve the algorithm.

4. Figure 2, bottom right: The dependency of the validation accuracy with the learning rate (LR) of Inception module. We show in this case the validation accuracy as the solid line and the dotted line the training accuracy to show that the model is not over-fitting. During back-propagation, the derivation of the Cost function with respect to the weights is calculated. At each epoch in training the weights are updated according to the calculated gradient. This "update" is regularised by the learning rate. A large learning rate with modified the weights in the direction of the gradient in a large way, before recalculating. The advantages of this is that it can learn quickly and will not get stuck in local minima. However, it can also not converge and never find the best fitting value. Whereas a small learning rate may never reach the best fitting value and may get stuck in local minima. This parameter is therefore extremely important. We test seven learning rates, in equal log-space. We find that $\log(LR) = -3$ gives the optimal validation accuracy and will therefore use this going forward. It should be noted these tests are carried out directly on the Inception model since the LR is dependant on the number of free parameters.

Increasing the sample size with transfer learning

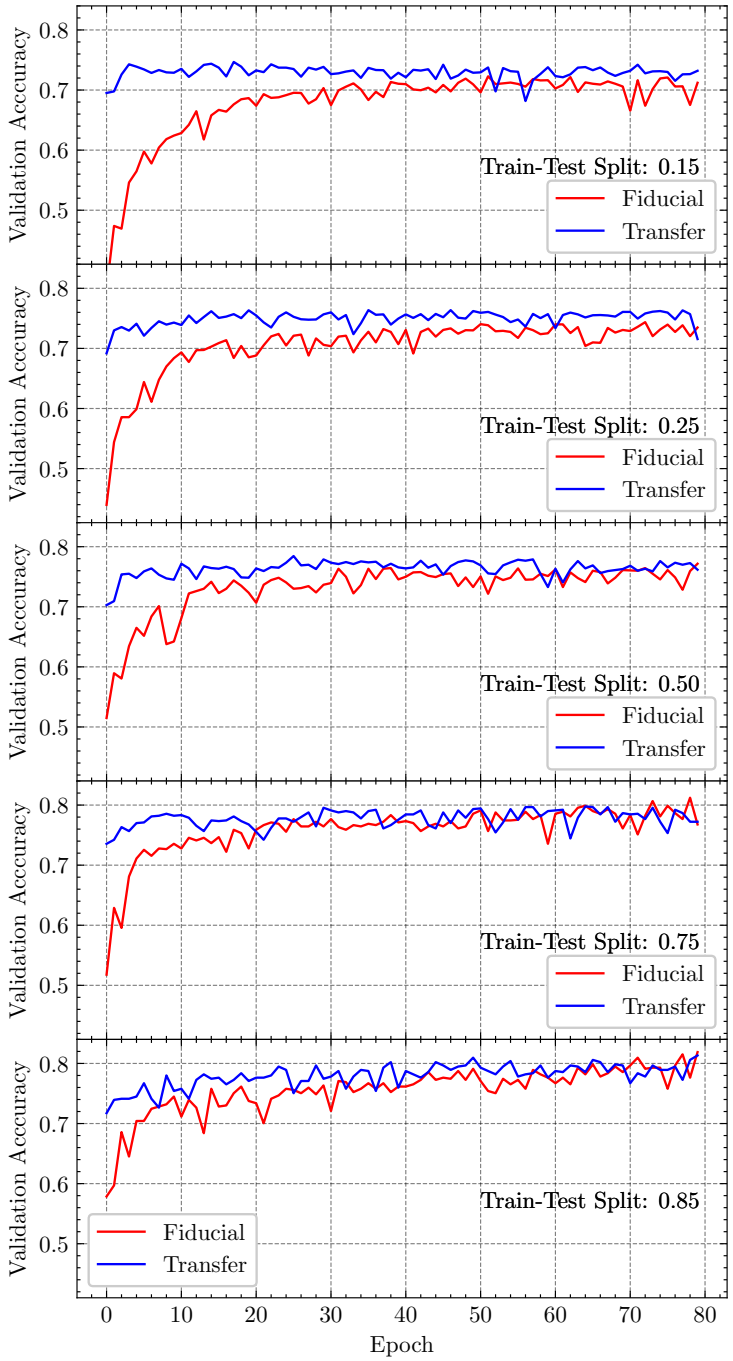
Given the small training set, we look at the possibility of using transfer learning to bolster the sample size. The idea behind transfer learning in this scenario is to train a model on a set of data to get estimates of the CNN weights, then to either freeze these weights and fine-tune an additional top layer to the model using a smaller training set or to fine-tune all the weights of the model.

The advantages of transfer learning is that if the training set is computationally expensive to produce (for example simulating galaxies), then we can simulate a simpler training set for the initial model (for example a dark matter only simulation), and then fine-tune on a smaller sub-sample of the more expensive data. We carry out these test on our data. We first train our model using a suite of dark matter only (DMO) simulations.

The slight complication here is that the final model will use two input channels (total matter and stellar), therefore the initial training on the DMO must also contain two input channels, which by definition is impossible since they only contain dark matter. We therefore emulate the stellar particles in the DMO with re-scaled collisionless particles. In other words, for CDM DMO we have two channels, dark matter and dark matter multiplied by the ratio of the cosmological dark matter density and the baryon density (Ω_M/Ω_B). For the interacting models, we use collisionless particles from the corresponding CDM simulation, i.e. SIDM1 dark matter plus CDM times the same ratio. This gives some estimate of the two channels. We then train these on the entire set of DMO simulations and then fine-tune them (without freezing any weights). We look at the impact of this as a function of the training set size. Figure 3 shows the results. Each panel shows the impact of transfer learning for a different final training sample size. For example, the top panel has a relatively small training set (15% of the total simulations). The red line shows the fiducial model and the blue line shows the model first trained on DMO simulations and then fine-tuned with the same training set as the fiducial model. We find for small training sizes transfer learning clearly improves the model. However at larger training sample sizes, this difference seems to disappear. It is difficult because with a larger training set, the test set is smaller and hence has more scatter. Indeed it also seems in the larger training sets that both the pre-trained model and the fiducial take the same number of epochs to reach the final accuracy, hence showing no improvement in training time. We therefore note this potentially interesting feature of training on cosmological models in low training sample sizes. However, for the training sets used here (0.85), we find no difference and therefore choose to train each model without the use of DMO simulations.

Supplementary Information 2.1 Tests that showed no improvement.

In a bid to improve the performance of our CNN we looked in to adding different pieces of information, however in each case they did not improve the model, only added more complexity. We briefly outline them here for completeness.



Supplementary Figure 3: Impact of transfer learning: We look at the improvement of initially training our model on dark matter only simulations and then fine-tuning on hydro-sims. We show the results as a function of size of training sample. We find in the limit of small training sample sizes (top panels), that initially training on DMO simulations and fine-tuning on cosmo-sims (blue line) improves the performance of the model over the fiducial training method (red line) by $\sim 5\%$. However, as the training sample sizes grow (bottom panels) this improvement reduces, and becomes negligible for training sets of $> 75\%$.

1. Added information - redshift and X-ray concentration. We postulated that adding information on the redshift and X-ray concentration (a proxy for the dynamical nature) of the cluster might help break some degeneracies with respect to each dark matter model. However, this did not improve the model. Indeed we find that the model does not perform better on relaxed or merging clusters when

trained on the entire population. This shows that in the future observational sample selections that may preferentially select one type of cluster will not impact the parameter inference.

2. Pre-trained models (e.g. ResNet) - we trialled taking different pre-trained models (for example ResNet [3] and EfficientNET [4]) and "fine-tuning" them. That is we froze the pre-trained weights of the model and simply optimised a final layer with our specific training set. We found these models exhibited significantly lower performance. This primarily came from interpolating the maps to the input shape that is required of these architecture. As such we decided to move away from these pre-trained models.

References

[1] Merten, J. *et al.* On the dissection of degenerate cosmologies with machine learning . *Mon. Not. R. Astron. Soc.* **487**, 104–122 (2019).

[2] Dieleman, S., Willett, K. W. & Dambre, J. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Mon. Not. R. Astron. Soc.* **450**, 1441–1459 (2015).

[3] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv e-prints* arXiv:1512.03385 (2015).

[4] Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv e-prints* arXiv:1905.11946 (2019).