APPENDIX A: Benchmark of the network on artificial data

Here we provide some additional information on the image phase space in which our master model is performing thoroughly. For that, we first analyze the absolute brightness and noise levels of our data for skyrmions and background by a gaussian fit $g(x) = \mathrm{CONST} \cdot \exp(-1/2 \cdot ({}^{x-\mu}/_{\sigma})^2)$ and obtain μ_{SK}, μ_{BG} and the noise strength $\boldsymbol{\sigma_{noise}} = {}^{\boldsymbol{\sigma_{SK}} + \boldsymbol{\sigma_{BG}}}/_{\mathbf{2}}$ per image. Fig.8 a)-d) is showing the process examplatory. Fig. 8 e) plots the standard deviation of the noise, to check that the noise level does not vary between classes. Fig. 8 f) shows the histogram of the average noise standard deviation in the sets.

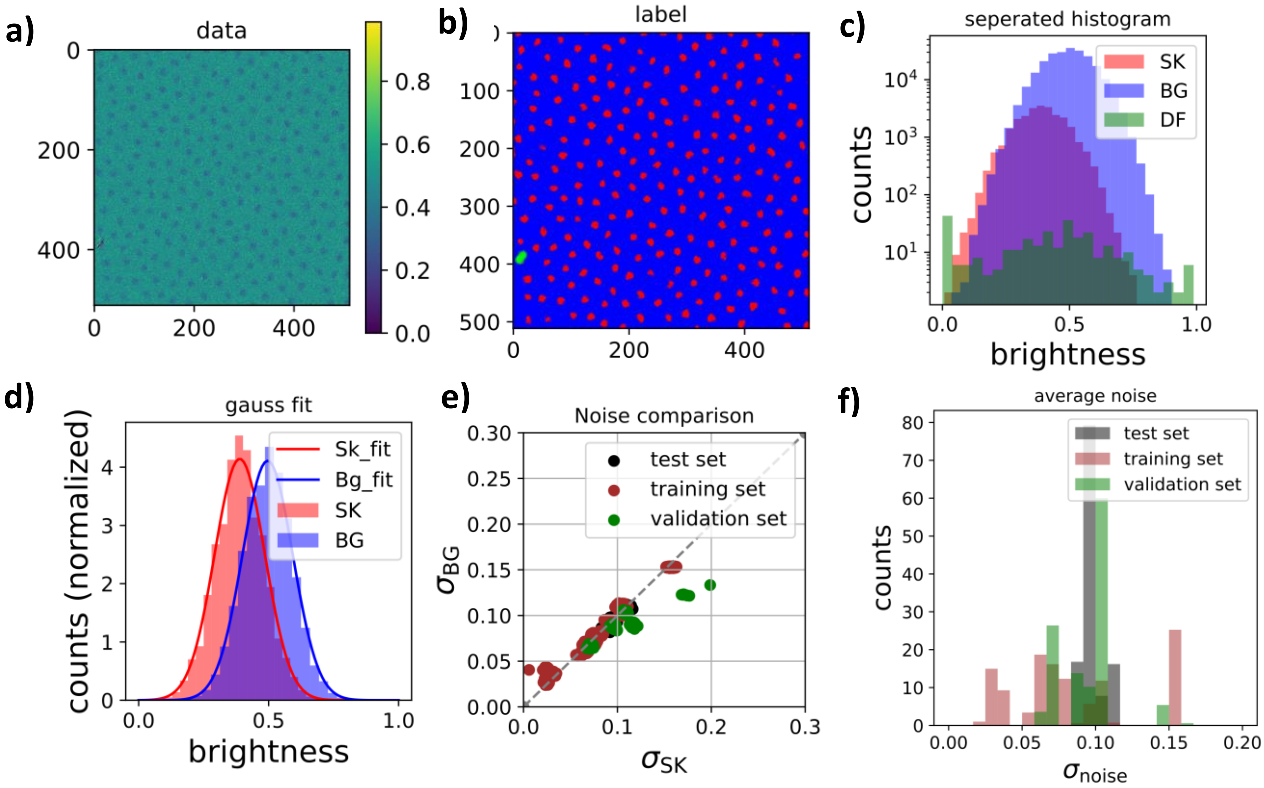

FIG. 8: Analysis of the datasets.a) -d) shows the evaluation exemplatory on one image: The true data a) is splitted based on their labels b) and plot the histograms classwise (SK skyrmions, BG background, DF defects) in c). We omit the defect class for the further analysis. SK & BG classes are normalized and gauss-fitted d). We compare noise levels of all images (classwise) in e) and find that the noise-level can be assumed class-independent. We define the noise level $\boldsymbol{\sigma_{noise}} = {}^{\boldsymbol{\sigma_{SK}} + \boldsymbol{\sigma_{BG}}}/_{\mathbf{2}}$ and show the noise histogram of our datasets in e).

As a next step, we generate artificial skyrmion data with the finite-size approach. The mask is generated with the following rule: $\boldsymbol{if\ m_z < 0.0: v_{mask} := 0, else\ v_{mask} := 1}$. We then sample a contrast level for skyrmions and background, such that the skyrmion level $\mu_{SK} < \mu_{BG}$, as we decided to have the skyrmions always darker than the background. We then add gaussian noise, sampled in the range that appears in the dataset – see Fig.8 f) -, clip values to $\boldsymbol{v_{mask}} \in [0,1]$ and obtained our artificial MOKE data. The defect class is omitted for this analysis, as no easy description for this class can be derived. Fig. 9 visualizes the process of producing the artificial data exemplary.

We have now benchmarked the master model that performed best on the validation set (*master model #1*) with1000 artificially generated samples and for different noise levels. The results can be seen in Fig. 10. The blue region corresponds to a high MCC and therefore a good classification of the (artificial) data.

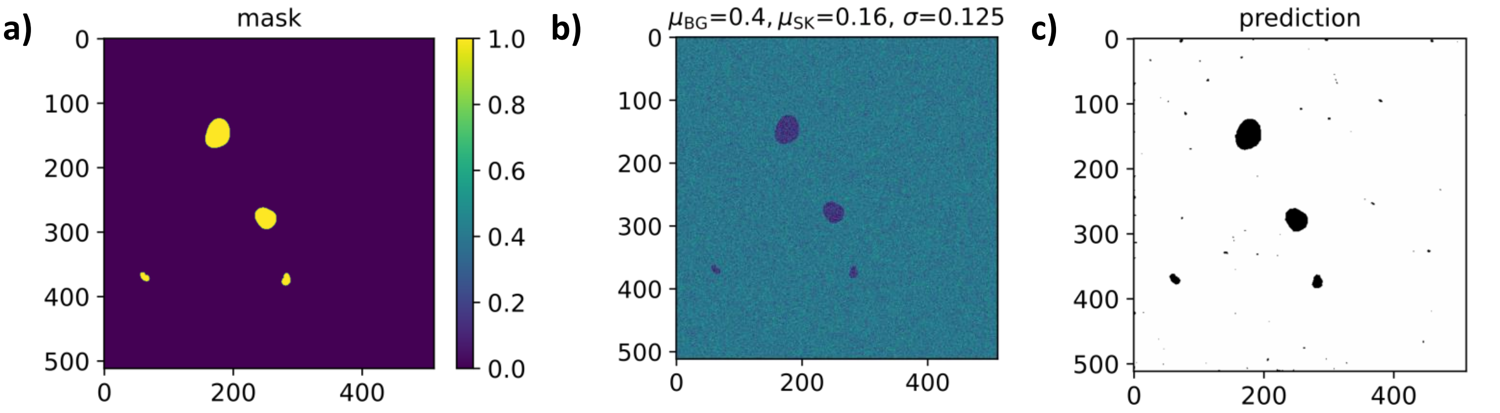

FIG. 9: The computer-generated skyrmion mask a) is varied, by changing skyrmion pixels to a sampled μ_{SK}, background pixel to a sampled μ_{BG}, such that $\mu_{SK} < \mu_{BG}$. We then add gaussian noise in a realistic range according to Fig. 8 f), $\sigma_{noise} < 0.2$, and obtain our artificial MOKE image. The noise level is digitized to values visible in Figure 10. The prediction c) is used to calculate the MCC for Figure 10.

We find that the network performs best when the skyrmions are only slightly darker than the background. This is reasonable, as our data is considered to be not easy separable. If the peaks of the two classes would be more distinct, one could perform the segmentation with an much more straight-forward approach. The additional black, green and brown points corresponds to the test, validation and training data, while the noise level from Fig.8 f) was digitized to the nearest value available in the sampling. We find that the network has good performance in the area of the phase space where our datasets are located, which is an indicator for a successful training. Fig. 11 shows the performance of the network on artificial data for fixed contrast values, but varying skyrmion sizes. We find that a minimum skyrmion size of ~10 pixels are needed for a reliable prediction. This can however vary with the contrast and noise level.

This analysis of the predictive power is useful for users to check if their data is in an area of good performance. If this is not the case, they can either a) shift the brightness level of the data, b) retrain the network with artificial or new data in the desired image phase space.

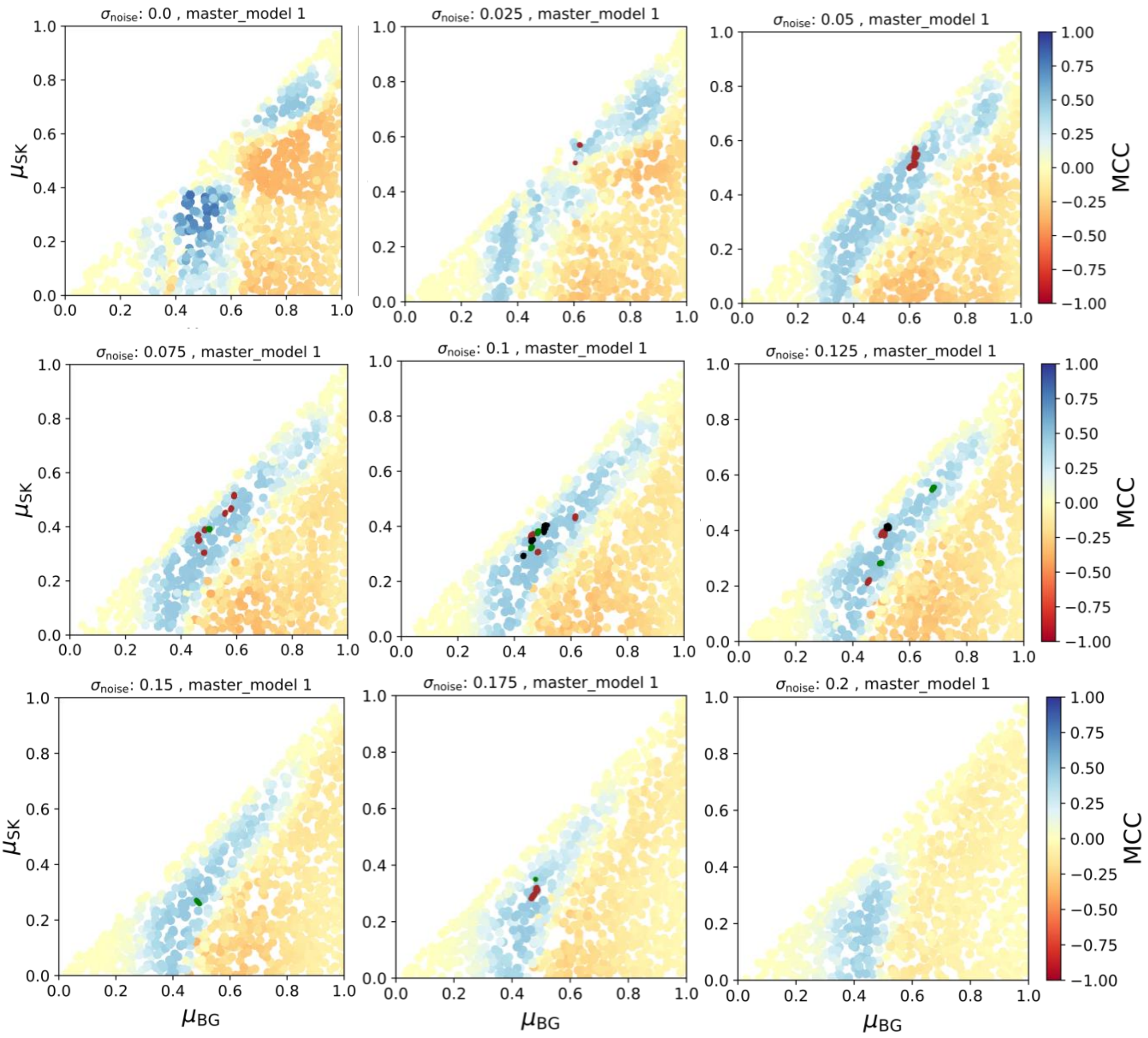

FIG. 10: Performance of the master model #1 on artificial skyrmion data, parametrized by μ_{SK}, μ_{BG} and the noise strength σ_{noise}. We find good performance in areas close to our training and validation set, and also in regions where μ_{SK} is only slightly smaller than μ_{BG}. We also plot the training (brown), validation (green) and test (black) set into the graphs, parametrized as shown in Fig. 8.

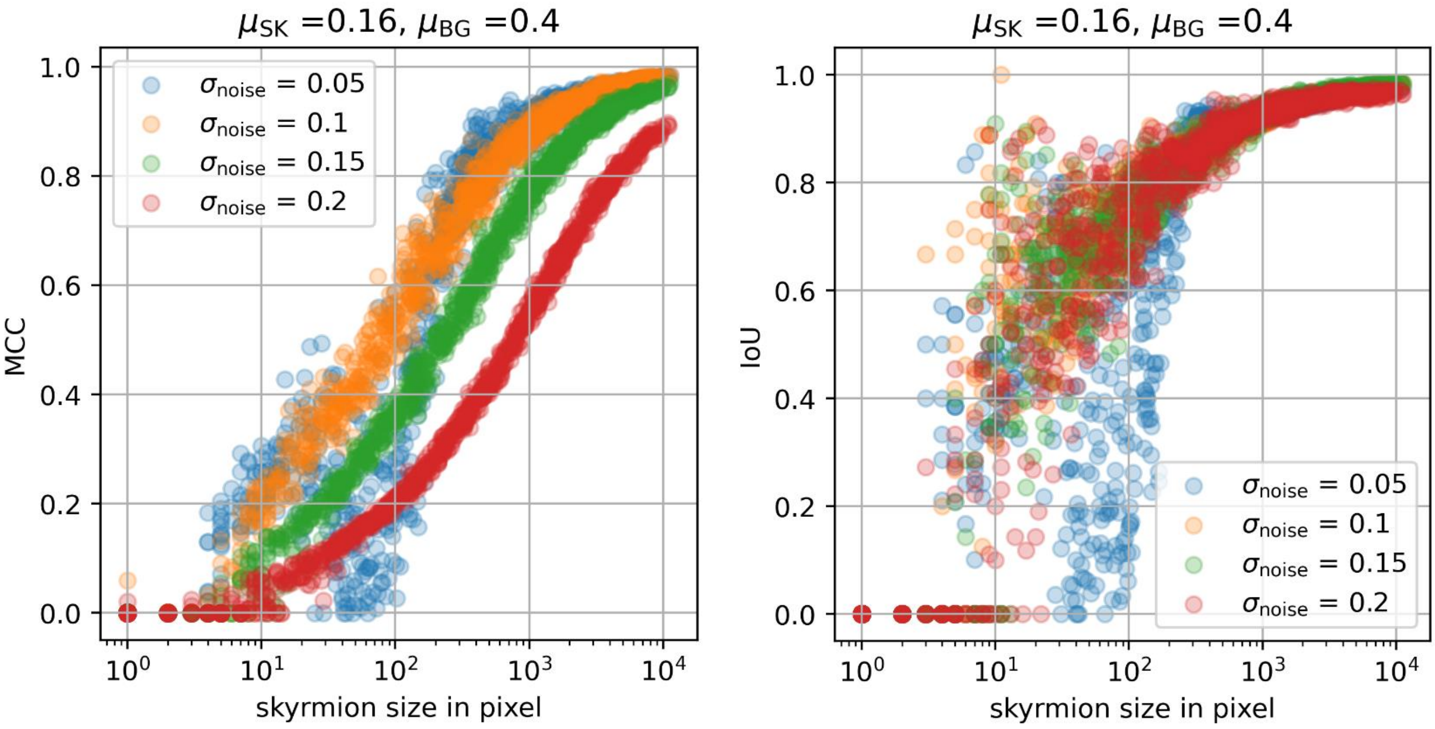

FIG 11: Artificial dataset with 1000 generated images with one skyrmion in each image generated. We plot a) the MCC and b) the Intersect over union (IoU) for the skyrmion predictions that totally or partially overlap with the label. We find that with increasing the noise the prediction becomes worse, and that one needs approximately a minimum skyrmion size of 30 pixels to obtain reliable prediction.

APPENDIX B: Material stacks and data acquisition

The data used for this study was collected from two sources: A huge part was collected from data archives, containing measurements which were performed to analyse or characterize material stacks or during the performance of experiments in the last decade in the AG Kläui lab. The used stacks were mainly Ta/CoFeBe/Ta/MgO/ stacks. Table 3 provides the material stack and its variety in the layer thicknesses within the data set. Typical skyrmion size in these type of stacks ranges from $\sim 500\ nm$ up to $\sim 5\ \mu m$.

Material	**Thickness range used**	**Effect of the layer**
Ta	5 nm	DMI generation
$Co_{60}Fe_{20}B_{20}$	0.8-1 nm	FM
Ta	0.07-0.09 nm	Dusting layer to tune anisotropy
MgO	0.8-1 nm	PMA generating layer
Ta	5 nm	Capping (prevent oxidation)

TABLE 3: receipe of a typical single layer stack used in the dataset.

The second part of the data, the defect set, was recorded especially for the purpose of the training. The stacks were also very similar to the ones aboves but focussing on parts of the sample with impurities and defects.

We would like to emphasize that no information of the material stack were included in the training, but the segmentation is only based on the image input. So, if a particular layer might decrease the absolute contrast of the sample, we do not use this information. Also, the absolute (true) skyrmion size is not relevant, as for each measurement different lenses and different optical zoom was used, leading to magnification factors between 1 and 400. The skyrmion size in pixels of our test and validation set is histogrammed in Figure 6. With that knowledge one can check if the network is suitable for other experimental skyrmion data

The advantage of this approach is that the network can be easily adapted to other material stacks and absolute skyrmion sizes and measurements using other magnetic microscopy techniques, as long as the contrast level is in the blue areas of Figure 10. If the prediction is not as desired, our model weights can be used as a pre-trained model to train on new data.