A&A, 682, A132 (2024)
https://doi.org/10.1051/0004-6361/202346826
© The Authors 2024

Astronomy & Astrophysics

The eROSITA Final Equatorial-Depth Survey (eFEDS): A machine learning approach to inferring galaxy cluster masses from eROSITA X-ray images

Sven Krippendorf[1,2], Nicolas Baron Perez[3,4], Esra Bulbul[3], Melih Kara[5], Riccardo Seppi[3], Johan Comparat[3], Emmanuel Artis[3], Yunus Emre Bahar[3], Christian Garrel[3], Vittorio Ghirardini[3], Matthias Kluge[3], Ang Liu[3], Miriam E. Ramos-Ceja[3], Jeremy Sanders[3], Xiaoyuan Zhang[3], Marcus Brüggen[4], Sebastian Grandis[6], and Jochen Weller[1,3]

[1] Universitäts-Sternwarte, LMU Munich, Scheinerstr. 1, 81679 München, Germany
e-mail: sven.krippendorf@physik.uni-muenchen.de
[2] Arnold Sommerfeld Center for Theoretical Physics, LMU Munich, Theresienstr. 37, 80333 München, Germany
[3] Max-Planck-Institut für extraterrestrische Physik, Gießenbachstraße 1, 85748 Garching, Germany
[4] Hamburg Observatory, University of Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany
[5] Institute for Astroparticle Physics, Karlsruhe Institute of Technology, 76021 Karlsruhe, Germany
[6] Institute for Astro- and Particle Physics, University of Innsbruck, Technikerstr. 25, 6020 Innsbruck, Austria

Received 5 May 2023 / Accepted 15 October 2023

ABSTRACT

We have developed a neural network-based pipeline to estimate masses of galaxy clusters with a known redshift directly from photon information in X-rays. Our neural networks were trained using supervised learning on simulations of eROSITA observations, focusing on the Final Equatorial Depth Survey (eFEDS). We used convolutional neural networks that have been modified to include additional information on the cluster, in particular, its redshift. In contrast to existing works, we utilized simulations that include background and point sources to develop a tool that is directly applicable to observational eROSITA data for an extended mass range – from group size halos to massive clusters with masses in between $10^{13}\,M_\odot < M < 10^{15}\,M_\odot$. Using this method, we are able to provide, for the first time, neural network mass estimations for the observed eFEDS cluster sample from Spectrum-Roentgen-Gamma/eROSITA observations and we find a consistent performance with weak-lensing calibrated masses. In this measurement, we did not use weak-lensing information and we only used previous cluster mass information, which was used to calibrate the cluster properties in the simulations. When compared to the simulated data, we observe a reduced scatter with respect to luminosity and count rate based scaling relations. We also comment on the application for other upcoming eROSITA All-Sky Survey observations.

Key words. methods: numerical – galaxies: clusters: intracluster medium – large-scale structure of Universe – X-rays: galaxies – X-rays: galaxies: clusters

1. Introduction

Improving our understanding of the mass function of galaxy clusters enables us to improve our inference with respect to key cosmological parameters. These parameters include Ω_M, the density parameter of matter in the Universe, and σ_8, which describes the dispersion of linear density fluctuations. The ongoing eROSITA (extended ROentgen Survey with an Imaging Telescope Array) All-Sky Survey (Predehl et al. 2021) on board the Spectrum Roentgen Gamma mission (Sunyaev et al. 2021) will provide us with the largest intra-cluster medium (ICM) selected galaxy clusters to date, which promises to provide tight constraints on cosmology through cluster abundance measurements (Merloni et al. 2012). A key ingredient in this analysis is to understand the cluster masses associated with a selected underlying sample (Bulbul et al. 2019). Traditionally this is performed with weak-lensing (WL) calibrated scaling relations in the context of the eROSITA cluster census or using dynamical mass measurements (Mamon et al. 2013; Old et al. 2014, 2015) in situations where the data allow for this approach. In the context of eROSITA, the former procedure has been demonstrated on the Final Equatorial Depth Survey (eFEDS) using the Hyper-Supreme Camera WL mass measurements (see Bahar et al. 2022; Chiu et al. 2022).

Cosmology analyses through cluster abundances detected in the X-ray or SZ surveys heavily rely on the availability of external WL mass measurements (Mantz et al. 2015; Bocquet et al. 2019; Grandis et al. 2019). This procedure requires the knowledge of cluster masses through WL surveys and introduces bias and scatter in the final cosmology contours if the survey data are not deep enough. Unaccounting for these biases and selection differences may affect the final cosmology measurements (Ramos-Ceja et al. 2022). Recently, applications of new machine learning (ML) tools and methods on large astronomy data and numerical simulations presented a promising method for reducing scatter in such cluster mass calibration using X-ray images (see Ntampaka et al. 2019; Green et al. 2019; Yan et al. 2020), SZ Compton y-maps (Cohn & Battaglia 2020; Wadekar et al. 2023a,b), and using optical data (Ntampaka et al. 2015; Ho et al. 2019, 2021, 2022; Kodi Ramanah et al. 2020).

In this work, we present a method that avoids the explicit knowledge of these WL measurements by using X-ray data and the redshift of clusters. In spirit, this is the same approach as using existing scaling relations on a new cluster sample. To calibrate (or, to put it differently, to train our ML model) we use simulations and the accuracy of these methods is determined by the cluster model in the training data. To reliably apply this method on new observations, we focus on training our ML model with a realistic cluster sample, namely, simulated clusters which represent our knowledge on clusters based on previous observations and represent the observational setting. In comparison to standard scaling relations, this ML model is more flexible as it can combine different features in a non-linear model. Furthermore, we consider models that utilize most of the information available, including the observation's energy and spatial information, rather than preprocessed features such as the luminosity (profiles) of a galaxy cluster. Given the success in other domains with similar data structures (e.g., in computer vision tasks on images Krizhevsky et al. 2017), a natural candidate for such models are convolutional neural networks (CNNs). The potential of these methods for estimating galaxy cluster masses has been previously demonstrated in Ntampaka et al. (2019), where a reduced mass scatter compared to luminosity-based methods was reported. In this work, we modify these methods to address a cluster sample at a larger redshift ($0.01 < z < 1.3$) and mass range ($10^{13} < M/M_\odot < 10^{15}$). Additionally, we account for emissions from other X-ray sources, for instance, active galactic nuclei (AGN), which are major contaminators in cluster analyses. Here, we present a method where additional filtering for such point sources is not required.

Finally, our neural network (NN) method incorporates a measure of uncertainty alongside the respective mass prediction. To estimate the uncertainty, we assume that the logarithm of our cluster masses is distributed according to some underlying Gaussian distribution with an associated mean and standard deviation. Both can be inferred using the log-likelihood associated with this normal distribution (cf. Sect. 3 for a detailed description). In addition, to account for the model uncertainty of our NN, we use a frequentist ensemble approach for our final mean and standard deviation. We train and validate our method on simulations of eROSITA galaxy clusters dedicated for eFEDS observations (Comparat et al. 2020; Liu et al. 2022b; Seppi et al. 2022).

This allows us to apply our method on the eFEDS cluster sample (Liu et al. 2022a) and provide, for the first time, the ML mass estimates on cluster observations. When comparing the performance of our mass estimates with those obtained from WL-calibrated scaling relations using count rate measurements (Chiu et al. 2022) on the simulations we find a reduced scatter. Our results on simulations are of similar scatter as using idealised luminosity information.

The paper is organized as follows: In Sect. 2, we describe the respective data products used in this work. Section 3 describes our machine learning approach and we discuss the results of our numerical work in Sect. 4. Our conclusions are presented in Sect. 5.

Throughout this paper, our simulated observations are obtained using a flat ΛCDM cosmology close to that of the Planck Collaboration (Planck Collaboration VI 2020), with $H_0 = 67.74\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_\mathrm{m} = 0.308900$, $\Omega_\mathrm{b} = 0.048206$, and $\sigma_8 = 0.8147$ as described in Comparat et al. (2020). Our masses M_{500c} refer to the mass included in the region with a mean density of 500 times the critical density.

2. eROSITA X-ray and simulated observations

This section presents the data we have used to train and test our machine learning method. We restricted ourselves to the data corresponding to the performance verification mini-survey of eROSITA, eFEDS (Brunner et al. 2022), the data analysis pipeline (Liu et al. 2022b), and the corresponding eFEDS simulations (see Comparat et al. 2019 for the procedure on how AGNs were simulated, Comparat et al. 2020 how galaxy clusters were painted for $M_{500c} > 10^{13.7}\,M_\odot$, and Seppi et al. 2022 for the extension to lower masses $M_{500c} > 10^{13}\,M_\odot$). We comment on the extension of our method to the eROSITA All-Sky survey (eRASS) observations in our conclusions in Sect. 5.

2.1. eROSITA X-ray Images

The 140 deg^2 eFEDS field, designed as a performance verification survey, had a uniform depth of 2.2 ks (1.2 ks after correcting for vignetting) approximately equal to the depth of the final eROSITA All-Sky Survey (Brunner et al. 2022). In this field, a total of 542 cluster candidates were detected with an extent likelihood threshold larger than six and detection likelihood larger than five (see Liu et al. 2022a, for details). Of these 542 candidates, 477 galaxy groups and clusters were confirmed with the follow-up optical data with redshift measurements (Klein et al. 2022). The clusters detected in the point source catalog were excluded in this analysis due to differences in the selection criteria (Bulbul et al. 2022). We used the subsample of 463 optically confirmed clusters, which have WL-calibrated features between $10^{13}\,M_\odot < M_{500} < 10^{15}\,M_\odot$. This selection was applied as this corresponds to the mass range on which our networks are trained on, namely, the cluster sample from the eFEDS simulations subsequently described.

To create X-ray images, we used the eROSITA Standard Analysis Software System (`eSASS` Brunner et al. 2022), version `eSASSusers_201009`. The calibrated event lists were corrected for good time intervals, dead times, corrupted events and frames, and bad pixels. Images were generated in ten equally spaced energy intervals of 205 eV each in the soft band for the range 0.25−2.30 keV, using the `eSASS` tool `evtool`. Multiple energy bands were selected to maximize the information on the X-ray images, taking advantage of the superb soft sensitivity of eROSITA. We kept X-ray photons in a fixed square of 300 pixels (corresponding to 1200″) centered on the X-ray centroid identified by `eSASS`.

2.2. eROSITA Simulated Images

The mock observations used in this study have the same exposure depth and field area to match the eFEDS observations. A method developed in Comparat et al. (2020) and Seppi et al. (2022) is employed to generate the mock photons for our training, validation, and test sets. A full-sky dark matter-only simulation provides the halo sample. Based on the properties of the dark matter halos, the X-ray properties of the sources are impainted using a Gaussian process model, which has been fit using previous cluster observations. These properties are then used to generate a source list passed to the SIXTE software (Dauser et al. 2019), which outputs the survey mock photons. It is worth stressing that these include not only cluster photons but, in addition, also point sources.

Within the eFEDS simulation, 18 realizations of the same eFEDS field were created to have enough sources for statistical analysis (see Liu et al. 2022a). All realizations together contain 148 833 clusters, whereas a single realization contains approximately 8000 clusters. To train our neural networks on a representative sample, we restricted ourselves to the same thresholds for cluster selection used for the eFEDS catalog (`eSASS` software version `eSASSusers_211214`). This gave a final sample of 7947 clusters. We found that these selection criteria improved our ML performance compared to using more clusters utilizing the ones

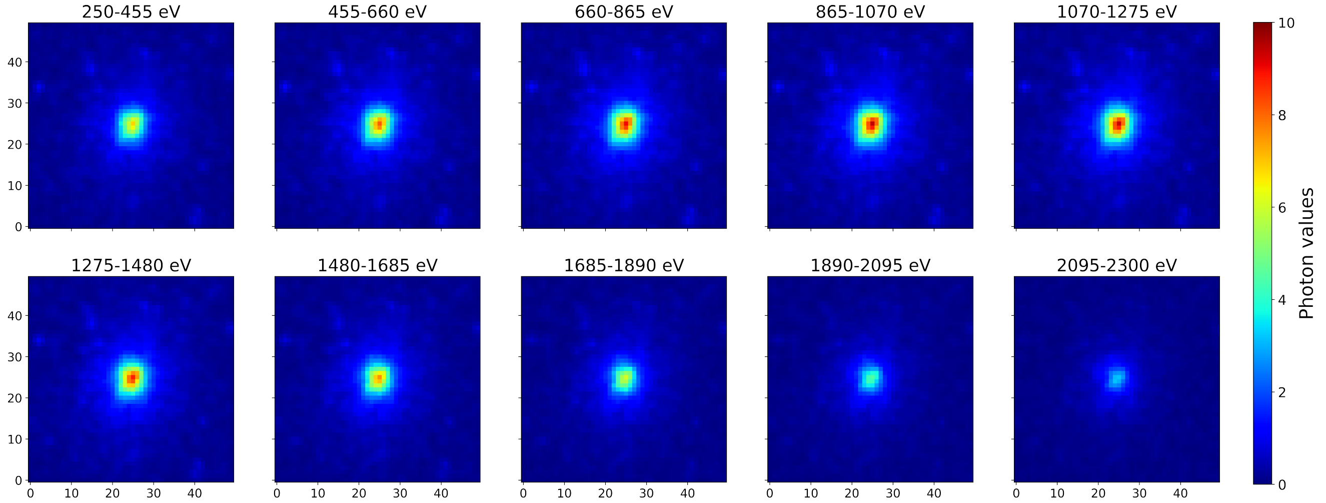

Fig. 1. Example X-ray image as input to our neural networks. All ten bands of the image of the galaxy cluster (SRC_ID: 10006566 from realization 5) of the *eFEDS* simulations. Each image has a dimension of 50 × 50. Due to our smoothing, the photon values are continuous. This cluster has a mass $\log(M_{500}/M_\odot) = 15$ and is located at a redshift of 0.11.

with smaller detection and extent likelihoods. As for the observations, we used X-ray photons in a fixed square of 300 pixels centered around the halo cluster center, noting no performance difference between the simulated cluster center and the eSASS detected center.

2.3. Neural network input datasets

These respective images were in a standard data format for images in machine learning processed as a three-dimensional (3D) array, where the first two dimensions carried the spatial information and the third dimension respectively carried the “color" information. We modified the images to make our machine-learning pipeline more efficient.

To render the input less sparse and to have fewer memory requirements, we scaled the boxes of size 300 × 300 × 10 down to a size of 50 × 50 × 10, and we applied Gaussian smoothing in all three directions, including the energy direction. The respective formula can be found in Appendix A.

We did not remove background photons or identified point sources, but we clipped the pixel number at 36 to avoid instabilities in our neural network training. An example of such an energy-band-image (EBI) can be found in Fig. 1. To resolve the ambiguity between a less luminous cluster at low redshift and a highly luminous cluster at high redshift, we also used the redshift information as input to our network. We optimized the spatial region and smoothing used for the EBI, ensuring that in almost all cases, the entire cluster is visible in the image. It is important to stress that these input images are independent of R_{500c} as such a selection would automatically include information about the cluster mass.

The respective mass and redshift distributions of clusters in eFEDS simulations and our `eSASS` selected sample are shown in Fig. 2. From all the realizations of the simulations, we end up with 7947 clusters from which we used 70% for our training, 15% for our validation, and 15% for our test set.

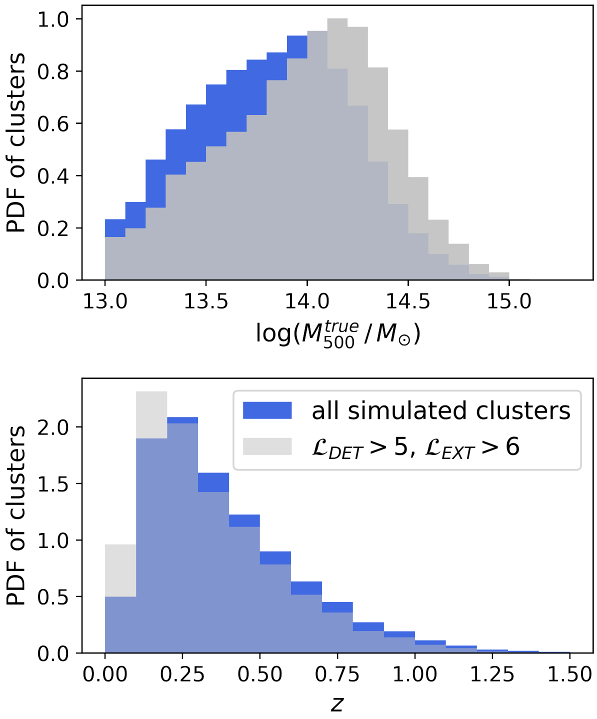

Fig. 2. Simulated cluster sample with and without the applied filters used for training and evaluation with their respective redshift and mass distribution. The total number of simulated and filtered clusters is 25 031 and 7947, respectively.

3. Machine learning method

As a proof of concept, we utilized a standard architecture using convolutional and pooling layers, followed by at least one dense layer. We provide the network with information about the source’s redshift at the first of these dense layers. We discuss the impact of uncertainty of this feature on the performance in Appendix C, which turns out to be irrelevant for our purposes. To avoid overfitting, we utilized preprocessing layers, which perform random rotations and flips during training, efficiently augmenting the training dataset. We found that this significantly improved performance. We leave a discussion on how other types of architectures, for instance, using geometric deep

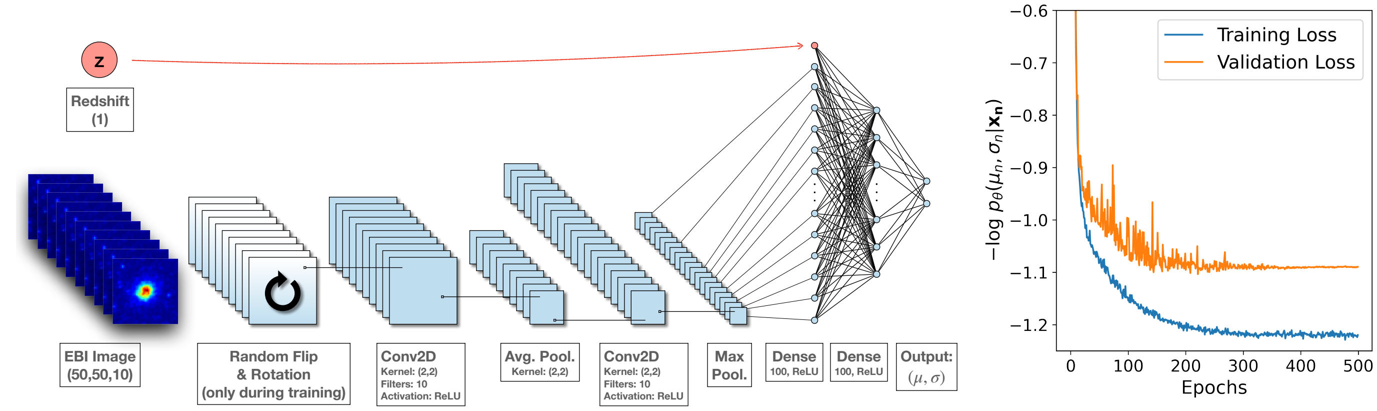

Fig. 3. Overview of our neural network architecture and training. *Left:* neural network architecture of our best performing model. We used a batch size of 100 and Adam as an optimizer with a starting learning rate of 10^{-4}. *Right:* training behavior of our network.

learning (see Bronstein et al. 2021), affect the performance for the future[1].

To enable stable training for a large variety of hyperparameters, we first trained our networks using several standard regression loss functions (e.g., mean squared error loss of the logarithmic mass). A few hundred epochs of training are typically sufficient. To assess the performance beyond the values for the losses, we check the scatter of masses on the respective training and validation sets for additional biases. More details on our choices and associated scans can be found in Appendix B.

In our hyperparameter scan, we identify several promising architectures and perform further analysis of these architectures. In particular, we train our network from scratch using a negative log-likelihood for each data sample to predict the mean and standard deviation of a Gaussian:

$$-\log\, p_\theta(y_n|\mathbf{x}_n) = \frac{\log\left(\sigma_\theta^2(\mathbf{x_n})\right)}{2} + \frac{(y_n - \mu_\theta(\mathbf{x_n}))^2}{2\sigma_\theta^2(\mathbf{x_n})} + \text{constant}, \quad (1)$$

where $\mathbf{x}_n$ denotes the data, y_n the data label, σ_θ, μ_θ are our predictions which depend on the neural network parameters θ. The relevant hyperparameters and the training curve for one of our well-performing models can be found in Fig. 3.

Finally, to address the systematic uncertainty in the neural network prediction, we opt for an ensemble method as presented in Lakshminarayanan et al. (2016) and leave Bayesian approaches for the future (Gal & Ghahramani 2015). In practice, we repeated the training procedure with N random weight initializations and the final predictions are calculated via:

$$\log \overline{M}_{500c}^{\mathrm{NN}} = \frac{1}{N}\sum_{i=1}^{N} \mu_{\theta_i}(x), \quad (2)$$

$$\sigma_*^2(x) = \frac{1}{N}\sum_{i=1}^{N} \sigma_{\theta_i}^2(x) + \mu_{\theta_i}^2(x) - \mu_*^2(x). \quad (3)$$

We now turn to a discussion of the results obtained using this approach for mass estimation.

4. Results

To analyze the performance of our neural networks on eFEDS simulations, we first compare the predicted and actual mass distributions in the simulations. As shown in Fig. 4, predictions on the test set and the scatter follow the ideal slope very closely. We observe a scatter of $\sigma = 0.188$ on the test set. Our mean error prediction is identical to this value with a mean error of $\langle\sigma_*\rangle = 0.188$. As shown in Fig. 5, we observe a bias for the mass range of $13.0 < \log M_{500}/M_\odot < 13.5$ where we are over-predicting the mass on average. In the mass range $14.5 < \log M_{500}/M_\odot < 15.0$ we are under-predicting the masses respectively. To interpret these biases, we performed two experiments:

To improve the quality of our training and test sample, we used a cluster sample that has a detection and extent likelihood larger than 60. This reduces the scatter to $\sigma = 0.159$ on the test set using the same likelihood cuts. In addition, we see a reduction of the bias for high-mass objects. Clearly this cut reduces the number of available clusters significantly (from 7947 to 1156 in total). In addition, it is very encouraging to see that our mean uncertainty also reduces to $\langle\sigma_*\rangle = 0.158$.

Furthermore, to change the number of clusters at high and low-mass respectively, we weigh our samples to effectively generate a uniform distribution in mass during training. This ensures, in particular, that the network is more strongly penalized when falsely predicting high-mass clusters. We find that the scatter is slightly increased but we reduce the bias for the high-mass clusters from −0.177 to −0.097 and for the low-mass clusters from 0.121 to 0.082. This is encouraging as we only know approximately the observed distribution of cluster masses and our method should be able to compensate for small differences in the distribution.

These respective scatters in the mass predictions have to be compared with the underlying probabilistic cluster model and the application of scaling relations. First of all, there is the intrinsic scaling relation in the data where in our case the luminosity-mass scaling relation has a scatter of $\sigma = 0.2$ (cf. Fig. 6 in Comparat et al. 2020) and the temperature-mass scaling relation which has a $\sigma = 0.07$. We see that the scatter in our method depends on the quality of the dataset, namely, when selecting clusters with high detection and extent likelihood, we reduced the scatter below the luminosity scaling relation. Next, when comparing our scatter with scaling relations, a natural caveat is whether the respective scaling relation provides similar results as a scaling relation which is calibrated on this cluster sample, for instance, using WL observables. To do this, we utilized scaling relations which have been calibrated for the eFEDS cluster sample. These scaling relations were calibrated using the three-year (S19A) WL data from the Hyper Suprime-Cam (HSC) Subaru Strategic Program survey, simultaneously fitting the count rate and shear profiles to obtain the best-fit scaling relations. We

[1] In this work, we utilize Keras (Chollet et al. 2015) and Tensorflow (Abadi et al. 2015) for our experiments.

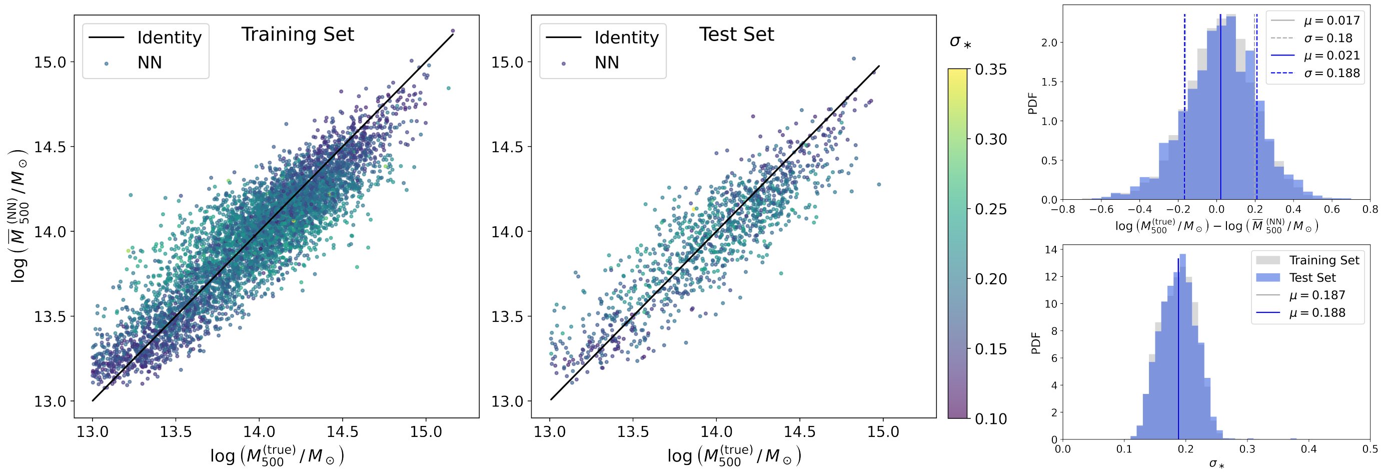

Fig. 4. NN on simulations: overview of mass estimation on eFEDS simulations using our ensemble of 30 convolutional neural networks trained with our likelihood loss from Eq. (1) (cf. Fig. 3 for the hyperparameters). *Left:* mass scatter between predicted mean masses and masses from the simulations on the training set. The colors indicate our predicted standard deviation. *Middle:* our mean mass predictions on the test set. *Right top:* distribution of our error $\log_{10}(\mu_*/M_\odot) - \log_{10}\left(M_{500c}^{\rm true}/M_\odot\right)$. *Right bottom:* distribution of our error estimates, which show a mean uncertainty of 0.189 on the test set.

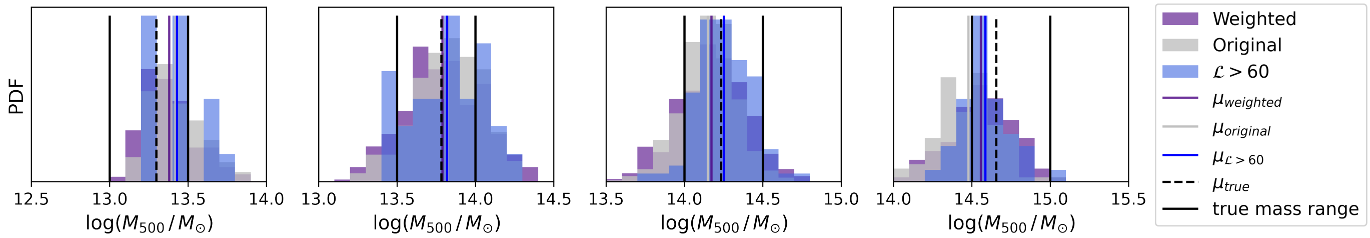

Fig. 5. NN biases on simulations: mass predictions on eFEDS simulation test sets for different mass ranges to evaluate respective biases between NNs trained on datasets with the three different datasets described in the main text.

refer the reader to Chiu et al. (2022) for a detailed description. When using the cluster luminosities and the luminosity mass scaling relations reported in Eq. (67) of Chiu et al. (2022), we recover a scatter of $\sigma = 0.197$ on our test dataset with detection likelihood larger than 5 and extent likelihood larger than 6, where we have used the actual luminosities in the simulation. This is comparable to the luminosity scatter in the simulation and 4.8% larger than the scatter we observe for our NN masses. When applying to the higher quality cluster sample, the scatter reduces to $\sigma = 0.186$ but is significantly above the NN scatter. To apply these scaling relations we have used an appropriate selection function based using the cluster luminosity and redshift, although we note only a small effect on the ensemble level. In this analysis, the luminosity is normalized with the factor:

$$N = \left[\frac{M_{500}}{M_{\rm piv}}\right]^{\left(\delta_{L_X}\ln\left[\frac{1+z}{(1+z_{\rm piv}}\right]\right)}\left[\frac{E(z)}{E(z_{\rm piv})}\right]^{C_{{\rm SS},L_X}}\left[\frac{1+z}{1+z_{\rm piv}}\right]^{\gamma_{L_X}}, \qquad (4)$$

where the evolution factor $E(z) = H(z)/H_0$, the pivotal mass $M_{\rm piv} = 1.4\times10^{14}\,M_\odot$ and the pivotal redshift $z_{\rm piv} = 0.35$. Moreover, the scaling relation parameters as calibrated in the analysis are $\delta_{L_X} = -0.07$, $C_{{\rm SS},L_X} = 2$ and $\gamma_{L_X} = -0.51$. For this scaling relation analysis, a fiducial flat ΛCDM cosmology was used with $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$, $\Omega_{\rm m} = 0.3$, $\Omega_{\rm b} = 0.05$, $\sigma_8 = 0.8$, and $n_s = 0.95$.

Furthermore, to provide an outlook on inference of cosmological parameters, the scatter is worse when using the count rate scaling relations on the data with a detection likelihood larger than 5 and extent likelihood larger than 6 with the measured count rate, where we find a scatter of $\sigma = 0.265$. We note that this is without applying the selection function, which (in light of the effect on the luminosity scaling relation) appears to have a minor effect in terms of changing the mass predictions for this sample. Such a selection function is currently not available for this sample. A more detailed comparison of our systematic uncertainties with systematic uncertainties appearing for the scaling relations between the count rate and WL mass as discussed in Grandis et al. (2021) is left for the future.

For both scaling relations we find a significant reduction in scatter. The amount of reduction depends significantly on the data used for training, this does not only dependent on the mass and redshift distribution.

One further advantage of the NN-based mass estimation is that the training networks use the full morphology information of the input clusters in the X-ray images (Ghirardini et al. 2022) compared to other methods and are not impacted by the line-of-sight structure or assumed 3D morphology of the source when estimating masses (ZuHone et al. 2023) or hydrostatic mass bias often a problem for X-ray mass measurements (Scheck et al. 2023).

We note that the predicted means and the respective standard deviations do not vary hugely on an ensemble level. In particular, we observe that the ratio of the individual σ-values for each network and the correspond ensemble prediction σ_* is given as $\langle\sigma/\sigma_*\rangle = 0.951 \pm 0.039$, where we quote the single standard deviation values and where we have averaged over all clusters in the test sample. On an individual level we report the clusters with the highest and smallest differences in the predicted masses (see

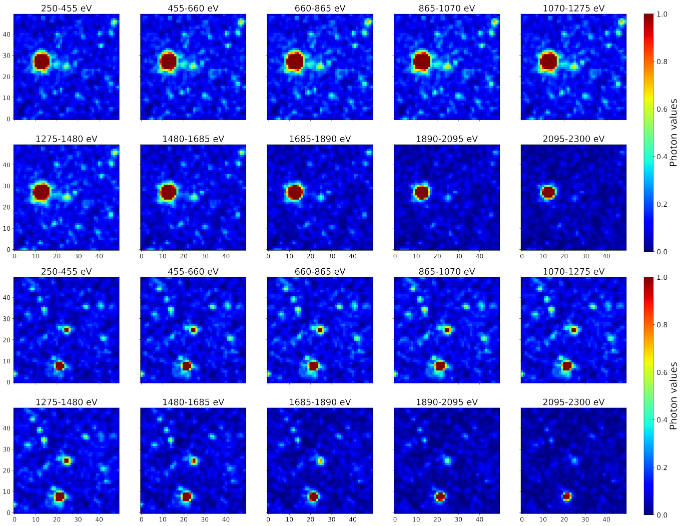

Fig. 6. All ten channels of the EBI of the galaxy cluster with the highest and lowest $\Delta\mu = \mu_{\max} - \mu_{\min}$, where $\mu_{\max}$ ($\mu_{\min}$) corresponds to the highest (smallest) predicted mass by the ensemble members. *Top:* this cluster, located as usual at the center of the EBI, has a mass of $\log M_{500}/M_{\odot} = 13.959$ and is at redshift $z = 0.211$ in the simulations (Object with SRC_ID: 10003763 from realization 18). The lowest and highest predicted mean values in our ensemble are $\mu_{\min} = 13.323$ and $\mu_{\max} = 14.587$, respectively. For illustration purposes, to make the actual cluster visible, we have clipped the photon values at 1. The final mass predictions are $\log M_{500}/M_{\odot} = 14.079 \pm 0.305$. *Bottom:* example of a cluster with little differences among the ensemble (SRC_ID: 10006556 from realization 9), it has a mass of $\log M_{500}/M_{\odot} = 13.873$ and is at redshift $z = 0.305$. We obtain $\mu_{\max} = 14.095$ and $\mu_{\min} = 13.994$ and our final prediction for this cluster is $\log M_{500}/M_{\odot} = 14.042 \pm 0.161$.

Fig. 6). We often find upon visual inspection that the largest differences in the predicted masses occur when other bright X-ray sources are present in the EBI and our cluster of interest is a less luminous source.

Having seen that our method provides sensible looking mass estimates on eFEDS simulations, we now estimate the masses for the eFEDS cluster sample with extent likelihood larger than 6 and detection likelihood larger than 5. We use the ensemble of neural networks which we have trained on data with the same selection criteria. We show the scatter between our NN predicted masses and the masses obtained using WL-calibrated luminosity scaling relations in Fig. 7. We use WL calibrated masses here because they are less affected by non-thermal astrophysical processes unlike hydrostatic masses. Additional cuts on the sample would be required to compare with clusters with hydrostatic masses from temperature measurements (cf., Bahar et al. 2022), which in addition is not accounted for in our selection function. Hence, we compare with the WL-calibrated masses, which are used for cosmology analysis from eROSITA cluster observations.

We observe that both predictions agree for clusters where our NN ensemble predicts a low uncertainty of $\sigma_* < 0.185$ (more visible points correspond to clusters with such a low uncertainty). We note that for the few clusters present in the eFEDS cluster sample with a mass below the range we have trained on, our neural network ensemble still predicts masses in the mass regime it was trained on and does not generalize for these data outside of the known regime.

Furthermore, to compare our predicted masses with the luminosity-mass estimates of the clusters, we show the scatter between the luminosity and our mass estimates on the right side of Fig. 7. Overall, we find a linear relation which is close to the slope identified via the WL-calibrated scaling relation. However, we find deviations from the WL-calibrated masses at high masses. Further analysis of the features being used by the NN ensemble, ultimately aiming at a data-driven scaling relation, is beyond the scope of this paper.

5. Conclusions

We have demonstrated that galaxy cluster masses can be estimated using NN ensemble predictions when applied to the eFEDS-field of eROSITA, both to the respective simulations and actual observations. Depending on the training data, we observe a significant reduction in scatter in comparison to luminosity based scaling relations from $\sigma = 0.186$ to $\sigma = 0.159$ on a sample with higher detection and extent likelihood and from $\sigma = 0.197$ to $\sigma = 0.188$ on the entire sample. Compared to count rate based scaling relations, the improvement is from $\sigma = 0.265$ to $\sigma = 0.188$. Our approach is applicable to clusters at different redshifts and we are not required to remove other clusters or point sources from the respective images to mimic a realistic observational set-up. Going beyond existing NN methods for cluster mass estimation, our method provides uncertainty measurements of the NN predicted masses for each cluster. Our ML approach can be integrated into a highly developed workflow for estimating cluster masses and their subsequent use for cosmological parameter inference. The interplay with each of these components is important to understand shortcomings and potentials for improved mass estimates in the future:

- WL and other additional measurements: given the dependence on our simulations of eFEDS clusters, our NN methods do not require in addition WL measurements for (a subset of) the X-ray selected cluster sample. Any constraints, for instance, for a subsequent cosmological analysis, arising from the requirement of availability of WL information can be circumvented.

  Our model can be easily expanded and improved by adding new features and observations, similar to using redshift information to the network. For instance, richness information coming from optical observations of clusters of galaxies, as being developed in the context of *Euclid*, promises to improve the NN-predictions in the high-mass end (e.g., Euclid Collaboration 2023). As in any other model, adding new multi-wavelength data requires appropriate calibration and will be used in future work.
- Simulations: as our ML approach is obtained from the underlying data model, it heavily depends on the data used for training. To make our method work, it is crucial that the training data are of sufficient quality. This requires, for instance, that the training data contains clusters in the appropriate mass regime and that the clusters in the training sample are ideally very close to the cluster sample the method is applied on. At this stage, a generalization beyond properties captured by the training data is not guaranteed. Throughout this project, we often encountered performance deterioration when including different cluster samples for training. Addressing the independence of the training data is a clear future goal but (as demonstrated here) can be circumvented by utilizing a dedicated training set. Implicitly, our method depends on the data used to shape the simulations and in particular on the underlying scaling relations. However, we crucially observe that the mass distribution of clusters in the training sample is not of high importance as showcased when

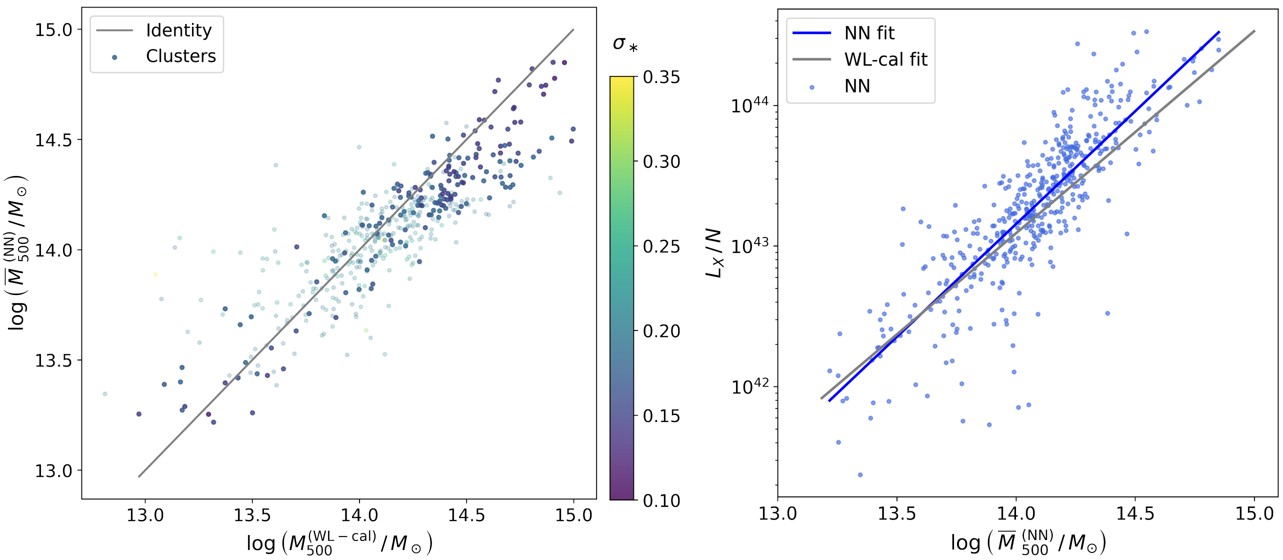

Fig. 7. NN on eFEDS observed data: comparison of WL calibrated mass estimates (using luminosity scaling relations as in Eq. (67) of Chiu et al. 2022) and masses obtained from our ensemble neural networks. *Left:* respective mass predictions on eFEDS clusters. The uncertainties on the mass predictions are color-coded and correspond to the NN uncertainties. *Right:* correlation between predicted masses on eFEDS observations and the measured luminosities as presented in Chiu et al. (2022).

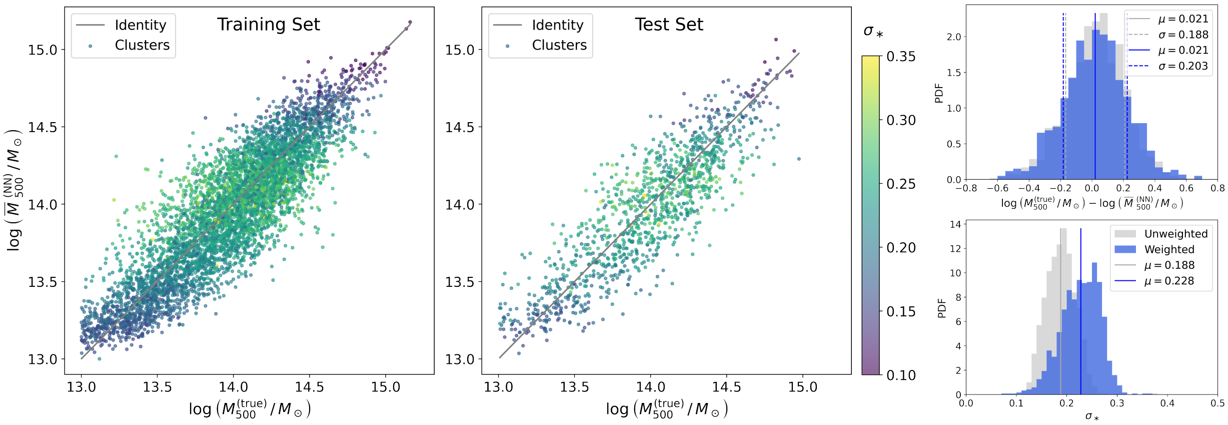

Fig. 8. NN trained on uniform cluster sample: overview of mass estimation on eFEDS simulations using our deep ensemble trained with class weights such that the effective mass distribution is close to uniform, illustrating the robustness to uncertainty in the underlying mass distribution. *Left:* mass scatter between predicted means and masses from the simulations on the training set. The colors indicate our predicted standard deviation. *Middle:* our mass predictions on the test set. *Right:* comparison of the residuals and distribution of standard deviations in our test set mass distributions between our normal training procedure with masses distributed as shown in Fig. 2 and our weighted clusters. The distribution of our error estimates shows a mean uncertainty of 0.228.

using a uniform distribution of masses (Fig. 8). This is particularly encouraging, as this allows for generalization across different mass distributions from different cosmologies.

– ML versus known astrophysical features: there are two approaches for predicting masses using ML; either using known astrophysical features (e.g., measured luminosities or count rates) as the input (cf., Green et al. 2019) or directly using the photon information. Here, we explore the latter and demonstrate that it provides competitive mass estimates. Future studies will provide more information on which method ultimately predicts the most accurate mass estimates. It would be very interesting to compare the ML features with previously identified features (e.g., using, appropriate dimensional reduction and symbolic regression; see Wadekar et al. 2023b for a work in this direction).

Finally, we summarize the advances from the ML-based mass predictions, presented here:

– We demonstrate, for the first time, that meaningful uncertainty measures can be provided with the mass estimates in X-ray cluster mass estimations with ML and, in particular, neural networks. This is a crucial requirement for integrating ML-based methods into cosmological analyses with cluster counts.

– As our simulations also include clusters with masses as low as $10^{13}\ M_\odot$, we are able to demonstrate for the first time that this neural network approach to X-ray cluster mass estimates also works in this mass regime without introducing large biases. A further successful extension to the low-mass regime would be very interesting and could dramatically increase the sample utilized for cosmology, we found that objects with a high detection and extent likelihood provide an avenue forward.

- Instead of a single channel that can only capture information about the total number of photons, we utilize an input format that also captures the energy information of photons. Our EBIs enable the neural networks, at least in principle, to utilize energy-dependent information such as the cluster temperature.

One of the immediate next objectives is to apply our method to other eROSITA cluster samples, in particular, the upcoming All-Sky Survey data. To make our method applicable to these observations, our pipeline follows the standard automatised pipeline for the detection of sources in eROSITA and the only change is that we need to ensure that the performance does not decrease due to the different exposure times for individual clusters in those samples, as the first All-Sky survey data are shallower than the eFEDS data used in this work. A further extension to observations of other X-ray telescopes (e.g., XMM, *Chandra*), despite being very interesting, would require dedicated datasets to train the ML method appropriately.

As this paper was in its final stage, the preprint Ho et al. (2023), which discusses a similar question, appeared on astro-ph. Our approach differs from Ho et al. (2023) by the use of the simulation data sets for training. The sample of simulated clusters used in this work represents the eROSITA cluster selection. Our method could thus be successfully and self-consistently applied to the eROSITA survey observations and compared with the observational WL mass measurements utilized for the same sample. Through this work, we also provide a clear path toward using ML-based masses in cosmological analyses. Additionally, we successfully utilize a likelihood loss for the first time, enabling uncertainty estimates, namely, a prerequisite for employing ML-based masses in future scaling relations and cosmology analyses.

Acknowledgements. This work is based on data from eROSITA, the soft X-ray instrument aboard SRG, a joint Russian-German science mission supported by the Russian Space Agency (Roskosmos), in the interests of the Russian Academy of Sciences represented by its Space Research Institute (IKI), and the Deutsches Zentrum für Luft und Raumfahrt (DLR). The SRG spacecraft was built by Lavochkin Association (NPOL) and its subcontractors and is operated by NPOL with support from the Max Planck Institute for Extraterrestrial Physics (MPE). The development and construction of the eROSITA X-ray instrument was led by MPE, with contributions from the Dr. Karl Remeis Observatory Bamberg & ECAP (FAU Erlangen-Nuernberg), the University of Hamburg Observatory, the Leibniz Institute for Astrophysics Potsdam (AIP), and the Institute for Astronomy and Astrophysics of the University of Tübingen, with the support of DLR and the Max Planck Society. The Argelander Institute for Astronomy of the University of Bonn and the Ludwig Maximilians Universität Munich also participated in the science preparation for eROSITA. The eROSITA data shown here were processed using the eSASS software system developed by the German eROSITA consortium. E.B., A.L., V.G., X.Z., and C.G. acknowledge financial support from the European Research Council (ERC) Consolidator Grant under the European Union's Horizon 2020 research and innovation programme (grant agreement CoG DarkQuest No 101002585). We used the following packages: Numpy (Harris et al. 2020), Tensorflow (Abadi et al. 2015), Astropy (Astropy Collaboration 2022), Pandas (McKinney 2010), Matplotlib (Hunter 2007), Scipy (Virtanen et al. 2020), Scikit-learn (Pedregosa et al. 2011).

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org
Astropy Collaboration (Price-Whelan, A. M., et al.) 2022, ApJ, 935, 167
Bahar, Y. E., Bulbul, E., Clerc, N., et al. 2022, A&A, 661, A7
Bocquet, S., Dietrich, J. P., Schrabback, T., et al. 2019, ApJ, 878, 55
Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. 2021, ArXiv e-prints [arXiv:2104.13478]
Brunner, H., Liu, T., Lamer, G., et al. 2022, A&A, 661, A1
Bulbul, E., Chiu, I. N., Mohr, J. J., et al. 2019, ApJ, 871, 50
Bulbul, E., Liu, A., Pasini, T., et al. 2022, A&A, 661, A10
Chiu, I. N., Ghirardini, V., Liu, A., et al. 2022, A&A, 661, A11
Chollet, F., et al. 2015, Keras, https://keras.io
Cohn, J. D., & Battaglia, N. 2020, MNRAS, 491, 1575
Comparat, J., Merloni, A., Salvato, M., et al. 2019, MNRAS, 487, 2005
Comparat, J., Eckert, D., Finoguenov, A., et al. 2020, Open J. Astrophys., 3, 13
Dauser, T., Falkner, S., Lorenz, M., et al. 2019, A&A, 630, A66
Euclid Collaboration (Bisigello, L., et al.) 2023, MNRAS, 520, 3529
Gal, Y., & Ghahramani, Z. 2015, ArXiv e-prints [arXiv:1506.02142]
Ghirardini, V., Bahar, Y. E., Bulbul, E., et al. 2022, A&A, 661, A12
Grandis, S., Mohr, J. J., Dietrich, J. P., et al. 2019, Mon. Not. Roy. Astron. Soc., 488, 2041
Grandis, S., Bocquet, S., Mohr, J. J., Klein, M., & Dolag, K. 2021, MNRAS, 507, 5671
Green, S. B., Ntampaka, M., Nagai, D., et al. 2019, ApJ, 884, 33
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Ho, M., Rau, M. M., Ntampaka, M., et al. 2019, ApJ, 887, 25
Ho, M., Farahi, A., Rau, M. M., & Trac, H. 2021, ApJ, 908, 204
Ho, M., Ntampaka, M., Rau, M. M., et al. 2022, Nat. Astron., 6, 936
Ho, M., Soltis, J., Farahi, A., et al. 2023, MNRAS, 524, 3289
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Klein, M., Oguri, M., Mohr, J. J., et al. 2022, A&A, 661, A4
Kodi Ramanah, D., Wojtak, R., Ansari, Z., Gall, C., & Hjorth, J. 2020, MNRAS, 499, 1985
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2017, Commun. ACM, 60, 84
Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2016, ArXiv e-prints [arXiv:1612.01474]
Liu, T., Merloni, A., Comparat, J., et al. 2022a, A&A, 661, A27
Liu, A., Bulbul, E., Ghirardini, V., et al. 2022b, A&A, 661, A2
Mamon, G. A., Biviano, A., & Boué, G. 2013, MNRAS, 429, 3079
Mantz, A. B., von der Linden, A., Allen, S. W., et al. 2015, MNRAS, 446, 2205
McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, eds. S. van der Walt, & J. Millman, 56
Merloni, A., Predehl, P., Becker, W., et al. 2012, ArXiv e-prints [arXiv:1209.3114]
Ntampaka, M., Trac, H., Sutherland, D. J., et al. 2015, ApJ, 803, 50
Ntampaka, M., ZuHone, J., Eisenstein, D., et al. 2019, ApJ, 876, 82
Old, L., Skibba, R. A., Pearce, F. R., et al. 2014, MNRAS, 441, 1513
Old, L., Wojtak, R., Mamon, G. A., et al. 2015, MNRAS, 449, 1897
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Planck Collaboration VI. 2020, A&A, 641, A6
Predehl, P., Andritschke, R., Arefiev, V., et al. 2021, A&A, 647, A1
Ramos-Ceja, M. E., Oguri, M., Miyazaki, S., et al. 2022, A&A, 661, A14
Scheck, D., Sanders, J. S., Biffi, V., et al. 2023, A&A, 670, A33
Seppi, R., Comparat, J., Bulbul, E., et al. 2022, A&A, 665, A78
Sunyaev, R., Arefiev, V., Babyshkin, V., et al. 2021, A&A, 656, A132
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nat. Methods, 17, 261
Wadekar, D., Thiele, L., Hill, J. C., et al. 2023a, MNRAS, 522, 2628
Wadekar, D., Thiele, L., Villaescusa-Navarro, F., et al. 2023b, Proc. Nat. Acad. Sci., 120, e2202074120
Yan, Z., Mead, A. J., Van Waerbeke, L., Hinshaw, G., & McCarthy, I. G. 2020, MNRAS, 499, 3445
ZuHone, J., Bahar, Y. E., Biffi, V., et al. 2023, A&A, 675, A150

Appendix A: Gaussian Kernel

X-ray astronomical maps correspond to the spatial distribution of the detected photons. Depending on the source luminosity, the amount of detected photons emitted by the source can be very sparse and can lead to a difficult recognition of the source shape. In fact, some EBI contain only a couple of source photons. In the EBI generation, we use Gaussian blur to support the learning process of our models. Gaussian blur is usually used in image processing to reduce the noise and detail. In our case, our aim is to smooth the sparse distribution of detected photons.

We utilize the kernel operation:

$$\tilde{I}(x, y, E) = \sum_{i=-X}^{X} \sum_{j=-Y}^{Y} \sum_{k=-Z}^{Z} I(x-i, y-j, E-k) G(i, j, k) , \qquad (A.1)$$

where I denotes our original EBI-array and $\tilde{I}$ the smoothed array. For the convolution, we use the following kernel $G(i, j, k)$

$$G(i, j, k) = \frac{1}{N} \exp\left(-\frac{(i/X)^2 + (j/Y)^2 + (k/Z)^2}{2\sigma^2}\right) , \qquad (A.2)$$

where N is the normalizing factor and σ the distributions standard deviation. We observed that smoothing also along the channel dimension, enhances the source structure in each channel of our EBIs and leads to improvements in our models' performance. The final choice of the filter size is $(3, 3, 11)$, which corresponds to $(X, Y, Z) = (1, 1, 5)$ in (A.1) and a standard deviation of $\sigma = 0.75$. We use `scipy.signal.convolve` (Virtanen et al. 2020) to perform this smoothing.

Appendix B: Hyperparameters

To identify well-performing neural networks we have performed a hyper-parameter search on which we provide an overview in this appendix.

- **CNN hyperparameters:** we run standard variations of the kernel size, and number of filters in convolutional blocks. We found that average pooling worked better than maxpooling. Furthermore we varied the activation functions (relu and leaky relu) and the number and dimensions of the final dense layers.
- **Different network architectures:** we have also observed that locally connected convolutional layers did not increase the performance. They were comparable to our CNN approach. In our hyper-parameter analysis, we have compare the behaviour of different pooling layers, and have varied the size and number of hidden layers, kernel sizes moderately. We find for a range of hyperparameters good performance.
- **EBI hyperparameters:** we have varied the extraction size among 100, 200, 300, and 500 pixels. 300 pixels showed the best performance and it corresponds to a size where essentially all cluster photons are contained within the extraction range. We also experimented with larger extraction sizes and could not find an improvement in performance. We found no difference between using the eSASS cluster center or the cluster center provided from the simulations as the center for our EBI image, that is, training with either of them resulted in no difference in performance. We also tried rescaling the EBI images according to redshift but could not identify better performance.
- **Regression losses:** besides our likelihood loss we have experimented with various standard regression losses (mean squared error, mean average percentage error, and mean squared logarithmic error). We find that they are generally lead to similar behaviour on the mass scatter.
- **Classification versus regression**: as classification tasks are sometimes easier learning problems than regression in machine learning, we have experimented with classification where the classes correspond to different mass bins. It turned out that this approach did not lead to improved performance in comparison to our current CNN-based approach.
- **Optimizers:** we have used Adam and generally found reducing the learning rate on plateau to be useful for performance and scanned through minimal learning rates. A batch size of a 100 was chosen, with variations not showing an increased performance.

Overall, we identify in this search several models with different hyperparameters, which do perform similarly. We think that this robustness is encouraging for these NN approaches.

Appendix C: Redshift uncertainty

Table C.1. Importance of redshift for our ensemble NN mass predictors.

Noise-level	Training Set	Test Set
$\sigma_N = 0$	$\sigma = 0.180$	$\sigma = 0.188$
$\sigma_N = 0.01$	$\sigma = 0.178$	$\sigma = 0.186$
$\sigma_N = 0.1$	$\sigma = 0.191$	$\sigma = 0.204$
$\sigma_N = 0.2$	$\sigma = 0.208$	$\sigma = 0.224$

Notes. We show the respective standard deviations between the true and predicted masses observed on the training and test set. We find that increasing the noise level leads to worse predictions.

We utilized the redshift as the input for our mass estimation network. Here, we discuss the relevance of this feature and how erroneous redshifts do affect our mass predictions. For this purpose, we compared a single CNN with and without redshift information on our best hyperparameter choice. In this case, we find a deterioration from $\sigma = 0.186$ to $\sigma = 0.253$ when providing no redshift information on the training set (respectively, it increases from $\sigma = 0.196$ to $\sigma = 0.274$ on the test set).

To estimate the effect of uncertain redshifts, we artificially added Gaussian noise with standard deviation σ_N of varying size on the redshifts and re-train our ensemble of networks with this data. The higher the level of this noise, the more our mass predictions worsen (as expected). Our observed results are summarized in Table C.1. We find that the expected uncertainty, which is well below 10%, is not relevant for our method. The small improvement is for a noise level of $\sigma_N = 0.1$ is attributed to fluctuations due to different NN initializations.