EPJ Web of Conferences **293**, 00019 (2024) https://doi.org/10.1051/epjconf/202429300019
mm Universe 2023

A machine learning method to infer clusters of galaxies mass radial profiles from mock Sunyaev-Zel'dovich maps with The Three Hundred clusters

A. Ferragamo[1,2,3,*], D. de Andres[4,5], A. Sbriglio[3], W. Cui[4,5,6], M. De Petris[3], G. Yepes[4,5], R. Dupuis[7], M. Jarraya[7], I. Lahouli[7], F. De Luca[8], G. Gianfagna[3,9], and E. Rasia[10,11]

[1]Instituto de Astrofísica de Canarias (IAC), C/ Vía Láctea s/n, E-38205 La Laguna, Tenerife, Spain
[2]Universidad de La Laguna, Departamento de Astrofísica, C/ Astrofísico Francisco Sánchez s/n, E-38206 La Laguna, Tenerife, Spain
[3]Dipartimento di Fisica, Sapienza Universitá di Roma, Piazzale Aldo Moro 5, I-00185 Roma, Italy
[4]Departamento de Física Téorica, Módulo 15, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain
[5]Centro de Investigación Avanzada en Física Fundamental (CIAFF), Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain
[6]Institute for Astronomy, University of Edinburgh, Edinburgh EH9 3HJ, United Kingdom
[7]EURANOVA, Mont-Saint-Guibert, Belgium
[8]Dipartimento di Fisica, Università di Roma Tor Vergata, Via della Ricerca Scientifica 1, I-00133 Roma, Italy
[9]INAF - Istituto di Astrofisica e Planetologia Spaziali, via Fosso del Cavaliere 100, I-00133 Roma, Italy
[10]IFPU - Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34014 Trieste, Italy
[11]INAF Osservatorio Astronomico di Trieste, via Tiepolo 11, I-34131, Trieste, Italy

Abstract. Our study introduces a new machine learning algorithm for estimating 3D cumulative radial profiles of total and gas mass in galaxy clusters from thermal Sunyaev-Zel'dovich (SZ) effect maps. We generate mock images from 2522 simulated clusters, employing an autoencoder and random forest in our approach. Notably, our model makes no prior assumptions about hydrostatic equilibrium. Our results indicate that the model successfully reconstructs unbiased total and gas mass profiles, with a scatter of approximately 10%. We analyse clusters in various dynamical states and mass ranges, finding that our method's accuracy and precision are consistent. We verify the capabilities of our model by comparing it with the hydrostatic equilibrium technique, showing that it accurately recovers total mass profiles without any bias.

1 Introduction

Galaxy clusters, the largest gravitationally bound structures in the Universe, play a pivotal role in our understanding of cosmology. They are primarily composed of dark matter, and baryons in the form of diffused hot gas known as the Intra Cluster Medium (ICM), and galaxies see [1]. These clusters are critical for probing cosmological parameters, such as the cosmic matter density and the amplitude of density fluctuations characterised by the σ_8 parameter,

*e-mail: antonio.ferragamo@uniroma1.it

which quantifies the linear effective overdensity within spheres of radius $8\,h^{-1}$Mpc see [2]. Accurately estimating the total mass of galaxy clusters is fundamental in this context. However, the inherent challenge lies in the fact that the dark matter component cannot be directly observed, necessitating the use of various methods that rely on observable baryonic components.

Several techniques based on the clusters self-similarity, such as the hydrostatic equilibrium assumption, have traditionally been employed to estimate cluster mass but can introduce biases. In the case of simulated clusters, these biases are typically within the range of 10-20% see [3]. To address this issue, a modern analysis approach utilising Machine Learning algorithms has emerged. These algorithms are trained to infer clusters mass, offering a promising avenue for obtaining mass estimates that are virtually free from bias [see 4]. For example, the determination of M_{200} was accomplished by [5] using simulated Sunyaev-Zel'dovich effect [SZ, 6] maps, while [7] employed optical, X-ray, and SZ images generated from the BAHAMAS simulation. On the other hand, [8] trained a Convolutional Neural Network (CNN) model to recover M_{500} from mock X-ray images in the Illustris TNG dataset. In this paper, we explore the application of Machine Learning techniques to tackle these challenges and present results that demonstrate the effectiveness of this approach in obtaining unbiased clusters mass profiles.

This paper is organised as follows. In section 2 we introduce The Three Hundred simulation. In 3 and 4 we describe the data-set the model we used for our analysis. In sec. 5 we present the results of our study, whereas section 6 shows the comparison with the classical HE method. Finally in 7 we present our conclusions.

2 The Three Hundred

The dataset utilised in this study comprises simulated galaxy clusters (GCs) selected from The Three Hundred (The300) [9]. These clusters were derived from a zoomed re-simulation of the 324 most massive Lagrangian regions within the MultiDark Planck 2 simulation (MDPL2) [10]. The MDPL2 simulation encompasses a volume of $1h^{-1}$Gpc and is populated with 3840^3 dark matter (DM) particles, each with a mass of $1.5 \times 10^9 h^{-1}\mathrm{M}_\odot$. This mass configuration is in accordance with the cosmological parameters of the Planck 2015 data release [11], which include $h = 0.678$, $n = 0.96$, $\sigma_8 = 0.823$, $\Omega_\Lambda = 0.693$, $\Omega_m = 0.307$, and $\Omega_b = 0.048$. To simulate these galaxy clusters, a region with a radius of $15h^{-1}$Mpc around the center of each of the 324 chosen Lagrangian regions was populated with gas particles, initially possessing a mass of $2.36 \times 10^8 h^{-1}\mathrm{M}_\odot$. Subsequently, these regions were re-simulated using various simulation codes. Our analysis in this study primarily focuses on the results obtained from the Gadget-X [12, 13] and GIZMO-SIMBA [14] hydrodynamical simulation codes, which incorporate diverse feedback mechanisms.

3 Dataset

The dataset employed in this study comprises 2522 clusters, evenly distributed across a total mass range spanning from $10^{13.5}$ to $10^{15.5}\,h^{-1}\mathrm{M}_\odot$ at six closely spaced redshifts, ranging from $z = 0$ to $z = 0.116$. This distribution was designed to create a nearly uniform population of clusters across a range of masses, although it is worth noting that there will inherently be fewer clusters at the higher end of the mass spectrum. For each individual cluster, we extract cumulative 3D radial profiles for both the total mass and the gas mass. These profiles are derived by summing the mass of all particles within concentric spheres, centred on the position determined by the highest density peak according to AHF, up to a distance of $2R_{200}$.

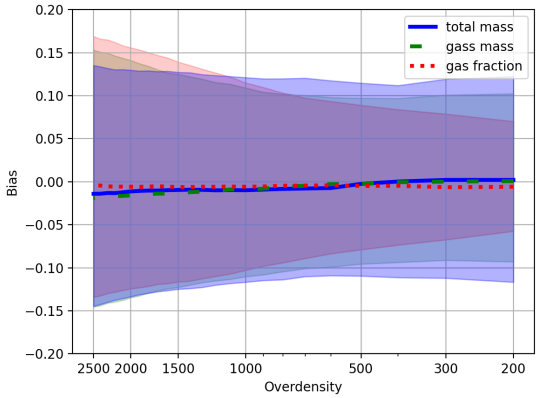

Figure 1. Median bias profiles for a cluster sample drawn from both GADGET-X and GIZMO-SIMBA, illustrating the relationship with overdensity for total mass (depicted by the blue solid line), gas mass (indicated by the orange dashed line), and gas fraction (represented by the green dotted line). The lightly shaded regions in blue, orange, and green correspond to the $16^{\text{th}} - 84^{\text{th}}$ percentiles for total mass, gas mass, and gas fraction, respectively.

To ensure comprehensive profile coverage, we selected 24 overdensities, uniformly spaced from 200 to 2500, allowing us to effectively sample the profiles. This approach enables the estimation of mass at common overdensities utilised in the literature, such as M_{200}, M_{500}, and M_{2500}. It distinguishes our work from previous studies, which often estimate cluster mass at a specific single aperture.
This sample was then divided into two subsets. One, consisting of 80% of the clusters, was used to train the network (traning set) while the remaining 20% was used to test the performance of the network (test set).

4 Model

The overall workflow consists of two key components: an autoencoder and a RF regressor. This approach revolves around the unsupervised extraction of features from SZ images, followed by the utilization of this representation for predicting mass profiles.
The autoencoder aims to establish a mapping between input data and an output, utilising an internal representation with reduced dimensionality. In our context, we employ an autoencoder to extract a representative feature vector from our input data, which consists of SZ maps, while preserving the original data fidelity.
Complementing the autoencoder, a random forest (RF) is incorporated. The RF comprises multiple decision trees each capable of making classifications or regressions based on input features. Individual decision trees can exhibit limited flexibility and overfitting. However, when combined into an RF, these issues are mitigated.

5 Results

In Section 3 we described the dataset used in this analysis. It is essential to remark that each simulation possesses its distinct characteristics, including variations in cosmology, resolu-

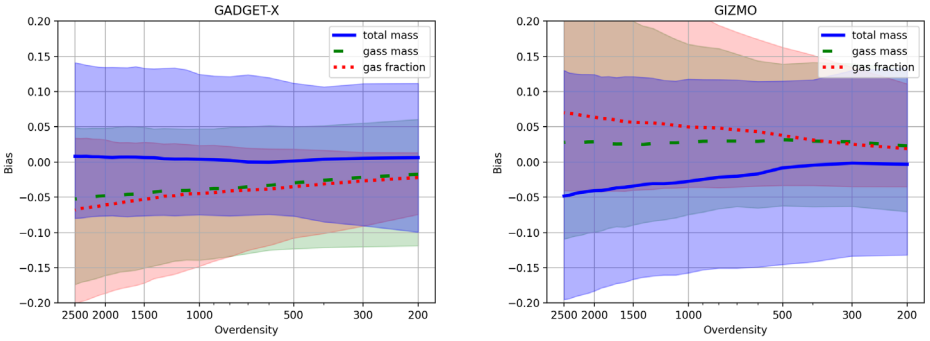

Figure 2. Median bias profiles for GADGET-X (left panel) and GIZMO-SIMBA (right panel) cluster datasets with respect to overdensity, showcasing the relationships for total mass (illustrated by the blue solid line), gas mass (depicted with the red dashed line), and gas fraction (represented by the green dotted line). The lightly shaded regions in blue, pink, and green correspond to the $16^{\text{th}} - 84^{\text{th}}$ percentiles for total mass, gas mass, and gas fraction, respectively.

tion, and baryonic physics. These disparities can potentially exert varying impacts on aspects like structure mass, mass profile shape, or SZ maps. Consequently, when discrepancies between products from different simulations are substantial, it may compromise the accuracy of machine learning models or conventional techniques such as scaling relations. To address this challenge, one approach within the machine learning context is to train a network using data from multiple simulations, as proposed by the Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) [15]. Building on this approach, we decided to retrain our model by incorporating clusters from both the GADGET-X and GIZMO-SIMBA simulation runs, with further details provided in [16]. As depicted in Figure 1, the application of this network to predict mass profiles within a test set composed of a combination of clusters from both GADGET-X and GIZMO-SIMBA simulations yielded results that align closely with those achieved by the network trained and tested exclusively on GADGET-X clusters. Specifically, the median bias for both mass profiles is approximately zero, and the scatter remains around 10% across all overdensities.

Of particular interest are the outcomes when applying the network to test sets comprised exclusively of clusters from each run separately. In the case of the test set containing solely GADGET-X clusters (as depicted in the left panel of Figure 2), the median bias for predicted total mass profiles (indicated by the blue solid line) is close to zero, and the scatter (represented by the blue shaded region) closely matches the results obtained with the network trained exclusively on GADGET-X clusters. However, for gas mass profiles (shown by the orange dashed line), a slight bias is observed, decreasing to approximately -5% toward the cluster core, while the scatter remains consistent with previous cases. In the right panel of Figure 2, we present the results of the network applied to the GIZMO-SIMBA clusters. In this scenario, the median bias for total mass profiles is essentially zero up to $\Delta \leq 500$, but it gradually increases to approximately 5% in the core. The scatter experiences a slight increase but remains below 20%. Notably, gas mass profiles exhibit a fairly constant bias of around 3%. However, the scatter for these profiles significantly rises in the inner regions of the clusters, potentially attributed to the strong AGN-feedback implemented in the GIZMO-SIMBA simulation. In contrast, predictions from GADGET-X remain more stable in the central regions.

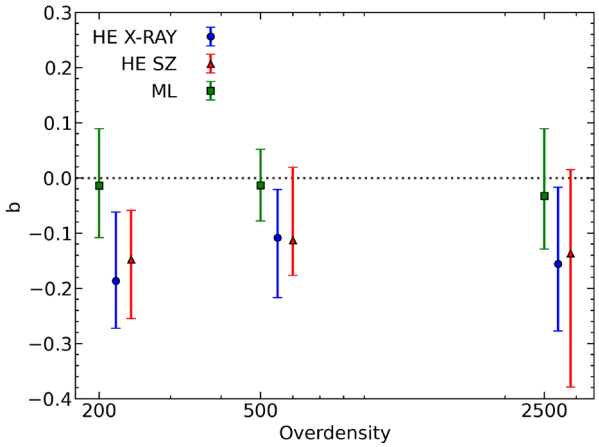

Figure 3. Comparison of median bias values obtained through the ML method (green squares) and mass bias values calculated using the HE both for x-ray (red triangles) and SZ (blue dots), at three shared overdensities for the same cluster selection within the The Three Hundred sample at $z = 0$, as analysed in [17]. The error bars denote the $16^{\text{th}} - 84^{\text{th}}$ percentiles.

6 Comparison with Hydrostatic Equilibrium mass estimates

Our machine learning model successfully retrieves mass radial profiles with a median bias close to zero, a notable achievement considering that no prior assumptions are made regarding the physical properties of the clusters. To assess the accuracy of our results, we conducted a comparison with the mass bias computed under the Hydrostatic Equilibrium (HE) approximation, as detailed in [17], using synthetic clusters derived from the The Three Hundred project. For this comparative analysis, we specifically considered the most massive clusters existing in each resimulated region at redshift $z = 0$, amounting to 53 objects within a mass range spanning from $1.3 \times 10^{14}\ h^{-1}\mathrm{M}_{\odot}$ to $3 \times 10^{15}\ h^{-1}\mathrm{M}_{\odot}$. The findings of this comparison are presented in Figure 3. In the case of the HE bias, when X-ray observables (illustrated by the red dots) such as electron gas temperature and density or SZ-derived pressure and density were employed (blue dots), the bias magnitude ranged between 10-20%. On the other hand, for our machine learning estimates (depicted as green dots), the bias is consistently below 1%, with a scatter of around 10%. Notably, the bias exhibits its lowest scatter at $\Delta = 500$.
While it is essential to acknowledge that the biases are within a compatible range considering the uncertainties, our machine learning-based mass estimates demonstrate a systematic advantage in terms of accuracy and lack of bias compared to the HE approach. Furthermore, the machine learning methodology results in a reduced scatter across the entire spectrum of overdensities.

7 Conclusions

In conclusion, this research marks a significant advancement in the estimation of cluster masses through the innovative application of machine learning techniques. We have achieved the simultaneous inference of integrated radial profiles for both gas and total mass from synthetic SZ images. Our machine learning model, integrating an autoencoder and a random forest regressor, showcases the capability to recover unbiased profiles, with biases consistently below 1% and a scatter of approximately 10%. Furthermore, we have derived the gas

fraction profile from the total and gas mass profiles, demonstrating a similarly low bias of less than 1% and a diminishing scatter, down to approximately 3% in the cluster outskirts. Crucially, our approach outperforms conventional methods, such as the Hydrostatic Equilibrium approximation, by delivering more accurate mass estimates that are free from hydrostatic mass bias. Moreover, our machine learning model, trained on a mixture of clusters from different runs of The300, displays the capacity to effectively accommodate diverse hydrodynamical simulations, yielding results consistent with single-simulation training. As we look forward to future applications, it is paramount to expand the dataset by incorporating data from several simulations. This step is essential to account comprehensively for a broad range of potential baryonic effects.

References

[1] A.V. Kravtsov, S. Borgani, Annual Review of Astronomy and Astrophysics **50**, 353 (2012)
[2] G.W. Pratt, M. Arnaud, A. Biviano, D. Eckert, S. Ettori, D. Nagai, N. Okabe, T.H. Reiprich, Space Sci. Rev. **215**, 25 (2019), 1902.10837
[3] G. Gianfagna, M. De Petris, G. Yepes, F. De Luca, F. Sembolini, W. Cui, V. Biffi, F. Kéruzoré, J. Macías-Pérez, F. Mayet et al., MNRAS **502**, 5115 (2021)
[4] D. Andres, W. Cui, F. Ruppin, M. De Petris, G. Yepes, G. Gianfagna, I. Lahouli, G. Aversano, R. Dupuis, M. Jarraya et al., Nature Astronomy **6**, 1325 (2022), 2209.10333
[5] N. Gupta, C.L. Reichardt, ApJ **900**, 110 (2020)
[6] R.A. Sunyaev, Y.B. Zeldovich, Ap&SS**7**, 3 (1970)
[7] Z. Yan, A.J. Mead, L. Van Waerbeke, G. Hinshaw, I.G. McCarthy, MNRAS **499**, 3445 (2020)
[8] M. Ntampaka, J. ZuHone, D. Eisenstein, D. Nagai, A. Vikhlinin, L. Hernquist, F. Marinacci, D. Nelson, R. Pakmor, A. Pillepich et al., ApJ **876**, 82 (2019)
[9] W. Cui, A. Knebe, G. Yepes, F. Pearce, C. Power, R. Dave, A. Arth, S. Borgani, K. Dolag, P. Elahi et al., MNRAS **480**, 2898 (2018)
[10] A. Klypin, G. Yepes, S. Gottlöber, F. Prada, S. Heß, MNRAS **457**, 4340 (2016), 1411.4001
[11] Planck Collaboration XIII, A&A **594**, A13 (2016)
[12] G. Murante, P. Monaco, M. Giovalli, S. Borgani, A. Diaferio, MNRAS **405**, 1491 (2010)
[13] E. Rasia, S. Borgani, G. Murante, S. Planelles, A.M. Beck, V. Biffi, C. Ragone-Figueroa, G.L. Granato, L.K. Steinborn, K. Dolag, ApJl **813**, L17 (2015)
[14] R. Davé, D. Anglés-Alcázar, D. Narayanan, Q. Li, M.H. Rafieferantsoa, S. Appleby, MNRAS **486**, 2827 (2019)
[15] F. Villaescusa-Navarro, D. Anglés-Alcázar, S. Genel, D.N. Spergel, R.S. Somerville, R. Dave, A. Pillepich, L. Hernquist, D. Nelson, P. Torrey et al., ApJ **915**, 71 (2021), 2010.00619
[16] W. Cui, R. Dave, A. Knebe, E. Rasia, M. Gray, F. Pearce, C. Power, G. Yepes, D. Anbajagane, D. Ceverino et al., MNRAS **514**, 977 (2022), 2202.14038
[17] G. Gianfagna, E. Rasia, W. Cui, M. De Petris, G. Yepes (2022), Vol. 257 of *EPJWC*, p. 00020, 2111.01903