Exploring the Universe with Artificial Intelligence (UniversAI)
Proceedings IAU Symposium No. 397, 2026
I. Liodakis & V. Efthymiou, eds.
doi:10.1017/S1743921325105097

Bridging observations and simulations: a machine learning approach to galaxy clusters

Efrain Gatuzz

Max-Planck-Institut für extraterrestrische Physik (MPE), Gießenbachstraße 1, 85748 Garching, Germany
email: egatuzz@mpe.mpg.de

Abstract. The intracluster medium (ICM) records the history of galaxy clusters through its complex dynamical properties. To effectively interpret these properties, robust methods are needed to compare observational data with theoretical models. We present a novel machine learning framework for comparing ICM line-of-sight velocity maps derived from X-ray observations. Our approach uses convolutional and Siamese neural networks to identify similarities between different kinematic fields. We outline the architecture of this framework and perform a series of sanity checks to validate its performance. These checks demonstrate the model's ability to correctly identify and quantify kinematic features, establishing a powerful new tool for future comparative studies of the ICM.

Keywords. X-rays: galaxies: clusters, galaxies: clusters: general, methods: N-body simulations

1. Introduction

The ICM, a hot, diffuse plasma, constitutes the dominant baryonic component of galaxy clusters (Andreon and Hurn 2010; Dai et al. 2010; Kravtsov and Borgani 2012). Its complex dynamics are driven by various astrophysical processes, including cluster mergers, AGN feedback, and large-scale structure formation, which generate turbulent motions, bulk flows, and coherent velocity patterns such as gas sloshing (Ascasibar and Markevitch 2006; Vazza et al. 2018; ZuHone et al. 2018; Bambic and Reynolds 2019; Ichinohe et al. 2019).

Recent advances in X-ray spectroscopy, particularly through the calibration technique introduced by Sanders et al. (2020), have enabled direct measurements of ICM line-of-sight (LOS) velocities with high precision. Applying this method to clusters such as Perseus, Virgo, and Ophiuchus has revealed a range of kinematic phenomena, including AGN-driven outflows and merger-induced bulk flows. Interpreting these observations requires comparison with predictions from cosmological hydrodynamical simulations (Gaspari et al. 2014; Zhuravleva et al. 2014; Mohapatra and Sharma 2019). However, the high dimensionality and inherent complexity of both observational and simulated datasets pose a significant challenge for direct comparison.

In this work, we present a novel methodology that employs machine learning to address this problem. Specifically, we use Convolutional Neural Networks (CNNs) and Siamese architectures to construct a scalable framework for comparing observed LOS velocity maps with synthetic counterparts derived from simulations. This approach builds upon successful applications of similar techniques in other areas of astrophysics (Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018; Pearson 2019; Brunel et al. 2019;

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1743921325105097 Published online by Cambridge University Press

Zhang et al. 2022), and is designed to identify both kinematic and morphological similarities between observed and simulated ICM structures. In the following sections, we outline the core components of this method.

2. Galaxy Cluster Velocity Maps

To test our machine learning framework, we used LOS velocity maps from several galaxy clusters: Virgo (Gatuzz et al. 2022b, 2023b), Centaurus (Gatuzz et al. 2022a, 2023d), Ophiuchus (Gatuzz et al. 2023c,e), and A3266 (Gatuzz et al. 2024, 2025). These maps were derived from *XMM-Newton* EPIC observations, processed using a standardized data reduction and analysis pipeline. Briefly, the observations were reduced with SAS (version 19.1.0), including the selection of single-pixel events and the filtering of flaring time intervals. A critical aspect of the process involved the application of updated calibration files and the use of instrumental background lines to achieve a LOS velocity precision better than 100 km/s at the Fe-K line, following the method described in Sanders et al. (2020). Point sources, including central AGN, were identified and excluded from the analysis. Spectral maps were generated by fitting X-ray spectra extracted from dynamically sized elliptical regions across each observation. The ICM emission was modeled using an `apec` thermal plasma component (Foster et al. 2019), with additional components introduced in cases requiring more complex modeling, such as the multi-temperature structure observed in the Centaurus cluster. This procedure, which accounts for Galactic absorption and instrumental background contributions, yielded the LOS velocity maps that serve as inputs for our machine learning framework. Further details on the reduction and spectral analysis techniques can be found in the original references.

3. Machine Learning Methodology

We present a deep learning framework based on a Siamese Convolutional Neural Network (CNN) designed to compare velocity maps of the ICM. The objective is to identify kinematically similar fields by learning a robust representation, or *embedding*, for each input velocity map. The Siamese CNN consists of two identical branches with shared weights, facilitating similarity learning. Each branch functions as a feature extractor, composed of multiple convolutional layers that capture hierarchical spatial features, followed by max-pooling layers to reduce spatial dimensions. After the final convolutional block, a flattening layer and a fully connected dense layer convert the feature maps into a compact, fixed-length numerical vector-the embedding-which encodes the essential characteristics of the input map.

The network is trained using the triplet loss function (Schroff et al. 2015). During training, the network is presented with triplets: an anchor image (A), a positive image (P) that is similar to the anchor, and a negative image (N) that is dissimilar. The loss function adjusts the network's weights via back propagation to minimize the distance between the anchor and positive embeddings while enforcing a margin between the anchor and negative embeddings. Once training is complete and the weights are optimized, the CNN branches act as fixed feature extractors. The observed velocity map from *XMM-Newton* is passed through one branch to generate its embedding vector, while each simulated velocity map is processed through the second branch to compute its corresponding embedding. A critical preprocessing step ensures all images are resampled to a uniform pixel resolution consistent with the *XMM-Newton* data.

To quantify similarity, we compute both the Euclidean distance and the cosine similarity between the embeddings of the observation and each simulation. The simulation with the smallest Euclidean distance (or, equivalently, the highest cosine similarity) is identified as the best match.

https://doi.org/10.1017/S1743921325105097 Published online by Cambridge University Press

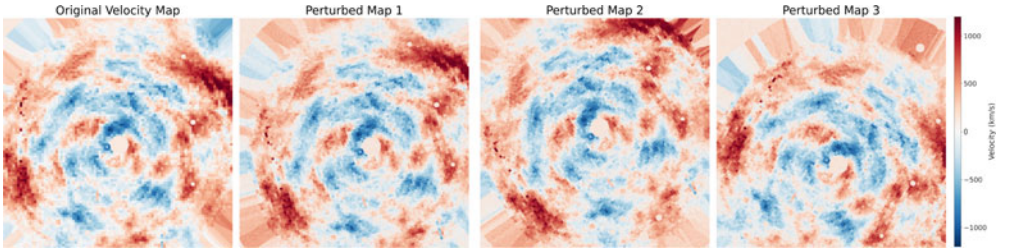


Figure 1. Examples from the artificial dataset generated for the Virgo cluster. These perturbed velocity maps were created by applying controlled transformations (e.g., rotation, noise) to the original *XMM-Newton* observation.

4. Validation Using Artificial Data

To assess the robustness and reliability of our Siamese CNN in identifying kinematic similarities, we conducted a series of validation tests using controlled artificial datasets. These tests serve as sanity checks to confirm that the network learns meaningful and distinctive representations of ICM velocity maps.

For each observed cluster-Virgo, Centaurus, Ophiuchus, and A3266-we generated a set of 1,000 artificial velocity maps. These maps were created by applying controlled perturbations to the original *XMM-Newton* observations, including random rotations, pixel-level noise additions, and minor distortions (see Figure 1). We then trained the Siamese CNN on a subset of this artificial dataset.

To evaluate performance, we queried the trained model with the original (unperturbed) observational map and searched for its closest match among a validation set containing the 1000 perturbed maps along with the original. In every case, the model correctly identified the original observation as the best match-returning the smallest embedding distance and highest similarity score. This outcome confirms that the Siamese CNN has successfully learned to recognize the intrinsic kinematic patterns of the ICM, even when subjected to small but realistic variations.

To further support these findings, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the embedding space. This dimensionality reduction technique projects high-dimensional embedding vectors into a 2D space while preserving local structures. In these t-SNE plots, each point represents the embedding of one artificial velocity map. As shown in Figure 2, the embeddings of the 1000 perturbed Virgo maps form a compact cluster tightly surrounding the embedding of the original observation. This visualization demonstrates that the network has learned a coherent representation of the underlying ICM kinematics-grouping similar structures together while separating dissimilar ones in the embedding space.

5. Conclusion

In this work, we have introduced a novel machine learning framework for the comparative analysis of ICM kinematics. Our approach leverages a Siamese CNN trained with a triplet loss function to learn meaningful and robust representations of velocity maps derived from X-ray observations. This architecture is specifically designed to identify and quantify similarities between different kinematic fields.

To validate the effectiveness of our method, we performed a series of controlled tests using artificially perturbed datasets based on real *XMM-Newton* observations. The model consistently identified the original, unperturbed maps as the closest match, demonstrating its ability to capture the essential kinematic features. These results were further supported by t-SNE visualizations, which showed that embeddings of the perturbed

https://doi.org/10.1017/S1743921325105097 Published online by Cambridge University Press

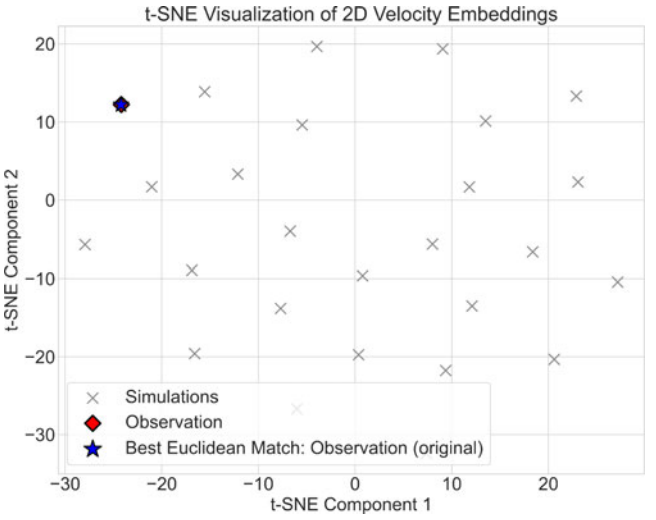


Figure 2. t-SNE projection of the embedding space for the Virgo cluster. The plot shows that the best matching correspond to the original observation, demonstrating that the Siamese CNN learns consistent kinematic representations.

maps clustered tightly around those of the originals-providing strong evidence that the network has learned to distinguish meaningful patterns in the data.

This framework offers a scalable and automated solution for matching observations with simulations, addressing a key challenge in the comparison of high-dimensional, complex astrophysical datasets. It serves as a powerful tool for linking X-ray observations with theoretical models of cluster formation and evolution.

In a forthcoming paper, we will apply this methodology to cosmological hydrodynamical simulations from the IllustrisTNG suite (Springel et al. 2018), providing a comprehensive analysis of how observed ICM velocity structures compare to those predicted by state-of-the-art simulations.

References

Andreon, S. & Hurn, M. A. 2010, *MNRAS*, 404, 1922
Ascasibar, Y. & Markevitch, M. 2006, *ApJ*, 650, 102
Bambic, C. J. & Reynolds, C. S. 2019, *ApJ*, 886, 78
Brunel, A., et al. 2019, *arXiv e-prints*, arXiv:1901.00461
Dai, X., et al. 2010, *ApJ*, 719, 119
Domínguez Sánchez, H., et al. 2018, *MNRAS*, 476, 3661
Foster, A., et al. 2019, *American Astronomical Society Meeting Abstracts*, 233 251.05
Gaspari, M., et al. 2014, *A&A*, 569, A67
Gatuzz, E., et al. 2022a *MNRAS*, 513a, 1932
Gatuzz, E., et al. 2022b *MNRAS*, 511b, 4511
Gatuzz, E., et al. 2023a *MNRAS*, 524a, 2945
Gatuzz, E., et al. 2023b *MNRAS*, 520b, 4793
Gatuzz, E., et al. 2023c *MNRAS*, 522, 2325
Gatuzz, E., et al. 2023d *MNRAS*, 525d, 6394
Gatuzz, E., et al. 2023e *MNRAS*, 526e, 396
Gatuzz, E., et al. 2024, *A&A*, 692, A108
Gatuzz, E., et al. 2025, *A&A*, 698, A94
Huertas-Company, M., et al. 2015,*ApJS*, 221, 8
Ichinohe, Y., et al. 2019, *MNRAS*, 483, 1744

https://doi.org/10.1017/S1743921325105097 Published online by Cambridge University Press

Kravtsov, A. V. & Borgani, S. 2012, *ARAA*, 50, 353
Mohapatra, R. & Sharma, P. 2019, *MNRAS*, 484, 4881
Pearson, K. A. 2019, *AJ*, 158, 243
Sanders, J. S., et al. 2020, *A&A*, 633, A42
Schroff, F., et al. 2015, In *2015 IEEE CVPR*, pp. 815
Springel, V., et al. 2018, *MNRAS*, 475, 676
Vazza, F., et al. 2018, *MNRAS*, 481, L120
Zhang, Z., et al. 2022, *Research in Astronomy and Astrophysics*, 22, 055002
Zhuravleva, I., et al. 2014, *Nature*, 515, 85
ZuHone, J. A., et al. 2018, *ApJ*, 853, 180

https://doi.org/10.1017/S1743921325105097 Published online by Cambridge University Press