Manifold Learning for Cosmic Structures

Ana Sofía M. Uzsoy
Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA
ana_sofia.uzsoy@cfa.harvard.edu

Claire Lamman
Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA
Center for Cosmology and AstroParticle Physics (CCAPP), Ohio State University, Columbus, OH

Melanie Weber
John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

Abstract

We present a scalable manifold learning approach to represent galaxies in a low-dimensional embedding space based on the geometry of their surrounding structure. We validate this method on a toy dataset consisting of points in balls and lines in space, and demonstrate its utility for astrophysics research on the realistic TNG100 galaxy simulation box. For both datasets, our method effectively captures the local structure around each galaxy. For the TNG100 simulations we show that our first embedding dimension correlates with halo mass and star-formation rate, which aligns with known physical relationships.

1 Introduction

Characterizing structures in the cosmic web, such as filaments or clusters of galaxies, is a key aspect of understanding the physical processes underlying the distribution of matter throughout the Universe. There are known relationships between the physical properties of galaxies and their surrounding structure; at low redshifts, galaxies in clusters are likely to be more massive, less star-forming, and have higher metallicity than galaxies in the field [1, 2, 3].

Previous work identifies galaxy clusters using clustering algorithms such as friends-of-friends [4, 5, 6] or DBSCAN [7] that yield binary cluster labels. Recent work has also utilized the eigendecomposition of the Hessian of the density field to identify nodes, filaments, and sheets in distributions of galaxies or stars [8, 9]. While the Hessian captures important second-order information that allows for detailed characterization of structures, it is computationally expensive to calculate for every galaxy in a dataset. Scalable methods are increasingly necessary in the age of large-scale astronomical surveys such as the Dark Energy Spectroscopic Instrument [DESI; 10], and the Vera C. Rubin Observatory [11], which aim to provide large 3D maps of tens of millions of galaxies.

Manifold learning is a branch of Geometric Machine Learning that aims to find a low-dimensional nonlinear representation of a higher-dimensional dataset such that key structural properties are preserved. Examples of such methods include Isomap [12], Locally Linear Embeddings [LLE; 13], and Laplacian Eigenmaps [14], all of which characterize the data's intrinsic geometry to learn low-dimensional representations optimized for downstream tasks. We present a new, scalable manifold learning approach for learning low-dimensional representations of galaxies based on their local geometric structure. In contrast to existing methods, our approach compares similarity distances between local galaxy neighborhoods to produce a continuous embedding space that captures

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

higher-order structures without expensive Hessian calculations. Our code is publicly available at https://github.com/asmuzsoy/galaxy-manifolds.

2 Methods

Given a dataset of n points in a 3D box, we aim to create an m-dimensional embedding space (where $m < 3$) in which points cluster based on the structure of their local neighborhoods. We first define a neighborhood radius r, which is necessarily much less than the box size. We then create an $n \times n$ distance matrix D_p (denoting "physical distances") that contains Euclidean distances between points if they are less than r and zeros otherwise. We use periodic boundary conditions when calculating physical distances between points in space to avoid edge effects. The size r of the local neighborhood is a tunable hyperparameter and allows for representation of structures at different scales.

Under this framework, every row $D_{p,i}$ denotes a vector of length n with the distances from point i to every other point that is within distance r, and zeros otherwise. D_p is then sorted along this axis, so that the distances are in order of smallest to largest, with most vectors having a significant number of leading zeros due to points being further than distance r away. We then calculate a new $n \times n$ matrix D_s (denoting "similarity distances"), containing the Euclidean distances between these sorted distance vectors.

Points with similar local structure will have more similar sorted distance vectors, and thus lower similarity distances than points with more different local structure. The number of nonzero values in each point i's distance vector ($D_{p,i}$) denotes the number of other points within distance r and serves as a proxy for local density. A comparison between two points' distance vectors with different numbers of nonzero values will result in a larger similarity distance, which is desired behavior as it denotes differences in their neighborhood structure.

From here, we use classical multidimensional scaling [MDS; 15], to create a low-dimensional representation of this data. First, the $n \times n$ matrix of pairwise similarity distances D_s is converted to a matrix of "centered" distances B by applying a centering matrix C on both sides:

$$B = -\frac{1}{2} C D_s C \qquad \text{where} \qquad C = \mathbb{I} - \frac{1}{n} J_n \tag{1}$$

where $\mathbb{I}$ is the identity matrix and J_n is an $n \times n$ matrix of ones.

The low-dimensional representation X is then determined by taking the eigendecomposition of B and scaling the top m eigenvectors by the square root of their corresponding eigenvalues, where m is the desired dimensionality of the final embedding space. In this work, we use $m = 1$ or 2.

Through this procedure, MDS optimizes a low-dimensional representation of the data by minimizing strain S:

$$S(x_1, x_2, ..., x_n) = \left(\frac{\sum_{i,j} (b_{ij} - x_i^T x_j)^2}{\sum_{i,j} b_{ij}^2} \right)^{1/2} \tag{2}$$

where b_{ij} denotes each element in the centered matrix B and x_i denotes the new low-dimensional representation of point i.

Our approach is reminiscent of Isomap [12], which constructs a similarity graph with edges between a point and its nearest neighbors; the shortest-path distances between nodes in the similarity graph are then used to approximate geodesic distances in a suitable low-dimensional representation space, which are then used as input to MDS. Our approach differs from Isomap in that the MDS routine is applied to the neighborhood similarity matrix D_s instead of the geodesic distance matrix.

3 Results

Synthetic Data For a simple test of our method, we create a synthetic "barbell" dataset of 2,200 points in a box of size 20 on each side. This dataset consists of 500 points in each of the two balls, 200 points in a line connecting the two balls, and 1,000 random points uniformly distributed throughout the entire box. Points in the balls are sampled from 3D Gaussian distributions with means of $\pm(3, 3, 3)$ and covariance matrices of $2\mathbb{I}_3$. We use a neighborhood size r = 0.5.

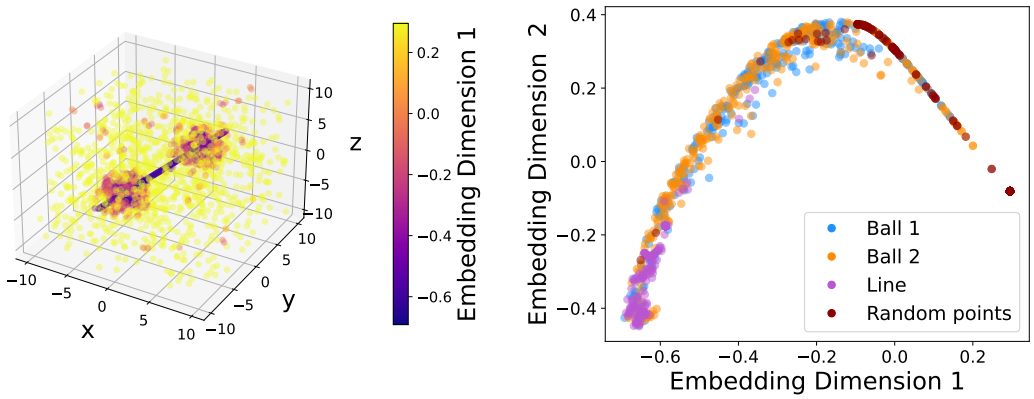

Figure 1: (left) The synthetic "barbell" dataset, colored by the first embedding dimension. (right) The points from the synthetic "barbell" dataset in the first two embedding dimensions, colored by the structures they belong to (either of the two balls, the line, or randomly distributed).

Figure 1 shows the results of our method on the "barbell" dataset; the three different structures (balls, line and random points) occupy different regions of the embedding space. The embedding of the two balls are indistinguishable with no positional dependence. Since the balls are Gaussian-distributed, points are more concentrated towards the center and more diffuse on the outsides. This can be seen in their embedding, where the points in the balls range from the embedding values of points in the line to those of the random points.

The neighborhood size r is a hyperparameter that specifies the scale of structure to be expressed in the embedding space. While we primarily use a value of $r = 0.5$ for this dataset (shown in Figure 1), we explored a variety of values that capture different neighborhood sizes around each point. Figure 2 shows the resulting embeddings calculated with $r = 0.1$ (left) and $r = 5$ (right) to show the diversity of scales that can be represented by this method. Neighorhood sizes that are too small or too big do not capture enough information about the structure around a point to effectively separate points with different surrounding geometries. As seen in Figure 2, the embeddings of the points in the balls are indistinguishable from those of the randomly distributed points at $r = 0.1$ and the points in the line at $r = 5$.

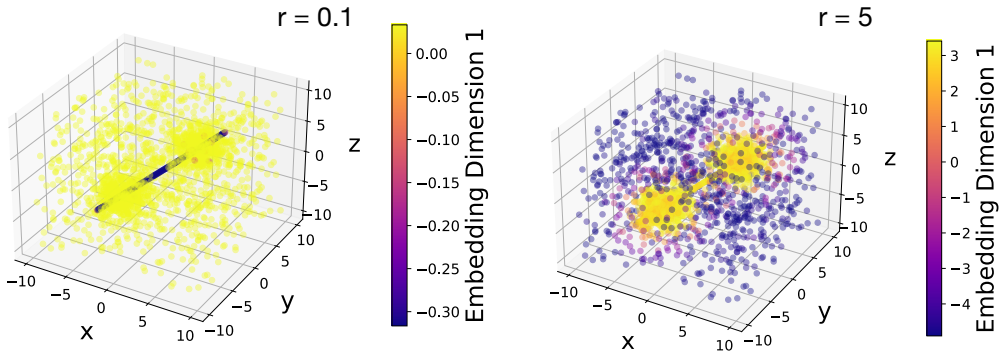

Figure 2: The synthetic "barbell" dataset, colored by the first embedding dimension, with embeddings calculated with a neighborhood size r of 0.1 (left) and 5 (right). The embeddings shown in Figure 1 use $r = 0.5$.

Realistic galaxy simulations The IllustrisTNG project is a realistic suite of galaxy formation simulations that capture how visible and dark matter evolve over time. We use the TNG100 Simulations

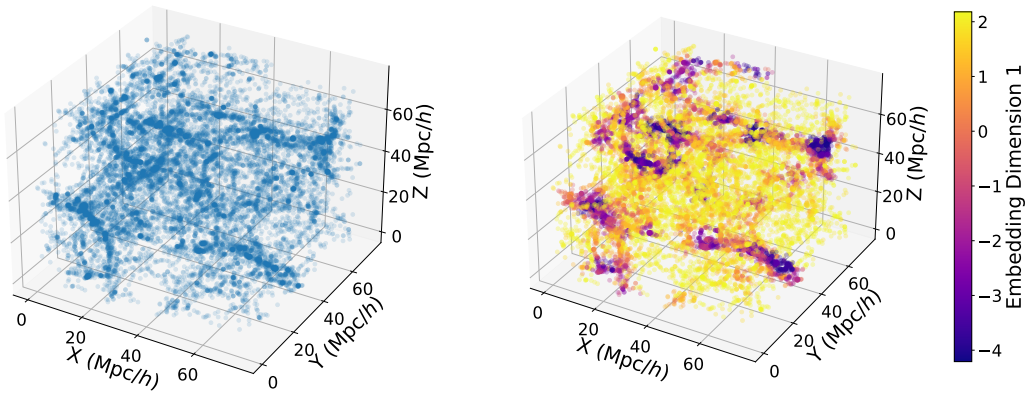

Figure 3: (left) The distribution of galaxies (subhalos) in the TNG100 dataset, in comoving coordinates. (right) The same dataset, colored by the first embedding dimension.

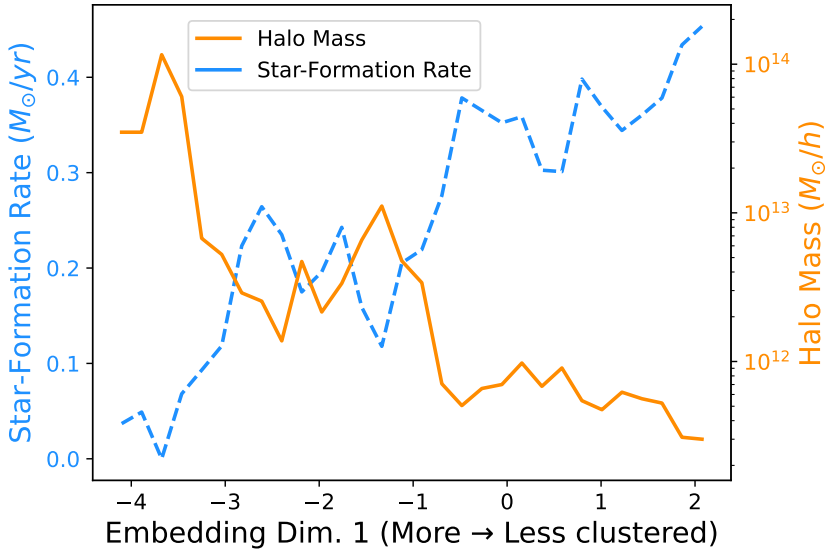

Figure 4: Median host dark matter halo mass (orange, right y-axis) and galaxy star-formation rate (dashed blue, left y-axis) from TNG100, calculated across 30 bins of the first embedding dimension.

[16, 17, 18, 19, 20], which have a box size of 75 Mpc[1] $/h$, and use the subhalo catalog at $z \approx 0$, which aims to simulate the universe at the present day. The TNG simulations use comoving distances (scaled by the dimensionless Hubble constant h), which factor out the expansion of the universe, and the size of the TNG100 box is small enough that we can approximate cosmological distances as Euclidean. We select confirmed galaxies that have stellar masses greater than $10^9 M_\odot$ and more than 1,000 star particles, yielding 17,611 galaxies. Here we use a neighborhood size of $r = 5$ Mpc/h.

Figure 3 shows the distribution of galaxies in the TNG100 simulation, and the resulting embeddings using our method. Just as in Figure 1, there is no positional dependence in the embedding values, and the embedding is able to separate points in clusters and filamentary structures from points in sparser and more diffuse environments. A comparison to DBSCAN can be found in Appendix A.

The first embedding dimension provides a continuous variable that encodes something akin to the "clusteriness" of an environment, and this can be useful in probing galaxy-environment interactions.

[1]Relevant astronomical units: 1 megaparsec (Mpc) ≈ 3.26 million light years ≈ 3.1×10^{19} km, 1 $M_\odot$ = the mass of the Sun = 2×10^{30} kg.

The TNG100 simulations include physical parameters of galaxies as well as their positions, and we can use our embedding to further examine the relationship between galaxy properties and their environments.

Figure 4 shows the relationship between the first embedding dimension, galaxy star-formation rate (`SubhaloSFR`) and host dark matter halo mass (`Group_M_Crit200`). Both of these quantities correlate with our first embedding dimension as we would expect from known physical relationships. Galaxies with lower embedding values, which denote more clustered regions, have higher halo masses and lower star-formation rates than galaxies with higher embedding values, denoting field regions.

In addition to reproducing known differences between cluster and field galaxies, our embedding allows us to probe environments between these two extremes. The halo mass and star-formation rate both seemingly plateau between embedding values of -1 and -3, which appear to represent more filamentary structures in Figure 3. A continuous embedding value allows us to examine the transitions between clusters, filaments, and field galaxies, and show how physical parameters can vary smoothly with environmental structure instead of just in discrete populations.

4 Discussion & Conclusion

Our manifold learning method improves on traditional discrete cluster-finding methods by providing a continuous value that captures more nuance in the structure of galaxy neighborhoods. By calculating similarity distances between points' sorted neighborhood distance vectors, we are able to identify many of the relevant structures in a dataset without having to do expensive Hessian calculations at each point. One limitation of our method is that we compute radial distances between points, so two neighborhoods with points at the same distances but distributed at different angles (i.e. colinear vs. spherical) would encode as the same distance vector.

This method is extremely scalable, requiring only a fraction of pairwise distances (those within a small neighborhood) to be computed. It includes only one eigendecomposition of a symmetric matrix (which can be leveraged for further computational speedup) to compute an embedding of the entire dataset. The combination of scalability and enhanced expressivity compared to discrete cluster identification or local density estimation makes this method particularly well-suited for use with data from large-scale astronomical surveys such as DESI and Rubin.

Potential areas for future work include creating a standardized low-dimensional representation space, within which different data sets' representations can be aligned using optimal transport. This would simplify comparison between different data sets. Moreover, void regions could be identified by slowly introducing uniformly distributed points and monitoring changes in the embedding space. While being extremely relevant for the characterization of the cosmic web, our method works on any point cloud, and could have additional applications in fields such as particle physics or atmospheric science.

Acknowledgments and Disclosure of Funding

ASMU was supported by a National Science Foundation Graduate Research Fellowship and would like to thank Vedant Chandra, Michelle Ntampaka, Douglas Finkbeiner, and Ashley Villar for helpful discussions and suggestions. MW was supported by NSF award CBET-2112085 and DMS-2406905, an Aramont Fellowship for Emerging Science Research, and an Alfred P. Sloan Research Fellowship in Mathematics. CL is supported by an NSF Postdoctoral Fellowship under award 2502789.

References

[1] D. Elbaz, E. Daddi, D. Le Borgne, M. Dickinson, D. M. Alexander, R. R. Chary, J. L. Starck, W. N. Brandt, M. Kitzbichler, E. MacDonald, M. Nonino, P. Popesso, D. Stern, and E. Vanzella. The reversal of the star formation-density relation in the distant universe. *Astronomy & Astrophysics*, 468(1):33–48, June 2007. doi: 10.1051/0004-6361:20077525.

[2] Michael C. Cooper, Christy A. Tremonti, Jeffrey A. Newman, and Ann I. Zabludoff. The role of environment in the mass-metallicity relation. *Monthly Notices of the Royal Astronomical Society*, 390(1):245–256, October 2008. doi: 10.1111/j.1365-2966.2008.13714.x.

[3] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Greg Snyder, Dylan Nelson, and Lars Hernquist. Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe. *Monthly Notices of the Royal Astronomical Society*, 444(2):1518–1547, October 2014. doi: 10.1093/mnras/stu1536.

[4] V. R. Eke, Carlton M. Baugh, Shaun Cole, Carlos S. Frenk, Peder Norberg, John A. Peacock, Ivan K. Baldry, Joss Bland-Hawthorn, Terry Bridges, Russell Cannon, Matthew Colless, Chris Collins, Warrick Couch, Gavin Dalton, Roberto de Propris, Simon P. Driver, George Efstathiou, Richard S. Ellis, Karl Glazebrook, Carole Jackson, Ofer Lahav, Ian Lewis, Stuart Lumsden, Steve Maddox, Darren Madgwick, Bruce A. Peterson, Will Sutherland, and Keith Taylor. Galaxy groups in the 2dFGRS: the group-finding algorithm and the 2PIGG catalogue. *Monthly Notices of the Royal Astronomical Society*, 348(3):866–878, March 2004. doi: 10.1111/j.1365-2966.2004.07408.x.

[5] Y. P. Jing. Accurate Fitting Formula for the Two-Point Correlation Function of Dark Matter Halos. *The Astrophysical Journal Letters*, 503(1):L9–L13, August 1998. doi: 10.1086/311530.

[6] Z. Xiong, P. Zhang, X. M. Yang, G. C. Liu, D. Liu, J. P. Li, and H. J. Tian. Performance Comparison of Three Clustering Algorithms in Open Cluster Member Star Identification. *Acta Astronomica Sinica*, 66(2):17, March 2025. doi: 10.15940/j.cnki.0001-5245.2025.02.006.

[7] Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1):1–30, 2019. doi: 10.18637/jss.v091.i01.

[8] Gabriella Contardo, David W. Hogg, Jason A. S. Hunt, Joshua E. G. Peek, and Yen-Chi Chen. The Emptiness Inside: Finding Gaps, Valleys, and Lacunae with Geometric Data Analysis. *The Astronomical Journal*, 164(5):226, November 2022. doi: 10.3847/1538-3881/ac961e.

[9] Behnam Darvish, Bahram Mobasher, D. Christopher Martin, David Sobral, Nick Scoville, Andra Stroe, Shoubaneh Hemmati, and Jeyhan Kartaltepe. Cosmic Web of Galaxies in the COSMOS Field: Public Catalog and Different Quenching for Centrals and Satellites. *The Astrophysical Journal*, 837(1):16, March 2017. doi: 10.3847/1538-4357/837/1/16.

[10] DESI Collaboration, M. Abdul-Karim, A. G. Adame, D. Aguado, J. Aguilar, S. Ahlen, S. Alam, G. Aldering, D. M. Alexander, R. Alfarsy, L. Allen, C. Allende Prieto, O. Alves, A. Anand, U. Andrade, E. Armengaud, S. Avila, A. Aviles, H. Awan, S. Bailey, A. Baleato Lizancos, O. Ballester, A. Bault, J. Bautista, S. BenZvi, L. Beraldo e Silva, J. R. Bermejo-Climent, F. Beutler, D. Bianchi, C. Blake, R. Blum, A. S. Bolton, M. Bonici, S. Brieden, A. Brodzeller, D. Brooks, E. Buckley-Geer, E. Burtin, R. Canning, A. Carnero Rosell, A. Carr, P. Carrilho, L. Casas, F. J. Castander, R. Cereskaite, J. L. Cervantes-Cota, E. Chaussidon, J. Chaves-Montero, S. Chen, X. Chen, T. Claybaugh, S. Cole, A. P. Cooper, M. C. Cousinou, A. Cuceu, T. M. Davis, K. S. Dawson, R. de Belsunce, R. de la Cruz, A. de la Macorra, A. de Mattia, N. Deiosso, J. Della Costa, R. Demina, U. Demirbozan, J. DeRose, A. Dey, B. Dey, J. Ding, Z. Ding, P. Doel, K. Douglass, M. Dowicz, H. Ebina, J. Edelstein, D. J. Eisenstein, W. Elbers, N. Emas, S. Escoffier, P. Fagrelius, X. Fan, K. Fanning, V. A. Fawcett, E. Fernández-García, S. Ferraro, N. Findlay, A. Font-Ribera, J. E. Forero-Romero, D. Forero-Sánchez, C. S. Frenk, B. T. Gänsicke, L. Galbany, J. García-Bellido, C. Garcia-Quintero, L. H. Garrison, E. Gaztañaga, H. Gil-Marín, O. Y. Gnedin, S. Gontcho A Gontcho, A. X. Gonzalez-Morales, V. Gonzalez-Perez, C. Gordon, O. Graur, D. Green, D. Gruen, R. Gsponer, C. Guandalin, G. Gutierrez, J. Guy, C. Hahn, J. J. Han, J. Han, S. He, H. K. Herrera-Alcantar, K. Honscheid, J. Hou, C. Howlett, D. Huterer, V. Iršič, M. Ishak, A. Jacques, J. Jimenez, Y. P. Jing, B. Joachimi, S. Joudaki,

R. Joyce, E. Jullo, S. Juneau, N. G. Karaçaylı, T. Karim, R. Kehoe, S. Kent, A. Khederlarian, D. Kirkby, T. Kisner, F. S. Kitaura, N. Kizhuprakkat, H. Kong, S. E. Koposov, A. Kremin, A. Krolewski, O. Lahav, Y. Lai, C. Lamman, T. W. Lan, M. Landriau, D. Lang, J. U. Lange, J. Lasker, J. M. Le Goff, L. Le Guillou, A. Leauthaud, M. E. Levi, S. Li, T. S. Li, K. Lodha, M. Lokken, Y. Luo, C. Magneville, M. Manera, C. J. Manser, D. Margala, P. Martini, M. Maus, J. McCullough, P. McDonald, G. E. Medina, L. Medina-Varela, A. Meisner, J. Mena-Fernández, A. Menegas, M. Mezcua, R. Miquel, P. Montero-Camacho, J. Moon, J. Moustakas, A. Muñoz-Gutiérrez, D. Muñoz-Santos, A. D. Myers, J. Myles, S. Nadathur, J. Najita, L. Napolitano, J. A. Newman, F. Nikakhtar, R. Nikutta, G. Niz, H. E. Noriega, N. Padmanabhan, E. Paillas, N. Palanque-Delabrouille, A. Palmese, J. Pan, Z. Pan, D. Parkinson, J. Peacock, W. J. Percival, A. Pérez-Fernández, I. Pérez-Ràfols, P. Peterson, J. Piat, M. M. Pieri, M. Pinon, C. Poppett, A. Porredon, F. Prada, R. Pucha, F. Qin, D. Rabinowitz, A. Raichoor, C. Ramírez-Pérez, S. Ramirez-Solano, M. Rashkovetskyi, C. Ravoux, A. H. Riley, A. Rocher, C. Rockosi, J. Rohlf, A. J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, C. G. Sabiu, K. Said, A. Saintonge, L. Samushia, E. Sanchez, N. Sanders, C. Saulder, E. F. Schlafly, D. Schlegel, D. Scholte, M. Schubnell, H. Seo, A. Shafieloo, R. Sharples, J. Silber, M. Siudek, A. Smith, D. Sprayberry, J. Suárez-Pérez, J. Swanson, T. Tan, G. Tarlé, P. Taylor, G. Thomas, R. Tojeiro, R. J. Turner, W. Turner, L. A. Ureña-López, R. Vaisakh, M. Valluri, M. Vargas-Magaña, L. Verde, M. Walther, B. Wang, M. S. Wang, W. Wang, B. A. Weaver, N. Weaverdyck, R. H. Wechsler, M. White, M. Wolfson, J. Yang, C. Yèche, S. Youles, J. Yu, S. Yuan, E. A. Zaborowski, P. Zarrouk, H. Zhang, C. Zhao, R. Zhao, Z. Zheng, R. Zhou, H. Zou, S. Zou, and Y. Zu. Data release 1 of the dark energy spectroscopic instrument, 2025. URL `https://arxiv.org/abs/2503.14745`.

[11] Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall,

Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, and David J. Mills. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

[12] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[13] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323. URL `http://www.sciencemag.org/cgi/content/abstract/290/5500/2323`.

[14] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 06 2003. ISSN 0899-7667. doi: 10.1162/089976603321780317. URL `https://doi.org/10.1162/089976603321780317`.

[15] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. doi: 10.1007/BF02289565.

[16] Federico Marinacci, Mark Vogelsberger, Rüdiger Pakmor, Paul Torrey, Volker Springel, Lars Hernquist, Dylan Nelson, Rainer Weinberger, Annalisa Pillepich, Jill Naiman, and Shy Genel. First results from the IllustrisTNG simulations: radio haloes and magnetic fields. *Monthly Notices of the Royal Astronomical Society*, 480(4):5113–5139, November 2018. doi: 10.1093/mnras/sty2206.

[17] Annalisa Pillepich, Dylan Nelson, Lars Hernquist, Volker Springel, Rüdiger Pakmor, Paul Torrey, Rainer Weinberger, Shy Genel, Jill P. Naiman, Federico Marinacci, and Mark Vogelsberger. First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. *Monthly Notices of the Royal Astronomical Society*, 475(1):648–675, March 2018. doi: 10.1093/mnras/stx3112.

[18] Volker Springel, Rüdiger Pakmor, Annalisa Pillepich, Rainer Weinberger, Dylan Nelson, Lars Hernquist, Mark Vogelsberger, Shy Genel, Paul Torrey, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: matter and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 475(1):676–698, March 2018. doi: 10.1093/mnras/stx3304.

[19] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Rüdiger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: the galaxy colour bimodality. *Monthly Notices of the Royal Astronomical Society*, 475(1):624–647, March 2018. doi: 10.1093/mnras/stx3040.

[20] Jill P. Naiman, Annalisa Pillepich, Volker Springel, Enrico Ramirez-Ruiz, Paul Torrey, Mark Vogelsberger, Rüdiger Pakmor, Dylan Nelson, Federico Marinacci, Lars Hernquist, Rainer Weinberger, and Shy Genel. First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium. *Monthly Notices of the Royal Astronomical Society*, 477(1):1206–1224, June 2018. doi: 10.1093/mnras/sty618.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Comparison to DBSCAN for identifying galaxy clusters in TNG100

As an additional comparison, we run the DBSCAN [7] clustering algorithm (using the implementation from `scikit-learn` [21]) on the TNG100 dataset to identify galaxy clusters. We set the `eps` parameter, which denotes the maximum distance between two points to be considered in the same neighborhood, to the same neighborhood size as in our method (5 Mpc/h). We set the `min_samples` parameter, denoting the minimum number of points in a cluster for a point to be considered a core

point, to 50. This DBSCAN implementation returns an integer value for each point with a cluster label, or a value of -1 if a point is not determined to belong to any clusters.

Figure 5 shows the 21 galaxy clusters identified by DBSCAN and how the in- and out-of-cluster galaxies occupy unique regions of embedding space in the top two embedding dimensions from our method on the same dataset. The clusters identified by DBSCAN could likely be reproduced with reasonable fidelity by simply imposing a cut on a threshold value of our first embedding dimension, while our embedding space allows for more expressivity of the galaxy structures outside of just binary cluster labels.

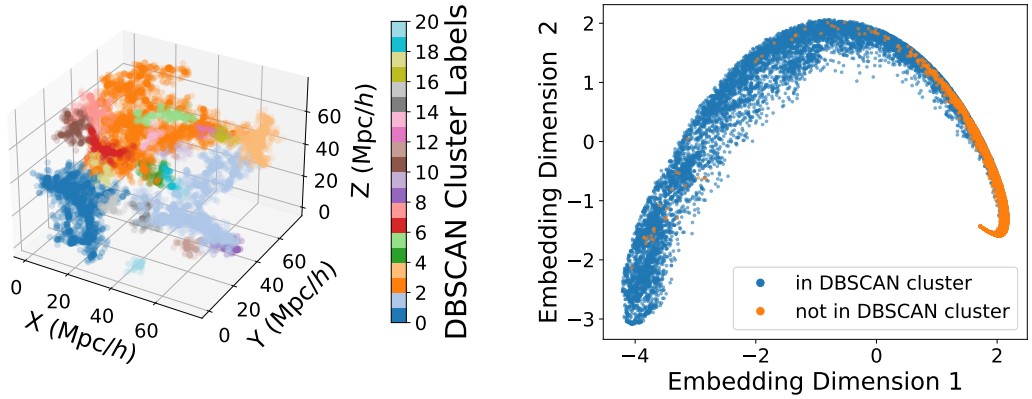

Figure 5: (left) 21 galaxy clusters identified by DBSCAN on the TNG100 dataset, each denoted by its own color. (right) The same dataset in the top two embedding dimensions from our method, with blue and orange points denoting galaxies inside and outside of clusters identified by DBSCAN, respectively.