Robust Halo Masses using HaloFlow with Domain Adaptation

Nikhil Garuda
Department of Astronomy
University of Texas at Austin
Austin, TX 78723
garuda@utexas.edu

ChangHoon Hahn
Department of Astronomy
University of Texas at Austin
Austin, TX 78723
changhoon.hahn@utexas.edu

Connor Bottrell
International Centre for Radio Astronomy Research
University of Western Australia
connor.bottrell@uwa.edu.au

Khee-Gan Lee
Kavli IPMU (WPI), UTIAS
The University of Tokyo
kglee@ipmu.jp

Abstract

Precise halo mass (M_h) measurements are crucial for cosmology and galaxy formation. HaloFlow [19] provides a new approach using simulation-based inference and state-of-the-art simulated galaxy images that can accurately measure M_h with significantly higher precision. However, HaloFlow requires a simulated training dataset, and is thus limited by domain shifts. In this work, we extend HaloFlow with unsupervised domain adaptation (DA) methods to improve the robustness of inferred M_h, HaloFlow$^{\rm DA}$. We implement two DA methods: a domain-adversarial network (DANN) and Maximum Mean Discrepancy (MMD) alignment. We test the performance of our DA methods on a suite of three different cosmological hydrodynamic simulations. To quantify the performance of DA improvement of HaloFlow we introduce a normalised bias metric beta, β. Our results show that DA significantly improves robustness: HaloFlow with MMD reduces β by 10-63%. These gains represent a key step toward applying HaloFlow$^{\rm DA}$ for reliable M_h inference on galaxy survey observations.

1 Introduction

Halo masses, M_h, play a key role in both cosmology and galaxy formation. The abundance of most massive halos that host galaxy clusters is sensitive to the expansion history and structure growth [34, 3, 25, 14], and serve as key observables for dark energy studies [2]. M_h also shapes galaxy properties and the circumgalactic medium [35, 29], making it critical for galaxy evolution studies and fast radio burst constraints on baryon content [24]. Yet, current methods for measuring M_h, such as weak lensing and satellite kinematics require deep imaging or rely on strong assumptions that introduce significant systematics [36].

To overcome these challenges, we use HaloFlow[1] [19], a simulation-based inference (SBI) framework that leverages photometric and morphological information of galaxies to estimate M_h. HaloFlow improves on earlier ML-based halo mass estimators [26, 33] by using neural density estimation for Bayesian inference and operating on state-of-the-art simulated galaxy images [7]. This forward modeling of galaxy images closely mimics real observations, enabling end-to-end inference pipelines.

Despite our use of highly realistic synthetic observations, they are based on simulations that ultimately assume some physical model, have finite resolution, and use subgrid approximations. When HaloFlow is applied to observations, limitations in our simulations will likely introduce domain

[1] https://github.com/galactic-ai/haloflow

shift — i.e., a mismatch between training and testing data distributions [21]. This mismatch can bias inferences and produce unreliable posteriors.

In this work, we aim to mitigate domain shift using two unsupervised domain adaptation (DA) techniques: domain adversarial neural networks (DANN, [15]) and maximum mean discrepancy (MMD, [17]). DANN uses adversarial training to align feature distributions between domains, while MMD minimises the distance between them. We demonstrate these techniques on a suite of three different cosmological hydrodynamical simulations: training on two and testing on the third. By addressing domain shift across simulations, this work provides a step toward applying HaloFlow$^{\mathrm{DA}}$ to real observational data, where domain mismatch is unknown.

2 Data

We use four cosmological hydrodynamical simulations to investigate domain shifts: TNG50 and TNG100 from the IllustrisTNG project [31, 27, 28], L100N1504 (Eagle100) from the EAGLE suite [30], and m100n1024 (Simba100) from the SIMBA simulations [12]. Each simulation evolves galaxy populations within large cosmological volumes with distinct subgrid models for star formation, stellar and AGN feedback, chemical enrichment, and gas cooling. These differences induce distribution shifts in galaxy properties across simulations. For the TNG simulations, we combine the TNG50 and TNG100 galaxies to form a TNG_ALL dataset[2] We focus on central galaxies at $z = 0.1$, selected based on the SUBFIND friends-of-friends (FoF) catalogs [11]. For each galaxy, we use stellar mass (M_*) and host halo mass (M_h) as target parameters θ for inference.

We generate realistic synthetic images of galaxies using a forward-modeling pipeline adapted from Bottrell et al. [7]. We apply the `SKIRT` [5, 9] dust radiative transfer code to the outputs of hydrodynamical simulations and produce noise-free rest-frame images by modeling stellar and gas emission within each galaxy. Stellar emission is modeled using Bruzual & Charlot [8] stellar population synthesis templates with a Chabrier [10] initial mass function. For young stars under 10 Myr, H-II region emission is included via the MAPPINGS III library [18], alongside a metallicity-dependent dust model assuming Milky Way grain properties. The `SKIRT` outputs are then processed with `RealSim` [6], which embeds galaxies into real Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP) backgrounds to add observational realism by including PSF convolution, pixel noise, and image blending to match the quality and systematics of HSC. For each galaxy, we construct four images based on four lines of sight arranged along the arms of a tetrahedron to capture orientation effects.

From the synthetic HSC images, we measure photometric and morphological features using `Galight` [13] as we would in observations. Each galaxy is fit with a single-component Sérsic profile to estimate magnitudes and effective radii. We also compute non-parametric morphology statistics: concentration (C), asymmetry (A), smoothness (S), Gini coefficient, M_{20}, residual asymmetry (A_{res}), and the inner surface brightness $\mu_{1\,\mathrm{kpc}}$. The final feature vector $\mathbf{x}$ has 55 dimensions and includes: (i) Sérsic magnitudes in the *grizy* bands, (ii) shape parameters such as $R_{\mathrm{eff}}, n, b/a$, and (iii) non-parametric morphology features. These features reflect realistic observational uncertainties, as they are measured directly from mock HSC-like images. The resulting dataset includes 34920 galaxies from all the simulations, which we use to train and evaluate our SBI models under various domain shift scenarios.

3 HaloFlow$^{\mathrm{DA}}$

We develop a pipeline[3] to infer robust posterior probability distribution of θ given observed features $\mathbf{x}$, $p(\theta \mid \mathbf{x})$. We extend the SBI framework in HaloFlow to incorporate a DA module that learns domain-invariant representations to improve robustness to domain shifts among simulations. We introduce two DA methods: adversarial training via DANN, and non-parametric alignment via MMD. Both these methods produce a compressed representation of $\mathbf{x}$, $c\mathbf{x}$, which is then used for HaloFlow$^{\mathrm{DA}}$, $p(\theta \mid c\mathbf{x})$.

HaloFlow: HaloFlow estimates the posterior using neural density estimators (NDEs). In particular, we use Masked Autoregressive Flows as implemented in the `sbi` Python package [16, 32]. The flow provides a bijective transformation between a simple Gaussian base distribution to the posterior distribution. It is trained by maximizing the total log-likelihood over simulated galaxy pairs (x, θ), using the Adam optimiser [22] [4]. During training, we apply early stopping after 20 epochs without

[2]We rescale TNG50 galaxies to match TNG100.

[3]DA models can run within a few minutes and NDEs take 18-24 hours to run on an M4 MacBook Air.

[4]This is equivalent to minimizing the Kulback-Leibler divergence between the NDE and the true posterior.

improvement on a held-out validation set (10% of the data). To finalize the NDE, we train 1000 flow configurations using `Optuna` [1] to tune flow depth, width, and learning rate. We select the top five based on validation performance and ensemble them. Ensembling flows with different hyperparameters improves the accuracy of the normalising flow [23, 4].

In addition to the NDE, we apply an implicit prior correction. In cosmological simulations, low-mass halos and galaxies are more abundant, so the learned posterior implicitly favors them. We correct this bias using a maximum entropy prior method [20] by reweighting posterior samples with importance weights proportional to $1/p(\theta)$. We refer readers to Hahn et al. [19] for further details.

Domain-Adversarial Neural Networks (DANNs): DANNs learn representations of $\mathbf{x}$ that predict θ while remaining invariant across simulation domains. It consists of a shared feature extractor, a parameter regressor, and a domain classifier. We train the extractor and regressor to minimize a weighted mean squared error loss $\mathcal{L}_{\rm reg}$[5] Meanwhile, the domain classifier minimizes a cross-entropy loss $\mathcal{L}_{\rm DA}$ to identify the simulation origin of each feature. A gradient reversal layer inverts the gradients from the domain classifier, encouraging the feature extractor to learn domain-invariant features. The total loss is $\mathcal{L} = \mathcal{L}_{\rm reg} + \lambda(\alpha)\mathcal{L}_{\rm DA}$, where $\lambda(\alpha) = \frac{2}{1+e^{-4.5\alpha}} - 1$ controls the influence of domain adaptation during training, and $\alpha \in [0, 1]$ represents normalized training progress, with α define as the current epoch divided by the total number of epochs.

Maximum Mean Discrepancy (MMD): MMD minimizes domain mismatch by aligning $\mathbf{x}$ distributions from the different simulations. We compute MMD using a Gaussian kernel, embedding $\mathbf{x}$ into a reproducing kernel Hilbert space (RKHS) and measuring the squared distance between their mean embeddings. During training, we compute the loss: $\mathcal{L}_{\rm DA} = \text{MMD}^2(A, C) + \text{MMD}^2(B, C)$, where A and B are batches from two simulations, and C is from a held-out third. This encourages alignment across all three domains, rather than overfitting to a single source-target pair. The total training loss is: $\mathcal{L} = \mathcal{L}_{\rm reg} + 0.5\ \mathcal{L}_{\rm DA}$, where $\mathcal{L}_{\rm reg}$ is the weighted MSE loss used in DANN. We use a constant scaling factor of 0.5 to balance regression and domain adaptation tasks.

4 Results

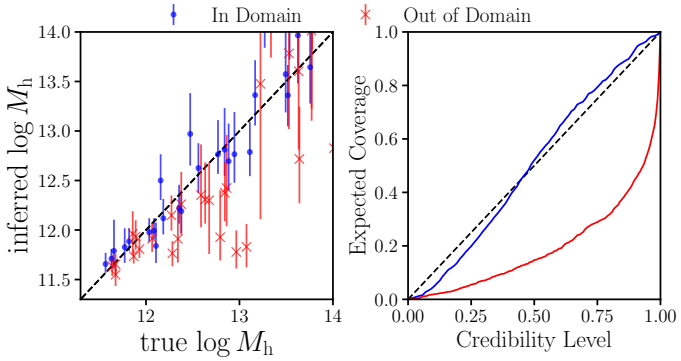

Figure 1: **Left:** Inferred vs. true M_h using HaloFlow trained on Simba100, evaluated on Simba100 (red) and TNG_ALL (blue). Predictions are accurate and well-calibrated in the in-domain case, but show systematic bias and increased scatter under domain shift. *Note: A subset of the test dataset is randomly sampled to visualize the results.* **Right:** Coverage plots for the same two scenarios. in-domain scenario shows near-ideal calibration, while out-of-domain shows poorer calibration, indicating overconfident and unreliable posteriors under simulation mismatch.

To evaluate the impact of model mismatch on posterior inference, we compare HaloFlow trained on Simba100 and tested on both Simba100 (in-domain) and TNG_ALL (out-of-domain) in Figure 1. HaloFlow accurately recovers M_h on the in-domain setting, but predictions systematically deviate on the out-of-domain setting, reflecting a clear model misspecification. Coverage plots confirms this and shows that the in-domain setting achieves near-ideal calibration, while the out-of-domain setting has under coverage, unreliable posteriors due to domain shift.

[5]We define $\mathcal{L}_{reg}$ as a weighted mean squared error, where each training sample is weighted by the inverse of its expected abundance. Specifically, weights combine inverse number counts from each simulation along with the inverse of Schechter function $\phi(M) = \phi_M (M/M_\star)^\alpha \exp(-M/M_\star)$ for each M_h.

To quantify this degradation, we define a normalised bias metric as $\beta = \frac{|m-\hat{m}|}{\sigma_{\hat{m}}}$ where $\hat{m}$ is the posterior median, m is the true M_h , and $\sigma_{\hat{m}}$ is the standard deviation of the posterior. We choose this metric because it standardizes the error by the model's uncertainty, offering a clear measure of prediction accuracy relative to confidence. A β near zero indicates accurate and well-calibrated predictions, while larger values reflect greater deviation or overconfidence. We compute the median β over the test set to summarize overall inference quality. For the in-domain setting, the median β is 0.56, indicating accurate predictions, while the out-of-domain setting yields a higher median β of 1.05, confirming that simulation mismatch significantly degrades inference.

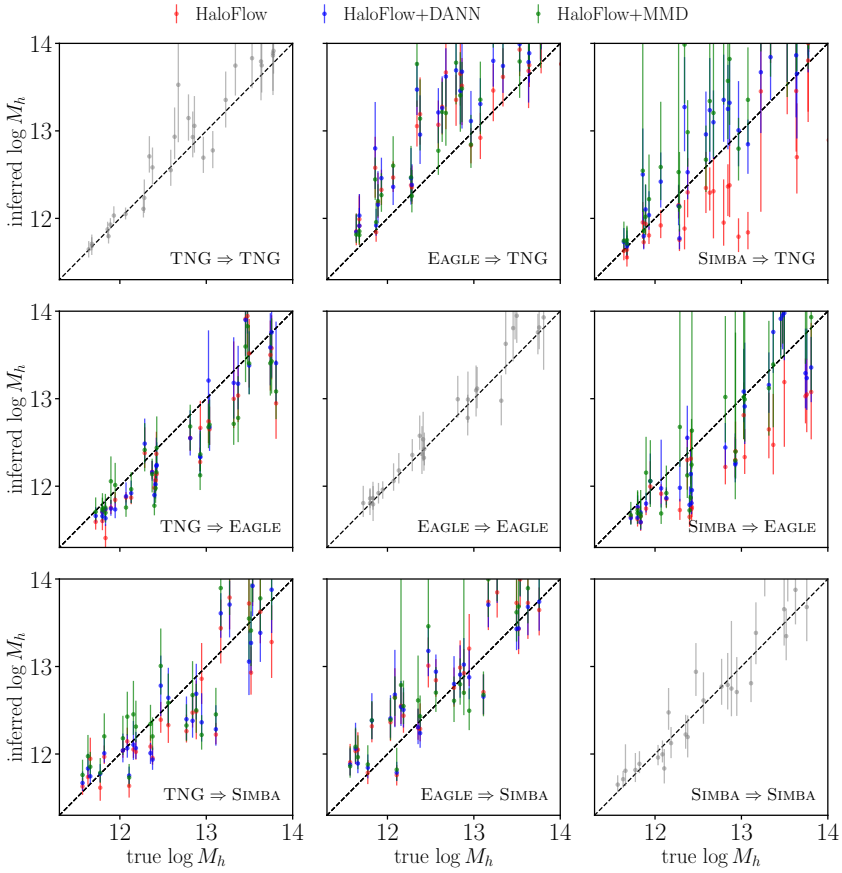

Figure 2: Inferred vs. true halo masses across all train/test simulation pairs. Each panel corresponds to one combination (column: train sim, row: test sim). Scatter points show inferred M_h from: HaloFlow only (red), HaloFlow + DANN (blue), and HaloFlow + MMD (green). Diagonal panels show accurate predictions with HaloFlow alone (gray). Off-diagonal panels reveal model mismatch for HaloFlow (red), while DANN and MMD (blue and green) improve robustness across simulations.

Figure 2 compares M_h inference across all train/test simulation pairs. Each panel shows one train→test combination, with scatter points representing posterior medians from three models: HaloFlow only (red), HaloFlow + DANN (blue), and HaloFlow + MMD (green). The diagonal panels show in-domain performance, where HaloFlow alone already achieves accurate predictions. Off-diagonal panels reveal domain shift effects, where HaloFlow (red points) shows large deviations from true halo masses. Both DANN and MMD (blue and green points) improve robustness by reducing these deviations across simulations. To quantify these improvements, we compute the median β for all train/test pairs and summarize the results in Table 1. Both DA methods significantly reduce β in cross-domain settings. MMD achieves the lowest β in all of the six off-diagonal cases, with the largest improvement seen in the Simba100 → Eagle100 setting, where MMD cuts the median β nearly in half, from 1.70 to 0.62 (63.5% improvement).

5 Summary and Next Steps

In this work, we demonstrated that domain shift between cosmological simulations can degrade the accuracy and calibration of simulation-based M_h inference using HaloFlow. By applying DA

Table 1: Median β for all train/test simulation pairs. Rows indicate test simulations, columns indicate training simulations, matching the layout of Figure 2. Values are reported as HaloFlow / HaloFlow + DANN / HaloFlow + MMD. Diagonal entries correspond to in-domain performance where domain adaptation is not applied. Bold values indicate best (lowest) β for each pair.

Test \ Train	**TNG_ALL**	**Eagle100**	**Simba100**
TNG_ALL	**0.57** / — / —	1.26 / 1.38 / **1.04**	1.05 / 0.86 / **0.57**
Eagle100	1.09 / 0.95 / **0.91**	**0.51** / — / —	1.70 / 1.14 / **0.62**
Simba100	1.77 / 1.26 / **0.82**	1.39 / 1.46 / **1.24**	**0.56** / — / —

techniques to compress observational features into domain-invariant representations, we substantially improve robustness under simulation mismatch, reducing bias and recovering more robust posteriors.

The ultimate goal of this work is to apply HaloFlow to real observational data from galaxy surveys and enabling accurate M_h inference. The HaloFlow forward modeling pipeline incorporates realistic observational and systematic effects by generating galaxy images from state-of-the-art simulations, closely mimicking real survey conditions. The DA extensions we introduce in this work, further improve the robustness of the SBI pipeline and bring us closer to deploying HaloFlow on observations.

In future work, we will further analyse and interpret the impact of the DA techniques on the final inferred M_h. As a next step, we will apply our method to galaxy images from the HSC survey and compare our inferred M_h with independent weak lensing estimates, validating our approach and benchmarking it against existing methods.

Acknowledgements

The authors acknowledge the Texas Advanced Computing Center (TACC)[6] at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper. This work was supported by resources provided by the Pawsey Supercomputing Research Centre's Setonix Supercomputer[7] and Acacia Object Storage[8] , with funding from the Australian Government and the Government of Western Australia.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, Anchorage AK USA, July 2019. ACM.

[2] Andreas Albrecht, Gary Bernstein, Robert Cahn, Wendy L. Freedman, Jacqueline Hewitt, Wayne Hu, John Huth, Marc Kamionkowski, Edward W. Kolb, Lloyd Knox, John C. Mather, Suzanne Staggs, and Nicholas B. Suntzeff. Report of the Dark Energy Task Force, September 2006. arXiv:astro-ph/0609591.

[3] Steven W. Allen, August E. Evrard, and Adam B. Mantz. Cosmological Parameters from Observations of Galaxy Clusters. *Annual Review of Astronomy and Astrophysics*, 49(1):409–470, September 2011.

[4] Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*, page stz1960, July 2019.

[5] Maarten Baes, Joris Verstappen, Ilse De Looze, Jacopo Fritz, Waad Saftly, Edgardo Vidal Pérez, Marko Stalevski, and Sander Valcke. EFFICIENT THREE-DIMENSIONAL NLTE DUST RADIATIVE TRANSFER WITH SKIRT. *The Astrophysical Journal Supplement Series*, 196(2):22, October 2011. Publisher: American Astronomical Society.

[6] `http://www.tacc.utexas.edu`

[7] `https://doi.org/10.48569/18sb-8s43`

[8] `https://doi.org/10.48569/nfe9-a426`

[6] Connor Bottrell, Maan H Hani, Hossen Teimoorinia, Sara L Ellison, Jorge Moreno, Paul Torrey, Christopher C Hayward, Mallory Thorp, Luc Simard, and Lars Hernquist. Deep learning predictions of galaxy merger stage and the importance of observational realism. *Monthly Notices of the Royal Astronomical Society*, 490(4):5390–5413, December 2019.

[7] Connor Bottrell, Hassen M Yesuf, Gergö Popping, Kiyoaki Christopher Omori, Shenli Tang, Xuheng Ding, Annalisa Pillepich, Dylan Nelson, Lukas Eisert, Hua Gao, Andy D Goulding, Boris S Kalita, Wentao Luo, Jenny E Greene, Jingjing Shi, and John D Silverman. IllustrisTNG in the HSC-SSP: image data release and the major role of mini mergers as drivers of asymmetry and star formation. *Monthly Notices of the Royal Astronomical Society*, 527(3):6506–6539, November 2023. Publisher: Oxford University Press (OUP).

[8] G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344(4):1000–1028, October 2003. arXiv:astro-ph/0309134.

[9] P. Camps and M. Baes. SKIRT 9: Redesigning an advanced dust radiative transfer code to allow kinematics, line transfer and polarization by aligned dust grains. *Astronomy and Computing*, 31:100381, April 2020. Publisher: Elsevier BV.

[10] Gilles Chabrier. Galactic Stellar and Substellar Initial Mass Function. *Publications of the Astronomical Society of the Pacific*, 115(809):763–795, July 2003. Publisher: IOP Publishing.

[11] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large-scale structure in a universe dominated by cold dark matter. *The Astrophysical Journal*, 292:371–394, May 1985.

[12] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H Rafieferantsoa, and Sarah Appleby. simba: Cosmological simulations with black hole growth and feedback. *Monthly Notices of the Royal Astronomical Society*, 486(2):2827–2849, June 2019. Publisher: Oxford University Press (OUP).

[13] Xuheng Ding, John Silverman, Tommaso Treu, Andreas Schulze, Malte Schramm, Simon Birrer, Daeseong Park, Knud Jahnke, Vardha N. Bennert, Jeyhan S. Kartaltepe, Anton M. Koekemoer, Matthew A. Malkan, and David Sanders. The Mass Relations between Supermassive Black Holes and Their Host Galaxies at $1 < z < 2$ with HST-WFC3. *The Astrophysical Journal*, 888(1):37, January 2020.

[14] Scott Dodelson, Katrin Heitmann, Chris Hirata, Klaus Honscheid, Aaron Roodman, Uroš Seljak, Anže Slosar, and Mark Trodden. Cosmic Visions Dark Energy: Science, April 2016. arXiv:1604.07626 [astro-ph].

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer International Publishing, Cham, 2017. Series Title: Advances in Computer Vision and Pattern Recognition.

[16] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic Posterior Transformation for Likelihood-Free Inference, May 2019. arXiv:1905.07488 [cs].

[17] Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A Kernel Method for the Two-Sample Problem, 2008. Version Number: 1.

[18] Brent Groves, Michael A. Dopita, Ralph S. Sutherland, Lisa J. Kewley, Jörg Fischera, Claus Leitherer, Bernhard Brandl, and Wil Van Breugel. Modeling the Pan-Spectral Energy Distribution of Starburst Galaxies. IV. The Controlling Parameters of the Starburst SED. *The Astrophysical Journal Supplement Series*, 176(2):438–456, June 2008. Publisher: American Astronomical Society.

[19] ChangHoon Hahn, Connor Bottrell, and Khee-Gan Lee. HaloFlow. I. Neural Inference of Halo Mass from Galaxy Photometry and Morphology. *The Astrophysical Journal*, 968(2):90, June 2024.

[20] Will Handley and Marius Millea. Maximum-Entropy Priors with Derived Parameters in a Specified Distribution. *Entropy*, 21(3):272, March 2019.

[21] Feng Hou, Jin Yuan, Ying Yang, Yang Liu, Yang Zhang, Cheng Zhong, Zhongchao Shi, Jianping Fan, Yong Rui, and Zhiqiang He. DomainVerse: A Benchmark Towards Real-World Distribution Shifts For Tuning-Free Adaptive Domain Generalization, March 2024. arXiv:2403.02714 [cs].

[22] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick, December 2015. arXiv:1506.02557 [stat].

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, November 2017. arXiv:1612.01474 [stat].

[24] Khee-Gan Lee, Ilya S. Khrykin, Sunil Simha, Metin Ata, Yuxin Huang, J. Xavier Prochaska, Nicolas Tejos, Jeff Cooke, Kentaro Nagamine, and Jielai Zhang. The FRB 20190520B Sight Line Intersects Foreground Galaxy Clusters. *The Astrophysical Journal Letters*, 954(1):L7, September 2023.

[25] Adam B. Mantz, Anja Von Der Linden, Steven W. Allen, Douglas E. Applegate, Patrick L. Kelly, R. Glenn Morris, David A. Rapetti, Robert W. Schmidt, Saroj Adhikari, Mark T. Allen, Patricia R. Burchat, David L. Burke, Matteo Cataneo, David Donovan, Harald Ebeling, Sarah Shandera, and Adam Wright. Weighing the giants – IV. Cosmology and neutrino mass. *Monthly Notices of the Royal Astronomical Society*, 446(3):2205–2225, January 2015.

[26] M. Ntampaka, H. Trac, D. J. Sutherland, N. Battaglia, B. Póczos, and J. Schneider. A MACHINE LEARNING APPROACH FOR DYNAMICAL MASS MEASUREMENTS OF GALAXY CLUSTERS. *The Astrophysical Journal*, 803(2):50, April 2015. Publisher: The American Astronomical Society.

[27] 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. Publisher: Oxford University Press (OUP).

[28] Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Rüdiger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, and Federico Marinacci. Simulating galaxy formation with the IllustrisTNG model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106, January 2018. Publisher: Oxford University Press (OUP).

[29] J Xavier Prochaska and Yong Zheng. Probing Galactic Halos with Fast Radio Bursts. *Monthly Notices of the Royal Astronomical Society*, January 2019.

[30] Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, January 2015. Publisher: Oxford University Press (OUP).

[31] 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. Publisher: Oxford University Press (OUP).

[32] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro Gonçalves, David Greenberg, and Jakob Macke. Sbi: A toolkit for simulation-based inference. *The Journal of Open Source Software*, 5:2505, August 2020.

[33] Pablo Villanueva-Domingo, Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, Federico Marinacci, David N. Spergel, Lars Hernquist, Mark Vogelsberger, Romeel Dave, and Desika Narayanan. Inferring halo masses with Graph Neural Networks, February 2023. arXiv:2111.08683 [astro-ph].

[34] G. Mark Voit. Tracing cosmic evolution with clusters of galaxies. *Reviews of Modern Physics*, 77(1):207–258, April 2005.

[35] Risa H. Wechsler and Jeremy L. Tinker. The Connection Between Galaxies and Their Dark Matter Halos. *Annual Review of Astronomy and Astrophysics*, 56(1):435–487, September 2018.

[36] R Wojtak, L Old, G A Mamon, F R Pearce, R de Carvalho, C Sifón, M E Gray, R A Skibba, D Croton, S Bamford, D Gifford, A von der Linden, J C Muñoz-Cuartas, V Müller, R J Pearson, E Rozo, E Rykoff, A Saro, T Sepp, and E Tempel. Galaxy Cluster Mass Reconstruction Project – IV. Understanding the effects of imperfect membership on cluster mass estimation. *Monthly Notices of the Royal Astronomical Society*, 481(1):324–340, November 2018.