Fermilab National Accelerator Labroatory

Bridging the gap between cosmological simulations with Graph Neural Networks and Domain Adaptation

Computational Science and AI Directorate

Andrea Roncoli

Date: 15 October 2023

Abstract

Deep learning models have been shown to outperform methods that rely on summary statistics, like the power spectrum, in extracting information from complex cosmological data sets. However, due to differences in the subgrid physics implementation and numerical approximations across different simulation suites, models trained on data from one cosmological simulation show a drop in performance when tested on another. Similarly, models trained on any of the simulations would also likely experience a drop in performance when applied to observational data. Training on data from two different suites of the CAMELS hydrodynamic cosmological simulations, we examine the generalization capabilities of Domain Adaptive Graph Neural Networks (DA-GNNs). By utilizing GNNs, we capitalize on their capacity to capture structured scale-free cosmological information from galaxy distributions. Moreover, by including unsupervised domain adaptation via Maximum Mean Discrepancy (MMD), we enable our models to extract domain-invariant features. We demonstrate that DA-GNN achieves higher accuracy and robustness on cross-dataset tasks (up to 28% better relative error and up to almost an order of magnitude better χ^2). Using data visualizations, we show the effects of domain adaptation on proper latent space data alignment. This shows that DA-GNNs are a promising method for extracting domain-independent cosmological information, a vital step toward robust deep learning for real cosmic survey data.

Contents

1 Introduction and Purpose

1.1 Background

The study of our cosmos, the vast expanse that surrounds us, has captivated human curiosity for centuries. Cosmology, the scientific endeavor to understand the fundamental structure, origin, evolution, and eventual fate of the universe, is a field that continually pushes the boundaries of human knowledge. Through millennia of observation, mathematical models, and increasingly sophisticated technology, we have gained remarkable insights into the workings of the universe.

In recent decades, technological advancements have ushered in a new era of cosmological research. The advent of powerful telescopes, high-performance computing, and data analysis techniques, including machine learning, have enabled a deeper understanding of the cosmos. Among these tools, deep learning has emerged as a formidable tool for data analysis and prediction in various domains, including cosmology.

1.2 The Need for Generalization through Domain Adaptation

Deep learning models have demonstrated exceptional capabilities in uncovering intricate patterns and extracting meaningful features from vast quantities of data. Within the realm of cosmology, these models have exhibited promising performance in extracting valuable insights from simulated cosmological datasets. However, a critical challenge lies in extending the applicability of these models to observational data, which represents the true nature of the universe. This challenge necessitates a critical need for generalization, leveraging the principles of domain adaptation and transfer learning.

1.2.1 Domain Adaptation

Domain adaptation addresses the disparity between the simulated (source) domain and the observational (target) domain. The source domain, typically the simulated data, contains distinct characteristics due to differences in subgrid physics implementation, numerical approximations, and other simulation-specific factors. On the other hand, the target domain, consisting of observational data, encompasses the complexities and nuances of the actual universe.

The core goal of domain adaptation is to bridge the gap between these domains, enabling the deep learning model to generalize its knowledge from the source to the target domain effectively. By minimizing the domain shift, where the source and target domains differ, domain adaptation ensures that the model's performance remains robust and accurate when applied to observational data.

1.3 The Quest for Understanding the Parameters of Our Universe

Understanding the fundamental parameters that define the universe is a central pursuit in cosmology. These parameters encompass a wide range of characteristics, such as the density of matter, the nature of dark energy, the initial conditions set at the Big Bang, and the distribution of matter and energy across the cosmos. Accurately inferring these parameters is essential for constructing comprehensive cosmological models that align with observational data, thereby advancing our comprehension of the universe's intricate tapestry.

The underlying motivation for this project is to devise methodologies that enable the extraction of precise and robust cosmological information from diverse datasets, including simulated and observational data. Achieving this goal is paramount in fortifying our understanding of the universe, bolstering the accuracy of cosmological models, and ultimately shedding light on the profound mysteries that govern our reality.

In the subsequent sections of this report, we delve into the approach undertaken, the methodology employed, and the results obtained in pursuit of this crucial endeavor.

2 Project Execution and Milestones

This section offers a chronological overview of the project's evolution, outlining the key milestones, challenges, and the overall timeline. It sheds light on how the project progressed, from its inception to its current state. The emphasis here is on providing a comprehensive understanding of how the project evolved over time, showcasing its growth and adaptation to various challenges.

The technical intricacies and detailed analysis of the project are comprehensively reported in the research paper that was submitted to the NeurIPS conference, which is presented in full in the subsequent section.

2.1 Understanding Graph Neural Networks and Domain Adaptation

In the initial phase of this project, a comprehensive study of relevant literature was conducted, focusing on Graph Neural Networks (GNNs), particularly Graph Convolutional Neural Networks (GCNs), and Domain Adaptation. Special attention was given to understanding Maximum Mean Discrepancy (MMD) and popular domain adaptation methods, such as Adversarial Discriminative Domain Adaptation (ADDA) and Domain Adaptive Neural Networks (DANN).

2.2 Code Familiarization and Infrastructure Setup

Upon acquiring a strong theoretical foundation, efforts transitioned to practical implementation. The CosmoGraphNet GitHub repository was pivotal in this regard, as our project is fundamentally an expansion of this previous work. The codebase was downloaded and meticulously studied to understand its inner workings, including model architectures, data preprocessing, and training processes. Moreover, part of the initial effort was invested in learning how to effectively utilize the Elastic Facility computing resources, made available through Fermilab. ChatGPT This step was crucial to harness the computational capabilities of high-performance GPUs, as training the models would have been impractical without this significant computational power.

2.3 Implementing Domain Adaptation with MMD

Implementing domain adaptation with Maximum Mean Discrepancy (MMD) required substantial modifications to the existing codebase. The primary objective was to enable the model to learn from samples across multiple simulations. Unlike traditional training that focuses on a single dataset, domain adaptation demands the model to generalize knowledge across different domains. This necessitated the incorporation of two distinct data loaders within the PyTorch training routine: one for each simulation.

The introduction of two data loaders allowed the model to simultaneously process samples from each simulation, facilitating a comprehensive understanding of the varying characteristics inherent to different simulations. Consequently, the training process involved optimizing a hybrid loss function. Alongside the pre-existing task-specific loss, an MMD-based loss was introduced to quantify the domain discrepancy. This supplementary loss was instrumental in guiding the model to align its learned features with domain-invariant information, a critical step towards achieving effective domain adaptation.

During backpropagation, the combined loss, comprising both the original task-specific loss and the newly introduced MMD-based loss, was utilized. The gradients from both components were computed and utilized to update the model's parameters. This intricate training scheme ensured that the model not only excelled in its primary task but also adapted effectively to the differing characteristics presented by distinct simulations, establishing a foundation for robust domain adaptation.

The successful implementation of this tailored training routine, integrating MMD as a guiding principle, significantly enhanced the model's ability to learn domain-invariant features, leading to superior generalization across diverse cosmological datasets.

2.4 Experimentation and Optimization Challenges

In the experimentation phase, significant effort was directed towards optimizing the models for superior performance. However, the initial attempts at running optimization on the models yielded results that were logically inconsistent, hinting at potential errors within the codebase. As is customary in intricate software development, encountering bugs and inconsistencies during the early stages is not uncommon. Recognizing this, it became evident that implementing a robust logging system was imperative to comprehensively track and analyze the evolution of the training curves and other pertinent statistics.

To address this, a detailed logging system was meticulously developed, offering a comprehensive view of model performance and aiding in identifying and rectifying the issues within the code. This logging system played a critical role in troubleshooting, allowing for a systematic exploration of potential errors and facilitating the necessary adjustments to the codebase.

Following this intensive debugging phase, the codebase was refined and stabilized, culminating in optimized and logically consistent model runs.

2.5 Paper Drafting and Conference Submissions

With the optimized models and promising results in hand, the focus transitioned towards presenting our findings to the academic community. We meticulously compiled the research and insights into a well-structured research paper, aiming to contribute to the field of cosmological data analysis and domain adaptation.

The research paper has been submitted to the esteemed NeurIPS conference, a leading platform for cutting-edge research in artificial intelligence. However, it's important to note that NeurIPS is currently in the process of reviewing submitted papers, and acceptance is yet to be confirmed. We are eagerly awaiting the results and the opportunity to share our work with the broader academic and research community.

In parallel, the research work gained recognition from the MLIAP conference, leading to the acceptance of our abstract for a full talk at the conference in Paris. This acknowledgment affirms the potential impact and relevance of our research, opening doors to engage with a diverse audience and foster collaboration within the scientific community.

3 Research Paper - Concise Technical Details

This section introduces the research paper, a product of this project and a submission to NeurIPS, a conference where submissions are limited to four pages. The paper is a dedicated exploration of the technical intricacies of the project, offering a detailed examination of methodologies, models, experiments, and results.

The technical details deliberately streamlined in the earlier sections, which primarily focused on the project's timeline and evolutionary journey, find their place here. This includes mathematical formulations of the losses, domain adaptation visualization plots, and various other technical aspects fundamental to the project.

The succinct nature of the paper, a result of adhering to the page limit set by the conference, does not compromise its depth. Instead, it provides a concentrated yet comprehensive understanding of the research, aiming to communicate the essence of the project's technical approach.

Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep learning models have been shown to outperform methods that rely on summary statistics, like the power spectrum, in extracting information from complex cosmological data sets. However, due to differences in the subgrid physics implementation and numerical approximations across different simulation suites, models trained on data from one cosmological simulation show a drop in performance when tested on another. Similarly, models trained on any of the simulations would also likely experience a drop in performance when applied to observational data. Training on data from two different suites of the CAMELS hydrodynamic cosmological simulations, we examine the generalization capabilities of Domain Adaptive Graph Neural Networks (DA-GNNs). By utilizing GNNs, we capitalize on their capacity to capture structured scale-free cosmological information from galaxy distributions. Moreover, by including unsupervised domain adaptation via Maximum Mean Discrepancy (MMD), we enable our models to extract domain-invariant features. We demonstrate that DA-GNN achieves higher accuracy and robustness on cross-dataset tasks (up to 28% better relative error and up to almost an order of magnitude better χ^2). Using data visualizations, we show the effects of domain adaptation on proper latent space data alignment. This shows that DA-GNNs are a promising method for extracting domain-independent cosmological information, a vital step toward robust deep learning for real cosmic survey data.

1 Introduction

Accurate determination of cosmological parameters using big data from astronomical surveys is a task of paramount importance in modern science. Historically, the extraction of valuable cosmological information has relied on computing summary statistics [32, 17, 16]. More recently, deep learning methods, such as 2D and 3D Convolutional Neural Networks (CNNs), showed great promise in extracting rich non-linear information that summary statistics struggle to capture [33, 40, 30]. However, CNNs lack scale-invariance, as their analysis is firmly anchored to the grid size of the convolutional kernels, while any information on scales below that is lost. Choosing a superfine grid to avoid information loss, though, would simply yield almost entirely zeros in case of sparse and irregular data, such as galaxy clusterings. Thus, CNNs result in an inadequate method for structured sparse data. In contrast, Graph Neural Networks (GNNs) [24, 4, 50, 47] can handle structured cosmic web data in a scale-free manner [42, 15]. As with any other model, the typical procedure is to train GNNs on labeled data (like simulations) and then infer cosmological parameters from unlabeled data (like observations). However, there is a significant risk of these models not generalizing in the presence of the domain shift between simulations and observations. Systematic biases have been demonstrated even in experiments that train and test on simulations with different subgrid physics [42]. Domain

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

adaptation (DA) techniques [12, 44, 19, 28] can be used to increase model robustness to this type of domain shift. Here we propose the use of Domain Adaptive Graph Neural Networks (DA-GNNs) and investigate the utility of distance-based DA losses i.e., Maximum Mean Discrepancy (MMD) [6]. MMD is an unsupervised DA technique because it does not require labeled data, which is paramount for future applications on observations. We show that our domain-adaptive models achieve stronger generalization across datasets than regular GNN models. Our work is a significant step towards building future models trained on simulations, yet robust enough to work on observational data.

Related Work GNNs have shown great potential for extracting information from large sparse datasets, such as the distribution of galaxies, galaxy clusters, and cosmic large-scale structure [26, 29, 42, 34, 43, 15]. Unfortunately, due to the complexity of most deep learning models, they often learn dataset-specific features, which renders them useless when testing on a different dataset (different simulations or astronomical observations). In astronomy, it has been shown that DA techniques applied to different types of CNNs can substantially improve model performance in cross-dataset applications [8, 11, 10, 38, 22, 2]. Recently, it has been shown that DA can be used on other types of deep learning algorithms such as GNNs [13, 25, 46, 48, 7, 45, 18]. However, DA on GNNs has not been used for any astrophysics or cosmology applications.

2 Data and Methods

Data We use galaxy catalogs from the CAMELS [39] magneto-hydrodynamic simulations, which follow the evolution of dark matter particles and fluid elements (baryons) from redshift z = 127 to z = 0. We use snapshots at z=0 from two different simulation suites: 1) IllusrisTNG [31] was generated with Arepo2 [1] and employs the IllustrisTNG subgrid physics model; 2) SIMBA [14] was generated with Gizmo3 [2] and employs the SIMBA subgrid physics model. Using two independent models and codebases to simulate galaxies, cosmic gas, and large-scale structure is critical to assess the generalization potential of the machine learning models. In particular, we use the LH set of both suites, which contains 1000 simulations evolved with different random seeds and different values of two cosmological parameters (total matter density Ω_m and the amplitude of density fluctuations σ_8) and four astrophysical parameters ($A_{SN1}, A_{SN2}, A_{AGN1}, A_{AGN2}$ related to supernovae efficiency and active galactic nuclei (AGN) feedback, respectively)[3]. We use the following features from the galaxy catalogs as input to our models: 3D positions, stellar mass, stellar radius, stellar metallicity, and maximum circular velocity.

Methods

Following [42], we generate graphs from 3D galaxy catalogs; these graphs are rotation and translation invariant with respect to the catalogs themselves. We later feed them as inputs to the DA-GNN, an architecture based on CosmoGraphNet with the addition of DA techniques. The model is composed of two parts. The first part is a graph encoder that transforms the graphs into a vector in the latent space through graph blocks [4]. The second part is a simple feedforward network that performs regression, predicting the posterior mean μ and standard deviation σ of the Ω_m cosmological parameter. This can be achieved by minimizing the following loss [27, 41]:

$$\mathcal{L}_{\mu,\sigma} = \log\Big(\sum_{i\in batch}(\Omega_{m,i}-\mu_i)^2\Big) + \log\Big(\sum_{i\in batch}((\Omega_{m,i}-\mu_i)^2-\sigma_i^2)^2\Big), \tag{1}$$

where $\Omega_{m,i}$ is the ground-truth value for the i-th sample in the training set batch, and μ_i and σ_i are the mean and standard deviation, respectively, predicted for sample i.

2.1 Domain Adaptation

Our objective is to create models that generalize across domains i.e., cosmology simulations with different subgrid physics implementations. To assess this, we train on IllustrisTNG and test on SIMBA – and vice versa. We experiment with the use of MMD, a distance-based DA technique. MMD measures the distance of two probability distributions, based on the notion of embedding probabilities in a reproducing kernel Hilbert space. We include an MMD-based component in the

[1]`https://arepo-code.org/`

[2]`http://www.tapir.caltech.edu/~phopkins/Site/GIZMO.html`

[3]CAMELS dataset documentation: `https://camels.readthedocs.io/en/latest/index.html`

network loss function, following [9, 49]. For two distributions Z^1 and Z^2 (with N samples each), this is calculated as:

$$\mathcal{L}_{MMD} = \log(\frac{1}{N-1}\sum_{i\neq j}^{N}[k(z_i^1, z_j^1) + k(z_i^2, z_j^2) - k(z_i^1, z_j^2) - k(z_i^2, z_j^1)]), \quad (2)$$

where k is the Gaussian Radial Basis Function kernel and z_q^p is the sample q of distribution p (Z^1 or Z^2) [6, 35, 23, 49, 9]. The loss is calculated on the latent space distributions produced by the graph encoder when processing samples from SIMBA and IllustrisTNG sets. Our final objective function is $\mathcal{L} = \mathcal{L}_{\mu,\sigma} + \lambda\mathcal{L}_{MMD}$, where $\lambda \geq 0$ controls the relative contribution of the MMD loss and is a hyperparameter of the model. We find that $\lambda \approx 0.1$ for the best-performing models in this work. The MMD component of the total loss causes the graph encoder to generate similar latent distributions for both simulations, which will improve the performance of the regressor on cross-dataset tasks.

Optimization and Computing Resources. We performed experiments on NVIDIA A100 40GB GPU. For each of the models, implemented using PyTorch Geometric [20], we perform a hyperparameter search using the Optuna library[1], with 50 trials per model. More details on code performance, model implementations, and selected hyperparameters can be found in the publicly available code[4].

2.2 Evaluation

We split both IllustrisTNG and SIMBA data into training/validation/testing sets with a proportion of 70%/15%/15%. During training, we save the final models at the epoch with the best validation score. For performance metrics, we use the mean relative error ϵ (reported in percentages), the coefficient of determination R^2, and the χ^2 (N = 150 test points), measured as:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}\frac{|\Omega_{m,i} - \mu_i|}{\Omega_{m,i}}, \quad R^2 = 1 - \frac{\sum_{i=1}^{N}(\Omega_{m,i} - \mu_i)^2}{\sum_{i=1}^{N}(\Omega_{m,i} - \overline{\Omega}_m)^2}, \quad \chi^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{(\Omega_{m,i} - \mu_i)^2}{\sigma_i^2}, \quad (3)$$

where $\overline{\Omega}_m$ is the mean of Ω_m value in the test set. A value of χ^2 close to 1 suggests that the standard deviations are correctly predicted and can be seen as minimizing the second term of Equation 1. A higher (lower) value can be seen as an underestimation (overestimation) of the uncertainties[3].

3 Results

DA-GNN achieves significantly better results (up to 28% better relative error ϵ and up to almost an order of magnitude better χ^2) on cross-domain generalization with respect to CosmoGraphNet, whilst achieving comparable results on the same domain test set [5], as shown in Table 1 and Figure 1. In [40], the authors were able to infer the value of Ω_m with higher cross-domain accuracy. However, that analysis utilizes the full matter surface density maps i.e., 2D images, instead of the full 3D galaxy distributions. In [15], the authors propose a GNN-based model that performs well cross-domain when trained on the Astrid simulation [5] alone. However, this apparent robustness is achieved by choosing Astrid as the training set and by using input features that are less subject to simulation code variability – galaxy positions and 1D velocities. When authors try training on other simulations or using more simulation-dependant parameters (e.g., stellar mass), cross-dataset performance drops significantly. Therefore, domain-shift robustness across different cosmological datasets requires DA.

Latent space organization Isomaps are two-dimensional projections of the multi-dimensional latent space [36]. Figure 1 shows the difference in the latent space structure without (top row) and with (bottom row) DA. Ellipses in the top right isomap highlight how the two distributions are encoded in

[4]GitHub repository will be added after the anonymous review stage.

[5]In [42], authors get slightly better results for the same domain, and slightly worse for the cross-domain tests. We impute these differences to choices such as batch sizes and optimization techniques we took due to computational and time constraints.

[6]In Appendix A, the IllustrisTNG counterpart of this plot is presented.

Table 1: Comparison of results: No Domain Adaptation (top) and MMD (bottom).

	I -> I			I -> S			S -> S			S -> I		
	R^2	ϵ	χ^2	R^2	ϵ	χ^2	R^2	ϵ	χ^2	R^2	ϵ	χ^2
NoDA	0.97	5.0	1.39	-1.04	43.8	59.43	0.97	5.2	1.79	0.22	25.0	185.54
MMD	0.97	4.7	1.12	**0.69**	**15.7**	**17.99**	0.97	5.9	1.54	**0.68**	**16.7**	**19.96**

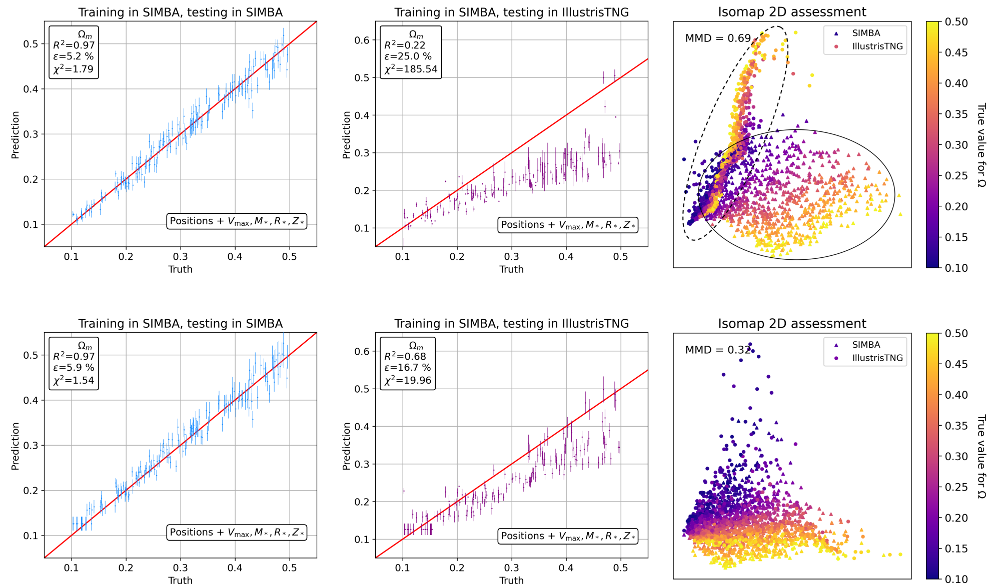

Figure 1: Comparison of models without (top row) and with DA (bottom row), trained on the SIMBA suite. From left to right, we report: scatter plot for the value of Ω_m on 1) same domain, 2) cross-domain and 3) the isomap showing how the GNN is encoding the two datasets in the latent space (SIMBA - triangles, IllustrisTNG - circles)[6]. In the non-domain adapted isomap, ellipses highlight regions where distributions lie, showing the difference between simulation encodings that leads to substantial drop in performance on the cross-domain task.

different regions of the latent space. Without the MMD loss, the model encodes samples with very different values of Ω_m close to each other, if they originate from different simulations (circles and triangles of different colors are overlapping). This scenario leads to the fragility of the regressor, which cannot learn to output different values for the same latent space encodings. On the contrary, the DA-GNN (bottom right plot) correctly encodes the samples in a domain-invariant way. Visually, circle and triangle distributions are overlapping, which indicates domain mixing. Furthermore, the direction in the color gradient shows that the DA-GNN encodes information such that the regressor can now more correctly predict cosmological parameters based on the encodings of both simulations.

4 Conclusions

We propose and demonstrate a method for unsupervised DA for cosmological inference with GNNs. We use an MMD-based loss to enable the domain-invariant encoding of features by the GNN. This approach enhances cross-domain robustness: compared to previous methods, DA-GNNs reduce prediction error and improve uncertainty estimates.

Limitations The cross-domain accuracy remains worse when compared to single-domain performance. Although reaching the same accuracy might not be possible, more flexible approaches such as adversarial-based DA techniques [21, 37], instead of distance-based ones such as MMD, might yield better results. Moreover, due to computational and time constraints, our models have been trained and tested only on two of the four available CAMELS simulation suites. Using more suites would yield better cross-domain efficacy and reliability at assessment time. These limitations will be addressed in future work.

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 and Data Mining*, 2019.

[2] Stephon Alexander, Sergei Gleyzer, Hanna Parul, Pranath Reddy, Marcos Tidball, and Michael W. Toomey. Domain Adaptation for Simulation-based Dark Matter Searches with Strong Gravitational Lensing. , 954(1):28, September 2023.

[3] Rene Andrae, Tim Schulze-Hartung, and Peter Melchior. Dos and don'ts of reduced chi-squared. *arXiv preprint arXiv:1012.3754*, 2010.

[4] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[5] Simeon Bird, Yueying Ni, Tiziana Di Matteo, Rupert Croft, Yu Feng, and Nianyi Chen. The ASTRID simulation: galaxy formation and reionization. , 512(3):3703–3716, May 2022.

[6] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[7] Ruichu Cai, Fengzhu Wu, Zijian Li, Pengfei Wei, Lingling Yi, and Kun Zhang. Graph domain adaptation: A generative view. *arXiv preprint arXiv:2106.07482*, 2021.

[8] A. Ćiprijanović, D. Kafkes, K. Downey, S. Jenkins, G. N. Perdue, S. Madireddy, T. Johnston, G. F. Snyder, and B. Nord. DeepMerge - II. Building robust deep learning algorithms for merging galaxy identification across domains. , 506(1):677–691, September 2021.

[9] A Ćiprijanović, Diana Kafkes, Kathryn Downey, Sudney Jenkins, Gabriel N Perdue, Sandeep Madireddy, Travis Johnston, Gregory F Snyder, and Brian Nord. Deepmerge–ii. building robust deep learning algorithms for merging galaxy identification across domains. *Monthly Notices of the Royal Astronomical Society*, 506(1):677–691, 2021.

[10] A. Ćiprijanović, A. Lewis, K. Pedro, S. Madireddy, B. Nord, G. N. Perdue, and S. M. Wild. DeepAstroUDA: semi-supervised universal domain adaptation for cross-survey galaxy morphology classification and anomaly detection. *Machine Learning: Science and Technology*, 4(2):025013, June 2023.

[11] Aleksandra Ćiprijanović, Diana Kafkes, Gregory Snyder, F. Javier Sánchez, Gabriel Nathan Perdue, Kevin Pedro, Brian Nord, Sandeep Madireddy, and Stefan M. Wild. DeepAdversaries: examining the robustness of deep learning models for galaxy morphology classification. *Machine Learning: Science and Technology*, 3(3):035007, September 2022.

[12] Gabriela Csurka. Domain Adaptation for Visual Applications: A Comprehensive Survey. *arXiv e-prints*, page arXiv:1702.05374, February 2017.

[13] Quanyu Dai, Xiao-Ming Wu, Jiaren Xiao, Xiao Shen, and Dan Wang. Graph transfer learning via adversarial domain adaptation with graph convolution. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4908–4922, 2022.

[14] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H. Rafieferantsoa, and Sarah Appleby. SIMBA: Cosmological simulations with black hole growth and feedback. , 486(2):2827–2849, June 2019.

[15] Natalí S. M. de Santi, Helen Shao, Francisco Villaescusa-Navarro, L. Raul Abramo, Romain Teyssier, Pablo Villanueva-Domingo, Yueying Ni, Daniel Anglés-Alcázar, Shy Genel, Elena Hernández-Martínez, Ulrich P. Steinwandel, Christopher C. Lovell, Klaus Dolag, Tiago Castro, and Mark Vogelsberger. Robust Field-level Likelihood-free Inference with Galaxies. , 952(1):69, July 2023.

[16] DES and SPT Collaborations, T. M. C. Abbott, M. Aguena, A. Alarcon, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, B. Ansarinejad, S. Avila, D. Bacon, and et al. Joint analysis of Dark Energy Survey Year 3 data and CMB lensing from SPT and Planck. III. Combined cosmological constraints. , 107(2):023531, January 2023.

[17] DES Collaboration, T. M. C. Abbott, M. Aguena, A. Alarcon, S. Allam, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, S. Avila, D. Bacon, and et. al. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. , 105(2):023520, January 2022.

[18] Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu. Graph adaptive knowledge transfer for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–52, 2018.

[19] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[20] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[22] Sankalp Gilda, Antoine de Mathelin, Sabine Bellstedt, and Guillaume Richard. Unsupervised Domain Adaptation for Constraining Star Formation Histories. *arXiv e-prints*, page arXiv:2112.14072, December 2021.

[23] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[24] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[25] Lukas Hedegaard Morsing, Omar Ali Sheikh-Omar, and Alexandros Iosifidis. Supervised Domain Adaptation using Graph Embedding. *arXiv e-prints*, page arXiv:2003.04063, March 2020.

[26] Yesukhei Jagvaral, François Lanusse, Sukhdeep Singh, Rachel Mandelbaum, Siamak Ravanbakhsh, and Duncan Campbell. Galaxies and haloes on graph neural networks: Deep generative modelling scalar and vector quantities for intrinsic alignment. , 516(2):2406–2419, October 2022.

[27] Niall Jeffrey and Benjamin D Wandelt. Solving high-dimensional parameter inference: marginal posterior densities & moment networks. *arXiv preprint arXiv:2011.05991*, 2020.

[28] Suruchi Kumari and Pravendra Singh. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *arXiv e-prints*, page arXiv:2308.01265, July 2023.

[29] T. Lucas Makinen, Tom Charnock, Pablo Lemos, Natalia Porqueres, Alan F. Heavens, and Benjamin D. Wandelt. The Cosmic Graph: Optimal Information Extraction from Large-Scale Structure using Catalogues. *The Open Journal of Astrophysics*, 5(1):18, December 2022.

[30] Michelle Ntampaka, Daniel J Eisenstein, Sihan Yuan, and Lehman H Garrison. A hybrid deep learning approach to cosmological constraints from galaxy redshift surveys. *The Astrophysical Journal*, 889(2):151, 2020.

[31] 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. , 473(3):4077–4106, January 2018.

[32] Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, and et al. Planck 2018 results. VI. Cosmological parameters. , 641:A6, September 2020.

[33] Dezső Ribli, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, and István Csabai. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society*, 490(2):1843–1860, 2019.

[34] Helen Shao, Francisco Villaescusa-Navarro, Pablo Villanueva-Domingo, Romain Teyssier, Lehman H Garrison, Marco Gatti, Derek Inman, Yueying Ni, Ulrich P Steinwandel, Mihir Kulkarni, et al. Robust field-level inference of cosmological parameters with dark matter halos. *The Astrophysical Journal*, 944(1):27, 2023.

[35] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.

[36] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000.

[37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[38] Ricardo Vilalta, Kinjal Dhar Gupta, Dainis Boumber, and Mikhail M. Meskhi. A General Approach to Domain Adaptation with Applications in Astronomy. , 131(1004):108008, October 2019.

[39] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. , 915(1):71, July 2021.

[40] Francisco Villaescusa-Navarro, Shy Genel, Daniel Angles-Alcazar, David N Spergel, Yin Li, Benjamin Wandelt, Leander Thiele, Andrina Nicola, Jose Manuel Zorrilla Matilla, Helen Shao, et al. Robust marginalization of baryonic effects for cosmological inference at the field level. *arXiv preprint arXiv:2109.10360*, 2021.

[41] Francisco Villaescusa-Navarro, Shy Genel, Daniel Angles-Alcazar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, et al. The camels multifield data set: Learning the universe's fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2):61, 2022.

[42] Pablo Villanueva-Domingo and Francisco Villaescusa-Navarro. Learning Cosmology and Clustering with Cosmic Graphs. , 937(2):115, October 2022.

[43] 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. *The Astrophysical Journal*, 935(1):30, 2022.

[44] Mei Wang and Weihong Deng. Deep Visual Domain Adaptation: A Survey. *arXiv e-prints*, page arXiv:1802.03601, February 2018.

[45] Man Wu, Shirui Pan, Chuan Zhou, Xiaojun Chang, and Xingquan Zhu. Unsupervised domain adaptive graph convolutional networks. In *Proceedings of The Web Conference 2020*, pages 1457–1467, 2020.

[46] Mengxi Wu and Mohammad Rostami. Unsupervised Domain Adaptation for Graph-Structured Data Using Class-Conditional Distribution Alignment. *arXiv e-prints*, page arXiv:2301.12361, January 2023.

[47] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[48] Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. *arXiv preprint arXiv:2306.04979*, 2023.

[49] Wen Zhang and Dongrui Wu. Discriminative joint probability maximum mean discrepancy (djp-mmd) for domain adaptation. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[50] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

A Additional Plots

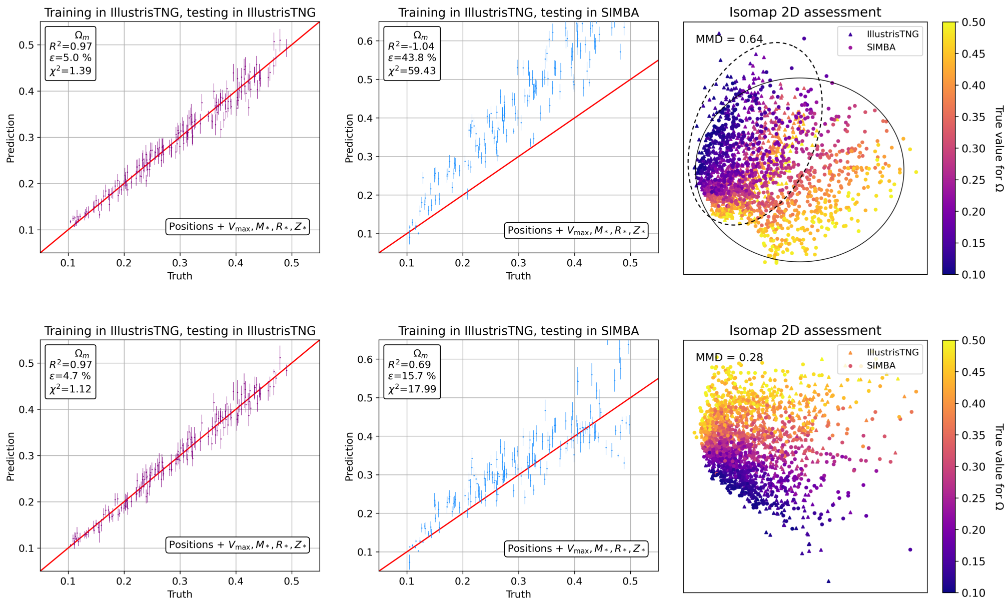

Figure 2: Comparison of models without (top row) and with DA (bottom row), trained on the IllustrisTNG suite. From left to right, we report: scatter plot for the value of Ω_m on 1) same domain, 2) cross-domain and 3) the isomap showing how the GNN is encoding the two datasets in the latent space (IllustrisTNG - triangles, SIMBA - circles). In the non-domain adapted isomap, ellipses highlight regions where distributions lie, showing the difference between simulation encodings that leads to substantial drop in performance on the cross-domain task.

4 Conclusions

This comprehensive report has highlighted the evolutionary journey and technical intricacies of our project, aiming to harness the potential of Domain Adaptive Graph Neural Networks (DA-GNNs) for robust cosmological data analysis. The project evolved through diligent stages, from extensive literature review and code familiarization to the implementation of domain adaptation techniques.

In our exploration, we studied Graph Neural Networks (GNNs) and various domain adaptation approaches, with a primary focus on Maximum Mean Discrepancy (MMD) as a key domain adaptation technique. The integration of MMD into our models allowed for the alignment of features across different cosmological simulations, aiding in the generalization of the models to diverse datasets.

The timeline and evolution of the project were meticulously outlined, emphasizing the significant challenges and subsequent optimizations encountered throughout. Debugging and optimization phases were pivotal in refining the models and achieving logically consistent results. A strong logging system was crucial to track training curves and aid in debugging.

The project culminated in the creation of a research paper, a condensed yet thorough technical documentation that encapsulates the essential aspects of the project. This paper was submitted to NeurIPS, presenting a focused view of the methodologies, results, and domain adaptation techniques employed. The concise format, adhering to the conference's page limit, underscored the need for clear and precise communication of technical details.

The acceptance of an abstract for a full talk at the MLIAP conference further affirms the project's significance and potential impact within the scientific community. This recognition serves as a stepping stone towards sharing our findings with a broader audience and fostering collaboration and knowledge exchange.

In conclusion, this project not only deepened our understanding of cutting-edge techniques in machine learning and domain adaptation but also showcased the potential of DA-GNNs in the domain of cosmological data analysis. The journey, marked by its challenges and triumphs, underscores the importance of innovation, collaboration, and relentless pursuit of scientific advancement.

5 Acknowledgements

I extend my heartfelt gratitude to Alexandra Ćiprijanović , whose kindness and friendliness have made this journey both enriching and enjoyable. As a colleague and someone to just have a chat with, your caring nature has been a constant source of encouragement.

This research endeavor has not only contributed to my growth as a researcher but also as an individual. I owe a significant part of this personal and professional development to the nurturing and supportive environment provided by the DeepSkies community. It has truly been an honor to be a part of this community. A special thanks goes to Brian Nord, whose unwavering support has been instrumental in shaping this community.

I would also like to express my gratitude to INFN and the University of Pisa for organizing the summer student project, and providing a platform for this enriching experience. Additionally, I extend my thanks to Fermilab, the institution I had the privilege to visit during this project, for their support and valuable resources.

Lastly, my appreciation goes to every individual with whom I had the privilege to share knowledge and experiences during various meetings, whether virtual or in person. These interactions have undeniably enriched this project, including the other Italian students who participated in the program and have become both great colleagues and friends and have been invaluable in making this experience something special.