Transfer Learning Beyond the Standard Model

Veena Krishnaraj[1] **Adrian E. Bayer**[2,1] **Christian Kragh Jespersen**[1] **Peter Melchior**[1]
[1]Princeton University, Princeton, NJ 08544, USA
[2]Flatiron Institute, New York, NY 10010, USA
{vk9342,abayer,ckragh,peter.melchior}@princeton.edu

Abstract

Machine learning enables powerful cosmological inference but typically requires many high-fidelity simulations covering many cosmological models. Transfer learning offers a way to reduce the simulation cost by reusing knowledge across models. We show that pre-training on the standard model of cosmology, ΛCDM, and fine-tuning on various beyond-ΛCDM scenarios—including massive neutrinos, modified gravity, and primordial non-Gaussianities—can enable inference with significantly fewer beyond-ΛCDM simulations. However, we also show that negative transfer can occur when strong physical degeneracies exist between ΛCDM and beyond-ΛCDM parameters. We consider various transfer architectures, finding that including bottleneck structures provides the best performance. Our findings illustrate the opportunities and pitfalls of foundation-model approaches in physics: pre-training can accelerate inference, but may also hinder learning new physics.

1 Introduction

Simulation-based inference (SBI) has been successfully adopted in cosmology to infer the standard model (ΛCDM) parameters from large-scale structure surveys [Hahn et al., 2024]. A key goal of Stage-IV surveys is to detect physics beyond the standard model—such as massive neutrinos, modified gravity, and primordial non-Gaussianities [DESI Collaboration et al., 2024]. Accurately testing beyond-ΛCDM extensions requires large suites of computationally expensive simulations, often far more expensive than their ΛCDM counterparts, creating a major bottleneck.

A promising way to alleviate this challenge is transfer learning, where knowledge acquired in one domain is reused to accelerate learning in another [Zhuang et al., 2020]. In cosmology, transfer learning has recently been applied between low and high fidelity simulations of the same underlying physics [Saoulis et al., 2025, Hikida et al., 2025, Thiele et al., 2025]. In this work, we ask a more ambitious question: **can transfer learning enable machine learning models to generalize to new physics?** Specifically, we investigate fine-tuning neural networks trained on ΛCDM to perform parameter inference beyond ΛCDM. In a sense, our study probes whether ΛCDM can serve as a foundation model upon which physics beyond the standard model can be fine-tuned.

Despite its promise, transfer learning has been shown to sometimes hinder performance in a phenomenon known as *negative transfer* [Zhang et al., 2023]. Whether transfer succeeds depends on the relationship between the source and target domains: if the target involves genuinely new physics not represented in the pre-trained model, or if strong parameter degeneracies obscure the relevant signals, transfer can fail or mislead. We thus explore different transfer architectures, including bottleneck or "dummy" units, to balance reuse of ΛCDM features with the flexibility to capture new physics.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

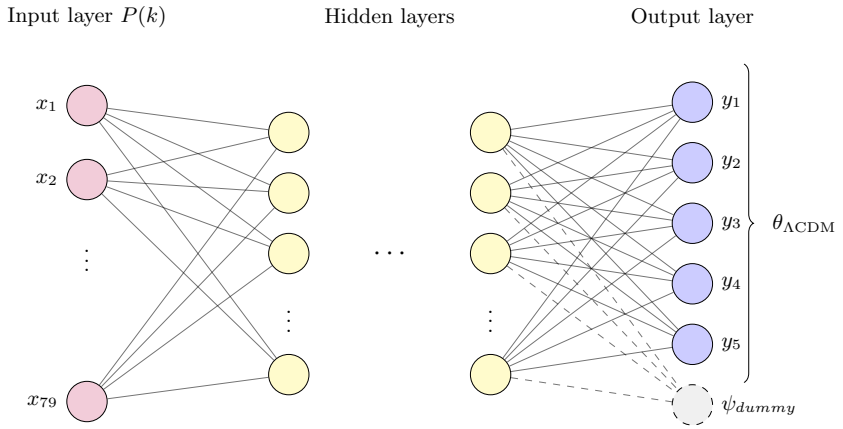


Figure 1: Dummy network architecture. The model takes the (marked) power spectrum $P(k)$ as input and outputs cosmological parameters $\theta_{\Lambda\mathrm{CDM}}$. Additional latent "dummy" nodes ψ_{dummy} are included in the output layer to provide extra representational capacity for fine-tuning.

2 Methods

We consider three different beyond-ΛCDM (fine-tuning) examples: massive neutrinos (M_ν), modified gravity ($f(R)$), and primordial non-Gaussianities ($f_{\mathrm{NL}}^{\mathrm{equilateral}}$ and $f_{\mathrm{NL}}^{\mathrm{local}}$). We use the Quijote simulations [Villaescusa-Navarro et al., 2020]. For the ΛCDM (pre-training) simulations, we vary 5 cosmological parameters: $\Omega_m \in [0.10,\ 0.50]$, $\Omega_b \in [0.02,\ 0.08]$, $h \in [0.50,\ 0.90]$, $n_s \in [0.80,\ 1.20]$, $\sigma_8 \in [0.60,\ 1.00]$, and fix $M_\nu = 0\mathrm{eV}$, $w = -1$, $f_{R0} = 0$, and $f_{\mathrm{NL}} = 0$. For each beyond-ΛCDM (fine-tuning) example, a separate Latin Hypercube of simulations is used where both the ΛCDM and beyond-ΛCDM parameters are varied (except in the case of local-f_{NL}, where only f_{NL} is varied, to assess the impact of a mismatch in the distribution of ΛCDM parameters during transfer). We provide a thorough description of the simulation setup in Appendix A.

We use a fully connected neural network to predict cosmological parameters from the matter power spectrum (or marked power spectrum [Massara et al., 2021]). The input to the network is a vector of 79 bins linearly spaced in the range $k \in [0.0089, 0.5]\ h/\mathrm{Mpc}$. All target parameters are linearly normalized to the range [0, 1]. The simulations for a given cosmology are divided into training, validation, and testing datasets, comprised of 70%, 15%, and 15% of the total dataset respectively. We further subsample the training set to investigate the performance as a function of the number of pre-training and fine-tuning simulations, while the validation and test sets remain fixed.

Our model consists of a fully connected neural network with up to three hidden layers, each consisting of a LeakyReLU activation (slope 0.2). A sigmoid activation function is applied to the output layer to match the $[0, 1]$ normalized targets. Training minimizes mean squared error (MSE) using the AdamW optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$), with batch size 32 and early stopping if validation loss does not improve by more than 10^{-6} after 50 epochs, with a maximum limit of 1000 epochs. We use Optuna [Akiba et al., 2019] to tune the number of layers (up to 3), neurons per hidden layer (4-500), learning rate, weight decay, and dropout, running 100 trials with TPE sampling and pruning.

We implement a two-stage transfer learning approach. First we train the network on the ΛCDM simulation set. Crucially, we include dummy nodes ψ_{dummy} in the pre-training network to output the same number of parameters as the corresponding beyond-ΛCDM model. For pre-training, the MSE is computed only using the ΛCDM parameters, thus the extra nodes are dummies. During pre-training we allow for learning rates in $[10^{-5}, 10^{-1}]$. In the second stage we fine-tune the network on the beyond-ΛCDM dataset with initialized weights from the pre-trained network, and using the dummy nodes for the beyond-ΛCDM parameters, reducing the learning rate range to $[10^{-6}, 10^{-3}]$. Fig 1 depicts our network setup.

Our choice of including dummy nodes in the pre-training is motivated by prior work in representation learning, where additional latent units, or bottleneck structures, can improve transferability and mitigate negative transfer [Yosinski et al., 2014, Bengio et al., 2014, Ho et al., 2023], and is conceptually related to the modular "head" architectures used in foundation models that enable flexible adaptation to diverse downstream tasks [Devlin et al., 2019, Radford et al., 2021]. We also investigated two

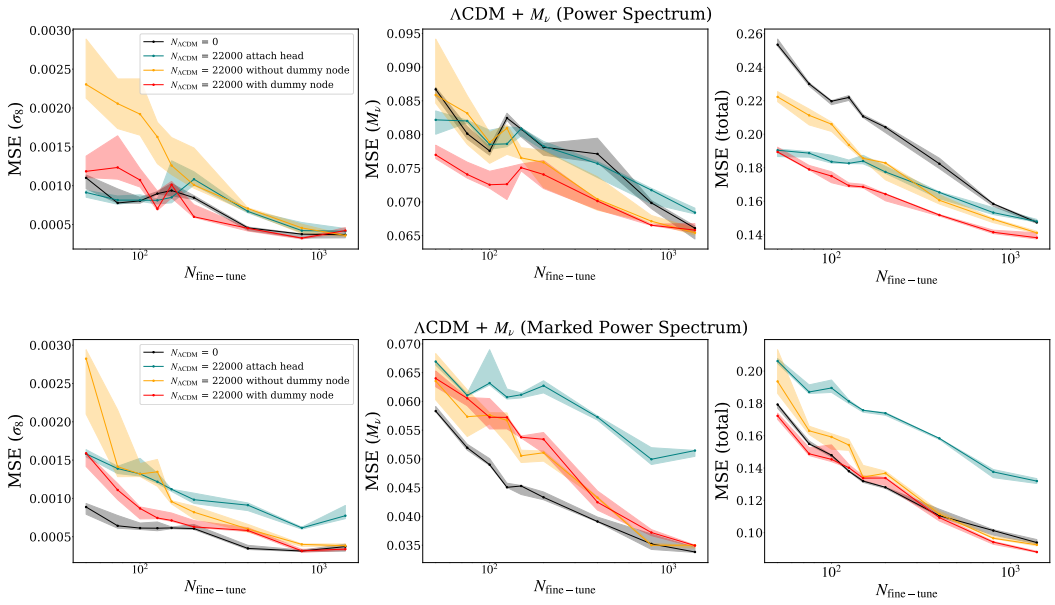


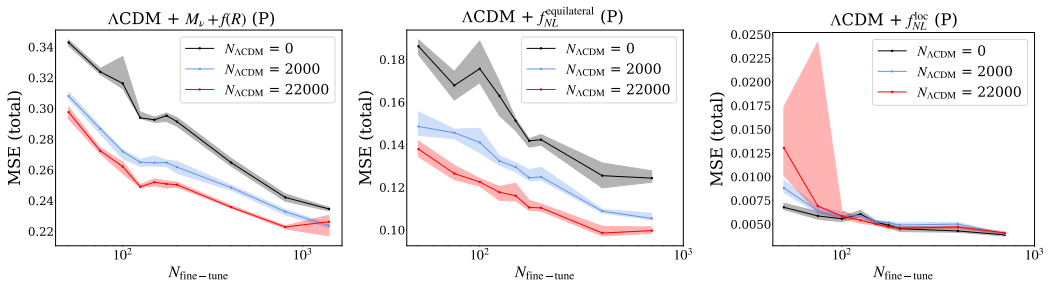
Figure 2: Test MSE as a function of the number of fine-tuning simulations for the massive neutrino cosmology using standard (top) and marked (bottom) power spectra for σ_8 (left), M_ν (center), and the total MSE across all normalized parameters (right). Transfer learning using a dummy node (red) always outperforms the result with no transfer learning (black) in terms of the total MSE, however negative transfer occurs for the marked power spectrum for σ_8 and M_ν due to the physical degeneracy between M_ν and σ_8. Other transfer learning architectures (teal, yellow) are suboptimal and result in more severe negative transfer.

other typical pre-training architectures which we found to be suboptimal: one without any dummy nodes, and one where we fix the pre-trained weights and attach a trainable inference head instead. Further details are provided in Appendix B.2.

3 Results

Fig 2 shows the test MSE as a function of the number of beyond-ΛCDM simulations used to train the fine-tuning network for the M_ν extension to ΛCDM. Each point represents the median MSE of the fine-tuning network's top 10 performing models, with error bars indicating the 16th and 84th percentiles. 22,000 simulations are used for pre-training. We consider the MSE on two individual parameters, σ_8 and M_ν, as well as the total MSE across all parameters. Transfer learning using a dummy node (red) always outperforms no transfer learning (black) in terms of the total MSE, with almost an order of magnitude less simulations required to achieve a given total MSE in the case of the power spectrum. However, in the case of the marked power, the MSE on σ_8 and M_ν is worse when performing transfer learning. This *negative transfer* occurs because of the physical degeneracy between σ_8 and M_ν [Bayer et al., 2021]—the pre-trained network has learned what features in the data to associate with σ_8 in the absence of neutrino mass, and then has to unlearn some of these features and associate them with M_ν upon fine-tuning. This occurs for the marked power which is very sensitive to σ_8 and M_ν [Massara et al., 2021], whereas the power spectrum alone is less sensitive [Bayer et al., 2021] and thus the introduction of $M_\nu > 0$ does not confuse the pre-trained network. We explicitly show this negative learning in by performing a feature analysis in Appendix B.3. Other transfer learning architectures—without the dummy node, or by attaching a head—perform worse in the limit of large number of simulations, and can cause an even larger negative transfer: in particular, head attachment suffers from negative transfer in terms of the total MSE, as the frozen weights of the pre-trained network enforce a representation which is too rigidly aligned with ΛCDM and thus the tuneable head is unable to transfer beyond it. However, for very few simulations ($< 10^2$) head attachment is of comparable quality to the other methods.

Having determined the dummy node approach to be best, we analyze the other beyond-ΛCDM scenarios with this method, exploring the effect of the number of pre-training simulations. Fig 3



(a) (b) (c)

Figure 3: Total MSE across all normalized parameters for modified gravity (left), equilateral (center), and local (right) primordial non-Gaussianity cosmologies. The colored lines represent different pre-training set sizes, which outperform the model trained directly on beyond-ΛCDM without transfer learning (black), except in the case of local $f_{\rm NL}$ due to the prior.

shows the total MSE for the remaining beyond-ΛCDM cosmologies. In the modified gravity case, results are similar to the massive neutrino case, with significant gains. There is also an increase in performance in the equilateral $f_{\rm NL}$ case, where degeneracies are mild. In the local non-Gaussianity case, the Quijote simulations do not vary the ΛCDM parameters, and only vary $f_{\rm NL}^{\rm local}$, thus transfer learning has little advantage—while this result is simply due to the simulation prior, we include the result to show the effect of different priors on the ΛCDM parameters between the pre-training and fine-tuning simulations. In all examples we find that even 2,000 pre-training simulations is enough to see benefits from transfer learning, with further improvements when using 22,000.

Full per-parameter MSE results are provided in Appendix B.1, Figs 4–6.

4 Discussion and Conclusions

In this work, we investigated the effectiveness of transfer learning for cosmological parameter inference beyond the standard ΛCDM model. Using a two-stage approach, we pre-trained neural networks on large ΛCDM simulation datasets and fine-tuned them on much smaller, computationally expensive beyond-ΛCDM simulations. We considered cosmologies with massive neutrinos, modified gravity, and primordial non-Gaussianities, using both power spectra and marked power spectra.

We find that transfer learning can reduce simulation requirements by up to an order of magnitude, but its success is dependent on the underlying parameter space. In models with fewer degeneracies—such as equilateral-type primordial non-Gaussianity—transfer learning improves inference across most parameters. By contrast, in scenarios with strong degeneracies, such as massive neutrino cosmologies where σ_8 and M_ν are entangled, transfer learning can lead to negative transfer, particularly when using a summary which is very sensitive to σ_8 and M_ν. Among the architectures tested, we found that introducing additional latent units, or dummy nodes, provided the most optimal performance. Multi-fidelity transfer may also improve performance [Thiele et al., 2025].

Our study focused on a simple fully connected network, but we expect qualitatively similar conclusions for more expressive architectures such as normalizing flows predicting the full posterior distribution, which would be a natural extension to test. Moreover, while we restricted our analysis to matter power spectra, applying transfer learning to observables such as galaxy clustering or weak lensing would be fruitful future work—in some cases this may yield greater gains, as, for example in the neutrino mass case, these observables have reduced sensitivity to M_ν [Bayer et al., 2022], making an easier transfer task. Looking beyond cosmology, this analysis could inform other areas of fundamental physics, such as learning extensions beyond the Standard Model of particle physics.

Overall, our results suggest that transfer learning can accelerate inference beyond the standard model, but its effectiveness hinges on parameter degeneracies, the choice of data summary, and the choice of architecture. More broadly, they illustrate both the promise and the pitfalls of foundation models for physics: pre-training on large standard-model datasets can dramatically reduce costs, but may also bias representations in ways that hinder the discovery of new physics if not carefully safeguarded.

References

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.

Adrian E. Bayer, Francisco Villaescusa-Navarro, Elena Massara, Jia Liu, David N. Spergel, Licia Verde, Benjamin D. Wandelt, Matteo Viel, and Shirley Ho. Detecting Neutrino Mass by Combining Matter Clustering, Halos, and Voids. *Astrophys. J.*, 919(1):24, 2021. doi: 10.3847/1538-4357/ac0e91.

Adrian E. Bayer, Arka Banerjee, and Uroš Seljak. Beware of fake ν 's: The effect of massive neutrinos on the nonlinear evolution of cosmic structure. *Physical Review D*, 105(12):123510, June 2022. doi: 10.1103/PhysRevD.105.123510.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014. URL `https://arxiv.org/abs/1206.5538`.

William R Coulton, Francisco Villaescusa-Navarro, Drew Jamieson, Marco Baldi, Gabriel Jung, Dionysios Karagiannis, Michele Liguori, Licia Verde, and Benjamin D. Wandelt. Quijote-png: Simulations of primordial non-gaussianity and the information content of the matter field power spectrum and bispectrum. *The Astrophysical Journal*, 943(1):64, January 2023. ISSN 1538-4357. doi: 10.3847/1538-4357/aca8a7. URL `http://dx.doi.org/10.3847/1538-4357/aca8a7`.

DESI Collaboration, A. G. Adame, J. Aguilar, S. Ahlen, S. Alam, D. M. Alexander, C. Allende Prieto, M. Alvarez, O. Alves, A. Anand, U. Andrade, E. Armengaud, S. Avila, A. Aviles, H. Awan, B. Bahr-Kalus, S. Bailey, C. Baltay, A. Bault, J. Behera, S. BenZvi, F. Beutler, D. Bianchi, C. Blake, R. Blum, M. Bonici, S. Brieden, A. Brodzeller, D. Brooks, E. Buckley-Geer, E. Burtin, R. Calderon, R. Canning, A. Carnero Rosell, R. Cereskaite, J. L. Cervantes-Cota, S. Chabanier, E. Chaussidon, J. Chaves-Montero, D. Chebat, S. Chen, X. Chen, T. Claybaugh, S. Cole, A. Cuceu, T. M. Davis, K. Dawson, A. de la Macorra, A. de Mattia, N. Deiosso, A. Dey, B. Dey, Z. Ding, P. Doel, J. Edelstein, S. Eftekharzadeh, D. J. Eisenstein, W. Elbers, A. Elliott, P. Fagrelius, K. Fanning, S. Ferraro, J. Ereza, N. Findlay, B. Flaugher, A. Font-Ribera, D. Forero-Sánchez, J. E. Forero-Romero, C. S. Frenk, C. Garcia-Quintero, L. H. Garrison, E. Gaztañaga, H. Gil-Marín, S. Gontcho A Gontcho, A. X. Gonzalez-Morales, V. Gonzalez-Perez, C. Gordon, D. Green, D. Gruen, R. Gsponer, G. Gutierrez, J. Guy, B. Hadzhiyska, C. Hahn, M. M. S Hanif, H. K. Herrera-Alcantar, K. Honscheid, C. Howlett, D. Huterer, V. Iršič, M. Ishak, R. Joyce, S. Juneau, N. G. Karaçaylı, R. Kehoe, S. Kent, D. Kirkby, H. Kong, S. E. Koposov, A. Kremin, A. Krolewski, O. Lahav, Y. Lai, T. W. Lan, M. Landriau, D. Lang, J. Lasker, J. M. Le Goff, L. Le Guillou, A. Leauthaud, M. E. Levi, T. S. Li, K. Lodha, C. Magneville, M. Manera, D. Margala, P. Martini, W. Matthewson, M. Maus, P. McDonald, L. Medina-Varela, A. Meisner, J. Mena-Fernández, R. Miquel, J. Moon, S. Moore, J. Moustakas, N. Mudur, E. Mueller, A. Muñoz-Gutiérrez, A. D. Myers, S. Nadathur, L. Napolitano, R. Neveux, J. A. Newman, N. M. Nguyen, J. Nie, G. Niz, H. E. Noriega, N. Padmanabhan, E. Paillas, N. Palanque-Delabrouille, J. Pan, S. Penmetsa, W. J. Percival, M. M. Pieri, M. Pinon, C. Poppett, A. Porredon, F. Prada, A. Pérez-Fernández, I. Pérez-Ràfols, D. Rabinowitz, A. Raichoor, C. Ramírez-Pérez, S. Ramirez-Solano, M. Rashkovetskyi, C. Ravoux, M. Rezaie, J. Rich, A. Rocher, C. Rockosi, N. A. Roe, A. Rosado-Marin, A. J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, L. Samushia, E. Sanchez, C. Saulder, E. F. Schlafly, D. Schlegel, M. Schubnell, H. Seo, A. Shafieloo, R. Sharples, J. Silber, A. Slosar, A. Smith, D. Sprayberry, T. Tan, G. Tarlé, P. Taylor, S. Trusov, R. Vaisakh, D. Valcin, F. Valdes, G. Valogiannis, M. Vargas-Magaña, L. Verde, M. Walther, B. Wang, M. S. Wang, B. A. Weaver, N. Weaverdyck, R. H. Wechsler, D. H. Weinberg, M. White, M. J. Wilson, L. Yi, J. Yu, Y. Yu, S. Yuan, C. Yèche, E. A. Zaborowski, P. Zarrouk, H. Zhang, C. Zhao, R. Zhao, R. Zhou, T. Zhuang, and H. Zou. Desi 2024 vii: Cosmological constraints from the full-shape modeling of clustering measurements, 2024. URL `https://arxiv.org/abs/2411.12022`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

ChangHoon Hahn et al. Cosmological constraints from non-Gaussian and nonlinear galaxy clustering using the SimBIG inference framework. *Nature Astron.*, 8(11):1457–1467, 2024. doi: 10.1038/s41550-024-02344-2.

Yuga Hikida, Ayush Bharti, Niall Jeffrey, and François-Xavier Briol. Multilevel neural simulation-based inference. 6 2025.

Matthew Ho, Xiaosheng Zhao, and Benjamin Wandelt. Information-ordered bottlenecks for adaptive semantic compression, 2023. URL https://arxiv.org/abs/2305.11213.

Wayne Hu and Ignacy Sawicki. Models of f(R) cosmic acceleration that evade solar system tests. *Physical Review D*, 76(6), September 2007. ISSN 1550-2368. doi: 10.1103/physrevd.76.064004. URL http://dx.doi.org/10.1103/PhysRevD.76.064004.

Elena Massara, Francisco Villaescusa-Navarro, Shirley Ho, Neal Dalal, and David N. Spergel. Using the Marked Power Spectrum to Detect the Signature of Neutrinos in Large-Scale Structure. *Phys. Rev. Lett.*, 126(1):011301, 2021. doi: 10.1103/PhysRevLett.126.011301.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Alex A. Saoulis, Davide Piras, Niall Jeffrey, Alessio Spurio Mancini, Ana M. G. Ferreira, and Benjamin Joachimi. Transfer learning for multifidelity simulation-based inference in cosmology, 2025. URL https://arxiv.org/abs/2505.21215.

Leander Thiele, Adrian E. Bayer, and Naoya Takeishi. Simulation-Efficient Cosmological Inference with Multi-Fidelity SBI. 7 2025.

Francisco Villaescusa-Navarro, ChangHoon Hahn, Elena Massara, Arka Banerjee, Ana Maria Delgado, Doogesh Kodi Ramanah, Tom Charnock, Elena Giusarma, Yin Li, Erwan Allys, Antoine Brochard, Cora Uhlemann, Chi-Ting Chiang, Siyu He, Alice Pisani, Andrej Obuljen, Yu Feng, Emanuele Castorina, Gabriella Contardo, Christina D. Kreisch, Andrina Nicola, Justin Alsing, Roman Scoccimarro, Licia Verde, Matteo Viel, Shirley Ho, Stephane Mallat, Benjamin Wandelt, and David N. Spergel. The quijote simulations. *The Astrophysical Journal Supplement Series*, 250(1):2, August 2020. ISSN 1538-4365. doi: 10.3847/1538-4365/ab9d82. URL http://dx.doi.org/10.3847/1538-4365/ab9d82.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014. URL https://arxiv.org/abs/1411.1792.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, February 2023. ISSN 2329-9274. doi: 10.1109/jas.2022.106004. URL http://dx.doi.org/10.1109/JAS.2022.106004.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2020. URL https://arxiv.org/abs/1911.02685.

A Data Description

Here we provide a thorough description of the Quijote simulations[1] [Villaescusa-Navarro et al., 2020]. For the ΛCDM (pre-training) simulation, we use the Big Sobol Sequence (BSQ) suite. BSQ consists of 32,768 simulations described by 5 cosmological parameters with varying values: $\Omega_m \in [0.10,\ 0.50]$, $\Omega_b \in [0.02,\ 0.08]$, $h \in [0.50,\ 0.90]$, $n_s \in [0.80,\ 1.20]$, $\sigma_8 \in [0.60,\ 1.00]$. In all simulations $M_\nu = 0\text{eV}$, $w = -1$, $f_{R0} =$, and $f_{\rm NL} = 0$.

The three different beyond-ΛCDM (fine-tuning) simulations setups are described as follows:

- For M_ν we use 2,000 Latin Hypercube simulations which vary M_ν in the range $M_\nu \in [0.01,\ 1.0]\text{eV}$ and w in the range $w \in [-1.3,\ -0.7]$. While w is varied, it cannot be constrained with a single redshift snapshot in real space, as it only affects the cosmological background, so we do not perform inference on w here. Initial tests confirmed that the

[1] https://quijote-simulations.readthedocs.io/

network failed to learn any meaningful information about w, and including it in the inference task only appeared to weaken performance. The ranges of the five ΛCDM parameters match those of the BSQ, with the exception of $\Omega_b \in [0.03,\ 0.07]$. We consider two summary statistics in this case, the power spectrum P and the marked power spectrum MP, as it has been shown that MP is much more constraining on M_ν compared to P Massara et al. [2021]: this enables comparison in transfer learning on the amount of information in the summary.

- In the case of modified gravity, the Quijote simulations use a Hu and Sawicki $f(R)$ model Hu and Sawicki [2007] where the Einstein-Hilbert action is extended by a function of the Ricci scalar, introducing a scalar degree of freedom that modifies gravity on large scales. For $f(R)$, we use 2,048 simulations with the ΛCDM parameters following the same ranges as in BSQ, with the addition of $M_\nu \in [0.01,\ 1.0]$eV and $f_{\rm R0} \in [-3 \times 10^{-4},\ 0]$. With modified gravity there are two definitions of σ_8: one corresponding to the GR underlying cosmology (σ_8(LCDM)) and another reflecting the full modified gravity model (σ_8(MG)). For the purpose of transfer learning, we perform inference on σ_8(LCDM) for consistency.
- For $f_{\rm NL}$ we consider local ($f_{\rm NL}^{\rm local}$) and equilateral ($f_{\rm NL}^{\rm equilateral}$) using the Quijote-PNG suite Coulton et al. [2023]. For each, we use a Latin hypercube with 1,000 simulations. The local set fixes $\Omega_m = 0.3175$, $\Omega_b = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $\sigma_8 = 0.834$, and $M_\nu = 0$ eV, while varying $f_{\rm NL}^{\rm local} \in [-300,\ 300]$. The equilateral set keeps $\Omega_b = 0.049$ and $M_\nu = 0$eV fixed, and varies the remaining ΛCDM parameters as in BSQ, along with $f_{\rm NL}^{\rm equilateral} \in [-600,\ 600]$. This allows us to test the effects of transfer learning when the prior on ΛCDM parameters differs between the two sets.

All simulations follow the evolution of 512^3 dark matter particles in a periodic comoving volume of $(1\ h^{-1}\text{Gpc})^3$, with initial conditions generated at $z = 127$ and evolved using Gadget-III. Simulations that include massive neutrinos add an additional 512^3 neutrino particles. Although the parameter ranges are mostly consistent across cosmologies, Ω_b varies slightly; for normalization consistency across models, we adopt the broader range defined by the ΛCDM dataset for normalizations.

B Additional Results

B.1 MSE for all parameters

Here we provide results and discussion of the MSE on all the cosmological parameters for all the different examples considered in the paper. We also test further options for the number of pre-training simulations.

For massive neutrinos with the standard power spectrum (Figure 4a), transfer learning modestly improves performance for some ΛCDM parameters when training data is exceptionally scarce. However, it offers little to no benefit for σ_8 and M_ν, even at low simulation counts. At larger training set sizes, training both with and without transfer learning yields similar results.

A similar trend is observed when using the marked power spectrum (Figure 4b), where all parameters show either some improvements or comparable performance when using transfer learning. However, unlike the standard power spectrum, transfer learning does not offer a significant advantage at any number of training simulations. Furthermore, at low numbers of beyond-ΛCDM simulations, transfer learning performs noticeably worse than training from scratch for M_ν and σ_8 in particular. This decline is likely driven by degeneracy between M_ν, σ_8, and Ω_m. During pre-training, the marked power spectrum learns a precise knowledge of σ_8 and Ω_m which then has to be unlearned in order to recognize the effects of M_ν when it is introduced. This is an example of *negative transfer*. We do not observe the same behavior for the power spectrum because it is not informative of M_ν on its own Bayer et al. [2022] and thus the introduction of $M_\nu > 0$ does not confuse the pre-trained network.

We now consider the combination of massive neutrinos and modified gravity in Figure 5a. Transfer learning provides a noticeable advantage, particularly when training data is limited, mirroring trends seen in 4a. For f_{R0}, M_ν and σ_8, performance is nearly identical with and without transfer learning – consistent with earlier power spectrum results for massive neutrinos, and again likely due to degeneracies between these parameters. However, unlike the marked power spectrum case, no significant performance drop is observed here, reinforcing the idea that the decline there stems from the marked power spectrum's added sensitivity to M_ν.

For equilateral-type primordial non-Gaussianity (Figure 5b), where all parameters except Ω_b are varied, transfer learning consistently outperforms the baseline for all parameters except σ_8 and Ω_b, where performance is similar. This suggests that while degeneracy between σ_8 and $f_{\rm NL}^{\rm equilateral}$ may limit gains for those parameters, the influence of $f_{\rm NL}$ on the other cosmological parameters is minimal, allowing transfer learning to provide positive results in those cases.

For the local primordial non-Gaussianity case (Figure 5c) – where all ΛCDM parameters are fixed across the fine-tuning data – transfer learning offers no benefit over direct training on the beyond-ΛCDM dataset, with nearly identical results across all parameters. Since the fine-tuning task only involves learning the effect of $f_{\rm NL}^{\rm local}$ on the power spectrum, transfer learning appears unnecessary and ineffective, but does also not hinder performance.

B.2 Alternative architectures

We also tested two alternative pre-training setups:

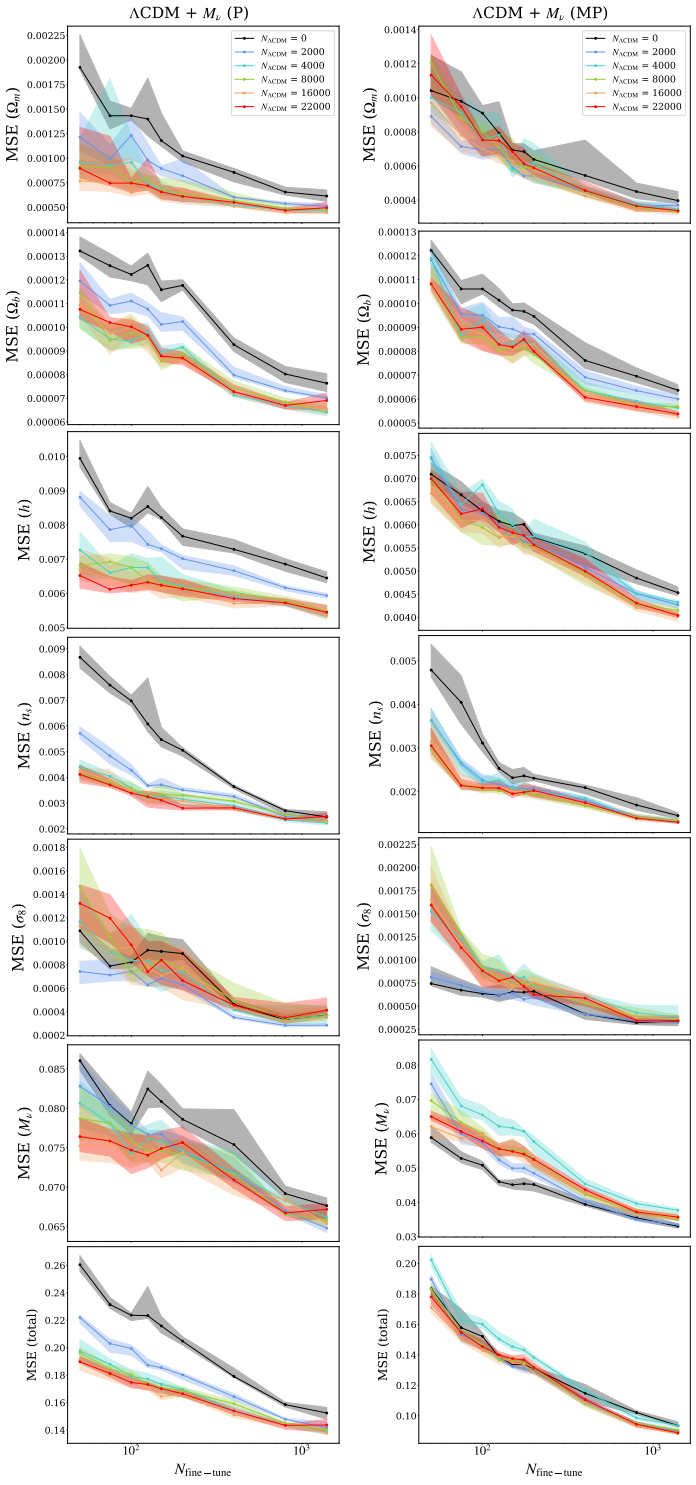
1. No-dummy pre-training: This network is identical to the setup described in Section 2, except that no dummy output node was included during pre-training. In this case, for fine-tuning only the ΛCDM parameters were initialized with pre-trained weights, while the additional parameters required for the beyond-ΛCDM models started from random initialization.
2. Attach a trainable inference head: Here we modified the pre-training network from Section 2 by constraining the final hidden layer to 10 neurons (finding this to be optimal). Once trained, we passed power spectra from the smaller beyond-ΛCDM datasets (those including massive neutrinos, modified gravity, or primordial non-Gaussianities) through the best performing ΛCDM model and extract the 10-dimensional output of the final hidden layer. This effectively reduces each beyond-ΛCDM power spectrum to a set of 10 latent features. In the second stage we train a new network on these reduced power spectra to predict the extended set of cosmological parameters of each beyond-ΛCDM cosmology.

Overall, we found that the latent-feature head case performed the worst, often exhibiting severe negative transfer. The weight-initialization case performed better, but the dummy-node setup described in Section 2 was the most effective, as shown in Fig. 6.

B.3 Parameter Degeneracy Analysis

To assess how degeneracies among Ω_m, σ_8, and M_ν may shape the network's performance, we study which parts of the marked power spectrum the model relies on to infer each parameter, as shown in Figure 7. We compare `SHAP beeswarm` plots for a pretrained model trained on $22{,}000$ ΛCDM simulations (Figure 7a) and a fine-tuned model trained on 50 massive-neutrino simulations using the marked power spectrum (Figure 7b), where negative transfer is most pronounced at low beyond-ΛCDM sample counts.

During pretraining, the network learns to attribute small-scale power-spectrum variations to Ω_m and σ_8, forming a LCDM-consistent mapping of those features to parameters. When M_ν, which physically suppresses small-scale power, is introduced, those same modes become predictive for M_ν, forcing the model to reassign small-scale sensitivity. This reallocation is consistent with the model treating M_ν-driven changes as if they were σ_8-like under the pretrained representation. Under fine-tuning σ_8 is still important at small scales but with a reversed sign and has a grater reliance on large-scale information. By contrast, while Ω_m loses some small-scale influence, its oscillatory sensitivity pattern at large scales remains comparatively unchanged, indicating more transferable structure. Overall, it appears that during fine-tuning the network has to effectively "unlearn" its σ_8 mapping at small scales and reallocate it to a combination of M_ν and σ_8 effects and "relearn" σ_8 from elsewhere, resulting in degradation in performance and negative transfer.



(a) (b)

Figure 4: Extension of Figure 2 showing the MSE for all individual parameters in the massive neutrino cosmology, but only for the dummy node architecture. Transfer learning provides improvements for some ΛCDM parameters when training data is very limited and when using the power spectrum (left), but offers little to no benefit for σ_8 and M_ν. In fact, for the marked power spectrum (right) it can even degrade performance (negative transfer) at low simulation counts.

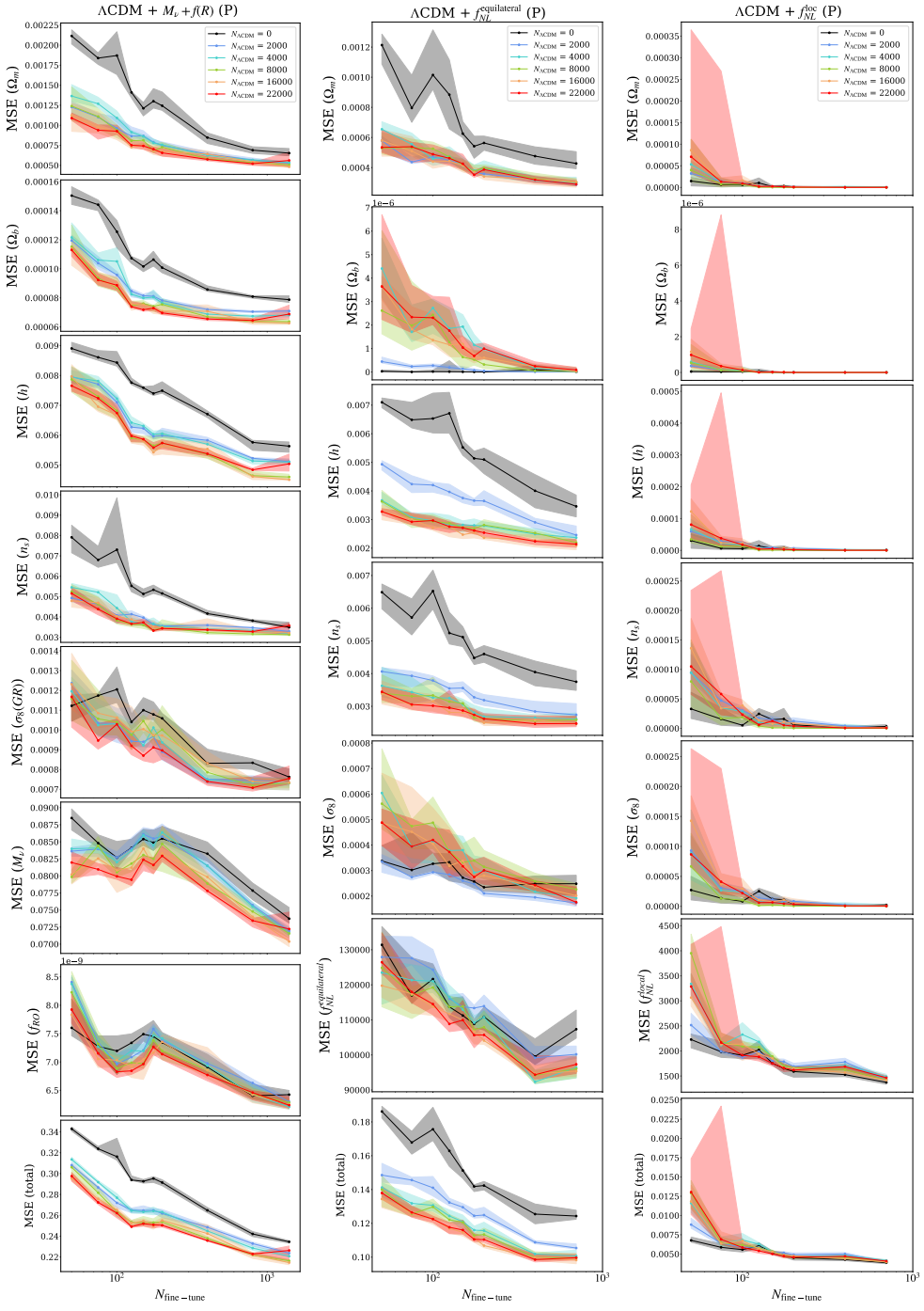


Figure 5: Same as Figure 3, but showing MSE for each individual parameter in the modified gravity, equilateral, and local non-Gaussianity cosmologies.

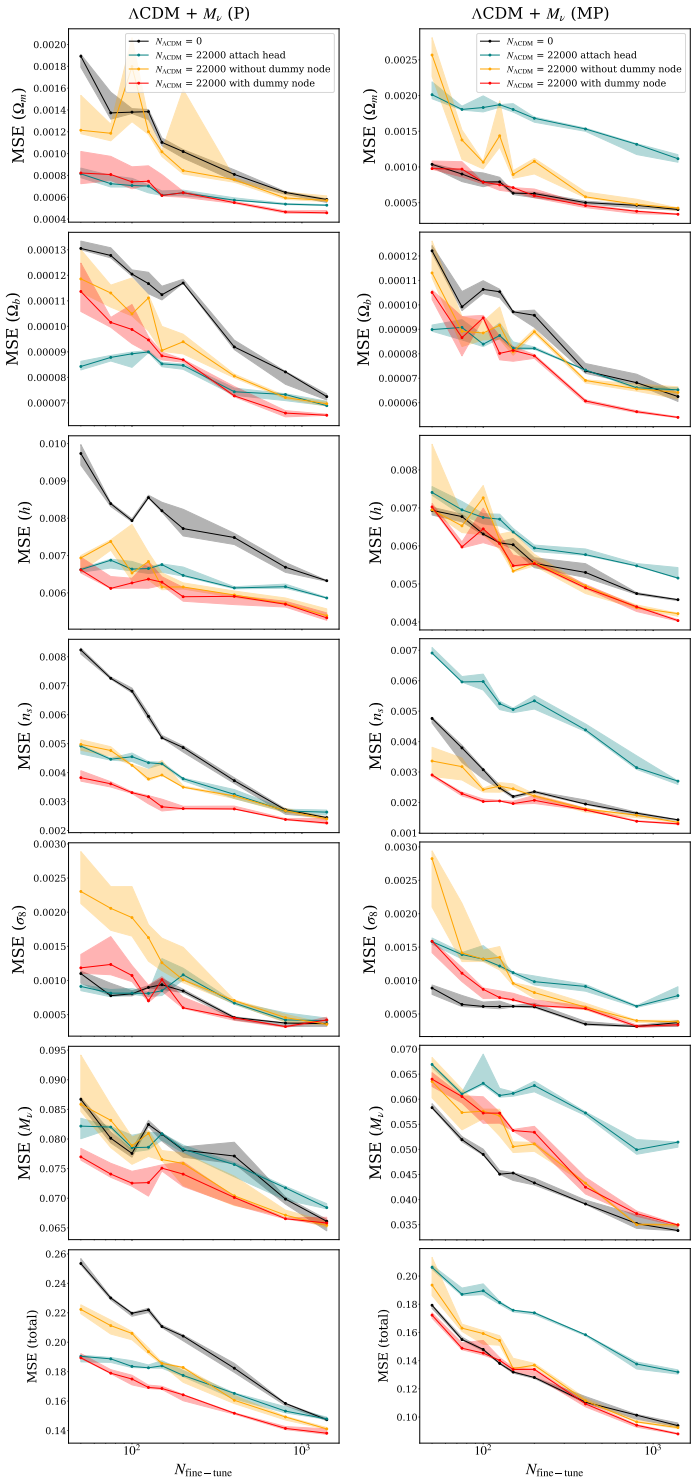
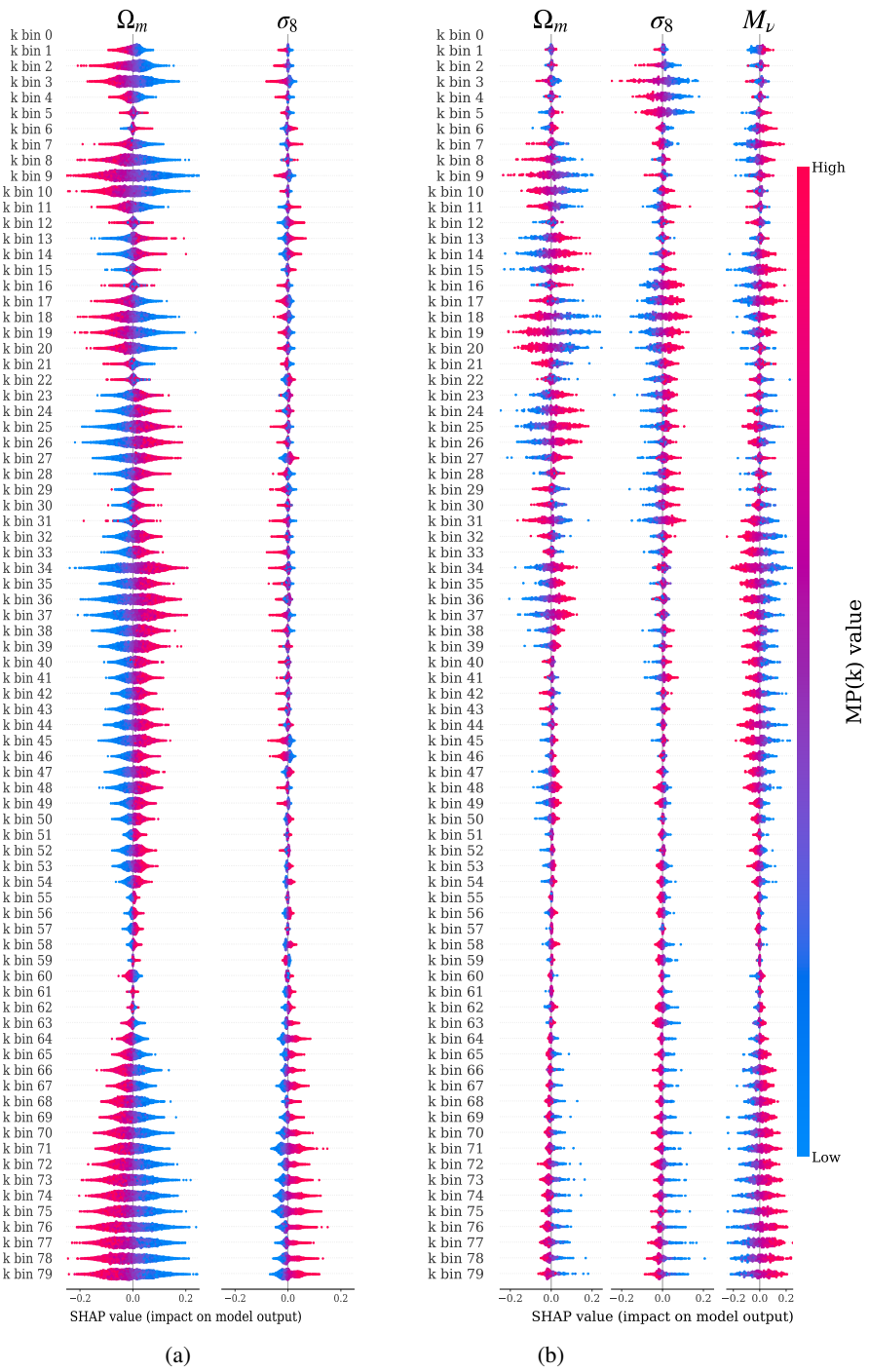


Figure 6: Same as Figure 2, but showing the MSE for all individual parameters in the massive neutrino cosmology. Provides full context for per-parameter behavior discussed in the main text in Section 3.



(a) (b)

Figure 7: `SHAP beeswarm` plots for Ω_m and σ_8 in the pretrained model (left) and for Ω_m, σ_8, and M_ν in the fine-tuned model (right), computed on the marked power spectrum MP(k). SHAP values quantify the local contribution of a feature (y-axis) to the model output relative to a baseline. The sign of the SHAP value (x-axis) indicates whether increasing that feature pushes the prediction up or down and the horizontal spread at a given feature indicates importance. Here we consider the power-spectrum k bins as the features, while the color represents the value of $P(k)$. In pretraining, small scales i.e. high-k bins carry substantial contribution for σ_8. After introducing M_ν, that small-scale influence is reassigned to M_ν, while σ_8 shows a sign flip at high k (i.e high power (pink) pushed the σ_8 prediction up (positive SHAP) and low power (blue) pushed the prediction down (negative SHAP), but during fine-tuning M_ν adopts this behavior and σ_8's is reversed) and its relative weight shifts toward larger scales. This pattern indicates that the model's initial small-scale σ_8 cue is "unlearned" and repurposed for M_ν — indicative of the σ_8–M_ν degeneracy that underlies the observed negative transfer.