

Transfer learning for multifidelity simulation-based inference in cosmology

Alex A. Saoulis,^{★,1,2} Davide Piras,^{3,4} Niall Jeffrey,¹ Alessio Spurio Mancini,⁵ Ana M. G. Ferreira,² Benjamin Joachimi¹

¹ Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, United Kingdom

² Department of Earth Sciences, University College London, 5 Gower Place, London, WC1E 6BS, United Kingdom

³ Département de Physique Théorique, Université de Genève, 24 quai Ernest Ansermet, 1211 Genève 4, Switzerland

⁴ Centre Universitaire d’Informatique, Université de Genève, 7 route de Drize, 1227 Genève, Switzerland

⁵ Department of Physics, Royal Holloway, University of London, Egham Hill, Egham, TW20 0EX, United Kingdom

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Simulation-based inference (SBI) enables cosmological parameter estimation when closed-form likelihoods or models are unavailable. However, SBI relies on machine learning for neural compression and density estimation. This requires large training datasets which are prohibitively expensive for high-quality simulations. We overcome this limitation with multifidelity transfer learning, combining less expensive, lower-fidelity simulations with a limited number of high-fidelity simulations. We demonstrate our methodology on dark matter density maps from two separate simulation suites in the hydrodynamical CAMELS Multifield Dataset. Pre-training on dark-matter-only N -body simulations reduces the required number of high-fidelity hydrodynamical simulations by a factor between 8 and 15, depending on the model complexity, posterior dimensionality, and performance metrics used. By leveraging cheaper simulations, our approach enables performant and accurate inference on high-fidelity models while substantially reducing computational costs.

Key words: cosmology: cosmological parameters, large-scale structure of Universe, dark matter – methods: statistical

1 INTRODUCTION

Cosmological inference is increasingly turning to machine learning (ML) techniques for improved precision, accuracy, and efficiency. In particular, simulation-based inference (SBI) has emerged as a tool to enable statistical analysis of the large-scale structure beyond traditional Gaussian likelihood-based analysis. These techniques have been applied to weak lensing measurements of cosmic shear such as the Kilo-Degree Survey (KiDS, von Wietersheim-Kramsta et al. 2025; Lin et al. 2023), the Dark Energy Survey (DES, Jeffrey et al. 2021; Gatti et al. 2024; Jeffrey et al. 2025), and the Sloan Digital Sky Survey-III Baryon Oscillation Spectroscopic Survey (SDSS-III: BOSS, Lemos et al. 2024; Hahn et al. 2024).

Inference on shear and clustering data beyond two-point statistics has gained importance for precision cosmology, particularly as upcoming surveys prepare to probe more non-linear scales (e.g., *Euclid*, Euclid Collaboration et al. 2025; the Vera Rubin Observatory, Ivezić et al. 2019; the Nancy Grace Roman Space Telescope, Gehrels et al. 2015). A wide body of research has now explored various strategies for advancing beyond-Gaussian analysis, incorporating e.g. field-level inference, such as Bayesian Origin Reconstruction from Galaxies (BORG, Jasche & Wandelt 2013; Jasche et al. 2015; Jasche & Lavaux 2019), using lognormal maps (Xavier et al. 2016; Leclercq & Heavens 2021; Boruah et al. 2022), or higher-order statistics, such as the three-point correlation function (Takada & Jain 2003; Schneider & Lombardi 2003; Halder et al. 2021; Hahn et al. 2024),

aperture mass (Jarvis et al. 2004; Semboloni et al. 2011; Martinet et al. 2021; Secco et al. 2022), scattering transforms (Cheng et al. 2020; Régaldo-Saint Blanchard et al. 2024; Cheng et al. 2025), and peak counts (Harnois-Déraps et al. 2021; Zürcher et al. 2022). Another important line of research, particularly for better understanding the widely discussed S_8 tension (see e.g. Abdalla et al. 2022, for a review) is probing the effects of systematics. For example, non-linear effects on the matter distribution become important over small scales, and become coupled with complex baryonic effects which are hard to model (McCarthy et al. 2018; Schneider et al. 2019, 2020). Large simulation efforts have been dedicated to probing the effect of baryonic feedback at small scales (McCarthy et al. 2018; Villaescusa-Navarro et al. 2021, 2022; Schaye et al. 2023; Ni et al. 2023; Elbers et al. 2025).

SBI uses ML models known as neural density estimators (NDEs) to model the probabilistic relationship between parameters and data empirically, allowing the method to drop the common assumption of a Gaussian likelihood¹. Examples of modelling choices include the posterior (Papamakarios & Murray 2016; Alsing et al. 2018, 2019; Greenberg et al. 2019; Deistler et al. 2022), the likelihood (Papamakarios et al. 2019; Lueckmann et al. 2019), or ratios of these quantities (Hermans et al. 2020; Durkan et al. 2020). This allows for various sources of error, including both measurement errors and systematic effects, to be incorporated in the likelihood by di-

[★] Contact e-mail: a.saoulis@ucl.ac.uk

¹ We define SBI as using ML-based models of the likelihood (or related quantites), as distinct from using a simulation-based pipeline to estimate the Gaussian covariance of the likelihood (e.g., Harnois-Déraps et al. 2024).

rectly simulating them. SBI offers an efficiency advantage over many traditional likelihood-based techniques, enabling significant reductions in the number of simulations required to model the posterior by interpolating between them (Alsing et al. 2018; Cranmer et al. 2020). In addition to this, ML-based neural compression has gained popularity for extracting summary statistics from high-dimensional observations, such as from ensembles of summary statistics, weak gravitational lensing convergence maps or dark matter density maps (Gupta et al. 2018; Ribli et al. 2019; Fluri et al. 2019, 2022; Matilla et al. 2020; Makinen et al. 2021; Jeffrey et al. 2021; Lu et al. 2023; Gatti et al. 2024; Lanzieri et al. 2024; Lemos et al. 2024). ML-based compression and density estimation have received further attention for their ability to significantly improve the constraining power of the observations (e.g., Jeffrey et al. 2021; Dai & Seljak 2024).

However, training robust and informative neural compression models is challenging, particularly when considering small datasets (for instance, fewer than $O(10^4)$ data examples, see e.g. Jeffrey et al. 2025; Bairagi et al. 2025; Park et al. 2025). Recent work has demonstrated that ML models fail to optimally compress low-dimensional power spectrum data in a data-limited regime (Bairagi et al. 2025). Neural compression of field-level data, on the other hand, often relies on deep learning techniques such as convolutional neural networks (CNNs), which are particularly data-hungry: for instance, Jeffrey et al. (2025) trained a large ensemble of CNN-based neural compression models in order to mitigate against their weak performance in the absence of a large training dataset. In addition, the density modelling of SBI is also hamstrung by a lack of training data. Prior work has shown that common neural density estimation techniques underperform with limited data, yielding inaccurate and poorly calibrated posteriors (Lueckmann et al. 2021; Hermans et al. 2022; Lemos et al. 2023a; Delaunoy et al. 2024; Tucci & Schmidt 2024; Krouglova et al. 2025). This makes extending SBI to more realistic cosmological models that require expensive, fine-grid hydrodynamical simulations challenging.

Our work aims to reduce the number of expensive simulations required to perform cosmological inference by leveraging cheaper simulators. In a recent example, Jia (2024a,b) use a pre-trained inference model (via neural quantile estimation , NQE) that is calibrated on a small target dataset using a quantile-shifting technique. In this study, we develop the use of transfer learning, a popular technique in the ML community that leverages data from one domain to improve performance in another (see e.g. Zhuang et al. 2020, for a survey). One example of transfer learning is domain adaptation, which has been used within cosmology to improve the robustness of inference with respect to uncertain physical processes; domain adaptation improves generalisation across datasets by aligning their feature representations. This can be achieved in a number of ways: introducing additional loss terms, such as maximum mean discrepancy (MMD, Roncoli et al. 2023); adversarially, by training a discriminator to minimise domain differences (Ganin & Lempitsky 2015; Jo et al. 2025; Andrianomena & Hassan 2025); or using optimal transport methods to explicitly map between latent distributions (Wehenkel et al. 2024; Andrianomena & Hassan 2025).

Our approach is straightforward: we perform transfer learning by first pre-training on a large corpus of cheaper, lower-fidelity data before training the model on a small set of accurate examples (a process known as fine-tuning). This widely used approach underpins foundation models, which are large, generic pre-trained models that can later be fine-tuned for specific tasks (He et al. 2016; Devlin et al. 2019; Dosovitskiy et al. 2021; Radford et al. 2021; Zhai et al. 2022; Kirillov et al. 2023). Pre-training allows models to learn generalisable features, which improves performance when adapting to new,

related datasets (Bengio 2012; Kornblith et al. 2019; Hoffmann et al. 2019; Mishra et al. 2022; Tahir et al. 2024; Lastufka et al. 2024). Some prior work has used this approach for cosmological inference (Sharma et al. 2024; Gondhalekar & Moriaki 2024), but there has been no comprehensive investigation into whether pre-training can substantially reduce the number of accurate simulations required to perform inference. Concurrent with this work, Krouglova et al. (2025) demonstrated that the exact same principle of transfer learning is effective for standard density estimation architectures such as neural spline flows, with applications directly to SBI.

This paper is structured as follows. Our methodology is described in Section 2. In Section 2.1 we describe the multifidelity simulation suites that we use for transfer learning. Section 2.2 presents the ML architectures and training procedures developed for this work, while the metrics used for model evaluation are introduced in Section 2.3. Section 3 presents the results of our multifidelity transfer learning methodology, and compares it with a high-fidelity-only approach for two examples: Section 3.1 explores a two-dimensional inference problem, and Section 3.2 addresses a more complicated five-dimensional inference problem, with a larger set of cosmological parameters and astrophysical nuisance parameters. Finally, we present a discussion of our results and our conclusions in Section 4.

2 METHODOLOGY

2.1 Data

This work introduces a simple framework for multifidelity inference on cosmological data. We utilise the CAMELS Multifield Dataset (CMD, Villaescusa-Navarro et al. 2022; Ni et al. 2023), a well-studied collection of simulations covering different fidelities and sub-grid physics models, to demonstrate our methodology. In particular, the CMD includes a gravity-only N -body simulation suite using GADGET-3 (Springel 2005), as well as several magnetohydrodynamical simulation suites. In this study, we use N -body simulations as the lower-fidelity dataset, and the IllustrisTNG CMD suites as high-fidelity simulations. These IllustrisTNG simulations were produced using the AREPO code (Springel 2010) to solve the same sub-grid physics models as the original IllustrisTNG simulations (Weinberger et al. 2016; Pillepich et al. 2018).

The CMD comprises thousands of simulations sampling universes with different cosmologies and astrophysical processes. These simulations are standardised to volumes of $(25 h^{-1}\text{Mpc})^3$. For each simulation, 15 current-time ($z = 0$) pseudo-independent 2D matter slices are extracted by considering 5 slices per dimension (of thickness $5 h^{-1}\text{Mpc}$); these are then pixelised into bins of approximate area $(0.1 h^{-1}\text{Mpc})^2$ to produce 2D images of the density fields with side of 256 pixels (see Villaescusa-Navarro et al. 2021 for the exact procedure). This work performs inference directly on density maps: we use 2D dark matter density maps, denoted as M_{cdm} , from the N -body simulations and the IllustrisTNG simulations. We then take the log of the maps and scale them such that the pixel statistics follow an approximate unit Gaussian distribution. These processed maps are then passed into the ML models. An example of these maps is given in Fig. 1, which shows the differences between two simulations, one hydrodynamical and one N -body, with identical parameters and initial conditions. Figure 1 explores the differences between the multifidelity simulations through a mass density ratio map, the power spectrum ratio, and a comparison of the peak statistics. The peak statistics were computed by first applying a four pixel aperture mass

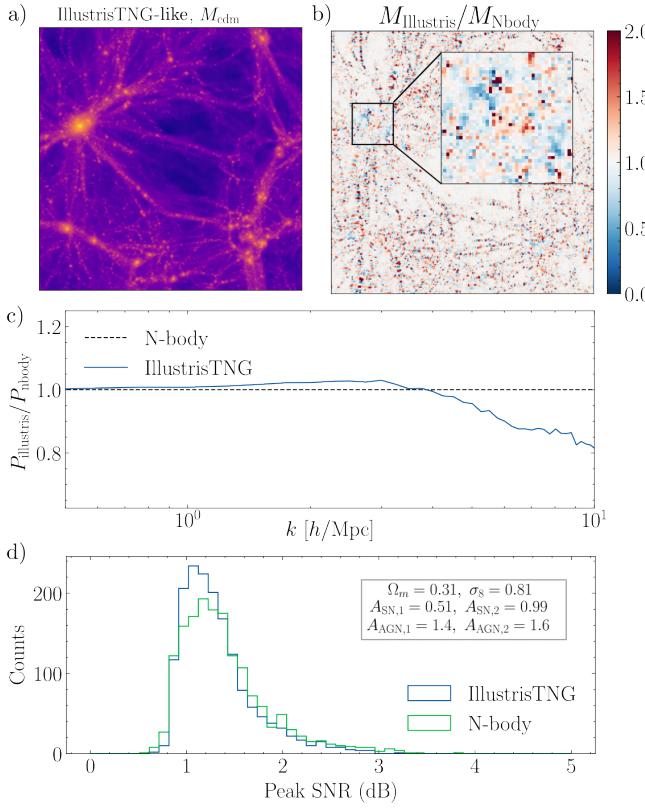


Figure 1. Comparison of the (scaled) dark matter density maps between paired simulations from the IllustrisTNG and N -body CMD simulation suites. Paired simulations have identical initial conditions as well as cosmological and astrophysical parameters. Panel a) shows a log-scaled dark matter density map, M_{cdm} , from the IllustrisTNG Latin hypercube (LH) suite. Panel b) gives the ratio between the IllustrisTNG M_{cdm} and the paired N -body M_{cdm} , with a zoomed-in inlay of a particularly high contrast region. Panel c) shows the power spectrum ratio between the two maps, and panel d) shows the peak count statistics, computed by applying a four pixel M_{ap} filter (see text for details) and computing the peak signal-to-noise ratio (SNR) over the background level. The exact cosmological and astrophysical feedback parameters are given in panel d).

(M_{ap}) filter following Schneider et al. (1998), implemented using lenspack².

We apply our methodology to two of the simulation suites from the CMD: (i) the Latin hypercube suite (LH, Villaescusa-Navarro et al. 2022), which varies the matter density fraction Ω_m and the amplitude of the matter density power spectrum, parametrised by σ_8 , alongside four astrophysical nuisance parameters ($A_{\text{SN},1}$, $A_{\text{SN},2}$, $A_{\text{AGN},1}$, $A_{\text{AGN},2}$); and, (ii) the Sobol28 suite (SB28, Ni et al. 2023), which varies Ω_m , σ_8 , the scalar spectral index n_s , the Hubble parameter h , and the baryonic density fraction Ω_b , in addition to 23 astrophysical nuisance parameters. The astrophysical parameters in the LH suite control the strength and behaviour of the stellar and active galactic nuclei (AGN) feedback in the simulations, while the SB28 suite varies a more detailed set of baryonic feedback processes controlling stellar and AGN feedback, supermassive black hole growth rates, star formation rates and stellar population modelling. It is worth noting that the small simulation box-size $25 (h^{-1}\text{Mpc})^3$ may restrict the

degree to which some of these effects, particularly AGN feedback, impact the simulations.

These suites contain 15000 and 30720 paired N -body and IllustrisTNG dark matter maps, respectively. In both cases we perform inference over only the cosmological parameters, implicitly marginalising over all nuisance parameters. For both suites, we use the last 200 cosmologies (corresponding to 3000 matter density maps) as holdout validation and test sets, ensuring no leakage between the training, validation, and test sets. We also augment the 2D matter map dataset by randomly flipping and rotating the images during training.

2.2 Model training

In this study, we focus on neural posterior estimation (NPE, Papamakarios & Murray 2016). We perform inference directly at the map level, and train a CNN-NDE neural network end-to-end to model the posterior distribution $p(\theta | x)$, where θ denotes the cosmological parameters and x the observation. The neural network parameters φ are trained to produce a model of the posterior $q_\varphi(\theta | x)$, which is generally achieved through the forward Kullback-Leibler (KL) divergence:

$$\begin{aligned} D_{\text{KL}}(p(\theta | x) \| q_\varphi(\theta | x)) &= \mathbb{E}_{p(\theta, x)} [\log p(\theta | x) - \log q_\varphi(\theta | x)] \\ &= \mathbb{E}_{p(\theta, x)} [-\log q_\varphi(\theta | x) + \text{const.}] . \end{aligned} \quad (1)$$

The log-posterior term $p(\theta | x)$ does not depend on the neural network parameters φ , and so can be ignored as a constant in the objective function:

$$\mathcal{L}(\varphi) = -\mathbb{E}_{p(\theta, x)} [\log q_\varphi(\theta | x)] . \quad (2)$$

Note that this formulation yields an identical objective to the Variational Mutual Information Maximisation (VMIM) approach, without the emphasis on learning an information-optimal summary statistic (Jeffrey et al. 2021).

The objective in Eq. (2) is used to first pre-train the network on N -body dark matter maps until convergence. It is then used again without modification to fine-tune the network on IllustrisTNG maps. We use this framework for simplicity, since only one network needs to be trained and all training can be done end-to-end. Extensions to e.g. neural likelihood estimation, and further variants, will be explored in future work. We do not envision these would require any significant modifications, but would need several training stages as in Jeffrey et al. (2021).

We performed an initial exploration of architectures for data compression and density estimation. We were partly motivated by the fact that common neural network architectures used for neural summarisation, such as CNNs, are well-suited for transfer between datasets. This stems from the inductive bias of CNNs, which encourages the learning of generic, transferable features that are then composed over a hierarchy of scales (Girshick et al. 2014; Yosinski et al. 2014; Kornblith et al. 2019). We found that the tailored CNN architecture from Villaescusa-Navarro et al. (2022) outperformed various standard architectures, such as ResNet (He et al. 2016) and ConvNext (Liu et al. 2022), both when pre-trained from natural image data or randomly initialised. We therefore used the CNN architecture from Villaescusa-Navarro et al. (2022) as our neural compression backbone, changing only the dimension of the final output layer to instead serve as a latent embedding. We found that using the CNN to compress matter density maps to larger latent dimension sizes slightly improved performance, so set the latent dimension to 128.

² <https://github.com/CosmoStat/lenspack>

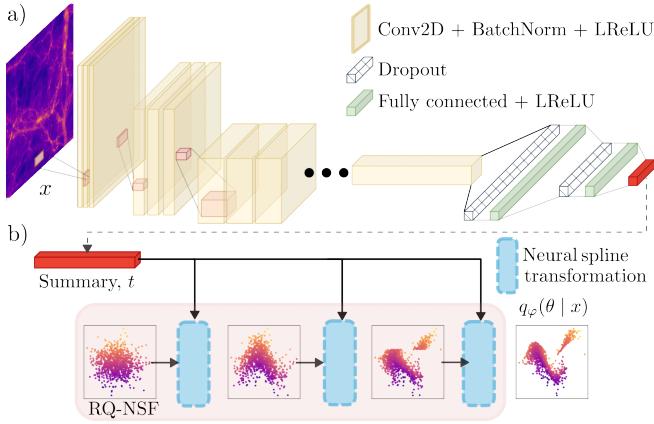


Figure 2. The neural network architecture used to perform NPE in this work. a) A CNN identical to that used in Villaescusa-Navarro et al. (2022) extracts informative features from the input dark matter map images. The CNN is built of blocks with three repeats of a 2D convolution, batch normalisation, followed by leaky ReLU non-linearities (LReLU, Maas et al. 2013). After each block, the spatial dimension of the activation maps is halved using a stride of 2, and the number of channels is doubled. The final convolutional layer is flattened and passed to a pair of feedforward layers. b) The extracted low-dimensional summary statistics, t , are fed into a rational-quadratic neural spline flow (RQ-NSF). The RQ-NSF uses the summary statistic as conditional information to transform a simple base distribution into the modelled posterior $q_\varphi(\theta | x)$.

We used a rational-quadratic neural spline flow (RQ-NSF, Durkan et al. 2019) as the NDE head of the network, implemented using the `sbi` Python package (Tejero-Cantero et al. 2022). We found this NDE architecture gave significant improvements over alternative popular choices, such as Masked Autoregressive Flows (MAFs, Papamakarios et al. 2017). We also found that inserting batch normalisation (Ioffe & Szegedy 2015; Santurkar et al. 2018) layers between spline flow blocks (similar to the CNN architecture) substantially improved performance, both in pre-training and fine-tuning. An overview of the architecture is presented in Fig. 2.

Once the architecture was selected, we performed end-to-end training of the network on the NPE objective in Eq. (2). We made a range of modifications that improved performance. We utilised a weight decay of 0.01, which regularises the neural network by adding a small penalty term to the network weight magnitudes (Krogh & Hertz 1991; Loshchilov & Hutter 2017). This was particularly important for the fine-tuning stage, where regularisation while training on very small datasets was greatly beneficial. We used a short learning rate (LR) warm-up period, which has been found to improve deep learning model training (He et al. 2016; Goyal et al. 2017; Vaswani et al. 2017) with a number of posited explanations (see e.g., Gotmare et al. 2019; Kalra & Barkeshli 2024). Larger batch-sizes (we selected 64) also improved performance for all models. We found that using a cyclic learning rate scheduler (Smith 2017) as in Villaescusa-Navarro et al. (2022) improved performance when training on large datasets, i.e. $> \mathcal{O}(10^5)$ maps. Baseline experiments and pre-training were performed with a LR of 2×10^{-4} . Fine-tuning was performed with a LR of 1×10^{-5} and an exponential decay scheduler. All models were trained using the AdamW optimizer (Loshchilov & Hutter 2017). All models pre-trained on N -body simulations used the entire training suite, and models with the lowest validation loss were saved for fine-tuning and evaluation.

2.3 Evaluation

2.3.1 Posterior Accuracy

Model evaluation in SBI is a well-studied task, and commonly used metrics include posterior-predictive checks (e.g., Papamakarios et al. 2017; Durkan et al. 2019), kernel-based distance tests such as MMD, and classifier 2-sample tests (Friedman 2004; Lopez-Paz & Oquab 2017). Lueckmann et al. (2021) presented a review and comparison between various evaluation choices. In our case, where the true posterior is unknown, one could construct a ‘‘high-quality’’ posterior by using the full simulation suite to train multiple models with different initialisations, and then averaging their inference results. This ‘‘ensemble’’-based strategy can be used to improve posterior quality by integrating over the model’s epistemic uncertainty (Hermans et al. 2022; Lin et al. 2023). Unfortunately, we found empirically that five-member model ensembles led to under-confident (conservative) posteriors. This finding is compatible with prior work (Hermans et al. 2022). We expect that this could be overcome with larger ensembles, but this approach becomes highly computationally demanding for deep learning-based models. We therefore avoid overreliance on metrics that require reference posteriors.

In the absence of a high-quality reference posterior, the most appealing headline metric to quantify model performance is simply the mean test posterior probability (MTPP) $\log q_\varphi(\theta | x)$ at the true parameter values θ . This is estimated by computing an expectation over the test dataset \mathcal{D} :

$$\text{MTPP} = \mathbb{E}_{(x, \theta) \sim \mathcal{D}} [\log q_\varphi(\theta | x)]. \quad (3)$$

This serves as a robust test for posterior quality given enough test examples (Lueckmann et al. 2021), though the scale of this metric is not particularly interpretable.

To complement this, we estimate the calibration of each model over the entire test set. This is achieved by running inference on each test data realisation, and estimating the frequency at which the truth lies within a given credibility level. This test allows us to identify modelling issues in the posteriors, such as bias and overconfidence (Hermans et al. 2022). For K credibility level bins, we estimate the observed frequency within a given bin \hat{p}_i , and compare it with the expected (ideal) frequency $p_i = 1/K$. We then compute a calibration error C , which is the relative mean squared error between the two quantities:

$$C = \frac{1}{K} \sum_{i=1}^K \left(\frac{\hat{p}_i - p_i}{p_i} \right)^2 = \frac{1}{K} \sum_{i=1}^K \left(\frac{\hat{p}_i}{p_i} - 1 \right)^2. \quad (4)$$

The right-hand side simply provides an equivalent expression for the calibration error C in terms of the ratios between observed and expected frequencies at given credibility levels, \hat{p}_i/p_i , which we refer to as ‘‘overcoverage.’’

We utilise Tests of Accuracy with Random Points (TARP, Lemos et al. 2023b) to estimate the coverage statistics efficiently. We bootstrap the estimated credibility level statistics produced by TARP 25 times and quote the mean of the estimated \hat{p}_i .

Together, the MTPP and calibration error C provide robust diagnostics of the fidelity of the learned posteriors. Once these diagnostics indicate well-calibrated and accurate inference, we can begin to assess how informative the posteriors are.

2.3.2 Constraining Power

Good calibration is a necessary pre-condition for a useful, reliable posterior model. Once this property is satisfied we can test the amount

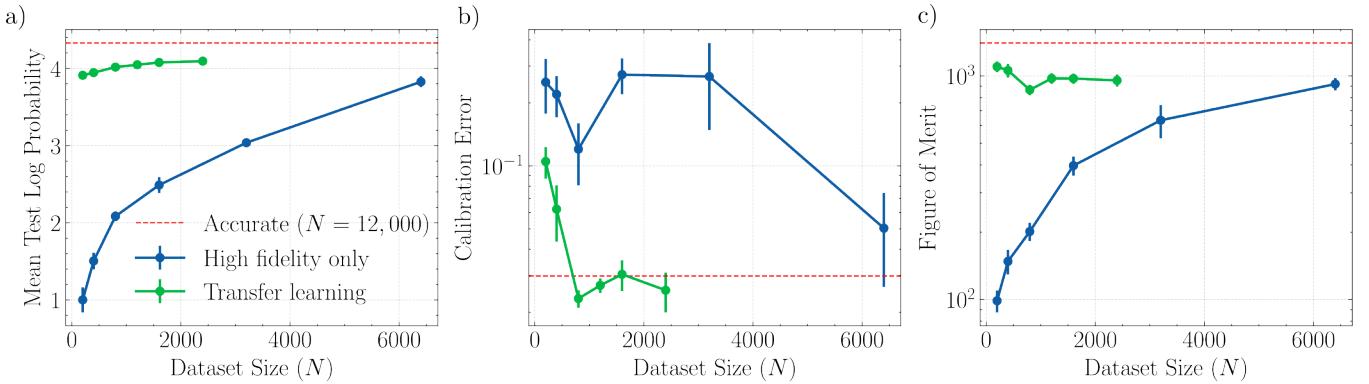


Figure 3. Inference results on the IllustrisTNG LH suite for a 2D posterior over Ω_m and σ_8 . We compare the performance of the two approaches: training with only high-fidelity maps (blue) against the transfer learning approach (green), which uses 13000 N -body simulations for pre-training. An accurate benchmark model trained on the entire IllustrisTNG LH suite training set is shown by the dashed red line. Panel a) shows the mean test posterior probability (MTPP), panel b) shows the calibration error C of the modelled posterior, defined in Eq. (4), and panel c) shows the FoM, all as a function of high-fidelity dataset size N . Panel c) should be interpreted with the proviso that the model must be well-calibrated before the FoM measures genuine constraining power. All results show the mean and standard error over six independent training runs (including independent pre-training runs on the N -body simulations).

of information that the model is capable of extracting from the dark matter density maps. We compute the Figure of Merit (FoM), which estimates the constraining power of each model. Assuming a flat prior on $[0, 1]$, the FoM can be computed as:

$$\text{FoM} = [\det \text{Cov}[\theta | x]]^{-1/n}, \quad (5)$$

where $\text{Cov}[\theta | x]$ is the covariance matrix of the posterior, and n is the dimensionality of the posterior. A higher FoM indicates a tighter constraint on the parameters, meaning the model is more informative (provided the model is unbiased). In practice, we transform the cosmological variables to a flat prior on $[0, 1]$ to compute the FoM and estimate the covariance using samples from the modelled posterior.

Finally, for a more interpretable metric of the posterior quality, we compute the mean squared error (MSE) between the modelled posterior mean $\hat{\theta}$ and the true parameters θ over the entire test set. We report these MSEs broken down by parameter to probe whether the posterior quality differs significantly between parameters.

3 RESULTS

We run a range of experiments comparing high-fidelity-only models against the transfer learning approach. We use N to denote the number of IllustrisTNG 2D dark matter density maps used during the training stage. Of the dataset size N , 90% is used as training data, while 10% is used for validation data. We repeat all training runs six times, changing the initialisation of the network and the (random) order in which the training data is passed to the network. All results report the mean performance on the test set, using the lowest validation loss model from each repeat.

3.1 CAMELS Multifield Dataset: LH suite

We compare the performance of each method over a range of IllustrisTNG dataset sizes. We present the results of the LH experiments in Fig. 3. We find that small IllustrisTNG dataset sizes lead to poor performance when training from random network initialisation, whereas pre-training on N -body maps leads to good performance with very

few simulations. For instance, a pre-trained model that is then fine-tuned with $N = 200$ IllustrisTNG maps has higher MTPP than the high-fidelity-only approach with $N = 6400$ maps.

One complicating factor is the posterior calibration, quantified in Fig. 3b. This demonstrates that despite the high test posterior probability, models fine-tuned with very few IllustrisTNG maps appear to be poorly calibrated. We find that acceptable calibration is achieved after $N = 800$ fine-tuning maps, while training with only high-fidelity maps requires $N = 6400$ for similar posterior calibration. We therefore find at least a factor of 8 reduction in the number of simulations required to produce a performant, well-calibrated model of the posterior.

We present calibration curves from a range of dataset sizes across the two approaches in Fig. 4. These show the standard cumulative distribution of observed credibility levels in the main panels, as well as the overcoverage distribution \hat{p}_i/p_i in the insets. These reaffirm the calibration issues identified in Fig. 3: training from scratch with fewer than $N = 6400$ maps leads to significantly overconfident posteriors. The high overcoverage at $\{0, 1\}$ (paired with the below-ideal coverage in the middle of the distribution) is a clear indicator of overconfidence, since it indicates that the true parameters occur at extreme credibility levels too often. On the other hand, the transfer learning models display a more minor form of bias and overconfidence until reaching around $N = 800$ maps.

The FoM performance as a function of dataset size is shown in Fig. 3c. When training using only high-fidelity simulations, low dataset sizes lead to low FoMs. This observation indicates that the features (or summary statistics) extracted by the CNN are not particularly informative. Additionally, the overconfident posteriors indicate limitations in the performance of the NDE, since it is incapable of producing trustworthy posteriors. We therefore conclude that both the CNN neural compressor and the density estimation model perform poorly with small training datasets.

On the other hand, the very high MTPP and FoMs of the transfer learning models (even from $N = 200$) indicate that the pre-trained CNNs produce highly informative features. However, the inferred posteriors for $N = [200, 400]$ are biased and overconfident, suggesting that the NDE needs at least $N = 800$ maps to correctly adjust the inferred posteriors to ensure good calibration. We found that the FoM for the N -body pre-training task was ~ 1400 , much greater than the

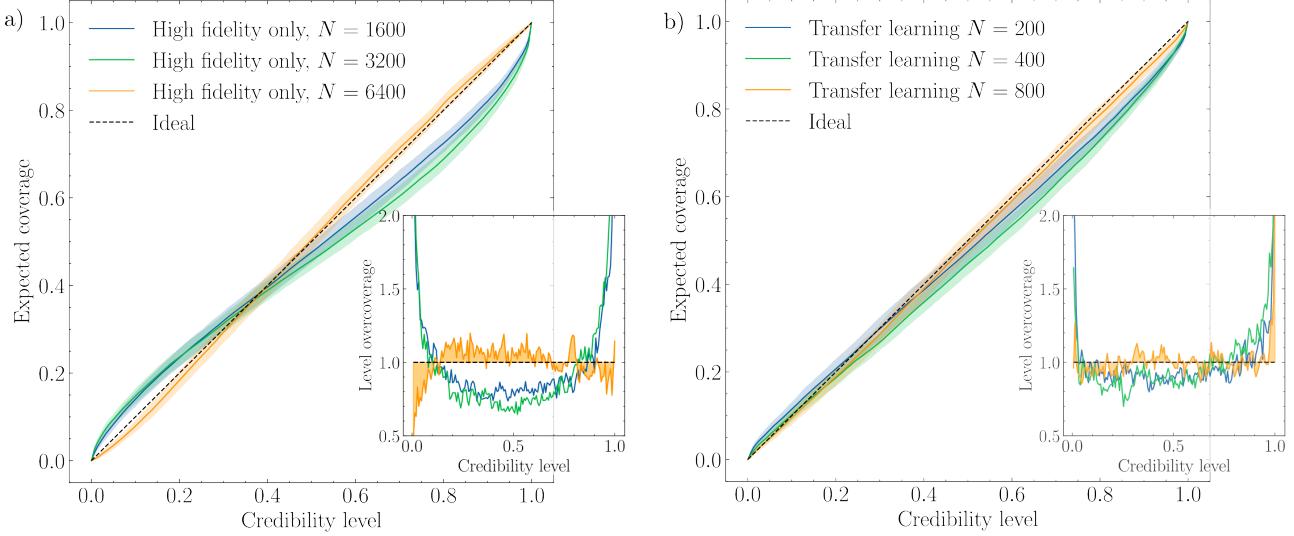


Figure 4. Main panels: cumulative calibration curves of the nominal credibility level distribution, assessing the posterior coverage quality as a function of dataset size. The ideal calibration curve is shown by the black dashed line. Shaded regions show the 2σ uncertainties derived from bootstrapping. Insets: the overcoverage values per credibility level, see Eq. (4) and the surrounding text for details. The shaded orange region highlights the discrepancy between the ideal and observed distribution of credibility levels, which is quantified by the calibration error metric introduced in Eq. (4). Panel a) shows training from high-fidelity-only simulations, where models with small dataset sizes display significant overconfidence, and better calibration (though now mildly underconfident) is achieved at $N = 6400$. Panel b) shows that transfer learning requires around $N = 800$ IllustrisTNG maps to achieve good calibration.

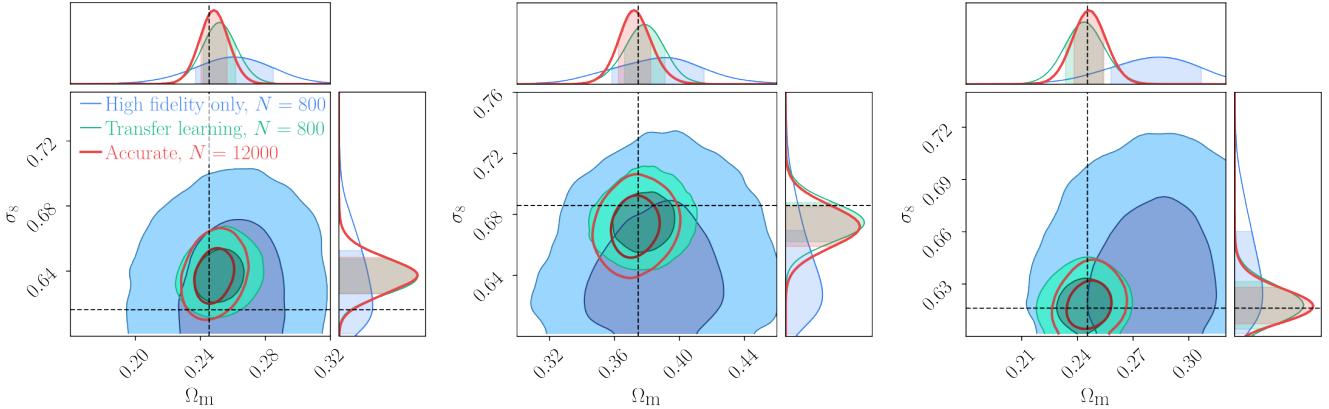


Figure 5. Three representative examples of inference from the LH simulation suite. The true cosmology is shown by the black dashed line. A model trained using transfer learning with $N = 800$ high-fidelity IllustrisTNG maps is compared against a high-fidelity-only model trained with $N = 800$ maps. The posteriors are compared with an ‘‘accurate’’ posterior model that was trained using the full simulation suite.

baseline and transfer learning models on the IllustrisTNG inference task. The dip in the FoM at $N = 800$ is thus potentially related to the reduced constraining power of extracted CNN features on the IllustrisTNG data finally becoming properly incorporated by the NDE. This could be due to the feature-shift between N -body simulations and IllustrisTNG, as well as the inherent greater uncertainties due to the more complex physics of the hydrodynamical simulation suite. These results indicate that for very small fine-tuning datasets, the performance bottleneck on this task is adapting the NDE. In future work we will explore whether the NDE head could be fine-tuned while preserving good calibration statistics with more advanced techniques (such as balanced SBI, e.g. Delaunoy et al. 2022).

Three representative examples of inference on test cosmology maps are shown in Fig. 5. We compare models produced using only

high-fidelity maps ($N = 800$), transfer learning ($N = 800$), and an ‘‘accurate’’ reference model trained on the full IllustrisTNG training set. These examples are qualitatively consistent with the analysis presented above. The high-fidelity-only $N = 800$ model gives very uninformative constraints compared to the other two posteriors. On the other hand, the fine-tuned model appears well-calibrated, and only slightly less constraining than the model trained with $\times 16.25$ more high-fidelity maps. We present a similar comparison with a high-fidelity model trained on $N = 3200$ in Appendix A, demonstrating that even for larger high-fidelity dataset sizes, pre-training yields significantly improved posteriors.

Appendix B presents a range of further tests into the model performance. We found that the pre-trained models performed very poorly on high-fidelity maps when no fine-tuning was performed (corre-

sponding to $N = 0$). We explored the quality of the fine-tuned CNN compressor by freezing the CNN and re-training the NDE with larger dataset sizes. This presented more evidence that the limiting factor at very low fine-tuning dataset sizes was the NDE. We also showed that the small performance gap between the “accurate” model and the transfer learning models with larger dataset sizes was caused by slightly worse compression.

Another consideration was the possibility that the paired aspect of the datasets was responsible for the significant performance gains from lower-fidelity pre-training. While we did make any explicit use of the simulation pairs, this could still have had an impact depending on the training dynamics of the network. We tested this in Appendix B and found strong evidence that the paired aspect of the source and target dataset has no impact on our results. We therefore conclude that our transfer learning approach does not depend on paired datasets. In principle, this means that large, pre-existing simulation suites could be used as multifidelity datasets without the need to pair initial conditions and cosmological parameters.

We found that fine-tuning with the entire IllustrisTNG suite produced a slightly weaker model than the “accurate” benchmark model (i.e. the transfer learning curve in Fig. 3a does not intersect with the “accurate” performance). This is due to the low learning rate (LR) used during fine-tuning, and a higher LR recovered the “accurate” performance. This is consistent with the intuitive notion that a higher LR allows the training procedure to escape from the slightly sub-optimal region of the weight-space that is reached during pre-training. These observations are perfectly compatible with the study of Sharma et al. (2024), who found no clear benefits of transfer learning when using a large high-fidelity dataset.

We present the degree of agreement between the posterior sample means $\hat{\theta}$ and the true cosmologies θ in Fig. 6, broken down for σ_8 and Ω_m . These indicate pre-training yields very large improvements in the posterior for both parameters. Ω_m and σ_8 are inferred with similar precision relative to the baseline over the entire test set.

3.2 CAMELS Multifield Dataset: SB28 suite

We repeat the experiments of Section 3.1 on the SB28 suite, this time performing inference over a 5-dimensional posterior. The very broad range of nuisance parameters, as well as the extra three cosmological parameters $\{n_s, h, \Omega_b\}$, lead to a more challenging inference problem. Prior work has only explored inferring Ω_m and σ_8 from this dataset, and has found that the larger set of cosmological and astrophysical parameters leads to much weaker constraints on σ_8 (Ni et al. 2023). Ni et al. (2023) also indicated that the Hubble constant h and the baryonic fraction Ω_b have minor effects on the simulations, indicating that these may be challenging to constrain. Note that the SB28 suite contains roughly double the number of simulations as the LH suite, enabling a more accurate ML-based reference model.

Fig. 7 displays the headline metrics comparing the two training approaches with a baseline “accurate” model that used the entire SB28 suite training set. Again, the transfer learning models significantly outperform models trained from scratch, and the MTPP score of the transfer learning experiments are only surpassed when training a high-fidelity-only model with $N = 12,800$ IllustrisTNG maps. Panels b) and c) in Fig. 7 present a similar pattern as in Section 3.1: all transfer learning models are more constraining and better calibrated than training from scratch (until $N = 12,800$). However, fine-tuned models display (minor) calibration issues until $N \geq 800$. The uptick in calibration error C for the transfer learning approach is very minor and largely within errors, so we do not attempt to interpret it. Interestingly, all fine-tuned models are better calibrated in the SB28

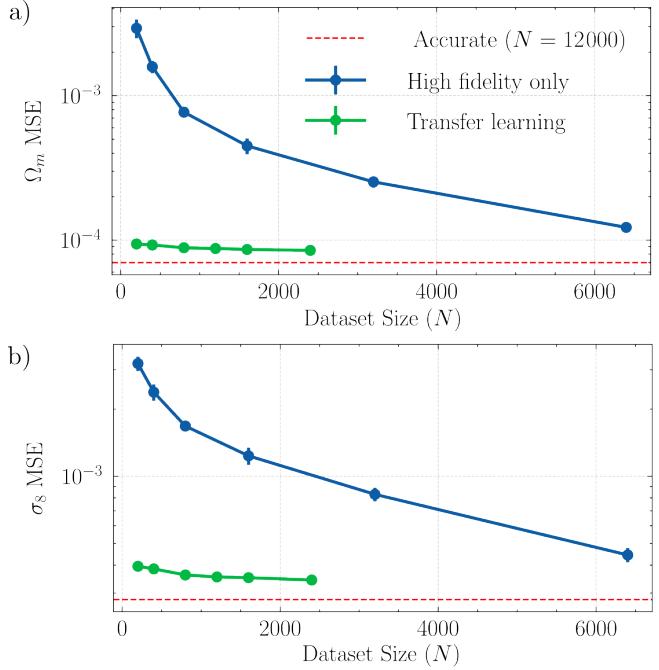


Figure 6. Mean squared error (MSE) between the inferred posterior mean $\hat{\theta}$ and the true cosmology θ . Results are broken down per-parameter, with panel a) showing Ω_m and panel b) showing σ_8 . Transfer learning models show very good consistency with the true cosmological parameters.

experiments than in the LH experiments (note the different y-axis scales).

The performance difference between high-fidelity-only models and transfer learning models is even larger than in Section 3.1. Depending on the exact MTPP performance desired, Fig. 7 indicates that pre-training on N -body simulations allows for a factor of 10 to 15 reduction in high-fidelity simulations to train an informative, well-calibrated model of the posterior.

We found that none of the models could constrain Ω_b and h far beyond the uniform prior. Given that this was a feature of the “accurate” baseline model, trained on 27720 maps, we may conclude that this a genuine feature of the simulations, at least up to the resolving power of the CNN-NDE architecture used to perform inference. This is reflected by the FoM results in Fig. 7 c), which are significantly lower for all models than the FoM in the LH suite experiment. We present two examples of inference in Fig. 8, showing only $\{\Omega_m, \sigma_8, n_s\}$. This implicitly marginalises over the poorly constrained Ω_b and h . Again, we compare training from scratch with $N = 800$ and fine-tuning with just $N = 800$ maps against an “accurate” baseline. We present examples of inference of the full 5-dimensional posterior in Appendix C.

Again, multifidelity transfer learning produces a model that significantly outperforms high-fidelity-only training. The difference in posterior quality is even more stark than in Fig. 5; the high-fidelity-only posteriors are very uninformative and fail to extract much useful cosmological information from the density maps. The overconfidence identified in Fig. 8 is apparent as a bias in the left panel of Fig. 5. On the other hand, the transfer learning approach recovers the key features of the accurate baseline posteriors, including both the location and width of the posterior contours. Both the “accurate” and transfer learning models yield a degeneracy between the amplitude

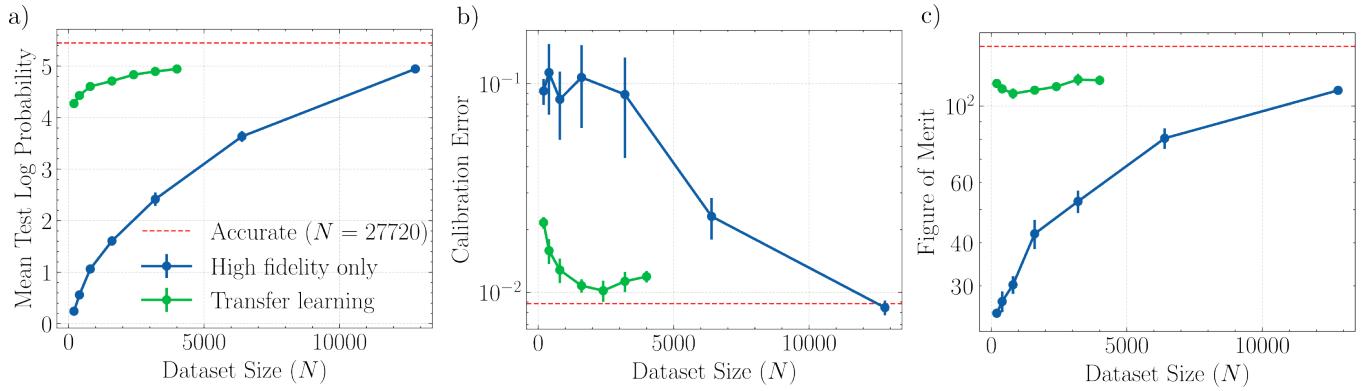


Figure 7. Inference results for the 5-dimensional posteriors on the IllustrisTNG SB28 suite. Training with only high-fidelity maps (blue) is compared against the transfer learning approach (green). An accurate benchmark model trained on the entire IllustrisTNG SB28 suite training set is shown by the dashed red line. Panel a) shows the MTPP, panel b) shows the calibration error and panel c) shows the FoM, all as a function of high-fidelity dataset size N . Panel c) should be interpreted with the proviso that the model must be well-calibrated before the FoM measures genuine constraining power. Again, all results show the mean and standard error over six independent training runs.

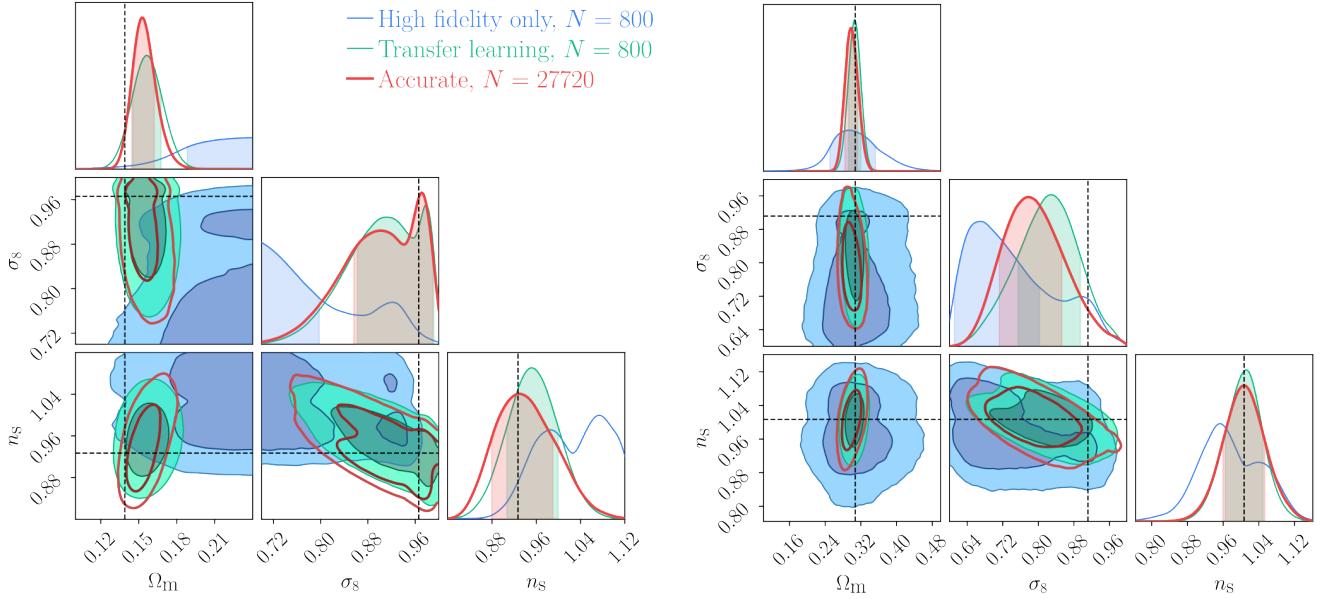


Figure 8. Two examples of posterior inference on IllustrisTNG dark matter maps from the SB28 test set. Contours are visualised over the three parameters that can be constrained by the data: $\{\Omega_m, \sigma_8, n_s\}$. The true cosmology is shown by the black dashed line. A model trained using transfer learning with $N = 800$ high-fidelity IllustrisTNG maps is compared against a high-fidelity-only model trained with $N = 800$ maps. The posteriors are compared with an “accurate” posterior model that was trained using the full training set.

of the matter density power spectrum σ_8 and the scalar spectral index n_s . This degeneracy is expected for two-point statistics at the very short scales probed in the CMD simulations, given both parameters have similar, difficult to distinguish marginal effects on the power spectrum. The results in Fig. 8 indicate that the non-linear effects probed by the CNN are insufficient to fully break this degeneracy.

The larger performance gains reported here relative to the LH suite in Section 3.1 likely result from the more complex task of modelling a 5-dimensional posterior and marginalising over a larger set of astrophysical parameters. We tentatively conclude that transfer learning may perform even better in more challenging inference

problems, particularly those involving higher-dimensional posteriors and a broader set of nuisance parameters. As the complexity of the target task increases, the value of incorporating prior knowledge through pre-training is likely to grow.

4 DISCUSSION AND CONCLUSIONS

In this study, we have demonstrated that leveraging multifidelity simulations can significantly reduce the number of expensive simulations required to perform cosmological inference with SBI. By

pre-training a neural inference model on a large set of lower-fidelity dark matter only simulations, we were able to perform informative and well-calibrated inference on IllustrisTNG hydrodynamical simulations with < 1000 high-fidelity dark matter maps. This is a substantial improvement over previous work, which had demonstrated that training neural compression algorithms with small datasets led to suboptimal compression and inference (Hermans et al. 2022; Park et al. 2025; Bairagi et al. 2025; Jeffrey et al. 2025). The relative simplicity of our framework makes this method broadly applicable across cosmology.

Prior work has explored various approximate Bayesian computation (ABC) methods for multifidelity inference (Prescott & Baker 2020, 2021), for instance by using low fidelity rejection sampling to improve the accurate simulation efficiency during inference. These have been extended to sequential multifidelity ABC (Warne et al. 2022), as well as to likelihood-free multifidelity inference by leveraging importance sampling (Prescott et al. 2024). Variance reduction strategies that capitalise on paired multifidelity simulations to isolate the statistical uncertainty, have also been used to improve estimates of cosmological observables (Chartier et al. 2021; Lee et al. 2024). Adaptation of these strategies directly for cosmological inference, particularly when dealing with significantly non-Gaussian posteriors that SBI is well suited for, remains an interesting avenue for future work. An additional line of work would be to explore tailoring our approach for transfer learning, for instance through architecture improvements that work with Fourier representations of the inputs (e.g., Yang & Soatto 2020; Mao et al. 2023; Bernardini et al. 2025) or specialised pre-training and transfer learning techniques (e.g., He et al. 2022; Oquab et al. 2024; Akhmetzhanova et al. 2024).

This work focused on matter density maps at different fidelities; there are many more possible observables that have previously been probed for performing cosmological inference, such as neutral hydrogen, gas temperature and metallicity maps (Hassan et al. 2020; Prelogović et al. 2022; Andrianomena & Hassan 2023, 2025; Gluck et al. 2024). Adaptation between observables could call for similar specialised approaches (e.g., Lian et al. 2025).

Future work could apply transfer learning to a wide variety of multifidelity datasets across cosmological inference. Recent work has demonstrated that neural compression even performs sub-optimally on lower dimensional data, such as power spectra (Bairagi et al. 2025) or ensembles of traditional summary statistics (Park et al. 2025), when dataset sizes are limited. There is a very wide array of methods for producing mock observations of varying fidelities: for instance, empirically-calibrated semi-analytic emulators (e.g., Takahashi et al. 2012; Mead et al. 2016, 2021), fast-executing lognormal dark matter simulations (e.g., Tessore et al. 2023; Lin et al. 2023; von Wietersheim-Kramsta et al. 2025) and ML-based emulators (e.g., Heitmann et al. 2009; Euclid Collaboration: Knabenhans et al. 2021; Aricò et al. 2021; Giri & Schneider 2021; Piras et al. 2023). These techniques could be used to build large mock pre-training datasets, allowing for a significant reduction in the computation time required for the production of high-fidelity simulation datasets for transfer learning. Similarly, computation budgets could be reoriented towards fewer high-fidelity simulations with more particles or larger simulation boxes. Either way, by enabling an order of magnitude reduction in high-fidelity simulations, this work demonstrates that multifidelity transfer learning has the potential to transform our approach to simulation-based inference in cosmology.

ACKNOWLEDGEMENTS

AAS was supported by the STFC UCL Centre for Doctoral Training in Data Intensive Science (grant ST/W00674X/1) and by departmental and industry contributions. AAS was also supported by the A. G. Leventis Foundation educational grant scheme. DP was supported by a Swiss National Science Foundation (SNSF) grant, and by the SNF Sinergia grant CRSIIS-193826 “AstroSignals: A New Window on the Universe, with the New Generation of Large Radio-Astronomy Facilities”. BJ acknowledges support by the ERC-selected UKRI Frontier Research Grant EP/Y03015X/1.

DATA AVAILABILITY

The Python software used to produce the results of this paper is available at <https://github.com/asaoulis/transfer-sbi>. All the simulation data used is publicly available via the CAMELS Multifield Dataset (Villaescusa-Navarro et al. 2021, 2022).

REFERENCES

- Abdalla E., et al., 2022, Journal of High Energy Astrophysics, 34, 49
- Akhmetzhanova A., Mishra-Sharma S., Dvorkin C., 2024, Monthly Notices of the Royal Astronomical Society, 527, 7459
- Alsing J., Wandelt B., Feeney S., 2018, Monthly Notices of the Royal Astronomical Society, 477, 2874
- Alsing J., Charnock T., Feeney S., Wandelt B., 2019, Monthly Notices of the Royal Astronomical Society, 488, 4440
- Andrianomena S., Hassan S., 2023, Journal of Cosmology and Astroparticle Physics, 2023, 051
- Andrianomena S., Hassan S., 2025, Astrophysics and Space Science, 370, 14
- Aricò G., Angulo R. E., Contreras S., Ondaro-Mallea L., Pellejero-Ibañez M., Zennaro M., 2021, *Monthly Notices of the Royal Astronomical Society*, 506, 4070
- Bairagi A., Wandelt B., Villaescusa-Navarro F., 2025, arXiv preprint arXiv:2503.13755
- Bengio Y., 2012, in Proceedings of ICML workshop on unsupervised and transfer learning, pp 17–36
- Bernardini M., et al., 2025, Monthly Notices of the Royal Astronomical Society, 538, 1201
- Borouah S. S., Rozo E., Fiedorowicz P., 2022, Monthly Notices of the Royal Astronomical Society, 516, 4111
- Carlini N., Liu C., Erlingsson Ú., Kos J., Song D., 2019, in 28th USENIX security symposium (USENIX security 19), pp 267–284
- Carlini N., Ippolito D., Jagielski M., Lee K., Tramer F., Zhang C., 2023, in The Eleventh International Conference on Learning Representations, https://openreview.net/forum?id=TatRHT_1cK
- Chartier N., Wandelt B., Akrami Y., Villaescusa-Navarro F., 2021, Monthly Notices of the Royal Astronomical Society, 503, 1897
- Cheng S., Ting Y.-S., Ménard B., Bruna J., 2020, Monthly Notices of the Royal Astronomical Society, 499, 5902
- Cheng S., Marques G. A., Grandón D., Thiele L., Shirasaki M., Ménard B., Liu J., 2025, Journal of Cosmology and Astroparticle Physics, 2025, 006
- Cranmer K., Brehmer J., Louppe G., 2020, Proceedings of the National Academy of Sciences, 117, 30055
- Dai B., Seljak U., 2024, Proceedings of the National Academy of Sciences, 121, e2309624121
- Deistler M., Goncalves P. J., Macke J. H., 2022, in Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 23135–23149, https://proceedings.neurips.cc/paper_files/paper/2022/file/9278abf072b58caf21d48dd670b4c721-Paper-Conference.pdf
- Delaunoy A., Hermans J., Rozet F., Wehenkel A., Louppe G., 2022, in Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 20025–20037, https://proceedings.neurips.cc/paper_files/paper/2022/file/9278abf072b58caf21d48dd670b4c721-Paper-Conference.pdf

- //proceedings.neurips.cc/paper_files/paper/2022/file/7e6288bfb68182db7d6e328b0afea89a-Paper-Conference.pdf
- Delaunoy A., Bonardeaux M. d. l. B., Mishra-Sharma S., Louppe G., 2024, arXiv preprint arXiv:2408.15136
- Devlin J., Chang M.-W., Lee K., Toutanova K., 2019, in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp 4171–4186
- Dosovitskiy A., et al., 2021, in International Conference on Learning Representations. <https://openreview.net/forum?id=YicbFdNTTy>
- Durkan C., Bekasov A., Murray I., Papamakarios G., 2019, in Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf
- Durkan C., Murray I., Papamakarios G., 2020, in III H. D., Singh A., eds, Proceedings of Machine Learning Research Vol. 119, Proceedings of the 37th International Conference on Machine Learning. PMLR, pp 2771–2781, <https://proceedings.mlr.press/v119/durkan20a.html>
- Elbers W., et al., 2025, Monthly Notices of the Royal Astronomical Society, 537, 2160
- Euclid Collaboration: Knabenhans M., et al., 2021, [Monthly Notices of the Royal Astronomical Society](#), 505, 2840
- Euclid Collaboration et al., 2025, [A&A](#), 697, A1
- Fluri J., Kacprzak T., Lucchi A., Refregier A., Amara A., Hofmann T., Schneider A., 2019, Physical Review D, 100, 063514
- Fluri J., Kacprzak T., Lucchi A., Schneider A., Refregier A., Hofmann T., 2022, Physical Review D, 105, 083518
- Friedman J., 2004, Technical report, On multivariate goodness-of-fit and two-sample testing. SLAC National Accelerator Laboratory (SLAC), Menlo Park, CA (United States)
- Ganin Y., Lempitsky V., 2015, in Bach F., Blei D., eds, Proceedings of Machine Learning Research Vol. 37, Proceedings of the 32nd International Conference on Machine Learning. PMLR, Lille, France, pp 1180–1189, <https://proceedings.mlr.press/v37/ganin15.html>
- Gatti M., et al., 2024, Physical Review D, 109, 063534
- Gehrels N., et al., 2015, arXiv preprint arXiv:1503.03757
- Giri S. K., Schneider A., 2021, [Journal of Cosmology and Astroparticle Physics](#), 2021, 046
- Girshick R., Donahue J., Darrell T., Malik J., 2014, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 580–587
- Gluck N., Oppenheimer B. D., Nagai D., Villaescusa-Navarro F., Anglés-Alcázar D., 2024, Monthly Notices of the Royal Astronomical Society, 527, 10038
- Gondhalekar Y., Moriwaki K., 2024, arXiv preprint arXiv:2411.14392
- Gotmare A., Keskar N. S., Xiong C., Socher R., 2019, in International Conference on Learning Representations. <https://openreview.net/forum?id=r14E0sCqKX>
- Goyal P., et al., 2017, arXiv preprint arXiv:1706.02677
- Greenberg D., Nonnenmacher M., Macke J., 2019, in Chaudhuri K., Salakhutdinov R., eds, Proceedings of Machine Learning Research Vol. 97, Proceedings of the 36th International Conference on Machine Learning. PMLR, pp 2404–2414, <https://proceedings.mlr.press/v97/greenberg19a.html>
- Gupta A., Matilla J. M. Z., Hsu D., Haiman Z., 2018, Physical Review D, 97, 103515
- Hahn C., et al., 2024, Physical Review D, 109, 083534
- Halder A., Friedrich O., Seitz S., Varga T. N., 2021, Monthly Notices of the Royal Astronomical Society, 506, 2780
- Harnois-Déraps J., Martinet N., Castro T., Dolag K., Giblin B., Heymans C., Hildebrandt H., Xia Q., 2021, Monthly Notices of the Royal Astronomical Society, 506, 1623
- Harnois-Déraps J., et al., 2024, Monthly Notices of the Royal Astronomical Society, 534, 3305
- Hassan S., Andrianomena S., Doughty C., 2020, Monthly Notices of the Royal Astronomical Society, 494, 5761
- He K., Zhang X., Ren S., Sun J., 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
- He K., Chen X., Xie S., Li Y., Dollár P., Girshick R., 2022, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 16000–16009
- Heitmann K., Higdon D., White M., Habib S., Williams B. J., Lawrence E., Wagner C., 2009, The Astrophysical Journal, 705, 156
- Hermans J., Begy V., Louppe G., 2020, in III H. D., Singh A., eds, Proceedings of Machine Learning Research Vol. 119, Proceedings of the 37th International Conference on Machine Learning. PMLR, pp 4239–4248, <https://proceedings.mlr.press/v119/hermans20a.html>
- Hermans J., Delaunoy A., Rozet F., Wehenkel A., Begy V., Louppe G., 2022, Transactions on Machine Learning Research, p. <https://openreview.net/forum?id=LHAbHkt6Aq>
- Hoffmann J., Bar-Sinai Y., Lee L. M., Andrejevic J., Mishra S., Rubinstein S. M., Rycroft C. H., 2019, Science advances, 5, eaau6792
- Ioffe S., Szegedy C., 2015, in Bach F., Blei D., eds, Proceedings of Machine Learning Research Vol. 37, Proceedings of the 32nd International Conference on Machine Learning. PMLR, Lille, France, pp 448–456, <https://proceedings.mlr.press/v37/ioffe15.html>
- Ivezic Ž., et al., 2019, The Astrophysical Journal, 873, 111
- Jarvis M., Bernstein G., Jain B., 2004, Monthly Notices of the Royal Astronomical Society, 352, 338
- Jasche J., Lavaux G., 2019, Astronomy & Astrophysics, 625, A64
- Jasche J., Wandelt B. D., 2013, Monthly Notices of the Royal Astronomical Society, 432, 894
- Jasche J., Leclercq F., Wandelt B. D., 2015, Journal of Cosmology and Astroparticle Physics, 2015, 036
- Jeffrey N., Alsing J., Lanusse F., 2021, Monthly Notices of the Royal Astronomical Society, 501, 954
- Jeffrey N., et al., 2025, Monthly Notices of the Royal Astronomical Society, 536, 1303
- Jia H., 2024a, arXiv preprint arXiv:2411.14748
- Jia H., 2024b, in Salakhutdinov R., Kolter Z., Heller K., Weller A., Oliver N., Scarlett J., Berkenkamp F., eds, Proceedings of Machine Learning Research Vol. 235, Proceedings of the 41st International Conference on Machine Learning. PMLR, pp 21731–21752, <https://proceedings.mlr.press/v235/jia24a.html>
- Jo Y., Genel S., Sengupta A., Wandelt B., Somerville R., Villaescusa-Navarro F., 2025, arXiv preprint arXiv:2502.13239
- Kalra D. S., Barkeshli M., 2024, in Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 111760–111801, https://proceedings.neurips.cc/paper_files/paper/2024/file/ca98452d4e9ecbc18c40da2aa0da8b98-Paper-Conference.pdf
- Kirillov A., et al., 2023, in Proceedings of the IEEE/CVF international conference on computer vision. pp 4015–4026
- Kornblith S., Shlens J., Le Q. V., 2019, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 2661–2671
- Krogh A., Hertz J., 1991, in Advances in Neural Information Processing Systems. Morgan-Kaufmann, https://proceedings.neurips.cc/paper_files/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf
- Krouglova A. N., Johnson H. R., Confavreux B., Deistler M., Gonçalves P. J., 2025, arXiv preprint arXiv:2502.08416
- Lanzieri D., Zeghal J., Makinen T. L., Boucaud A., Starck J.-L., Lanusse F., 2024, arXiv preprint arXiv:2407.10877
- Lastufka E., et al., 2024, [arXiv e-prints](#), p. [arXiv:2409.11175](#)
- Leclercq F., Heavens A., 2021, Monthly Notices of the Royal Astronomical Society: Letters, 506, L85
- Lee M. E., et al., 2024, The Astrophysical Journal, 968, 11
- Lemos P., Cranmer M., Abidi M., Hahn C., Eickenberg M., Massara E., Yallup D., Ho S., 2023a, Machine Learning: Science and Technology, 4, 01LT01
- Lemos P., Coogan A., Hezaveh Y., Perreault-Levasseur L., 2023b, in Proceedings of the 40th International Conference on Machine Learning. PMLR, pp 19256–19273, <https://proceedings.mlr.press/v202/lemos23a.html>
- Lemos P., et al., 2024, Physical Review D, 109, 083536
- Lian W., Lindblad J., Micke P., Sladoje N., 2025, arXiv preprint arXiv:2503.09826
- Lin K., von Wietersheim-Kramsta M., Joachimi B., Feeney S., 2023, Monthly Notices of the Royal Astronomical Society, 534, 3305

- Notices of the Royal Astronomical Society, 524, 6167
- Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S., 2022, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11976–11986
- Lopez-Paz D., Oquab M., 2017, in International Conference on Learning Representations. <https://openreview.net/forum?id=SJkXfE5xx>
- Loshchilov I., Hutter F., 2017, arXiv preprint arXiv:1711.05101
- Lu T., Haiman Z., Li X., 2023, Monthly Notices of the Royal Astronomical Society, 521, 2050
- Lueckmann J.-M., Bassetto G., Karaletsos T., Macke J. H., 2019, in Symposium on Advances in Approximate Bayesian Inference, pp 32–53
- Lueckmann J.-M., Boelts J., Greenberg D., Goncalves P., Macke J., 2021, in International conference on artificial intelligence and statistics, pp 343–351
- Maas A. L., Hannun A. Y., Ng A. Y., et al., 2013, in Proceedings of the 30th International Conference on Machine Learning. https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf
- Makinen T. L., Charnock T., Alsing J., Wandelt B. D., 2021, Journal of Cosmology and Astroparticle Physics, 2021, 049
- Mao X., Liu Y., Liu F., Li Q., Shen W., Wang Y., 2023, in Proceedings of the AAAI Conference on Artificial Intelligence, pp 1905–1913
- Martinet N., Harnois-Déraps J., Jullo E., Schneider P., 2021, Astronomy & Astrophysics, 646, A62
- Matilla J. M. Z., Sharma M., Hsu D., Haiman Z., 2020, Physical Review D, 102, 123506
- McCarthy I. G., Bird S., Schaye J., Harnois-Déraps J., Font A. S., Van Waerbeke L., 2018, Monthly Notices of the Royal Astronomical Society, 476, 2999
- Mead A., Heymans C., Lombriser L., Peacock J., Steele O., Winther H., 2016, Monthly Notices of the Royal Astronomical Society, 459, 1468
- Mead A., Brieden S., Tröster T., Heymans C., 2021, Monthly Notices of the Royal Astronomical Society, 502, 1401
- Mishra S., Panda R., Phoo C. P., Chen C.-F. R., Karlinsky L., Saenko K., Saligramma V., Feris R. S., 2022, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9194–9204
- Ni Y., et al., 2023, The Astrophysical Journal, 959, 136
- Oquab M., et al., 2024, Transactions on Machine Learning Research Journal, pp 1–31
- Papamakarios G., Murray I., 2016, in Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf
- Papamakarios G., Pavlakou T., Murray I., 2017, in Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf
- Papamakarios G., Sterratt D., Murray I., 2019, in Chaudhuri K., Sugiyama M., eds, Proceedings of Machine Learning Research Vol. 89, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. PMLR, pp 837–848, <https://proceedings.mlr.press/v89/papamakarios19a.html>
- Park M., Gatti M., Jain B., 2025, Physical Review D, 111, 063523
- Pillepich A., et al., 2018, Monthly Notices of the Royal Astronomical Society, 473, 4077
- Piras D., Joachimi B., Villaescusa-Navarro F., 2023, Monthly Notices of the Royal Astronomical Society, 520, 668
- Prelogović D., Mesinger A., Murray S., Fiameni G., Gillet N., 2022, Monthly Notices of the Royal Astronomical Society, 509, 3852
- Prescott T. P., Baker R. E., 2020, SIAM/ASA Journal on Uncertainty Quantification, 8, 114
- Prescott T. P., Baker R. E., 2021, SIAM/ASA Journal on Uncertainty Quantification, 9, 788
- Prescott T. P., Warne D. J., Baker R. E., 2024, Journal of Computational Physics, 496, 112577
- Radford A., et al., 2021, in Proceedings of the 38th International Conference on Machine Learning. PMLR, pp 8748–8763, <https://proceedings.mlr.press/v139/radford21a.html>
- Régaldo-Saint Blanchard B., et al., 2024, Physical Review D, 109, 083535
- Ribli D., Pataki B. Á., Zorrilla Matilla J. M., Hsu D., Haiman Z., Csabai I., 2019, Monthly Notices of the Royal Astronomical Society, 490, 1843
- Roncoli A., Ćiprijanović A., Voetberg M., Villaescusa-Navarro F., Nord B., 2023, arXiv preprint arXiv:2311.01588
- Santurkar S., Tsipras D., Ilyas A., Madry A., 2018, in Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2018/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf
- Schaye J., et al., 2023, Monthly Notices of the Royal Astronomical Society, 526, 4978
- Schneider P., Lombardi M., 2003, Astronomy & Astrophysics, 397, 809
- Schneider P., Van Waerbeke L., Jain B., Kruse G., 1998, Monthly Notices of the Royal Astronomical Society, 296, 873
- Schneider A., Teyssier R., Stadel J., Chisari N. E., Le Brun A. M., Amara A., Refregier A., 2019, Journal of Cosmology and Astroparticle Physics, 2019, 020
- Schneider A., Stoira N., Refregier A., Weiss A. J., Knabenhan M., Stadel J., Teyssier R., 2020, Journal of Cosmology and Astroparticle Physics, 2020, 019
- Secco L. F., et al., 2022, Physical Review D, 105, 103537
- Semboloni E., Schrabback T., van Waerbeke L., Vafaei S., Hartlap J., Hilbert S., 2011, Monthly Notices of the Royal Astronomical Society, 410, 143
- Sharma D., Dai B., Seljak U., 2024, Journal of Cosmology and Astroparticle Physics, 2024, 010
- Smith L. N., 2017, in 2017 IEEE winter conference on applications of computer vision (WACV), pp 464–472
- Springel V., 2005, Monthly notices of the royal astronomical society, 364, 1105
- Springel V., 2010, Monthly Notices of the Royal Astronomical Society, 401, 791
- Tahir J., Ganguli S., Rotskoff G. M., 2024, arXiv preprint arXiv:2410.08194
- Takada M., Jain B., 2003, Monthly Notices of the Royal Astronomical Society, 340, 580
- Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, The Astrophysical Journal, 761, 152
- Tejero-Cantero A., Boelts J., Deistler M., Lueckmann J.-M., Durkan C., Gonçalves P., Greenberg D., Macke J., 2022, sbi: Simulation-based inference toolkit, <https://github.com/mackelab/sbi>
- Tessore N., Loureiro A., Joachimi B., von Wietersheim-Kramsta M., Jeffrey N., 2023, The Open Journal of Astrophysics, 6
- Tucci B., Schmidt F., 2024, Journal of Cosmology and Astroparticle Physics, 2024, 063
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., 2017, in Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a84aa-Paper.pdf
- Villaescusa-Navarro F., et al., 2021, The Astrophysical Journal, 915, 71
- Villaescusa-Navarro F., et al., 2022, The Astrophysical Journal Supplement Series, 259, 61
- Warne D. J., Prescott T. P., Baker R. E., Simpson M. J., 2022, Journal of Computational Physics, 469, 111543
- Wehenkel A., Gamella J. L., Sener O., Behrmann J., Sapiro G., Cuturi M., Jacobsen J.-H., 2024, arXiv preprint arXiv:2405.08719
- Weinberger R., et al., 2016, Monthly Notices of the Royal Astronomical Society, 465, 3291
- Xavier H. S., Abdalla F. B., Joachimi B., 2016, Monthly Notices of the Royal Astronomical Society, 459, 3693
- Yang Y., Soatto S., 2020, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4085–4095
- Yeom S., Giacomelli I., Fredrikson M., Jha S., 2018, in 2018 IEEE 31st computer security foundations symposium (CSF), pp 268–282
- Yosinski J., Clune J., Bengio Y., Lipson H., 2014, in Advances in Neural Information Processing Systems. Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2014/file/532a2f85b6977104bc93f8580abb330-Paper.pdf
- Zhai X., Kolesnikov A., Houlsby N., Beyer L., 2022, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp

- 12104–12113
 Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., Xiong H., He Q., 2020,
Proceedings of the IEEE, 109, 43
 Zürcher D., et al., 2022, *Monthly Notices of the Royal Astronomical Society*,
 511, 2075
 von Wietersheim-Kramsta M., Lin K., Tessore N., Joachimi B., Loureiro A.,
 Reischke R., Wright A. H., 2025, *Astronomy & Astrophysics*, 694, A223

APPENDIX A: FURTHER POSTERIOR COMPARISONS

We present a further comparison between high-fidelity-only training, with $N = 3200$, and the transfer learning approach for the LH suite 2-parameter inference problem in Fig. A1. Despite the factor of $\times 4$ increase in the number of high-fidelity maps used during training, the multifidelity approach yields significantly tighter posteriors that better match the “accurate” model. In addition, the right-most panel in Fig. A1 is suggestive of the overconfidence issue identified in Section 3.1; high-fidelity-only models trained with fewer than $N = 6400$ IllustrisTNG maps exhibit a large degree of overconfidence, per Fig. 4.

APPENDIX B: PROBING MODEL PERFORMANCE

We performed a range of experiments to better understand the model performance. The results of these experiments are summarised in Fig. B1. We explored the impact of the paired aspect of the multifidelity simulation suite (i.e. that each simulation in the lower-fidelity N -body suite is paired with a high-fidelity simulation with identical cosmological parameters and initial conditions). This could potentially improve performance, perhaps due to implicit memorisation of the (pre-)training data (a well-studied phenomenon in deep learning, see e.g. Yeom et al. 2018; Carlini et al. 2019, 2023). We tested this by ensuring different (i.e. unpaired) cosmologies were used during pre-training and fine-tuning, and found that pairing had no discernable impact on performance.

We further investigated the neural network inference model by splitting it into two components: the neural compression performed by the CNN, and the density estimation of the NDE. In order to better disambiguate the role of each component, we took the best transfer learning models from Section 3.1 and froze the CNN. This fixed the summary statistics that were extracted from the dark matter density maps for a given transfer learning size N . We then retrained the NDE with the entire training dataset ($N = 12000$) with the frozen neural compression model. The resulting performance is shown in red in Fig. B1. We found that for very low compression training dataset size ($N = 200$), retraining the NDE with much more data led to improved MTPP performance even without more informative summary statistics. This presents more evidence that for very low fine-tuning dataset sizes, the NDE is an important bottleneck of performance. However, as the fine-tuning dataset size increases, the model performance reverts to the standard transfer learning baseline. This indicates that for larger fine-tuning datasets, the difference between an “accurate” model trained on the full LH training suite and the transfer learning models is driven by slightly poorer neural compression. These results suggest that while pre-training on N -body simulations encourages highly informative summaries, there are likely subtle differences in the high-fidelity IllustrisTNG simulations (that are useful for slightly improving cosmological constraints) which the CNN fails to discover during fine-tuning.

Fig. B1 also shows the $N = 0$ transfer learning case, where only the N -body simulation pre-training is performed and no high-fidelity

maps are used. The extremely poor performance indicates that there are significant differences between the different simulation fidelities, and a fine-tuning step is necessary.

Finally, we test whether transfer learning for (the more observationally important) total matter density M_{tot} behaves any differently. A key concern is that the improved performance from transfer learning could be largely due to the strong similarity between the IllustrisTNG dark matter density M_{cdm} and that of dark matter-only N -body simulations. Figure B2 demonstrates that multifidelity transfer learning performs just as well when fine-tuning on M_{tot} .

We compare the results from Section 3.1 with inference on the M_{tot} field with an identical methodology. We find that transfer learning still leads to up to an order-of-magnitude reduction in the number of high-fidelity maps required to train an accurate, trustworthy inference model, compared with high-fidelity-only training.

The small downward shift of all M_{tot} inference performance curves (“accurate”, transfer learning and high-fidelity-only) on the MTPP metric from Section 3.1a indicates that inference using the M_{tot} maps is slightly more challenging. However, there is also a slightly larger gap between the “accurate” M_{tot} model and transfer learned models (and low N high-fidelity-only models) compared with inference results on M_{tot} . This suggests that: i) some features in M_{tot} require a large number of training maps ($N > 6400$) for the CNN to learn to extract, more-so than in the M_{cdm} case, and ii) N -body pre-training gives slightly less informative features than in the M_{cdm} case, perhaps for similar reasons as i).

APPENDIX C: UNCONSTRAINED PARAMETERS IN SB28

The posterior estimation models in Section 3.2 were trained to perform 5-dimensional inference on $\{\Omega_m, \sigma_8, n_s, h, \Omega_b\}$. However, we found that h and Ω_b could not be constrained by the data (or, partially, by the CNN-NDE architecture). Here we present some more details on these unconstrained parameters.

Figure C1 shows the posterior sample ensemble mean MSE for two cosmological parameters: σ_8 and Ω_b . The posterior recovery of σ_8 behaves similarly to Section 3.1, with very good performance relative to the “accurate” baseline. On the other hand, we find that pre-training on N -body simulations leads to no improvement over the high-fidelity-only training approach for Ω_b . In one sense this is expected: N -body simulations do not provide a strong probe of how Ω_b affects dark matter maps (beyond the initial matter power spectrum), and so there should not be much direct transfer of knowledge. In addition, Fig. C2 demonstrates that there is little constraining information on Ω_b in the dark matter density maps.

However, the fact that the high quality pre-trained summary statistics cannot be adapted to improve inference of Ω_b is, at least naively, somewhat surprising. This suggests that the features relevant for inferring Ω_b are disjoint from those governed by variations in $\{\Omega_m, \sigma_8, n_s, h\}$ for N -body simulations, resulting in a representation mismatch that prevents effective transfer from pre-training. Ni et al. (2023) demonstrated that the values of Ω_b explored in the CMD SB28 simulation suite had a minor effect on both the star formation rate density and the gas power spectra of the simulations, smaller even than several of the astrophysical nuisance parameters. In addition to this, since Ω_b primarily modulates the amount of gas available for star formation and black hole accretion (Elbers et al. 2025), these signatures may be occluded by the wide range of nuisance parameters affecting baryonic feedback in the hydrodynamical simulations.

On the other hand, we found that while h could not be properly constrained by the data, transfer learning yielded similar constraints

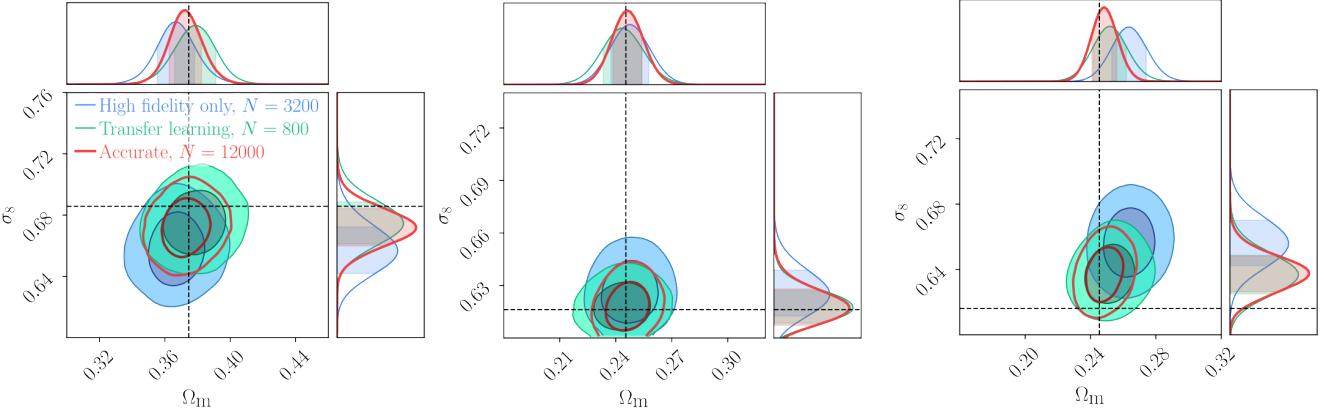


Figure A1. Three representative examples of inference from the LH simulation suite. The true cosmology is shown by the black dashed line. A model trained using transfer learning with $N = 800$ high-fidelity IllustrisTNG maps is compared against a high-fidelity-only model trained with $N = 3200$ maps. The posteriors are compared with an “accurate” posterior model that was trained using the full simulation suite.

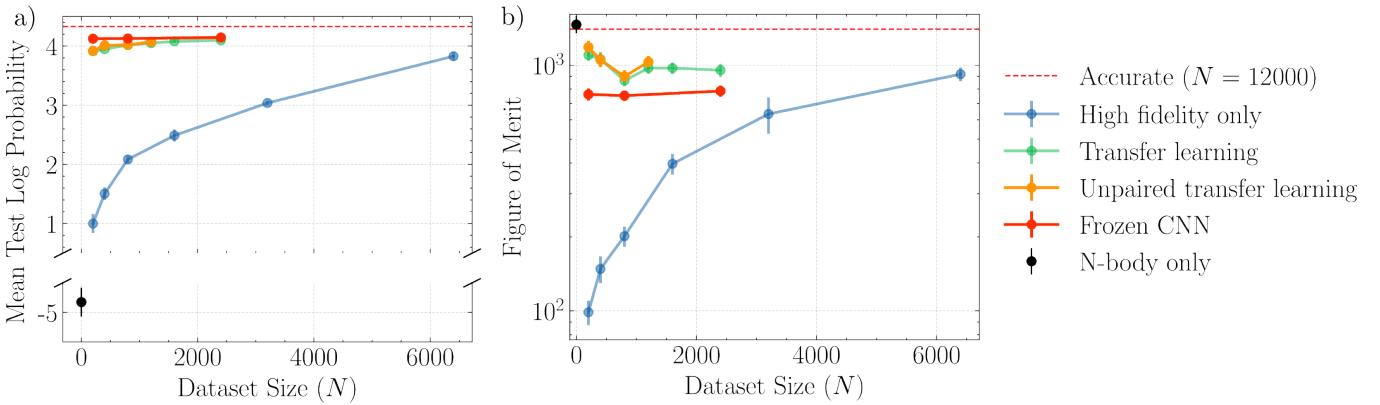


Figure B1. Comparing inference results on the IllustrisTNG LH suite for various experiments. We reproduce the results from Fig. 3 for a) MTPP (note that the y-axis scale has been shortened for enhanced visualisation) and b) FoM. We also show results from experiments: pre-training and fine-tuning without any paired data (orange); training a neural compression model with N IllustrisTNG maps and then freezing the CNN compression to train an NDE with the full ($N = 12000$) LH training suite (red); and the performance of models only pre-trained on N -body simulations (black, corresponding to $N = 0$ IllustrisTNG maps).

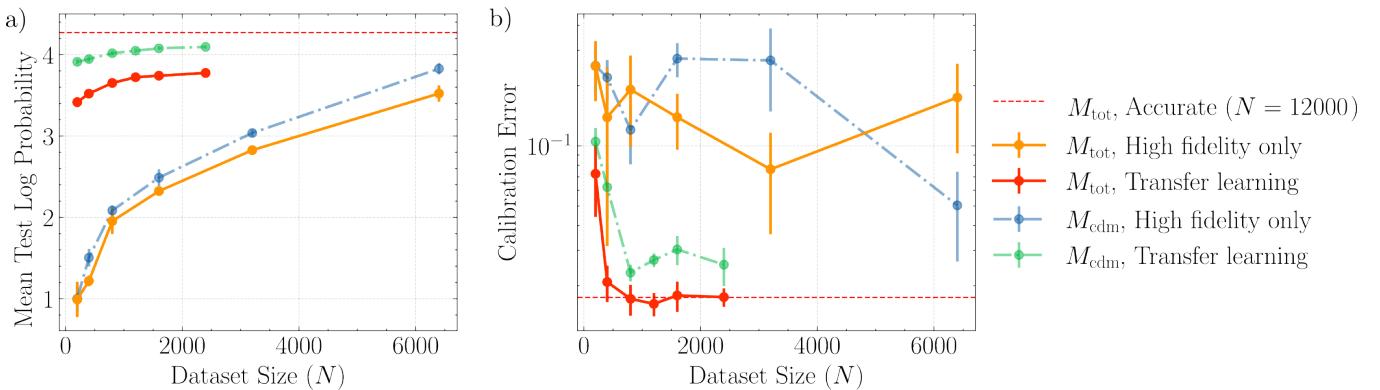


Figure B2. Comparing transfer learning results on the IllustrisTNG LH suite using M_{cdm} (dot-dashed lines) and M_{tot} (solid lines). We reproduce the M_{cdm} results from Fig. 3 for a) MTPP and b) calibration error. Pre-training on N -body simulations gives near equivalent improvements over training with only high-fidelity maps when performing inference on M_{tot} fields.

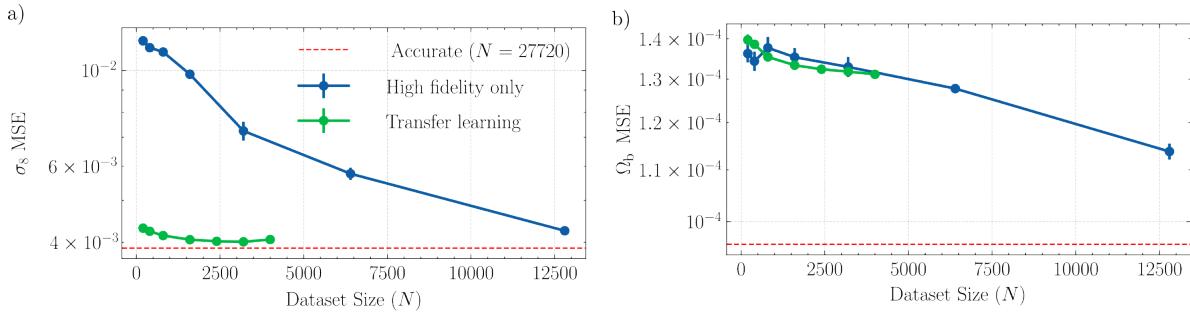


Figure C1. Mean squared error (MSE) between the inferred posterior mean $\hat{\theta}$ and the true cosmology θ . Results are broken down per-parameter, with panel a) showing σ_8 and panel b) showing Ω_b . While transfer learning yields a significant improvement in σ_8 , we find negligible impact on Ω_b , which has little effect on the N -body simulations.

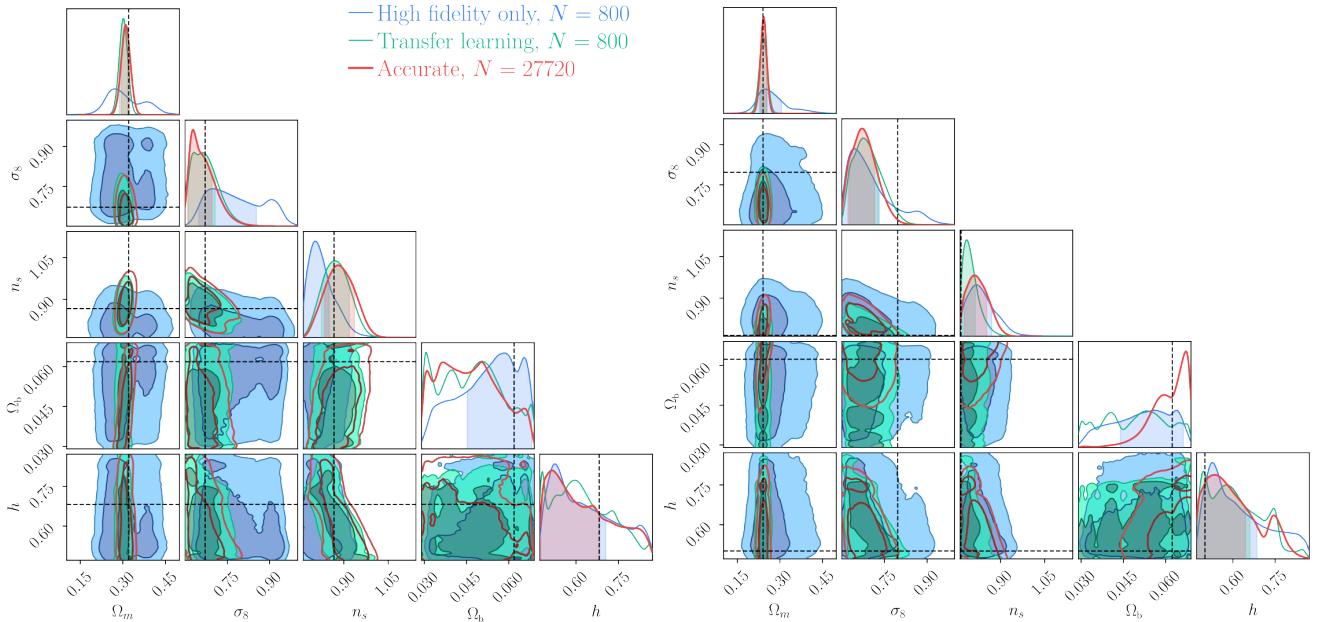


Figure C2. Two examples of posterior inference on IllustrisTNG dark matter maps from the SB28 test suite over the full 5-dimensional posterior. The true cosmology is shown by the black dashed line. A model trained using transfer learning with $N = 800$ high-fidelity IllustrisTNG maps is compared against a high-fidelity-only model trained with $N = 3200$ maps. The posteriors are compared with an ‘‘accurate’’ posterior model that was trained using the full simulation suite. We find that for most 2D dark matter density maps, Ω_b and h are unconstrained.

to the ‘‘accurate’’ baseline (as opposed to Ω_b , which was poorly constrained *and* transfer learning gave no benefit).

Two examples of 5-dimensional posterior inference are given in Fig. C2. In the first example, none of the models can constrain h and Ω_b much beyond the uniform prior. We can highlight two key qualitative features: the fine-tuned approach gives significantly better agreement with the baseline than training from scratch, and it is statistically consistent with the accurate posterior. In the second example, the ‘‘accurate’’ posterior gives a (weak) constraint on Ω_b , while the other models fail to provide any constraints. A small but not insignificant fraction of the inferred posteriors follows this second pattern, which is consistent with the minor improvement in constraining power of Ω_b shown in Fig. C1. We found that these examples tended to coincide with extreme cosmologies (at the boundaries of the prior volume), and particularly for large values of Ω_b , as is the

case in Fig. C2. We reserve a more systematic analysis as a potential avenue for future work.