Multiscale Flow for Robust and Optimal Cosmological Analysis

Biwei Dai [1 2] **Uroš Seljak** [1 2]

Abstract

We propose Multiscale Flow, a generative Normalizing Flow that creates samples and models the field-level likelihood of two dimensional cosmological data such as weak lensing, thus enabling Simulation Based Likelihood Inference. Multiscale Flow uses hierarchical decomposition of cosmological fields via a wavelet basis, and then models different wavelet components separately as Normalizing Flows. This decomposition allows us to separate the information from different scales and identify distribution shifts in the data such as unknown scale-dependent systematics. The resulting likelihood analysis can not only identify these types of systematics, but can also be made optimal, in the sense that the Multiscale Flow can learn the full likelihood at the field without any dimensionality reduction.

1. Introduction

The late-time cosmological fields are highly non-Gaussian with no tractable likelihood functions. Extracting information from these non-Gaussian fields has been mainly attempted through a limited set of summary statistics (Peebles, 1980; Jain & Van Waerbeke, 2000; Allys et al., 2020; Fluri et al., 2018). However, these analyses have the same underlying issues of summary statistics being ad-hoc and potentially sub-optimal.

Recently, Dai & Seljak (2022) proposed learning the field-level data likelihood with Normalizing Flows (NFs). This approach does not require compressing the data into a low-dimensional summary statistic, and instead tries to extract all the information in the data from the field level likelihood. They show that NF likelihood agrees well with analytical solution on Gaussian Random Fields, and it leads to significant improvement over the standard power spectrum analysis on nonlinear matter fields from N-body simulations.

Despite the differences in these LSS analyses methods, they all face the same challenge of robustness: how do we know which information is reliable, and which is not, if it is corrupted by effects that are ignored or inaccurately modeled? How do we detect distribution shifts in the actual data that were not in the training data? While marginalizing over the baryon parameters, subgrid models and various systematic effects are helpful and necessary, there is no guarantee that current baryon and systematic models span all potential realistic scenarios.

One way to mitigate the impact of such modeling uncertainties is by separation of scales, with very small-scale information likely being contaminated by many astrophysical nuisance effects and observation systematics, and large-scale information likely being more robust. This strategy is widely used in current cosmological survey analyses of power spectrum or correlation function (Krause et al., 2017; Doux et al., 2021). In this paper we apply the scale separation idea to the field-level likelihood modeling with NFs. Specifically, we use a set of scale-separated basis functions to represent the pixelized data, and decompose the data likelihood function into the contributions from different scales. Performing consistency checks between the different scales enables us to decide what scale to include and what to exclude. Furthermore, our hierarchical analysis also combines likelihood information from different scales to achieve optimality in the limit of sufficient training data. In this work we use wavelet basis, which is localized in both real space and Fourier space, allowing us to easily handle the survey mask and to separate the signals from different physical scales. Such decomposition is also known as Multiresolution Analysis (MRA) in image processing.

2. Multiresolution Analysis with Fast Wavelet Transform

In this section we briefly introduce Multiresolution Analysis (MRA), which hierarchically decomposes the data into components at different scales, allowing us to separate the information from different scales and study them individually. MRA is usually performed with Fast Wavelet Transform

[1]Berkeley Center for Cosmological Physics and Department of Physics, University of California, Berkeley, CA 94720, USA [2]Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA 94720, USA. Correspondence to: Biwei Dai <biwei@berkeley.edu>.

ICML 2023 Workshop on Machine Learning for Astrophysics, Honolulu, Hawaii, USA. PMLR 202, 2023.

(FWT) (Mallat, 1989). We focus on decimated wavelet transform, which preserves the dimensionality of the data and can be viewed as a special kind of bijective NF transform.

The basic idea of FWT is to recursively apply low-pass filters (also called scaling functions) and high-pass filters (also called wavelet functions) to the data. In each iteration, the data x_{2^n} with resolution 2^n is decomposed into a low-resolution approximation $x_{2^{n-1}}$, and detail coefficients of the remaining signal $x^d_{2^{n-1}}$:

$$x_{2^{n-1}} = (\phi * x_{2^n}) \downarrow 2 \quad (1)$$
$$x^d_{2^{n-1}} = (\psi * x_{2^n}) \downarrow 2 \quad (2)$$

where ϕ is the low pass filter (scaling function), ψ is the high pass filter (wavelet function), $*$ is convolution operation, and $\downarrow 2$ is the operator to downsample the data by a factor of 2: $(x \downarrow 2)_{i,j} = x_{2i,2j}$. For a 2D map x_{2^n}, there are three high pass filters and the dimension of $x^d_{2^{n-1}}$ is $3 \times 2^{n-1} \times 2^{n-1}$. The low-resolution data $x_{2^{n-1}}$ is passed to the next iteration and treated as input for further decomposition (Figure 1).

In this work we consider Haar wavelet (Haar, 1910), the simplest and the most spatially localized wavelet function. The scaling function and wavelet function of Haar wavelet can be represented by 2×2 kernels in real space:

$$\phi = \frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \psi_1 = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad (3)$$
$$\psi_2 = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \psi_3 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \quad (4)$$

With MRA, the log likelihood of a map x_{2^n} with resolution 2^n can be rewritten into an auto-regressive form as

$$\begin{aligned} \log p(x_{2^n}|y) &= \log p(x_{2^{n-1}}, x^d_{2^{n-1}}|y) \\ &= \log p((x_{2^{n-1}}|y) + \log p(x^d_{2^{n-1}}|x_{2^{n-1}}, y) \\ &= \log p(x_{2^{n-2}}|y) + \log p(x^d_{2^{n-2}}|x_{2^{n-2}}, y) \\ &\quad + \log p(x^d_{2^{n-1}}|x_{2^{n-1}}, y) \\ &= \dots \\ &= \log p(x_{2^k}|y) + \textstyle\sum_{m=k}^{n} \log p(x^d_{2^m}|x_{2^m}, y), \end{aligned} \quad (5)$$

where 2^k is the scale where we stops the decomposition. In practice, we can choose k such that it corresponds to the scale that either has extracted all the information from the data, or is large enough not to be affected by unknown small scale systematic effects.

3. Multiscale Flow

3.1. Normalizing Flows

Flow-based models provide a powerful framework for density estimation (Dinh et al., 2017; Papamakarios et al., 2017) and sampling (Kingma & Dhariwal, 2018). These models map the data x to latent variables z through a sequence of invertible transformations $f = f_1 \circ f_2 \circ ... \circ f_n$, such that $z = f(x)$ and z is mapped to a base distribution $\pi(z) = \mathcal{N}(0, I)$. The probability density of data x can be evaluated using the change of variables formula: $p(x) = \pi(f(x)) \left|\det\left(\frac{\partial f(x)}{\partial x}\right)\right|$. To sample from $p(x)$, one first samples latent variable z from $\pi(z)$, and then transform variable z to x through $x = f^{-1}(z)$.

In cosmological analysis we are interested in the likelihood function $p(x|y)$, which can be estimated using conditional Normalizing Flows (NFs). In conditional NFs the flow transformation is dependent on the conditional parameters y, i.e., $f = f_y$. We discuss below how we parametrize the conditional flow f_y.

3.2. Multiscale Flow

With the likelihood decomposition Equation 5, our task now is to build NFs to model different likelihood terms separately. For simplicity we will drop the subscript 2^m in this section. The model described here is similar to Wavelet Flow (Yu et al., 2020), even though they are developed independently. Following Glow (Kingma & Dhariwal, 2018), our flow transformation $f(x|y)$ consists of multiple block flows, where each block consists of an actnorm, an invertible 1×1 convolution, and an affine coupling layer.

Actnorm: The actnorm layer applies an affine transformation per channel, similar to batch normalization (Ioffe & Szegedy, 2015), but its scale and bias parameters are initialized such that the output has zero mean and unit variance per channel given an initial minibatch of data, and then these parameters are treated as regular trainable parameters.

Invertible 1×1 convolution: The invertible 1×1 convolution is a learnable $C \times C$ matrix (where C is the number of channels) that linearly mixes different channels.

Affine coupling: The affine coupling layer firstly splits the data x^d to x^{d1} and x^{d2} based on the channels, and then applies pixelwise affine transformation to x^{d2}, with scale and bias given by x^{d1}:

$$(\log s,\ t) = \text{CNN}(x^{d1}, x, y) \quad (6)$$
$$z^{d2} = \exp(\log s) \cdot x^{d2} + t, \quad (7)$$

where $\log s$ and t are scale and bias coefficient maps with the same dimensionality as x^{d2}, and CNN is a learned function parametrized by a convolutional neural network. The dependence of conditional parameter y is modeled by introducing gating into CNN, i.e., each channel of CNN is scaled by a value between 0 and 1 which is determined by parameter y. The gating allows the conditional variable y to determine the relative weights between different features (channels). The output of the affine coupling layer is the concatenation of x^{d1} and z^{d2}. In this paper we consider 2D maps, so at each scale x^d contains 3 channels. We set the

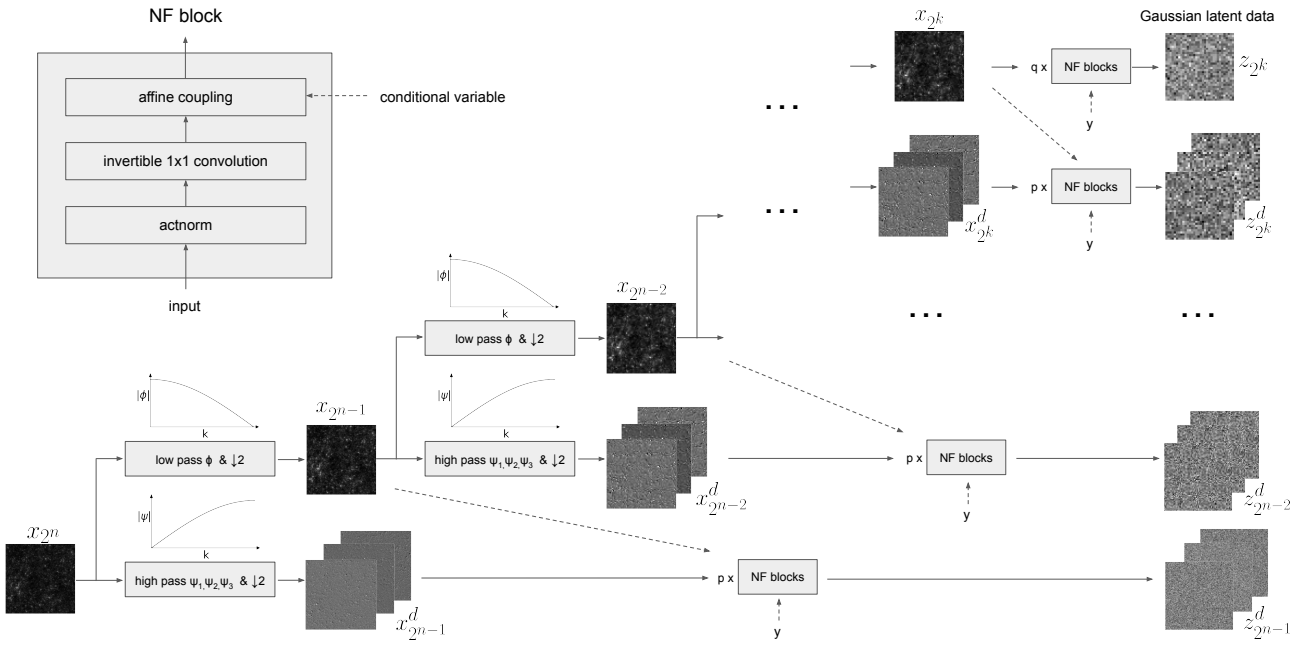

Figure 1. Illustration of Multiscale Flow model. The input map x_{2^n} with resolution 2^n is iteratively processed with a set of low pass filters (ϕ), high pass filters (ψ_1, ψ_2, ψ_3) and downsampling ($\downarrow 2$), resulting in a series of detailed maps $x^d_{2^{n-1}}, x^d_{2^{n-2}}, \cdots, x^d_{2^k}$ and an approximation map x_{2^k}. These maps are then transformed by several NF blocks to Gaussian latent maps $z^d_{2^{n-1}}, z^d_{2^{n-2}}, \cdots, z^d_{2^k}, z_{2^k}$, where each NF block is composed of an actnorm layer, an invertible 1×1 convolution, and an affine coupling layer (Equation 6, 7), as shown on the top left of this figure. The NF transformation is conditioned on conditional variable y and approximation maps, which are represented by dashed arrows in the illustration. The log likelihood of the input map x_{2^n} can be calculated with Equation 5.

first channel to be x^{d1}, and the other two channels to be x^{d2}.

In this work the large-scale term $\log p(x_{2^k}|y)$ is modeled by q flow blocks, and each other term $\log p(x^d_{2^m}|x_{2^m}, y)$ is modeled with p flow blocks, where p and q are hyperparameters in the model. See Figure 1 for an Illustration of the Multiscale Flow model. All of these NFs can be trained independently in parallel to speed up the training process. The details of the training can be found in the Appendix.

4. Results

4.1. Cosmological constraints from weak lensing maps

We apply Multiscale Flow to $3.5 \times 3.5\text{deg}^2$ mock weak lensing convergence maps (Ribli et al., 2019) for field-level inference. We decompose the 512^2 resolution map to four scales, with likelihood decomposition $\log p(x_{512}|y) = \log p(x_{64}|y) + \log p(x^d_{64}|x_{64}, y) + \log p(x^d_{128}|x_{128}, y) + \log p(x^d_{256}|x_{256}, y)$. The constraining power of Multiscale Flow of different galaxy shape noise level is shown in Table 1. We list the figure of merit of maps with different resolutions, and compare them with summary statistics power spectrum, peak count, scattering transform (Cheng et al., 2020), and statistics learned by CNNs (Ribli et al., 2019). Multiscale Flow achieves the best performance among all methods, outperforming power spectrum by factors of 3, 5 and 9 on galaxy densities $n_g = 10, 30, 100\text{arcmin}^{-2}$, respectively. Multiscale Flow also achieves 2 - 3 higher constraining power when compared to peak counts, CNN, and scattering transform. This is mainly because Multiscale Flow models the field-level likelihood function and tries to learn all the information in the data, while power spectrum, peak count, and scattering transform compress the data and only learn partial information. CNN is a powerful data compressor and has been shown to achieve optimal compression on Gaussian random fields and log-normal fields with sufficient training data (Makinen et al., 2021). However, in the regime where training data is expensive to generate and insufficient, there is evidence that generative models (Multiscale flow) perform better and is less prone to overfitting compared to discriminative models (CNN) (Ng & Jordan, 2001). We plan to study this topic to better understand their difference in our future work.

4.2. Impact of baryons

Next we apply Multiscale Flow to mock weak lensing maps with baryonic physics included (Lu et al., 2022). Similar to the previous experiment, these maps also have a resolution of 512^2, and we adopt the same likelihood decomposition. For these maps we have 4 additional baryon parameters

Table 1. Comparison of the constraining power between different methods. The figure of merit is measured by the reciprocal of the 1σ confidence area on the (Ω_m, σ_8) plane, using a 3.5×3.5 deg^2 convergence map.

Method	$n_g = 10\text{arcmin}^{-2}$	$n_g = 30\text{arcmin}^{-2}$	$n_g = 100\text{arcmin}^{-2}$
Multiscale Flow $p(x_{512}\|y)$	**149**	**375**	**957**
Multiscale Flow $p(x_{256}\|y)$	143	353	858
Multiscale Flow $p(x_{128}\|y)$	108	281	706
Multiscale Flow $p(x_{64}\|y)$	72	175	391
power spectrum	30 (30)	52 (51)	81 (79)
peak count	(40)	(85)	(137)
CNN	(44)	(121)	(292)
scattering transform $s_0 + s_1 + s_2$	($\lesssim 50$)	($\lesssim 140$)	($\lesssim 329$)

1. Unless specified with Multiscale Flow, the analysis of other approaches are performed on maps with resolution 512^2.
2. The numbers in parenthesis are estimated using maps with 1 arcmin smoothing. We expect this smoothing to have little effect on constraining power estimation, because the small-scale modes are dominated by shape noise. This is also explicitly verified in the case of power spectrum, where we present FoM with and without smoothing.
3. The FoM of scattering transform is estimated using Fisher matrix, which is an upper limit of the true FoM according to the Cramér-Rao inequality. It has been shown that Fisher forecast could potentially overestimate the parameter constraints, due to the non-Gaussian distribution of the statistics (Park et al., 2022).

Table 2. Similar to Table 1, but on a different dataset with baryon physics included. Here we show the constraining power after marginalizing over baryonic effects.

Method	$n_g = 10\text{arcmin}^{-2}$	$n_g = 20\text{arcmin}^{-2}$	$n_g = 50\text{arcmin}^{-2}$	$n_g = 100\text{arcmin}^{-2}$
Multiscale Flow $p(x_{512}\|y)$	**149**	**220**	**362**	**521**
Multiscale Flow $p(x_{256}\|y)$	147	213	341	494
Multiscale Flow $p(x_{128}\|y)$	112	166	269	398
Multiscale Flow $p(x_{64}\|y)$	75	113	183	259
power spectrum	34(33)	48(48)	68(65)	84 (78)

Lu et al. (2022) measures the Figure of merit of CNN on a much larger map (1500deg^2), so it's hard to perform a direct comparison. However, Lu et al. (2022) shows that CNN improves about 60% compared to the power spectrum, while our approach leads to a $3-5$ times improvement compared to the power spectrum.

(Aricò et al., 2020). In Table 2 we compare the constraining power of Multiscale Flow and power spectrum on (Ω_m, σ_8) plane, after marginalizing over the baryon parameters. With the presence of baryon physics, Multiscale Flow has $4-6$ times higher constraining power on cosmological parameters when compared to power spectrum. Note that because these maps are generated with different ray-tracing pipelines compared to Ribli et al. (2019), we cannot directly compare the results between Table 1 and Table 2.

4.3. Identifying distribution shifts with scale-dependent posterior analysis

Identifying distribution shifts from unknown effects that are present in the data, but not in the training simulations, is one of the great challenges of modern Machine Learning. Here we propose to use consistency of information as a function of scale to identify such shifts. As a simple example, we train the Multiscale Flow with dark-matter-only convergence maps (Ribli et al., 2019), and apply the model to convergence maps with baryon physics included (Lu et al., 2022). We show the posterior distributions from different scales in the left panel of Figure 2. The baryon physics

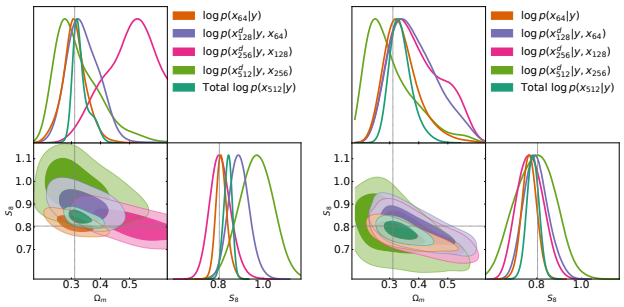

Figure 2. Scale-dependent posterior analysis of a baryon-corrected convergence map using Multiscale Flow trained on dark-matter-only maps (left panel), and Multiscale Flow trained on BCM maps (right panel, no distribution shift).

modifies the matter distribution on small scales and bias the posterior constraints from small scales. In this case naively combining all of the scales leads to a posterior constraint that is 2σ biased (dark green contour). The inconsistency of posterior between different scales suggests a presence of an unknown systematics (baryon physics) that is not modeled in the training data. If we remove the small scale informa-

tion (because we believe the large scales are less likely to be affected by systematics), we can recover an unbiased constraint of cosmological parameters (orange contour).

As a comparison, in the middle panel of Figure 2 we show the posteriors from Multiscale Flow trained using maps with baryon physics. There is no distribution shift in this case and the information from the different scales is consistent.

5. Discussion

In this paper our main focus is optimal and robust field-level likelihood analysis (also see Appendix for sample generation). We expect many applications of Multiscale Flow, such as 21cm and other intensity maps, weak lensing maps, projected galaxy clustering, X-ray and thermal SZ maps etc. Multiscale Flow can also be used to model 3D galaxy field or 1D spectrum data like Lyman alpha forest.

Acknowledgements

We thank the Columbia Lensing group (http://columbialensing.org) for making their simulations available. B.D. thanks Xiangchong Li for helpful discussions on wavelet transform. This work is supported by U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Contract No. DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory to enable research for Data-intensive Machine Learning and Analysis.

References

Allys, E., Marchand, T., Cardoso, J.-F., Villaescusa-Navarro, F., Ho, S., and Mallat, S. New interpretable statistics for large-scale structure analysis and generation. *Physical Review D*, 102(10):103506, 2020.

Aricò, G., Angulo, R. E., Hernández-Monteagudo, C., Contreras, S., Zennaro, M., Pellejero-Ibañez, M., and Rosas-Guevara, Y. Modelling the large-scale mass density field of the universe as a function of cosmology and baryonic physics. *Monthly Notices of the Royal Astronomical Society*, 495(4):4800–4819, July 2020. doi: 10.1093/mnras/staa1478.

Cheng, S., Ting, Y.-S., Ménard, B., and Bruna, J. A new approach to observational cosmology using the scattering transform. *Monthly Notices of the Royal Astronomical Society*, 499(4):5902–5914, December 2020. doi: 10.1093/mnras/staa3165.

Dai, B. and Seljak, U. Translation and rotation equivariant normalizing flow (TRENF) for optimal cosmological analysis. *Monthly Notices of the Royal Astronomical Society*, 516(2):2363–2373, October 2022. doi: 10.1093/mnras/stac2010.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Doux, C., Baxter, E., Lemos, P., Chang, C., Alarcon, A., Amon, A., Campos, A., Choi, A., Gatti, M., Gruen, D., et al. Dark energy survey internal consistency tests of the joint cosmological probes analysis with posterior predictive distributions. *Monthly Notices of the Royal Astronomical Society*, 503(2):2688–2705, 2021.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Fluri, J., Kacprzak, T., Refregier, A., Amara, A., Lucchi, A., and Hofmann, T. Cosmological constraints from noisy convergence maps through deep learning. *Physical Review D*, 98(12):123518, 2018.

Haar, A. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen.*, 69(3):331–371, 1910. ISSN 0025-5831.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. Averting A crisis in simulation-based inference. *CoRR*, abs/2110.06581, 2021. URL https://arxiv.org/abs/2110.06581.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jain, B. and Van Waerbeke, L. Statistics of Dark Matter Halos from Gravitational Lensing. *The Astrophysical Journal Letters*, 530(1):L1–L4, February 2000. doi: 10.1086/312480.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10236–10245, 2018.

Krause, E., Eifler, T. F., Zuntz, J., Friedrich, O., Troxel, M. A., Dodelson, S., Blazek, J., Secco, L. F., MacCrann, N., Baxter, E., Chang, C., Chen, N., Crocce, M., DeRose, J., Ferte, A., Kokron, N., Lacasa, F., Miranda, V., Omori,

Y., Porredon, A., Rosenfeld, R., Samuroff, S., Wang, M., Wechsler, R. H., Abbott, T. M. C., Abdalla, F. B., Allam, S., Annis, J., Bechtol, K., Benoit-Levy, A., Bernstein, G. M., Brooks, D., Burke, D. L., Capozzi, D., Carrasco Kind, M., Carretero, J., D'Andrea, C. B., da Costa, L. N., Davis, C., DePoy, D. L., Desai, S., Diehl, H. T., Dietrich, J. P., Evrard, A. E., Flaugher, B., Fosalba, P., Frieman, J., Garcia-Bellido, J., Gaztanaga, E., Giannantonio, T., Gruen, D., Gruendl, R. A., Gschwend, J., Gutierrez, G., Honscheid, K., James, D. J., Jeltema, T., Kuehn, K., Kuhlmann, S., Lahav, O., Lima, M., Maia, M. A. G., March, M., Marshall, J. L., Martini, P., Menanteau, F., Miquel, R., Nichol, R. C., Plazas, A. A., Romer, A. K., Rykoff, E. S., Sanchez, E., Scarpine, V., Schindler, R., Schubnell, M., Sevilla-Noarbe, I., Smith, M., Soares-Santos, M., Sobreira, F., Suchyta, E., Swanson, M. E. C., Tarle, G., Tucker, D. L., Vikram, V., Walker, A. R., and Weller, J. Dark Energy Survey Year 1 Results: Multi-Probe Methodology and Simulated Likelihood Analyses. *arXiv e-prints*, art. arXiv:1706.09359, June 2017.

Lu, T., Haiman, Z., and Zorrilla Matilla, J. M. Simultaneously constraining cosmology and baryonic physics via deep learning from weak lensing. *Monthly Notices of the Royal Astronomical Society*, 511(1):1518–1528, March 2022. doi: 10.1093/mnras/stac161.

Makinen, T. L., Charnock, T., Alsing, J., and Wandelt, B. D. Lossless, scalable implicit likelihood inference for cosmological fields. *Journal of Cosmology and Astroparticle Physics*, 2021(11):049, 2021.

Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pp. 841–848. MIT Press, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.html.

Papamakarios, G., Murray, I., and Pavlakou, T. Masked autoregressive flow for density estimation. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2338–2347, 2017.

Park, C. F., Allys, E., Villaescusa-Navarro, F., and Finkbeiner, D. P. Quantification of high dimensional non-gaussianities and its implication to fisher analysis in cosmology. *arXiv preprint arXiv:2204.05435*, 2022.

Peebles, P. J. E. *The large-scale structure of the universe*. 1980.

Ribli, D., Pataki, B. Á., Zorrilla Matilla, J. M., Hsu, D., Haiman, Z., and Csabai, I. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society*, 490(2):1843–1860, December 2019. doi: 10.1093/mnras/stz2610.

Yu, J. J., Derpanis, K. G., and Brubaker, M. A. Wavelet flow: Fast training of high resolution normalizing flows. *Advances in Neural Information Processing Systems*, 33:6184–6196, 2020.

A. Training of Multiscale Flow

Following (Dai & Seljak, 2022), we adopt a two-stage training strategy in this work: we first train the NF with the generative loss, which minimizes the negative log-likelihood and is the standard loss function of NF:

$$\mathcal{L}_{\rm g} = -\frac{1}{N}\sum_{i=1}^{N}\log p(x_i|y_i). \tag{8}$$

The generative loss is suitable for sampling and density estimation, but may lead to biased or overconfident posterior as is shown in Figure 7 of (Dai & Seljak, 2022). To solve this issue they propose further optimizing the posteriors by training the model with the discriminative loss, $\mathcal{L}_{\rm d} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i) = -\frac{1}{N}\sum_{i=1}^{N}\left[\log p(x_i|y_i) + \log p(y_i) - \log p(x_i)\right]$, where the evidence $p(x)$ is estimated using Importance Sampling (IS): $\log p(x) \approx \log \frac{1}{M}\sum_{y_j\sim q(y|x)}^{M}\frac{p(x|y_j)p(y_j)}{q(y_j|x)}$, and $q(y|x)$ is chosen to be a Gaussian distribution with learned mean and fixed covariance matrix. However, we find that IS becomes inefficient when the number of parameters y gets large and when the posterior becomes non-Gaussian. Furthermore, training on posteriors may not always solve the overconfidence problem (Hermans et al., 2021).

In this work, we propose adding another loss term that forces the posteriors to be well calibrated. Our proposed loss term is based on the notion that properly calibrated Bayesian posteriors have to result in correct frequency of events. If the prior is flat the posterior is entirely determined by the likelihood. We thus enforce that the average likelihood over the training data equals the likelihood averaged over the posterior. The loss is defined as

$$\begin{aligned}\tilde{\mathcal{L}}_{\rm d} =& \frac{1}{N}\sum_{i=1}^{N}\log p(x_i|y_i) - \int p(y|x_i)\log p(x_i|y)dy \\ \approx& \frac{1}{N}\sum_{i=1}^{N}\log p(x_i|y_i) - \frac{1}{M}\sum_{y_j\sim p(y|x_i)}^{M}\log p(x_i|y_j).\end{aligned} \tag{9}$$

After the generative training we add this loss to the generative loss with a hyperparameter λ, $\mathcal{L} = \frac{1}{1+w\lambda}\mathcal{L}_{\rm g} + \frac{w\lambda}{1+w\lambda}\tilde{\mathcal{L}}_{\rm d}$, where $w = \frac{d_x}{d_y}$ is a prefactor to balance the dimension difference between the data and the parameter space, and we divide the loss by $1 + w\lambda$ to normalize the weights. In Figure 3 we show the percentage of outliers in our posterior analysis with different λ values. For very small λ the posterior is too narrow (underestimated errors) and the loss is dominated by the first loss term (generative loss). For $\lambda > 0.1$ the posterior is well calibrated due to the second term $\tilde{\mathcal{L}}_{\rm d}$. In this paper we use large λ to calibrate the posterior (shown in Table 3).

In the second line of Eq. 9 we replaced the integral with Monte Carlo estimates, which requires samples from the posterior $p(y|x_i)$. We obtain the posterior samples by running a Hamiltonian Monte Carlo (HMC) sampler (Duane et al., 1987) before the discriminative training. In this paper we generate $M = 5 - 10$ HMC samples for each data with 200 HMC steps. These samples are saved, and then updated with 5 - 20 HMC steps every epoch of training. An advantage of this method is that instead of evaluating the evidence term $\log p(x) = \log\int p(x|y)p(y)dy$, we now evaluate $\int \log p(x|y)p(y|x)dy$. The estimation of the former usually comes with large variance, while the latter can be estimated with only a few HMC samples.

B. Multiscale Flow Hyperparameters

We use $p = 12$ block flows to model the large-scale term $\log p(x_{64}|y)$, and $q = 4$ block flows to model each of the three small-scale terms. The CNN in Equation 6 is chosen to be a convolutional residual neural network with 2 residual blocks and 64 hidden channels in the residual blocks.

C. Empirical coverage probability of posteriors

On weak lensing maps with baryon physics, we apply Multiscale Flow to test data with fiducial parameters, and in Table 3 we report the percentage of test data with true cosmological parameters fall in 68% and 95% confidence regions. In most cases the percentages are larger than the 68% and 95% expectation, suggesting that our posterior constraint is conservative.

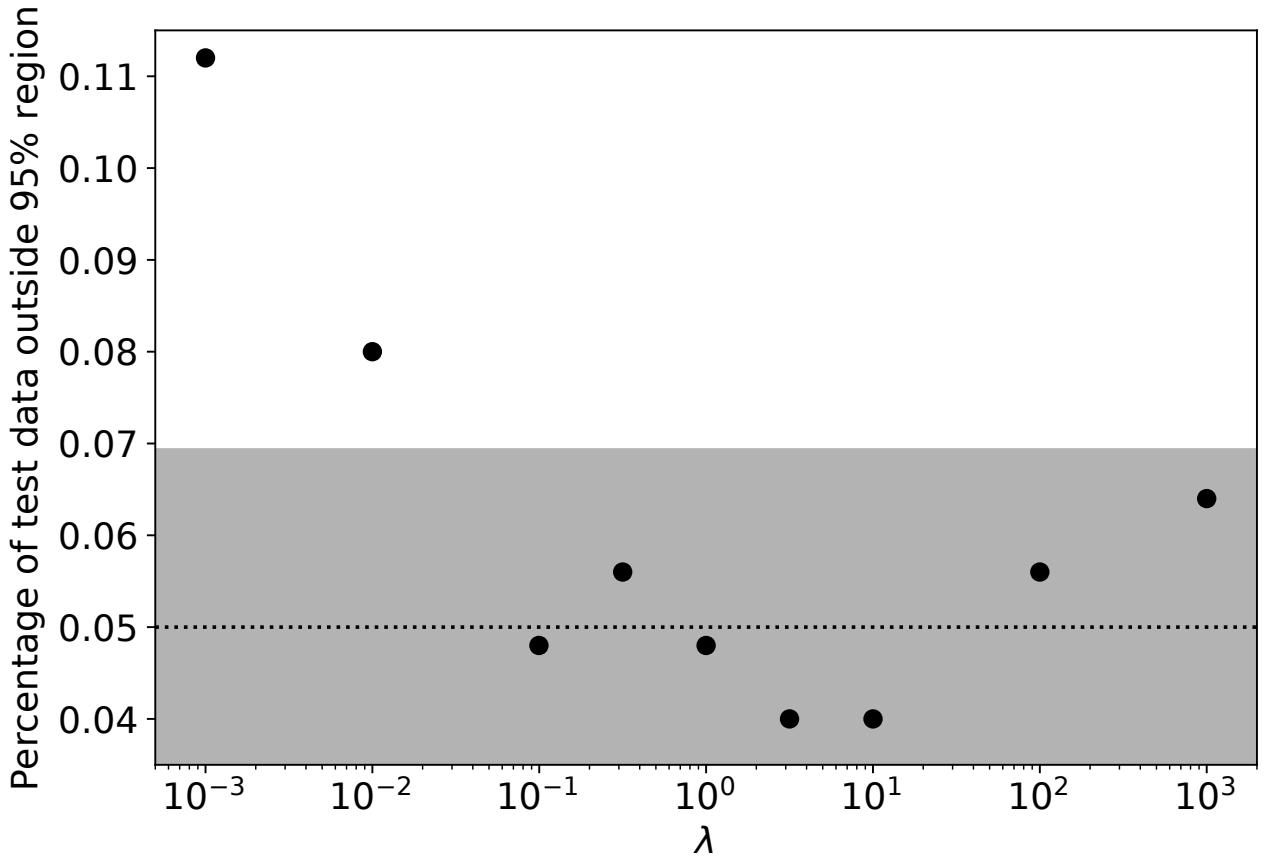

Figure 3. Percentage of test data that fall outside 95% confidence region for different λ values. A perfectly calibrated posterior has 5% outliers. The shaded region shows the uncertainty due to finite number of test data. This measurement is made on weak lensing maps with 64^2 resolution and $n_g = 30\text{arcmin}^{-2}$ galaxy density.

Table 3. Empirical coverage probability of posteriors after marginalizing over baryon parameters. We report the percentage of test data that falls within 68% confidence region and 95% confidence regions. A perfectly calibrated posterior should have 68% and 95% test data that falls in these two regions, respectively.

Method	$n_g = 10\text{arcmin}^{-2}$	$n_g = 20\text{arcmin}^{-2}$	$n_g = 50\text{arcmin}^{-2}$	$n_g = 100\text{arcmin}^{-2}$
Multiscale Flow $p(x_{512}\|y)$	$72.8\%,\ 96.8\%$	$74.4\%,\ 95.2\%$	$73.6\%,\ 97.6\%$	$66.4\%,\ 97.6\%$
Multiscale Flow $p(x_{256}\|y)$	$70.4\%,\ 96.0\%$	$76.8\%,\ 95.2\%$	$74.4\%,\ 97.6\%$	$68.8\%,\ 96.8\%$
Multiscale Flow $p(x_{128}\|y)$	$76.0\%,\ 96.0\%$	$74.4\%,\ 97.6\%$	$76.0\%,\ 97.6\%$	$73.6\%,\ 96.8\%$
Multiscale Flow $p(x_{64}\|y)$	$80.8\%,\ 95.2\%$	$70.4\%,\ 94.4\%$	$72.0\%,\ 95.2\%$	$74.4\%,\ 95.2\%$

D. Multiscale Flow Posteriors

The posterior distribution of different scales on 20 test maps with galaxy number density $n_g = 30\text{arcmin}^{-2}$ is shown in Figure 4. The posterior constraints of all scales are consistent with the true cosmological parameters, which are shown with black stars.

The posterior distributions of Multiscale Flow and power spectrum on baryon maps with $n_g = 20\text{arcmin}^{-2}$ are shown in Figure 5. Unfortunately, due to the small area of the lensing map, all these methods cannot constrain baryon parameters very well (see also Figure 5 of (Lu et al., 2022) for CNN constraints), and the posterior is dominated by the prior bounds, especially in the cases of high shape noise.

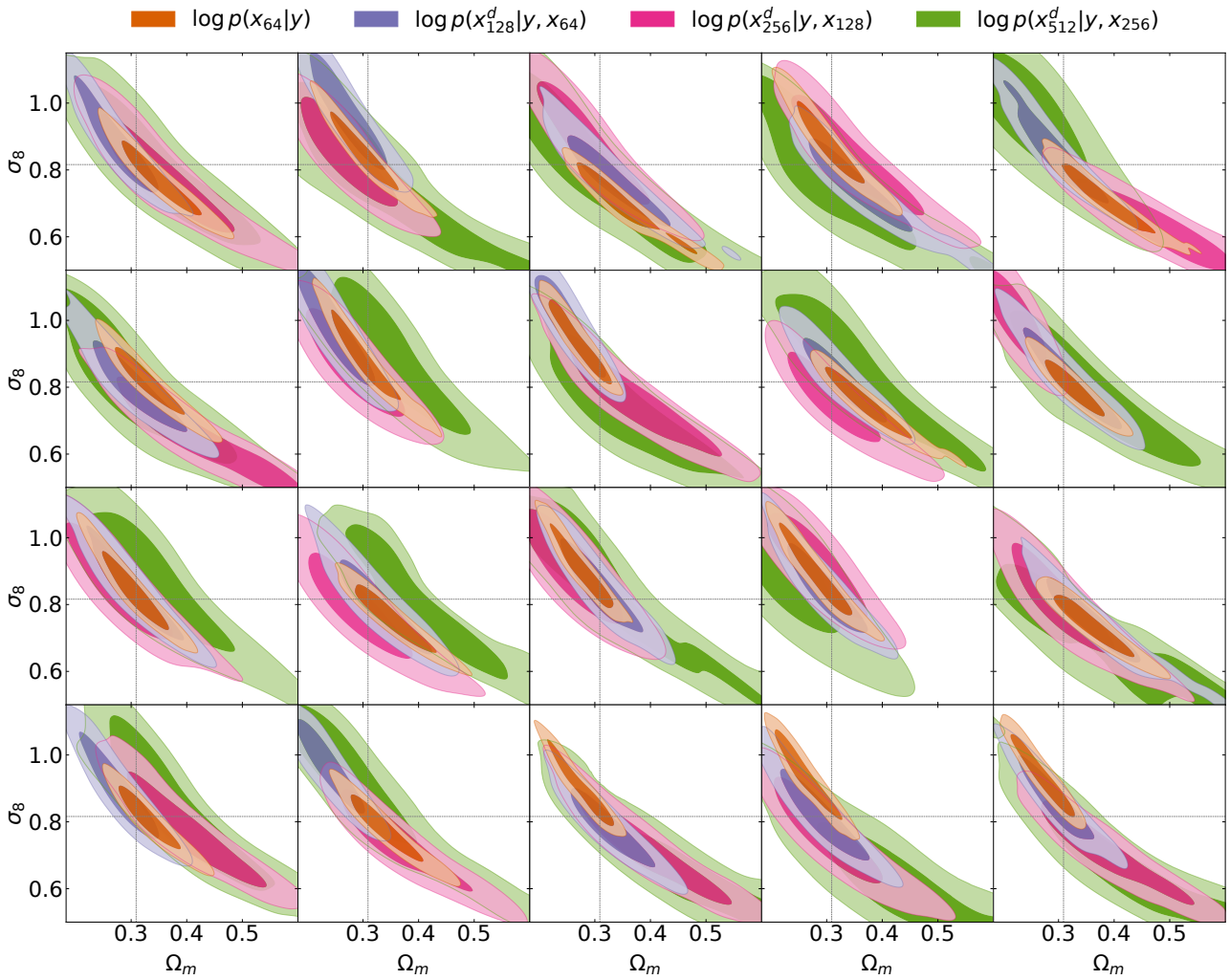

Figure 4. Multiscale Flow posterior comparison of different scales on 20 test data with galaxy number density $n_g = 30\text{arcmin}^{-2}$.

E. Sample generation and super-resolution

We show an example of sample generation with Multiscale Flow in Figure 6. The process can also be viewed as iterative super-resolution of the low-resolution samples. In Figure 7 we show that Multiscale Flow samples and test data agree well in terms of power spectrum and pixel probability distribution function. This demonstrates that Multiscale Flow samples can be used in lieu of expensive N-body simulations and ray tracing as a fast generator of mock data.

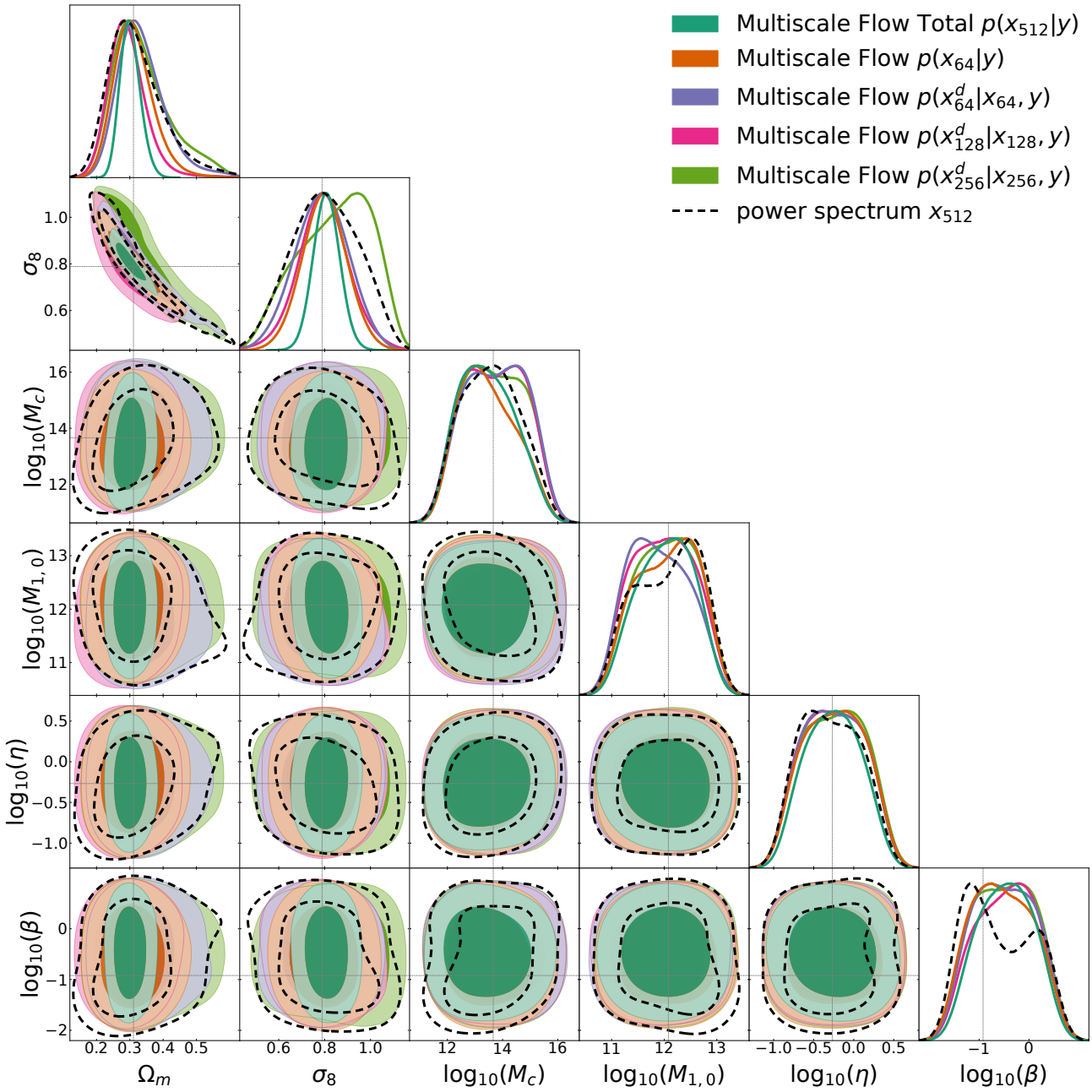

Figure 5. Comparison of posterior distributions between different scales of Multiscale Flow and power spectrum on a $3.5 \times 3.5\text{deg}^2$ convergence map with $n_g = 20\text{arcmin}^{-2}$.

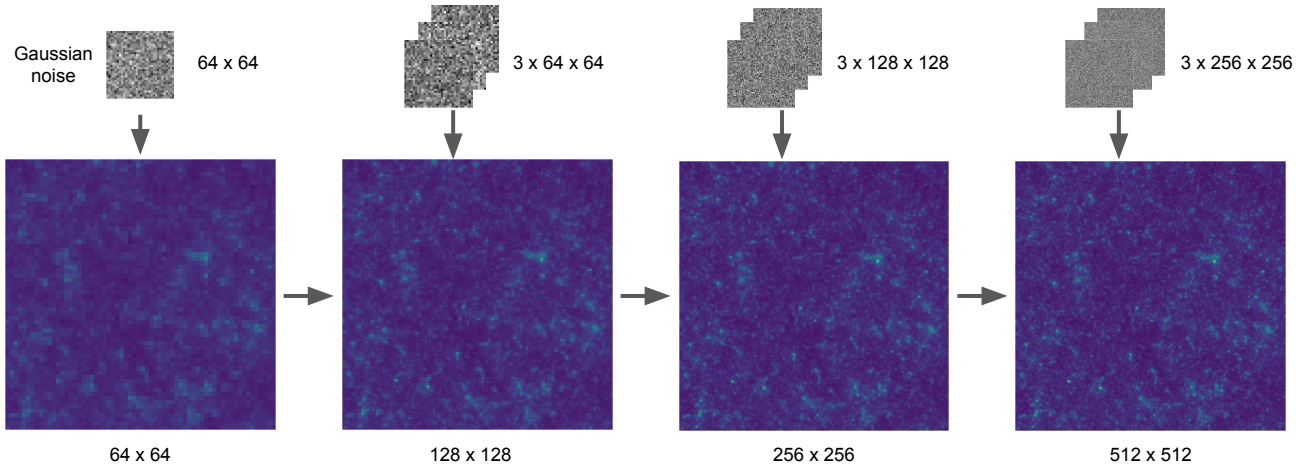

Figure 6. Illustration of Multiscale Flow sample generation (the reverse of Figure 1). The sample of the lowest resolution is firstly generated, and then small-scale information are gradually added. This process can also be viewed as super-resolution.

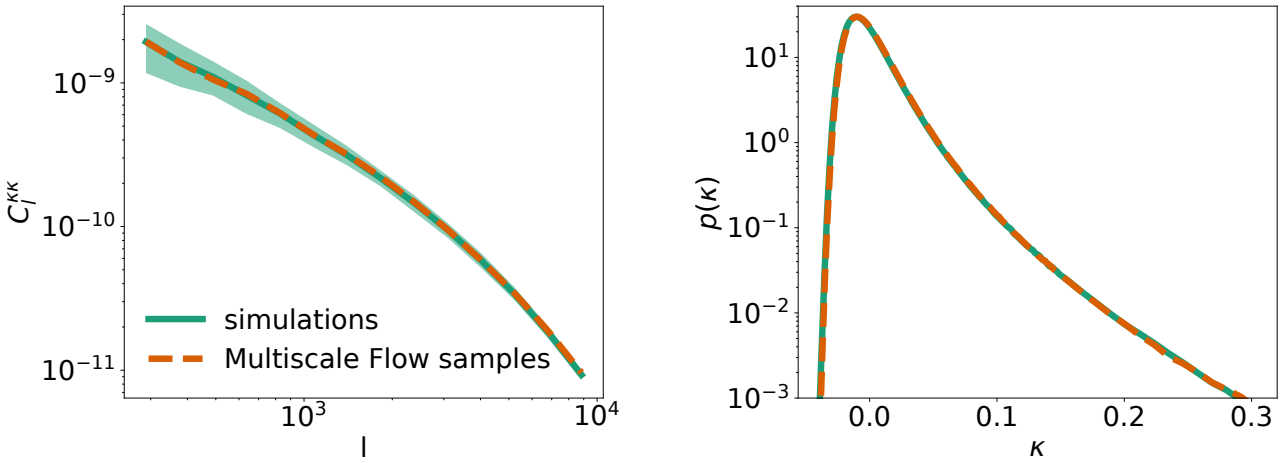

Figure 7. Comparison of power spectrum (left) and pixel probability distribution function (right) between simulations and Multiscale Flow samples at fiducial cosmology.