Hybrid Summary Statistics

T. Lucas Makinen[1*] **Ce Sui**[2*] **Benjamin D. Wandelt**[3,4]
Natalia Porqueres[5] **Alan Heavens**[1]
[1]Imperial College London [2]Tsinghua University [3]Sorbonne Université
[4]Center for Computational Astrophysics, Flatiron Institute
[5]Oxford University
l.makinen21@imperial.ac.uk
suic20@mails.tsinghua.edu.cn

Abstract

We present a way to capture high-information posteriors from training sets that are sparsely sampled over the parameter space for robust simulation-based inference. In physical inference problems, we can often apply domain knowledge to define traditional summary statistics to capture some of the information in a dataset. We show that augmenting these statistics with neural network outputs to maximise the mutual information improves information extraction compared to neural summaries alone or their concatenation to existing summaries and makes inference robust in settings with low training data. We introduce 1) two loss formalisms to achieve this and 2) apply the technique to two different cosmological datasets to extract non-Gaussian parameter information.

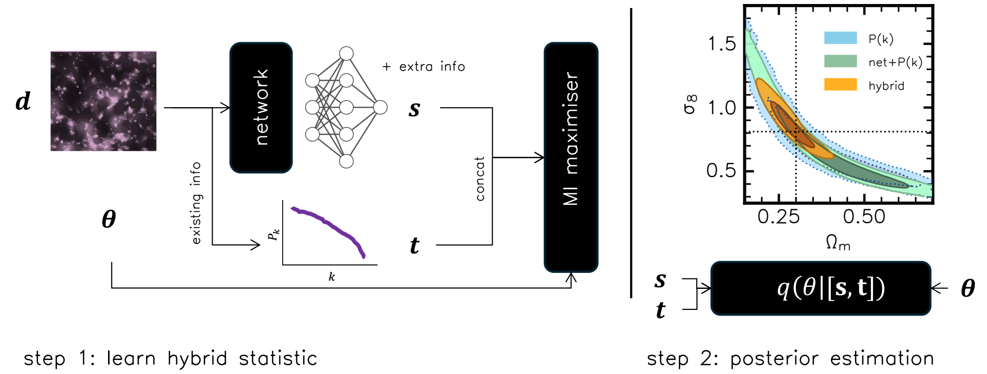

Figure 1: Schematic for learning a hybrid summary statistic **s** with a choice of mutual information maximiser and existing static **t**. Black boxes denote neural functions to be learned.

1 Introduction

Implicit (simulation-based) inference makes solving otherwise intractable inverse problems possible by employing neural compressors which can flexibly map big data vectors into informative summaries (Cranmer et al., 2020; Hoffmann & Onnela, 2023). These mappings can be made optimal (Charnock et al., 2018; Makinen et al., 2021, 2024; Lanzieri et al., 2024; Jeffrey et al., 2020), but often require large numbers of simulations to achieve convergence.

[*] Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

For many physics problems, such as large N-body and hydrodynamical solvers, forward simulations are expensive to generate, so networks tasked to compress these data for parameter inference must learn informative features using sparsely-sampled training sets, especially when large numbers of parameters are varied. Introducing informative priors in data feature (Battaglia et al., 2018; Ivanov et al., 2024, e.g.) or likelihood (Modi & Philcox, 2023) space can simplify the task of a neural network to learn an objective.

Main Contributions. We present a way to obtain highly informative summaries over parameter space in low-training data settings that *boost* information extraction from data beyond existing statistics and the mere concatenation of neural summaries learned separately and traditional summaries. These “hybrid" statistics are neural summaries that are learned to maximise the mutual information (MI) beyond an existing summary of the data and the parameters of interest.

Summary of Results. We present two equivalent objectives that maximise MI information between a new, neural summary, and an existing summarisation of the data over parameter space. We apply this technique to two different cosmological problems where information can be lost to existing summary functions–21cm Epoch of Reionisation (21cm) and weak gravitational lensing (WL) parameter inference. We show that hybridised summaries are far more robust in settings with low available training simulations, indicating improved network optimisation over parameter space.

2 Formalism

In inference problems, we aim to infer quantities of interest θ from data $\mathbf{d}$. In the physical sciences we often have domain knowledge which allows us to design a summary statistic of the data $\mathbf{t}(\mathbf{d})$ that capture some, but not all, of the information in a dataset. To enhance the information extraction, we aim to learn additional summaries $\mathbf{s}(\mathbf{d})$ that are complementary to $\mathbf{t}$. We can formalise the extra information beyond $\mathbf{t}$ captured by $\mathbf{s}$ as the conditional mutual information $I(\mathbf{s};\theta|\mathbf{t})$, which measures how much the uncertainty about θ is reduced by knowing $\mathbf{s}$ given a static $\mathbf{t}$. Evaluating the conditional mutual information can be difficult. However, using the chain rule of MI we can write

$$I(\mathbf{s};\theta|\mathbf{t}) = I([\mathbf{s},\mathbf{t}];\theta) - I(\mathbf{t};\theta) \tag{1}$$

where $I(\mathbf{t};\theta)$ is a constant and $[a,b]$ denotes concatenation. We can then maximise the additional information captured by $\mathbf{s}$ by maximising $I([\mathbf{t},\mathbf{s}];\theta)$. There are various ways to maximize MI; here, we focus on two specific objectives, detailed in Appendix A. The first, referred to as the Posterior Entropy (EPE) objective, is given by:

$$\min_{\mathbf{s},q} \mathcal{L} = -\mathbb{E}_{p(\theta,\mathbf{d})}\Big[\log q\left(\theta\middle|[\mathbf{s}(\mathbf{d}),\mathbf{t}(\mathbf{d})]\right)\Big], \tag{2}$$

where $\mathbf{s}(\mathbf{d})$ is a neural network and q is a neural density estimator. Hoffmann & Onnela (2023) demonstrate that this loss unifies many information-theoretic losses into a stable objective. This mutual information objective can also be parameterised as a cross-entropy classification problem by employing the Jensen-Shannon divergence (Chen et al., 2021; Devon Hjelm et al., 2018; Nowozin et al., 2016):

$$\min_{\mathbf{s},c} \mathcal{L}(\mathbf{s},c) = \mathbb{E}_{p(\theta,\mathbf{d})}\left[\mathrm{sp}(-c(\theta,[\mathbf{s}(\mathbf{d}),\mathbf{t}(\mathbf{d})]))\right] + \mathbb{E}_{p(\theta)p(\mathbf{d})}\left[\mathrm{sp}(c(\theta,[\mathbf{s}(\mathbf{d}),\mathbf{t}(\mathbf{d})]))\right], \tag{3}$$

where s is the summarizer, c is a classifier tasked with distinguishing between data from $p(\theta,x)$ and $p(\theta)p(x)$ and $\mathrm{sp}(z) = \log(1+e^z)$ is the softplus function. We refer to this as the Cross Entropy (CE) loss.

To test the information capture in learned summaries we employ a two-step process, illustrated in Fig. 1. We first optimise a neural compression (embedding network) to a few additional numbers conditional on the specified (not learned) existing summary $\mathbf{t}$, using either EPE or CE losses (denoted MI maximiser in Fig. 1). For the EPE loss we train a simple mixture density network (MDN) with two small hidden layers to approximate $q(\theta|[\mathbf{s},\mathbf{t}])$. For the CE loss we train a classifier fully-connected network with hidden sizes $[128, 64, 64]$ to one output. We then take the static learned and existing summaries and parameterise a separate posterior estimator using a masked autoregressive flow (MAF) to minimise $p(\theta|[\mathbf{s},\mathbf{t}])$ from the `LtU-ILI` package[2] (Ho et al., 2024). We feed comparison summaries (power spectrum and competing network schemes) into the same MAF architecture to obtain a consistent comparison of information capture. We detail exact configurations in Appendix B.

[2] https://github.com/maho3/ltu-ili

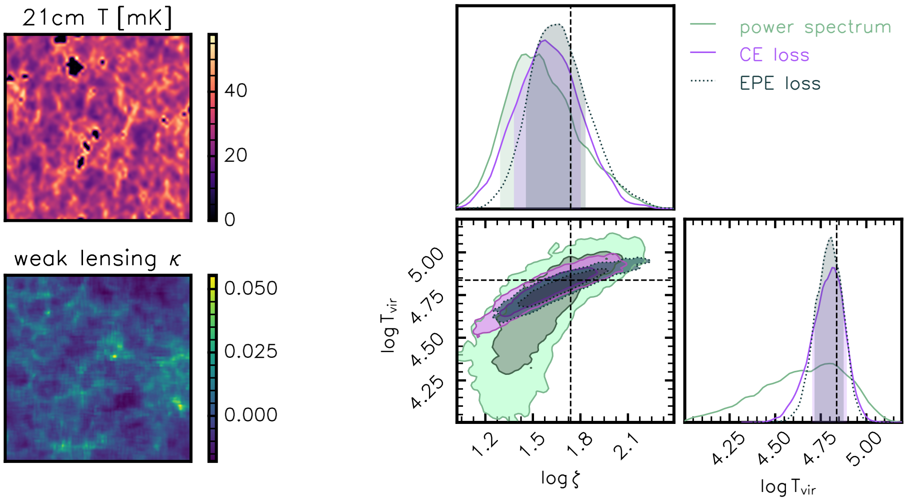

Figure 2: Both 21cm and weak lensing data exhibit non-Gaussian features (upper and lower left panels). Both EPE and CE loss formalisms result in consistent, tighter posteriors than power spectrum alone indicating information extraction from non-Gaussian features in the 21cm data (right). The black dashed line represents the true parameter values.

3 Experiments

21cm Parameter Inference & Loss Comparison. The 21 cm signal is non-Gaussian due to reionization patchiness. Therefore, the power spectrum alone cannot fully capture the information contained in images of the 21 cm signal. While many previous studies have focused on designing new summaries, we show that adding just a few supplementary features to the power spectrum can significantly enhance the extracted information content.

The reionization parameters we vary are (1) ζ, the ionizing efficiency. It primarily determines the timing of the EoR, with higher values leading to earlier ionization of the Universe. We vary ζ as $10 \leq \zeta \leq 250$, and (2) $T_{\rm vir}$, the minimum virial temperature of halos that host ionizing sources. $T_{\rm vir}$ controls the timing of astrophysical epochs and influences the scales of heated and ionized regions. We vary this parameter as $4 \leq \log_{10}(T_{\rm vir}/{\rm K}) \leq 6$. More details can be found in Appendix B.

In this initial experiment, we compare tree types of summaries, assuming a sufficiently large training set (10,000 samples): 1) Power spectrum only (11 k-bins). 2) Hybrid method: Power spectrum + two learned supplementary features (EPE Loss). 3) Hybrid method: Power spectrum + two learned supplementary features (CE Loss). One of the resulting inferences is shown in Figure 2. The results show that both approaches capture non-Gaussian information and improve inference performance, yielding consistent posteriors. The slight difference in posterior recovery is likely due to differences between classifier and MDN network architectures for each loss, but each yielded similar convergence times in training. This suggests we can learn two additional parameters instead of new summaries, maintaining power spectrum interpretability while achieving near-optimal inference. The agreement also confirms both loss functions are effective, converging to the same results.

Tomographic Weak Lensing Inference & Ablation Study. Weak gravitational lensing (WL) alters the trajectories of photons as they pass through massive structures of visible and dark matter. This observable is sensitive to (Ω_m, S_8), parameters that control the universe's matter content and dark matter clustering, respectively. Here we test hybrid summaries on noisy tomographic WL convergence image data of shape $(128, 128, 4)$ presented in Makinen et al. (2024) varied over a wide uniform prior. As an existing summary, we repeat Makinen et al. (2024)'s procedure and histogram all auto- and cross-power spectra for each redshift bin into a vector of 60 numbers for each simulation. The simulations are also subject to additive shape noise, which we add to the noise-free simulations on-the-fly during network training (details in Appendix B).

We perform an ablation study to demonstrate the effectiveness of the hybrid statistics over neural-only methods and display results in Fig. 3. The embedding network in the hybrid setting is the lightweight

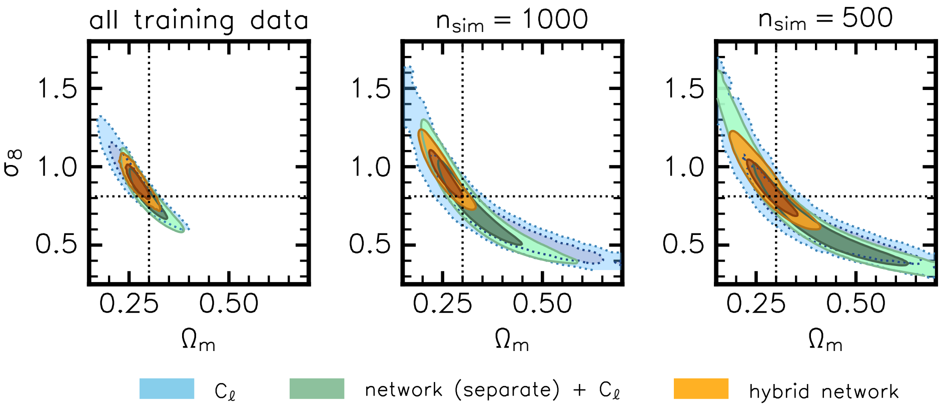

Figure 3: Hybrid summaries are able to capture more non-Gaussian parameter information in settings with both high (left) and low (right) numbers of training simulations than summaries from a larger network trained with the same loss separately from and then concatenated to the C_ℓ statistic for the final inference in step 2. When reducing the total available simulation volume all inferences suffer, but the hybrid statistics are most robust to this change, indicating that the MI objective improves capture of non-power spectrum features.

CNN with symmetric kernels adapted from Makinen et al. (2024) to output 3 additional numbers alongside the C_ℓs. For comparison, we train a larger, more expressive CNN embedding network under the same EPE loss *without* access to the C_ℓ vector to output 3 numbers, and then for the density estimation (step 2) concatenate its outputs to the C_ℓs (green contours; network (separate) + C_ℓ). Step 2 probes the information content captured in the static network and C_ℓ summaries. We train the networks and density estimators from scratch first using all 5000 simulations available (split into 70% train and 30% validation sets). We then reduce the total number of available simulations to 1000 and further to 500 and re-learn the embeddings and posteriors from scratch. When the simulation budget is reduced, all inferences suffer, but the hybrid statistic formalism encourages the smaller network to find non-Gaussian features in the dataset that are more robust to this change. We also note that the more expressive, separately-trained CNN inference degrades almost to the level of the C_ℓ contour in the lowest-data setting, even when concatenated to the C_ℓ vector in the density estimation step.

4 Conclusions & Outlook

We detailed a method to learn compressed summary statistics that explicitly complement an existing summary of the data through mutual information maximisation. We show that these techniques can capture non-Gaussian information in two cosmological applications using two different loss criteria to significantly improve parameter information capture.

We additionally demonstrate that these hybridised summaries improve information capture when the training simulation budget is limited. This suggests that requiring a network to find patterns in the data that are explicitly complementary to a provided summary “tells it where to look" and improves the compression optimisation in smaller datasets over wide parameter space.

We note that MI can also be used as a static metric to quantify the information content in arbitrary summaries as in Sui et al. (2023). This technique could be extended towards exhaustive information studies to measure how much more information might be unlocked with multiple traditional or neural statistics.

5 Data Availability

All code, tutorials, and relevant simulations can be found at `https://github.com/tlmakinen/hybridStats`.

6 Acknowledgements

We thank Justin Alsing, Tom Charnock, Niall Jeffrey, and David Spergel for conversations that inspired this work. This work was supported by the Simons Collaboration on “Learning the Universe”. TLM acknowledges the support of the Imperial College London President’s Scholarship and G-Research for conference travel costs. CS is supported by the National SKA Program of China (grant No. 2020SKA0110401) and NSFC (grant No. 11821303).

References

Barber, D., & Agakov, F. 2004, Advances in neural information processing systems, 16, 201

Battaglia, P. W., Hamrick, J. B., Bapst, V., et al. 2018, Relational inductive biases, deep learning, and graph networks. https://arxiv.org/abs/1806.01261

Charnock, T., Lavaux, G., & Wandelt, B. D. 2018, Physical Review D, 97, doi: 10.1103/physrevd.97.083004

Chen, Y., Zhang, D., Gutmann, M., Courville, A., & Zhu, Z. 2021, Neural Approximate Sufficient Statistics for Implicit Models. https://arxiv.org/abs/2010.10079

Cranmer, K., Brehmer, J., & Louppe, G. 2020, Proceedings of the National Academy of Sciences, 117, 30055, doi: 10.1073/pnas.1912789117

Devon Hjelm, R., Fedorov, A., Lavoie-Marchildon, S., et al. 2018, arXiv e-prints, arXiv:1808.06670, doi: 10.48550/arXiv.1808.06670

Ho, M., Bartlett, D. J., Chartier, N., et al. 2024, LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology. https://arxiv.org/abs/2402.05137

Hoffmann, T., & Onnela, J.-P. 2023, Minimising the Expected Posterior Entropy Yields Optimal Summary Statistics. https://arxiv.org/abs/2206.02340

Ivanov, M. M., Cuesta-Lazaro, C., Mishra-Sharma, S., Obuljen, A., & Toomey, M. W. 2024, Full-shape analysis with simulation-based priors: constraints on single field inflation from BOSS. https://arxiv.org/abs/2402.13310

Jeffrey, N., Alsing, J., & Lanusse, F. 2020, Monthly Notices of the Royal Astronomical Society, 501, 954, doi: 10.1093/mnras/staa3594

Lanzieri, D., Zeghal, J., Makinen, T. L., et al. 2024, Optimal Neural Summarisation for Full-Field Weak Lensing Cosmological Implicit Inference. https://arxiv.org/abs/2407.10877

Makinen, T. L., Charnock, T., Alsing, J., & Wandelt, B. D. 2021, Journal of Cosmology and Astroparticle Physics, 2021, 049, doi: 10.1088/1475-7516/2021/11/049

Makinen, T. L., Heavens, A., Porqueres, N., et al. 2024, Hybrid summary statistics: neural weak lensing inference beyond the power spectrum. https://arxiv.org/abs/2407.18909

Mesinger, A., & Furlanetto, S. 2007, The Astrophysical Journal, 669, 663, doi: 10.1086/521806

Mesinger, A., Furlanetto, S., & Cen, R. 2011, Monthly Notices of the Royal Astronomical Society, 411, 955, doi: 10.1111/j.1365-2966.2010.17731.x

Modi, C., & Philcox, O. H. E. 2023, Hybrid SBI or How I Learned to Stop Worrying and Learn the Likelihood. https://arxiv.org/abs/2309.10270

Nowozin, S., Cseke, B., & Tomioka, R. 2016, arXiv e-prints, arXiv:1606.00709, doi: 10.48550/arXiv.1606.00709

Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., & Tucker, G. 2019, arXiv e-prints, arXiv:1905.06922, doi: 10.48550/arXiv.1905.06922

Sui, C., Zhao, X., Jing, T., & Mao, Y. 2023, arXiv e-prints, arXiv:2307.04994, doi: 10.48550/arXiv.2307.04994

A Mutual information maximization

Conditional mutual information (MI) is defined as

$$I(\mathbf{s};\theta|\mathbf{t}) = \iiint p(\theta,\mathbf{s},\mathbf{t}) \log \frac{p(\theta|\mathbf{s},\mathbf{t})}{p(\theta|\mathbf{t})} d\mathbf{s}d\mathbf{t}d\theta, \tag{4}$$

which measures how much the uncertainty about θ is reduced by knowing $\mathbf{s}$ given $\mathbf{t}$. Using the chain rule of MI we can write

$$I(\mathbf{s};\theta|\mathbf{t}) = I([\mathbf{s},\mathbf{t}];\theta) - I(\mathbf{t};\theta) \tag{5}$$

where $I(\mathbf{t};\theta)$ is a constant and $[a,b]$ denotes concatenation. We can then maximise the additional information captured by $\mathbf{s}$ by maximising $I([\mathbf{t},\mathbf{s}];\theta)$. Mutual information is defined as

$$I(\theta;\mathbf{z}) = D_{\mathrm{KL}}(p(\theta,\mathbf{z})\|p(\theta)p(\mathbf{z})) = \mathbb{E}_{p(\theta,\mathbf{z})}\left[\log\frac{p(\theta|\mathbf{z})}{p(\theta)}\right]. \tag{6}$$

where we package $\mathbf{z} = [\mathbf{s},\mathbf{t}]$. Most of the time, the true posterior $p(\theta|\mathbf{z})$ is unknown. To address this, we approximate the posterior with a neural density estimator $q(\theta|\mathbf{z})$, which provides a tractable lower bound on the MI (Barber & Agakov, 2004; Poole et al., 2019):

$$I(\theta;\mathbf{z}) = \mathbb{E}_{p(\theta,\mathbf{z})}\left[\log\frac{q(\theta|\mathbf{z})}{p(\theta)}\right] + \mathbb{E}_{p(\mathbf{z})}\left[D_{\mathrm{KL}}(p(\theta|\mathbf{z})\|q(\theta|\mathbf{z}))\right] \geq \mathbb{E}_{p(\theta,\mathbf{z})}[\log q(\theta|\mathbf{z})] + h(\theta). \tag{7}$$

This leads to a calculable loss function for maximizing mutual information, with $h(\theta)$ fixed, leading to the objective in Equation 2.

Additionally, since the exact value of mutual information is not required, we can use alternative divergences that may offer better robustness and efficiency. By using the Jensen-Shannon divergence and the variational representation from Nowozin et al. (2016), we can derive a lower bound that leads to a cross-entropy loss (Equation 3), frequently used in representation learning (Chen et al., 2021; Devon Hjelm et al., 2018).

B Experimental Details

B.1 21cm Simulations

The 21 cm signals are simulated using the publicly available code `21cmFAST`[3] (Mesinger & Furlanetto, 2007; Mesinger et al., 2011). The simulations were performed on a cubic box of 128 comoving Mpc on each side, with 64^3 grid cells. For this work, we use coeval boxes at redshift 12, and extract a single slice from each cube to form a 2D dataset.

Network details. To obtain hybrid summaries, we use a CNN with three convolutional layers (32, 64, and 128 filters) followed by max pooling. The flattened output is processed through two fully connected layers, with ReLU activations throughout. For classification, we use an FCN with hidden layers of sizes 128, 64, and 64, each followed by ReLU activation. For the EPE loss, a mixture density network (MDN) with a 64-unit hidden layer and output layers for mixing coefficients, standard deviations, and means is employed. Mixing coefficients use softmax, standard deviations are exponentiated, and means are directly outputted, with five components used.

Posterior Coverage. In Figure 2, we present the posterior for a single test sample. To demonstrate that these results are not due to overfitting, we also show calibration results for a test set of 2,048 samples. Posterior coverage is used as a validation metric, as shown in Figure 4. The predicted percentiles closely match the empirical percentiles, indicating that our summaries are robust and the SBI inference is neither overly confident nor conservative in any case.

B.2 Weak Lensing Simulations

Here we test hybrid summaries on noisy tomographic WL simulations presented in Makinen et al. (2024). The simulations were generated using the `pmwd` particle mesh code and then collected into four redshift bins to form convergence image data of shape $(128, 128, 4)$. The simulations are

[3] https://github.com/andreimesinger/21cmFAST

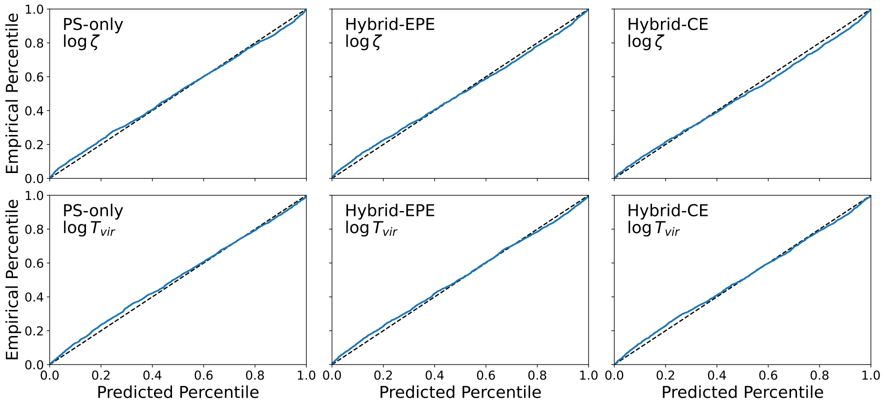

Figure 4: Posterior coverage for the 21 cm example across three summaries and two reionization parameters. The blue lines represent the actual calibration using 2,048 test samples, while the black dashed line indicates perfect calibration.

generated over a wide uniform prior in the parameters $p(\Omega_m, S_8) = \mathcal{U}([0.15, 0.7] \times [0.35, 1.52])$. As an existing summary, we repeat Makinen et al. (2024)'s procedure and histogram all auto- and cross-power spectra for each redshift bin into a vector of 60 numbers for each simulation. The simulations are also subject to additive shape noise, which we add to the noise-free simulations on-the-fly during network training.

Network details. For obtaining hybrid summaries the convolutional neural network with symmetric Multipole Kernels (MPK) was adapted from Makinen et al. (2024). The lightweight network is initialised without pretraining for this analysis and contains $1,615$ learnable parameters. For the large non-hybrid CNN we apply a 3×3 kernel to embed the field into 16 filters, and then down-sample with stride-2 convolutions with output filters $[32, 64, 128]$. The network is then mean-pooled in the spatial axes and the flattened filters are passed to a dense network with a specified output size to be fed into the mutual information maximiser (here a mixture density network for the EPE loss). This results in $97,971$ learnable parameters. For both embedding networks we employ the `smooth_leaky` activation function from Makinen et al. (2024). For the EPE loss configuration, a mixture density network (MDN) with hidden layers of size $[70, 70]$ and output layers for mixing coefficients, standard deviations, and means is employed to parameterise a four-component mixture.