Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Received: *July 29, 2024*
Revised: *October 18, 2024*
Accepted: *December 4, 2024*
Published: *January 21, 2025*

Hybrid summary statistics: neural weak lensing inference beyond the power spectrum

T. Lucas Makinen,[a] **Alan Heavens**,[a] **Natalia Porqueres**,[b] **Tom Charnock**,[f] **Axel Lapel**[c,d] **and Benjamin D. Wandelt**[c,e]

[a]*Imperial Centre for Inference and Cosmology (ICIC) & Imperial Astrophysics,*
Imperial College London, Blackett Laboratory,
Prince Consort Road, London SW7 2AZ, U.K.

[b]*Department of Physics, University of Oxford, Denys Wilkinson Building,*
Keble Road, Oxford OX1 3RH, U.K.

[c]*Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris,*
98 bis boulevard Arago, 75014 Paris, France

[d]*Sorbonne Université, Université Paris Diderot, Sorbonne Paris Cité,*
CNRS, Laboratoire de Physique Nucléaire et de Hautes Energies (LPNHE),
4 place Jussieu, F-75252, Paris Cedex 5, France

[e]*Center for Computational Astrophysics, Flatiron Institute,*
162 5th Avenue, New York, NY 10010, U.S.A.

[f]*Freelance consultant in statistical modelling*

E-mail: l.makinen21@imperial.ac.uk, a.heavens@imperial.ac.uk, natalia.porqueres@physics.ox.ac.uk, tom@charnock.fr, axel.lapel@iap.fr, bwandelt@iap.fr

Abstract: Cosmological inference relies on compressed forms of the raw data for analysis, with traditional methods exploiting physics knowledge to define summary statistics, such as power spectra, that are known to capture much of the information. An alternative approach is to ask a neural network to find a set of informative summary statistics from data, which can then be analysed either by likelihood- or simulation-based inference. This has the advantage that for non-Gaussian fields, they may capture more information than two-point statistics. However, a disadvantage is that the network almost certainly relearns that two-point statistics are informative. In this paper, we introduce a new hybrid method, which combines the best of both: we use our domain knowledge to define informative physics-based summary statistics, and explicitly ask the network to augment the set with extra statistics that capture information that is not already in the existing summaries. This yields a new, general loss formalism that reduces both the number of simulations and network size needed to extract useful non-Gaussian information from cosmological fields, and guarantees that the resulting summary statistics are at least as informative as the power spectrum. In combination, they can then act as powerful inputs to implicit inference of model parameters. We use a generalisation of

https://doi.org/10.1088/1475-7516/2025/01/095

Information Maximising Neural Networks (IMNNs) to obtain the extra summaries, and obtain parameter constraints from simulated tomographic weak gravitational lensing convergence maps. We study several dark matter simulation resolutions in low- and high-noise regimes. We show that i) the information-update formalism extracts at least 3× and up to 8× as much information as the angular power spectrum in all noise regimes, ii) the network summaries are highly complementary to existing 2-point summaries, and iii) our formalism allows for networks with extremely lightweight architectures to match much larger regression networks with far fewer simulations needed to obtain asymptotically optimal inference.

Keywords: cosmological parameters from LSS, Machine learning , weak gravitational lensing, cosmic web

ArXiv ePrint: 2407.18909

Contents

1 Introduction

Weak gravitational lensing alters the trajectories of distant photons as they pass through the large-scale structure of visible and dark matter to our detectors. These deflections alter the observed shapes of galaxies, whose patterns can be used to trace the matter distribution in between, and are sensitive to parameters that describe the expansion history and structure formation of the Universe. The inference of these parameters from cosmological weak lensing surveys is usually performed using two-point statistics of the lensing images, such as shear correlation functions or power spectra. Recent analyses include the Dark Energy Survey [1, 2], the Kilo-Degree Survey [KiDS, 3, 4] and the Hyper Suprime-Cam survey [5]. However, two-point statistics do not fully describe the rich non-Gaussian features present in large-scale structure, where more cosmological information might be found.

Implicit inference (also known as simulation-based inference or likelihood-free inference) has made it possible to utilise higher-order statistics derived from simulations (see e.g. [6] for a review), and circumvent the need for an explicit likelihood function, which can be challenging to compute via Bayesian Hierarchical Models [7–12]. In weak gravitational lensing for example, even (incorrectly) assuming that the underlying cosmological density is Gaussian, the two-point statistics that describe this field can themselves have significantly non-Gaussian sampling distributions [7, 13, 14]. The question arises as to which additional statistics contain significant extra information about the parameters, beyond that which is present in the two-point functions. Higher-order statistics are an obvious choice, but they suffer from a lack of knowledge of their sampling distributions, and the very large number of statistics make them cumbersome for implicit inference. However, advances in deep learning have made it possible to learn highly-informative neural compressions for massive simulations automatically, and in some cases losslessly [15–17]. These compressions yield radically smaller summary spaces, which are ideal for implicit inference, and which can be used for Bayesian posterior estimation via accept-reject or density estimation strategies [18].

These new advances have made simulation-based studies for weak lensing a popular avenue of research in recent years. This includes studies of peak counts [19–22], higher-order statistics [23] such as wavelet scattering transformations [24, 25], Fourier-space normalising flows [26], and field-based convolutional neural networks [27–30]. The ultimate goal is to obtain statistics that exhaust the information content of the observed weak lensing field. This can be done explicitly (assuming a likelihood for shear or convergence voxels) via field-level sampling [7–11, 31–35].

Impicit inference approaches have matured enough for real-data analysis, beginning with [36], who analysed map-level KiDS data with an assumed Gaussian summary likelihood, and more recently [14], who reproduced C_ℓ constraints with an implicit likelihood. Ref. [37] presented a Dark Energy Survey data analysis using a convolutional neural network compression of the full mass map and demonstrated marked improvement over existing power spectrum and peak count constraints on the same dataset.

This work seeks to demonstrate an improved optimisation strategy and add a new layer of interpretability to this growing body of literature. A common question for deep learning and implicit inference practitioners is what features are being learned from the data by neural approaches. Ref. [17] showed explicitly that neural networks trained on halo catalogues identified features that could be linked to explicitly-understood cosmological distributions such as halo mass and correlation functions. Here we respond to this question by modifying our optimisation criterion such that a network only outputs statistics obtained from the data that work alongside to an existing statistic, in this case the weak lensing angular power spectrum. To be explicit, we train the network to maximise the extra Fisher information that is not already present in the power spectrum. We term these neural summaries “hybrid” statistics since they combine new and existing functions of the data. We make our constraint comparison within a completely simulation-based setup to interrogate the information content of the respective summaries in a sampling distribution-agnostic way.

This paper is organised as follows: in sections 2 and 3 we present our general formalism for finding optimal new summaries from simulated data given some existing descriptive statistics, and describe how the strategy can be implemented automatically with Information Maximising Neural Networks [IMNNs, 15, 16]. In section 4, we describe our mock weak lensing formalism and present the simulation suite details. In section 5, we present our angular C_ℓ compression scheme as the existing statistic in the information-update formalism. We then present our physics-inspired, lightweight neural network architecture designed to find optimal additional summaries. In section 6.1, we make comparisons of information gain over the power spectrum as a function of resolution and increased shape noise, and show that our optimisation scheme captures physical features in realistic noise regimes.

2 Implicit inference

The goal of most science experiments is to obtain data $\mathbf{d}$ with which to test models that describe the data generation process. In cosmology, this often boils down to obtaining a posterior distribution for some model parameters $\boldsymbol{\theta}$: $p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, which requires knowledge of the likelihood $p(\mathbf{d}|\boldsymbol{\theta})$, and the assumption of a suitable prior $p(\boldsymbol{\theta})$. This data-generating distribution is often too complicated to evaluate for inference, or too complex to write down analytically.

2.1 Density estimation

Simulation-based inference circumvents the need for a tractable likelihood $p(\mathbf{d}|\boldsymbol{\theta})$, and instead seeks to parameterise the underlying, implicit likelihood or posterior present in forward simulations of the data. Neural density estimators [NDEs; e.g. 38] use neural networks that give some estimate $q(\boldsymbol{\theta}, \mathbf{d}; \varphi)$ of the desired posterior (or likelihood) by varying weights and biases (parameterised as φ) to minimize the loss

$$U(\varphi) = -\sum_{i=1}^{N} \ln q(\boldsymbol{\theta}_i|\mathbf{d}_i; \varphi), \tag{2.1}$$

over batches of parameter-data samples drawn from the joint distribution $(\boldsymbol{\theta}_i, \mathbf{d}_i) \curvearrowleft p(\boldsymbol{\theta}, \mathbf{d}_i)$. This loss is equivalent to minimising the Kullback-Leibler divergence between the target distribution and q [39]. In this work we employ Masked Autoregressive Flows [MAF; 40] to model the posterior distribution directly. We detail our implementation in section 5.3. Density estimation also makes posterior predictive and coverage tests far easier to perform. We show examples of these tests in appendix A.

2.2 Data compression

Accept-reject and density estimation schemes become more difficult to compare to a target, observed data vector $\mathbf{d}_{\text{obs}}$ the larger the dimensionality $\dim(\mathbf{d})$, which posits the need for data compression to some smaller summary space $\mathbf{x}$. We would like a function $f : \mathbf{d} \mapsto \mathbf{x}$ that is ideally maximally informative about the parameters $\boldsymbol{\theta}$. Under certain conditions, $f(\mathbf{d})$ can yield a sufficient statistic, for which the dimension $\mathbf{x}$ is equal to the number of parameters, e.g. $\dim(\mathbf{x}) = \dim(\boldsymbol{\theta})$. Ref. [41] introduced Massively Optimised Parameter

Estimation and Data (MOPED) compression, which gives optimal score compression for cases where the likelihood and sampling distributions are Gaussian, and this was generalised to other forms of score compression by [42–44]. Neural compression is a popular scheme for learning mappings agnostic to sampling distributions, for which several optimisation schemes have been proposed. Regression-style approaches learn a compression $f(\mathbf{d}; w)$ parameterised by (neural) weights w via a loss, for example quadratic, over parameter-data pairs over a prior, using variants of the mean square error (MSE), or in some cases learning f and the neural posterior (via eq. (2.1)) simultaneously, dubbed Variational Mutual Information Maximisation [VMIM; 45]. Ref. [30] recently compared these losses paired with convolutional networks for a separate weak lensing simulation suite.

This work builds upon the Information Maximising Neural Network (IMNN) approach [15, 16], which prescribes a neural compression that maximises the determinant of the Fisher matrix of summaries around a *local* point in parameter space. A compression is i) learned at a fiducial point from a set of dedicated fiducial and derivative simulations and then ii) applied to simulations over a prior for posterior construction. This approach has numerous advantages, namely:

1. An asymptotically-optimal compression can be learned from simulations around a single point in parameter space.

2. The compression automatically and simultaneously gives Fisher posterior constraint forecasts.

3. Priors used for density estimation are decoupled from the compression step and can be chosen after learning the compression.

4. Adding additional parameters of interest to the compression learning scheme only requires relatively small numbers of extra simulations for the new derivatives; e.g. the distribution $p(\boldsymbol{\theta}, \mathbf{d})$ need not be re-simulated.

In the following section we will extend this approach to find new (neural) data compressions that only increase information about parameters above what is already present in a set of existing statistics such as the power spectrum.

3 How to choose an optimal new summary

Consider some data $\mathbf{d} \in \mathbb{R}^N$ created from parameters $\boldsymbol{\theta}$ that can be summarised in a compressed summary vector via a function $h : \mathbf{d} \mapsto \mathbf{t}$ with $\mathbf{t} \in \mathbb{R}^{n_t}$ where $n_t < N$. We can estimate the covariance matrix of the summaries $\mathbf{C}_t$, and the mean $\boldsymbol{\mu}_t$ from simulations, along with derivatives with respect to parameters of the mean $\boldsymbol{\mu}_{,\theta_i}$. Assuming for now that the summaries have a Gaussian sampling distribution (this assumption is temporary and is dropped in the inference phase), we can compute the Fisher information of the observables via

$$[\mathbf{F}_t]_{ij} = \boldsymbol{\mu}_{,\theta_i}^T \mathbf{C}_t^{-1} \boldsymbol{\mu}_{,\theta_j} \tag{3.1}$$

where we introduce the notation $\boldsymbol{y}_{,\theta_i} \equiv \partial \boldsymbol{y} / \partial \theta_i$ for partial derivatives with respect to parameters. The Fisher information matrix here describes how much information $h(\mathbf{d})$

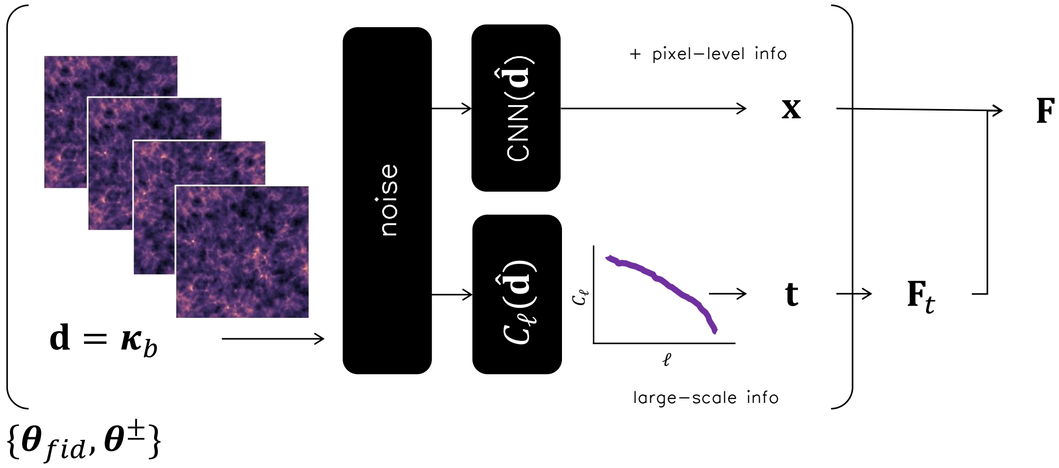
+ pixel−level info
$\mathrm{CNN}(\hat{\mathbf{d}})$
noise
$\mathbf{x}$
$\mathbf{F}$
$C_\ell(\hat{\mathbf{d}})$
C_ℓ
ℓ
$\mathbf{t}$
$\mathbf{F}_t$
$\mathbf{d} = \boldsymbol{\kappa}_b$
large−scale info
$\{\boldsymbol{\theta}_{fid}, \boldsymbol{\theta}^{\pm}\}$

Figure 1. Hybrid summary network schematic, illustrated for weak gravitational lensing. Noisy data (weak lensing $\boldsymbol{\kappa}_b$ with shape noise) are passed in parallel to an existing summary function (tomographic C_ℓ with optional MOPED compression) to produce summaries $\mathbf{t}$, and a network (CNN) to output an additional set of summaries $\mathbf{x}$, described in section 5 and illustrated in figure 2. To train the network the Fisher information is first calculated for $\mathbf{t}$ and then updated via equation (3.10) to yield $\mathbf{F}$, for the loss eq. (3.13).

contains about the model parameters, and is given as the second moment of the score of the likelihood with respect to h, assuming a parameter-independent, Gaussian covariance of the statistic $\mathbf{t}$.[1] A large Fisher information for a function of the data indicates that the mapping to $\mathbf{t} = h(\mathbf{d})$ is very informative about the model parameters used to generate the realisation of data $\mathbf{d}$. Fisher forecasting for a given model is made possible by the information inequality and the Cramér-Rao bound [46, 47], which states that the minimum variance of the value of an estimator $\boldsymbol{\theta}$ is given by

$$\langle(\theta_i - \langle\theta_i\rangle)^2\rangle \geq (\mathbf{F}^{-1})_{ii}, \tag{3.2}$$

with no summation over i.

We would next like to add new summary statistics to increase the information over what is present in $\mathbf{t}$. For clarity, we begin by adding a single summary x, before generalising to multiple additional summaries. We do this via some function $f : \mathbf{d} \mapsto x$. This new number has variance σ_x^2 and when concatenated to the old observables gives the mean vector

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_t, \langle x\rangle]^T. \tag{3.3}$$

For mean-subtracted quantities Δt and Δx, the covariance vectors between old and new observables can be computed (e.g. from simulations) as $[\mathbf{u}]_i = \langle\Delta t_i \Delta x\rangle$, which yields the

[1] Note that the Gaussian assumption is used here only to define a compression; once the summaries are defined, SBI no longer assumes Gaussianity. If the compressed summaries are not Gaussian-distributed, the compression will be suboptimal but the downstream SBI analysis will implicitly determine and use their true sampling distribution.

full covariance matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_t & \mathbf{u} \\ \mathbf{u}^T & \sigma_x^2 \end{pmatrix}. \tag{3.4}$$

Notice here that the smaller the values of $\mathbf{u}$, the less correlated x is with $\mathbf{t}$. The full updated Fisher matrix is then

$$F_{ij} = \boldsymbol{\mu}_{,\theta_i}^T \mathbf{C}^{-1} \boldsymbol{\mu}_{,\theta_j}. \tag{3.5}$$

With some rearrangement, we obtain the fast Information-Update Formula (IUF):

$$\mathbf{F} = \mathbf{F}_t + \frac{1}{s}\mathbf{v}\mathbf{v}^T, \tag{3.6}$$

with $[\mathbf{v}]_i = \langle x \rangle_{,\theta_i} - \boldsymbol{\mu}_{,\theta_i}^T \mathbf{C}_t^{-1}\mathbf{u}$ and $s = \sigma_x^2 - \mathbf{u}^T\mathbf{C}_t^{-1}\mathbf{u}$. This calculation is only $\mathcal{O}(n_{\text{params}}^2 + n_{\text{params}} d + d^2)$ operations where $d = \dim(\mathbf{t})$, which is asymptotically d times faster than eq. (3.5) when $d \gg n_{\text{params}}$. This formalism also yields a fast update for the determinant of the new Fisher matrix:

$$\ln\det\mathbf{F} = \ln\det\mathbf{F}_t + \ln\left(1 + \frac{1}{s}\mathbf{v}^T\mathbf{F}_t^{-1}\mathbf{v}\right). \tag{3.7}$$

Interpretation. The updated Fisher information in equation (3.6) clearly separates the information contribution from the existing observables in the first Fisher term and the new observables in the second term. An optimal, "complementary" new observable x adds a lot of information if it has highly correlated measurement error with the existing summaries $\mathbf{t}$, but changes with respect to parameters in a way that is as distinguishable as possible from how x and $\mathbf{t}$ change together.

Multiple new observables. The IUF can be naturally extended to a vector of new summaries $\mathbf{x}$. We promote $\mathbf{v}$ to a matrix

$$[\boldsymbol{V}]_{ij} = \langle \mathbf{x}_j \rangle_{,\theta_i} - \boldsymbol{\mu}_{,\theta_i}^T \mathbf{C}_t^{-1}\mathbf{u}_j, \tag{3.8}$$

and the scalar s generalises to the matrix

$$\boldsymbol{\Sigma} = \mathbf{C}_x - \mathbf{U}^T\mathbf{C}_t^{-1}\mathbf{U} \tag{3.9}$$

where $[\mathbf{U}]_{ij} = [\mathbf{u}_j]_i$ and $\mathbf{C}_x$ is the covariance of network outputs $\mathbf{x}$. Altogether the updated Fisher matrix for a vector of extended summaries is

$$\mathbf{F} = \mathbf{F}_t + \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^T. \tag{3.10}$$

3.1 Finding a new summary with a neural network

We can find a new observable $\mathbf{x}$ by optimising the IUF equation ((3.6)) with a neural network $f : \mathbf{d} \mapsto \mathbf{x}$ that operates on the data. This formalism folds neatly into the IMNN formalism [15–17]. We illustrate this procedure for weak lensing data in a schematic in figure 1. We can choose to optimise equation (3.5) directly, but equation (3.10) is less computationally expensive for large covariance matrices and more than one additional summary. The ingredients needed to compute the components of the loss are a suite of n_s simulations at a fiducial value of parameters $\{\mathbf{d}\}_{i=1}^{n_s}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{fid}}}$ and a set of n_d seed-matched simulations at perturbed values of each parameter $\boldsymbol{\theta}_i^{\pm} = \boldsymbol{\theta}_i \pm \Delta\boldsymbol{\theta}_i$ holding all other parameters fixed at their fiducial values. Using this finite difference gradient dataset the partial derivatives of a data summary function $Q(\mathbf{d})$ with respect to parameters is

$$\left(\frac{\partial \hat{\mu}_i}{\partial \theta_\alpha}\right) \approx \frac{1}{n_d} \sum_{i=1}^{n_d} \frac{Q(\mathbf{d}_i^+) - Q(\mathbf{d}_i^-)}{\theta_\alpha^+ - \theta_\alpha^-}. \tag{3.11}$$

For n_{params} summaries, this method requires $n_d \times n_{\text{params}} \times 2$ simulations with n_d unique random seeds alongside the n_s simulations at the fiducial point required for the covariance. This is done for the mean of both existing and new summary statistics, consolidated as $\mathbf{y} = [\mathbf{t}, \mathbf{x}]$. The covariance of the (existing and new) summaries is estimated from the data as well, using n_s simulations at $\boldsymbol{\theta}_{\text{fid}}$:

$$\hat{\mathbf{C}}_{\alpha\beta} = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (\mathbf{y}_i - \bar{\mathbf{y}})_\alpha (\mathbf{y}_i - \bar{\mathbf{y}})_\beta, \tag{3.12}$$

where $\bar{\mathbf{y}}$ is the average over the simulations at the fiducial point. The full covariance can be broken down into (or estimated separately by) its components $\mathbf{C}_t, \mathbf{C}_x$, and $\mathbf{u}$ according to eq. (3.4). Note that this covariance is assumed to be independent of the parameters, which, whilst not strictly true, is enforced by regularisation during the fitting of a network. If it does not hold, it simply makes the compression suboptimal elsewhere in the parameter space. Crucially, both old and new summaries and their statistics must be computed on the same (noisy) simulations to correctly distinguish between noise fluctuations and newly-informative features of the data during optimisation.

Optimising a neural function $\mathbf{x} = f(\mathbf{d}; w)$ to maximise the determinant of $\mathbf{F}$ from eq. (3.10) forces the new summaries $\mathbf{x}$ to add complementary information to the existing summaries' Fisher contributions. As described in [16] and [48], the Fisher information is invariant to nonsingular linear transformations of the summaries. To remove this ambiguity, a term penalising the network summary covariance $\mathbf{C}_x$ is added. This gives the loss function

$$\Lambda = -\ln\det\mathbf{F} + \lambda\frac{1}{2}\operatorname{tr}\mathbf{C}_x \tag{3.13}$$

where λ is a regularising coefficient. This scalar loss function can be optimised via gradient descent to update weights w for the network's contributions to the combined Fisher information. The network can do no worse at summarising the data than the existing summary, since the loss only optimises the second term of eq. (3.10). With the updated Fisher information we can also compute quasi-maximum likelihood estimates (MLE) for the parameters for a given mean-subtracted summary vector $\Delta = [\Delta\mathbf{t}, \Delta\mathbf{x}]$ [42]:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{fid}} + \mathbf{F}^{-1}\boldsymbol{\mu}_{,\theta_i}\mathbf{C}^{-1}\Delta^T. \tag{3.14}$$

We then use these hybrid statistics (as they are functions of the data) as our highly-informative and extremely compressed data set, ideal for simulation-based or implicit inference. The resulting compression is *asymptotically* optimal at the fiducial point in parameter space, but for smoothly-varying data manifolds results in a smooth summary space that can be exploited for neural posterior estimation away from the fiducial point as described in [15, 16]. A useful aspect of learning this local compression is that a prior for posterior estimation can be specified after the compression network is trained, unlike regression networks.

4 Weak gravitational lensing

4.1 Formalism

The effect of weak lensing (WL) on a source field is defined by its shear, γ, which captures the distortions in the shapes of observed galaxies. In the flat-sky limit in Fourier space, this observable can be related to the convergence field κ observable, which describes variation in angular size:

$$\tilde{\gamma}(\boldsymbol{\ell}) = \frac{(\ell_1 + i\ell_2)^2}{\ell^2}\tilde{\kappa}(\boldsymbol{\ell}) \tag{4.1}$$

where $\boldsymbol{\ell} = (\ell_1, \ell_2)$ is the complex wavevector. The convergence field can be connected to the underlying dark matter field by integrating the fractional overdensity along the line-of-sight to give [49]:

$$\kappa(\boldsymbol{\vartheta}) = \frac{3H_0^2\Omega_m}{2c^2}\int_0^{r_{\rm lim}} \frac{r dr}{a(r)} g(r)\delta^f(r\boldsymbol{\vartheta}, r), \tag{4.2}$$

where $\boldsymbol{\vartheta}$ denotes the coordinate on the sky, r is the comoving distance, $r_{\rm lim}$ is the galaxy survey's maximum comoving distance, δ^f is the dark matter overdensity field at scale factor a, and, for a flat Universe

$$g(r) = \int_r^{r_{\rm lim}} dr' n(r')\frac{r - r'}{r}, \tag{4.3}$$

is the integration of the redshift distribution $n(r)$ in the given comoving shell. In real-data analyses the data will be the cosmic shear, but here we restrict our analysis to noisy convergence maps. The forward model to generate $\boldsymbol{\kappa}$ consists of a cosmological parameter draw, $\boldsymbol{\theta}$, which is used to generate primordial fluctuations, $\delta^{\rm ic}$. Here the initial conditions are a Gaussian random field governed by the [50] cosmological power spectrum $P(k;\theta)$, which includes baryonic acoustic oscillations (BAO). The cosmic initial conditions are then evolved forward via a specified non-linear gravity model $G(\delta^{\rm ic})$, which describes the growth of the large-scale structure (LSS). The evolved dark matter field δ^f is then used to generate the convergence field. Using the Born Approximation, we implement a discrete version of equation (4.2) using a summation over voxels to approximate the radial line-of-sight integrals:

$$\kappa^b_{mn} = \frac{3H_0^2\Omega_m}{2c^2}\sum_{j=0}^{N}\delta^f_{mnj}\left[\sum_{s=j}^{N}\frac{(r_s - r_j)}{r_s}n^b(r_s)\Delta r_s\right]\frac{r_j\Delta r_j}{a_j}, \tag{4.4}$$

where b indexes the tomographic bin, and m, n index the spatial pixels on the sky. The index j indicates the voxel along the line-of-sight at the comoving distance r_j. The total number of voxels along the line-of-sight, N, is obtained from a ray tracer. Δr_j is the length of the line segment inside voxel j, and δ^f_{mnj} is the discretized dark matter overdensity field. The comoving radial distance r_s is the distance to the source. Each tomographic bin has a source redshift distribution $n^b(z_s)$. Once κ^b_{mn} is computed, the convergence field $\mathbf{d} = \{\hat{\kappa}^b_{mn}\}$ is obtained by adding uncertainties equivalent to the shape noise (and measurement error) in the shear field. This is captured by zero-centred Gaussian white noise added pixel-wise with variance

$$\sigma_n^2 = \sigma_\epsilon^2 \frac{N_{\text{tomo}}}{n_{\text{gal}} A^b_{\text{pixel}}}, \tag{4.5}$$

where σ_ϵ^2 is the total galaxy intrinsic ellipticity dispersion, n_{gal} is the source galaxy density on the sky, and A^b_{pixel} is the angular size of the pixel in each tomographic bin. For Stage-IV weak lensing surveys like Euclid n_{gal} will be ~ 30 arcmin^{-2} and $\sigma_\epsilon \simeq 0.3$ [51]. For network training purposes we introduce an amplitude scaling parameter $\sigma_n' = A\sigma_n$ that we report in terms of effective source galaxy density.

4.2 Simulation details

We analyse several simulation suites at different resolutions to conduct our experiments. In all cases our physical box size is kept fixed at $L_x = L_y = 250$ Mpc h^{-1} and $L_z = 4000$ Mpc h^{-1} in comoving units, in a pixel grid of shape $(N_x, N_y, N_z) = (N, N, 512)$, where we vary N to probe changing gravity solver scales. We utilise `pmwd` particle mesh (PM) simulations [52] integrated for 63 timesteps to generate the nonlinear dark matter overdensity field for $N = [64, 128]$ resolution and 100 timesteps for $N = 192$ resolution. This controls the particle spacings L/N which probe increasingly nonlinear scales described by the particle-mesh (PM) simulations. We compute the line-of-sight integral in comoving units before binning the L_z dimension in redshift bins converted to comoving units via the cosmology-dependent change of variable. For this analysis we do not include lightcone effects. We choose our four tomographic redshift bins to be Gaussian, centred at $z = [0.5, 0.75, 1.0, 1.25]$ with width $\sigma_z = 0.14$, following [8]. The resulting convergence fields span a 3.58×3.58 deg^2 field of view. Shape noise is added to the noise-free simulations before computing two-point or network statistics as described below. We generate two distinct datasets to i) construct a locally optimal compression and ii) perform posterior density estimation.

Compression simulations. For a given resolution we generate two equally-sized datasets for training and validation of our network compression. To calculate network and two-point covariances described below we simulate $n_s = 1500$ simulations at a fiducial cosmology $\boldsymbol{\theta}_{\text{fid}} = (\Omega_m, S_8) = (0.3, 0.8)$. For finite-difference derivatives we simulate $n_d \times 2 \times n_p = 375 \times 2 \times 2$ seed-matched simulations at a perturbed parameter set $\boldsymbol{\theta} \pm \Delta\boldsymbol{\theta}_{\text{fid}} = \boldsymbol{\theta}_{\text{fid}} \pm (0.0115, 0.01)$. All other cosmological parameters were held fixed at Planck 2018 parameters [53]. The total number of simulations used for optimal compression is thus 4500.

Density estimation simulations. Because our compression from the convergence field data is learned locally, we are free to choose our prior guided by the compression method's Fisher forecast. We simulate 5000 simulations over a uniform prior in (Ω_m, S_8), whose width is chosen according to the strategy described in section 5.3.

5 Finding hybrid weak lensing statistics

The information-update formula is perfectly suited to improving weak lensing $\Omega_m - \sigma_8$ parameter constraints with neural summaries. We would like to know if more cosmological parameter information can be extracted from the convergence field beyond a simulation-based tomographic angular C_ℓ statistic analysis, and in what resolution regimes. We present the MOPED scheme for angular C_ℓ compression and information-update neural network architecture.

5.1 MOPED angular C_ℓ compression

Here our existing summaries are either binned angular C_ℓ or MOPED-compressed vectors $\mathbf{t}$ (without the optional Gram-Schmidt orthogonalisation employed in [41]). We outline our setup below and display a schematic of the architecture in figure 1.

We compute empirical auto- and cross-spectra C_ℓ across the four noisy tomographic bins, resulting in 10 C_ℓ vectors. To test scale-dependent information, we optionally apply a maximum $\ell_{\rm cut}$ to each vector to mimic existing survey analyses. To reduce the number of simulations needed to estimate the covariance matrix, we bin each spectrum into six evenly-spaced ℓ bins weighted by C_ℓ value. We can estimate the covariance of the C_ℓ vector using the n_s fiducial simulations, and the finite difference derivatives with respect to each parameter via eq. (3.11). Together, the Fisher matrix for these summaries is computed with eq. (3.1). With these ingredients we can then perform the MOPED compression from mean-subtracted vectors evaluated at a fiducial set of parameters $\Delta = \left(\hat{C}_\ell - \langle C_\ell \rangle_{\rm fid}\right)$ down to score summaries $\mathbf{t}$:

$$\mathbf{t} = \boldsymbol{\theta}_{\rm fid} + \boldsymbol{\mu}_{,\theta_i} \mathbf{C}_t^{-1} \Delta^T. \tag{5.1}$$

which can then be rescaled by the Fisher matrix to obtain an MLE of the parameters [42]:

$$\hat{\boldsymbol{\theta}}_{\rm MOPED} = \boldsymbol{\theta}_{\rm fid} + \mathbf{F}_t^{-1} \mathbf{t}. \tag{5.2}$$

In practice, we replace $\mathbf{t}$ with $\hat{\boldsymbol{\theta}}_{\rm MOPED}$ as our existing MOPED statistics, which has covariance $\mathbf{C}_t = \mathbf{F}_t^{-1}$. These compressed summaries are the default C_ℓ-based summaries that we feed into the normalising flow posterior estimation scheme (section 5.3), as normalising flows are not guaranteed to work well with large inputs e.g. the 60-dimensional binned C_ℓ vector [6]. For network optimisation however, the longer, binned power spectrum vector can be used to find $\hat{\boldsymbol{\theta}}_{\rm network}$. Changes in the ℓ bins with respect to noise and parameters increases the number of cross-correlations ($\mathbf{u}$) with network summaries and encourage improved information extraction. We explore this option for training in noisy settings in section 6.2. Both choices of statistics fit neatly into the existing statistic formalism described in section 3.

5.2 Physically-informed neural network architecture

We design a novel, lightweight 2D convolutional neural network that is constrained by physics knowledge, namely that a) physical laws are translation-invariant and direction-invariant, and b) the non-Gaussian cosmic shear information is found in clustering patterning on small scales. This motivates both the CNN approach, and the use of a convolution kernel whose

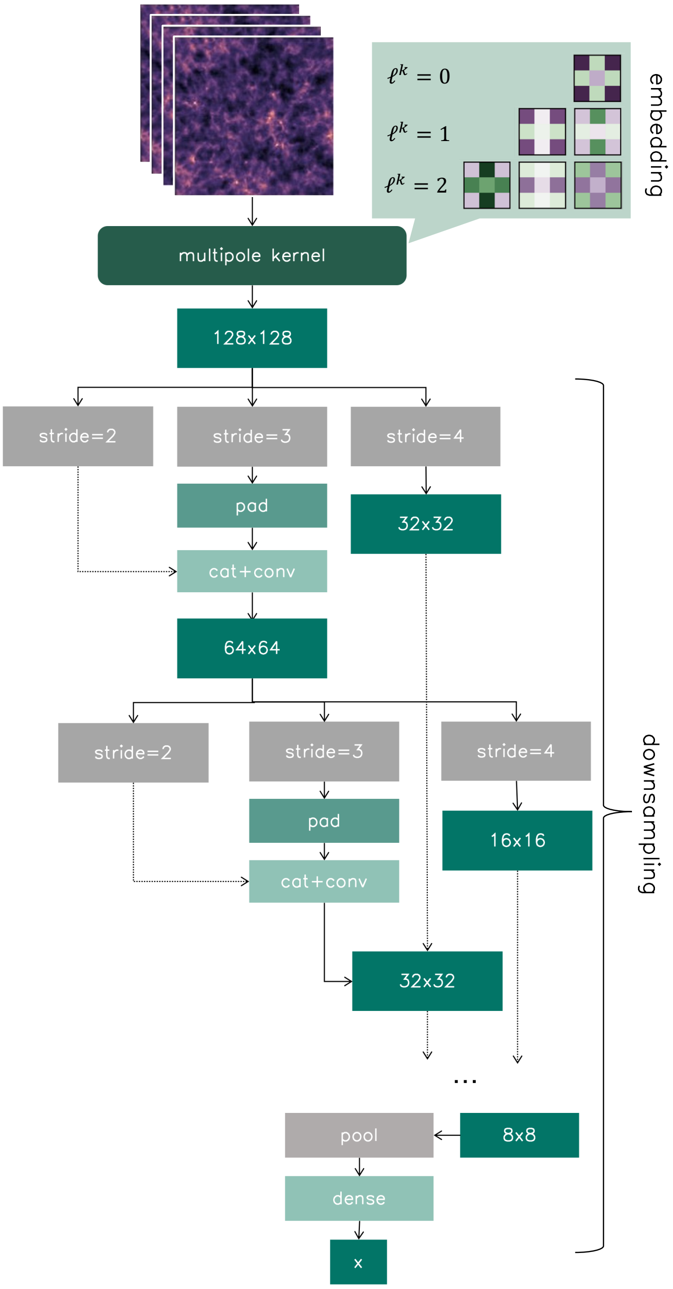

Figure 2. We use a small convolutional neural network that exploits the data symmetries to compress κ fields down to additional summaries. Input data (here of shape $(128, 128, 4)$) are passed through a residual multipole kernel layer (shared colour indicates shared weights) and then subsequently passed to convolutional blocks with varying strides with small 2×2 kernels to capture fluctuations on different scales. All linear layers are followed by a nonlinear activation function. Dashed arrows indicate feature concatenation at the same spatial resolution. This downsampling continues until the spatial resolution of the data reaches 8×8, after which the output tensor is mean-pooled along the spatial axes and passed to three dense layers. The output from the network is a pair of numbers.

complexity is truncated to a multipole expansion around a circularly-symmetric kernel. We present this network layer-by-layer and display a schematic in figure 2.

The inputs to the network are the log-transformation of the convergence maps at the four specified tomographic bins, adapted from [54–56]:

$$\boldsymbol{\kappa}^b = \kappa_o \ln\left[1 + \boldsymbol{\kappa}^b/\kappa_o\right] \tag{5.3}$$

where $\kappa_o = |\kappa^b_{\rm min}| + 0.01$, where $\kappa^b_{\rm min}$ is the minimum convergence value for the tomographic bin b.

Multipole kernel embedding. For convergence maps we can target clustering information by learning convolution functions of the data with certain, enforced symmetries. [57] showed via neural emulation of dark matter simulations that learned convolutional weights tend to be distributed in spherically-symmetric and close-to-spherically-symmetric ways. Although the physics laws suggest restriction to circularly-symmetric kernels in 2D, as employed by [58] and [59] for modelling halo bias corrections, [57] found that networks were able to extract more information by mild breaking of this symmetry, perhaps simply associated with the grid pixelisation of the kernel. This motivates ordering kernel complexity by increasingly breaking from rotational symmetry. Here we expand on [59]'s implementation and explicitly encode low-order multipole expansion symmetries in CNN kernels. Convolutional kernel weights are shared for kernel pixels equidistant from the centre of a 3D or 2D kernel, associated to the spherical harmonic coefficients $Y^{\ell_k}_m(\theta, \phi)$. Here we make use of these multipole kernels (MPK) in a 2D setting for information capture, embedding the convergence field using a smaller number of neural weights. This choice of embedding is also likely to improve performance in the presence of (white) noise, as noise artefacts are not distributed with the same rotational symmetry as convergence clustering features.

We first embed each log-transformed tomographic bin into six filters corresponding to the $\ell_k = [0, 1, 2]$ multipole moments for a 7×7 kernel per tomographic bin, which for e.g. $N = 128$ corresponds to a 0.19 deg^2 receptive field. We found empirically that including higher ℓ_k did not improve information capture. We illustrate a cartoon example of these symmetric kernels for a 3×3 kernel in figure 2. This output is then passed to a nonlinearity and then to another multipole kernel for each input filter, which are then summed along the filter axis at each multipole kernel to yield six output channels. We learn the residual from the first embedding layer l to the next, e.g. $x^{l+1} = \texttt{mpk\_layer}(x^l + \texttt{mpk\_layer}(x^l))$. This choice of data embedding layer drastically reduces the number of learnable weights and forces the network to learn physically-symmetric functions of the data in its first layer. The largest model considered here contains just 6,904 trainable parameters, which is 0.08% the footprint of the ResNet18 employed e.g. by [30, 60].

Incept-stride tree network. The embedded data are then passed to an inception-style network [61] with one important difference: instead of varying kernel sizes, we keep the kernel shape fixed to 2×2 and vary the *stride* that each layer takes in parallel passes over the data. The objective of this section of the network is to downsample the embedded data by combining information from different scales so that the only features on informative scales are strongly activated and pushed through the learned network to the output summaries using the fewest independent kernel weights possible.

The data is passed to stride-2, stride-3, and stride-4 downsampling layers followed by a nonlinear activation function and a subsequent stride-1 convolution. The outputs of the stride-3 block are padded periodically in the spatial dimension and concatenated to the output of the stride-2 block, and then passed to another stride-1 convolution. The stride-4 outputs are kept aside until the data has been passed to the next inception block and the data has reached the same spatial resolution. We continue downsampling in this tree-like fashion until the data reach a spatial resolution of 8×8. We then mean-pool the features along the spatial axes and pass the resulting flattened filter axis to a final linear layer that outputs the desired additional summaries. Every layer is followed by a new, bijective `smooth_leaky` activation function, which we found empirically extracted information most consistently across our experiments:

$$\texttt{smooth\_leaky}(x) = \begin{cases} x, & x \leq -1 \\ -|x|^3/3, & -1 \leq x < 1 \\ 3x & x > 1. \end{cases} \tag{5.4}$$

The intuition here is that unlike natural image data, lensing shear maps are relatively smooth functions, so are best linked to smoother activation functions following convolutions, in contrast to natural images with sharp features like feature borders, for which typical disjoint activations like the `leaky_ReLU` were developed [62].

Training setup. To train the network we split our dataset into equally-sized validation and training sets, with the same n_s number of fiducial and n_d seed-matched derivative simulations. Every epoch a new noise realisation is added onto the noisefree convergence maps and a random rotation is performed. These transformations are seed-matched for each derivative simulation index. We use the `adam` optimiser with a fixed learning rate of 0.0005 with gradient clipping at a value of 1.0, and a weight decay penalty of 0.0005 added to the loss function. These two modifications to the optimisation routine "smooths" the loss landscape and prevents the network from overfitting to the training data, respectively. Training is halted when the validation loss stops decreasing significantly for a `patience` number of epochs.

Noise hardening. All networks are first trained at a low noise level, $A_{\rm noise} = 0.125$, after which the noise level is increased in increments of $\Delta A_{\rm noise} = 0.05$ for a minimum of 100 epochs subject to a patience setting of 75 epochs at each setting. This can be thought of as "domain-transfer" learning on-the-fly. Slowly increasing the noise allows physical features (e.g. convergence patterns) to be embedded early in training, such that the network outputs are already concentrated on the informative distribution of the data when the shape noise increases.

5.3 Neural density estimation

To measure the information capture in both MOPED and network summaries we employ a neural posterior estimation scheme to parameterise the amortised summary-parameter posterior $p(\boldsymbol{\theta}|\mathbf{y})$, where $\mathbf{y}(\mathbf{d})$ is either the MOPED summary or updated summary set of

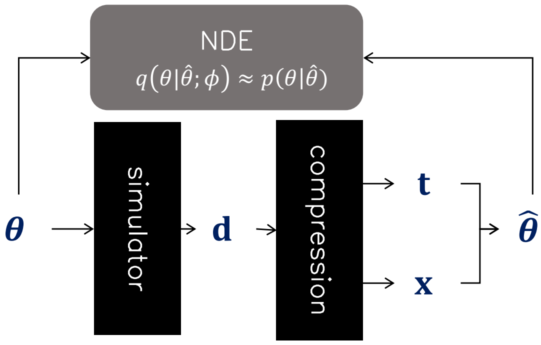
NDE
$q(\theta|\hat{\theta};\phi) \approx p(\theta|\hat{\theta})$
$\boldsymbol{\theta}$
simulator
$\mathbf{d}$
compression
$\mathbf{t}$
$\mathbf{x}$
$\widehat{\boldsymbol{\theta}}$

Figure 3. Cartoon of density estimation scheme with fixed compression (network or MOPED). Parameters $\boldsymbol{\theta}$ are drawn from a prior and MLE estimates $\hat{\boldsymbol{\theta}}$ are produced from data $\mathbf{d}$ for either (fixed) compression method using eq. (3.14) and fed to a MAF neural density estimator for the amortised posterior distribution, trained under the loss in eq. (2.1).

MLE parameter estimates using eq. (3.14). We employ an identical ensemble of masked autoregressive flows [MAFs; 18, 40] for each set of summaries using the LtU-ILI codebase [63]. We opt for networks with 50 hidden units and 12 transformations. We chose this high level of complexity such that the posterior density parameterisations in all cases were sufficiently descriptive. A unique aspect of our network training scheme is that a joint parameter-data prior distribution can be chosen after learning the network and MOPED compression, displayed as a cartoon in figure 3. We generated 5000 simulations for each of two wide uniform priors in the S_8 formalism: $p^{(1)}(\Omega_m, S_8) = \mathcal{U}[[0.15, 0.35] \times [0.7, 1.52]]$ and $p^{(2)}(\Omega_m, S_8) = \mathcal{U}[[0.15, 0.35] \times [0.5, 1.0]]$. We opt for the smaller prior in cases where the 3σ Fisher posterior estimate for the observable considered falls within the support of $p^{(2)}$.

6 Results

6.1 Low-noise regime

We first investigate the information extraction as a function of dark matter simulation resolution with a small amount of shape noise, and compare information extraction to two-point statistics at different scales. We construct particle mesh simulations with varying numbers of pixels $N_x = N_y \in [64, 128, 192]$. Intuitively we expect more information beyond the two-point statistic to be found at higher resolutions, which can be interpreted as the descriptiveness of the underlying gravity model.

Scale cutoff. We first explored the effect of a scale cutoff at resolution $N = 128$ for the C_ℓ summaries with a low noise setting. We construct MOPED summaries from C_ℓs subject to a Stage-III survey-like cut at $\ell_{\text{cut}} = 1500$, as well as summaries from all available ℓ modes at the given resolution. The highly-compressed MOPED summaries give almost identical posteriors to using the full set of C_ℓ values as the data vector. The network is tasked with finding complementary summaries in the $\ell_{\text{cut}} = \ell_{\max}$ case. We display the constraints obtained on the same target simulation in figure 4 and in table 1. The network extracts up to 5 times more information than the two-point function in a low-noise setting with all modes and 8.3 times more information in high-noise settings, as measured by the determinant of the Fisher matrix.

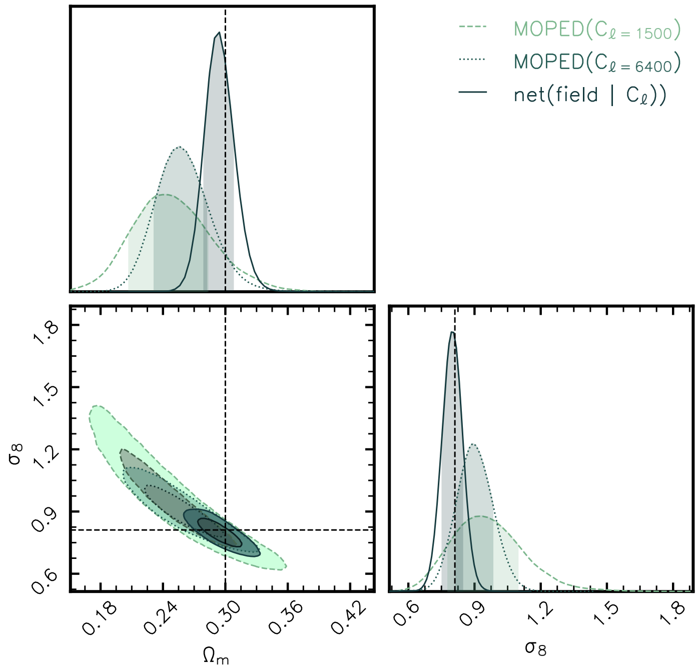

Figure 4. Using information-update network summaries (green) drastically improves $\Omega_m - \sigma_8$ constraints beyond MOPED C_ℓ summaries in a low-noise setting. We compare the posteriors obtained from a KiDS-like survey truncation at $\ell_{\rm cut} = 1500$ (blue) to the constraints from all available modes $\ell_{\rm cut} = 6400$ at the given resolution (green). The network's additional summaries (dark green) is able to improve information extraction by a factor of 5 beyond the $\ell_{\rm cut} = 6400$ and a factor of 8 above $\ell_{\rm cut} = 1500$.

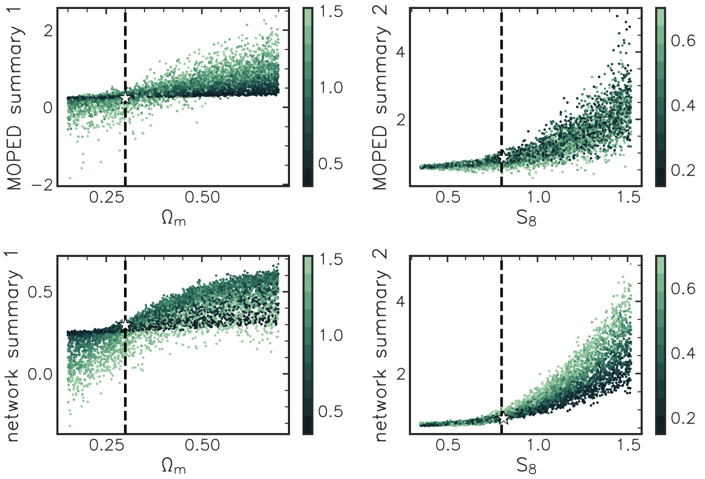

Figure 5. Information-update network (bottom) makes simulations more distinguishable in summary space than C_ℓ compression (top). Points in parameter-summary space are coloured by the opposite parameter's value. The network finds patterns that separate these summaries in a complementary fashion even away from the fiducial point $(\Omega_m, S_8) = (0.3, 0.8)$. We display a 3D view of this four-dimensional space in figure 11.

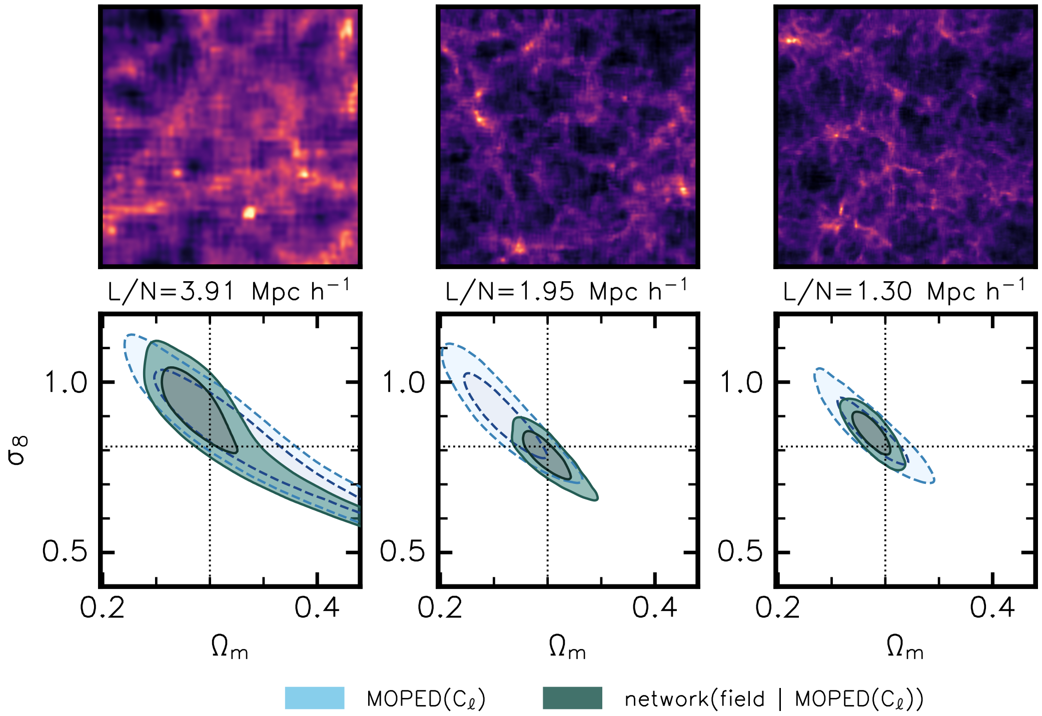

Figure 6. Computing additional complementary summaries from the convergence field improves parameter constraints (green) over the two-point information (blue) as the field becomes more nonlinear in the low-noise regime. For $N = 64$ fields the information gain above the C_ℓ constraints is modest, but improves as more nonlinear scales are included at the level of the field as resolution increases.

	resolution	$H(C_\ell)$	$H(\mathrm{net})$	ratio
low noise	$N = 64$	6.9	7.4	**2.9**
	$N = 128$	6.9	7.7	**5.0**
	$N = 192$	7.6	8.5	**5.1**
high noise	$N = 128$	5.3	6.0	**4.5**
	$N = 192$	5.2	6.3	**8.3**

Table 1. Summary of parameter Shannon information ($\mathrm{H} = \frac{1}{2} \ln \det F$) from MOPED and information-update networks for low noise ($n_{\mathrm{gal}} = 1900$) and high noise ($n_{\mathrm{gal}} = 83$) scenarios. The ratios of Fisher determinants are shown in the last column. For the noisy $N = 192$ case we optimise networks against the binned C_ℓ vectors as opposed to the MOPED summaries.

Summary scatter. The information-update loss scheme asks the network to use data features such that output summaries complement existing C_ℓ-based summaries. We can interpret the statistics learned by the network by visualising a summary scatter over the suite of prior simulations. Figure 5 shows the network and MOPED outputs versus true parameter, coloured by the opposite parameter's value for summaries used to generate the network and $\ell_{\rm max} = 6400$ posteriors in figure 4. Remarkably, even though the network is only trained at the fiducial cosmology (dashed vertical line in each plane), the information-update loss allows the network to find useful features with which to distinguish parameters in a smoother summary space (less scatter) than the MOPED compression. This increased structure is then harnessed by the density estimation scheme to provide tighter parameter constraints than MOPED. The complementary nature of the information from the network-updated statistics to the original statistics decreases away from the fiducial point in both dimensions, but does so smoothly, i.e. the information about the parameters coming from the four statistics is mixed away from the fiducial point.

Resolution dependence. We next compare constraints as a function of PM simulation resolution, which effectively controls the nonlinearity of the dark matter gravity solver. Here we wish to measure the parameter information gain that using a nonlinear network to probe nonlinear scales adds to the power spectrum. We generate three suites of fiducial and finite-difference convergence maps to learn the compression with $A_{\rm noise} = 0.125$. Figure 6 shows constraints at each resolution. More non-Gaussian information is extracted for the higher resolution simulations, indicated by the increase in network constraining power over the C_ℓs, since these simulations probe smaller, more nonlinear scales accessed by the network. We report our network and MOPED Fisher constraints in table 1. In this low-noise setting we observe an information increase of a factor of 2.9 for $N = 64$ and a factors of 4-5 for $N = [128, 192]$ high-resolution simulations, aligning with our intuition.

6.2 High-noise regime

The information-update formalism displays promising results in the presence of increased systematics such as galaxy shape noise. Here we start with a network trained on the lowest noise setting and slowly increase the noise amplitude (equivalently decreasing the galaxy density). The network is able to increase its relative performance against the two-point statistic as we increase resolution (figure 7) and shape noise (figure 8). Figure 7 shows that with increased simulation resolution the network has access to more nonlinear scales and can compensate for the shot noise that dominates the C_ℓ calculation at high ℓ values. For a fixed resolution ($N = 192$), the additional summaries found by the network appear robust to noise; keeping the parameter constraints consistent as increased shape noise pushes the C_ℓ constraints towards the prior edge in Ω_m. This is especially promising for higher noise cases for galaxy density $n_{\rm gal} < 50$ arcmin^{-2}, as this falls between the capabilities of Euclid [51] and Roman [64] telescopes.

Optimising with the binned power spectrum. For our high-resolution simulations we also explored optimising the information-update formalism (eq. (3.10)) with respect to the full binned C_ℓ vector. This "stretches" the off-diagonals $\mathbf{u}$ of the full summary covariance

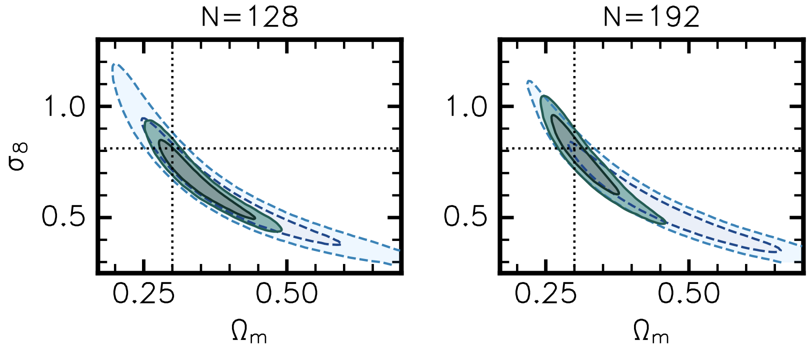

Figure 7. Additional neural summaries (green) are robust to shape noise ($n_{\rm gal} = 120$) at different resolutions. Constraints from C_ℓ-only summaries (blue) suffer from the increased noise due to shot-noise contributions to the high-ℓ bins.

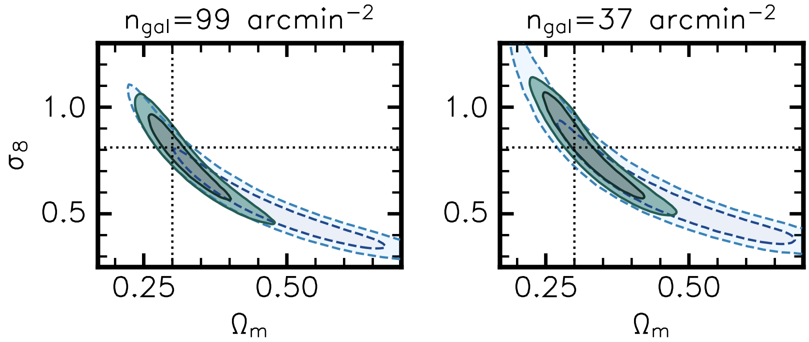

Figure 8. Neural summaries (green) are robust to increased shape noise, controlled by the galaxy density parameter $n_{\rm gal}$. Increased shot noise at small scales inhibits angular C_ℓ constraints (blue). Here we display an inference on the same $N = 192$ resolution simulation subject to increased shape noise. We optimise the network using the binned C_ℓ vectors.

(eq. (3.4)), such that the IUF forces the network to respond to explicit fluctuations in these C_ℓ bins with respect to noise. Here we posit that the network will be able to find summaries that complement fluctuations at different ℓ scales more efficiently. We find that indeed this choice of optimisation allows the network to extract 8.3 times more information than the two point function in the noisiest ($n_{\rm gal} = 30$) setting, compared to a 5.6 times improvement when optimising against the MOPED-compressed summaries. We visualise the joint covariance of learned and binned C_ℓ summaries (eq. (3.4)) in appendix B.

7 Discussion & conclusions

In this paper, we present an implicit inference technique to extract neural summary statistics from field-level data, specifically weak lensing maps, that are designed to match or increase automatically the Fisher information about the cosmological parameters over a set of pre-defined summaries, typically traditional two-point statistics. We apply this method to find summary statistics from tomographic convergence maps that explicitly complement the angular power spectrum estimates. This powerful hybrid mixture of physics-based and neural network derived summary statistics is guaranteed to improve the two-point parameter constraints and allows for networks with small physics-informed architecture to achieve similar results to larger regression networks. We demonstrated that this approach extracts between a factor 3 and 8 more information than the angular power spectrum, as measured by the determinant of the Fisher matrix. For weak lensing, the main gain is a substantial reduction in the credible region for Ω_m, with a smaller improvement in the S_8 error.

Other studies have previously combined power spectrum and network summaries in weak lensing analyses. Ref. [37] for example feed in both sets of independently-obtained summaries into an NDE for posterior estimation to obtain a $\sim 2\times$ improvement in information extraction in $\Omega_m - S_8$. Here we show that coordinating the field-level network optimisation with an existing summary can give us even more efficient extraction.

The hybrid summary formalism presented in this work is not limited to weak lensing data, and it can be generalised to any dataset to identify the features from which the information captured by large neural networks comes. This technique might also reduce the need for large convolutional networks to learn the large-scale correlations in larger dark matter and galaxy simulations [65]. In future work, we will apply this formalism to find the summary that complements the information from more than one pre-determined summary statistic, such as angular power spectrum and peak count summaries. This has the potential to improve the cosmology constraints from implicit likelihood analyses of weak lensing such as [36] and [37].

It is worth making a final comment on the black-box nature of the neural network. Although the additional summary statistics are obtained by the network based on information theoretic considerations, the process might still be regarded as a black box. For the purposes of Bayesian inference, however, it does not matter how the network found the summaries; it only matters that they are informative. The (a priori unknown) sampling distribution of the summaries is not used in simulation-based inference, which can be applied in a fully Bayesian way. However, our new multipole network architecture does allow the user to probe this information capture as a function of kernel complexity. Here we found empirically that e.g. truncating kernels to $\ell_k = [0, 1]$ captured less information than the $\ell_k = [0, 1, 2]$ kernels used in the presented analysis. We leave a thorough comparison to a future work.

8 Code availability

The code for this analysis will be made available at https://github.com/tlmakinen/hybridStatsWL. All custom networks and simulation tools were written in `Jax` [66] and `flax` [67] and were run on a single NVIDIA v100 32Gb GPU. Posterior density estimation was performed locally on a laptop CPU using the LtU-ILI code [63].

Acknowledgments

TLM acknowledges the Imperial College London President's Scholarship fund for support of this study, and thanks Niall Jeffrey, Justin Alsing, David Spergel, and Maximilian von Wietersheim-Kramsta for insightful discussions. NP is supported by the Beecroft Trust. BDW. acknowledges support by the ANR BIG4 project, grant ANR-16-CE23-0002 of the French Agence Nationale de la Recherche; and the Labex ILP (reference ANR-10-LABX-63)

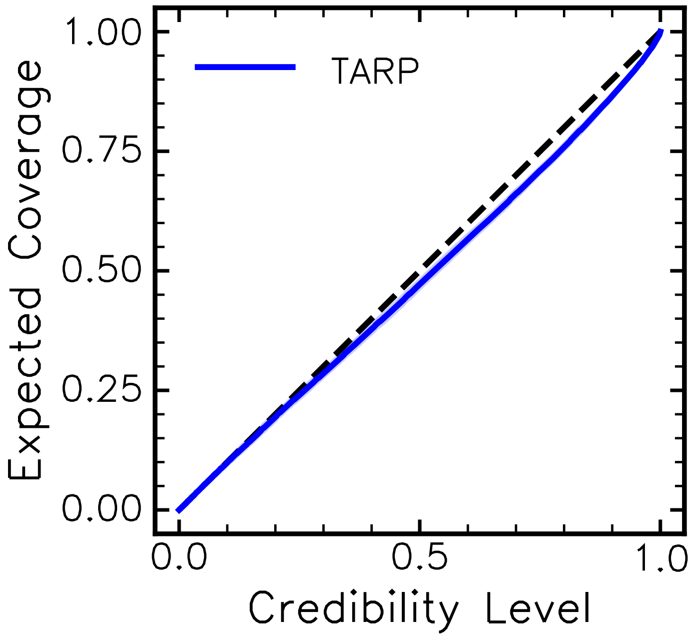

Figure 9. Example coverage test result for low-noise inference with $N = 192$ resolution (rightmost panel in figure 6) using `TARP` [68]. Using our amortised parameter-summary posterior $p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, we can do repeated mock data parameter inference over the prior, and measure which fraction of true values from the appropriate credible intervals matches the expected fraction. The blue line traces 100 "distances to random points" (DRP), which is accelerated using the `TARP` framework within LtU-ILI [63]. The DRP line (blue) traces the truth line (dashed), indicating a successful test.

part of the Idex SUPER, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme Investissements d'avenir under the reference ANR-11-IDEX-0004-02. TLM acknowledges helpful conversations facilitated by the Learning the Universe Collaboration. This work was conducted within the Aquila Consortium. The Flatiron Institute is supported by the Simons Foundation.

A Posterior coverage tests

One of the distinct advantages of SBI neural density estimation is the immediate availability of coverage tests. In this work we trained an estimator for the posterior distribution given some point-estimates for the parameters via MOPED or the hybrid-summary network: $p(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$. This density estimator is an *amortised* posterior, meaning the posterior density for any given summaries $\hat{\boldsymbol{\theta}}$ is immediately available without having to do MCMC sampling with a likelihood. We can then do repeated mock data parameter inference over the prior using this posterior density, and calculate how many true parameter vlaues from the credible intervals match the expected fraction, forming a posterior "coverage" test. We display one such test in figure 9 making use of the `TARP` coverage test framework presented in [68].

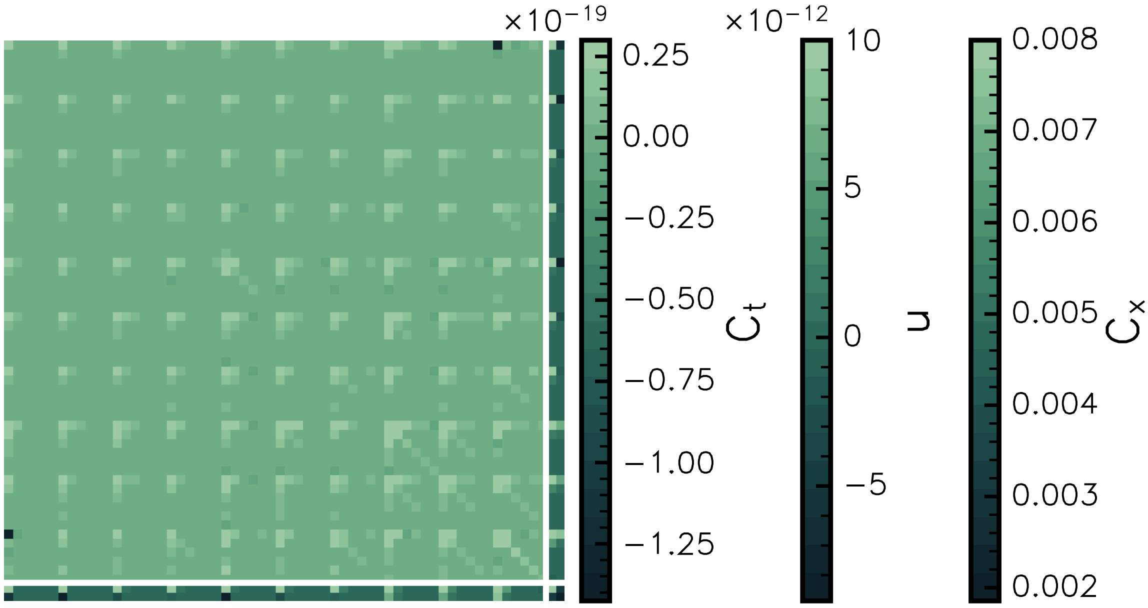

Figure 10. Example joint C_ℓ-network summary covariance visualisation (eq. (B.1)) for a network optimised against the binned power spectrum. We separate the 60×60 C_ℓ covariance structure (upper left corner) from network summaries (lower right corner) with the set of intersecting white lines. The learned network summaries exhibit a non-zero correlation structure with the ℓ bins, illustrated by the $\mathbf{u}$ off-diagonal matrix vectors on the bottom and right-hand edges. Here it is obvious that only one of the two network summaries is highly correlated with the binned power spectrum modes.

B Learned covariance matrix visualisation

In figure 10 we illustrate the full joint covariance,

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_t & \mathbf{u} \\ \mathbf{u}^T & \mathbf{C}_x \end{pmatrix}, \tag{B.1}$$

of the binned tomographic C_ℓ statistic and two learned network summaries, clearly separated by the white intersecting lines. We plot each component of this structure with a separate colourbar. The cross-correlation row-matrices $\mathbf{u}$ indicate that the learned summaries exhibit non-trivial correlation structure with the binned ℓ-modes, which contributes to information capture according to the hybrid statistics formalism.

C Summary scatter

Here we display a three-dimensional view of the four-dimensional joint distribution of compressed summaries and parameters $p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ obtained from MOPED and network compressions in a low-noise setting. The information update formalism tells the convolutional network during optimisation to make use of the nonlinear information on the smaller (pixel-level) scales that it has access to in a way that is complementary to the power spectrum. Although optimised at a fiducial point, the mapping learned is smooth as a function of data $\mathbf{d}(\boldsymbol{\theta})$ away from the training point, resulting in more structure in the four-dimensional joint distribution space $p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ than MOPED, allowing the summaries (z-axis and colour) to respond more

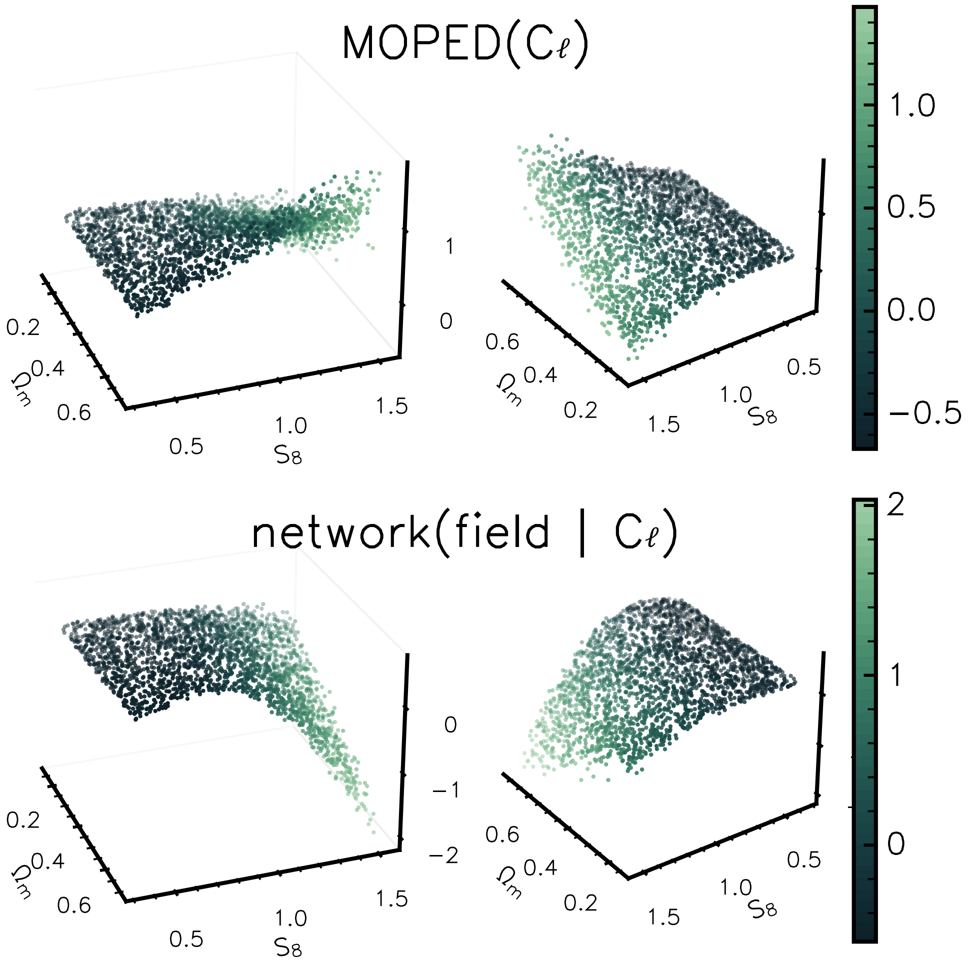

Figure 11. Computing additional information from the data endows the network summaries (bottom row) with more structure as a function of parameters (Ω_m, S_8) over a prior than MOPED two-point summaries (top row). Summaries for each method $\hat{\Omega}_m$ and $\hat{S}_8$ are indicated by z-direction and colourbar, respectively. It is highly visible via the increased structure in joint space that the network is able to capture information from the smaller scales it has access to. More scatter in z or colour at a particular parameter value in the MOPED summaries indicates a less informative compression of the simulated convergence data to from power spectrum summaries.

smoothly and rapidly as a function of parameters $(x, y) = (\Omega_m, S_8)$. This smoother joint distribution surface can then be harnessed by the NDE scheme to produce tighter posteriors in an amortised fashion.

References

[1] DES collaboration, *Dark Energy Survey year 3 results: cosmology from cosmic shear and robustness to data calibration*, *Phys. Rev. D* **105** (2022) 023514 [arXiv:2105.13543] [INSPIRE].

[2] DES collaboration, *Dark Energy Survey year 3 results: cosmology from cosmic shear and robustness to modeling uncertainty*, *Phys. Rev. D* **105** (2022) 023515 [arXiv:2105.13544] [INSPIRE].

[3] S.-S. Li et al., *KiDS-1000: cosmology with improved cosmic shear measurements*, *Astron. Astrophys.* **679** (2023) A133 [arXiv:2306.11124] [INSPIRE].

[4] KiDS collaboration, *KiDS-1000 cosmology: cosmic shear constraints and comparison between two point statistics*, *Astron. Astrophys.* **645** (2021) A104 [arXiv:2007.15633] [INSPIRE].

[5] R. Dalal et al., *Hyper Suprime-Cam year 3 results: cosmology from cosmic shear power spectra*, *Phys. Rev. D* **108** (2023) 123519 [arXiv:2304.00701] [INSPIRE].

[6] K. Cranmer, J. Brehmer and G. Louppe, *The frontier of simulation-based inference*, *Proc. Nat. Acad. Sci.* **117** (2020) 30055 [arXiv:1911.01429] [INSPIRE].

[7] J. Alsing, A.F. Heavens and A.H. Jaffe, *Cosmological parameters, shear maps and power spectra from CFHTLenS using Bayesian hierarchical inference*, *Mon. Not. Roy. Astron. Soc.* **466** (2017) 3272 [arXiv:1607.00008] [INSPIRE].

[8] N. Porqueres, A. Heavens, D. Mortlock and G. Lavaux, *Bayesian forward modelling of cosmic shear data*, *Mon. Not. Roy. Astron. Soc.* **502** (2021) 3035 [arXiv:2011.07722] [INSPIRE].

[9] N. Porqueres, A. Heavens, D. Mortlock and G. Lavaux, *Lifting weak lensing degeneracies with a field-based likelihood*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 3194 [arXiv:2108.04825] [INSPIRE].

[10] N. Porqueres et al., *Field-level inference of cosmic shear with intrinsic alignments and baryons*, arXiv:2304.04785 [INSPIRE].

[11] A. Loureiro et al., *Almanac: weak lensing power spectra and map inference on the masked sphere*, *Open J. Astrophys.* **6** (2022) 2023 [arXiv:2210.13260] [INSPIRE].

[12] E. Sellentin et al., *Almanac: MCMC-based signal extraction of power spectra and maps on the sphere*, *Open J. Astrophys.* **6** (2023) 31 [arXiv:2305.16134] [INSPIRE].

[13] E. Sellentin and A.F. Heavens, *On the insufficiency of arbitrarily precise covariance matrices: non-Gaussian weak lensing likelihoods*, *Mon. Not. Roy. Astron. Soc.* **473** (2018) 2355 [arXiv:1707.04488] [INSPIRE].

[14] M. von Wietersheim-Kramsta et al., *KiDS-SBI: Simulation-Based Inference analysis of KiDS-1000 cosmic shear*, arXiv:2404.15402 [INSPIRE].

[15] T.L. Makinen, T. Charnock, J. Alsing and B.D. Wandelt, *Lossless, scalable implicit likelihood inference for cosmological fields*, *JCAP* **11** (2021) 049 [*Erratum ibid.* **04** (2023) E02] [arXiv:2107.07405] [INSPIRE].

[16] T. Charnock, G. Lavaux and B.D. Wandelt, *Automatic physical inference with information maximizing neural networks*, *Phys. Rev. D* **97** (2018) 083004 [arXiv:1802.03537] [INSPIRE].

[17] T.L. Makinen et al., *The cosmic graph: optimal information extraction from large-scale structure using catalogues*, *Open J. Astrophys.* **5** (2022) [arXiv:2207.05202] [INSPIRE].

[18] J. Alsing, T. Charnock, S. Feeney and B. Wandelt, *Fast likelihood-free cosmology with neural density estimators and active learning*, *Mon. Not. Roy. Astron. Soc.* **488** (2019) 4440 [arXiv:1903.00007] [INSPIRE].

[19] A. Peel et al., *Cosmological constraints with weak lensing peak counts and second-order statistics in a large-field survey*, *Astron. Astrophys.* **599** (2017) A79 [arXiv:1612.02264] [INSPIRE].

[20] DES collaboration, *Dark Energy Survey year 3 results: cosmology with peaks using an emulator approach*, *Mon. Not. Roy. Astron. Soc.* **511** (2022) 2075 [arXiv:2110.10135] [INSPIRE].

[21] J.M. Kratochvil, Z. Haiman and M. May, *Probing cosmology with weak lensing peak counts*, *Phys. Rev. D* **81** (2010) 043519 [arXiv:0907.0486] [INSPIRE].

[22] LSST Dark Energy Science (LSST DESC) collaboration, *Forecasting the power of higher order weak-lensing statistics with automatically differentiable simulations*, *Astron. Astrophys.* **679** (2023) A61 [arXiv:2305.07531] [INSPIRE].

[23] Euclid collaboration, *Euclid preparation. XXVIII. Forecasts for ten different higher-order weak lensing statistics*, *Astron. Astrophys.* **675** (2023) A120 [arXiv:2301.12890] [INSPIRE].

[24] S. Cheng, Y.-S. Ting, B. Ménard and J. Bruna, *A new approach to observational cosmology using the scattering transform*, *Mon. Not. Roy. Astron. Soc.* **499** (2020) 5902 [arXiv:2006.08561] [INSPIRE].

[25] S. Cheng et al., *Cosmological constraints from weak lensing scattering transform using HSC Y1 data*, arXiv:2404.16085 [INSPIRE].

[26] B. Dai and U. Seljak, *Multiscale flow for robust and optimal cosmological analysis*, *Proc. Nat. Acad. Sci.* **121** (2024) e2309624121 [arXiv:2306.04689] [INSPIRE].

[27] D.H. Ribli et al., *Weak lensing cosmology with convolutional neural networks on noisy data*, *Mon. Not. Roy. Astron. Soc.* **490** (2019) 1843 [arXiv:1902.03663] [INSPIRE].

[28] J. Fluri et al., *Cosmological constraints from noisy convergence maps through deep learning*, *Phys. Rev. D* **98** (2018) 123518 [arXiv:1807.08732] [INSPIRE].

[29] J. Fluri et al., *Cosmological constraints with deep learning from KiDS-450 weak lensing maps*, *Phys. Rev. D* **100** (2019) 063514 [arXiv:1906.03156] [INSPIRE].

[30] D. Sharma, B. Dai and U. Seljak, *A comparative study of cosmological constraints from weak lensing using convolutional neural networks*, *JCAP* **08** (2024) 010 [arXiv:2403.03490] [INSPIRE].

[31] D.K. Ramanah, G. Lavaux, J. Jasche and B.D. Wandelt, *Cosmological inference from Bayesian forward modelling of deep galaxy redshift surveys*, *Astron. Astrophys.* **621** (2019) A69 [arXiv:1808.07496] [INSPIRE].

[32] F. Leclercq, *Bayesian large-scale structure inference and cosmic web analysis*, Ph.D. thesis, U. Paris-Saclay, Orsay, France (2015) [arXiv:1512.04985] [INSPIRE].

[33] J. Jasche, F. Leclercq and B.D. Wandelt, *Past and present cosmic structure in the SDSS DR7 main sample*, *JCAP* **01** (2015) 036 [arXiv:1409.6308] [INSPIRE].

[34] S.S. Boruah and E. Rozo, *Map-based cosmology inference with weak lensing — information content and its dependence on the parameter space*, *Mon. Not. Roy. Astron. Soc.* **527** (2023) L162 [arXiv:2307.00070] [INSPIRE].

[35] E. Tsaprazi, J. Jasche, G. Lavaux and F. Leclercq, *Higher-order statistics of the large-scale structure from photometric redshifts*, arXiv:2301.03581 [INSPIRE].

[36] J. Fluri et al., *Full wCDM analysis of KiDS-1000 weak lensing maps using deep learning*, *Phys. Rev. D* **105** (2022) 083518 [arXiv:2201.07771] [INSPIRE].

[37] N. Jeffrey et al., *Dark Energy Survey year 3 results: likelihood-free, simulation-based wCDM inference with neural compression of weak-lensing map statistics*, *Mon. Not. Roy. Astron. Soc.* **536** (2025) 1303 [arXiv:2403.02314].

[38] C. Bishop, *Mixture density networks*, technical report NCRG/94/004, Aston University, U.K. (1994).

[39] S. Kullback and R.A. Leibler, *On information and sufficiency*, *Annals Math. Statist.* **22** (1951) 79 [INSPIRE].

[40] G. Papamakarios, T. Pavlakou and I. Murray, *Masked autoregressive flow for density estimation*, in *Advances in Neural Information Processing Systems*, I. Guyon et al. eds., volume 30, Curran Associates Inc., (2017) [arXiv:1705.07057] [INSPIRE].

[41] A. Heavens, R. Jimenez and O. Lahav, *Massive lossless data compression and multiple parameter estimation from galaxy spectra*, *Mon. Not. Roy. Astron. Soc.* **317** (2000) 965 [astro-ph/9911102] [INSPIRE].

[42] J. Alsing and B. Wandelt, *Generalized massive optimal data compression*, *Mon. Not. Roy. Astron. Soc.* **476** (2018) L60 [arXiv:1712.00012] [INSPIRE].

[43] J. Carron and I. Szapudi, *Optimal non-linear transformations for large scale structure statistics*, *Mon. Not. Roy. Astron. Soc.* **434** (2013) 2961 [arXiv:1306.1230] [INSPIRE].

[44] T. Hoffmann and J.-P. Onnela, *Minimising the expected posterior entropy yields optimal summary statistics*, arXiv:2206.02340.

[45] N. Jeffrey, J. Alsing and F. Lanusse, *Likelihood-free inference with neural compression of DES SV weak lensing map statistics*, *Mon. Not. Roy. Astron. Soc.* **501** (2021) 954 [arXiv:2009.08459] [INSPIRE].

[46] H. Cramér, *Mathematical methods of statistics*, The University Press (1946).

[47] C.R. Rao, *Information and the accuracy attainable in the estimation of statistical parameters*, *Bull. Calcutta Math. Soc.* **37** (1945) 81 [DOI:10.1007/978-1-4612-0919-5_16].

[48] F. Livet, T. Charnock, D. Le Borgne and V. de Lapparent, *Catalog-free modeling of galaxy types in deep images. Massive dimensional reduction with neural networks*, *Astron. Astrophys.* **652** (2021) A62 [arXiv:2102.01086].

[49] M. Kilbinger, *Cosmology with cosmic shear observations: a review*, *Rept. Prog. Phys.* **78** (2015) 086901 [arXiv:1411.0115] [INSPIRE].

[50] D.J. Eisenstein and W. Hu, *Power spectra for cold dark matter and its variants*, *Astrophys. J.* **511** (1997) 5 [astro-ph/9710252] [INSPIRE].

[51] Euclid collaboration, *Euclid preparation. XV. Forecasting cosmological constraints for the Euclid and CMB joint analysis*, *Astron. Astrophys.* **657** (2022) A91 [arXiv:2106.08346] [INSPIRE].

[52] Y. Li et al., *pmwd: a differentiable cosmological particle-mesh N-body library*, arXiv:2211.09958.

[53] Planck collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6 [*Erratum ibid.* **652** (2021) C4] [arXiv:1807.06209] [INSPIRE].

[54] F. Simpson et al., *Enhancing the cosmic shear power spectrum*, *Mon. Not. Roy. Astron. Soc.* **456** (2016) 278 [arXiv:1507.04862] [INSPIRE].

[55] H.-J. Seo et al., *Re-capturing cosmic information*, *Astrophys. J. Lett.* **729** (2011) L11 [arXiv:1008.0349] [INSPIRE].

[56] B. Joachimi, A.N. Taylor and A. Kiessling, *Cosmological information in Gaussianised weak lensing signals*, *Mon. Not. Roy. Astron. Soc.* **418** (2011) 145 [arXiv:1104.1399] [INSPIRE].

[57] D.K. Ramanah, T. Charnock, F. Villaescusa-Navarro and B.D. Wandelt, *Super-resolution emulator of cosmological simulations using deep physical models*, *Mon. Not. Roy. Astron. Soc.* **495** (2020) 4227 [arXiv:2001.05519] [INSPIRE].

[58] T. Charnock et al., *Neural physical engines for inferring the halo mass distribution function*, *Mon. Not. Roy. Astron. Soc.* **494** (2020) 50 [arXiv:1909.06379] [INSPIRE].

[59] S. Ding, G. Lavaux and J. Jasche, *PineTree: a generative, fast, and differentiable halo model for wide-field galaxy surveys*, Astron. Astrophys. **690** (2024) A236 [arXiv:2407.01391] [INSPIRE].

[60] D. Lanzieri et al., *Optimal neural summarisation for full-field weak lensing cosmological implicit inference*, arXiv:2407.10877 [INSPIRE].

[61] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, *Inception-v4, Inception-ResNet and the impact of residual connections on learning*, arXiv:1602.07261 [INSPIRE].

[62] B. Xu, N. Wang, T. Chen and M. Li, *Empirical evaluation of rectified activations in convolutional network*, arXiv:1505.00853 [INSPIRE].

[63] M. Ho et al., *LtU-ILI: an all-in-one framework for implicit inference in astrophysics and cosmology*, Open J. Astrophys. **7** (2024) 001c.120559 [arXiv:2402.05137] [INSPIRE].

[64] D. Spergel et al., *Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 report*, arXiv:1503.03757 [INSPIRE].

[65] SIMBIG collaboration, *Field-level simulation-based inference of galaxy clustering with convolutional neural networks*, Phys. Rev. D **109** (2024) 083536 [arXiv:2310.15256] [INSPIRE].

[66] J. Bradbury et al., *JAX: composable transformations of Python+NumPy programs GitHub repository*, https://github.com/jax-ml/jax (2018).

[67] J. Heek et al., *Flax: a neural network library and ecosystem for JAX GitHub repository*, https://github.com/google/flax (2020).

[68] P. Lemos, A. Coogan, Y. Hezaveh and L. Perreault-Levasseur, *Sampling-based accuracy testing of posterior estimators for general inference*, *40th International Conference on Machine Learning* **202** (2023) 19256 [arXiv:2302.03026].