Predictive uncertainty on astrophysics recovery from multifield cosmology

Sambatra Andrianomena\textsuperscript{a,b,*} and Sultan Hassan\textsuperscript{c,b,d,1}

\begin{itemize}
\item \textsuperscript{a}South African Radio Astronomy Observatory (SARAO), Black River Park, Observatory, Cape Town, 7925, South Africa
\item \textsuperscript{b}Department of Physics, University of the Western Cape, Bellville, Cape Town 7535, South Africa
\item \textsuperscript{c}Center for Cosmology and Particle Physics, Department of physics, New York University, 726 Broadway, New York, NY 10003, U.S.A.
\item \textsuperscript{d}Flatiron Institute Center for Computational Astrophysics, 162 5th Ave 5th floor, New York, NY 10010, U.S.A.
\end{itemize}

E-mail: andrianomena@gmail.com, sultanier@gmail.com

Received August 19, 2022
Revised May 10, 2023
Accepted June 7, 2023
Published June 23, 2023

Abstract. We investigate how the constraints on the density parameter ($\Omega_m$), the power spectrum amplitude ($\sigma_8$) and the supernova feedback parameters ($A_{SN1}$ and $A_{SN2}$) vary when exploiting information from multiple fields in cosmology. We make use of a convolutional neural network to retrieve the salient features from different combinations of field maps from IllustrisTNG in the CAMELS project. The fields considered are neutral hydrogen (HI), gas density (Mgas), magnetic fields (B) and gas metallicity (Z). We estimate the predictive uncertainty — sum of the squares of aleatoric and epistemic uncertainties — of the parameters inferred by our model by using Monte Carlo dropout, a Bayesian approximation. Results show that in general, the performance of the model improves as the number of channels of its input is increased. In the best setup which includes all fields (four channel input, Mgas-HI-B-Z) the model achieves $R^2 > 0.96$ on all parameters. Similarly, we find that the predictive uncertainty, which is dominated by the aleatoric uncertainty, decreases as more fields are used to train the model in general. The uncertainties obtained by dropout variational inference are overestimated on all parameters in our case, in that the predictive uncertainty is much larger than the actual squared error, which is the square of the difference between the ground truth and prediction. After calibration, which consists of a simple $\sigma$ scaling method, the average deviation of the predictive uncertainty from the actual error goes down to 25% at most (on $A_{SN1}$).

Keywords: Bayesian reasoning, cosmological simulations, hydrodynamical simulations, Machine learning

ArXiv ePrint: 2208.08927

\textsuperscript{1}NHFP Hubble fellow.

*Corresponding author.
1 Introduction

In order to retrieve the relevant information from observables (e.g. galaxy survey, cosmic shear) for predicting underlying cosmology, summary statistics, such as power spectrum, have been used extensively (e.g. [1–12]). The regime in those studies is where the bias is less scale-dependent such that the theory (i.e. power spectrum) can still be leveraged to harvest the cosmological information. However, inferring the cosmological parameters in a highly non-linear regime, where the complex baryonic physics can not be neglected, is a non-trivial task. For instance, [13] demonstrated the increasing effect of baryons on weak lensing power spectrum on small scales such that precision cosmology on those scales will need to account for those effects. In addition, [14] suggested that an elaborate model of Active Galactic Nuclei (AGN) outflows is required to explain their effect on the matter power spectrum. These works show how small scale processes can affect fluctuations on cosmological scales. Nevertheless, it has been well established that the power spectrum is not an optimal summary statistics to capture information out of highly non-linear fields (e.g. two different fields could share the same power spectrum [15]). For this reason, some works have focused on the use of higher order statistics, such as bispectrum, to capture more information, including non-gaussianity. However, these higher order statistics are challenging to measure, compute, and interpret (e.g. [16–19]). The optimal way to prevent loss of information is to perform inference at the field level, where machine learning methods have been very successful to constraining cosmology and astrophysics (e.g see [20] for a recent review). These methods can directly build a mapping between the distribution of matter and the corresponding cosmology. After the ImageNet competition,\(^1\) it is well known now that the Convolutional Neural Networks (CNNs) achieve the state-of-the-art performance to extract most information from fields. This is due to their ability to simultaneously capture local features (high frequencies) as well as global features (low frequencies) out of highly non linear fields. For instance, [21] and [22] made use of CNNs to predict \(\Omega_m\) and \(\sigma_8\) from simulated 3D distribution of dark matter. They reported that their neural network achieved better accuracy in comparison with power spectral analyses. Ref. [23] and [24] utilized the predictive power of CNN to extract cosmological information from weak lensing maps, provided the wealth of information the latter.

\(^1\)https://www.image-net.org/challenges/LSVRC/.
carry. Ref. [25], who also resorted to deep learning to constrain cosmology using weak lensing maps, found that the small scale influence of baryons degraded the cosmological constraints.

However, more stringent cosmological and astrophysical constraints are expected from future HI surveys. HI intensity mapping has been proven to be a powerful cosmological probe with the advent of SKA experiment [26], since mapping out the HI distribution traces the underlying density and large scale structure. Few examples of CNN’s success in the context of HI maps are the following. Using HI 2D maps from SimFast21 [27–29], a semi-numerical model to evolve reionization on cosmological scales, [30] trained two different network architectures to infer the cosmological and astrophysical parameters from the HI field. Ref. [31] and [32] used similar network architectures to identify reionization sources, and to reconstruct the reionization history respectively. Using different semi-numerical model, [33] employed CNNs to learn galaxy properties from 21 cm lightcones, and [34] recovered the model astrophysical parameters. Ref. [35] painted HI on the matter field from N-body simulations. Ref. [36] removed astrophysical effects. Ref. [37] provided optimal summary statistics for simulation-based inference. Most of these works focused on HI maps during reionization, whereas ours is more focused on the local intensity maps, including HI.

Within the context of inferring the cosmological and astrophysical parameters directly from a field distribution (3D, 2D maps) using deep learning, uncertainty on the model predictions has not been carefully estimated, to the best of our knowledge, in that the models that have been considered only generate point estimates so far. In a real world scenario, the frequentist approach in deep learning is prone to overconfidence. Given its prediction which is point-wise estimate, assessing what the model is ignorant about (e.g. out-of-distribution instances) poses a challenge. Unlike Bayesian deep learning in which prediction is associated with uncertainty, wrong prediction corresponds to large uncertainty, reflecting what the model doesn’t know. In this study, we model the predictive uncertainty (also denoted as total uncertainty) of our network which results from the combination of the epistemic and aleatoric uncertainties. The former, also known as model uncertainty, encodes the uncertainty in the network weights and can be reduced by having more data for the training. The latter accounts for the intrinsic noise in the data (e.g. resolution) and cannot be decreased by having more examples in the training set. In our study, we turn our frequentist model into a Bayesian model by resorting to Monte Carlo Dropout (MCD). Ref. [38] demonstrated that a neural network with dropout layers and an approximation to a Bayesian model like Gaussian process are mathematically equivalent. MCD consists of enabling the dropout layers at test time and sampling predictions by running $T$ number of forward passes in order to estimate the uncertainties.

For the first time, [39] extracted cosmological and astrophysical parameters from multifield 2D maps, consisting of a set of fields where each one corresponds to a channel at the input of a deep regressor. They showed that their deep architecture, which consists of chaining up convolutional layers in seven different stages to extract the relevant features from the input, is capable of marginalizing over the contamination by the astrophysical effects at the field level to constrain both the cosmology and astrophysics with a relatively high accuracy. Interestingly, they found that the parameter constraints greatly improved using 11 channels at the input to train the network model compared to only using a single channel.

In this work, which is based on the results found in [39], we investigate the amount of information gained from considering different sets of fields when constraining cosmology and astrophysics. As a proof of concept, we restrict ourselves to a simple scenario with only four fields that [39] proved to be good. More importantly, our main contribution is to estimate
the predictive uncertainty and investigate how it varies as a function of the set of fields at the model input. The data used in this work and the network model we consider are presented in section 2. Results are shown in section 3 where the effects of the set of fields used for the training on both the model performance and the total uncertainty for each parameters are detailed in 3.1 and 3.2 respectively. The use of a scaling-based approach to mitigate the miscalibration of the uncertainties produced by the Bayesian network is discussed in 3.3. We finally conclude in section 4.

2 Data and model

The Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) project [40–42] consists of 4,233 numerical simulations which comprise dark matter only and state-of-the-art hydrodynamic simulations. The latter consider two different implementations of the subgrid physics models, namely, IllustrisTNG [43] and SIMBA [44]. CAMELS, whose main goal is to understand the theory that explains observables for a given cosmology and astrophysics, is specifically designed for training machine learning algorithms. The multifield dataset that we choose in this work is from IllustrisTNG simulation suite in the CAMELS project, and comprises 195000 2D maps as a whole, yielding 15000 maps for each of the thirteen fields (e.g. gas density, neutral hydrogen density). The maps were produced at $z = 0$ from varying cosmological ($\Omega_m, \sigma_8$), astrophysical parameters ($A_{SN1}, A_{SN2}, A_{AGN1}, A_{AGN2}$) and the initial seed number in thousands of simulation runs. For the stellar feedback processes, which drive the galactic winds and hence impact on the metallicity of the environment, $A_{SN1}$ controls the energy transfer and $A_{SN2}$ conditions the wind speed. Whereas $A_{AGN1}$ and $A_{AGN2}$ control the power injected and jet speed in AGN feedback processes respectively. It is noted that the implementation and the physical interpretation of these four astrophysical parameters are different in SIMBA. In the latter, $A_{SN1}$ directly modulates the mass loading factor and although the wind speed is characterized by $A_{SN2}$, the implementation is different from IllustrisTNG [40]. $A_{AGN1}$ and $A_{AGN2}$ condition the total momentum flux and maximum jet velocity respectively in SIMBA [40]. For easy reference, we present in table 1 the prior range of each parameter, and refer the interested reader to [40] for more details. Each 2D map of $25 \, h^{-1}\text{Mpc}$ in size has $256 \times 256$ pixels resolution. For our analyses we only select four fields, namely gas density ($M_{\text{gas}}$), neutral hydrogen density (HI), magnetic field (B) and gas metallicity (Z). The fields have been selected based on how sensitive they are to the variation of the parameters of interest in this study (see [39]). Figure 1 shows the four fields considered in this study in three different realizations which are related to three

| Parameter | Range |
|-----------|-------|
| $\Omega_m$ | [0.1, 0.5] |
| $\sigma_8$ | [0.6, 1.0] |
| $A_{SN1}$ | [0.25, 4] |
| $A_{SN2}$ | [0.5, 2] |
| $A_{AGN1}$ | [0.25, 4] |
| $A_{AGN2}$ | [0.5, 2] |

Table 1. Prior range of each parameter.
different sets of cosmological and astrophysical parameters, with different varying feedback strengths. The four fields — Z (first row), B (second row), HI (third row) and Mgas (fourth row) — in the same column correspond to one realization. As this set of parameters are highly degenerate (e.g. low $A_{SN2}$ and high $A_{AGN2}$ would boost star formation in the outskirts of galaxies and push gas to low dense regions, see figure 2 in [39]), it is challenging to determine the leading effect. However, the following combination of $\Omega_m = 0.296, \sigma_8 = 0.963, A_{SN1} = 0.261, A_{AGN1} = 0.865, A_{SN2} = 0.519, A_{AGN2} = 0.801$ (left column) shows the scenario where gas is being pushed out to less dense regions. The right column shows the combination of $\Omega_m = 0.263, \sigma_8 = 0.663, A_{SN1} = 3.536, A_{AGN1} = 0.577, A_{SN2} = 1.794, A_{AGN2} = 1.017$, where the gas is concentrated at the high dense regions. The middle column has a moderate feedback strength between the left and right columns with parameters $\Omega_m = 0.157, \sigma_8 = 0.842, A_{SN1} = 0.611, A_{AGN1} = 0.807, A_{SN2} = 0.887, A_{AGN2} = 1.002$.

In order to build a mapping between the input (2D maps) and the underlying cosmology and astrophysics, we resort to a deep convolutional network, whose architecture, which is quite similar to VGG-16 [45], is provided in table 2. The feature extractor consists of having 4 stages, where the output of each stage is downsampled using Maxpooling. Two convolutional layers are chained up in each of the first two stages (where the model learns large scales of the images), whereas three are used in each of the two last stages (where fine details on smaller scales are learned). The batch normalization layer after each convolutional layer is for regularization. The extracted features are then averaged and flattened before being fed into two dense layers for predictions. Compared to both VGG-16 and the architecture used in [39], our network is both shallower and narrower but its design is such that it has a good expressive power for this particular problem. Increasing the model capacity of our network, which slows down the training due to the higher number of network parameters, does not significantly improve the results. To estimate the predictive uncertainty via Monte Carlo Dropout [38] we have two dropout layers, each with probability of 0.2. The choice of the dropout rate is based on [38]. The first 6 components from the network outputs are the mean of parameters $\Omega_m, \sigma_8, A_{SN1}, A_{SN2}, A_{AGN1}, A_{AGN2}$ which we denote by $\mu_i \ (i = 1, \ldots, 6)$ and the remaining components are the standard deviation for each parameter $\theta_i$, denoted by $\sigma_i$. This construction is such that the aleatoric uncertainty, which is the other component of the total uncertainty, can be estimated. Following the prescription in [39], the optimization (using Adam optimizer) via mini-batch gradient descent is achieved by using the loss function

$$
\mathcal{L} = \frac{1}{N_{\text{param}}} \sum_{i=1}^{N_{\text{param}}} \log \left( \frac{1}{M_{\text{batch}}} \sum_{j=1}^{M_{\text{batch}}} (\theta_{i,j} - \mu_{i,j})^2 \right) + \frac{1}{N_{\text{param}}} \sum_{i=1}^{N_{\text{param}}} \log \left( \frac{1}{M_{\text{batch}}} \sum_{j=1}^{M_{\text{batch}}} \left( (\theta_{i,j} - \mu_{i,j})^2 - \sigma_{i,j}^2 \right)^2 \right),
$$

(2.1)

where $N_{\text{param}}, M_{\text{batch}}$ and $\theta_{i,j}$ are the number of parameters (6), batch size (64) and the ground truth respectively. The first term in Equation 2.1 is the mean squared error whereas the second term enforces the aleatoric uncertainty to be consistent with the squared error such that it is well calibrated.

We use 80% of the dataset for training which is run for 200 epochs with batch size of 64 and the remaining instances are used for validation and testing. In total, there are four training runs for four different setups.
$\Omega_m = 0.296, \sigma_8 = 0.963$  $\Omega_m = 0.157, \sigma_8 = 0.842$  $\Omega_m = 0.263, \sigma_8 = 0.663$

$A_{SN1} = 0.261, A_{AGN1} = 0.865$ $A_{SN1} = 0.611, A_{AGN1} = 0.807$ $A_{SN1} = 3.536, A_{AGN1} = 0.577$

$A_{SN2} = 0.519, A_{AGN2} = 0.801$ $A_{SN2} = 0.887, A_{AGN2} = 1.002$ $A_{SN2} = 1.794, A_{AGN2} = 1.017$

Figure 1. Each map of size $25 \, h^{-1}\text{Mpc}$ is an example of each field that is considered in this study. Mgas (fourth row), HI (third row), B (second row) and Z (first row) indicate gas density maps, neutral hydrogen density maps, magnetic fields maps and gas metallicity maps respectively. The maps in the same column correspond to the same realization. The first column, second column and third column correspond to weak, modest and strong stellar feedback.
| Layer                | (in channel, out channel, kernel, stride) |
|----------------------|------------------------------------------|
| 1 Batchnorm          | \((n, n, -, -)\)                         |
| 2 ConvConv           | \((n, 32, 3 \times 3, 1)\)               |
| 3 Maxpooling         | \((32, 32, 3 \times 3, 2)\)             |
| 4 ConvConv           | \((32, 64, 3 \times 3, 1)\)             |
| 5 Maxpooling         | \((64, 64, 3 \times 3, 2)\)             |
| 6 ConvConvConv       | \((64, 128, 3 \times 3, 1)\)            |
| 7 Maxpooling         | \((128, 128, 3 \times 3, 2)\)           |
| 8 ConvConvConv       | \((128, 256, 3 \times 3, 1)\)           |
| 9 AdaptiveAveragepooling | \((256, 256, -, 2)\)          |
| 10 Flatten           | –                                        |
| 11 Dropout           | –                                        |
| 12 Fully Connected Layer | \((1024, 512, -, -)\)         |
| 13 ReLU               | –                                        |
| 14 Dropout           | –                                        |
| 15 Fully Connected Layer | \((512, 12, -, -)\)           |

Table 2. The architecture of the network in this work. The letter \(n\) indicates the number of channels (or also maps) of each input, ConvConv and ConvConvConv denote chaining two and three Convolution+Batch Normalization+ReLU layers respectively.

3 Results

In this section, the performance of the model is assessed using coefficient of determination \(R^2\), a metric that indicates how well the model is capable of explaining the proportion of variance in the predicted parameters. The coefficient of determination is given by \(^2\)

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (\theta_{i}^{\text{true}} - \theta_{i}^{\text{predicted}})^2}{\sum_{i=1}^{n} (\theta_{i}^{\text{true}} - \bar{\theta}_{i}^{\text{true}})^2},
\]

where \(\theta_{i}^{\text{predicted}}, \theta_{i}^{\text{true}}\) and \(\bar{\theta}_{i}^{\text{true}}\) are the prediction, the ground truth value and the mean of the ground truth values in the test set respectively. Following [39], we also use the relative accuracy which is defined as \(\langle \delta \theta_i / \theta_{i}^{\text{predicted}} \rangle\), where \(\delta \theta_i = | \theta_{i}^{\text{true}} - \theta_{i}^{\text{predicted}} |\) and the ensemble average\(^3\) is computed over the test set for each parameter.

3.1 Performance

We show in figure 2 the predictions for all 6 parameters in each setup. The errorbars denote the absolute difference between ground truth and prediction, and the solid grey line is the

\(^2\theta\) denotes the ground truth in general but here we use the same notation to indicate both the true and predicted values with superscript “true” and “predicted” respectively.

\(^3\)Which is denoted by the angle brackets.
For better visualization, the plots in figure 2 are obtained by selecting random examples from the test set and their corresponding predictions by the network. It can be noticed how the performance of the model differs between setups, which are indicated by different markers.

Results suggest that inferring $A_{\text{AGN1}}$ and $A_{\text{AGN2}}$ are challenging for the model, as indicated by the predictions, with larger errors, fluctuating around the same value in all setups. We find that $R^2$ values are -0.05 for $A_{\text{AGN1}}$ and 0.50 for $A_{\text{AGN2}}$, indicative of weak correlation (or no correlation at all) between their actual values and predictions. This trend is consistent with what was found by [39] and is attributed to the fact that the maps are not very sensitive to the variation of $A_{\text{AGN1}}$ and $A_{\text{AGN2}}$ (see figure 7). Together with that, it can also be argued that the model capacity is such that it fails to retrieve the relevant information. Overall, the results presented in figure 2 are in good agreement with those obtained in [39]. For the rest of our analyses, we restrict ourselves to the four cosmological and astrophysical parameters of interest, i.e. $\Omega_m$, $\sigma_8$, $A_{\text{SN1}}$ and $A_{\text{SN2}}$.

We present in figure 3 the variation of $R^2$ and that of accuracy as a function of the type of inputson the left and right columns respectively for the parameters ($\Omega_m$, $\sigma_8$, $A_{\text{SN1}}$, $A_{\text{SN2}}$). For each metric, we plot the cosmology and astrophysics in different panels to clearly show how the metrics vary with the setups. In general, it is clear that the more information is fed into the model the better its performance is. This is evidenced by the increasing $R^2$ value (see figure 3 left column) with an increasing number of the input channels. The cosmological parameters, especially the matter density, appear to be less sensitive to the number of fields (or channels), whereas the model performance on the astrophysical parameters greatly improves with more fields stacked at the model input. The $R^2$ value related to $A_{\text{SN1}}$ goes from 0.5, when only HI maps are used as inputs, to 0.98 with the Mgas-HI-B-Z setup. $A_{\text{SN1}}$ proves to be more difficult than $A_{\text{SN2}}$ to extract overall. This can attributed to the fact that the latter controls the wind speed which influences the topology of the environment (as particles are dispersed) which translates into salient features in the maps. In that sense, the fields maps are more sensitive to $A_{\text{SN2}}$, compared to $A_{\text{SN1}}$. The variation of accuracy as a function of the setup is consistent with that of the coefficient of determination as indicated by the increase/decrease in $R^2$ values on the left panels of figure 3 which corresponds to decrease/increase of the accuracy on the right panels. In other words, the predictions are more accurate when more information is provided to the model. It appears that the model accuracy on $\Omega_m$ is best with the Mgas-HI setup, corresponding to the lowest relative error. But overall it’s a small fluctuation, provided that the difference between the best and worst accuracy is $\sim 0.4\%$, which is insignificant. We argue that the overall increase in performance of the regressor has begun to reach a plateau with the Mgas-HI-B-Z setup ($R^2 > 0.96$ for all parameters). Ref. [39] investigated the performance of their network when combining the 11 fields for the training. By comparing the accuracy of our network prediction with only four fields/channel inputs — $\langle \delta \theta / \theta \rangle [\%] = 2.2, 2.7, 4.5, 8.6$ for $\sigma_8$, $\Omega_m$, $A_{\text{SN2}}$, $A_{\text{SN1}}$ respectively — with theirs which was obtained from stacking eleven fields at the input — $\langle \delta \theta / \theta \rangle [\%] = 2.3, 2.5, 4.8, 13.0$ for $\sigma_8$, $\Omega_m$, $A_{\text{SN2}}$, $A_{\text{SN1}}$ respectively — the results are promising.

Ref. [46] demonstrated how their model, trained on maps of total mass density $M_{\text{tot}}$ (gas + dark matter + stars + black holes) generated from one simulation, was capable of inferring the cosmological parameters ($\Omega_m$, $\sigma_8$) of $M_{\text{tot}}$ maps from a completely different simulation to a relatively high accuracy. This demonstrates a good level of generalization of their network and points towards good model predictions on out-of-distribution sample in the context of cosmological inference. In order to understand the capacity of our network, we also train it
Figure 2. Predictions for each parameter. Each type of marker corresponds to predictions in one setup (e.g. 2D maps of neutral hydrogen density as inputs). The errorbars indicate the absolute difference between the target and network output ($\mu_i$). The network performs well on the cosmological ($\Omega_m, \sigma_8$) and stellar feedback ($A_{SN1}$ and $A_{SN2}$) parameters with Mgas-HI-B-Z setup. However constraining the AGN feedback parameters is challenging.

on Mtot maps from one simulation suite in CAMELS to predict $\Omega_m$ and $\sigma_8$ on Mtot maps from another simulation suite. Table 3 summarizes our results. The network is perfectly capable of predicting the cosmological parameters from out-of-distribution maps, i.e. trained
Figure 3. Performance of the model. Panels on the left column show the coefficient of determination for each parameter as a function of the type of input. Panel on the right column present the variation of the average relative accuracy for each parameter as function of the setup. The cosmological and astrophysical parameters are plotted separately for better visualization. It is clear that extracting the cosmology and astrophysics improves with more information (more fields are stacked) at the input.

Table 3. Comparing the performance of our model on out-of-distribution sample with its performance on in-distribution sample. Values in each cell are $R^2$ and the accuracy $\langle \delta \theta/\theta \rangle$ (in round brackets).
Figure 4. Distributions of predictive, epistemic and aleatoric uncertainties for each parameter in each setup. The panels in the same column are associated with one setup, and those on the same row are the results for one parameter. Each colored box and their whiskers denote the interquartile range (IQR) and minimum/maximum respectively. The dots represent the outliers in the distribution. The x-axis of each panel indicates the type of uncertainty whereas its y-axis denotes the value of the uncertainty related to a parameter. Overall, it can be seen, especially for the astrophysical parameters, that the uncertainties improve with the increasing number of stacked fields.

on IllustrisTNG maps and tested on SIMBA maps and vice versa. This is evidenced by both the relatively high values of $R^2 (> 0.9)$ and accuracy on all parameters for the two cases (IllustrisTNG — SIMBA and SIMBA — IllustrisTNG). Overall the values of the two metrics ($R^2$, $\langle \delta \theta / \theta \rangle$) in the case of out-of-distribution samples and in-distribution samples are comparable, which is indicative of a good generalization capability of our model in this specific setup, i.e. inferring cosmological parameters using Mtot maps.
3.2 Uncertainty

We build our network such that its outputs comprise the mean $\mu$ and standard deviation $\sigma$, giving a total number of 12 outputs. Following [47], the predictive uncertainty (also known as total uncertainty), resulting from adding the epistemic and aleatoric uncertainties in quadrature, is given by

$$
\sigma^2_{\text{predictive}} = \frac{1}{T} \sum_{t=1}^{T} \mu_t^2 - \left( \frac{1}{T} \sum_{t=1}^{T} \mu_t \right)^2 + \frac{1}{T} \sum_{t=1}^{T} \sigma_t^2,
$$

(3.2)

where $\mu_t$ and $\sigma_t$ are samples from $T$ number of forward passes which is set to 200 in our case. The first two terms in Equation 3.2 denote the epistemic uncertainty, whereas the last term is the aleatoric uncertainty.

Figure 4 shows the distributions of all types of uncertainty for each of the four parameters in each setup. All panels in the same row show the distributions in each setup for one

---

We mentioned in the previous section that $A_{\text{AGN1}}$ and $A_{\text{AGN2}}$ would be left out for the rest of the work. However that does not change the number of the outputs of our network, which always includes all 6 parameters.
| Parameter | relative change in the mean of total uncertainty [%] |
|-----------|-----------------------------------------------|
| $\Omega_m$ | 9.72                                          |
| $\sigma_8$ | -14.37                                        |
| $A_{SN1}$  | -69.50                                        |
| $A_{SN2}$  | -42.77                                        |

Table 4. Relative change in mean total uncertainty between one channel input (HI) to four channel input. Negative value means decrease in uncertainty, in other words tighter constraint on average, and positive value means weaker constraint on average.

parameter. Based on the median value of the distribution, results suggest that the aleatoric uncertainty is a larger contributor to the uncertainty budget (total uncertainty). The predictive, epistemic and aleatoric uncertainties corresponding to the cosmological parameters ($\Omega_m$, $\sigma_8$) are less sensitive to the amount of information provided to the network in that their change is marginal if any with the increasing number of channels of the input. However the positive effect of stacking more field maps for the training on the uncertainties is noticeable in the case of the astrophysical parameters ($A_{SN1}$, $A_{SN2}$). As the channel number is increased, both the IQR and the median value of each type of distribution get smaller for the astrophysical parameters. A more compact representation of the distribution of each type of uncertainty as a function of the setup is shown in figure 5. The latter presents the mean uncertainty, indicated by a marker, and the dispersion of uncertainty, denoted by shaded area, of each parameter in each setup. The cosmological and astrophysical parameters are plotted separately. The constraints get tighter with more fields used to train the model and the improvement is more significant for the astrophysical parameters, especially $A_{SN1}$. The mean predictive uncertainties (and also the mean aleatoric ones) related to the matter density appear to slightly increase with the number of channels at the model input. However, by taking the variability of the uncertainty into account, that increase is only marginal. Provided the dynamic range of each parameter, we further investigate the change in the mean value of predictive uncertainty ($\bar{\sigma}_{\text{setup}}$) by computing its relative increase/decrease from only using HI maps to using Mgas-HI-B-Z maps to train the model. Table 4 presents the relative change in mean of predictive uncertainty according to

$$\Delta \sigma_{\text{predictive}} = 100 \times \frac{\sigma_{\text{Mgas-HI-B-Z\ predictive}} - \sigma_{\text{HI\ predictive}}}{\sigma_{\text{HI\ predictive}}}. \quad (3.3)$$

The relative change in the mean predictive uncertainty is relatively small for the cosmological parameters. The constraints on the matter density parameter, with a corresponding $\Delta \sigma_{\text{predictive}} = 9.72\%$, seems marginally tighter with HI setup than with Mgas-HI-B-Z (table 4) on average, however the overall dispersion of the predictive uncertainty (see figure 5 bottom left panel) shows that resulting total uncertainty on $\Omega_m$ prediction is not really sensitive to the selected setup for the training. The constraints on $\sigma_8$, with a corresponding $\Delta \sigma_{\text{predictive}} = -14.37\%$, slightly improve on average with Mgas-HI-B-Z setup. Unlike with the case of $\Omega_m$, the downward trend (along with the channel number) of the predictive uncertainty on $\sigma_8$ is a bit more noticeable. Similar to the performance of the model when including all four fields for the training, the generated uncertainties on the cosmology seem to be less impacted by the information that the gas density, magnetic field and gas metallic-
ity carry. However, as indicated in table 4, the model has gained relevant extra information from the three other channels to significantly tighten the constraints on $A_{SN2}$ and $A_{SN1}$. The relatively large improvement in the constraints on the astrophysical parameters that control the stellar feedback ($A_{SN2}$, $A_{SN1}$) can be attributed mainly to the information added by the gas metallicity to which they are related. As a consistency check, we compute the relative change in the mean predictive uncertainty between using gas metallicity only as input and the Mgss-HI-B-Z setup and find an improvement by only 10.33% and 7.59% for the constraints on $A_{SN1}$ and $A_{SN2}$ on average respectively.

3.3 Calibrating uncertainty

Predictive uncertainty that is estimated using variational inference with MC Dropout is subject to miscalibration [48]. It is either underestimated or overestimated depending on the specific task at hand. The poorly calibrated uncertainty estimates can be explained by model misspecification which is characterized by the choice of a parameterized model that does not fully capture the data distribution $p(\theta|x)$ (where $\theta$ and $x$ are the target and input respectively). For instance, for regression problem, the likelihood is in general assumed to be Gaussian by considering mean squared error (MSE) as the loss function, which can be problematic if the assumption of the gaussianity in the conditional $p(\theta|x)$ does not hold in reality. Miscalibration of the uncertainties also results from the use of approximate inference [49] to estimate them, which is our case. In our analyses, we assess how well calibrated the uncertainties produced by the model are and use a simple approach to mitigate the calibration issue. In what follows, we further restrict ourselves to the Mgss-HI-B-Z setup, as it corresponds to the best constraints for all four parameters overall, but the same approach can be done for the other setups.

Ref. [48] prescribed the expected uncertainty calibration error (UCE) to evaluate the miscalibration of uncertainty for regression tasks in deep learning. To calculate UCE for a test set, the errors (between predictions and actual values) and their corresponding uncertainties are first binned and the mean value of each quantity within a given bin is computed. And the deviation of the estimated uncertainties from the errors is defined as the weighted average of the absolute difference between the mean error and mean uncertainty in all bins and given by

$$\text{UCE}[%] = \frac{1}{N} \sum_{m=1}^{M} \frac{|B_m|}{|B_m|} |\text{err}(B_m) - \text{uncert}(B_m)|,$$

where $N$ is the total number of instances in the test set, $M$ indicates the number of bins which is set to 10 in our case, $B_m$ denotes the set of instances with a given bin, $\text{err}(B_m)$ the error in each bin given by

$$\text{err}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} ||\mu_i - \theta_i||^2,$$

and the average uncertainty in each bin $\text{uncert}(B_m)$ is

$$\text{uncert}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} (\sigma_{\text{predictive}})_i.$$

A perfect calibration is when UCE = 0%, i.e. the error against the uncertainty plot lies on the identity line. It is worth noting that [50] and [51] opted for $M = 15$ in their analyses but
Figure 6. Calibration of the model predictive uncertainty for $\Omega$, $\sigma_8$, $A_{SN1}$ and $A_{SN2}$, considering the combination of four different 2D maps as input, i.e. Mgas-HI-B-Z setup. The values related to the cosmological parameters are in $10^{-3}$ for better visualization. The uncertainties are overestimated, and applying $\sigma$ scaling method greatly improves the uncertainty calibration. It is worth noting that the number of bins $M$ is equal to that of point. By choosing a larger value of $M$, it is possible that some bins are empty so that the resulting number of points is smaller than $M$.

| Parameter | UCE$_{\text{modified}}$[%], uncalibrated | UCE$_{\text{modified}}$[%], calibrated |
|-----------|-----------------------------------------|-------------------------------------|
| $\Omega_m$ | 575.04 | 18.25 |
| $\sigma_8$ | 155.90 | 15.81 |
| $A_{SN1}$ | 348.02 | 25.87 |
| $A_{SN2}$ | 250.42 | 12.81 |

Table 5. Uncertainty calibration.

we consider smaller bin number in order to capture the trend while avoiding having empty bins. We consider the weighted average of the relative difference between the uncertainty and the error, such that

$$UCE_{\text{modified}}[\%] = \frac{1}{M} \sum_{m=1}^{M} \left| \frac{B_m \cdot \text{err}(B_m) - \text{uncert}(B_m)}{\text{err}(B_m)} \right|.$$  \hspace{1cm} (3.7)

In other words, the relative deviation of the uncertainty with respect to the error in each bin is computed and the metric which is used to measure the miscalibration is the weighted average of that relative deviation. The main idea behind using $UCE_{\text{modified}}$ in Equation 3.7 is to be able to compare the miscalibration related to each parameter.

Bottom row in figure 6 shows the calibration plots for all parameters for uncalibrated uncertainty. On the same test set, we run five inferences, each with 200 forward passes, to investigate the variation of the uncertainties in each bin. Each red dot represents the mean
of average error

$$\text{err}(B_m) = \frac{1}{R} \sum_{r=1}^{R} \{\text{err}(B_m)\}_r$$

(3.8)

as a function the mean of the average uncertainty in a bin

$$\text{uncert}(B_m) = \frac{1}{R} \sum_{r=1}^{R} \{\text{uncert}(B_m)\}_r,$$

(3.9)

where $R$ is the number of runs. The blue shaded area denotes the variation in each bin and the black dashed line is the identity line. It is clear that the predictive uncertainty is poorly calibrated, as indicated by the values of UCE_{modified} in the second column of table 3.7 which are all well above 100%. It is overestimated as it is larger than the actual error in prediction in all bins.

There are various methods for calibrating uncertainty, ranging from involved ones such as building a mapping between the uncertainty and the error by using a multilayer perceptron [52] to simpler scaling based approach [48, 53]. As a simple illustration of mitigating miscalibration of uncertainties generated by our Bayesian network, we opt for $\sigma$ scaling [48] which is an approach that consists of computing a scalar $S$ that scales down the uncertainties. For each parameter, we have that [48]

$$S = \sqrt{\frac{1}{N_{\text{valid}}} \sum_{i=1}^{N_{\text{valid}}} \frac{||\mu_i - \theta_i||^2}{(\sigma^2_{\text{predictive}})_i}},$$

(3.10)

where $N_{\text{valid}}$ is the number of instances in the validation set which we use and find that $S = 0.38, 0.63, 0.48, 0.51$ for $\Omega_m, \sigma_8, A_{SN1}, A_{SN2}$ respectively. For calibration, the uncertainties produced by the model on the test set are then multiplied by $S$. Top row in figure 6 shows the calibrated uncertainties using $\sigma$ scaling. The resulting UCE_{modified} is given in the third column of table 5. Results shows that after calibration, the predictive uncertainty deviates on average by $\sim 25\%$ (on $A_{SN1}$) from the actual error at most and its smallest deviation is about $12\%$ (on $A_{SN2}$). As can be seen, the choice of a simple scaling based approach is suboptimal especially if the required accuracy of the parameter estimation is at the percent level like in precision cosmology. In that case, a more involved calibration approach like the one which consists of using neural network to predict the error from the uncertainty [52] could be a potential choice to improve the uncertainty estimation. However, it is on a case-by-case basis, as in some specific tasks a more powerful recalibration model (such as neural network) can be outperformed by a simple $\sigma$ scaling (see [48]).

Calibrated uncertainties will tell us how reliable the corresponding model predictions are in a real world scenario where the ground truth is not available. In other words, the prediction error (squared error) is expected to be of the same order of magnitude of the calibrated uncertainty. Thus, any prediction which corresponds to high uncertainty is erroneous and can then be discarded.

4 Conclusion

We have shown in this work, which is inspired from the results obtained in [39], the exploitation of information from combining different fields to better understand the underlying
cosmology and astrophysics by using Bayesian network. To this end, we have trained a deep convolutional neural network to retrieve the salient features from 2D multifield maps in order to constrain the cosmological ($\Omega_m$, $\sigma_8$) and astrophysical ($A_{SN1}$, $A_{SN2}$) parameters. In their work, [39] showed great improvements on the parameter constraints by stacking 11 fields (as opposed to only considering one field per training) at the input of their model. In the light of that, our goal in this work is twofold: 1) assessing the performance of the model as a function of a set of fields as model input, and 2) modeling the predictive uncertainty which plays a key role in a real world scenario where the ground truth is not available. We have opted for a simple scenario (based on the results in [39]) in which the fields considered are neutral hydrogen (HI), gas surface density ($M_{gas}$), magnetic field (B) and gas metallicity (Z) maps generated by the IllustrisTNG simulation suite in the CAMELS project. To investigate how the amount of information extracted from the fields impacts on both the performance of the model and the predictive uncertainty it produces, we have selected four different setups. In four separate training runs, we have selected: HI maps (HI), stacking Mgas and HI maps (Mgas-HI), stacking Mgas, HI and B maps (Mgas-HI-B), and stacking Mgas, HI, B and Z maps (Mgas-HI-B-Z) as inputs to the network. We have modeled the aleatoric uncertainty by having a mean ($\mu_i$) and standard deviation ($\sigma_i$) for each parameter at the output of our network. The total uncertainty is estimated by using MC dropout method, running 200 stochastic forward passes at inference time. We have also assessed the miscalibration of the uncertainties generated by our network and used a simple scaling based approach to mitigate it.

Overall, the network is capable of retrieving the salient features from the maps to predict the cosmological parameters and the astrophysical ones ($A_{SN1}$, $A_{SN2}$) which control the stellar feedback. The two parameters that encode the AGN feedback, however, pose a challenge for the model. This can be accounted for by either the model complexity or the fact that the fields don’t contain enough information for the mapping. The regressor only requires the amount of information contained in maps of neutral hydrogen to determine the matter density parameter to a good level of accuracy, as indicated by an $R^2 \sim 0.98$ and accuracy $\langle \delta \theta/\theta \rangle \sim 2.9$. This is expected since neutral hydrogen is a biased tracer of the underlying dark matter. Our finding comes to corroborate the results obtained in [39]. By increasing the number of channels (or fields) at the input of the network, the latter only marginally improves its performance on predicting $\Omega_m$, achieving $R^2 \sim 0.99$ and $\langle \delta \theta/\theta \rangle \sim 2.7\%$. The results however suggest that the performance of the network on predicting $\sigma_8$, $A_{SN1}$ and $A_{SN2}$ increases to varying degrees as more fields are considered for the training. The network gains great amount of information from including the magnetic fields and gas metallicity in the input so as to constrain $A_{SN1}$ and $A_{SN2}$, as was already shown by [39]. By combining all the fields considered in this work, the performance of the model on all four parameters ($\Omega_m$, $\sigma_8$, $A_{SN1}$ and $A_{SN2}$) is the best, as evidenced by $R^2 > 0.96$.

Based on [46] where, using total mass density field, it was shown that their neural network was capable of marginalizing the baryonic effects from SIMBA datasets although trained on IllustrisTNG (and vice versa), we have also assessed the capacity of our network by doing predictions on out-of-distribution sample and compare them with predictions on in-distribution test set. Using the same set of hyperparameters for the main setups in this work, we train the model with the total matter density $M_{tot}$ maps from IllustrisTNG and infer $\Omega_m$ and $\sigma_8$ on test sets from IllustrisTNG and SIMBA separately, and vice versa. Results show that the performance of the network on in-distribution sample is only marginally better than its performance on out-of-distribution sample. This points towards the fact that our model is able to attain a promising level of generalization, identifying the relevant features from the total matter density maps to predict the cosmology, with the amount of data used for training.
We find a larger contribution of the aleatoric uncertainty to the total uncertainty budget on average. In all setups, for all four parameters, the median (or also the mean) value of the aleatoric uncertainty distribution is higher than that of epistemic uncertainty. However, except with the case of $A_{\text{SN1}}$, the IQR of the latter is broader than the former in general. It is interesting to see that the epistemic uncertainty (encoding the model ignorance), which can be mitigated by using more data, decreases on average for the astrophysics by combining more fields, which does not augment the number of examples for the training but only provides more extra key features to the network. The decrease is more pronounced for $A_{\text{SN1}}$. This implies that, in our case, the effect of adding more channels is to help the model better understand regions of parameter space with few available data. This is the same effect of adding more data (more examples) in order to minimize the epistemic uncertainty in less explored regions. The aleatoric uncertainty also decreases but more rapidly with more channels for the parameters that encode the galactic winds; owing to the fact that the model is able to beat down the inherent noise (such as resolution) in the maps by exploiting the extra information each field provides. The constraints on $\Omega_m$ appear to be insensitive to having more channels at the input of the model. However the constraints on the amplitude of density fluctuations, the galactic winds parameters $A_{\text{SN1}}$ and $A_{\text{SN2}}$ in the Mgas-HI-B-Z setup improve by $\sim 14\%$, $\sim 69\%$ and $\sim 42\%$ compared to those in the HI setup respectively.

The predictive uncertainties on all parameters are all poorly calibrated. Results show that on average the relative deviation of the total uncertainty ($\text{UCE}_{\text{modified}}$) on all parameters are at least twice as big and can even reach $6 \times$ larger than the squared error on the setup that corresponds to the best constraints on all four parameters, i.e. Mgas-HI-B-Z. This indicates that the predictive uncertainty is overestimated. To remedy this miscalibration of the uncertainty, we adopt a simple scaling based approach which consists of computing the overall ratio $S$ between the squared error and the total uncertainty using the validation set. Calibration amounts to multiplying the uncertainties generated by the model when predicting the parameters of the test set by $S$. After calibration, the average deviation of the predictive uncertainty from the squared error goes down to about $25\%$ at most (on $A_{\text{SN1}}$). The best calibrated uncertainty, or the smallest $\text{UCE}_{\text{modified}}$, corresponds to $A_{\text{SN2}} (12\%)$.

Within the context of parameter inference in multifield cosmology, the results in this work look promising, but there are limitations that need to be highlighted. The maps used to train our network are noiseless, in that the systematics from instruments and observations like in a realistic scenario are not considered, e.g. foreground noise in HI intensity mapping. It is quite interesting to see how the performance of the network and the uncertainties will be impacted by the systematics and we defer this for future work. The field are selected based on the amount of information they carry about the cosmological and astrophysical parameters, as demonstrated in [39]. The objective is to investigate the improvement on the constraints by combining maps, but we have only considered four setups from a few selected fields as a first step. For future work, based on the results obtained in this work, it is possible to search for the optimized combination(s) that provide the maximum amount of relevant information that the model can exploit to further tighten the constraints. And the question: “what minimum subset of these fields enables us to get a fraction of those constraints?” [41] will also be addressed.

Acknowledgments

The authors acknowledge constructive comments from the anonymous referee which have improved the paper quality significantly. The authors are thankful to Francisco Villaescusa-
Navarro for the useful discussions. SA acknowledges financial support from the South African Radio Astronomy Observatory (SARAO). SH acknowledges support from Simons Foundation, and support for Program number HST-HF2-51507 provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, incorporated, under NASA contract NAS5-26555.

A Parameters degeneracy

We here attempt to discuss how sensitive our considered maps are with respect to variation in parameters. This might shed some light on the interpretability of the results. To do so, we make use of the 1P set in the CAMELS dataset. The 1P set refers to simulation runs where a single parameter is varied each time, while keeping the rest of parameters fixed. As a first order estimate, we choose the 1-dimensional probability density function (PDF) of our quantities, namely the magnetic field ($B$), metallicity $Z/Z_\odot$, gas mass $M_{\text{gas}}$, and neutral fraction $f_{\text{HI}}$, at the particle level (i.e. we directly use the provided quantities of gas particles from the snapshots). We show the PDFs of all quantities in figure 7 in different panels. From this figure alone, we can conclude the following:

• Stronger variations in PDFs are seen when cosmological ($\Omega_m, \sigma_8$) or stellar feedback parameters ($A_{\text{SN1}}, A_{\text{SN2}}$) are varied. This indicates that cosmological and stellar feedback parameters have the strongest impact on the four considered quantities. This explains the higher accuracy in parameter recovery as reported by the network (clear difference is easier to learn and distinguish).

• Almost no variation is seen when AGN feedback parameters ($A_{\text{AGN1}}, A_{\text{AGN2}}$) are varied. This explains why the network fails to recovery the AGN feedback parameters (degeneracy is more difficult to learn).

• Stronger PDF variation is observed for $f_{\text{HI}}$ when $\Omega_m$ is varied. This explains why HI map alone is able to constrain $\Omega_m$ with a similar accuracy when all four fields combined.

It is worth mentioning that some loss of local information is expected at the PDF level (when collapsing all particles in the box into a global 1D statistics), and hence quantifying parameter degeneracy might be more accurate at the field or grid level. For instance, the PDF variation in the case of $M_{\text{gas}}$ is minimal regardless of parameter variations. This shows that the 1D PDF does not capture enough (local) information to discriminate between different models/parameters. Nevertheless, the PDF statistic still captures some information that are somewhat illuminating as a first order metric.
Figure 7. Parameters variation impact on the distribution of metallicity, magnetic field, neutral fraction, and gas mass.
References

[1] M.M. Ivanov, M. Simonović and M. Zaldarriaga, Cosmological parameters from the BOSS galaxy power spectrum, *JCAP* 05 (2020) 042 [arXiv:1909.05277] [InSPIRE].

[2] BOSS collaboration, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological implications of the Fourier space wedges of the final sample, *Mon. Not. Roy. Astron. Soc.* 467 (2017) 2085 [arXiv:1607.03143] [InSPIRE].

[3] BOSS collaboration, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample, *Mon. Not. Roy. Astron. Soc.* 470 (2017) 2617 [arXiv:1607.03155] [InSPIRE].

[4] BOSS collaboration, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: on the measurement of growth rate using galaxy correlation functions, *Mon. Not. Roy. Astron. Soc.* 469 (2017) 1369 [arXiv:1607.03148] [InSPIRE].

[5] eBOSS collaboration, Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory, *Phys. Rev. D* 103 (2021) 083533 [arXiv:2007.08991] [InSPIRE].

[6] W.J. Percival et al., The shape of the SDSS DR5 galaxy power spectrum, *Astrophys. J.* 657 (2007) 645 [astro-ph/0608636] [InSPIRE].

[7] M.M. Ivanov, Cosmological constraints from the power spectrum of eBOSS emission line galaxies, *Phys. Rev. D* 104 (2021) 103514 [arXiv:2106.12580] [InSPIRE].

[8] M.M. Ivanov et al., Constraining early dark energy with large-scale structure, *Phys. Rev. D* 102 (2020) 103502 [arXiv:2006.11235] [InSPIRE].

[9] O.H.E. Philcox, M.M. Ivanov, M. Simonović and M. Zaldarriaga, Combining full-shape and BAO analyses of galaxy power spectra: a 1.6% CMB-independent constraint on $H_0$, *JCAP* 2020 (2020) 032 [arXiv:2002.04035].

[10] G. D’Amico, L. Senatore, P. Zhang and H. Zheng, The Hubble tension in light of the full-shape analysis of large-scale structure data, *JCAP* 05 (2021) 072 [arXiv:2006.12420] [InSPIRE].

[11] T. Colas et al., Efficient cosmological analysis of the SDSS/BOSS data from the effective field theory of large-scale structure, *JCAP* 06 (2020) 001 [arXiv:1909.07951] [InSPIRE].

[12] Y. Kobayashi, T. Nishimichi, M. Takada and H. Miyatake, Full-shape cosmology analysis of the SDSS-III BOSS galaxy power spectrum using an emulator-based halo model: a 5% determination of $\sigma_8$, *Phys. Rev. D* 105 (2022) 083517 [arXiv:2110.06969] [InSPIRE].

[13] Y.P. Jing et al., The influence of baryons on the clustering of matter and weak lensing surveys, *Astrophys. J. Lett.* 640 (2006) L119 [astro-ph/0512426] [InSPIRE].

[14] R. Levine and N.Y. Gnedin, AGN outflows and the matter power spectrum, *Astrophys. J. Lett.* 649 (2006) L57 [astro-ph/0604308] [InSPIRE].

[15] E. Sefusatti and R. Scoccimarro, Galaxy bias and halo-occupation numbers from large-scale clustering, *Phys. Rev. D* 71 (2005) 063001 [astro-ph/0412626] [InSPIRE].

[16] H. Gil-Marín et al., The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies, *Mon. Not. Roy. Astron. Soc.* 465 (2017) 1757 [arXiv:1606.00439] [InSPIRE].

[17] R. Scoccimarro, Fast estimators for redshift-space clustering, *Phys. Rev. D* 92 (2015) 083532 [arXiv:1506.02729] [InSPIRE].

[18] O.H.E. Philcox and D.J. Eisenstein, Computing the small-scale galaxy power spectrum and bispectrum in configuration-space, *Mon. Not. Roy. Astron. Soc.* 492 (2020) 1214 [arXiv:1912.01010] [InSPIRE].
[19] O.H.E. Philcox, A faster Fourier transform? Computing small-scale power spectra and bispectra for cosmological simulations in $O(N^2)$ time, *Mon. Not. Roy. Astron. Soc.* 501 (2021) 4004 [arXiv:2005.01739] [inSPIRE].

[20] M. Huertas-Company and F. Lanusse, The DAWES review 10: the impact of deep learning for the analysis of galaxy surveys, *Publ. Astron. Soc. Austral.* 40 (2023) e001 [arXiv:2210.01813] [inSPIRE].

[21] S. Ravanbakhsh et al., Estimating cosmological parameters from the dark matter distribution, in the proceedings of the International Conference on Machine Learning, (2016), p. 2407 [arXiv:1711.02033] [inSPIRE].

[22] S. Pan et al., Cosmological parameter estimation from large-scale structure deep learning, *Sci. China Phys. Mech. Astron.* 63 (2020) 110412 [arXiv:1908.10590] [inSPIRE].

[23] A. Gupta, J.M.Z. Matilla, D. Hsu and Z. Haiman, Non-Gaussian information from weak lensing data via deep learning, *Phys. Rev. D* 97 (2018) 103515 [arXiv:1802.01212] [inSPIRE].

[24] D.H. Ribli, B.A. Pataki and I. Csabai, An improved cosmological parameter inference scheme motivated by deep learning, *Nature Astron.* 3 (2019) 93 [arXiv:1806.05959] [inSPIRE].

[25] 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].

[26] M.G. Santos et al., Cosmology from a SKA HI intensity mapping survey, *PoS AASKA14* (2015) 019 [arXiv:1501.03989] [inSPIRE].

[27] M.G. Santos et al., Cosmic reionization and the 21 cm signal: comparison between an analytical model and a simulation, *Astrophys. J.* 689 (2008) 1 [arXiv:0708.2424] [inSPIRE].

[28] M.G. Santos et al., Fast and large volume simulations of the 21 cm signal from the reionization and pre-reionization epochs, *Mon. Not. Roy. Astron. Soc.* 406 (2010) 2421 [arXiv:0911.2219] [inSPIRE].

[29] S. Hassan, R. Davé, K. Finlator and M.G. Santos, Simulating the 21 cm signal from reionization including non-linear ionizations and inhomogeneous recombinations, *Mon. Not. Roy. Astron. Soc.* 457 (2016) 1550 [arXiv:1510.04280] [inSPIRE].

[30] S. Hassan, S. Andrianomena and C. Doughty, Constraining the astrophysics and cosmology from 21 cm tomography using deep learning with the SKA, *Mon. Not. Roy. Astron. Soc.* 494 (2020) 5761 [arXiv:1907.07787] [inSPIRE].

[31] S. Hassan, A. Liu, S. Kohn and P. La Plante, Identifying reionization sources from 21 cm maps using convolutional neural networks, *Mon. Not. Roy. Astron. Soc.* 483 (2019) 2524 [arXiv:1807.03317] [inSPIRE].

[32] T. Mangena, S. Hassan and M.G. Santos, Constraining the reionization history using deep learning from 21 cm tomography with the Square Kilometre Array, *Mon. Not. Roy. Astron. Soc.* 494 (2020) 600 [arXiv:2003.04905] [inSPIRE].

[33] D. Prelogović et al., Machine learning astrophysics from 21 cm lightcones: impact of network architectures and signal contamination, *Mon. Not. Roy. Astron. Soc.* 509 (2021) 3852 [arXiv:2107.00018] [inSPIRE].

[34] N. Gillet et al., Deep learning from 21 cm tomography of the cosmic dawn and reionization, *Mon. Not. Roy. Astron. Soc.* 484 (2019) 282 [arXiv:1805.02699] [inSPIRE].

[35] D. Wadekar, F. Villaescusa-Navarro, S. Ho and L. Perreault-Levasseur, HInet: generating neutral hydrogen from dark matter with neural networks, *Astrophys. J.* 916 (2021) 42 [arXiv:2007.10340] [inSPIRE].

[36] P. Villanueva-Domingo and F. Villaescusa-Navarro, Removing astrophysics in 21 cm maps with neural networks, *Astrophys. J.* 907 (2021) 44 [arXiv:2006.14305] [inSPIRE].
X. Zhao, Y. Mao, C. Cheng and B.D. Wandelt, *Simulation-based inference of reionization parameters from 3D tomographic 21 cm light-cone images*, Astrophys. J. **926** (2022) 151 [arXiv:2105.03344] [arXiv:2105.03344].

Y. Gal and Z. Ghahramani, *Dropout as a bayesian approximation: representing model uncertainty in deep learning*, in the proceedings of the international conference on machine learning, (2016), p. 1050.

F. Villaescusa-Navarro et al., *Multifield cosmology with artificial intelligence*, arXiv:2109.09747 [arXiv:2109.09747].

CAMELS collaboration, *The CAMELS project: Cosmology and Astrophysics with Machine Learning Simulations*, Astrophys. J. **915** (2021) 71 [arXiv:2010.00619] [arXiv:2010.00619].

CAMELS collaboration, *The CAMELS multifield data set: learning the universe’s fundamental parameters with artificial intelligence*, Astrophys. J. Supp. **259** (2022) 61 [arXiv:2109.10915] [arXiv:2109.10915].

CAMELS collaboration, *The CAMELS project: public data release*, Astrophys. J. Suppl. **265** (2023) 54 [arXiv:2201.01300] [arXiv:2201.01300].

D. Nelson et al., *The IllustrisTNG simulations: public data release*, Comput. Astrophys. Cosmol. **6** (2019) 1.

R. Davé et al., *Simba: cosmological simulations with black hole growth and feedback*, Mon. Not. Roy. Astron. Soc. **486** (2019) 2827 [arXiv:1901.10203] [arXiv:1901.10203].

K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv:1409.1556 [arXiv:1409.1556].

F. Villaescusa-Navarro et al., *Robust marginalization of baryonic effects for cosmological inference at the field level*, arXiv:2109.10360 [arXiv:2109.10360].

A. Kendall and Y. Gal, *What uncertainties do we need in bayesian deep learning for computer vision?*, Adv. Neural Inf. Process. Syst. **30** (2017) 1. [arXiv:1703.04977].

M.-H. Laves et al., *Well-calibrated regression uncertainty in medical imaging with deep learning*, in the proceedings of the Medical imaging with deep learning, (2020), p. 393.

M.R. Cervera et al., *Uncertainty estimation under model misspecification in neural network regression*, arXiv:2111.11763.

C. Guo, G. Pleiss, Y. Sun and K.Q. Weinberger, *On calibration of modern neural networks*, in the proceedings of the International conference on machine learning, (2017), p. 1321.

M.-H. Laves, S. Ihler, K.-P. Kortmann and T. Ortmayer, *Well-calibrated model uncertainty with temperature scaling for dropout variational inference*, arXiv:1909.13550.

V. Kuleshov, N. Fenner and S. Ermon, *Accurate uncertainties for deep learning using calibrated regression*, in the proceedings of the International conference on machine learning, (2018), p. 2796.

D. Levi, L. Gispan, N. Giladi and E. Fetaya, *Evaluating and calibrating uncertainty prediction in regression tasks*, arXiv:1905.11659.