

main advantages of BNNs are avoiding overfitting and providing a measure of uncertainty for the predictions. If we pass a sample repeatedly through the network we obtain a distribution instead of a single prediction for the classification result.

BaCoN is based on the architecture of convolutional neural networks (CNNs) so it combines Bayesian inference with image analysis. A CNN does not fully connect all the nodes of adjacent layers. A convolutional layer convolves the image layer with a smaller set of filters and the recognised features are saved in activation maps.

For BaCoN we use a matter power spectrum at four redshifts as the input image. The network performs its classification based on visual features in the spectrum. The patterns recognised in the convolutional layers are in our case shapes in the power spectrum. This makes the substructure of the graph more important than the overall amplitude. Hence, our choice of network architecture influences what the network is most sensitive to, which will differ from an established cosmological MCMC analyses. This also alters our approach to the noise model which we will discuss in [subsection 4.1](#). In general, it had been expected that CNNs do not keep track of the position of a structure in the input image, but it was shown recently by [Islam et al. \(2019\)](#) that the convolutional layers may implicitly learn absolute position information. This would explain the high accuracy we obtain when analysing spectral data. Furthermore, [Zhou et al. \(2016\)](#) have developed Class Activation Maps (CAMs) that are able to visualise the areas of the image that were most influential on the classification decision of a CNN. [Zhong et al. \(2022\)](#) have shown that they can be deployed successfully for spectra retrieving position information of class-specific features in the input.

Our network architecture ends with a final fully connected layer to be able to output probabilities for 5 classes. The actual classification is decided by a probability threshold. If one class has a probability above 50 % the spectrum counts as being classified. The final output for a test of a trained BaCoN network is a confusion matrix which shows the percentages of all correctly, wrongly and not classified spectra. This is ideally a fully diagonal matrix and we can use the off-diagonal entries to interpret degeneracies between classes. The contribution to true positives, false positives, true negatives and false negatives are broken down by the specific classes. We evaluate all our networks using the full confusion matrix for the test data but we will sometimes only state the total test accuracy (true positive rate over all classes) for conciseness.

The architecture and machine learning techniques used here are the same as in our previous work. We point the reader to [Mancarella et al. \(2020\)](#) for the details.

## 2.2 Halo model reaction

To create our training data we use the approach of [Cataneo et al. \(2019\)](#), which models the non-linear power spectrum as

$$P_{\text{NL}}(k, z) = \mathcal{R}(k, z) P_{\text{pseudo}}(k, z). \quad (1)$$

$\mathcal{R}$  is the halo model reaction function and quantifies corrections to the pseudo spectrum coming from the non-linear, non-standard physics in the beyond- $\Lambda$ CDM universe. We refer the reader to [Bose et al. \(2021\)](#) for the analytic formulae on which this function is based. It can be efficiently computed using the publicly available ReACT code ([Bose et al. 2020, 2022, 2023](#)). The accuracy of this function was found to be at the 1%-level in [Cataneo et al. \(2019\)](#) for  $k \leq 1 \text{ h/Mpc}$  for both modified gravity and dynamical dark energy models.

$P_{\text{pseudo}}(k, z)$  is called the pseudo power spectrum and is defined as a non-linear spectrum in a  $\Lambda$ CDM universe but whose initial conditions are tuned so that the linear clustering matches the beyond-

$\Lambda$ CDM universe at a given redshift,  $z$ . The purpose of using the pseudo is to ensure the halo mass functions in both beyond- $\Lambda$ CDM and pseudo universes are similar, which gives a better transition between linear and non-linear regimes. This quantity can be modelled using non-linear formulas such as HMcode ([Mead et al. 2015, 2016b, 2020](#)) or halofit ([Takahashi et al. 2012](#)), which require the specification of a linear power spectrum, allowing the user to provide the modified linear clustering while keeping the non-linear clustering based on  $\Lambda$ CDM physics. The drawback is that these fitting formulae introduce a significant inaccuracy in the calculation. These inaccuracies have been qualified at 5% in both cases for  $k \leq 1 \text{ h/Mpc}$ , but with HMcode typically achieving a higher  $\sim 2.5\%$  accuracy for most cosmologies ([Mead et al. 2020; Takahashi et al. 2012](#)). These inaccuracies both dominate the error budget of [Equation 1](#).

In [Mancarella et al. \(2020\)](#), the authors used the halofit formula to generate the training set, and account for theoretical uncertainties using a constant systematic. This was highly underestimated as we highlight in [subsection 4.1](#). In this work we improve upon the ‘old’ training data by using two, improved prescriptions for our predictions.

The first employs the improved HMcode formula of [Mead et al. \(2020\)](#)

$$P_{\text{NL}}^{\text{HMcode2020}}(k, z) = \mathcal{R}(k, z) P_{\text{pseudo}}^{\text{HMcode2020}}(k, z). \quad (2)$$

We expect this prediction to be  $\sim 6\%$  accurate for  $k \leq 1 \text{ h/Mpc}$  as we consider a wide range of cosmologies. In particular, the pseudo cosmologies may have a fairly large amplitude of linear clustering due to enhancements from modified gravity forces. This will affect the accuracy of the pseudo as noted in [Atayde et al. \(2024\); Tsedrik et al. \(2024\)](#). This accuracy increases to  $\geq 9\%$  for  $k \leq 3 \text{ h/Mpc}$  as estimated from [Cataneo et al. \(2019\)](#), which did not include massive neutrinos.

The second also employs HMcode, but as a ratio of modified-to- $\Lambda$ CDM predictions, i.e., a boost. This effectively factors out some of the inaccuracy inherent in the HMcode prediction. To get  $P_{\text{NL}}$  we then multiply the HMcode-based boost with a highly accurate prediction for  $P_{\text{NL}}^{\Lambda\text{CDM}}$ . This is available via the sophisticated  $\Lambda$ CDM power spectrum emulator, EuclidEmulator2 ([Knabenhans et al. 2021](#)) (EE2), which efficiently emulates  $N$ -body predictions for  $\Lambda$ CDM cosmologies. This ‘optimal’ accuracy version for the spectra predictions is then given as

$$P_{\text{NL}}^{\text{EE2}}(k, z) = B^{\text{HMcode2020}}(k, z) \times P_{\Lambda\text{CDM}}^{\text{EE2}}, \quad (3)$$

where

$$B^{\text{HMcode2020}}(k, z) \equiv \frac{\mathcal{R}(k, z) P_{\text{pseudo}}^{\text{HMcode2020}}(k, z)}{P_{\Lambda\text{CDM}}^{\text{HMcode2020}}(k, z)}. \quad (4)$$

EE2 is  $\sim 1\%$  accurate for  $k \leq 10 \text{ h/Mpc}$  for  $\Lambda$ CDM cosmologies. The HMcode boost was found to have an accuracy of  $\sim 2\%$  for a range of gravitational and dark energy theories including massive neutrinos, for  $k \leq 1 \text{ h/Mpc}$  ([Bose et al. 2021](#)). This gives us an estimated accuracy of 3 – 4% for  $k \leq 1 \text{ h/Mpc}$ . This degrades to  $\sim 6\%$  for  $k \leq 3 \text{ h/Mpc}$  ([Bose et al. 2021](#)). From these references we estimate an accuracy of 5% for  $k \leq 2.5 \text{ h/Mpc}$  which will be used in [section 4](#).

In this work we consider both [Equation 2](#) and [Equation 3](#) to train our network, and importantly use both of these predictions to calibrate the theoretical error assumed in the network’s predictions. The effects of massive neutrinos and beyond- $\Lambda$ CDM physics are primarily included in  $\mathcal{R}$ , but also in the linear spectrum that goes into  $P_{\text{pseudo}}$ . We also look to include the effects of baryonic physics, which is known to greatly affect the matter power spectrum at non-linear scales ([Schneider et al. 2019, 2020b,a](#)). This can now easily

be included via  $p_{\text{pseudo}}^{\text{HMcode2020}}$  through the single-parameter baryonic feedback modelling available within HMcode (Mead et al. 2020). This is a less comprehensive modelling of feedback processes than the more sophisticated emulators such as those of Aricò et al. (2021); Giri & Schneider (2021), but will serve as a simple first test of our network's capacity to distinguish between baryonic physics, massive neutrinos and non-standard physics. We leave more sophisticated feedback modelling to a future work.

### 2.3 Data Sets

We will consider 5 classes of cosmological and gravitational models, following Mancarella et al. (2020):

- (i)  $\Lambda$ CDM which assumes general relativity as the underlying gravitational model.
- (ii) The Hu-Sawicki  $f(R)$  gravity model (Hu & Sawicki 2007), parametrised by the value of the additional scalar field today,  $f_{R0}$ . This model exhibits the Chameleon screening mechanism (Khoury & Weltman 2004). We assume a  $\Lambda$ CDM background for this model.
- (iii) The Dvali-Gabadadze-Porrati (DGP) brane-world model (Dvali et al. 2000), parametrised by  $\Omega_{rc} = 1/(4r_c^2 H_0^2)$ , where  $H_0$  is the Hubble constant and  $r_c$  is a scale governing where gravitational interactions begin to dilute into the 5-dimensional bulk. This model exhibits the Vainshtein screening mechanism (Vainshtein 1972). We assume a  $\Lambda$ CDM background for this model.
- (iv) An evolving dark energy model as parameterised in Chevalier & Polarski (2001); Linder (2003) ( $w$ CDM), with the parameter pair  $\{w_0, w_a\}$  giving the value of the dark energy equation of state today and its time evolution respectively, where  $w(a) = w_0 + (1 - w_0)a$ ,  $a$  being the scale factor of the FLRW metric. Here the background is distinct from  $\Lambda$ CDM.
- (v) A random class as considered in Mancarella et al. (2020), but with slightly different settings as described in Sec. A. This class aims to capture any unknown and unconsidered models of gravity or energy whose phenomenology is largely distinct from the other classes.

For each of these scenarios we consider different sets of baseline  $\Lambda$ CDM cosmological parameters,  $\{\Omega_m, \Omega_b, H_0, n_s, A_s\}$  - the total matter fraction, the total baryonic matter fraction, the Hubble constant, the spectral index and the primordial amplitude of perturbations respectively. In addition to these, we also include the effects of massive neutrinos, parametrised by the sum of the neutrino masses  $\sum m_\nu$ , as well as baryonic feedback effects, parametrised by the  $T_{\text{AGN}}$  parameter of HMcode.

We sample these parameters to generate large sets of power spectra for each scenario. The parameters are sampled from Gaussian distributions with the following means and standard deviations:

$\{\Omega_m, \Omega_b, H_0, n_s, A_s\}$  are sampled using the Planck 2018 (Aghanim et al. 2020b) best fits as a mean and we take the standard deviation forecasted for the combination of weak lensing and spectroscopic clustering probes of the Euclid mission (Blanchard et al. 2020a). When using Equation 3 we also impose the hard EE2 priors, which particularly restrict  $\Omega_b \in [0.04, 0.06]$ .

$\{\Omega_{rc}, f_{R0}\}$  are sampled with their  $\Lambda$ CDM-limits as a mean ( $\Omega_{rc} = f_{R0} = 0$ ) and a standard deviation taken to be the  $3\sigma$  constraint forecasted for an LSST-like survey in Bose et al. (2020) using only linear scales.

$\{w_0, w_a\}$  are sampled using their  $\Lambda$ CDM-limits ( $\{w_0, w_a\} = \{-1, 0\}$ ) and the standard deviation is taken to be the value forecasted for the combination of weak lensing and spectroscopic clus-





tering probes of the Euclid mission (Blanchard et al. 2020a). We also impose the following hard limits to ensure stability of the ReACT code,  $w_0 \in [-1.3, -0.7]$  and  $w_a \in [-1.5, 0.3]$ .


$\sum m_\nu$  is taken to have the same standard deviation and mean as the fiducial value assumed in Blanchard et al. (2020a), a lower bound estimate based on observations from neutrino oscillation experiments (Esteban et al. 2020). In parameter files we quote the massive neutrino energy density fraction today,  $\Omega_\nu$ , with  $\sum m_\nu = \Omega_\nu h^2 93.14$  eV.

$\log_{10}[T_{\text{AGN}}]$  is taken to have the default mean value of HMcode, 7.8, and a standard deviation covering the fits to the BAHAMAS simulations (McCarthy et al. 2017) given in Mead et al. (2020).

These choices aim to give an estimate of what the BNN can achieve given the data from forthcoming galaxy surveys such as Euclid and LSST, while remaining consistent with the Planck CMB observations. The parameter ranges are summarised in Table 1. We demonstrate the effects of dynamical dark energy, modified gravity, massive neutrinos and baryonic feedback on the matter power spectrum in section B. Figure B1 shows the characteristic changes of  $P_{\text{NL}}$  when the model parameter is varied, keeping the baseline cosmological parameters constant.

We generate power spectra for all these scenarios using the public codes (see Equation 2 and Equation 3):

- (i)  $\mathcal{R}$  is computed using ReACT .
- (ii)  $p_{\text{pseudo}}^{\text{HMcode2020}}$ ,  $B^{\text{HMcode2020}}$  and the baryonic boost are computed using HMcode .
- (iii)  $P_{\text{ACDM}}^{\text{EE2}}$  is computed using EE2 .
- (iv) The modified linear spectra with massive neutrinos are computed using MGCAMB  for the  $\Lambda$ CDM,  $f(R)$  and  $w$ CDM scenarios.
- (v) The DGP linear spectra were generated using a private, modified version of CLASS (Lesgourgues 2011; Blas et al. 2011) which was also employed in Frusciante et al. (2023).

On the repository  we include the pipelines used to generate spectra for all scenarios and using both Equation 2 and Equation 3.

Each power spectrum data file is generated using a parameter set sampled from the ranges detailed in Table 1. These data files, used for both training and testing of the network, comprise of a 5 columns  $\times$  500 rows. The first column is simply the values of the Fourier mode sampled in  $h/\text{Mpc}$ , where we sample logarithmically in the range  $[0.01, 10] h/\text{Mpc}$ . The 2nd to 5th columns are the values of the power spectrum at those Fourier modes, with each column corresponding to the following redshifts  $z \in \{1.5, 0.785, 0.478, 0.1\}$ . These redshifts are chosen to roughly sample the range of tomographic bins Euclid's weak lensing survey will be making measurements at (Laureijs et al. 2011; Blanchard et al. 2020b), with an omission of the highest bins which can go beyond  $z = 2$ . We do not expect strong beyond- $\Lambda$ CDM effects at high redshift, but aim to consider this in a more comprehensive future iteration of the network where we train directly on the observables and not the matter power spectrum.

The complete list of data sets available here is summarised in Table 2.

## 3 TRAINING THE NETWORK

The whole process from data generation to training and testing the classification network is displayed as a scheme in Figure 1. We will describe all the involved stages in this section. Some steps are marked with blue boxes in the graphic and are discussed in the following

**Table 1.** Parameter ranges for new BaCoN data. These are sampled assuming a Gaussian distribution with means and standard deviations given below. The fiducial (mean) cosmology (without baryons but with massive neutrinos) gives a  $\sigma_8(z=0) = 0.812$ , which is the Planck 2018 best fit. We normalise with a  $\Lambda$ CDM spectrum with  $\sum m_\nu = 0$ , where we use  $A_s = 2.025 \times 10^{-9}$  to get  $\sigma_8 = 0.812$ . Hard limits are placed on  $w_0$  and  $w_a$  as described in the main text to ensure stability of ReACT. When using EE2,  $\Omega_b$  is also restricted to the emulator range of  $[0.04, 0.06]$ .

Parameter	Mean	Std. Dev.	Reference & Notes
$\Omega_m$	0.3158	0.009	Table. 1 of <a href="#">Aghanim et al. (2020b)</a> (Plik, best fit) & Table. 9 of <a href="#">Blanchard et al. (2020b)</a> (GC <sub>s</sub> + WL, pessimistic)
$\Omega_b$	0.0494	0.016	We use $\sigma_{\text{Planck}}/2$ as large $\Omega_b$ leads to computational issues in ReACT.
$H_0$	67.32	0.41	
$n_s$	0.966	0.007	
$A_s$	$2.199 \times 10^{-9}$	$2.199 \times 10^{-11}$	Mean cosmology corresponds to $\sigma_8 = 0.812$ . We convert % error on $\sigma_8$ to $A_s$ .
$ f_{R0} $	$10^{-10}$	$10^{-5.5}$	$\Lambda$ CDM-limit mean and $3\sigma$ , $\ell_{\text{max}} = 500$ from Table. 2 of <a href="#">Bose et al. (2020)</a>
$\Omega_{\text{rc}}$	$10^{-10}$	0.173	$\Lambda$ CDM-limit mean and $3\sigma$ , $\ell_{\text{max}} = 500$ taken from chains of <a href="#">Bose et al. (2020)</a>
$\{w_0, w_a\}$	$\{-1, 0\}$	$\{0.097, 0.32\}$	$\Lambda$ CDM-limit mean and & Table. 11 of <a href="#">Blanchard et al. (2020b)</a> (GC <sub>s</sub> + WL, pessimistic)
$\sum m_\nu$	0.06	0.06	Fiducial of <a href="#">Blanchard et al. (2020b)</a> and take same for Std. Dev.
$\log_{10}[T_{\text{AGN}}]$	7.8	0.2	Range of BAHAMAS simulation from Table. 4 of <a href="#">Mead et al. (2020)</a> .

**Table 2.** Data sets available ([here](#)) and used in this work.

P(k) prescription	Classes	Set size	Notes
HMcode-based	All	20,000	Main training set
HMcode-based	All	1,000	Main test set
EE2-based	All	20,000	Main training set
EE2-based	All	1,000	Main test set
EE2-based	$\Lambda$ CDM	20,000	No baryons or massive neutrinos
EE2-based	All	1,000	No baryons or massive neutrinos
halofit-based with halofit	All	19,900	No baryons or massive neutrinos, used in <a href="#">Mancarella et al. (2020)</a>
EE2-based (with halofit?)	All	19,900	No baryons or massive neutrinos

subsections. We will go through the specific influences of these marked parts in the result section.

### 3.1 Data Generation

The theoretical background of the data generation has been described in [subsection 2.2](#). Here, we will only address the practical choices. We start by drawing cosmological parameters from the priors for  $\Lambda$ CDM. This set of parameters is then fed into a code to produce a non-linear  $\Lambda$ CDM matter power spectrum (either EE2, or HMcode, or halofit). Then we choose the cosmological class that we want to generate matter power spectra for. This can be  $f(R)$ , DGP,  $w$ CDM or  $\Lambda$ CDM and draw from the class-specific parameter distribution. If we want to include effects of massive neutrinos and baryonic feedback, we include their respective parameters as well. All parameters are then passed on to ReACT and HMcode to calculate the boost. Finally, this is combined with the  $\Lambda$ CDM spectrum to a non-linear matter power spectrum for the selected class.

### 3.2 Noise Model

We normalise the power spectra data with a generic matter power spectrum to reduce the dynamic range of the data presented to the BaCoN network. Throughout this work, we use the same  $\Lambda$ CDM normalisation spectrum. We have tested the influence of different normalisation spectra including a linear matter power spectrum without Baryonic Acoustic Oscillations (BAOs). There is no noticeable effect as long as the same normalisation is used for the training and the testing phases. We set the default normalisation spectrum as an EE2  $\Lambda$ CDMspectrum with the Planck cosmology ([Aghanim et al. 2020b](#)) without baryonic feedback or massive neutrinos.

For every normalised spectrum we produce 10 realisations sampled from our *noise model*. This model has two components:

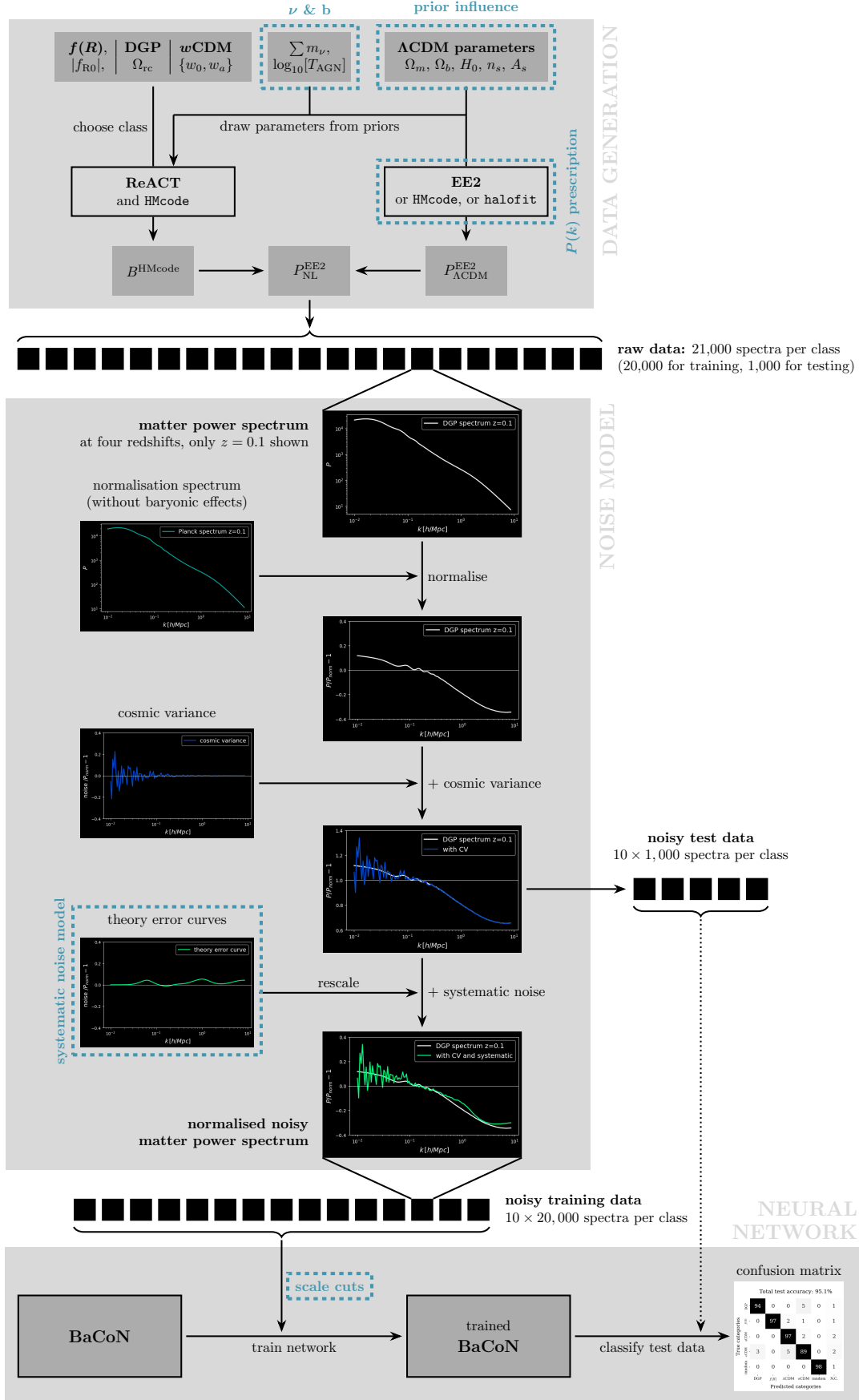
The first component represents noise coming from a Stage IV-like survey's cosmic variance. This component is given by

$$\sigma_p(k) = \sqrt{\frac{4\pi^2}{k^2 \Delta k V(z)}} \times P(k), \quad (5)$$

where  $V(z)$  is the volume probed at a given redshift and  $\Delta k$  is the bin-width. The Stage IV-like volumes we adopt are:  $V(1.5) = 10.43 \text{ Gpc}^3/h^3$ ,  $V(0.785) = 6.27 \text{ Gpc}^3/h^3$ ,  $V(0.478) = 3.34 \text{ Gpc}^3/h^3$  and  $V(0.1) = 0.283 \text{ Gpc}^3/h^3$  ([Laureijs et al. 2011](#); [Blanchard et al. 2020b](#)). We do not consider shot noise as we are using dark matter spectra, whose associated particle density is very large.

The second component is used to model the theoretical uncertainty in our power spectra predictions. As stated in [subsection 2.2](#), all terms used in the power spectra model have associated inaccuracies that contribute to the overall theoretical error of our non-linear matter power spectra predictions. This becomes very apparent when we look at the difference of power spectra predictions based on EE2 and HMcode as displayed in [Figure 2](#). Both codes produce characteristic fingerprints in the spectrum and deviate from each other by up to 4%. To avoid a fitting of the neural network to these prescription-specific errors, we generate curves with similar features to model such theoretical errors in the training process. The development of our theoretical error models will be discussed in detail in [subsection 4.1](#). These theory error curves are rescaled with a factor which we treat as a parameter to vary, and which accounts for potential errors of different amplitudes.

Note that while we add cosmic variance to the test data before passing through the network, we not add the theoretical error component. This is because the test data should be treated the same way as observational data that will be passed through the network for classification.



**Figure 1.** Flowchart of the full pipeline from data generation to the final classification result. The noise model is demonstrated on a DGP spectrum. See [section 3](#) for a detailed explanation.