The e-MANTIS emulator: fast predictions of the non-linear matter power spectrum in $f(R)$CDM cosmology

Iñigo Sáez-Casares [ORCID],[1]★ Yann Rasera[1] and Baojiu Li [ORCID][2]

[1]*Laboratoire Univers et Théories, Université Paris Cité, Observatoire de Paris, Université PSL, CNRS, F-92190 Meudon, France*
[2]*Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*

ABSTRACT

In order to probe modifications of gravity at cosmological scales, one needs accurate theoretical predictions. N-body simulations are required to explore the non-linear regime of structure formation but are very time consuming. In this work, we release a new public emulator, dubbed E-MANTIS, that performs an accurate and fast interpolation between the predictions of $f(R)$ modified gravity cosmological simulations, run with ECOSMOG. We sample a wide 3D parameter space given by the current background scalar field value $10^{-7} < |f_{R_0}| < 10^{-4}$, matter density $0.24 < \Omega_m < 0.39$, and primordial power spectrum normalization $0.6 < \sigma_8 < 1.0$, with 110 points sampled from a Latin hypercube. For each model we perform pairs of $f(R)$CDM and ΛCDM simulations covering an effective volume of $(560\,h^{-1}\,\mathrm{Mpc})^3$ with a mass resolution of $\sim 2 \times 10^{10} h^{-1} M_\odot$. We build an emulator for the matter power spectrum boost $B(k) = P_{f(R)}(k)/P_{\Lambda\mathrm{CDM}}(k)$ using a Gaussian process regression method. The boost is mostly independent of h, n_s, and Ω_b, which reduces the dimensionality of the relevant cosmological parameter space. Additionally, it is more robust against statistical and systematic errors than the raw power spectrum, thus strongly reducing our computational needs. According to our dedicated study of numerical systematics, the resulting emulator has an estimated maximum error of 3 per cent across the whole cosmological parameter space, for scales $0.03\,h\,\mathrm{Mpc}^{-1} < k < 7\,h\,\mathrm{Mpc}^{-1}$, and redshifts $0 < z < 2$, while in most cases the accuracy is better than 1 per cent. Such an emulator could be used to constrain $f(R)$ gravity with weak lensing analyses.

Key words: gravitation – methods: numerical – dark energy – dark matter – large-scale structure of Universe – cosmology: theory.

1 INTRODUCTION

The nature of the late-time accelerated expansion of the Universe remains one of the main unsolved puzzles of modern cosmology. The unknown component responsible for such an acceleration of the expansion rate is commonly referred to as dark energy (DE). In the standard cosmological model, ΛCDM, DE is modelled by a cosmological constant term Λ in the Einstein field equations. Other alternative scenarios are also usually considered, such as the quintessence models, where DE is made of a dynamical scalar field (see for example Amendola & Tsujikawa 2010, for a review of different DE models). Another possibility is that DE comes from a modification of general relativity (GR) at cosmological scales, such as in several modified gravity (MG) theories (see Clifton et al. 2012, for an extensive review of MG theories in a cosmological context).

Cosmic structure formation is the result of the interplay between gravitational collapse and the expansion of the Universe. Therefore, the large scale structure (LSS) of the Universe offers a unique window to probe simultaneously the nature of DE and the laws of gravity at cosmological scales. This is precisely one of the main objectives of current and future next-generation photometric and spectroscopic

galaxy surveys, such as DESI (DESI Collaboration 2016), LSST (Ivezić et al. 2019), and *Euclid* (Amendola et al. 2018).

Accurate theoretical predictions will be required to fully exploit the potential of those experiments. One of the basic theoretical tools to characterize the growth of cosmic structures is the matter power spectrum. Even if not directly observable, it serves as a building block for other observables. However, getting accurate predictions for the power spectrum in the non-linear regime of structure formation requires the use of computationally expensive N-body simulations. Constraining theoretical models with observations requires searching a large parameter space formed by cosmological and DE/MG parameters. Therefore, predictions from N-body simulations cannot directly be used in an observational analysis, which would require to run simulations for hundreds of thousands of cosmological models.

In order to overcome this limitation, the concept of cosmic emulators was introduced in the pioneering work of Habib et al. (2007). An emulator is calibrated on a reduced number of simulations. It is then able to produce accurate and fast predictions for any cosmological model within the parameter space covered by the training simulations. In the past years, several emulators have been developed to predict the matter power spectrum in ΛCDM and some extended DE models (e.g. Lawrence et al. 2010; Agarwal et al. 2014; Lawrence et al. 2017; Euclid Collaboration 2019; Knabenhans et al. 2019; Angulo et al. 2021).

★ E-mail: inigo.saez-casares@obspm.fr

When it comes to modified gravity theories, cosmological simulations require significantly more computing resources than in standard gravity. Indeed, modified gravity simulations require to solve for the dynamics of at least one extra degree of freedom, which often exhibits a non-linear equation of motion (see Llinares 2018, for a recent review on MG simulations).

In this work, we focus on a particular MG model, the *f(R)* gravity theory (see Clifton et al. 2012, for other MG models). It is part of the scalar–tensor family of MG theories, where the gravitational dynamics are described by the usual metric tensor and an extra scalar field. More precisely, we work with a particular form of *f(R)* gravity introduced by Hu & Sawicki (2007). More details are given in Section 2. The Hu & Sawicki *f(R)* gravity model is of cosmological interest since it is able to produce cosmic acceleration without an explicit cosmological constant. Additionally, it exhibits the so-called chameleon screening mechanism (Khoury & Weltman 2004), which hides the deviations from GR in dense environments such as the solar system. Such a mechanism is required to pass laboratory and solar system scale precision tests of gravity.

Several parallel and adaptive N-body codes have been developed over the past years to run cosmological simulations in *f(R)* gravity, such as ECOSMOG (Li et al. 2012a), MG-GADGET (Puchwein, Baldi & Springel 2013), ISIS (Llinares, Mota & Winther 2014), a modified version of AREPO (Arnold, Leo & Li 2019; Hernández-Aguayo et al. 2021) and MG-GLAM (Hernández-Aguayo et al. 2022; Ruan et al. 2022). A first emulator for the matter power spectrum in *f(R)* gravity was introduced by Ramachandra et al. (2021), named MGEMU. However, their emulator is based on COmoving Lagrangian Acceleration (COLA) simulations, which are less accurate than N-body methods in the non-linear regime. Very recently, generic pipelines to build emulators for the matter power spectrum in *f(R)* gravity and other MG theories using COLA simulations have been developed (Fiorini, Koyama & Baker 2023; Mauland, Winther & Ruan 2023). To date, the only *f(R)* emulator based on N-body simulations is the FORGE emulator (Arnold et al. 2022), which is built on top of cosmological simulations run with AREPO.

The purpose of this paper is to present a new emulator, dubbed as E-MANTIS[1] (Emulator for Multiple observable ANalysis in extended cosmological TheorIeS), for the matter power spectrum boost in *f(R)* gravity, defined as $B(k) = P_{f(R)}(k)/P_{\Lambda CDM}(k)$. The boost is less affected by statistical and systematic errors and has a smoother cosmological dependence than the power spectrum. Therefore, by emulating the boost instead of the raw power spectrum, we significantly reduce our computational needs. The simulations used to calibrate E-MANTIS are run with the modified gravity code ECOSMOG. Our emulation strategy and simulation code are different than the ones used in FORGE. We therefore expect both emulators to be complementary and we will investigate the differences within this article. We also make a detailed and careful study of numerical systematic which is generally not present in publications dedicated to modified gravity emulators.

The paper is structured as follows. Section 2 presents some theoretical aspects about *f(R)* gravity and the Hu & Sawicki model in the context of cosmological perturbations. In Section 3, we introduce the methodology and numerical codes used to run the N-body simulations and build the emulator. Section 4 estimates the accuracy of the emulator against statistical and systematic errors in the training data as well as pure emulation errors. Finally, in Section 5,

we compare the predictions from E-MANTIS to other prescriptions in the literature and give an example of usage.

2 F(R) GRAVITY

In this section, we briefly introduce the basic equations of *f(R)* gravity applied to cosmological perturbations and their phenomenology. For more details, we refer the reader to reviews on the subject (e.g. Sotiriou & Faraoni 2010; Clifton et al. 2012). In *f(R)* gravity the Einstein–Hilbert part of the action is modified by the addition of a new arbitrary function *f* of the Ricci scalar *R*

$$S = \frac{1}{2\kappa^2} \int d^4x \sqrt{-g} \left[R + f(R) \right] + S_m, \tag{1}$$

where $\kappa^2 = 8\pi G/c^4$ with G being Newton's gravitational constant and c the speed of light in vacuum, g is the determinant of the space–time metric $g_{\mu\nu}$ and S_m is the action describing the matter content of the Universe. By extremizing the total action with respect to the metric, we obtain the modified field equations

$$G_{\mu\nu} + f_R R_{\mu\nu} - \left(\frac{f}{2} - \Box f_R \right) g_{\mu\nu} - \nabla_\mu \nabla_\nu f_R = \kappa^2 T_{\mu\nu}, \tag{2}$$

where $R_{\mu\nu}$ is the Ricci curvature tensor, $G_{\mu\nu}$ is the Einstein tensor, $T_{\mu\nu}$ is the matter stress–energy tensor, ∇_μ is the covariant derivative compatible with the metric, and $\Box \equiv g^{\mu\nu} \nabla_\mu \nabla_\nu$. We have introduced the derivative of the *f* function with respect to the Ricci scalar

$$f_R \equiv \frac{df(R)}{dR}. \tag{3}$$

This quantity is a new dynamical scalar field introduced by the *f(R)* term in the modified gravitational action. It is sometimes referred to as the scalaron. The equation of motion for this scalar field can be obtained by taking the trace of equation (2)

$$\Box f_R = \frac{1}{3} \left(R - f_R R + 2f + \kappa^2 T \right), \tag{4}$$

where $T = T_\mu^\mu$ is the trace of the stress–energy tensor. We are interested in modelling the late-time evolution of the Universe. Therefore, we make the usual assumption that the Universe is filled with a pressure-less non-relativistic fluid called cold dark matter (CDM). In such a case, as long as we focus on sufficiently small scales, we can write $T = -\rho c^2$, where ρ is the mass density of CDM (Chisari & Zaldarriaga 2011).

2.1 Cosmological perturbations

We want to solve the field equations for cosmological scalar perturbations around an homogeneous Friedmann–Lemaître–Robertson–Walker (FLRW) background. In order to do so, we use the Newtonian gauge where the line element is given by

$$ds^2 = -\left(1 + 2\psi/c^2\right) c^2 dt^2 + a^2(t)\left(1 - 2\phi/c^2\right) \delta_{ij} dx^i dx^j, \tag{5}$$

where ψ and ϕ are the two gauge invariant Bardeen potentials (Bardeen 1980) and a is the scale factor. We work in the weak field limit, i.e. $|\psi/c^2|, |\phi/c^2|, |f_R| \ll 1$. We also adopt the quasi-static approximation, which means that we neglect all time derivatives of the three scalar fields with respect to their spatial ones. The validity of this approximation for *f(R)* gravity cosmological simulations has been confirmed by Bose, Hellwing & Li (2015). Under these approximations and by subtracting the background component from equation (4), we obtain

$$\frac{1}{a^2} \vec{\nabla}^2 f_R = \frac{1}{3} \left[\delta R(f_R) - \frac{8\pi G}{c^2} \delta\rho \right], \tag{6}$$

where $\delta R \equiv R - \bar{R}$ and $\delta \rho \equiv \rho - \bar{\rho}$, the over-bars denoting background quantities, and $\vec{\nabla}$ is the gradient operator in Euclidean space with respect to the comoving coordinate $\vec{x}$. In order to obtain equation (6), we have used the fact that $\delta f \sim f_R \delta R \ll \delta R$. From the 00 component of equation (2), we derive the equation for the lapse function ϕ,

$$\frac{1}{a^2}\vec{\nabla}^2\phi = \frac{8\pi G}{3}\delta\rho + \frac{c^2}{6}\delta R\,(f_R)\,, \tag{7}$$

where we have neglected terms of order $H^2\phi/c^2 \ll \partial_i^2\phi/a^2$. Such an approximation remains valid as long as we focus on small scales such that $k \gg aH/c$, where k is the comoving wavenumber. In the same way, from the trace of the ij component of equation (2) we obtain

$$\frac{2}{a^2}\vec{\nabla}^2(\psi - \phi) = 8\pi G\delta\rho - c^2\delta R\,(f_R) + \frac{c^2}{a^2}\vec{\nabla}^2 f_R. \tag{8}$$

We can now combine the last three equations, which gives

$$\frac{1}{a^2}\vec{\nabla}^2\psi = \frac{16\pi G}{3}\delta\rho - \frac{c^2}{6}\delta R\,(f_R)\,. \tag{9}$$

Finally, by combining equations (7) and (9), one can obtain

$$\frac{1}{a^2}\vec{\nabla}^2(\psi + \phi) = 8\pi G\delta\rho. \tag{10}$$

This equation shows that the relation between the potential in lensing studies and the matter distribution is the same as in GR. More details about the derivations of these equations can be found for example in Hu & Sawicki (2007), Oyaizu (2008), and Bose, Hellwing & Li (2015).

2.2 The Hu and Sawicki model

From now on we focus on a particular $f(R)$ model, proposed by Hu & Sawicki (2007), which uses the functional form

$$f(R) = -m^2\frac{c_1\left(R/m^2\right)^n}{c_2\left(R/m^2\right)^n + 1}\,, \tag{11}$$

where c_1, c_2, and n are dimension-less parameters and m is a curvature scale. This scale is taken to be

$$m^2 \equiv \frac{\Omega_m H_0^2}{c^2}\,, \tag{12}$$

where Ω_m is the current fractional matter density and H_0 is the Hubble parameter today.

It is interesting to perform an expansion of the $f(R)$ function at high curvature with respect to m^2

$$\lim_{R \gg m^2} f(R) \simeq -\frac{c_1}{c_2}m^2 + \frac{c_1}{c_2^2}m^2\left(\frac{m^2}{R}\right)^n. \tag{13}$$

First, we see that the Hu & Sawicki model is equivalent to a cosmological constant in the limit where $c_1/c_2^2 \to 0$ and c_1/c_2 is kept constant. Therefore, this model is able to produce cosmic acceleration. In order to closely match the ΛCDM expansion history one can fix

$$\frac{c_1}{c_2} = 6\frac{\Omega_\Lambda}{\Omega_m}\,, \tag{14}$$

where Ω_Λ is the current fractional density of dark energy.

Secondly, at finite c_1/c_2^2 and fixed n, the deviation from a pure cosmological constant vanishes as the density increases. This is an example of a screening mechanism (see for example Brax et al. 2022, for a recent review), and in particular the so-called chameleon

mechanism (Khoury & Weltman 2004). Such a mechanism allows the Hu & Sawicki model to exhibit significant deviations from ΛCDM at cosmological scales while recovering GR in the Solar System, where the laws of gravity are more tightly constrained. The remaining free parameters n and c_1/c_2^2 control the efficiency of the chameleon screening.

The scalar field f_R can be written in terms of the scalar curvature R as

$$f_R = -n\frac{c_1\left(\frac{R}{m^2}\right)^{n-1}}{\left[c_2\left(\frac{R}{m^2}\right)^n + 1\right]^2} \simeq -n\frac{c_1}{c_2^2}\left(\frac{m^2}{R}\right)^{n+1}. \tag{15}$$

The second equality is an approximation valid in the high curvature regime $R \gg m^2$. Such an approximation remains valid up to $z = 0$ for the choice of m^2 given by equation (12) as explained in Hu & Sawicki (2007) and Oyaizu (2008). In practice, it is useful to replace the parameter c_1/c_2^2 by the background value of the scalaron field today f_{R_0}. Indeed, by combining equations (13), (14), and (15), one can write in the high curvature limit,

$$f(R) \simeq -2\Lambda + \frac{f_{R_0}}{n}\frac{R_0^{n+1}}{R^n}\,, \tag{16}$$

where R_0 is the current background value of the Ricci scalar and Λ is the cosmological constant.

At the background level, the deviation from an equation of state $w = -1$ is of the order of f_{R_0} (Hu & Sawicki 2007). In this work, the strongest modification of gravity that we consider is $f_{R_0} = -10^{-4}$. As a consequence, for all practical purposes, the background expansion history is identical to ΛCDM. Additionally, for such values of f_{R_0}, the effect of $f(R)$ gravity on the evolution of the matter density field at high redshift is negligible. The differences between $f(R)$ and ΛCDM will arise at the level of the late-time non-linear matter clustering.

We restrict our study to the case $n = 1$, which is the most common case in the literature. Performing cosmological simulations with $n \neq 1$ is also much more cpu-time consuming (see Section 3.1). We refer the reader to other papers such as Ramachandra et al. (2021) and Ruan et al. (2022) for the case $n \neq 1$.

3 METHODS

3.1 Codes

The suite of simulation was performed with the automated pipeline introduced in Blot et al. (2015) and further refined in Blot et al. (2021). In this study, we have extended it to $f(R)$ modified gravity. Note that as stated above, the effect of modified gravity on the structure formation at high redshift (initial condition), and its impact on the background evolution are both negligible. The main changes are therefore limited to the dynamical solver.

The linear matter power spectrum was computed using the CAMB Boltzmann solver (Lewis, Challinor & Lasenby 2000). Initial conditions were generated using MPGRAFIC (Prunet et al. 2008) which assumes Gaussian random field and second-order Lagrangian perturbation theory (2LPT) (Crocce, Pueblas & Scoccimarro 2006). The initial redshift was chosen as low as possible while avoiding particle crossing (in practice $z_i \sim 50$). With this choice, numerical errors remain limited (Michaux et al. 2021).

Particles are evolved in a periodic cubic box using the $f(R)$ version of the RAMSES N-body code (Teyssier 2002) called ECOSMOG (Li et al. 2012a). RAMSES implements a particle-mesh (PM) method with adaptive mesh refinements (AMR) in overdense regions. A mesh is

refined (i.e. divided into eight sub-cells), whenever the number of particles inside it exceeds a given threshold $m_{\rm ref}$, up to a maximum refinement level $l_{\rm max}$. The (linear) modified Poisson equation for the Newtonian potential (equation 9) is solved through a standard multigrid approach (Guillet & Teyssier 2011). On the other hand, the non-linear equation of motion for the extra scalar field (equation 6) was solved using a Newton–Gauss–Seidel method embedded in the multigrid approach (Li et al. 2012a). The solver has been further optimized in Bose et al. (2017) thanks to the existence of analytical solutions for the particular case $n = 1$. We have therefore used this optimized ECOSMOG version. An important parameter is the required precision for the non-linear scalar field solver: we have set it to $\epsilon_{\rm sf} = 10^{-7}$, and we have checked that with this choice, the errors on the power spectrum boost are much smaller than 1 per cent in the studied range of wavenumbers.

The matter power spectrum was estimated using POWERGRID (Prunet et al. 2008) with a Cloud-In-Cell assignment (CIC). We use a mesh grid with 64 times more cells than the RAMSES coarse grid. We have dropped the matter power spectrum values for wavenumbers beyond half the Nyquist frequency of the POWERGRID grid to avoid aliasing.

3.2 Emulation

3.2.1 The matter power spectrum boost

Instead of emulating the matter power spectrum itself, we decide to focus on the matter power spectrum boost defined as

$$B(k) = P_{f(R)}(k)/P_{\Lambda\rm CDM}(k). \tag{17}$$

This quantity encodes the enhancement of the matter power spectrum due to $f(R)$ gravity with respect to a ΛCDM scenario. It brings several benefits that simplify the emulation procedure. First, the dynamical range of the power spectrum boost is significantly smaller than the one of the full power spectrum. We go from emulating a quantity spanning several orders of magnitude to one that is of the order of $\mathcal{O}(1)$ at all scales, which is a simpler task to perform (Angulo et al. 2021).

Secondly, we have verified that the power spectrum boost is mostly independent of the baryon density parameter Ω_b, the reduced Hubble parameter h^2, and the primordial spectral index n_s, even at deeply non-linear scales. This is in agreement with previous studies (Winther et al. 2019). In order to build an emulator, we only need to consider the 3D parameter space given by the background value of the scalar field today f_{R_0}, the total matter density parameter $\Omega_{\rm m}$, and the primordial power spectrum normalization[3] σ_8. The dependence of the power spectrum boost on the cosmological parameters is discussed in more details in Section 3.2.2.

Thirdly, by taking the ratio of the power spectrum obtained in $f(R)$ gravity and ΛCDM from the same exact initial conditions, we obtain a significant cancellation of cosmic variance and mass resolution errors, which dominate the error budget at large and small scales, respectively. With this method, we can build an accurate emulator without having to perform large volume and very high resolution

---

[2]Defined as $H_0 = 100h$ km s^{-1} Mpc^{-1}, where H_0 is the Hubble constant today.

[3]We follow the convention of previous works in the field, and by σ_8 we refer to the value obtained assuming a linear ΛCDM evolution, even for $f(R)$CDM cosmologies. The parameter σ_8 is used as an indirect normalization of the primordial power spectrum, which is therefore the same in both ΛCDM and $f(R)$CDM for a given set of cosmological parameters.

Table 1. Cosmological models used in the *cosmo* simulation suite. The first line (in italics) is our reference model F5($\star$). Models 4–13 vary one of the 5 ΛCDM parameters by ± 25 per cent. Models 2 (F4) and 3 (F6) vary the f_{R_0} parameter. The parameters in bold are the ones that are being varied by a specific model. The values of Ω_Λ are not explicitly given, but they are fixed by the flatness condition $\Omega_\Lambda = 1 - \Omega_m - \Omega_r$. The radiation density parameter is fixed by the CMB temperature to $\Omega_r \sim 10^{-4}$. For each model in this table we run the corresponding ΛCDM simulations in addition to the $f(R)$ ones.

Model	$\lvert f_{R_0}\rvert$	Ω_m	σ_8	h	Ω_b	n_s
F5($\star$)	*10^{-5}*	*0.3153*	*0.8111*	*0.6736*	*0.049302*	*0.9649*
F4	**10^{-4}**	0.3153	0.8111	0.6736	0.049302	0.9649
F6	**10^{-6}**	0.3153	0.8111	0.6736	0.049302	0.9649
Ω_m^+	10^{-5}	**0.3941**	0.8111	0.6736	0.049302	0.9649
Ω_m^-	10^{-5}	**0.2365**	0.8111	0.6736	0.049302	0.9649
σ_8^+	10^{-5}	0.3153	**1.0140**	0.6736	0.049302	0.9649
σ_8^-	10^{-5}	0.3153	**0.6083**	0.6736	0.049302	0.9649
h^+	10^{-5}	0.3153	0.8111	**0.8420**	0.049302	0.9649
h^-	10^{-5}	0.3153	0.8111	**0.5052**	0.049302	0.9649
Ω_b^+	10^{-5}	0.3153	0.8111	0.6736	**0.061620**	0.9649
Ω_b^-	10^{-5}	0.3153	0.8111	0.6736	**0.036980**	0.9649
n_s^+	10^{-5}	0.3153	0.8111	0.6736	0.049302	**1.2060**
n_s^-	10^{-5}	0.3153	0.8111	0.6736	0.049302	**0.7237**

simulations, as it would be necessary for the raw matter power spectrum. More details are given in Section 4.1.

An important caveat of emulating the boost is that in order to get the full power spectrum in $f(R)$ models, it is necessary to combine the predictions of E-MANTIS with an independent ΛCDM emulator. However, current state-of-the-art ΛCDM emulators usually achieve per cent level accuracy (Angulo et al. 2021; Euclid Collaboration 2021; Moran et al. 2022).

3.2.2 Defining the parameter range

Before fixing the emulation parameter space, we explore the dependence of the matter power spectrum boost with the six cosmological parameters $\{f_{R_0}, \Omega_m, \sigma_8, h, \Omega_b, n_s\}$. Our aim is to show that in order to emulate the matter power spectrum boost, we only need to consider a subset of these parameters.

We define two reference cosmological models (one with usual gravity and the other one with modified gravity): the ΛCDM parameters take the best-fitting values from Planck Collaboration VI (2020), while the corresponding $f(R)$CDM parameters takes the same values but with $f_{R_0} = -10^{-5}$. From this reference we build pairs of models where each ΛCDM parameter is varied by ± 25 per cent at a time. We also study models with $f_{R_0} = -10^{-4}$ (F4) and $f_{R_0} = -10^{-6}$ (F6) to investigate the effect of f_{R_0} (corresponding to a variation of ± 20 per cent for $\log |f_{R_0}|$ while matching standard values found in the literature for comparison). We always consider a flat universe such that $\Omega_\Lambda = 1 - \Omega_m - \Omega_r$ and take a constant radiation density parameter fixed by the CMB temperature to $\Omega_r \sim 10^{-4}$. The parameter values of these models are summarized in Table 1.

For each model, we evolve $N_{\rm part} = 512^3$ DM particles in a periodic cubic box of length size $L_{\rm box} = 328.125\,h^{-1}$Mpc in a ΛCDM and a $f(R)$ simulation. Both simulations are initialized using the exact same initial conditions and same resolution. We perform such pairs of simulations for five independent initial conditions, which are shared across the different models. In total, we realize a set of 65 pairs of ΛCDM and $f(R)$ simulations to study the cosmological dependence

Table 2. Table summarizing the main characteristics of the EMULATOR simulation suite. L_{box} is the side length of the simulation box, N_{part} the number of DM particles, m_{part} the mass resolution (which varies depending on the particular cosmology, here are given the two most extreme values), l_{min} the coarse refinement level (corresponding to a coarse grid of 512^3), l_{max} the maximum refinement level, m_{ref} the refinement threshold, Δx the spatial resolution of the maximum refinement level, ϵ_p and ϵ_{sf} are the convergence criteria for the modified Poisson equation (equation 9) and the extra scalar field equation of motion (equation 6) respectively, N_{real} the number of realizations per model, N_{cosmo} the number of cosmological models (the $\times 2$ factor accounts for the corresponding ΛCDM simulations performed for each $f(R)$ model) and z_i the approximate initial redshift (which is also cosmology dependent).

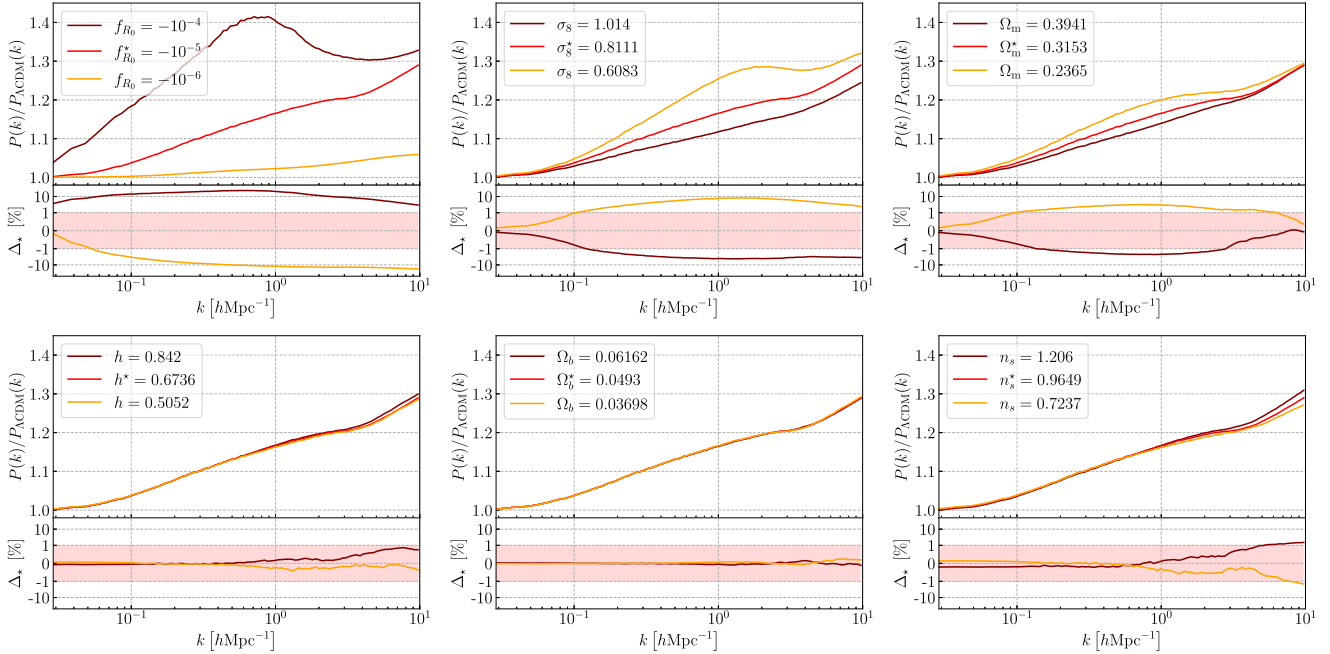| $L_{box}(h^{-1}\mathrm{Mpc})$	N_{part}	$m_{part}(h^{-1}\,\mathrm{M}_\odot)$	l_{min}	l_{max}	m_{ref}	$\Delta x(h^{-1}\mathrm{kpc})$	$\epsilon_p(\epsilon_{sf})$	N_{real}	N_{cosmo}	z_i
328.125	512^3	$(1.73–2.88)\times 10^{10}$	9	15	14	10	$10^{-4}(10^{-7})$	5	$110(\times 2)$	~ 50

Figure 1. Cosmological dependence of the matter power spectrum boost at $z = 0$. In each subplot, one of the six cosmological parameters (as indicated in the inset) is varied by ± 25 per cent around the reference model F5($\star$) from Table 1. The lower panels give the relative difference with respect to the reference. The light bands mark the ± 1 per cent variation limit. Inside it the y axis scaling is linear and outside it is logarithmic. Only the parameters f_{R_0}, σ_8, and Ω_m have an impact greater than 1 per cent on the matter power spectrum boost. We can therefore neglect the effect of the other parameters when building the emulator and keep them fixed to their reference values.

of the matter power spectrum boost. We refer to this ensemble of simulations as the *cosmo* simulation suite. From this simulation suite, we measure the matter power spectrum boost for each cosmology and average it over the 5 realizations. The numerical parameters of these simulations are the same as the *emulator* simulation suite that is introduced in Section 3.3 (see Table 2).

Fig. 1 shows the dependence of the matter power spectrum boost with each cosmological parameter. Each subplot varies one of the six parameters independently from the others. The lower panels show the relative difference with respect to the reference model F5($\star$). These results show that the variation of the boost with the parameters h, Ω_b, and n_s is very weak up to small scales of the order of $k \sim 10\,h\mathrm{Mpc}^{-1}$. For Ω_b and h, the relative variation of the boost remains smaller than 0.2 per cent and 1 per cent, respectively, at all scales considered, i.e. for $0.03\,h\,\mathrm{Mpc}^{-1} < k < 10\,h\,\mathrm{Mpc}^{-1}$. In the case of n_s, the relative variation is smaller than 1 per cent except at the smallest scales, where it starts to cross the 1 per cent line. We stress here the fact that we have performed very strong variations of the cosmological parameters. If we had considered only variations within the current observational constraints (Planck Collaboration VI 2020), the effect on the boost would have been completely negligible. The weak dependence on

h shown in Fig. 1 explains why Harnois-Déraps et al. (2023) find, in their forecast, that the weak lensing convergence power spectrum places very weak constraint on this parameter.

As expected the parameter with the strongest impact is f_{R_0}, which for the F4 model, can lead to an enhancement of the matter power spectrum of up to 40 per cent with respect to the ΛCDM scenario. The parameters σ_8 and Ω_m lead to variations of up to 8 per cent and 3 per cent, respectively. These results are in qualitative agreement with previous works such as Winther et al. (2019).

One can interpret why σ_8 and Ω_m play an important role in terms of the screening mechanism. For instance, if the amount of matter in the Universe increases, the overall density does the same, and because of the chameleon screening mechanism, the effect of $f(R)$ gravity gets weaker. Indeed, we can see in Fig. 1 that the matter power spectrum boost is a decreasing function of Ω_m, the other parameters being fixed. The same goes for σ_8. With a higher σ_8 the amount of clustering in the Universe increases and overdensities are globally denser. The screening is therefore stronger and the effect of $f(R)$ gravity is weaker.

We have established the impact of each individual cosmological parameter on the matter power spectrum boost around the reference

F5(⋆)model. In particular, the parameters n_s, h and Ω_b have an effect smaller than 1 per cent, even for very large variations of those parameters. We could therefore ignore them in the construction of the emulator. However, the simultaneous variation of multiple parameters might produce a non-negligible effect. In Appendix A, we investigate the evolution of errors in the most sensitive parts of the parameter space. We find that the cost of ignoring those parameters on the accuracy of the power spectrum boost determination is still smaller than 1 per cent in most cases. In the worst cases, mostly around the edges of our wide cosmological parameter space, the induced error can reach the $2-3$ per cent level. As we show in Section 4.1.1, this is less than the systematic errors in our training data.

From this study, we can conclude that in order to build an emulator, with a few per cent level accuracy, for the matter power spectrum boost, it is enough to focus on the three parameters $\{f_{R_0}, \Omega_m, \sigma_8\}$. Considering a 3D parameter space, instead of the full 6D dimensional, reduces the number of simulations we need to perform in order to achieve a given emulation accuracy. Indeed, performing an interpolation in 3D requires less samples than in 6D. Given the computational cost of carrying out N-body simulations in $f(R)$ gravity, this is an important aspect.

Ideally, we would build an emulator that covers all possible cosmological models within this 3D parameter space. However, we can only emulate a finite volume of it. This volume needs to be large enough to accommodate all reasonable cosmological models as well as some margin around them, so that the emulator can be properly used in any theoretical, numerical, or observational analysis. We decide to consider variations of Ω_m and σ_8 of up to 25 per cent around the current best-fitting values in ΛCDM from Planck Collaboration VI (2020). This range is enough to explore very conservative deviations of order $\sim 5\sigma$ from the constraints obtained in $f(R)$ gravity with the galaxy clustering ratio (Bel et al. 2015) for Ω_m and with galaxy clusters (Cataneo et al. 2015) for σ_8. Both studies exclude models with $-f_{R_0} \geq 10^{-4}$ at the 2σ level. We will use this value as the upper bound for f_{R_0}. For the lower bound we take $f_{R_0} = -10^{-7}$, which corresponds to an almost vanishing modification of the matter power spectrum with respect to ΛCDM. In the end the emulator covers the parameter range given by

$$-7 \leq \log |f_{R_0}| \leq -4, \tag{18}$$

$$0.2365 \leq \Omega_m \leq 0.3941, \tag{19}$$

$$0.6083 \leq \sigma_8 \leq 1.0140. \tag{20}$$

Such large range of variation of the parameters is also important to be able to measure their impact on the boost. This is key to accurately estimate the derivatives of the power spectrum boost with respect to the cosmological parameters, so as to make an accurate interpolation. In practice E-MANTIS works with $-\log |f_{R_0}|$ instead of f_{R_0}, since it improves its numerical accuracy. We fix the other parameters to their best-fitting values from Planck Collaboration VI (2020): $h = 0.6736$, $\Omega_b = 0.049302$, and $n_s = 0.9649$.

We conclude this section by noting that, as mentioned in Section 3.1, we limit our study to the case where the modified gravity parameter $n = 1$. This is mainly due to computational limitations, since performing N-body simulations for other values of this parameter is much more time consuming (Bose et al. 2017). However, Ramachandra et al. (2021) shows that a variation of ± 25 per cent around the value $n = 1$ has an effect smaller than 1 per cent on the matter power spectrum boost. Therefore, our emulator would also be valid in that regime, at least for scales $k \lesssim 1 \, h \, \mathrm{Mpc}^{-1}$, which is the

range of validity of their predictions. Moreover this work suggests that the most sensitive part of the boost with respect to a variation of n is the one at intermediate scales near $k \simeq 0.5 \, h \, \mathrm{Mpc}^{-1}$.

3.2.3 Sampling the parameter space

We now detail the strategy employed to sample the model parameter space with training points. This ensemble of training models is usually referred to as the experimental design (ED). Since N-body simulations are highly time consuming, we are limited by the number that can be performed. Usually, cosmological emulators use a number of training models of the order of $\mathcal{O}(10-100)$ (Nishimichi et al. 2019; Euclid Collaboration 2021). In order to efficiently sample the parameter space with a reduced number of points, it is common to use a Latin hypercube sampling (LHS) method. It has been shown to give good performances when combined with a Gaussian process (GP) emulation (Habib et al. 2007).

A Latin hypercube design (LHD) is generated by first dividing the D-dimensional parameter space into a N^D regular grid. Then N grid cells are randomly selected with the condition that each cell cannot have any common coordinate with the other ones. Finally, within each selected grid cell a point is randomly sampled (Garud, Karimi & Kraft 2017). Usually, additional criteria are imposed. For instance, one can generate a set of k independent LHD and then select the one that maximizes the minimal distance between sample points (maximin-distance condition). This kind of method ensures that the ED optimally covers the whole parameter space in a homogeneous way.

One possible modification of the LHS method is the maximin-distance Sliced LHD (SLHD) introduced by Ba, Myers & Brenneman (2015). A LHD is divided into sub-samples with an equal number of points each. The particularity is that each sub-sample, also referred to as a *slice*, verifies the LHD condition on its own. A version of the maximin-distance condition is also applied, where the maximized distance is a mixture of the distances between points in each slice and in the whole sample. We refer the reader to the original paper for more details about the precise implementation. The SLHD sampling technique was used in Nishimichi et al. (2019) in order to separate the ED points into a training and a validation sample. In this way, the validation points are homogeneously distributed over the parameter space, while at the same time being as far as possible from the training points. In this work, we use the R package SLHD[4] to generate a SLHD with 5 slices of 20 points each (hereafter named *primary* SLHD). We have performed cosmological simulations only for the first four slices (i.e. 80 cosmological models). Indeed, we found that it was more beneficial to sample additional points by targeting specific regions of the parameter space, instead of adding an additional set of homogeneously distributed models (see the next paragraph). The left panel of Fig. 2 shows the distribution of models of the first four slices from the *primary* SLHD. Each slice is represented by a different colour. We have sampled Ω_m and σ_8 with an homogeneous prior. For f_{R_0}, we sample homogeneously in $\log |f_{R_0}|$ instead.

Sampling training points with a LHS method is a good strategy when we have no a priori knowledge of how the signal that we want to emulate varies over the whole parameter space. However, if the gradient (or its derivatives) of the signal is stronger in a particular region, it will require a higher density of training points to get an accurate emulation. In this work, we use a strategy similar to the one followed by Angulo et al. (2021) in order to add training points in the regions where emulation is more difficult. We start by building
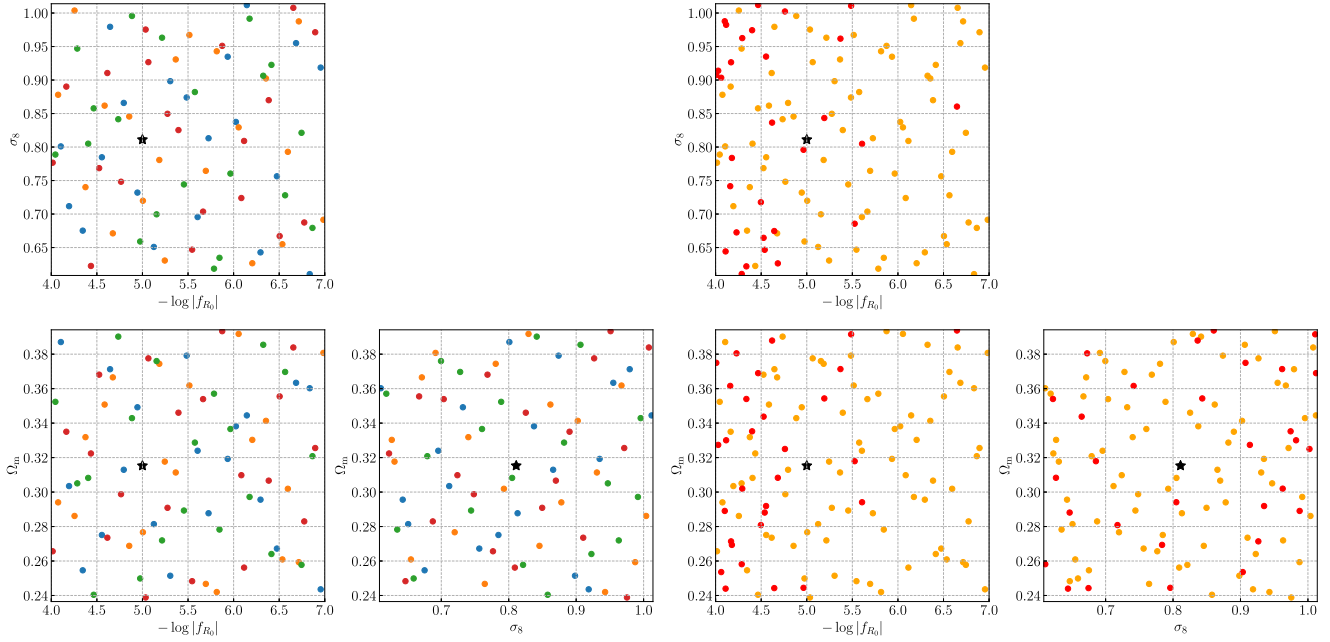
---

[4]https://CRAN.R-project.org/package=SLHD

Figure 2. 2D projections of the model points used to train and test the emulator. Left: each colour represents one of the first four slices of the primary sliced Latin hypercube design (SLHD). Each slice is composed of 20 points, forming a LHD on the 3D parameter space $\{\Omega_{\mathrm{m}}, \sigma_8, f_{R_0}\}$. The black star corresponds to our reference F5 cosmology. Right: training points of the final emulator. The points in orange correspond to the four slices of the primary SLHD. The red ones are the additional training models drown from the refinement SLHD by the iterative procedure described at the end of Section 3.2.3. They are mostly distributed towards the high values of f_{R_0}, while staying uniform in Ω_{m} and σ_8. This is the region where the emulation errors are larger.

a preliminary emulator using three slices (i.e. a total of 60 training models) from the primary SLHD. We can use the remaining slice (i.e. 20 validation models) to assess the accuracy of this emulator and decide whether or not it is necessary to add new training points.

Since additional training points are needed, we generate a second SLHD (hereafter the *refinement* SLHD) with 100 models per slice. In order to sample new training points from these slices, we follow an iterative procedure. First, we randomly select 10 points from the first slice, with a selection probability proportional to the emulation error. We use the interpolation error directly given by the GP itself. Even though this estimation is not exact, it gives a good order of magnitude for our purpose here. We then re-train the emulator with the newly selected training models. Finally, we assess once again the accuracy of the emulator with the validation models. We repeat this procedure until the desired accuracy is obtained.

In the end, we add a total of 30 training models using this method. With the obtained gain in emulation accuracy, the emulation errors are now smaller than 1 per cent even for the most extreme models (see Appendix B for more details). The final version of the emulator is trained using all available simulated models, i.e. the 80 models from the first four slices of the primary SLHD and the additional 30 refinement models. The final distribution of the 110 training points is shown in the right panel of Fig. 2.

3.2.4 Smoothing the power spectrum boost

For each cosmological node, we measure, from our simulations, the binned matter power spectrum in $f(R)$ gravity and the corresponding ΛCDM cosmology. These raw power spectra need some treatment before they can be used to train the emulator. Indeed, they are noisy

due to the fine binning in k and the finite volume of the simulations. Such noise reduces the accuracy of the emulator.

In order to smooth out the training data it is common to use filtering techniques (e.g. Ramachandra et al. 2021; Arnold et al. 2022). In this work, we adopt a different strategy. We start by rebinning the power spectra into larger k bins. We adopt a constant linear binning with $\Delta k \simeq 0.009 h \mathrm{Mpc}^{-1}$ from $k = 0.024 h\,\mathrm{Mpc}^{-1}$ to $k = 0.18 h\,\mathrm{Mpc}^{-1}$ (in terms of bin edges). For the remaining scales we use a constant logarithmic binning up to $k = 10 h\,\mathrm{Mpc}^{-1}$. In order to avoid a discontinuity in the bin widths, the first logarithmic bin has the same size as the linear bins. The rebinned power spectra have a number of bins $N_k = 90$ instead of the initial $N_{\mathrm{raw}} = 1040$ bins. We then compute the boosts by taking the ratio of the power spectra in $f(R)$CDM with ΛCDM. Additionally, we simulate a number $N_{\mathrm{real}} = 5$ of independent realizations per cosmological model and average the boosts over them. In the end, we obtain a signal that is smooth enough to train the emulator and with enough bins in k. Appendix B shows how the number of realizations impacts the emulation accuracy.

3.2.5 Emulation strategy

We detail in this section the procedure adopted to interpolate between the training cosmological models. We focus initially on a single redshift. The strategy to extend the emulator to arbitrary redshifts is explained at the end.

Our training data are the smoothed power spectrum boosts from the cosmological nodes. They each have $N_k = 90$ bins in k. Emulating every bin independently is a quite inefficient strategy. Indeed, the different k-bins are highly correlated with each other with respect to the variation of the cosmological parameters. We can therefore use a principal component analysis (PCA) to reduce the dimensionality

of the training data. In a PCA, the training data is decomposed in a basis of orthogonal functions

$$B(\boldsymbol{\theta}; k) = \sum_{i=1}^{N_{\mathrm{PCA}}} \alpha_i(\boldsymbol{\theta}) \phi_i(k) + \epsilon, \tag{21}$$

where $\boldsymbol{\theta}$ represents the vector of cosmological parameters, the ϕ_i are a set of orthogonal basis functions, and N_{PCA} is the number of basis functions used to describe the signal. Such a decomposition is obtained by diagonalizing the covariance matrix of the training data. The quantity ϵ represents the part of B that is not captured by the truncated decomposition. By setting $N_{\mathrm{PCA}} = N_k$, we would have $\epsilon = 0$ identically. In practice, only a reduced number of PCA coefficients will be required to describe most of our signal. The base functions ϕ_i are independent of the cosmological parameters and all the cosmological dependence is carried by the α_i coefficients. Furthermore, these coefficients are now fully independent from each other, which was not the case of the k bins. We find that with $N_{\mathrm{PCA}} = 5$, we are able to describe more than 99.99 per cent of the total variance contained in the power spectrum boosts. By increasing the number of PCA coefficients, we would mostly capture the noise in the training signals. Additionally, we have verified that, given our strategy of binning in k, and the number of PCA coefficients that we keep, the absolute error introduced by this decomposition is on average smaller than 0.1 per cent.

We now build an independent emulator for each of the N_{PCA} coefficients. We use a Gaussian process (GP) regression to interpolate the PCA coefficients between training nodes. GPs are commonly used in cosmological emulators (e.g. Habib et al. 2007; Euclid Collaboration 2019; Angulo et al. 2021; Ramachandra et al. 2021; Arnold et al. 2022). One of the most used kernels is the radial-basis function (RBF) kernel. This kernel states that the covariance of the signal decreases as a squared exponential with the distance in the parameter space. It is characterized by a length scale in the isotropic case, or a length scale per parameter space dimension in the anisotropic case. The RBF kernel is most efficient at emulating smooth quantities, since it is infinitely differentiable. For our emulator we use a generalization of the RBF kernel to less smooth functions, the Matern kernel. We find that an anisotropic Matern-5/2 kernel gives the best emulation accuracy given our training data. We refer the reader to Rasmussen & Williams (2006) for more in depth details about GPs and kernels. It is worth mentioning that before performing the PCA, we standardize the training data and model points. This means that we remove the mean and rescale them in order to obtain a unit variance distribution. This simple step simplifies the usage of the PCA and GP emulation and improves their performance. We use the Python package SCIKIT-LEARN[5] (Pedregosa et al. 2011) to perform the standardization of the training data, the PCA, and the GP regression.

We have just described the procedure to build the emulator at a given redshift. We have a total of 19 redshift nodes between $0 < z < 2$ at our disposal (see Section 3.3). In practice, we build an independent emulator for each redshift node. Then, for any arbitrary redshift, we linearly interpolate the predicted power spectrum boost between the two closest redshift nodes. We have checked that this method is sufficient to obtain interpolation errors smaller than 1 per cent at all scales and for all cosmological models. More details are given in Appendix C. Such linear interpolation in redshift is also used in other modified gravity emulators (Ramachandra et al. 2021; Arnold et al. 2022).

[5] https://scikit-learn.org/stable/index.html

3.3 Emulator simulation set

We perform N-body simulations for each of the $N_{\mathrm{cosmo}} = 110$ cosmological models sampled in Section 3.2.3. We use the numerical simulation chain described in Section 3.1. Each simulation evolves 512^3 DM particles in a cubic periodic box of side length $L_{\mathrm{box}} = 328.125 h^{-1}\mathrm{Mpc}$, which gives a mass resolution of the order of $\sim 2 \times 10^{10} h^{-1}\,\mathrm{M_\odot}$ (depending on the particular cosmology), from an initial redshift of $z_i \sim 50$ to $z = 0$. For each of the N_{cosmo} models, we simulate $N_{\mathrm{real}} = 5$ independent random realizations, covering a total effective volume of $(560 h^{-1}\mathrm{Mpc})^3$. For each $f(R)$ cosmological model and realization, we perform the equivalent ΛCDM simulation. In the end, this gives a total of 1100 cosmological simulations. We refer to this large set of simulations as the emulator simulation set. The CPU-time usage for this simulation suite was of $\sim$3.5 million core-hours using the AMD Irene ROME partition of the Joliot-Curie supercomputer hosted at the Très Grand Centre de Calcul (TGCC). The main characteristics of our emulator simulation set are summarized in Table 2. We save the matter power spectra at 19 redshift outputs per simulation: $z = 0, 0.05, 0.1, 0.15, 0.25, 0.3, 0.42, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.25, 1.4, 1.5, 1.74, 2$.

3.4 Simulation set for the evaluation of systematic and statistical errors

The finite volume and mass resolution of the N-body simulations produce non-negligible errors on the measured power spectrum boosts. A proper estimation of such effects would require us to perform a dedicated set of simulations with fixed mass resolution and varying volume to assess the finite volume and sample variance errors. Another set with fixed volume and varying number of DM particles would be needed to estimate the resolution errors. However, it would require a significant amount of computing resources. This is particularly true in $f(R)$ gravity, since, as we will show in the next section, the resolution errors are highly dependent on the value of f_{R_0}. It is therefore important to assess the accuracy of our simulations for different $f(R)$ models, which again increases the computational cost.

We resort instead to a more approximate method. We perform a set of simulations where we vary the simulation box-size while keeping the number of DM particles fixed. The simulations used to build our emulator evolve $N_{\mathrm{part}} = 512^3$ DM particles in a cubic periodic box of side $L_{\mathrm{box}} = 328.125\,h^{-1}\mathrm{Mpc}$. We refer to this box-size and resolution as the *standard resolution* (SR). In order to assess the systematic errors at small scales due to the finite mass resolution we consider a second class of simulations with $L_{\mathrm{box}} = 164.0625\,h^{-1}\mathrm{Mpc}$. We label them as the *high resolution* (HR) simulations. The HR box has a volume eight times smaller than the SR box and the same number of DM particles, which corresponds to a mass resolution eight times higher. The HR box-size is relatively small and will suffer from an important cosmic variance. As long as it only affects the largest scales, we can still use it to test the convergence of the SR simulations at small scales. We also run simulations with larger box-sizes in order to assess the impact of the simulation box length. More precisely we consider box-sizes of $L_{\mathrm{box}} = 656.25$ and $1312.5\,h^{-1}\mathrm{Mpc}$, which we refer to as *low resolution* (LR) and *very low resolution* (VLR), respectively. These simulation boxes have a volume 8 and 64 times bigger than the SR, respectively, and the same number of DM particles. We use them to estimate the accuracy of our measurements at large scales. More precisely, with this method we are probing a combination of sample variance and finite volume effects.

For each box-size, we run a single realization of the models F4, F5, and F6 as well as the corresponding ΛCDM cosmology. In order

Table 3. Table summarizing the different simulation boxes used to study the large scale and mass resolution errors of the matter power spectrum boost. L_{box} is the side length of the simulation box and m_{part} the mass of an N-body particle. The number of DM particles in each simulation is 512^3. The other numerical parameters are identical to those described in Table 2. The models F4, F5, and F6 correspond to the cosmological parameters given in Table 1. Inside the parentheses are given the number of realizations performed per $f(R)$ model. For each $f(R)$ model and realization, we also perform the corresponding ΛCDM simulation.

Name	$L_{\text{box}}(h^{-1}\text{Mpc})$	$m_{\text{part}}\left(h^{-1}\,\text{M}_\odot\right)$	Models
High Res. (HR)	164.0625	2.9×10^9	F4, F5, F6 ($\times$ 1)
Standard Res. (SR)	328.125	2.3×10^{10}	F4, F6 ($\times$ 1), F5 ($\times$ 30)
Low Res. (LR)	656.25	1.8×10^{11}	F4, F5, F6 ($\times$ 1)
Very Low Res. (VLR)	1312.5	1.5×10^{12}	F4, F5, F6 ($\times$ 1)

to study the impact of the number of realizations, we also simulate an ensemble of 30 independent realizations with the SR box for the F5 model (and the equivalent ΛCDM cosmology). The volumes and mass resolutions of this suite of simulations is given in Table 3. The other numerical parameters are identical to the ones used in the *emulator* suite (see Table 2).

4 NUMERICAL CONVERGENCE STUDIES

The final accuracy of the emulator is affected by both the errors on the training data and the pure emulation errors of the GP regression. In this section, we estimate the amplitude of both sources of errors.

4.1 Systematic and statistical simulation errors

In this section, we study the impact of the finite simulation volume and mass resolution on the matter power spectrum boost. In particular, we show that by computing the matter power spectrum boost we get more accurate results, both at large and small scales, than with the full power spectrum. We use the simulation suite described in Section 3.4.

4.1.1 Small scale convergence: impact of the mass resolution

We start by comparing the measured matter power spectra and boosts from the different boxes at small scales. In this test we use the HR simulations as a reference. For the SR, LR, and VLR boxes we compute the quantity

$$\Delta_{\text{res}}(z, k) = \frac{Y - Y_{\text{HR}}}{Y_{\text{HR}}}, \qquad (22)$$

where Y is a placeholder for the power spectrum P or the boost B. The results at $z = 0$ for the F5 model are shown in Fig. 3. The left panel shows Δ_{res} for the total matter power spectrum. The resolution drop in the SR box starts around $k \simeq 2h\,\text{Mpc}^{-1}$. At smaller scales the SR power spectrum is systematically underestimated, reaching an error of ~ 10 per cent at $k = 10\,h\,\text{Mpc}^{-1}$. Such finite resolution effects are well known. They are in agreement with previous ΛCDM simulations run with RAMSES and the same mass resolution (Rasera et al. 2022). The fluctuations at larger scales are mainly due to the large cosmic variance of the HR box and the difference in realizations from one box-size to another.

We now turn to the power spectrum boost. The corresponding resolution error Δ_{res} is shown in the right panel of Fig. 3. The lower resolution runs tend to overestimate the boost at intermediate to small scales. In the case of the SR box, this effect induces an error smaller than 2 per cent. As explained in Li et al. (2012b), a low resolution simulation is not able to properly resolve the screening mechanism in

high density environments. As a consequence, the clustering in $f(R)$ gravity, and therefore the matter power spectrum, is overestimated in a scale-dependent manner. In addition to this overestimation at intermediate to small scales, the LR and VLR boosts show a severe drop at the very small scales. However, this drop is not visible in the SR run, where $|\Delta_{\text{res}}|$ is smaller than 1 per cent at $k = 10h\,\text{Mpc}^{-1}$. In any case, this comparison shows that the mass resolution error is reduced by almost an order of magnitude when going from the total power spectrum in $f(R)$ gravity to its boost with respect to ΛCDM.

In addition to the wavenumber k, the resolution errors also depend on redshift and the parameter f_{R_0}. This is shown in Fig. 4 for the SR simulation box. The top left panel gives Δ_{res} for the three models F4, F5, and F6 and at two different redshifts $z = 0, 2$. The overestimation of the boost at intermediate scales is more important for the F4 model and vanishes for F6. Indeed, the screening mechanism enters into play in denser environments for higher values of $\left|f_{R_0}\right|$. At $z = 2$, the power spectrum boost suffers from a resolution drop at very small scales, which also depends on the value of f_{R_0}. Indeed, at $z = 2$, $f(R)$ gravity has just started having an effect on matter clustering. This effect being small, the lower resolution simulations are not able to resolve it, while the higher resolution ones predict it more accurately. On the other hand, the overestimation at intermediate scales is smaller.

The remaining panels of Fig. 4 give $|\Delta_{\text{res}}|$ at all redshifts ($0 < z < 2$) and scales of interest for the models F4, F5, and F6. For F6 the resolution error on the boost is smaller than 1 per cent most of the time and smaller than 2 per cent virtually for all redshifts ($0 < z < 2$) and scales considered ($k < 10h\,\text{Mpc}^{-1}$). In the case of F5, the accuracy is better than 3 per cent in most of the $z - k$ plane. However, the resolution error in the $z \gtrsim 1$ and $k \gtrsim 7h\,\text{Mpc}^{-1}$ region is of the order of $3 - 6$ per cent. Finally, for the F4 model the accuracy is of the order of $1 - 2$ per cent in most of the $z - k$ plane and only reaches the 3 per cent level at very low redshifts and small scales as well as at high redshifts and very small scales. Overall, we estimate that the systematic resolution errors in our power spectrum boost measurements are always smaller than 3 per cent in the range $0 < z < 2$ and $k \lesssim 7\,h\,\text{Mpc}^{-1}$.

4.1.2 Large scale convergence: impact of the simulation box length

We now perform the same kind of analysis but focusing on large scales. This time the reference box-size is the VLR and we compute for the LR, SR, and HR simulation boxes the quantity

$$\Delta_{\text{vol}}(z, k) = \frac{Y - Y_{\text{VLR}}}{Y_{\text{VLR}}}, \qquad (23)$$

where, as in the previous section, Y stands for the power spectrum P or its boost B.
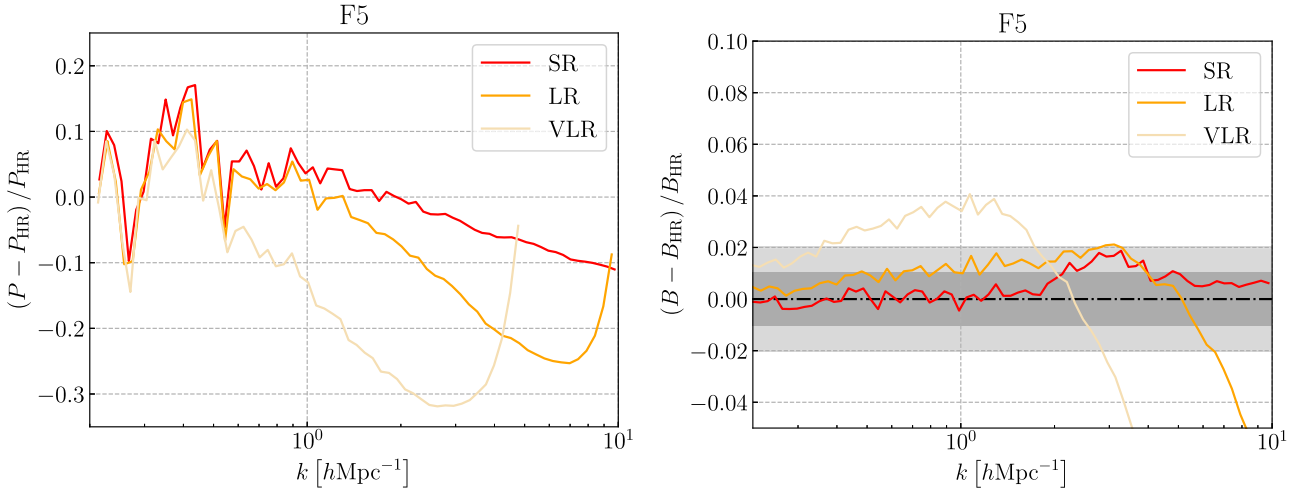
Figure 3. Results of the mass resolution convergence test at $z = 0$ for the F5 model. The HR box is used as a reference and the measurements from the SR, LR, and VLR boxes are compared to it. The left panel shows the relative difference of the total matter power spectrum. The systematic drop of the SR power spectrum starts around $k \simeq 2h\,\mathrm{Mpc}^{-1}$ and reaches the 10 per cent level at $k = 10h\,\mathrm{Mpc}^{-1}$. The right panel shows the same comparison for the matter power spectrum boost. This time the resolution error in the SR box remains under the 2 per cent level at all considered scales. The mass resolution errors are up to an order of magnitude smaller in the power spectrum boost than in the full power spectrum.

The results for the model F5 at $z = 0$ are shown in Fig. 5. At the largest scales available for the SR and HR resolutions, the relative difference on the power spectrum is of the order of 50 per cent and 100 per cent, respectively. However, when considering the boost, the large scale errors remain under the 1 per cent level even for the fundamental mode of the HR box. We have checked that this results hold for the three models F4, F5, and F6 and all redshifts $0 < z < 2$.

With this test, we are really probing a combination of cosmic variance and finite volume errors. However, the purpose here is not to carry a detailed investigation of such effects, which would require a more dedicated study. These results simply show that by taking the ratio of the matter power spectrum in $f(R)$ gravity with the corresponding ΛCDM cosmology, there is a very large cancellation of large scale errors (of the order of ~ 100 per cent). Such a large cancellation happens because the $f(R)$ and ΛCDM simulations are evolved from the same initial conditions. Furthermore, the impact of $f(R)$ gravity on the matter distribution is weak at linear scales for the values of f_{R_0} considered in this work. As a consequence, the large scale random fluctuations remain almost the same in the $f(R)$ and ΛCDM simulations up to $z = 0$. This is another advantage of considering the boost instead of the full power spectrum. However, we are still limited by the fundamental mode of our simulation box and cannot resolve features in the power spectrum that are smaller than $\Delta k \sim k_{\mathrm{F}} \simeq 0.018h\,\mathrm{Mpc}^{-1}$.

4.1.3 Sample variance: impact of the number of realizations

As a complementary study to the one presented in the previous section, we assess the impact of the number of realizations on the matter power spectrum boost measurements, i.e. pure sample variance errors. In order to do so, we use the set of 30 independent realizations for our reference cosmological model F5 and the corresponding ΛCDM cosmology, run with the SR simulation box (see Section 3.4). We start by measuring the matter power spectrum for each realization of the F5 model as well as the corresponding boosts with respect to ΛCDM. These quantities, along with their means computed over the 30 realizations, are shown in the upper panels

of Fig. 6. The lower panels show the standard errors of the means computed over a number $N_{\mathrm{real}} = 1, 5, 10$ of independent realizations. They are divided by the mean over the full set of realizations in order to obtain a relative standard error.

We recover a similar result as in the last section. There is a large cancellation of cosmic variance in the matter power spectrum boost with respect to the full power spectrum. Indeed, for the boost, the relative standard error is of the order of $\mathcal{O}\,(0.01$ per cent$)$ on the largest scales available, while for the full power spectrum it is of the order of $\mathcal{O}\,(10$ per cent$)$. This is true even with a single realization. The reasons behind this are the same as explained in the last section.

The cosmic variance cancellation is weaker at small scales. In fact, Fig. 6 shows that for the boost cosmic variance is an overall increasing function of k. For our emulator, we have decided to perform $N_{\mathrm{real}} = 5$ realizations for each training model. We can see that this is enough to keep statistical errors below 1 per cent at the 3σ level at all scales. In Fig. 6, we only show the results at $z = 0$. However, we have checked that the conclusions from this section hold at all redshifts relevant for our emulator, i.e. for $0 < z < 2$.

4.2 Emulation errors

In this section, we assess the emulation errors that are purely due to the GP interpolation between training nodes. For that we make use of the sliced structure of our experimental design by following the same procedure as in Nishimichi et al. (2019). We train the emulator using only three slices from the primary SLHD and the 30 refinement models. The remaining slice (i.e. 20 cosmological models) is used as a validation sample. The advantage of sampling the parameter space with a maximin-distance SLHD is that the testing set of cosmologies covers efficiently the whole parameter space while minimizing the overlap with the training points. This way the comparison of the predictions from the emulator against the measured boosts from the validation sample gives a representative estimation of the emulation accuracy across the whole parameter space. The results from this test are presented in Fig. 7. It shows the comparison of the predictions from the emulator and the simulation measurements for the 20
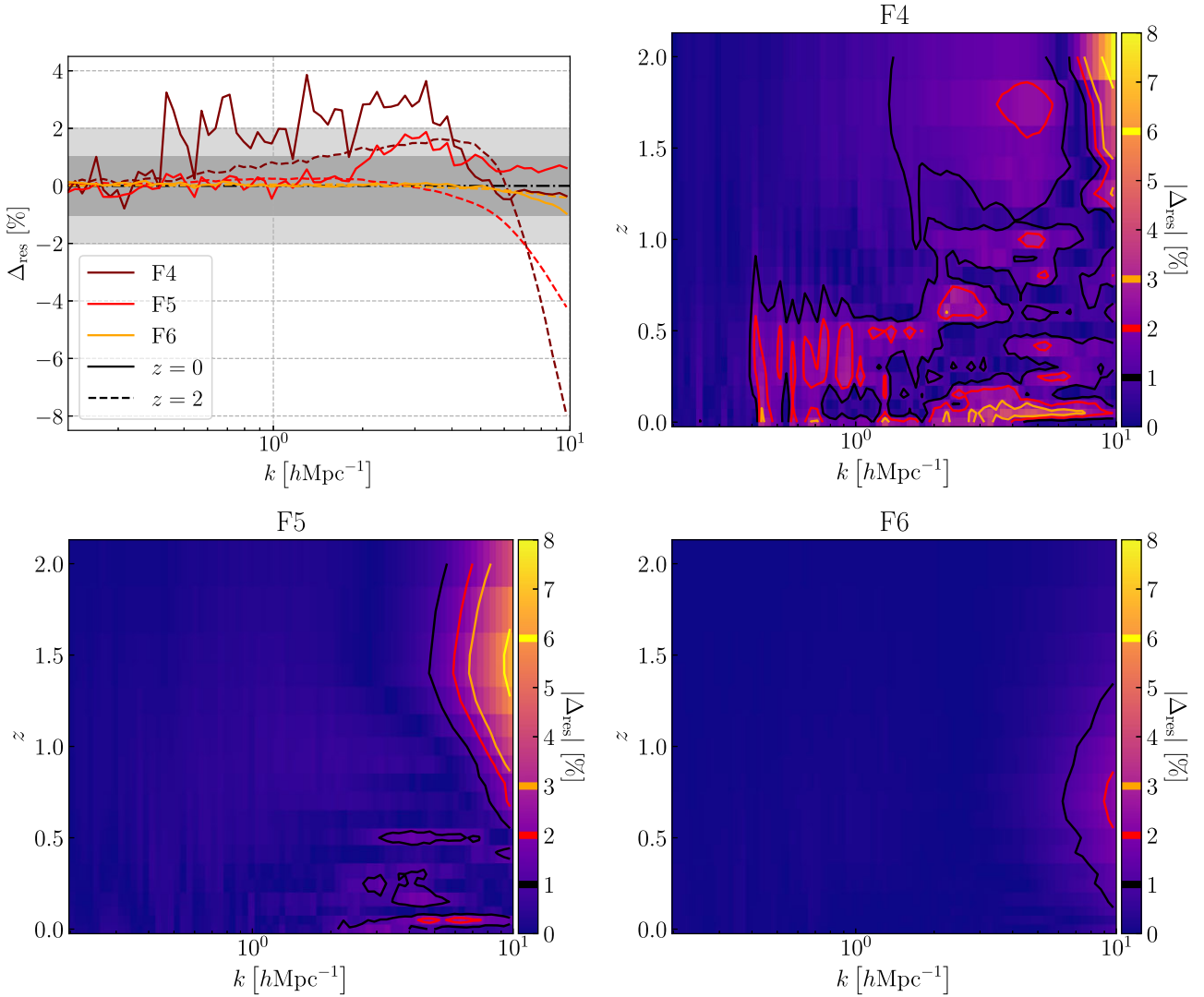
Figure 4. Results of the mass resolution convergence test for the matter power spectrum boost in the SR simulation box for different scales, redshifts and $f(R)$ models. The top left panels shows the resolution error Δ_{res} as a function of the wavenumber k for three different $f(R)$ models (F4, F5, and F6) and two redshifts ($z = 0$ with solid lines and $z = 2$ with dashed lines). The remaining panels give $|\Delta_{\text{res}}|$ as a function of redshift and scale for F4 (top right), F5 (bottom left) and F6 (bottom right). Overall, the systematic errors of the power spectrum boost are smaller than 3 per cent for all models in the range $0 < z < 2$ and $k \lesssim 7$ $h\,\text{Mpc}^{-1}$.

cosmological models of the validation sample at $z = 0$. The majority of models have relative errors that stay under the 0.5 per cent limit at all scales. In the worst cases the emulation errors are still smaller than 1 per cent. We have also computed the root-mean-squared-relative-error (RMSRE)

$$\text{RMSRE}(k) = \sqrt{\frac{1}{N_{\text{models}}} \sum_{i=0}^{N_{\text{models}}-1} \left(\frac{B_{\text{emu}}(k; \theta_i)}{B_{\text{sim}}(k; \theta_i)} - 1 \right)^2}, \quad (24)$$

where N_{models} is the number of validation models, θ_i is the cosmological parameter vector of the validation model i, B_{emu} is the boost predicted by the emulator, and B_{sim} is the boost measured from the simulation. The RMSRE, given by the black line in the lower panel of Fig. 7, is smaller than 0.5 per cent for all values of k. We have verified that these results remain valid for the 19 redshift nodes of our training set. Appendix C presents the same accuracy test at $z = 1$ and $z = 2$. The final version of our emulator is trained using the

full set of 110 models, and therefore, we expect the real emulation errors to be smaller than this estimation.

This type of comparison is an assessment of the emulation errors exclusively. To estimate the total error budget of E-MANTIS, it is important to also take into account the accuracy of the training data. In Section 4.1.1, we have estimated a maximum mass resolution error of 3 per cent in the range $0 < z < 2$ and $k \lesssim 7\,h\,\text{Mpc}^{-1}$. The pure emulation errors are therefore negligible with respect to the systematic errors in the training data. This is made possible by the large number of models in our experimental design, and, in particular, by the refinement strategy. More quantitative details are given in Appendix B. The large scale errors estimated in Sections 4.1.2 and 4.1.3 are smaller than 1 per cent and, as a consequence, they can be neglected with respect to the mass resolution systematic effects. Appendix C shows that the redshift interpolation errors are also negligible. The final accuracy of E-MANTIS is exclusively driven by the mass resolution errors in the training data. As shown in
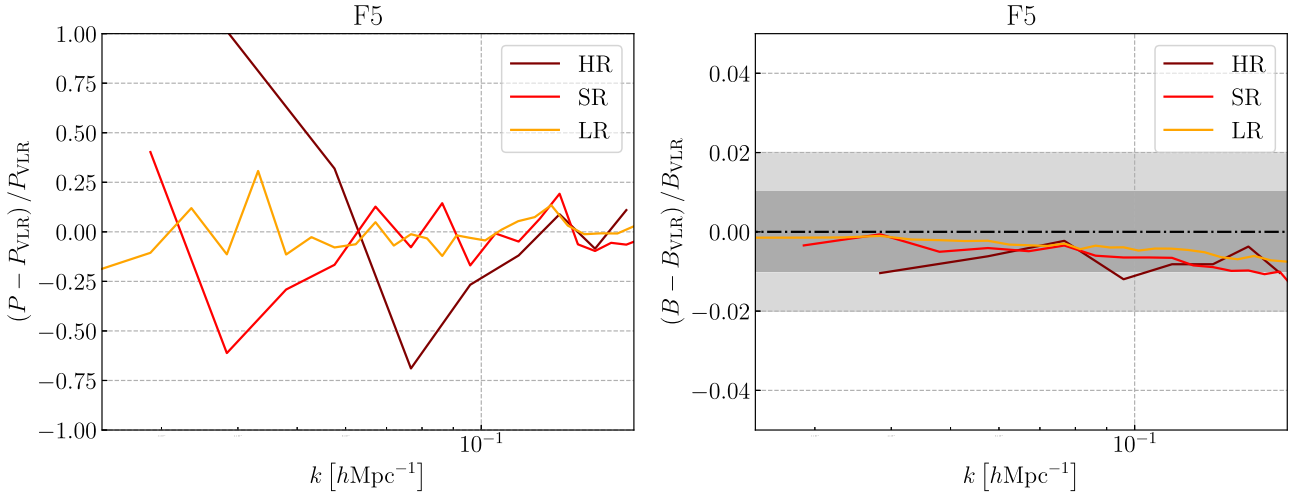
Figure 5. Results of the large-scale convergence test at $z = 0$ for the F5 model. The VLR box is used as a reference and the measurements from the HR, SR, and LR boxes are compared to it. The left panel shows the relative difference in the total matter power spectrum. The error of the SR box gets as high as 50 per cent at the largest scales covered by the emulator. The right panel shows the same test for the matter power spectrum boost. In this case, the large scale error in the SR box is smaller than 1 per cent even for the first k bin.
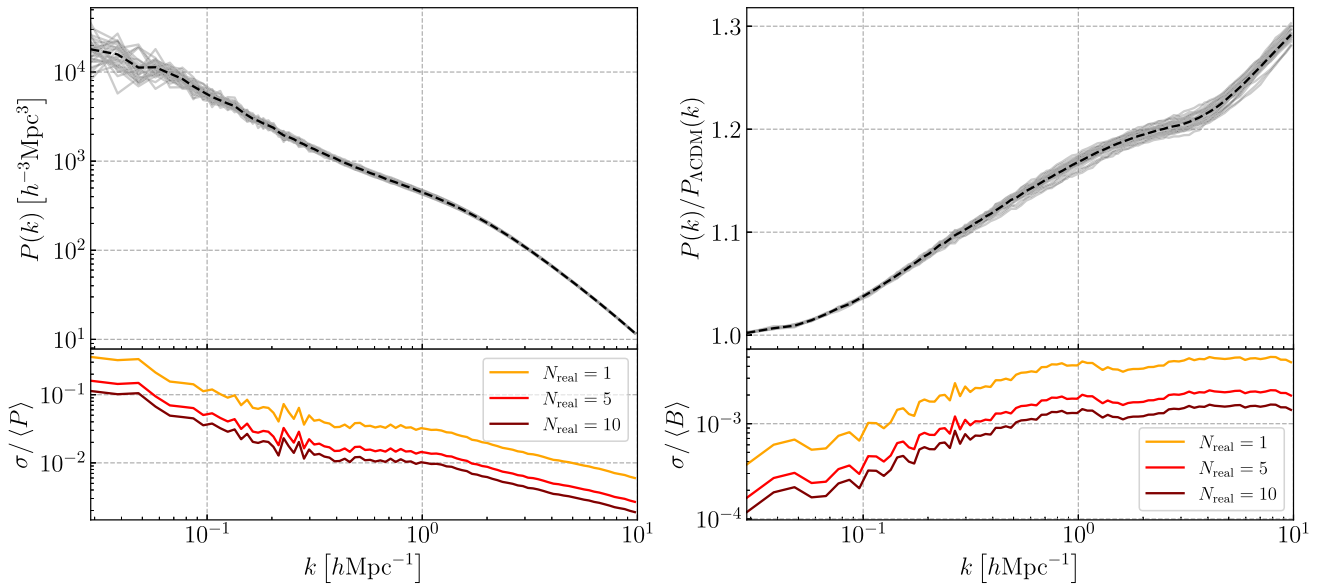
Figure 6. Upper panels: Matter power spectrum (left) for our reference cosmological model F5 and the corresponding boost (right) with respect to ΛCDM for 30 independent realizations (grey lines). The mean over the different realizations is shown as a black dashed line. Lower panels: Relative standard errors of the means computed over N_{real} independent realizations. Cosmic variance induces large scale errors of the order of $\mathcal{O}\,(10\ \mathrm{per\ cent})$ in the full power spectrum, while reducing to $\mathcal{O}\,(0.01\ \mathrm{per\ cent})$ in the case of the boost. This is true even with a single realization.

Section 4.1.1, such errors are always smaller than 3 per cent for all cosmological models considered and for $0 < z < 2$ and $\lesssim 7\,h\,\mathrm{Mpc}^{-1}$. This is a conservative bound and, in most cases, the resolution errors remain smaller than 1 per cent.

5 RESULTS

5.1 Comparison with other predictions

In this section, we compare the output of E-MANTIS to existing predictions in the literature for the matter power spectrum boost in $f(R)$ gravity. In Winther et al. (2015), the authors perform a

comparison of different modified gravity simulation codes. They show that the predicted matter power spectrum boosts in $f(R)$ gravity agrees within 1 per cent, at least up to $k \lesssim 7h\,\mathrm{Mpc}^{-1}$, for three different N-body simulation codes: ECOSMOG (Li et al. 2012a), MG-GADGET (Puchwein, Baldi & Springel 2013), and ISIS (Llinares, Mota & Winther 2014). Fig. 8 compares the power spectrum boosts they obtain with ECOSMOG with the predictions from E-MANTIS. For this comparison, we set $\Omega_m = 0.269$ and $\sigma_8 = 0.8$, which are the cosmological parameters used in Winther et al. (2015). Their simulations have a slightly better resolution than ours. Indeed, they evolve the same number of DM particles (512^3) in a smaller simulation box ($L_{\mathrm{box}} = 250$ versus $328.125h^{-1}\mathrm{Mpc}$), giving a
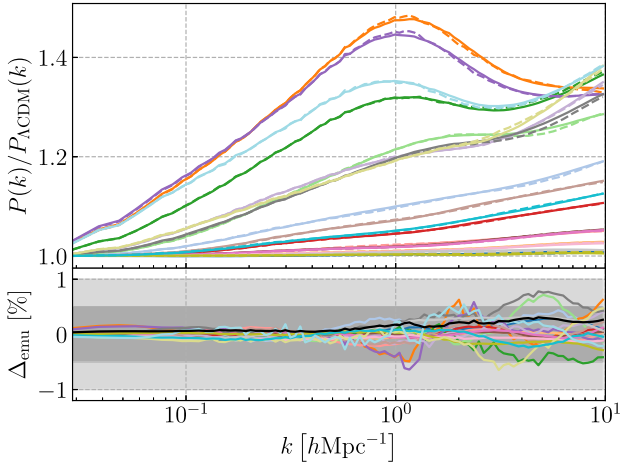
Figure 7. Results of the emulation accuracy test at $z = 0$. The emulator is trained using all available models except one slice from the primary SLHD, which is used as a validation sample. The top panel shows the matter power spectrum boost predicted by the emulator for each test model (solid lines) and the measurements from the simulations (dashed lines). The bottom panel shows the relative differences between the predictions of the emulator and the simulations. The solid black line gives the corresponding root-mean-square-relative-error (RMSRE), as defined by equation (24). The dark and light grey bands mark the 0.5 per cent and 1 per cent accuracy limits, respectively. All the validation models have an emulation error smaller than 1 per cent, even for the most extreme ones. The RMSRE is smaller than 0.5 per cent level at all the scales covered by the emulator.
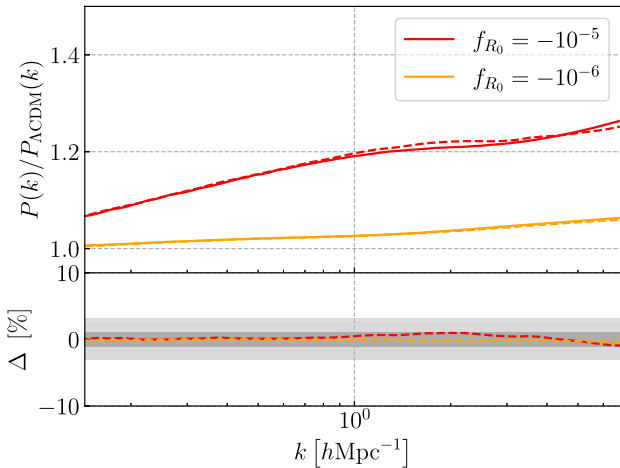
Figure 8. Comparison of the predicted boost by E-MANTIS (solid lines) with the ECOSMOG simulation data from Winther et al. (2015) (dashed lines) at $z = 0$. The top panel gives the power spectrum boosts and the lower panel their relative difference with respect to E-MANTIS. The light and dark grey bands mark the 1 per cent and 3 per cent levels, respectively. The agreement is better than 1 per cent at all scales available ($k \lesssim 7h\,\mathrm{Mpc}^{-1}$) and for both models F5 and F6. This comparison serves as a validation of E-MANTIS against an external simulation run with the same N-body code, which the authors of Winther et al. (2015) have shown to be in good agreement with other simulations codes such as MG-GADGET and ISIS.
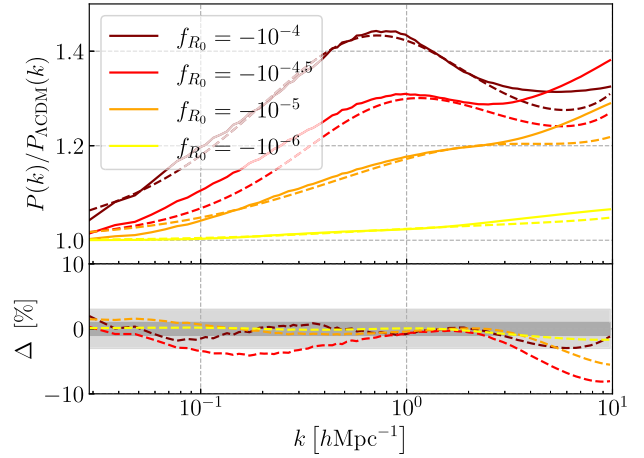

Figure 9. Comparison with the fitting formula from Winther et al. (2019) at $z = 0$, which, like E-MANTIS, is based on N-body simulations run with ECOSMOG. The top panel shows the predicted power spectrum boosts by E-MANTIS (solid lines) and the fitting formula (dashed lines) for different values of f_{R_0}. The bottom panel gives the relative difference of the fitting formula prediction with respect to E-MANTIS. The light and dark grey bands mark the 1 per cent and 3 per cent levels respectively. For the models F4 and F6 the agreement is better than 3 per cent at all scales considered. In the case F5 there is a 3 per cent agreement for $k \lesssim 5 - 6\,h\,\mathrm{Mpc}^{-1}$ and then the difference increases up to ~ 5 per cent. For F4.5, which is not one of the fitting nodes of Winther et al. (2019), the difference is more important. The fitting formula interpolates the boost between three values of f_{R_0}, while E-MANTIS is trained on 110 different cosmological models on a 3D latin hypercube.

mass resolution of $m_{\mathrm{part}} = 8.71 \times 10^9 h^{-1}\,\mathrm{M}_\odot$ (versus an average of $m_{\mathrm{part}} \sim 2.3 \times 10^{10} h^{-1}\,\mathrm{M}_\odot$ for our emulation suite). They use a smaller refinement particle criterion ($m_{\mathrm{ref}} = 8$ versus 14) and also an additional refinement level ($l_{\mathrm{max}} = 16$ versus 15) for F5. In spite of these differences, the predictions from E-MANTIS agree within 1 per cent for both models F5 and F6 and at all available scales ($k \lesssim 7\,h\,\mathrm{Mpc}^{-1}$).

Another available prediction for the matter power spectrum boost in $f(R)$ gravity is the fitting formula introduced by Winther et al. (2019). The authors perform a fit of the boost using the ELEPHANT simulations (Cautun et al. 2018). These simulations have been done with ECOSMOG and evolve 1024^3 DM particles in a cubic box of side $L_{\mathrm{box}} = 1024\,h\,\mathrm{Mpc}^{-1}$ with a mass resolution of $7.78 \times 10^{10} h^{-1}\,\mathrm{M}_\odot$. The fitting formula is calibrated from three different $f(R)$ models: $|f_{R_0}| = \{10^{-4}, 10^{-5}, 10^{-6}\}$. The other cosmological parameters are kept fixed. We compare the predicted boost from E-MANTIS to the fitting formula for the models F4, F5, and F6. The accuracy of the fitting formula should be the best for these three models, since they are the fitting nodes. We also carry a comparison for F4.5 in order to test the interpolation of the fitting formula. We set $\Omega_m = 0.281$ and $\sigma_8 = 0.82$, which are the values used in Winther et al. (2019). Fig. 9 gives the resulting comparison. The agreement between both predictions is at the 3 per cent level at all scales for F4 and F6. For F5 the 3 per cent agreement holds for $k \lesssim 5$–$6\,h\,\mathrm{Mpc}^{-1}$. At smaller scales, the prediction of the fitting formula drops below that of E-MANTIS, up to a difference of ~ 5 per cent. One possible explanation for such a difference could be that our training simulations have a higher mass resolution than the ELEPHANT simulations. In any case, such differences are well within the estimated 3 per cent accuracy of E-MANTIS combined with that of the fitting formula. Indeed, in Winther et al. (2019), the authors
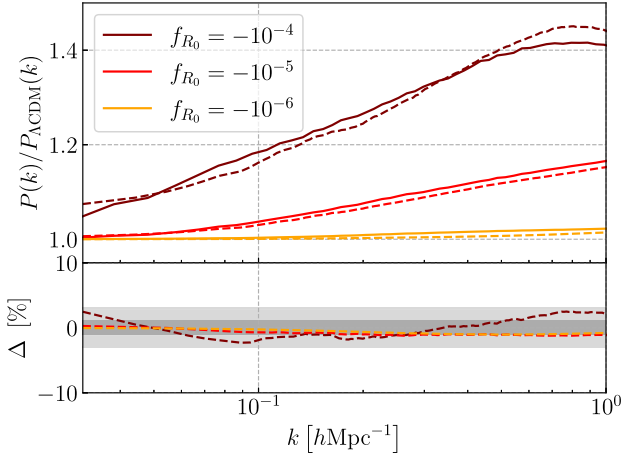
Figure 10. Comparison with the MGEMU emulator (Ramachandra et al. 2021), based on COLA simulations, at $z = 0$. The top panel shows the predicted power spectrum boosts by E-MANTIS (solid lines) and MGEMU (dashed lines) for different values of f_{R_0}. The bottom panel gives the relative difference of the MGEMU predictions with respect to E-MANTIS. The light and dark grey bands mark the 1 per cent and 3 per cent levels, respectively. The models F6 and F5 agree to the 1 per cent level at all scales considered. In the case of F4, the agreement is at the 3 per cent level. We restrict the comparison to scales $k \leq 1h\,\mathrm{Mpc}^{-1}$, which is the limit of validity of the COLA predictions.

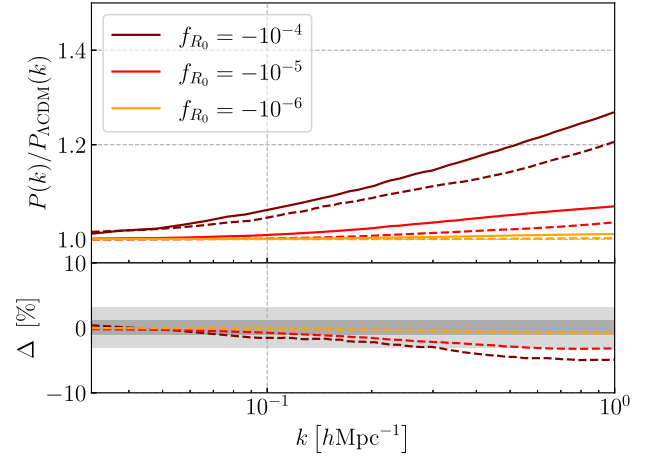**Figure 11.** Same as Fig. 10, but at $z = 1$. We notice that MGEMU underestimates the power spectrum boost with respect to E-MANTIS, by up to ~ 5 per cent for F4 and ~ 3 per cent for F5. As pointed out in Ramachandra et al. (2021), the COLA simulations implement an approximate solver for the MG screening, which produces this effect. As in Fig. 10, we limit the comparison to the scales covered by MGEMU, i.e. $k \leq 1h\,\mathrm{Mpc}^{-1}$.

cite an accuracy of 1 per cent for $k \lesssim 1h\,\mathrm{Mpc}^{-1}$ and of 5 per cent for $1 < k < 10h\,\mathrm{Mpc}^{-1}$. In the case of F4.5, the difference between both predictions is slightly larger. There is a ~ 3 per cent difference at intermediate scales. Additionally, the prediction from the fitting formula drops by ~ 8 per cent at $k = 10h\,\mathrm{Mpc}^{-1}$. Both predictions are based on simulations run with the same N-body code and using a similar numerical resolution. However, we expect the emulation of the boost by E-MANTIS to be more accurate than that of the fitting formula. Indeed, E-MANTIS is trained on 110 different cosmological models on a 3D latin hypercube, while in Winther et al. (2019) they simply interpolate between three values of f_{R_0}.

We perform a third comparison with the MGEMU emulator introduced in Ramachandra et al. (2021). In this work, the authors build an emulator for the matter power spectrum boost in $f(R)$ gravity using COmoving Lagrangian Acceleration (COLA) simulations. We refer the reader to the aforementioned paper for more specific details. The comparison at $z = 0$ between both emulators is given in Fig. 10. We set set $n_s = 0.9649$ and $n = 1$ in the COLA emulator to match the values used in our training data set. We also use $\Omega_m = 0.3153$ and $\sigma_8 = 0.8111$, which are the best-fitting values from Planck Collaboration VI (2020). We only consider scales $k \leq 1h\,\mathrm{Mpc}^{-1}$ since the usage of the COLA emulator for larger values of k is not recommended by its authors. Indeed, the COLA simulations are less accurate than N-body ones at small scales. Both emulators agree at the 1 per cent level for the models F5 and F6. For F4, the agreement is at the 3 per cent level. It is important to keep in mind that the model F4 is at the edge of the emulated parameter space of both emulators, which might lead to larger emulation errors in both predictions. The COLA emulator has a reported accuracy of 5 per cent and, therefore, the differences with E-MANTIS are well within the expected error bars. In Fig. 11, we show the same comparison, but at a redshift of $z = 1$. It can be seen that MGEMU tends to underestimate the power spectrum boost with respect to EMANTIS. The difference reaches the ~ 3 per cent level for F5 and ~ 5 per cent for F4. In Ramachandra et al. (2021),

the authors explain that the COLA simulations used to train MGEMU implement an approximate method to solve for the screening of the $f(R)$ extra field. This produces an underestimation of the boost at high redshifts, with respect to N-body simulations, which solve for the full MG equation of motion.

We perform a last comparison with the FORGE emulator introduced by Arnold et al. (2022). This emulator is built using cosmological N-body simulations run with the modified gravity AREPO code (Arnold, Leo & Li 2019). FORGE emulates the power spectrum boost in $f(R)$ gravity with respect to HALOFIT (Takahashi et al. 2012) instead of a N-body ΛCDM prediction. It would be incorrect to directly compare the outputs of both emulators, since they do not emulate the same base quantity. The first cosmological model of FORGE [node-0 from table 1 of Arnold et al. (2022)] is a ΛCDM model, which we can use to get $P_{\Lambda\mathrm{CDM}}/P_{\mathrm{HALOFIT}}$. We use this quantity to rescale the boost predicted by FORGE and compare it to the output of E-MANTIS. In order to minimize the emulation errors when removing the HALOFIT component from the FORGE prediction, we set the ΛCDM parameters to the cosmology of the FORGE node-0: $\Omega_m = 0.31315$, $\sigma_8 = 0.82172$, and $h = 0.6737$. Fig. 12 shows the comparison between FORGE and E-MANTIS for different $f(R)$ models at $z = 0$. For the model F6, there is a 1 per cent agreement at all the emulated scales. However, for models F5 and F4.5, the difference is larger, reaching 3 per cent and 5 per cent, respectively at intermediate to small scales. The reported accuracy of FORGE is of the order of 2 per cent for most of the models covered by it and in particular the ones used in this comparison. Combined with the 3 per cent estimated accuracy of E-MANTIS, the total error budget in this test is of 3.6 per cent. In the case of F4.5, the difference between both emulators is larger. The fact that we have used the FORGE node-0 to remove the HALOFIT component from its prediction might introduce its own additional source of errors. However, this is the best comparison we can perform given the output quantities of both emulators. Moreover the FORGE emulator does not benefit from the cancellation of numerical systematics that is obtained by considering the boost instead of the power spectrum. Additionally, the F4.5 models is at the edge of the FORGE emulated space, which might lead to increased emulation errors. Furthermore, there might be some systematic difference between the N-body predictions of
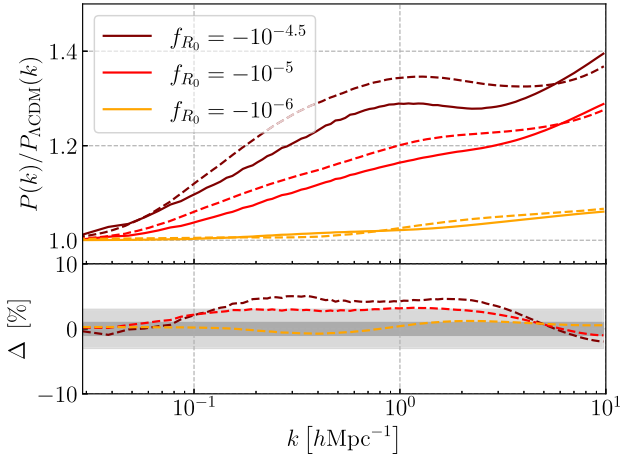
Figure 12. Comparison with the FORGE emulator (Arnold et al. 2022) at $z = 0$. The top panel shows the power spectrum boost for different $f(R)$ as predicted by FORGE (dashed lines) and E-MANTIS (solid lines). The bottom plot gives the relative difference of the FORGE prediction with respect to E-MANTIS. The light and dark grey bands mark the 1 per cent and 3 per cent levels, respectively. For F6, there is a 1 per cent level agreement at all scales considered. In the case of F5 and F4.5, there is a 3 per cent and 5 per cent difference respectively at intermediate to small scales. The FORGE predictions have been corrected to compensate for the effect of HALOFIT, since FORGE predicts $P_{f(R)}/P_{\mathrm{HALOFIT}}$ instead of the boost $P_{f(R)}/P_{\Lambda\mathrm{CDM}}$, which is the robust quantity of interest in this article (see Section 3.2.1).

ECOSMOG and AREPO that are not accounted for in the estimated accuracy of the emulators. To conclude, the two emulators are complementary as they are sensitive to different kind of systematics. A more in depth study would require a detailed comparison of both simulation codes, which is out of the scope of this work.

Overall there is a good agreement between the predictions of E-MANTIS and other existing predictions from the literature.

5.2 Example of application

In this section, we present a toy example of how to use the E-MANTIS emulator. We aim at recovering the cosmological parameters from a $f(R)$CDM simulation by running a Markov Chain Monte Carlo (MCMC) on the measured matter power spectrum.

The emulator presented in this work is able to compute the boost of the matter power spectrum $B(k) = P_{f(R)}/P_{\Lambda\mathrm{CDM}}$. In order to get the full power spectrum in $f(R)$ gravity, such quantity needs to be combined with a non-linear ΛCDM prediction. For the purpose of this section, we use the BACCO emulator (Angulo et al. 2021). Other ΛCDM emulators such as COSMICEMU (Moran et al. 2022) or EUCLIDEMULATOR2 (Euclid Collaboration 2021) would also be valid alternatives.

In order to run the MCMC, we need simulation data to be used as the data vector. The volume and mass resolution of the simulations presented throughout this work have been designed to predict the matter power spectrum boost with a few per cent errors. However, they are not able to estimate the full power spectrum with the same level of accuracy. Fig. 6 shows that reducing statistical errors on the full power spectrum under the 1 per cent level at scales $0.1h\,\mathrm{Mpc}^{-1} < k < 1h\,\mathrm{Mpc}^{-1}$ would require performing at least $\mathcal{O}(100)$ $f(R)$ realizations with the same box size. Alternatively, an accurate modelling of the non-linear covariance at those scales would avoid getting biased constraints even in the presence of

Table 4. Table summarizing the truth values and the flat prior ranges used in the MCMC of Section 5.2 for each parameter. The best-fitting values found by the MCMC, with their 1-σ confidence levels, are also given.

Parameter	True value	Prior range	Best fit
$-\log\left|f_{R_0}\right|$	5	[4, 7]	$5.04^{+0.11}_{-0.17}$
Ω_{m}	0.3071	[0.2365, 0.3941]	$0.307^{+0.013}_{-0.016}$
σ_8	0.8224	[0.73, 0.9]	0.8210 ± 0.0089
n_s	0.96641	[0.92, 1.01]	0.965 ± 0.024
h	0.6803	[0.6, 0.8]	0.706 ± 0.047
Ω_b	0.048446	[0.04, 0.06]	$0.0513^{+0.0072}_{-0.0038}$

significant errors in the data vector. However, computing a fully non-linear covariance requires an even higher number of simulations, at least of the order of $\mathcal{O}\left(10^3 - 10^4\right)$ (Blot et al. 2015). Such a numerical effort goes beyond the scope of this section. We resort to building a composite power spectrum in $f(R)$ gravity using the same boost strategy as in the construction of our emulator. We compute the full power spectrum in ΛCDM using a series of 384 independent realizations, each evolving 512^3 dark matter particles in a simulation volume of $\left(328.125\,h^{-1}\mathrm{Mpc}\right)^3$. This set of simulations is part of an ongoing simulation series which will be presented in a future work. The values of the cosmological parameters used for these simulations are given in Table 4. These simulations are run with the same numerical parameters previously described (see Table 2). Additionally, we perform five realizations with the same volume and mass resolution for the F5 model in order to compute the power spectrum boost. Fig. 6 shows that the statistical errors on the power spectrum boost with five realizations is much smaller than 1 per cent. We build a full power spectrum for F5 by multiplying the ΛCDM power spectrum by the F5 boost. For simplicity, we use the same binning in k as in E-MANTIS and described in Section 3.2.4. Such procedure gives us a full power spectrum in $f(R)$CDM cosmology with per cent level accuracy, while minimizing the numerical effort.

We assume a linear Gaussian covariance for the power spectrum data. Additionally, it is important to also take into account the errors from the emulator predictions, which at scales small enough dominate over the Gaussian covariance. The authors of Angulo et al. (2021) estimate the BACCO emulator to be accurate at the 2 per cent level in the case of ΛCDM. According to the convergence studies presented in Section 4, we consider a 3 per cent error on the power spectrum boost predicted by E-MANTIS. The total relative systematic error introduced by the emulators has therefore a value of $\sigma_{\mathrm{sys}} = 0.036$, which we add quadratically to the diagonal Gaussian covariance. The final covariance between two modes k_1 and k_2 takes the form

$$\mathrm{cov}(k_1, k_2) = \left[\frac{2}{N_{k_1}} + \sigma_{\mathrm{sys}}^2\right] P^2(k_1)\delta_{k_1 k_2}, \qquad (25)$$

where $N_{k_i} = k_i^2 \Delta k_i V/(2\pi^2)$ is the number of independent Gaussian variables in the bin centred on k_i and of width Δk_i, V is the simulation volume and $\delta_{k_1 k_2}$ is Kronecker delta. The first term inside the brackets corresponds to the linear Gaussian contribution [see for example Scoccimarro, Zaldarriaga & Hui (1999)].

We combine the binned matter power spectrum data described earlier, the predictions from the emulators and the covariance from equation (25) into a Gaussian likelihood function. We assume flat priors for the parameters $\left\{-\log\left|f_{R_0}\right|, \Omega_{\mathrm{m}}, \sigma_8, n_s, h, \Omega_b\right\}$. The variations we explore for each parameter correspond to the widest range allowed by the intersection of E-MANTIS and BACCO and are given in Table 4. We use the Python package EMCEE (Foreman-
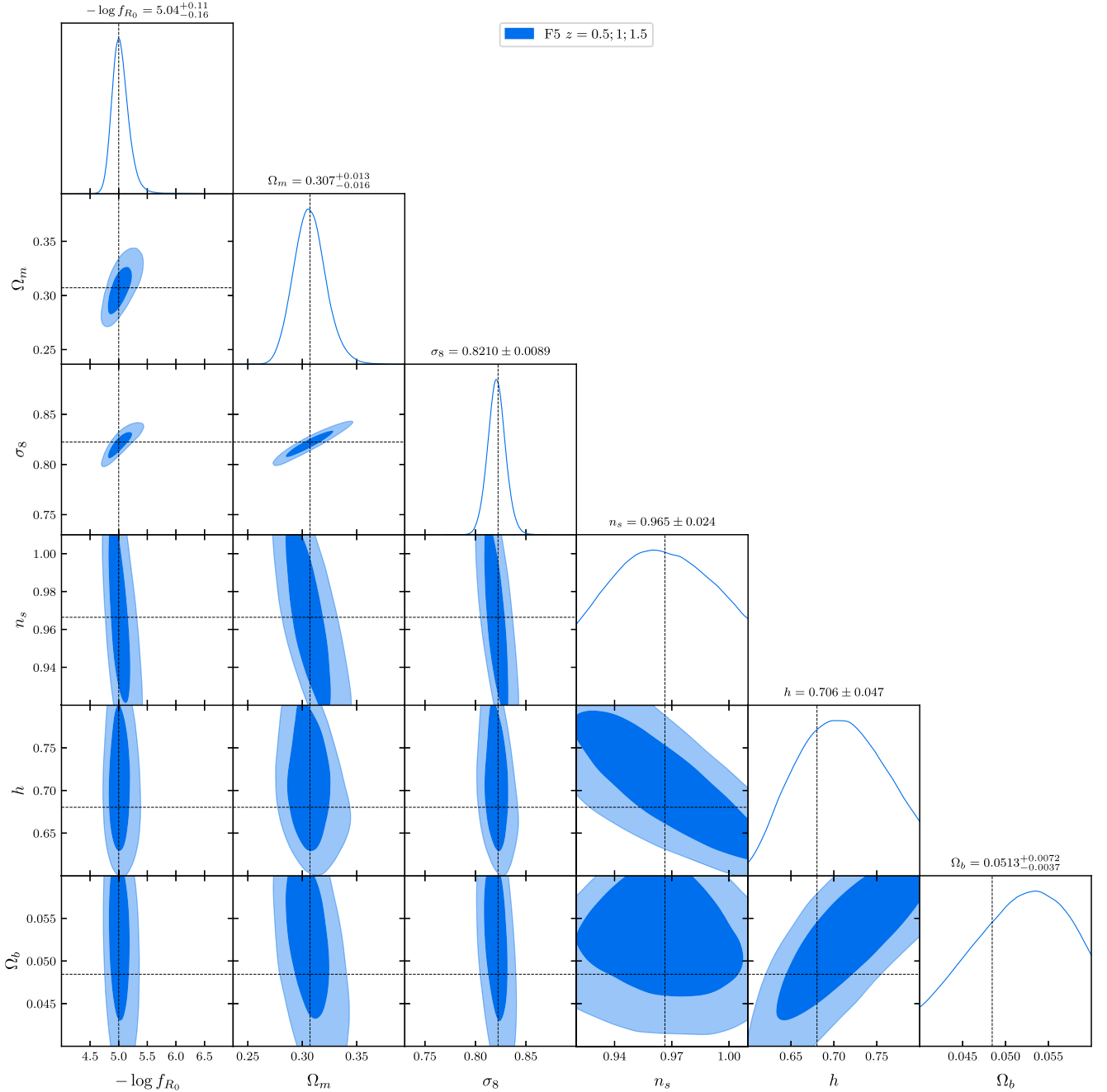
Figure 13. Marginalized 1D and 2D posterior distributions obtained from the MCMC described in Section 5.2 for the six cosmological parameters $\left\{-\log|f_{R_0}|, \Omega_m, \sigma_8, n_s, h, \Omega_b\right\}$. The dark and light contours give the 1-σ and 2-σ confidence levels, respectively. The fiducial values for each parameters, marked by the black dotted lines, always lay within the 1-σ contours. The matter power spectrum $P(k)$ is obtained from simulation snapshots at $z = 0.5, 1, 1.5$, while we use E-MANTIS for the predictions of the boost combined with BACCO for the ΛCDM predictions.

Mackey et al. 2013) to run the MCMC. We employ the 'stretch move' sampling method from Goodman & Weare (2010), which is the default algorithm used by EMCEE.

Our initial tests show that when using a single redshift the posterior distribution exhibits small multimodal features. These are due to the degeneracy between the parameters f_{R_0}, Ω_m, and σ_8. In order to break such degeneracies, we combine data from multiple redshifts by assuming that the data sets at different redshifts are independent from each other. E-MANTIS is able to make predictions up to $z = 2$. However, BACCO is limited to $z < 1.5$. In practice, we combine the

matter power spectra from $z = 0.5, 1$, and 1.5 into a single data vector and repeat the covariance from equation (25) for each redshift.

We run an MCMC with 256 walkers or independent chains and around ~100 000 steps for each one of them, which took ~10 h to complete on a single thread of a laptop 11th Gen Intel Core i7-1165G7 @ 2.80Ghz cpu. The resulting posterior distribution of the cosmological parameters is shown in Fig. 13. The true values for each parameter, marked with the dotted black lines, always lay within the 1-σ confidence contours (see also Table 4). This simple toy use case illustrates how to use the boost predicted by E-MANTIS

in combination with an independent ΛCDM prediction. It serves as an additional cross-validation of the newly built emulator.

6 CONCLUSIONS

In this work, we present a new emulator, named E-MANTIS, that predicts the matter power spectrum boost in $f(R)$ gravity with respect to ΛCDM, $B(k) = P_{f(R)}/P_{\Lambda CDM}$. We run N-body cosmological simulations with the modified gravity code ECOSMOG in order to predict such quantity for different cosmological models. Building an emulator for the boost instead of the raw power spectrum presents several advantages, which we summarize here.

Section 3.2.2 shows that the power spectrum boost mainly depends on only three cosmological parameters: f_{R_0}, Ω_m, and σ_8. Indeed, the variation of $B(k)$ with h, n_s, and Ω_b is smaller than 1 per cent up to $k = 10h\,\mathrm{Mpc}^{-1}$, even for variations of these parameters of ± 25 per cent around their best-fitting values from Planck Collaboration VI (2020). In Appendix A, we have checked that simultaneous variations of those parameters have an effect which remains smaller than $2 - 3$ per cent in the worst cases, mostly around the edges of the parameter space. This error is less than the maximum systematic error in our training data. Therefore, we can neglect the influence of n_s, h, and Ω_b on the boost when building the emulator with a few per cent accuracy. We then only need to sample a 3D parameter space with N-body simulations instead of a 6D one. As a consequence, the required number of simulations to reach a given emulation accuracy is reduced.

Section 4.1 shows that the boost is less affected by statistical and systematic errors than the raw power spectrum. Indeed, there is a very large cancellation of cosmic variance and large scale errors. This is due to the fact that the $f(R)$ and ΛCDM simulations are evolved from the same initial conditions and that $f(R)$ gravity has a weak effect at large scales. The same kind of cancellation happens with the systematic errors at small scales due to the finite mass resolution of the simulations. In the end, we require less accurate simulations to resolve the boost to a given precision that what would be needed for the raw power spectrum. We find that with our simulations, which cover an effective volume of $(560\,h^{-1}\mathrm{Mpc})^3$ with a particle mass resolution of $m_{\mathrm{part}} \sim 2 \times 10^{10}h^{-1}\,\mathrm{M}_\odot$, we are able to compute the power spectrum boost with an accuracy better than 3 per cent for $0.03\ h\,\mathrm{Mpc}^{-1} < k < 7\ h\,\mathrm{Mpc}^{-1}$ and $0 < z < 2$. In practice, as shown in Fig. 4, the systematic error on the power spectrum boost measurement depends on the value of f_{R_0}, redshift and scale. However, the 3 per cent estimation is a conservative value for most models, scales, and redshifts. Even though in most cases the actual accuracy is better than 1 per cent.

We sample the 3D cosmological parameter space $\{f_{R_0}, \Omega_m, \sigma_8\}$ with 80 models selected from a SLHD using the method introduced by Ba, Myers & Brenneman (2015). We also sample an additional 30 models in the regions of the parameter space where the emulation errors are the highest. For each of the 110 cosmological models we run five independent realizations both in $f(R)$ gravity and the corresponding ΛCDM cosmology. From this training data we build an emulator using a GP regression. Section 4.2, along with Appendix C, shows that the average emulation errors are smaller than 0.5 per cent for all scales and redshifts. The accuracy of E-MANTIS is therefore dominated by the systematic errors in the training data.

Overall, building an emulator for the boost instead of the raw power spectrum greatly reduces our computational needs. This is particularly relevant for $f(R)$ simulations, since they are more time consuming than ΛCDM ones. Indeed, we find that with our

specifications a $f(R)$ simulation is between two to ten times slower than a ΛCDM one, depending on the value of f_{R_0}.

In Section 5.1, we have compared the predictions of E-MANTIS to other existing predictions for the matter power spectrum boost in the literature. First, we have considered the N-body simulations results from Winther et al. (2015), where they compare the boost predicted by different state-of-the-art modified gravity N-body codes. This comparison serves as a validation, and shows that the predictions of E-MANTIS are in good agreement with the outputs of ECOSMOG, ISIS, and MG-GAGDET. Then, we have compared E-MANTIS to the fitting formula from Winther et al. (2019), which is calibrated on N-body simulations run with the same code and using a similar resolution. We have seen that there is a good agreement for F4, F5, and F6, which are the models the fitting formula interpolates from. The E-MANTIS emulator improves upon the predictions of this fitting formula by expanding the interpolation to the parameters Ω_m and σ_8 and by using 110 models (instead of 3) to emulate the boost.

We have also compared our predictions to the MGEMU emulator introduced in Ramachandra et al. (2021), which is based on COLA simulations. MGEMU is able to give predictions taking into account more cosmological parameters, such as n_s and the modified gravity parameter n. However, the range of validity of their predictions is restricted to scales $k \leq 1h\,\mathrm{Mpc}^{-1}$, since COLA simulations are less accurate than N-body. The overall agreement is good at $z = 0$ and for those scales. Additionally, we have seen in our comparison that MGEMU tends to underestimate the boost at higher redshifts. The authors of Ramachandra et al. (2021) explain that this is due to the approximate implementation of the screening mechanism in the COLA simulations. This effect can be larger than the error induced by neglecting n_s alone. Therefore, E-MANTIS can give more accurate predictions in the case of $n = 1$ even though the effect of n_s is not taken into account.

Finally, we have performed a comparison with the FORGE emulator presented in Arnold et al. (2022). To date, this is the only emulator for the matter power spectrum boost in $f(R)$ gravity based on N-body simulations. However, they use a different N-body code, the modified gravity version of AREPO (Arnold, Leo & Li 2019). They consider a slightly larger parameter space than E-MANTIS, since they also include the parameter h. However, we have shown in Section 3.2.2, that the largest error comes from neglecting n_s. Another difference, is that E-MANTIS emulates the boost with respect to a ΛCDM simulation, which we have shown in Section 4.1 allows for a large cancellation of systematic and statistical errors. On the other side, FORGE emulates the boost with respect to HALOFIT and therefore does not benefit from this effect. Our comparison shows a good agreement between both emulators for F6. However, there seems to be a slight disagreement for larger values of the f_{R_0} parameter, which is not expected given the emulation errors of both emulators. Such difference could come from the usage of HALOFIT as a reference which makes the comparison difficult, an underestimate of the error bars in one of the emulators, or a systematic difference between ECOSMOG and AREPO [which was not included in the comparison of Winther et al. (2015)]. A more detailed investigation would be required to solve this issue. Future generation surveys will likely require running their analyses using complementary theoretical predictions, in order to avoid getting biased cosmological constraints. For this reason, it is important that several emulators based on different numerical codes, such as FORGE and E-MANTIS, are available to the community.

In Section 5.2, we have used E-MANTIS to recover the cosmological parameters of a numerical simulation by running an MCMC on its matter power spectrum. This simple toy example illustrates how to combine the boost predicted by E-MANTIS with an independent

ΛCDM prediction in order to get the full matter power spectrum prediction in *f(R)*CDM cosmology.

In this work we have ignored the effect of baryonic physics on the matter distribution. Its impact on the matter power spectrum boost in *f(R)* gravity is non-negligible for scales $k \gtrsim 2\,h\,\mathrm{Mpc}^{-1}$ (Arnold, Leo & Li 2019). Because of this, the E-MANTIS predictions should only be used on their own for larger scales. In order to get accurate predictions for smaller scales, some correction taking into account baryonic feedback is required. Currently, there are no available emulators for this type of correction in *f(R)* gravity. However, Arnold, Leo & Li (2019) has shown that the full physics *f(R)* power spectrum boost can be recovered by applying a GR baryonic correction to the dark matter only prediction. Therefore, the E-MANTIS boost can be combined with a ΛCDM baryonic emulator, such as BCEMU (Giri & Schneider 2021), in order to get accurate predictions at scales smaller than $k \gtrsim 2\,h\,\mathrm{Mpc}^{-1}$. Alternatively, if a baryonic correction emulator in *f(R)* gravity is developed in the near future, the non-linear dark matter only boost predicted by E-MANTIS (or another equivalent tool) will still be required in order to get the full matter power spectrum. For this reasons, we believe that the emulator presented in this work could be a useful tool in order to constrain *f(R)* gravity with the next-generation of weak lensing surveys, such as LSST and *Euclid*.

In the future, we plan to extend E-MANTIS to other observables, such as dark matter halo statistics and profiles. We also plan to extend the current parameter space and cover alternative dark energy models.

ACKNOWLEDGEMENTS

DATA AVAILABILITY

The emulator presented in this work, as well as the training power spectrum boosts, are publicly available at the following address: https://doi.org/10.5281/zenodo.7738362. The rest of the data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Agarwal S., Abdalla F. B., Feldman H. A., Lahav O., Thomas S. A., 2014, MNRAS, 439, 2102
Amendola L., Tsujikawa S., 2010, Dark Energy: Theory and Observations. Cambridge Univ. Press, Cambridge
Amendola L. et al., 2018, Living Rev. Relativ., 21, 2
Angulo R. E., Zennaro M., Contreras S., Aricò G., Pellejero-Ibañez M., Stücker J., 2021, MNRAS, 507, 5869
Arnold C., Leo M., Li B., 2019, Nat. Astron., 3, 945
Arnold C., Li B., Giblin B., Harnois-Déraps J., Cai Y.-C., 2022, MNRAS, 515, 4161
Ba S., Myers W. R., Brenneman W. A., 2015, Technometrics, 57, 479
Bardeen J. M., 1980, Phys. Rev. D, 22, 1882
Bel J., Brax P., Marinoni C., Valageas P., 2015, Phys. Rev. D, 91, 103503
Blot L., Corasaniti P. S., Alimi J. M., Reverdy V., Rasera Y., 2015, MNRAS, 446, 1756
Blot L., Corasaniti P.-S., Rasera Y., Agarwal S., 2021, MNRAS, 500, 2532
Bose S., Hellwing W. A., Li B., 2015, J. Cosmol. Astropart. Phys., 2015, 034
Bose S., Li B., Barreira A., He J.-h., Hellwing W. A., Koyama K., Llinares C., Zhao G.-B., 2017, J. Cosmol. Astropart. Phys., 2017, 050
Brax P., Casas S., Desmond H., Elder B., 2022, Universe, 8, 11
Cataneo M. et al., 2015, Phys. Rev. D, 92, 044009
Cautun M., Paillas E., Cai Y.-C., Bose S., Armijo J., Li B., Padilla N., 2018, MNRAS, 476, 3195
Chisari N. E., Zaldarriaga M., 2011, Phys. Rev. D, 83, 123505
Clifton T., Ferreira P. G., Padilla A., Skordis C., 2012, Phys. Rep., 513, 1
Crocce M., Pueblas S., Scoccimarro R., 2006, MNRAS, 373, 369
DESI Collaboration 2016, preprint (arXiv:1611.00036)
Euclid Collaboration 2019, MNRAS, 484, 5509
Euclid Collaboration 2021, MNRAS, 505, 2840
Fiorini B., Koyama K., Baker T., 2023, preprint (arXiv:2310.05786)
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Garud S. S., Karimi I. A., Kraft M., 2017, Comput. Chem. Eng., 106, 71
Giri S. K., Schneider A., 2021, J. Cosmol. Astropart. Phys., 2021, 046
Goodman J., Weare J., 2010, Comm. App. Math. Com. Sci., 5, 65
Guillet T., Teyssier R., 2011, J. Comput. Phys., 230, 4756
Habib S., Heitmann K., Higdon D., Nakhleh C., Williams B., 2007, Phys. Rev. D, 76, 083503
Harnois-Déraps J., Hernandez-Aguayo C., Cuesta-Lazaro C., Arnold C., Li B., Davies C. T., Cai Y.-C., 2023, MNRAS, 525, 6336
Harris C. R. et al., 2020, Nature, 585, 357
Hernández-Aguayo C., Arnold C., Li B., Baugh C. M., 2021, MNRAS, 503, 3867
Hernández-Aguayo C., Ruan C.-Z., Li B., Arnold C., Baugh C. M., Klypin A., Prada F., 2022, J. Cosmol. Astropart. Phys., 01, 048
Hu W., Sawicki I., 2007, Phys. Rev. D, 76, 064004
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ivezić Ž. et al., 2019, ApJ, 873, 111
Khoury J., Weltman A., 2004, Phys. Rev. D, 69, 044026
Knabenhans M. et al., 2019, MNRAS, 484, 5509
Lawrence E., Heitmann K., White M., Higdon D., Wagner C., Habib S., Williams B., 2010, ApJ, 713, 1322
Lawrence E. et al., 2017, ApJ, 847, 50
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Li B., Zhao G.-B., Teyssier R., Koyama K., 2012a, J. Cosmol. Astropart. Phys., 1, 051
Li B., Hellwing W. A., Koyama K., Zhao G.-B., Jennings E., Baugh C. M., 2012b, MNRAS, 428, 743
Llinares C., 2018, Int. J. Mod. Phys. D, 27, 1848003
Llinares C., Mota D. F., Winther H. A., 2014, A&A, 562, A78
Mauland R., Winther H. A., Ruan C.-Z., 2023, preprint (arXiv:2309.13295)
Michaux M., Hahn O., Rampf C., Angulo R. E., 2021, MNRAS, 500, 663
Moran K. R. et al., 2023, MNRAS, 520, 3443
Nishimichi T. et al., 2019, ApJ, 884, 29
Oyaizu H., 2008, Phys. Rev. D, 78, 123523
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Planck Collaboration VI 2020, A&A, 641, A6
Prunet S., Pichon C., Aubert D., Pogosyan D., Teyssier R., Gottloeber S., 2008, ApJS, 178, 179
Puchwein E., Baldi M., Springel V., 2013, MNRAS, 436, 348
Ramachandra N., Valogiannis G., Ishak M., Heitmann K., 2021, Phys. Rev. D, 103, 123525
Rasera Y. et al., 2022, A&A, 661, A90
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. The MIT Press, Cambridge
Ruan C.-Z., Hernández-Aguayo C., Li B., Arnold C., Baugh C. M., Klypin A., Prada F., 2022, J. Cosmol. Astropart. Phys., 05, 018
Scoccimarro R., Zaldarriaga M., Hui L., 1999, ApJ, 527, 1

Sotiriou T. P., Faraoni V., 2010, Rev. Mod. Phys., 82, 451
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Teyssier R., 2002, A&A, 385, 337
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Winther H. A. et al., 2015, MNRAS, 454, 4208
Winther H. A., Casas S., Baldi M., Koyama K., Li B., Lombriser L., Zhao G.-B., 2019, Phys. Rev. D, 100, 123540

APPENDIX A: IMPACT OF VARYING MULTIPLE COSMOLOGICAL PARAMETERS ON THE MATTER POWER SPECTRUM BOOST

In Section 3.2.2, we have shown that the individual effect of the parameters n_s, h, and Ω_b on the matter power spectrum boost, around the reference model F5, is smaller than 1 per cent. Because of this, we have decided to neglect the effect of those parameters when building our emulator. We then only need to interpolate in a 3D space instead of a 6D one, which reduces the number of training simulations required to reach a certain emulation accuracy level. However, simultaneous variations of multiple of those parameters might produce a non-negligible effect. Additionally, the effect of those parameters in other regions of the E-MANTIS parameter space might also be stronger. Indeed, in our study of Section 3.2.2, we have only considered variations around the reference model F5, which is placed towards the centre of the emulation parameter space.

Probing all possible combinations of cosmological parameters would effectively require building an emulator in a 6D parameter space. This would be computationally expensive, specially given the cost of running $f(R)$ gravity simulations. Instead, we extend the *cosmo* simulation suite with a few selected models in order to get an estimation of the error that E-MANTIS makes by neglecting the effect of n_s, h, and Ω_b. All the additional simulations discussed in this appendix are run with the same characteristics as in the *cosmo* suite, except that we use one single random realization instead of five.

First, we study the impact of varying two of the neglected parameters simultaneously around the reference cosmological model F5. According to Fig. 1, the two neglected parameters with the largest impact are n_s and h. We run two simulations, which vary n_s and h by +25 per cent and −25 per cent, respectively. Fig. A1 shows that the effects of both parameters mostly add up linearly (see Fig. 1). The error made by neglecting both parameters is smaller than 1 per cent for $k \leq 2h\,\mathrm{Mpc}^{-1}$ and reaches ~ 2 per cent at $k = 7h\,\mathrm{Mpc}^{-1}$. We stress that this is a worst case scenario, since we have varied both parameters by extreme values. Fig. 1 shows that the impact of Ω_b is much smaller than 1 per cent at all scales considered. Therefore, we do not expect these results to significantly change when Ω_b takes different values.

We now assess the error made by individually neglecting the parameters n_s and h around other values of the main E-MANTIS parameters than the reference model F5. We are interested in some worst case scenarios. For instance we want to estimate the error made by E-MANTIS, when neglecting the individual impact of n_s and h around the edges of the parameter space in σ_8 and Ω_m. In order to do so, we run simulations by varying one of the main parameters as well as one of the neglected ones. Fig. A2 shows the relative error made by neglecting n_s or h for some extreme values of σ_8 and Ω_m. We also perform the same study around F6, since it is a model of observational interest for future surveys. One can see that
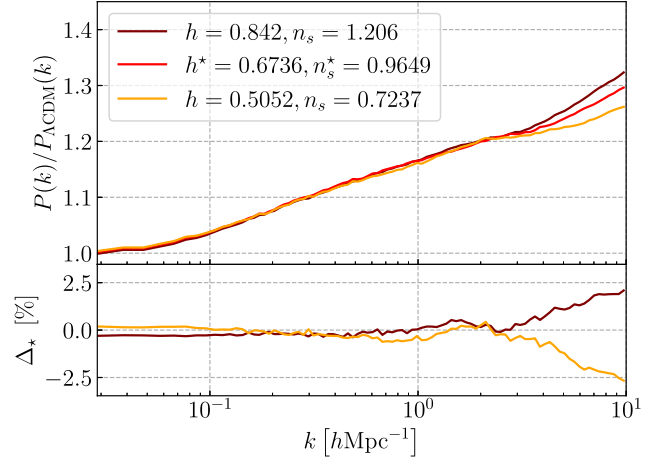

Figure A1. Effect of a simultaneous variation of the parameters n_s and h on the matter power spectrum boost. Both parameters are varied by +25 per cent and −25 per cent at the same time around the reference model F5. The top panel shows the boosts and the bottom panel the relative difference with respect to F5. The effect on the boost is smaller than 1 per cent up to $k \leq 2h\,\mathrm{Mpc}^{-1}$, and reaches ~ 2 per cent at $k = 7\,h\,\mathrm{Mpc}^{-1}$. We see that with respect to the results shown in Fig. 1, the individual effects of n_s and h mostly add up linearly.
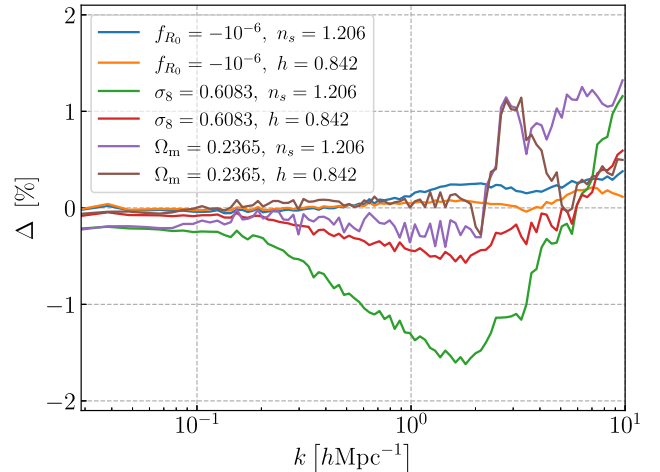

Figure A2. Error made by neglecting the effect of n_s and h for some selected values of the main E-MANTIS parameters different than the reference F5. The maximum error is around ~ 1.5 per cent, while in most cases it stays smaller than 1 per cent.

the relative error is smaller than 1 per cent in most cases and reaches the ~ 1.5 per cent level in the worst considered case. This should be compared to Fig. 1: by changing the reference model the maximum errors have slightly increased from 1 per cent to 1.5 per cent.

By considering both results from Figs A1 and A2, we find that the maximum error on the boost made by neglecting the parameters n_s, h, and Ω_b is 2 per cent. An extrapolation to even more complex combinations suggests a possible upper limit of around 3 per cent (if some of the deviations add up). To conclude, the maximum errors are expected at the 2–3 per cent level. They mostly happen around the edges of the parameter space. In most of the cases, the error is however smaller than 1 per cent.

APPENDIX B: INFLUENCE OF THE NUMBER OF REALIZATIONS AND THE REFINEMENT MODELS ON THE EMULATION ACCURACY

In this appendix, we show how both the number of realizations and the refinement models impact the final accuracy of the emulator. We perform the same kind of test as in Section 4.2. One slice of the primary SLHD is left out of the training models to be used as a validation sample. The emulator is built with different training data set configurations. Then, for each configuration, the predictions of the emulator are compared to the simulations measurements from the validation models. We consider four different configurations, where the emulator is trained using:

(i) 60 models from the primary SLHD and one realization per cosmological model,

(ii) 60 models from the primary SLHD, the 30 refinement models and one realization per cosmological mode,

(iii) 60 models from the primary SLHD and five independent realizations per cosmological model,

(iv) 60 models from the primary SLHD, the 30 refinement models and five independent realizations per cosmological model.

The results of this test at $z = 0$ are given in Fig. B1. The refinement models as well as the five independent realizations per cosmological model make it possible to get an average emulation error smaller than 0.5 per cent, and therefore negligible with respect to the systematic errors in the training data as estimated in Section 4.1.1. Even for the worst validation models, the emulation errors remains well below 1 per cent at all relevant scales. As expected, the signal is also smoother with five realizations than with a single one.

APPENDIX C: EMULATION ACCURACY AT DIFFERENT REDSHIFTS

Our $f(R)$ matter power spectrum boost emulator is built using training data from 19 redshift nodes (see Section 3.3). In Section 4.2 we have
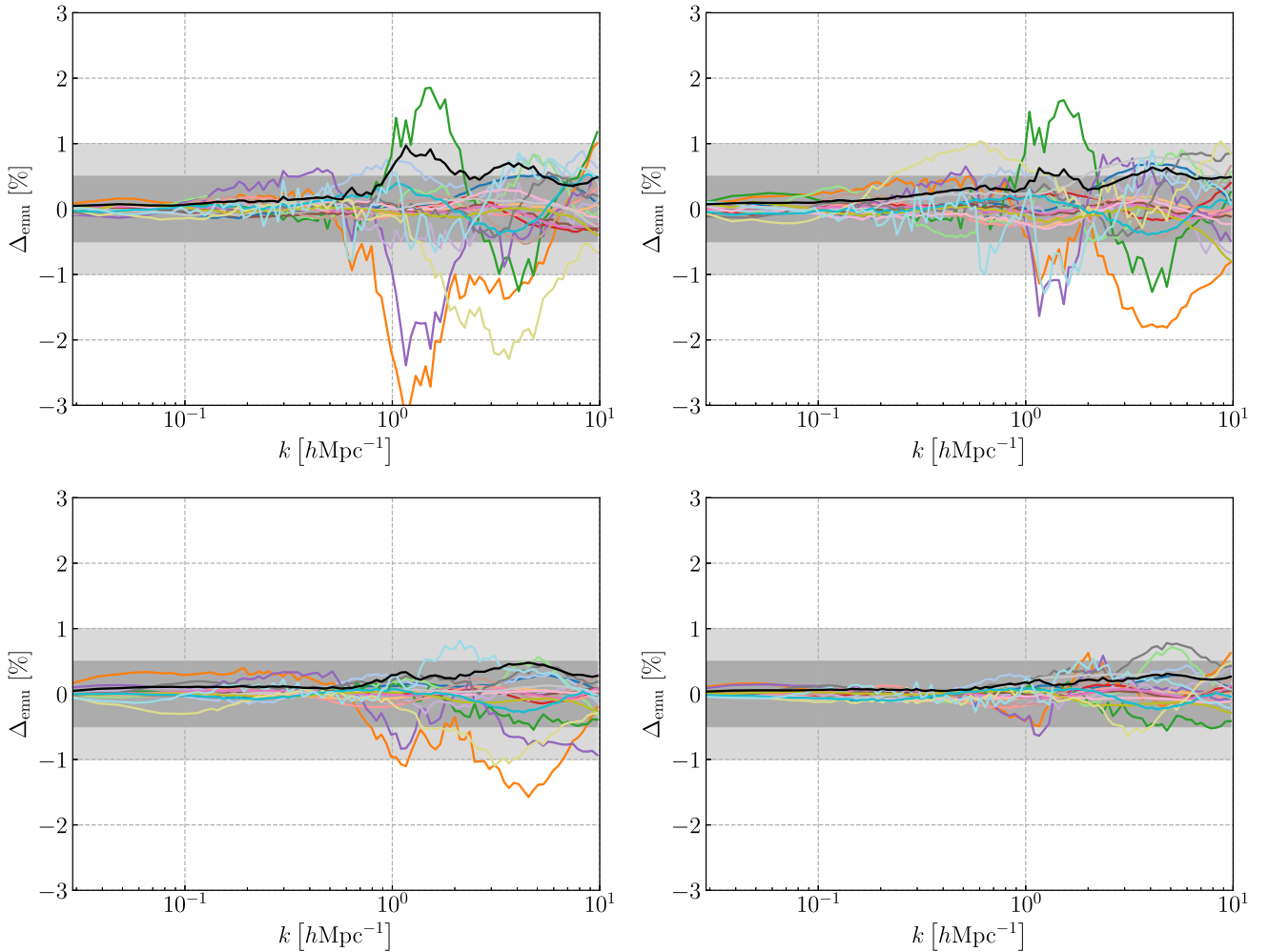

Figure B1. Emulation errors at $z = 0$ for different training data set configurations. Top left: the emulator is trained on three slices of the primary SLHD (60 models) and using a single realization per cosmological model. Top right: the 30 refinement models are added to the training data set. Botton left: the emulator is trained without refinement models but using five independent realizations per cosmological model. Bottom right: the 30 refinement models are once again added to the training data set. In all the cases, the remaining slice from the primary SLHD is used as a validation sample. Each coloured line shows the emulation error for one of the 20 validation models. The black line give the RMSRE, as defined by equation (24). The dark and light grey bands mark the 0.5 and 1 per cent accuracy limits. The 30 refinement models as well as the five realizations per cosmological model are needed to get emulation errors smaller than 1 per cent for all validation models and scales. In such a case, the RMSRE is well below the 0.5 per cent level at all scales.
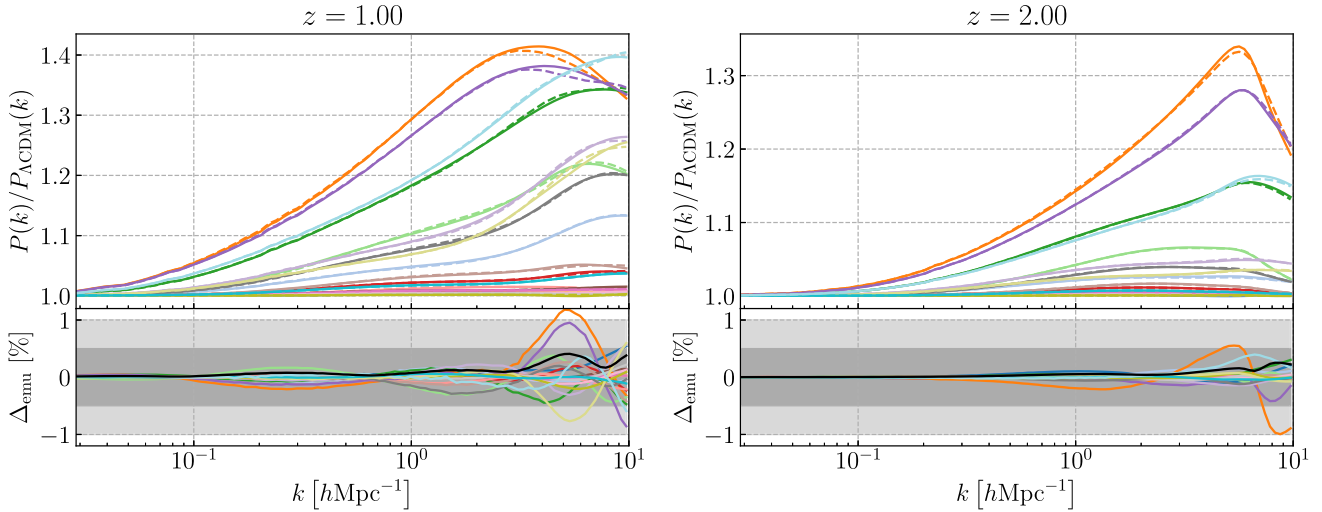
Figure C1. Results of the emulation accuracy test for the emulator at $z = 1$ (left) and $z = 2$ (right). The emulator is trained using the refinement models and three slices from the primary SLHD. The remaining slice is used as a validation set. The top panels give the predicted boost by the emulator for each test model (solid lines). The dashed lines are measurements from the simulations. The bottom panels give the relative emulation errors for each validation model. The solid black line corresponds to the RMSRE as defined by equation (24). The dark and light grey bands give the 0.5 per cent and 1 per cent levels, respectively. At both redshifts the emulation errors are smaller than 0.5 per cent for most models and scales. Only for a few extreme models the emulation errors reach the 1 per cent level at the smallest scales. The RMSRE is always smaller than 0.5 per cent.
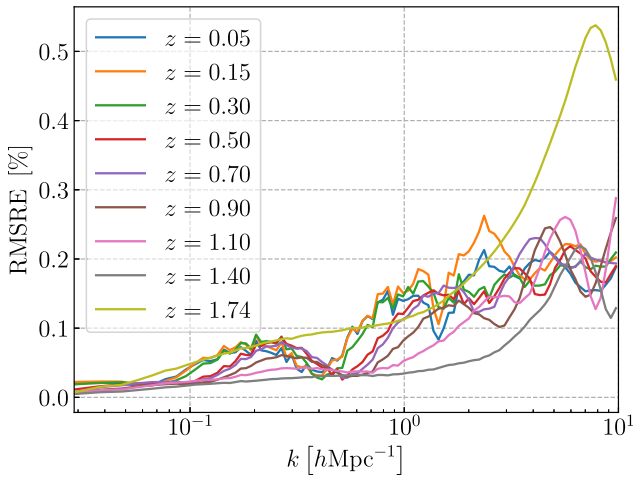

Figure C2. Redshift interpolation accuracy test. The emulator is trained using only 10 redshift nodes. The remaining nine redshift nodes are used as a validation sample. For each validation redshift we compare the predictions from the emulator to the simulation measurements for the 80 models of the primary SLHD. We give here the RMSRE, as defined by equation (C1), which for most redshifts is smaller than 0.3 per cent at all scales. The redshift interpolation errors in the final emulator, which uses twice as many redshift nodes as in this test, are expected to be smaller.

estimated the emulation errors at $z = 0$ by splitting the simulation data into a training and validation set. Fig. C1 gives the same results but at $z = 1$ and $z = 2$. The conclusions drawn for $z = 0$ are still valid at those redshifts. More precisely, the emulation errors for the 20 validation models only reach the 1 per cent level for the most extreme models and the RMSRE, defined in equation (24), remains under 0.5 per cent at all scales. We have verified that these results hold for the 19 redshift nodes of E-MANTIS between $0 < z < 2$.

E-MANTIS is also able to perform predictions for redshift values outside the training nodes. We proceed by building an independent emulator at each redshift node. Then, for any arbitrary redshift, we perform a linear interpolation of the predicted boost between the two neighbouring redshift nodes. In order to estimate the contribution to the emulation error budget of such a procedure, we train our emulator using the full set of available cosmological modes but only 10 redshift nodes. The remaining nodes are used as a validation sample. On these nine validation nodes, we compare the prediction of the emulator to the simulation measurements for the 80 models of the primary SLHD. We exclude the refinement models from this comparison in order to have an homogeneous distribution of models across the cosmological parameter space. We compute the corresponding RMSRE at each validation redshift, defined as

$$\mathrm{RMSRE}(k, z) = \sqrt{\frac{1}{N_{\mathrm{models}}} \sum_{i=0}^{N_{\mathrm{models}}-1} \left(\frac{B_{\mathrm{emu}}(k, z; \theta_i)}{B_{\mathrm{sim}}(k, z; \theta_i)} - 1 \right)^2}, \quad \text{(C1)}$$

where N_{models} is the number of models used in the comparison, θ_i is the cosmological parameter vector of model i, B_{emu} is the power spectrum boost predicted by the emulator and B_{sim} is the boost measured from the simulations. The obtained RMSREs are shown in Fig. C2. For most redshifts the RMSRE is smaller than 0.3 per cent at all scales. Only for $z = 1.74$ the RMSRE reaches the 0.5 per cent level. The final emulator has twice as many redshift nodes as in this test. We therefore expect the real redshift interpolation errors to be smaller that the present estimation. In any case, these errors are negligible with respect to the mass resolution errors, which can reach the 3 per cent level.

This paper has been typeset from a TeX/LaTeX file prepared by the author.