APPENDICES
Horizon-AGN virtual observatory – 1. SED-fitting performance and forecasts for future imaging surveys

C. Laigle[1⋆], I. Davidzon[2], O. Ilbert[3], J. Devriendt[1], D. Kashino[4], C. Pichon[5,6,7], P. Capak[2], S. Arnouts[3], S. de la Torre[3], Y. Dubois[5], G. Gozaliasl[8,9,10], D. Le Borgne[5], S. Lilly[4], H. J. McCracken[5], M. Salvato[11], A. Slyz[1]

[1] *Sub-department of Astrophysics, University of Oxford, Keble Road, Oxford OX1 3RH*
[2] *IPAC, Mail Code 314-6, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA*
[3] *Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France*
[4] *Department of Physics, ETH Zürich, Wolfgang-Pauli-strasse 27, CH-8093 Zürich, Switzerland*
[5] *Sorbonne Universités, CNRS, UMR 7095, Institut dÁstrophysique de Paris, 98 bis bd Arago, 75014 Paris*
[6] *Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, United Kingdom*
[7] *Korea Institute for Advanced Study (KIAS), 85 Hoegiro, Dongdaemun-gu, Seoul, 02455, Republic of Korea*
[8] *Finnish centre for Astronomy with ESO (FINCA), Quantum, Vesilinnantie 5, University of Turku, FI-20014, Turku, Finland;*
[9] *Department of Physics, University of Helsinki, PO Box 64, FI-00014 Helsinki, Finland*
[10] *Helsinki Institute of Physics, University of Helsinki, P.O. Box 64, FI-00014, Helsinki, Finland*
[11] *Max-Planck-Institut für extraterrestrische Physik, Garching, Germany*

25 April 2019

APPENDIX A: MOCKS ADDITIONAL FEATURES

Let us provide more details about how the virtual photometry has been computed.

A1 Calibration of the dust attenuation

A1.1 Dust-to-metal mass ratio

Dust attenuation is implemented assuming that the distribution of gas metallicity is a good proxy for the dust distribution. This computation implies to choose a value for the dust-to-metal mass ratio, i.e. to define which fraction of metals are locked into dust grains. For the sake of simplicity, the dust-to-metal mass ratio is assumed constant, though some works have shown that it could vary with redshift or within a same galaxy as a function of metallicity (e.g. Galametz et al. 2011; Mattsson et al. 2012; De Cia et al. 2013; Fisher et al. 2014). Most of the time, the implementation of dust attenuation in simulations uses a dust-to-metal ratio of 0.4 (e.g. Jonsson 2006), which is the Milky-Way value (Dwek 1998). Nevertheless, there is no evidence that this factor, derived from high-resolution models of our Galaxy, should be used at face value in the simulation, especially because the spatial resolution ($\sim$ 1 physical kpc) of the simulation implies that dust scattering and absorption occurs at the subgrid scale. One expects therefore that the emergent dust attenuation will depend on the smaller scale distribution of dust and metals, which are not resolved. Given the low resolution of the simulation, we do not implement a two-component dust attenuation models to account separately for dust obscuration in both birth clouds and diffuse interstellar medium, although this has been successfully implemented elsewhere (Trayford et al. 2015).
In addition, prior to converting the metal mass into dust mass, it is important to reproduce the correct gas phase metallicity in the simulation. As discussed in Kaviraj et al. (2017), the relatively low resolution reached in HORIZON-AGN implies a delayed enrichment of star-forming clouds, which underestimates the gas phase metallicity compared to observations. To correct for this, a redshift-dependent boosting factor (varying from 4 at $z \sim 0$ to 2.4 at $z \sim 3$) has been computed in order to brings the simulated mass-gas phase metallicity relation in agreement with obsevations from Mannucci et al. (2010) at $z = 0$, 0.7, 2.5 and Maiolino et al. (2008) at $z = 3.5$. However, the normalization of the mass-gas metallicity relation undergoes large variations depending on the chosen observable used to measure gas metallicity, up to a factor of 5 (see e.g. Andrews & Martini 2013). In particular, the renormalisation of HORIZON-AGN gas phase metallicity tends to align onto the highest values derived from observations (namely using the R23 method, see e.g. Lian et al. 2015). As a consequence, one can expect the boosting factor derived in Kaviraj et al. (2017) to be an upper limit. In fact, while comparing the simulated galaxy counts with COSMOS2015 in various bands as a function of redshift after renormalizing the gas-phase metallicity, we find that a dust-to-metal mass ratio of 0.4 is excessive as it results in too few galaxy counts compared to observations. Therefore, we empirically choose a dust-to-metal mass ratio of 0.2, which yields an overall better agreement with the observed counts in all bands at all redshifts.

A1.2 Attenuation curve

Beyond the dust mass distribution in the galaxy, the amount of extinction at a given wavelength will be dependent upon the chosen attenuation curve. In this work, the $R_V = 3.1$ Milky Way dust grain model by Weingartner & Draine (2001) is used for post-processing the simulated galaxies. This model includes in particular the prominent 2175 Å-graphite bump. In the simulation, the spectrum of

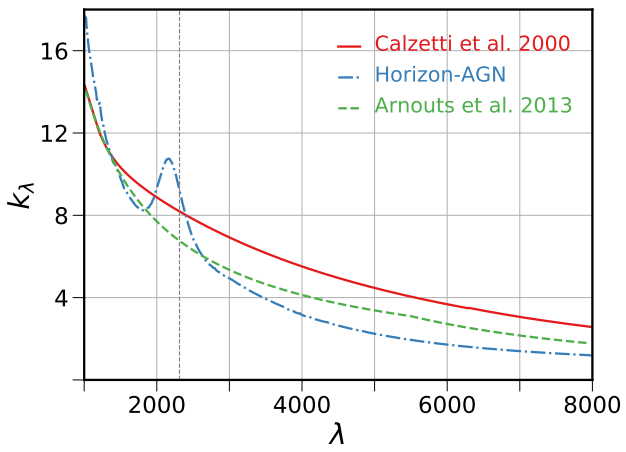

Figure A1. Extinction curves used when fitting the photometry to get the physical properties (*red solid line* and *green dashed line*) and the average extinction for ∼ 1000 HORIZON-AGN galaxies randomly selected (*blue dashed-dotted line*). The vertical dashed line indicates the median wavelength of the *NUV* filter.

each stellar particle, assumed to be a SSP, is attenuated with this model and knowing the specific dust column density in front of each particle. However, when summing up the contribution of all the SSPs, we find that the overall attenuation curve of the resulting galaxy spectrum is less steep and the bump tends to reduce. As discussed in Fischera & Dopita (2011), turbulence can further reduce the bump and flatten the extinction curve. Fig. A1 presents the extinction curve in HORIZON- AGN as averaged over one thousand galaxies randomly selected, and the two curves used in LEPHARE for the computation of stellar masses and SFR. k_λ is defined as $A_\lambda/\mathrm{E}(B-V)$. None of the two curves used in LEPHARE can correctly reproduce the one in HORIZON-AGN, and this discrepancy is likely to be the reason for the bimodality in the SFR, as further discussed in Appendix B.

A2 IGM absorption

In order to implement the IGM absorption, the Lyman-α forest is implemented on each galaxy line-of-sight from the gas density, velocity and temperature in the IGM in front of the galaxy.

Let us consider the line-of-sight of a background source emitting at the observed wavelength λ_0. Here $\lambda_\alpha = 1215.7$ Å corresponds to the transition from the ground state to the first excited state of the Hydrogen atom. The wavelength of the photons emitted by the background source spectra are redshifted by a factor of $(1+z)$. At some point the light from the source will be redshifted at $\lambda_\alpha = 1215.7$ Å. At this point, it may be absorbed by HI from the IGM. The probability of transmission of the light at the observed frequency ν_0 will be given by:

$$F(\nu_0) = e^{-\tau_\alpha(\nu_0)}, \tag{A1}$$

where $\tau_\alpha(\nu_0)$, the Ly-α optical depth at the observed frequency ν_0 is given by

$$\tau_\alpha(\nu_0) = \int_0^{x_s} \mathrm{d}x \frac{\sigma_\alpha n_{\mathrm{HI}}(x,z)}{1+z}, \tag{A2}$$

where x is the comoving coordinate of the comoving point varying along the line of sight between the observer ($x = 0$) and the source ($x = x_s$), z is the corresponding redshift, n_{HI} the neutral hydrogen density at point x and redshift z, and σ_α the Ly-α cross-section. Here σ_α is a function of the frequency ν of the photon with respect to the rest frame of the neutral hydrogen at position x. Here $\nu = \nu_0(1+z)(1+v/c)$, where v is the peculiar velocity along the line-of-sight. So σ_α may be then written as

$$\sigma_\alpha = \frac{\sigma_{\alpha,0}\, c}{b(x,z)\sqrt{\pi}} e^{-\frac{\left(v(x,z)(1+z)\nu_0 - c\nu_\alpha + c(1+z)\nu_0\right)^2}{\nu_\alpha^2 b^2(x,z)}},$$

where $b(x,z) = \sqrt{2k_B T(x,z)/m_p}$, $\sigma_{\alpha,0} = (3\pi\sigma_{\mathrm{T}}/8)^{1/2} f \lambda_\alpha$, with $\sigma_{\mathrm{T}} = 6.25\times10^{-25}\mathrm{cm}^2$ the Thomson cross-section, and $f = 0.4162$ the oscillator strength. As we do not save the neutral hydrogen outputs for HORIZON-AGN, the neutral hydrogen density is computed in post-processing by considering that the fraction $x_{\mathrm{HI}} = n_{\mathrm{HI}}/n_{\mathrm{H}}$ is a balance between photo-ionizations, collisional ionizations and recombinations. At equilibrium with the cosmic UV background field, it yields:

$$\alpha(T) n_{\mathrm{e}}(1-x_{\mathrm{HI}}) = \gamma(T) n_{\mathrm{e}} x_{\mathrm{HI}} + \Gamma x_{\mathrm{HI}}, \tag{A3}$$

where α and γ are the collisional recombination and ionisation rates, Γ is the photoionisation rate, and n_e is the free electron number density. Considering a uniform background radiation field as implemented in the simulation, the photoionisation rate is assumed to be spatially uniform. Its overall normalization is quite uncertain and is adjusted in order to match the PDF of the transmitted flux at $z = 1.5$, $z = 2$, $z = 2.5$ and $z = 3$ (see and e.g. Theuns et al. 1998; Bolton et al. 2005; Lukić et al. 2015; Bolton et al. 2017). The chemical composition of the IGM is close to primordial, so it can safely be assumed that n_e receives contribution only from ionised Hydrogen and Helium (assumed entirely in its fully ionized form), which allows to determine its composition (Choudhury et al. 2001). Prescriptions from Black (1981) are used to determine the collisional recombination and ionisation rate as a function of gas temperature.

In order to take into account the full Lyman series absorption, we assume that the only difference between Lyman-α and other Lyman transition comes from different absorption cross-sections. This modelling hence neglect different widths of Lorentz profiles from one transition to an other, but this is expected to be a secondary effect (Iršič & Viel 2014).

Eventually, galaxy spectra are then multiplied with the Lyman-series absorption lines. Fig. A2 shows the median absorption by the IGM in the u and B bands as a function of redshift and compares with the litterature (Meiksin 2005; Inoue et al. 2014). The median absorption in a given band will also depends on the hardness of the galaxy UV spectrum, and on the presence or absence of Lyman-limit systems (Meiksin 2005). On overall, our implementation of IGM absorption matches well the litterature. As noted in e.g. Inoue et al. (2014), the Madau model tends to slightly overestimate the correction at $z > 3$ compared to observations, and therefore also overestimate the correction of our virtual galaxies.

A3 Flux error implementation

Implementing realistic errors on the flux is crucial for the accuracy of our forecasts to retrieve correct redshift and masses. To this end, we compute flux errors for our mock galaxies and perturb their fluxes accordingly.

For the COSMOS-like catalogue, this implementation is done in each band while relying on the COSMOS2015 catalogue. The real

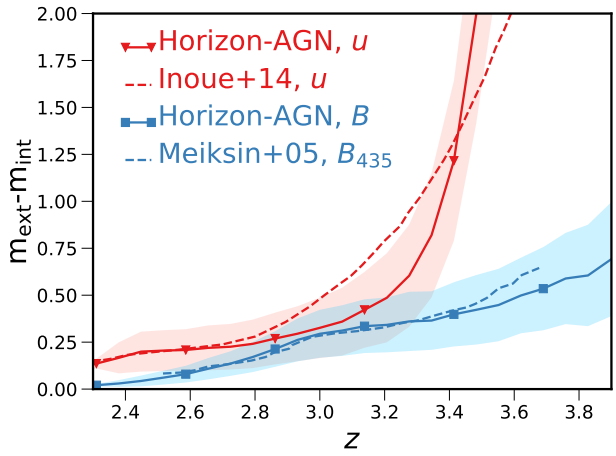

Figure A2. Median absorption by the IGM in the u, and B bands as a function of redshift for HORIZON-AGN galaxies. Here $m_{\rm int}$ and $m_{\rm ext}$ are respectively the intrinsic magnitudes and magnitudes after IGM absorption.

dataset is split in small bins of flux[1]. In each bin the flux error distribution is fitted with a Gaussian function. The HORIZON-AGN catalogue is then divided in the same way. For each simulated galaxy in a given bin, and for each COSMOS filter, an error is randomly chosen according to the COSMOS2015 error distribution. Galaxy apparent fluxes, which initially corresponded "exactly" to the star particles' content, are now perturbed according to their 1σ error. We note that for this implementation, the intrinsic fluxes from HORIZON-AGN are confronted to the -already perturbed- observed ones. As a result, the horizontal width of the faint-end tail tends to be larger in the simulation as in observations.

For the *Euclid*-like, LSST-like and *Euclid*+LSST catalogues, errors are implemented according to the COSMOS2015 flux error distribution in the closest filters pass-bands, and shifted according to the expected depth of the surveys at similar number of σ. We emphasize that, although reasonable, these errors might no reflect the specific noise of the survey, and in any case it does not take into account possibly systematics in the photometry.

Fig. A4 shows the distribution of 1σ errors, as a function of total magnitudes, in several bands in HORIZON-AGN (grey area) and compares them to observed data (red area) when available (for the COSMOS-like sample). The adopted depths in all bands are summarised in Table A1. Fig. A3 presents the simulated galaxy count in the r and $K_{\rm s}$ bands in bins of redshift and compares them to the counts from COSMOS2015. Several features can be noticed. First of all the HORIZON-AGN simulated catalogue is mass limited while COSMOS2015 is not, which explains the drop-off of the count at faint magnitudes in the $K_{\rm s}$ band at low-redshift. Part of the discrepancy at bright magnitudes is driven by the smaller area covered by HORIZON-AGN (1deg^2) compared to COSMOS2015 (~1.4deg^2). Finally, we find that the galaxy counts are overestimated at the faint end of the distribution. This effect has already been discussed in Kaviraj et al. (2017) and is mostly driven by stellar feedback not being strong enough in HORIZON-AGN.

[1] In COSMOS2015 the total fluxes are computed from the corrected aperture magnitudes (using the offset defined in Equation 4 of L16), and then accounting for the Galactic foreground extinction (following their Equation 10).

Survey	band	depth
	u	26.6
	B	27.0
	V	26.2
	r	26.5
	i^+	26.2
COSMOS-like	z^{++}	25.9
3σ depth,	Y	25.3
extended sources	J	24.9
	H	24.6
	$K_{\rm s}$	24.7
	IB	25-26
	riz	24.5
Euclid-like	Y	24.0
5σ depth,	J	24.0
extended sources	H	24.0
	g	24.6
DES-like	r	24.1
5σ depth,	i	24.0
extended sources	z	23.9
	u	26.3
	g	27.5
LSST-like	r	27.7
5σ depth,	i	27.0
extended sources	z	26.2
	y	24.9

Table A1. A summary of the adopted depths in all bands. The depths of intermediate bands (IB) are detailed in Laigle et al. (2016).

A4 Stellar mass loss

Let us discuss the prescriptions used in both hydrodynamical simulations and stellar population synthesis models to take into account stellar mass losses due to galactic winds, remnants and supernovae. The impact on the comparison between simulated and observed data is not negligible if the two samples rely on different prescriptions.

In HORIZON-AGN, the stellar mass loss is modelled as a function of time and metallicity assuming that stars are distributed with a Salpeter (1955) IMF and supernovae type Ia occur with the frequency described in Greggio & Renzini (1983), assuming a binary fraction of 5 percent. To compare to SED-fitting estimates, one may either rescale the M_* values in the simulation by a factor ~1/1.7 (the usual conversion from Salpeter to Chabrier IMF, e.g. Santini et al. 2011) or perform the SED fitting with BC03 templates that assume a Salpeter IMF. Neither of these solutions is sufficient to fully remove the bias because even when the IMF is the same, the resulting fraction of ejected stellar mass may significantly differ. In other words, the mass of HORIZON-AGN stellar particle also account for remnant mass. Different SSP models will implement their formation differently, and therefore at a given time and metallicity the remnant mass will change from one model to the other. Fig. A5 presents the mass evolution of a SSP as a function of time in HORIZON-AGN and using the Maraston (2005) and BC03 models.

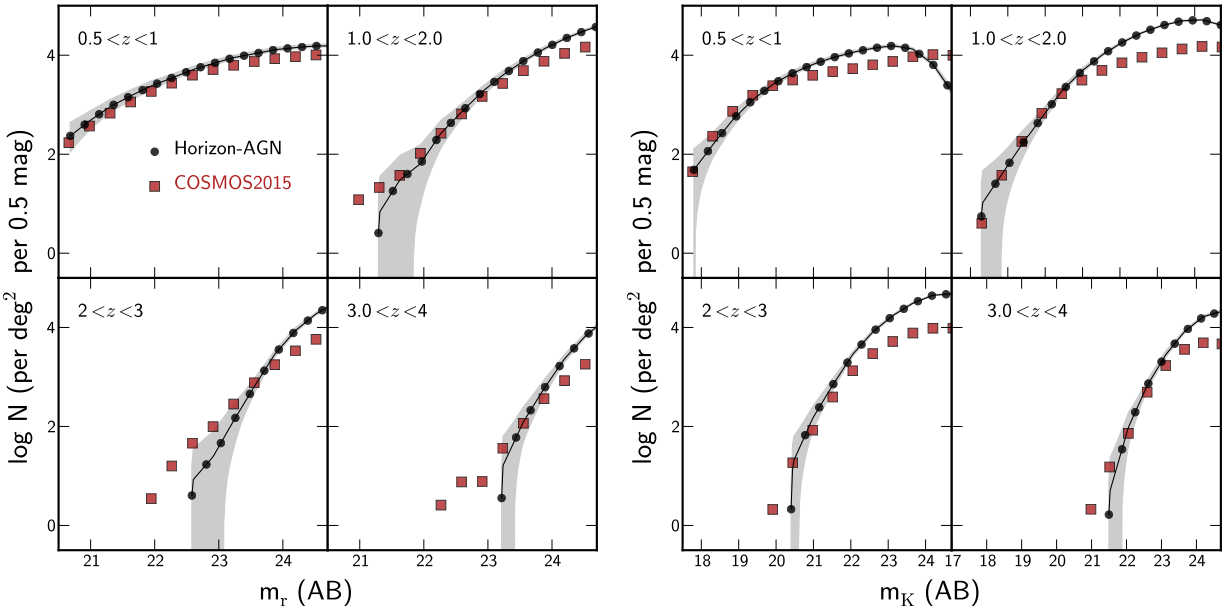

Figure A3. Galaxy total magnitude count in the r-band (*left*) and K_s band (*right*) as a function of redshift in COSMOS2015 (red squares) and HORIZON-AGN (sold line). COSMOS2015 photometry is corrected for Galactic extinction.

To account for this discrepancy, one possible solution is to choose a SED-fitting library based on a stellar population synthesis model whose features are in better agreement with that used in the simulation. However, we cannot modify LEPHARE as we want to compare our findings to COSMOS2015's and therefore be consistent with the set-up used for that catalogue. Instead, we prefer to correct the HORIZON-AGN virtual photometry by matching a BC03 SSP to each hydrodynamical particle at the time of its formation. We let the SSP evolve so that at any age of the stellar particle we can compute the mass loss fraction according to the BC03 model.

We stress that these details, related *a-priori* to the "sub-grid physics", should be reconciled before comparing the simulation to real data (e.g., Henriques et al. 2015). However, although most of the studies take into account the IMF conversion, it is difficult to find in the literature comparisons that correct for the different stellar mass loss parametrisation (see e.g., Davidzon et al. 2018).

A5 Limitations of our modelling

It should be finally emphasized that our end-to-end modelling still has shortcomings, which are listed below.

No systematics in the photometry Although statistical photometric errors are consistently implemented in the virtual dataset (see Section A3), systematics arising when extracting the photometry from the images (blending of objects, clumpy objects possibly split at the extraction, lensing magnification and PSF, image artefacts) are not accounted for here and will be the topic of future work.

Spatially constant IMF and stellar mass loss prescriptions Our modelling also ignores specific aspects of galaxy evolution which can modify the photometry. For example, nebular emission lines are not implemented in the photometry. Furthermore, the pipeline implicitly assumes that the IMF does not spatially vary within the galaxies, and is perfectly known at the SED-fitting stage. A Chabrier IMF (Chabrier 2003) is *de facto* used both when computing the photometry from the simulated catalogue and to build the galaxy template library for SED fitting. Fitting the galaxy photometry with a different IMF from the one chosen to compute this photometry would obviously lead to new systematics in the stellar mass computation. In addition to the choice of the IMF, which controls the amount of stars formed as a function of their mass (and therefore in particular the overall mass-to-light ratio), the chosen prescription for stellar mass losses as a function of time and metallicity is important. In the work presented here, the simulated photometry is computed with BC03 SSP templates (see Section A4), and fitted with a SED library which includes a higher diversity of stellar mass loss prescriptions. Therefore, we do not assume to know *a priori* the stellar mass loss prescriptions of the simulated galaxies, and effectively the best-fit is not always a BC03 template. However, in practice the simulated galaxies have all the same stellar mass loss prescriptions, and these prescriptions are spatially constant within the galaxies: this is unlikely to be the case in the real Universe. Therefore the simulated galaxy population present less diversity than the observed one, and one could expect the SED-fitting to perform much better on the HORIZON-AGN Universe than on the real one.

In despite of these limitations, we found that the $z_{\rm phot}$ accuracy of the virtual catalogue is comparable to that of COSMOS2015. This suggests that varying IMF and stellar mass loss prescriptions should not dramatically impact redshift reconstruction, which indeed relies on galaxy colours (i.e. the relative values of flux in different bands, and not the absolute value of the flux). Only the redshift errors are lower in the simulated catalogue with respect to the real one. However a much larger and systematic impact on stellar mass and SFR is expected, and the reconstruction accuracy quoted in the main paper should be understood as an optimistic case. On the other hand, we have shown that with a very good photometry stellar masses are well retrieved (within 0.12 dex) with the COSMOS-like configuration when the IMF is fixed. Therefore, one can expect that all additional systematics will be driven by the un-

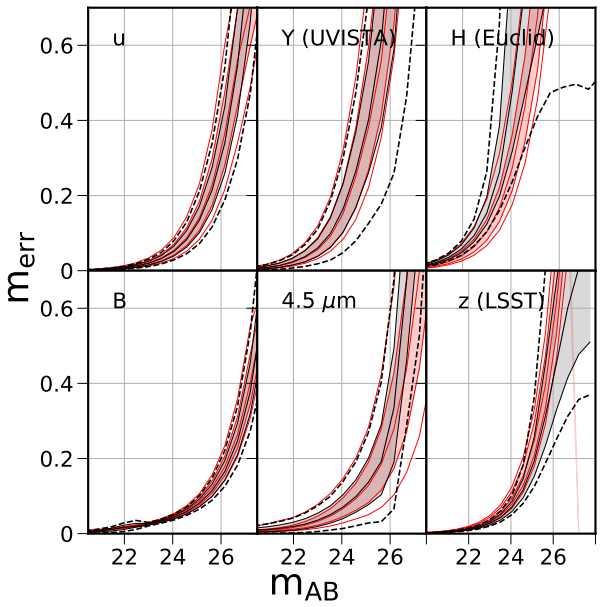

Figure A4. Comparison between observed and simulated apparent magnitude errors (m_{err}) in u, B, Y, IRAC [4.5] bands from COSMOS, H from *Euclid* and z from LSST. In each panel, *grey* area shows 68 percent of the galaxies from HORIZON-AGN and the *red* one shows 68 percent of the galaxies from COSMOS. *Solid* lines are the median, and outer lines encompass 95 percent of the m_{err} distribution. In the case of *Euclid* and LSST bands, the observed errors are those in COSMOS in the corresponding filter passbands, but shifted to match the magnitude limits at completion. H from *Euclid* will be at completion $\sim$ 0.6 dex shallower than the current H-band from UVISTA, and z-band from LSST will be $\sim$ 0.3 dex deeper than the current z^{++} band from Subaru.

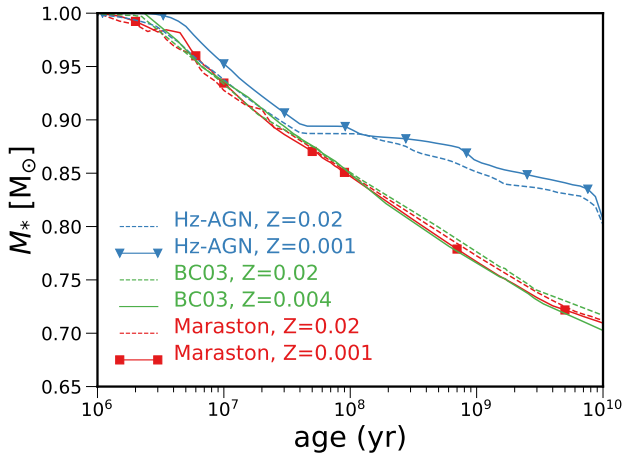

Figure A5. Evolution of the stellar mass as a function of time for a stellar particle in HORIZON-AGN (*blue* lines) and for the SSPs produced by BC03 (*green* lines) and Maraston (2005, *red* lines) stellar population synthesis model, for solar (*solid* lines) and sub-solar (*dashed* lines) metallicity. Salpeter IMF is assumed here.

Table B1. Statistical errors (NMAD) and percentage of catastrophic errors (η) in different i magnitude bins, without absorption, with IGM absorption and with both dust and IGM absorption.

i band	No absorption		IGM		IGM+dust	
mag	NMAD	η (%)	NMAD	η (%)	NMAD	η (%)
(22,23]	0.009	0.0	0.009	0.0	0.008	0.0
(23,24]	0.013	0.0	0.013	0.0	0.014	0.0
(24,25]	0.022	0.6	0.021	0.4	0.026	0.5
(25,26]	0.049	8.7	0.045	7.6	0.052	9.2

certainty on the IMF. Comparing the mass from the SED-fitting with an independent measurement (kinematics, small-scale lensing) can therefore be a way to constrain the IMF.

APPENDIX B: IMPACT OF ABSORPTION ON SED

Dust attenuation is implemented in the virtual photometry as post-processing at the scale of $\sim$1 kpc, following the prescription detailed in Appendix A. In particular, at equal dust mass, the resulting attenuation of the total spectrum will depend on the geometry of the galaxy and the angle under which it is seen. Conversely, at the SED-fitting stage, dust attenuation is kept as simple as possible and therefore does not depend on galaxy geometry: all the SSPs are assumed to undergo the same attenuation given a specific attenuation curve. In a similar way, IGM absorption is implemented in the HORIZON-AGN lightcone independently for each galaxy, knowning the foreground distribution of HI. However, at the SED-fitting stage, IGM absorption is accounted for with an average redshift-dependent relation for all galaxies. These differences can play a role in driving the scatter and the systematic trends observed in mass and star-formation rate comparison. Further details on this effect are presented now.

B1 z_{phot} performance

LEPHARE is run with 3 different input photometric catalogues, in order to show the impact of dust and IGM on z_{phot}. The main run (presented in Section **??**) includes both inter-galactic and intra-galactic (dust) absorption. An additional run is performed on an input photometry that does not include absorption from either components, and the last one is performed on an input photometry that includes only inter-galactic absorption. The comparison of the performance of these runs directly tells us about the impact of inhomogenous IGM and spatially varying dust absorption on the SED-fitting performance.

Results of the z_{phot} performance for no absorption, IGM-absorption only and IGM+dust absorption are presented in Table B1. On overall IGM slightly helps to constrain the z_{phot}. However, as discussed in the main text, we note that overestimating the IGM absorption at the SED-fitting stage can make a large fraction of the catastrophic outliers falling in $[z_{sim} > 2.5] \cap [z_{obs} < 1.5]$. Adding dust to the photometry reduces slightly the performance. Table B1 has to be interpreted with caution however. In a given magnitude bin, the dust-free catalogue probes a galaxy population which can overlap a fainter magnitude bin in the dusty catalog. In other words, not only the presence/absence of dust drives the difference in the performance of the z_{phot} in a given magnitude bin, but also the possible differences in the intrinsic SED of the galaxies.

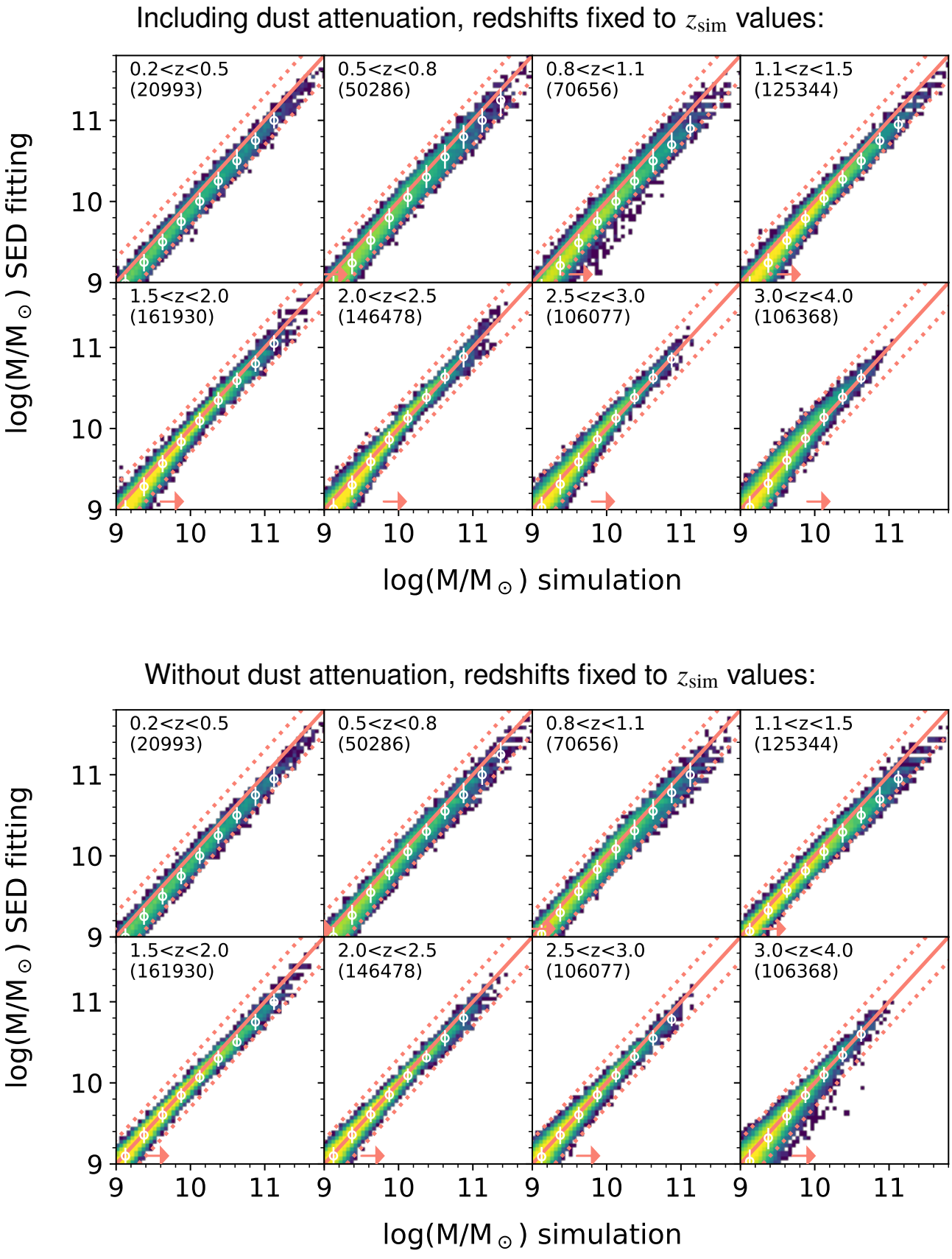

Figure B1. Comparison, in different $z_{\rm sim}$-bins, between stellar masses estimated through SED fitting and intrinsic values. Dust attenuation is included in the *upper* set of panels, while in the *lower* plots dust-free photometry has been used in the SED fitting. In both cases, the redshift is fixed to its intrinsic value ($z_{\rm sim}$) during the computation. In each panel, the *solid* line is the 1:1 relation while *dashed* lines show ±0.3 dex offset from it.

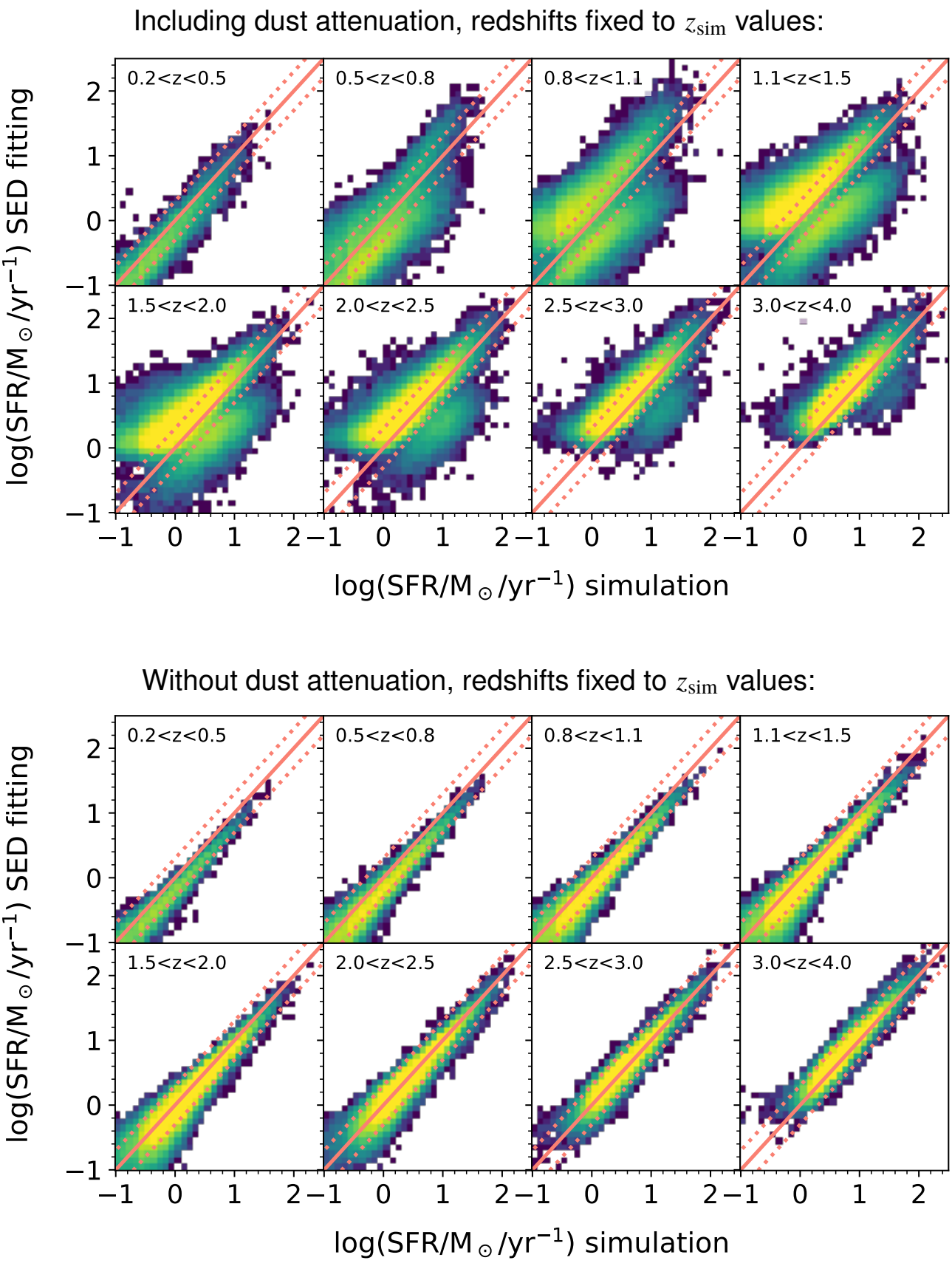

Figure B2. Comparison, in different $z_{\rm sim}$-bins, between SFR estimated through SED fitting and intrinsic values. Dust attenuation is included in the *upper* set of panels, while in the *lower* plots dust-free photometry has been used in the SED fitting. In both cases, the redshift is fixed to its intrinsic value ($z_{\rm sim}$) during the computation. In each panel, the *solid* line is the 1:1 relation while *dashed* lines show ±0.3 dex offset from it.

B2 Stellar mass estimation

As the impact of IGM is limited at the very high redshift, the subsequent analysis focus on the impact of absorption as a whole on stellar mass and SFR estimates. In order to isolate the effect of absorption, the physical parameter estimation is done at the intrinsic redshift of the galaxies (i.e. the mass and SFR uncertainties do not include the $z_{\rm phot}$ uncertainties) on two input photometric catalogues (with and without absorption).

Fig. B1 presents the comparison between the intrinsic and observed stellar mass, for galaxies selected at $K_{\rm s} < 24.7$. Horizontal orange arrows indicate the mass completeness in each redshift bin. Without absorption, mass completeness is naturally better as galaxies are brighter. Dust attenuation induces a negligible overestimation of the observed stellar mass especially at high redshift and for massive galaxies. This is expected, as the mass estimates is essentially provided by the NIR bands which are barely affected by the dust. However, when only optical bands are available (e.g. in the LSST-only configuration), dust is expected to have a more dramatic effect, similar to the one observed in the SFR computation (see the discussion below).

Without any kind of absorption (bottom panel), there is a persistent underestimation of stellar mass by at most 0.1 dex at low redshift, owing to the simplistic SFHs and metallicity distributions in the template library. As noted before (see Section **??**) the effect of dust (overestimation) and SFHs (underestimation) can act in opposite direction and therefore tend to compensate each other.

B3 SFR estimation

Fig. B2 presents the comparison between the intrinsic and observed SFR, for galaxies selected at $K_{\rm s} < 24.7$ and $0.2 < z < 4$. The effect of dust is dramatic: it drives a large scatter and a bimodality, with a systematic overestimation of the SFR for a fraction of the population up to at least $z \sim 2.5$. This is a consequence of the degeneracy between dust and SFR. Fig. B3 isolates what in the dust modelling drives the effect. The comparison between SFR estimated from the best-fitted template and the intrinsic SFR is presented at $1.1 < z < 1.5$. The diagram is colour-coded by the extinction curve used in the fit (*left* panel) and $\Delta_{A_{NUV}}$ (*right* panel), where $\Delta_{\rm A_{NUV}} = {\rm A}^{\rm sim}_{NUV} - {\rm A}^{\rm phot}_{NUV}$. Qualitatively, when the best-fit template is derived using Arnouts et al. (2013)'s extinction curve, A_{NUV} is overestimated and the SFR is consequently overestimated. The reverse happens when the best-fit template is attenuated using Calzetti et al. (2000)'s extinction curve.

APPENDIX C: ZERO-POINT MAGNITUDE OFFSETS

It is common in the literature to apply an offset to apparent magnitudes in order to correct for systematics due to calibration discrepancies between different filters (e.g., Ilbert et al. 2013). These offsets are computed by LePhare using a spectroscopic sub-sample: the code fits galaxy SEDs after fixing their redshifts at the spectroscopic value, then it compares the magnitude observed in each filter to the one of the best-fit model (i.e., the template minimizing the reduced χ^2). An offset is added in each filter to reduce the difference between predicted and observed magnitudes. The code iterates the procedure until convergence, finding the final values of the zero-point offsets that will be add by default in the next LePhare runs.

Although such a procedure is quite efficient at improving SED-fitting results, they might also introduce a bias if the used spectroscopic sub-sample is not representative of the whole population. Moreover zero-point offsets correct, at least partly, for possible incompleteness in the template library. For this reason some authors prefer not apply them, e.g. when deriving stellar masses, because besides solving calibration issues they may affect the physical interpretation of the SED fitting results (see Moutard et al. 2016).

We compute the zero-point offsets in the COSMOS filters, not for a spectroscopic-subsample but for the whole galaxy catalogue. All the offsets are found to be smaller than 0.02 mag, which is of the order of the minimal photometric errors, with the exception of 0.029 and 0.039 mag in the z^{++} and $K_{\rm s}$ band respectively. In COSMOS2015 these offsets are generally much larger (see e.g. Table 4 in Laigle et al. 2016), an indication that in real datasets they are mainly coming from calibration issues. Because such issues are not present in the virtual photometry, there is no need to include zero-point offsets to our virtual magnitudes.

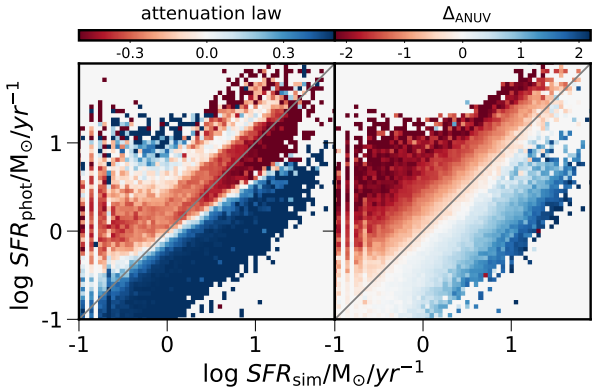

Figure B3. Comparison between SFR estimated from SED-fitting and intrinsic values at $1.1 < z < 1.5$. *Left*: The diagram is colour-coded by the attenuation law used in the best-fit template. Values of 0.5 and -0.5 are given to galaxies fitted with Arnouts et al. (2013) and the modified Calzetti et al. (2000) respectively. *Right*: The diagram is colour-coded by the difference in the attenuation computed in the *NUV* band between the one of the best-fit template, and the intrinsic one: $\Delta_{A_{NUV}} = {\rm A}^{\rm sim}_{NUV} - {\rm A}^{\rm phot}_{NUV}$.

APPENDIX D: ESTIMATING THE REDSHIFT ERRORS

D1 Robustness of $\sigma_z^{\rm fit}$

By measuring the cumulative distribution of $|z_p - z_s|/(1+\sigma_z^{\rm fit})$ in bins of magnitudes (figure 13 in L16) for the high-confidence spectroscopic sub-sample in COSMOS, L16 concluded that $\sigma_z^{\rm fit}$ derived by LePhare underestimates $\sigma_z^{\rm true}$, with a trend increasing with fainter magnitudes. In an effort to better quantify this trend, Figure C displays $\log \sigma_z^{\rm true}/\sigma_z^{\rm fit}$ in bins of magnitude and redshift for the COSMOS spectroscopic sample (*extreme left* panel) and the Horizon-AGN (*middle left* panel) simulated sample. This plot highlights that Horizon-AGN presents also an underestimation of $\sigma_z^{\rm fit}$ at bright magnitudes in the redshift range $1 < z < 2.5$. This underestimation might be due either to a remaining underestimation of magnitude errors, or to a lack of representativeness of the template library. The distribution of the reduced χ^2 in this plane (*middle right* panel) speaks in favour of this claim, as the

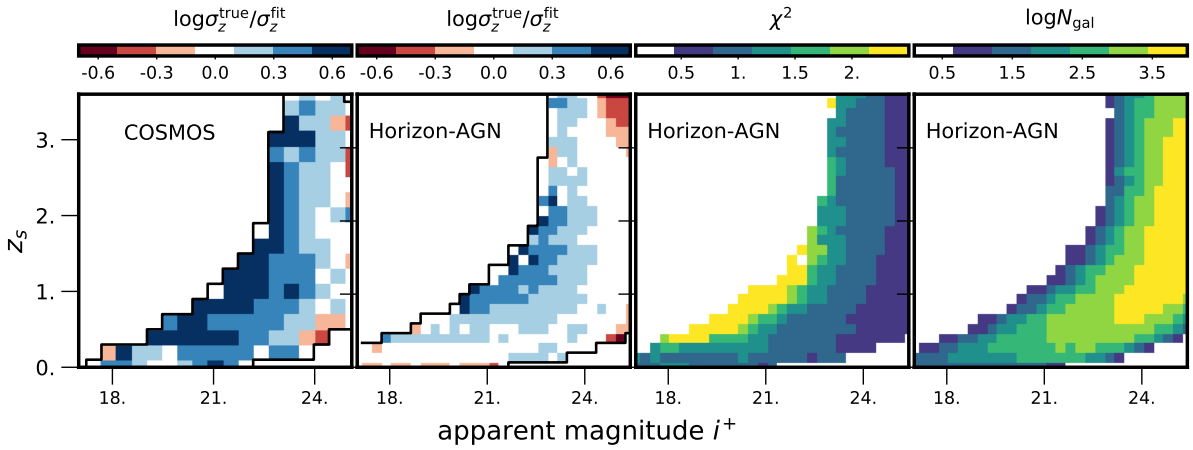

Figure C1. The plane z_s (either spectroscopic redshift for COSMOS or intrinsic redshift in HORIZON-AGN) versus apparent magnitude in the i^+-band color-coded by $\log(\sigma_z^{true}/\sigma_z^{fit})$ in COSMOS2015 (*extreme-left*) and HORIZON-AGN (*middle-left*), the reduced χ^2 in HORIZON-AGN (*middle right*) and $\log(N_{gal})$ in HORIZON-AGN.

regions with higher χ^2 broadly match those where σ_z^{fit} is underestimated the most. The underestimation is more severe in COSMOS, because the real photometry presents more diversity than the simulated one.

D2 Catastrophic outliers

Let us now investigate what causes the higher fraction of catastrophic outliers in the observed zCOSMOS sample (Lilly et al. 2007) with respect to the virtual photometric catalogue, as displayed in Fig. 2 in the main text. It is important to check if this discrepancy is driven by additional observational limitations (spectroscopic redshift misidentification[2], crowed photometry, etc.) independent on photometry modeling, or if the virtual photometric catalogue actually misses some essential components of the real galaxy population, which would make it a poor predicator of the accuracy of SED-fitting performance.
Let us therefore focus on the failure for the objects marked as red squares on Fig. 2 in the main text. After individual inspection of the spectroscopy and the photometry for each of these objects, we conclude that in the large majority of the cases, the failure arises because of one of the two reasons:

Uncertain photometry This case happens for $\sim$ 35 per cent of the outliers. The photometry extraction might be uncertain because of clumpy galaxies which might be over-split or even identified as two different objects, or on the contrary because of blended objects (which is an issue particularly severe for *NUV* or IRAC bands given the confusion limit). In the latter case, removing the IRAC bands in the SED fitting improves the match with the spectroscopic redshifts. This process does not impact the virtual catalogue because the identification of the virtual galaxies is done directly in 3D.

Spectroscopic misidentification For another ~35 per cent of the outliers, a second object on the VIMOS slit of a length of $> 10''$ is misidentified as the target in the zCOSMOS-Deep observations. Therefore, the spectroscopic redshift attributed to the original target is erroneous. In Figure D1, an example of the misidentification issue is shown. We found that these z_{spec} often agree well with the photometric objects of the second on-slit objects. The full description on the zCOSMOS-Deep sample and the redshift evaluation will be given in a future paper (Lilly et al., in prep).
Finally note that for a few outliers in COSMOS at low redshift, removing the *NUV* photometry in the SED-fitting improves the match with the spectroscopic redshifts. Recall that the *NUV* band is not used in the SED fitting of the HORIZON-AGN COSMOS-like catalogue, which could also be a reason for the lower fraction of outliers.

[2] This is a potential issue specific to the zCOSMOS-Deep sample, zCOSMOS-Bright being much more secure.

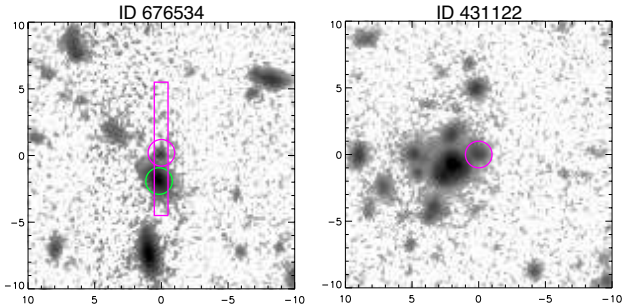

Figure D1. Examples of spectroscopic or photometric issues, not reproduced in our simulation, leading to z_{phot} catastrophic failures. Images from the Subaru i band are shown here as $20'' \times 20''$ postage stamps centred on the target galaxy (purple circle). ID numbers are from COSMOS2015. *Left:* The purple frame indicates the $10''$-long VIMOS slit ($1''$ width) and the green circle is the on-slit misidentified observed object. *Right:* blending issue where a nearby, brighter galaxy contaminates the photometry of the target

REFERENCES

Andrews B. H., Martini P., 2013, ApJ, 765, 140
Arnouts S., et al., 2013, A&A, 558, A67
Black J. H., 1981, MNRAS, 197, 553

Bolton J. S., Haehnelt M. G., Viel M., Springel V., 2005, MNRAS, 357, 1178
Bolton J. S., Puchwein E., Sijacki D., Haehnelt M. G., Kim T.-S., Meiksin A., Regan J. A., Viel M., 2017, MNRAS, 464, 897
Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, ApJ, 533, 682
Chabrier G., 2003, PASP, 115, 763
Choudhury T. R., Srianand R., Padmanabhan T., 2001, ApJ, 559, 29
Davidzon I., Ilbert O., Faisst A. L., Sparre M., Capak P. L., 2018, ApJ, 852, 107
De Cia A., Ledoux C., Savaglio S., Schady P., Vreeswijk P. M., 2013, A&A, 560, A88
Dwek E., 1998, ApJ, 501, 643
Fischera J., Dopita M., 2011, A&A, 533, A117
Fisher D. B., et al., 2014, Nature, 505, 186
Galametz M., Madden S. C., Galliano F., Hony S., Bendo G. J., Sauvage M., 2011, A&A, 532, A56
Greggio L., Renzini A., 1983, A&A, 118, 217
Henriques B. M. B., White S. D. M., Thomas P. A., Angulo R., Guo Q., Lemson G., Springel V., Overzier R., 2015, MNRAS, 451, 2663
Ilbert O., McCracken H. J., Le Fevre O., Capak P., Dunlop J., et al. 2013, A&A, 556, A55
Inoue A. K., Shimizu I., Iwata I., Tanaka M., 2014, MNRAS, 442, 1805
Iršič V., Viel M., 2014, J. Cosmology Astropart. Phys., 12, 024
Jonsson P., 2006, MNRAS, 372, 2
Kaviraj S., et al., 2017, MNRAS, 467, 4739
Laigle C., et al., 2016, ApJS, 224, 24
Lian J. H., Li J. R., Yan W., Kong X., 2015, MNRAS, 446, 1449
Lilly S. J., et al., 2007, ApJS, 172, 70
Lukić Z., Stark C. W., Nugent P., White M., Meiksin A. A., Almgren A., 2015, MNRAS, 446, 3697
Maiolino R., et al., 2008, A&A, 488, 463
Mannucci F., Cresci G., Maiolino R., Marconi A., Gnerucci A., 2010, MNRAS, 408, 2115
Maraston C., 2005, MNRAS, 362, 799
Mattsson L., Andersen A. C., Munkhammar J. D., 2012, MNRAS, 423, 26
Meiksin A., 2005, MNRAS, 356, 596
Moutard T., et al., 2016, A&A, 590, A103
Salpeter E. E., 1955, ApJ, 121, 161
Santini P., et al., 2011, eprint arXiv, 1111, 5728
Theuns T., Leonard A., Efstathiou G., Pearce F. R., Thomas P. A., 1998, MNRAS, 301, 478
Trayford J. W., et al., 2015, MNRAS, 452, 2879
Weingartner J. C., Draine B. T., 2001, ApJ, 548, 296