



Universidad Autónoma  
de Madrid

**Biblos-e Archivo**  
Repositorio Institucional UAM

**Repositorio Institucional de la Universidad Autónoma de Madrid**

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:  
This is an **author produced version** of a paper published in:

Monthly Notices of the Royal Astronomical Society 510.4 (2022): 5392-5407

**DOI:** <https://doi.org/10.1093/mnras/stac006>

**Copyright:** © 2021 The Author(s) Published by Oxford University Press on behalf of Royal Astronomical Society

El acceso a la versión del editor puede requerir la suscripción del recurso  
Access to the published version may require subscription

# UNITSIM-Galaxies: data release and clustering of emission-line galaxies

Alexander Knebe,<sup>1,2,3\*</sup> Daniel Lopez-Cano,<sup>1,9</sup> Santiago Avila,<sup>1,4</sup> Ginevra Favole,<sup>5</sup>

Adam R. H. Stevens,<sup>3</sup> Violeta Gonzalez-Perez,<sup>1,2</sup> Guillermo Reyes-Peraza,<sup>1,4</sup>

Gustavo Yepes,<sup>1,2</sup> Chia-Hsun Chuang,<sup>6</sup> Francisco-Shu Kitaura<sup>7,8</sup>

<sup>1</sup>*Departamento de Física Teórica, Módulo 15, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

<sup>2</sup>*Centro de Investigación Avanzada en Física Fundamental (CIAFF), Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

<sup>3</sup>*International Centre for Radio Astronomy Research, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia*

<sup>4</sup>*Instituto de Física Teórica, (UAM/CSIC), Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*

<sup>5</sup>*Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland*

<sup>6</sup>*Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305*

<sup>7</sup>*Instituto de Astrofísica de Canarias (IAC), Calle Vía Lactea s/n, 38200, La Laguna, Tenerife, Spain*

<sup>8</sup>*Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38206, La Laguna, Tenerife, Spain*

<sup>9</sup>*Donostia International Physics Center (DIPC), Paseo Manuel de Lardizabal, 4, 20018 Donostia-San Sebastián, Spain*

Last updated 2015 May 22; in original form 2013 September 5

## ABSTRACT

New surveys such as ESA’s Euclid mission are planned to map with unprecedented precision the large-scale structure of the Universe by measuring the 3D positions of tens of millions of galaxies. It is necessary to develop theoretically modelled galaxy catalogues to estimate the expected performance and to optimise the analysis strategy of these surveys. We populate two pairs of  $(1h^{-1}\text{Gpc})^3$  volume dark-matter-only simulations from the UNIT project with galaxies using the SAGE semi-analytic model of galaxy formation, coupled to the photoionisation model GET\_EMLINES to estimate their  $\text{H}\alpha$  emission. These catalogues represent a unique suite that includes galaxy formation physics and – thanks to the fixed-pair technique used – an effective volume of  $\sim (5h^{-1}\text{Gpc})^3$ , which is several times larger than the Euclid survey. We present the performance of these data and create five additional emission-line galaxy (ELG) catalogues by applying a dust attenuation model as well as adjusting the flux threshold as a function of redshift in order to reproduce Euclid-forecast  $dN/dz$  values. As a first application, we study the abundance and clustering of those model  $\text{H}\alpha$  ELGs: for scales greater than  $\sim 5h^{-1}\text{Mpc}$ , we find a scale-independent bias with a value of  $b \sim 1$  at redshift  $z \sim 0.5$ , that can increase nearly linearly to  $b \sim 4$  at  $z \sim 2$ , depending on the ELG catalogue. Model galaxy properties, including their emission-line fluxes (with and without dust extinction) are publicly available.

**Key words:** methods: numerical – galaxies: formation – galaxies: high-redshift – galaxies: abundances – cosmology: theory – large-scale structure of the universe

## 1 INTRODUCTION

During the last few decades, numerous projects have been aimed at creating large cartographic maps of galaxies, such as 2dFGRS (Cole et al. 2005), SDSS (York et al. 2000; Eisenstein et al. 2005), WiggleZ (Drinkwater et al. 2010; Parkinson et al. 2012), BOSS (Dawson et al. 2012; Alam et al. 2017), eBOSS (Dawson et al. 2016; eBOSS Collaboration et al. 2020) or DES (The Dark Energy Survey Collaboration 2005; Abbott et al. 2018b). They have been carried out with the objective of trying to better understand the large-scale structure of the Universe, to estimate the different parameters that regulate the formation of structures, to determine the expansion history of the Universe, to study how galaxies form, to reconstruct their star formation histories, and to impose constraints upon different models that currently exist for dark energy and for alternative theories of gravity. While advances have certainly been made, all these grand

topics remain open areas of investigation, and likely will for years to come.

New surveys such as Euclid (Laureijs et al. 2011; Amendola et al. 2013), the Nancy Grace Roman Space Telescope (Spergel et al. 2013, 2015), the Dark Energy Spectroscopic Instrument (DESI, Collaboration et al. 2016), and the 4-metre Multi-Object Spectroscopic Telescope (4MOST, de Jong et al. 2012) are planned to map with unprecedented precision the large-scale structure of the Universe by measuring the 3D positions of tens of millions of galaxies. These missions are expected to start operating in the coming years, providing the scientific community with wider, deeper, and more accurate data, which may be used to impose stronger constraints upon theoretical models and to provide more accurate estimates for some of the aforementioned parameters relevant in cosmology. Some of these forthcoming missions (e.g., Euclid) will focus on conducting spectroscopic surveys of galaxies using near-infrared grisms in order to determine the positions of galaxies by observing their emission lines such as  $\text{H}\alpha$ . The wavelength of the observed emission lines will serve to determine the redshifts of the detected objects. Such observations

\* Contact e-mail: alexander.knebe@uam.es

have already been undertaken in the past. There are, for instance, the High- $z$  Emission Line Survey (HiZELS, [Geach et al. 2008](#)) and the Wide Field Camera 3 Infrared Spectroscopic Parallels survey (WISP, [Atek et al. 2010](#)). The WISP survey, for instance, has been used by [Colbert et al. \(2013\)](#) to measure the number density evolution of  $H\alpha$  emitters; [Sobral et al. \(2016\)](#) employed the HiZELS data (and additional follow-up observations) to quantify the evolution of the  $H\alpha$  luminosity function. But all previous efforts lack the volumes to be probed by future missions.

Observational campaigns need to be complemented by cosmological simulations: a cornerstone of large-scale structure analysis. Cosmological simulations inform and validate galaxy clustering models. They are also used to test and optimise different estimators and analysis pipelines, to estimate covariance matrices, and to compare with measurements from data. Smaller scales (i.e. below 1 Mpc) are known to contain many more Fourier modes than larger ones and hence constraining power. However, they are heavily affected by the physics of galaxy formation. Since the spatial volumes that the aforementioned surveys seek to study are notoriously large, it is still necessary to rely on dark-matter-only simulations in which galaxies are introduced in post-processing either by halo occupation distribution (HOD, e.g. [Berlind et al. 2003](#), as well as the [Euclid Flagship mock galaxy catalogue](#)), (sub-)halo abundance matching (SHAM, e.g. [Vale & Ostriker 2004](#)) or semi-analytic models (SAM) (e.g. the MultiDark-Galaxies,<sup>1</sup> [Knebe et al. 2018a](#)). While there are efforts to push the limits of ‘full physics’ hydrodynamical simulations to larger and larger volumes (e.g. [Lee et al. 2020](#)), it still remains more feasible to match the volumes that missions like Euclid will cover with gravity-only simulations.

The demand for large volumes modelled with sufficiently high resolution is also the reason why, during the last years, alternatives to running such demanding simulations have been explored. For instance, the technique developed by [Angulo & Pontzen \(2016\)](#) dramatically reduces the variance arising from the sparse sampling of wavemodes in cosmological simulations. The method uses two simulations that are ‘fixed’ and ‘paired’, i.e. the initial Fourier mode amplitudes are fixed to the ensemble average power spectrum and their phases are shifted by  $\pi$ . This approach has been adopted by the UNIT collaboration<sup>2</sup> ([Chuang et al. 2019](#)) where it has been shown that the effective volume of such fixed-and-paired simulations can be several times larger than the actual volume simulated: in [Chuang et al. \(2019\)](#) we have shown that the original four ( $1h^{-1}\text{Gpc}$ )<sup>3</sup> simulations correspond to a total effective volume of ca. ( $5h^{-1}\text{Gpc}$ )<sup>3</sup>, i.e.  $\sim 7$  times of the survey volume of Euclid or DESI. We use the same two pairs of simulations for our study here. Our simulations include the large scales with an accuracy greater than expected by these surveys, and here we have populated them with galaxies using a semi-analytical model that includes all the relevant physical processes for galaxy formation. In terms of galaxy clustering statistics, each pair can be as precise on (non-)linear scales as an average over approximately 150 traditional simulations. They therefore are suitable to statistically study matter–galaxy interplay and galaxy clustering alongside its bias.

In this work we present and use galaxy catalogues for simulations that were generated by applying the SAGE semi-analytic model ([Croton et al. 2016](#)) to the aforementioned gravity-only UNIT simulations. These SAGE galaxies have then been processed with the

GET\_EMLINES code ([Orsi et al. 2014](#)) in order to obtain emission-line galaxies (ELGs). Using the resulting ELG catalogues, we study the predicted number density evolution of  $H\alpha$  emitters and compare it to other theoretical models as well as observational data. We also generate additional ELG catalogues by imposing certain flux threshold and/or even apply a dust attenuation model. All catalogues are used to study the clustering of our  $H\alpha$  galaxies and their linear bias with respect to the dark matter field, a quantity first studied by [Kaiser \(1984\)](#) for Abell clusters and developed in theoretical detail by [Bardeen et al. \(1986\)](#). The bias is a key parameter and a result of not only halo formation but also the varied physics of galaxy formation that can cause the spatial distribution of baryons to differ from that of dark matter. The bias connects the observed statistics to theoretical predictions and has recently been the target of many theoretical studies in light of ELGs (e.g. [Geach et al. 2012](#); [Cochrane et al. 2017](#); [Merson et al. 2019](#); [Tutusaus et al. 2020](#)). Our results add to these and may be used to make forecasts for Euclid and related studies for which both the abundance and bias of  $H\alpha$  ELGs is an input.

There already exist previous works based upon the UNIT simulations and the modelling of ELGs in them ([Zhai et al. 2021](#); [Zhai et al. 2019](#)). However, the important difference to our work is that in those papers only one of the UNIT simulations has been used, as opposed to all four here. Further, Zhai et al. applied a completely different modelling for the ELGs, namely the GALACTICUS semi-analytic model ([Benson 2012](#)), coupled to the CLOUDY photoionisation code ([Ferland et al. 2013](#)) for the calculation of emission line properties. Further, their dust model was tuned as a function of redshift to match observations of the  $H\alpha$  luminosity function in the redshift range  $z \in [0.8, 2.3]$ . And while Zhai et al. also studied galaxy clustering in the later work, they have not investigated the bias. Our work therefore extends those previous studies and should be viewed as complementary. We further have made our galaxy catalogues publicly available.

The structure of this article is as follows. In Section 2 the methods used to generate the ELG catalogues are presented, namely the  $N$ -body UNIT simulations (Section 2.1), the SAGE semi-analytic model (Section 2.2) and the emission-line modelling (Section 2.3). Next, in Section 3, we present a series of figures to validate the galaxy catalogues generated by SAGE by comparing key properties with observational results. Then in Section 4 we examine the validity of the modelling for the emission lines of the galaxies. Afterwards, in Section 5, the results obtained by studying the two-point correlation function and the bias obtained for the ELGs in the Euclid range of redshifts will be presented. Finally, in Section 6, the conclusions derived from this work will be outlined.

## 2 THE METHODS

### 2.1 The UNIT Simulations

As a basis for this work, four gravity-only simulations that have been developed within the UNIT project have been employed. The names for the two pairs of simulations that we use throughout this work are UNITSIM1 (U1), UNITSIM1-Inverted Phase (U1IP), UNITSIM2 (U2), and UNITSIM2-Inverted Phase (U2IP). The procedure followed for generating these simulations as well as an analysis of the resulting correlation properties is discussed in [Chuang et al. \(2019\)](#). For this particular study we have used the two pairs of simulations in which the code GADGET ([Springel et al. 2001](#)) has been used to study the behavior of a total of  $4096^3$  particles in a volume of  $1h^{-3}\text{Gpc}^3$  per simulation, thus obtaining a mass resolution of  $1.2 \times 10^9 h^{-1}M$  per simulation particle.

<sup>1</sup> Galaxy catalogues based upon three distinct SAMs can be downloaded from [CosmoSim](#).

<sup>2</sup> <http://www.unitsims.org>

In [Chuang et al. \(2019\)](#) it is also explained how the ROCKSTAR halo catalogues and the corresponding CONSISTENTTREES merger trees have been generated for each of the gravity-only simulations using the publicly available codes from [Behroozi et al. \(2012\)](#). All the data corresponding to the UNIT simulations are publicly available at <http://www.unitsims.org>. By making the galaxy catalogues and their emission-line properties available too, this work further adds to the community.

## 2.2 Semi-analytic galaxy modelling via SAGE

SAGE (Semi-Analytic Galaxy Evolution, [Croton et al. 2016](#)) is a modular, publicly available<sup>3</sup> semi-analytic model of galaxy formation, branched from the Munich family of models (specifically from [Croton et al. 2006](#)). Haloes (in this case, from the UNIT simulations) are initially seeded with ‘hot’ gas based on the cosmic baryon fraction (modulo a reionization factor at higher redshift and in low-mass haloes). Cooling/accretion of this gas onto the central galaxy is based on the two-mode (hot and cold) model of [White & Frenk \(1991\)](#). Star formation in the disc occurs once the gas is above a critical average surface density (see [Kennicutt 1989](#); [Kauffmann 1996](#)). Metals are immediately injected and gas recycled into the inter-stellar medium (ISM), where a constant mass-loading factor is also applied to re-heat gas out of the disc, some of which will end up in an ejected component if the energy budget allows it. A parametrized fraction of the ejected gas (connected to the virial velocity) is reincorporated into the halo on a dynamical time-scale. Satellite galaxies are tracked in the merger trees until merged or unresolved. Once their subhaloes become unresolved, satellites are either disrupted (where their baryons are placed in intracluster reservoirs) or immediately merged with the central, dependent on how long they survived as a satellite. SAGE, therefore, does not have orphan galaxies. Mergers and disc instabilities trigger starbursts, drive stars into the bulge, and cause gas to be accreted onto the central black hole. This triggers quasar-mode active galactic nuclei (AGN) feedback, which reheats gas from the disc. When galaxies have sufficiently (super)massive black holes, cooling is also suppressed by radio-mode AGN activity (both past and present), modelled by a phenomenological ‘heating’ radius that can only grow with time, within which gas cannot cool.

This is the same SAGE model that was also applied to the MultiDark simulation MDPL2 ([Knebe et al. 2018a](#)). The model was calibrated for that simulation by fitting visually first the  $z = 0$  stellar mass function ([Baldry et al. 2008](#)), and secondarily using the stellar metallicity–mass relation ([Tremonti et al. 2004](#)), baryonic Tully–Fisher relation ([Stark et al. 2009](#)), black hole–bulge mass relation ([Scott et al. 2013](#)), and cosmic star formation rate density ([Somerville et al. 2001](#)). The model has not been re-calibrated here as both the UNIT and MDPL2 simulations were run with the same cosmological parameters ([Planck Collaboration et al. 2015](#)) and have the same box size. However, the mass resolution is marginally better for UNITSIM, due to the 20 per cent larger number of particles. For the general performance of the SAGE model we refer the reader to the results presented in [Knebe et al. \(2018a\)](#), as the calibration plots change minimally when going from MPDL2 to UNIT (see also [Fig. 1](#), in this paper). The calibration does not include constraints for emission-line galaxies.

For a more detailed description of the model we refer the reader to [Croton et al. \(2016\)](#) and section 2.4 of [Knebe et al. \(2018a\)](#).

## 2.3 Emission-line galaxy modelling

Once we have populated the dark matter haloes from the UNIT simulations with the semi-analytic galaxies generated by SAGE we obtain values for the intensity of the most relevant emission lines such as  $H\alpha$ , [OIII]4959, [OIII]5007, [NII]6548 and [NII]6584 for each of the model galaxies. In this study we focus on the  $H\alpha$  line – with a particular focus on the Euclid mission. The other emission lines are left for future work.

**GET\_EMLINES code.** In order to reproduce the intensity of  $H\alpha$  emission lines of our galaxies, we have used the method presented in [Orsi et al. \(2014\)](#), i.e. the publicly available GET\_EMLINES code.<sup>4</sup> This code is based on the algorithm MAPPINGS-III described in [Groves et al. \(2004\)](#) and [Allen et al. \(2008\)](#), which relates the ionization parameter of gas in galaxies,  $q$ , to their cold-gas metallicity  $Z_{\text{cold}}$  as:

$$q(Z) = q_0 \left( \frac{Z_{\text{cold}}}{Z_0} \right)^{-\gamma}, \quad (1)$$

where  $q_0$  is the ionisation parameter of a galaxy that has cold gas metallicity  $Z_0$  and  $\gamma$  is the exponent of the power law. We adopted the suggested values of  $q_0 = 2.8 \times 10^7 \text{ cm s}^{-1}$  and  $\gamma = 1.3$ , which were found to yield  $H\alpha$  luminosities for star-forming galaxies in good agreement with observations ([Orsi et al. 2014](#)). Cold gas metallicity is defined as the ratio between the cold gas mass in metals to the total cold-gas mass:

$$Z_{\text{cold}} = \frac{M_{Z,\text{cold}}}{M_{\text{cold}}}. \quad (2)$$

The other relevant component is the star formation rate (SFR).<sup>5</sup> Note that SAGE provides this quantity averaged over the previous time-step in the merger trees, despite this interval being broken into sub-time-steps in the code. But the model ideally requires as inputs the instantaneous SFR and cold gas metallicity of galaxies. However, [Favole et al. \(2020\)](#) have shown that for galaxies that are not too bright the differences are negligible. To be able to properly compare our results to observations, we convert the luminosities to fluxes and also apply a dust extinction to the luminosities of the model galaxies.

Please note that when applying the GET\_EMLINES code to the SAGE catalogues,<sup>6</sup> we rejected all galaxies with a star formation rate equal to zero. One might be inclined to therefore claim that our emission-line galaxies are ‘star-forming galaxies’, but usually a threshold on the specific star formation of order 0.01/Gyr (and hence clearly larger than 0) is assumed to separate ‘passive’ and ‘star-forming’ galaxies. Therefore, our ELGs are based upon SAGE galaxies that do form stars, but also include ‘passive’ galaxies in the conventional sense.

**Dust extinction.** We use here a Cardelli extinction law implemented following [Favole et al. \(2020\)](#)<sup>7</sup>, but we also summarize it here. The attenuation from interstellar dust is added to the intrinsic  $H\alpha$  luminosity using:

$$L(\lambda_j)^{\text{att}} = L(\lambda_j)^{\text{intr}} 10^{-0.4A_\lambda(\tau_\lambda^{\tilde{\tau}}, \theta)}, \quad (3)$$

<sup>4</sup> [https://github.com/aaorsi/get\\_emlines](https://github.com/aaorsi/get_emlines)

<sup>5</sup> Total SFR in SAGE is the sum of the SfrDisk and SfrBulge fields.

<sup>6</sup> We are using the plural here when referring to the catalogues as we will always have at our disposal the four catalogues coming from the two pairs of UNIT simulations.

<sup>7</sup> <https://github.com/gfavole/dust>

<sup>3</sup> <https://github.com/darrencroton/sage>

where the attenuation coefficient, as a function of the galaxy optical depth  $\tau_\lambda^z$  and the dust scattering angle  $\theta$ , is defined as (Osterbrock 1989; Draine 2003; Izquierdo-Villalba et al. 2019; Favole et al. 2020):

$$A_\lambda(\tau_\lambda^z, \theta) = -2.5 \log_{10} \frac{1 - \exp(-a_\lambda \sec \theta)}{a_\lambda \sec \theta}. \quad (4)$$

In Eq. (4),  $a_\lambda = \sqrt{1 - \omega_\lambda} \tau_\lambda^z$ , and  $\omega_\lambda$  is the dust albedo. We assume  $\cos \theta = 0.30$  and  $\omega_\lambda = 0.56$ , meaning that the scattering is not isotropic but forward-oriented, and about 60 per cent of the extinction is caused by scattering.

The galaxy optical depth is defined as (Hatton et al. 2003; De Lucia & Blaizot 2007):

$$\tau_\lambda^z = \left( \frac{A_\lambda}{A_V} \right)_Z \left( \frac{Z_{\text{cold}}}{Z} \right)^{1.6} \left( \frac{\langle N_H \rangle}{2.1 \times 10^{21} \text{ atoms cm}^{-2}} \right), \quad (5)$$

in terms of the cold gas metallicity  $Z_{\text{cold}}$  defined in Eq. (2) and the extinction curve for solar metallicity:  $Z = 0.0134$  (Asplund et al. 2009). We assume the Cardelli et al. (1989) extinction law:

$$\left( \frac{A_\lambda}{A_V} \right) = a(x) + b(x)/R_V, \quad (6)$$

where  $x \equiv \lambda^{-1}$ ,  $R_V \equiv A_V/E(B - V) = 3.1$  is the ratio of total to selective extinction for the diffuse interstellar medium in the Milky Way, and

$$\begin{aligned} a(x) &= 1 + 0.17699y - 0.50447y^2 - 0.02427y^3 + \\ &\quad 0.72085y^4 + 0.01979y^5 - 0.77530y^6 + 0.32999y^7, \\ b(x) &= 1.41338y + 2.28305y^2 + 1.07233y^3 - 5.38434y^4 \\ &\quad - 0.62251y^5 + 5.30260y^6 - 2.09002y^7, \end{aligned} \quad (7)$$

with  $y = (x - 1.82)$ . The quantity  $\langle N_H \rangle$  in Eq. (5) is the mean hydrogen column density defined as (Hatton et al. 2003; De Lucia & Blaizot 2007):

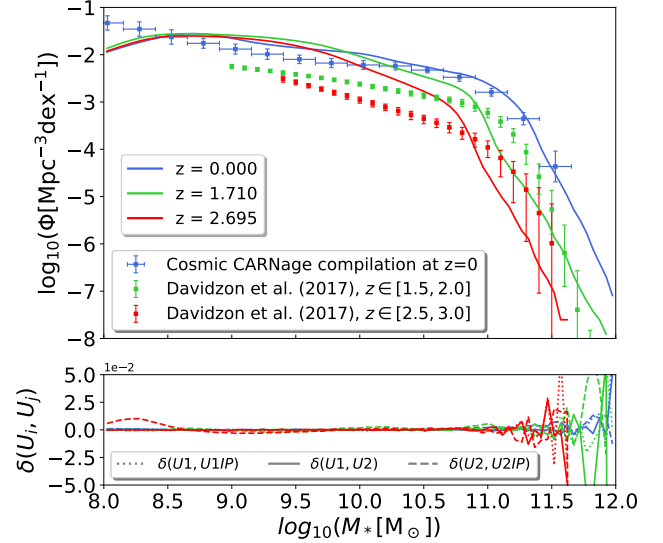
$$\langle N_H \rangle = \frac{M_{\text{cold}}^{\text{disc}}}{1.4 m_p \pi (1.68 R_{1/2}^{\text{disc}})^2} \text{ atoms cm}^{-2}, \quad (8)$$

where  $M_{\text{cold}}^{\text{disc}}$  is the cold-gas mass of the disc,  $m_p = 1.67 \times 10^{-27}$  kg is the proton mass, and  $R_{1/2}^{\text{disc}}$  is the half-mass radius of the disc.

We caution that emission lines are expected to be more attenuated than the continuum, e.g. De Barros et al. (2016), which is the model used here.

### 3 THE SAGE GALAXIES

The aim of this section is to validate how well our theoretically modelled SAGE galaxies perform with respect to the quantities that enter into the calculation of the emission-line properties. This involves a) stellar mass, b) star formation rates, c) metallicities, and d) disc lengths. We will further focus on redshifts in the range  $z \in [1, 2]$  and compare to observational data where possible. For comparisons of other properties to observations and the calibration plots, respectively, we refer the reader to Knebe et al. (2018a) where SAGE has been applied to the MultiDark simulation MDPL2. Note that in this Section we are using the complete SAGE galaxy catalogue, not restricting any results to ELGs. However, we provide in the Appendix all the corresponding plots for our model ELGs.



**Figure 1.** Stellar mass function. In the upper panel we compare the results for the modelled galaxies at various redshifts (solid lines) to observational data (points with error bars). The lower panel shows the fractional difference of U1 to the other UNIT simulations. Note that the  $z = 0$  SMF has been used to calibrate the SAGE model whereas the results for higher redshift are a prediction of the model.

#### 3.1 Stellar Mass Function

The stellar mass function (SMF) is one of the most significant properties that can be inferred from galaxy surveys since this function represents the number of galaxies in stellar-mass bins, normalized to the volume of the survey/simulation and to the bin width. Its simplicity yet fundamental importance resides in the fact that the SMF is often employed for calibrating semi-analytic models such as SAGE used here.

In the main panel of Fig. 1 the results obtained for the SMF computed from the SAGE galaxies modelled over the UNITSIM1 simulation are presented for three different redshifts  $z = [0.0, 1.710, 2.695]$ . Together with the results obtained from our simulation, a series of observational results obtained for a range of redshifts similar to those simulated are also represented in the same figure. The compilation for redshift  $z = 0$  is taken from the so-called ‘CARNage calibration’ data set described in great detail in section 3.3 and appendix A of Knebe et al. (2018b)<sup>8</sup>. The observations for the higher redshifts are taken from Davidzon et al. (2017) and are based on the UltraVISTA near-infrared survey of the COSMOS field. In the bottom panel of Fig. 1 the variation in SMF between UNITSIM1 and the three other UNIT simulations is shown, i.e. the y-axis represents<sup>9</sup>

$$\delta(U_i, U_j) = \frac{\text{SMF}(U_i)}{\text{SMF}(U_j)} - 1, \quad (9)$$

where  $U_i$  refers to the one of our four UNIT simulations (and  $U_j$  to another, different one).

For all the simulations conducted, the results produced for the SMF qualitatively follow the observational trends. This outcome is in line with previous results such as those presented in Favole et al. (2020)

<sup>8</sup> The ‘CARNage calibration’ set is available for download from <http://popia.ft.uam.es/public/CARNageSet.zip>.

<sup>9</sup> Note that we use the same strategy for presenting the variations across the four UNIT simulations in practically all plots.

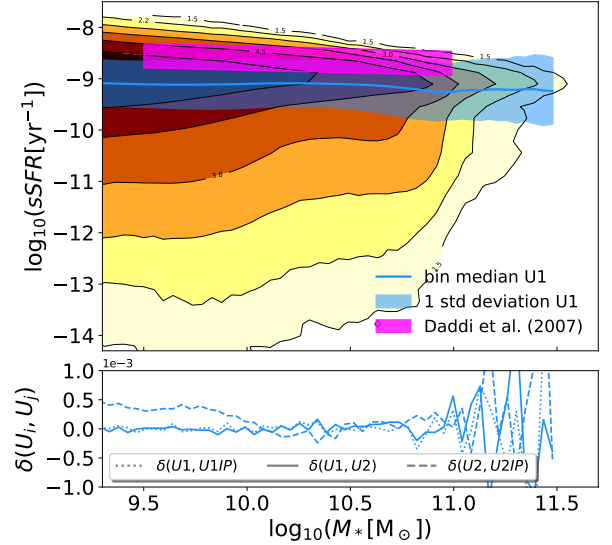
and Asquith et al. (2018). The results obtained at redshift  $z = 0$  agree almost seamlessly with the observational data. This is readily explained by the fact that the SAGE model was pre-calibrated to very similar data. When studying the behavior at higher redshifts (which is a prediction of the model) certain discrepancies start to show up. For stellar masses below  $10^{11}M_{\odot}$  the SMF calculated for the SAGE galaxies exceeds the observational points, while the opposite is true for masses higher than  $10^{11}M_{\odot}$ . This is related to the condition that getting both the SMF at  $z = 0$  and the cosmic star-formation history to simultaneously agree with the observations demands that stars that should have been formed in haloes below this simulation’s resolution limit must instead be formed as extra stars in the haloes that are resolved. This inevitably leads to resolved high- $z$  galaxies having too much stellar mass (and star-formation rates that are too high) in the model. It also changes how galaxies acquire stellar mass through mergers (as fewer mergers are resolved), which might help explain why there are too few galaxies with  $M_{*} > 10^{11}h^{-1}M_{\odot}$  at higher  $z$  in the model. Additionally, the deviations observed here for high redshifts – especially at the low-mass end – are similarly found when studying the SMF produced by other semi-analytic models, as extensively discussed in Asquith et al. (2018). Our explanation is hence generic and not only applies to SAGE. Therefore, despite the discrepancies seen in Fig. 1, the results obtained are reasonably accurate for us to say that the modelled SAGE galaxies fairly depict the behaviour of the SMF that could be expected in the redshift range for which Euclid is planned to operate.

Another important aspect worth mentioning in this section is that due to resolution limitations in our simulations, galaxies whose stellar mass is lower than  $10^9M_{\odot}$  have not been considered. Please refer to Knebe et al. (2018b,a) for a justification of this threshold, but we can also see in Fig. 1 how the number of galaxies starts to decline for stellar masses below that threshold due to numerical limitations. Therefore, to produce the results presented in the following sections we will discard all those galaxies whose mass is inferior to this threshold. This is not a cause for concern in this work though, as the vast majority of relevant ELGs have stellar masses above this threshold (see Appendix A2.1).

### 3.2 Star Formation

With respect to the star formation (SF) in galaxies, which is also used as an input to the GET\_EMLINES code, we only present the relation between specific star formation (i.e. SF per unit stellar mass) and stellar mass at redshift  $z \sim 2$ . We find that SAGE makes a prediction for this relation that is in excellent agreement with the observations of Daddi et al. (2007): in the main panel of Fig. 2 the specific SF rate (sSFR) of U1-SAGE galaxies is plotted against the stellar mass  $M_{*}$  for redshift  $z = 2.028$ . We show both the contours of a 2D histogram of this scatter plot as well as the median of the values obtained for sSFR within a series of bins along the  $x$ -axis. As is customary, in the bottom panel of the Fig. 2 the variations between simulations with respect to the other UNITSIM-SAGE galaxies have been represented. When comparing our results to observational data extracted from Daddi et al. (2007), we find sufficient agreement, at least within the  $1\sigma$  regions. Though not explicitly shown here, we also confirm that our SAGE results are in excellent agreement with observational data for the sSFR (as provided by Elbaz et al. 2011) as a function of stellar mass at redshift  $z = 0$ . These results, in turn, are also compatible to those shown in Favole et al. (2020) for redshift  $z = 0.1$ .

For a comparison of the star formation rate (SFR) function to observational data at redshift  $z = 0.14$  and the redshift evolution



**Figure 2.** Specific star formation rate vs. stellar mass at redshift  $z \sim 2$  in comparison to observations by Daddi et al. (2007). We show both the median (solid blue line) and the contour levels of the scatter relation. This relation is a prediction of the SAGE model.

of the cosmic star formation rate density, we refer the reader to Knebe et al. (2018a). While the SFR function is compatible with the observational data at low redshift – as seen for the MultiDark galaxies and also confirmed for the UNITSIM galaxies (though not explicitly presented here) – it is worth mentioning that for SFR values greater than  $\sim 10^{1.6}M_{\odot}/\text{yr}$ , the number of galaxies generated with SAGE seems to underestimate the observed number (see fig. 2 in Knebe et al. 2018a). As we will see later in Section 4.2 this is going to leave an imprint on the abundance of (dust-attenuated) ELGs, especially at high redshifts. We finally like to remark again that the relation between sSFR and stellar mass as shown here is a prediction of the SAGE model.

Based on these results, we can say that our galaxies sufficiently reproduce the behaviour of the sSFR that would be expected for a sample of real galaxies in Euclid’s operating range of redshifts.

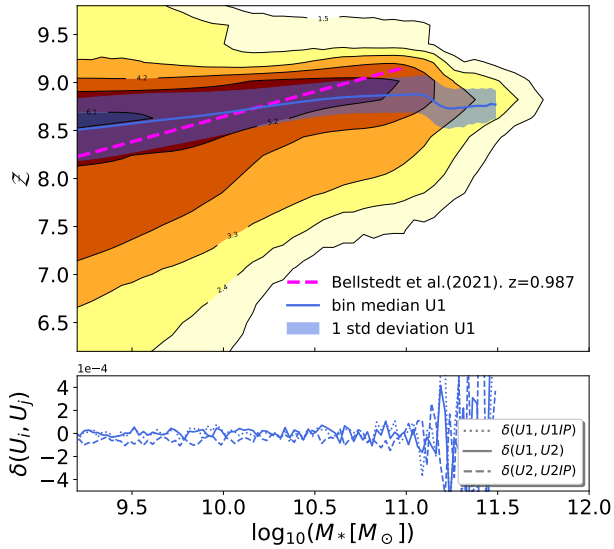
### 3.3 The mass–metallicity relation

Another aspect of galaxies to be considered for the emission-line modelling is the chemical composition, since – depending on the fraction of metals that a galaxy may contain – its SFR may be substantially modified due to the fact that a higher metal content favours cooling mechanisms. This property is explicitly taken into account by the GET\_EMLINES code and has to be provided as an input, respectively.

Since SF is regulated by the collapse of cold gas clouds, in Fig. 3 we study the relation that exists between the total mass of metals contained in such clouds and the total mass of cold gas in a given galaxy throughout the parameter  $\mathcal{Z}$  which is calculated as (Favole et al. 2020; Knebe et al. 2018a):

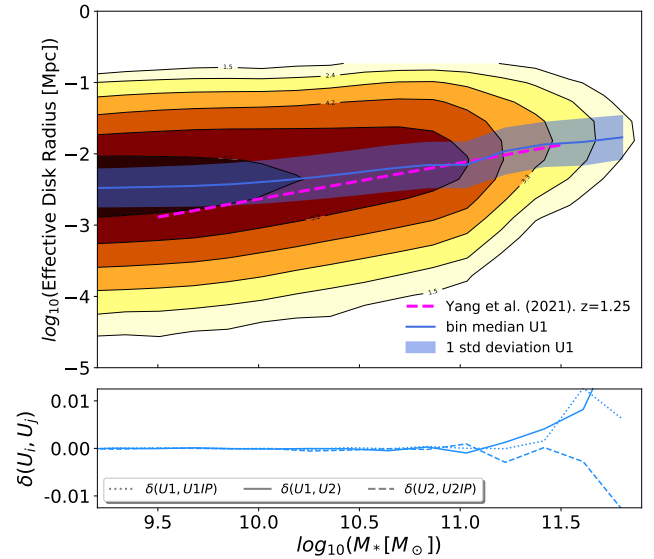
$$\mathcal{Z} = 8.69 + \log_{10}(Z_{\text{cold}}) - \log_{10}(Z) \quad (10)$$

where  $Z_{\text{cold}}$  was previously defined in Eq. (2), and we use the same  $Z = 0.0134$  as already in Eq. (5). Note that this quantity  $\mathcal{Z}$  is meant to be a proxy for  $12 + \log(\text{O}/\text{H})$ .



**Figure 3.** Cold gas metallicity vs. stellar mass. The relation shown here for redshift  $z \sim 1$  is a prediction of the SAGE model. We also show the relation as found in Bellstedt et al. (2021) (dashed line).

In the main panel of Fig. 3 we show the correlation between  $Z$  and stellar mass as contours alongside the median (solid blue line) for redshift  $z \sim 1$ . The lower panel shows again the fractional difference with respect to the other UNITSIM model. The relation is as expected, i.e. larger mass galaxies have larger metallicities, with a strength comparable to the one observed for lower redshifts. This relation – as observed at redshift  $\sim 0.1$  – is used during the calibration of the SAGE model; its extension to  $z \sim 1$  shown here nevertheless is a clear prediction. We also show the relation as expected from observations by using the best-fit function presented in Bellstedt et al. (2021, eq. 6). This fitting function was obtained by applying the spectral-energy-distribution-fitting code ProSpect to galaxies from the Galaxy and Mass Assembly (GAMA) survey at  $z < 0.06$ ; comparing with observations of gas-phase metallicity over a large range of redshifts, they then showed that their best-fit evolving mass–metallicity relationship is consistent with observations at all epochs and hence used here by us at redshift  $z \sim 1$ . We only show the Bellstedt et al. function out to  $M_* = 10^{11} M_\odot$  which was their limit for obtaining the best-fit parameters. The predictions of the SAGE model are in fair agreement with the Bellstedt et al. function. If one were to extrapolate the Bellstedt results, we would find a deficit of cold gas metallicity for the highest mass galaxies with  $M_* > 10^{11} M_\odot$ . Even though there is no observational data in that regime, one possible explanation could be that the cold gas in those galaxies comes from mergers rather than accretion/cooling. I.e. AGN feedback might have shut off cooling entirely, so enriched gas in the circumgalactic medium will not get back to the inter-stellar medium. Instead, we might just be seeing the low-metallicity gas from now-cannibalised low-mass galaxies dominating most of the cold gas in the galaxy. But it yet remains unclear if the the drop in metallicity predicted for SAGE galaxies at  $M_* = 10^{11} M_\odot$  will also be seen in observations. While the redshift  $z \sim 1$  is relevant for the Euclid mission, it also appears to be important to verify the mass-metallicity relation for even higher redshifts as it plays an important role in the estimation of emission lines. The Bellstedt et al. (2021) function can also be used to obtain results at, for instance,  $z = 2$ . There also exists a best-fit relation derived from actual observations at  $z = 2.2$  (Maiolino et al. 2008, eq. 2 together with table 5). We



**Figure 4.** Effective disc radius as a function of stellar mass (contours and blue solid line with  $1\sigma$  error region). This is a prediction of the SAGE model. We also show the relation as reported for late-type galaxies in Yang et al. (2021) at  $z = 1.25$  (dashed line).

refrain from showing the corresponding plot here, but confirm that our SAGE galaxies reproduce those two observations equally well as seen here for  $z = 1$ .

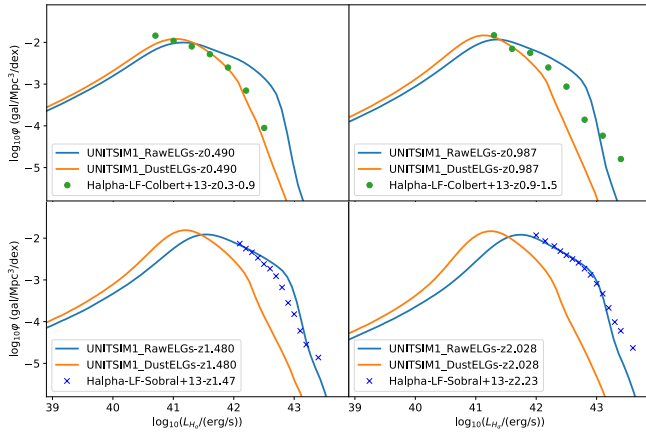
### 3.4 The disc size–mass relation

The last relevant quantity to validate for our SAGE galaxies is the size of the disc. While it is not important for GET\_EMLINES it nevertheless enters into our dust attenuation model via Eq. (8). We therefore show in Fig. 4 the correlation of the effective disc radius (i.e. exponential scale radius, as calculated by SAGE) with stellar mass at redshift  $z = 1.22$ . For comparison we use the best-fit relation as reported by Yang et al. (2021, eq. 1) for late-type galaxies at redshift  $z = 1.25$  and as derived from the complete Hubble Frontier Fields data set. While the agreement is very good for higher mass galaxies, the disc sizes predicted by SAGE for galaxies with mass  $M_* < 10^{10} M_\odot$  are systematically larger than the observed ones. However, this does not significantly affect our results here as our ELGs preferentially have stellar masses  $M_* > 10^{10.5} M_\odot$  (see Fig. A2).

Given all the results presented throughout this particular section, with the majority even being predictions of the SAGE model, we are confident that our UNITSIM-SAGE galaxies meet all the requirements to be used for the emission-line modelling, which is discussed in great detail in the following section.

## 4 SAGE’S EMISSION-LINE GALAXIES (ELGS)

The results presented in the previous section indicate that our SAGE model galaxies are in sufficient agreement with a range of observations, in particular those properties that are used as an input for the model that calculates spectral emission lines. Here we now focus on the ELGs and contrast additional properties with a set of



**Figure 5.** Evolution of the  $H\alpha$  luminosity function for our *RawELGs* (blue) and *DustELGs* catalogues (orange) ELGs in comparison to observational data: the Colbert et al. (2013) data (red points) are best compared to *DustELGs* whereas the Sobral et al. (2016) data (blue crosses) to *RawELGs* (see main text for details). Redshifts are (clockwise starting in the upper left panel)  $z = 0.490, 0.987, 1.480,$  and  $2.018$ .

observations.<sup>10</sup> To this extent, we start with generating two distinct ELG catalogues, constructed from the full list of SAGE galaxies: one set will be obtained by simply applying the `GET_EMLINES` code (*RawELGs*) and another one by additionally modelling dust extinction (*DustELGs*). These value-added properties are included in the publicly available catalogues. However, in order to compare to existing observations and to make predictions for Euclid, we apply a redshift-independent flux cut of  $F_{\text{cut}} = 2 \times 10^{-16} \text{ erg s}^{-1} \text{ cm}^{-2}$ , which corresponds to the limit of the Euclid satellite.

#### 4.1 The luminosity function of $H\alpha$ -ELGs

We start with comparing the  $H\alpha$  luminosity functions (LFs) – as obtained by `GET_EMLINES` – at various redshifts of interest to observational data. The results can be viewed in Fig. 5 for the two base catalogues *RawELGs* and *DustELGs* at  $z = 0.49, 0.987, 1.48,$  and  $2.23$ . For the first two redshifts we contrast our theoretical LFs to observations as found in Colbert et al. (2013). The data are taken from their table 3, where we removed again the [NII] contamination; as the data have not been corrected for dust extinction, they are best compared against our *DustELGs*. For the latter two redshifts, the observations from Sobral et al. (2016) are used. We used the data as provided in their table 4, noting that here they corrected for dust extinction, and hence those curves should be compared against our *RawELGs*. We actually find that our ELGs match the observations fairly well, though there are some obvious discrepancies at redshift  $z \sim 1$ : both *DustELGs* and even *RawELGs* do not provide enough high-luminosity ELGs. This eventually translates into a too-low (integrated) abundance, as we will see below. But we are not too concerned about that as the relevant redshift range for Euclid is  $z \in [0.9, 1.8]$ , and the match of our *RawELGs* galaxies with the Sobral et al. (2016) observations is rather good for  $z \sim 1.5$ , i.e. the centre of that interval.

In the Introduction we mentioned that Zhai et al. (2019) also model  $H\alpha$  ELGs using the GALACTICUS SAM coupled to the single

UNITSIM1 simulation. But their catalogue was constructed such that the SAM parameters were tuned to best reproduce – amongst other properties – the  $H\alpha$  LFs, and in particular the observed ones shown here for redshifts  $z = 1.48$  and  $2.23$  (see their fig. 1). They accomplish this by – in practice – adjusting  $A_\lambda(\tau_\lambda^z, \theta)$  (as also found in our Eq. (3)) as a free parameter, tuning it until they match the observed  $H\alpha$  LF at a given redshift. Our value for  $A_\lambda(\tau_\lambda^z, \theta)$  is based upon physical properties of the underlying galaxies whose values change as a function of redshift (leading to an implicit redshift dependence of our dust model). Meaning, we actually use a physically motivated  $A_\lambda$  and hence the LFs seen here are a clear prediction of our modelling.

#### 4.2 Abundance evolution of flux-selected $H\alpha$ -ELGs

We show in Fig. 6 the redshift evolution of the number density for our *RawELGs* and *DustELGs* catalogues, after applying the redshift-independent flux cut of  $F_{\text{cut}} = 2 \times 10^{-16} \text{ erg s}^{-1} \text{ cm}^{-2}$ , in comparison to observational data from Colbert et al. (2013) and Bagley et al. (2020). We also show two of the three models of Pozzetti et al. (2016, P16). By fitting to observed luminosity functions from existing  $H\alpha$  surveys, P16 build three distinct models for the  $H\alpha$  number density evolution. Different fitting methodologies, functional forms for the luminosity function, subsets of the empirical input data, and treatment of systematic errors were considered to explore the robustness of the results. Functional forms and model parameters were made available<sup>11</sup> (and are being used here), along with the counts and redshift distributions up to  $z \sim 2.5$  for a range of limiting fluxes bracketing the sensitivity of Euclid. Their models are named ‘Pozzetti model #1, #2, and #3’, with model #1 being the most optimistic and model #3 the most pessimistic for Euclid.<sup>12</sup> Both these models are shown here, also for a flux cut of  $2 \times 10^{-16} \text{ erg s}^{-1} \text{ cm}^{-2}$ .

We can see in Fig. 6 how, for  $z < 1$ , our *DustELGs* follow the same trends as the P16 models, but show a substantial lack of objects at higher redshift. By comparison, our *RawELGs* clearly overpredict the abundance of ELGs for the applied redshift-independent flux cut (at least for  $z > 0.5$ ). A similar discrepancy between semi-analytic galaxies and the P16 models can also be seen in fig. 5 of P16, where their three models are compared against the results from two other SAMs. It should also be mentioned that a more recent study of the observed number density evolution of  $H\alpha$  ELGs indicates a possible decline beyond redshift  $z \sim 1.4$  (Bagley et al. 2020, lower right panel of their fig. 7), although it is not as pronounced as the dip found for our *DustELGs*. To highlight this we have added those data points<sup>13</sup> to our plot, too. While there is agreement between the observations of Bagley et al. and P16’s model #3 in the redshift range  $z \in [1, 1.5]$ , the observational data drop more steeply at higher redshifts and are more in line with our *DustELGs* prediction. However, Bagley et al.

<sup>11</sup> The P16 data can be downloaded from here: <http://www.bo.astro.it/~pozzetti/Halpha/Halpha.html>

<sup>12</sup> P16 called the models that way themselves, based upon the fact that if you have more galaxies, you reduce the shot-noise. Hence, Pozzetti model #1 is more optimistic for Euclid’s figure-of-merits than #3, as we will have smaller error bars in the cosmological parameters.

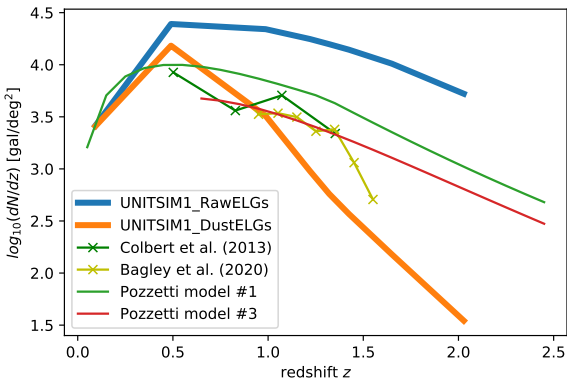
<sup>13</sup> The Bagley et al. (2020) data are based upon completeness-corrected measurements of the blended  $H\alpha$  and NII fluxes, while our fluxes include only  $H\alpha$ . We have therefore ‘corrected’ the Bagley et al. (2020) data points – as obtained with PlotDigitizer – by reversing their adjustment to model #3 of P16 to account for the combined fluxes. This was done by finding the shift needed to bring the digitized data points into the same kind of agreement with the original Pozzetti model #3, as seen in the lower right panel of Bagley’s fig. 7 for the blended Pozzetti model #3.

<sup>10</sup> The same validation plots as shown in Section 3 for the SAGE galaxies can be found for the ELGs in Appendix A2.



**Table 1.** Average number density and flux cuts as a function of redshift  $z$  (first column). Columns 2–3 (*RawELGs*) and 9–10 (*DustELGs*) list the mean and standard deviation (across the four UNIT simulations) of the number density of ELGs with an applied redshift-independent flux cut of  $F_{\text{cut}} = 2 \times 10^{-16}$  erg  $\text{s}^{-1} \text{cm}^{-2}$ . Columns 4–5 and 6–7 give the target number density (taken from table 3 in P16) and average flux cut applied to reach it (the standard deviation is smaller than the reported accuracy and hence left out for clarity) for *RawELGs-Poz1* and *RawELGs-Poz3*, respectively. Columns 10,11 and 12,13 provide the same information for *DustELGs-Poz1* and *DustELGs-Poz3*.

$z$	<i>RawELGs</i> ( $F_{\text{cut}} = 2$ )		<i>RawELGs-Poz1</i>		<i>RawELGs-Poz3</i>		<i>DustELGs</i> ( $F_{\text{cut}} = 2$ )		<i>DustELGs-Poz1</i>		<i>DustELGs-Poz3</i>	
	$\langle dN/dz \rangle$	$\sigma$	$dN/dz$	$\langle F_{\text{cut}} \rangle$	$dN/dz$	$\langle F_{\text{cut}} \rangle$	$\langle dN/dz \rangle$	$\sigma$	$dN/dz$	$\langle F_{\text{cut}} \rangle$	$dN/dz$	$\langle F_{\text{cut}} \rangle$
0.490	24652	47	9946	6.441	–	–	15262	30	9946	2.85	–	–
0.987	22015	89	7353	4.864	3779	7.080	3238	12	7353	1.37	3779	1.88
1.220	17709	94	5097	4.600	2518	6.300	957	7	5097	1.13	2518	1.50
1.321	15809	98	4281	4.452	2148	5.759	577	4	4281	1.05	2148	1.35
1.425	13988	77	3447	4.343	1817	5.353	370	3	3447	0.98	1817	1.22
1.650	10277	57	2253	3.930	1279	4.564	153	3	2253	0.84	1279	1.00
2.028	5294	38	1006	3.330	616	3.687	35	1	1006	0.67	616	0.77



**Figure 6.** Redshift evolution of the number density of dust-attenuated ELGs (*DustELGs*, orange) and the initial *RawELGs* catalogue (i.e. no dust modelling, blue), both for a redshift-independent flux cut at  $2 \times 10^{-16}$  erg  $\text{s}^{-1} \text{cm}^{-2}$ , in comparison to the observational data of Colbert et al. (2013) and Bagley et al. (2020). We also show model #1 and #3 of Pozzetti et al. (2016) for the same flux threshold (not to be confused with our catalogues *PozMod1* and *PozMod3* that were designed to match these number densities). Only UNITSIM1 ELGs are shown for clarity.

(2020) also say that their higher redshift points are in the region where the sensitivity of their instrument could be degraded.

This discrepancy between ours and the Pozzetti ELG number densities is also reflected in Table 1, where we list as a function of redshift the number density of ELGs in our reference *RawELGs* and *DustELGs* catalogues (as averaged over the four UNIT simulations, also providing the standard deviation). While we have to acknowledge that both our *RawELGs* and *DustELGs* do not reproduce the predictions of P16, we also have to remark again that it yet remains unclear what the correct abundance evolution  $dN/dz$  should look like.

### 4.3 Flux-adjusted catalogues

Taking the models of P16 as the reference, we now construct four additional catalogues that are designed to match the P16  $dN/dz$  curves as shown in Fig. 6. We take *RawELGs* as the starting point and adjust the flux threshold until reaching the target  $dN/dz$  values as given by P16’s models #1 and #3, providing us with the two models *RawELGs-Poz1* and *RawELGs-Poz3*. We use the same approach for *DustELGs*,

providing two more models: *DustELGs-Poz1* and *DustELGs-Poz3*. We used this methodology with all four UNITSIM catalogues. The means of the required flux cuts to our data are listed in columns 4–7, and 10–13 of Table 1 (we omit error estimates as they are below the reported accuracy). The remaining columns – 2, 3, 8, and 9 – are the mean number densities (and its standard deviation) for the *RawELGs* and *DustELGs* catalogues, respectively, when using a redshift independent flux threshold of  $F_{\text{cut}} = 2 \times 10^{-16}$  erg  $\text{s}^{-1} \text{cm}^{-2}$ . Our methodology for constructing ELGs eventually leaves us with six distinct catalogues<sup>14</sup>

***RawELGs*:** directly coming from GET\_EMLINES (with a flux threshold of  $F_{\text{cut}} = 2 \times 10^{-16}$  erg  $\text{s}^{-1} \text{cm}^{-2}$  across all redshifts, when used here),

***RawELGs-Poz1*:** variable flux threshold applied to *RawELGs* to match the number density of Pozzetti’s model #1 at each redshift,

***RawELGs-Poz3*:** variable flux threshold applied to *RawELGs* to match the number density of Pozzetti’s model #3 at each redshift,

***DustELGs*:** passing the *RawELGs* ELGs through our dust model (with a flux threshold of  $F_{\text{cut}} = 2 \times 10^{-16}$  erg  $\text{s}^{-1} \text{cm}^{-2}$  across all redshifts, when used here),

***DustELGs-Poz1*:** variable flux threshold applied to *DustELGs* to match the number density of Pozzetti’s model #1 at each redshift,

***DustELGs-Poz3*:** variable flux threshold applied to *DustELGs* to match the number density of Pozzetti’s model #3 at each redshift,

where we note that all the ELGs are, by construction, a subset of the full SAGE catalogue used in the previous section. Likewise, the four additional ‘*ELGs-Poz*’ catalogues are sub-sets of the public *RawELGs* and *DustELGs*, respectively.

Instead of introducing a redshift-dependent flux cut – which might be considered counter-intuitive, as Euclid will have a fixed flux threshold – we could have also taken the *RawELGs* model as the starting point and tuned our dust extinction parameters until we match the P16  $dN/dz$  values, akin to what Zhai et al. (2019) have done. But finding the best possible dust model is beyond the scope of this work and hence we prefer to adhere to the former approach. The main idea here is to restrict the model ELGs to the brightest ones that are still observable. And we have seen in Fig. 5 that applying the

<sup>14</sup> We need to state here again that the public versions of *RawELGs* and *DustELGs* are not subjected to any flux cut: they contain all ELGs as provided by GET\_EMLINES.

dust model basically just shifts the LF towards lower luminosities, especially at high redshift and for the brightest ELGs (e.g. Sobral et al. 2016). Therefore, adjusting the luminosity threshold will still select the brightest galaxies. Moreover, one could also re-calibrate SAGE, the `GET_EMLINES` code or choose a different dust model beyond a Cardelli law, all of which can affect the number density of ELGs. But exploring all these possibilities is beyond the scope of the present work. We prefer to work with minimal variations to the existing models and codes.

We also like to emphasize that our ‘-Poz1’ and ‘-Poz3’ models are *not* the two models #1 and #3 of P16. They are ELG catalogues where we adjusted the number densities to match those of model #1 and #3 of Pozzetti, respectively. We did this to correct for the mismatch of ELGs with respect to the Pozzetti models seen in Fig. 6. We further refrain from showing their abundance evolution as they match – by construction – the curves from P16.

Given the results presented in this section, we conclude that our UNITSIM-SAGE-ELGs provide a fair sample and can be used for further analysis. The *RawELGs* and *DustELGs* galaxies will serve as the two base catalogues, with the four additional catalogues acting as our best predictions for Euclid. As a particular application we employ them now for a study of galaxy clustering and the related bias.

## 5 CLUSTERING OF ELGS

Quantifying the clustering of galaxies is one of the main objectives of ongoing and upcoming galaxy surveys such as the Euclid satellite mission. Clustering measurements probe the fluctuations of the underlying dark matter from the positions of galaxies, and they encode geometric, model-dependent cosmological information. Using the positions of our theoretical UNITSIM ELGs, we now study the two-point correlation function  $\xi_{\text{ELGs}}(r)$  and its redshift evolution. We further use the positions of  $10^7$  randomly selected dark matter particles from the total  $4096^3$  particles present in each of the UNIT gravity-only simulations to calculate  $\xi_{\text{DM}}(r)$ .<sup>15</sup> This allows us to also infer the bias that we define here as

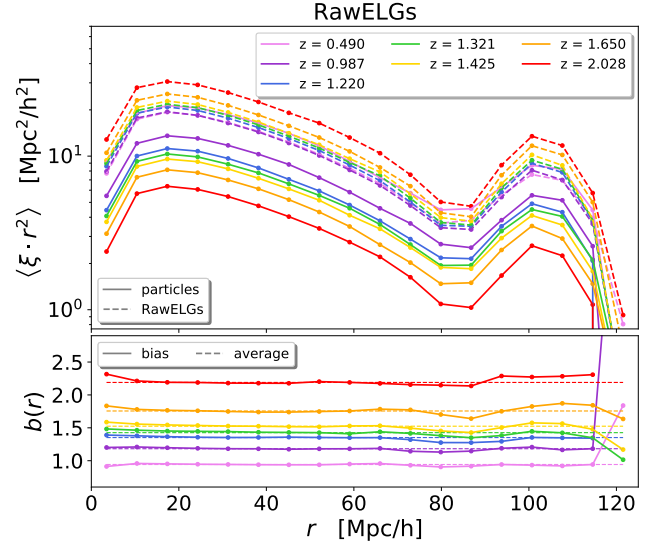
$$b(r) = \sqrt{\frac{\xi_{\text{ELGs}}(r)}{\xi_{\text{DM}}(r)}} \quad (11)$$

between both populations and study its evolution across redshift. The bias  $b$ , i.e. the statistical relation between the distribution of galaxies and matter, needs to be taken into account when interpreting galaxy surveys; it describes how galaxies trace the underlying dark matter distribution. The biased galaxy formation scenario (e.g. Dekel & Rees 1987) implies that galaxies are not uniformly distributed in the Universe, but primarily form in the peaks of the matter density field. Galaxies are therefore biased tracers of it, sampling only the overdense regions (see Desjacques et al. 2018, for a recent review). The particular bias of ELGs, i.e. a sub-class of all galaxies, will be of greatest importance for surveys such as Euclid.

All two-point correlation functions (2PCFs) have been obtained with the CUTE<sup>16</sup> software presented in Alonso (2012). In addition, for the results that we will present throughout this section, we have taken the average of the values computed for the 2PCF over

<sup>15</sup> We confirm that the resulting 2PCFs have converged and will not change when using more particles. Further, this number of dark matter particles is comparable to the number of ELGs, at least at redshifts  $z \leq 1$ .

<sup>16</sup> <https://github.com/damonge/CUTE>



**Figure 7.** Top panel: 2PCF of SAGE ELGs with flux greater than  $2 \times 10^{-16}$  erg  $s^{-1} \text{cm}^{-2}$  (*RawELGs* galaxies, dashed lines) and collisionless trace particles (solid lines) for various redshifts. Bottom panel: associated bias as defined by Eq. (11).

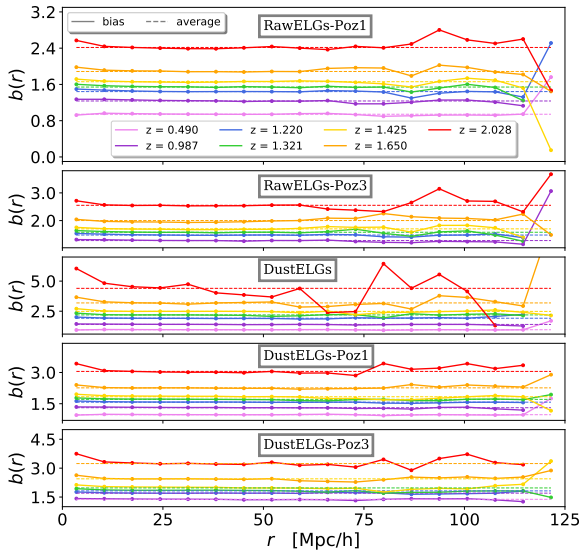
the four simulations UNITSIM1, UNITSIM1-InvertedPhase, UNITSIM2, and UNITSIM2-InvertedPhase.

In the top panel of Fig. 7 we present the 2PCF computed for the *RawELGs* (dashed lines) and dark matter (solid lines). The lower panel of the same figure shows the bias  $b(r)$  defined via Eq. (11). In order to better verify the scale-dependence of the bias, we also calculate the ‘average’ bias

$$\langle b \rangle = \frac{1}{N_{\text{bin}} - 1} \sum_2^{N_{\text{bin}}} b_i, \quad (12)$$

where  $N_{\text{bin}}$  is the number of bins and  $b_i = b(r_i)$  the value of the bias in distance bin  $r_i$ . This average bias  $\langle b \rangle$  is shown as a dashed horizontal line in the lower panel of Fig. 7. Note that we exclude the first bin in this calculation since for such small distances the bias is certainly scale-dependent (see Fig. 10 below). It is also obvious that the data for this particular model become rather noisy at high redshifts due to the very low number of objects above the reference flux cut of  $F_{\text{cut}} = 2 \times 10^{-16}$  erg  $s^{-1} \text{cm}^{-2}$  (see Table 1). But we can nevertheless appreciate that for distances  $r \gtrsim 5h^{-1} \text{Mpc}$  the bias is remarkably constant, something we will quantify in more detail below.

An equivalent analysis has been conducted for our other ELG catalogues, but we decided to only show here in Fig. 8 the results for the bias and not also the 2PCFs. Once more we can see that we get fairly noisy results at redshift  $z = 2.028$  due to the reduced number of galaxies at that redshift. We also observe that at scales  $\sim 120h^{-1} \text{Mpc}$  the bias behaves more erratic, which can be explained by the fact that the 2PCF crosses zero at  $r \sim 130h^{-1} \text{Mpc}$  (Sánchez et al. 2008; Prada et al. 2011): taking the numerical ratio between two numbers close to zero then introduces noise. But the most important point is that the bias of ELGs (at least for  $z < 2$ ) in all our catalogues remains constant on scales  $r \in [5, 100]h^{-1} \text{Mpc}$  (in line with the findings of, for instance, Abbott et al. 2018a). Below  $5h^{-1} \text{Mpc}$  it is obvious that the mixture contribution between the one- and two-halo terms will introduce non-linear effects which in turn will cause the bias to no longer behave independently with scale. On larger scale we have



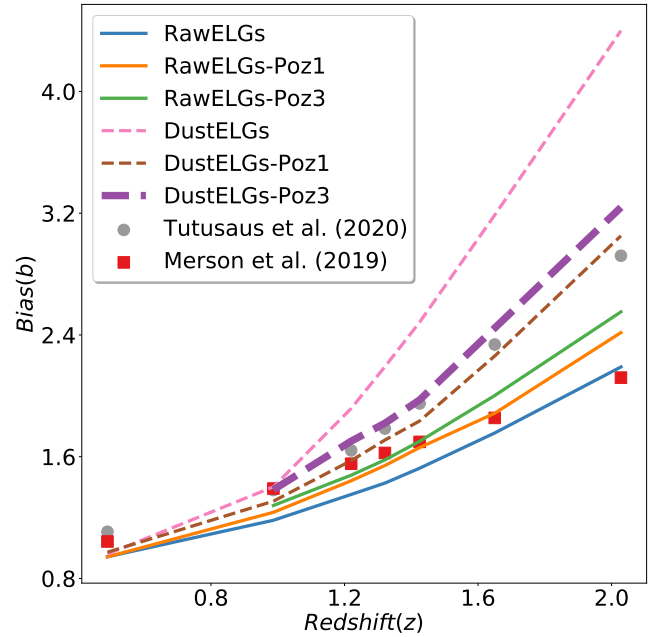
**Figure 8.** The bias for the *RawELGs-Poz1*, *RawELGs-Poz3*, *DustELGs*, *DustELGs-Poz1*, and *DustELGs-Poz3* galaxies (in that order from top to bottom), using the same  $r$ -range and redshift colouring as for Fig. 7.

already seen above that the zero-crossing of the 2PCF is introducing noise and hence the results for the bias are expected to be affected by this, too. We further note that the bias clearly is a function of redshift. But this is also expected, as the mass of the haloes hosting ELGs will change with redshift (see Fig. A1 in the Appendix). Not only that, but haloes of the same mass or luminosity at different redshifts will also have a different bias. It therefore only appears natural that the bias increases with redshift as, for instance, modelled analytically by Basilakos et al. (2008) or found in other cosmological simulations (e.g. Merson et al. 2019; Tutusaus et al. 2020).

Fig. 9 now quantifies the evolution of the average bias  $\langle b \rangle$  (obtained from the results presented in Fig. 7 and Fig. 8) as a function of redshift for all our catalogues. This figure is accompanied by Table 2 that lists the plotted values. We find that for all our galaxies the bias systematically increases with redshift, despite showing different growth rates, especially for the two base catalogues *RawELGs* and *DustELGs*. We also acknowledge that the strength of this  $b(z)$  relation for our four ‘-Poz’ galaxies – especially the ones based upon *DustELGs* – is in excellent agreement with the relation presented in Tutusaus et al. (2020, eq. 11), shown as circles in Fig. 9. The  $b(z)$  function given in Tutusaus et al. is derived from studying the bias in the Euclid *Flagship* simulation,<sup>17</sup> which is also just based upon dark matter. But the way in which the dark matter haloes are populated with galaxies is quite distinct to our approach: they have applied a Halo Occupation Distribution (HOD) that does not take into account the merger trees of the haloes (for a comparison of these two different techniques see, for instance, Knebe et al. 2015, 2018b).<sup>18</sup> Merson

<sup>17</sup> [https://www.euclid-ec.org/?page\\_id=4133](https://www.euclid-ec.org/?page_id=4133)

<sup>18</sup> While there is no reference paper for this galaxy catalogue, we nevertheless like to mention that it is based upon the MICE HOD (Carretero et al. 2015). The clustering is fit to SDSS galaxies as a function of magnitude and colour at low redshift. Then, most of the properties are assumed to depend on redshift only via their SEDs/color evolution, allowing for correlations between many observables.



**Figure 9.** Bias values averaged for scales larger than  $5 h^{-1}$  Mpc computed for our six models. The grey points are the best fit  $b(z)$  for the bias found in Euclid’s *Flagship* simulation (Tutusaus et al. 2020), and the red squares the results as reported by Merson et al. (2019).

et al. (2019) also forecast the redshift evolution of the linear bias for H $\alpha$ -emitting galaxies in a similar redshift range. Their data are shown here as squares. Like Tutusaus et al. (2020), they also used a HOD for which they calibrated the dust attenuation to reproduce observed H $\alpha$  counts. Merson et al. (2019) now predict lower biases than Tutusaus et al. (2020) and our dust-based ‘-Poz’ galaxies, more in line with the results we obtain for our *RawELGs* catalogue and its derivatives. The comparison of these three different  $b(z)$  predictions for ELGs indicates that the theoretical models have not yet converged. There are degeneracies and uncertainties that still require more detailed and refined investigations before any final conclusion could be drawn. But we finally remark that our findings for the redshift evolution of the bias  $b(z)$  are also in agreement with those of Favole et al. (2017, right panel of their fig. 6), who used a SHAM model. However, in their work, the bias increases more mildly, as the SDSS redshift range studied there is very much reduced compared to ours.

We further recognize in Fig. 9 that the bias is sensitive to the particulars of our modelling, especially at high redshift. This certainly relates to how we treat the dust extinction and select the observable ELGs from *RawELGs*, respectively. But this is known and can also be appreciated when comparing the bias predictions from Tutusaus et al. (2020) and Merson et al. (2019) where similar discrepancies are seen. We particularly notice the degeneracy between dust modelling and flux selection: first applying our extinction prescription and then matching a preset  $dN/dz$  by varying the flux threshold always leads to larger bias than not employing a dust model at all. Even though we argued before that the dust-attenuated luminosities – as seen in Fig. 5 – are a shifted version of the raw values (at least for luminous ELGs; see also Sobral et al. 2016, where a constant luminosity offset was applied to model dust extinction), here we realize that their relation is not that simple. But we have clearly seen that fixing the abundance of H $\alpha$  ELGs, the differences substantially reduce. Nevertheless, we like to stress again that designing a new dust extinction model is beyond

**Table 2.** Bias values averaged for scales larger than  $5h^{-1}\text{Mpc}$  computed for all our ELG catalogues. The values listed here correspond to the lines presented in Fig. 9.

$z$	<i>RawELGs</i>	<i>RawELGs-Poz1</i>	<i>RawELGs-Poz3</i>	<i>DustELGs</i>	<i>DustELGs-Poz1</i>	<i>DustELGs-Poz3</i>
0.490	0.94	0.94	–	0.96	0.97	–
0.987	1.18	1.23	1.28	1.40	1.31	1.38
1.220	1.35	1.44	1.48	1.92	1.57	1.70
1.321	1.42	1.54	1.58	2.19	1.71	1.82
1.425	1.53	1.66	1.70	2.48	1.83	1.97
1.650	1.76	1.89	2.00	3.19	2.26	2.45
2.028	2.19	2.42	2.55	4.40	3.05	3.24

the scope of this work and hence we leave a more detailed study of this to a future work. Note that in this work we primarily aim at presenting the publicly available data, discussing its scope and possible limitations.

So far we have mainly focused on large scales, but to conclude this section we also present how the bias varies for *small* scales. In Fig. 10 we present the bias  $b(z)$  for various redshifts and all our catalogues out to  $r \approx 20h^{-1}\text{Mpc}$  using logarithmic binning. We observe that for redshifts  $z < 2$  the bias remains constant down to scales  $r \approx 3h^{-1}\text{Mpc}$  and then starts to mildly drop. It is actually around this distance that we expect the contribution from the one-halo term to start to become relevant. However, this behaviour weakens for higher redshifts and possibly reverses for  $z = 2$ . Something similar has also been observed by Nuzi et al. (2012, fig. 10) for BOSS CMASS galaxies, but there the inversion was already seen at redshift  $z \approx 0.53$  (and one needs to bear in mind that CMASS galaxies and ELGs are not directly comparable as they are different types of galaxies, where the latter are mostly star-forming and the former could be dominated by passive galaxies).

## 6 CONCLUSIONS

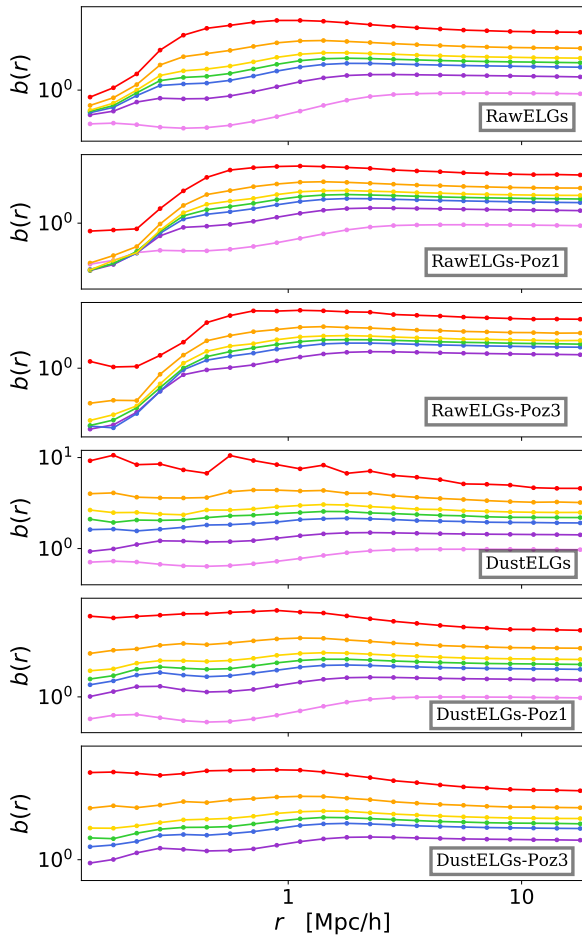
Realistic simulations are a necessary tool to optimise and validate the methodology that will be used to extract cosmological constraints from future surveys. Indeed, they are used to estimate the theoretical error budget on surveys (for example, for the eBOSS-ELG analysis, see Alam et al. 2021). In this work we have employed the UNIT simulations, which model the evolution of dark matter within a  $1h^{-1}\text{Gpc}$  box at a mass resolution of  $1.2 \times 10^9 h^{-1}M_{\odot}$  per particle (Chuang et al. 2019). Given the large volume of these simulations together with the fixed-and-paired technique of Angulo & Pontzen (2016) that enhances the effective volume of the simulations, the resulting galaxy mocks that we have produced represent a unique resource for model testing based on a semi-analytic model of galaxy formation. We used our ELG catalogues to make predictions for the galaxy statistics that the Euclid experiment is expected to obtain for redshifts between  $0.9 < z < 1.8$ . Note that the simulations presented here cover an effective survey volume of about seven times the effective survey volume of Euclid (Chuang et al. 2019). And having the galactic physics included is key, since the complicated relation between haloes and galaxies can modify the clustering of ELGs significantly, even at scales used to put cosmological constraints when working in Fourier space (see, for instance, Gonzalez-Perez et al. 2020; Avila et al. 2020).

For this work we have generated six synthetic catalogues of emission-line galaxies of which the two base ones (i.e. *RawELGs*

and *DustELGs*, without any flux cuts applied) are publicly available. The galaxies were first obtained by applying the semi-analytic galaxy formation model SAGE (Croton et al. 2016) to the gravity-only UNIT simulations. They were then subjected to the emission line modelling with the GET\_EMLINES code (Orsi et al. 2014) and an additional dust attenuation model (following Favole et al. 2020). This left us with the two base ELG catalogues *RawELGs* and *DustELGs*, in addition to the general SAGE galaxy catalogues. As argued throughout Section 3, the properties associated with our UNITSIM-SAGE galaxies reproduce observed properties of galaxies with  $0 \leq z < 2$ . Here we have focused on those properties that are most relevant for the construction of ELGs catalogues, i.e. stellar mass, star formation rate, metallicity, and disc size. In particular, we find that the (evolution of the) mass–metallicity relation agrees sufficiently well with observations. However, we have seen in Knebe et al. (2018a) that the SAGE model underpredicts the number of galaxies with high SFRs. This then affects the abundance of our (dust-attenuated) ELGs as seen in Fig. 6. While we presented the validation plots in the main body of the paper only for the full set of SAGE galaxies, the corresponding plots for the *RawELGs* and *DustELGs* ELGs can be found in Appendix A2.

In Section 4 we adjusted the number densities of our two base UNITSIM-ELG sets by applying distinct flux thresholds to them (using the Euclid-models as given by P16), eventually comparing the redshift evolution of their abundance to observations. When studying the density of galaxies per  $\text{deg}^2$  with fluxes greater than  $2 \times 10^{-16} \text{erg s}^{-1} \text{cm}^{-2}$  as a function of redshift we observe that the density obtained for the raw ELG galaxies is above both the observations and other theoretical modelling. That means that some additional selection needs to be applied to end up with a more realistic ELG catalogue. We have addressed this in several ways. We first applied a dust-attenuation (a Cardelli law, following Favole et al. 2020), which led to a possible underestimation of the expected density of galaxies  $dN/dz$  observed from redshift  $z \sim 1.4$  onwards. Nevertheless, the most recent study by Bagley et al. (2020) suggests that the observational value for  $dN/dz$  could be closer to our results than predicted by P16. We also designed additional catalogues where we instead varied the flux threshold for the selection of galaxies from the *RawELGs* catalogue; those fluxes were adjusted to reproduce number densities as predicted by P16.

The linear bias is a key parameter to understand the cosmological power of Euclid and can help construct forecasts that inform the optimisation of both observational and analysis strategies. The bias of  $H\alpha$  galaxies may be particularly relevant for forecasts on studies such as primordial non-Gaussianities or relativistic effects. We therefore studied the clustering of all our six samples listed in Table 1: two with the Euclid flux cut applied and four in which the



**Figure 10.** The bias  $b(r)$  for all six models in logarithmic  $r$  bins, focusing on the small scales up to  $r < 20h^{-1}\text{Mpc}$ , using the same redshift colouring as for Fig. 7.

flux cuts are adjusted to follow the predictions by two of the models presented in P16. We measure the linear bias as a function of redshift by averaging  $\xi_{\text{ELGs}}/\xi_{\text{DM}}$  for scales  $r > 5h^{-1}\text{Mpc}$ . For the samples whose abundances are matched to the to P16 predictions, we find a  $b(z)$  in line with that reported in Tutusaus et al. (2020) for the Euclid *Flagship* simulation (and mildly in agreement with the same results reported by Merson et al. 2019). This is striking, as the *Flagship* mock was constructed following a very different methodology (Carretero et al. 2015). Additionally, we report the clustering at small scales, that becomes scale-dependent. These measurements can be used to test the robustness of different large-scale structure models to extract cosmological information from the small scales, that have the highest signal-to-noise ratio but at the same time are the most difficult to model.

We close with the remark that an improved dust attenuation modelling might be the most physical approach for choosing the ELGs so that the observed  $dN/dz$  will be recovered. This would,

however, only affect the catalogues that are based upon *DustELGs*; it will leave *RawELGs* untouched, which is the primary ELG catalogue made available publicly. Therefore, while we have shown throughout this work that the particulars of the dust extinction have an effect on the the results, the published data contain all that is required for the community to apply their favourite post-processing models for dust and emission lines from star-forming regions. Or put differently, the base catalogue *RawELGs* is certainly inclusive, i.e. a superset of the ELGs of interest. A better understanding of the process of selecting observable ELGs from that base catalogue and developing an improved dust attenuation model will be left for a future work. The public data can already been used for a great variety of studies and have extensive applications like, for example, informing Halo Occupation Distribution models. Indeed, we will study the properties of  $\text{H}\alpha$  ELG HOD models in a follow-up paper.

## ACKNOWLEDGEMENTS

We thank the anonymous referee for a detailed and constructive report that helped to improve the content of the paper and its presentation.

AK and GY are supported by the MICIU/FEDER through grant number PGC2018-094975-C21. AK further thanks Chasing Dorothea for all i want. SA is supported by the MICUES project, funded by the EU H2020 Marie Skłodowska-Curie Actions grant agreement no. 713366 (InterTalentum UAM). DLC acknowledge the support of the ERC-StG number 716151 (BACCO). GF acknowledges financial support from the SNF 175751 “Cosmology with 3D Maps of the Universe” research grant. VGP and SA are supported by the Atracción de Talento Contract no. 2019-T1/TIC-12702 granted by the Comunidad de Madrid in Spain. ARHS acknowledges receipt of the Jim Buckee Fellowship at ICRAR-UWA. FSK acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (MINECO) under the Severo Ochoa program SEV-2015-0548, and for the grants RYC2015-18693 and AYA2017-89891-P. GR is grateful for the financial support granted by the Ministry of Education of Spain (FPI, MEC) with reference PRE2018-087035, for the I+D project with reference to SEV -2016-0597 - 18-2 IFT (UAM-CSIC).

The UNIT simulations have been produced in the MareNostrum Supercomputer, hosted by the Barcelona Supercomputing Center, Spain, under the PRACE project number 2016163937.

This research has made use of NASA’s Astrophysics Data System and the arXiv preprint server. This work was created by making use of the following software: PYTHON, MATPLOTLIB (Hunter 2007), NUMPY (van der Walt et al. 2011), SCIPY (Virtanen et al. 2019), and ASTROPY (Astropy Collaboration et al. 2013, 2018).

## DATA AVAILABILITY

The data are available at <http://www.unitsims.org>. There one finds the halo catalogues and merger trees of the underlying dark matter-only simulations as well as the SAGE galaxies alongside their emission-line information. These latter ELG files contain the two base catalogues *RawELGs* and *DustELGs*, but without applying any flux cut.

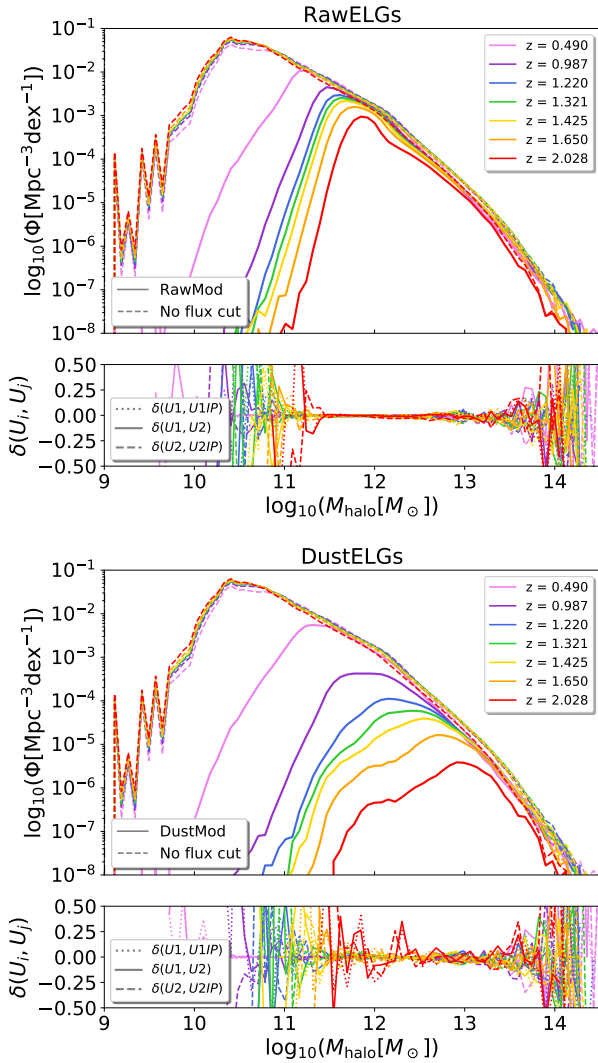
All files are in hdf5 format. Sample reading routines that specify the properties stored inside the files (and their units) are available there too, but can also be downloaded from here:

[http://popia.ft.uam.es/public/read\\_SAGE.py](http://popia.ft.uam.es/public/read_SAGE.py)  
[http://popia.ft.uam.es/public/read\\_ELGs.py](http://popia.ft.uam.es/public/read_ELGs.py)

The data are directly available for those redshifts used in this paper; for other redshifts the data will be available upon request.

## REFERENCES

- Abbott T. M. C., et al., 2018a, *Phys. Rev. D*, **98**, 043526  
 Abbott T. M. C., et al., 2018b, *The Astrophysical Journal Supplement Series*, **239**, 18  
 Alam S., et al., 2017, *MNRAS*, **470**, 2617  
 Alam S., et al., 2021, *MNRAS*, **504**, 4667  
 Allen M. G., Groves B. A., Dopita M. A., Sutherland R. S., Kewley L. J., 2008, *The Astrophysical Journal Supplement Series*, **178**, 20–55  
 Alonso D., 2012, preprint, ([arXiv:1210.1833](https://arxiv.org/abs/1210.1833))  
 Amendola L., et al., 2013, *Living Reviews in Relativity*, **16**  
 Angulo R. E., Pontzen A., 2016, *Monthly Notices of the Royal Astronomical Society: Letters*, **462**, L1–L5  
 Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, *ARA&A*, **47**, 481  
 Asquith R., et al., 2018, *Monthly Notices of the Royal Astronomical Society*, **480**, 1197  
 Astropy Collaboration et al., 2013, *A&A*, **558**, A33  
 Astropy Collaboration et al., 2018, *AJ*, **156**, 123  
 Atek H., et al., 2010, *ApJ*, **723**, 104  
 Avila S., et al., 2020, *MNRAS*, **499**, 5486  
 Bagley M. B., et al., 2020, *ApJ*, **897**, 98  
 Baldry I. K., Glazebrook K., Driver S. P., 2008, *MNRAS*, **388**, 945  
 Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, *ApJ*, **304**, 15  
 Basilakos S., Plionis M., Ragone-Figueroa C., 2008, *ApJ*, **678**, 627  
 Behroozi P. S., Wechsler R. H., Wu H.-Y., 2012, *The Astrophysical Journal*, **762**, 109  
 Bellstedt S., et al., 2021, arXiv e-prints, [p. arXiv:2102.11514](https://arxiv.org/abs/2102.11514)  
 Benson A. J., 2012, *New Astronomy*, **17**, 175  
 Berlind A. A., et al., 2003, *ApJ*, **593**, 1  
 Cardelli J. A., Clayton G. C., Mathis J. S., 1989, *ApJ*, **345**, 245  
 Carretero J., Castander F. J., Gaztañaga E., Crocce M., Fosalba P., 2015, *MNRAS*, **447**, 646  
 Chuang C.-H., et al., 2019, *Monthly Notices of the Royal Astronomical Society*, **487**, 48–59  
 Cochrane R. K., Best P. N., Sobral D., Smail I., Wake D. A., Stott J. P., Geach J. E., 2017, *MNRAS*, **469**, 2913  
 Colbert J. W., et al., 2013, *ApJ*, **779**, 34  
 Cole S., et al., 2005, *MNRAS*, **362**, 505  
 Collaboration D., et al., 2016, *The DESI Experiment Part I: Science, Targeting, and Survey Design* ([arXiv:1611.00036](https://arxiv.org/abs/1611.00036))  
 Croton D. J., et al., 2006, *MNRAS*, **365**, 11  
 Croton D. J., et al., 2016, *ApJS*, **222**, 22  
 Daddi E., et al., 2007, *ApJ*, **670**, 156  
 Davidzon I., et al., 2017, *A&A*, **605**, A70  
 Dawson K. S., et al., 2012, *The Astronomical Journal*, **145**, 10  
 Dawson K. S., et al., 2016, *The Astronomical Journal*, **151**, 44  
 De Barros S., Reddy N., Shivaee I., 2016, *ApJ*, **820**, 96  
 De Lucia G., Blaizot J., 2007, *MNRAS*, **375**, 2  
 Dekel A., Rees M. J., 1987, *Nature*, **326**, 455  
 Desjacques V., Jeong D., Schmidt F., 2018, *Phys. Rep.*, **733**, 1  
 Draine B. T., 2003, *Annual Review of Astronomy and Astrophysics*, **41**, 241  
 Drinkwater M. J., et al., 2010, *Monthly Notices of the Royal Astronomical Society*, **401**, 1429–1452  
 Eisenstein D. J., et al., 2005, *ApJ*, **633**, 560  
 Elbaz D., et al., 2011, *Astronomy & Astrophysics*, **533**, A119  
 Favole G., Rodríguez-Torres S. A., Comparat J., Prada F., Guo H., Klypin A., Montero-Dorta A. D., 2017, *MNRAS*, **472**, 550  
 Favole G., et al., 2020, *MNRAS*, **497**, 5432  
 Ferland G. J., et al., 2013, *Rev. Mex. Astron. Astrofis.*, **49**, 137  
 Geach J. E., Smail I., Best P. N., Kurk J., Casali M., Ivison R. J., Coppin K., 2008, *MNRAS*, **388**, 1473  
 Geach J. E., Sobral D., Hickox R. C., Wake D. A., Smail I., Best P. N., Baugh C. M., Stott J. P., 2012, *MNRAS*, **426**, 679  
 Gonzalez-Perez V., et al., 2020, *MNRAS*, **498**, 1852  
 Groves B. A., Dopita M. A., Sutherland R. S., 2004, *The Astrophysical Journal Supplement Series*, **153**, 75–91  
 Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, *MNRAS*, **343**, 75  
 Hunter J. D., 2007, *Computing in Science & Engineering*, **9**, 90  
 Izquierdo-Villalba D., Bonoli S., Spinoso D., Rosas-Guevara Y., Henriques B. M. B., Hernández-Monteagudo C., 2019, *MNRAS*, **488**, 609  
 Kaiser N., 1984, *ApJ*, **284**, L9  
 Kauffmann G., 1996, *MNRAS*, **281**, 475  
 Kennicutt Jr. R. C., 1989, *ApJ*, **344**, 685  
 Knebe A., et al., 2015, *MNRAS*, **451**, 4029  
 Knebe A., et al., 2018a, *MNRAS*, **474**, 5206  
 Knebe A., et al., 2018b, *MNRAS*, **475**, 2936  
 Laureijs R., et al., 2011, *Euclid Definition Study Report* ([arXiv:1110.3193](https://arxiv.org/abs/1110.3193))  
 Lee J., et al., 2020, arXiv e-prints, [p. arXiv:2006.01039](https://arxiv.org/abs/2006.01039)  
 Maiolino R., et al., 2008, *A&A*, **488**, 463  
 Merson A., Smith A., Benson A., Wang Y., Baugh C., 2019, *MNRAS*, **486**, 5737  
 Nuza S. E., Sanchez A. G., Prada F., Klypin A., Schlegel D. J., Gottloeber S., Montero-Dorta A. D., Manera M. e. a., 2012, preprint, ([arXiv:1202.6057](https://arxiv.org/abs/1202.6057))  
 Orsi A., Padilla N., Groves B., Cora S., Tecce T., Gargiulo I., Ruiz A., 2014, *Monthly Notices of the Royal Astronomical Society*, **443**, 799–814  
 Osterbrock D. E., 1989, *Astrophysics of gaseous nebulae and active galactic nuclei*  
 Parkinson D., et al., 2012, *Phys. Rev. D*, **86**, 103518  
 Planck Collaboration et al., 2015, preprint, ([arXiv:1502.01589](https://arxiv.org/abs/1502.01589))  
 Pozzetti L., et al., 2016, *A&A*, **590**, A3  
 Prada F., Klypin A., Yepes G., Nuza S. E., Gottloeber S., 2011, arXiv e-prints, [p. arXiv:1111.2889](https://arxiv.org/abs/1111.2889)  
 Sánchez A. G., Baugh C. M., Angulo R. E., 2008, *MNRAS*, **390**, 1470  
 Scott N., Graham A. W., Schombert J., 2013, *ApJ*, **768**, 76  
 Sobral D., Stroe A., Koyama Y., Darvish B., Calhau J., Afonso A., Kodama T., Nakata F., 2016, *MNRAS*, **458**, 3443  
 Somerville R. S., Primack J. R., Faber S. M., 2001, *MNRAS*, **320**, 504  
 Spergel D., et al., 2013, *Wide-Field InfraRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA Final Report* ([arXiv:1305.5422](https://arxiv.org/abs/1305.5422))  
 Spergel D., et al., 2015, *Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report* ([arXiv:1503.03757](https://arxiv.org/abs/1503.03757))  
 Springel V., Yoshida N., White S. D., 2001, *New Astronomy*, **6**, 79–117  
 Stark D. V., McGaugh S. S., Swaters R. A., 2009, *AJ*, **138**, 392  
 The Dark Energy Survey Collaboration 2005, arXiv e-prints, [pp astro-ph/0510346](https://arxiv.org/abs/astro-ph/0510346)  
 Tremonti C. A., et al., 2004, *ApJ*, **613**, 898  
 Tutusaus I., et al., 2020, arXiv e-prints, [p. arXiv:2005.00055](https://arxiv.org/abs/2005.00055)  
 Vale A., Ostriker J. P., 2004, *MNRAS*, **353**, 189  
 Virtanen P., et al., 2019, arXiv e-prints, [p. arXiv:1907.10121](https://arxiv.org/abs/1907.10121)  
 White S. D. M., Frenk C. S., 1991, *ApJ*, **379**, 52  
 Yang L., Roberts-Borsani G., Treu T., Birrer S., Morishita T., Bradač M., 2021, *MNRAS*, **501**, 1028  
 York D. G., et al., 2000, *AJ*, **120**, 1579  
 Zhai Z., Benson A., Wang Y., Yepes G., Chuang C.-H., 2019, *Monthly Notices of the Royal Astronomical Society*, **490**, 3667–3678  
 Zhai Z., Chuang C.-H., Wang Y., Benson A., Yepes G., 2021, *MNRAS*, **501**, 3490  
 de Jong R. S., et al., 2012, *Ground-based and Airborne Instrumentation for Astronomy IV*  
 de los Reyes M. A., et al., 2015, *AJ*, **149**, 79  
 eBOSS Collaboration et al., 2020, arXiv e-prints, [p. arXiv:2007.08991](https://arxiv.org/abs/2007.08991)  
 van der Walt S., Colbert S. C., Varoquaux G., 2011, *Computing in Science Engineering*, **13**, 22



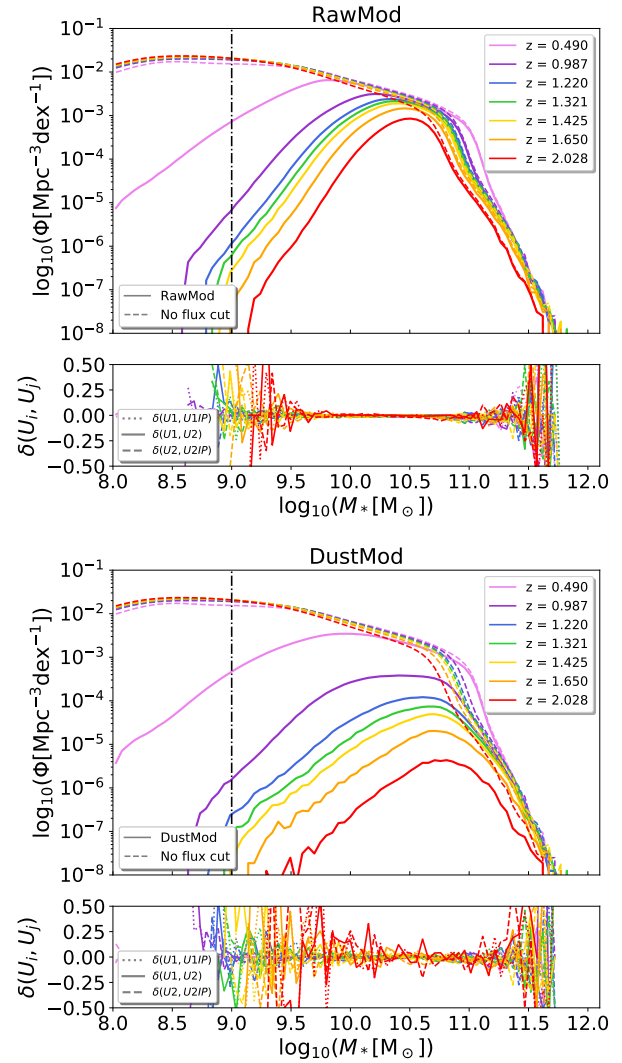
**Figure A1.** Halo mass function of all ELGs (dashed lines) and the flux-selected samples (solid lines) for *RawELGs* (top) and *DustELGs* (bottom).

## APPENDIX A: ADDITIONAL VALIDATION PLOTS

In this Appendix we provide supplementary plots that further show the validity and properties of the SAGE and ELG galaxies used throughout this study.

### A1 Halo Mass Function of flux-selected ELGs

Applying a SAM will eventually lead to selecting a sub-sample of the underlying dark matter haloes as galaxies, i.e. while every halo contains a galaxy, some might be too small to be detectable. To better understand which haloes host our ELGs, we show their halo mass functions for the two base models *RawELGs* and *DustELGs* for various redshifts in Fig. A1. The dashed lines are without applying any flux cut, whereas the solid lines use the Euclid-inspired cut  $F_{\text{cut}} = 2 \times 10^{-16} \text{ erg s}^{-1} \text{ cm}^{-2}$ . We can see that the flux cut primarily affects low-mass haloes, i.e. the less luminous ELGs also live in lower mass host haloes. We further observe a shift of this ‘cut-off’ halo mass with redshift; while at  $z \sim 0.5$  it is approximately  $10^{11} M_{\odot}$ , it increases to  $\sim 10^{12} M_{\odot}$  at  $z \sim 2$  for *RawELGs* and even  $\sim 10^{13} M_{\odot}$  for *DustELGs*.



**Figure A2.** Stellar mass function of all ELGs (dashed lines) and the flux-selected samples (solid lines) for *RawELGs* (top) and *DustELGs* (bottom). The vertical dot-dashed line shows our lower stellar mass limit.

### A2 Baryonic properties of flux selected ELGs

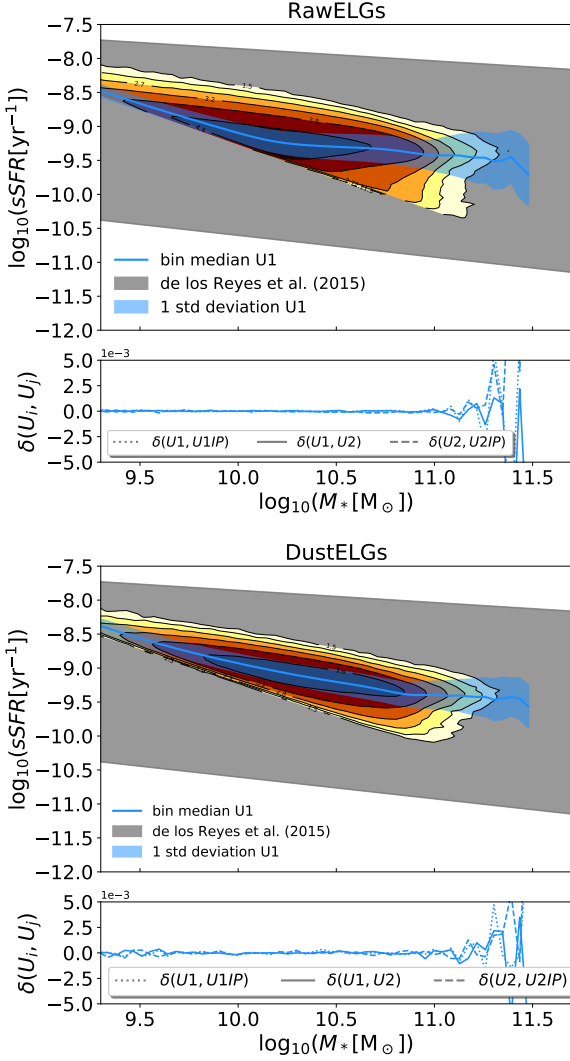
In Section 3 we presented baryonic relations for the full set of SAGE galaxies, focusing on those properties that are relevant for the dust attenuation modelling. Here we now like to provide counterparts of those plots for the ELGs.

#### A2.1 Stellar Mass Function

In order to view the effect of the flux selection and its relation to the stellar masses of the resulting sub-sample of ELGs, we show in Fig. A2 both the SMF of all ELGs (i.e. no flux cut, dashed lines) and the flux-selected samples of ELGs (solid lines) for various redshifts. We restrict the results again to the two base models *RawELGs* and *DustELGs*. We appreciate that the majority of ELGs coincide with the most massive galaxies.

#### A2.2 Specific star formation rate

In Fig. 2 we show the specific star formation rate of all our SAGE galaxies in comparison to the observations of Daddi et al. (2007) at



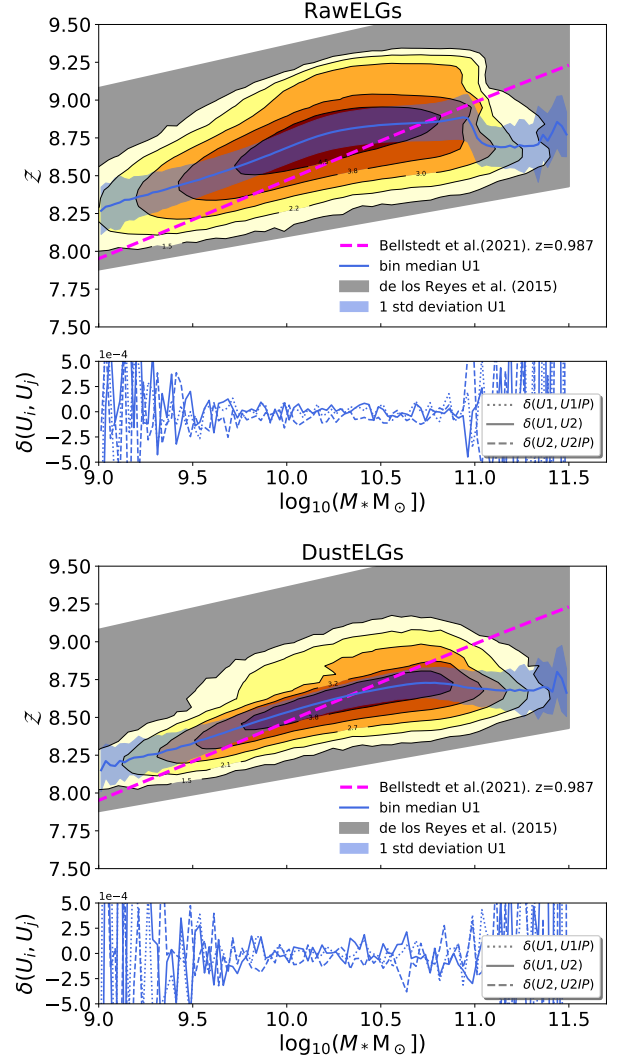
**Figure A3.** Specific star formation rate of the *RawELGs* (top) and *DustELGs* (bottom) ELGs at redshift  $z \sim 1$  in comparison to the best-fit relation as found by de los Reyes et al. (2015) at  $z \sim 0.8$ , shown as grey-shaded region. This figure is a reproduction of Fig. 2, but this time for our model ELGs.

redshift  $z \sim 2$ . Here we now present in Fig. A3 another version of that plot, this time using the (flux-cut) ELGs of the *RawELGs* and *DustELGs* catalogues. We further show results for  $z \sim 1$  and add the best-fitting correlation for  $H\alpha$  emitting galaxies, as found by de los Reyes et al. (2015, eq. 3).<sup>19</sup>

### A2.3 The mass–metallicity relation

Here we reproduce Fig. 3 for the *RawELGs* and *DustELGs* catalogues, additionally adding the best-fit relation for  $H\alpha$ -emitting galaxies, as reported by de los Reyes et al. (2015, eq. 4). The results can be viewed in Fig. A4, which shows that the SAGE-ELGs follow the observations sufficiently well.

<sup>19</sup> de los Reyes et al. (2015) studied 299  $H\alpha$ -selected galaxies at redshift  $z \sim 0.8$ .



**Figure A4.** Cold gas metallicity vs. stellar mass for all ELGs for *RawELGs* (top) and *DustELGs* (bottom) at redshift  $z \sim 1$ . This figure is a reproduction of Fig. 3, but this time for our model ELGs, but we also added the best-fitting relation as found by de los Reyes et al. (2015) at  $z \sim 0.8$ , shown as grey-shaded region.

### A2.4 The disc size–mass relation

At last we turn to the effective disc size of our *RawELGs* and *DustELGs* galaxies, shown in Fig. 4 for all SAGE galaxies. The results can be viewed in Fig. A5, again in comparison to the general results of Yang et al. (2021).

## APPENDIX B: CONVERSION OF NUMBER DENSITIES

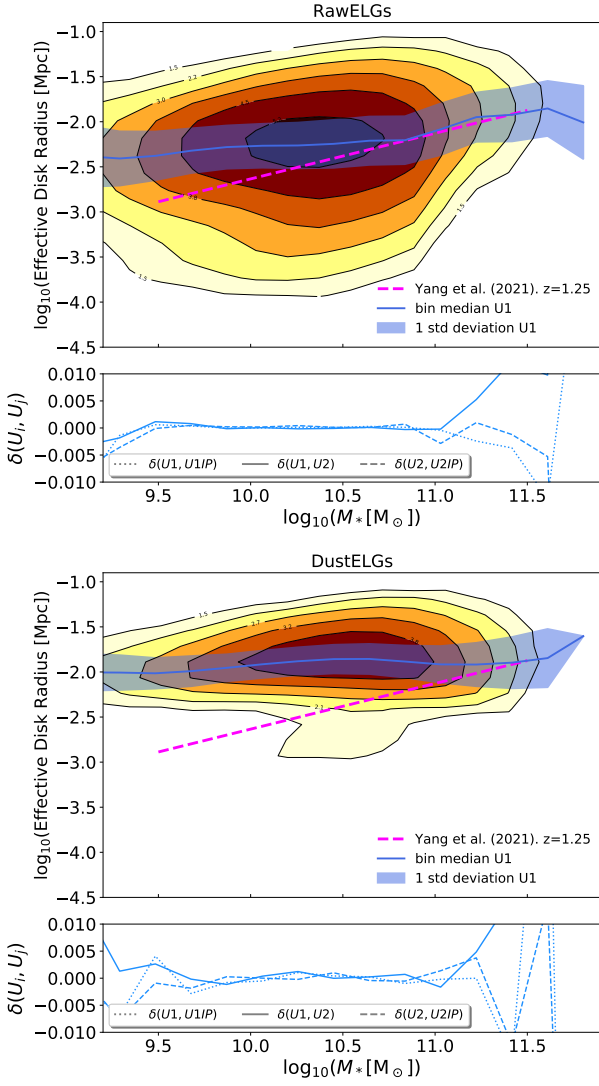
Here we show the steps necessary to go from volumetric number density

$$n = \frac{dN}{dV} \quad (\text{B1})$$

to the angular and redshift density

$$\eta = \frac{dN}{d\Omega dz} \quad (\text{B2})$$





**Figure A5.** Effective disc radius as a function of stellar mass at redshift  $z = 1.25$  for all *RawELGs* (top) and *DustELGs* (bottom) galaxies. This figure is a reproduction of Fig. 4, but this time for our model ELGs.

Taking into account

$$dV = d\Omega r^2 dr \quad (\text{B3})$$

where  $d\Omega$  is the solid angle in steradians, we then get

$$\eta = n \cdot r^2 \frac{dr}{dz}. \quad (\text{B4})$$

Therefore, to go from number density  $n = N/V$  of galaxies to number density of galaxies per square degree and redshift interval we find

$$\eta = n r^2(z) \frac{dr}{dz} \left( \frac{\pi}{180^\circ} \right)^2, \quad (\text{B5})$$

where  $r(z)$  is the comoving distance

$$r(z) = \frac{c}{H_0} \int_0^z \frac{ds}{E(s)} \quad (\text{B6})$$

with

$$E^2(z) = \frac{1}{(\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{k,0}(1+z)^2 + \Omega_{\Lambda,0})},$$

(B7)

where  $\Omega_X$  are the usual density parameters of radiation ( $X = r$ ), matter ( $X = m$ ), curvature ( $X = k$ ), and cosmological constant ( $X = \Lambda$ ) at present time. We note that the derivative of  $r(z)$  with respect to  $z$  as needed in Eq. (B5) is simply

$$\frac{dr}{dz} = \frac{c}{H_0} \frac{1}{E(z)}. \quad (\text{B8})$$

Note that in the main body of the paper  $\eta$  is referred to as  $dN/dz$ , which is not fully consistent with the terminology used here, but compliant with how other workers in the field refer to this quantity.  $N$  as used in the main part is ‘number of galaxies per unit area’, whereas here it simply means ‘number of galaxies’.

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.