

# A Galaxy-Halo Model for Multiple Cosmological Tracers

Philip Bull<sup>1,2</sup>★

<sup>1</sup>California Institute of Technology, Pasadena, CA 91125, USA

<sup>2</sup>Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California, USA

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

The information extracted from large galaxy surveys with the likes of DES, DESI, Euclid, LSST, SKA, and WFIRST will be greatly enhanced if the resultant galaxy catalogues can be cross-correlated with one another. Predicting the nature of the information gain, and developing the tools to realise it, depends on establishing a consistent model of how the galaxies detected by each survey trace the same underlying matter distribution. Existing analytic methods, such as halo occupation distribution (HOD) modelling, are not well-suited for this task, and can suffer from ambiguities and tuning issues when applied to multiple tracers. We construct a simple alternative that provides a common model for the connection between galaxies and dark matter halos across a wide range of wavelengths (and thus tracer populations). This is based on a chain of parametrised statistical distributions that model the connection between (a) halo mass and bulk physical properties of galaxies, such as star-formation rate; and (b) those same physical properties and a variety of emission processes. The result is a flexible parametric model that allows analytic halo model calculations to be carried out for multiple tracers, as well as providing semi-realistic galaxy properties for fast mock catalogue generation.

**Key words:** galaxies – cosmology: theory

## 1 INTRODUCTION

The viability of large-scale structure surveys as a cosmological probe rests on our ability to understand the connection between galaxies and the dark matter distribution that they inhabit. Galaxies are the luminous ‘tracers’ of the dark matter field that we actually observe, while the clustering of the invisible dark matter, much of which has collapsed into discrete halos, bears the bulk of the cosmological information. Without a sufficiently accurate model of their relationship, forthcoming galaxy surveys are limited in their ability to return accurate, precision cosmological constraints (Reid et al. 2014; Clerkin et al. 2015). As a result, there has been a concerted effort to develop theoretical models of the galaxy-dark matter connection, and validate them against simulations and existing observations across a wide range of frequencies (e.g. Berlind et al. 2003; Zehavi et al. 2005; Zheng et al. 2009; Skibba & Sheth 2009; White et al. 2011).

A defining feature of forthcoming surveys is their coverage of large fractions of the sky. A number of the datasets will therefore cover substantially overlapping cosmological volumes, enabling cross-correlation analyses to be performed. The relationship between different galaxy samples (often seen at different wavelengths) yields valuable additional information beyond what a single tracer can provide, both about the galaxies themselves and the large-scale matter distribution (e.g. Gaztañaga et al. 2012). Cross-correlations can be used to side-step the cosmic variance limit on certain observables, for example (McDonald & Seljak

2009); to defeat otherwise difficult systematic effects; or to uncover the physical properties and formation processes of the galaxies themselves (Johnston et al. 2007).

While extremely promising, such analyses also require the application of a suitable galaxy-halo model – with the added complication that it must be *consistent* between tracers. This is challenging for methods designed primarily for the single-tracer case. Existing multi-wavelength data are often limited in depth, or by sample variance, making it hard to empirically calibrate the models against two or more tracers simultaneously, while including the necessary correlation information between tracers.

In this paper, we construct a simple, modular analytic model for the connection between galaxies and halos that provides consistent predictions across multiple wavelengths and tracer populations. We achieve this by first using empirical scaling relations to connect the host halo mass to bulk physical properties of the galaxies, such as stellar mass and star-formation rate. These relationships are described in a statistical manner, by constructing appropriately-parametrised probability distributions around the mean scaling relations (which are also parametrised). The physical properties are then related to a variety of emission processes by another set of scaling relations, each of which is relevant to some wavelength regime. Correlations between different types of emission (and thus different frequency bands) arise because of their shared dependence on the basic physical properties of the galaxy.

The ‘building blocks’ of the model – a chain of interconnected parametric statistical distributions (Fig. 1) – are mostly based on established relations that are found in the literature, such as the star-forming main sequence, or stellar mass-halo mass re-

★ E-mail: philip.j.bull@jpl.nasa.gov

lation. By *jointly* fitting the parameters of the model to various datasets, rather than relying on empirical calibrations of individual components from the literature, we can ensure that the components of the model all connect together in a consistent way. Each component can be replaced or upgraded as more realistic models become available, and uncertainties on model parameters can be straightforwardly propagated through the full system to determine their effects on observables, using standard Monte Carlo techniques or even analytic marginalisation.

The model is differentiated from similar approaches in the literature by its analytic construction, meaning that it can be used for rapid parameter estimation and forecasting, as well as the more usual task of painting galaxy properties onto dark matter-only simulations to create mock catalogues. A reference implementation, written in Python, is made publicly-available with this paper.

As a proof of concept, in this paper we will construct a model for optical and radio continuum emission from ‘normal’ (non-AGN) galaxies at  $z \approx 0$ , and calibrate it off multi-wavelength luminosity function data (plus a semi-analytic simulation for one model component). Simple extensions of the model can be made to generalise it to higher redshifts and several other tracer populations. Other novel features include a new set of analytic fitting functions that connect optical magnitudes to star-formation rate and stellar mass, and a simple dust attenuation model.

The paper is organised as follows. In Sect. 2 we briefly review various approaches that have been used to model galaxy populations. We then define the components of our model in Sect. 3, and calibrate it against simulations and observations in Sect. 4. We also analyse how well the model fits the input data, and discuss some of the other observables it can predict. We then conclude in Sect. 5 with a discussion of the model’s limitations, and ways that it could be improved and extended.

We denote the natural logarithm by  $\log$  and the base-10 logarithm by  $\log_{10}$  throughout, and assume the [Planck Collaboration \(2016\)](#) best-fit  $\Lambda$ CDM cosmology with  $h = 0.67$  and  $\Omega_M = 0.32$ .

## 2 GALAXY POPULATION MODELLING

In this section, we briefly review some of the methods for modelling galaxy populations. These generally fall into one of three main categories:

- Hydrodynamical simulations;
- Semi-analytic models;
- Halo model approaches.

*Hydrodynamical simulations* (e.g. [Hopkins et al. 2014](#); [Vogelsberger et al. 2014](#)) are typically used to make samples of galaxies with highly realistic properties. By explicitly modelling the physical processes relevant to galaxy formation and evolution, very detailed simulations can be constructed, albeit at very high computational expense. This typically limits the simulations to small spatial volumes, so that sufficiently high spatial resolutions can be achieved to model the necessary processes. ‘Sub-grid’ models can be used to incorporate processes that happen on unresolved length scales, although these typically introduce a number of free parameters that must be tuned to match observations. While useful for building a detailed understanding of galaxies, hydrodynamical simulations are generally too expensive to use for applications such as creating large numbers of mock galaxy catalogues for the analysis of large-scale structure surveys.

*Semi-analytic models* (SAMs) are commonly used to populate dark matter-only N-body simulations with galaxies. They also work by explicitly modelling the various physical processes responsible for determining observable galaxy properties, subject to

some simplifying assumptions designed to reduce computational overhead ([Somerville & Primack 1999](#); [Kauffmann & Haehnelt 2000](#); [Springel et al. 2005](#); [Benson 2012](#); [Baugh 2013](#)). The simplified physical models in SAMs often have tunable parameters that can be calibrated against observations. While intended to be cheaper than hydrodynamical simulations but much more realistic than halo model approaches, SAMs are still computationally intensive due to the need to repeatedly solve systems of differential equations to determine the properties of each galaxy. This makes them unwieldy for performing Monte Carlo parameter studies for example, which require many different realisations of the galaxy population to be constructed for different sets of parameter values (although such studies have been attempted; see e.g. [Henriques et al. 2009](#)). Minimal SAMs do exist, which can be used to more rapidly model some subset of galaxy properties (e.g. [Cohn 2016](#)).

*Halo model approaches* are a much lighter alternative than either of the previous two methods. These work by assigning galaxy properties to dark matter halos, using a set of scaling relations that depend on properties such as the halo mass. Halo models have the advantage of often being simple enough to evaluate analytically, making them suitable for cheaply populating mock catalogues (e.g. [Manera et al. 2013](#); [Koda et al. 2016](#)), or performing statistical analyses (e.g. [Reid et al. 2014](#)).

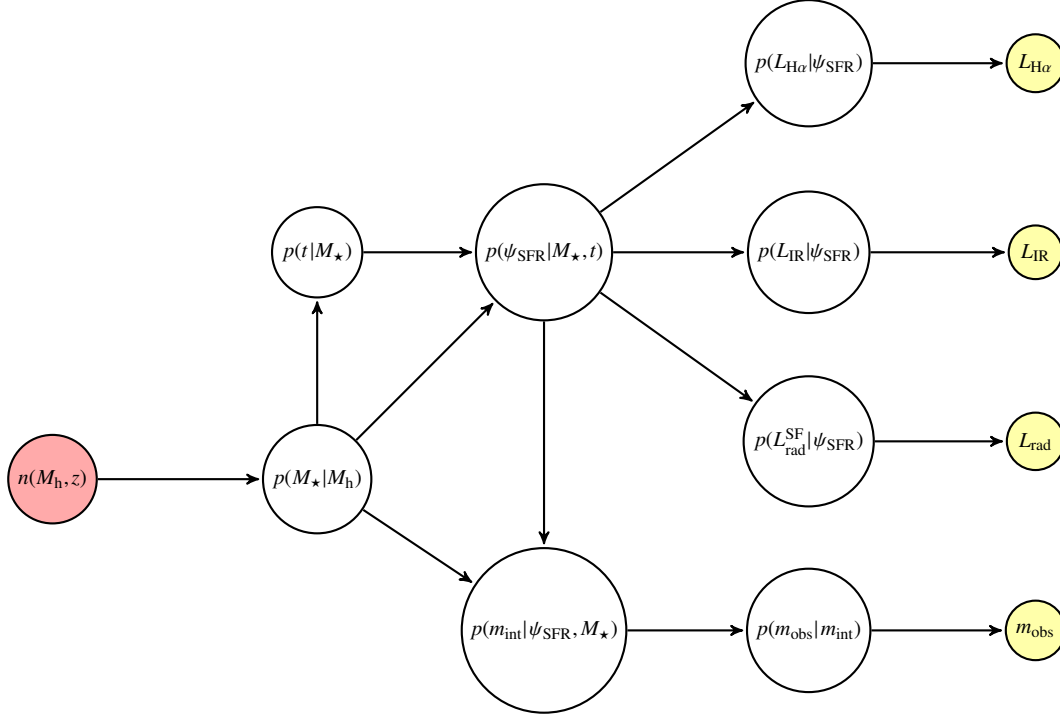
Halo model methods come in a number of different flavours. Halo occupation distribution (HOD; [Peacock & Smith 2000](#); [Berlind & Weinberg 2002](#); [Berlind et al. 2003](#)) modelling predicts the mean number of galaxies of a given type that reside in a halo of given mass,  $\langle N \rangle(M_h)$ . HODs are typically calibrated empirically, with the resulting fits being extrapolated to model the distribution of the same type of galaxies in future surveys. While HODs can be simultaneously defined for multiple galaxy populations (e.g. [Krumpe et al. 2010](#); [Miyaji et al. 2011](#); [Krause et al. 2013](#)), the correlation between each pair of populations must be modelled (and calibrated) separately, e.g. using the observed cross-correlation function. Ambiguities may also arise when trying to ascertain which galaxies appear in more than one survey – HODs calibrated at different wavelengths can predict different numbers of galaxies to exist in a given halo, so some galaxies will be ‘missing’ in one band, but not in another.

An alternative approach is to construct conditional luminosity functions (CLFs; [Yang et al. 2003](#); [Vale & Ostriker 2004](#); [Cooray 2006](#)), which describe the expected luminosity distribution of galaxies hosted in halos of a given mass. These methods are again calibrated empirically against a given population of galaxies, but can be extended to other wavelengths by assuming an appropriate spectral energy distribution (SED). Correlations between different types of tracer must again be added by hand though.

A unified framework for implementing halo model methods such as these is presented in [Hearin et al. \(2016\)](#).

Several recent works have explicitly sought to model galaxy populations across multiple wavelengths. Some are empirical, in the sense that they create realisations of mock galaxy samples that are consistent with a set of observed luminosity functions (LFs) by construction ([Jouvel et al. 2009](#); [Schreiber et al. 2016](#)). The data for these are generally obtained from deep but narrow surveys (e.g. COSMOS), so the redshift dependence and faint end of the luminosity function are well measured, but large-scale clustering information tends to be minimal. A related approach is to use machine learning techniques to create a predictive model of a multi-wavelength galaxy sample, again by training it off an existing galaxy catalogue ([Xu et al. 2013](#)).

Alternatively, one can take scaling relations between galaxy properties (e.g. star-formation rate and stellar mass) and combine them to reproduce observables. This is the basic approach taken



**Figure 1.** Graph of the probabilistic model linking input parameters (halo mass  $M_h$  and redshift  $z$ , drawn from a halo mass function) to observable luminosities/magnitudes in various bands. Only the radio and optical bands are considered in this paper; the IR and H $\alpha$  bands are shown for illustration.

in this paper, and was also used by Li et al. (2016) and Schreiber et al. (2016). This has the advantage of allowing several different observables to be constructed from a common set of components, and is justified by the empirical discovery of a number of appropriate scaling relations. A possible problem is inconsistency between the various components, which may be calibrated off different datasets with different assumptions. We will avoid this problem here by performing global model fits, rather than calibrating components individually.

### 3 MODEL DEFINITION

In this section, we specify each component of the model in detail, according to the basic structure illustrated in Fig. 1.

#### 3.1 Halo mass function

The initial input to the model is a distribution of dark matter (DM) halos in mass, position, and redshift. This is represented by the halo mass function,

$$n(M_h, z) \equiv \frac{dN(z)}{d \log M_h dV dz}, \quad (1)$$

where  $M_h$  is the halo mass and  $V$  is the comoving volume. This distribution ultimately sets the overall abundance and clustering properties of galaxies, and the dependence on the underlying cosmological model. As a simplifying assumption, we will assume that DM halos are characterised solely by their mass and redshift, ignoring other intrinsic properties such as angular momentum and concentration. Furthermore, we will make no explicit distinction between parent and sub-halos, although in reality some galaxy properties do depend on this. Finally, we will assume that each halo (or sub-halo) hosts precisely one galaxy – there are no empty halos, and no multiply-occupied ones (in contrast to HOD, where satellites are explicitly modelled).

In most of what follows we will use an analytic form for the halo mass function, but it is important to note that the model can be applied to *any* representation of a set of DM halos. For example, the chain of statistical distributions shown in Fig. 1 can just as well be applied to a DM halo catalogue taken from an N-body simulation, and used to populate mock catalogues of galaxies using a Monte Carlo method.

For the analytic calculations in the remainder of the paper, we will adopt the Tinker et al. (2008) mass function,

$$n(M_h, z) = -f(\sigma) \frac{\rho_m}{M_h} \frac{d \log \sigma(M_h)}{d \log M_h} \quad (2)$$

$$f(\sigma) = A \left[ 1 + \left( \frac{\sigma(M_h)}{b} \right)^{-a} \right] \exp(-c/\sigma^2), \quad (3)$$

where  $\rho_m$  is the background matter density and  $\sigma(M)$  is the rms density fluctuation within a sphere of radius  $R = (3M/4\pi\rho_m)^{1/3}$ , from linear theory. The dependence on cosmological parameters is implicit in  $\sigma(M)$ , an integral over the matter power spectrum,  $P(k)$ , which we take from CAMB (Lewis et al. 2000). The best-fit parameters, calibrated against simulations, are  $A = 0.186$ ,  $a = 1.47$ ,  $b = 2.57$ , and  $c = 1.19$ , all at  $z = 0$ .

For calculations of the clustering, one must also specify a halo bias function,  $b(M_h)$ . The (Eulerian) halo bias is

$$b(M_h) = 1 + (a\nu_1 - 1)/\delta_1 + \frac{2p/\delta_1}{1 + (a\nu_1)^p}, \quad (4)$$

where  $\nu_1 = [\delta_1/\sigma(M_h)]^2$  and  $\delta_1 \approx \delta_c$ .

#### 3.2 Stellar mass – halo mass relation

As a first step, it is necessary to link dark matter halos to their baryonic content. We model this through the relationship between halo mass and stellar mass, as described by the conditional mass function (CMF). This is related to the conditional probability to find a stellar mass of  $M_*$  in a halo of mass  $M_h$ :  $p(M_* | M_h)$ .

Halo mass and stellar mass are expected to be strongly related as a simple consequence of hierarchical structure formation; as more massive halos collapse, they trap a correspondingly larger mass of baryonic matter, some fraction of which forms stars as the gas settles and cools in the centre of the halo potential (White & Rees 1978). A relatively tight relationship between the two is indeed seen observationally (Yang et al. 2007; Moster et al. 2010; Behroozi et al. 2010), and is well fit by a broken powerlaw,

$$\overline{M}_\star(M_h) = \frac{2M_h A_\star}{\left(\frac{M_h}{M_1^\star}\right)^{-\beta_\star} + \left(\frac{M_h}{M_1^\star}\right)^{\gamma_\star}}, \quad (5)$$

where  $A_\star$  is the overall normalisation,  $M_1^\star$  is a mass scale, and  $\beta_\star$  and  $\gamma_\star$  are the low- and high-mass end slopes respectively. The conditional pdf is modelled as log-normal,

$$p(M_\star|M_h) = \frac{1}{\sqrt{2\pi}M_\star\sigma_{M_\star}} \exp\left(-\frac{\log^2(M_\star/\overline{M}_\star)}{2\sigma_{M_\star}^2}\right), \quad (6)$$

where  $\sigma_{M_\star}$  is the scatter, modelled by Moster et al. (2010) as

$$\frac{\sigma_{M_\star}}{\log 10} = \sigma_\infty^\star + \sigma_1^\star \left(1 - \frac{2}{\pi} \arctan\left[\xi_\star \log_{10}(M_h/M_2^\star)\right]\right). \quad (7)$$

While Moster et al. (2010) provide separate fits for both central and satellite galaxies, we use only the central relation as, for simplicity, our model ignores this distinction.

### 3.3 Star-formation rate – stellar mass relation

The star-formation rate (SFR),  $\psi_{\text{SFR}}$ , is a key quantity that predicts many of a galaxy’s other observable properties. It is thought to depend on a number of factors, including the availability of cold molecular gas (the raw material of star formation); the recent merger history of the galaxy (mergers disturb cold gas, inducing its collapse to form stars); and quenching processes such as feedback from accretion onto the central supermassive black hole or supernovae. See Somerville et al. (2015) and references therein for a more thorough overview.

As we are only interested in the aggregate properties of a large galaxy sample, we adopt stellar mass as a sufficiently robust predictor of the SFR. A connection between the two is well established – a *star-forming main sequence* (SFMS) is found in the  $\psi_{\text{SFR}} - M_\star$  plane, along which the bulk of actively star-forming galaxies lie (Daddi et al. 2007; Elbaz et al. 2007; Noeske et al. 2007; Salim et al. 2007; Peng et al. 2010; Karim et al. 2011; Wang et al. 2013). A second population of *quiescent* or *passive* galaxies lies below the SFMS, which typically contain older, redder stars, and less (but not necessarily negligible) star formation. This population may be arranged in a similar, but more scattered, sequence to the SFMS, or else in a much looser, relatively uniform distribution in SFR below the main sequence; see Lagos et al. (2011) for a comparison of semi-analytic model predictions. Other populations may be identified in the  $\psi_{\text{SFR}} - M_\star$  plane, such as starburst galaxies (which lie above the SFMS), but we will work only with the simple active vs. passive categorisation here.

Motivated by the above, we will model the distributions of passive and SFMS galaxies separately. The first step is to assign a probability that a galaxy will belong to one population or the other. Several studies have attempted to estimate the fraction of passive galaxies from recent surveys (e.g. Fontana et al. 2009; Peng et al. 2010; Moustakas et al. 2013); we adopt a modified version of the relation from Behroozi et al. (2013), which they attributed to Brammer et al. (2011):

$$f_{\text{passive}}(M_\star) = c + \frac{1-c}{1 + \left(\frac{M_\star}{10^{a_f} M_\odot}\right)^{\beta_f}}, \quad (8)$$

where we reparametrise  $c \equiv \frac{1}{2}(1 + \tanh \zeta_f)$  to ensure that  $f_{\text{passive}} \in [0, 1]$ . Factors other than stellar mass can also be important for determining whether a galaxy is passive or not (e.g. environment; see Peng et al. 2010). We assume that these factors have been marginalised over here, leaving  $M_\star$  as the only relevant variable. The resulting pdf,  $p(t|M_\star)$ , where  $t$  denotes the galaxy type, is a uniform distribution partitioned at  $f_{\text{passive}}$ .

For the SFMS, we adopt a mean relation with a similar form to the one found in Wang et al. (2013),

$$\overline{\psi}_{\text{SFR}}(M_\star) = 10^{\alpha_{\text{SFMS}}} \left(\frac{M_\star}{10^{10} M_\odot}\right)^{\beta_{\text{SFMS}}} M_\odot \text{yr}^{-1}. \quad (9)$$

The scatter in the SFMS is incorporated by modelling it as a log-normal distribution,

$$p_{\text{SFMS}}(\psi_{\text{SFR}}|M_\star) = \frac{1}{\sqrt{2\pi}\psi_{\text{SFR}}\sigma_{\text{SFMS}}} \exp\left(-\frac{\log^2(\psi_{\text{SFR}}/\overline{\psi}_{\text{SFR}})}{2\sigma_{\text{SFMS}}^2}\right). \quad (10)$$

There is more freedom in deciding how to model the passive population. The results of Wang et al. (2013) suggest a relatively uniform distribution of galaxies below the SFMS, while Moustakas et al. (2013) show a similar log-normal distribution to the SFMS, but shifted to higher  $M_\star$  and lower  $\psi_{\text{SFR}}$ . Semi-analytic models show an even wider range of behaviours, as discussed in Lagos et al. (2011). For simplicity, we will use a log-normal distribution of the same form as Eq. (10), but with the mean from Eq. (9) shifted by some factor, and a different scatter,

$$\begin{aligned} \overline{\psi}_{\text{SFR}} &\rightarrow a_{\text{pass}} \overline{\psi}_{\text{SFR}} \\ \sigma_{\text{SFMS}} &\rightarrow \sigma_{\text{pass}}. \end{aligned} \quad (11)$$

This more closely follows the observations of Moustakas et al. (2013), with the apparent shift in  $M_\star$  caused by the transition  $f_{\text{passive}} \rightarrow 1$  as  $M_\star$  increases.

### 3.4 Star-formation – luminosity relations

Star formation activity releases energy through several processes that cause (typically bright) emission across many bands (Kennicutt 1998). Young, high-mass stars emit UV radiation, which excites gas in the surrounding inter-stellar medium (ISM). As well as the UV emission itself, this generates line radiation, for example through the de-excitation of hydrogen through  $\text{H}\alpha$  and  $\text{H}\beta$  emission. Young stars are short-lived, exploding as supernovae on timescales of order 10 Myr. Supernovae inject high-energy electrons into the ISM, which interact with the galactic magnetic field to emit synchrotron radiation at radio frequencies (Condon 1992). The process of star formation itself takes place in dense clouds of gas and dust, and the latter, warmed by the collapse process, emits in the infrared.

Star formation-related emission dominates the luminosity of many galaxies. This has led to the development of many “SFR indicator” relations (Kennicutt 1998; Bell 2003; Moustakas et al. 2006) that can be written in the form  $L \approx A \psi_{\text{SFR}}^\beta$ , with  $\beta \approx 1$ . In this paper, we will consider only one SFR-related type of emission as an example. This is continuum radio emission, from synchrotron radiation associated with supernova remnants. Continuum surveys with the SKA and its precursors are expected to detect tens of millions of galaxies from  $z \sim 0-5$ , which can be used for weak lensing and 2D clustering studies (Jarvis et al. 2015). Assuming a powerlaw spectrum for the emission,  $S \propto (\nu/\nu_{\text{ref}})^\alpha$ , with  $\alpha \approx -0.7$ , one can use the SFR indicator relation at 1.4 GHz from Bell (2003) to obtain

$$L_{1.4\text{GHz}} = 1.812 \times 10^{28} \left(\frac{\psi_{\text{SFR}}}{M_\odot \text{yr}^{-1}}\right) \text{erg s}^{-1} \text{Hz}^{-1}. \quad (12)$$



Stellar mass–halo mass relation, $p(M_\star M_h)$		
$A_\star$	Overall normalisation	0.018
$\beta_\star$	Slope at low-mass end	Eq. 5 1.321
$\gamma_\star$	Slope at high-mass end	0.596
$\log_{10} M_1^\star$	Mass scale of transition in mean relation ( $M_\odot$ )	12.498
$\sigma_1^\star$	Overall dispersion amplitude at low masses	0.557
$\sigma_\infty^\star$	Dispersion at high-mass end	Eq. 7 0.031
$\xi_\star$	Width of transition in dispersion <sup>†</sup>	4.250
$\log_{10} M_2^\star$	Mass scale of transition in dispersion ( $M_\odot$ ) <sup>†</sup>	11.800
Galaxy type (passive fraction), $f_{\text{pass}}(M_\star)$		
$\alpha_f$	Mass scale of transition (exponent)	10.804
$\beta_f$	Determines width of transition	Eq. 8 -2.436
$\zeta_f$	Determines passive fraction at low mass	-1.621
Stellar mass–SFR relation (SFMS galaxies), $p(\psi_{\text{SFR}} M_\star)$		
$\alpha_{\text{SFMS}}$	Overall normalisation of mean relation (exponent)	Eq. 9 -0.077
$\beta_{\text{SFMS}}$	Power-law index of scaling with stellar mass	1.037
$\sigma_{\text{SFMS}}$	Dispersion of SFMS	Eq. 10 0.391
Stellar mass–SFR relation (passive galaxies), $p(\psi_{\text{SFR}} M_\star)$		
$a_{\text{pass}}$	Determines normalisation of mean relation	Eq. 11 0.0011
$\sigma_{\text{pass}}$	Dispersion of passive sequence	0.029

**Table 1.** Summary of the best-fit parameter values for the main components of the model. The best-fit parameters for the optical magnitude relations are given in Table 2. Parameters marked with (†) were fixed during the MCMC sampling procedure.

We will omit scatter in this relation, treating the scaling as being deterministic for simplicity. Unlike some other SFR indicators (e.g. H $\alpha$ ; Kennicutt 1998; Pozzetti et al. 2016), the radio emission does not suffer from extinction, but can be contaminated by (e.g.) free-free and emission from AGN jets. We assume here that free-free is subdominant, and that AGN-dominated and passive galaxies can be selected out, leaving only normal star-forming galaxies.

### 3.5 Optical magnitude relations

The connection between bulk galaxy properties and emission in optical bands is more convoluted. Optical emission is primarily sourced by the aggregate of all of the stars in the galaxy, which form over an extended period of time, and evolve at different rates depending on their initial mass. The observed optical luminosity is therefore a probe of the star-formation history of the galaxy, not just its present state.

Semi-analytic models typically reconstruct the star-formation history explicitly, evolving an assumed initial mass function forward in time with a stellar population synthesis (SPS) model, and taking into account later bursts of star formation caused by mergers and other events. In the absence of the actual merger history from a simulation, Monte Carlo realisations of plausible merger histories can also be substituted (e.g. Parkinson et al. 2008).

A similar approach would be possible here, at the cost of adding significant computational complexity and relying on a ‘black box’ SPS code. Instead, we attempt a simpler approxi-

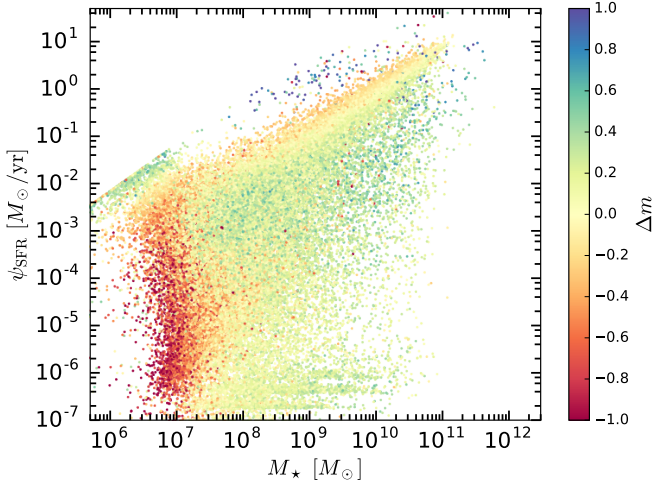
mate treatment to preserve the analytic, parametric nature of the model. We begin by hypothesising that the stellar population can be characterised by two main components: newly-formed stars, dominated by the high-mass end of the IMF; and an older population that formed long before, dominated by lower-mass stars on the main sequence. The former will be bluer, the latter redder.

Next, we propose an ansatz for the mean absolute magnitude of a galaxy in a given optical band (labelled by  $\nu$ ):

$$\begin{aligned} \bar{m}_\nu(\psi_{\text{SFR}}, M_\star) = & -c_0^{(\nu)} + A_\star^{\text{opt}} \left( c_\star^{\text{opt}} + \left( \frac{M_\star}{10^9 M_\odot} \right)^{\beta_\star^{\text{opt}}} \right) \\ & + A_\times^{(\nu)} \left( \frac{M_\star}{10^9 M_\odot} \right)^{\beta_\times^{\text{opt}}} (\psi_{\text{SFR}})^{\gamma_\times^{\text{opt}}}. \end{aligned} \quad (13)$$

A rough physical interpretation of each term is as follows. The first is an offset that determines a characteristic galaxy magnitude in a given band. The second characterises the total emission from the older stellar population, using stellar mass as a proxy. The final term represents the contribution from the younger population, with SFR as a proxy. Since there is a relationship between SFR and stellar mass, this term must also include an  $M_\star$ -dependent factor to model the ‘mixing’ between the two. The amplitude of the final term also depends on the band; star formation contributes less to the total luminosity in redder optical bands.

To calibrate the relation, we use the public semi-analytic catalogues of Guo et al. (2011), which provide *ugriz* absolute magnitudes with and without a dust extinction correction. The residuals



**Figure 2.** Difference between  $u$ -band magnitudes at  $z = 0$  (without dust attenuation) in the Guo et al. (2011) simulation, and those predicted by our best-fit relation, Eq. 13, as a function of SFR and stellar mass. Resolution effects in the simulation are apparent at high and low mass, and low SFR.

of the best fit to the unattenuated  $u$ -band magnitudes at  $z = 0$  are shown in Fig. 2, as a function of  $M_*$  and  $\psi_{\text{SFR}}$ . The distribution of residuals is reasonably simple, with a mean close to zero and only a low level of structure present. The standard deviation of the residuals across the whole plane is  $\sigma(\Delta m_v) \sim 0.3$  mag in all 5 bands, which is quite small considering the simplicity of this approach; it has only five global parameters and two free parameters per band. Ignoring the dependence of the residuals on  $M_*$  and  $\psi_{\text{SFR}}$ , we find that they are well modelled by a shifted log-normal distribution,

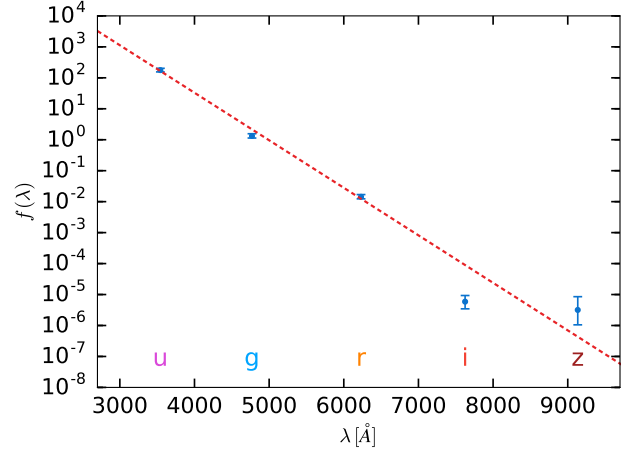
$$p(m_v | \psi_{\text{SFR}}, M_*, z) = \frac{1}{\sqrt{2\pi}x\sigma_m^{(v)}} \exp\left(-\frac{1}{2}\left(\frac{\log x - \mu^{(v)}}{\sigma_m^{(v)}}\right)^2\right), \quad (14)$$

where  $x \equiv \bar{m}_v - m_v + b_v$ ,  $x > 0$ . Here,  $\log b_v \equiv \mu^{(v)} + (\sigma_m^{(v)})^2/2$ , where  $b_v$  is the mean of the shifted log-normal distribution and  $\mu^{(v)}$  and  $\sigma_m^{(v)}$  are free parameters in each band. These parameters are calibrated by performing a least-squares fit of a log-normal pdf to the histogram of residuals, independent of  $\psi_{\text{SFR}}$  and  $M_*$ .

As shown in Fig. 2, the residuals do have some dependence on  $\psi_{\text{SFR}}$  and  $M_*$ . The residuals can also be correlated between bands. While we have neglected structure like this in our model as a simplifying assumption, such correlations are partially responsible for the observational patterns seen in ‘colour-magnitude’ diagrams, which are often used to select different populations of galaxies. In its current state, the model will therefore be unable to reproduce features such as the red sequence or blue cloud (e.g. Strateva et al. 2001) for example.

### 3.6 Optical dust extinction

Some fraction of the optical emission from galaxies is absorbed by the dust that they contain. The amount of absorption depends on a number of factors, including wavelength (bluer colours are more strongly absorbed); morphology and inclination angle (light from the bulge and disk components is affected differently); and the dust content and its distribution inside the galaxy. Intrinsic dust absorption is difficult to simulate or characterise with observations, which is problematic – accounting for dust is a necessary step in connecting optical emission to the physical properties of galaxies, as well as being important for modelling infrared luminosities.



**Figure 3.** Amplitude of the optical dust attenuation as a function of wavelength,  $f(\lambda)$ . The blue datapoints show the median and 68% credible intervals for  $f(\lambda_j)$  in each band  $j$ , after conditioning on the best-fit values of all other parameters. The red dashed line shows the relation from Eq. 16 for the best-fit values of  $\tau_0$ ,  $\kappa$ , and  $\lambda_0$ .

There have been numerous attempts to construct attenuation models from observations and simple considerations such as galaxy morphological properties (e.g. Tuffs et al. 2004; Driver et al. 2007). Several dust models that have been used in semi-analytic simulations are compared in Fontanot et al. (2009). Many such models use an empirically-determined spectral dependence from Calzetti et al. (2000), which has been found to offer a good description of the corrections required to get SFR determinations at different wavelengths to agree (e.g. Pannella et al. 2009).

In keeping with our analytic approach, we will adopt a simpler parametric model. An ansatz for the optical depth due to internal dust extinction is

$$\tau(\lambda) = f(\lambda) \left( \frac{M_*}{10^{10} M_\odot} \right)^{\beta_\tau} (1 + a_{\text{disk}} \sin \theta), \quad (15)$$

where  $f(\lambda)$  is the attenuation amplitude as a function of (rest-frame) wavelength,  $a_{\text{disk}}$  is the effective attenuation due to the disk, and  $\theta$  is the inclination angle of the disk with respect to the line of sight. The frequency dependence (see Fig. 3) can be modelled approximately as

$$f(\lambda) = \tau_0 \exp[-\kappa(\lambda - \lambda_0)], \quad (16)$$

where  $\tau_0$  is an overall normalisation parameter,  $\kappa$  determines the scaling with wavelength, and  $\lambda_0$  is a reference wavelength. The change in magnitude due to the dust attenuation is

$$\Delta m \equiv m_{\text{obs}} - m_{\text{int}} = 5 \log_{10} \left( \frac{S_{\text{int}} e^{-\tau}}{S_{\text{int}}} \right) = 1.086 \tau, \quad (17)$$

where  $S_{\text{int}}$  is the intrinsic flux before attenuation.

The two factors in parentheses in Eq. (15) are motivated as follows. First, the amount of dust in the galaxy is assumed to scale with the stellar mass to some power. Dust builds up in the interstellar medium as successive populations of stars evolve and die. While star formation is associated with thermal dust emission, this only represents one component of the total dust content of a galaxy; stellar mass should be more indicative of the integrated star formation history that has actually generated the dust. There is observational support for a relatively steep power-law scaling of the dust attenuation with stellar mass (e.g. Pannella et al. 2009; Buat et al. 2012; Heinis et al. 2014), but we will not enter into a discussion here.

Band:	u	g	r	i	z
$A_\star$			3876.6		
$\beta_\star$			-0.000243		
$c_\star$			-1.011		
$A_\times^{(v)}$	-3.612	-2.700	-2.016	-1.726	-1.564
$\beta_\times$			-0.263		
$\gamma_\times$			0.290		
$c_0^{(v)}$	-27.318	-25.945	-25.284	-24.982	-24.810
$\mu^{(v)}$	-0.066	-0.054	-0.019	-0.009	-0.009
$\sigma_m^{(v)}$	0.281	0.252	0.247	0.253	0.272
$\tau_0$			1.041		
$\lambda_0$			4977.2		
$\kappa$			0.004		
$a_{\text{disk}}$			1.429		
$\beta_\tau$			5.478		

**Table 2.** Best-fit values for the parameters of the optical magnitude model, including dust attenuation. The non-attenuation parameters were fitted to the  $z = 0$  simulated galaxy catalogue of Guo et al. (2011). The attenuation parameters  $a_{\text{disk}}$  and  $\beta_\tau$  were fitted to the  $g$  and  $z$ -band GAMA LF data simultaneously with the non-optical model parameters (see below). The attenuation parameters  $\tau_0$ ,  $\lambda_0$ , and  $\kappa$  were fitted using a maximum likelihood method on the full 5-band GAMA LF data (conditioned on all other model parameters).

Secondly, extinction will be greater when the disk is seen edge-on, i.e. when  $\theta = \pi/2$ . The relative optical depth of the disk (compared to the bulge) is represented by a scaling factor,  $a_{\text{disk}}$ . For random galaxy orientations,  $\sin\theta$  follows a uniform distribution. If we assume that inclination angle is the only random variable in the optical depth expression, it follows that

$$p(m_v^{\text{obs}}|m_v^{\text{int}}) = \text{Uniform}[m_v^{\text{int}} + \Delta m(\theta = 0), m_v^{\text{int}} + \Delta m(\theta = \pi/2)].$$

In the absence of explicit modelling of other stochastic contributions to the optical depth, it is likely that the inclination angle term will absorb other sources of scatter in the observations. As such, the  $a_{\text{disk}}$  parameter should not be interpreted directly as a physical quantity.

Since one only ever observes the dust-attenuated flux in a given band, it is useful to immediately marginalise out the intrinsic magnitude,

$$p(m_v^{\text{obs}}|M_\star, \psi_{\text{SFR}}) = \int p(m_v^{\text{obs}}|m_v^{\text{int}}) p(m_v^{\text{int}}|M_\star, \psi_{\text{SFR}}) dm_v^{\text{int}}. \quad (18)$$

The integral can be evaluated analytically for our model pdfs;

$$p(m_v^{\text{obs}}|M_\star, \psi_{\text{SFR}}) = \frac{1}{2} \frac{\text{erf}(y(m_v^{\text{obs}}, \pi/2)) - \text{erf}(y(m_v^{\text{obs}}, 0))}{\Delta m_v(\pi/2) - \Delta m_v(0)}, \quad (19)$$

where the denominator comes from the normalisation of the uniform distribution, and the argument of the error functions is

$$y(m_v^{\text{obs}}, \theta) = \frac{\log(x_r(\theta)) - \mu^{(v)}}{\sqrt{2} \sigma_m^{(v)}}, \quad (20)$$

$$x_r(\theta) = \max\{\bar{m}_v^{\text{int}} + \Delta m_v(\theta) + b_v - m_v^{\text{obs}}, 0\}. \quad (21)$$

The dependence on  $M_\star$  and  $\psi_{\text{SFR}}$  enters through  $\bar{m}_v^{\text{int}}$  and  $\Delta m_v$ .

It should be noted that this is a rather basic statistical model for intrinsic dust extinction. We have assumed that the optical depth in all galaxies follows the same mean relation, and (effectively) differs only in inclination angle. The other parameters of the model will surely also vary from galaxy to galaxy though, especially depending on its morphology – passive elliptical galaxies do not have separate bulge and disk components, for example. A simple extension to the model would be to choose  $a_{\text{disk}}$  separately for star-forming and passive galaxies, but we do not pursue this here. As such, this aspect of the model should be considered preliminary – it is able to fit the observations, as we shall see in the next section, but any physical interpretation of the fits should proceed with caution.

#### 4 MODEL CALIBRATION AND PARAMETER DEPENDENCE

A set of parameter values is needed as an input to the model. While best-fit parameters for most of the conditional distributions in Fig. 1 are available separately in the literature, suitable parameter values can also be obtained by fitting the entire model to a set of diverse observational (or simulated) data. This has the advantage of ensuring consistency between the various components of the model. We will pursue the latter approach here, using a Monte Carlo sampling method to derive the joint posterior distribution of the model parameters given several datasets from galaxy surveys at different wavelengths.

In the first part of this section, we describe the data used to perform the fits: recent constraints on the radio luminosity function of star-forming galaxies, and the luminosity function of optically-detected galaxies, all at  $z \approx 0$ . We also describe the sampling method, and cuts that were used on the data. We then report the parameter constraints that were obtained, and analyse them for consistency using simple goodness-of-fit statistics. We also compare with the best-fit values that were found for several of the model components independently by previous studies.

##### 4.1 Input data

We use the following data to constrain the model:

- The luminosity function of star-forming radio galaxies cross-identified in NVSS and 6dFGS (Mauch & Sadler 2007). This mostly constrains the SFMS and halo mass-stellar mass pdfs, and does not require a dust attenuation correction.
- The *ugriz* optical luminosity functions from GAMA ( Loveday et al. 2012). These are corrected for extinction in the Milky Way, but not for the internal extinction in the target galaxies. We fit both the  $g$  and  $z$ -band data to the full model; the  $z$  band is least affected by extinction, while the  $g$  band is strongly affected, allowing some of the parameters of the dust attenuation model to be constrained. Data from the other bands are not used in the general fits, but they are used to constrain the dust extinction amplitude parameter (conditioned on the best-fit parameters of the general fit; see below).
- The joint distribution of stellar mass, star-formation rate, and intrinsic (zero extinction) optical magnitudes from the Guo et al. (2011) semi-analytic simulation. This is used to calibrate the optical magnitude conditional distributions only, as described in Sect. 3.5. The resulting best-fit parameters are then fixed throughout the rest of the analysis.

These datasets were chosen because they are sufficient to constrain all of the key parameters of the model at  $z \approx 0$ . Other datasets

or simulations could also have been included, but this is beyond the scope of the present work. Note that most of these datasets assume different background cosmologies; we applied corrections to convert them to our fiducial cosmology as appropriate.

We construct independent, approximate likelihoods for each dataset, based on the binned data reported in the literature. The binned GAMA luminosity functions are reported with symmetric error bars. We assume that these data are Gaussian distributed, and that the errors are independent, yielding a likelihood of the form

$$\log \mathcal{L}(\theta) = -\frac{1}{2} \sum_i \left( \frac{\Phi_i - \Phi(\theta)}{\sigma_i} \right)^2, \quad (22)$$

where  $i$  labels the magnitude bins of the luminosity function,  $\Phi = dn/dm$ , and  $\theta$  denotes the set of parameters for which the model luminosity function is evaluated. The optical luminosity function is calculated from the model as

$$\begin{aligned} \Phi(m_v^{\text{obs}}) \equiv \frac{dn}{dm_v^{\text{obs}}} &= \int d\psi_{\text{SFR}} \int dM_{\star} \int d\log M_h \\ &\times n(M_h) p(M_{\star}|M_h) p(\psi_{\text{SFR}}|M_{\star}) \\ &\times p(m_v^{\text{obs}}|\psi_{\text{SFR}}, M_h). \end{aligned} \quad (23)$$

The NVSS/6dFGS radio luminosities are reported with asymmetric errorbars. We were unable to find published posterior distributions for the data, and so we use an approximate likelihood for data with asymmetric errors from an unknown distribution, taken from Barlow (2003):

$$\log \mathcal{L}(\theta) \approx -\frac{1}{2} \sum_i x_i^2(\theta) (1 - 2A_i x_i(\theta) + 5A_i^2 x_i^2(\theta)); \quad (24)$$

$$x_i(\theta) = \frac{y_i - \hat{y}_i(\theta)}{\sigma_i}; \quad \sigma = \frac{\sigma_+ + \sigma_-}{2}; \quad A = \frac{\sigma_+ - \sigma_-}{\sigma_+ + \sigma_-}, \quad (25)$$

where  $y_i$  is the observed value,  $\sigma_{\pm}$  are the upper and lower asymmetric errorbars on  $y_i$ , and  $\hat{y}_i(\theta)$  is the model value for a given set of parameters. This limits to a Gaussian likelihood for symmetric error bars. Since the studies in question report asymmetric errors on the logarithm of the radio luminosity function, we will work directly with  $y_i = \log \Phi_i$  as the random variate. Again, this is an approximate form for the likelihood *in the absence of more information* about the posterior distributions. If we had access to the actual posteriors, this treatment would be unnecessary. The radio luminosity function is calculated in our model as

$$\begin{aligned} \Phi(L_{\text{rad}}) \equiv \frac{dn}{d\log L} &= \int d\psi_{\text{SFR}} \int dM_{\star} \int d\log M_h \\ &\times n(M_h) p(M_{\star}|M_h) p_{\text{SFMS}}(\psi_{\text{SFR}}|M_{\star}) \\ &\times L_{\text{rad}} p(L_{\text{rad}}|\psi_{\text{SFR}}), \end{aligned} \quad (26)$$

where only the SFMS is used; passive galaxies are assumed to have been selected out.

The final log-likelihood is simply the sum of the log-likelihoods for each dataset, i.e. two optical bands and the radio.

## 4.2 Sampling method

We use the emcee affine-invariant Markov Chain Monte Carlo (MCMC) sampler (Foreman-Mackey et al. 2013) to sample from the posterior distribution. We run this for 2000 samples for each of 128 workers, leaving 5500 samples after discarding burn-in and thinning the chains.

Seventeen (17) parameters were sampled by the MCMC (see Table 1 for a summary). We did not sample the parameters of the optical luminosity relation defined in Sect. 3.5, as this would allow too much freedom in the model. The posterior distributions of other parameters could shift if these were allowed to vary, although

this would likely also lead to degeneracies. We did sample three of the dust attenuation parameters however (see Table 2). The starting position of the chains was chosen based on a rough visual fit to the input data, which aimed to find parameter values broadly similar to the best-fit values from the literature. This procedure was necessary to help avoid local maxima of the likelihood.

Simple prior bounds were chosen to keep the walkers in a physically reasonable region of parameter space. Certain features can be reproduced by allowing the SFMS to be too broad or too narrow for example, and the passive and star-forming sequences can switch position in the SFR- $M_{\star}$  plane if the amplitude parameters are not kept in check. Negative dispersion parameters should also be disallowed, as otherwise the pdfs become ill-defined. To avoid problems like these, we chose the following priors:

$$0.05 \leq \sigma_{\text{SFMS}} \leq 1 \quad (27)$$

$$0.01 \leq \sigma_{\text{pass}} \leq 2 \quad (28)$$

$$0.02 \leq \sigma_1^{\star}, \sigma_{\infty}^{\star} \quad (29)$$

$$0.001 \leq a_{\text{pass}} \leq 0.9 \quad (30)$$

$$11.6 \leq \log_{10} M_1^{\star} \leq 12.5. \quad (31)$$

The last prior in this set was chosen to restrict the transition scale of the stellar mass-halo mass relation to be the same order of magnitude as the one found by Moster et al. (2010). Allowing this to be completely free often resulted in the walkers finding local minima with substantially different best-fit parameters from others in the literature.

We identified the burn-in period by plotting the mean of the log-likelihood over all 128 workers as a function of sample number. The mean log-likelihood stops improving as the chains enter the maximum likelihood region, which was seen as a flattening of the plot after a certain number of steps. To thin the chains, we calculated the integrated autocorrelation time (Sokal 1997) for the chain from each worker as a function of thinning factor,

$$\tau_{\text{int}}(f) = \frac{1}{2} \sum_{\tau} \rho(\tau, f), \quad (32)$$

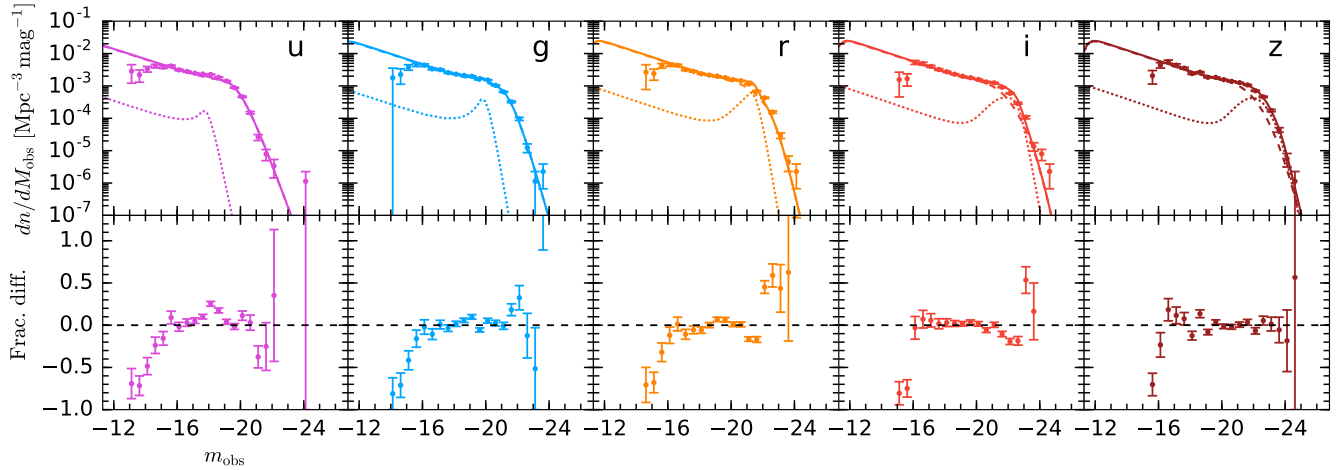
where  $\rho(\tau, f)$  is the normalised autocorrelation function at lag  $\tau$  and  $f$  is the thinning factor. We chose the smallest thinning factor that reduced the autocorrelation time below unity, and then combined the thinned chains from all workers into a single chain. As jumps between workers are allowed by ensemble samplers, the chains from each worker are correlated with one another, so this procedure does not fully decorrelate the samples. We verified that the sample means and standard deviations for a few parameters did not shift significantly when different thinning factors or burn-in periods were used though, or when different numbers of workers were kept in the final chains. The posterior distribution for all 17 parameters is shown in Fig. A, and the processed chain is available to download from <http://www.philbull.com/ghost>.

Once a best-fit model has been obtained from the MCMC, we then fit the frequency-dependent dust attenuation parameters  $\tau_0$ ,  $\kappa$ , and  $\lambda_0$  (see Eq. 16) to all five optical bands simultaneously, using a maximum likelihood method. The likelihood for the optical luminosity functions is evaluated by varying these parameters while keeping all other parameters fixed to their best-fit values. The results of this procedure are shown in Table 2 and Fig. 3.

## 4.3 Analysis of the model fits

The MCMC procedure described above results in a set of samples from the posterior distribution of the model and chosen data. The parameter values of the best-fit (i.e. maximum likelihood) model from the chains are summarised in Table 1. In this section, we provide a brief analysis of how well the global best-fit model fits the





**Figure 4.** *Upper panels:* Optical luminosity functions in the *ugriz* bands at  $z \approx 0$ , fitted to data from GAMA (points with errorbars) and the NVSS radio luminosity function. Only the *g* and *z* bands were used in the fitting procedure. The solid coloured lines show the total luminosity functions, including the dust attenuation correction; dashed lines show the LF from the star-forming main sequence only; and dotted lines show the passive galaxy LF. *Lower panels:* Fractional difference between the best-fit optical luminosity function from our model and the GAMA datapoints.

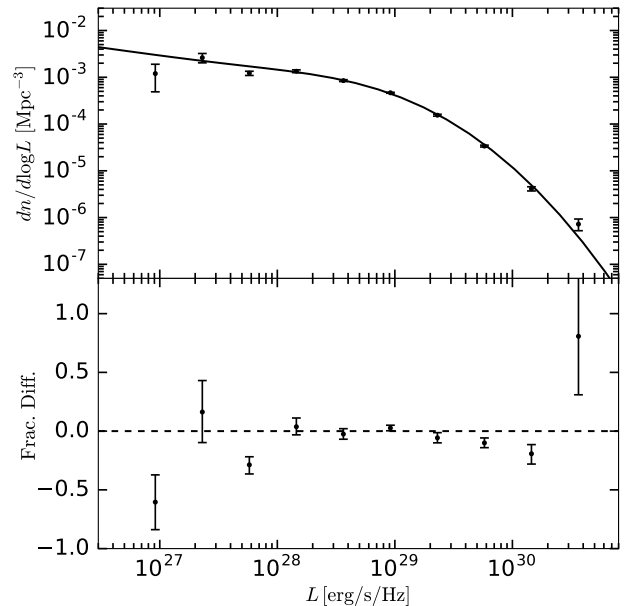
input data: the optical and radio luminosity functions from GAMA and NVSS respectively. (The fit to the simulated optical magnitude data was discussed in Sect. 3.6.)

Fig. 5 shows the luminosity function of star-forming radio galaxies at  $z \approx 0$  from the global best-fit model, along with the NVSS data. The residuals are relatively small, with no large deviations across the full luminosity range. The model has a  $\chi^2$  of 30.1 for 10 datapoints however, which is formally a poor fit (although recall that we used an approximate, non-Gaussian likelihood function, and have assumed that the errors are independent). This must be compared to the performance of other methods to model luminosity functions though, which also tend to produce formally poor fits. For example, if we fit the phenomenological Schechter-like luminosity function suggested by Mauch & Sadler (2007) to the radio LF only, using the same likelihood function, we obtain a best-fit model with  $\chi^2 = 18.6$  (for 4 free parameters), which gives a probability-to-exceed of 0.005 – a better, but still poor fit. We conclude that our model fit is acceptable for simple modelling purposes, due to the lack of significant deviations, but that it does not provide a complete, statistically acceptable description of the data.

The best-fit model for the optical LFs is shown in Fig. 4. The residuals are again reasonably small, except for noticeable deviations at the extreme bright and faint ends in most bands, and a bump feature at  $m_{\text{obs}} \approx -18$  in the *u*-band. The goodness-of-fit for the bands that were used in the MCMC were again relatively acceptable but formally poor, with  $\chi^2 = 39.5$  (49.2) and 34.1 (34.1) for the *g* and *z* bands, over 17 and 16 datapoints respectively (where the numbers in parentheses denote the  $\chi^2$  after the frequency-dependent dust attenuation parameters have been fitted to all five bands). The *u*, *r*, and *i* bands, which were not included in the MCMC, have  $\chi^2 = 195.2, 156.7, 87.8$  (161.0, 156.0, 85.0) respectively, with 17 datapoints each.

The deviations at the faint end of the LFs in Fig. 4 are likely due to completeness effects, where galaxies have gone undetected near the flux limit of the survey. While a completeness correction was applied to the data by Loveday et al. (2012), a systematic drop remains in the last few datapoints in all bands, which indicates a residual completeness systematic. We discarded the last 3 points in each band in the MCMC to avoid this skewing our results.

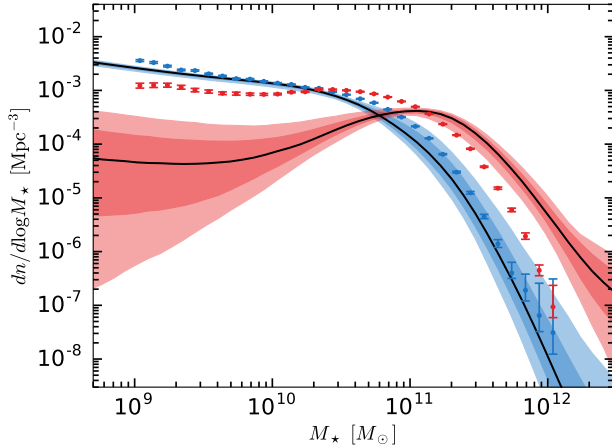
The deviations at the bright end are less severe, due to the larger errorbars there. These could be caused by a model error;



**Figure 5.** *Upper panel:* Luminosity function of star-forming radio galaxies from the best-fit model (solid black line) to the GAMA and NVSS data, with the latter data shown as black points. *Lower panel:* The fractional difference between the best-fit model and the NVSS datapoints.

we do not explicitly include a bright starburst galaxy population for example, but this would contribute most significantly at the bright end. The deviation could also be explained by selection effects, e.g. due to bright galaxies being intrinsically rarer and thus more likely to be under-sampled in surveys that cover limited volumes (as is always the case at  $z \approx 0$ ). Intrinsic dust attenuation also affects bright galaxies most significantly, so a model error or a selection effect related to this could also cause a discrepancy.

The bump feature in the *u*-band is significant, even if the fit is otherwise (qualitatively) a good one. There is a slight hint of such a feature at around the same position (i.e. a little fainter than the faint-end transition in the LF) in all of the bands. The best-fit model tries to explain this by allowing the LF of the sub-dominant passive sequence to form a relatively sharp peak at around this



**Figure 6.** Stellar mass function for SFMS (blue) and passive (red) galaxies. Datapoints are from SDSS/GALEX at  $z \approx 0.1$  (Moustakas et al. 2013). Solid black lines with coloured bands show the predictions of our global best-fit model, including 68% and 95% credible regions from the MCMC. Our model fits did not use the SDSS/GALEX data, or any other SMF data; these curves are predictions.

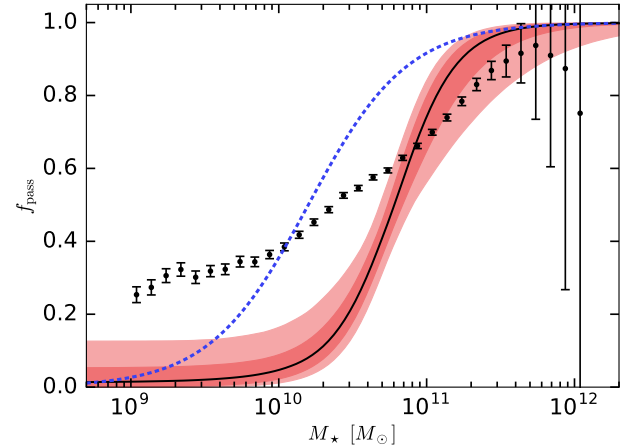
magnitude. This is achieved by driving the width of the passive sequence,  $\sigma_{\text{pass}}$ , to low values. Assuming that this bump is a real feature of the data, the failure of our  $u$ -band LF to reproduce it could be due to the mean optical magnitude model, Eq. 13. A systematic trend in the width and height of the bump in the passive LF can be seen in Fig. 4; both decrease towards bluer wavelengths. The data could likely be explained better if the bump was higher in the  $u$ -band, which may be achievable with a slightly different set of optical magnitude relation parameters for this band, or by taking into account the structure of the residuals in the  $\psi_{\text{SFR}} - M_*$  planes (see Fig. 2).

For the dataset that was directly fitted for (i.e. the  $g$  and  $z$  band optical LFs and radio LF), we found an effective  $\chi^2$  of 103.7 for 43 datapoints and 17 free parameters, giving roughly 26 degrees of freedom when the datapoints are assumed independent. For the most optimistic hypothesis, that the data and errorbars are correctly estimated and free of systematic errors, and that our model is the correct underlying description of the data, the probability to exceed this  $\chi^2$  is negligible ( $p \lesssim 3 \times 10^{-11}$ ), implying that this hypothesis is incorrect. For the full dataset, including the other three optical bands and the maximum likelihood-fitted frequency-dependent dust attenuation parameters, the total  $\chi^2$  is 515.5 over (effectively) 19 free parameters and 94 datapoints.

Our overall conclusion is that the data are qualitatively well-fit by the model, which produces luminosity functions that closely follow the shape and amplitude of the LF data. If one assumes that the errors on the data are independent, the goodness-of-fit is technically poor however. Numerous effects could contribute to cause this, one of which is likely to be the fact that the model is rather simple. Nevertheless, the results we obtained will be sufficient for many modelling purposes.

#### 4.4 Consistency with previous results

In this section, we compare our global best-fit model with a small selection of constraints on individual components of the model from the literature. Since we use only the radio and optical LFs, and semi-analytic model fits to the optical magnitude relation, to constrain the model, the constraints we place on other components are indirect. Comparing with direct constraints should then give us a good idea of the model’s consistency.



**Figure 7.** The predicted fraction of passive galaxies as a function of stellar mass. Red bands show the 68% and 95% credible intervals from our model, when fitted to the GAMA and NVSS optical/radio luminosity functions. The purple dashed line shows the simpler relation assumed in Behroozi et al. (2013). The black datapoints show the passive fraction from SDSS/GALEX, derived by dividing the reported quiescent stellar mass function by the sum of the quiescent and star-forming SMFs, and then estimating the errors using simple Gaussian error propagation.

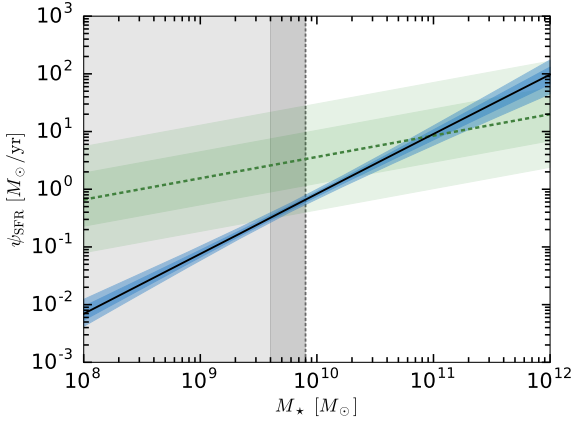
Fig. 6 shows the stellar mass function for the SFMS and passive populations for the global best-fit model, compared with data from SDSS/GALEX at  $z \approx 0.1$  (Moustakas et al. 2013). Neither the SFMS or passive SMF model predictions are a good fit to the data. The SFMS stellar mass function has a similar shape to the data, and fits well at  $M_* \sim 10^{10} M_\odot$ , which is encouraging given that this is a prediction based on a fit to very different data. The amplitude of the high-mass end is systematically low however, by 2-3 standard deviations per datapoint, which accumulates to yield a poor overall goodness of fit. The passive SMF is a much worse fit, despite the relatively large uncertainty at high and low stellar mass from the MCMC.

This discrepancy could be caused by a number of different effects. For example, the categorisation into star-forming and passive galaxies assumed by our model could differ markedly from the selection used by Moustakas et al. (2013), as the distinction is not particularly sharp. Fig. 7 shows the passive fraction, along with the SDSS/GALEX data and another model from the literature. Neither of the models nor the data agree with one another. An increase in the passive fraction at low mass would ease the discrepancy with the passive SMF, but make the discrepancy with the SFMS SMF worse. It is not clear that a globally acceptable fit could be found for both the SFMS and passive SMFs when taking the data at face value.

Residual systematic effects could also be present in the data. Fig. 4 of Moustakas et al. (2013) shows the total SMF from SDSS/GALEX plotted alongside four previous determinations of the SMF from the literature. The difference between the curves is several times the errorbars in most cases, suggesting that at least some of the five datasets are inconsistent with one another.

Finally, while the SDSS/GALEX errorbars shown in Fig. 6 are small, there still appears to be relatively little scatter in the positions of neighbouring datapoints, indicating that the errors are correlated. We were unable to find information in Moustakas et al. (2013) to quantify this (e.g. a correlation matrix), but if the correlation is strong, this will substantially reduce the amount of independent information in the data, and thus the significance of the discrepancy with our model prediction.

Fig. 8 shows the best-fit model for the mean SFR relation of

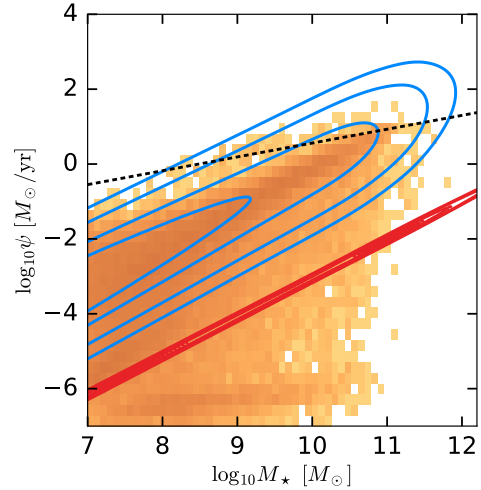


**Figure 8.** The mean star-formation rate as a function of stellar mass. The blue bands show 68% and 95% credible intervals for our model from the MCMC. The green dashed line shows the best-fit Wang et al. (2013) relation in the redshift range  $z = [0.2, 0.5]$  (corresponding to  $\alpha = -3.14$ ,  $\beta = 0.37$ ) along with the reported errors on the fit (green bands). The grey shaded regions show stellar masses that lie below the completeness limit of some (or all) of the data used by Wang et al. (2013) to fit their model.

the star-forming main sequence (Eq. 9), along with 68% and 95% credible intervals from our MCMC chains, and for a previous fit of a power-law relation to other data by Wang et al. (2013) (in the redshift bin  $z = [0.2, 0.5]$ ). Roughly following the trend with redshift found in Wang et al. (2013), one would expect an extrapolation of their model to  $z = 0$  to reduce its amplitude, but leave the slope relatively unchanged. The slope of our model is clearly much steeper. The comparison should only be performed above the completeness limit of the data that they used however, as denoted by the grey bands in Fig. 8. The errors on their fit are relatively large, and essentially cover the posterior of our relation above the completeness limit. Wang et al. (2013) do report that their best-fit model has a reduced  $\chi^2$  of 0.2 though, which likely indicates that the errorbars are overestimated.

The number density distribution in the  $\psi_{\text{SFR}} - M_{\star}$  plane is plotted in Fig. 9 for both the SFMS and passive population. These are compared with the distribution at  $z = 0$  from the Guo et al. (2011) simulation. The SFMS aligns very well with the one found in the simulation, with essentially the same slope and a similar width. Our model appears to overestimate the density of galaxies at high  $\psi_{\text{SFR}}$  and  $M_{\star}$ , but this could also be due to under-sampling in the simulation (which has a relatively small box size of  $62.5 h^{-1} \text{Mpc}$ ). The passive distribution is entirely different from the one in the simulation however; the simulated one has no sequence-like structure, while our best-fit model prefers a very narrow sequence. One can imagine that a much larger value of  $\sigma_{\text{pass}}$  would have been able to approximately reproduce the simulated result if this was preferred by the data, but it is not. As discussed in Sect. 4.3, part of the reason for this preference might be the need to reproduce a small, narrow feature in the optical LFs. The various semi-analytic models for the passive population also greatly differ from one another (see Sect. 3.3), and so the one that we chose (and/or the one in the simulation) may simply be inconsistent with the data. This could also cause the large discrepancy with the passive SMF that was seen in Fig. 6.

In summary, our model appears to reproduce previous results reasonably well, yielding qualitatively good fits to the optical and radio LFs, reasonable overlap with a previous determination of the star-forming main sequence, and only a mild underestimate of the stellar mass function of star-forming galaxies. The passive sequence of the best-fit model appears to be strongly discrepant with simulations and some previous results however (although we have



**Figure 9.** Predicted number density distribution in the  $M_{\star} - \psi_{\text{SFR}}$  plane,  $dn/d\log M_{\star} d\log \psi_{\text{SFR}}$ , for star-forming main sequence (blue) and passive (red) galaxies. The model shown is the best-fit model to the GAMA and NVSS optical/radio luminosity function data. Contours are shown at values of  $\{10^{-3}, 10^{-4}, 10^{-6}, 10^{-8}\} \text{Mpc}^{-3}$  (SFMS) and  $\{10^{-4}, 10^{-6}, 10^{-8}\} \text{Mpc}^{-3}$  (passive) respectively. The density plot (orange) shows the same quantity measured from the Guo et al. (2011) semi-analytic mock catalogue at  $z = 0$ . This simulation was not used in the fits, except to define the optical magnitude parameters in Eq. 13. The dashed line is the same as in Fig. 8.

noted that different simulations give a broad spread of predictions for the shape of the passive population in the SFR- $M_{\star}$  plane). A detailed investigation to construct a more accurate model for the passive population is left for future work.

#### 4.5 Joint luminosity functions

We now give an example of the kind of multiple-tracer prediction that the model is designed to calculate. Given two samples of galaxies in different wavebands, it is useful to quantify what fraction of the galaxies can be detected in both bands simultaneously. This can be used to predict how efficiently sources can be cross-matched between surveys, for example, or for calculating correlated shot noise in the cross-correlation of the surveys. The basic quantity required for these calculations is the joint luminosity function,  $dn(L_1, L_2)/d\log L_1 d\log L_2$ . As an example, we will calculate the joint luminosity function for one optical band and the 1.4 GHz radio continuum emission from star-forming galaxies. This can be written as

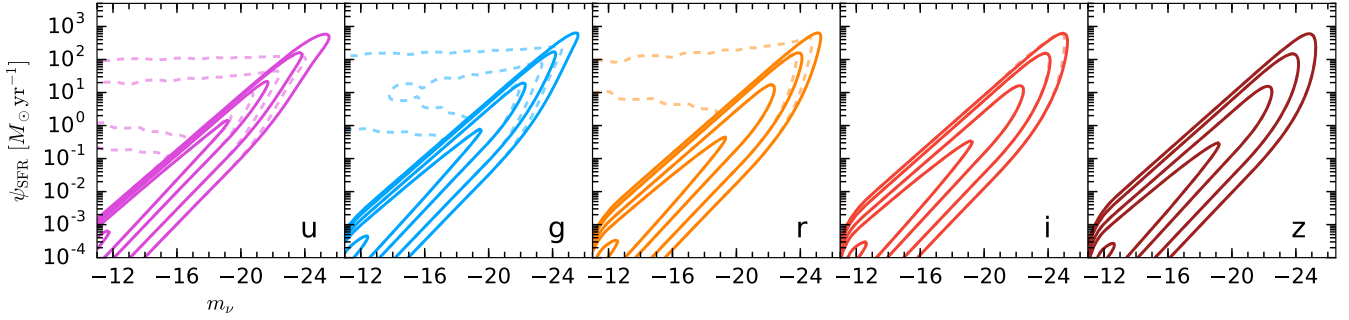
$$\begin{aligned} \frac{dn(L_{\text{rad}}, m_{\nu}^{\text{obs}})}{dL_{\text{rad}} dm_{\nu}^{\text{obs}}} &= \int d\psi_{\text{SFR}} \int dM_{\star} \int dM_{\text{h}} \frac{dn(M_{\text{h}})}{dM_{\text{h}}} \\ &\times p(L_{\text{rad}}, m_{\nu}^{\text{obs}} | M_{\star}, \psi_{\text{SFR}}) \\ &\times p(\psi_{\text{SFR}} | M_{\star}) \cdot p(M_{\star} | M_{\text{h}}). \end{aligned} \quad (33)$$

All of these distributions were defined above, with the exception of the joint luminosity/magnitude pdf, which we model as

$$p(L_{\text{rad}}, m_{\nu}^{\text{obs}} | M_{\star}, \psi_{\text{SFR}}) = p(L_{\text{rad}} | \psi_{\text{SFR}}) \cdot p(m_{\nu}^{\text{obs}} | M_{\star}, \psi_{\text{SFR}}). \quad (34)$$

This assumes that the two types of emission are linked only by their common dependence on  $\psi_{\text{SFR}}$  and  $M_{\star}$ , i.e. that they are independent when we condition on SFR and stellar mass. We have also used the fact that the radio luminosity depends only on SFR. In fact, our model assumes a deterministic relationship between SFR and  $L_{\text{rad}}$  (Eq. 12), so that

$$p(L_{\text{rad}} | \psi_{\text{SFR}}) = \delta^{\text{D}}(\psi_{\text{SFR}} - \psi_{\text{SFR}}^{\text{rad}}(L_{\text{rad}})), \quad (35)$$



**Figure 10.** Joint luminosity function of galaxies on the SFMS at  $z = 0$ , for a tracer of the SFR (e.g. radio continuum at 1.4 GHz) and five optical bands. The solid contours show the luminosity function without dust attenuation in the optical, while the dashed lines show the result with attenuation. The values of the contours are  $dn/d \log \psi_{\text{SFR}} dm_v = \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}\} \text{ Mpc}^{-3}$  (from outermost to innermost).

where  $\delta^D$  is the Dirac delta function. Eq. 33 then simplifies to

$$\frac{dn(L_{\text{rad}}, m_v^{\text{obs}})}{d \log L_{\text{rad}} dm_v^{\text{obs}}} = L_{\text{rad}} \int dM_{\star} \frac{dn(M_{\star})}{dM_{\star}} p(\psi_{\text{SFR}}^{\text{rad}}(L_{\text{rad}}) | M_{\star}) \times p(m_v^{\text{obs}} | M_{\star}, \psi_{\text{SFR}}^{\text{rad}}(L_{\text{rad}})), \quad (36)$$

where we have also changed variables to  $\log L_{\text{rad}}$ .

With the joint luminosity function in hand, it is straightforward to evaluate the number density of galaxies that are detected by both optical and radio surveys. Assuming luminosity and magnitude limits  $L_{\star}$  and  $m_{\star}$  respectively, we have

$$n(L_{\text{rad}} \geq L_{\star}, m_v^{\text{obs}} \leq m_{\star}) = \int_{\log L_{\star}}^{\infty} \int_{-\infty}^{m_{\star}} \frac{dn(L, m_v)}{d \log L dm_v} d \log L dm_v. \quad (37)$$

The fraction of optically-detected galaxies that also have radio counterparts (and vice versa) is then

$$f(\text{rad} \in \text{opt}) = \frac{n(L_{\text{rad}} \geq L_{\star}, m_v^{\text{obs}} \leq m_{\star})}{n(L_{\text{rad}} \geq 0, m_v^{\text{obs}} \leq m_{\star})} \quad (38)$$

$$f(\text{opt} \in \text{rad}) = \frac{n(L_{\text{rad}} \geq L_{\star}, m_v^{\text{obs}} \leq m_{\star})}{n(L_{\text{rad}} \geq L_{\star}, m_v^{\text{obs}} \leq \infty)}, \quad (39)$$

where the denominators are simply the total number density of optical/radio galaxies in each survey respectively, regardless of whether they have radio/optical counterparts or not.

Fig. 10 shows the joint luminosity function for optical bands and a general direct SFR tracer (like radio continuum) as predicted by our model, for the best-fit parameter values described above. As expected, galaxies that are most actively star-forming are brightest in the optical, although there is not a 1:1 correspondence even in the case where dust attenuation is neglected. This is simply a consequence of the optical magnitudes also tracing the stellar mass of the galaxies; the spread in  $M_{\star}$  for fixed  $\psi_{\text{SFR}}$  and the scatter in the optical magnitude pdf broaden the joint luminosity function. The correlation is tighter in the bluer bands, where the emission is dominated by younger stars that trace recent star formation.

With attenuation included, the picture becomes more complicated. For the bluest bands, in which attenuation is the strongest, a branch of high-SFR galaxies appears that spreads broadly down to fainter optical magnitudes. The spreading effect is reduced at lower SFR however, mostly because our model of attenuation scales with stellar mass; since  $\psi_{\text{SFR}}$  also scales with  $M_{\star}$ , lower SFR corresponds, on average, to lower stellar mass and thus weaker attenuation. Note that the number density of galaxies in this branch is relatively low, and that the main effect of attenuation is the shifting of the bright end of the joint luminosity function to fainter optical magnitudes.

## 5 DISCUSSION

We have presented a modular, analytic statistical model that connects host dark matter halo properties to bulk properties of the galaxies that reside within them, which are then linked to observables such as galaxy luminosity across a range of bands. This model was motivated by the existence of various scaling relations between these quantities, and the need to model multiple tracer populations across wavelength regimes when analysing forthcoming multi-survey large-scale structure datasets.

We showed that conditional distributions representing various scaling relations – many of which were already known in the literature – can be combined in a consistent way to simultaneously model the luminosity functions of low redshift galaxies in the optical and radio, with good precision. Consistency is achieved by jointly fitting the model parameters to multi-wavelength data, instead of combining relations that were separately fitted to various datasets. We also reported on the uncertainties and correlations between model parameters that resulted from this fitting procedure, using the results of an MCMC sampling procedure.

As presented here, our model predicts only a limited set of galaxy properties that are important for future surveys. We did not attempt to model (e.g.) the neutral gas content or emission from nuclear activity, even though these will be important for radio surveys for example. We posit that many such properties can be included as straightforward extensions to the model, by adding suitable conditional distributions to the probabilistic graph shown in Fig. 1. This will generally involve finding appropriate scaling relations between galaxy properties and then promoting them to (parametrised) statistical distributions.

Many such scaling relations have been seen observationally or in simulations, or result from simple physical arguments. Properties of active galactic nuclei (AGN) such as their radio and X-ray luminosity may be modelled through scaling relations involving the mass of their central black hole, stellar mass, or halo mass, for example (Ferrarese 2002; Fontanot et al. 2011; Aird et al. 2012). Relatively simple models of the neutral hydrogen (HI) content of galaxies also exist (e.g. Obreschkow et al. 2009; Bagla et al. 2010; Villaescusa-Navarro et al. 2014; Padmanabhan & Refregier 2016), often involving a scaling between HI mass and halo mass only. Aside from using the simple star-forming/passive categorisation to assign disk/elliptical labels to the galaxies, more sophisticated morphological properties can also be modelled using scaling relations (e.g. see the treatment in Schreiber et al. 2016). Adding these properties into the model, plus luminosity relations for other bands, is left for future work, but it should be clear that extensions like this are supported in a natural way.

While our model does explicitly include per-band optical



magnitudes, including dust attenuation, there is some further structure in the relationships between different bands that is not modelled – namely, colours. Galaxy colours are commonly used as selection criteria in surveys, as structures exist in colour-magnitude diagrams that can reliably distinguish between different types of galaxy. The existence of a red sequence in the colour-magnitude diagram (Bell et al. 2004) can be used to identify the large red galaxies that typically lie at the centre of galaxy clusters, for example (e.g. Rykoff et al. 2014). Our model fails to reproduce the red sequence however. This is because the optical pdfs (Eq. 14) are assumed independent; the magnitudes in different bands are linked only through the mutual dependence of the mean relations (Eq. 13) on stellar mass and SFR. Sequences in the colour-magnitude diagram require correlations in the scatter between two bands, however. This could be straightforwardly added by modelling the correlations between bands in Eq. 14.

While sufficient for some applications, this would also limit the model. Real galaxies have complex optical SEDs that depend on a host of other factors, in addition to bulk properties like stellar mass and SFR. The star-formation history of the galaxy and its initial mass function (IMF) are two such ingredients, both of which were used in the semi-analytic simulations from which our optical magnitude relations were calibrated. In future work, it would be desirable to remove the dependence of this part of the model on SAMs, and all of the assumptions that underlie them. One way of achieving this would be to build the optical magnitudes from bandpass integrals over distributions of template galaxy spectra, such as those derived from the stellar population synthesis models of Bruzual & Charlot (2003). This method has been used in a number of galaxy mock catalogue generation algorithms. The downside is that it introduces significant extra complexity, still requires a number of assumptions (e.g. about SFR history), and destroys the analytic nature of our model. A middle ground might be to use sets of empirical galaxy spectrum templates, such as those generated by certain photometric redshift estimation methods (e.g. Masters et al. 2015), although this also adds significant complexity. For applications that depend on the details of optical SEDs, then, it seems likely that more sophisticated modelling methods, such as semi-analytics, should be used instead.

The model is incomplete in another important respect – we have calibrated it using only  $z \approx 0$  observations. The redshift dependence of several of the constituent conditional distributions has been studied previously (e.g. see Behroozi et al. (2010) for the redshift dependence of the stellar mass-halo mass relation), and so redshift-dependent parametrisations are readily available. These simply add more parameters to the model that can be constrained with additional data at other redshifts. Other components are new to this work, however. We have tentatively confirmed that some of these components have redshift dependences that are amenable to simple parametrisations, by performing fits on semi-analytic simulations (e.g. for Eq. 13). A full treatment, including constraining the model from higher-redshift data, is left to future work.<sup>1</sup>

We have also neglected several observational effects that are important in modelling the observed distribution of galaxy luminosities. A small fraction of galaxies are strongly lensed, for example, which enhances their apparent luminosity. This causes a magnification bias (Kochanek 2006), which can be especially important at the bright end of the luminosity function, where intrinsically bright objects may be rarer than the lensed population of

fainter objects. As such, our luminosity functions should be convolved with a lensing magnification pdf (e.g. Marra et al. 2013) to account for this effect.

A useful aspect of the model is its modular nature. While we had to make a number of assumptions in the construction of the model, many of these can be tested by proposing alternatives for some of the conditional distributions shown in Fig. 1. Through statistical tests, such as Bayesian model selection procedures, one can then choose components that most faithfully describe various observations, while penalising overly-flexible models that lack predictive power. These can then be swapped into the default configuration of the model, without requiring it to be completely rebuilt.

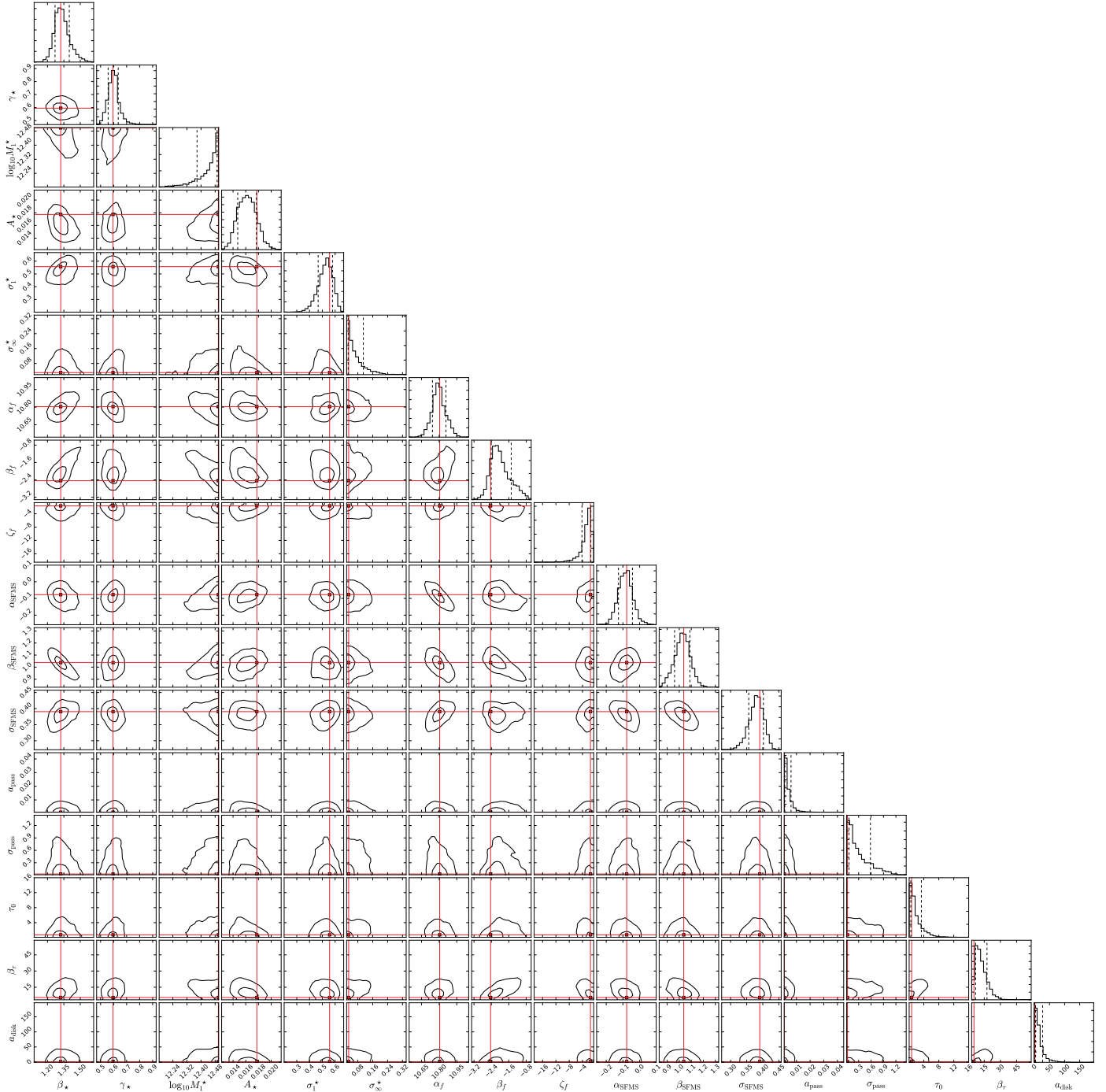
The type of model we have presented here fills a gap between detailed, computationally-expensive galaxy modelling methods like semi-analytics and hydrodynamical simulations on the one hand, and single-population halo models (e.g. HOD models) on the other. It has sufficient structure to allow consistent modelling of multiple galaxy populations across multiple wavelengths, but is simple enough that it can be explored analytically, allowing MCMC analyses to be performed without extensive computational resources. An open-source reference implementation, written in Python, and making use of the `numpy` and `scipy` (Jones et al. 2001) packages, is available from <https://github.com/philbull/ghost>.

*Acknowledgements:* I am grateful to O. Doré, B. Hensley, E. Huff, A. Merson, P. Serra, M. Viero, and H.-Y. Wu for valuable discussions, comments, and encouragement, and to A. Brown for valuable work on a summer project involving this model. I gratefully acknowledge use of computing resources at the University of Oslo. PB’s research was supported by an appointment to the NASA Postdoctoral Program at the Jet Propulsion Laboratory, California Institute of Technology, administered by Universities Space Research Association under contract with NASA.

## References

- Aird, J., Coil, A. L., Moustakas, J., et al. 2012, *ApJ*, 746, 90
- Bagla, J. S., Khandai, N., & Datta, K. K. 2010, *MNRAS*, 407, 567
- Barlow, R. 2003, *arXiv [physics/0306138]*
- Baugh, C. M. 2013, *Publ. Astron. Soc. Australia*, 30, e030
- Behroozi, P. S., Conroy, C., & Wechsler, R. H. 2010, *ApJ*, 717, 379
- Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013, *ApJ*, 770, 57
- Bell, E. F. 2003, *ApJ*, 586, 794
- Bell, E. F., Wolf, C., Meisenheimer, K., et al. 2004, *ApJ*, 608, 752
- Benson, A. J. 2012, *New Astron.*, 17, 175
- Berlind, A. A. & Weinberg, D. H. 2002, *ApJ*, 575, 587
- Berlind, A. A., Weinberg, D. H., Benson, A. J., et al. 2003, *ApJ*, 593, 1
- Brammer, G. B., Whitaker, K. E., van Dokkum, P. G., et al. 2011, *ApJ*, 739, 24
- Bruzual, G. & Charlot, S. 2003, *MNRAS*, 344, 1000
- Buat, V., Noll, S., Burgarella, D., et al. 2012, *A&A*, 545, A141
- Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, *ApJ*, 533, 682
- Clerkin, L., Kirk, D., Lahav, O., Abdalla, F. B., & Gaztañaga, E. 2015, *MNRAS*, 448, 1389
- Cohn, J. D. 2016, *ArXiv e-prints*, 1609.03956
- Condon, J. J. 1992, *ARA&A*, 30, 575
- Cooray, A. 2006, *Mon. Not. Roy. Astron. Soc.*, 365, 842
- Daddi, E., Dickinson, M., Morrison, G., et al. 2007, *ApJ*, 670, 156
- Driver, S. P., Popescu, C. C., Tuffs, R. J., et al. 2007, *MNRAS*, 379, 1022
- Elbaz, D., Daddi, E., Le Borgne, D., et al. 2007, *A&A*, 468, 33
- Ferrarese, L. 2002, *ApJ*, 578, 90
- Fontana, A., Santini, P., Grazian, A., et al. 2009, *A&A*, 501, 15
- Fontanot, F., Pasquali, A., De Lucia, G., et al. 2011, *MNRAS*, 413, 957
- Fontanot, F., Somerville, R. S., Silva, L., Monaco, P., & Skibba, R. 2009, *MNRAS*, 392, 553
- Foreman-Mackey, D. 2016, *The Journal of Open Source Software*, 24

<sup>1</sup> Ideally, this would result in every component having a redshift dependence that is well-described by simple, few-parameter extensions to the model. In the worst case is should still be possible to find a new set of best-fit parameters for every redshift slice, however.



**Figure A.** 2D posterior distributions of the 17 parameters that were varied in the MCMC. The contours are 1 and  $2\sigma$  bounds, and the marker shows the best-fit model. The 1D marginal distribution for each parameter is shown on the diagonal. Some of the parameters with strongly non-Gaussian 1D marginal distributions are restricted by their prior bounds (e.g.  $\sigma_\infty^*$ ), while others simply prefer a logarithmic distribution (e.g.  $a_{\text{disk}}$ ). We used the *corner* software (Foreman-Mackey 2016) to make this plot.

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306  
 Gaztañaga, E., Eriksen, M., Crocce, M., et al. 2012, *MNRAS*, 422, 2904  
 Guo, Q., White, S., Boylan-Kolchin, M., et al. 2011, *Mon. Not. Roy. Astron. Soc.*, 413, 101  
 Hearin, A., Campbell, D., Tollerud, E., et al. 2016, *ArXiv e-prints*, 1606.04106  
 Heinis, S., Buat, V., Béthermin, M., et al. 2014, *MNRAS*, 437, 1268  
 Henriques, B. M. B., Thomas, P. A., Oliver, S., & Roseboom, I. 2009, *MNRAS*, 396, 535  
 Hopkins, P. F., Kereš, D., Oñorbe, J., et al. 2014, *MNRAS*, 445, 581

Jarvis, M., Bacon, D., Blake, C., et al. 2015, *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, 18  
 Johnston, D. E., Sheldon, E. S., Wechsler, R. H., et al. 2007, *ArXiv e-prints*, 0709.1159  
 Jones, E., Oliphant, T., Peterson, P., et al. 2001, *SciPy: Open source scientific tools for Python*, [Online; accessed 2016-10-27]  
 Jovel, S., Kneib, J.-P., Ilbert, O., et al. 2009, *A&A*, 504, 359  
 Karim, A., Schinnerer, E., Martínez-Sansigre, A., et al. 2011, *ApJ*, 730, 61  
 Kauffmann, G. & Haehnelt, M. 2000, *MNRAS*, 311, 576  
 Kennicutt, Jr., R. C. 1998, *ARA&A*, 36, 189

- Kochanek, C. S. 2006, Saas-Fee Advanced Course 33: Gravitational Lensing: Strong, Weak and Micro
- Koda, J., Blake, C., Beutler, F., Kazin, E., & Marin, F. 2016, MNRAS, 459, 2118
- Krause, E., Hirata, C. M., Martin, C., Neill, J. D., & Wyder, T. K. 2013, MNRAS, 428, 2548
- Krumpe, M., Miyaji, T., & Coil, A. L. 2010, ApJ, 713, 558
- Lagos, C. D. P., Lacey, C. G., Baugh, C. M., Bower, R. G., & Benson, A. J. 2011, MNRAS, 416, 1566
- Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473
- Li, S.-J., Zhang, Y.-C., Yang, X.-H., et al. 2016, Research in Astronomy and Astrophysics, 16, 013
- Loveday, J., Norberg, P., Baldry, I. K., et al. 2012, MNRAS, 420, 1239
- Manera, M., Scoccimarro, R., Percival, W. J., et al. 2013, MNRAS, 428, 1036
- Marra, V., Quartin, M., & Amendola, L. 2013, Phys. Rev. D, 88, 063004
- Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53
- Mauch, T. & Sadler, E. M. 2007, MNRAS, 375, 931
- McDonald, P. & Seljak, U. 2009, JCAP, 0910, 007
- Miyaji, T., Krumpe, M., Coil, A. L., & Aceves, H. 2011, ApJ, 726, 83
- Moster, B. P., Somerville, R. S., Maubetsch, C., et al. 2010, ApJ, 710, 903
- Moustakas, J., Coil, A. L., Aird, J., et al. 2013, ApJ, 767, 50
- Moustakas, J., Kennicutt, Jr., R. C., & Tremonti, C. A. 2006, ApJ, 642, 775
- Noeske, K. G., Weiner, B. J., Faber, S. M., et al. 2007, ApJ, 660, L43
- Obreschkow, D., Croton, D., De Lucia, G., Khochfar, S., & Rawlings, S. 2009, ApJ, 698, 1467
- Padmanabhan, H. & Refregier, A. 2016, ArXiv e-prints, 1607.01021
- Pannella, M., Carilli, C. L., Daddi, E., et al. 2009, ApJ, 698, L116
- Parkinson, H., Cole, S., & Helly, J. 2008, MNRAS, 383, 557
- Peacock, J. A. & Smith, R. E. 2000, MNRAS, 318, 1144
- Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, ApJ, 721, 193
- Planck Collaboration. 2016, A&A, 594, A13
- Pozzetti, L., Hirata, C. M., Geach, J. E., et al. 2016, ArXiv e-prints, 1603.01453
- Reid, B. A., Seo, H.-J., Leauthaud, A., Tinker, J. L., & White, M. 2014, MNRAS, 444, 476
- Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, ApJ, 785, 104
- Salim, S., Rich, R. M., Charlot, S., et al. 2007, ApJS, 173, 267
- Schreiber, C., Elbaz, D., Pannella, M., et al. 2016, ArXiv e-prints, 1606.05354
- Skibba, R. A. & Sheth, R. K. 2009, MNRAS, 392, 1080
- Sokal, A. 1997, in Functional integration (Springer), 131–192
- Somerville, R. S., Popping, G., & Trager, S. C. 2015, MNRAS, 453, 4337
- Somerville, R. S. & Primack, J. R. 1999, MNRAS, 310, 1087
- Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Nature, 435, 629
- Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
- Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709
- Tuffs, R. J., Popescu, C. C., Völk, H. J., Kylafridis, N. D., & Dopita, M. A. 2004, A&A, 419, 821
- Vale, A. & Ostriker, J. P. 2004, MNRAS, 353, 189
- Villaescusa-Navarro, F., Viel, M., Datta, K. K., & Choudhury, T. R. 2014, J. Cosmology Astropart. Phys., 9, 050
- Vogelsberger, M., Genel, S., Springel, V., et al. 2014, Nature, 509, 177
- Wang, L., Farrah, D., Oliver, S. J., et al. 2013, MNRAS, 431, 648
- White, M., Blanton, M., Bolton, A., et al. 2011, ApJ, 728, 126
- White, S. D. M. & Rees, M. J. 1978, MNRAS, 183, 341
- Xu, X., Ho, S., Trac, H., et al. 2013, ApJ, 772, 147
- Yang, X., Mo, H. J., van den Bosch, F. C., et al. 2007, ApJ, 671, 153
- Yang, X.-h., Mo, H. J., & van den Bosch, F. C. 2003, Mon. Not. Roy. Astron. Soc., 339, 1057
- Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2005, ApJ, 630, 1
- Zheng, Z., Zehavi, I., Eisenstein, D. J., Weinberg, D. H., & Jing, Y. P. 2009, ApJ, 707, 554