

Sub-millimeter galaxies in hierarchical models: revisiting the need for a top-heavy stellar initial mass function with Bayesian optimisation

Edward J. Elliott,^{1,2,3} Carlton M. Baugh,^{1,2}★ Cedric G. Lacey,¹

¹*Institute for Computational Cosmology, Department of Physics, Science Laboratories, South Road, Durham, DH1 3LE, UK*

²*Institute for Data Science, Science Laboratories, South Road, Durham, DH1 3LE, UK*

³*Engelhart, 7th Floor, Berkeley Square House, 4-19 Berkeley Square, London, W1J 6BR, UK*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The properties of high-redshift sub-millimetre galaxies (SMGs) remain controversial within hierarchical structure formation models. We revisit whether a top-heavy stellar initial mass function (IMF) in starbursts is required to reproduce both SMG observations and local galaxy properties. Using Bayesian optimisation, we perform an extensive search of the 15-dimensional parameter space of the GALFORM semi-analytical model. This efficient approach converges to optimal parameter values in fewer than 200 model evaluations, representing orders of magnitude fewer runs than traditional methods. We test whether GALFORM can simultaneously match three key observational constraints: the $z = 0$ K -band luminosity function, the SMG number counts at 850 μm , and the SMG redshift distribution. We consider two model variants: one with a universal solar neighbourhood IMF for all star formation, and another allowing the IMF slope in starbursts to vary as a free parameter. When assuming a universal Chabrier IMF, we find no parameter combination that simultaneously reproduces all three datasets. The model either matches the SMG constraints while grossly overpredicting the local K -band luminosity function, or matches the local luminosity function while severely underpredicting SMG counts by factors of 3–100. In contrast, allowing a top-heavy IMF in starbursts enables excellent simultaneous fits to all constraints. The best-fitting model prefers an IMF slope parameter $x \approx 0.7$ (where $\text{dn}/\text{dlog}m \propto m^{-x}$), somewhat more top-heavy than recent models but less extreme than early proposals. Our comprehensive parameter space exploration definitively confirms that, within the GALFORM framework, a top-heavy IMF in starbursts is necessary to reconcile high-redshift dusty star-forming galaxies with local galaxy populations.

Key words: methods:statistical – methods:numerical – galaxies:formation

1 INTRODUCTION

Observations of high-redshift galaxies provide powerful constraints on galaxy formation models, particularly when combined with measurements of local galaxy populations. Sub-millimetre galaxies (SMGs), first discovered using the SCUBA instrument on the James Clerk Maxwell Telescope (Smail et al. 1997; Hughes et al. 1998), represent a population undergoing intense dust-obscured star formation at high redshift (for a review, see Casey et al. 2014). The sub-millimetre emission from these galaxies benefits from a negative k -correction, such that their observed flux remains nearly constant over a wide range of redshifts for a fixed total infrared luminosity. This property provides a unique window for studying galaxy evolution across cosmic time.

Bright SMGs contribute up to half the star formation rate density at $z \sim 2$ –3 despite their low space density of $\sim 10^{-5} \text{ cMpc}^{-3}$ (Chapman et al. 2005; Smith et al. 2017). They possess large stellar masses of $\sim 10^{11} M_{\odot}$ (Swinbank et al. 2004; Da Cunha et al. 2015) and reside in massive dark matter halos of $\sim 10^{13} M_{\odot}$ (Blain et al. 2004; Hickox et al. 2012). Assuming their emission is powered by star formation with a solar neighbourhood IMF yields intense star formation rates,

suggesting that the SMG phase could build a substantial fraction of the stellar mass in present-day massive elliptical galaxies.

A viable galaxy formation model should reproduce both the number counts and redshift distribution of SMGs while simultaneously matching observations of local galaxies. This has proven remarkably challenging. Early semi-analytic models underestimated SMG number counts by more than an order of magnitude (Granato et al. 2000; Somerville et al. 2012). Baugh et al. (2005) proposed that adopting a stellar initial mass function in starbursts with a larger proportion of massive stars than in the solar neighbourhood (specifically, $\text{dn}/\text{dlog}m \propto m^{-x}$ with $x = 0$; compared to $x = 1.35$ above one solar mass for a Chabrier or Salpeter IMF) allowed the model to match both SMG number counts and redshift distributions without compromising low-redshift predictions. In this model, both dust absorption of starlight and the spectral energy distribution of dust emission were calculated self-consistently using a physical model that included the dependence of dust temperature on dust mass. Later model iterations (Lacey et al. 2016; Baugh et al. 2019; Cowley et al. 2019), which include AGN heating feedback to regulate the abundance of bright galaxies, reproduce SMG and local galaxy properties with a somewhat less extreme but still top-heavy IMF in bursts, with $x \sim 1$.

McAlpine et al. (2019) presented predictions from the EAGLE hydrodynamic simulation (Schaye et al. 2015), post-processed with

★ E-mail: c.m.baugh@durham.ac.uk

a dust radiative transfer code, assuming a Chabrier IMF for all star formation. These authors found some agreement between the model and the SMG redshift distribution for $z > 1$, but the simulation greatly underpredicted number counts, with essentially no sources brighter than 5 mJy (see comparison in Cowley et al. 2019). The SIMBA hydrodynamic simulation (Davé et al. 2019; Lovell et al. 2021), also assuming a Chabrier IMF, similarly produces reasonable agreement with the SMG redshift distribution but underpredicts counts by factors of 3–10 at the bright end (depending on assumptions about source blending) and by more than an order of magnitude at the faint end. Hayward et al. (2021) compared sub-millimetre predictions for the Illustris and IllustrisTNG simulations (Nelson et al. 2015; Pillepich et al. 2018) using the fitting formula from Hayward et al. (2011). Illustris reproduces number counts reasonably well but predicts a significantly lower median redshift for bright sources than observed, while IllustrisTNG produces a better redshift distribution but underpredicts SMG number counts by more than an order of magnitude at bright fluxes.

Here we revisit the need for a top-heavy IMF in hierarchical galaxy formation models. Allowing the IMF to vary with star formation mode has been controversial (see, for example, Bastian et al. 2010). Nevertheless, observational evidence for IMF variations exists. Using emission line strengths and Balmer decrements from star-forming galaxies in the Galaxy and Mass Assembly Survey, Gunawardhana et al. (2011) argued that the IMF slope varies with star formation rate, becoming more top-heavy as the rate increases, reaching $x \approx 0.9$ in the most intense star-forming galaxies—similar to the value adopted by Lacey et al. (2016). Romano et al. (2017) inferred a top-heavy IMF with $x \approx 0.95$ in nearby starburst galaxies based on CNO isotopic ratios in the ISM, extended by Zhang et al. (2018) to starbursts at $z \sim 2$ –3. Schneider et al. (2018), studying massive stars in the Large Magellanic Cloud, found $x = 0.9 \pm 0.3$, with subsequent analysis by Farr & Mandel (2018) finding $x = 1.05 \pm 0.14$, both top-heavy compared to the solar neighbourhood IMF. While many studies provide evidence for IMF variations, the exact nature remains uncertain. Some studies (e.g., Conroy & van Dokkum 2012; Smith 2020) find evidence for bottom-heavy IMFs in high-mass galaxies (noting that an IMF can be simultaneously both top-heavy and bottom-heavy compared to a Chabrier IMF), while Weidner et al. (2013) propose a time-dependent IMF favouring massive stars at early times and low-mass stars at late times.

All galaxy formation models—whether hydrodynamic simulations or semi-analytic codes—involve numerous parameters governing sub-grid processes such as star formation and feedback from supernovae and active galactic nuclei (Bauha 2006; Benson & Bower 2010a; Crain et al. 2015; Somerville & Davé 2015). These parameters must be calibrated to match some subset of observations before predictions can be made for other observables. One criticism of including a top-heavy IMF in GALFORM is that, given the relatively large parameter space, another unexplored parameter combination might fit observations without this assumption. We aim to robustly determine whether a parameter set exists that allows the model, assuming a universal solar neighbourhood IMF, to match SMG observational constraints while maintaining reasonable fits to low-redshift datasets such as the K -band luminosity function. As a secondary aim, we investigate the required IMF slope in starbursts: will a more extensive parameter space search reveal the possibility of a less extreme IMF in bursts?

To achieve these aims, we employ Bayesian optimisation techniques (see e.g. Frazier 2018). This approach calibrates model parameters by searching for the best-fitting parameter set, judged by how well the model reproduces target datasets. The fitting metric mea-

sures the distance between model predictions and calibration data. This represents a “black-box” problem: we do not know how the metric depends on model parameters. Bayesian optimisation is designed for computationally expensive models requiring optimisation methods that use few model evaluations. In Bayesian optimisation, the metric is estimated at every point in parameter space with varying uncertainty. We use Gaussian processes (see e.g. Rasmussen & Williams 2006) to describe the metric. As more model evaluations are performed, the Gaussian process is updated. Once no further significant improvement occurs in the metric, the best-fitting model has the lowest metric value.

We show that this method can find suitable fits in a small fraction of the model evaluations required by other, often more elaborate, approaches (see e.g. Kampakoglou et al. 2008; Henriques et al. 2009; Bower et al. 2010; Vernon et al. 2010; Benson & Bower 2010b; Lu et al. 2011, 2012; Ruiz et al. 2015; Martindale et al. 2017; van der Velden et al. 2019, 2021; Elliott et al. 2021; Madar et al. 2024). However, this approach provides limited information about uncertainty in the best-fit parameters in exchange for requiring an order of magnitude fewer full model evaluations. Bayesian optimisation has been successfully applied to other computationally expensive astronomical problems involving high-dimensional parameter spaces, including interpreting supernovae light curves (Leclercq 2018), analysing the Lyman- α forest power spectrum (Rogers et al. 2019; Takhtaganov et al. 2021), and searching for signatures of inflation in cosmic microwave background data (Hamann & Wons 2022).

Semi-analytical model calibration and parameter space exploration has generally taken two forms: (1) direct exploration using full model calculations, and (2) emulation, where a small number of full model runs are performed with the bulk of parameter sampling done by a computationally cheap surrogate. Although semi-analytical models are much cheaper than hydrodynamic simulations, direct parameter space exploration remains computationally expensive due to the large number of model runs required for formal searches of high-dimensional spaces.

Examples of the first approach using full model runs include Kampakoglou et al. (2008), who implemented Markov Chain Monte Carlo (MCMC) techniques to calibrate a semi-analytic model to multiple observational datasets. Henriques et al. (2009) used MCMC to calibrate the L-GALAXIES model to several datasets, finding that the choice of datasets altered the values of the best-fitting parameters, pointing to model deficiencies. Lu et al. (2011, 2012) constrained the parameter space providing acceptable fits to the K -band luminosity function, expanding this to include the HI mass function in Lu et al. (2014). Ruiz et al. (2015) used particle swarm optimisation to calibrate a semi-analytic model to the K -band luminosity function.

The second class of methods involves constructing a statistical emulator of the semi-analytic model that can be evaluated orders of magnitude faster than running the model itself, at the cost of being approximate. Bower et al. (2010) and Vernon et al. (2010) employed a Bayesian emulation technique (developed by Goldstein & Wooff 2007) to constrain the GALFORM parameter space region providing reasonable fits to the K - and b_J -band luminosity functions (see also Benson & Bower 2010b). The same approach was applied by Rodrigues et al. (2017) to calibrate GALFORM to the stellar mass function in the local Universe and its evolution, and by van der Velden et al. (2021) to calibrate the Meraxes semi-analytic model at high redshift. The Bayes linear approach used by Bower et al. (2010) makes approximations about the functions being minimised and is not strictly a black-box method like Bayesian optimisation. The Bower et al. (2010) method involved searching the parameter space in waves, with the space redefined at each wave to make the

search adaptive. Bower et al. (2010) used 5500 runs of GALFORM in their search of a parameter space similar in size to that considered here; we use fewer than 200 runs of the full model in one sweep of the parameter space, making the search fully automatic.

We presented a new framework for automated GALFORM calibration using local observations in Elliott et al. (2021). The first step was sensitivity analysis to determine which model parameters most strongly shaped predictions for the calibration data (see Oleskiewicz & Baugh 2020 for the first application of sensitivity analysis to GALFORM). This identified a subset of 10 parameters as most important for determining the calibration data. We ran GALFORM for 1000 parameter sets sampled from this 10-dimensional space using a Latin hypercube (Stein 1987). These runs allowed us to build an emulator of GALFORM using an artificial neural network. The emulator enabled extensive MCMC searches of the parameter space, returning best-fitting parameter sets for chosen calibration data along with the range of acceptable model predictions. The speed of this approach allowed us to explore best-fitting models resulting from different calibration dataset combinations. Differences in resulting best-fitting models point to possible model deficiencies or to incompatibilities between observational measurements. Madar et al. (2024) applied a similar framework to the calibration of GALFORM to reproduce the number counts of H_α emitters.

Here we face a more challenging problem. First, the parameter space is larger than in Elliott et al. (2021), with 15 dimensions instead of 10. The focus of the parameter space search is now: “Can we find any example of a model that works under these assumptions?” rather than finding a best-fitting parameter set with its associated range of acceptable models. Second, the computational overhead for each parameter set is much higher. We need predictions for galaxy number counts and redshift distributions, which require running GALFORM at many redshifts rather than just $z = 0$. Moreover, some predictions are sensitive to rare events such as starbursts, requiring many more dark matter halo merger histories to obtain robust predictions.

To overcome these challenges, we investigate a new approach to model calibration using Bayesian optimisation. This global optimisation technique efficiently searches the parameter space using a Gaussian process prior and is capable of efficiently searching high-dimensional parameter spaces for global minima. We use this method to test whether a parameter set exists that can match SMG observations and low-redshift constraints simultaneously without including a top-heavy IMF. Past parameter space explorations, though usually performed manually, have suggested this is not possible when assuming a universal solar neighbourhood IMF. We aim to test this conclusion with a more sophisticated parameter search over a comprehensive list of relevant parameters and a wide search space. This approach represents an enormous improvement over the old-fashioned one-parameter-at-a-time searching originally used to argue for a top-heavy IMF.

This paper is structured as follows: in §2, we briefly introduce GALFORM, focusing on processes and parameters relevant to our study. In §3, we review the theory and practical considerations behind Bayesian optimisation. We list the observational datasets used for calibration in §3.5 and validate our method on a surrogate model in §3.6. In §4, we present the model calibration for different assumptions about the IMF and about the importance placed on reproducing various datasets. We give our conclusions in §5.

2 GALAXY FORMATION MODEL

We use the GALFORM semi-analytical model of galaxy formation introduced by Cole et al. (2000) (see also Bower et al. 2006 and Lacey et al. 2016). GALFORM models the key physical process thought to shape the formation and evolution of galaxies in the cold dark matter cosmology (for reviews of the semi-analytical approach to modelling galaxy formation see Baugh 2006 and Benson & Bower 2010a). The model starts from a sample of dark matter halos along with their formation and merger histories, which is typically extracted from a cosmological simulation of the dark matter evolution. The model then tracks the transfer of mass and metals between different reservoirs of baryons, with each halo typically containing both a central galaxy and satellite galaxies. It predicts the full star formation history of each galaxy in a halo, along with the chemical evolution of the gas. Two modes of star formation are considered: quiescent star formation in galaxy disks and bursts of star formation triggered by galaxy mergers or by bar instabilities in disks. GALFORM combines this information with a stellar population synthesis model to predict the luminosities of galaxies in different bands (see Gonzalez-Perez et al. 2014). Here we use the model implemented in the P-Millennium cosmological N -body simulation of the evolution of the dark matter distribution (Baugh et al. 2019).

Below we introduce the parameters that are varied in the calibration and explain how they affect various processes. For a full list of the model parameters and a complete description of the model, see Lacey et al. (2016). The list of model parameters varied, and the ranges considered for each parameter are given in Table 1.

2.1 Quiescent star formation in disks

GALFORM uses an empirical star formation law formulated by Blitz & Rosolowsky (2006) based on observations of local star-forming disk galaxies, which was implemented by Lagos et al. (2011). The SFR in the disk is given by

$$\psi_{\text{disk}} = \nu_{\text{SF}} M_{\text{mol, disk}}, \quad (1)$$

where $M_{\text{mol, disk}}$ is the mass of molecular gas in the disk, and ν_{SF} is a constant which we treat as an adjustable parameter within a range suggested by the observations (Bigiel et al. 2011). The fraction of cold gas in the molecular phase depends on the gas pressure in the mid-plane of the disc (see Lagos et al. 2011).

2.2 Galaxy mergers

In the model, galaxy mergers can trigger bursts of star formation and destroy galactic disks. Mergers occur between satellite and central galaxies, on a timescale controlled by dynamical friction on the satellite due to the main halo. We calculate this merger timescale using the improved galaxy merger model of Simha & Cole (2017), in which satellite subhalos are tracked in the dark matter simulation, up until a subhalo is longer resolved, after which the remaining time for the satellite to merge is calculated analytically based on dynamical friction arguments (see Campbell et al. 2015 for details of the implementation of this model in GALFORM). We define two thresholds, f_{ellip} and f_{burst} , which determine the outcome when a satellite galaxy with baryonic mass $M_{\text{b, sat}}$ merges with a central galaxy with baryonic mass $M_{\text{b, cen}}$. First, if $M_{\text{b, sat}}/M_{\text{b, cen}} \geq f_{\text{ellip}}$ the merger is classified as a *major* merger, and the disk component of the galaxy is destroyed and is added to the spheroid component of the galaxy if one exists, or a new spheroid is formed. The cold gas in the disk is assumed to be consumed in a burst of star formation and is

Table 1. The parameters explored in this work and the range of values over which they are varied. The first column gives the symbol for the parameter (and units where relevant), the second gives the process or effect the parameter is involved in, and the third column gives the range of parameter values probed.

Name	Process	Range
ν_{SF} (Gyr^{-1})	Quiescent star formation	0.1 - 4.0
f_{ellip}	Galaxy mergers	0.2 - 0.5
f_{burst}	Galaxy mergers	0.01 - 0.3
F_{stab}	Disk instability	0.5 - 1.2
$\tau^{*}_{\text{burst, min}}$ (Gyr)	Burst star formation	0.01 - 0.05
x	IMF slope in bursts	0 - 1.35
$V_{\text{SN, disk}}$ (km s^{-1})	SN feedback	10 - 800
$V_{\text{SN, burst}}$ (km s^{-1})	SN feedback	10 - 800
γ_{SN}	SN feedback	1.0 - 4.0
α_{reheat}	SN feedback	0.2 - 3.0
f_{SMBH}	BH growth	0.001 - 0.05
α_{cool}	AGN feedback	0.2 - 4.0
f_{cloud}	Dust	0.2 - 0.8
t_{esc} (Gyr)	Dust	0.0001 - 0.01
β_{burst}	Dust	1.5 - 1.2

added to the spheroid. Second, if $M_{\text{b, sat}}/M_{\text{b, cen}} < f_{\text{ellip}}$, the merger is classified as *minor*, and the disk survives. In this case, the cold gas in the disk is consumed in a starburst which adds stars to the spheroid if a second condition is met, namely if $M_{\text{b, sat}}/M_{\text{b, cen}} \geq f_{\text{burst}} \cdot f_{\text{burst}}$ and f_{ellip} are free parameters.

2.3 Disk instabilities

Galactic disks that are dominated by rotational motion can become unstable to bar formation if their degree of self-gravity is too large. The GALFORM model follows [Efstathiou et al. \(1982\)](#), and assumes that disks become unstable if the following criterion is met:

$$\frac{V_{\text{c}}(r_{\text{disk}})}{(1.68 G M_{\text{disk}}/r_{\text{disk}})^{1/2}} \leq F_{\text{stab}}, \quad (2)$$

where M_{disk} is the total disk mass and r_{disk} is the disk half-mass radius. Numerical simulations of exponential stellar disks by [Efstathiou et al. \(1982\)](#) found a value of $F_{\text{stab}} \approx 1.1$, while [Christodoulou et al. \(1995\)](#) found a value of 0.9 for gaseous disks. A value of 0.61 or below corresponds to universally stable disks, since this is the value of the left hand side of Eqn. 2 for a completely self-gravitating disk. We allow the parameter F_{stab} to vary within a reasonable range (see Table 1). We assume that disks that become unstable are disrupted by bar instabilities on a sub-resolution timescale, such that all the mass is instantaneously transferred to the spheroid and any gas present takes part in a burst of star formation, adding stars to the spheroid.

2.4 Starbursts

Bursts of star formation, triggered by mergers or bar instabilities in dynamically unstable disks, are assumed to form stars at a rate

$$\psi_{\text{burst}} = \nu_{\text{SF, burst}} M_{\text{cold, burst}} = \frac{M_{\text{cold, burst}}}{\tau^{*}_{\text{burst}}}, \quad (3)$$

where the star formation timescale is given by

$$\tau^{*}_{\text{burst}} = \max[f_{\text{dyn}} \tau_{\text{dyn, bulge}}, \tau^{*}_{\text{burst, min}}]. \quad (4)$$

Here the bulge dynamical time is defined as $\tau_{\text{dyn, bulge}} = r_{\text{bulge}}/V_{\text{c}}(r_{\text{bulge}})$, where the velocity is the circular velocity at the half-mass radius of the bulge. The minimum timescale of a burst,

$\tau^{*}_{\text{burst, min}}$, is treated as an adjustable parameter in the range 1-100 Myr. f_{dyn} is held at the value of 20 used by [Lacey et al. \(2016\)](#).

2.5 Supernova feedback

Supernova explosions lead to gas being ejected from galaxies and out of the halo. The model assumes that the mass ejection rate is proportional to the instantaneous star formation rate, ψ , with a mass loading factor dependent on the circular velocity of the galaxy, V_{c} :

$$\dot{M}_{\text{eject}} = \left(\frac{V_{\text{c}}}{V_{\text{SN}}} \right)^{-\gamma_{\text{SN}}} \psi, \quad (5)$$

where both V_{SN} and γ_{SN} are model parameters. The circular velocity used is that at the half-mass radius of the disk for quiescent star formation or the circular velocity at the half-mass radius of the spheroid for starbursts. Separate values of V_{SN} can be specified for star formation taking place in the disk ($V_{\text{SN, disk}}$) and bulge ($V_{\text{SN, burst}}$) components of a galaxy. These parameters have generally been assumed to be the same in most previous versions of the model (although see [Benson & Bower 2010b](#) for a counter-example). Gas ejected from the halo is assumed to return from a reservoir beyond the halo's virial radius to the hot gas reservoir in the main halo at a rate given by

$$\dot{M}_{\text{return}} = \alpha_{\text{ret}} \frac{M_{\text{res}}}{\tau_{\text{dyn, halo}}}, \quad (6)$$

where $\tau_{\text{dyn, halo}}$ is the dynamical time of the halo, M_{res} is the mass in the reservoir of ejected gas beyond the virial radius, and α_{ret} is a free parameter.

2.6 SMBH growth and AGN feedback

Supermassive black holes (SMBH) can inject energy into the halo gas, inhibiting gas cooling. Hot halo accretion, BH-BH mergers, as well as starbursts, can all increase the mass of the black hole ([Bower et al. 2006](#); [Croton et al. 2006](#); [Lagos et al. 2008](#)). In the case of starbursts, the mass accreted onto the SMBH is a fraction f_{SMBH} of the mass of stars formed, where f_{SMBH} is an adjustable parameter. AGN heating is assumed to occur if both of the following conditions are met: (1) the gas halo is in quasi-hydrostatic equilibrium, that is the condition:

$$\tau_{\text{cool}}/\tau_{\text{ff}} > 1/\alpha_{\text{cool}}, \quad (7)$$

is met, where τ_{cool} is the cooling time of the gas, τ_{ff} is the free-fall time, and α_{cool} is an adjustable parameter; (2) the AGN power required to balance the radiative cooling luminosity of the hot halo gas is below a fraction f_{Edd} of the Eddington luminosity of the SMBH (which is held fixed at $f_{\text{Edd}} = 0.01$). The implementation of the growth of SMBH in GALFORM is described in [Lacey et al. \(2016\)](#) and [Griffin et al. \(2019\)](#).

2.7 Stellar initial mass function

The stellar initial mass function (IMF) gives the mass distribution of newly formed stars and strongly affects the evolution of the total luminosity, as well as the metal and gas content of the galaxy. The IMF, $\Phi(m)$ is defined such that the number of stars with mass m is $dN = \Phi(m) d\ln m$, and $\Phi(m)$ is normalised such that $\int_{m_{\text{L}}}^{m_{\text{H}}} m \Phi(m) d\ln m = 1$ between some minimum (m_{L}) and maximum (m_{H}) stellar mass.

In some of the models considered later, we will allow the slope of the IMF to depend on the mode of star formation, assuming a solar

neighbourhood IMF for quiescent star formation that takes place in disks and a power law IMF for bursts of star formation. The slope of the IMF power law, x , is then a model parameter:

$$\Phi(m) := dN/d\ln m \propto m^{-x}. \quad (8)$$

To accommodate this change, we must self-consistently calculate the recycled fraction (i.e. the fraction of mass returned to the ISM from mass lost by stars over their lifetime), given by

$$R = \int_{1M_{\odot}}^{m_H} (m - m_{\text{rem}}(m)) \Phi(m) d\ln m, \quad (9)$$

where m_{rem} is the mass of the remnant left by a dying star of birth mass m . We also calculate the yield, p , i.e. the fraction of the initial mass synthesised into metals and ejected, as

$$p = \int_{1M_{\odot}}^{m_H} p_Z(m) m \Phi(m) d\ln m, \quad (10)$$

where $p_Z(m)$ is the fraction of mass ejected as newly produced metals by a star of mass m .

For reference, the solar neighbourhood IMF (Chabrier 2003) assumed in quiescent star formation tends to a power law slope of $x = 1.35$ above one solar mass and turns over with a log-normal form below one solar mass. Most published variants of GALFORM have adopted a solar neighbourhood IMF in all modes of star formation, typically the Kennicutt IMF (Kennicutt 1983). Different forms of the IMF are compared in fig. 2 of Lacey et al. (2016).

2.8 Stellar Population Synthesis

We calculate galaxy spectral energy distributions (SEDs) by building a composite stellar population from simple stellar populations (SSPs) of age t and metallicity Z , for the adopted IMF, using the predicted star formation and chemical evolution histories of the progenitors of a galaxy. To calculate the SED for each SSP we use the stellar population synthesis (SPS) model developed by Conroy et al. (2009).

We consider two variants of GALFORM. In the first case, we assume that the IMF is the same in starbursts and quiescent star formation, and has the solar neighbourhood form proposed by Chabrier (2003). In the second case, quiescent star formation again takes place following a solar neighbourhood IMF, but the IMF in bursts takes on a power-law form, with the slope of the power law being a model parameter. The integral to find the luminosity per unit wavelength L_{λ} is performed separately over stars formed in bursts and those formed in the disk. The results for the disk and bulge are then added to get the total galaxy luminosity.

The FSPS model (Conroy 2013) includes a flexible treatment for the contribution to the SED from thermally-pulsating asymptotic giant branch (TP-AGB) stars, a stellar phase which is difficult to model accurately. Conroy (2013) introduces a parameterisation to adjust the effect of the TP-AGB contribution. Adjusting these parameters, we find that this has little effect on the datasets that we consider, so we use the default values. Gonzalez-Perez et al. (2014) show the effect of choosing different SPS models on the GALFORM model predictions. Gonzalez-Perez et al. found that the model predictions for the luminosity function in the ultra-violet and optical are insensitive to the choice of the near-infrared SPS model, whereas the evolution of the luminosity function in the near-infrared does depend on the treatment of TP-AGB stars. The FSPS model allows new SSP tables to be generated for any IMF.

2.9 Absorption and reradiation of starlight by dust

Within galaxy disks, a two-component dust model is assumed with diffuse and molecular cloud components. In both cases, the dust is mixed in with the stars. The mass of dust in a galaxy is assumed to be a constant fraction δ_{dust} of the mass of metals in the cold gas, $M_{\text{dust}} = \delta_{\text{dust}} Z_{\text{cold}} M_{\text{cold}}$, where $\delta_{\text{dust}} = 0.334$ (Silva et al. 1998). The dust grain size distribution is assumed to be the same as in the solar neighbourhood. The optical depth of dust in the disk is then calculated as

$$\tau_{\text{dust},\lambda} = 0.043 \left(\frac{k_{\lambda}}{k_V} \right) \left(\frac{\Sigma_{\text{gas}}}{M_{\odot} \text{pc}^{-2}} \right) \left(\frac{Z_{\text{cold}}}{0.02} \right), \quad (11)$$

where Σ_{gas} is the surface density of the gas, k_{λ} represents the extinction curve (being the wavelength-dependent extinction coefficient normalized to some reference wavelength), and k_V is the extinction coefficient in the V-band, which is used as the normalization reference for k_{λ} .

The model assumes that a fraction f_{cloud} of the gas and dust is in clouds of mass $m_{\text{cloud}} = 10^6 M_{\odot}$ and radius $r_{\text{cloud}} = 16$ pc, based on observations of local galaxies (Granato et al. 2000). Stars are assumed to form inside molecular clouds and escape over an adjustable timescale t_{esc} . The optical depth in each cloud therefore scales as $\tau_{\text{cloud}} \propto m_{\text{cloud}}/r_{\text{cloud}}^2$, and the optical depth of the diffuse component as $\tau_{\text{diffuse}} \propto (1 - f_{\text{cloud}}) M_{\text{cold}} Z_{\text{cold}}/r_{\text{diffuse}}^2$. Here, r_{diffuse} is taken to be the disk radius, r_{disk} , for stars in the disk, and r_{bulge} , for stars in the bulge. Attenuation by diffuse dust is calculated by interpolating the tabulated results of radiative transfer runs by Benson (2018) using the HYPERION code of Robitaille (2011). These models are higher resolution versions of the dust extinction tables previously used in GALFORM, which were based on the radiative transfer calculations by Ferrara et al. (1999), and extend over a wider range of optical depth values.

The dust re-radiates the energy it absorbs from the starlight at infrared and sub-millimetre wavelengths. In the model, the emission from the diffuse and cloud components are treated separately. We calculate the total stellar luminosity absorbed by the dust in a galaxy, and assume the dust reradiates this energy as a modified black body:

$$L_{\lambda}^{\text{dust}} \propto M_{\text{dust}} \kappa_d(\lambda) B_{\lambda}(T_{\text{dust}}), \quad (12)$$

where M_{dust} is the mass and T_{dust} the temperature of the dust component, $B_{\lambda}(T)$ is the Planck function, and κ_d is the dust opacity per unit mass. Integrating $L_{\lambda}^{\text{dust}}$ over wavelength and setting the result equal to the absorbed luminosity allows us to solve for the dust temperature, T_{dust} , for each component. The dust opacity in the IR/sub-mm is approximated as a broken power law

$$\kappa_d \propto \begin{cases} \lambda^{-2} & \lambda < \lambda_b \\ \lambda^{-\beta_b} & \lambda > \lambda_b, \end{cases} \quad (13)$$

where we allow the adjustable exponent β_b for bursts to vary within the range 1.5 - 2.0 (see e.g. Silva et al. 1998). For quiescent disks, $\lambda_b = \infty$ and for bursts $\lambda_b = 100 \mu\text{m}$.

3 BAYESIAN OPTIMISATION

We use Bayesian optimisation (see Frazier 2018 and Garnett 2023 for reviews) to set the model parameters so that GALFORM reproduces as closely as possible the observational datasets used in the model calibration. Here we explain why we take this approach and set out its background. We start in § 3.1 with an overview of the problem, the exploration of a high dimensional parameter space of a computationally expensive model, and explain how Bayesian optimisation

helps us to address this challenge. The nature and role of Gaussian processes, the tool used to encapsulate our knowledge of the parameter space is described in § 3.2. The kernel function is an important part of the Gaussian process description and is explained in § 3.3. The method used to choose where in the parameter space to add new calculations with the full model is described in § 3.4. The datasets used to calibrate GALFORM are listed in § 3.5. A simple validation of our approach is given in § 3.6. We close in § 3.7 by describing the practical application of Bayesian optimisation to GALFORM.

3.1 An overview of Bayesian optimization

Calibrating a galaxy formation model involves finding parameter values that make the model predictions match observations as closely as possible. We measure this match using a metric, f , which quantifies the discrepancy between model predictions and observational data—the smaller the value of f , the better the match.

In a traditional calibration approach, we would evaluate the model at many different points in the parameter space (where each point represents a specific choice of parameter values) and rank them according to f . For a small number of parameters, we could simply try every possible combination on a grid. However, in the application here GALFORM has 15 adjustable parameters, making a complete grid search computationally impossible. Even sophisticated methods like Markov Chain Monte Carlo (MCMC; Robert 2015) require thousands of model evaluations to thoroughly explore such a high-dimensional space.

The fundamental challenge is that each GALFORM evaluation has a modest computational cost that soon becomes prohibitive if a large number of runs is needed. For the predictions used here, we must run the model at many different redshifts to obtain galaxy number counts and redshift distributions, making each evaluation costly. We need a method that can find good parameter values with far fewer model runs.

Bayesian optimisation provides a solution by using a scheme to decide where to evaluate the model next. Instead of blindly sampling the parameter space, it builds up knowledge about the metric f across the entire parameter space and uses this knowledge to decide where the next evaluation is most likely to improve our understanding or find a better fit. A simple analogy is looking for a missing set of keys: think of it as the difference between randomly searching for your keys versus systematically checking the places they are most likely to be, and updating your search strategy based on where you have already looked.

The key insight is that we do not need to know the exact value of f everywhere in the parameter space. Instead, we maintain a probabilistic description of what f might be at each location, with varying degrees of certainty. Where we have evaluated the full model, our certainty is high. Where we have not explored, our uncertainty is larger. This probabilistic description is provided by a Gaussian process (GP), which we explain in the next subsection.

After an initial set of model evaluations (typically about 30 for our 15-dimensional space, obtained using Latin hypercube sampling), Bayesian optimisation begins an iterative process: (1) use the GP to estimate f everywhere in the parameter space, (2) identify where the next evaluation would be most valuable, (3) run GALFORM at that location, and (4) update the GP with this new information. As we show later, this approach typically finds excellent parameter values within just 100–200 model evaluations—orders of magnitude fewer than traditional methods would require.

This is technically classified as a black-box, derivative-free, global optimisation problem (Frazier 2018). “Black-box” means we do not

assume any particular mathematical form for f ; “derivative-free” means we do not compute how f changes with respect to each parameter (information that could guide a gradient-based search); and “global optimisation” means we aim to find the best solution across the entire parameter space, not just a local minimum.

3.2 Gaussian processes: describing our knowledge of the parameter space

A Gaussian process (GP) provides a mathematical framework for describing what we know about the metric function f throughout the parameter space (Rasmussen & Williams 2006). Crucially, the GP not only gives an estimate of f at any location, but also quantifies our uncertainty in that estimate.

To understand GPs intuitively, imagine you are trying to map the elevation of a landscape. If you have measured the height at a few specific locations, you can make reasonable guesses about the height at nearby points—elevation typically changes smoothly rather than jumping erratically. Points very close to your measurements can be estimated with high confidence, while points far from any measurement are more uncertain.

Quantitatively, we consider the function $f(\mathbf{x})$, where $\mathbf{x}$ represents a specific set of values for all 15 parameters. The GP assumes that the joint probability distribution for the values of f at a set of n points in this parameter space is a multivariate Gaussian, with mean $\mu(\mathbf{x})$, and covariance $K(\theta, \mathbf{x}_i, \mathbf{x}_j)$ between points $\mathbf{x}_i$ and $\mathbf{x}_j$ that also depends on parameters θ . Mathematically, this can be written as $f \sim \mathcal{N}(\mu, K)$. In our application, we set the unconstrained mean $\mu(\mathbf{x})$ to zero because our target is to minimize f (make it as close to zero as possible). If the values of f at a set of points $\mathbf{x}_1, \dots, \mathbf{x}_n$ are already known (through exact evaluations of the model), then the probability distribution for the value of f at a point $\mathbf{x}'$ is given by the conditional probability distribution. This conditional probability distribution is also a Gaussian, with constrained mean $\mu(\mathbf{x}')$ and constrained variance $\sigma^2(\mathbf{x}')$ that depend on the values of f at the points $\mathbf{x}_1, \dots, \mathbf{x}_n$. The mean value $\mu(\mathbf{x}')$ is our best estimate of $f(\mathbf{x}')$, while $\sigma(\mathbf{x}')$ measures the uncertainty in this estimate.

The covariance matrix $K(\theta, \mathbf{x}_i, \mathbf{x}_j)$ encodes the key assumption that underlies GPs: nearby points in the parameter space should have similar values of f . This matrix describes how strongly the value of f at one location $\mathbf{x}_i$ is correlated with its value at another location $\mathbf{x}_j$. The closer two points are in the parameter space, the more strongly correlated their f values should be. The parameter θ controls the length scale over which correlations persist—essentially, how rapidly f can change as we move through the parameter space.

To illustrate how the length scale affects the GP, Fig. 1 shows several random draws from GPs with different correlation lengths. Each curve represents a possible realization of what f might look like. When the correlation length is short (left panel), the function varies rapidly. When it is long (right panel), the function varies smoothly over larger distances. The GP learns appropriate length scales from the data.

As we evaluate GALFORM at more points in the parameter space, the GP is updated with this new information. The uncertainty shrinks to nearly zero at the locations where we have run the full model, while remaining larger in unexplored regions. This updated GP then guides where we should evaluate the model next, as we describe in Section 3.4.

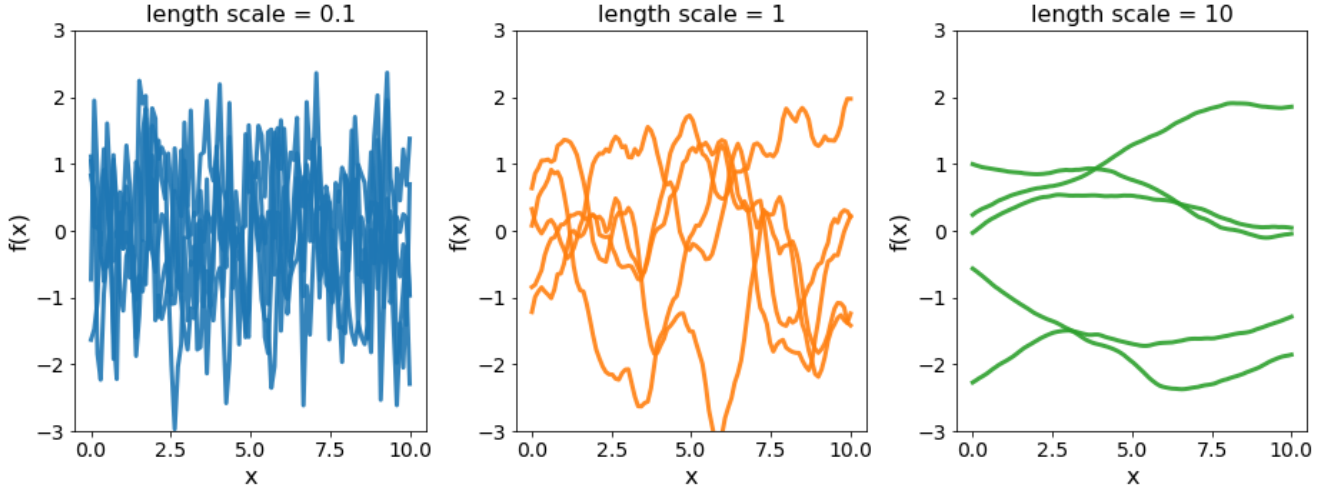


Figure 1. A demonstration of the effect of the length scale adopted in the kernel function on the appearance of a Gaussian process (GP). Each panel shows several realisations or draws from a GP. In each case, the process has zero mean. However, the hyperparameter that governs the scales over which values of f are correlated varies between panels. The left panel shows the shortest correlation length scale, with $\theta = 0.1$, the middle panel shows 1.0, and the right panel 10.0. A shorter length scale corresponds to a function which changes rapidly with small changes to the input parameters.

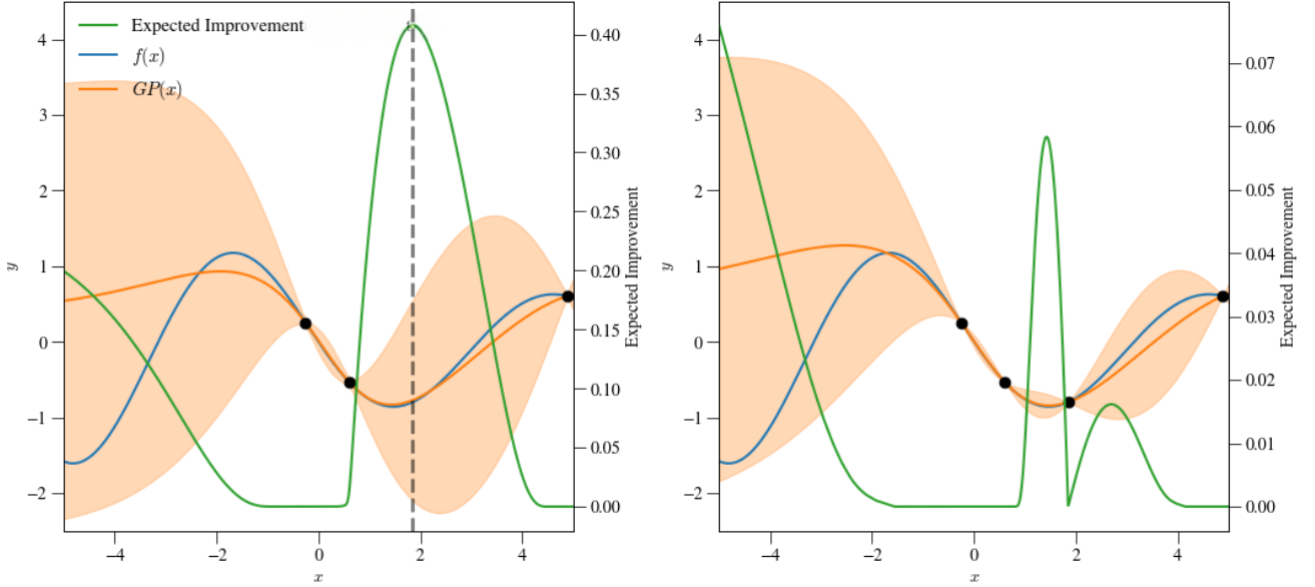


Figure 2. An illustration of one iteration of the expected improvement (EI) algorithm. The objective is to find the minimum value of the target function shown by the blue curve. Two iterations are plotted, $n = 3$ (left panel) and $n = 4$ (right panel). The Gaussian process (GP) posterior (orange solid line) is an estimate of the target function. The orange-shaded region shows the 3σ confidence interval of the GP. The left panel shows the GP after 3 evaluations of the function (shown by the black solid points). The green curve shows the EI (right axis), which corresponds to the expectation integral of the GP posterior below the minimum evaluation so far (i.e. how much we expect to improve upon the current minimum evaluation at each point x). The EI suggests that a new evaluation just below $x = 2$ will bring the biggest improvement in our knowledge of the target function. The right panel shows the updated GP posterior and EI curve after evaluating the function for $n = 4$ at the point of maximum expected improvement, as shown by the black dashed line in the left panel. At this point, the next evaluation would be chosen to be below $x = -4$.

3.3 The kernel function: encoding smoothness

The kernel function (also called the covariance function) is the mathematical tool that implements the correlation structure of the GP. It determines how the value of f at one location in the parameter space relates to its value at another location.

For our application, we use the Matérn 5/2 kernel, which has proven effective for optimisation problems in many fields. The covariance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the parameter space is given

by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \left(1 + \sqrt{5}r + \frac{5r^2}{3} \right) \exp(-\sqrt{5}r), \quad (14)$$

where

$$r = \sqrt{\sum_{m=1}^d \frac{(x_{im} - x_{jm})^2}{l_m^2}}. \quad (15)$$

Here, d is the number of dimensions in the parameter space (15 in

our case), x_{im} denotes the value of parameter m at location $\mathbf{x}_i$, and l_m is the length scale for parameter m . The parameter σ^2 controls the overall variance.

A crucial feature of this kernel is that it includes a separate length scale l_m for each parameter dimension. This property, called *automatic relevance determination*, allows the GP to learn that changes in some parameters affect f more strongly than changes in others. Parameters with long length scales contribute less to determining f —the metric changes slowly as these parameters vary. Conversely, parameters with short length scales are highly relevant—small changes in these parameters lead to significant changes in how well the model matches observations.

The Matérn 5/2 kernel has several advantages for our application:

- It assumes that f is smooth (twice differentiable), which is physically reasonable for our model
- It is flexible enough to capture complex behaviour
- It is computationally efficient to evaluate
- The automatic relevance determination helps identify which parameters most strongly affect the model fits

By learning appropriate length scales from the GALFORM evaluations, the GP builds an increasingly accurate representation of the metric function across the parameter space.

3.4 Choosing where to evaluate next: the acquisition function

At each step of the optimisation, we must decide where in the parameter space to evaluate GALFORM next. This decision involves a fundamental trade-off. On one hand, we want to evaluate the model where we expect to find parameter values that give a better fit to the data (lower f). On the other hand, we want to explore regions where our uncertainty is high, as these unexplored regions might contain even better solutions that we have not yet discovered. This is the classic *exploitation versus exploration* dilemma.

The acquisition function provides a mathematical framework for balancing this trade-off. It assigns a score to every possible location in the parameter space, indicating how valuable it would be to evaluate the model there. We then choose to evaluate GALFORM at the location with the highest acquisition function value.

We use the *expected improvement* (EI) acquisition function, which has a clear intuitive interpretation. Let $f_{\min}^n$ denote the smallest (best) value of the metric we have found after n evaluations of GALFORM. For any new location $\mathbf{x}$, we can ask: what improvement over $f_{\min}^n$ can we expect if we evaluate the model at $\mathbf{x}$? The EI quantifies this expected improvement, accounting for both the GP's predicted value of $f(\mathbf{x})$ and the uncertainty in this prediction.

More formally, we define a utility function, u :

$$u(\mathbf{x}) = \max(0, f_{\min}^n - f(\mathbf{x})). \quad (16)$$

This utility is zero if the new evaluation would not improve upon our current best result (i.e. $f(\mathbf{x}) > f_{\min}^n$), and positive if it would yield a better fit. Since the GP provides a probability distribution for $f(\mathbf{x})$ at any location (rather than a single value), we compute the expected utility by integrating over all possible values weighted by their probabilities:

$$a_{\text{EI}}(\mathbf{x}) = \mathbb{E}[u(\mathbf{x})] = \int_{-\infty}^{\infty} u(\mathbf{x}) \text{GP}(\mathbf{x}) d\mathbf{x}, \quad (17)$$

where $\text{GP}(\mathbf{x})$ denotes the probability distribution given by the Gaussian process at location $\mathbf{x}$.

Fortunately, this integral has a closed-form solution. Let $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ denote the conditional mean and conditional standard deviation

predicted by the GP at location $\mathbf{x}$. Define the standardized improvement:

$$Z(\mathbf{x}) = \frac{f_{\min}^n - \mu(\mathbf{x})}{\sigma(\mathbf{x})}, \quad (18)$$

assuming $\sigma(\mathbf{x}) > 0$. Then the expected improvement can be written as:

$$a_{\text{EI}}(\mathbf{x}) = \sigma(\mathbf{x}) [Z(\mathbf{x})\Phi(Z(\mathbf{x})) + \phi(Z(\mathbf{x}))], \quad (19)$$

where Φ is the cumulative distribution function of the standard normal distribution and ϕ is its probability density function.

This formula (Eqn. 20) explicitly reveals how exploitation and exploration contribute to the expected improvement. Both terms inside the brackets are multiplied by the uncertainty $\sigma(\mathbf{x})$, which means that *both contributions vanish in regions where we are certain about f* :

- **Exploitation contribution:** The term $Z(\mathbf{x})\Phi(Z(\mathbf{x}))$ is large when the GP predicts a mean value $\mu(\mathbf{x})$ much lower than the current best $f_{\min}^n$ (making Z large and positive). The cumulative distribution function $\Phi(Z)$ gives the probability that sampling at $\mathbf{x}$ will yield an improvement. This term favors regions where we expect good results.

- **Exploration contribution:** The term $\phi(Z(\mathbf{x}))$ is the standard normal probability density, which is largest when $Z \approx 0$ —that is, when the predicted mean $\mu(\mathbf{x})$ is close to the current best value. This term ensures that regions with high uncertainty are explored even when the predicted mean is only moderately promising.

Crucially, the factor $\sigma(\mathbf{x})$ outside the brackets means that *both* exploitation and exploration are scaled by uncertainty. At points we have already evaluated, where $\sigma(\mathbf{x}) = 0$, the EI is exactly zero—there is no value in re-evaluating the model at the same parameters, regardless of how good those parameters were. Conversely, regions with large uncertainty receive higher scores, even if the mean prediction alone would not make them the most promising location.

The optimal balance between exploitation and exploration emerges automatically from this mathematical structure. A region with very low $\mu(\mathbf{x})$ (good predicted performance) but small $\sigma(\mathbf{x})$ (we are already confident about this region) may score lower than a region with moderate $\mu(\mathbf{x})$ but large $\sigma(\mathbf{x})$ (significant chance of discovering something unexpectedly better).

Fig. 2 illustrates how the EI acquisition function works for a simple one-dimensional example. In the left panel, we show an unknown target function (blue curve) that we are trying to minimize. We have evaluated this function at three locations (black points). The orange line shows the GP's current estimate of the function, with the shaded region indicating the uncertainty (specifically, the 3σ confidence interval). Notice that the uncertainty is zero at the evaluated points and grows larger in unexplored regions.

The green curve shows the EI acquisition function. It is highest slightly to the right of $x = 2$, balancing two factors: (1) the GP predicts relatively low values of f in this region (exploitation), and (2) the uncertainty is large enough that we might discover even lower values (exploration). The black dashed line marks where we would evaluate the function next.

The right panel shows the updated GP and EI after evaluating the function at the location identified in the left panel. The uncertainty has decreased around the new evaluation point, and the EI now identifies a different location (far left) as the most promising for the next evaluation.

This is an iterative process: GALFORM is evaluated at the location with the highest EI, the GP is updated with the new result, the EI is recalculated across the parameter space, and the next location

at which to evaluate the model is found. After sufficient iterations, the algorithm converges to a parameter set that gives the best fit to the observational data, and which subsequently is not significantly improved upon, as we demonstrate in Section 3.6.

3.5 Dataset selection for parameter calibration

We aim to test the ability of GALFORM to reproduce observations of dusty star-forming galaxies, SMGs, detected at sub-mm wavelengths, which tend to be high-redshift objects, as well as matching the properties of the low-redshift galaxy population. In Elliott et al. (2021) we used nine observational datasets measured for local galaxies to calibrate GALFORM. To make the analysis simpler to follow, here we restrict ourselves to the $z = 0$ K -band luminosity function and supplement this with the number counts and redshift distribution of SMGs at $850\ \mu\text{m}$. We investigate if the best fitting models can reproduce the local $850\ \mu\text{m}$ luminosity function and other selected $z = 0$ galaxy observations in § 4.3, but emphasize that these datasets are not used in the calibration.

Below we list the observational datasets used to constrain the model parameters:

- For the local K -band LF, we compare to the observational estimate from Kochanek et al. (2001).
- For the SMG number counts, at the bright end we compare to Stach et al. (2018), based on ALMA $870\ \mu\text{m}$ follow-up observations of a sample of sources detected at $850\ \mu\text{m}$ by SCUBA, and at the faint end we compare to Chen et al. (2013).
- For the SMG redshift distribution, we compare the model to the measurements from Dudzeviciute et al. (2020). These authors used the same $850\ \mu\text{m}$ SCUBA and $870\ \mu\text{m}$ ALMA data as Stach et al. (2018), and then used ancillary data at other wavelengths in conjunction with the MAGPHYS code (Da Cunha et al. 2015) to obtain photometric redshifts. We apply a flux limit $S_{870} > 4.0\ \text{mJy}$ (as in Stach et al. 2018) to the Dudzeviciute et al. (2020) sample, and then weight the individual sources by the scaling factors as a function of $850\ \mu\text{m}$ flux from Geach et al. (2017), which account for varying depth and incompleteness in the original SCUBA survey. We then calculate the redshift distribution by stacking the MAGPHYS-derived redshift probability distributions for the individual sources.

It is important to note that the redshift distribution of SMGs that we obtained using the data from Dudzeviciute et al. (2020) has a somewhat *higher* median redshift than previous estimates used to calibrate the GALFORM models. Wardlow et al. (2011) reported a median photometric redshift for their sample of $z = 2.2 \pm 0.2$ when including only SMGs with robustly detected optical counterparts, but this increased to $z = 2.5 \pm 0.5$ when a slightly higher flux limit $S_{850} > 4\text{mJy}$ was applied and allowing for the likely redshifts of SMGs with no identifiable optical counterpart. For comparison, applying the same flux limit $S_{850} > 4\text{mJy}$ to the sample of Dudzeviciute et al. (2020) gives a median redshift of $z = 2.8 \pm 0.09$. The two estimates of the SMG redshift distribution for sources brighter than $4\ \text{mJy}$ are shown in Fig. 3 (where the distribution plotted for the Wardlow et al. (2011) sample only includes SMGs with robust optical identifications). The model calibrations performed in Lacey et al. (2016) and Baugh et al. (2019) thus used an SMG redshift distribution with a somewhat lower median redshift than the calibration data employed here (note that Lacey et al., applied a brighter flux limit to the SMG redshift distribution, but only considered SMGs from Wardlow et al. with optical counterparts). It is therefore interesting to find out if the model with a top-heavy IMF can still reproduce the higher median redshift found by Dudzeviciute et al. (2020), and if

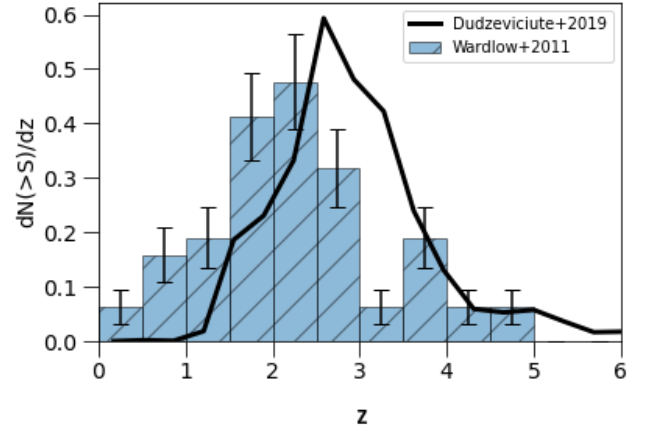


Figure 3. Comparison between the redshift distribution for SMGs brighter than $4\ \text{mJy}$ inferred by Dudzeviciute et al. (2020) (black solid line) and Wardlow et al. (2011) (blue histogram, which only includes SMGs with robust optical counterparts). Both redshift distributions have been normalized to unit area under the curve. Here, we calibrate GALFORM to the redshift distribution estimated by Dudzeviciute et al. (2020).

this has implications for the value of the recovered slope of the IMF in starbursts.

We will see later that the GALFORM predictions for the redshift distribution of SMGs show a series of spikes. The spikes at intermediate and high redshifts arise due to the limited number of halo merger trees used in the calculation. Using many more halos would smooth out these spikes, but at the expense of greatly increasing the computational cost for each full model evaluation. The spike at low redshift ($z \sim 0.1$) is different, being robust to using more halos, and results from the local population of quiescently star-forming galaxies. This local SMG population appears to be under-represented in the SMG sample analysed by Dudzeviciute et al. (2020), possibly because the survey field was chosen to avoid nearby galaxies so as to target the high-redshift population. However, the local SMG population is robustly detected observationally in surveys that measure the local $850\ \mu\text{m}$ LF, as discussed in § 4.3. To remove any issue with the low redshift spike, we excluded both both model and observed SMG at $z < 0.8$ when we calculated the MAE for the SMG redshift distribution in the parameter fitting.

3.6 Validation of the optimisation approach

Before running the Bayesian optimisation on the GALFORM model, we first assess whether this method is likely to succeed. We also need to gain some insight into the convergence properties of the optimisation process (or parameter calibration) so that we have some guidance as to when to end the search. To do this, we build an approximate emulator of GALFORM so that we can get a better understanding of the optimisation routine, without requiring many time-consuming evaluations of the GALFORM model.

To carry out this test, we first build a neural network emulator to predict all three observational datasets described in § 3.5 from 500 runs of the full GALFORM model, with samples drawn from the space defined by the parameter ranges given in Table 1 using a Latin hypercube. The trained neural network provides us with a fast alternative to running GALFORM, and allows us to test our optimisation strategy before applying it to the full model. This is similar to the approach taken in Elliott et al. (2021). One might ask why we cannot

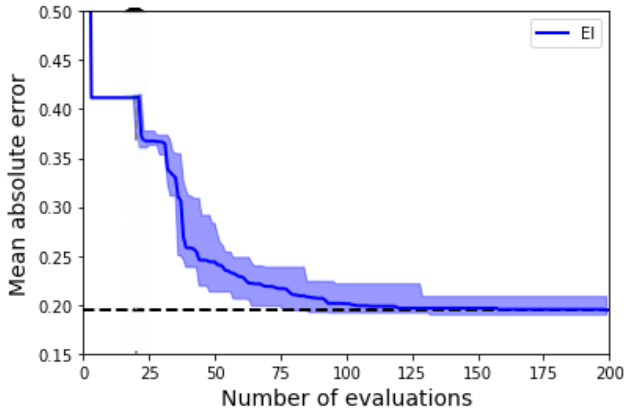


Figure 4. Performance of the Expected Improvement (EI) Bayesian Optimisation algorithm on a neural network emulator of GALFORM. The solid blue line shows the median over 30 separate runs. The shaded region shows the minimum to maximum range for these runs. The dashed horizontal line shows the global minimum found using MCMC, with 20 chains of 10 000 steps each.

simply use this emulator to perform the model calibration instead of using Bayesian optimisation. The observables considered here are more complicated for the neural network to learn than those used to calibrate the model in Elliott et al.; in particular, the redshift distribution of SMGs is hard to reproduce accurately, without running many more full GALFORM evaluations. Hence, if we used the trained neural network to give a set of parameters deemed to be the best fitting ones, these may not lead to the full GALFORM model reproducing the calibration data closely. Also, as commented above, because we are predicting statistics that are computed from the model output at many redshifts, each model evaluation has a higher computational overhead than in Elliott et al. Hence, whilst the emulator is too approximate to use to calibrate GALFORM, it has similar properties to the model and serves to let us test the optimisation process.

Here, we use a neural network model from the TensorFlow software library (Abadi et al. 2015). From our full GALFORM runs, we have sets of parameters $\mathbf{x}$, and an associated output $\mathbf{y}$, which corresponds to the GALFORM predictions for the datasets we are considering (for example, the values of the K -band luminosity function in different magnitude bins). We then calculate the error or distance between the GALFORM predictions and the calibration datasets as given by the mean absolute error, MAE, which for a single dataset is

$$\text{MAE} = \frac{1}{n} \sum_i |y_i - y_{\text{obs},i}|, \quad (20)$$

where y_{obs} corresponds to the observational calibration dataset which has n bins. Importantly, we linearly rescale each dataset so that y_{obs} lies within the range $[0,1]$, so that we can combine the errors defined in this way for different datasets in a consistent way into a single distance value or metric. In the case of the K -band LF and the SMG number counts it is the logarithm of these quantities that is rescaled. For the SMG redshift distribution, we normalise both predicted and observed distributions over the same redshift range (here taken to be $z > 0.8$). We then calculate the total MAE to be minimised as a weighted mean of the MAEs for the individual observational datasets. This allows us to give more weight to some datasets than others, if desired.

We train the neural network to predict the error between the GALFORM predictions and the observational data, given the parameter values $\mathbf{x}$ so that we have a fast-to-evaluate but approximate emulator of GALFORM on which to test our Bayesian optimisation strategies.

By building an emulator which we can evaluate quickly, we allow ourselves to find the approximate global minimum error in a short amount of time using MCMC. Of course, the emulated model will not correspond exactly to running the full GALFORM model, but it will be of similar complexity and so allows us to assess the optimisation methods, ready to apply to the full GALFORM model.

Having trained our neural network emulator to produce an approximate error between the GALFORM predictions for the K -band LF, the SMG redshift distribution and number counts, and the corresponding observational datasets, given a set of input parameters, we first find an approximate global optimum using an extensive MCMC search. We use a simple implementation of the Metropolis-Hastings algorithm (e.g. Robert 2015), using 20 chains, each 10 000 steps in length and starting from a randomly chosen location in the parameter space, to locate the set of parameters which returns the lowest approximate error for the neural network emulator, as judged by the MAE metric.

Next, we test our optimisation strategies to see if we can find a set of best-fitting parameters that gives a comparable error or MAE to that obtained with MCMC. To do this, we initially perform a pseudo-random sampling of the parameter space using Sobol sampling (e.g. see Oleśkiewicz & Baugh 2020), drawing 2 samples per dimension of the parameter space to be searched (in our case 30 initial evaluations for a 15 dimension parameter space). We evaluate the neural network emulator at these points to return approximate errors, fit the GP to the parameter-error pairs, and begin the expected improvement search to decide where to evaluate the neural network emulator next in the parameter space.

Fig. 4 shows the result of applying the EI Bayesian optimisation to the neural network emulator of GALFORM. The y-axis gives the MAE returned by the neural network, with the x-axis showing the total number of evaluations of the emulator carried out up to that point. The first 30 runs are chosen through pseudo-random Sobol sampling. The Bayesian optimisation starts after the 30th evaluation. The horizontal dashed line corresponds to the minimum error found using the more exhaustive MCMC search of the parameter space with the emulator, which is the target MAE for this exercise. We ran many trials of the optimisation, as indicated by the shading, each time with a different random seed. We find that the EI algorithm can locate an error close to the approximate global minimum within just 100 to 150 full model evaluations, with some runs converging on the minimum value in as few as 75 evaluations. We see very little improvement beyond 150 evaluations of the emulator. Thus, using Bayesian optimisation allows us to find a best-fitting model with roughly one thousand times fewer evaluations than MCMC (recall that the MCMC approach used 200 000 emulator evaluations).

Although the emulator is an approximation of GALFORM, it still represents a complex 15-dimensional function with properties similar to the full model, and the encouraging convergence of the EI algorithm allows us to apply the methodology to GALFORM with confidence.

3.7 Applying Bayesian optimisation to GALFORM

To apply Bayesian optimisation to GALFORM, we first evaluate the model n times at points chosen in the parameter space using Sobol sampling (see Oleśkiewicz & Baugh 2020), where n is equal to twice the number of parameters we are investigating. In our case, since we are varying 15 model parameters, we draw an initial sample of 30 runs across the parameter space. Once these 30 GALFORM runs are in place we begin the optimisation process by applying the EI algorithm as described above to decide the location in parameter space at which to run new model calculations. We look to the experiment in the last subsection for guidance as to when to stop the optimisation, and

allow the runs to progress until at least 150 full model evaluations have been carried out. Once 150 runs have been completed, we decide to stop the optimisation if the MAE has not improved significantly compared to the MAE obtained from the best evaluation over the previous 25 steps.

4 RESULTS

Here we present the main results of calibrating GALFORM using Bayesian optimisation, applied to different combinations of calibration datasets (the local K -band LF, the number counts of SMGs and the redshift distribution of bright SMGs). Throughout we consider two model variants, with the main difference being whether we fix the stellar IMF in bursts of star formation to be a solar neighbourhood IMF (resulting in a 14 dimensional parameter space, referred to as the universal IMF variant) or treat the power law slope of the IMF in bursts as a model parameter (giving a 15 dimension parameter space, called the dual IMF model). In § 4.1, we present the best-fitting models for these two variants, treating all the calibration datasets equally (the corresponding parameter values are listed in Table 2). In § 4.2 we show how the model fits change if more weight is placed on reproducing the local calibration data. § 4.3 presents model predictions for observational datasets that were not used in the parameter calibration.

4.1 Calibrations

The results of the model calibration are shown in Fig 5 for the case in which each of the three datasets is given equal weight in the MAE metric. The associated parameter values are listed in Table 2. (Note these plots also show a special case in which the local K -band luminosity function is given extra weight; this case is discussed in § 4.2.) The left panel of Fig. 5 shows the GALFORM predictions of the SMG redshift distribution, the middle panel shows the low redshift K -band LF, and the right panel the SMG number counts. We also include, for reference, the predictions from the Baugh et al. (2019) model (which assumes an IMF in starbursts with slope $x = 1$), which are shown by the black dashed lines in each panel. This model has, in general, a lower median redshift for SMGs than the best-fitting models we find here, as it was calibrated to the redshift distribution from Wardlow et al. (2011), before the redshift distribution from Dudzeviciute et al. (2020) became available. The Baugh et al. model is a recalibration of the model from Lacey et al. (2016) following its implementation in the P-Millennium N-body simulation. This simulation has updated cosmological parameters and a superior mass resolution compared with the N-body simulation used by Lacey et al. However, the recalibration carried out by Baugh et al. focussed on the local K -band luminosity function, but not the number counts or redshift distribution of SMGs. Furthermore, the recalibration carried out by Baugh et al. was essentially a perturbation of the Lacey et al. fit, considering a much smaller parameter space of only three parameters, without a framework for an extensive search of the space.

We find that the model with a universal Chabrier IMF and an equal weighting of the calibration datasets is unable to match the local K -band LF, the SMG redshift distribution, and the SMG number counts *at the same time*. When the calibration data does not include SMG observations, the universal IMF variant can produce a good match to the $z = 0$ K -band LF (Elliott et al. 2021). However, when the calibration datasets include SMG observations, this variant returns a poor match to the observed K -band LF, with large excesses at the faint- and bright-ends. At $M_K - 5 \log h = -23$, this model predicts

ten times more galaxies than are observed: this difference is many times greater than the uncertainty in the observational estimate. At the faint end, the model overpredicts the number of galaxies by at least a factor of three. Interestingly, this model prefers very low values for both f_{burst} and F_{stab} , as can be seen in Table 2. The parameter f_{burst} is defined as the mass ratio (accreted satellite galaxy over central galaxy) threshold for a burst of star formation to occur following a merger (with the universal IMF model preferring a value of $f_{\text{burst}} \approx 0.05$), whereas F_{stab} sets the threshold for a burst of star formation caused by the galactic disk becoming dynamically unstable to bar formation. The preferred value of $F_{\text{stab}} = 0.53$ corresponds to a model in which there are no disk instabilities (in the formulation by Efstathiou et al. 1982, the ratio on the left-hand side of the expression in Eqn. 2 is equal to 0.61 for a self-gravitating disk; in the model only disks for which Eqn. 2 exceeds this value but is less than the adopted value of F_{stab} are allowed to become unstable and experience bursts). This combination of parameters allows the model galaxies to retain larger reservoirs of gas, which are used up in starbursts following mergers rather than disk instabilities.

Though both models prefer a higher value of v_{SF} , the parameter which controls the rate of quiescent star formation in disks, than is suggested by observations of galactic disks in the local Universe, the value of this parameter for the universal IMF model is significantly higher ($v_{\text{SF}} = 3.48 \text{ Gyr}^{-1}$) than the value preferred in the other two models listed in Table 2. This value is about a factor of 8 higher than suggested by local measurements ($v_{\text{SF}} = 0.43 \text{ Gyr}^{-1}$, with a $1 - \sigma$ range of $0.25 - 0.74 \text{ Gyr}^{-1}$ Bigiel et al. 2011). This combination of parameters, with disk instabilities in effect turned off, a high rate of minor-merger driven bursts, and much stronger quiescent star formation rates, allows the model to generate the star formation necessary to match the SMG redshift distribution and counts. However, this behaviour means that the low redshift K -band LF is not matched adequately, since it leads to an excess of large, bright disk galaxies which have not undergone a recent merger, and which now make up the bright end of the LF, and which have lower mass SMBH, due to the lack of SMBH growth through disk instabilities.

For the dual IMF variant, the added flexibility of varying the IMF in starbursts allows the model to match the local K -band LF, as well as producing a realistic SMG redshift distribution and number counts. In this model, we see more typical values for the disk instability parameter ($f_{\text{stab}} = 0.75$), and a smaller number of bursts triggered by minor mergers due to a higher value of $f_{\text{burst}} = 0.17$. The model prefers an IMF slope parameter, $x = 0.7$, which is *more* top-heavy than that assumed in Lacey et al. (2016), who adopted $x = 1$ (though those papers were considering a larger number of calibration datasets), but is less top-heavy than assumed in early models (Baugh et al. 2005).

The optimisation of the two GALFORM variants is shown in Fig. 6, in which we plot the MAE metric as a function of the number of full runs of the model carried out. A solution is found for the universal IMF variant after 90 runs of the full GALFORM model, after which there is no change in the value of the MAE. In the case of the dual IMF variant, 140 model runs are needed to find a best-fitting model. The minimum MAE for the dual IMF variant is almost a factor of two smaller than that for the universal IMF model, confirming that this model gives a better reproduction of the calibration data.

4.2 Enforcing low-redshift agreement

When optimising the models in the previous subsection the three calibration datasets were weighted equally in the MAE metric. The variant with a dual IMF matches all three calibration datasets rea-

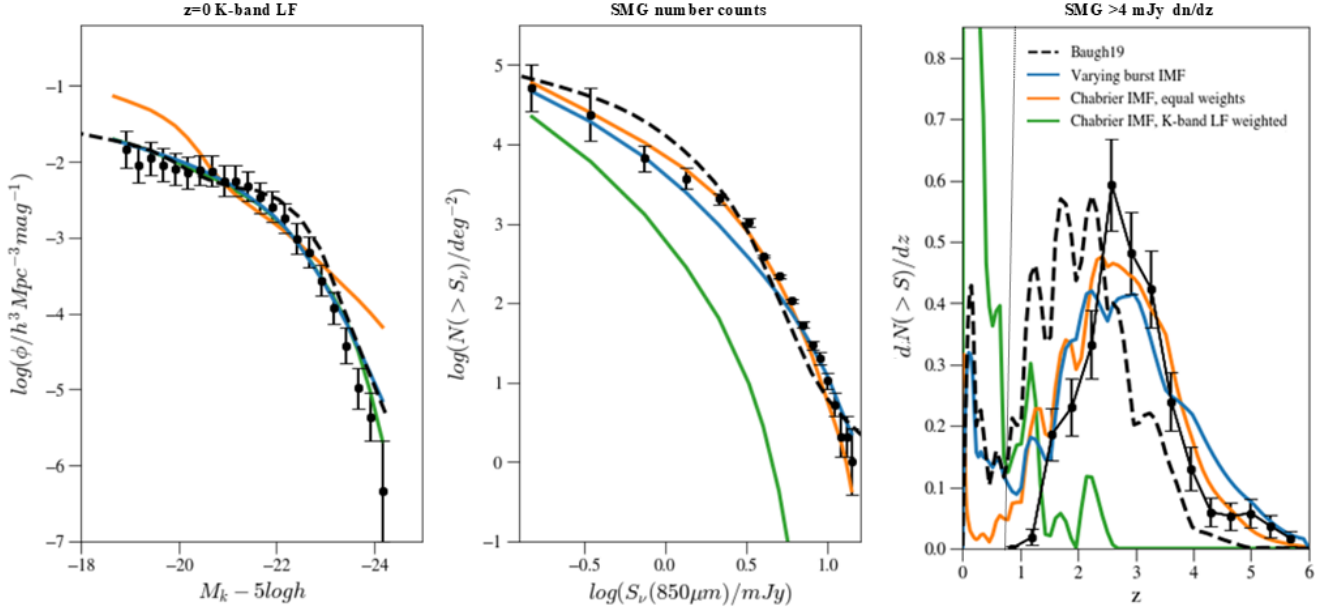


Figure 5. A comparison of the model predictions with the three calibration datasets under consideration (the parameters of these models are given in Table 2). Left: The $z = 0$ K -band LF. Center: the SMG number counts at $870\ \mu\text{m}$. Right: the normalized SMG redshift distribution for $S_{870} > 4\ \text{mJy}$. The spikes in the model predictions for the redshift distribution are artefacts due to the limited number of halos simulated. In each case, the black points with error bars show the observational data. For the SMG redshift distribution, we calibrate to data from [Dudzeviciute et al. \(2020\)](#), using $z > 0.8$ as indicated by the vertical dotted grey line. For the local K -band LF, we calibrate to data from [Kochanek et al. \(2001\)](#), and for the SMG number counts, we calibrate to data from [Stach et al. \(2018\)](#) at the bright end, and [Chen et al. \(2013\)](#) at the faint end. The orange solid curves show the model which assumes a universal Chabrier IMF in all modes of star formation. The green lines show the predictions from a model that also adopts a universal Chabrier IMF, but which is calibrated to give an improved fit to the low-redshift K -band LF by increasing the weight given to this dataset in the parameter optimisation. The blue lines show a model in which the IMF slope in bursts is allowed to vary according to $dn/d\ln m \propto m^{-x}$, where x is an adjustable parameter. For reference, the black dashed line shows the GALFORM model from [Baugh et al. \(2019\)](#); this model was calibrated using an earlier measurement of the SMG redshift distribution from [Wardlow et al. \(2011\)](#), which has a lower median redshift than the [Dudzeviciute et al. \(2020\)](#) data.

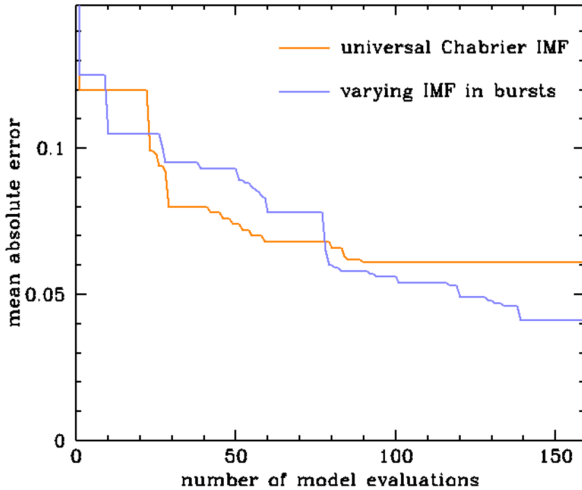


Figure 6. The minimum mean absolute error, MAE, of the GALFORM predictions as a function of the number of full model evaluations carried out, for the three calibration datasets: the $z = 0$ K -band LF, the SMG number counts, and the SMG redshift distribution. The blue line shows the universal IMF model and the orange line shows the variable IMF model, in which the slope of the IMF in bursts is also a parameter; this variant gives a better (smaller) MAE. The optimisation is terminated after 165 model runs are reached and there is no significant improvement in the MAE over the preceding 25 runs.

sonably well (blue curves in Fig. 5). However, the calibration of the simplest variant with a universal Chabrier IMF returned a model that matched the counts and redshift distribution of SMGS but gave a poor reproduction of the local K -band LF (orange curves in Fig. 5).

It is interesting to know how the predictions for the SMG number counts and redshift distribution are degraded if we enforce better agreement with the $z = 0$ K -band LF. We can achieve this by recalibrating the universal Chabrier IMF variant with a triple weighting applied to the low-redshift K -band LF in the MAE, and single weightings for each of the SMG calibration datasets.

Running the optimisation with a higher weighting for the $z = 0$ K -band LF, we find that the model no longer produces a realistic SMG redshift distribution or counts while simultaneously matching the K -band LF constraint. The green curves in Fig. 5 show this case. Despite a much-improved match to the K -band LF, as expected, the predictions for the number counts are too low at all fluxes, with the deficit ranging from a factor of three at the faint end to more than a factor of a hundred at $3\ \text{mJy}$ and brighter; the predicted median redshift is much lower than the observed one. In this weighting case, the best-fitting model parameters are more similar to the model which assumes a different IMF slope in bursts, with $F_{\text{stab}} \approx 0.7$ for example.

4.3 Further predictions at low redshift

Having calibrated the model to the $z = 0$ K -band LF, the SMG numbers counts, and the SMG redshift distribution, we now explore the predictions for other low-redshift properties of the galaxy popula-

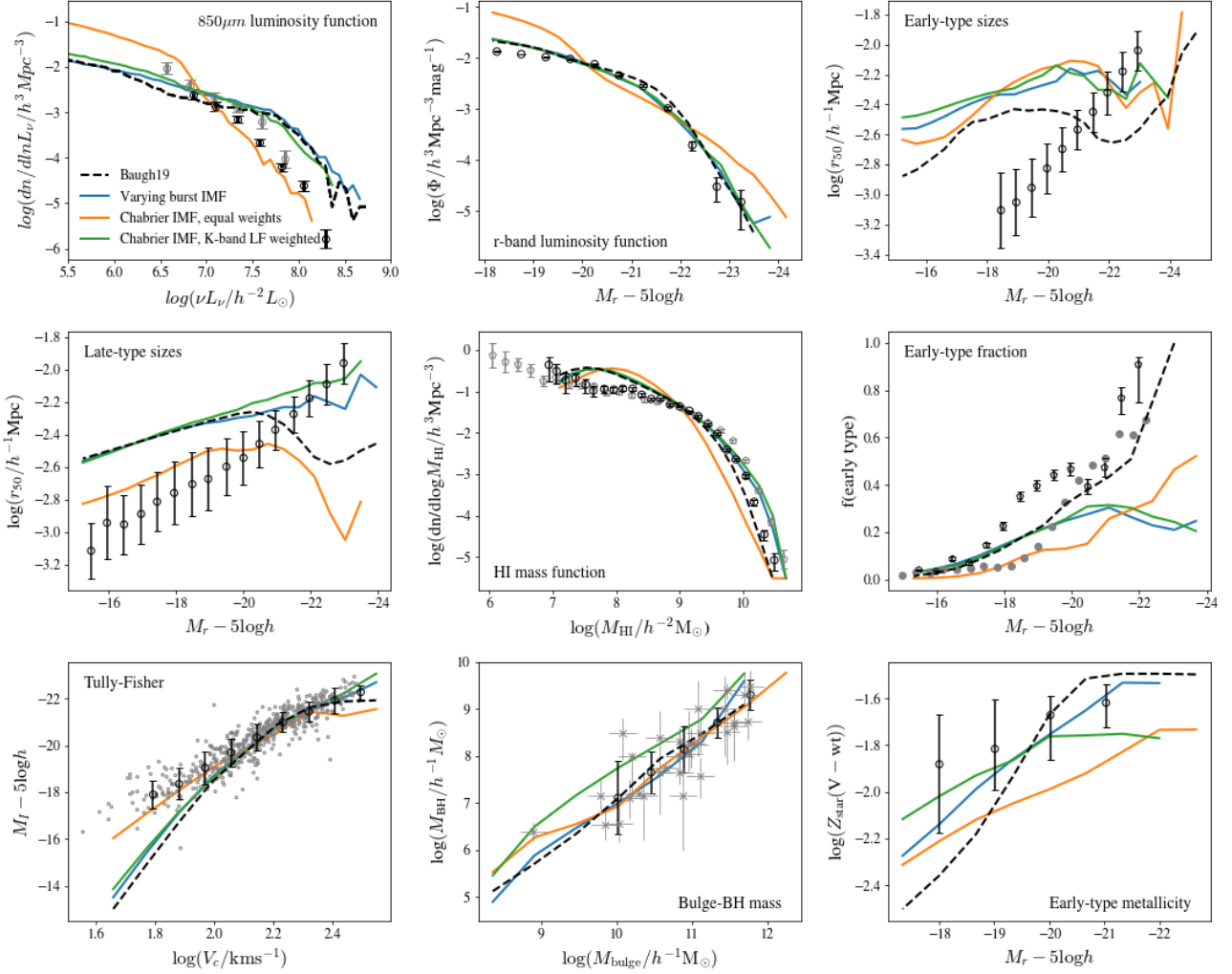


Figure 7. Low redshift predictions for the three GALFORM calibrations: the model with a variable IMF slope in bursts (blue), the equal-weighted calibration assuming a universal Chabrier IMF (orange), and the calibration with a higher weight for the $z = 0$ K -band LF, again assuming a universal Chabrier IMF (green). The black dashed lines show the model calibration from [Baugh et al. \(2019\)](#). $850\,\mu\text{m}$ LF: data from [Vlahakis et al. \(2005\)](#) (grey circles) and [Dunne et al. \(2000\)](#) (black circles). r -band LF: data from [Driver et al. \(2012\)](#). Early- and late-type sizes: data from [Shen et al. \(2003\)](#). HI mass function: data from [Zwaan et al. \(2005\)](#) (black circles) and [Martin et al. \(2010\)](#) (grey circles). Early-type fraction: data derived from [Moffett et al. \(2016\)](#) (black symbols; A. Moffett, private communication), and from [González et al. \(2009\)](#) (grey symbols). For the Tully-Fisher relation, we compare to a sample of Sb-Sd galaxies from the [Mathewson et al. \(1992\)](#) catalogue selected by [de Jong & Lacey \(2000\)](#) (grey points show individual galaxies, black points with bars show the binned median and 10-90 percentile range). For the bulge-BH mass relation, we compare to data from [Haring & Rix \(2004\)](#), and for the early-type metallicity, we compare to data from [Smith et al. \(2009\)](#). Note that the models were not calibrated to the datasets plotted here.

tion. In particular, we explore the predictions for the $z = 0$ $850\,\mu\text{m}$ and r -band LFs, the early- and late-type galaxy sizes, the HI mass function, the dependence of the early-type fraction on r -band magnitude, the I -band Tully-Fisher relation, the black hole (BH)-bulge mass relation, and the stellar metallicities of early-type galaxies.

The model predictions for the above datasets are shown in Fig. 7. The blue solid lines show the predictions for the dual IMF GALFORM variant, which allowed a variable IMF slope in bursts, whereas the orange and green lines show the predictions for variants which assumed a universal Chabrier IMF. The orange line represents model predictions with equal weightings applied to the three calibration datasets, whereas the model shown by the green line used a higher weighting for the low-redshift K -band LF during calibration. Interestingly, we see that the $z = 0$ predictions for the dual IMF model and the low-redshift K -band LF weighted models are very similar.

As discussed in [Elliott et al. \(2021\)](#), the low-redshift K -band LF is most sensitive to the choice of supernova feedback parameters ($V_{\text{SN,disk}}, V_{\text{SN,burst}}$), which are similar in the case of these two calibrations. These parameters also dominate many of the other properties, leading to very similar predictions for the $z = 0$ properties. For example, the Tully-Fisher relation, and late-type sizes (again as shown in [Elliott et al. 2021](#)) are dominated by the effects of the supernova feedback parameters, at least at lower luminosities. We see that the blue and green lines are almost identical for these datasets, where the orange line better matches the faint end of both relationships.

All of the models fail to match the early-type fraction brighter than $M_r - 5 \log h = -20$, with the low-redshift weighted Chabrier model, and the dual IMF model predicting a constant or gently decreasing early-type fraction at the brightest magnitudes. The models are therefore not producing enough high-mass early-type galaxies. This

Table 2. The best-fitting parameters for the three optimisation cases. The second column shows the parameter values for the dual IMF variant, which treated the IMF slope in bursts, x , as another parameter. The third column shows the best-fitting parameters for the variant with a universal Chabrier IMF with equal weighting attributed to each dataset, and the model in the fourth column also assumes a universal Chabrier IMF, but with an increased weighting applied to the low-redshift K -band LF. * indicates that this parameter was held fixed. x gives the slope of the IMF above $1M_{\odot}$.

Parameter	variant		
	Dual IMF (IMF slope x in bursts)	Universal IMF	Universal IMF (extra weight to K LF)
F_{stab}	0.746	0.535	0.77
γ_{SN}	3.55	1.48	3.06
α_{cool}	2.99	3.8	2.90
α_{reheat}	1.83	2.14	2.11
$V_{\text{SN, disk}} \text{ (km s}^{-1}\text{)}$	293	774	383
$V_{\text{SN, burst}} \text{ (km s}^{-1}\text{)}$	349	399	194
f_{ellip}	0.383	0.385	0.212
f_{burst}	0.166	0.056	0.261
$\nu_{\text{SF}} \text{ (Gyr)}^{-1}$	1.94	3.48	2.186
f_{SMBH}	0.014	0.037	0.025
$\tau_{\text{burst, min}} \text{ (Gyr)}$	0.145	0.157	0.176
f_{cloud}	0.452	0.404	0.427
$t_{\text{esc}} \text{ (Gyr)}$	0.006	0.006	0.009
β_{burst}	1.53	1.76	1.50
x	0.67	1.35*	1.35*

is due to a more complex interplay of parameters, and our method does not give us the tools to investigate this thoroughly. The Baugh et al. calibration, however, can match both the early-type fraction, as well as the three calibration datasets adequately while assuming a top-heavy IMF in bursts. No calibration assuming a Chabrier IMF was able to simultaneously match the three calibration datasets.

Here, we have chosen to calibrate our model to the SMG redshift distribution at $z > 0.8$, as McAlpine et al. (2019) did when investigating the SMG population in EAGLE (Schaye et al. 2015). The local $850 \mu\text{m}$ LF therefore offers a constraint on this population at low redshift (which, at the luminosities observed, consists mainly of quiescently star-forming galaxies). We find, interestingly, that the equal-weighted Chabrier IMF fit better matches the local $850 \mu\text{m}$ LF, while producing very poor fits to the K - and r -band LFs. On the other hand, the GALFORM variant with a dual IMF over-predicts the bright end of the local $850 \mu\text{m}$ LF.

5 CONCLUSIONS

We have assessed whether the GALFORM galaxy formation model can match observations of sub-millimetre galaxies (SMGs) at $850 \mu\text{m}$ —typically high-redshift galaxies—while simultaneously reproducing properties of local galaxies. In particular, we tested the assertion by Baugh et al. (2005) that properties of both SMG and local galaxy populations can only be matched if the IMF in starbursts is assumed to be top-heavy.

Our analysis contains several advances over previous work. We used Bayesian optimisation to thoroughly search a parameter space with many dimensions (15 dimensions when the IMF slope is allowed to vary in starbursts). This method allows comprehensive parameter space searches without requiring enormous numbers of computa-

tionally expensive GALFORM runs, which are particularly costly when making predictions for number counts and redshift distributions. We formally quantified how well models reproduce observational data using a metric. This replaces the old-fashioned one-parameter-at-a-time variation and “chi-by-eye” judgement used to conclude the need for a top-heavy IMF in bursts in our previous work (Baugh et al. 2005; Lacey et al. 2016). With Bayesian optimisation, we definitively address whether the model can reproduce both SMG and local galaxy populations with a solar neighbourhood IMF.

We used three observational datasets to calibrate model parameters: the $z = 0$ K -band luminosity function, the SMG number counts, and the SMG redshift distribution. We attempted to reproduce these datasets using two model variants: (i) a universal IMF model imposing a solar neighbourhood IMF on all star formation, and (ii) a dual IMF model with a solar neighbourhood IMF for disk star formation and a power-law IMF in bursts, with the slope treated as a model parameter.

Within the GALFORM framework, even when varying more parameters than explored in previous work, the model with a universal Chabrier IMF is unable to produce acceptable matches to the SMG constraints and the local K -band luminosity function simultaneously. This confirms the conclusion of Baugh et al. (2005) and subsequent studies (e.g., Lacey et al. 2008; Lacey et al. 2016). Here we have demonstrated the need for a top-heavy IMF with a comprehensive, automated parameter space search. We also explored calibration to the most recent SMG redshift distribution data (Dudzeviciute et al. 2020), which has a higher median redshift than previous datasets (e.g., Wardlow et al. 2011). When assuming a universal Chabrier IMF and enforcing a good fit to the low-redshift K -band luminosity function, we find similar results to those obtained with EAGLE (McAlpine et al. 2019): the best-fitting GALFORM model severely underpredicts SMG number counts and produces almost no bright sources at high redshift ($z > 1$).

Bayesian optimisation found good fits to the data using relatively few GALFORM evaluations, typically 100–150. In Elliott et al. (2021), around 1000 GALFORM runs were used in an emulator-based approach (though calibrating to more datasets). Earlier works, such as Henriques et al. (2020), used even more full model runs; our MCMC test to assess Bayesian optimisation used around 200 000 emulator evaluations. We conclude that Bayesian optimisation is significantly faster than other methods explored for calibrating semi-analytic galaxy formation models, converging in fewer than 200 full model evaluations.

The downside of Bayesian optimisation is reduced information about model performance across the parameter space and parameter interactions. In previous emulator-based works (e.g., Henriques et al. 2009; Bower et al. 2010; Vernon et al. 2010; Elliott et al. 2021), emulators built from large numbers of runs were assessed for accuracy across the entire parameter space. This allows comprehensive parameter exploration, and calibration can be easily extended to include additional datasets. In Elliott et al. (2021), we explored implications of calibrating to diverse dataset combinations and tensions between them with only the overhead cost of the initial 1000 runs. We could also run full MCMC explorations of parameter ranges producing acceptable matches to observational datasets. Here, however, we do not obtain a complete sense of parameter ranges matching observational datasets, as the routine searches for the global optimum rather than inferring posterior parameter distributions.

Nevertheless, we confidently conclude that we could not find any parameter set for a model with a universal Chabrier IMF that simultaneously matched the local K -band luminosity function and the SMG redshift distribution and number counts. This aligns with conclusions from manual searches (Baugh et al. 2005; Lacey et al. 2016). Within

GALFORM, including the flexibility of a top-heavy IMF in bursts allows simultaneous fits to these datasets. The best-fitting dual IMF model prefers an IMF slope parameter $x = 0.7$, more top-heavy than assumed in Lacey et al. (2016) (who adopted $x = 1$, though considering more calibration datasets) but less top-heavy than early models (Baugh et al. 2005).

We also examined predictions for additional low-redshift galaxy properties not used in calibration, including the local 850 μm luminosity function, the r -band luminosity function, early- and late-type galaxy sizes, the HI mass function, the early-type fraction versus luminosity, the Tully-Fisher relation, the black hole–bulge mass relation, and early-type stellar metallicities. The dual IMF model and the universal IMF model with enhanced K -band weighting produce similar predictions for most of these properties, as they prefer similar supernova feedback parameters that dominate these observables. However, all models fail to match the early-type fraction at the brightest magnitudes, suggesting areas for future model improvement.

In future work, we will extend Bayesian optimisation to recover parameter ranges for acceptable models rather than single best-fitting models, and include more observational datasets in the calibration process to improve model constraints and uncover tensions between datasets. Our work demonstrates that comprehensive, automated parameter space searches are now feasible for semi-analytic galaxy formation models, providing robust tests of fundamental assumptions such as the universality of the stellar initial mass function.

ACKNOWLEDGEMENTS

EJE was supported by a PhD Studentship from the Durham Centre for Doctoral Training in Data Intensive Science, funded by the Science and Technology Facilities Council (STFC, ST/P006744/1) and Durham University. CMB and CGL acknowledge support from STFC (ST/T000244/1, ST/X001075/1). We used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

DATA AVAILABILITY

The observational datasets, GALFORM outputs, parameter values, and best-fitting parameter values underlying this article may be shared on reasonable request to the corresponding author.

REFERENCES

Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, <https://www.tensorflow.org/>
 Bastian N., Covey K. R., Meyer M. R., 2010, *ARA&A*, **48**, 339
 Baugh C. M., 2006, *Reports on Progress in Physics*, **69**, 3101
 Baugh C. M., Lacey C. G., Frenk C. S., Granato G. L., Silva L., Bressan A., Benson A. J., Cole S., 2005, *MNRAS*, **356**, 1191
 Baugh C. M., et al., 2019, *MNRAS*, **483**, 4922
 Benson A., 2018, *Research Notes of the AAS*, **2**, 188
 Benson A. J., Bower R., 2010a, *MNRAS*, **405**, 1573
 Benson A. J., Bower R., 2010b, *MNRAS*, **405**, 1573
 Bigiel F., et al., 2011, *ApJLetters*, **730**, 1
 Blain A. W., Chapman S. C., Smail I., Ivison R., 2004, *ApJ*, **611**, 725
 Blitz L., Rosolowsky E., 2006, *ApJ*, **650**, 933

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, *MNRAS*, **370**, 645
 Bower R. G., Vernon I., Goldstein M., Benson A. J., Lacey C. G., Baugh C. M., Cole S., Frenk C. S., 2010, *MNRAS*, **407**, 2017
 Campbell D. J. R., et al., 2015, *MNRAS*, **452**, 852
 Casey C. M., Narayanan D., Cooray A., 2014, *Phys. Rep.*, **541**, 45
 Chabrier G., 2003, *PASP*, **115**, 763
 Chapman S. C., Blain A. W., Smail I., Ivison R. J., 2005, *ApJ*, **622**, 772
 Chen C. C., Cowie L. L., Barger A. J., Casey C. M., Lee N., Sanders D. B., Wang W. H., Williams J. P., 2013, *ApJ*, **762**, 1
 Christodoulou D. M., Shlosman I., Tohline J. E., 1995, *ApJ*, **443**, 551
 Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, *MNRAS*, **319**, 168
 Conroy C., 2013, *ARA&A*, **51**, 393
 Conroy C., van Dokkum P. G., 2012, *ApJ*, **760**, 71
 Conroy C., Gunn J. E., White M., 2009, *ApJ*, **699**, 486
 Cowley W. I., Lacey C. G., Baugh C. M., Cole S., Frenk C. S., Lagos C. D. P., 2019, *MNRAS*, **487**, 3082
 Crain R. A., et al., 2015, *MNRAS*, **450**, 1937
 Croton D. J., et al., 2006, *MNRAS*, **365**, 11
 Da Cunha E., et al., 2015, *ApJ*, **806**
 Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, *MNRAS*, **486**, 2827
 Driver S. P., et al., 2012, *MNRAS*, **427**, 3244
 Dudzeviciute U., et al., 2020, *MNRAS*, **494**, 3828
 Dunne L., Eales S., Edmunds M., Ivison R., Alexander P., Clements D. L., 2000, *MNRAS*, **315**, 115
 Efsthathiou G., Lake G., Negroponte J., 1982, *MNRAS*, **199**, 1069
 Elliott E. J., Baugh C. M., Lacey C. G., 2021, *MNRAS*, **506**, 4011
 Farr W. M., Mandel I., 2018, *Science*, **361**, aat6506
 Ferrara A., Bianchi S., Cimatti A., Giovanardi C., 1999, *ApJS*, **123**, 437
 Frazier P. I., 2018, *arXiv e-prints*, p. [arXiv:1807.02811](https://arxiv.org/abs/1807.02811)
 Garnett R., 2023, Bayesian Optimization. Cambridge University Press
 Geach J. E., et al., 2017, *MNRAS*, **465**, 1789
 Goldstein M., Wooff D., 2007, Bayes linear statistics : theory and methods. Wiley series in probability and statistics, John Wiley, Chichester, England
 Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, *MNRAS*, **439**, 264
 González J. E., Lacey C. G., Baugh C. M., Frenk C. S., Benson A. J., 2009, *MNRAS*, **397**, 1254
 Granato G. L., Lacey C. G., Silva L., Bressan A., Baugh C. M., Cole S., Frenk C. S., 2000, *ApJ*, **542**, 710
 Griffin A. J., Lacey C. G., Gonzalez-Perez V., Lagos C. d. P., Baugh C. M., Fanidakis N., 2019, *MNRAS*, **487**, 198
 Gunawardhana M. L. P., et al., 2011, *MNRAS*, **415**, 1647
 Hamann J., Wons J., 2022, *J. Cosmology Astropart. Phys.*, **2022**, 036
 Häring N., Rix H.-W., 2004, *ApJ*, **604**, L89
 Hayward C. C., Kere D., Jonsson P., Narayanan D., Cox T. J., Hernquist L., 2011, *ApJ*, **743**
 Hayward C. C., et al., 2021, *MNRAS*, **502**, 2922
 Henriques B. M., Thomas P. A., Oliver S., Roseboom I., 2009, *MNRAS*, **396**, 535
 Henriques B. M. B., Yates R. M., Fu J., Guo Q., Kauffmann G., Srisawat C., Thomas P. A., White S. D. M., 2020, *MNRAS*, **491**, 5795
 Hickox R. C., et al., 2012, *MNRAS*, **421**, 284
 Hughes D. H., et al., 1998, *Nature*, **394**, 241
 Kampakoglou M., Trotta R., Silk J., 2008, *MNRAS*, **384**, 1414
 Kennicutt R. C. J., 1983, *ApJ*, **272**, 54
 Kochanek C. S., et al., 2001, *ApJ*, **20**, 566
 Lacey C. G., Baugh C. M., Frenk C. S., Silva L., Granato G. L., Bressan A., 2008, *MNRAS*, **385**, 1155
 Lacey C. G., et al., 2016, *MNRAS*, **462**, 3854
 Lagos C. D. P., Cora S. A., Padilla N. D., 2008, *MNRAS*, **388**, 587
 Lagos C. d. P., Lacey C. G., Baugh C. M., Bower R. G., Benson A. J., 2011, *MNRAS*, **416**, 1566
 Leclercq F., 2018, *Phys. Rev. D*, **98**, 063511
 Lovell C. C., Geach J. E., Davé R., Narayanan D., Li Q., 2021, *MNRAS*, **502**, 772
 Lu Y., Mo H. J., Weinberg M. D., Katz N., 2011, *MNRAS*, **416**, 1949

- Lu Y., Mo H. J., Katz N., Weinberg M. D., 2012, [MNRAS](#), 421, 1779
- Lu Y., Mo H. J., Lu Z., Katz N., Weinberg M. D., 2014, [MNRAS](#), 443, 1252
- Madar M. S., Baugh C. M., Shi D., 2024, [MNRAS](#), 535, 3324
- Martin A. M., Papastergis E., Giovanelli R., Haynes M. P., Springob C. M., Stierwalt S., 2010, [ApJ](#), 723, 1359
- Martindale H., Thomas P. A., Henriques B. M., Loveday J., 2017, [MNRAS](#), 472, 1981
- Mathewson D. S., Ford V. L., Buchhorn M., 1992, [ApJS](#), 81, 413
- McAlpine S., et al., 2019, [MNRAS](#), 488, 2440
- Moffett A. J., et al., 2016, [MNRAS](#), 457, 1308
- Nelson D., et al., 2015, [Astronomy and Computing](#), 13, 12
- Oleśkiewicz P., Baugh C. M., 2020, [MNRAS](#), 493, 1827
- Pillepich A., et al., 2018, [MNRAS](#), 473, 4077
- Rasmussen C. E., Williams C. K. I., 2006, Gaussian processes for machine learning. Adaptive computation and machine learning, MIT Press
- Robert C. P., 2015, [arXiv e-prints](#), p. [arXiv:1504.01896](#)
- Robitaille T. P., 2011, [A&A](#), 536, A79
- Rodrigues L. F. S., Vernon I., Bower R. G., 2017, [MNRAS](#), 466, 2418
- Rogers K. K., Peiris H. V., Pontzen A., Bird S., Verde L., Font-Ribera A., 2019, [J. Cosmology Astropart. Phys.](#), 2019, 031
- Romano D., Matteucci F., Zhang Z. Y., Papadopoulos P. P., Ivison R. J., 2017, [MNRAS](#), 470, 401
- Ruiz A. N., et al., 2015, [ApJ](#), 801
- Schaye J., et al., 2015, [MNRAS](#), 446, 521
- Schneider F. R. N., et al., 2018, [Science](#), 359, 69
- Shen S., Mo H. J., White S. D., Blanton M. R., Kauffmann G., Voges W., Brinkmann J., Csabai I., 2003, [MNRAS](#), 343, 978
- Silva L., Granato G. L., Bressan A., Danese L., 1998, [ApJ](#), 509, 103
- Simha V., Cole S., 2017, [MNRAS](#), 472, 1392
- Smail I., Ivison R. J., Blain A. W., 1997, [ApJ](#), 490, L5
- Smith R. J., 2020, [ARA&A](#), 58, 577
- Smith R. J., Lucey J. R., Hudson M. J., 2009, [MNRAS](#), 400, 1690
- Smith D. J., Hayward C. C., Jarvis M. J., Simpson C., 2017, [MNRAS](#), 471, 2453
- Somerville R. S., Davé R., 2015, [ARA&A](#), 53, 51
- Somerville R. S., Gilmore R. C., Primack J. R., Domínguez A., 2012, [MNRAS](#), 423, 1992
- Stach S. M., et al., 2018, [ApJ](#), 860, 161
- Stein M., 1987, [Technometrics](#), 29, 143
- Swinbank A. M., Smail I., Chapman S. C., Blain A. W., Ivison R. J., Keel W. C., 2004, [ApJ](#), 617, 64
- Takhtaganov T., Lukić Z., Müller J., Morozov D., 2021, [ApJ](#), 906, 74
- Vernon I., Goldstein M., Bower R. G., 2010, Bayesian analysis., 05, 619
- Vlahakis C., Dunne L., Eales S., 2005, [MNRAS](#), 364, 1253
- Wardlow J. L., et al., 2011, [MNRAS](#), 415, 1479
- Weidner C., Ferreras I., Vazdekis A., La Barbera F., 2013, [MNRAS](#), 435, 2274
- Zhang Z.-Y., Romano D., Ivison R. J., Papadopoulos P. P., Matteucci F., 2018, [Nature](#), 558, 260
- Zwaan M. A., Meyer M. J., Staveley-Smith L., Webster R. L., 2005, [MNRAS: Letters](#), 359, 1
- de Jong R. S., Lacey C., 2000, [ApJ](#), 545, 781
- van der Velden E., Duffy A. R., Croton D., Mutch S. J., Sinha M., 2019, [ApJS](#), 242, 22
- van der Velden E., Duffy A. R., Croton D., Mutch S. J., 2021, [ApJS](#), 253, 50

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