Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **518,** 2123–2163 (2023) https://doi.org/10.1093/mnras/stac2633
Advance Access publication 2022 October 16

TRINITY I: self-consistently modelling the dark matter halo–galaxy–supermassive black hole connection from $z = 0$–10

Haowen Zhang (张昊文)[1★] Peter Behroozi[1,2] Marta Volonteri,[3] Joseph Silk[3,4,5] Xiaohui Fan,[1] Philip F. Hopkins[6] Jinyi Yang (杨锦怡)[1†] and James Aird[7,8]

[1]*Steward Observatory, University of Arizona, 933 N Cherry Ave., Tucson, AZ 85721, USA*
[2]*Division of Science, National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan*
[3]*Institut d'Astrophysique de Paris (UMR 7095: CNRS & Sorbonne Universite), 98 bis Bd. Arago, F-75014 Paris, France*
[4]*Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA*
[5]*BIPAC, Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK*
[6]*TAPIR, California Institute of Technology, Mailcode 350-17, Pasadena, CA 91125, USA*
[7]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK*
[8]*Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK*

Accepted 2022 September 9. Received 2022 September 8; in original form 2021 May 21

ABSTRACT

We present TRINITY, a flexible empirical model that self-consistently infers the statistical connection between dark matter haloes, galaxies, and supermassive black holes (SMBHs). TRINITY is constrained by galaxy observables from $0 < z < 10$ [galaxies' stellar mass functions, specific and cosmic star formation rates (SFRs), quenched fractions, and UV luminosity functions] and SMBH observables from $0 < z < 6.5$ (quasar luminosity functions, quasar probability distribution functions, active black hole mass functions, local SMBH mass–bulge mass relations, and the observed SMBH mass distributions of high-redshift bright quasars). The model includes full treatment of observational systematics [e.g. active galactic nucleus (AGN) obscuration and errors in stellar masses]. From these data, TRINITY infers the average SMBH mass, SMBH accretion rate, merger rate, and Eddington ratio distribution as functions of halo mass, galaxy stellar mass, and redshift. Key findings include: (1) the normalization and the slope of the SMBH mass–bulge mass relation increases mildly from $z = 0$ to $z = 10$; (2) The best-fitting AGN radiative+kinetic efficiency is ∼0.05–0.06, but can be in the range ∼0.035–0.07 with alternative input assumptions; (3) AGNs show downsizing, i.e. the Eddington ratios of more massive SMBHs start to decrease earlier than those of lower mass objects; (4) The average ratio between average SMBH accretion rate and SFR is $\sim 10^{-3}$ for low-mass galaxies, which are primarily star-forming. This ratio increases to $\sim 10^{-1}$ for the most massive haloes below $z \sim 1$, where star formation is quenched but SMBHs continue to accrete.

Key words: galaxies: evolution – galaxies: haloes – quasars: sumpermassive black holes.

1 INTRODUCTION

It is widely accepted that supermassive black holes (SMBHs) exist in the centres of most galaxies (Kormendy & Richstone 1995; Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000; Tremaine et al. 2002; Ho 2008; Gültekin et al. 2009; Kormendy & Ho 2013; Heckman & Best 2014). SMBHs are called active galactic nuclei (AGNs) during phases when they are accreting matter and releasing tremendous amounts of energy. With their potential for high-energy output, SMBHs are leading candidates to regulate both the star formation of their host galaxies and their own mass accretion (Silk & Rees 1998; Bower et al. 2006; Somerville et al. 2008; Sijacki et al. 2015). At the same time, galaxies may also influence SMBH growth via the physics of how gas reaches the central SMBH as well as via galaxy mergers. Hence, it is possible for both SMBHs and their host galaxies to influence each others' growth, also known as 'coevolution'. As a result, constraining the interaction between SMBHs and their host galaxies is critical to our understanding of both galaxy and SMBH assembly histories (see e.g. Hopkins et al. 2007b; Ho 2008; Alexander & Hickox 2012; Kormendy & Ho 2013; Heckman & Best 2014; Brandt & Alexander 2015).

The coevolution scenario is consistent with two key observations. First, relatively tight scaling relations (∼0.3 dex scatter) exist between SMBH masses, $M_\bullet$, and host galaxy dynamical properties (e.g. velocity dispersion, σ, or bulge mass, $M_{\rm bulge}$, at $z \sim 0$; see Häring & Rix 2004; Gültekin et al. 2009; Kormendy & Ho 2013; McConnell & Ma 2013; Savorgnan et al. 2016). Second, the cosmic SMBH accretion rate (CBHAR) density tracks the cosmic star formation rate (CSFR) density over $0 < z < 4$, with a roughly constant CBHAR/CSFR ratio between 10^{-4} and 10^{-3} (Merloni, Rudnick & Di Matteo 2004; Silverman et al. 2008; Shankar, Weinberg & Miralda-Escudé 2009; Aird et al. 2010; Delvecchio et al. 2014; Yang et al. 2018). At the same time, other predictions of the coevolution model (e.g. tight galaxy–SMBH property relationships at higher redshifts) have remained more difficult to verify.

In the local Universe, galaxy–SMBH scaling relations (e.g. $M_\bullet$–$M_{\rm bulge}$ or $M_\bullet$–σ) have been measured via high spatial resolution spectroscopy and dynamics modelling (e.g. Magorrian et al.

★ E-mail: hwzhang0595@arizona.edu
† Strittmatter Fellow

1998; Ferrarese & Ford 2005; McConnell & Ma 2013). Total (i.e. active+dormant) SMBH mass functions can be obtained by convolving these scaling relations with the distributions of galaxy properties, such as galaxy bulge mass function or velocity dispersion functions (e.g. Salucci et al. 1999; Marconi et al. 2004). Beyond the local Universe, lower spatial resolution makes it impractical to measure individual SMBH masses in the same way. Hence, SMBH mass measurements at $z > 0$ rely on indirect methods such as reverberation mapping (Blandford & McKee 1982; Peterson 1993) and empirical relations between SMBH mass, spectral line width, and AGN luminosity (i.e. 'virial' estimates; Vestergaard & Peterson 2006). All such indirect methods work only on actively accreting SMBHs, which: (1) imposes a selection bias on the SMBHs included, and (2) makes it difficult to measure host galaxy masses at the same time. As a result, it has been even harder to obtain unbiased measurements of the galaxy–SMBH mass connection beyond $z = 0$.

There has also been great interest in measuring SMBH luminosity distributions, as these carry information about mass accretion rates. At $z > 0$, surveys have been carried out in X-ray, optical, infrared, and radio bands to identify AGNs and study their collective properties (see Hopkins, Richards & Hernquist 2007a, Shen et al. 2020, and references therein). As redshift increases (e.g. at $z \gtrsim 2$), the AGN sample is biased towards brighter and rarer objects, due to the evolution of AGN populations and/or limited instrument capability. None the less, for lower luminosity AGNs, it is often possible to measure both the SMBH luminosity and the mass of the host galaxy (e.g. Bongiorno et al. 2012; Aird, Coil & Georgakakis 2018).

Besides observational efforts, the galaxy–SMBH connection is a key ingredient in galaxy formation theory. Supernova feedback becomes inefficient in massive haloes; hence, to reproduce these haloes' low observed star formation rates (SFRs), AGN feedback is widely implemented in hydrodynamical simulations and semi-analytic models (SAMs) for galaxy evolution (see e.g. Croton et al. 2006; Somerville et al. 2008; Dubois et al. 2012; Schaye et al. 2015; Sijacki et al. 2015; Weinberger et al. 2017). These simulations allow studying the evolution of the galaxy–SMBH connection for individual galaxies. However, numerical simulations must make assumptions about physical mechanisms below their resolution limits, which complicates the interpretation of their results (see e.g. Habouzit et al. 2021).

Empirical models are a complementary tool to study SMBHs. Instead of assuming specific physics, these models use observations to self-consistently and empirically characterize the properties of SMBHs and/or their connection with host galaxies. There are broadly two different categories of empirical models involving SMBHs.

The first group of models solves the continuity equation for the SMBH mass function, linking the mass growth histories of SMBHs to their energy outputs. By comparing the local cosmic BH mass density with the total AGN energy output, these models provide estimates of the average radiative efficiency, duty cycles, and Eddington ratio distributions of AGNs (see e.g. Sołtan 1982; Small & Blandford 1992; Cavaliere & Vittorini 2000; Yu & Tremaine 2002; Steed & Weinberg 2003; Marconi et al. 2004; Yu & Lu 2004; Merloni & Heinz 2008; Shankar et al. 2009; Shankar, Weinberg & Miralda-Escudé 2013; Aversa et al. 2015; Tucci & Volonteri 2017).

The second group of models focuses on the galaxy–SMBH or galaxy–AGN connection (e.g. Conroy & White 2013; Caplar, Lilly & Trakhtenbrot 2015, 2018; Yang et al. 2018; Comparat et al. 2019; Georgakakis et al. 2019; Carraro et al. 2020; Shankar et al. 2020a,b; Allevato et al. 2021). Some of these models jointly infer the galaxy–SMBH mass scaling relation and SMBH accretion rate distributions. Previous models differ in terms of the flexibility in connecting the accretion rate distribution and the galaxy properties, as well as the data sets they try to fit. For example, Veale, White & Conroy (2014) used quasar luminosity functions (QLFs) to constrain several halo–galaxy–SMBH models, e.g. assigning AGN luminosities based on SMBH masses or accretion rates, and assuming lognormal or truncated power-law Eddington ratio distributions. They found that all these models could fit QLFs nearly equally well over $1 < z < 6$. This model degeneracy implies the need for data constraints beyond QLFs to fully characterize the galaxy–SMBH connection.

In this paper, we present TRINITY, an empirical model connecting dark matter haloes, galaxies, and SMBHs from $z = 0$–10; TRINITY extends the empirical DM halo–galaxy model from Behroozi, Wechsler & Conroy (2013). Compared to previous empirical models, TRINITY is constrained by a larger compilation of galaxy and AGN data, including not only QLFs, but also quasar probability distribution functions (QPDFs), active black hole mass functions (ABHMFs), the local bulge mass–SMBH mass relations, the observed SMBH mass distribution of high-redshift bright quasars, galaxy stellar mass functions (SMFs), galaxy UV luminosity functions (UVLFs), galaxy quenched fractions (QFs), galaxy specific star formation rates (SSFRs), and CSFRs. The enormous joint constraining power of this data set allows TRINITY to have both a more flexible parametrization as well as better constraints on the model parameters. In addition, TRINITY features more realistic modelling of AGN observables by including, e.g. SMBH mergers and kinetic AGN luminosities in the model.

Similar to the model in Behroozi et al. (2013), TRINITY is built upon population statistics from a dark matter *N*-body simulation. Specifically, the model makes a guess for how haloes, galaxies, and SMBHs evolve over time. This guess is then applied to the haloes in the simulation, resulting in a mock universe. This mock universe is compared with the real Universe in terms of the observables above, quantified by a Bayesian likelihood. With this likelihood, a Markov Chain Monte Carlo (MCMC) algorithm is used to explore model parameter space until convergence. The resultant parameter posterior distribution tells us the optimal way to connect galaxies and SMBHs to their host haloes, as well as the uncertainties therein.

This work is the first in a series of TRINITY papers, and it covers the TRINITY methodology. The second paper (Paper II) discusses QLFs, the radiative versus kinetic energy output from AGNs, and the build-up of SMBHs across cosmic time; the third paper (Paper III) provides predictions for quasars and other SMBHs at $z > 6$; the fourth paper (Paper IV) discusses the SFR–BHAR correlation as a function of halo mass, galaxy mass, and redshift; and the fifth paper (Paper V) covers SMBH merger rates and TRINITY's predictions for gravitational wave experiments. The sixth (Paper VI) and seventh (Paper VII) papers present the AGN autocorrelation functions and AGN–galaxy cross-correlation functions from TRINITY, respectively. They also discuss whether/how well AGN clustering signals can be used to constrain models like TRINITY. Mock catalogues containing full information about haloes, galaxies, and SMBHs will be introduced in the sixth paper.

The paper is organized as follows. In Section 2, we describe the methodology. Section 3 covers the simulation and observations used in TRINITY. Section 4 presents the results of our model, followed by the discussion and comparison with other models in Section 5. Finally, we discuss the caveats of and the future directions for TRINITY in Section 6, and present conclusions in Section 7. In this work, we adopt a flat Λ cold dark matter cosmology with parameters ($\Omega_{\rm m} = 0.307$, $\Omega_\Lambda = 0.693$, $h = 0.678$, $\sigma_8 = 0.823$, $n_s = 0.96$) consistent with *Planck* results (Planck Collaboration XIII 2016). We use data sets that adopt the Chabrier stellar initial mass

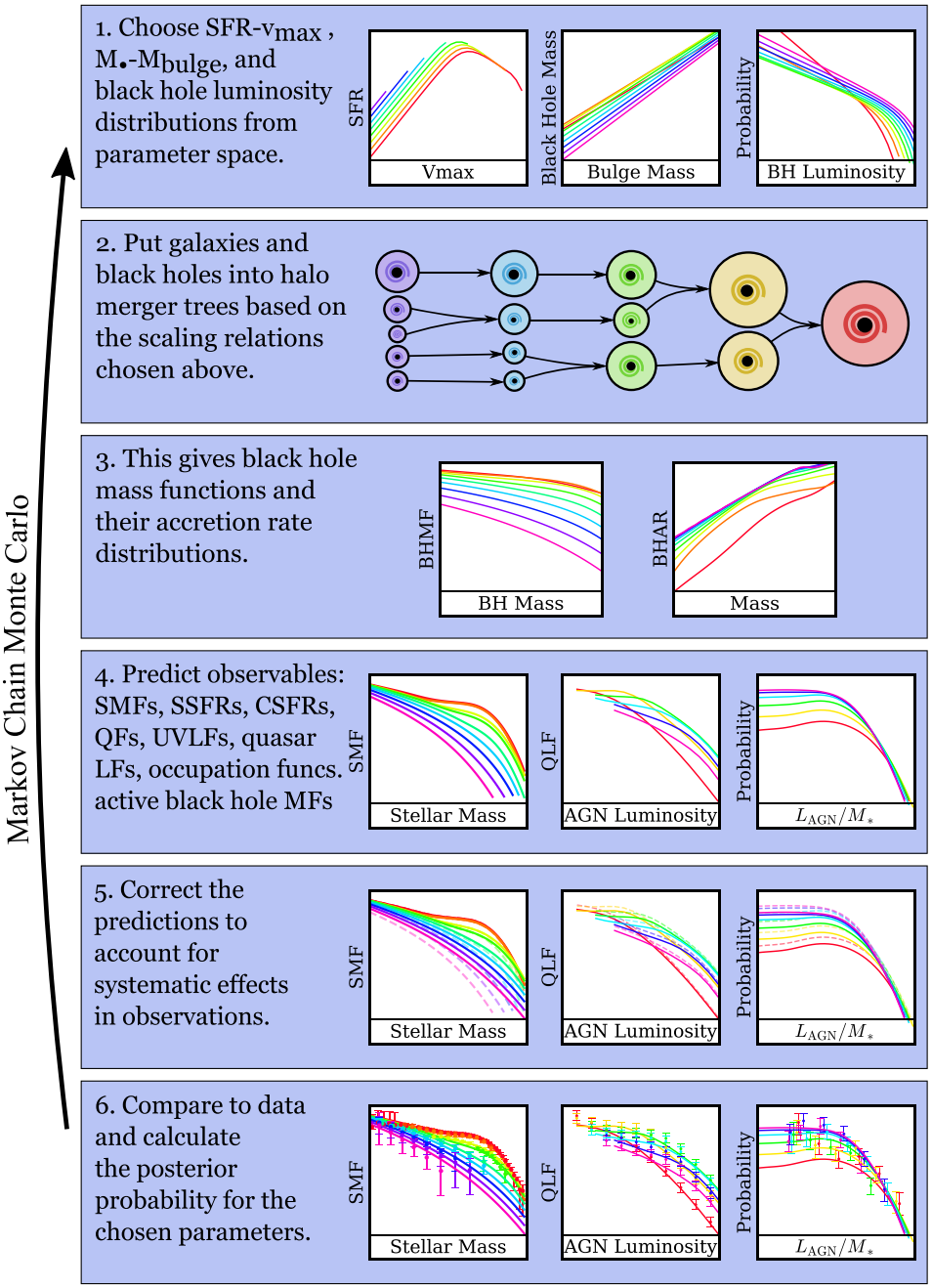

Figure 1. Visual summary of the methodology used to constrain the halo–galaxy–SMBH connection. See Section 2 for details.

function (IMF, Chabrier 2003), the Bruzual & Charlot (2003) stellar population synthesis (SPS) model, and the Calzetti dust attenuation law (Calzetti et al. 2000). Halo masses are calculated following the virial overdensity definition from Bryan & Norman (1998).

2 METHODOLOGY

2.1 Overview

TRINITY is an empirical model that self-consistently infers the halo–galaxy–SMBH connection from $z = 0$–10. In TRINITY, we make this statistical connection in several steps (Fig. 1). We first parametrize the SFR as a function of halo mass and redshift. For a given choice in this parameter space (for the full description of our model parameters, see Table 1), we can integrate the resulting SFRs over average halo assembly histories to get the stellar mass–halo mass (SMHM) relation (Section 2.2). We then convert total galaxy mass to *bulge* mass with a scaling relation from observations. Next, we connect SMBHs with galaxies by parametrizing the redshift evolution of the SMBH mass–bulge mass ($M_\bullet$–$M_{\rm bulge}$) relation (Section 2.4). A given choice of this relation will determine average SMBH accretion rates, because average galaxy growth histories are set by the SFR–halo relationship. Lastly, we parametrize the Eddington ratio distributions and mass-to-energy conversion efficiency, which determines how SMBH growth translates to the observed distribution of SMBH luminosities. In brief, this modelling process gives the distribution

of galaxy and SMBH properties. After modelling AGN radiative and kinetic luminosities (Section 2.7) as well as correcting for systematic effects, these properties are used to predict the galaxy and AGN observables (Section 2.8). We compare these predictions to observations to compute a likelihood function, and use an MCMC algorithm to obtain the posterior distribution of model parameters that are consistent with observations. Each choice of model parameters fully specifies the halo–galaxy–SMBH connection, and the posterior distribution provides the plausible range of uncertainties in this connection given observational constraints.

Of note, TRINITY models ensemble populations of haloes, galaxies, and SMBHs by following different halo mass bins along average halo growth tracks (as in Behroozi et al. 2013), instead of tracking *individual* halo and galaxy histories (as in the UNIVERSEMACHINE; Behroozi et al. 2019). Given this statistical nature, TRINITY is not yet able to provide object-specific growth histories. For calculation of average star formation histories (SFHs) in different halo mass bins, we refer readers to appendix B of Behroozi et al. (2013). In Appendix C, we lay out the procedure to calculate SMBH masses that: (1) were inherited from the most massive progenitor (MMP) haloes; (2) came in with infalling satellite haloes. While a future version of TRINITY will be integrated into the UNIVERSEMACHINE, the present version requires only halo population statistics (i.e. halo mass functions and merger rates) from dark matter simulations (like Grylls et al. 2019), as opposed to individual halo merger trees. As a result, TRINITY allows extremely efficient computation of observables, and hence, rapid model exploration.

2.2 Connecting galaxies to haloes

We adopt a very similar parametrization for the halo–galaxy connection in TRINITY as was shown to work successfully in the UNIVERSEMACHINE (Behroozi et al. 2019). Although simpler parametrizations exist, this choice makes future integration with the UNIVERSEMACHINE easier. The UNIVERSEMACHINE modelled star-forming and quiescent haloes individually, but TRINITY models halo population averages, and we maintain this parametrization in TRINITY. In practice, however, TRINITY only depends on the total SFR of all haloes in a given mass bin, which depends almost exclusively on the SFR for star-forming galaxies and the quiescent fraction as a function of halo mass and redshift.

Our model assumes that the median SFRs of star-forming galaxies are a function of both the host halo mass and redshift. In this work, we adopt the maximum circular velocity of the halo ($v_{\rm max} = \max(\sqrt{GM(<R)/R})$ at the time when it reaches its peak mass, $v_{\rm Mpeak}$, as a proxy for the peak halo mass $M_{\rm peak}$. This choice reduces the sensitivity to pseudo-evolution in halo mass definitions and to spikes in $v_{\rm max}$ during mergers (Behroozi et al. 2019). Our parametrization is

$$\mathrm{SFR_{SF}} = \frac{\epsilon}{v^{\alpha} + v^{\beta}} \tag{1}$$

$$v = \frac{v_{\rm Mpeak}}{V \cdot \mathrm{km\ s^{-1}}} \tag{2}$$

$$a = \frac{1}{1+z} \tag{3}$$

$$\log_{10}(V) = V_0 + V_a(a-1)) + V_{z1}\ln(1+z) + V_{z2}z \tag{4}$$

$$\log_{10}(\epsilon) = \epsilon_0 + \epsilon_1(a-1) + \epsilon_{z1}\ln(1+z) + \epsilon_{z2}z \tag{5}$$

$$\alpha = \alpha_0 + \alpha_a(a-1) + \alpha_{z1}\ln(1+z) + \alpha_{z2}z \tag{6}$$

$$\beta = \beta_0 + \beta_a(a-1) + \beta_z z. \tag{7}$$

The *median* SFRs of star-forming galaxies ($\mathrm{SFR_{SF}}$) are a power law with slope $-\alpha$ for $v_{\rm Mpeak} \ll V$, and another power law with slope $-\beta$ for $v_{\rm Mpeak} \gg V$. The parameter ϵ is the characteristic SFR when $v_{\rm Mpeak} \sim V$. We remove the Gaussian boost in SFR at $v_{\rm Mpeak} \sim V$ in the UNIVERSEMACHINE, because the UNIVERSEMACHINE's posterior distribution of model parameters suggested no need for such a boost.

We adopt the following parametrization for the fraction of quiescent galaxies, $f_{\rm Q}$, as a function of redshift and $v_{\rm Mpeak}$:

$$f_{\rm Q} = 1 - \frac{1}{1+\exp(x)} \tag{8}$$

$$x = \frac{\log_{10}(v_{\rm Mpeak}) - v_{\rm Q}}{w_{\rm Q}} \tag{9}$$

$$v_{\rm Q} = v_{\rm Q,0} + v_{\rm Q,}a(a-1) + v_{\rm Q,}z z \tag{10}$$

$$w_{\rm Q} = w_{\rm Q,0} + w_{\rm Q,}a(a-1) + w_{\rm Q,}z z. \tag{11}$$

For quiescent galaxies, we assign a median SSFR of $10^{-11.8}$ yr^{-1} to match SDSS values (Behroozi et al. 2015). We also set the lognormal scatter of the SFRs in star-forming and quiescent galaxies to be $\sigma_{\rm SFR,SF} = 0.30$ dex and $\sigma_{\rm SFR,Q} = 0.42$ dex, respectively (Speagle et al. 2014). Thus, the *average total* SFR in each given $M_{\rm peak}$ (or $v_{\rm Mpeak}$) bin is simply

$$\begin{aligned}\mathrm{SFR_{tot}} = {} & \mathrm{SFR_{SF}} \times (1-f_{\rm Q}) \times \exp(0.5(\sigma_{\rm SFR,SF}\ln 10)^2) \\ & + \mathrm{SSFR_Q} \times M_* \times f_{\rm Q} \times \exp(0.5(\sigma_{\rm SFR,Q}\ln 10)^2),\end{aligned} \tag{12}$$

where the exponentials reflect the difference between the *average* and *median* values of lognormal distributions.

Aside from star formation, galaxies also gain stellar mass via mergers, where stars from incoming galaxies are transferred to central galaxies. In this work, we assume that a certain fraction, $f_{\rm merge}$, of the stars from incoming galaxies are merged into the central galaxies. As in Behroozi et al. (2019), we assume $f_{\rm merge}$ to be independent of halo mass due to the approximately self-similar nature of haloes. We also assume $f_{\rm merge}$ to be redshift-independent. The average stellar mass in a given halo mass bin at a given redshift z is correspondingly

$$\begin{aligned}M_*(t) = {} & \int_0^t (1 - f_{\rm loss}(t-t'))\mathrm{SFR_{tot}}(t')\mathrm{d}t' \\ & + f_{\rm merge}\int_0^t\int_0^{t'} (1 - f_{\rm loss}(t-t''))\dot{M}_{*,\rm inc}(t', t'')\mathrm{d}t''\mathrm{d}t'\end{aligned} \tag{13}$$

$$f_{\rm loss}(T) = 0.05\ln\left(1 + \frac{T}{1.4\ \mathrm{Myr}}\right), \tag{14}$$

where $f_{\rm loss}(T)$ is the stellar mass-loss fraction as a function of stellar age T from Behroozi et al. (2013), $\mathrm{SFR_{tot}}$ is the total average SFR from equation (12), and $\dot{M}_{*,\rm inc}(t',\ t'')$ is the rate at which the incoming satellite galaxies merge into central galaxies, as a function of the time of disruption t' and the time that the stellar population formed, t''. For a given halo mass bin around the descendant halo mass, $M_{\rm desc}$, $\dot{M}_{*,\rm inc}(t',\ t'')$ can be calculated by convolving the halo merger rates from the UNIVERSEMACHINE ($\mathrm{d}^2N(M_{\rm desc}, \theta, z(t'))/(\mathrm{d}\log\theta\ \mathrm{d}z)$, see Appendix A) with the SFHs of merged satellite haloes:

$$\begin{aligned}\dot{M}_{*,\rm inc}(t', t'') = {} & \int_0^1 \frac{\mathrm{d}^2N(M_{\rm desc},\ \theta,\ z(t'))}{\mathrm{d}\log\theta\ \mathrm{d}z}\mathrm{SFR}(M_{\rm sat}, t'')\mathrm{d}\log\theta \\ & \times\frac{\mathrm{d}z}{\mathrm{d}t'},\end{aligned} \tag{15}$$

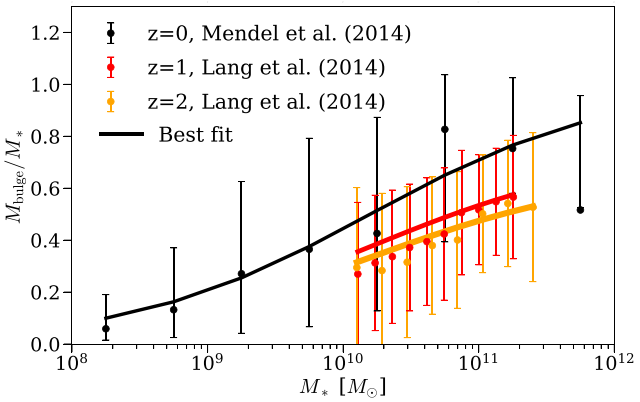

Figure 2. The fit to the median galaxy bulge mass–total mass relation for $z = 0$–2 (solid lines, equations 16–17). Observed data points are from Mendel et al. (2014) and Lang et al. (2014). The error bars from Mendel et al. (2014) represent the 16–84th percentile range of the $M_{\rm bulge}/M_*$ ratios from the SDSS catalogue, and those from Lang et al. (2014) are based on the 68 per cent confidence intervals of bulge-to-total ratio (B/T) as a function of stellar mass. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

where $M_{\rm sat}$ is the mass of the satellite halo, and $\theta = M_{\rm sat}/M_{\rm desc}$ is the mass ratio between the satellite halo and the descendant halo.

It is also well known that there is scatter in stellar mass at fixed halo mass (see e.g. Wechsler & Tinker 2018). We parametrize this scatter as a lognormal distribution with a width σ_* that is redshift-independent, with a flat prior on $\sigma_{*,0}$ of 0–0.3 dex.

The galaxy–SMBH connection is made via the SMBH mass–bulge mass ($M_\bullet$–$M_{\rm bulge}$, Section 2.4) relation. To make the halo–galaxy–SMBH connection, we need to convert *total* galaxy mass M_* to the *bulge* mass $M_{\rm bulge}$. In this work, we fit the median bulge mass–total mass relations from SDSS (Mendel et al. 2014) and CANDELS (Lang et al. 2014) galaxies with

$$M_{\rm bulge} = \frac{f_z(z) M_*}{1 + \exp\{k_{\rm SB}[\log_{10}(M_*/M_{\rm SB})]\}} \tag{16}$$

$$f_z(z) = \frac{z+2}{2z+2}, \tag{17}$$

where $k_{\rm SB} = -1.13$ determines how fast $M_{\rm bulge}$ converges to M_* at the massive end, and $M_{\rm SB} = 10^{10.2} M_\odot$ is a characteristic stellar mass. This fit is shown in Fig. 2. It should be noted that no data points exist beyond $z = 2.5$, so equation (16) is extrapolated at $z > 2.5$. With the functional form chosen here, $M_{\rm bulge}/M_*$ asymptotes at high redshifts to half the value of $M_{\rm bulge}/M_*$ at $z = 0$. We discuss how alternative assumptions for the $M_{\rm bulge}$–M_* relation would affect our results in Appendix E2.

Disc–bulge decompositions are sensitive to the fitting method used, and it is also difficult to estimate how much of the scatter in bulge-to-total mass ratios is intrinsic versus observational. As a result, we subsume the scatter in the bulge-to-total mass relation into the scatter of the $M_\bullet$–$M_{\rm bulge}$ relation, as the two scatters are degenerate in TRINITY given current data constraints.

At $0 < z < 8$, SMFs primarily constrain the halo–galaxy connection. Beyond $z = 8$, SMFs are not available, so we constrain the halo–galaxy connection with galaxy UVLFs instead. This requires generating UV luminosities from SFRs as a function of host halo mass and redshift. To do so, we fit the median UV magnitude, $\widetilde{M}_{\rm UV}$, and the lognormal scatter, $\sigma_{M_{\rm UV}}$, as functions of SFR, $M_{\rm peak}$, and redshift from the output of the UNIVERSEMACHINE:

$$\widetilde{M}_{\rm UV} = k_{\rm UV} \times \log_{10} {\rm SFR} + b_{\rm UV} \tag{18}$$

$$\sigma_{M_{\rm UV}} = k_{\sigma_{\rm UV}} \times \log_{10} M_{\rm peak} + b_{\sigma_{\rm UV}} \tag{19}$$

$$k_{\rm UV} = a_k (\log_{10} M_{\rm peak})^2 + b_k \log_{10} M_{\rm peak} + c_k (a-1) + d_k \tag{20}$$

$$b_{\rm UV} = a_b (\log_{10} M_{\rm peak})^2 + b_b \log_{10} M_{\rm peak} + c_b (a-1) + d_b \tag{21}$$

$$k_{\sigma_{\rm UV}} = a_{k_\sigma} z + b_{k_\sigma} \tag{22}$$

$$b_{\sigma_{\rm UV}} = a_{b_\sigma} z + b_{b_\sigma}. \tag{23}$$

Details of the fitting process are shown in Appendix B. UNIVERSEMACHINE models UV luminosities using the Flexible Stellar Population Synthesis code (FSPS; Conroy & White 2013), and equations (18)–(23) provide a rapid way to obtain statistically equivalent results. We hence use these scaling relations to assign UV magnitude distributions to haloes given their masses, SFRs, and redshifts, allowing us to calculate UVLFs at $z = 9$ and $z = 10$.

2.3 Observational systematics for galaxies

Following Behroozi et al. (2019), we model several observational systematics when predicting galaxy observables. We include a mass-independent systematic offset μ between the observed ($M_{*,\rm obs}$) and the true stellar mass ($M_{*,\rm true}$) to model uncertainties from the IMF, SPS model, the dust model, SFH model, assumed metallicities, and redshift errors:

$$\log_{10}\left(\frac{M_{*,\rm obs}}{M_{*,\rm true}}\right) = \mu. \tag{24}$$

The offset μ has the following redshift scaling:

$$\mu = \mu_0 + \mu_a (a-1). \tag{25}$$

Following Behroozi et al. (2013), we set the prior width on μ_0 and μ_a to 0.14 and 0.24 dex, respectively (see Table 2).

As described in appendix C of Behroozi et al. (2019), there are systematic offsets between observed and true SSFRs that peak near $z \sim 2$, which are most evident when comparing observed SSFRs to the evolution of observed SMFs. As in Behroozi et al. (2019), we include another redshift-dependent offset κ to account for this systematic offset in SFRs. The total offset between the observed ($\rm SFR_{*,obs}$) and true SFRs ($\rm SFR_{*,true}$) is

$$\log_{10}\left(\frac{\rm SFR_{*,obs}}{\rm SFR_{*,true}}\right) = \mu + \kappa \exp\left(-\frac{(z-2)^2}{2}\right). \tag{26}$$

The prior width on κ is set to 0.24 dex (Table 2), again from Behroozi et al. (2019).

We also model a redshift-dependent, lognormal scatter in the measured stellar mass relative to the true mass:

$$\sigma(z) = \min(\sigma_0 + \sigma_z z, 0.3). \tag{27}$$

This scatter causes an Eddington bias (Eddington 1913) in the SMF, which enhances the number density of massive galaxies because there are more small galaxies that can be scattered up than massive galaxies that can be scattered down. Following Conroy & White (2013), we fix $\sigma_0 = 0.07$ dex. We adopt a Gaussian prior on σ_z with centre 0.05 and width 0.015 dex, respectively (see Table 2), following Behroozi et al. (2019).

Finally, the correlation between scatter in the SFR and scatter in the stellar mass at fixed halo mass affects the calculation of SSFRs

as a function of stellar mass. To account for this correlation ρ, we adopt the following formula from Behroozi et al. (2013):

$$\rho(a) = 1 + (4\rho_{0.5} - 3.23)a + (2.46 - 4\rho_{0.5})a^2, \quad (28)$$

where $\rho_{0.5}$ is a free parameter that represents the correlation between the SSFR and stellar mass at $z = 1$ (i.e. $a = 0.5$). The details of this correction are in appendix C.2 of Behroozi et al. (2013). Following Behroozi et al. (2013), we set the prior on $\rho_{0.5}$ to be a uniform distribution between 0.23 and 1.0 (Table 2).

2.4 Connecting SMBHs to galaxies

2.4.1 *SMBH occupation fractions*

In the real Universe, not every halo and galaxy host central SMBHs. That is, the occupation fraction of SMBHs, $f_{\rm occ}$, is likely below unity. At $z = 0$, we find that most massive galaxies host central SMBHs, but it is still debated how many smaller and/or earlier galaxies are SMBH-occupied (see Greene, Strader & Ho 2020 and references therein). Theoretical studies suggest that $f_{\rm occ}$ could be a sigmoid function of halo mass with potential redshift evolution (e.g. Volonteri 2010; Bellovary et al. 2011; Dunn et al. 2018). Therefore, we adopt this functional form in TRINITY, and allow the following redshift dependence of (1) a minimum SMBH occupation fraction, $f_{\rm occ,min}$, (2) the characteristic halo mass, $M_{\rm h,c}$; and (3) the (log-)halo mass range, $w_{\rm h,c}$, over which $f_{\rm occ}$ changes significantly:

$$f_{\rm occ} = \frac{\exp(x)}{1 + \exp(x)} \times (1 - f_{\rm occ,min}) + f_{\rm occ,min} \quad (29)$$

$$x = \frac{\log_{10}(M_{\rm peak}) - \log_{10}(M_{\rm h,c})}{w_{\rm h,c}} \quad (30)$$

$$\log_{10}(f_{\rm occ,min}) = f_{\rm occ,min,0} + f_{\rm occ,min,a}(a - 1) \quad (31)$$

$$\log_{10}(M_{\rm h,c}) = M_{\rm h,c,0} + M_{\rm h,c,a}(a - 1) \quad (32)$$

$$w_{\rm h,c} = w_{\rm h,c,0} + w_{\rm h,c,a}(a - 1). \quad (33)$$

$f_{\rm occ,min}$ is motivated by the calculation of characteristic $M_\bullet$ for *host* galaxies, where $f_{\rm occ}$ is used as a denominator (see equation 39).

However, all the posterior parameter distributions of TRINITY models – the fiducial models and the variants covered in the Appendix – predict $f_{\rm occ} \sim 1$ in the halo/galaxy mass ranges covered by TRINITY. The physical reason is that, without new SMBH seeds at lower redshifts, $f_{\rm occ}$ at a fixed halo mass can only decrease as less massive, unseeded haloes grow in mass. On the other hand, a uniformly high $f_{\rm occ}$ down to $M_{\rm peak} \sim 10^{11} M_\odot$ in the local universe is required to explain AGN observations such as ABHMFs. As a result, $f_{\rm occ}$ can only be higher at $z > 0$ for $M_{\rm peak} > 10^{11} M_\odot$, which leads to $f_{\rm occ} \sim 1$. This result is also consistent with earlier simulations of SMBH formation (e.g. Habouzit, Volonteri & Dubois 2017; Tremmel 2017), which found $f_{\rm occ} \sim 1$ in haloes with $M_{\rm peak} > 10^{11} M_\odot$.

2.4.2 *Redshift-dependent $M_\bullet$–M_{bulge} relation*

There are multiple known empirical scaling relations between $M_\bullet$ and galaxy properties, with strong debate over which is most fundamental (Ferrarese & Merritt 2000; Ferrarese 2002; Novak, Faber & Dekel 2006; Aller & Richstone 2007; Hu 2008; Beifiori et al. 2012; Shankar et al. 2016; van den Bosch 2016). Here, we parametrize the relation between SMBHs and galaxy bulge mass. Specifically, the *median* $M_\bullet$–$M_{\rm bulge}$ relation is a redshift-dependent power law:

$$\log_{10} \widetilde{M}_\bullet = \beta_{\rm BH} + \gamma_{\rm BH} \log_{10} \left(\frac{M_{\rm bulge}}{10^{11} M_\odot} \right) \quad (34)$$

$$\beta_{\rm BH} = \beta_{\rm BH,0} + \beta_{{\rm BH},a}(a - 1) + \beta_{{\rm BH},z} z \quad (35)$$

$$\gamma_{\rm BH} = \gamma_{\rm BH,0} + \gamma_{{\rm BH},a}(a - 1) + \gamma_{{\rm BH},z} z. \quad (36)$$

We set Gaussian priors on $\beta_{\rm BH,0}$ and $\gamma_{\rm BH,0}$ from constraints on the local $M_\bullet$–$M_{\rm bulge}$ relation, which will be discussed in Section 3.2.2. With equations (34)–(36), some parameter values could result in unphysical (i.e. negative) growth of SMBHs; we hence exclude such parts of parameter space from MCMC exploration.

There is also lognormal scatter in SMBH mass at fixed bulge mass ($\sigma_{\rm BH}$). We assume $\sigma_{\rm BH}$ to be redshift-independent. This is because a redshift-dependent $\sigma_{\rm BH}$ will be unphysically small in the early Universe, if the Poisson prior probability of detecting low-mass bright quasars at $z \sim 6$ is applied. See Section 3.2.2 for more details.

Since the scatter in bulge mass at fixed stellar mass is subsumed in $\sigma_{\rm BH}$, this is in effect the scatter in SMBH mass at fixed *total* stellar mass. We also note that this scatter is effectively the combined scatter that accounts for both the variance in the intrinsic $M_\bullet$–$M_{\rm bulge}$ relation, as well as random error in direct SMBH mass measurements (e.g. dynamical modelling or reverberation mapping, but not virial estimates). Combining the scatter in SMBH mass at fixed stellar mass with the scatter in stellar mass at fixed halo mass, the scatter in SMBH mass at fixed halo mass is

$$\sigma_{\rm tot} = \sqrt{(\sigma_* \times \gamma_{\rm BH})^2 + \sigma_{\rm BH}^2}. \quad (37)$$

Such a calculation effectively assumes that the bulge mass fraction of galaxies is fixed at fixed halo mass. This lognormal scatter results in a difference between the *mean* ($\overline{M_\bullet}$) and *median* SMBH masses ($\widetilde{M_\bullet}$) at fixed halo mass:

$$\overline{M_\bullet} = \widetilde{M}_\bullet \times \exp(0.5(\sigma_{\rm tot} \ln 10)^2). \quad (38)$$

We note that the median and average $M_\bullet$'s calculated above are for *all* the galaxies, whether they host SMBHs or not. Generally, these masses are different from those for SMBH *host galaxies*. With an SMBH occupation fraction $f_{\rm occ} < 1$, the median and average $M_\bullet$'s for SMBH host galaxies, $\overline{M_{\bullet,\rm host}}$ and $\widetilde{M_{\bullet,\rm host}}$, would be

$$\overline{M_{\bullet,\rm host}} = \frac{\overline{M_\bullet}}{f_{\rm occ}} \quad (39)$$

$$\widetilde{M_{\bullet,\rm host}} = \frac{\widetilde{M_\bullet}}{f_{\rm occ}}. \quad (40)$$

However, as we noted in Section 2.4.1, all the posterior parameter distributions of TRINITY models predict $f_{\rm occ} \sim 1$ for $M_{\rm peak} > 10^{11} M_\odot$. Therefore, equation (39) results in effectively identical SMBH properties for *all* versus *host* haloes/galaxies, so we do not provide separate results for all versus host haloes/galaxies in the rest of this work.

2.5 SMBH mergers and accretion

Similar to their host galaxies, SMBHs grow in mass via accretion and mergers. We parametrize the fraction of SMBH growth due to mergers as $f_{\rm merge,BH}$, the formula for which is provided later in this section. The average black hole merger rate ($\overline{\rm BHMR}$) for a certain halo mass bin is by definition

$$\overline{\rm BHMR} \cdot \Delta t = (\text{Average BH Mass Now} - \text{Average BH Mass Inherited from Most Massive Progenitors}) f_{\rm merge,BH}, \quad (41)$$

where Δt is the time interval between two consecutive snapshots, and the inherited and new BH masses are calculated using the halo–galaxy–SMBH connection (see Appendix C for full details). Similarly, the average black hole accretion rate ($\overline{\mathrm{BHAR}}$) for a certain halo mass bin is

$$\begin{aligned}\overline{\mathrm{BHAR}} \cdot \Delta t = &(\text{Average BH Mass Now}\\ &- \text{Average BH Mass Inherited from}\\ &\text{Most Massive Progenitors})(1 - f_{\mathrm{merge,BH}}).\end{aligned} \tag{42}$$

In this work, we assume that the fractional merger contribution to the total SMBH growth ($f_{\mathrm{merge,BH}}$) is proportional to the fraction of galaxy growth due to mergers:

$$f_{\mathrm{merge,BH}} = f_{\mathrm{scale}} \times \frac{f_{\mathrm{merge}}\dot{M}_{*,\mathrm{inc}}}{\mathrm{SFR} + f_{\mathrm{merge}}\dot{M}_{*,\mathrm{inc}}}, \tag{43}$$

where f_{merge} is the fraction of the incoming satellite galaxies' mass that is merged into central galaxies, and $\dot{M}_{*,\mathrm{inc}}$ is the mass rate at which satellite galaxies are disrupted in mergers (see equation 13). The proportionality factor, f_{scale}, has the following redshift dependency:

$$\log_{10}(f_{\mathrm{scale}}) = f_{\mathrm{scale},0} + f_{\mathrm{scale},1}(a - 1). \tag{44}$$

While we do not exclude $f_{\mathrm{scale}} > 1$ when exploring parameter space, we find f_{scale} to be consistently smaller than unity in the posterior distribution (see Appendix H for model extremes where $f_{\mathrm{scale}} = 0$ or $f_{\mathrm{scale}} = 1$).

In TRINITY, not all infalling SMBH mass merges with the central SMBH immediately. Physically, this could be due to several reasons: (1) some SMBHs orbit with the disrupted satellite (i.e. in a tidal stream) outside the host galaxy and have very long dynamical friction time-scales, (2) some SMBHs experience recoils and are ejected from the central galaxy; (3) some SMBHs may stall in the final parsec before merging with the central SMBH; or (4) some SMBHs may remain in the host galaxy but stay offset from the centre. Given the lack of direct observational evidence, we cannot distinguish between these possible scenarios here. Instead, we label all such objects as 'wandering SMBHs' for the rest of this work. The average mass in wandering SMBHs ($\overline{M}_{\bullet,\mathrm{wander}}$) for each halo mass bin is thus

$$\overline{M}_{\bullet,\mathrm{wander}} = \text{Total Infalling BH Mass} - \int_0^t \overline{\mathrm{BHMR}} \cdot \mathrm{d}t. \tag{45}$$

Although wandering SMBHs do not contribute to the observed $M_\bullet$–M_{bulge} relation, we assume that they do contribute to QLFs during their formation. For full details about calculating the average inherited SMBH mass from the previous time-step ($\overline{M}_{\bullet,\mathrm{inherit}}$) and the average infalling SMBH mass ($\overline{M}_{\bullet,\mathrm{infall}}$), see Appendix C.

2.6 AGN duty cycles, Eddington ratio distributions, and energy efficiencies

As noted in Section 2.1, TRINITY is not designed to follow the growth histories of individual haloes, galaxies, or SMBHs. Instead, TRINITY gives their *average* growth histories. To model AGN accretion rate distributions, it is therefore necessary to parametrize both the AGN duty cycles (i.e. the fraction of galaxies that host active SMBHs, f_{duty}) and the shapes of their Eddington ratio distributions. f_{duty} is a function of M_{peak} and z:

$$f_{\mathrm{duty}}(M_{\mathrm{peak}}, z) = \min\left\{\left(\frac{M_{\mathrm{peak}}}{M_{\mathrm{duty}}}\right)^{\alpha_{\mathrm{duty}}}, 1\right\} \tag{46}$$

$$M_{\mathrm{duty}} = M_{\mathrm{duty},0} + M_{\mathrm{duty},z}\log(1 + z) \tag{47}$$

$$\alpha_{\mathrm{duty}} = \alpha_{\mathrm{duty},0} + \alpha_{\mathrm{duty},z}\log(1 + z). \tag{48}$$

In this work, we define f_{duty} to be the fraction of active SMBH hosts relative to *all* galaxies. But given that the posterior distributions of all TRINITY models predict $f_{\mathrm{occ}} \sim 1$ at $M_{\mathrm{peak}} > 10^{11}$ and $0 \leq z \leq 10$, f_{duty} is effectively the fraction of SMBH *host* galaxies whose SMBHs are active.

At a fixed *halo* mass, the Eddington ratio distribution function (ERDF) is assumed to have a double power-law shape:

$$P(\eta|\eta_0, c_1, c_2) = f_{\mathrm{duty}}\frac{P_0}{\left(\frac{\eta}{\eta_0}\right)^{c_1} + \left(\frac{\eta}{\eta_0}\right)^{c_2}} + (1 - f_{\mathrm{duty}})\delta(\eta) \tag{49}$$

$$c_1 = c_{1,0} + c_{1,a}(a - 1) \tag{50}$$

$$c_2 = c_{2,0} + c_{2,a}(a - 1), \tag{51}$$

where η is the Eddington ratio, P_0 is the normalization of the ERDF for *active* SMBHs, c_1 and c_2 are the two power-law indices, η_0 is the break point of the double power-law, and $\delta(\eta)$ is the ERDF for dormant SMBHs, which is a Dirac delta function centred at $\eta = 0$. The constant of proportionality P_0 is calculated such that

$$\int_0^\infty \frac{P_0}{\left(\frac{\eta}{\eta_0}\right)^{c_1} + \left(\frac{\eta}{\eta_0}\right)^{c_2}} d\log\eta = 1. \tag{52}$$

This functional form is flexible enough to approximate many past assumptions for the shape of the ERDF (e.g. Gaussian distributions and Schechter functions).

The characteristic Eddington ratio η_0 in equation (49) is *not* a free parameter, but is constrained by the parametrizations in equations (46)–(49). Letting $\overline{\eta}$ be the average Eddington ratio, we have from equation (49) that

$$\overline{\eta} = f_{\mathrm{duty}}\int_0^\infty \eta P(\eta|\eta_0, c_1, c_2) d\log\eta, \tag{53}$$

and by definition

$$\overline{\eta} = \frac{\epsilon_{\mathrm{tot}}\overline{\mathrm{BHAR}} \times 4.5 \times 10^8\ \mathrm{yr}}{(1 - \epsilon_{\mathrm{tot}})\overline{M}_\bullet}, \tag{54}$$

where $\overline{M}_\bullet$ and $\overline{\mathrm{BHAR}}$ (equations 38 and 42) are the average SMBH mass and black hole accretion rate, respectively. The parameter ϵ_{tot} is the efficiency of releasing energy (both radiative and kinetic) through accretion. We hence solve for η_0 by combining equations (53) and (54). In this work, $\log_{10}(\epsilon_{\mathrm{tot}})$ is assumed to be redshift-independent.

Given the non-zero scatter in SMBH mass at fixed halo mass (equation 37), different SMBHs with the same host halo mass may have different Eddington ratio distributions. Without joint observational constraints as a function of SMBH mass and galaxy mass, we assume that SMBHs with the same host halo mass share the same Eddington ratio distribution *shapes* (equation 49), but can have different average Eddington ratios. To quantify the systematic change in average Eddington ratio with $M_\bullet$ at fixed halo mass, we parametrize the correlation coefficient between $\overline{\mathrm{BHAR}}$ and $M_\bullet$ as a function of redshift:

$$\begin{aligned}\log_{10}\overline{\mathrm{BHAR}}(M_\bullet|M_{\mathrm{peak}}) = &\overline{\mathrm{BHAR}}\left(\overline{M}_\bullet(M_{\mathrm{peak}})\right)\\ &+ \rho_{\mathrm{BH}}\log_{10}\left(\frac{M_\bullet}{\overline{M}_\bullet(M_{\mathrm{peak}})}\right)\end{aligned} \tag{55}$$

$$\rho_{\mathrm{BH}} = \rho_{\mathrm{BH},0} + \rho_{\mathrm{BH},a}(a - 1) + \rho_{\mathrm{BH},z}z. \tag{56}$$

For example, $\rho_{\mathrm{BH}} = 1$ means that different SMBHs at fixed halo mass share identical *Eddington ratio distribution*, while $\rho_{\mathrm{BH}} = 0$

means that these SMBHs have identical *absolute accretion rate* distributions. Here, we allow $\rho_{\rm BH}$ to take a value within $[-1, 1]$. Any $\rho_{\rm BH}$ above(below) $1(-1)$ is capped at $1(-1)$.

2.7 Kinetic and radiative Eddington ratios

SMBH accretion produces both radiative and kinetic energy (see e.g. Merloni & Heinz 2008), and the latter dominates the total energy output at low accretion rates. The radiative and kinetic luminosities depend on the efficiency of mass conversion into the two different forms of energies, $\epsilon_{\rm rad}$ and $\epsilon_{\rm kin}$. In analogy with this, we can recast the Eddington ratio in terms of its radiative and kinetic components. To forward model these observables, we adopt the following empirical relation between the total Eddington ratio η and its radiative component $\eta_{\rm rad}$:

$$\eta_{\rm rad} = \begin{cases} \eta^2/0.03, & \eta \leq 0.03 \\ \eta, & 0.03 < \eta \leq 2 \\ 2[1+\ln(\eta/2)], & \eta > 2 \end{cases}. \quad (57)$$

For $\eta \leq 2$, the scaling between $\eta_{\rm rad}$ and η is similar to the one used by Merloni & Heinz (2008). Merloni & Heinz (2008) adopted a more complex scaling relation between AGN radiative luminosity, X-ray luminosity, and SMBH mass that had substantial scatter. Rather than using the same complex model, we choose to adopt the simpler, more transparent scaling in equation (57). For $\eta \geq 2$, we adopt a logarithmic scaling to account for the fact that at such high accretion rates, the accretion disc becomes thick, trapping part of the outgoing radiation (Mineshige et al. 2000). The kinetic component $\eta_{\rm kin}$ is, by definition,

$$\eta_{\rm kin} = \eta - \eta_{\rm rad}\ ,\ \eta < 0.03\ . \quad (58)$$

At a given $\eta < 0.03$, equation (58) produces $\sim$0.3–0.5 dex more kinetic energy than Merloni & Heinz (2008). We also ignore the kinetic energy output from active SMBHs with $\eta > 0.03$, due to a lack of observational constraints. Thus, the AGN radiative and kinetic efficiencies are

$$\epsilon_{\rm rad} = \epsilon_{\rm tot} \times \begin{cases} \eta/0.03, & \eta \leq 0.03 \\ 1, & 0.03 < \eta \leq 2 \\ 2/\eta\,[1+\ln(\eta/2)], & \eta > 2 \end{cases} \quad (59)$$

and

$$\epsilon_{\rm kin} = \begin{cases} \epsilon_{\rm tot}(1-\eta/0.03), & \eta < 0.03 \\ 0, & \eta > 0.03 \end{cases}, \quad (60)$$

respectively. The radiative and kinetic luminosities and Eddington ratio distributions are

$$\frac{L_{(\cdot)}}{\rm erg\,s^{-1}} = 10^{38.1} \times \frac{M_\bullet}{M_\odot} \times \eta_{(\cdot)} \quad (61)$$

$$P(\eta_{(\cdot)}) = P(\eta)\frac{{\rm d}\log\eta}{{\rm d}\log\eta_{(\cdot)}}, \quad (62)$$

where $(\cdot)$ is either 'rad' or 'kin' and $d\log\eta/d\log\eta_{(\cdot)}$ is calculated using equations (57)–(58).

2.8 Calculating AGN observables

Having specified SMBH growth histories and ERDFs, we can now predict AGN observables. Although there are different observables in our data compilation, all of them involve counting the number densities of the host haloes/galaxies of SMBHs with certain properties.

The SMBH mass function at each redshift is the number density of haloes that host SMBHs of a given mass:

$$\phi_{\rm BH}(M_\bullet, z) = \int_0^\infty \phi_{\rm h}(M_{\rm peak}, z) P(M_\bullet|M_{\rm peak}, z){\rm d}M_{\rm peak}, \quad (63)$$

where $\phi_{\rm h}(M_{\rm peak}, z)$ is the halo mass function at redshift z, and $P(M_\bullet|M_{\rm peak}, z)$ is specified by the halo–galaxy–SMBH connection (see Section 2.2 and Section 2.4).

To model ABHMFs from Schulze & Wisotzki (2010) and Schulze et al. (2015), we apply the same selection criteria and remove SMBHs with radiative Eddington ratios below 0.01. Thus, the ABHMF is

$$\phi_{\rm ABH}(M_\bullet, z) = \int_0^\infty \int_{\eta_{\rm rad,min}=0.01}^\infty \phi_{\rm h}(M_{\rm peak}, z) P(M_\bullet|M_{\rm peak}, z) \times P(\eta_{\rm rad}|M_\bullet, M_{\rm peak}, z){\rm d}\eta_{\rm rad}\,{\rm d}M_{\rm peak}. \quad (64)$$

For the type I quasar SMBH mass functions from Kelly & Shen (2013), we include all SMBHs with $\eta > 0$. This is because modelling of the underlying $M_{\rm BH}$–$L_{\rm bol}$ distributions showed little incompleteness induced by the SDSS luminosity cut at $\log_{10}M_\bullet \gtrsim 9.5$, and we only use data above this mass. To account for obscured type II quasars, we use an empirical formula for the obscured fraction $F_{\rm obs}$ as a function of X-ray luminosity from Merloni et al. (2014):

$$F_{\rm obs}(L_{\rm X}) = 0.56 + \frac{1}{\pi}\arctan\left(\frac{43.89 - \log L_{\rm X}}{0.46}\right). \quad (65)$$

Thus, the type I quasar BHMF is

$$\phi_{\rm ABH'}(M_\bullet, z) = \int_0^\infty \int_0^\infty \phi_{\rm h}(M_{\rm peak}, z) P(M_\bullet|M_{\rm peak}, z) \times P(\eta_{\rm rad}|M_\bullet, M_{\rm peak}, z) \times (1 - F_{\rm obs}(L_{\rm X}))d\eta_{\rm rad}\,{\rm d}M_{\rm peak}, \quad (66)$$

where L_X is the X-ray luminosity that is calculated using the bolometric correction from Ueda et al. (2014):

$$L_{\rm X} = \frac{L_{\rm bol}}{k_{\rm bol}(L_{\rm bol})} \quad (67)$$

$$L_{\rm bol}/{\rm erg\,s^{-1}} = 10^{38.1} \cdot M_\bullet \cdot \eta_{\rm rad} \quad (68)$$

$$k_{\rm bol}(L_{\rm bol}) = 10.83\left(\frac{L_{\rm bol}}{10^{10}L_\odot}\right)^{0.28} + 6.08\left(\frac{L_{\rm bol}}{10^{10}L_\odot}\right)^{-0.020}. \quad (69)$$

Similarly, QLFs are given by the number density of haloes hosting SMBHs with a given luminosity:

$$\phi_{\rm L}(L_{\rm bol}, z) = \int_0^\infty \phi_{\rm h}(M_{\rm peak}) P(L_{\rm bol}|M_{\rm peak}, z){\rm d}M_{\rm peak}, \quad (70)$$

where $P(L_{\rm bol}|M_{\rm peak}, z)$ is calculated by counting the number density of SMBHs with the corresponding Eddington ratio:

$$P(L_{\rm bol}|M_{\rm peak}, z) = \int_0^\infty P(\eta_{\rm rad}(L_{\rm bol}, M_\bullet|M_\bullet, M_{\rm peak}, z) \times P(M_\bullet|M_{\rm peak}, z){\rm d}M_\bullet. \quad (71)$$

Finally, for QPDFs, Aird et al. (2018) expressed *Compton-thin* QPDFs in terms of the specific $L_{\rm X}$ ($sL_{\rm X}$):

$$sL_{\rm X} = \frac{L_{\rm X}/{\rm erg\,s^{-1}}}{1.04\times10^{34}\times M_*/M_\odot}. \quad (72)$$

The distribution of $sL_{\rm X}$ at fixed stellar mass and redshift is

$$P(sL_{\rm X}|M_*, z) = (1 - f_{\rm CTK}(L_{\rm X}, z)) \times P(L'_{\rm bol}|M_*, z) \quad (73)$$

$$L'_{\rm bol} = \frac{L_{\rm bol}}{\xi} \quad (74)$$

Table 1. Summary of parameters.

Symbol	Description	Equation	Parameters	Section
$V(z)$	Characteristic $v_{\rm Mpeak}$ in SFR–$v_{\rm Mpeak}$ relation	4	4	2.2
$\epsilon(z)$	Characteristic SFR in SFR–$v_{\rm Mpeak}$ relation	5	4	2.2
$\alpha(z)$	Low-mass slope of the SFR–$v_{\rm Mpeak}$ relation	6	4	2.2
$\beta(z)$	Massive-end slope of the SFR–$v_{\rm Mpeak}$ relation	7	3	2.2
$v_Q(z)$	Typical $v_{\rm Mpeak}$ for star formation quenching, in dex	10	3	2.2
$w_Q(z)$	Typical width in $v_{\rm Mpeak}$ for star formation quenching, in dex	11	3	2.2
$f_{\rm merge}$	Fraction of incoming satellite galaxy mass that is merged into central galaxies	–	1	2.2
σ_*	Scatter in true stellar mass at fixed halo mass, in dex	–	1	2.2
$\mu(z)$	Systematic offset between true and observed stellar masses, in dex	25	2	2.3
$\kappa(z)$	Additional systematic offset in observed versus true SFRs, in dex	26	1	2.3
$\sigma(z)$	Scatter between measured and true stellar masses, in dex	27	1	2.3
$\rho_{0.5}$	Correlation between SFR and stellar mass at fixed halo mass at $z = 1$ ($a = 0.5$)	28	1	2.3
$f_{\rm occ,min}(z)$	Minimum SMBH occupation fraction	31	2	2.4.1
$M_{\rm h,c}(z)$	Characteristic halo mass where SMBH occupation fraction changes significantly	32	2	2.4.1
$w_{\rm h,c}(z)$	Log-halo mass range over which SMBH occupation fraction changes significantly	33	2	2.4.1
$\beta_{\rm BH}(z)$	Median SMBH mass for galaxies with $M_{\rm bulge} = 10^{11} M_\odot$, in dex	35	3	2.4.2
$\gamma_{\rm BH}(z)$	Slope of the SMBH mass–bulge mass ($M_\bullet$–$M_{\rm bulge}$) relation	36	3	2.4.2
$\sigma_{\rm BH}$	Scatter in SMBH mass at fixed bulge mass, in dex	–	1	2.4.2
$f_{\rm scale}(z)$	Ratio between the fractions of SMBH and galaxy growth coming from mergers	44	2	2.5
$f_{\rm duty}(M_{\rm peak}, z)$	AGN duty cycle	46	4	2.6
$c_1(z)$, $c_2(z)$	Faint- and bright-end slopes of the AGN Eddington ratio distribution functions	50,51	4	2.6
$\epsilon_{\rm tot}$	Total energy efficiency (radiative and kinetic) of mass accretion on to SMBHs	–	1	2.6
$\rho_{\rm BH}(z)$	Correlation coefficient between SMBH accretion rate and mass at fixed halo mass	56	3	2.6
ξ	Systematic offset in Eddington ratio when calculating AGN probability distribution functions, in dex	74	1	2.8
Total number of galaxy parameters			28	
Total number of SMBH parameters			28	
Total number of parameters			56	

Notes. $v_{\rm Mpeak}$: the maximum circular velocity at the time when the halo reaches its peak mass (see Section 2.2).

$$L_{\rm bol}/{\rm erg\cdot s^{-1}} = 1.04\times10^{34}\times M_*/M_\odot \times {\rm s}L_{\rm X}\times k_{\rm bol}(L_{\rm bol}) \quad (75)$$

$$P(L_{\rm bol}|M_*,z) = \int_0^\infty {\rm d}M_{\rm peak}\int_0^\infty {\rm d}M_\bullet P(\eta_{\rm rad}(L_{\rm bol},M_\bullet)|M_{\rm peak},z) \times P(M_\bullet|M_*,z)P(M_*|M_{\rm peak}z), \quad (76)$$

where the Compton-thick fraction $f_{\rm CTK}(L_X, z)$ and the bolometric correction $k_{\rm bol}(L_{\rm bol})$ are both given by Ueda et al. (2014) (see Appendix D2 for full details about $f_{\rm CTK}$), and ξ is the systematic offset in bolometric luminosity when calculating the AGN probability distribution functions in terms of sL_X. This free parameter accounts for a residual inconsistency between the QPDFs from Aird et al. (2018) and the QLFs from Ueda et al. (2014) after the data point downsampling and exclusion as described in Appendix D4.

2.9 Methodology summary

Here, we summarize the major steps to constrain the halo–galaxy–SMBH connection as shown in Fig. 1:

(1) Choose a point in parameter space (Table 1), which fully specifies the halo–galaxy–SMBH connection (Section 2.2, Section 2.4), SMBH merger contributions (Section 2.5), and the BHAR–AGN luminosity conversion (Sections 2.6 and 2.7).

(2) Put galaxies and SMBHs into haloes accordingly, which determines galaxy and SMBH growth histories.

(3) Calculate SMBH mass functions and Eddington ratio distributions (Section 2.6).

(4) Predict galaxy and AGN observables (Section 2.8 and Table 3).

(5) Correct these predictions for systematic effects in real observations, e.g. systematic offsets in measured versus true stellar masses (Section 2.3) as well as Compton-thick obscuration (Section 2.8 and Appendix D).

(6) Compare these predictions with real data to calculate the posterior probability $P(\boldsymbol{\theta}|\mathbf{d}) = \pi(\boldsymbol{\theta})\times\mathcal{L}(\boldsymbol{\theta}|\mathbf{d})$ of the parameters $\boldsymbol{\theta}$ given the observational constraints $\mathbf{d}$. The likelihood $\mathcal{L}(\boldsymbol{\theta}|\mathbf{d})$ is calculated with the $\chi^2(\boldsymbol{\theta}|\mathbf{d})$ from the comparison between our predictions with real data: $\mathcal{L}(\boldsymbol{\theta}|\mathbf{d})\propto\exp[-\chi^2(\boldsymbol{\theta}|\mathbf{d})]$.

(7) Repeat steps 1–6, using an MCMC algorithm to determine the posterior distribution of the model parameters.

In this work, we use a custom implementation of the adaptive Metropolis MCMC method (Haario, Saksman & Tamminen 2001). A chain length of 2×10^6 steps was chosen to ensure the convergence of the posterior distribution. We have verified that this choice of chain length is at least ∼50 times longer than the autocorrelation length for every model parameter.

3 SIMULATIONS AND DATA CONSTRAINTS

3.1 Dark matter halo statistics

As noted in Section 2.1, TRINITY requires only halo population statistics from dark matter simulations, as opposed to individual

Table 2. Summary of priors.

Symbol	Description	Equation	Prior
$\sigma_{*,0}$	Value of σ_* at $z = 0$, in dex	–	$U(0, 0.3)$
μ_0	Value of μ at $z = 0$, in dex	25	$G(0, 0.14)$
μ_a	Redshift scaling of μ, in dex	25	$G(0, 0.24)$
κ	Additional systematic offset in observed versus true SFRs, in dex	26	$G(0, 0.24)$
σ_z	Redshift scaling of σ, in dex	27	$G(0.05, 0.015)$
$\rho_{0.5}$	Correlation between SFR and stellar mass at fixed halo mass at $z = 1$ ($a = 0.5$)	28	$U(0.23, 1)$
$\beta_{\rm BH,0}$	SMBH mass at $M_{\rm bulge} = 10^{11} M_\odot$ and $z = 0$	35	$G(8.46, 0.20)$
$\gamma_{\rm BH,0}$	Slope of the $M_\bullet$–$M_{\rm bulge}$ relation at $z = 0$	36	$G(1.05, 0.14)$

Notes. $G(\mu, \sigma)$ denotes a Gaussian with median μ and width σ, and $U(x_1, x_2)$ denotes a uniform distribution between x_1 and x_2.

Table 3. Summary of observational constraints.

Type	Redshifts	Primarily constrains	References
Stellar mass functions	0–8	SFR–$v_{\rm Mpeak}$ relation	Table 4
Galaxy quenched fractions	0–4	Quenching–$v_{\rm Mpeak}$ relation	Table 5
Cosmic star formation rates	0–10	SFR–$v_{\rm Mpeak}$ relation	Table 6
Specific star formation rates	0–9	SFR–$v_{\rm Mpeak}$ relation	Table 7
Galaxy UV luminosity functions	9–10	SFR–$v_{\rm Mpeak}$ relation	Table 8
Quasar luminosity functions	0–5	Total SMBH accretion	Ueda et al. (2014)
Quasar probability distribution functions	0–2.5	AGN duty cycle, BHAR distributions	Aird et al. (2018)
Active SMBH mass functions	0–5	AGN energy efficiency	Table 9
SMBH mass – bulge mass relation	0	Galaxy–SMBH connection	Table 10
Observed SMBH mass distribution of bright quasars	5.8–6.5	Galaxy-SMBH connection	Shen et al. (2019)

Notes. $v_{\rm Mpeak}$ is the maximum circular velocity of the halo at the time when it reaches its peak mass, $M_{\rm peak}$. This is used as a proxy for the halo mass in TRINITY. BHAR is the SMBH accretion rate.

halo merger trees. We use the peak historical mass ($M_{\rm peak}$) halo mass functions from Behroozi et al. (2013) for the cosmology specified in the introduction. These mass functions are based on central halo mass functions from Tinker et al. (2008), with adjustments to include satellite halo number densities as well as to use $M_{\rm peak}$ instead of the present-day mass. These adjustments were based on the Bolshoi & Consuelo simulations (Klypin, Trujillo-Gomez & Primack 2011). We refer readers to appendix G of Behroozi et al. (2013) for full details. With these calibrations, the halo statistics used in this work are suitable for studying the evolution of haloes from $10^{10} M_\odot$ to $10^{15} M_\odot$. For average halo mass accretion histories, we use the fitting formulae in appendix H of Behroozi et al. (2013). For halo mergers, we fit merger rates from the UNIVERSEMACHINE (Behroozi et al. 2019), with full details and formulae in Appendix A.

3.2 Observational data constraints

We have compiled galaxy and AGN observables from $z = 0$–10, which are summarized in Table 3. The following sections provide brief descriptions of these data.

3.2.1 Galaxy data

Five different observables are used to constrain the halo–galaxy connection in TRINITY: SMFs (Table 4), QFs (Table 5), CSFRs, (Table 6), SSFRs (Table 7), and UVLFs (Table 8). In this work, we adopt the compilation of these observables from Behroozi et al. (2019). Here, we briefly introduce the data sources and the conversions made to ensure consistent physical assumptions across different data sets. For full details, we refer readers to appendix C of Behroozi et al. (2019).

SMFs at $z = 0$–8 come from the following surveys: the Sloan Digital Sky Survey (SDSS, York et al. 2000), the PRIsm MUlti-object Survey (PRIMUS, Coil et al. 2011; Cool et al. 2013), UltraVISTA (McCracken et al. 2012), the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS, Grogin et al. 2011; Koekemoer et al. 2011), and the FourStar Galaxy Evolution Survey (ZFOURGE, Straatman et al. 2016). Data points were converted to be consistent with the Chabrier (2003) IMF, the Bruzual & Charlot (2003) SPS model, and the Calzetti et al. (2000) dust model. Additional corrections were made to homogenize photometry for massive galaxies (see appendix C of Behroozi et al. 2019).

Constraints on galaxy QFs as a function of stellar mass are taken from Bauer et al. (2013), Moustakas et al. (2013), and Muzzin et al. (2013). Each group calculated QFs in a different way, but we assume that they all refer to galaxies with negligible global SFRs (see Section 2.2). Although this results in some uncertainty in the

Table 4. Observational constraints on galaxy stellar mass functions.

Publication	Redshifts	Wavebands	Area (deg^2)
Baldry et al. (2012)	0.002–0.06	*ugriz*	143
Moustakas et al. (2013)	0.05–1	UV-MIR	9
Tomczak et al. (2014)	0.2–3	UV-K_S	0.08
Ilbert et al. (2013)	0.2–4	UV-K_S	1.5
Muzzin et al. (2013)	0.2–4	UV-K_S	1.5
Song et al. (2016)	4–8	UV-MIR	0.08

Table 5. Observational constraints on galaxy quenched fractions.

Publication	Redshifts	Definition of quenching
Bauer et al. (2013)	0–0.3	Observed SSFR
Moustakas et al. (2013)	0.2–1	Observed SSFR
Muzzin et al. (2013)	0.2–4	UVJ diagram

Table 6. Observational constraints on the cosmic star formation rate.

Publication	Redshifts	Waveband	Area (deg^2)
Robotham & Driver (2011)	0–0.1	UV	833
Salim et al. (2007)	0–0.2	UV	741
Gunawardhana et al. (2013)	0–0.35	Hα	144
Ly et al. (2011a)	0.8	Hα	0.8
Zheng et al. (2007)	0.2–1	UV/IR	0.46
Rujopakarn et al. (2010)	0–1.2	FIR	0.4–9
Drake et al. (2015)	0.6–1.5	[OII]	0.63
Shim et al. (2009)	0.7–1.9	Hα	0.03
Sobral et al. (2014)	0.4–2.3	Hα	0.02–1.7
Magnelli et al. (2011)	1.3–2.3	IR	0.08
Karim et al. (2011)	0.2–3	Radio	2
Santini et al. (2009)	0.3–2.5	IR	0.04
Ly et al. (2011b)	1–3	UV	0.24
Kajisawa et al. (2010)	0.5–3.5	UV/IR	0.03
Schreiber et al. (2015)	0–4	FIR	1.75
Planck Collaboration XXX (2014)	0–4	FIR	2240
Dunne et al. (2009)	0–4	Radio	0.8
Cucciati et al. (2012)	0–5	UV	0.6
Le Borgne et al. (2009)	0–5	IR-mm	varies
van der Burg, Hildebrandt & Erben (2010)	3–5	UV	4
Yoshida et al. (2006)	4–5	UV	0.24
Finkelstein et al. (2015)	3.5–8.5	UV	0.084
Kistler, Yuksel & Hopkins (2013)	4–10.5	GRB	varies

Notes. The technique of Le Borgne et al. (2009) (parametric derivation of the cosmic SFH from counts of IR-sub mm sources) uses multiple surveys with different areas. Kistler et al. (2013) used GRB detections from the *Swift* satellite, which has fields of view of ∼3000 deg^2 (fully coded) and ∼10 000 deg^2 (partially coded).

Table 7. Observational constraints on galaxy average specific star formation rates.

Publication	Redshifts	Type	Area (deg^2)
Salim et al. (2007)	0–0.2	UV	741
Bauer et al. (2013)	0–0.35	Hα	144
Whitaker et al. (2014)	0–2.5	UV/IR	0.25
Zwart et al. (2014)	0–3	Radio	1
Karim et al. (2011)	0.2–3	Radio	2
Kajisawa et al. (2010)	0.5–3.5	UV/IR	0.03
Schreiber et al. (2015)	0–4	FIR	1.75
Tomczak et al. (2016)	0.5–4	UV/IR	0.08
Salmon et al. (2015)	3.5–6.5	SED	0.05
Smit et al. (2014)	6.6–7	SED	0.02
Labbé et al. (2013)	7.5–8.5	UV/IR	0.04
McLure et al. (2011)	6–8.7	UV	0.0125

Table 8. Observational constraints on galaxy UV luminosity functions.

Publication	Redshifts	Area (deg^2)
Bouwens et al. (2019)	8–9	0.24
Ishigaki et al. (2018)	8–9	0.016
Oesch et al. (2018)	10	0.23

interpretation of galaxy QFs, it does not affect the main analysis, which only depends on the average SFR as a function of halo mass.

SSFRs and CSFRs at $0 < z < 10.5$ are obtained from multiple surveys (including SDSS, GAMA, UltraVISTA, CANDELS, and ZFOURGE) and techniques (UV, IR, radio, Hα, SED fitting, and gamma-ray bursts). These data points were only corrected to ensure the same initial mass function (the Chabrier 2003 IMF), because aligning other physical assumptions does not improve the self-consistency between SFRs and the growth of SMFs (Madau & Dickinson 2014; Leja et al. 2015; Tomczak et al. 2016).

In this work, we also use UVLFs from Ishigaki et al. (2018), Oesch et al. (2018), and Bouwens et al. (2019) at $z = 9$–10 to constrain the halo–galaxy connection beyond the redshift coverage of SMFs.

In this paper, we have assumed a non-evolving IMF from Chabrier (2003). With IMFs from Kroupa (2001) and Salpeter (1955), the inferred stellar masses would be factors of 1.07 and 1.7 higher than using the Chabrier (2003) IMF, respectively. For SFRs, these factors are 1.06 and 1.58, respectively (Salim et al. 2007). More generally, a top-heavy IMF would produce a higher fraction of massive stars, decreasing the mass-to-UV light ratios of galaxies, and ultimately the inferred stellar masses and SFRs from SPS. There is some observational evidence that the IMF becomes more top-heavy with increasing SFR (e.g. Gunawardhana et al. 2011), but it remains an open issue whether IMF varies with environment or redshift (Conroy, Gunn & White 2009; Bastian, Covey & Meyer 2010; van Dokkum & Conroy 2012; Krumholz 2014; Lacey et al. 2016). Therefore, we opt to use a universal IMF in this paper; for discussion on the potential effects of non-universal IMFs, we refer readers to appendix G of Behroozi et al. (2019).

3.2.2 *SMBH data*

There are five different kinds of SMBH observables in our compiled data set: QLFs, QPDFs, ABHMFs, the local SMBH mass–bulge mass ($M_\bullet$–$M_{\rm bulge}$) relation, and the observed SMBH mass distribution of high-redshift bright quasars. These SMBH data are summarized in Table 9 (QLFs, QPDFs, and ABHMFs) and Table 10 ($M_\bullet$–$M_{\rm bulge}$).

We have used bolometric QLFs at $z = 0$–5 from Ueda et al. (2014), which are based on a series of X-ray surveys. There are also QLFs based on observations in other wavebands (e.g. UVLFs from Kulkarni, Worseck & Hennawi 2019), but we use those from X-ray surveys due to their uniformity in AGN selection and robustness against (moderate) obscuration. We adopted the empirical correction scheme from Ueda et al. (2014) to account for Compton-thick AGN populations (see Appendix D2 for full details). We also tested using bolometric QLFs from multiple wavebands from Shen et al. (2020), and found no qualitative changes in our results. The posterior distribution of model parameters does change significantly if assuming QLFs and Compton-thick corrections from Ananna et al. (2019). However, there is strong inconsistency between these luminosity functions and the QPDFs from Aird et al. (2018). In light of this, we do not adopt Ananna et al. QLFs in the main text. For further details, we refer readers to Appendix D2.

QLFs constrain the total radiative energy output of active SMBHs (Conroy & White 2013; Caplar et al. 2015). To constrain the mass-dependence of AGN luminosity distributions, we included QPDFs from Aird et al. (2018). These functions are expressed as the conditional probability distributions of $sL_X \equiv L_X/(1.04 \times 10^{34} {\rm erg\,s^{-1}} \times M_*/M_\odot)$. These distributions are given as functions of stellar mass (M_*) and redshift, and are obtained by modelling the X-ray luminosities of galaxies in the CANDELS and UltraVISTA surveys. Aird et al. (2018) did not correct for the presence of Compton-thick AGNs in their modelling, so we adopted the empirical scheme given by Ueda et al. (2014) to correct our predicted QPDFs for this selection bias (see Appendix D2 for more details).

Table 9. Observational constraints on AGNs.

Publication	Type	Redshifts	Waveband	Area (deg^2)
Ueda et al. (2014)	Luminosity functions	0–5	X-ray	0.12–34 000
Aird et al. (2018)	AGN probability distribution functions	0.1–2.5	X-ray	0.22–1.6
Schulze & Wisotzki (2010)	Active black hole mass functions	0–0.3	Optical	9500
Schulze et al. (2015)	Active black hole mass functions	1–2	Optical	0.62–6250
Kelly & Shen (2013)	Active black hole mass functions	1.5–5	Optical	6250
Shen et al. (2019)	Observed SMBH mass distribution of bright quasars	5.8–6.5	Optical	14 000

Notes. 'Waveband' indicates the waveband used to measure SMBH properties. Aird et al. (2018) additionally used UV, optical, and IR data to constrain host galaxy properties.

Table 10. Observational constraints on the SMBH mass–bulge mass ($M_\bullet$–$M_{\rm bulge}$) relation at $z = 0$.

Publication	$\beta_{\rm BH}$	$\gamma_{\rm BH}$
Häring & Rix (2004)	8.20	1.12
Beifiori et al. (2012)	8.25	0.79
Kormendy & Ho (2013)	8.69	1.15
McConnell & Ma (2013)	8.46	1.05
Savorgnan et al. (2016)	8.55	1.05
Median	8.46	1.05
Standard deviation	0.20	0.14

Notes. The median $M_\bullet$–$M_{\rm bulge}$ relation is assumed to be a power law: $\log_{10}(M_\bullet/M_\odot) = \beta_{\rm BH} + \gamma_{\rm BH}\log_{10}(M_{\rm bulge}/10^{11}M_\odot)$.

In modelling how AGN luminosity connects to SMBH growth, there is a degeneracy between the SMBH accretion rate and the radiative efficiency. To break this degeneracy, we include (1) ABHMFs from $z = 0.2$–5 from Schulze & Wisotzki (2010), Kelly & Shen (2013), and Schulze et al. (2015); and (2) the local $M_\bullet$–$M_{\rm bulge}$ relation to constrain the total amount of SMBH mass accreted over cosmic time. Given the different sample selection criteria and data reduction schemes used by different groups, we decided not to use individual data points for the $M_\bullet$–$M_{\rm bulge}$ relation. Instead, we picked five commonly used local $M_\bullet$–$M_{\rm bulge}$ relations and calculated the medians and standard deviations of their slopes and intercepts (see Table 10). We then apply Gaussian priors on both the slope and the intercept at $z = 0$ in TRINITY, with the centres and widths set to these medians and standard deviations.

Given the capability of contemporary telescopes, the sample of $z \gtrsim 5$ AGNs is likely biased against faint objects. However, the observed SMBH mass distribution of these high-redshift quasars still provides useful constraints on TRINITY. Specifically, we know from observations that few quasars with $L_{\rm bol} > 10^{47}$ erg s^{-1} at $5.8 < z < 6.5$ have *observed* $M_\bullet < 10^8 M_\odot$ (Shen et al. 2019). Therefore, the expected number of these quasars in TRINITY, $N_{\rm exp}$, should also be small. Assuming Poisson statistics, the prior probability that we detect no low-mass bright quasars with a survey like SDSS is

$$P(N_{\rm obs} = 0|N_{\rm exp}) = \exp\left(-N_{\rm exp}\right) \tag{77}$$

$$\begin{aligned} N_{\rm exp} = & \int_0^{10^8} P(M_{\bullet,\rm obs}|M_{\bullet,\rm int}){\rm d}M_{\bullet,\rm obs} \\ & \times \int_0^\infty {\rm d}M_{\bullet,\rm int} \int_{10^{47}}^\infty {\rm d}L_{\rm bol}\, P(L_{\rm bol}|M_{\bullet,\rm int})\phi_{\rm BH}(M_{\bullet,\rm int}) \\ & \times \mathcal{S}_{\rm SDSS} \times \Delta z \end{aligned} \tag{78}$$

$$\begin{aligned} P(M_{\bullet,\rm obs}|M_{\bullet,\rm int}) = & \frac{1}{\sqrt{2\pi}\sigma_{\rm BH,obs}} \\ & \times \exp\left[-\frac{(\log M_{\bullet,\rm obs} - \log M_{\bullet,\rm int})^2}{2\sigma^2_{\rm BH,obs}}\right], \end{aligned} \tag{79}$$

where $M_{\bullet,\rm int}$ and $M_{\bullet,\rm obs}$ are the intrinsic and observed SMBH masses, respectively, and $\sigma_{\rm BH,obs} = 0.4$ dex is the random scatter in SMBH mass as induced by virial estimates (Park et al. 2012). $\mathcal{S}_{\rm SDSS} = 14\,000$ deg^2 is the survey area of SDSS. Here, we take $\Delta z = 6.5$–$5.8 = 0.7$ to keep consistency with Shen et al. (2019). In the MCMC process, we included this prior to prevent TRINITY from producing too many low-mass and super-Eddington quasars, which are not supported by observations (e.g. Mazzucchelli et al. 2017; Trakhtenbrot, Volonteri & Natarajan 2017).

In the process of compiling these data, we found systematic discrepancies between some observational data sets, which are addressed in Appendices D4 (quasar X-ray luminosities) and D5 (ABHMFs).

4 RESULTS

We present the best-fitting parameters and the comparisons to observations in Section 4.1, as well as results for the evolution of the $M_\bullet$–$M_{\rm bulge}$ relation in Section 4.2, black hole accretion rates and Eddington ratio distributions in Section 4.3, the SMBH mass function in Section 4.4, SMBH mergers in Section 4.5, AGN energy efficiency as well as systematic uncertainties in Section 4.6, and the correlation coefficient between average SMBH accretion rate and $M_\bullet$ at fixed halo mass in Section 4.7.

4.1 Best-fitting parameters and comparison to observables

We obtained the posterior distribution of model parameters with an MCMC algorithm (Section 2.9). The best-fitting model was found by the following two-step procedure: (1) calculate the weighted average of the 2000 highest-probability points in the MCMC chain; (2) starting from this weighted average, run a gradient descent optimization over each dimension of the parameter space, until the model χ^2 stops changing.

Our best-fitting model is able to fit all the data in our compilation (Section 3), including SMFs (Fig. 3, left-hand panel), QFs (Fig. 3, right-hand panel), CSFRs (Fig. 4, left-hand panel), SSFRs (Fig. 4, right-hand panel), galaxy UVLFs (Fig. 5), QLFs (Fig. 6), ABHMFs (Figs 7 and 8), QPDFs (Fig. 9), and the local $M_\bullet$–$M_{\rm bulge}$ relation (Fig. 10). For 1189 data points and 56 parameters, the naive reduced χ^2 is 0.66, which suggests a reasonable fit. The best-fitting model and 68 per cent confidence intervals for parameters are presented in Appendix H.

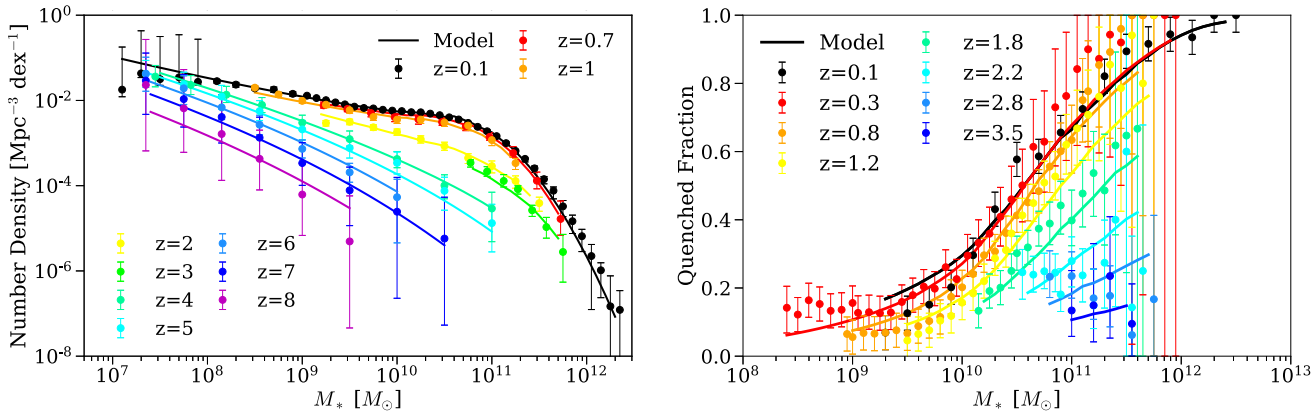

Figure 3. Left-hand panel: Comparison between observed galaxy SMFs and our best-fitting model from $z = 0$–8. The observed SMFs are listed in Table 4. Right-hand panel: Comparison between observed galaxy QFs and our best-fitting model from $z = 0$–4. The observed QFs are listed in Table 5. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

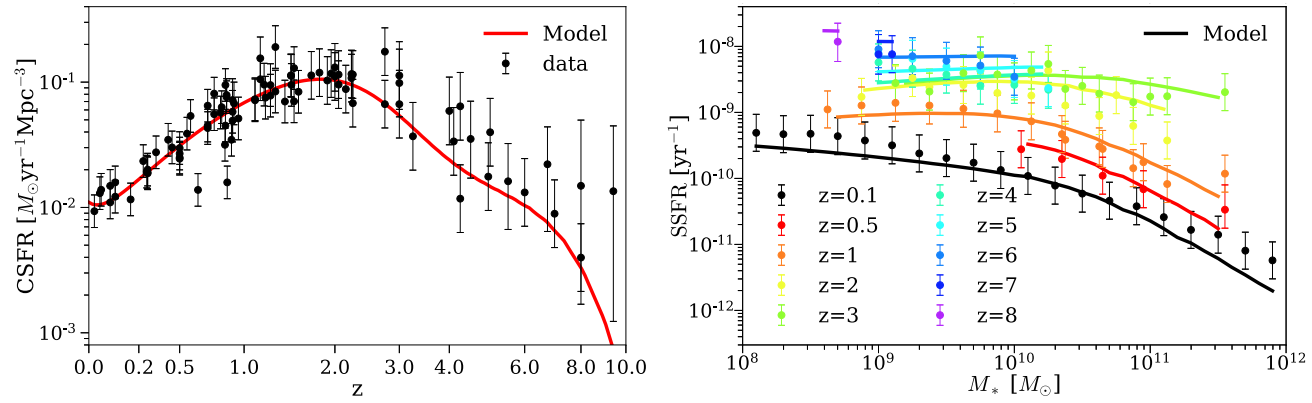

Figure 4. Left-hand panel: Comparison between observed CSFRs and our best-fitting model from $z = 0$–10. The references for observations are listed in Table 6. Right-hand panel: Comparison between observed galaxy SSFRs as a function of stellar mass and our best-fitting model from $z = 0$–8. The references for observations are listed in Table 7. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

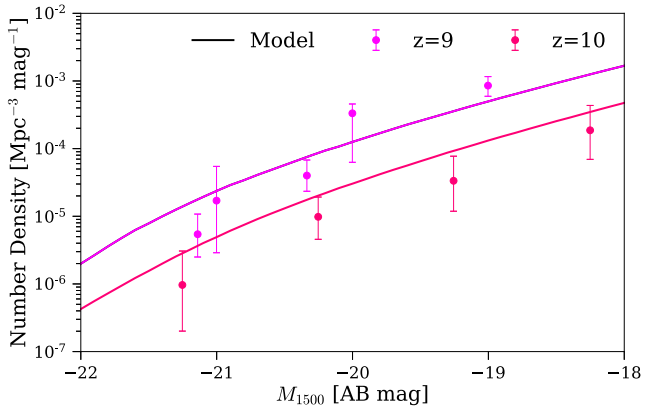

Figure 5. Comparison between observed galaxy UVLFs and our best-fitting model from $z = 9$–10. The references for observations are listed in Table 8. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

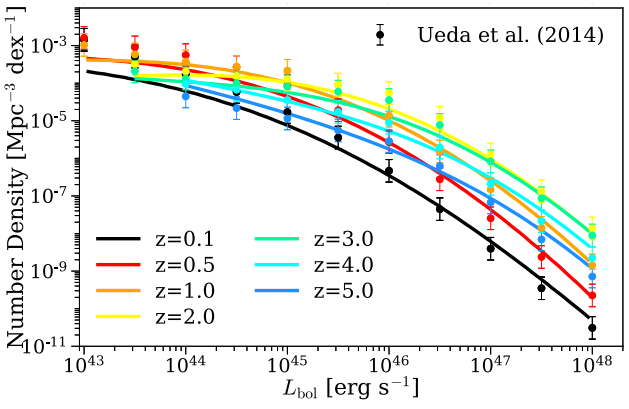

Figure 6. Comparison between the observed QLFs from Ueda et al. (2014) and our best-fitting model from $z = 0$–5. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

As shown in Fig. 9, TRINITY largely reproduces the mass-dependence of the QPDFs from Aird et al. (2018), but it does not fully recover the QPDF shape for galaxies with $M_* < 10^{10} M_\odot$. Specifically, TRINITY tends to overpredict active AGNs in these low-mass galaxies at $z > 1$. Given the complexity of the models adopted by Aird et al. (2018) to calculate these QPDFs, we did not add additional free

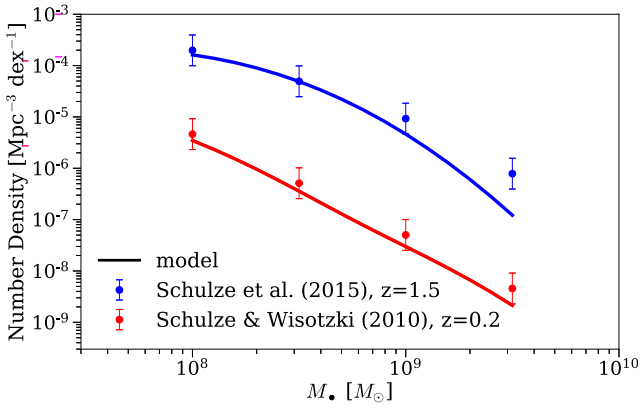

Figure 7. Comparison between the observed ABHMFs from Schulze & Wisotzki (2010) and Schulze et al. (2015), and our best-fitting model at $z = 0.2$ and $z = 1.5$. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

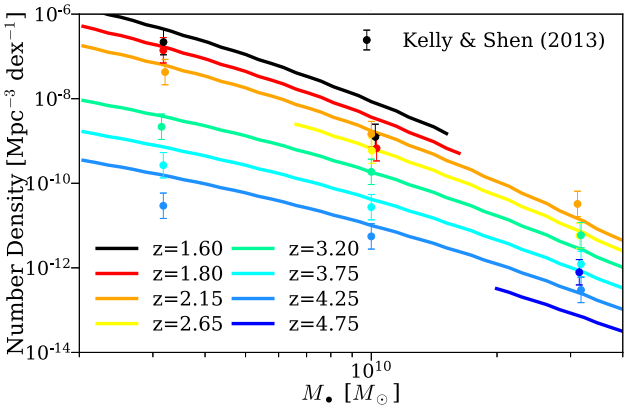

Figure 8. Comparison between the observed ABHMFs from Kelly & Shen (2013) and our best-fitting model from $z = 1.5$–5. The data points and the best-fitting models in each higher redshift bin are shifted downwards by 0.5 dex incrementally for the sake of clarity. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

parameters to fully reproduce their shapes, which reduces the risk of overfitting.

4.2 The $M_\bullet$–$M_{\rm bulge}$ relation for $z = 0$ to $z = 10$

In Fig. 11, we show the redshift evolution of the *median* SMBH mass–bulge mass ($M_\bullet$–$M_{\rm bulge}$) relation (top panel) along with the lognormal scatter (bottom panel) from $z = 0$–10. We find that both the slope and the normalization of the median $M_\bullet$–$M_{\rm bulge}$ relation increase mildly from $z = 0$–10. From $z = 0$–3, the evolution in the median $M_\bullet$ at fixed $M_{\rm bulge}$ is at most ∼0.3 dex, which is within the typical SMBH mass uncertainties. The median $M_\bullet$–$M_{\rm bulge}$ relation beyond $z = 0$ is jointly constrained by the QLFs, QPDFs, ABHMFs, and the galaxy SMFs. Specifically, QLFs and QPDFs jointly constrain the Eddington ratio distributions and duty cycles of SMBHs. On the other hand, ABHMFs specify the abundances of *active* SMBHs as a function of their masses. Combined with the Eddington ratio distributions and duty cycles, this information helps TRINITY infer the number density of *active* + *dormant* SMBHs at different masses, i.e. the *total* SMBH mass functions. Reproducing these SMBH mass functions given the observed number density of galaxies (i.e. their SMFs) places strong constraints on the $M_\bullet$–$M_{\rm bulge}$ relation. At $z \geq 8$ (shown in Fig. 11 as dashed lines), the median $M_\bullet$ at fixed bulge mass is lower compared to the $z = 0$ values, but consistent within the statistical uncertainties from MCMC. Without existing SMBH data at this cosmic era, we expect that future observations [by e.g. the *James Webb Space Telescope* (*JWST*)] will test our predictions. It is likely that many future observations can only probe the most massive SMBHs at such high redshifts, but they will still provide useful tests as to whether their number densities are consistent with the median $M_\bullet$–$M_{\rm bulge}$ relation and the scatter around it.

The scatter around the median $M_\bullet$–$M_{\rm bulge}$ relation is $\sigma_{\rm BH} \approx 0.27$ dex. As described in Section 2.5, a lognormal scatter of $\sigma_{\rm BH}$ causes an offset between the *median* and *mean* SMBH masses (Section 2.5) at fixed stellar mass. Mean SMBH masses directly influence average BHARs, which are constrained by observed QLFs and QPDFs. Consequently, $\sigma_{\rm BH}$ is primarily constrained by (a) the evolution of the median $M_\bullet$–$M_{\rm bulge}$ relation; and (b) the average BHARs inferred from QLFs and QPDFs. Another constraint comes from the shape of ABHMFs, since bigger scatter would produce more overmassive SMBHs in low-mass galaxies than undermassive SMBHs in high-mass galaxies. Therefore, flatter ABHMFs implies a bigger scatter around the $M_\bullet$–$M_{\rm bulge}$ relation.

In Fig. 12, we show the evolution of the *mean* $M_\bullet$–$M_{\rm bulge}$ relation from $z = 0$–10. With $\sigma_{\rm BH} \approx 0.27$ dex, the mean relation is offset from the median relation by a constant factor of $0.5\sigma_{\rm BH}^2 \ln 10 \approx 0.08$ dex.

Fig. 13 shows the best-fitting median SMBH mass–galaxy total stellar mass ($M_\bullet$–M_*) relation. Our $z = 0$ $M_\bullet$–M_* relation is consistent with measurements by Greene et al. (2016) using water megamaser disc observations. This relation is qualitatively similar to the $M_\bullet$–$M_{\rm bulge}$ relation mainly because of the approximate proportionality between $M_{\rm bulge}$ and M_* (equation 16). Quantitatively, the evolution of the $M_\bullet$–M_* relation in the range $0 < z < 2$ is less significant than that of the $M_\bullet$–$M_{\rm bulge}$ relation, due to lower $M_{\rm bulge}/M_*$ ratios at higher redshifts, which is also consistent with observational studies like Ding et al. (2020). The evolution of the $M_\bullet$–M_* relation causes the median $M_\bullet/M_*$ ratio (Fig. 13, bottom panel) to decrease with redshift. Overall, the mild evolution is consistent with observational studies that found no significant redshift dependence in the $M_\bullet$–$M_{\rm bulge}$ and $M_\bullet$–M_* relations in the range $0 < z < 2$ (e.g. Schramm & Silverman 2013; Sun et al. 2015; Suh et al. 2020).

Fig. 14 shows the best-fitting median SMBH mass–halo peak mass ($M_\bullet$–$M_{\rm peak}$) relation. At $z \lesssim 5$, the $M_\bullet$–$M_{\rm peak}$ relation can be approximated as a double power law, connected by a knee at $M_{\rm peak} \sim 10^{12} M_\odot$. Above $z = 5$, it is roughly a single power law due to the lack of massive haloes. This halo mass dependence is inherited from the well-known SMHM (M_*–$M_{\rm peak}$) relation, because of the approximate single power-law shapes of the $M_\bullet$–M_* connection (Fig. 13; see also Kormendy & Ho 2013).

The top panel of Fig. 15 shows the median SMBH mass ($\widetilde{M}_\bullet$) as a function of $M_{\rm peak}$ and z. From $z = 0$–10, SMBH masses in haloes with $M_{\rm peak} \sim 10^{11} M_\odot$ remain consistently low. But SMBHs do grow in mass along with their host haloes/galaxies, as indicated by the halo growth curves (white solid lines).

The bottom panel of Fig. 15 shows the $\widetilde{M}_\bullet$ histories along the growth histories of different haloes. At all halo masses, SMBH growth is very fast in the early universe, and slows down towards lower redshifts. However, the fast-growth phase ends earlier for more massive black holes. This is consistent with the phenomenon called 'AGN downsizing' (e.g. Merloni 2004; Barger et al. 2005), and we discuss this further in Section 4.3 and Section 5.3.

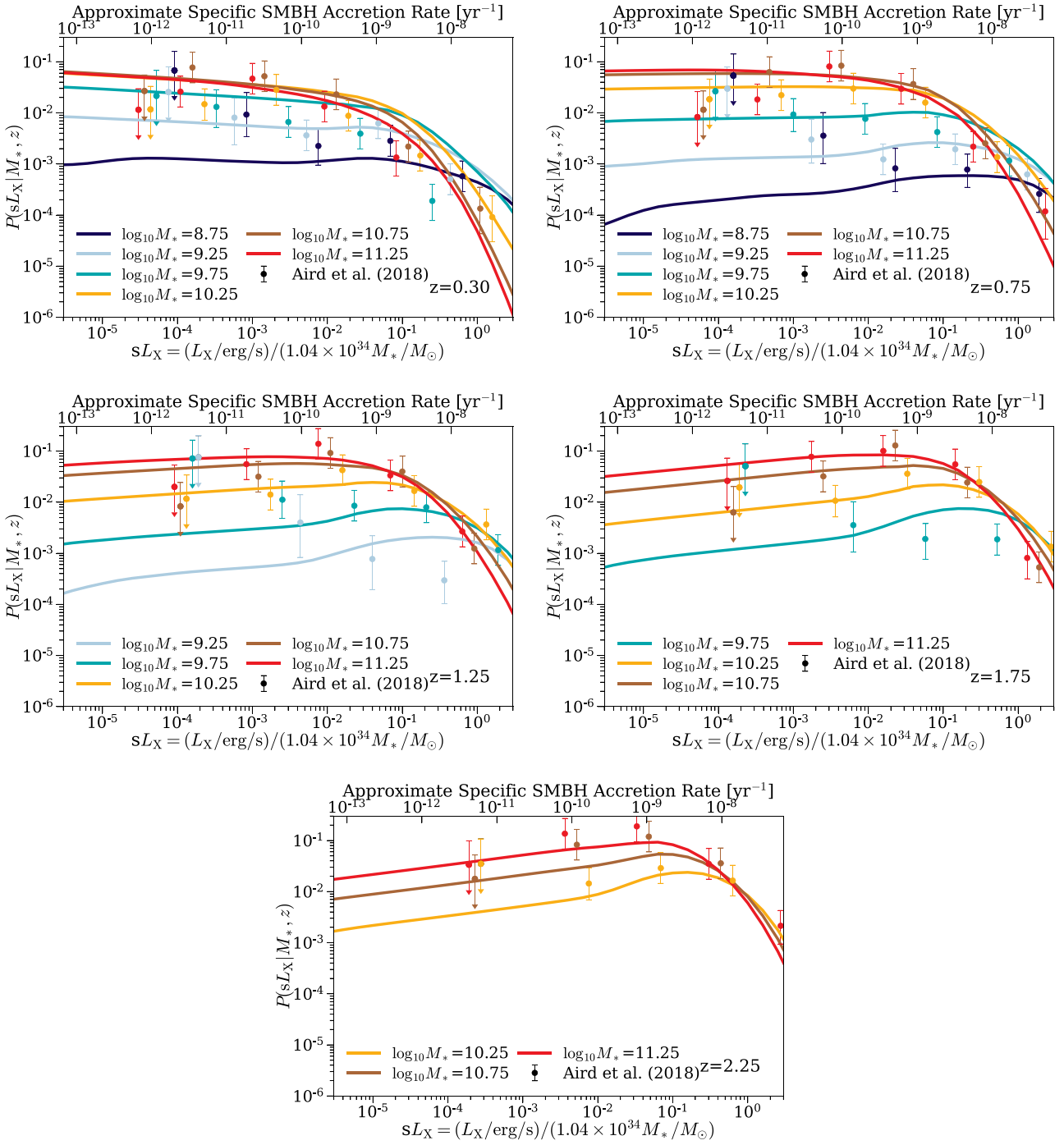

Figure 9. The comparison between the observed QPDFs from Aird et al. (2018) and our best-fitting model from $z = 0$–2.5. The data points include Compton-thin AGNs only, so the model values are corrected for direct comparison. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

4.3 Average black hole accretion rates and Eddington ratio distributions

The top panel of Fig. 16 shows the average black hole accretion rate ($\overline{\mathrm{BHAR}}$) as a function of M_peak and z. In general, BHARs peak at $M_\mathrm{peak} \sim 10^{12}M_\odot$, and decrease towards lower and higher masses. Below $z \sim 2$ and $M_\mathrm{peak} \sim 10^{13.5}M_\odot$, BHARs decrease with time at fixed mass. At $z \sim 2$, there is also a slight increase in BHAR towards higher halo mass. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_SF is 0.5 as a function of redshift. Below (above) this dashed line, the mass growth of SMBHs occurs primarily in star-forming (quenched) galaxies. In TRINITY, average BHARs are constrained by the total energy output from AGNs, which is mainly inferred from the QPDFs and ABHMFs.

The bottom panel of Fig. 16 shows the average BHAR histories of haloes with different masses at $z = 0$. At all halo masses, average BHARs keep rising in the early universe, and then peak and decrease towards lower redshifts. The BHARs of more massive haloes peak at

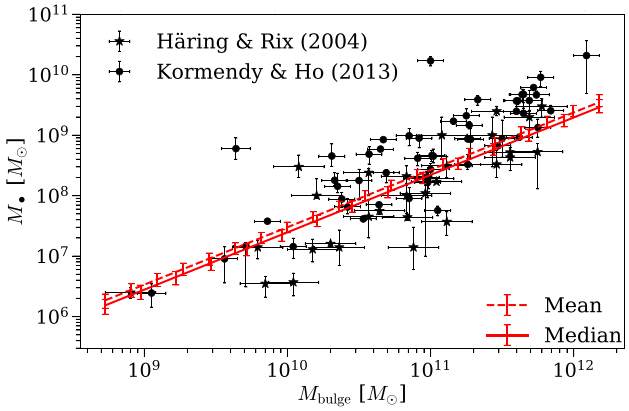

Figure 10. The local $M_\bullet$–$M_{\rm bulge}$ relation. The filled circles are the data compiled by Kormendy & Ho (2013), and the stars are those compiled by Häring & Rix (2004). The red solid line is the median $M_\bullet$–$M_{\rm bulge}$ mass relation, and the red dashed line is the mean relation. These lines are offset because lognormal distributions are positively skewed, with the mean being greater than the median. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

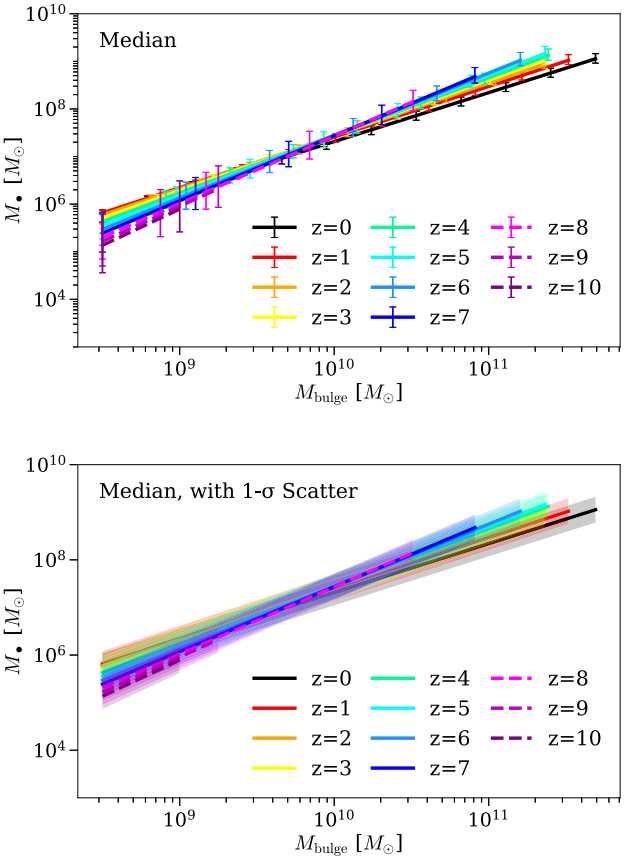

Figure 11. The evolution of the median $M_\bullet$–$M_{\rm bulge}$ relation and the corresponding lognormal scatter from $z = 0$–10. Top panel: the median relations (see Section 4.2). The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. Bottom panel: The same median relations, except that the shaded regions show the *lognormal scatter* around the median relations. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by e.g. *JWST*). All the data used to make this plot can be found here.

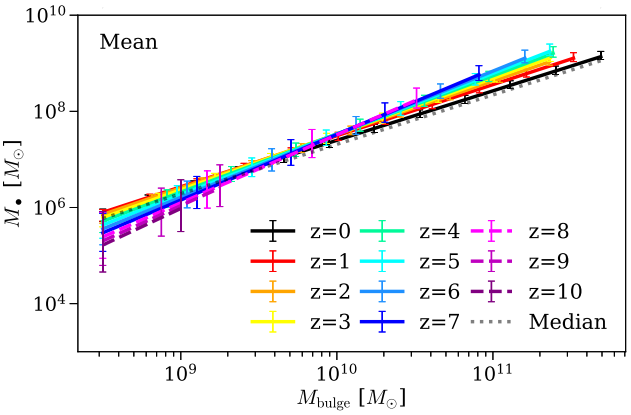

Figure 12. The evolution of the mean $M_\bullet$–$M_{\rm bulge}$ relation from $z = 0$–10 (see Section 4.2). The grey dotted line shows the median relation at $z = 0$ for comparison. The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by e.g. *JWST*). All the data used to make this plot can be found here.

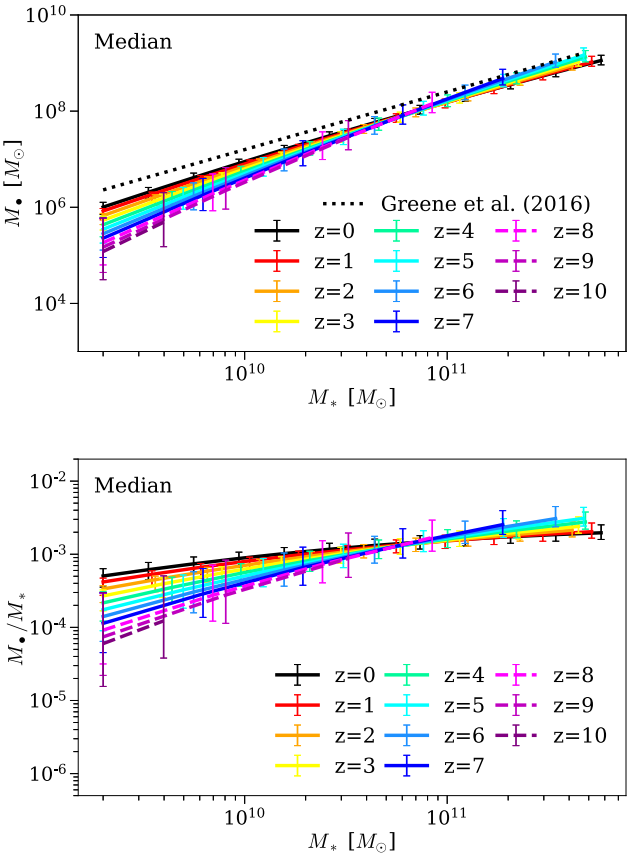

Figure 13. Top panel: the best-fitting median $M_\bullet$–M_* relation from $z = 0$–10 (solid lines, see Section 4.2), and the observed $z = 0$ $M_\bullet$–M_* relation from Greene et al. (2016) (dotted line). Bottom panel: the best-fitting median $M_\bullet/M_*$ ratios as a function of M_* and z. The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by e.g. *JWST*). All the data used to make this plot can be found here.

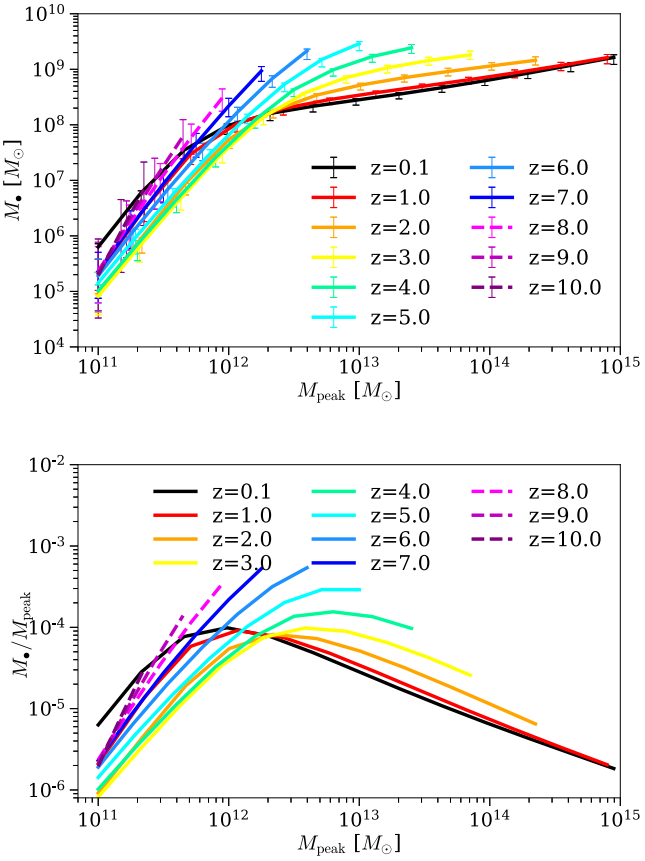

Figure 14. Top panel: the best-fitting median $M_\bullet$–$M_{\rm peak}$ (peak halo mass) relation from $z = 0$–10 (see Section 4.2). Bottom panel: the best-fitting $M_\bullet/M_{\rm peak}$ ratios as a function of $M_{\rm peak}$ and z. The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by e.g. *JWST*). All the data used to make this plot can be found here.

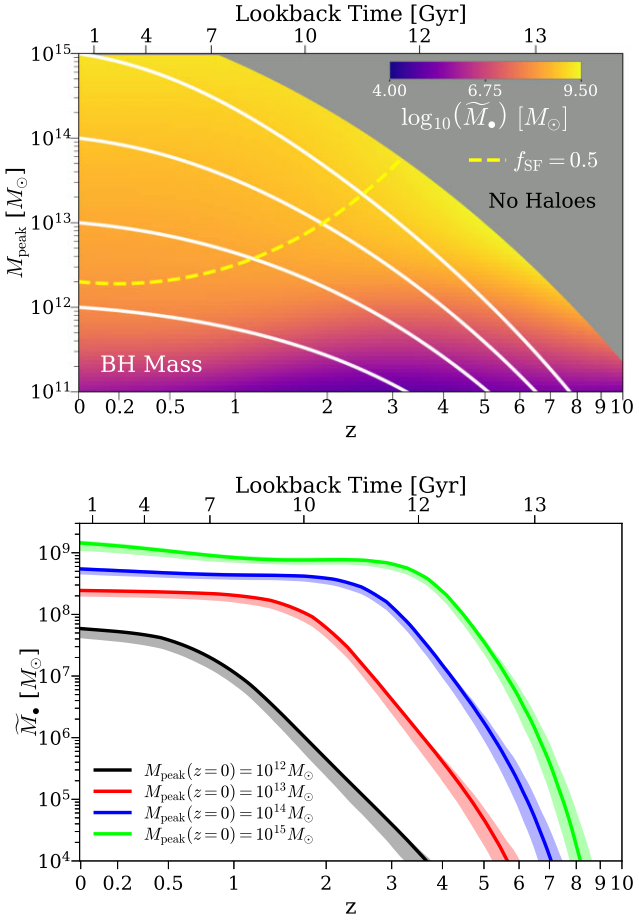

Figure 15. Top panel: the median SMBH mass ($\widetilde{M}_\bullet$) as a function of $M_{\rm peak}$ and z (see Section 4.2). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. Bottom panel: The $\widetilde{M}_\bullet$ histories as a function of halo mass at $z = 0$. The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

higher redshifts. There is also an increase in BHAR with time below $z \sim 2$ among the most massive haloes. This is mainly constrained by the increase in AGN luminosities with stellar mass, as indicated by the low-redshift QPDFs from fig. 5 of Aird et al. (2018).

Fig. 17 shows the average galaxy SFRs as a function of $M_{\rm peak}$ and z. The $M_{\rm peak}$ and z dependencies of SFR are similar to those of BHAR below $M_{\rm peak} \sim 10^{14} M_\odot$. Above $M_{\rm peak} \sim 10^{14} M_\odot$, however, SFR decreases monotonically with halo mass at all redshifts, whereas the massive black holes still have detectable accretion rates. In other words, BHARs follow SFRs mainly among less-massive haloes, where star-forming galaxies dominate the population. For massive galaxies at lower redshifts, they are much more likely to be quiescent in their SFRs, but still have significant SMBH activity. This difference between small and large galaxy populations is hidden when we compare the *cosmic* BHARs and SFRs, where less massive objects ($M_{\rm peak} \sim 10^{12} M_\odot$) dominate the demographics.

The top panel of Fig. 18 shows the ratios between the average BHAR and SFR, $\overline{\rm BHAR}/\overline{\rm SFR}$, as a function of $M_{\rm peak}$ and z. At $z \gtrsim 6$, $\overline{\rm BHAR}/\overline{\rm SFR}$ increases with increasing $M_{\rm peak}$. Towards lower redshifts, $\overline{\rm BHAR}/\overline{\rm SFR}$ grows more slowly for all haloes, and shows a plateau at $\overline{\rm BHAR}/\overline{\rm SFR} \sim 10^{-3}$. More massive haloes reach this plateau at higher redshifts, which is consistent with the downsizing of SMBH growth. Below $z \sim 2$, however, the mass dependency gets stronger again, in the sense that more massive haloes have higher $\overline{\rm BHAR}/\overline{\rm SFR}$. Physically, this is because massive galaxies are strongly quenched towards lower redshifts, but the mass accretion of massive black holes is not suppressed as much. The bottom panel of Fig. 18 shows the $\overline{\rm BHAR}/\overline{\rm SFR}$ histories of different halo populations. At $z \lesssim 2$, $\overline{\rm BHAR}/\overline{\rm SFR}$ either stays at a similar level as $z \gtrsim 2$, or increases with time for essentially all halo populations, indicating that SMBHs are catching up with galaxies in their growth.

The top panel of Fig. 19 shows the average SMBH total Eddington ratio ($\overline{\eta}$) as a function of $M_{\rm peak}$ and z. At $z \gtrsim 7$, all SMBHs have $0.1 < \overline{\eta} < 1$ regardless of host halo mass. At lower redshifts, the average Eddington ratio decreases, with stronger trends for higher halo masses. In other words, SMBHs are less active in massive haloes and/or at later cosmic times. A similar trend can be seen when we follow the growth of different haloes, as shown by the white solid curves. In the bottom panel, we see all SMBHs accreting rapidly at high redshifts, with average Eddington ratios of unity at $z \sim 10$. Below $z = 10$, Eddington ratios drop with time for all SMBHs, but the exact patterns differ among halo populations. For more massive

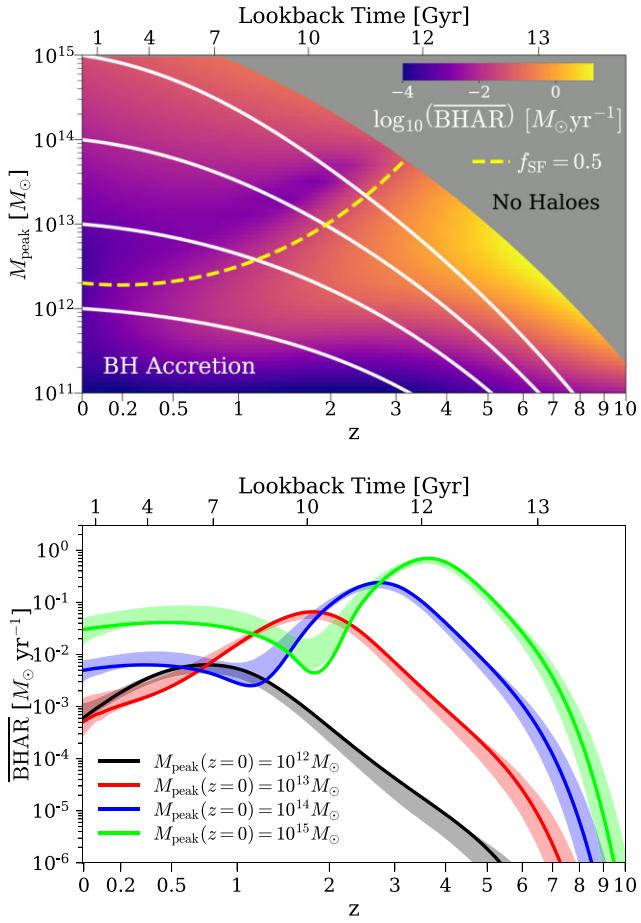

Figure 16. Top panel: average black hole accretion rate ($\overline{\mathrm{BHAR}}$) as a function of $M_{\rm peak}$ and z (see Section 4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. Bottom panel: $\overline{\mathrm{BHAR}}$ histories as a function of halo mass at $z = 0$. The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

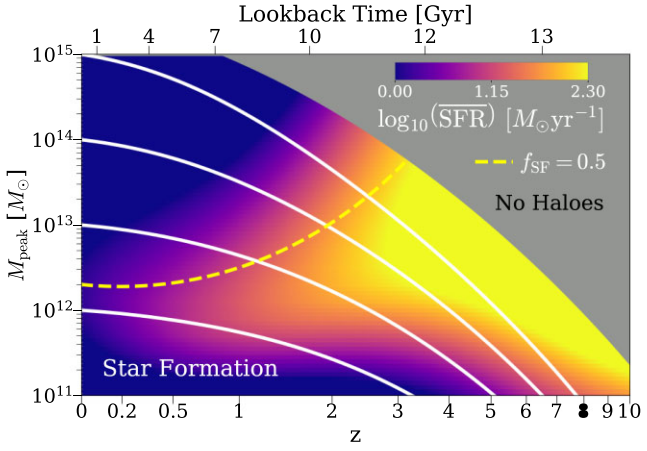

Figure 17. The average star formation rates ($\overline{\mathrm{SFR}}$) as a function of $M_{\rm peak}$ and z (see Section 4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

haloes with $M_{\rm peak} > 10^{13} M_\odot$, the average Eddington ratios experience a two-phase decline before the final slight rejuvenation: an initial, slower decrease, and a later, faster drop. Haloes with $M_{\rm peak} = 10^{12}$–$10^{13} M_\odot$ at $z = 0$ do not experience the final flattening phase in Eddington ratio. Below $z \sim 4$, more massive haloes experience the final and faster decline in Eddington ratios earlier compared to less massive ones. As the bottom panel of Fig. 15 shows, this also reflects the same 'AGN downsizing' phenomenon: SMBH activity starts to decline earlier in more massive haloes/galaxies.

It should be pointed out that the 'AGN downsizing' effect exists not only when we look at different halo populations, but also when we look at SMBHs with different masses. Fig. 20 shows the average SMBH *total* (i.e. radiative+kinetic) Eddington ratio, $\overline{\eta}$, as a function of $M_\bullet$ and z. Again, we see that at high redshifts, SMBHs of different masses accrete at similar Eddington ratios. Below $z \sim 3$, the activity level among more massive black holes starts to decline earlier. Consequently, we see that $\overline{\eta}$ decreases towards higher $M_\bullet$.

4.4 SMBH mass functions

Fig. 21 shows the total black hole mass functions (BHMFs) for $0 \leq z \leq 10$. Similar to the galaxy SMFs, the 'knee' in the BHMF becomes less and less significant towards higher redshifts. This is because, in the early universe, the M_*–$M_{\rm peak}$ relation, and therefore the $M_\bullet$–$M_{\rm peak}$ relation, can be approximated as a single power law. We also see strong evolution in the BHMF above $z \gtrsim 5$ *regardless of SMBH mass*. This directly results from the universally high Eddington ratios at high redshifts. (see also Section 4.3). At $z < 3$, the AGN downsizing effect slows down the evolution of the total BHMF at the massive end. In the meantime, moderately massive SMBHs with $10^8 < M_\bullet < 10^9 M_\odot$ grow significantly. This continued growth builds up the 'knee' in the BHMF in the low-redshift universe.

4.4.1 *The host haloes of $M_\bullet > 10^{9.5} M_\odot$ SMBHs*

In Fig. 22, we show the total BHMFs at $z = 0.0$, 0.5, and 1.0, decomposed into contributions from different host halo masses. Similar to equation (63), the BHMF contributed by haloes in the mass range ($M_{\rm peak,min}$, $M_{\rm peak,max}$) is

$$\phi(M_\bullet, M_{\rm peak,min}, M_{\rm peak,max}, z) = \int_{M_{\rm peak,min}}^{M_{\rm peak,max}} \phi(M_{\rm peak}, z) \times P(M_\bullet | M_{\rm peak}, z) \mathrm{d}M_{\rm peak}, \quad (80)$$

where $\phi(M_{\rm peak}, z)$ is the halo mass function and $P(M_\bullet|M_{\rm peak}, z)$ is the probability distribution of $M_\bullet$, given the host halo mass $M_{\rm peak}$ at redshift z. In TRINITY, $P(M_\bullet|M_{\rm peak}, z)$ is a lognormal distribution with the median and scatter determined from the halo–galaxy–SMBH connection (Section 2.2 and Section 2.4). Given the flat $M_\bullet$–$M_{\rm peak}$ relation at the massive end (see Fig. 14), $P(M_\bullet|M_{\rm peak}, z)$ only changes slightly with increasing halo mass. On the other hand, there are many fewer haloes with $M_{\rm peak} > 10^{14} M_\odot$ than $M_{\rm peak} < 10^{14} M_\odot$, due to the exponential decrease in halo number density. Hence, the haloes with $10^{13} M_\odot < M_{\rm peak} < 10^{14} M_\odot$, rather than those with $10^{14} M_\odot < M_{\rm peak} < 10^{15} M_\odot$, dominate the BHMF for $M_\bullet > 10^{9.5} M_\odot$ at $z = 1.0$. In other words, when looking at an $M_\bullet$-selected sample with large $M_\bullet$, we are more likely to observe less massive haloes

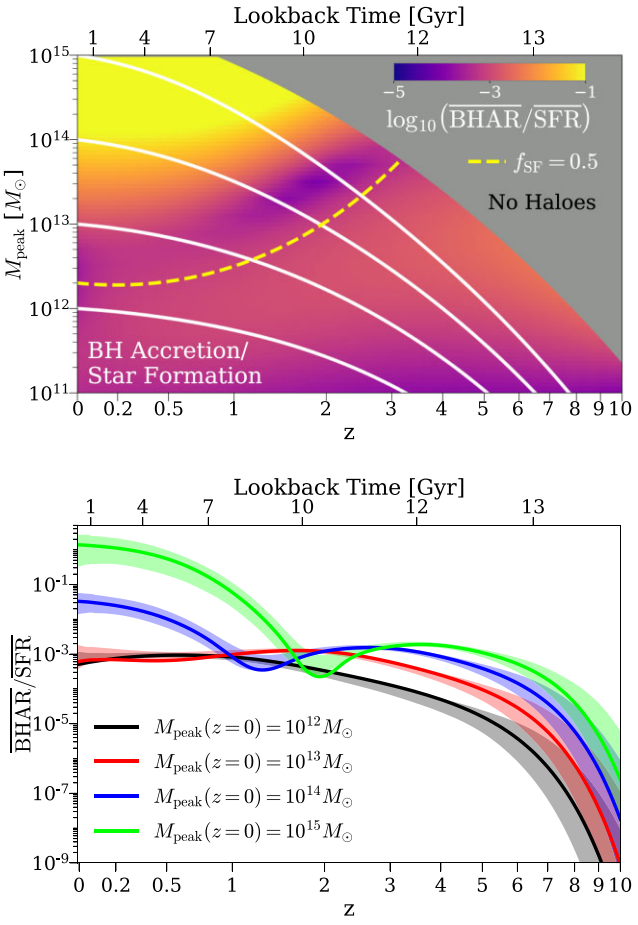

Figure 18. Top panel: the $\overline{\mathrm{BHAR}}/\overline{\mathrm{SFR}}$ ratio as a function of redshift and M_{peak} for our best-fitting model (see Section 4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\mathrm{peak}} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. Bottom panel: the $\overline{\mathrm{BHAR}}/\overline{\mathrm{SFR}}$ ratio histories as a function of M_{peak} at $z = 0$. The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

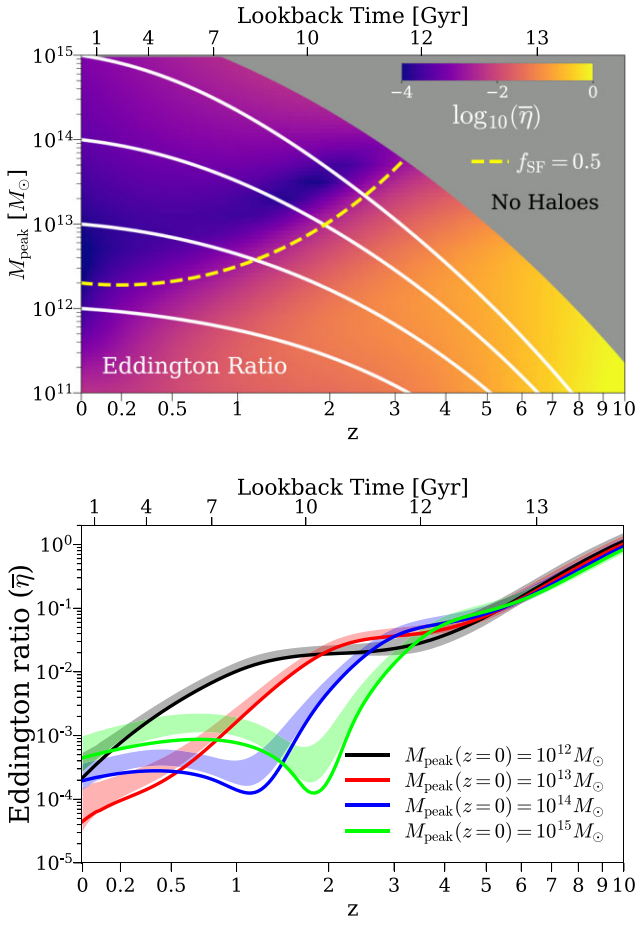

Figure 19. Top panel: average SMBH *total* (i.e. radiative+kinetic) Eddington ratio ($\overline{\eta}$) as a function of M_{peak} and z (see Section 4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\mathrm{peak}} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. Bottom panel: $\overline{\eta}$ histories as a function of halo mass at $z = 0$. The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

than indicated by the median $M_\bullet$–M_{peak} relation. This bias is also discussed in Lauer et al. (2007). Towards lower redshifts, more and more massive haloes emerge with time. As a result, the high-mass BHMF in the local universe is composed almost equally of haloes with $13 < \log_{10} M_{\mathrm{peak}} < 14$ and $14 < \log_{10} M_{\mathrm{peak}} < 15$. In short, cluster-scale haloes ($\log_{10} M_{\mathrm{peak}} > 14$) are too rare to dominate the massive end of low-redshift BHMFs, mainly due to their own rarity and the flat $M_\bullet$–M_{peak} at these redshifts.

4.5 SMBH mergers

The top panel of Fig. 23 shows the average black hole merger rates (BHMRs) as a function of M_{peak} and z. Note that in this paper, we define BHMR as the *SMBH growth rate due to mergers*, instead of the number of SMBH mergers per unit SMBH, per unit redshift, and per unit (log-) SMBH mass ratio (as presented in Paper V). In general, BHMRs increase monotonically with M_{peak} and z. The same conclusion holds when we look at the average BHMR histories as a function of M_{peak} at $z = 0$, which is shown in the bottom

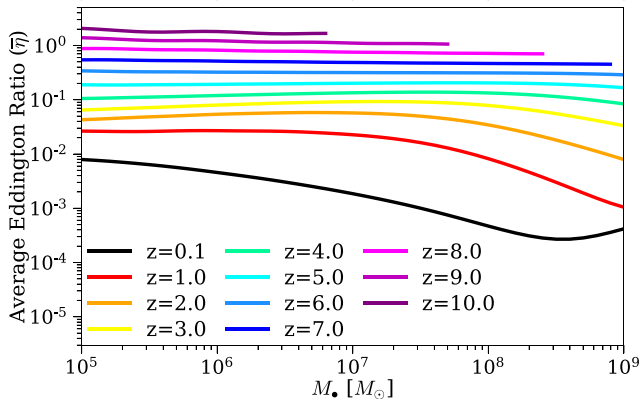

Figure 20. Average SMBH *total* (i.e. radiative+kinetic) Eddington ratio ($\overline{\eta}$) as a function of $M_\bullet$ and z. See Section 4.3. All the data used to make this plot can be found here.

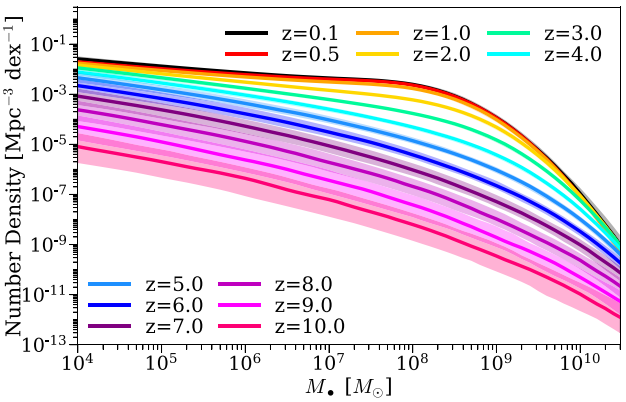

Figure 21. The total BHMF in the range $0 \leq z \leq 10$ (see Section 4.4). The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

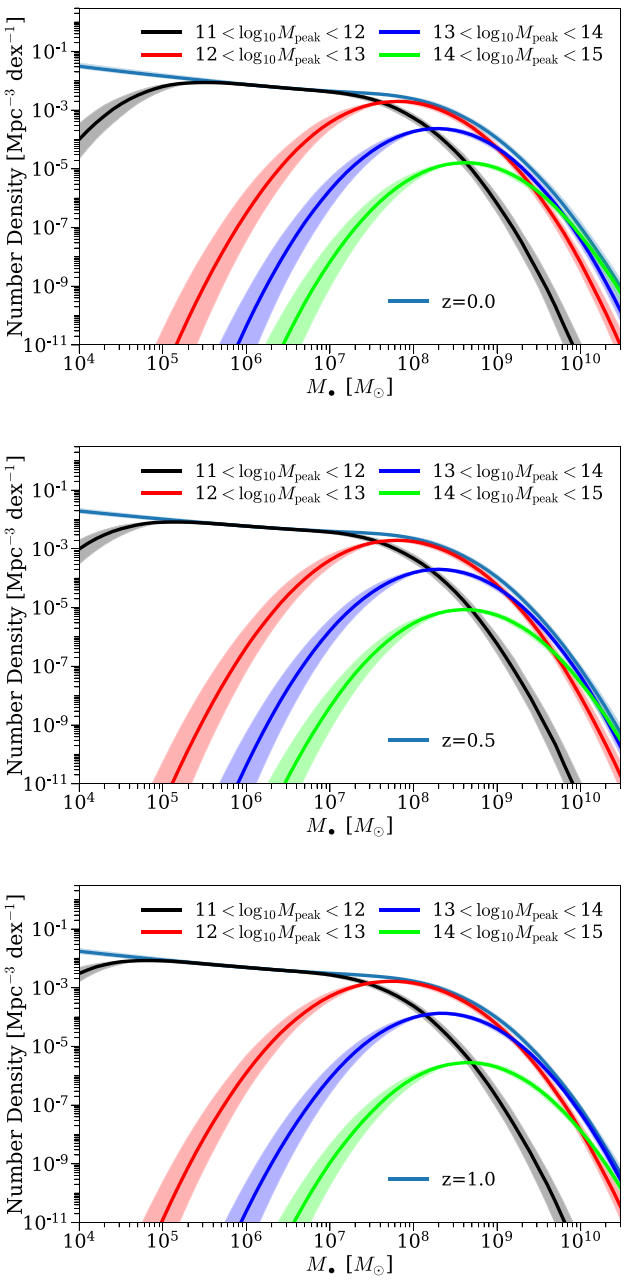

Figure 22. Total BHMFs at $z = 0.0$, 0.5, and 1.0 (the top, middle, and bottom panels), split into the contributions from different host dark matter halo mass bins (see Section 4.4.1). The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

panel of Fig. 23. The best-fitting model lies on the upper edges of the 68 per cent confidence intervals. Although the best-fitting model uses a significant amount of mergers to fit the data, the dominance of SMBH growth via smooth accretion (see Paper V) means that parameter sets with lower merger rates also fit the data well. As mentioned in Section 2.5, BHMRs are calculated by allowing a fraction of galaxy mergers (the free parameter $f_{\rm scale}$) to result in mergers of their SMBHs. This is done due to continuing uncertainty about SMBH merger time-scales (e.g. Tremmel et al. 2018). Therefore, these BHMRs are constrained by the combination of: (a) SMBH total growth rates, which are given by the evolution of active and total BHMFs; and (b) average black hole accretion rates, which are constrained by the QLFs and probability distribution functions. The best-fitting TRINITY model predicts $f_{\rm scale}$ to be $\log_{10}(f_{\rm scale}) = -0.192^{+0.126}_{-2.285} + (-0.000^{+1.970}_{-0.523})(a-1)$. This means that, for example, when the fractional merger contribution to instantaneous galaxy growth is 10 per cent, the merger contribution to SMBH growth would be 10 per cent $\times\ 10^{-0.192} \approx 6.4$ per cent. In Appendix E3, we also show the results of models with alternate assumptions about SMBH mergers. Further discussion about SMBH mergers in TRINITY and predictions for gravitational wave experiments are presented in Paper V.

4.6 AGN energy efficiency and systematic uncertainties

As described in Section 2.3 and Section 2.8, we modelled systematic uncertainties in stellar mass, SFRs, and SMBH Eddington ratios. These uncertainties are propagated into our model predictions, and their values quantify the degree of tension between different data sets. In TRINITY, the best-fitting values (see Appendix H) of the galaxy systematics are all consistent with those given by Behroozi et al. (2019). The systematic offset in SMBH Eddington ratios is motivated by the discrepancy between the QLFs from Ueda et al. (2014) and the QPDFs from Aird et al. (2018) (see Appendix D4). This discrepancy can be caused by different assumptions for: (1) differences in M_* estimates used by Aird et al. (2018) and those in our galaxy data compilation (Section 3.2.1); (2) the ways in which X-ray photons are counted, including how galaxy contributions are subtracted; (3) the functional forms used to fit the observational data. The net effect is $\eta' - \eta \sim 0.5$ dex, where η is the intrinsic Eddington ratio, and η' is the Eddington ratio used to calculate the observed QPDFs in Aird et al. (2018).

The total AGN energy efficiency from TRINITY is $\log_{10}\epsilon_{\rm tot} = -1.318^{+0.115}_{-0.009}$. In other words, the best-fitting model is consistent with a redshift-independent ~ 5 per cent mass-to-energy conversion efficiency. However, the exact value of $\epsilon_{\rm tot}$ is affected by various input assumptions, such as AGN bolometric corrections, Compton-thin/Compton-thick obscured fractions, and/or the assumed local $M_\bullet$–$M_{\rm bulge}$ scaling relation (if ever assumed). These assumptions alter the amount of radiation to be produced by SMBH accretion, which systematically changes the best-fitting $\epsilon_{\rm tot}$. In Appendices D1, D2, D3, and E2, we carry out experiments with different bolometric corrections, Compton-thick/Compton-thin obscuration fractions,

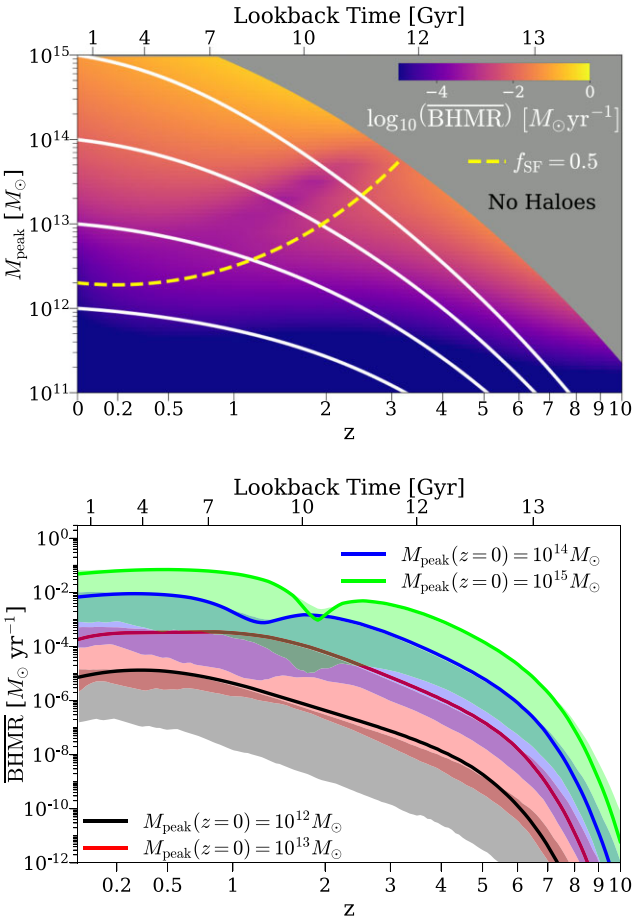

Figure 23. Top panel: the average black hole merger rates ($\overline{\rm BHMR}$) as a function of $M_{\rm peak}$ and z (see Section 4.5). The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. Bottom panel: $\overline{\rm BHMR}$ histories as a function of halo mass at $z = 0$. The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

fixed local $M_\bullet$–$M_{\rm bulge}$ scaling relations and compare with the fiducial TRINITY model. When varying these input assumptions, the best-fitting AGN energy efficiency can change from ∼0.035 to 0.07, i.e. a factor of 2 (or 0.3 dex). In this work, we opt not to allow a systematic offset in the normalization of the $M_\bullet$–$M_{\rm bulge}$ relation, $\beta_{\rm BH}$, due to its complete degeneracy with the AGN energy efficiency. Thus, the best-fitting value of the energy efficiency $\epsilon_{\rm tot}$ should be viewed as *a combination of* the intrinsic average efficiency and any potential systematic offset in $\beta_{\rm BH}$. We emphasize that this energy efficiency quantifies how effectively gravitational energy is converted into *radiation and kinetic energy*. Thus, there is no unique link between our efficiency and the average SMBH spin value.

4.7 Correlation coefficient ($\rho_{\rm BH}$) between average SMBH accretion rate and $M_\bullet$ at fixed halo mass

Fig. 24 shows the redshift evolution of $\rho_{\rm BH}$ from the best-fitting model. At $z \gtrsim 8$, the average SMBH accretion rate and $M_\bullet$ are highly correlated at fixed host halo mass. In other words, high-redshift SMBHs share the same *Eddington ratio* distributions, if they are hosted by haloes with similar masses. This correlation fades towards lower redshifts. By $z = 0$, there is essentially no correlation between

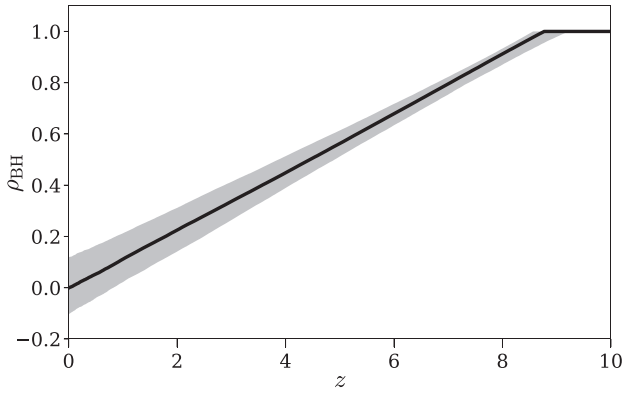

Figure 24. The correlation coefficient, $\rho_{\rm BH}$, between average SMBH accretion rate and $M_\bullet$ at fixed halo mass. See Section 4.7. The shaded region shows the 68 per cent confidence intervals inferred from the model posterior distribution. The data used to make this plot can be found here.

average SMBH accretion rate and $M_\bullet$, i.e. different SMBHs have the same *absolute* accretion rate distributions, if hosted by similar haloes. Overall, this evolution makes large SMBHs less and less active compared to their smaller counterparts (measured by difference in average Eddington ratio) in the same halo mass bin. Consequently, AGN downsizing effects apply not only to SMBHs in *different host haloes* (as shown in Section 4.3), but also to those hosted by *similar haloes and galaxies*. Although this conclusion holds qualitatively in all the model variants covered in the Appendix, the exact $\rho_{\rm BH}$ value at $z = 0$ does change significantly in some of these models (see Appendices D1 and D3).

5 COMPARISON WITH PREVIOUS STUDIES AND DISCUSSION

In this section, we compare TRINITY with hydrodynamical simulations as well as discuss the potential physical mechanisms that could reproduce the redshift evolution of the $M_\bullet$–$M_{\rm bulge}$ relation (Section 5.1); present the cosmic SMBH mass density as a function of redshift (Section 5.2); and discuss the physical implications of the best-fitting TRINITY model (Section 5.3).

5.1 Evolution of the galaxy–SMBH scaling relation

The growth of SMBHs and their feedback on host galaxies are important physical mechanisms to capture in hydrodynamical simulations. Although different simulations find similar local $M_\bullet$–$M_{\rm bulge}$ (or M_*) relations, they differ in the relation's redshift evolution. For example, the IllustrisTNG (Pillepich et al. 2018) and SIMBA (Davé et al. 2019) simulations predicted increasing normalizations of the scaling with time, whereas the Illustris (Vogelsberger et al. 2014), Horizon-AGN (Dubois, Volonteri & Silk 2014; Dubois et al. 2016), and EAGLE simulations (Schaye et al. 2015) predicted the opposite (Habouzit et al. 2021). This diversity in the redshift evolution results from different sub-grid physics adopted by each simulation.

TRINITY infers the redshift evolution of this scaling relation by extracting information directly from observational data, without any assumptions about the underlying physics. This can help determine which sub-grid physics models give results that are more consistent with observations. We show the $M_\bullet$–M_* relations at different redshifts from TRINITY and IllustrisTNG100 (Pillepich et al. 2018; Habouzit et al. 2021) in Fig. 25. Despite the offset, both mass scalings show increasing normalizations with time at $M_* \leq 10^{11} M_\odot$. This implies that

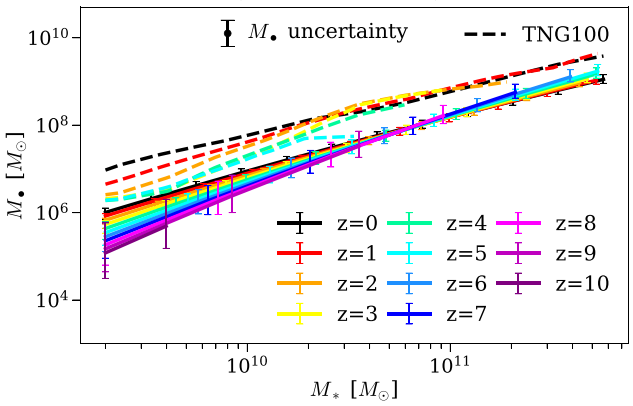

Figure 25. The median $M_\bullet$–M_* relations as functions of z for TRINITY (*solid lines*) and the IllustrisTNG100 simulation (*dashed lines*; Pillepich et al. 2018; Habouzit et al. 2021). See Section 5.1. The typical uncertainty in the measurement of $M_\bullet$, 0.3 dex, is shown by the black solid dot. At $z \geq 3$, the dynamical ranges of $M_\bullet$ and M_* in TNG100 are smaller than in TRINITY, due to the smaller simulation box size. All the data used to make this plot (including those from Illustris TNG and our best-fitting model) can be found here.

SMBH growth becomes increasingly efficient compared to galaxy growth at lower redshifts. For the hydrodynamical simulations listed in Habouzit et al. (2021), the following sub-grid physics models succeeded in reproducing this trend: (a) the strong supernova feedback in low-mass galaxies at high redshifts that reduces early SMBH growth in IllustrisTNG (Dubois et al. 2015; Bower et al. 2017; Pillepich et al. 2018); and (b) the low accretion AGN feedback mode that quenches galaxies but favours further SMBH growth in SIMBA (Davé et al. 2019). That said, SMBH masses depend on many different aspects of sub-grid physics, including cooling, star formation, supernova feedback, magnetic fields, etc. beyond those directly related to the growth of the SMBH. Hence, the success of a given sub-grid recipe at matching properties of SMBHs cannot be taken as evidence in support of its correctness without the context of the recipe's successes and failures at matching other non-SMBH observations.

5.2 Cosmic SMBH mass density

Fig. 26 shows the cosmic SMBH mass density as a function of redshift from TRINITY compared to previous studies. Unlike previous studies that tried to solve the continuity equation, in TRINITY, we assume that wandering SMBHs also contribute to QLFs during their growth. Thus, we include the cosmic wandering SMBH mass density in Fig. 26 for a fair comparison. We also show the cosmic wandering SMBH density separately in cyan, which accounts for ~ 15 per cent of the total SMBH mass density at $z = 0$. This is broadly consistent with the results from Volonteri, Haardt & Madau (2003) based on a semi-analytical model and Ricarte et al. (2021) based on the ROMULUS simulations.

Below $z \sim 2$, the offsets in the mass density between different studies are mostly driven by the different AGN energy efficiencies. Above $z \sim 2$, the systematic difference with Marconi et al. (2004) increases with redshift. The reason is that Marconi et al. (2004) forward modelled AGN evolution assuming that all SMBH growth occurred at $z < 3$. These initial conditions did not consider SMBH assembly histories at higher redshifts, and hence give different SMBH mass functions at $z \sim 3$ from TRINITY, in which SMBHs are modelled to start growing from $z = 15$.

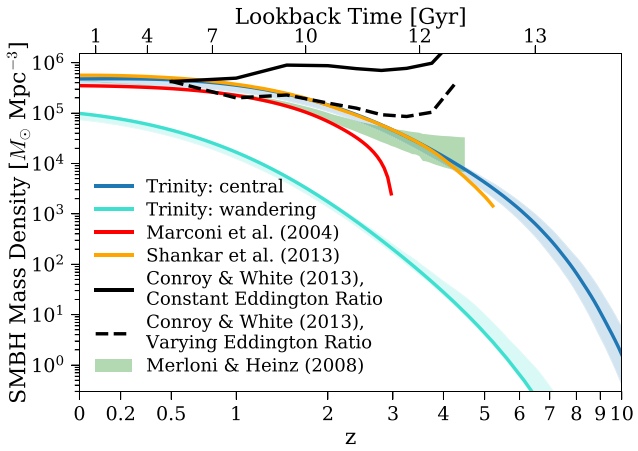

Figure 26. Cosmic SMBH mass density as a function of z (see Section 5.2). The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot (including those from previous studies and our best-fitting model) can be found here.

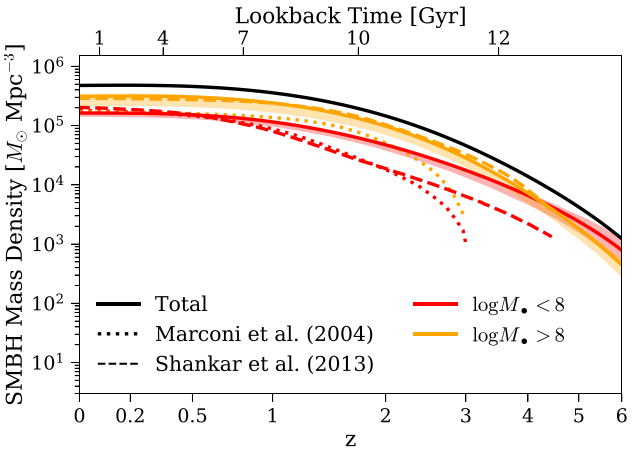

Figure 27. Cosmic SMBH mass densities split in different SMBH mass bins as functions of z, from TRINITY (solid lines), Marconi et al. (2004) (dotted lines), and Shankar et al. (2013) (dashed lines). See Section 5.2. All the data used to make this plot (including those from previous studies and our best-fitting model) can be found here.

Compared to other studies, Conroy & White (2013) inferred quite different SMBH mass density histories. They assumed a mass-independent Eddington ratio distribution and a linear $M_\bullet$–M_* relation, and tried to fit the QLFs *at each individual redshift* with two free parameters: (1) the normalization of the $M_\bullet$–M_* relation, and (2) the AGN duty cycle. The SMBH mass density at each redshift was then obtained by convolving the galaxy SMF with the $M_\bullet$–M_* relation. This method does not enforce any continuity equation for SMBH mass. As a result, it cannot guarantee the consistency between the inferred cosmic SMBH mass growth rates and the QLFs. This is shown in Fig. 26, where the SMBH mass density from Conroy & White (2013) decreases with time at some points in cosmic history for all variations considered. In light of this, we do not make further comparison with Conroy & White (2013) here.

Fig. 27 shows the cosmic SMBH mass density histories of different SMBH populations from TRINITY (solid lines), Marconi et al. (2004) (dotted lines), and Shankar et al. (2013) (dashed lines). The main difference between the results from TRINITY and these two studies is the cosmic times when low-mass SMBHs ($M_\bullet \leq 10^8 M_\bullet$)

experience major growth. Specifically, SMBHs below $10^8 M_\bullet$ nearly stop growing below $z \sim 1$ in TRINITY, but grow significantly from $z = 1$ to $z = 0$ in the Marconi et al. and Shankar et al. models. One possible reason for this is that TRINITY is required to fit the QPDFs for low-mass galaxies at lower redshifts from Aird et al. (2018), which limit the growth of low-mass black holes. However, neither Marconi et al. (2004) nor Shankar et al. (2013) had access to these QPDFs, so their predictions are not necessarily consistent with these data. Another difference exists at $z > 1$: at a fixed redshift, these low-mass SMBHs also make up a larger share of the cosmic SMBH mass density in TRINITY. This is likely due to TRINITY's self-consistent inference of SMBH growth history from $z = 15$, which results in non-negligible cosmic SMBH mass densities at the starting redshifts in the Marconi et al. and Shankar et al. models (i.e. $z \sim 3$ and $z \sim 5$, respectively).

5.3 Physical implications: AGN downsizing and AGN feedback on galaxy populations

In Section 4.3 and Section 4.7, we confirmed the 'AGN downsizing' effect, in the sense that more massive black holes become less active earlier compared to smaller black holes, whether they are in the same host halo mass bin or not. This is true when the SMBH activity is measured by Eddington ratio (see Figs 19 and 20). If we instead measure SMBH activity with absolute accretion rate, we see a slight increase in BHAR towards higher masses at $z \lesssim 2$ (see Fig. 16). As mentioned earlier, this is required by the QPDFs from Aird et al. (2018). Physically, this is consistent with AGN feedback (Croton et al. 2006; Somerville et al. 2008). That is, in massive haloes, SMBHs still show ongoing accretion, but become less active *relative to their masses* and radiatively inefficient. The energy from their mass accretion is mainly released in the form of kinetic jets and/or outflows, which serves to maintain quenching in their host galaxies. This picture is also supported by Fig. 18, where the $\overline{\rm BHAR}/\overline{\rm SFR}$ ratio increases towards higher mass and lower redshifts. Although cooling flows are known to exist in massive haloes (Fabian 1994), Fig. 18 suggests that the ratio of cold gas reaching the SMBH compared to the galaxy increases for more massive haloes. The same amount of gas also causes much more relative mass growth for SMBHs than galaxies, given their contrast in mass. Other possible fuelling channels include gas recycling from stellar mass-loss. Regardless of the source, SMBHs in massive haloes plausibly have sufficient material to continue growing (and generating feedback) even as the host galaxy itself is not able to grow.

Fig. 18 also shows that below $z \sim 6$, $\overline{\rm BHAR}$ and $\overline{\rm SFR}$ have relatively fixed average ratios for the haloes in which most star formation occurs. This is consistent with a picture in which the SMBH and the galaxy regulate each others' growth, but it is also consistent with a process in which a separate mechanism (e.g. mass accretion on to the halo) jointly feeds both galaxy and SMBH growth. Regardless of the mechanism, it must qualitatively change in haloes above masses of 10^{12}–$10^{13} M_\odot$ to reproduce the clear upturn in $\overline{\rm BHAR}/\overline{\rm SFR}$ for massive haloes.

6 CAVEATS AND FUTURE DIRECTIONS FOR EMPIRICAL MODELLING OF THE HALO–GALAXY–SMBH CONNECTION

In this section, we discuss caveats in the current version of TRINITY, which motivates its future incorporation into UNIVERSEMACHINE.

6.1 Bright quasars at $5.7 < z < 6.5$ below $M_\bullet = 10^8 M_\odot$

As described in Section 3.2.2, we applied a Poisson prior on the number of high-redshift bright quasars with masses below $M_\bullet = 10^8 M_\odot$. This is motivated by the fact that few such objects are found in real observations. However, our best-fitting model still predicts ~3.5 such objects in the same area as covered by SDSS, in contrast to current observations. By checking the intrinsic and observed BHMFs of bright quasars produced by TRINITY, we found that most of these objects have intrinsically high black hole masses but have lower observed masses due to the random scatter in virial estimates (see Section 3.2.2). Therefore, even if there are no intrinsically low-mass bright quasars at $z \gtrsim 6$, some should still exist in the observed sample.

6.2 Future directions

Currently, TRINITY makes only *statistical* halo–galaxy–SMBH connections. In the future, we plan to incorporate TRINITY into the UNIVERSEMACHINE by modelling SMBHs in *individual* haloes and galaxies. This will allow: (a) constraining the correlation between individual galaxy growth and SMBH growth, (b) more flexibility in terms of the distributions of physical properties; (c) direct modelling of AGN duty cycle time-scales; (d) study of the environmental effects on galaxy–SMBH coevolution; (e) use of more data constraints, including separate probability distribution functions for star-forming and quiescent galaxies as well as quasar correlation functions; and (f) enable the generation of more realistic halo–galaxy–SMBH mock catalogues for the whole community.

7 CONCLUSIONS

In this work, we introduce TRINITY, which is an empirical model that parametrizes the statistical halo–galaxy–SMBH connection. (Section 2). Compared to previous studies that are typically focused on one or two kinds of observables, TRINITY self-consistently matches a comprehensive set of observational data for galaxies and SMBHs from $z = 0$–10 (Section 3, Section 4.1). These joint constraints enable TRINITY to break degeneracies present in past studies. Key results are as follows:

(i) The normalization and the slope of the median $M_\bullet$–$M_{\rm bulge}$ relation increase slightly from $z = 0$ to $z = 10$. At all redshifts, the mild evolution of the median $M_\bullet$ at fixed galaxy total/bulge mass is consistent with existing observational measurements (Section 4.2, Fig. 11).

(ii) The AGN mass-to-energy conversion efficiency $\epsilon_{\rm tot}$ is ~0.05. However, the exact value of AGN efficiency depends on the adopted AGN bolometric correction, Compton-thin/Compton-thick obscured fractions, and the assumed local $M_\bullet$–$M_{\rm bulge}$ relation. When these input assumptions are changed, $\epsilon_{\rm tot}$ can vary from ~0.035 to 0.07, i.e. a factor of 2, or 0.3 dex. (Section 4.6, Appendices D1, D2, D3, and E2).

(iii) Average SMBH Eddington ratios are between 0.1 and 1 at $z \gtrsim 6$. This is consistent with the scenario that different SMBH populations at high redshifts are growing at close to the Eddington rate. Towards lower redshifts, their Eddington ratios (and thus specific accretion rates) decline. Therefore, total BHMFs show a strong increase in normalization at all masses from $z \sim 10$ to $z \sim 5$, and the evolution slows down towards lower redshifts. (Section 4.3, Fig. 19, Section 4.4, Fig. 21).

(iv) AGNs experience downsizing, in the sense that average Eddington ratios start to decrease earlier for more massive SMBHs. This applies to SMBHs hosted by either similar haloes/galaxies, or

in different host mass bins. However, this AGN downsizing *does not* hold for average SMBH accretion rates, which do not decrease towards higher masses at low redshifts (Section 4.3, Section 4.7, Figs 16, 19, 20, and 24).

(v) The ratio between average SMBH accretion rate and galaxy SFR is $\sim 10^{-3}$ for low-mass haloes, where star-forming galaxies dominate the population. This ratio increases in massive haloes (and galaxies) towards lower redshifts, where galaxies are more likely to be quiescent even as their SMBHs are still growing (Section 4.3, Fig. 18).

(vi) Sub-grid physics recipes that qualitatively reproduce the $M_\bullet$–$M_{\rm bulge}$ redshift evolution include but are not limited to: (a) strong supernova feedback in high-redshift, low-mass galaxies (IllustrisTNG, Dubois et al. 2015; Bower et al. 2017; Pillepich et al. 2018); (b) a low accretion feedback mode that keeps SMBH growing but quenches galaxies (SIMBA, Davé et al. 2019). See Section 5.1 and Fig. 25.

(vii) Forbidding super-Eddington accretion as well as non-unity occupation fractions prevents SMBHs from growing sufficiently to match the local $M_\bullet$–$M_{\rm bulge}$ relation. In this scenario, an AGN energy efficiency of ~ 24 per cent is needed to explain observations like QLFs and QPDFs at high redshifts (Appendix E1, Fig. E1).

(viii) Forbidding redshift evolution of the $M_\bullet$–$M_{\rm bulge}$ relation results in a best-fitting $M_\bullet$–$M_{\rm bulge}$ relation that is consistent with the fiducial model, (Appendix E2.1, Fig. E5), but a much higher correlation coefficient between SMBH accretion rate and BH mass at fixed halo mass ($\rho_{\rm BH}$) is required to reproduce AGN data (Fig. E4).

(ix) During galaxy mergers, central SMBHs are unlikely to quickly consume all the infalling satellite SMBHs, otherwise black hole accretion rates would experience a precipitous decline towards lower redshift and higher masses (Appendix E3.1, Fig. E6). Hence, a significant number of 'wandering' black holes are necessary.

(x) The following models make qualitatively consistent predictions with the fiducial TRINITY model: (a) no SMBH mergers take place; (b) the fractional growth contribution to SMBH growth is always the same as that for galaxy growth (Appendix E3.2, Figs E7 and E8).

This work is the first in a series of TRINITY papers. Paper II (Zhang et al., in preparation) discusses QLFs and the build-up of SMBHs across cosmic time; Paper III (Zhang et al., in preparation) presents predictions for quasars and other SMBHs at $z > 6$; Paper IV (Zhang et al., in preparation) discusses the SFR–BHAR correlation as a function of halo mass, galaxy mass, and redshift; and paper V (Zhang et al., in preparation) covers BHMRs and TRINITY's predictions for gravitational wave experiments. Paper VI (Knox, Zhang, and Behroozi, in preparation) and Paper VII (Zhang, Zhang, and Behroozi, in preparation) present the AGN autocorrelation functions and AGN–galaxy cross-correlation functions from TRINITY, respectively.

ACKNOWLEDGEMENTS

We thank Stacey Alberts, Rachael Amaro, Gurtina Besla, Haley Bowden, Jane Bright, Katie Chamberlain, Alison Coil, Ryan Endsley, Sandy Faber, Hayden Foote, Dan Foreman-Mackey, Nico Garavito-Camargo, Nickolay Gnedin, Richard Green, Jenny Greene, Kate Grier, Melanie Habouzit, Kevin Hainline, Elaheh Hayati, Andrew Hearin, Julie Hlavacek-Larrondo, Luis Ho, Allison Hughes, Yun-Hsin Huang, Raphael Hviding, Victoria Jones, Stephanie Juneau, Ryan Keenan, Oddisey Knox, David Koo, Andrey Kravtsov, Daniel Lawther, Rixin Li, Joseph Long, Jianwei Lyu, Chung-Pei Ma, Garreth Martin, Karen Olsen, Feryal Özel, Vasileios Paschalidis, Ekta Patel, Dimitrios Psaltis, Joel Primack, Yujing Qin, Eliot Quataert, George Rieke, Marcia Rieke, Paolo Salucci, Jan-Torge Schindler, Spencer Scott, Xuejian Shen, Yue Shen, Dongdong Shi, Irene Shivaei, Rachel Somerville, Fengwu Sun, Wei-Leong Tee, Yoshihiro Ueda, Marianne Vestergaard, Feige Wang, Ben Weiner, Christina Williams, Charity Woodrum, Jiachuan Xu, Minghao Yue, Dennis Zaritsky, Huanian Zhang, Xiaoshuai Zhang, and Zhanbo Zhang for very valuable discussions.

Support for this research came partially via program number *HST*-AR-15631.001-A, provided through a grant from the Space Telescope Science Institute under NASA contract NAS5-26555. PB was partially funded by a Packard Fellowship, Grant #2019-69646. PB was also partially supported by a Giacconi Fellowship from the Space Telescope Science Institute. Finally, PB was also partially supported through program number *HST*-HF2-51353.001-A, provided by NASA through a Hubble Fellowship grant from the Space Telescope Science Institute, under NASA contract NAS5-26555.

Data compilations from many studies used in this paper were made much more accurate and efficient by the online WEBPLOTDIGITIZER code.[1] This research has made extensive use of the arXiv and NASA's Astrophysics Data System.

This research used the Ocelote supercomputer of the University of Arizona. The allocation of computer time from the UA Research Computing High Performance Computing (HPC) at the University of Arizona is gratefully acknowledged. The Bolshoi–Planck simulation was performed by Anatoly Klypin within the Bolshoi project of the University of California High-Performance AstroComputing Center (UC-HiPACC; PI Joel Primack).

DATA AVAILABILITY

The parallel implementation of TRINITY, the compiled data sets (Section 3.2), data for all figures, and the posterior distribution of model parameters (Section 4.1, Appendix H) are available online.

REFERENCES

Aird J. et al., 2010, MNRAS, 401, 2531
Aird J., Coil A. L., Georgakakis A., 2018, MNRAS, 474, 1225
Alexander D. M., Hickox R. C., 2012, New Astron. Rev., 56, 93
Aller M. C., Richstone D. O., 2007, ApJ, 665, 120
Allevato V., Shankar F., Marsden C., Rasulov U., Viitanen A., Georgakakis A., Ferrara A., Finoguenov A., 2021, ApJ, 916, 34
Ananna T. T. et al., 2019, ApJ, 871, 240
Ananna T. T. et al., 2022, ApJS, 261, 9
Aversa R., Lapi A., de Zotti G., Shankar F., Danese L., 2015, ApJ, 810, 74
Baldry I. K. et al., 2012, MNRAS, 421, 621
Barger A. J., Cowie L. L., Mushotzky R. F., Yang Y., Wang W. H., Steffen A. T., Capak P., 2005, AJ, 129, 578
Bastian N., Covey K. R., Meyer M. R., 2010, ARA&A, 48, 339
Bauer A. E. et al., 2013, MNRAS, 434, 209
Behroozi P. S., Wechsler R. H., Conroy C., 2013, ApJ, 770, 57
Behroozi P. S. et al., 2015, MNRAS, 450, 1546
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143
Beifiori A., Courteau S., Corsini E. M., Zhu Y., 2012, MNRAS, 419, 2497
Bellovary J., Volonteri M., Governato F., Shen S., Quinn T., Wadsley J., 2011, ApJ, 742, 13
Blandford R. D., McKee C. F., 1982, ApJ, 255, 419
Bongiorno A. et al., 2012, MNRAS, 427, 3103

[1] https://apps.automeris.io/wpd/

Bouwens R. J., Stefanon M., Oesch P. A., Illingworth G. D., Nanayakkara T., Roberts-Borsani G., Labbé I., Smit R., 2019, ApJ, 880, 25
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., Schaye J., Frenk C. S., Theuns T., Schaller M., Crain R. A., McAlpine S., 2017, MNRAS, 465, 32
Brandt W. N., Alexander D. M., 2015, A&AR, 23, 1
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Buchner J. et al., 2015, ApJ, 802, 89
Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, ApJ, 533, 682
Caplar N., Lilly S. J., Trakhtenbrot B., 2015, ApJ, 811, 148
Caplar N., Lilly S. J., Trakhtenbrot B., 2018, ApJ, 867, 148
Carraro R. et al., 2020, A&A, 642, A65
Cavaliere A., Vittorini V., 2000, ApJ, 543, 599
Chabrier G., 2003, PASP, 115, 763
Coil A. L. et al., 2011, ApJ, 741, 8
Comparat J. et al., 2019, MNRAS, 487, 2005
Conroy C., White M., 2013, ApJ, 762, 70
Conroy C., Gunn J. E., White M., 2009, ApJ, 699, 486
Cool R. J. et al., 2013, ApJ, 767, 118
Croton D. J. et al., 2006, MNRAS, 365, 11
Cucciati O. et al., 2012, A&A, 539, A31
Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, MNRAS, 486, 2827
Delvecchio I. et al., 2014, MNRAS, 439, 2736
Ding X. et al., 2020, ApJ, 888, 37
Drake A. B. et al., 2015, MNRAS, 454, 2015
Dubois Y., Devriendt J., Slyz A., Teyssier R., 2012, MNRAS, 420, 2662
Dubois Y., Volonteri M., Silk J., 2014, MNRAS, 440, 1590
Dubois Y., Volonteri M., Silk J., Devriendt J., Slyz A., Teyssier R., 2015, MNRAS, 452, 1502
Dubois Y., Peirani S., Pichon C., Devriendt J., Gavazzi R., Welker C., Volonteri M., 2016, MNRAS, 463, 3948
Dunn G., Bellovary J., Holley-Bockelmann K., Christensen C., Quinn T., 2018, ApJ, 861, 39
Dunne L. et al., 2009, MNRAS, 394, 3
Duras F. et al., 2020, A&A, 636, A73
Eddington A. S., 1913, MNRAS, 73, 359
Fabian A. C., 1994, ARA&A, 32, 277
Ferrarese L., 2002, ApJ, 578, 90
Ferrarese L., Ford H., 2005, Space Sci. Rev., 116, 523
Ferrarese L., Merritt D., 2000, ApJ, 539, L9
Finkelstein S. L. et al., 2015, ApJ, 810, 71
Gebhardt K. et al., 2000, ApJ, 539, L13
Georgakakis A., Comparat J., Merloni A., Ciesla L., Aird J., Finoguenov A., 2019, MNRAS, 487, 275
Greene J. E. et al., 2016, ApJ, 826, L32
Greene J. E., Strader J., Ho L. C., 2020, ARA&A, 58, 257
Grogin N. A. et al., 2011, ApJS, 197, 35
Grylls P. J., Shankar F., Zanisi L., Bernardi M., 2019, MNRAS, 483, 2506
Gültekin K. et al., 2009, ApJ, 698, 198
Gunawardhana M. L. P. et al., 2011, MNRAS, 415, 1647
Gunawardhana M. L. P. et al., 2013, MNRAS, 433, 2764
Haario H., Saksman E., Tamminen J., 2001, Bernoulli, 7, 223
Habouzit M., Volonteri M., Dubois Y., 2017, MNRAS, 468, 3935
Habouzit M. et al., 2021, MNRAS, 503, 1940
Häring N., Rix H.-W., 2004, ApJ, 604, L89
Heckman T. M., Best P. N., 2014, ARA&A, 52, 589
Hlavacek-Larrondo J. et al., 2015, ApJ, 805, 35
Ho L. C., 2008, ARA&A, 46, 475
Hopkins P. F., Richards G. T., Hernquist L., 2007a, ApJ, 654, 731
Hopkins P. F., Bundy K., Hernquist L., Ellis R. S., 2007b, ApJ, 659, 976
Hu J., 2008, MNRAS, 386, 2242
Ilbert O. et al., 2013, A&A, 556, A55
Ishigaki M., Kawamata R., Ouchi M., Oguri M., Shimasaku K., Ono Y., 2018, ApJ, 854, 73
Kajisawa M., Ichikawa T., Yamada T., Uchimoto Y. K., Yoshikawa T., Akiyama M., Onodera M., 2010, ApJ, 723, 129
Karim A. et al., 2011, ApJ, 730, 61
Kelly B. C., Shen Y., 2013, ApJ, 764, 45
Kistler M. D., Yuksel H., Hopkins A. M., 2013, preprint (arXiv:1305.1630)
Klypin A. A., Trujillo-Gomez S., Primack J., 2011, ApJ, 740, 102
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Kormendy J., Ho L. C., 2013, ARA&A, 51, 511
Kormendy J., Richstone D., 1995, ARA&A, 33, 581
Kroupa P., 2001, MNRAS, 322, 231
Krumholz M. R., 2014, Phys. Rep., 539, 49
Kulkarni G., Worseck G., Hennawi J. F., 2019, MNRAS, 488, 1035
Labbé I. et al., 2013, ApJ, 777, L19
Lacey C. G. et al., 2016, MNRAS, 462, 3854
Lang P. et al., 2014, ApJ, 788, 11
Lauer T. R., Tremaine S., Richstone D., Faber S. M., 2007, ApJ, 670, 249
Le Borgne D., Elbaz D., Ocvirk P., Pichon C., 2009, A&A, 504, 727
Leja J., van Dokkum P. G., Franx M., Whitaker K. E., 2015, ApJ, 798, 115
Ly C., Lee J. C., Dale D. A., Momcheva I., Salim S., Staudaher S., Moore C. A., Finn R., 2011a, ApJ, 726, 109
Ly C., Malkan M. A., Hayashi M., Motohara K., Kashikawa N., Shimasaku K., Nagao T., Grady C., 2011b, ApJ, 735, 91
Madau P., Dickinson M., 2014, ARA&A, 52, 415
Magnelli B., Elbaz D., Chary R. R., Dickinson M., Le Borgne D., Frayer D. T., Willmer C. N. A., 2011, A&A, 528, A35
Magorrian J. et al., 1998, AJ, 115, 2285
Marconi A., Risaliti G., Gilli R., Hunt L. K., Maiolino R., Salvati M., 2004, MNRAS, 351, 169
Mazzucchelli C. et al., 2017, ApJ, 849, 91
McConnell N. J., Ma C.-P., 2013, ApJ, 764, 184
McCracken H. J. et al., 2012, A&A, 544, A156
McDonald M., McNamara B. R., Calzadilla M. S., Chen C.-T., Gaspari M., Hickox R. C., Kara E., Korchagin I., 2021, ApJ, 908, 85
McLure R. J. et al., 2011, MNRAS, 418, 2074
Mendel J. T., Simard L., Palmer M., Ellison S. L., Patton D. R., 2014, ApJS, 210, 3
Merloni A., 2004, MNRAS, 353, 1035
Merloni A., Heinz S., 2008, MNRAS, 388, 1011
Merloni A., Rudnick G., Di Matteo T., 2004, MNRAS, 354, L37
Merloni A. et al., 2014, MNRAS, 437, 3550
Mineshige S., Kawaguchi T., Takeuchi M., Hayashida K., 2000, PASJ, 52, 499
Moustakas J. et al., 2013, ApJ, 767, 50
Muzzin A. et al., 2013, ApJ, 777, 18
Novak G. S., Faber S. M., Dekel A., 2006, ApJ, 637, 96
Oesch P. A., Bouwens R. J., Illingworth G. D., Labbé I., Stefanon M., 2018, ApJ, 855, 105
Park D. et al., 2012, ApJ, 747, 30
Peterson B. M., 1993, PASP, 105, 247
Pillepich A. et al., 2018, MNRAS, 475, 648
Planck Collaboration XXX, 2014, A&A, 571, A30
Planck Collaboration XIII, 2016, A&A, 594, A13
Reines A. E., Volonteri M., 2015, ApJ, 813, 82
Ricarte A., Tremmel M., Natarajan P., Zimmer C., Quinn T., 2021, MNRAS, 503, 6098
Robotham A. S. G., Driver S. P., 2011, MNRAS, 413, 2570
Rujopakarn W. et al., 2010, ApJ, 718, 1171
Salim S. et al., 2007, ApJS, 173, 267
Salmon B. et al., 2015, ApJ, 799, 183
Salpeter E. E., 1955, ApJ, 121, 161
Salucci P., Szuszkiewicz E., Monaco P., Danese L., 1999, MNRAS, 307, 637
Santini P. et al., 2009, A&A, 504, 751
Savorgnan G. A. D., Graham A. W., Marconi A. r., Sani E., 2016, ApJ, 817, 21
Schaye J. et al., 2015, MNRAS, 446, 521
Schramm M., Silverman J. D., 2013, ApJ, 767, 13
Schreiber C. et al., 2015, A&A, 575, A74
Schulze A., Wisotzki L., 2010, A&A, 516, A87

Schulze A. et al., 2015, MNRAS, 447, 2085
Shankar F., Weinberg D. H., Miralda-Escudé J., 2009, ApJ, 690, 20
Shankar F., Weinberg D. H., Miralda-Escudé J., 2013, MNRAS, 428, 421
Shankar F. et al., 2016, MNRAS, 460, 3119
Shankar F. et al., 2020a, Nat. Astron., 4, 282
Shankar F. et al., 2020b, MNRAS, 493, 1500
Shen Y. et al., 2019, ApJ, 873, 35
Shen X., Hopkins P. F., Faucher-Giguère C.-A., Alexander D. M., Richards G. T., Ross N. P., Hickox R. C., 2020, MNRAS, 495, 3252
Shim H., Colbert J., Teplitz H., Henry A., Malkan M., McCarthy P., Yan L., 2009, ApJ, 696, 785
Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575
Silk J., Rees M. J., 1998, A&A, 331, L1
Silverman J. D. et al., 2008, ApJ, 679, 118
Small T. A., Blandford R. D., 1992, MNRAS, 259, 725
Smit R. et al., 2014, ApJ, 784, 58
Sobral D., Best P. N., Smail I., Mobasher B., Stott J., Nisbet D., 2014, MNRAS, 437, 3516
Sołtan A., 1982, MNRAS, 200, 115
Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481
Song M. et al., 2016, ApJ, 825, 5
Speagle J. S., Steinhardt C. L., Capak P. L., Silverman J. D., 2014, ApJS, 214, 15
Steed A., Weinberg D. H., 2003, preprint (arXiv:astro-ph/0311312)
Straatman C. M. S. et al., 2016, ApJ, 830, 51
Suh H., Civano F., Trakhtenbrot B., Shankar F., Hasinger G., Sanders D. B., Allevato V., 2020, ApJ, 889, 32
Sun M. et al., 2015, ApJ, 802, 14
Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
Tomczak A. R. et al., 2014, ApJ, 783, 85
Tomczak A. R. et al., 2016, ApJ, 817, 118
Trakhtenbrot B., Volonteri M., Natarajan P., 2017, ApJ, 836, L1
Tremaine S. et al., 2002, ApJ, 574, 740
Tremmel M., 2017, PhD thesis, University of Washington
Tremmel M., Governato F., Volonteri M., Quinn T. R., Pontzen A., 2018, MNRAS, 475, 4967
Tucci M., Volonteri M., 2017, A&A, 600, A64
Ueda Y., Akiyama M., Hasinger G., Miyaji T., Watson M. G., 2014, ApJ, 786, 104
van den Bosch R. C. E., 2016, ApJ, 831, 134
van der Burg R. F. J., Hildebrandt H., Erben T., 2010, A&A, 523, A74
van Dokkum P. G., Conroy C., 2012, ApJ, 760, 70
Veale M., White M., Conroy C., 2014, MNRAS, 445, 1144
Vestergaard M., Peterson B. M., 2006, ApJ, 641, 689
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Volonteri M., 2010, A&AR, 18, 279
Volonteri M., Haardt F., Madau P., 2003, ApJ, 582, 559
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
Weinberger R. et al., 2017, MNRAS, 465, 3291
Whitaker K. E. et al., 2014, ApJ, 795, 104
Yang G. et al., 2018, MNRAS, 475, 1887
York D. G. et al., 2000, AJ, 120, 1579
Yoshida M. et al., 2006, ApJ, 653, 988
Yu Q., Lu Y., 2004, ApJ, 602, 603
Yu Q., Tremaine S., 2002, MNRAS, 335, 965
Zheng X. Z., Bell E. F., Papovich C., Wolf C., Meisenheimer K., Rix H.-W., Rieke G. H., Somerville R., 2007, ApJ, 661, L41
Zwart J. T. L., Jarvis M. J., Deane R. P., Bonfield D. G., Knowles K., Madhanpall N., Rahmani H., Smith D. J. B., 2014, MNRAS, 439, 1459

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

reduced_mcmc_submit_fiducial.dat
best_fit_submit_fiducial.dat

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: HALO MERGER RATES

In TRINITY, SMBH mergers are directly linked to galaxy mergers. As shown in equation (15), halo merger rates are needed in the calculation of galaxy merger rates. Hence, we use the halo merger rates from the UNIVERSEMACHINE, where satellite galaxies will disrupt when their $v_{\rm max}/v_{\rm Mpeak}$ ratios reach a certain threshold (see Section 2.2 for the definitions of $v_{\rm max}$ and $v_{\rm Mpeak}$). We refer readers to section 3.3 and appendix B of Behroozi et al. (2019) for full details. Here, we fit these merger rates with a set of analytical formulae. Letting $a = 1/(1 + z)$ be the scale factor, $M_{\rm desc}$ the mass of the descendant halo, $M_{\rm sat}$ the mass of the satellite halo, and $\theta = M_{\rm sat}/M_{\rm desc}$ the mass ratio, the merger rate is expressed as the number of mergers per unit descendant halo, per unit redshift per log interval in mass ratio:

$$-\frac{{\rm d}^2N(M_{\rm desc},\theta,z)}{{\rm d}z{\rm d}\log_{10}\theta} = 10^{A(M_{\rm desc},a)}\theta^{B(a)}\exp(-3.162\theta) \quad \text{(A1)}$$

$$A(M_{\rm desc},a) = A_0(M_{\rm desc}) + A_1(a) \quad \text{(A2)}$$

$$A_0(M_{\rm desc}) = 0.148\log_{10}\left(\frac{M_{\rm desc}}{10^{12}M_\odot}\right) - 0.291 \quad \text{(A3)}$$

$$A_1(a) = -1.609 + 3.816a + (-2.152)a^2 \quad \text{(A4)}$$

$$B(a) = -1.114 + 1.498a + (-0.757)a^2. \quad \text{(A5)}$$

We show the quality of these fits in Fig. A1. Compared to Behroozi et al. (2013), these merger rates are lower by 15–40 per cent due to the presence of orphan galaxies in the UNIVERSEMACHINE.

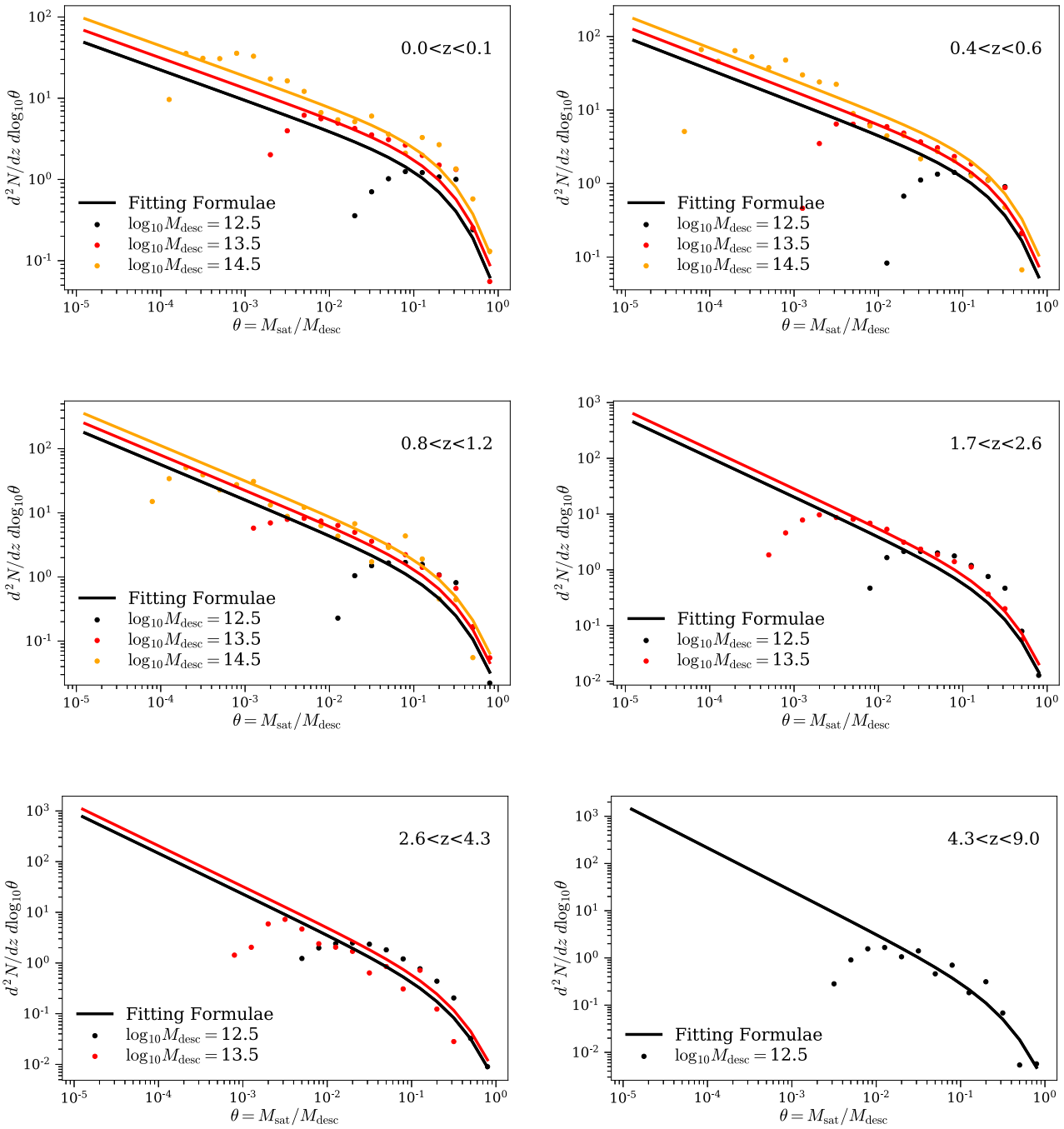

Figure A1. The rate of satellite galaxy disruption in host haloes in the UNIVERSEMACHINE as a function of z, descendant mass $M_{\rm desc}$, and satellite-to-descendant mass ratio $\theta = M_{\rm sat}/M_{\rm desc}$. The solid symbols are the binned estimates of merger rates, and the solid lines are the fitted results. See Appendix A. All the data used to make this plot (including the individual data points and our best-fitting model) can be found here.

APPENDIX B: MEDIAN GALAXY UV MAGNITUDES AND SCATTER AS FUNCTIONS OF HALO MASS AND STAR FORMATION RATES

To constrain the high-redshift halo–galaxy connection in TRINITY, we use the median galaxy UV magnitudes and the corresponding lognormal scatter from the UNIVERSEMACHINE as functions of redshift, halo mass ($M_{\rm peak}$), and SFRs to calculate galaxy UVLFs at $z = 9$ and $z = 10$. Here, we show the best-fitting parameters for these scaling relations, as well as the goodness of fitting.

The median galaxy UV magnitudes $\widetilde{M}_{\rm UV}$ have the following dependence on redshift, $M_{\rm peak}$, and SFR:

$$\widetilde{M}_{\rm UV} = k_{\rm UV} \times \log_{10}{\rm SFR} + b_{\rm UV} \tag{B1}$$

$$k_{\rm UV} = 0.154(\log_{10} M_{\rm peak})^2 + (-2.876)\log_{10} M_{\rm peak} + (-2.378)(a-1) + 9.478 \tag{B2}$$

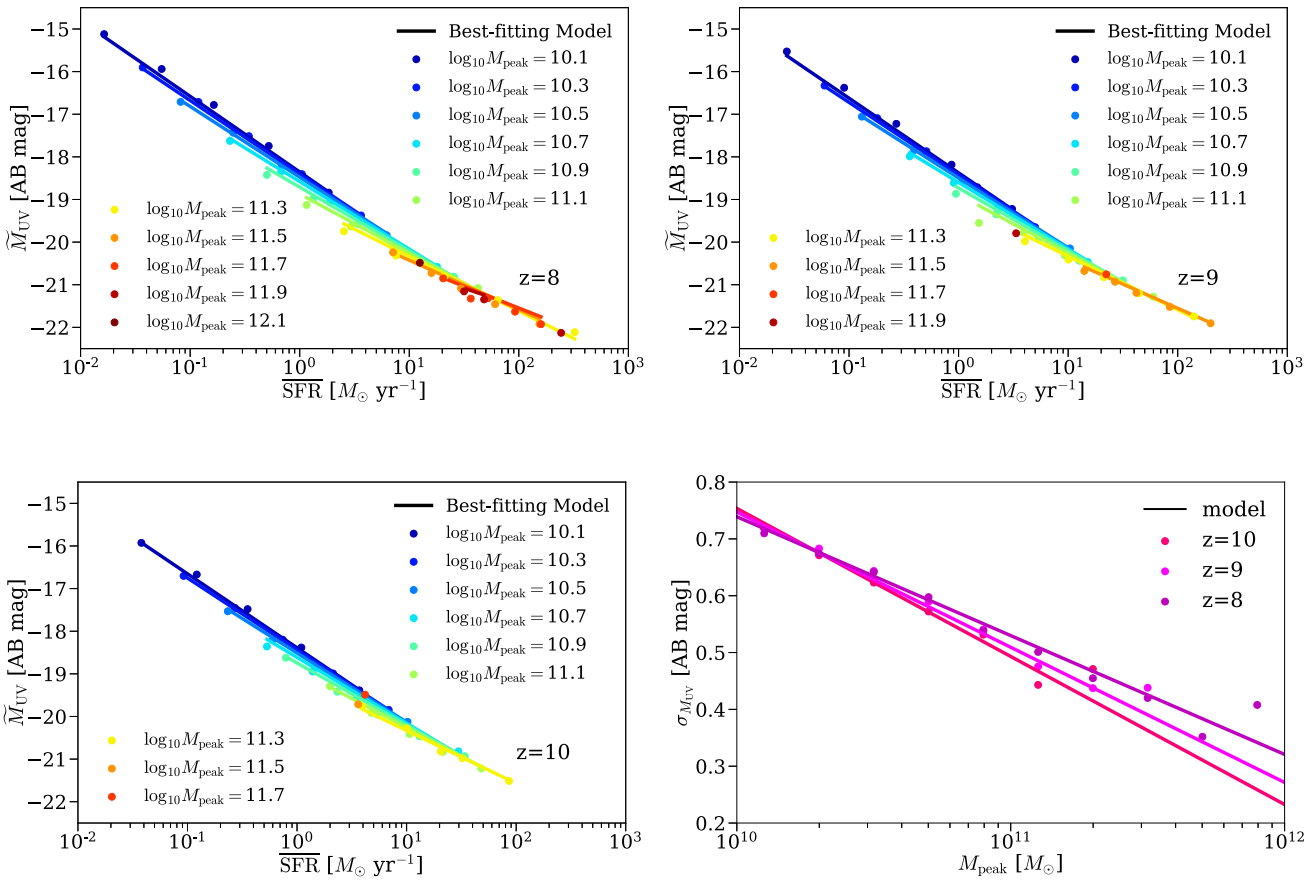

Figure B1. The fits to median UV magnitude, $\widetilde{M}_{\rm UV}$, as a function of $M_{\rm peak}$, $\overline{\rm SFR}$, and z, and the corresponding scatter, $\sigma_{M_{\rm UV}}$, as a function of $M_{\rm peak}$ and z, from the UNIVERSEMACHINE. The filled circles are the data points from the UNIVERSEMACHINE, and the solid lines are the best-fitting models in equations (B1)–(B6). See Appendix B. All the data used to make this plot (including the individual data points and our best-fitting model) can be found here.

$$b_{\rm UV} = (-0.347)(\log_{10} M_{\rm peak})^2 + 6.853 \log_{10} M_{\rm peak} + 1.993(a-1) + (-50.344). \quad \text{(B3)}$$

The lognormal scatter $\sigma_{\rm UV}$ has the following redshift and $M_{\rm peak}$ dependency:

$$\sigma_{\rm UV} = k_{\sigma_{\rm UV}} \times \log_{10} M_{\rm peak} + b_{\sigma_{\rm UV}} \quad \text{(B4)}$$

$$k_{\sigma_{\rm UV}} = -0.031z + 0.042 \quad \text{(B5)}$$

$$b_{\sigma_{\rm UV}} = 0.319z + 0.241. \quad \text{(B6)}$$

Fig. B1 shows the goodness of fit for equations (B1)–(B6) to both $\widetilde{M}_{\rm UV}$ and $\sigma_{\rm UV}$ from $z = 8$–10. Using these fitting functions, TRINITY produces SFRs and galaxy UV luminosities that are both consistent with the UNIVERSEMACHINE.

APPENDIX C: CALCULATING INHERITED AND INFALLING SMBH MASSES FROM MERGER TREE STATISTICS

In TRINITY, we assign SMBH masses to haloes at all redshifts and then calculate black hole growth rates by differentiation. This is different from how we model galaxies (where we directly model galaxy growth rates and integrate to obtain stellar masses), because the functional forms for galaxy growth rates in haloes are better known than the functional forms for SMBH growth rates in galaxies. Here, we detail how we calculate the masses of the inherited and infalling (see Section 2.5) SMBHs.

In TRINITY, haloes inherit both central and wandering SMBHs from their MMPs. For the jth halo mass bin at the ith snapshot, the average central SMBH mass inherited from MMPs is

$$\overline{M}^{j}_{\bullet,{\rm inherit},i} = \sum_k P^{j,k}_{{\rm MMP},i}\, \overline{M}^{k}_{\bullet,i-1}, \quad \text{(C1)}$$

where $P^{j,k}_{{\rm MMP},i}$ is the probability that haloes in the jth halo mass bin at the ith snapshot have MMPs in the kth mass bin at the $(i-1)$th snapshot. This probability is calculated based on the average halo growth curves from N-body simulations (see Section 3.1). $\overline{M}^{k}_{\bullet,i-1}$ is the average *central* SMBH mass of the haloes in the kth mass bin at the $(i-1)$th snapshot, determined by the halo–galaxy–SMBH connection.

As for infalling SMBHs, they come from: (1) wandering SMBHs inherited from MMPs; (2) *all* the SMBHs from infalling satellite haloes. The average mass of infalling SMBHs for the jth halo mass bin at the ith snapshot is then, by definition,

$$\overline{M}^{j}_{\bullet,{\rm infall},i} = \sum_k P^{j,k}_{{\rm MMP},i}\, \overline{M}^{k}_{\bullet,{\rm wandering},i-1} + \sum_k \mathcal{R}^{j,k}_{{\rm merger},i}\, \overline{M}^{k}_{\bullet,i-1}, \quad \text{(C2)}$$

where $\overline{M}^{k}_{\bullet,{\rm wandering},i-1}$ is the average total *wandering* SMBH mass of the haloes in the kth mass bin at the $(i-1)$th snapshot, and $\mathcal{R}^{j,k}_{{\rm merger},i}$ is the merger rate of satellite haloes in the kth mass bin into the descendant haloes in the jth mass bin at the ith snapshot. This rate is

calculated by integrating equation (A1) over the redshift dimension:

$$\mathcal{R}^{j,k}_{\mathrm{merger},i} = \int_{10^{-0.5\Delta\log_{10}M_{\mathrm{peak}}}M^{k}_{\mathrm{peak},i}/M^{j}_{\mathrm{peak},i}}^{10^{0.5\Delta\log_{10}M_{\mathrm{peak}}}M^{k}_{\mathrm{peak},i}/M^{j}_{\mathrm{peak},i}} \times \frac{\mathrm{d}^2N(M_{\mathrm{peak}},\theta,z)}{\mathrm{d}\log\theta\mathrm{d}z}\Big|_{z=z_i}^{M_{\mathrm{peak}}=M^{j}_{\mathrm{peak},i}}\mathrm{d}\theta, \quad \text{(C3)}$$

where z_i is the redshift of the ith snapshot, and $M^{j}_{\mathrm{peak},i}$ is the peak mass of the halo in the jth mass bin at the ith snapshot.

APPENDIX D: CORRECTIONS, EXCLUSIONS, AND UNCERTAINTIES FOR AGN DATA

D1 Bolometric corrections

Different bolometric corrections (BC) for the same quasar sample produce different *bolometric* QLFs, which, in principle, could lead to systematic differences in the inferred SMBH properties. Here, we investigate how the systematic difference in bolometric corrections would impact our results in Section 4.

Fig. D1 shows the different resulting bolometric QLFs at $z = 2$ produced by correcting Ueda et al. (2014) QLFs with BCs from Ueda et al. (2014) (filled circles, 'UedaBC') and Duras et al. (2020) (stars, 'DurasBC'). Due to smaller BC values at high X-ray luminosities, the 'DurasBC' gives many fewer bright quasars. At the less massive end, the two BCs result in consistent quasar number densities. The low number densities of bright quasars suppress the abundance of more massive SMBHs, because only the latter can produce so much energy with reasonable Eddington ratios. Ultimately, this forces TRINITY to choose $M_\bullet$–M_{bulge} relations with lower normalizations (β_{BH}) and slopes (γ_{BH}), as shown in Fig. D2. With the decrease in both the total energy output and the $M_\bullet$–M_{bulge} normalization, the AGN energy efficiency only decreases by ∼0.02 dex if the 'DurasBC' is adopted.

However, we do find significantly higher values of the correlation coefficient between average SMBH accretion rate and $M_\bullet$ at fixed host halo mass, ρ_{BH} (Section 4.7), when adopting the 'DurasBC' (Fig. D3). This is because TRINITY still has to reproduce similar numbers of quasars with $L_{\mathrm{bol}} \sim 10^{45}$ erg s^{-1} as in the 'UedaBC' case, but with lower $M_\bullet$. If ρ_{BH} stays as low as in the 'UedaBC' case, TRINITY will inevitably produce more(fewer) low-(high-)mass active black holes with Eddington ratios of $\eta > 0.01$. This would be

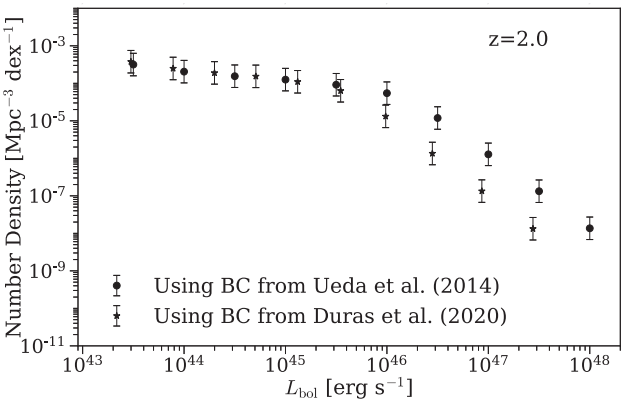

Figure D1. The comparison of the QLFs at $z = 2$ from Ueda et al. (2014), when using bolometric corrections (BC) from Ueda et al. (2014) (filled circles) and from Duras et al. (2020) (stars). See Appendix D1. All the data used to make this plot can be found here.

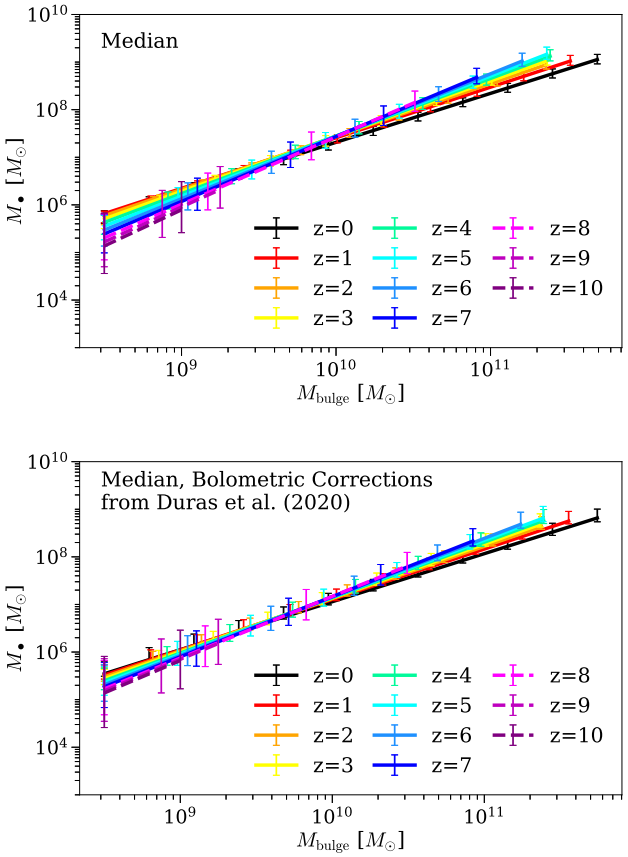

Figure D2. Top panel: the best-fitting median $M_\bullet$–M_{bulge} relation from $z = 0$–10 assuming the bolometric corrections from Ueda et al. (2014). Bottom panel: the best-fitting median $M_\bullet$–M_{bulge} relation from $z = 0$–10 assuming the bolometric corrections from Duras et al. (2020). The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by e.g. *JWST*). All the data used to make this plot can be found here.

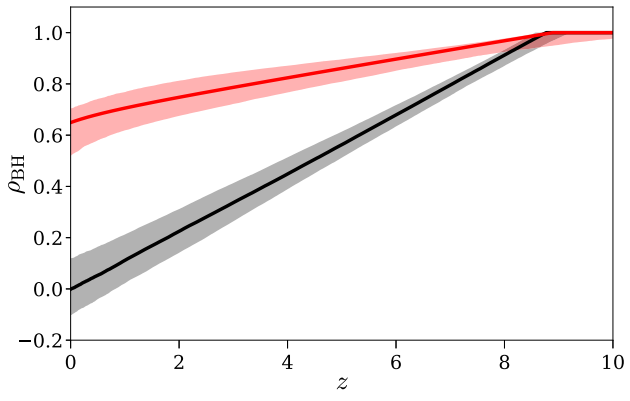

Figure D3. The correlation coefficient between average SMBH accretion rate and $M_\bullet$ at fixed host halo mass, ρ_{BH}, assuming the bolometric corrections from Ueda et al. (2014) (black solid line) and Duras et al. (2020) (red solid line). All the data used to make this plot can be found here.

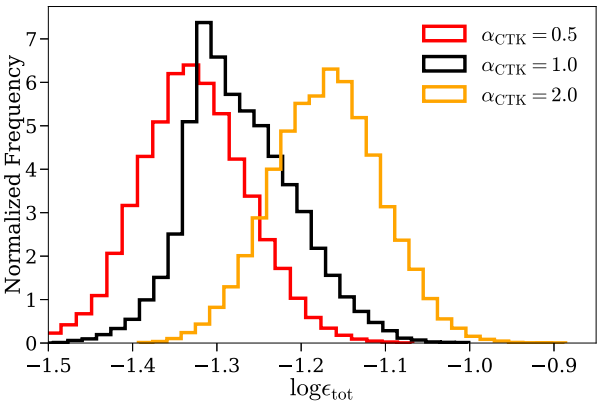

Figure D4. The comparison of the posterior distributions of SMBH efficiency $\epsilon_{\rm tot}$ between models with $\alpha_{\rm CTK}=0.5$, 1.0, and 2.0. See Appendix D2. All the data used to make this plot can be found here.

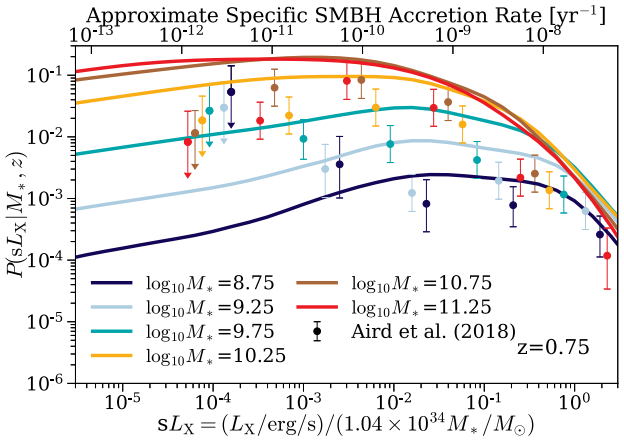

Figure D5. The comparison between the observed QPDFs from Aird et al. (2018) and the best-fitting model with QLFs and Compton-thick obscuration corrections from Ananna et al. (2019), at $z=0.75$. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

inconsistent with the ABHMFs from Schulze & Wisotzki (2010) and Schulze et al. (2015).

Other than $\rho_{\rm BH}$, using the bolometric corrections from either Ueda et al. (2014) or Duras et al. (2020) does not make any qualitative differences in our main results.

D2 Compton-thick correction

As mentioned in Section 3.2, we have adopted QLFs from Ueda et al. (2014) to constrain the total AGN energy budget. However, Ueda et al. did not include Compton-thick obscured AGNs in their QLF plots. Hence, we applied the following empirical correction given by Ueda et al. (2014) to convert from Compton-thin-only QLFs to total QLFs:

$$\begin{aligned}\Phi_{L,\rm tot}(L_X,z) &= \Phi_{L,\rm CTN}(L_X,z)\times(1+\alpha_{\rm CTK}\psi(L_X,z))\\ \psi(L_X,z) &= \min\left[0.84,\ \max\left[\psi_{43.75}(z)-0.24L_{43.75},\ \psi_{\rm min}\right]\right]\\ \psi_{43.75}(z) &= \begin{cases}0.43(1+z)^{0.48}[z<2.0]\\ 0.43(1+2)^{0.48}[z\geq2.0]\end{cases}\\ L_{43.75} &= \log_{10}(L_X/\mathrm{erg\ s^{-1}})-43.75,\end{aligned}\tag{D1}$$

where $\psi(L_X,z)$ is the fraction of Compton-thin absorbed AGN, and $\alpha_{\rm CTK}$ is the number ratio between Compton-thick and Compton-thin AGN. Ueda et al. adopted $\alpha_{\rm CTK}=1$ in their main analysis, but their analysis of the cosmic X-ray background radiation shows that there is a ±50 per cent uncertainty in $\alpha_{\rm CTK}$. In light of this, we ran TRINITY with $\alpha_{\rm CTK}=0.5$ and 2.0, aside from the fiducial model where $\alpha_{\rm CTK}=1.0$. The *only* model parameter that shows significant differences is the SMBH total efficiency ($\epsilon_{\rm tot}$, Fig. D4). A higher $\alpha_{\rm CTK}$ implies a larger Compton-thick AGN population, and thus higher QLFs at all redshifts. Consequently, TRINITY needs a higher AGN efficiency to account for the larger AGN number densities.

Since Ueda et al. (2014), several studies updated the absorption functions, i.e. the probability distribution of gas column density as a function of X-ray luminosity and redshift, and found much higher Compton-thick obscured fractions, especially for bright AGNs (Buchner et al. 2015; Ananna et al. 2019). According to Ananna et al. (2019), $\gtrsim$80 per cent of the AGNs with $L_{\rm X,\,2-10\,KeV}\gtrsim10^{45}$ are Compton-thick obscured. This is significantly higher than ~20 per cent as suggested by Ueda et al. (2014). To explore the potential impact of different Compton-thick corrections on TRINITY results, we ran a model with QLFs and Compton-thick obscuration corrections from Ananna et al. (2019). In this experiment, we found significant inconsistency between Ananna et al. (2019) results and other AGN data. Specifically, the high Compton-thick fractions at the bright end produces too many bright quasars. In this case, TRINITY is unable to reproduce the bright end of the luminosity function with only SMBHs in massive galaxies, given their small number densities. Consequently, TRINITY is forced to make SMBHs overmassive in lower mass galaxies to reproduce the luminosity functions. This ultimately leads to inconsistency with the QPDFs for low-mass galaxies from Aird et al. (2018) (see Fig. D5). The best-fitting model with Ananna et al. (2019) luminosity functions and Compton-thick corrections give a $\chi^2\approx844.62$, which is significantly worse compared to the fiducial model with data and corrections from Ueda et al. (2014) ($\chi^2\approx746.70$). We note that such a strong inconsistency is present even when the systematic offset in Eddington ratio, ξ, is allowed to vary in the MCMC (see Section 2.8). Given this inconsistency with other AGN data, we choose to keep using the QLFs and Compton-thick corrections from Ueda et al. (2014) in the main text. From this experiment, we have shown that TRINITY does have the ability to place upper limits on Compton-thick AGN fractions based on inter-data set consistency. Further discussion into this topic is beyond the scope of this paper, and is thus deferred to a future investigation.

D3 Obscured fraction

In the fiducial TRINITY model, we adopted the correction for obscured AGN from Merloni et al. (2014) for ABHMFs. We did not adopt the Compton-thin obscured fraction from Ueda et al. (2014) due to the reported inconsistency between the optical type-I versus type-II and X-ray obscured versus unobscured AGNs (Merloni et al. 2014). Here, we show the quantitative changes in the best-fitting model if the Compton-thin obscured fraction from Ueda et al. (2014) (i.e. $\psi(L_X,z)$ in Appendix D2) is also applied to ABHMFs.

Fig. D6 shows the difference in the obscured fraction, $F_{\rm obs}$, as a function of X-ray luminosity. We only show the comparison at $z=3$ as an example, and there is no qualitative difference at any other relevant redshift. The obscured fraction from Ueda et al. (2014) is higher than that from Merloni et al. (2014) at any fixed X-ray luminosity above $L_{\rm X}\sim3\times10^{43}$ erg s^{-1}. This leaves fewer

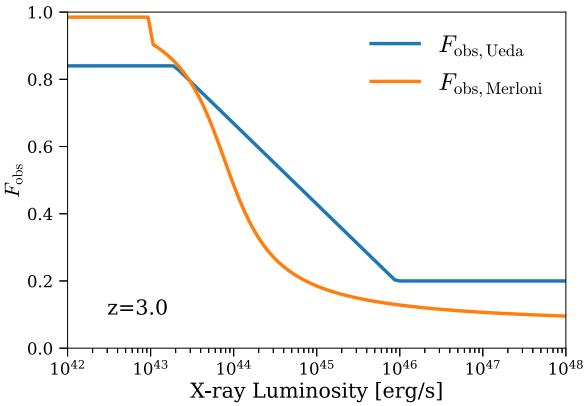

Figure D6. The obscured fractions of AGNs as functions of X-ray luminosity from Ueda et al. (2014) (blue solid line) and Merloni et al. (2014) (orange solid line). To save space, we only show the fractions at $z = 3$, since there are no qualitative differences across the relevant redshift range. All the data used to make this plot can be found here.

unobscured AGNs in the type I AGN mass function. To compensate for this deficit, TRINITY needs to increase the radiative efficiency from ~ 5 per cent to ~ 10 per cent to make more bright AGNs. However, only increasing efficiency will also increase the normalization of QLFs and QPDFs. Thus, TRINITY has to simultaneously adjust the redshift evolution of the $M_\bullet$–$M_{\rm bulge}$ relation, as shown in Fig. D7. Compared to the fiducial model, we no longer see significant evolution in the slope of the $M_\bullet$–$M_{\rm bulge}$ relation, whereas its normalization decreases slightly towards higher redshifts. These changes lead to less(more) growth of low-(high-)mass SMBHs, and thus, less(more) contribution to QLFs and QPDFs from low-(high-)mass SMBHs. The ultimate net result is that QLFs and QPDFs are still reproduced, while the ABHMFs are corrected by a larger $F_{\rm obs}$.

D4 AGN probability distribution functions from Aird et al. (2018)

To use QPDFs from Aird et al. (2018) to constrain our model, we had to account for two factors as below.

First, Aird et al. (2018) modelled the AGN probability distribution functions for each stellar mass and redshift bin as a finite series of gamma distributions. The function values in their public release[2] were evaluated with these model functions over a dense grid of s$L_{\rm X}$. Thus, naively taking all the points in their data release would artificially increase the weight of this data set. To avoid this, we downsampled their modelled AGN probability distribution functions with 1 dex spacing. This choice is based on the fact that the spacing between two neighbouring gamma distributions is 0.2 dex, and that an extra prior was applied to ensure smoothness of the probability distribution functions across neighbouring gamma distributions.

Secondly, in the process of compiling different data sets, we found that there is significant inconsistency between the QLFs from Ueda et al. (2014) and the high-s$L_{\rm X}$ and high-z (i.e. $z > 2.5$) end of AGN probability distribution functions from Aird et al. (2018). This may be due to the massive end of the AGN probability distribution functions being affected by the smoothness prior. To ensure consistency between these two data sets, we excluded AGN probability distribution function points with $z > 2.5$ or s$L_{\rm X}$>1 from Aird et al.

[2]Available at https://zenodo.org/record/1009605.

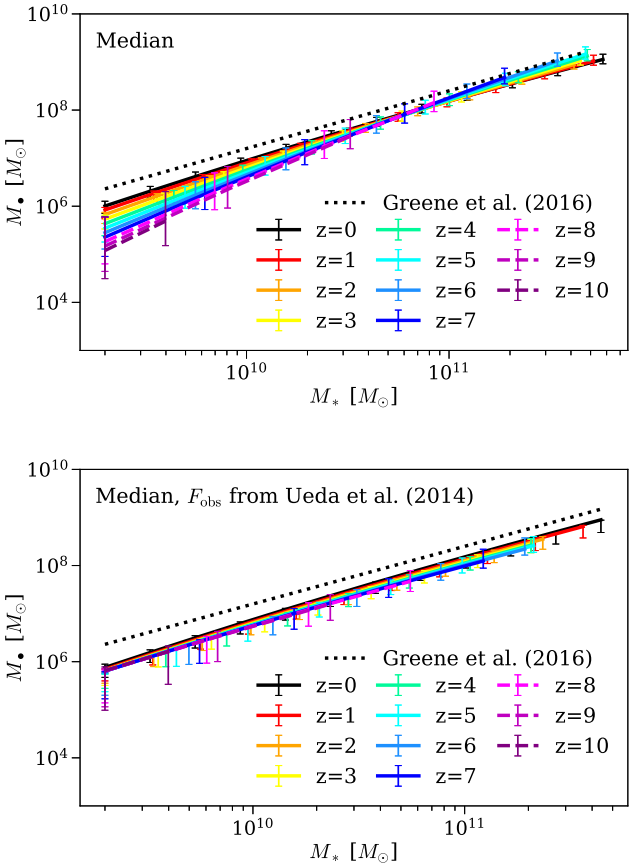

Figure D7. The median $M_\bullet$–M_* relations from $z = 0$–10 from the fiducial model (top panel) and the model where obscured fractions of AGNs as functions of X-ray luminosity from Ueda et al. (2014) are applied to the ABHMFs (bottom panel). The error bars show the 68 per cent confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found here.

(2018). After removing the most inconsistent data points, residual inconsistencies of the order of 0.3 dex persist between these two data sets. To address this, we further enlarged the uncertainties in the AGN probability distribution functions to 0.3 dex, and included an extra free parameter ξ to describe the systematic offset in the Eddington ratio in the calculation of probability distribution functions in terms of s$L_{\rm X}$ (see equation 74 in Section 2.8).

D5 Active black hole functions

D5.1 Active black hole functions from Schulze & Wisotzki (2010) and Schulze et al. (2015)

In TRINITY, we use ABHMFs at $z = 0.2$ and $z = 1.5$ from Schulze & Wisotzki (2010) and Schulze et al. (2015). However, two issues were addressed before using these ABHMFs as constraints. First, as is shown in Fig. 22 of Schulze et al. (2015), the massive end of the ABHMF varies with different model assumptions due to the different significance of Eddington bias. To avoid this model dependence, we chose to only use the data points in the region where the ABHMF estimate is independent of their model assumptions, i.e. $\log_{10}M_\bullet \lesssim 9.8$. Secondly, Schulze & Wisotzki (2010) used virial BH mass estimates that are on average smaller by 0.2 dex than those used in Schulze et al. (2015). To account for this, we applied a mass shift

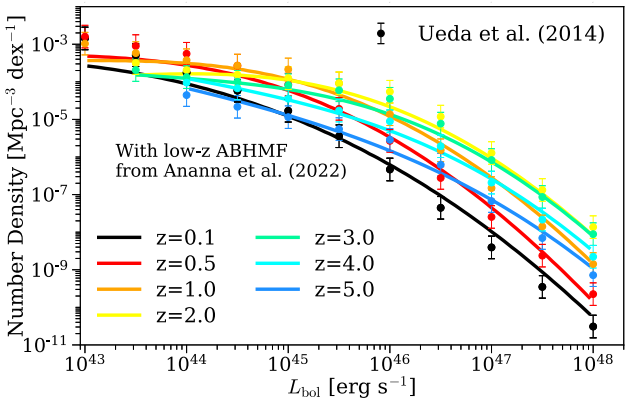

Figure D8. Comparison between the observed QLFs from Ueda et al. (2014) and our model prediction from $z = 0$–5, with $z \sim 0.2$ ABHMFs from Schulze & Wisotzki (2010) replaced by the Ananna et al. (2022) results. Higher redshift ABHMFs are the same as the fiducial model. All the data used to make this plot (including individual data points and our best-fitting model) can be found here.

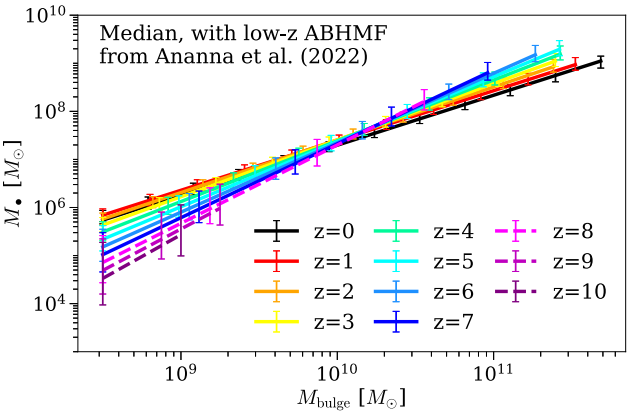

Figure D9. The evolution of the median $M_\bullet$–$M_{\rm bulge}$ relation, with $z \sim 0.2$ ABHMFs from Schulze & Wisotzki (2010) replaced by the Ananna et al. (2022) results. Higher redshift ABHMFs are the same as the fiducial model. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by e.g. *JWST*). All the data used to make this plot can be found here.

of +0.2 dex for all the ABHMF data points at $z = 0.2$ to keep consistency with those at $z = 1.5$.

D5.2 Systematic uncertainties in ABHMFs

Despite the corrections and exclusions for ABHMFs from Schulze & Wisotzki (2010) and Schulze et al. (2015), significant systematic differences remain among ABHMFs from different studies. For example, Ananna et al. (2022) obtained much higher $z \sim 0.2$ ABHMFs compared to Schulze & Wisotzki (2010). The potential causes for such differences include the different wavebands and bolometric corrections that were used (X-ray versus optical), different ways of correcting for obscured AGN, etc. We note that ABHMFs do provide important constraints on SMBH masses in TRINITY. Without any ABHMF data, TRINITY would yield a $M_\bullet$–$M_{\rm bulge}$ normalization with $\beta_{\rm BH,0} = 8.47$, and a too low AGN energy efficiency of $\epsilon_{\rm tot} \sim 3$ per cent. This is because the prior constraint on the local $M_\bullet$–$M_{\rm bulge}$ relation is not stringent enough as the sole constraint on SMBH masses, given the large inter-publication scatter (see Table 10). Therefore, we decided to keep ABHMF data in our data constraints.

To show the potential effects of adopting different ABHMF measurements, we did an experiment with the fiducial TRINITY model, replacing the low-redshift ABHMF from Schulze & Wisotzki (2010) with the one from Ananna et al. (2022). As shown in Fig. D9, the resulting redshift evolution of the $M_\bullet$–$M_{\rm bulge}$ relations is still consistent with the fiducial TRINITY model, although the difference is more significant at $z = 8$–10, where we do not have any AGN data. On the other hand, TRINITY needs to produce many more active SMBHs to match much higher number densities as required by Ananna et al. (2022). Consequently, a higher AGN efficiency of $\epsilon_{\rm tot} \sim 6.3$ per cent is adopted. Such a combination of $M_\bullet$–$M_{\rm bulge}$ relations and AGN efficiency naturally produces higher QLFs at $z \lesssim 2$ compared to the fiducial TRINITY results, as shown in Fig. D8, but the difference is well within the QLF uncertainties. Finally, a higher correlation coefficient between average SMBH accretion rate and $M_\bullet$ at fixed halo mass, $\rho_{\rm BH}$, is also needed to match the higher ABHMF at the massive end. Other than these quantitative changes, all the qualitative results remain invariant.

APPENDIX E: ALTERNATE MODEL PARAMETRIZATIONS

E1 Eddington-limited SMBH growth

In the fiducial model, we do not set any upper limit on the specific SMBH accretion rate. We also tested an alternate model where SMBHs cannot accrete at super-Eddington rates (hereafter called the 'Eddington-limited model'). Fig. E1 shows the comparison between the local $M_\bullet$–$M_{\rm bulge}$ relation with observations (top panel), and its redshift evolution (bottom panel). Given the limit in Eddington ratios, SMBHs cannot grow as fast as in the fiducial model. This results in a local $M_\bullet$–$M_{\rm bulge}$ relation that lies significantly below the observed values, and an increase in the normalization with increasing redshift. With limited accretion rates, TRINITY is also forced to recruit much higher AGN energy efficiencies – as high as 24 per cent – to get as many close-to-Eddington objects and reproduce the observations expressed in luminosities. Given the inconsistency with the observations, we do not adopt this model in the main text.

E2 Alternative galaxy–SMBH connections

In the fiducial TRINITY model, we make the galaxy–SMBH connection with redshift-dependent $M_{\rm bulge}$–M_* and $M_\bullet$–$M_{\rm bulge}$ relations. Given the observational uncertainties in these scaling relations, it is necessary to verify the robustness of our main results against these uncertainties. Therefore, we have run TRINITY with the following alternative assumptions: (a) the $M_{\rm bulge}$–M_* relation is redshift-independent and set to the observed one at $z = 0$ ('Const BMSM'); (b) the normalization of the $M_{\rm bulge}$–M_* relation is lower(higher) by setting $M_{\rm SB} = 11.5(9.0)$ (see equation 16, 'Small BMSM' and 'Big BMSM'); (c) the $M_{\rm bulge}$–M_* relation is steeper(flatter) by setting $k_{\rm SB} = 2.0(0.2)$ (also see equation 16, 'Steep BMSM' and 'Flat BMSM'[3]); (d) The $z = 0$ $M_{\rm bulge}$–M_* relation is fixed to the ones from either Häring & Rix (2004) or Kormendy & Ho (2013) ('Häring BHBM' and 'Kormendy BHBM'); (e) The galaxy–

[3]These alternative $M_{\rm SB}$ and $k_{\rm SB}$ values are chosen to cover the full range of 1σ uncertainties of the observed $M_{\rm bulge}$–M_* relation. See Fig. 2.

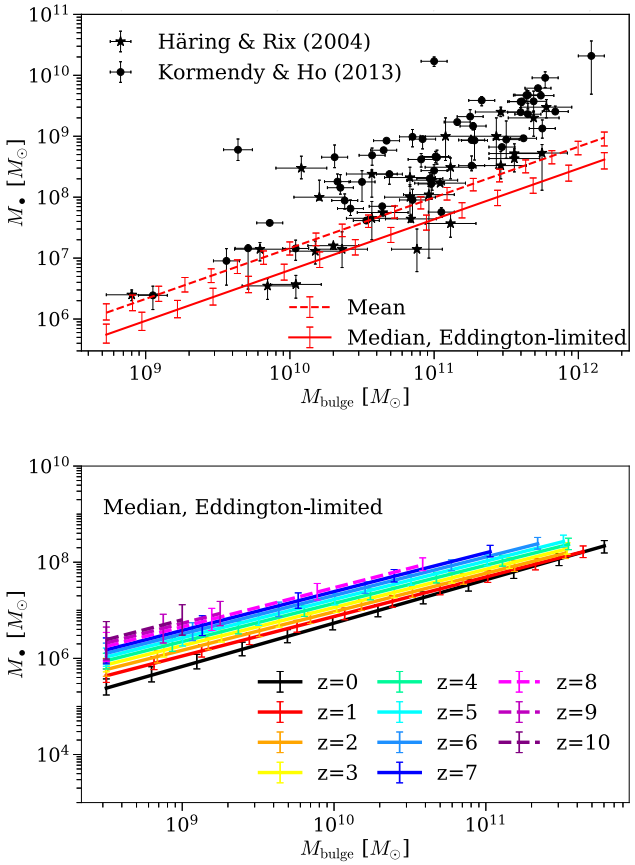

Figure E1. Top panel: The comparison between the $z = 0$ $M_\bullet$–$M_{\rm bulge}$ relation from the 'Eddington-limited' model and real data. Bottom panel: The redshift evolution of the $M_\bullet$–$M_{\rm bulge}$ relation from the 'Eddington-limited' model, where SMBH accretion is Eddington-limited. See Appendix E1. All the data used to make this plot (including the individual data points and our best-fitting model) can be found here.

SMBH connection is built by a redshift-dependent power-law $M_\bullet$–M_* relation, i.e. replacing $M_{\rm bulge}$ with M_* in equation (34) (BHSM); (f) The galaxy–SMBH connection is built by a *redshift-independent* power-law $M_\bullet$–M_* relation (Const BHSM); (g) The normalization of the $M_\bullet$–M_* relation has a redshift evolution as given by Merloni et al. (2014), and its slope is redshift-independent (Merloni BHSM). As shown in Fig. E2, most of these alternative models yield mutually consistent $M_\bullet$–M_* relations even before taking the inter-publication scatter of 0.2 dex (Table 10) into account.[4] The only exceptions are the 'Kormendy BHBM' and the 'BHSM' models. The 'Kormendy BHBM' model is consistent with the rest of the models when the inter-publication spread is included. We do note that the $M_\bullet$–$M_{\rm bulge}$ relation from Kormendy & Ho (2013) implies extremely massive black holes at fixed stellar mass. When constrained by galaxy SMFs and QPDFs, TRINITY would overproduce ABHMFs. In this sense, the $M_\bullet$–$M_{\rm bulge}$ from Kormendy & Ho (2013) is inconsistent with the galaxy data and ABHMFs in our data compilation. But to see the effect of an overall $M_\bullet$ offset on TRINITY results, we tried adding an offset in SMBH mass of 8.7–8.343 = 0.357 dex (where 8.343 is the normalization of the local $M_\bullet$–$M_{\rm bulge}$ relation given by the best-fitting fiducial model, also see Appendix H) to all the ABHMF data points, which effectively assumes that the Kormendy & Ho $M_\bullet$–$M_{\rm bulge}$ relation had been used to calibrate SMBH masses in the ABHMFs. With this offset, the 'Kormendy BHBM' model gives an AGN energy efficiency of $\epsilon_{\rm tot} \sim 3.5$ per cent. Such a smaller efficiency than that given by the fiducial model comes from more total SMBH mass with the same total AGN energy constraints from QLFs. Except for the systematic offset in AGN efficiency and the normalization of SMBH growth histories, the main results in this work are not affected. However, this systematic change in the inferred AGN energy with the normalization of the inferred/assumed local $M_\bullet$–$M_{\rm bulge}$ relation demonstrates that assuming a certain fixed SMBH mass normalization could induce inconsistency with other observational data sets. This further justifies our choice to use the *distribution* of $z = 0$ $M_\bullet$–$M_{\rm bulge}$ relations among different studies as prior constraints. As is pointed out by Reines & Volonteri (2015), the stellar mass measurements in Kormendy & Ho (2013) could be underestimated, leading to an overestimated $M_\bullet$–$M_{\rm bulge}$ normalization by ∼0.33 dex. The difference between TRINITY's best-fitting $M_\bullet$–$M_{\rm bulge}$ normalization with the Kormendy & Ho (2013) value, 0.357 dex, is also in line with this explanation. Given the potential inconsistency issue and bias in stellar mass measurements, we choose to present the results of the 'Kormendy BHBM' model in this appendix, instead of the main text of this work.

As for the 'BHSM' model, significantly higher values for $M_\bullet$ appear below $M_\bullet \sim 10^7 M_\odot$, compared to models that parametrize the $M_\bullet$–$M_{\rm bulge}$ relation. This is due to the 'BHSM' parametrization's inability to simultaneously reproduce the following with a single power law: (1) AGN observations constraining the massive end; and (2) The steeper $M_\bullet$–M_* slope at the low-mass end as in the $M_\bullet$–$M_{\rm bulge}$ parametrizations. We also note that such inter-model differences are more pronounced at $z = 8$–10, where no data exist. At these redshifts, our model results are pure extrapolations based on model assumptions and lower redshift data. At $z = 8$–10, the variance in $M_\bullet$-M_* relations from different models highlights the importance of upcoming high-z observations in constraining early galaxy–SMBH connections.

Although the 'Const BHSM' and the 'Merloni BHSM' models have fixed (non-)evolution with redshift, it is still worth checking if they predict qualitatively consistent SMBH accretion rates with the fiducial TRINITY model. As shown in Fig. E3, the 'Const BHSM' and the 'Merloni BHSM' models both predict average SMBH accretion rates and Eddington ratios as functions of $M_{\rm peak}$ and z. These predictions are qualitatively consistent with the fiducial TRINITY model.

Based on these experiments, we therefore argue that our results are relatively independent of the way that the galaxy–SMBH mass connection is parametrized.

E2.1 Redshift-independent SMBH mass–bulge mass relations

In the fiducial model, we assume a redshift-dependent $M_\bullet$–$M_{\rm bulge}$ relation. Here, we show the results from the 'constant $M_\bullet$–$M_{\rm bulge}$' model, where the redshift dependence is dropped. The best-fitting 'constant $M_\bullet$–$M_{\rm bulge}$' model gives $\log_{10}\widetilde{M}_\bullet = 8.378^{+0.161}_{-0.079} + 1.076^{+0.034}_{-0.034}\log_{10}(\frac{M_{\rm bulge}}{10^{11}M_\odot})$, which is consistent with the one from the fiducial model: $\log_{10}\widetilde{M}_\bullet = 8.342^{+0.091}_{-0.089} + 1.028^{+0.053}_{-0.035}\log_{10}(\frac{M_{\rm bulge}}{10^{11}M_\odot})$ (also see Appendix H). However, these two models differ in the correlation coefficient between SMBH average accretion rate and $M_\bullet$ at fixed host halo mass, $\rho_{\rm BH}$. As shown in Fig. E4, the 'constant $M_\bullet$–$M_{\rm bulge}$' model predicts significantly stronger correlation than the

[4]The 'Const BHSM' and 'Merloni BHSM' models (dotted lines) have pre-determined redshift evolution, and thus are included only for completeness.

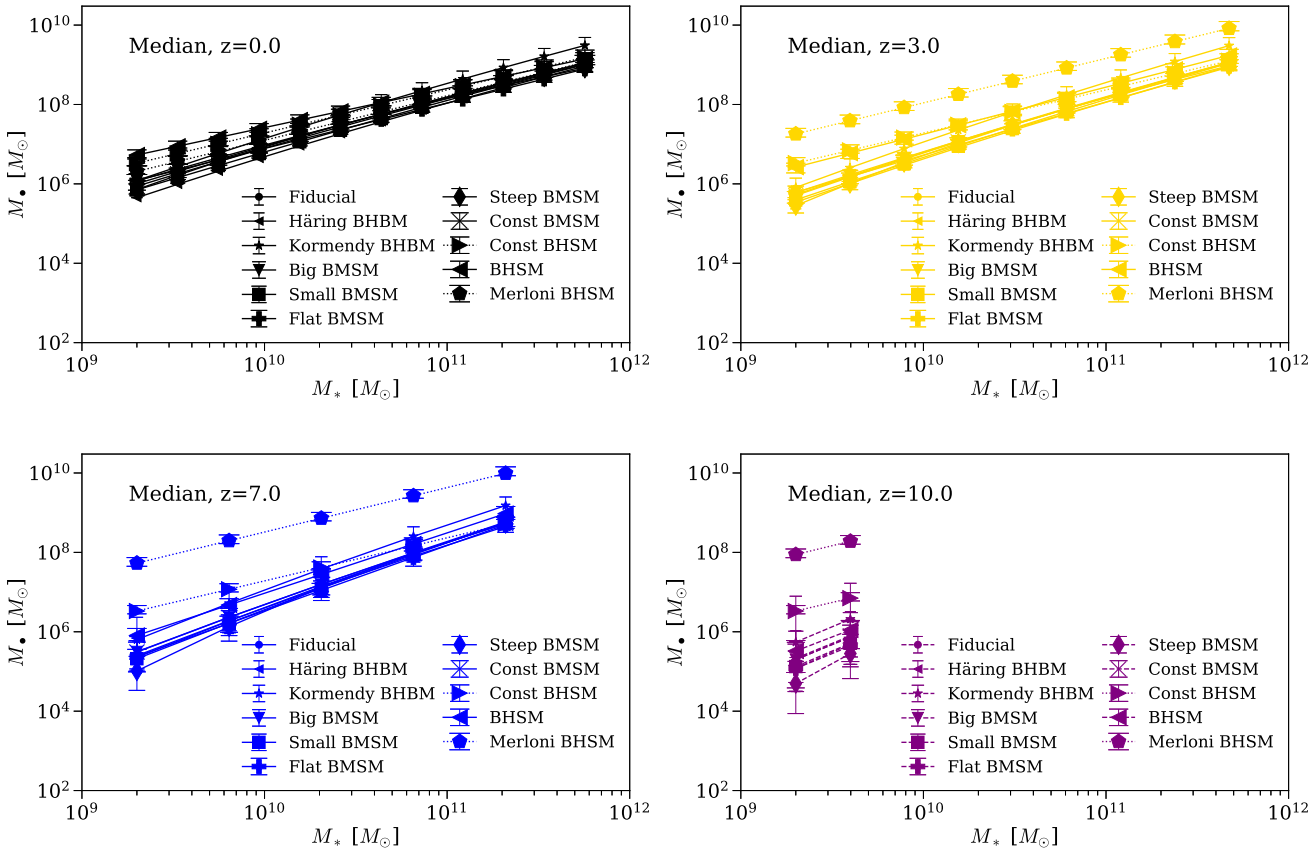

Figure E2. The median $M_\bullet$–M_* relations from different variant models at $z = 0$, 3, 7, and 10. See Appendix E2. The 'Const BHSM' and 'Merloni BHSM' models (dotted lines) have pre-determined redshift evolution in the median $M_\bullet$–M_* relation, and are thus shown only for completeness. For all the other redshifts, see here.

fiducial model. This is because in the fiducial model, the slope of the $M_\bullet$–$M_{\rm bulge}$ relation grows slightly towards higher redshifts, which naturally assigns more accretion to more massive SMBHs. Without this degree of freedom, the 'constant $M_\bullet$–$M_{\rm bulge}$' model needs higher $\rho_{\rm BH}$ values to reproduce the AGN data from massive galaxies. Fig. E5 shows the average $M_\bullet$, BHAR, Eddington ratio, and BHMR as functions of $M_{\rm peak}$ and z. The results are qualitatively consistent with the fiducial results. Quantitatively, the 'constant $M_\bullet$–$M_{\rm bulge}$' model predicts lower SMBH accretion rates and Eddington ratios at $M_{\rm peak} \gtrsim 10^{13} M_\odot$ and $z \lesssim 3$.

E3 Different assumptions about galaxy/BH mergers

Several previous studies opted to ignore mergers (e.g. Marconi et al. 2004), or made simple assumptions by linking SMBH mergers to halo mergers (e.g. Shankar et al. 2013). Here, we show the main results from TRINITY with alternate assumptions about SMBH mergers.

E3.1 Instant SMBH coalescence following halo mergers

One extreme case is the 'instant mergers' scenario, i.e. there is little delay between halo mergers and the coalescence of SMBHs. In this case, the central SMBH consumes *all* infalling SMBHs, regardless of how much of the infalling stellar mass is merged into the central galaxy versus the intracluster light (Section 2.2). Fig. E6 shows the average BHAR (left-hand panel) and BHMR (right-hand panel) from the 'instant mergers' model. It is clear that by forcing all the infalling satellite SMBHs to merge with central SMBHs, the vast majority of massive black hole growth at low redshifts must have been due to mergers, leaving little room for accretion. As a result, we see a precipitous drop in BHAR above $M_{\rm peak} \sim 10^{13} M_\odot$ below $z \sim 4$. Given that these low BHARs are in conflict with observations like Hlavacek-Larrondo et al. (2015) and McDonald et al. (2021) that show significant massive black hole accretion, we do not show other results from this model.

E3.2 No SMBH mergers or identical fractional merger contributions to SMBH and galaxy growth

In the fiducial model, we assume that the fractional merger contribution to SMBH and galaxy growth are proportional to each other. From the posterior parameter distribution, we found that the merger contribution to SMBH growth is smaller than the contribution to galaxy growth, i.e. $0 < f_{\rm scale} < 1$. Here, we consider two extreme cases. First, if the delay between galaxy mergers and the ensuing SMBH coalescence is sufficiently long, SMBH mergers would be rare, and the merger contribution to central SMBH growth becomes negligible. In this extreme case, we can assume that no SMBH mergers take place, and all central SMBH growth comes from accretion. In this 'no mergers' model, $f_{\rm scale} \equiv 0$ for all galaxies. The second extreme case we consider is if the fractional merger contributions to SMBH and galaxy growth are identical, i.e. $f_{\rm scale} \equiv 1$. In the following text, we call this scenario the 'same mergers' model.

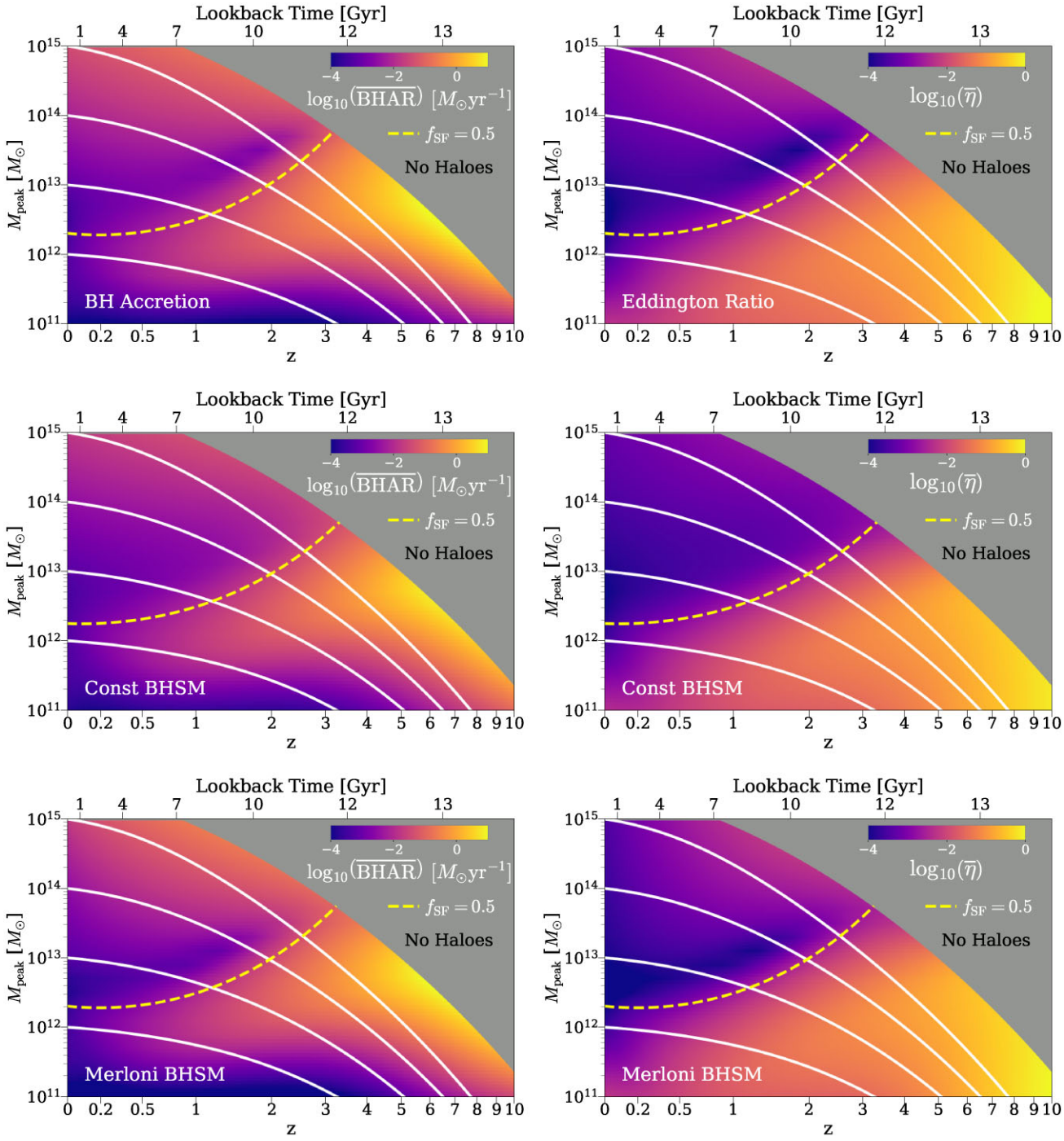

Figure E3. The average SMBH accretion rates (left column) and average Eddington ratios (right column) as functions of $M_{\rm peak}$ and z, from the fiducial (top panels), the 'Const BHSM' (middle panels), and the 'Merloni BHSM' models (bottom panels). See Appendix E2. All the data used to make this plot can be found here.

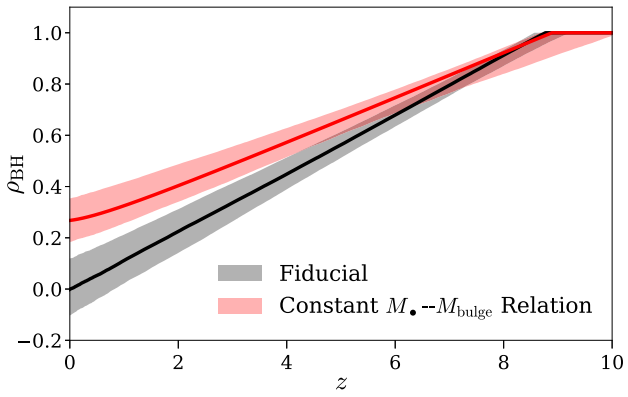

Figure E4. The correlation coefficient, $\rho_{\rm BH}$, between average SMBH accretion rate and $M_\bullet$ at fixed halo mass from the best-fitting model (black solid line) and the 'constant $M_\bullet$–$M_{\rm bulge}$' model. See Appendix E2.1. The shaded regions show the 68 per cent confidence intervals inferred from the model posterior distribution. The data used to make this plot can be found here.

Fig. E7 shows the resulting $M_\bullet$–$M_{\rm bulge}$ relations as functions of z from the 'no mergers' model (top panel), the fiducial model (middle panel), and the 'same mergers' model (bottom panel). The redshift evolution from all three models is largely consistent at $M_{\rm bulge} \gtrsim 10^{10.5} M_\odot$. Below $M_{\rm bulge} \sim 10^{10.5} M_\odot$, the 'same mergers' model predicts quantitatively higher $M_\bullet$ at fixed $M_{\rm bulge}$ (or M_*), and thus less SMBH mass growth. The bigger merger fraction depletes wandering SMBHs in low-mass galaxies before the predicted SMBH merger rates are fully accounted for, if the total SMBH growth is kept the same. Therefore, the total SMBH mass growth must be decreased to avoid such depletion.

Fig. E8 shows the average Eddington ratios as functions of $M_{\rm peak}$ and z from the 'no mergers' model (top panel), the fiducial model (middle panel), and the 'same mergers' model (bottom panel). The main difference between these three models is the average Eddington ratios of haloes with $M_{\rm peak} \gtrsim 10^{14} M_\odot$ below $z \sim 2$. From the top panel to the bottom panel, TRINITY attributes more and more SMBH growth to mergers among these haloes, producing lower and lower average Eddington ratios. However, the general 'downsizing' picture holds qualitatively across all these models.

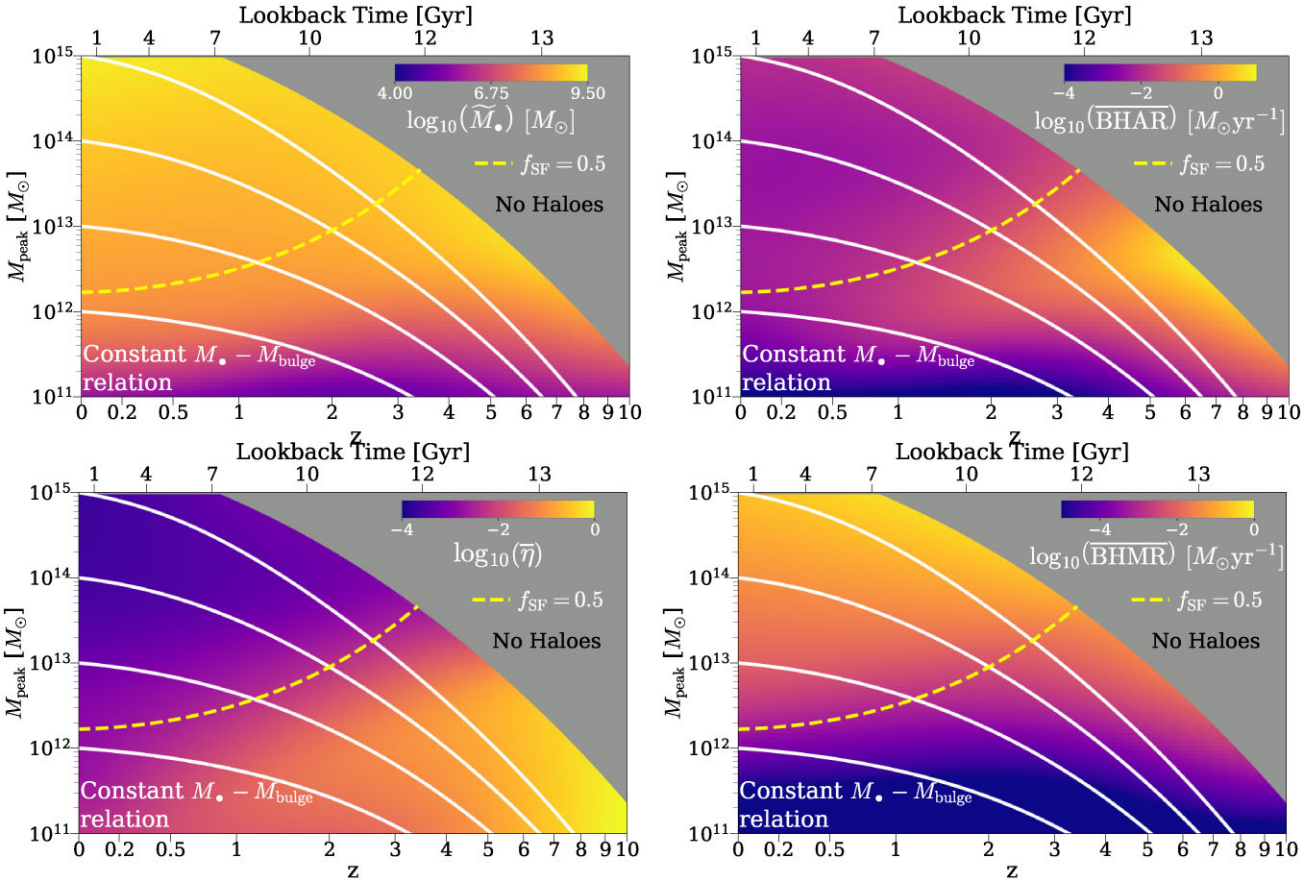

Figure E5. The average $M_\bullet$, BHAR, Eddington ratio, and BHMR as functions of $M_{\rm peak}$ and z from the 'constant $M_\bullet$–$M_{\rm bulge}$' model, where the $M_\bullet$–$M_{\rm bulge}$ relation is redshift-independent (see Appendix E2.1). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

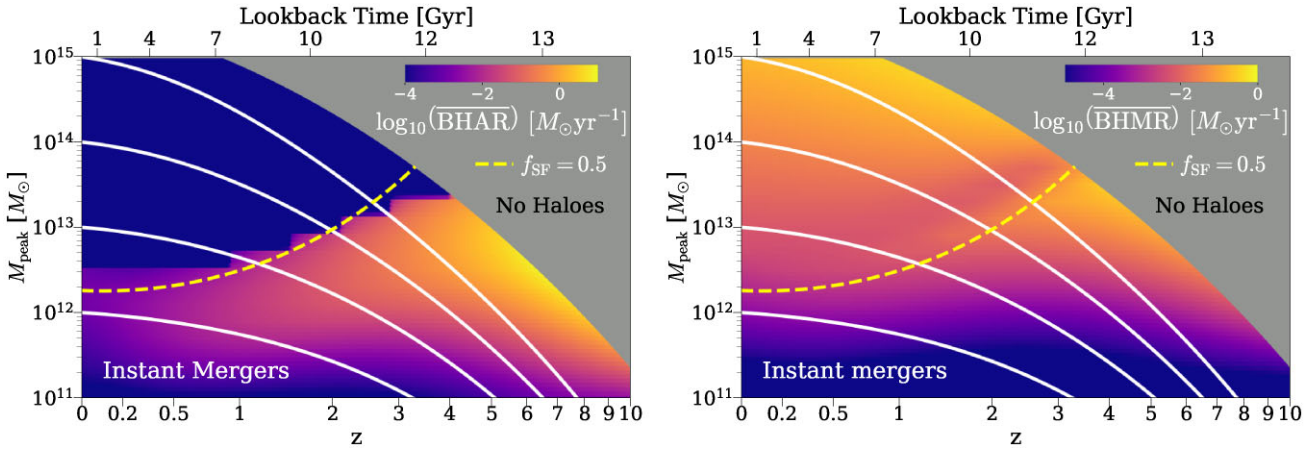

Figure E6. The average BHAR ($\overline{\rm BHAR}$, left-hand panel) and average BHMR ($\overline{\rm BHMR}$, right-hand panel) as a function of $M_{\rm peak}$ and z from the 'instant mergers' model (see Appendix E3.1). 'Instant mergers' means that all the infalling SMBHs in galaxy mergers are consumed immediately by the central SMBHs. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

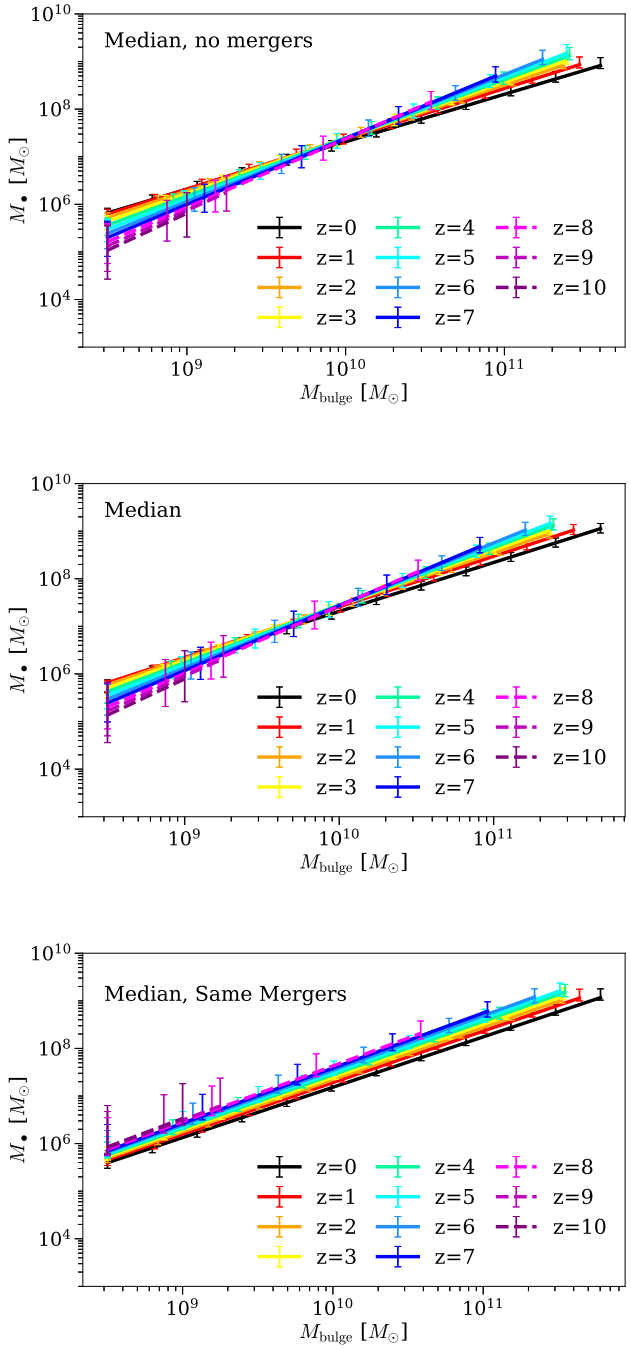

Figure E7. The median $M_\bullet$–$M_{\rm bulge}$ relations as a function of z from the 'no mergers' model (top panel, no SMBH mergers take place), the fiducial model (middle panel), and the 'same mergers' model (bottom panel, the fractional merger contribution to SMBH growth being the same as that for galaxy growth). See Appendix E3.2. All the data used to make this plot can be found here.

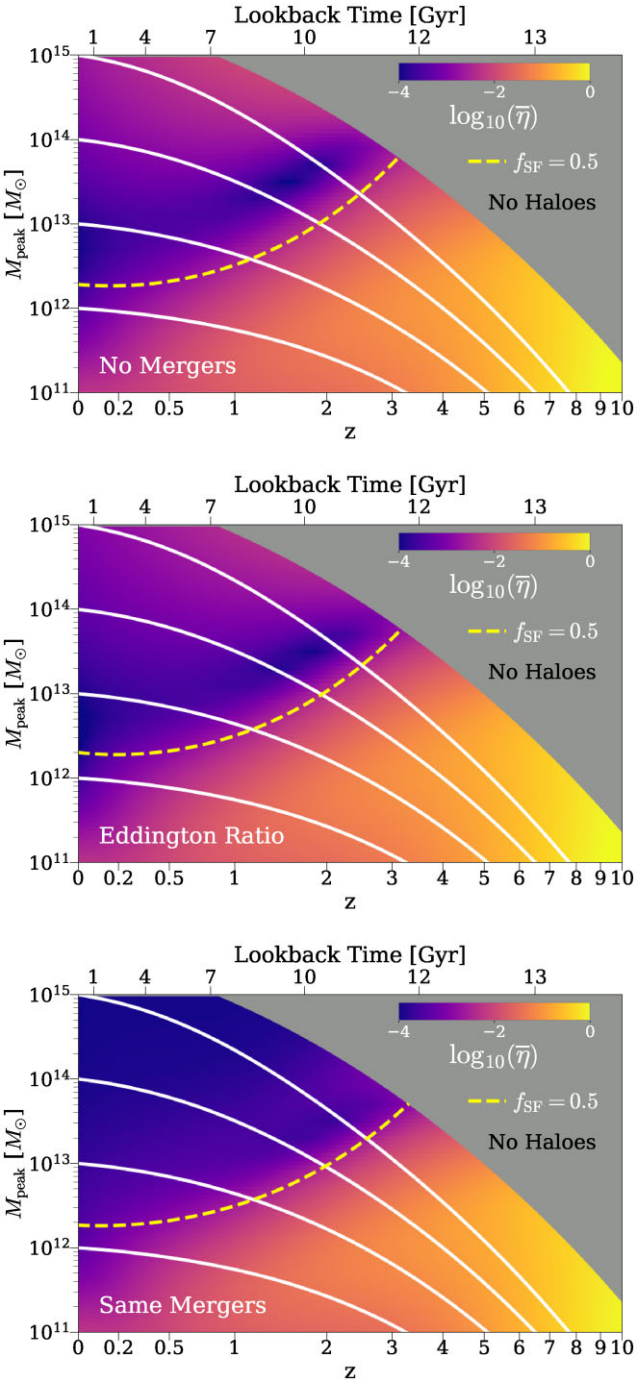

Figure E8. The average Eddington ratios as functions of $M_{\rm peak}$ and z from the 'no mergers' model (top panel), the fiducial model (middle panel), and the 'same mergers' model (bottom panel). See Appendix E3.2. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak}=10^{12}$, 10^{13}, 10^{14}, and $10^{15}M_\odot$ at $z=0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

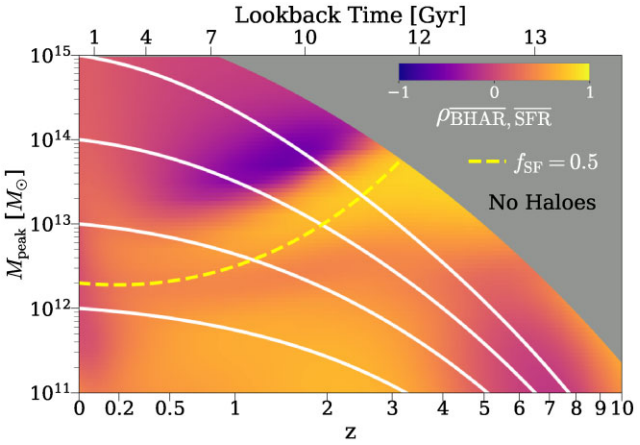

Figure F1. The correlation coefficient between average SMBH accretion and average galaxy star formation rate, $\rho_{\overline{\rm BHAR},\overline{\rm SFR}}$, as functions of $M_{\rm peak}$ and z from the fiducial model. See Appendix F. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

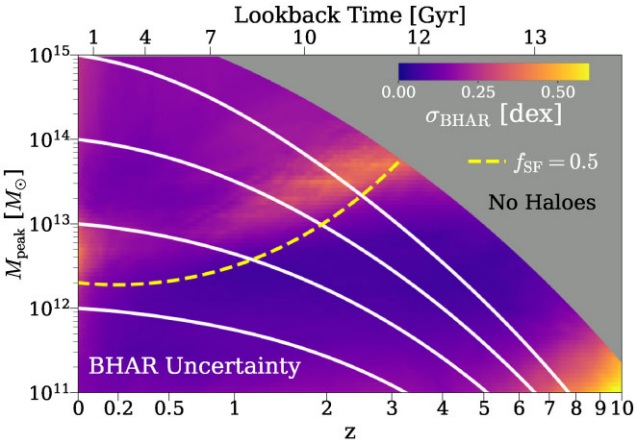

Figure F2. The 1σ uncertainty (from MCMC) in average SMBH accretion rate, $\sigma_{\rm BHAR}$ (in dex), as a function of $M_{\rm peak}$ and z from the fiducial model. See Appendix F. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction $f_{\rm SF}$ is 0.5 as a function of z. The white solid lines are the average mass growth curves of haloes with $M_{\rm peak} = 10^{12}$, 10^{13}, 10^{14}, and $10^{15} M_\odot$ at $z = 0$. The grey area shows where the number densities of dark matter haloes are negligible, and is therefore labelled as 'No Haloes'. All the data used to make this plot can be found here.

APPENDIX F: THE SYSTEMATIC EFFECT OF VARYING STAR FORMATION HISTORIES ON SMBH GROWTH HISTORIES

In TRINITY, we construct the galaxy–SMBH connection such that $M_\bullet$ is a function of the galaxy stellar mass. Stellar masses are calculated by integrating over galaxies' assembly histories. Consequently, a systematic change in the SFHs could in principle alter the SMBH growth histories from TRINITY. To quantify the sensitivity of SMBH accretion rates to the change in galaxy SFRs, we: (1) calculate average BHARs and SFRs as functions of $M_{\rm peak}$ and z for a representative subset of the MCMC chain; and then (2) calculate the correlation coefficient between the log of average BHAR and the log of average SFR, as a function of $M_{\rm peak}$ and z (Fig. F1).

At $0 < z \lesssim 3$ and $M_{\rm peak} < 13$, there is a moderate positive correlation between the average BHAR and SFR. This is because in this regime, systematically increasing SFR leads to larger galaxy stellar masses. To reproduce higher QPDF values in more massive stellar mass bins, as suggested by Aird et al. (2018) (Section 2.8), the BHAR needs to increase as well. Over $3 < z \lesssim 5$, we technically do not have QPDF constraints for different galaxy mass bins. Therefore, the positive correlation degrades towards higher redshifts. Around $z \sim 6$, the correlation becomes negligible. This is likely because we do not have any observational constraints at such a high redshift, except for the prior against super-Eddington quasars (Section 2.8). With such prior knowledge, TRINITY would not be forced to adjust BHAR along with any SFR change in this cosmic era.

It is also worth noting that at $M_{\rm peak} \gtrsim 10^{13} M_\odot$, there is a region with apparent negative correlation between the average BHAR and SFR. However, this is also the region where it is hard to robustly constrain SFRs as a function of halo mass. Thus, without better data constraints, we refrain from trying to explain its origin.

Fig. F2 shows the scatter in average SMBH accretion rate as a function of $M_{\rm peak}$ and z. Below $M_{\rm peak} \sim 10^{13} M_\odot$ and $z \sim 6$, the scatter in BHAR remains around 0.1 dex. Above $M_{\rm peak} \sim 10^{13} M_\odot$, the M_*–$M_{\rm peak}$ relation flattens, and thus galaxies with similar M_* can be found in a broader range of halo mass bins. This weakens the QPDFs' ability to constrain BHAR at fixed halo mass, because QPDFs are divided in different M_* bins. Ultimately, the uncertainties in BHAR are higher among more massive haloes. On the other hand, we do not have any constraints for AGNs at $z \gtrsim 6$. Thus, we see a significant increase in $\sigma_{\rm BHAR}$ with redshift in the range $6 < z < 10$.

APPENDIX G: TECHNICAL DETAILS ABOUT THE CALCULATION OF χ^2

Here, we introduce the details of the χ^2 calculation for any given model parameter set. In TRINITY, we first convert data points and their uncertainties into log units if they are in linear units. For the i-th data point with a value of $y_i^{+e'_{{\rm high},i}}_{-e'_{{\rm low},i}}$, we then convolve the error bars with a calculation tolerance of 0.01 dex:

$$e_{{\rm low/high},i} = \sqrt{e'^2_{{\rm low/high},i} + 0.01^2}\,. \tag{G1}$$

This calculation tolerance is set to prevent the model from overfitting to data points with very small confidence intervals. For this data point, suppose we have a model prediction, $\hat{y}_i$. If $|\hat{y}_i - y_i| \leq \epsilon_{\rm fit} \equiv 0.02$, then we assume that the model reproduces the data point sufficiently well, and ignore its contribution to the total χ^2. This error threshold is effectively a tolerance for the deviation of the analytical parametrizations from the actual scaling relations. If $|\hat{y}_i - y_i| > \epsilon_{\rm fit}$, we define

$$\Delta y_i = \begin{cases} \hat{y}_i - y_i - \epsilon_{\rm fit}, & \hat{y}_i > y_i \\ \hat{y}_i - y_i + \epsilon_{\rm fit}, & \hat{y}_i < y_i \end{cases}, \tag{G2}$$

and the χ^2_i for this data point is

$$\chi^2_i = \begin{cases} \left(\Delta y_i / e_{{\rm low},i}\right)^2, & \Delta y_i < -e_{{\rm low},i} \\ \left(\Delta y_i / e_{{\rm high},i}\right)^2, & \Delta y_i > e_{{\rm high},i} \\ \left(\Delta y_i / e_{{\rm med},i}\right)^2, & \text{otherwise} \end{cases}, \tag{G3}$$

where $e_{{\rm med},i}$ is a linear function of Δy_i:

$$e_{{\rm med},i}(\Delta y_i) = e_{{\rm low},i} + \frac{\Delta y_i + e_{{\rm low},i}}{e_{{\rm high},i} + e_{{\rm low},i}} \cdot (e_{{\rm high},i} - e_{{\rm low},i}). \tag{G4}$$

This definition is adopted to account for asymmetry in error bars, such that $e_{\mathrm{med},i} = e_{\mathrm{low},i}$ when $\Delta y_i = -e_{\mathrm{low},i}$ and $e_{\mathrm{med},i} = e_{\mathrm{high},i}$ when $\Delta y_i = e_{\mathrm{high},i}$. The total χ^2 is a summation of χ_i^2 over all the data points and the priors listed in Table 2:

$$\chi^2 = \sum_i \chi_i^2 + \mathrm{priors}. \tag{G5}$$

APPENDIX H: BEST-FITTING PARAMETER VALUES

The resulting best-fitting and 68 per cent confidence intervals for the posterior distributions follow:

Median star formation rates:

Characteristic v_{Mpeak} [km s^{-1}] (equation 4):

$$\log_{10}(V) = 2.289^{+0.017}_{-0.051} + \left(1.548^{+0.197}_{-0.221}\right)(a-1) + \left(1.218^{+0.147}_{-0.142}\right)\ln(1+z) + \left(-0.087^{+0.021}_{-0.021}\right)z$$

Characteristic SFR [$M_\odot$ yr^{-1}] (equation 5):

$$\log_{10}(\epsilon) = 0.556^{+0.045}_{-0.246} + \left(-0.944^{+1.133}_{-0.481}\right)(a-1) + \left(-0.042^{+0.887}_{-0.325}\right)\ln(1+z) + \left(0.418^{+0.054}_{-0.132}\right)z$$

Faint-end slope of SFR–v_{Mpeak} relation (equation 6):

$$\alpha = -3.907^{+0.148}_{-0.362} + \left(32.223^{+2.456}_{-1.724}\right)(a-1) + \left(20.241^{+1.627}_{-1.117}\right)\ln(1+z) + \left(-2.193^{+0.166}_{-0.175}\right)z$$

Massive-end slope of SFR–v_{Mpeak} relation (equation 7):

$$\beta = 0.329^{+0.239}_{-0.849} + \left(2.342^{+1.205}_{-0.953}\right)(a-1) + \left(0.492^{+0.190}_{-0.154}\right)z$$

Quenched fractions:

Characteristic v_{max} for quenching [km s^{-1}] (equation 10):

$$\log_{10}(v_{\mathrm{Q}}) = 2.337^{+0.013}_{-0.030} + \left(0.316^{+0.059}_{-0.143}\right)(a-1) + \left(0.283^{+0.022}_{-0.038}\right)z$$

Width in log-v_{max} for quenching [dex] (equation 11):

$$w_{\mathrm{Q}} = 0.193^{+0.018}_{-0.030} + \left(0.256^{+0.060}_{-0.126}\right)(a-1) + \left(0.062^{+0.018}_{-0.028}\right)z$$

Galaxy mergers:

Fraction of merging satellites that are transferred to the central galaxy (equation 13):

$$\log_{10}(f_{\mathrm{merge}}) = -0.748^{+0.066}_{-0.147}$$

The halo–galaxy connection:

M_* scatter at fixed M_{peak} [dex]:

$$\sigma_* = 0.279^{+0.004}_{-0.028}$$

Correlation coefficient between SSFR and M_* at fixed halo mass at $a = 0.5$ (i.e. $z = 1$) (equation 28):

$$\rho_{0.5} = 0.423^{+0.071}_{-0.100}$$

Systematics in stellar masses:

Offset between the true and the measured M_* [dex] (equation 25):

$$\mu = -0.111^{+0.127}_{-0.023} + \left(0.159^{+0.053}_{-0.043}\right)(a-1)$$

Additional systematic offset between the true and the measured SFRs (equation 26):

$$\kappa = 0.259^{+0.035}_{-0.025}$$

Scatter between the observed and the true M_* [dex] (equation 27):

$$\sigma = \min\{0.07 + 0.044^{+0.010}_{-0.008}(z-0.1),\ 0.3\}$$

Galaxy–SMBH connection:

Minimum SMBH occupation fraction (equation 31):

$$\log_{10}(f_{\mathrm{occ,min}}) = -2.640^{+2.285}_{-1.053} + \left(0.089^{+0.577}_{-1.277}\right)(a-1)$$

Characteristic halo mass and mass width where the SMBH occupation fraction changes significantly (equations 32–33):

$$\log_{10}(M_{\mathrm{h,c}}) = 10.804^{+2.792}_{-0.366} + \left(-14.220^{+6.976}_{-5.409}\right)(a-1)$$

$$w_{\mathrm{h,c}} = 3.355^{+0.266}_{-2.276} + \left(-0.574^{+4.948}_{-0.048}\right)(a-1)$$

Slope and zero-point of the SMBH mass–bulge mass ($M_\bullet$–M_{bulge}) relation (equations 35–36):

$$\gamma_{\mathrm{BH}} = 1.028^{+0.053}_{-0.035} + \left(0.036^{+0.043}_{-0.125}\right)(a-1) + \left(0.052^{+0.023}_{-0.033}\right)z$$

$$\beta_{\mathrm{BH}} = 8.343^{+0.091}_{-0.089} + \left(-0.173^{+0.047}_{-0.012}\right)(a-1) + \left(0.044^{+0.025}_{-0.013}\right)z$$

Scatter in the $M_\bullet$–M_{bulge} relation [dex] (equation 37):

$$\sigma_{\mathrm{BH}} = 0.269^{+0.051}_{-0.022}$$

SMBH mergers:

The fraction of SMBH growth due to mergers, relative to the fraction of galaxy growth due to mergers (equation 44):

$$\log_{10}(f_{\mathrm{scale}}) = -0.192^{+0.127}_{-1.494} + \left(0.000^{+1.640}_{-0.316}\right)(a-1)$$

AGN properties:

AGN duty cycles (equations 47–48):

$$\log_{10}(M_{\mathrm{duty}}) = 11.200^{+0.178}_{-0.003} + 1.269^{+0.049}_{-0.132}\ln(1+z)$$

$$\alpha_{\mathrm{duty}} = 4.692^{+0.175}_{-0.531} + \left(-2.723^{+0.313}_{-0.162}\right)\ln(1+z)$$

Power-law indices of the Eddington ratio distributions (equations 50 and 51):

$$c_1 = 0.527^{-0.023}_{-0.201} + \left(1.261^{+0.070}_{-0.308}\right)(a-1)$$

$$c_2 = 2.970^{-0.015}_{-0.339} + \left(-1.151^{+0.285}_{-0.215}\right)(a-1)$$

AGN energy efficiencies (equation 54):

$$\log_{10}(\epsilon_{\mathrm{tot}}) = -1.318^{+0.114}_{-0.010}$$

Correlation coefficient between SMBH accretion rate and mass at fixed halo mass (equation 56):

$$\rho_{\rm BH} = 0.001^{+0.117}_{-0.105} + \left(0.071^{+0.025}_{-0.160}\right)(a-1) + \left(0.123^{+0.005}_{-0.026}\right)z$$

AGN systematics:

Offset in the Eddington ratio between Ueda et al. (2014) and Aird et al. (2018) [dex] (equation 74):

$$\xi = -0.497^{+0.101}_{-0.058}$$

APPENDIX I: PARAMETER CORRELATIONS

Fig. I1 shows the rank correlation coefficients between all the model parameters, with the darker shades indicating stronger (positive or negative) correlations. It is natural to see correlations between different redshift evolution terms of the same parameter (e.g. ϵ_a and ϵ_{z1}), as each of them can partially mimic the behaviour of others at certain redshift intervals. In other words, different redshift evolution terms are not orthogonal to each other.

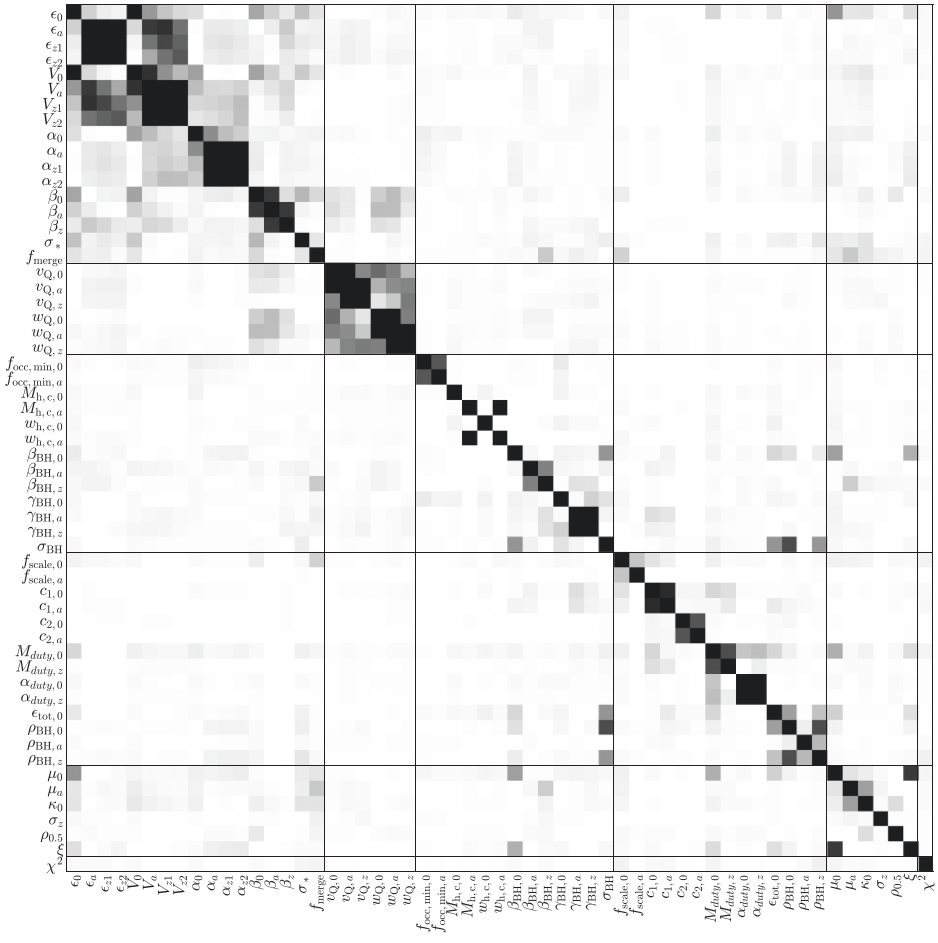

Figure I1. Rank correlation coefficients in the model posterior distribution. The darker shades indicate higher *absolute values* of correlation coefficients (both positive and negative). See Appendix I. All the data used to make this plot can be found here.

This paper has been typeset from a TEX/LATEX file prepared by the author.