

TRINITY I: Self-Consistently Modeling the Dark Matter Halo–Galaxy–Supermassive Black Hole Connection from $z = 0 – 10$

Haowen Zhang,¹[★] Peter Behroozi,¹ Marta Volonteri,² Joseph Silk,^{2,3,4}
 Xiaohui Fan,¹ Philip F. Hopkins,⁵ Jinyi Yang,^{1,*} and James Aird^{6,7}

¹University of Arizona, 933 N Cherry Ave., Tucson, AZ 85721, USA

²Institut d’Astrophysique de Paris (UMR 7095: CNRS & Sorbonne Université), 98 bis Bd. Arago, F-75014, Paris, France

³Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA

⁴BIPAC, Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK

⁵TAPIR, Mailcode 350-17, California Institute of Technology, Pasadena, CA 91125, USA

⁶Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK

⁷Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK

*Strittmatter Fellow

Accepted XXX. Received YYY; in original form ZZZ

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 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., 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 of the SMBH mass–bulge mass relation increases only mildly from $z = 0$ to $z = 3$, but decreases more strongly from $z = 3$ to $z = 10$; 2) The AGN radiative+kinetic efficiency is $\sim 0.04 – 0.07$, and does not show significant redshift dependence given the existing data constraints; 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: haloes – galaxies: evolution – quasars: supermassive 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_{bulge} , at $z \sim 0$; see

* E-mail: hwzhang0595@email.arizona.edu

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 $10^{-4} - 10^{-3}$ (Merloni et al. 2004; Silverman et al. 2008; Shankar et al. 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., $M_\bullet - M_{\text{bulge}}$ or $M_\bullet - \sigma$) have been measured via high spatial resolution spectroscopy and dynamics modeling (e.g., Magorrian et al. 1998; Ferrarese & Ford 2005; McConnell & Ma 2013). 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 hard 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 et al. 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. Nonetheless, 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 et al. 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, 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; Sijacki et al. 2015; Schaye 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. 2020).

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 and Eddington ratio of AGNs (see, e.g., Soltan 1982; Yu & Tremaine 2002; Merloni & Heinz 2008; Shankar et al. 2009; Tucci & Volonteri 2017).

The second group of models focuses on the galaxy–SMBH connection (e.g., Conroy & White 2013; Caplar et al. 2015, 2018; Yang et al. 2018). 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 datasets they try to fit. For example, Veale et al. (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 log-normal 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 et al. (2013). Compared to previous empirical models, TRINITY is constrained by a larger compilation of galaxy and AGN data, including not only quasar luminosity functions (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 cosmic star formation rates (CSFRs). The enormous joint constraining power of this dataset allows TRINITY to have both a more flexible parameterization as well as better constraints on the model parameters. In addition, TRINITY features more realistic modeling 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 quasar luminosity functions and the buildup 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 §2, we describe the methodology. §3 covers the simulation and observations used in TRINITY. §4 presents the results of our model, followed by the

discussion and comparison with other models in §5. Finally, we discuss the caveats of and the future directions for TRINITY in §6, and present conclusions in §7. In this work, we adopt a flat Λ CDM cosmology with parameters ($\Omega_m = 0.307$, $\Omega_\Lambda = 0.693$, $h = 0.678$, $\sigma_8 = 0.823$, $n_s = 0.96$) consistent with *Planck* results (Planck Collaboration et al. 2016). We use datasets that adopt the Chabrier stellar initial mass 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). Finally, we assume that every galaxy hosts a central SMBH.

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 star formation rate (SFR) as a function of halo mass and redshift. For a given choice in this parameter space, we can integrate the resulting star formation rates (SFRs) over average halo assembly histories to get the stellar mass–halo mass (SMHM) relation (§2.2). We then convert total galaxy mass to *bulge* mass with a scaling relation from observations. Next, we connect SMBHs with galaxies by parameterizing the redshift evolution of the SMBH mass–bulge mass (M_\bullet – M_{bulge}) relation (§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 parameterize the Eddington ratio distributions (i.e., specific SMBH accretion rates) and mass-to-energy conversion efficiency, which determines how SMBH growth translates to the observed distribution of SMBH luminosities. In brief, this modeling process gives the distribution of galaxy and SMBH properties. After modeling AGN radiative and kinetic luminosities (§2.7) as well as correcting for systematic effects, these properties are used to predict the galaxy and AGN observables (§2.8). We compare these predictions to observations to compute a likelihood function, and use a Markov Chain Monte Carlo (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 (as in Behroozi et al. 2013) instead of following individual halo and galaxy histories (as in the UNIVERSEMAchine; Behroozi et al. 2019). 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, 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 parameterization for the halo–galaxy connection in TRINITY as was shown to work successfully in the

UNIVERSEMAchine (Behroozi et al. 2019). Although simpler parameterizations exist, this choice makes future integration with the UNIVERSEMAchine easier. The UNIVERSEMAchine modeled star-forming and quiescent haloes individually, but TRINITY models halo population averages, and we maintain this parameterization in TRINITY. In practice, however, TRINITY only depends on the total star formation rate of all haloes in a given mass bin, which depends almost exclusively on the star formation rate for star-forming galaxies and the quiescent fraction as a function of halo mass and redshift.

Our model assumes that the median star formation rates (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 ($v_{\text{max}} = \max(\sqrt{GM(<R)/R})$) at the time when it reaches its peak mass, v_{Mpeak} , as a proxy for the peak halo mass M_{peak} . This choice reduces the sensitivity to pseudo-evolution in halo mass definitions and to spikes in v_{max} during mergers (Behroozi et al. 2019). Our parameterization is:

$$\text{SFR}_{\text{SF}} = \frac{\epsilon}{v^\alpha + v^\beta} \quad (1)$$

$$v = \frac{v_{\text{Mpeak}}}{V \cdot \text{km s}^{-1}} \quad (2)$$

$$\alpha = \frac{1}{1+z} \quad (3)$$

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

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

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

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

The *median* SFRs of star-forming galaxies (SFR_{SF}) are a power-law with slope $-\alpha$ for $v_{\text{Mpeak}} \ll V$, and another power-law with slope $-\beta$ for $v_{\text{Mpeak}} \gg V$. The parameter ϵ is the characteristic SFR when $v_{\text{Mpeak}} \sim V$. We remove the Gaussian boost in SFR at $v_{\text{Mpeak}} \sim V$ 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_Q , as a function of redshift and v_{Mpeak} :

$$f_Q = 1 - \frac{1}{1 + \exp(x)} \quad (8)$$

$$x = \frac{\log_{10}(v_{\text{Mpeak}}) - v_Q}{w_Q} \quad (9)$$

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

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

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

$$\begin{aligned} \text{SFR}_{\text{tot}} = & \text{SFR}_{\text{SF}} \times (1 - f_Q) \times \exp(0.5(\sigma_{\text{SFR,SF}} \ln 10)^2) \\ & + \text{SSFR}_Q \times M_* \times f_Q \times \exp(0.5(\sigma_{\text{SFR,Q}} \ln 10)^2), \end{aligned} \quad (12)$$

where the exponentials reflect the difference between the *average* and *median* values of log-normal 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_{merge} ,

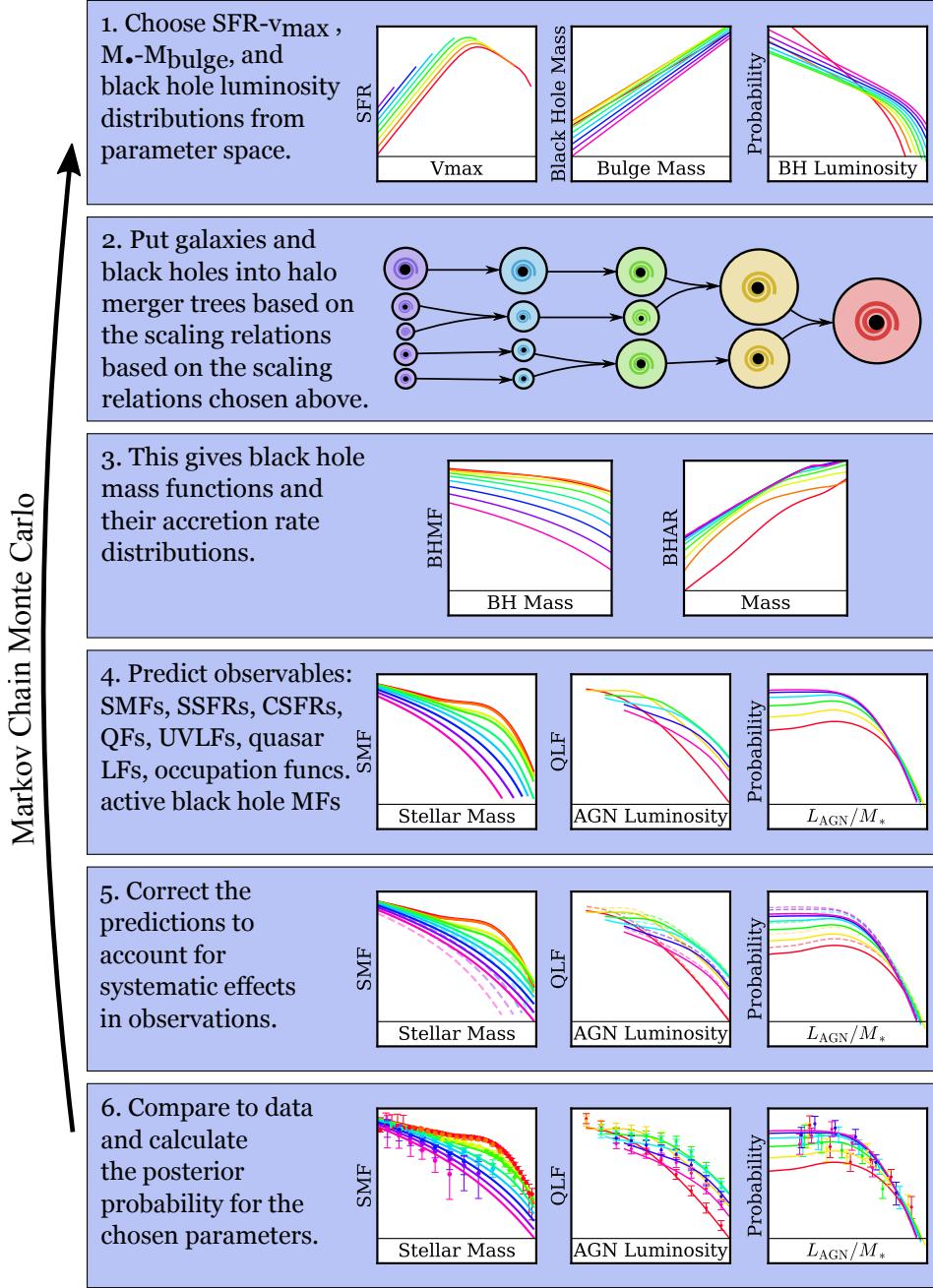


Figure 1. Visual summary of the methodology used to constrain the halo–galaxy–SMBH connection. See §2 for details.

of the stars from incoming galaxies are merged to the central galaxies. As in Behroozi et al. (2019), we assume f_{merge} to be independent of halo mass due to the approximately self-similar nature of haloes. We also assume f_{merge} to be redshift-independent. The average stellar mass in a given halo mass bin at a given redshift z is correspondingly:

$$\begin{aligned} M_*(t) &= \int_0^t (1 - f_{\text{loss}}(t-t')) \text{SFR}_{\text{tot}}(t') dt' \\ &+ f_{\text{merge}} \int_0^t \int_0^{t'} (1 - f_{\text{loss}}(t-t'')) \dot{M}_{*,\text{inc}}(t', t'') dt'' dt' \quad (13) \\ f_{\text{loss}}(T) &= 0.05 \ln \left(1 + \frac{T}{1.4 \text{ Myr}} \right), \quad (14) \end{aligned}$$

where $f_{\text{loss}}(T)$ is the stellar mass loss fraction as a function of stellar age T from Behroozi et al. (2013), SFR_{tot} is the total average SFR from Eq. 12, and $\dot{M}_{*,\text{inc}}(t', t'')$ is the rate at which the incoming satellite galaxies disrupt, as a function of the time of disruption t' and the time that the stellar population formed, t'' .

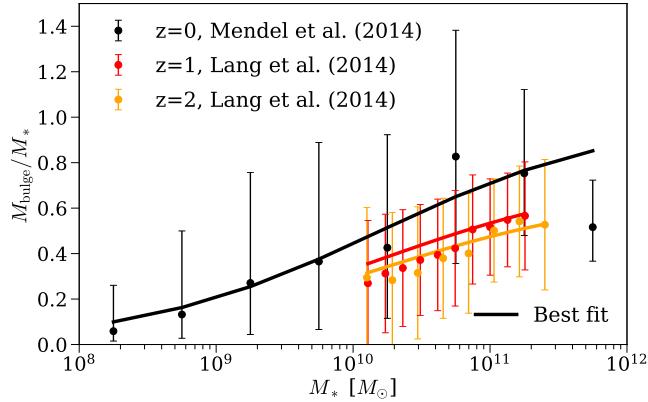


Figure 2. The fit to the median galaxy bulge mass–total mass relation for $z = 0 – 2$ (solid lines, Eqs. 15–16). Observed data points are from Mendel et al. (2014) and Lang et al. (2014). The error bars from Mendel et al. (2014) represent the 16–84th percentile range of the M_{bulge}/M_* ratios from the SDSS catalog, and those from Lang et al. (2014) are based on the 68% 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](#).

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 log-normal 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 relation (M_\bullet – M_{bulge} , §2.4). To make the halo–galaxy–SMBH connection, we need to convert *total* galaxy mass M_* to the *bulge* mass M_{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:

$$M_{\text{bulge}} = \frac{f_z(z)M_*}{1 + \exp(-1.13(\log_{10}M_* - 10.2))} \quad (15)$$

$$f_z(z) = \frac{z+2}{2z+2}. \quad (16)$$

This fit is shown in Fig. 2. It should be noted that no data points exist beyond $z = 2.5$, so Eq. 15 is extrapolated at $z > 2.5$. With the functional form chosen here, M_{bulge}/M_* asymptotes at high redshifts to half the value of M_{bulge}/M_* at $z = 0$.

Disk–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 vs. observational. As a result, we subsume the scatter in the bulge-to-total mass relation into the scatter of the M_\bullet – M_{bulge} relation, as the two scatters are degenerate.

At $0 < z < 8$, stellar mass functions (SMFs) primarily constrain the halo–galaxy connection. Beyond $z = 8$, SMFs are not available, so we constrain the halo–galaxy connection with galaxy UV luminosity functions 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, \tilde{M}_{UV} , and the log-normal scatter, $\sigma_{M_{\text{UV}}}$, as functions of SFR, M_{peak} , and redshift from the output of

the UNIVERSE MACHINE:

$$\tilde{M}_{\text{UV}} = k_{\text{UV}} \times \log_{10} \text{SFR} + b_{\text{UV}} \quad (17)$$

$$\sigma_{M_{\text{UV}}} = k_{\sigma_{\text{UV}}} \times \log_{10} M_{\text{peak}} + b_{\sigma_{\text{UV}}} \quad (18)$$

$$k_{\text{UV}} = a_k (\log_{10} M_{\text{peak}})^2 + b_k \log_{10} M_{\text{peak}} + c_k (a-1) + d_k \quad (19)$$

$$b_{\text{UV}} = a_b (\log_{10} M_{\text{peak}})^2 + b_b \log_{10} M_{\text{peak}} + c_b (a-1) + d_b \quad (20)$$

$$k_{\sigma_{\text{UV}}} = a_{k_\sigma} z + b_{k_\sigma} \quad (21)$$

$$b_{\sigma_{\text{UV}}} = a_{b_\sigma} z + b_{b_\sigma}. \quad (22)$$

Details of the fitting process are shown in Appendix D. UNIVERSE MACHINE models UV luminosities using the Flexible Stellar Population Synthesis code (FSPS; Conroy & White 2013), and Eqs. 17–22 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 ($M_{*,\text{obs}}$) and the true stellar mass ($M_{*,\text{true}}$) to model uncertainties from the IMF, SPS model, the dust model, the star formation history (SFH) model, assumed metallicities, and redshift errors:

$$\log_{10}\left(\frac{M_{*,\text{obs}}}{M_{*,\text{true}}}\right) = \mu. \quad (23)$$

The offset μ has the following redshift scaling:

$$\mu = \mu_0 + \mu_a (a-1). \quad (24)$$

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 specific star formation rates that peak near $z \sim 2$, which are most evident when comparing observed specific star formation rates 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 star formation rates. The total offset between the observed ($\text{SFR}_{*,\text{obs}}$) and true SFRs ($\text{SFR}_{*,\text{true}}$) is:

$$\log_{10}\left(\frac{\text{SFR}_{*,\text{obs}}}{\text{SFR}_{*,\text{true}}}\right) = \mu + \kappa \exp\left(-\frac{(z-2)^2}{2}\right). \quad (25)$$

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, log-normal scatter in the measured stellar mass relative to the true mass:

$$\sigma(z) = \min(\sigma_0 + \sigma_z z, 0.3). \quad (26)$$

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 star formation

rate 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):

$$\rho(a) = 1 + (4\rho_{0.5} - 3.23)a + (2.46 - 4\rho_{0.5})a^2, \quad (27)$$

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

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 et al. 2006; Aller & Richstone 2007; Hu 2008; Beifiori et al. 2012; Shankar et al. 2016; van den Bosch 2016). Here, we parameterize the relation between SMBHs and galaxy bulge mass. Unlike the modeling of star formation rates (§2.2), we assign SMBH *masses* first, and then infer their growth rates. This approach is necessary because the relationship between SMBH mass and galaxy mass is much less uncertain than the relationship between SMBH mass growth and galaxy mass or growth.

We parameterize the *median* M_\bullet – M_{bulge} relation as a redshift-dependent power-law:

$$\log_{10} \tilde{M}_\bullet = \beta_{\text{BH}} + \gamma_{\text{BH}} \log_{10} \left(\frac{M_{\text{bulge}}}{10^{11} M_\odot} \right) \quad (28)$$

$$\beta_{\text{BH}} = \beta_{\text{BH},0} + \beta_{\text{BH},a}(a-1) + \beta_{\text{BH},z}z \quad (29)$$

$$\gamma_{\text{BH}} = \gamma_{\text{BH},0} + \gamma_{\text{BH},a}(a-1) + \gamma_{\text{BH},z}z. \quad (30)$$

We set Gaussian priors on $\beta_{\text{BH},0}$ and $\gamma_{\text{BH},0}$ from constraints on the local M_\bullet – M_{bulge} relation, which will be discussed in §3.2.2. With Eqs. 28–30, 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 log-normal scatter in SMBH mass at fixed bulge mass (σ_{BH}). We assume σ_{BH} to be redshift-independent. This is because a redshift dependent σ_{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 §3.2.2 for more details.

Since the scatter in bulge mass at fixed stellar mass is subsumed in σ_{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_{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:

$$\sigma_{\text{tot}} = \sqrt{(\sigma_* \times \gamma_{\text{BH}})^2 + \sigma_{\text{BH}}^2}. \quad (31)$$

This log-normal scatter results in a difference between the *mean* (\bar{M}_\bullet) and *median* SMBH masses (\tilde{M}_\bullet) at fixed halo mass:

$$\overline{M}_\bullet = \tilde{M}_\bullet \times \exp(0.5(\sigma_{\text{tot}} \ln 10)^2). \quad (32)$$

2.5 SMBH mergers and accretion

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

$$\begin{aligned} \overline{\text{BHMR}} \cdot \Delta t = & (\text{Average BH Mass Now} \\ & - \text{Average BH Mass Inherited from Previous Timestep}) \\ & \times f_{\text{merge,BH}}, \end{aligned} \quad (33)$$

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 ($\overline{\text{BHAR}}$) for a certain halo mass bin is:

$$\begin{aligned} \overline{\text{BHAR}} \cdot \Delta t = & (\text{Average BH Mass Now} \\ & - \text{Average BH Mass Inherited from Previous Timestep}) \\ & \times (1 - f_{\text{merge,BH}}). \end{aligned} \quad (34)$$

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

$$f_{\text{merge,BH}} = f_{\text{scale}} \times \frac{f_{\text{merge}} \dot{M}_{*,\text{inc}}}{\text{SFR} + f_{\text{merge}} \dot{M}_{*,\text{inc}}}, \quad (35)$$

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

$$\log_{10}(f_{\text{scale}}) = f_{\text{scale},0} + f_{\text{scale},1}(a-1). \quad (36)$$

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

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 timescales, 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 ($\bar{M}_{\bullet,\text{wander}}$) for each halo mass bin is thus:

$$\begin{aligned} \bar{M}_{\bullet,\text{wander}} = & \text{Total Infalling BH Mass} \\ & - \int_0^t \overline{\text{BHMR}} \cdot dt. \end{aligned} \quad (37)$$

Although wandering SMBHs do not contribute to the observed M_\bullet – M_{bulge} relation, they do contribute to quasar luminosity functions during their formation. This is most important when comparing our results with previous studies (e.g., §5.2), which have often not modeled wandering SMBHs. For full details about calculating the av-

verage inherited SMBH mass from the previous timestep ($\bar{M}_{\bullet,\text{inherit}}$) and the average infalling SMBH mass ($\bar{M}_{\bullet,\text{infall}}$), see Appendix C.

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

In the real Universe, only a subset of SMBHs are actively accreting at a given time. In TRINITY, we parametrize the fraction of galaxies that host active SMBHs (i.e., the AGN duty cycle, f_{duty}) as a function of M_{\bullet} and z :

$$f_{\text{duty}}(M_{\bullet}, z) = f_1(M_{\bullet}) \times f_2(z) \quad (38)$$

$$f_1(M_{\bullet}) = \frac{\exp(x)}{1 + \exp(x)} \quad (39)$$

$$x = \frac{\log_{10}(M_{\bullet}) - M_{\bullet,c}}{w_{\bullet}} \quad (40)$$

$$f_2(z) = f_{2,0} + f_{2,a}(a - 1), \quad (41)$$

where $f_1(M_{\bullet})$ is a sigmoid function of SMBH mass, and $f_2(z)$ is a linear function of the scale factor $a = 1/(1+z)$. In this work, we define f_{duty} to be the fraction of active SMBH host galaxies among *all* galaxies, i.e., regardless of whether they actually host SMBHs. With the assumption that the SMBH occupation fraction is $f_{\text{occ}} \equiv 1$, this is equivalent to the fraction of active SMBHs among all the SMBHs hosted by galaxies. In §6.2, we discuss the potential effect on AGN Eddington ratio distributions of a lower f_{occ} .

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

$$P(\eta|\eta_0, c_1, c_2) = f_{\text{duty}} \frac{P_0}{\left(\frac{\eta}{\eta_0}\right)^{c_1} + \left(\frac{\eta}{\eta_0}\right)^{c_2}} + (1 - f_{\text{duty}})\delta(\eta) \quad (42)$$

$$c_1 = c_{1,0} + c_{1,a}(a - 1) \quad (43)$$

$$c_2 = c_{2,0} + c_{2,a}(a - 1), \quad (44)$$

where η is 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

$$\int_0^{\infty} \frac{P_0}{\left(\frac{\eta}{\eta_0}\right)^{c_1} + \left(\frac{\eta}{\eta_0}\right)^{c_2}} d\log\eta = 1. \quad (45)$$

This functional form is flexible enough to approximate many past assumptions for the shape of the ERDF (e.g., Gaussian distributions and Schechter functions). Given the non-zero scatter in SMBH mass at fixed halo mass (Eq. 31), more massive SMBHs in a certain halo mass bin have higher accretion rates than their smaller counterparts.

The characteristic Eddington ratio η_0 in Eq. 42 is *not* a free parameter, but is constrained by the parametrizations in Eqs. 38–42. Letting $\bar{\eta}$ be the average Eddington ratio, we have from Eq. 42 that:

$$\bar{\eta} = \int_0^{\infty} \eta P(\eta|\eta_0, c_1, c_2) d\log\eta, \quad (46)$$

and by definition

$$\bar{\eta} = \frac{\epsilon_{\text{tot}} \overline{\text{BHAR}} \times 4.5 \times 10^8 \text{ yrs}}{(1 - \epsilon_{\text{tot}}) \bar{M}_{\bullet}}, \quad (47)$$

where \bar{M}_{\bullet} (Eq. 28) and $\overline{\text{BHAR}}$ (Eq. 34) are the average SMBH mass and black hole accretion rate as functions of halo mass and redshift, respectively. The parameter ϵ_{tot} is the efficiency to release

energy (both radiative and kinetic) through accretion. We hence solve for η_0 by combining Eqs. 46 and 47. In this work, $\log_{10}(\epsilon_{\text{tot}})$ is assumed to scale linearly with the scale factor $a = 1/(1+z)$:

$$\log_{10}(\epsilon_{\text{tot}}) = \epsilon_{\text{tot},0} + \epsilon_{\text{tot},a}(a - 1). \quad (48)$$

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, ϵ_{rad} and ϵ_{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 η_{rad} :

$$\eta_{\text{rad}} = \begin{cases} \eta^2/0.03, & \eta \leq 0.03 \\ \eta, & 0.03 < \eta \leq 2 \\ 2[1 + \ln(\eta/2)], & \eta > 2 \end{cases}. \quad (49)$$

For $\eta \leq 2$, the scaling between η_{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 Eq. 49. For $\eta \geq 2$, we adopt a logarithmic scaling to account for the fact that at such high accretion rates, the accretion disk becomes thick, trapping part of the outgoing radiation (Minehige et al. 2000). The kinetic component η_{kin} is, by definition:

$$\eta_{\text{kin}} = \eta - \eta_{\text{rad}}, \quad \eta < 0.03. \quad (50)$$

At a given $\eta < 0.03$, Eq. 50 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:

$$\epsilon_{\text{rad}} = \epsilon_{\text{tot}} \times \begin{cases} \eta/0.03, & \eta \leq 0.03 \\ 1, & 0.03 < \eta \leq 2 \\ 2/\eta[1 + \ln(\eta/2)], & \eta > 2 \end{cases}, \quad (51)$$

and:

$$\epsilon_{\text{kin}} = \begin{cases} \epsilon_{\text{tot}}(1 - \eta/0.03), & \eta < 0.03 \\ 0, & \eta > 0.03 \end{cases}, \quad (52)$$

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

$$\frac{L_X}{\text{erg/s}} = 10^{38.1} \times \frac{M_{\bullet}}{M_{\odot}} \times \eta_X \quad (53)$$

$$P(\eta_X) = P(\eta) \frac{d\log\eta}{d\log\eta_X}, \quad (54)$$

where X is either “rad” or “kin” and where $d\log\eta/d\log\eta_X$ is calculated using Eqs. 49–50.

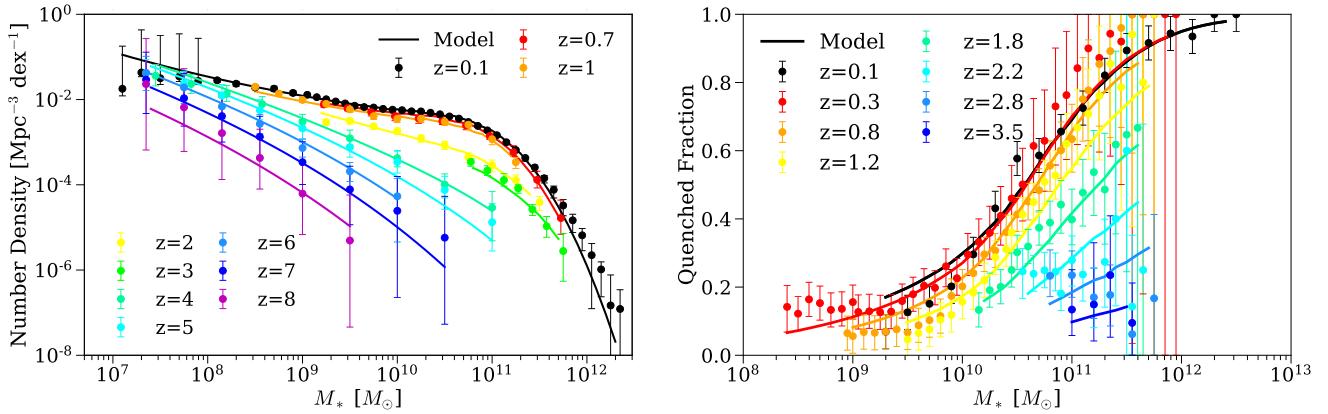


Figure 3. Left Panel: Comparison between observed galaxy stellar mass functions (SMFs) and our best-fitting model from $z = 0 - 8$. The observed stellar mass functions are listed in Table 4. **Right Panel:** Comparison between observed galaxy quenched fractions (QFs) and our best-fitting model from $z = 0 - 4$. The observed quenched fractions 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](#).

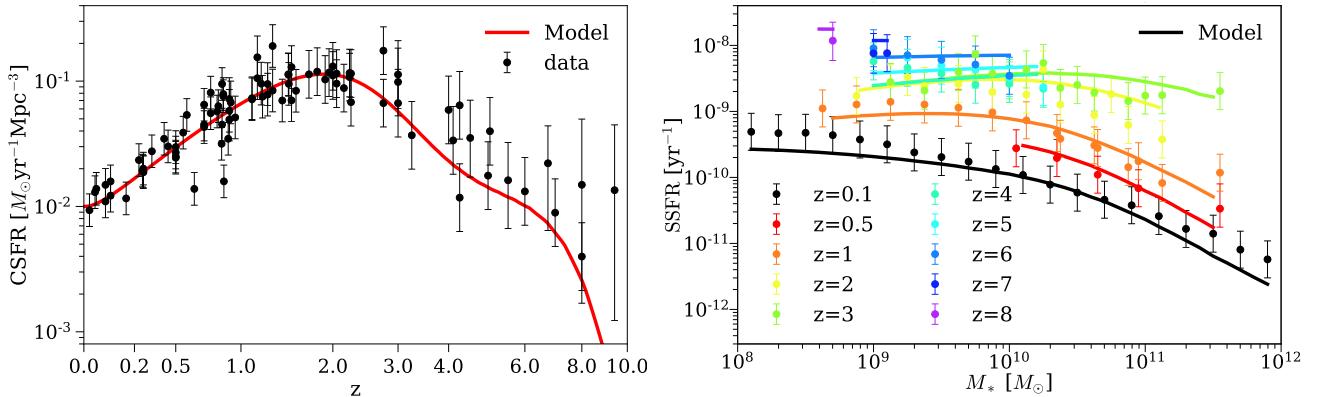


Figure 4. Left Panel: Comparison between observed cosmic star formation rates (CSFRs) and our best-fitting model from $z = 0 - 10$. The references for observations are listed in Table 6. **Right Panel:** Comparison between observed galaxy specific star formation rates (SSFRs) as a function of stellar 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](#).

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:

$$\phi_{\text{BH}}(M_\bullet, z) = \int_0^\infty \phi_h(M_{\text{peak}}, z) P(M_\bullet | M_{\text{peak}}, z) dM_{\text{peak}}, \quad (55)$$

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

To model active black hole mass functions 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 active black hole mass function is:

$$\phi_{\text{ABH}}(M_\bullet, z) = \int_0^\infty \int_{\eta_{\text{rad},\min}=0.01}^\infty \phi_h(M_{\text{peak}}, z) P(M_\bullet | M_{\text{peak}}, z) \times P(\eta_{\text{rad}} | M_\bullet, M_{\text{peak}}, z) d\eta_{\text{rad}} dM_{\text{peak}}. \quad (56)$$

For the type I quasar SMBH mass functions from Kelly & Shen (2013), we include all SMBHs with $\eta > 0$. This is because modeling of the underlying $M_{\text{BH}} - L_{\text{bol}}$ 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_{obs} as a function of X-ray luminosity from Merloni et al. (2014):

$$F_{\text{obs}}(L_X) = 0.56 + \frac{1}{\pi} \arctan\left(\frac{43.89 - \log L_X}{0.46}\right). \quad (57)$$

Table 1. Summary of Parameters

Symbol	Description	Equation	Parameters	Section
$V(z)$	Characteristic v_{Mpeak} in SFR– v_{Mpeak} relation	4	4	2.2
$\epsilon(z)$	Characteristic SFR in SFR– v_{Mpeak} relation	5	4	2.2
$\alpha(z)$	Low-mass slope of the SFR– v_{Mpeak} relation	6	4	2.2
$\beta(z)$	Massive-end slope of the SFR– v_{Mpeak} relation	7	3	2.2
$v_Q(z)$	Typical v_{Mpeak} for star formation quenching, in dex	10	3	2.2
$w_Q(z)$	Typical width in M_{peak} for star formation quenching, in dex	11	3	2.2
f_{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 the observed stellar masses, in dex	23	2	2.3
$\kappa(z)$	Additional systematic offset in observed vs. true SFRs, in dex	25	1	2.3
$\sigma(z)$	Scatter between measured and true stellar masses, in dex	26	1	2.3
$\rho_{0.5}$	Correlation between SFR and stellar mass at fixed halo mass at $z = 1$ ($a = 0.5$)	27	1	2.3
$\beta_{\text{BH}}(z)$	Median SMBH mass for galaxies with $M_{\text{bulge}} = 10^{11} M_{\odot}$, in dex	29	3	2.4
$\gamma_{\text{BH}}(z)$	Slope of the SMBH mass–bulge mass (M_{\bullet} – M_{bulge}) relation	30	3	2.4
σ_{BH}	Scatter in SMBH mass at fixed bulge mass, in dex	-	1	2.4
$f_{\text{scale}}(z)$	Ratio between the fractions of SMBH and galaxy growth coming from mergers	36	2	2.5
$f_{\text{duty}}(M_{\bullet}, z)$	AGN duty cycle	38	4	2.6
$c_1(z), c_2(z)$	Faint- and bright-end slopes of the AGN Eddington ratio distribution functions	43,44	4	2.6
ϵ_{tot}	Total energy efficiency (radiative and kinetic) of mass accretion onto SMBHs	48	2	2.6
ξ	Systematic offset in Eddington ratio when calculating AGN probability distribution functions, in dex	66	1	2.8
Total Number of Galaxy Parameters				28
Total Number of SMBH Parameters				20
Total Number of Parameters				48

Notes. v_{Mpeak} : the maximum circular velocity at the time when the halo reaches its peak mass (See §2.2).

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	24	$G(0, 0.14)$
μ_a	Redshift scaling of μ , in dex	24	$G(0, 0.24)$
κ	Additional systematic offset in observed vs. true SFRs, in dex	25	$G(0, 0.24)$
σ_z	Redshift scaling of σ , in dex	26	$G(0.05, 0.015)$
$\rho_{0.5}$	Correlation between SFR and stellar mass at fixed halo mass at $z = 1$ ($a = 0.5$)	27	$U(0.23, 1)$
$\beta_{\text{BH},0}$	SMBH mass at $M_{\text{bulge}} = 10^{11} M_{\odot}$ and $z = 0$	29	$G(8.46, 0.20)$
$\gamma_{\text{BH},0}$	Slope of the M_{\bullet} – M_{bulge} relation at $z = 0$	30	$G(1.05, 0.14)$

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 .

Thus, the type I quasar BHMF is:

$$\phi_{\text{ABH}'}(M_{\bullet}, z) = \int_0^{\infty} \int_0^{\infty} \phi_h(M_{\text{peak}}, z) P(M_{\bullet}|M_{\text{peak}}, z) \times P(\eta_{\text{rad}}|M_{\bullet}, M_{\text{peak}}, z) \times (1 - F_{\text{obs}}(L_X)) d\eta_{\text{rad}} dM_{\text{peak}}, \quad (58)$$

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

$$L_X = \frac{L_{\text{bol}}}{k_{\text{bol}}(L_{\text{bol}})} \quad (59)$$

$$L_{\text{bol}}/\text{erg} \cdot \text{s}^{-1} = 10^{38.1} \cdot M_{\bullet} \cdot \eta_{\text{rad}} \quad (60)$$

$$k_{\text{bol}}(L_{\text{bol}}) = 10.83 \left(\frac{L_{\text{bol}}}{10^{10} L_{\odot}} \right)^{0.28} + 6.08 \left(\frac{L_{\text{bol}}}{10^{10} L_{\odot}} \right)^{-0.020} \quad (61)$$

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

$$\phi_L(L_{\text{bol}}, z) = \int_0^{\infty} \phi_h(M_{\text{peak}}) P(L_{\text{bol}}|M_{\text{peak}}, z) dM_{\text{peak}}, \quad (62)$$

where $P(L_{\text{bol}}|M_{\text{peak}}, z)$ is calculated by counting the number den-

sity of SMBHs with the corresponding Eddington ratio:

$$P(L_{\text{bol}}|M_{\text{peak}}, z) = \int_0^{\infty} P(\eta_{\text{rad}}(L_{\text{bol}}, M_{\text{BH}})|M_{\bullet}, M_{\text{peak}}, z) \times P(M_{\bullet}|M_{\text{peak}}, z) dM_{\text{BH}}. \quad (63)$$

Finally, for quasar probability distribution functions, Aird et al. (2018) expressed Compton-thin QPDFs in terms of the specific L_X (sL_X):

$$sL_X = \frac{L_X/\text{erg} \cdot \text{s}^{-1}}{1.04 \times 10^{34} \times M_{\ast}/M_{\odot}}. \quad (64)$$

The distribution of sL_X at fixed stellar mass and redshift is:

$$P(sL_X|M_{\ast}, z) = (1 - f_{\text{CTK}}(L_X, z)) \times P(L'_{\text{bol}}|M_{\ast}, z) \quad (65)$$

$$L'_{\text{bol}} = \frac{L_{\text{bol}}}{\xi} \quad (66)$$

$$L_{\text{bol}}/\text{erg} \cdot \text{s}^{-1} = 1.04 \times 10^{34} \times M_{\ast}/M_{\odot} \times sL_X \times k_{\text{bol}}(L_{\text{bol}}) \quad (67)$$

$$P(L_{\text{bol}}|M_{\ast}, z) = \int_0^{\infty} dM_{\text{peak}} \int_0^{\infty} dM_{\bullet} P(\eta_{\text{rad}}(L_{\text{bol}}, M_{\bullet})|M_{\text{peak}}, z) P(M_{\bullet}|M_{\ast}, z) P(M_{\bullet}|M_{\text{peak}}, z), \quad (68)$$

where the Compton-thick fraction $f_{\text{CTK}}(L_X, z)$ and the bolometric correction $k_{\text{bol}}(L_{\text{bol}})$ are both given by Ueda et al. (2014) (see Ap-

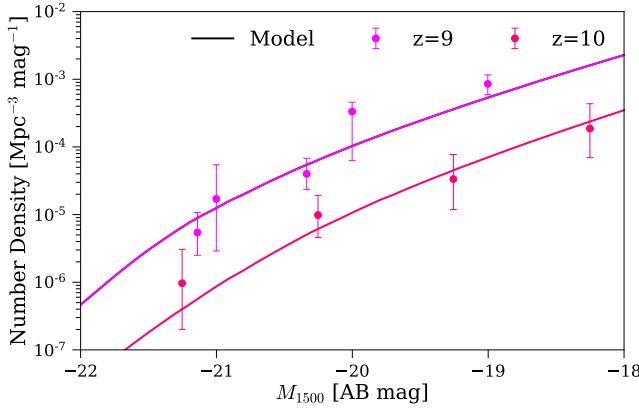


Figure 5. Comparison between observed galaxy UV luminosity functions (UVLFs) and our best-fitting model from $z = 9 - 10$. The references for observations are listed in Table 8. All the data used to make this plot (including individual data points and our best-fitting model) can be found [here](#).

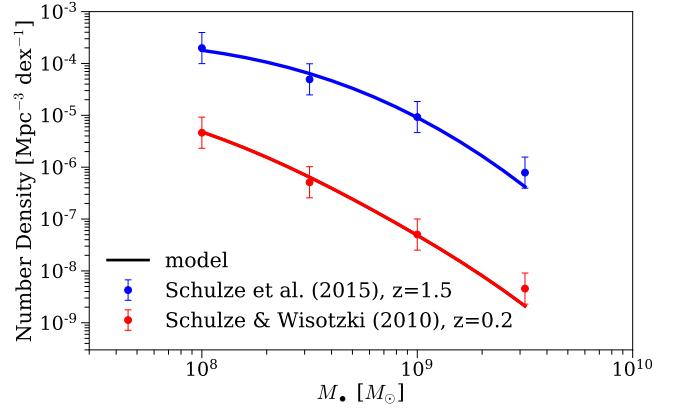


Figure 7. Comparison between the observed active black hole mass functions (ABHMFs) from [Schulze & Wisotzki \(2010\)](#), [Schulze et al. \(2015\)](#), and our best-fitting model at $z = 0.2$ and $z = 1.5$. All the data used to make this plot (including individual data points and our best-fitting model)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

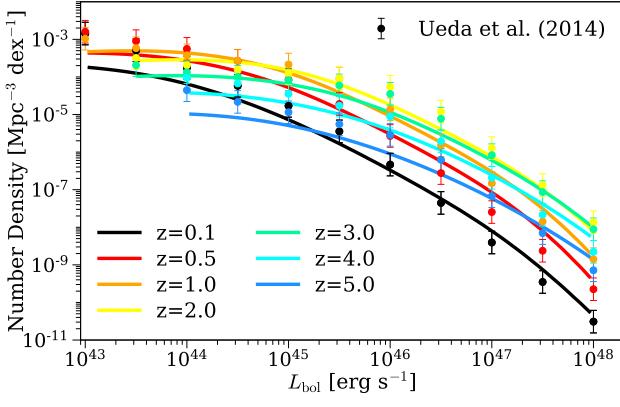


Figure 6. Comparison between the observed quasar luminosity functions (QLFs) from [Ueda et al. \(2014\)](#) and our best-fitting model from $z = 0 - 5$. All the data used to make this plot (including individual data points and our best-fitting model) can be found [here](#).

pendix E1 for full details about f_{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 downampling and exclusion as described in Appendix E2.

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 (§2.2, §2.4), SMBH merger contributions (§2.5), and the BHAR-AGN luminosity conversion (§2.6, 2.7).
2. Put galaxies and SMBHs into haloes accordingly, which determines galaxy and SMBH growth histories.

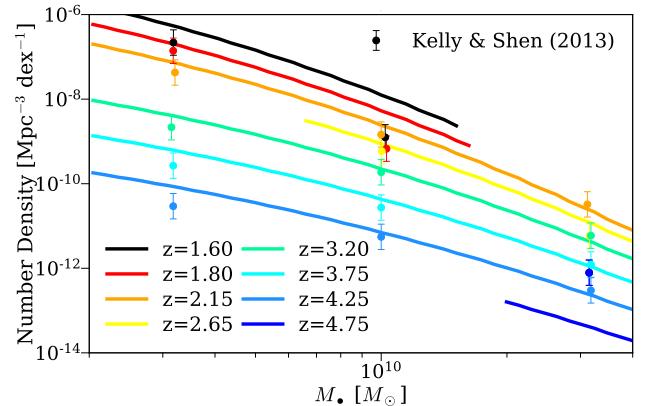


Figure 8. Comparison between the observed active black hole mass functions (ABHMFs) from [Kelly & Shen \(2013\)](#) and our best-fitting model from $z = 1.5 - 5$. The data points and the best fitting models in each higher redshift bin are shifted downwards by 0.5 dex incrementally for the sake of clarity.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

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

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

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

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

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

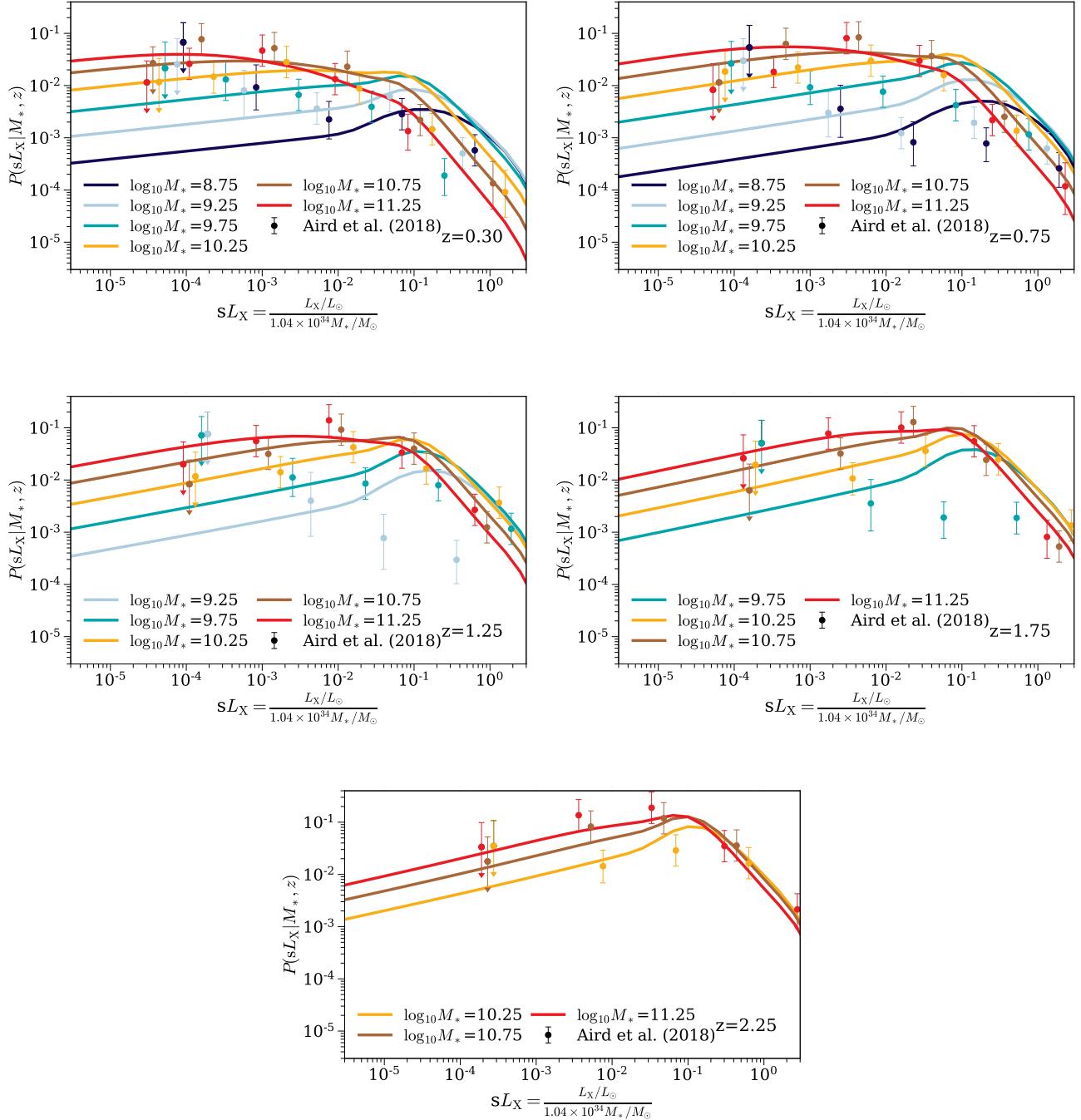


Figure 9. The comparison between the observed quasar probability distribution functions (QPDFs) from Aird et al. (2018) and our best-fitting model from $z = 0 – 2.5$. The data points include Compton-thin AGNs only, so the model values are corrected for direct comparison. All the data used to make this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

Metropolis MCMC method (Haario et al. 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 §2.1, TRINITY requires only halo population statistics from dark matter simulations, as opposed to individual halo merger trees. We use the peak historical mass (M_{peak}) halo mass functions from Behroozi et al. (2013) for the cosmology specified in the introduction. These mass functions are based on central

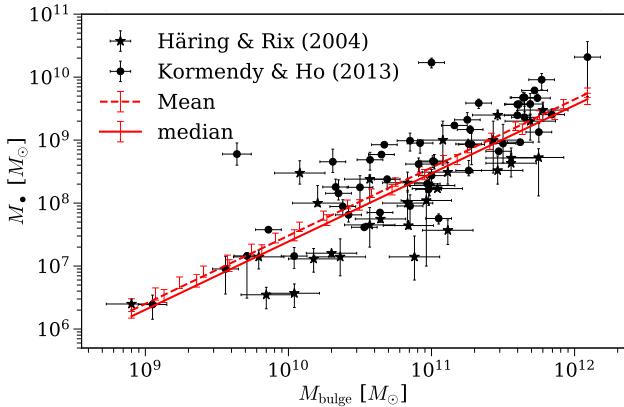


Figure 10. The local M_{\bullet} – M_{bulge} relation. The filled circles are the data compiled by Kormendy & Ho (2013), and the stars are those compiled by Häring & Rix (2004). The red solid line is the median SMBH–bulge mass relation, and the red dashed line is the mean relation. These lines are offset because log-normal 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](#).

halo mass functions from Tinker et al. (2008), with adjustments to include satellite halo number densities as well as to use M_{peak} instead of the present day mass. These adjustments were based on the Bolshoi & Consuelo simulations (Klypin et al. 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 halos 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 UNIVERSE MACHINE (Behroozi et al. 2019), with full details and formulae in Appendix B.

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: stellar mass functions (SMFs, Table 4), quenched fractions (QFs, Table 5), cosmic star formation rates (CSFRs, Table 6), specific star formation rates (SSFRs, Table 7), and UV luminosity functions (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 datasets. For full details, we refer readers to Appendix C of Behroozi et al. (2019).

Stellar mass functions 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 quenched fractions 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 quenched fractions in a different way, but we assume that they all refer to galaxies with negligible global star formation rates (see §2.2). Although this results in some uncertainty in the interpretation of galaxy quenched fractions, it does not affect the main analysis, which only depends on the average star formation rate 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, H α , 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 UV luminosity functions 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 factor 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 stellar population synthesis. There is some observational evidence that the IMF becomes more top-heavy with increasing SFR (e.g., Gunnawardhana et al. 2011), but it remains an open issue whether IMF varies with environment or redshift (Conroy et al. 2009; Bastian et al. 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 Supermassive black hole data

There are five different kinds of SMBH observables in our compiled dataset: quasar luminosity functions (QLFs), quasar probability distribution functions (QPDFs), active black hole mass functions (ABHMFs), the local SMBH mass–bulge mass (M_{\bullet} – M_{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_{bulge}).

We have used bolometric quasar luminosity functions (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., UV luminosity functions from Kulkarni et al. 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 E1 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.

QLFs constrain the total radiative energy output of active

Table 3. Summary of Observational Constraints

Type	Redshifts	Primarily Constrains	References
Stellar mass functions	0-8	SFR– v_{Mpeak} relation	Table 4
Galaxy quenched fractions	0-4	Quenching– v_{Mpeak} relation	Table 5
Cosmic star formation rates	0-10	SFR– v_{Mpeak} relation	Table 6
Specific star formation rates	0-9	SFR– v_{Mpeak} relation	Table 7
Galaxy UV luminosity functions	9-10	SFR– v_{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_{Mpeak} is the maximum circular velocity of the halo at the time when it reaches its peak mass, M_{peak} . This is used as a proxy for the halo mass in TRINITY. BHAR is the SMBH accretion rate.

Table 4. Observational Constraints on Galaxy Stellar Mass Functions

Publication	Redshifts	Wavebands	Area (deg 2)
Baldry et al. (2012)	0.002-0.06	<i>ugriz</i>	143
Moustakas et al. (2013)	0.05-1	UV-MIR	9
Tomczak et al. (2014)	0.2-3	UV-K $_{\text{S}}$	0.08
Ilbert et al. (2013)	0.2-4	UV-K $_{\text{S}}$	1.5
Muzzin et al. (2013)	0.2-4	UV-K $_{\text{S}}$	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	H α	144
Ly et al. (2011a)	0.8	H α	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	H α	0.03
Sobral et al. (2014)	0.4-2.3	H α	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 et al. (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 et al. (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 et al. (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 $\sim 3000 \text{ deg}^2$ (fully coded) and $\sim 10000 \text{ deg}^2$ (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	H α	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

SMBHs (Conroy & White 2013; Caplar et al. 2015). To constrain the mass-dependence of AGN luminosity distributions, we included quasar probability distribution functions (QPDFs) from Aird et al. (2018). These functions are expressed as the conditional probability distributions of $sL_X \equiv L_X / (1.04 \times 10^{34} \text{ erg s}^{-1} \times M_*)$. These distributions are given as functions of stellar mass (M_*) and redshift, and are obtained by modeling 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 modeling, so we adopted the empirical scheme given by Ueda et al. (2014) to correct our predicted QPDFs for this selection bias (see Appendix E1 for more details).

In modeling 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 active black hole mass functions (ABHMFs) from $z = 0.2 – 5$ from Schulze & Wisotzki (2010); Kelly & Shen (2013); Schulze et al. (2015), and the local $M_{\bullet} – M_{\text{bulge}}$ 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 $M_{\bullet} – M_{\text{bulge}}$ relation. Instead, we picked five commonly-used local $M_{\bullet} – M_{\text{bulge}}$ relations and calculated the medians and standard deviations of their slopes and intercepts (see Table 10). We then apply Gaus-

Table 9. Observational Constraints on AGNs

Publication	Type	Redshifts	Waveband	Area (deg ²)
Ueda et al. (2014)	Luminosity functions	0-5	X-ray	0.12-34000
Aird et al. (2018)	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	14000

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_{bulge}) relation at $z = 0$

Publication	β_{BH}	γ_{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_{bulge} relation is assumed to be a power-law: $\log_{10}(M_{\bullet}) = \beta_{\text{BH}} + \gamma_{\text{BH}} \log_{10}(M_{\text{bulge}}/10^{11} M_{\odot})$.

sian 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 $L_{\text{bol}} > 10^{47}$ erg/s 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_{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:

$$\begin{aligned} P(N_{\text{obs}} = 0 | N_{\text{exp}}) &= \exp(-N_{\text{exp}}) \\ N_{\text{exp}} &= \int_0^{10^8} P(M_{\bullet,\text{obs}} | M_{\bullet,\text{int}}) dM_{\bullet,\text{obs}} \\ &\times \int_0^{\infty} dM_{\bullet,\text{int}} \int_{10^{47}}^{\infty} dL_{\text{bol}} P(L_{\text{bol}} | M_{\bullet,\text{int}}) \Phi_{\text{BH}}(M_{\bullet,\text{int}}) \\ &\times S_{\text{SDSS}} \times \Delta z \\ P(\log M_{\bullet,\text{obs}} | M_{\bullet,\text{int}}) &= \frac{1}{\sqrt{2\pi}\sigma_{\text{BH,obs}}} \\ &\times \exp\left[-\frac{(\log M_{\bullet,\text{obs}} - \log M_{\bullet,\text{int}})^2}{2\sigma_{\text{BH,obs}}^2}\right], \end{aligned}$$

where $M_{\bullet,\text{int}}$ and $M_{\bullet,\text{obs}}$ are the intrinsic and observed SMBH masses, respectively, and $\sigma_{\text{BH,obs}} = 0.4$ dex is the random scatter in SMBH mass as induced by virial estimates (Park et al. 2012). $S_{\text{SDSS}} = 14000$ deg² 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 (Mazzucchelli et al. 2017; Trakhtenbrot et al. 2017).

In the process of compiling these data, we found systematic discrepancies between some observational datasets, which are addressed in Appendices E2 (quasar X-ray luminosities) and E3 (active black hole mass functions).

4 RESULTS

We present the best fitting parameters and the comparisons to observations in §4.1, as well as results for the evolution of the M_{\bullet} – M_{bulge} relation in §4.2, black hole accretion rates and Eddington ratio distributions in §4.3, the SMBH mass function in §4.4, SMBH mergers in §4.5, and AGN energy efficiency as well as systematic uncertainties in §4.6.

4.1 Best fitting parameters and comparison to observables

We obtained the posterior distribution of model parameters with an MCMC algorithm (§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 (§3), including stellar mass functions (SMFs, Fig. 3, left panel), quenched fractions (QFs, Fig. 3, right panel), cosmic star formation rates (CSFRs, Fig. 4, left panel), specific star formation rates (SSFRs, Fig. 4, right panel), galaxy UV luminosity functions (UVLFs, Fig. 5), quasar luminosity functions (QLFs, Fig. 6), active black hole mass functions (ABHMFs, Figs. 7 and 8), quasar probability distribution functions (QPDFs, Fig. 9) and the local M_{\bullet} – M_{bulge} relation (Fig. 10). For 1192 data points and 48 parameters, the naive reduced χ^2 is 0.86, which suggests a reasonable fit. The best fitting model and 68% confidence intervals for parameters are presented in Appendix G.

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. Given the complexity of the models adopted by Aird et al. (2018) to calculate these QPDFs, we did not add additional free parameters to fully reproduce their shapes, which reduces the risk of over-fitting.

4.2 The M_{\bullet} – M_{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_{bulge}) relation (top panel) along with the log-normal scatter (bottom panel) from $z = 0$ –10. We find that the median SMBH mass shows only mild evolution at $M_{\text{bulge}} \gtrsim 10^{10} M_{\odot}$, but increases strongly with time below this mass. From $z = 0$ –3, the evolution in the median M_{\bullet} at fixed M_{bulge} is at most ~ 0.3 dex, which is within the typical SMBH mass uncertainties. The median M_{\bullet} – M_{bulge} relation beyond $z = 0$ is jointly constrained by the quasar luminosity functions (QLFs), quasar proba-

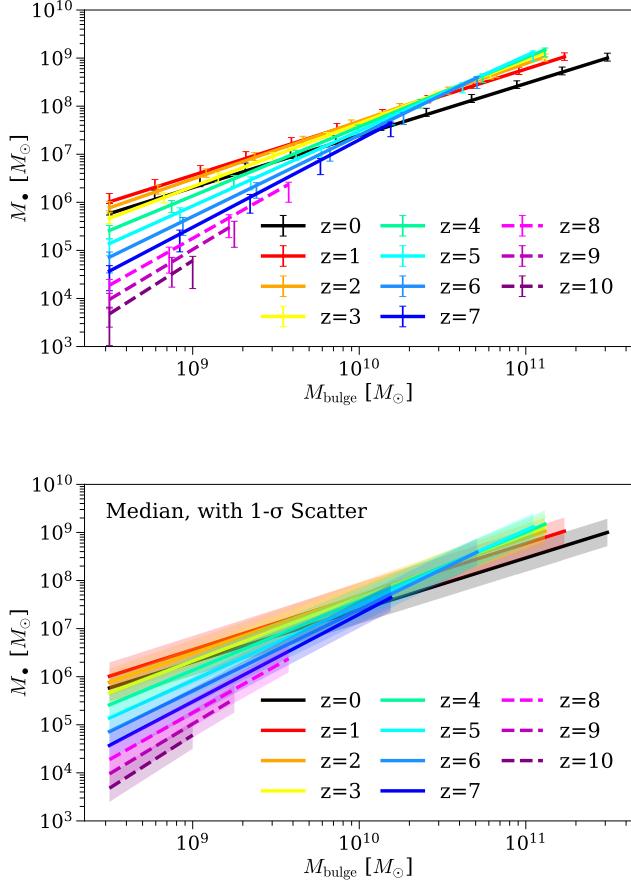


Figure 11. The evolution of the median M_\bullet – M_{bulge} relation and the corresponding log-normal scatter from $z = 0 – 10$. **Top Panel:** the median relations (see §4.2). The error bars show the 68% confidence intervals inferred from the model posterior distribution. **Bottom Panel:** The same median relations, except that the shaded regions show the *log-normal 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). Note that at $z \gtrsim 9$, the median SMBH masses in most galaxies seem to lie below the conventional lower limit for *supermassive* black holes, i.e., $10^5 M_\odot$. This may imply a breakdown of the assumption that every galaxy hosts a central SMBH (i.e., $f_{\text{occ}} \equiv 1$) at higher redshifts, which would lower the median and average M_\bullet for all galaxies. All the data used to make this plot can be found [here](#).

bility distribution functions (QPDFs), active black hole mass functions (ABHMFs), and the galaxy stellar mass functions (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_{bulge} relation. At $z \geq 8$ (shown in Fig. 11 as dashed lines), there are currently no SMBH data at all. We expect that future observations (by, e.g., the *James Webb Telescope* [JWST]) will test these predictions for the early Universe.

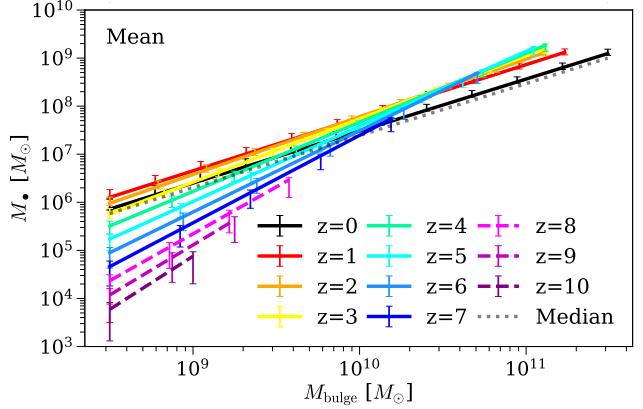


Figure 12. The evolution of the mean M_\bullet – M_{bulge} relation from $z = 0 – 10$ (see §4.2). The grey dotted line shows the median relation at $z = 0$ for comparison. The error bars show the 68% confidence intervals inferred from the model posterior distribution. The scaling relations at $z \geq 8$ are shown in dashed lines, which remain to be verified by future observations (by, e.g., JWST). At $z \gtrsim 9$, many galaxies have average $M_\bullet < 10^5 M_\odot$, i.e., below the conventional mass limit for SMBHs. See the caption of Fig. 11 and §4.2 for details. All the data used to make this plot can be found [here](#).

In Fig. 11, it seems that many galaxies at $z \gtrsim 9$ have median $M_\bullet < 10^5 M_\odot$, which is the conventional lower limit for *supermassive* black holes. We note that these small black hole masses are obtained assuming that every single galaxy hosts a central SMBH, i.e., the SMBH occupation fraction is $f_{\text{occ}} \equiv 1$. When $f_{\text{occ}} < 1$, the median M_\bullet for *all* galaxies will be smaller than that of the galaxies *that actually host central SMBHs*. Therefore, these small M_\bullet could be a sign that $f_{\text{occ}} < 1$ for high redshift and/or low mass galaxies. Based on the same argument, the strong evolution of the M_\bullet – M_{bulge} relation at $z \gtrsim 3$ may also be (at least partly) driven by the decrease of f_{occ} towards higher redshifts. Despite the low average/median black hole masses at high redshifts, TRINITY does predict a sufficient number of SMBHs with $M_\bullet > 10^9$ and $10^{10} M_\odot$ at $z \gtrsim 7$ to match observed quasar samples, which is discussed in paper III.

The scatter around the median M_\bullet – M_{bulge} relation is $\sigma_{\text{BH}} \approx 0.29$ dex. As described in §2.5, a log-normal scatter of σ_{BH} causes an offset between the *median* and *mean* SMBH masses (§2.5) at fixed stellar mass. Mean SMBH masses directly influence average BHARs, which are constrained by observed QLFs and QPDFs. Consequently, σ_{BH} is primarily constrained by (a) the evolution of the median M_\bullet – M_{bulge} relation; and (b) the average BHARs inferred from QLFs and QPDFs.

In Fig. 12, we show the evolution of the *mean* M_\bullet – M_{bulge} relation from $z = 0 – 10$. With $\sigma_{\text{BH}} \approx 0.29$ dex, the mean relation is offset from the median relation by a constant factor of $0.5\sigma_{\text{BH}}^2 \ln 10 \approx 0.10$ 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 disk observations. This relation is qualitatively similar to the M_\bullet – M_{bulge} relation, mainly because of the approximate proportionality between M_{bulge} and M_* (Eq. 15). Quantitatively, the evolution of the M_\bullet – M_* relation between $0 < z < 2$ is less significant than that of the M_\bullet – M_{bulge} relation, due to lower M_{bulge}/M_* ratios at higher redshifts, which is also consistent with observational studies like Ding et al. (2020).

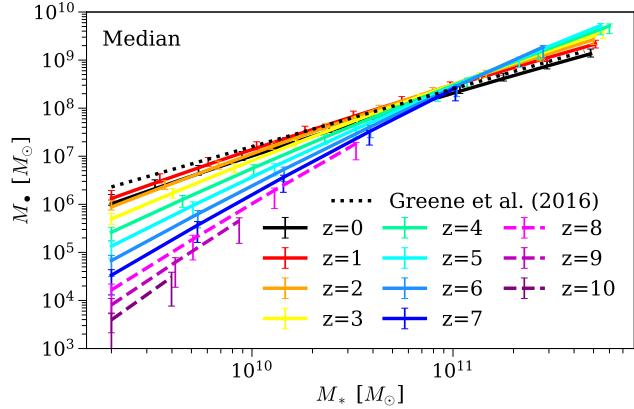


Figure 13. Top Panel: the best-fitting median M_\bullet - M_* relation from $z = 0 - 10$ (solid lines, see §4.2), and the observed $z = 0$ M_\bullet - M_* relation from Greene et al. (2016) (dotted line). **Bottom Panel:** the best-fitting median M_\bullet/M_* ratios as a function of M_* and z . The error bars show the 68% 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). At $z \gtrsim 9$, many galaxies have average $M_\bullet < 10^5 M_\odot$, i.e., below the conventional mass limit for SMBHs. See the caption of Fig. 11 and §4.2 for details. All the data used to make this plot can be found [here](#).

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 -bulge and M_\bullet - M_* relations between $0 < z < 2$ (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_{peak}) relation. At $z \lesssim 5$, the M_\bullet - M_{peak} relation can be approximated as a double power-law, connected by a knee at $M_{\text{peak}} \sim 10^{12} M_\odot$. 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 stellar mass–halo mass (M_* - M_{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 (\tilde{M}_\bullet) as a function of M_{peak} and z . From $z = 0 - 10$, SMBH masses in haloes with $M_{\text{peak}} \sim 10^{11} M_\odot$ 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).

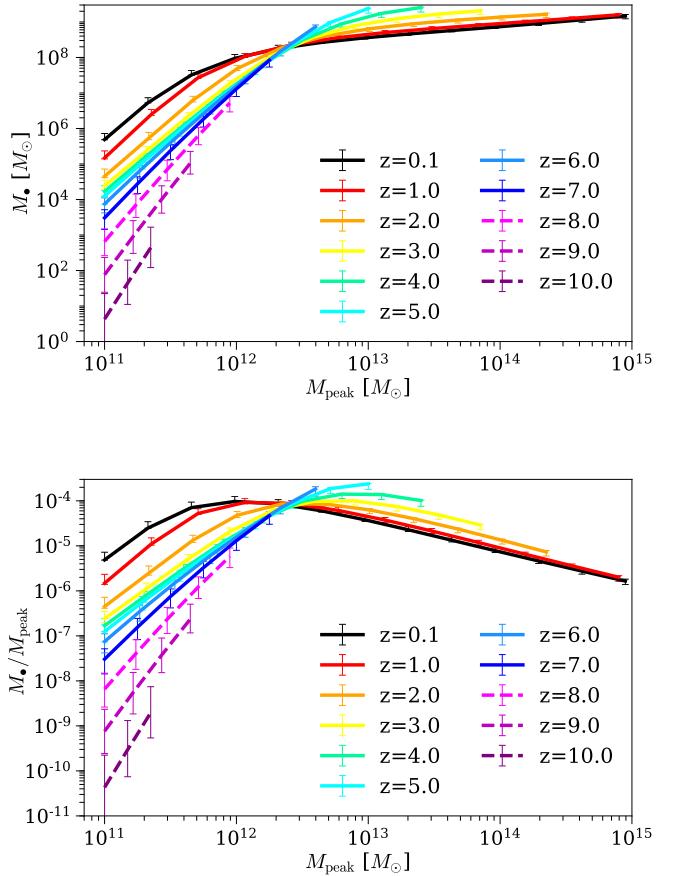


Figure 14. Top Panel: the best-fitting median M_\bullet - M_{peak} (peak halo mass) relation from $z = 0 - 10$ (see §4.2). **Bottom Panel:** the best-fitting $M_\bullet/M_{\text{peak}}$ ratios as a function of M_{peak} and z . The error bars show the 68% confidence intervals inferred from the model posterior 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). At $z \gtrsim 9$, many haloes have average $M_\bullet < 10^5 M_\odot$, i.e., below the conventional mass limit for SMBHs. See the caption of Fig. 11 and §4.2 for details. All the data used to make this plot can be found [here](#).

The bottom panel of Fig. 15 shows the \tilde{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” (Merloni 2004; Barger et al. 2005), and we discuss this further in §4.3 and §5.3.

4.3 Average black hole accretion rates and Eddington ratio distributions

The top panel of Fig. 16 shows the average black hole accretion rate (BHAR) as a function of M_{peak} and z . In general, BHARs peak at $M_{\text{peak}} \sim 10^{12} M_\odot$, and decrease towards lower and higher masses. Below $z \sim 2$ and $M_{\text{peak}} \sim 10^{13.5} M_\odot$, 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

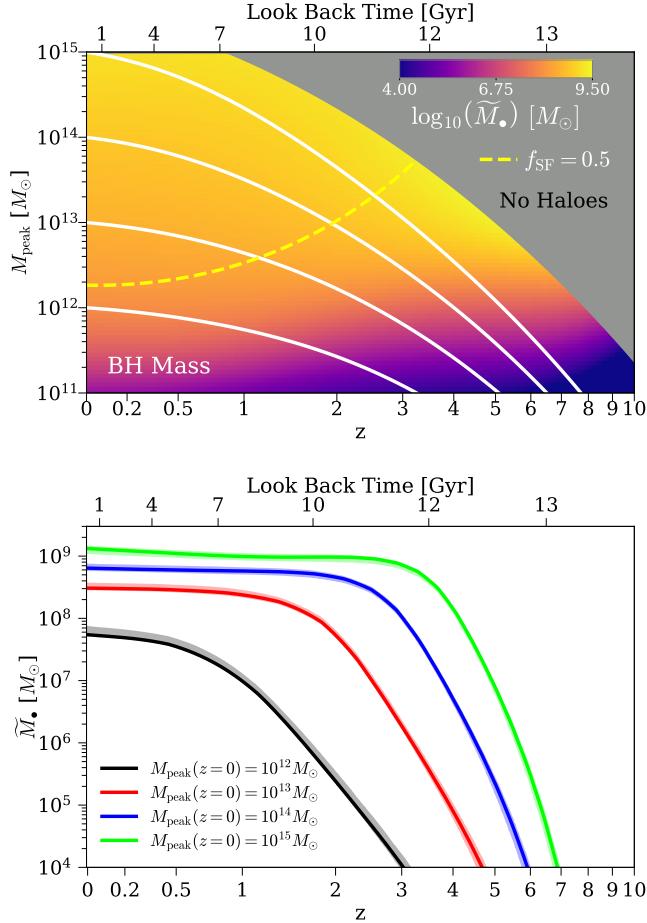


Figure 15. Top Panel: the median SMBH mass (\tilde{M}_*) as a function of M_{peak} and z (see §4.2). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” **Bottom Panel:** The \tilde{M}_* histories as a function of halo mass at $z = 0$. The shaded regions show the 68% confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found [here](#).

of SMBHs occurs primarily in star-forming (quenched) galaxies, respectively. 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 higher redshifts. There is also a slight increase in BHAR with time below $z \sim 1$ 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 star formation rates (SFRs) as a function of M_{peak} and z . The M_{peak} and z dependencies of SFR are similar to those of BHAR below $M_{\text{peak}} \sim 10^{14} M_\odot$. Above $M_{\text{peak}} \sim 10^{14} M_\odot$, however, SFR decreases monotonically with halo mass at all redshifts, whereas the massive black holes still have

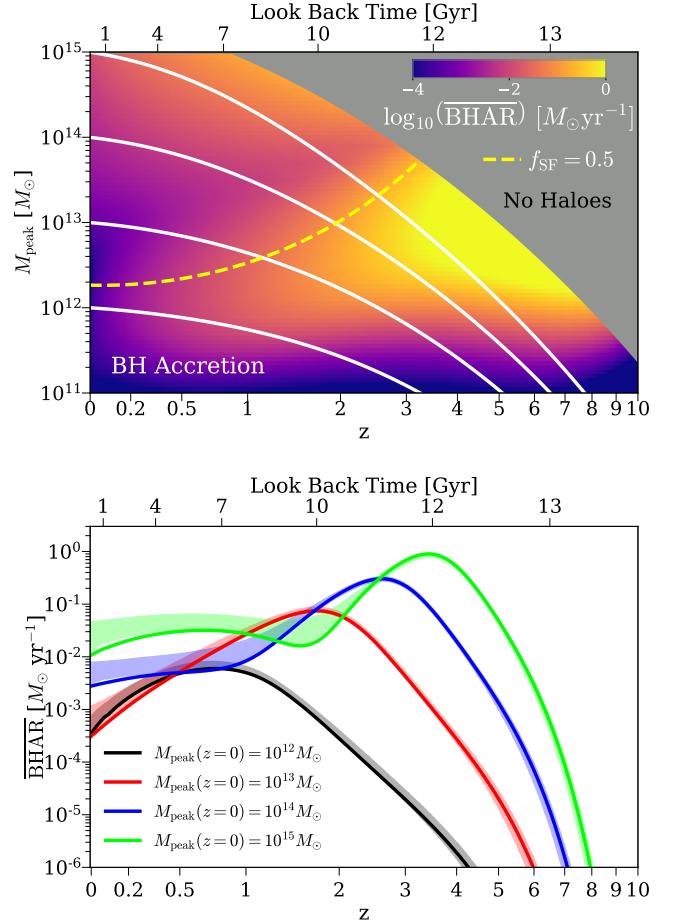


Figure 16. Top Panel: average black hole accretion rate ($\overline{\text{BHAR}}$) as a function of M_{peak} and z (see §4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” **Bottom Panel:** BHAR histories as a function of halo mass at $z = 0$. The shaded regions show the 68% confidence intervals inferred from the model posterior distribution.. All the data used to make this plot can be found [here](#).

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 ($M_{\text{peak}} \sim 10^{12} M_\odot$) dominate the demographics.

The top panel of Fig. 18 shows the ratios between the average BHAR and SFR, $\overline{\text{BHAR}}/\overline{\text{SFR}}$, as a function of M_{peak} and z . At $z \gtrsim 6$, $\overline{\text{BHAR}}/\overline{\text{SFR}}$ increases strongly with increasing M_{peak} and decreasing z . The latter effect results from the increasing normalization of the M_{\bullet} – M_{bulge} relation. Towards lower redshifts, $\overline{\text{BHAR}}/\overline{\text{SFR}}$ grows more slowly for all haloes, and shows a plateau at $\overline{\text{BHAR}}/\overline{\text{SFR}} \sim 10^{-3}$. 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

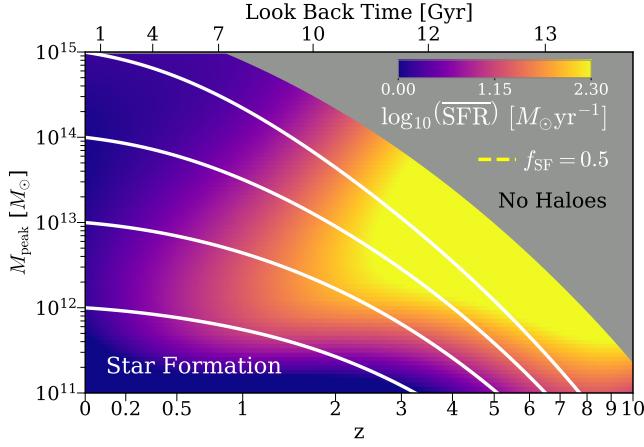


Figure 17. The average star formation rates ($\overline{\text{SFR}}$) as a function of M_{peak} and z (see §4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” All the data used to make this plot can be found [here](#).

higher $\overline{\text{BHAR}}/\overline{\text{SFR}}$. 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 $\overline{\text{BHAR}}/\overline{\text{SFR}}$ histories of different halo populations. Below $z \sim 2$, $\overline{\text{BHAR}}/\overline{\text{SFR}}$ either stays flat 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 ($\bar{\eta}$) as a function of M_{peak} and z . At $z \gtrsim 7$, the average SMBH accretion rate is around the Eddington rate regardless of host 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 haloes with $M_{\text{peak}} > 10^{12} M_{\odot}$, the average Eddington ratios experience a two-phase decline before the final flattening: an initial, slower decrease, and a later, faster drop. Haloes with $M_{\text{peak}} = 10^{12} - 10^{13} M_{\odot}$ 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, $\bar{\eta}$, 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 $\bar{\eta}$ decreases towards higher M_{\bullet} .

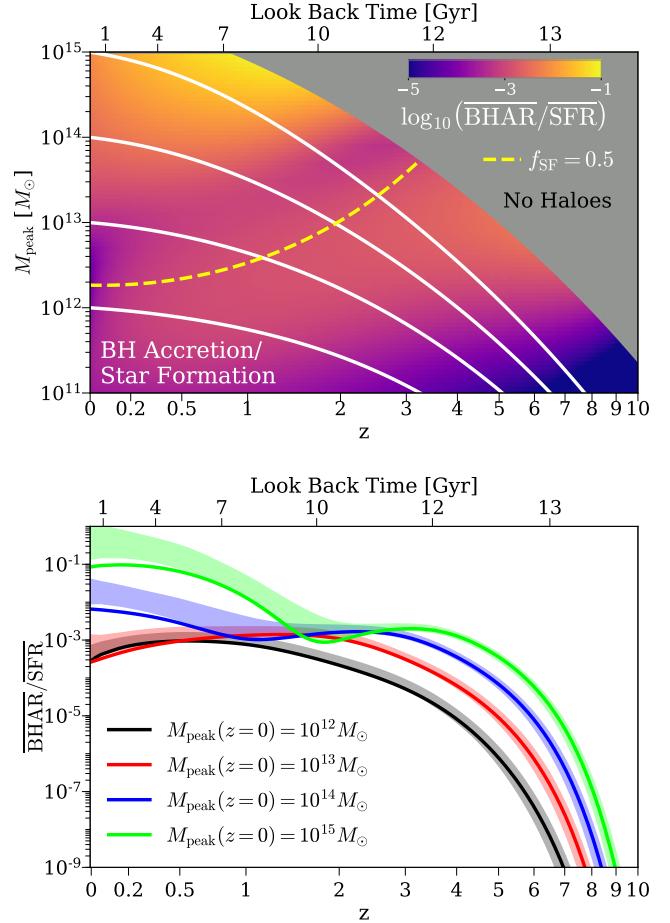


Figure 18. **Top** panel: the $\overline{\text{BHAR}}/\overline{\text{SFR}}$ ratio as a function of redshift and M_{peak} for our best fitting model (see §4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, and $10^{15} M_{\odot}$ at $z = 0$. **Bottom** panel: the $\overline{\text{BHAR}}/\overline{\text{SFR}}$ ratio histories as a function of M_{peak} at $z = 0$. The shaded regions show the 68% 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 labeled as “No Haloes.” All the data used to make this plot can be found [here](#).

4.4 SMBH mass functions

Fig. 21 shows the total black hole mass functions (BHMF) between $0 \leq z \leq 10$. Similar to the galaxy stellar mass functions, the “knee” in the black hole mass function becomes less and less significant towards higher redshifts. This is because, in the early Universe, the $M_{*}-M_{\text{peak}}$ relation, and therefore the $M_{\bullet}-M_{\text{peak}}$ relation, can be approximated as a single power-law. We also see strong evolution in the black hole mass function above $z \gtrsim 5$ *regardless of SMBH mass*. This directly results from universal accretion around the Eddington rate (see also §4.3). However, any potential change in the occupation fraction (f_{occ}) will affect both the amplitude and shape of the high redshift black hole mass functions. Specifically, the relative abundance of massive vs. less massive black holes will be higher if f_{occ} is lower. Therefore, further interpretation of black hole mass functions at $z \gtrsim 7$ is dependent on our knowledge of f_{occ} , which, unfortunately, is currently very limited.

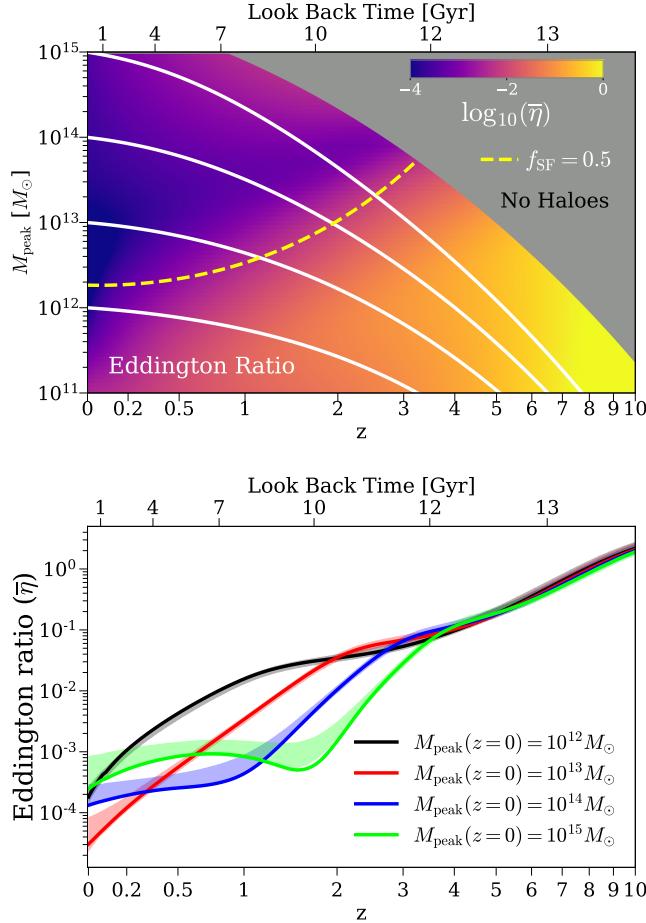


Figure 19. **Top Panel:** average SMBH *total* (i.e., radiative+kinetic) Eddington ratio ($\bar{\eta}$) as a function of M_{peak} and z (see §4.3). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” **Bottom Panel:** $\bar{\eta}$ histories as a function of halo mass at $z = 0$. The error bars show the 68% confidence intervals inferred from the model posterior distribution. All the data used to make this plot can be found [here](#).

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 Eq. 55, the BHMF contributed by haloes in the mass range ($M_{\text{peak,min}}, M_{\text{peak,max}}$) is:

$$\Phi(M_{\bullet}, M_{\text{peak,min}}, M_{\text{peak,max}}, z) = \int_{M_{\text{peak,min}}}^{M_{\text{peak,max}}} \Phi(M_{\text{peak}}, z) P(M_{\bullet}|M_{\text{peak}}, z) dM_{\text{peak}}, \quad (72)$$

where $\Phi(M_{\text{peak}}, z)$ is the halo mass function and $P(M_{\bullet}|M_{\text{peak}}, z)$ is the probability distribution of M_{\bullet} , given the host halo mass M_{peak} at redshift z . In TRINITY, $P(M_{\bullet}|M_{\text{peak}}, z)$ is a log-normal distribution with the median and scatter determined from the halo–galaxy–SMBH connection (§2.2 and §2.4). Given the flat M_{\bullet} – M_{peak} relation at the massive end (see Fig. 14), $P(M_{\bullet}|M_{\text{peak}}, z)$ only changes slightly with increasing halo mass. On the other

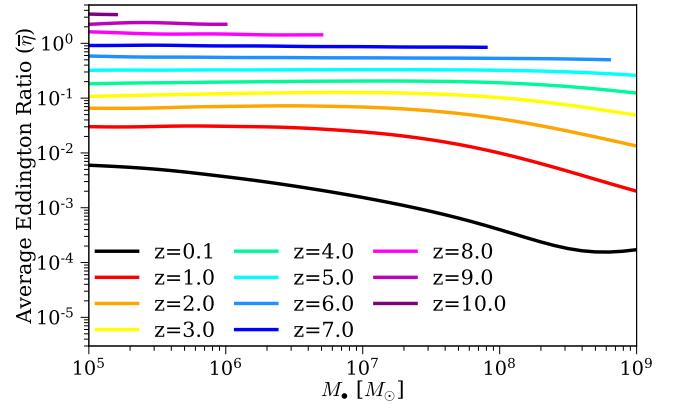


Figure 20. Average SMBH *total* (i.e., radiative+kinetic) Eddington ratio ($\bar{\eta}$) as a function of M_{\bullet} and z . See §4.3. All the data used to make this plot can be found [here](#).

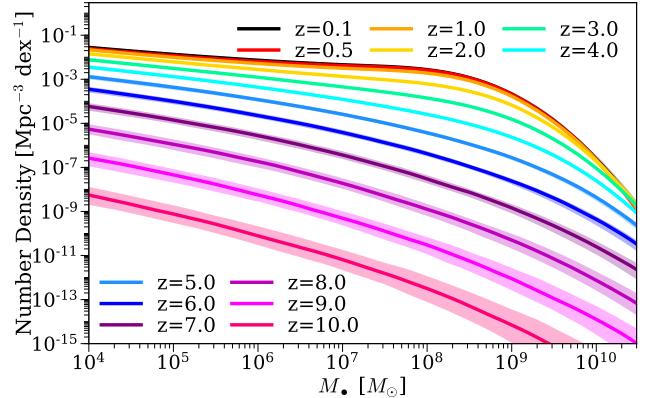


Figure 21. The total black hole mass function between $0 \leq z \leq 10$ (see §4.4). The shaded regions show the 68% confidence intervals inferred from the model posterior distribution.. All the data used to make this plot can be found [here](#).

hand, there are many fewer haloes with $M_{\text{peak}} > 10^{14} M_{\odot}$ than $M_{\text{peak}} < 10^{14} M_{\odot}$, due to the exponential decrease in halo number density. Hence, the haloes with $10^{13} M_{\odot} < M_{\text{peak}} < 10^{14} M_{\odot}$, rather than those with $10^{14} M_{\odot} < M_{\text{peak}} < 10^{15} M_{\odot}$, dominate the BHMF for $M_{\bullet} > 10^{9.5} M_{\odot}$ at $z = 0.5$ and 1.0 . In other words, when looking at a M_{\bullet} -selected sample with large M_{\bullet} , we are more likely to observe less massive haloes 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 $13 < \log_{10} M_{\text{peak}} < 14$ and $14 < \log_{10} M_{\text{peak}} < 15$. In short, cluster-scale haloes ($\log_{10} M_{\text{peak}} > 14$) 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.

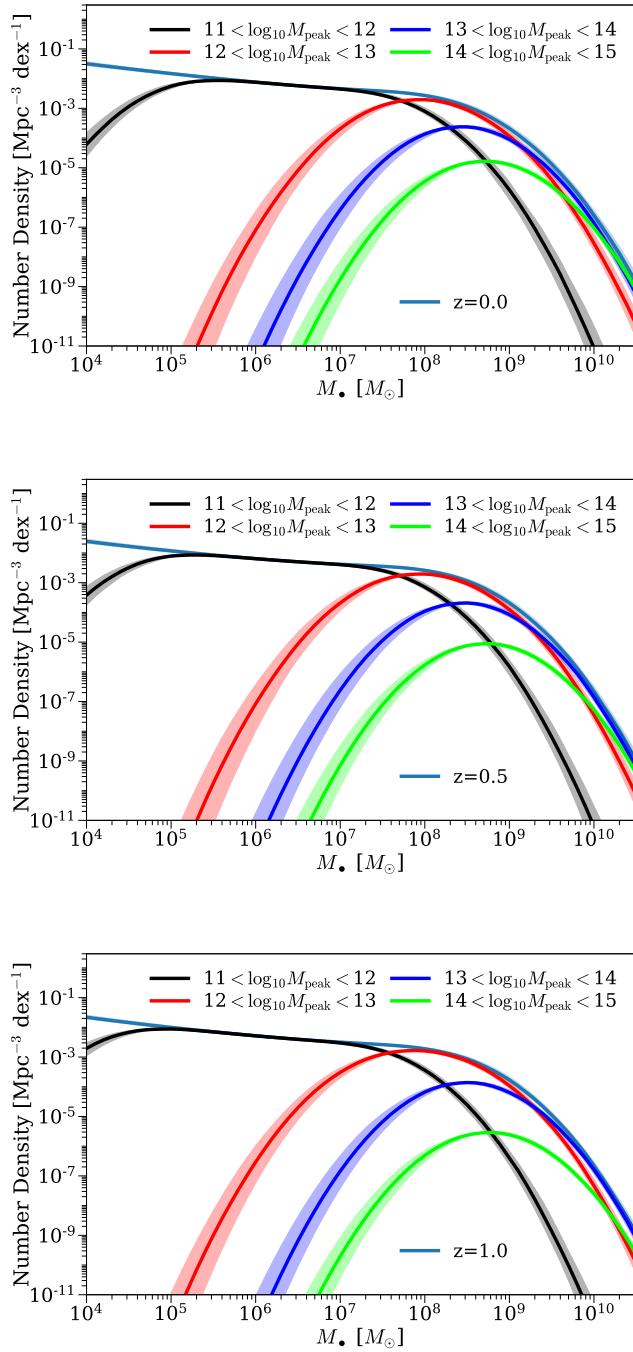


Figure 22. Total black hole mass functions at $z = 0.0, 0.5$, and 1.0 (the top, middle, and bottom panels), split into the contributions from different host dark matter halo mass bins (see §4.4.1). The shaded regions show the 68% confidence intervals inferred from the model posterior distribution.. All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

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 Pa-

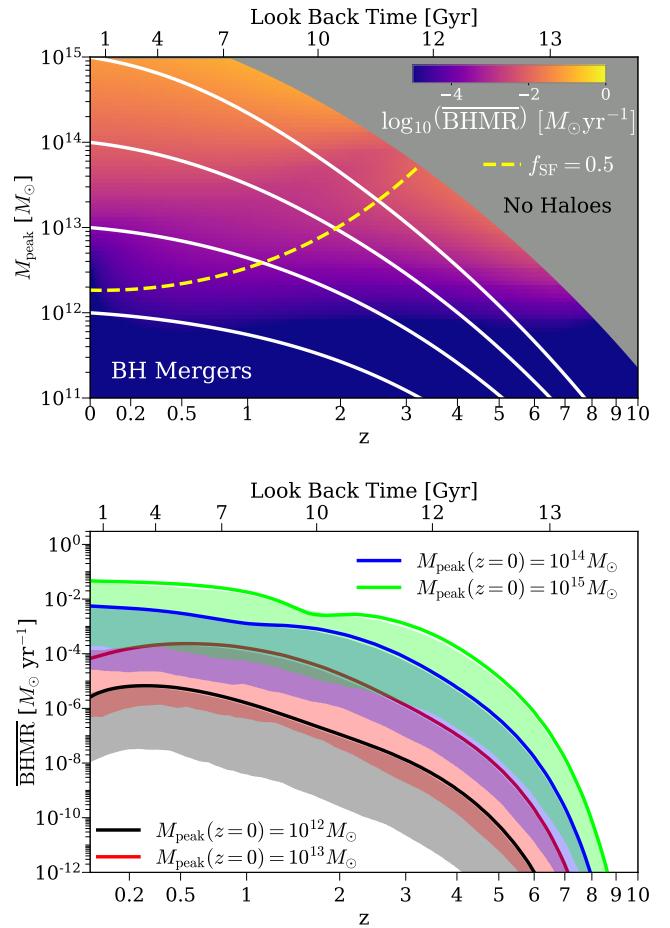


Figure 23. Top Panel: the average black hole merger rates ($\overline{\text{BHMR}}$) as a function of M_{peak} and z (see §4.5). The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, and $10^{15} M_{\odot}$ at $z = 0$. The grey area shows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

per 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 panel of Fig. 23. The best-fitting model lies on the upper edges of the 68% 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 §2.5, BHMRs are calculated by allowing a fraction of galaxy mergers (the free parameter f_{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 black hole mass functions; and b) average black hole accretion rates, which are constrained by the quasar luminosity functions and probability distribution functions. In Appendix A3, we also show the results of models with alternate assumptions about SMBH mergers.

Further discussion about SMBH mergers in TRINITY and pre-

dictions for gravitational wave experiments are presented in Paper V.

4.6 AGN energy efficiency and systematic uncertainties

As described in §2.3 and §2.8, we modeled systematic uncertainties in stellar mass, star formation rates, and SMBH Eddington ratios. These uncertainties are propagated into our model predictions, and their values quantify the degree of tension between different datasets. In TRINITY, the best fitting values (see Appendix G) 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 quasar luminosity functions (QLFs) from Ueda et al. (2014) and the quasar probability distribution functions (QPDFs) from Aird et al. (2018) (see Appendix E2). This discrepancy could be due to the fact that the two groups used different functional forms to fit the observational data. The net effect is $\eta' - \eta \sim 0.3$ dex, where η is the intrinsic Eddington ratio, and η' is the Eddington ratio to calculate the observed QPDFs.

The total AGN energy efficiency from TRINITY is $\log_{10} \epsilon_{\text{tot}} = -1.161^{+0.006}_{-0.231} + 0.025^{+0.040}_{-0.318}(a - 1)$. In other words, the best-fitting model is consistent with a redshift-independent $\sim 7\%$ mass-to-energy conversion efficiency; however, a factor-of-2 increase towards higher redshifts is allowed in the model posterior distribution. In this work, we opt not to allow a systematic offset in the normalization of the M_{\bullet} – M_{bulge} relation, β_{BH} , due to its complete degeneracy with the AGN energy efficiency. Thus, the best-fitting value of the energy efficiency ϵ_{tot} should be viewed as a *combination* of the intrinsic average efficiency and any potential systematic offset in β_{BH} .

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_{bulge} relation (§5.1); present the cosmic SMBH mass density as a function of redshift (§5.2); and discuss the physical implications of the best-fitting TRINITY model (§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_{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 et al. 2014, 2016), and EAGLE simulations (Schaye et al. 2015) predicted the opposite (Habouzit et al. 2020). 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 the M_{\bullet} – M_* relations at different redshifts from TRINITY and IllustrisTNG300 (Pillepich et al.

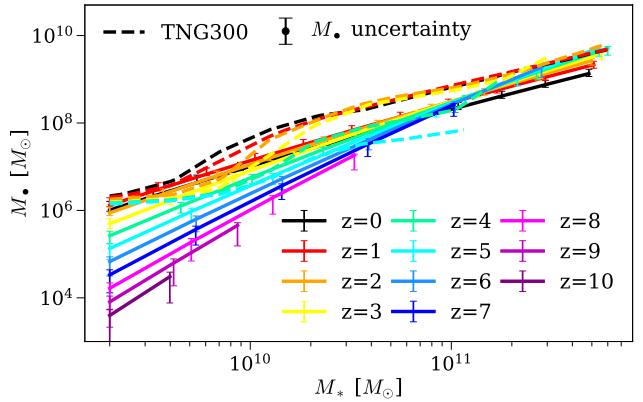


Figure 24. The median M_{\bullet} – M_* relations as functions of z for TRINITY (solid lines) and the IllustrisTNG300 simulation (dashed lines; Pillepich et al. 2018; Habouzit et al. 2020). See §5.1. The typical uncertainty in the measurement of M_{\bullet} , 0.3 dex, is shown by the black solid dot. All the data used to make this plot (including those from IllustrisTNG and our best-fitting model) can be found [here](#).

2018; Habouzit et al. 2020) in Fig. 24. Despite the offset, both mass scalings show increasing normalizations with time. This implies that SMBH growth becomes increasingly efficient compared to galaxy growth at lower redshifts. For the hydrodynamical simulations listed in Habouzit et al. (2020), 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 favors 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. 25 shows the cosmic SMBH mass density as a function of redshift from TRINITY compared to previous studies. The cosmic SMBH mass density from TRINITY includes both SMBHs in *galaxy centres* and *wandering* SMBHs (see §2.5), because the latter SMBHs also contributed to quasar luminosity functions during their growth. We also show the cosmic wandering SMBH density in green, which accounts for $\sim 15\%$ of the total SMBH mass density at $z = 0$. This is broadly consistent with the results from Volonteri et al. (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 modeled 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 differ-

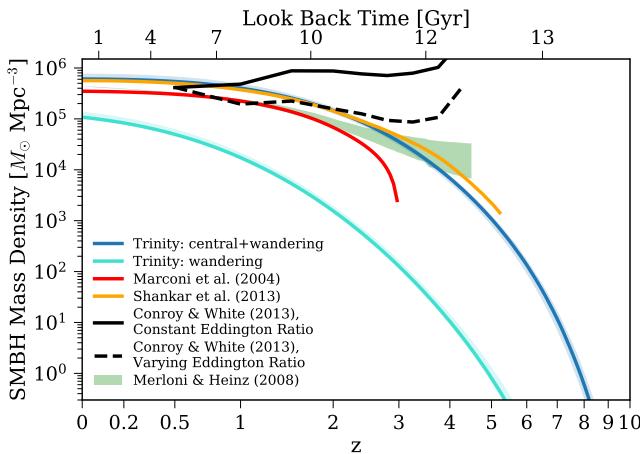


Figure 25. Cosmic SMBH mass density as a function of z (see §5.2). The shaded regions show the 68% confidence intervals inferred from the model posterior distribution. All the data used to make this plot (including those from previous studies and our best-fitting model) can be found [here](#).

ent SMBH mass functions at $z \sim 3$ from TRINITY, in which SMBHs are modeled to start growing from $z = 15$.

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 quasar luminosity functions 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 stellar mass function 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 quasar luminosity functions. This is shown in Fig. 25, 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. 26 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 different SMBHs 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. model. 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. For SMBHs above $10^8 M_\bullet$, TRINITY predicts slightly more late growth and less early growth compared to both Marconi et al. (2004) and Shankar et al. (2013). The reason for the discrepancy at high redshift is two-fold. First, at high redshifts, TRINITY predicts that the majority of active SMBHs are accreting at around the Eddington rate (see Paper II), but in Marconi et al. (2004) and Shankar et al. (2013) super-Eddington accretion is not allowed. Thus, to reproduce high- z QLFs, TRINITY would require a lower abundance of massive black holes compared to Shankar et al. (2013), and thus massive black holes do not have to grow as

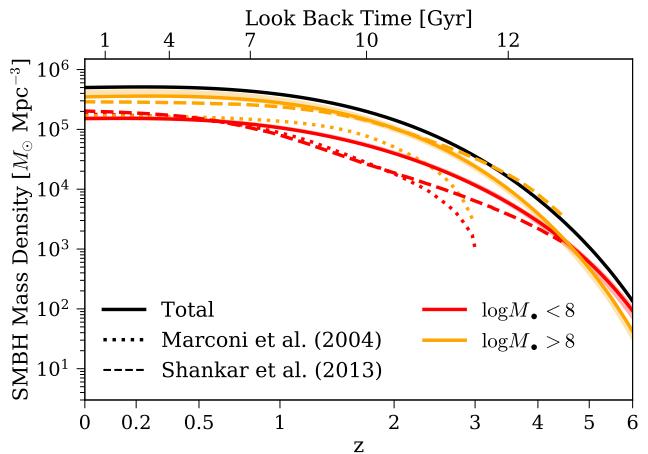


Figure 26. Cosmic SMBH mass densities split in different SMBH mass bins as functions of z , from TRINITY (solid lines), Marconi et al. (2004) (dotted lines), and Shankar et al. (2013) (dashed lines). See §5.2. All the data used to make this plot (including those from previous studies and our best-fitting model) can be found [here](#).

fast. At low redshifts, TRINITY finds that mergers contribute significantly to the growth of massive black holes, whereas Marconi et al. (2004) and Shankar et al. (2013)¹ did not account for the merger contributions in their published cosmic SMBH mass densities. Consequently, massive black holes in TRINITY grow slightly faster via mergers compared to those in Shankar et al. (2013).

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

In §4.3 we confirmed the ‘‘AGN downsizing’’ effect, in the sense that more massive black holes become less active earlier compared to smaller black holes. 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 quasar probability distribution functions from Aird et al. (2018). Physically, this is consistent with AGN feedback (Somerville et al. 2008; Croton et al. 2006). 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 BHAR/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 fueling 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.

¹ Shankar et al. (2013) did discuss the impact of mergers, but did not show the resulting SMBH mass densities.

Fig. 18 also shows that below $z \sim 6$, $\overline{\text{BHAR}}$ and $\overline{\text{SFR}}$ 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 onto 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 $\overline{\text{BHAR}}/\overline{\text{SFR}}$ for massive haloes.

6 CAVEATS AND FUTURE DIRECTIONS FOR EMPIRICAL MODELING 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 §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. Our best-fitting model predicts that only $\sim 8\%$ of the bright quasars are powered by low mass ($M_\bullet < 10^8 M_\odot$) black holes between $5.7 < z < 6.5$, but still predicts ~ 35 such objects in the observable Universe, in contrast to current observations. By checking the intrinsic and observed BHMFs of these 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 §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.

One potential reason for this discrepancy could be our model assumptions. In TRINITY, SMBHs with different M_\bullet at fixed M_{peak} are assumed to have the same Eddington ratio distribution. However, in the real Universe, it is unclear if less massive SMBHs (that are more likely to be downscattered below $M_\bullet = 10^8 M_\odot$) could be less active, and thus less likely to enter the observed sample. In other words, the statistical nature of TRINITY may not be able to fully capture the differential evolution of different SMBHs. This limitation will be alleviated by a future version of TRINITY that is coupled with the UNIVERSEMACHINE, where *individual objects* are traced across snapshots, and no fixed correlation between BHAR and SMBH mass needs to be assumed.

6.2 Potential impact of the assumption that every galaxy hosts a central SMBH

Throughout this work, we assume that every galaxy hosts a central SMBH. In other words, the SMBH occupation fraction is $f_{\text{occ}} \equiv 1$ for all galaxies across cosmic time. In this section, we discuss the potential impacts of this assumption on our results. In TRINITY, we calculate two kinds of statistics: (1) average SMBH properties of different halo or galaxy populations, like average SMBH accretion rates; and (2) the probability distributions of certain SMBH properties, like black hole mass functions. The average properties are calculated for *all* galaxies, regardless of whether all of them host central SMBHs. Consequently, the values of these average properties are not sensitive to changes in f_{occ} . However, the average SMBH properties of those galaxies that *host central*

SMBHs do depend on f_{occ} . Generally, if the average value of a certain SMBH property, X , is \bar{X} for *all* galaxies, then the average X for galaxies that *host central SMBHs* would be \bar{X}/f_{occ} . In addition, the probability distributions of SMBH mass-related properties (like M_\bullet or Eddington ratio) also depend on the occupation fraction. For example, the average M_\bullet for 10 galaxies would be the same for the following two scenarios: (1) they each host $10^5 M_\odot$ SMBHs; and (2) one of them hosts a $10^6 M_\odot$ SMBH, whereas the rest do not host any SMBHs. Clearly, the black hole mass distributions are different in these two cases. In general, smaller f_{occ} entails more massive black holes in fewer galaxies, which results in a less negative slope for black hole mass functions.

The change of average M_\bullet among SMBH host galaxies due to the change in f_{occ} also affects the Eddington ratio distributions from TRINITY. With $f_{\text{occ}} < 1$, the actual AGN duty cycle among SMBH host galaxies will be larger than that among *all* galaxies, which decreases the typical Eddington ratio and shifts the whole Eddington ratio distribution towards lower values (see Eqs. 42 and 46). In other words, if we assume $f_{\text{occ}} \equiv 1$ but in fact $f_{\text{occ}} < 1$, then Eddington ratios will be overestimated. We will present more detailed discussion on this in Paper II, where the Eddington ratio distributions at different redshifts are examined.

In principle, the change in f_{occ} value(s) may affect both the fitting process and the predictions from TRINITY. However, the black hole mass functions in the data compilation do not probe the mass range between $10^5 – 10^6 M_\odot$, where the occupation fraction is more uncertain. Thus, the fitting process is less likely to be affected significantly by the potential deviation of f_{occ} from unity. The lack of data constraints on f_{occ} in low mass regimes also prevents us from making alternate assumptions. That said, we note that our predictions of black hole mass functions between $10^5 – 10^6 M_\odot$ should be treated with caution, as they are effectively extrapolations to low mass regimes assuming $f_{\text{occ}} \equiv 1$.

6.3 Future directions

Currently, TRINITY makes only *statistical* halo–galaxy–SMBH connections. In the future, we plan to incorporate TRINITY into the UNIVERSEMACHINE by modeling 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 modeling of AGN duty cycle timescales; 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.

The uncertainty in the SMBH occupation fraction, f_{occ} , also affects some results from TRINITY, e.g., the low-mass end of BHMFs and AGN Eddington ratio distributions. In light of this, we will also keep collecting observational constraints on f_{occ} and explore the possibility of its parametrization in future versions of TRINITY. This will help us better understand the evolution of SMBH populations, especially those that are low-mass at high redshifts.

7 CONCLUSIONS

In this work, we introduce TRINITY, which is an empirical model that parametrizes the statistical halo–galaxy–SMBH connec-

tion. (§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 $z = 0 - 10$ (§3, §4.1). These joint constraints enable TRINITY to break degeneracies present in past studies. **Key results are as follows:**

- The normalization of the median $M_{\bullet} - M_{\text{bulge}}$ relation changes by $\lesssim 0.3$ dex from $z = 0 - 3$ and decreases with increasing redshift at $z > 3$ (§4.2, Fig. 11).
- The black hole–halo mass ($M_{\bullet} - M_{\text{peak}}$) relation also changes relatively little from $z = 0 - 3$ (§4.2, Fig. 14). This relation can be approximated as a double power-law at $z \lesssim 5$, and a single power-law at $z \gtrsim 5$. Similar to the $M_{\bullet} - M_{\text{bulge}}$ and $M_{\bullet} - M_{*}$ relations, the normalization of the $M_{\bullet} - M_{\text{peak}}$ relation decreases significantly towards higher redshifts. This could be indicative of a smaller SMBH occupation, f_{occ} , in the early Universe.
- The AGN mass-to-energy conversion efficiency ϵ_{tot} is ~ 0.07 . Current observations do not support more redshift evolution in ϵ_{tot} than about 0.3 dex (§4.6).
- Average SMBH Eddington ratios are around unity at $z \gtrsim 6$. This is consistent with the scenario that different SMBH populations at high redshifts are growing at around the Eddington rate. Towards lower redshifts their Eddington ratios (and thus specific accretion rates) decline. Therefore, we see total black hole mass functions (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. (§4.3, Fig. 19, §4.4, Fig. 21).
- AGNs experience downsizing, in the sense that average Eddington ratios start to decrease earlier for more massive SMBHs. This *does not* hold for average SMBH accretion rates, which do not decrease towards higher masses at low redshifts (§4.3, Figs. 16 and 19).
- 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 (§4.3, Fig. 18).
- Sub-grid physics recipes that qualitatively reproduce the $M_{\bullet} - M_{\text{bulge}}$ 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 §5.1 and Fig. 24.
- Forbidding super-Eddington accretion as well as non-unity occupation fractions prevents SMBHs from growing sufficiently to match the local $M_{\bullet} - M_{\text{bulge}}$ relation. In this scenario, an AGN energy efficiency of $\sim 50\%$ is needed to explain observations like QLFs and QPDFs at high redshifts (Appendix A1, Fig. A1).
- Forbidding redshift evolution of the $M_{\bullet} - M_{\text{bulge}}$ relation results in a best-fitting $M_{\bullet} - M_{\text{bulge}}$ relation that is $\sim 2\sigma$ higher than real observations at $z = 0$. (Appendix A2, Fig. A2).
- 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 A3.1, Fig. A3). Hence, a significant number of “wandering” black holes are necessary.
- 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 al-

ways the same as that for galaxy growth (Appendix A3.2, Figs. A4 and A5).

This work is the first in a series of TRINITY papers. Paper II (H. Zhang et al., in prep.) discusses quasar luminosity functions and the buildup of SMBHs across cosmic time; Paper III (H. Zhang et al., in prep.) presents predictions for quasars and other SMBHs at $z > 6$; Paper IV (H. Zhang et al., in prep.) discusses the SFR-BHAR correlation as a function of halo mass, galaxy mass, and redshift; and paper V (H. Zhang et al., in prep) covers black hole merger rates and TRINITY’s predictions for gravitational wave experiments. Paper VI (O. Knox et al, in prep) and Paper VII (Huinan Zhang et al., in prep) present the AGN auto-correlation functions and AGN–galaxy cross-correlation functions from TRINITY, respectively.

DATA AVAILABILITY

The parallel implementation of TRINITY, the compiled datasets (§3.2), data for all figures, and the posterior distribution of model parameters (§4.1, Appendix G) are available [online](#).

ACKNOWLEDGEMENTS

We thank Stacey Alberts, Rachael Amaro, Gurtina Besla, Haley Bowden, Jane Bright, Katie Chamberlain, Alison Coil, Ryan Endsley, Sandy Faber, Dan Foreman-Mackey, Nico Garavito-Camargo, Nickolay Gnedin, Richard Green, Jenny Greene, Kate Grier, Melanie Habouzit, Kevin Hainline, Andrew Hearin, Julie Hlavacek-Larrondo, Luis Ho, Allison Hughes, Yun-Hsin Huang, Raphael Hviding, Victoria Jones, Stephanie Juneau, Ryan Keenan, David Koo, Andrey Kravtsov, Daniel Lawther, Rixin Li, Joseph Long, Jianwei Lyu, Chung-Pei Ma, Garrett Martin, Karen Olsen, Feryal Öznel, Vasileios Paschalidis, Dimitrios Psaltis, Joel Primack, Yujing Qin, Eliot Quataert, George Rieke, Marcia Rieke, Jan-Torge Schindler, Spencer Scott, Xuejian Shen, Yue Shen, Dongdong Shi, Irene Shvarei, Rachel Somerville, Fengwu Sun, Wei-Leong Tee, Yoshihiro Ueda, Marianne Vestergaard, Feige Wang, Ben Weiner, Christina Williams, Charity Woodrum, Jiachuan Xu, Jinyi Yang, 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 WEBPLOT-DIGITIZER code.² 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

² <https://apps.automeris.io/wpd/>

simulation was performed by Anatoly Klypin within the Bolshoi project of the University of California High-Performance Astro-Computing Center (UC-HiPACC; PI Joel Primack).

APPENDIX A: ALTERNATE MODEL PARAMETRIZATIONS

A1 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. A1 shows the comparison between the local M_\bullet – M_{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_{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 50%—to get as many close-to-Eddington objects and reproduce the observations expressed in luminosities. Given the inconsistency with the observations and the unphysically high AGN efficiencies, we do not adopt this model in the main text.

As discussed in §6.2, Eddington ratios will be artificially shifted higher by the assumption that $f_{\text{occ}} \equiv 1$, if in fact $f_{\text{occ}} < 1$. Therefore, keeping this assumption might lead to overly strong constraints on Eddington ratios. However, without further observational constraints, we opt to keep this assumption. With better constraints on f_{occ} in the future, we will explore the possibility of parametrizing f_{occ} in the improved versions of TRINITY, and revisit the possibility to explain observations without super-Eddington accretions.

A2 Redshift-independent SMBH mass–Bulge Mass relations

In the fiducial model, we assume a redshift-dependent M_\bullet – M_{bulge} relation. Here, we show the results from the “constant M_\bullet – M_{bulge} ” model, where the redshift dependence is dropped. Fig. A2 shows the average M_\bullet , BHAR, Eddington ratio, and BHMR as functions of M_{peak} and z . The results are qualitatively consistent with the fiducial results, except that the “constant M_\bullet – M_{bulge} ” model gives a local normalization of the M_\bullet – M_{bulge} relation $\beta_{\text{BH},0}$ of 8.83. This is 2σ higher than the observational constraints (Table 10). The physical reason is that without redshift evolution of the M_\bullet – M_{bulge} relation, SMBH growth results purely from the evolution of the halo–galaxy bulge connection. Thus, a larger normalization is needed to provide the same amount of SMBH growth and explain the observed data. In light of this, we choose to put this “constant M_\bullet – M_{bulge} ” model in the Appendix, instead of the main text.

A3 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.

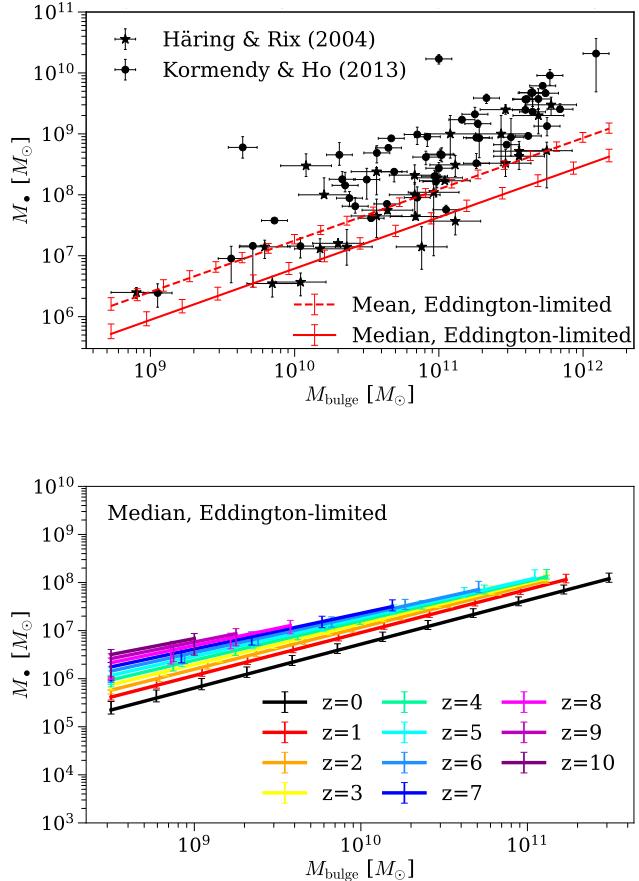


Figure A1. Top Panel: The comparison between the $z = 0$ M_\bullet – M_{bulge} relation from the “Eddington-limited” model and real data. **Bottom Panel:** The redshift evolution of the M_\bullet – M_{bulge} relation from the “Eddington-limited” model, where the SMBH accretions are Eddington-limited. See Appendix A1. All the data used to make this plot (including the individual data points and our best-fitting model) can be found [here](#).

A3.1 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 vs. the intracluster light (ICL) (§2.2). Fig. A3 shows the average BHAR (top panel) and BHMR (bottom 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 $M_{\text{peak}} \sim 10^{13} M_\odot$ 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.

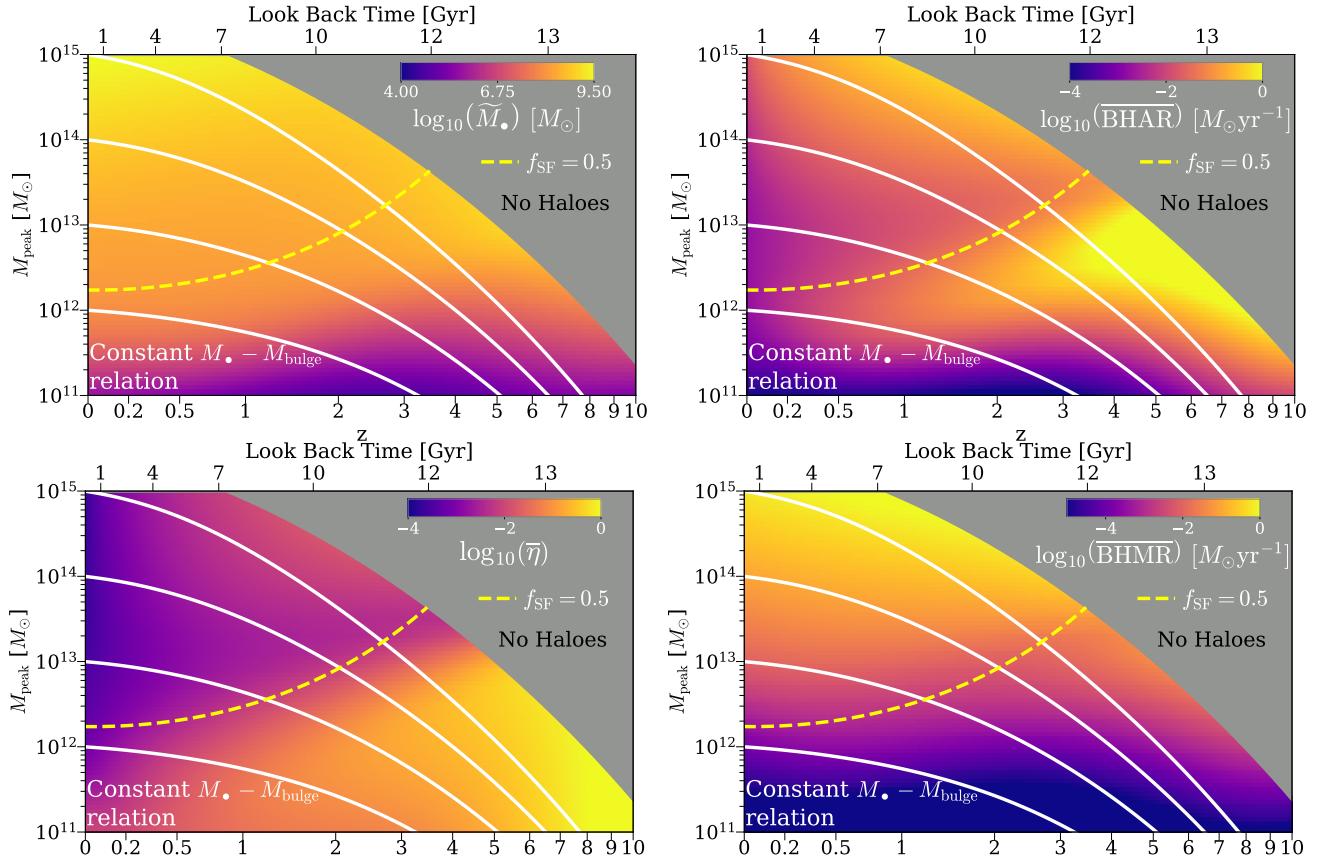


Figure A2. The average M_* , BHAR, Eddington ratio, and BHMR as functions of M_{peak} and z from the “constant M_* – M_{bulge} ” model, where the M_* – M_{bulge} relation is redshift-independent (see Appendix A2). The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” All the data used to make this plot can be found [here](#).

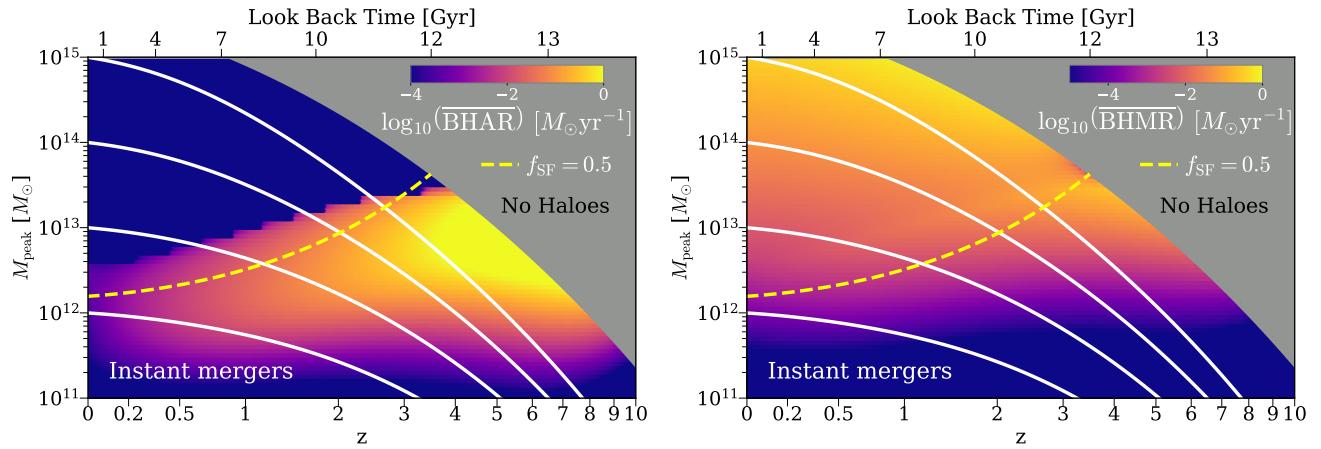


Figure A3. The average BHAR ($\overline{\text{BHAR}}$, top panel) and average BHMR ($\overline{\text{BHMR}}$, bottom panel) as a function of M_{peak} and z from the “instant mergers” model (see Appendix A3.1). “Instant mergers” means that all the infalling SMBHs in galaxy mergers are consumed immediately by the central SMBHs. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” All the data used to make this plot can be found [here](#).

A3.2 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., $0 < f_{\text{scale}} < 1$. 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, $f_{\text{scale}} \equiv 0$ for all galaxies. The second extreme case we consider is if the fractional merger contributions to SMBH and galaxy growth are identical, i.e., $f_{\text{scale}} \equiv 1$. In the following, we call this scenario the “same mergers” model.

Fig. A4 shows the resulting M_{\bullet} – M_{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 are largely consistent with each other. Compared to the “no mergers” and fiducial models, the “same mergers” model has slightly higher M_{\bullet} at high redshifts, where f_{scale} is significantly below unity in the first two models. At the massive end, the “same mergers” model also produces less evolution in M_{\bullet} , which is also due to a larger f_{scale} .

Fig. A5 shows the average Eddington ratios as functions of M_{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 halos with $M_{\text{peak}} \gtrsim 10^{14} M_{\odot}$ below $z \sim 2$. From the top panel to the bottom panel, TRINITY attributes more and more SMBH growth to mergers among these halos, producing lower and lower average Eddington ratios. However, the general “downsizing” picture holds qualitatively across all these models.

APPENDIX B: GALAXY MERGER RATES

In TRINITY, SMBH mergers are directly linked to galaxy mergers. Here, we use the galaxy merger rates from the UNIVERSEMACHINE, where satellite galaxies will disrupt when their $v_{\text{max}}/v_{M\text{peak}}$ ratios reach a certain threshold (see §2.2 for the definitions of v_{max} and $v_{M\text{peak}}$). We refer readers to §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_{desc} the mass of the descendant halo, M_{sat} the mass of the satellite halo, and $\theta = M_{\text{sat}}/M_{\text{desc}}$ 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:

$$-\frac{d^2N(M_{\text{desc}}, \theta, z)}{dz d \log_{10} \theta} = 10^{A(M_{\text{desc}}, a)} \theta^{B(a)} \exp(-3.162\theta) \quad (\text{B1})$$

$$A(M_{\text{desc}}, a) = A_0(M_{\text{desc}}) + A_1(a) \quad (\text{B2})$$

$$A_0(M_{\text{desc}}) = 0.148 \log_{10}\left(\frac{M_{\text{desc}}}{10^{12} M_{\odot}}\right) - 0.291 \quad (\text{B3})$$

$$A_1(a) = -1.609 + 3.816a + (-2.152)a^2 \quad (\text{B4})$$

$$B(a) = -1.114 + 1.498a + (-0.757)a^2. \quad (\text{B5})$$

We show the quality of these fits in Fig. B1. Compared to Behroozi et al. (2013), these merger rates are lower by 15–40% due to the presence of orphan galaxies in 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 (BHGRs) 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 §2.5) SMBHs.

In TRINITY, haloes inherit both central and wandering SMBHs from their most massive progenitors (MMPs). The average inherited central SMBH mass is:

$$\overline{M}_{\bullet, \text{inherit}}(M_{\text{peak}}, z) = \int_0^{\infty} P(M_{\text{MMP}}|M_{\text{peak}}, z) M_{\bullet, \text{MMP}}(M_{\text{MMP}}, z) dM_{\text{MMP}}, \quad (\text{C1})$$

where $P(M_{\text{MMP}}|M_{\text{peak}}, z)$ is the conditional probability that a halo with M_{peak} has a MMP with mass M_{MMP} , which is calculated based on the adopted N-body simulations (see §3.1). $M_{\bullet, \text{MMP}}$ is the average SMBH mass of the MMPs at a progenitor mass of M_{MMP} at the previous snapshot, which is a function of M_{MMP} and z , 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 is then, by definition:

$$\begin{aligned} \overline{M}_{\bullet, \text{infall}}(M_{\text{peak}}, z) = & \int_0^{\infty} \mathcal{R}(M_{\text{peak}}, M_{\text{sat}}, z) M_{\bullet, \text{sat}} dM_{\text{sat}} \\ & + \int_0^{\infty} P(M_{\text{MMP}}|M_{\text{peak}}, z) M_{\bullet, \text{MMP, wandering}}(M_{\text{MMP}}, z) dM_{\text{MMP}}, \end{aligned} \quad (\text{C2})$$

where M_{sat} is the peak halo masses of satellite haloes, and $\mathcal{R}(M_{\text{peak}}, M_{\text{sat}}, z)$ is the merger rate of satellite haloes into the descendant haloes. This rate is calculated by integrating Eq. B1 over the redshift dimension:

$$\mathcal{R}(M_{\text{peak}}, M_{\text{sat}}, z) = \int_z^{z+\Delta z} \frac{d^2N(M_{\text{peak}}, \theta, z)}{d \log \theta dz} dz, \quad (\text{C3})$$

where Δz is the redshift interval between the two consecutive snapshots, and $\theta = M_{\text{sat}}/M_{\text{peak}}$ is the mass ratio between the satellite and the descendant haloes.

APPENDIX D: 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 log-normal scatter from the UNIVERSEMACHINE as functions of redshift, halo mass (M_{peak}), and star formation rates to calculate galaxy UV luminosity functions 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 \tilde{M}_{UV} have the following dependence on redshift, M_{peak} , and SFR:

$$\tilde{M}_{\text{UV}} = k_{\text{UV}} \times \log_{10} \text{SFR} + b_{\text{UV}} \quad (\text{D1})$$

$$\begin{aligned} k_{\text{UV}} = & 0.154 \left(\log_{10} M_{\text{peak}} \right)^2 + (-2.876) \log_{10} M_{\text{peak}} \\ & + (-2.378)(a-1) + 9.478 \end{aligned} \quad (\text{D2})$$

$$\begin{aligned} b_{\text{UV}} = & (-0.347) \left(\log_{10} M_{\text{peak}} \right)^2 + 6.853 \log_{10} M_{\text{peak}} \\ & + 1.993(a-1) + (-50.344) \end{aligned} \quad (\text{D3})$$

The log-normal scatter σ_{UV} has the following redshift and M_{peak} dependency:

$$\sigma_{\text{UV}} = k_{\sigma_{\text{UV}}} \times \log_{10} M_{\text{peak}} + b_{\sigma_{\text{UV}}} \quad (\text{D4})$$

$$k_{\sigma_{\text{UV}}} = -0.031z + 0.042 \quad (\text{D5})$$

$$b_{\sigma_{\text{UV}}} = 0.319z + 0.241. \quad (\text{D6})$$

Fig. D1 shows the goodness of fit for Eqs. (D1)-(D6) to both \tilde{M}_{UV} and σ_{UV} from $z = 8 - 10$. Using these fitting functions, TRINITY produces SFRs and galaxy UV luminosities that are both consistent with the UNIVERSE MACHINE.

APPENDIX E: AGN DATA: CORRECTIONS, EXCLUSIONS, AND UNCERTAINTIES

E1 Compton-thick correction

As mentioned in §3.2, we have adopted quasar luminosity functions (QLFs) from Ueda et al. (2014) to constrain the total AGN energy budget. However, the Ueda et al. data points do not include Compton-thick obscured AGNs. Hence, we applied the following empirical correction given by Ueda et al. (2014) to convert from Compton-thin-only QLFs to total QLFs:

$$\begin{aligned} \Phi_{L,\text{tot}}(L_X, z) &= \Phi_{L,\text{CTN}}(L_X, z) \times (1 + \alpha_{\text{CTK}} \psi(L_X, z)) \\ \psi(L_X, z) &= \min[0.84, \max[\psi_{43.75}(z) - 0.24L_{43.75}, \psi_{\min}]] \\ \psi_{43.75}(z) &= \begin{cases} 0.43(1+z)^{0.48} & [z < 2.0] \\ 0.43(1+2)^{0.48} & [z \geq 2.0] \end{cases} \\ L_{43.75} &= \log_{10}(L_X/\text{erg s}^{-1}) - 43.75, \end{aligned} \quad (\text{E1})$$

where $\psi(L_X, z)$ is the fraction of Compton-thin absorbed AGN, and α_{CTK} is the number ratio between Compton-thick and Compton-thin AGN. Ueda et al. adopted $\alpha_{\text{CTK}} = 1$ in their main analysis, but their analysis of the cosmic X-ray background radiation shows that there is a $\pm 50\%$ uncertainty in α_{CTK} . In light of this, we ran TRINITY with $\alpha_{\text{CTK}} = 0.5$ and 2.0 , aside from the fiducial model where $\alpha_{\text{CTK}} = 1.0$. The *only* model parameters that show significant differences are the $z = 0$ SMBH total efficiency ($\epsilon_{\text{tot},0}$, Eq. 48, Fig. E1) and the duty cycle ($f_{2,0}$, Eq. 41, Fig. E2). A higher α_{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. Additionally, larger α_{CTK} also increases the amplitude of quasar QPDFs from Aird et al. (2018), which are also corrected for Compton-thick AGNs. From Eqs. 42 and 46, we see that with all the parameters fixed, the characteristic Eddington ratio of AGN, η_0 , would increase with the average Eddington ratio, $\bar{\eta}$. But such an increase will cause a horizontal shift of QPDFs, which undermines the goodness of fitting to QPDFs (Fig. 9). Therefore, the AGN duty cycle $f_{2,0}$ should increase with $\epsilon_{\text{tot},0}$ to (1) cancel the horizontal shift in η_0 ; and (2) elevate the model QPDFs to fit the data points that are shifted upwards due to a larger α_{CTK} . In Fig. E2, we see such a shift in $f_{2,0}$. Fig. E3 shows that the change in $\epsilon_{\text{tot},0}$ and $f_{2,0}$ are modulated such that the characteristic Eddington ratio η_0 remains invariant, which is constrained by the shape of QPDFs from Aird et al. (2018).

E2 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.

Firstly, Aird et al. (2018) modeled 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 release³ were evaluated with these model functions over a dense grid of sL_X . Thus, naively taking all the points in their data release would artificially increase the weight of this dataset. To avoid this, we downsampled their modeled AGN probability distribution functions with 1 dex spacing. This choice is based on the fact that the spacing between two neighboring 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 datasets, we found that there is significant inconsistency between the QLFs from Ueda et al. (2014) and the high- sL_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 datasets, we excluded AGN probability distribution function points with $z > 2.5$ or $sL_X > 1$ from Aird et al. (2018). After removing the most inconsistent data points, residual inconsistencies on the order of 0.3 dex persist between these two datasets. 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 sL_X (see Eq. 66 in §2.8).

E3 Active black hole functions from Schulze & Wisotzki (2010) and Schulze et al. (2015)

In TRINITY, we use active black hole mass functions (ABHMFs) at $z = 0.2$ and $z = 1.5$ from Schulze & Wisotzki (2010) and Schulze et al. (2015). However, two factors were addressed before using these ABHMFs as constraints. Firstly, 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 of +0.2 dex for all the ABHMF data points at $z = 0.2$ to keep consistency with those at $z = 1.5$.

APPENDIX F: 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 firstly convert data points and their uncertainties into log units if they are in linear

³ available at <https://zenodo.org/record/1009605>.

units. For the i -th data point with a value of $y_i - e'_{\text{low},i}^{+e'_{\text{high},i}}$, we then convolve the error bars with a calculation tolerance of 0.01 dex:

$$e_{\text{low/high},i} = \sqrt{e'^2_{\text{low/high},i} + 0.01^2}. \quad (\text{F1})$$

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 $|\hat{y}_i - y_i| \leq \epsilon_{\text{fit}} \equiv 0.02$, 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 $|\hat{y}_i - y_i| > \epsilon_{\text{fit}}$, we define:

$$\Delta y_i = \begin{cases} \hat{y}_i - y_i - \epsilon_{\text{fit}}, & \hat{y}_i > y_i \\ \hat{y}_i - y_i + \epsilon_{\text{fit}}, & \hat{y}_i < y_i \end{cases}, \quad (\text{F2})$$

and the χ^2 for this data point is:

$$\chi_i^2 = \begin{cases} \left(\Delta y_i / e_{\text{low},i}\right)^2, & \Delta y_i < -e_{\text{low},i} \\ \left(\Delta y_i / e_{\text{high},i}\right)^2, & \Delta y_i > e_{\text{high},i} \\ \left(\Delta y_i / e_{\text{med},i}\right)^2, & \text{otherwise} \end{cases}, \quad (\text{F3})$$

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

$$e_{\text{med},i}(\Delta y_i) = e_{\text{low},i} + \frac{\Delta y_i + e_{\text{low},i}}{e_{\text{high},i} + e_{\text{low},i}} \cdot (e_{\text{high},i} - e_{\text{low},i}). \quad (\text{F4})$$

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

$$\chi^2 = \sum_i \chi_i^2 + \text{priors}. \quad (\text{F5})$$

APPENDIX G: BEST FITTING PARAMETER VALUES

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

Median Star Formation Rates:

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

$$\log_{10}(V) = 2.203_{-0.017}^{+0.033} + (1.258_{-0.185}^{+0.210})(a-1) + (1.033_{-0.134}^{+0.158})\ln(1+z) + (-0.041_{-0.028}^{+0.018})z$$

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

$$\log_{10}(\epsilon) = 0.373_{-0.110}^{+0.131} + (0.845_{-0.759}^{+0.880})(a-1) + (1.720_{-0.616}^{+0.623})\ln(1+z) + (0.161_{-0.101}^{+0.096})z$$

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

$$\alpha = -4.616_{-0.349}^{+0.161} + (31.968_{-1.731}^{+0.909})(a-1) + (20.338_{-0.988}^{+0.689})\ln(1+z) + (-2.059_{-0.105}^{+0.108})z$$

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

$$\beta = -0.682_{-0.193}^{+0.454} + (2.453_{-0.938}^{+0.323})(a-1) + (0.854_{-0.186}^{+0.070})z$$

Quenched Fractions:

Characteristic v_{max} for quenching [km/s] (Eq. 10):

$$\log_{10}(v_Q) = 2.324_{-0.030}^{+0.011} + (0.216_{-0.114}^{+0.056})(a-1) + (0.254_{-0.027}^{+0.026})z$$

Width in $\log-v_{\text{max}}$ for quenching [dex] (Eq. 11):

$$w_Q = 0.189_{-0.032}^{+0.010} + (0.187_{-0.102}^{+0.062})(a-1) + (0.039_{-0.017}^{+0.027})z$$

Galaxy Mergers :

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

$$\log_{10}(f_{\text{merge}}) = -1.061_{-0.245}^{+0.047}$$

The Halo–Galaxy Connection:

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

$$\sigma_* = 0.242_{-0.004}^{+0.024}$$

Correlation coefficient between SSFR and M_* at fixed halo mass at $a = 0.5$ (i.e., $z = 1$) (Eq. 27):

$$\rho_{0.5} = 0.370_{-0.059}^{+0.098}$$

Systematics in Stellar Masses:

Offset between the true and the measured M_* [dex] (Eq. 23):

$$\mu = -0.111_{-0.039}^{+0.130} + (0.292_{-0.036}^{+0.040})(a-1)$$

Additional systematic offset between the true and the measured SFRs (Eq. 25):

$$\kappa = 0.319_{-0.027}^{+0.028}$$

Scatter between the observed and the true M_* [dex] (Eq. 26):

$$\sigma = \min\{0.07 + 0.053_{-0.011}^{+0.005}(z-0.1), 0.3\}$$

Galaxy–SMBH Connection:

Slope and zero point of the SMBH mass – bulge mass (M_{\bullet} – M_{bulge}) relation (Eq. 28):

$$\gamma_{\text{BH}} = 1.082_{-0.057}^{+0.013} + (0.221_{-0.034}^{+0.066})(a-1) + (0.133_{-0.013}^{+0.026})z$$

$$\beta_{\text{BH}} = 8.470_{-0.049}^{+0.113} + (-0.553_{-0.066}^{+0.063})(a-1) + (0.023_{-0.050}^{+0.006})z$$

Scatter in the M_{\bullet} – M_{bulge} relation [dex] (Eq. 31):

$$\sigma_{\text{BH}} = 0.285_{-0.025}^{+0.035}$$

SMBH Mergers:

The fraction of SMBH growth due to mergers, relative to galaxy mergers (Eq. 35):

$$\log_{10}(f_{\text{scale}}) = -0.048_{-2.285}^{+0.008} + (0.600_{-0.701}^{+1.300})(a-1)$$

AGN Properties:

AGN duty cycles (Eqs. 38–41):

$$M_{\text{BH,c}} = 7.827_{-0.216}^{+0.280}$$

$$w_{\text{BH,c}} = 0.739_{-0.060}^{+0.124}$$

$$f_2 = 0.308_{-0.144}^{+0.099} + (-0.697_{-0.157}^{+0.148})(a-1)$$

Power-law indices of the Eddington ratio distributions (Eqs. 43 and 44):

$$c_1 = -0.071_{-0.184}^{+0.121} + (0.796_{-0.245}^{+0.172})(a-1)$$

$$c_2 = 1.398_{-0.012}^{+0.146} + (0.126_{-0.059}^{+0.213})(a-1)$$

AGN energy efficiencies (Eq. 47):

$$\log_{10}(\epsilon_{\text{rad}}) = -1.161^{+0.006}_{-0.231} + 0.025^{+0.040}_{-0.318}(a-1)$$

AGN Systematics:

Offset in the Eddington ratio between Ueda et al. (2014) and Aird et al. (2018) [dex] (Eq. 66):

$$\xi = -0.315^{+0.051}_{-0.020}$$

APPENDIX H: PARAMETER CORRELATIONS

Fig. H1 shows the rank correlation coefficients between all the model parameters, with 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 behavior of others at certain redshift intervals. In other words, different redshift evolution terms are not orthogonal to each other.

Strong correlations are seen between the normalization of the median local M_{\bullet} - M_{bulge} relation, $\beta_{\text{BH},0}$, and the random scatter around the relation, σ_{BH} . Based on Soltan argument, the cosmic SMBH mass density is determined by the integral of quasar luminosity functions across cosmic time. To reproduce this mass density, a lower *median* M_{\bullet} - M_{bulge} relation must be paired with a bigger σ_{BH} to produce higher *mean* SMBH masses. We also see strong correlation between the redshift evolution of the M_{\bullet} - M_{bulge} relation, $\beta_{\text{BH},a}$, and that of the systematic offset between the observed vs. true stellar mass, μ_a . This is because a stronger redshift evolution in μ implies weaker evolution in *intrinsic* galaxy stellar mass functions. Without any change in the M_{\bullet} - M_{bulge} relation, this weaker redshift evolution will be propagated into black hole mass functions, resulting in too little SMBH accretion to account for the observed quasar luminosity functions. To offset this effect, we need a M_{\bullet} - M_{bulge} relation that evolves more strongly, i.e., whose $\beta_{\text{BH},a}$ is more negative.

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

REFERENCES

- 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., Labb   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 et al., 2014, *A&A*, **571**, A30
- Planck Collaboration et al., 2016, *A&A*, **594**, A13
- Ricarte A., Tremmel M., Natarajan P., Zimmer C., Quinn T., 2021, arXiv e-prints, p. arXiv:2103.12124
- 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
- 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-Escud   J., 2009, *ApJ*, **690**, 20
- Shankar F., Weinberg D. H., Miralda-Escud   J., 2013, *MNRAS*, **428**, 421
- Shankar F., et al., 2016, *MNRAS*, **460**, 3119
- Shen Y., et al., 2019, *ApJ*, **873**, 35
- Shen X., Hopkins P. F., Faucher-Gigu  re 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
- 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
- Soltan 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
- 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., Gottl  ber 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, *ApJL*, **836**, L1
- Tremaine S., et al., 2002, *ApJ*, **574**, 740
- 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
- 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., 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., 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, *ApJL*, **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
- van Dokkum P. G., Conroy C., 2012, *ApJ*, **760**, 70
- 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

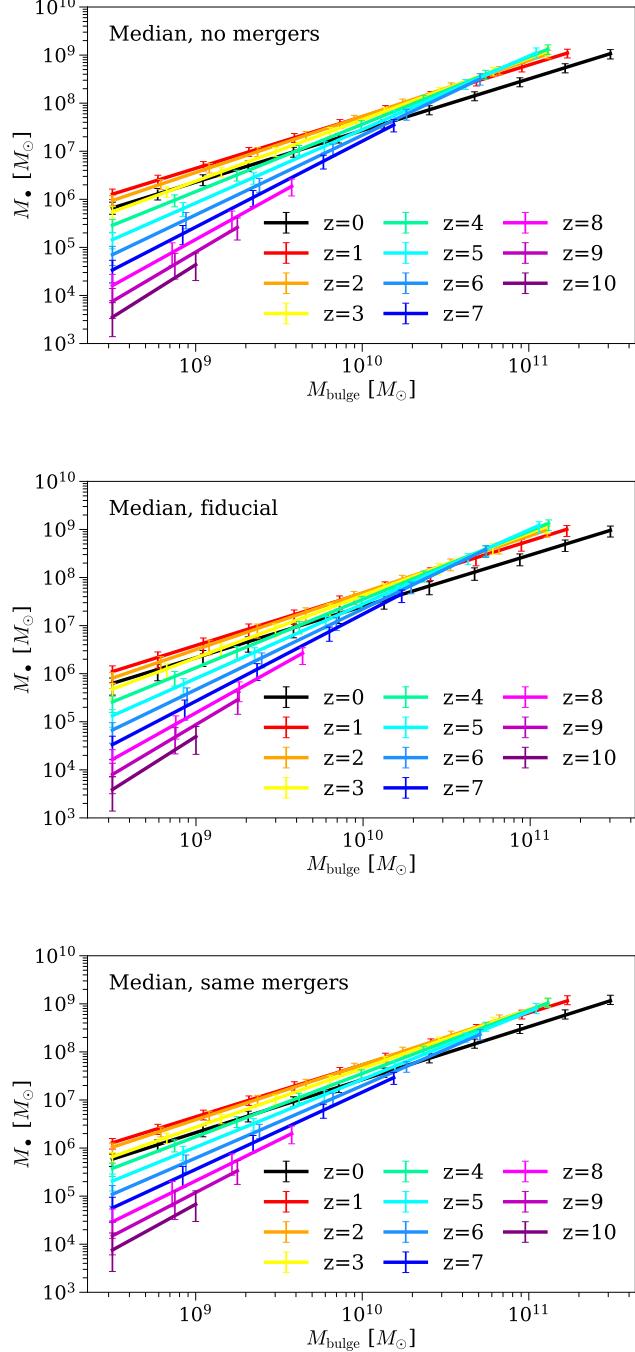


Figure A4. The median M_{\bullet} - M_{bulge} relations as a function of z from the “no mergers” model (top panel, no SMBH mergers take place), the fiducial model (middle panel), and the “same mergers” model (bottom panel, the fractional merger contribution to SMBH growth being the same as that for galaxy growth). See Appendix A3.2. All the data used to make this plot can be found [here](#).

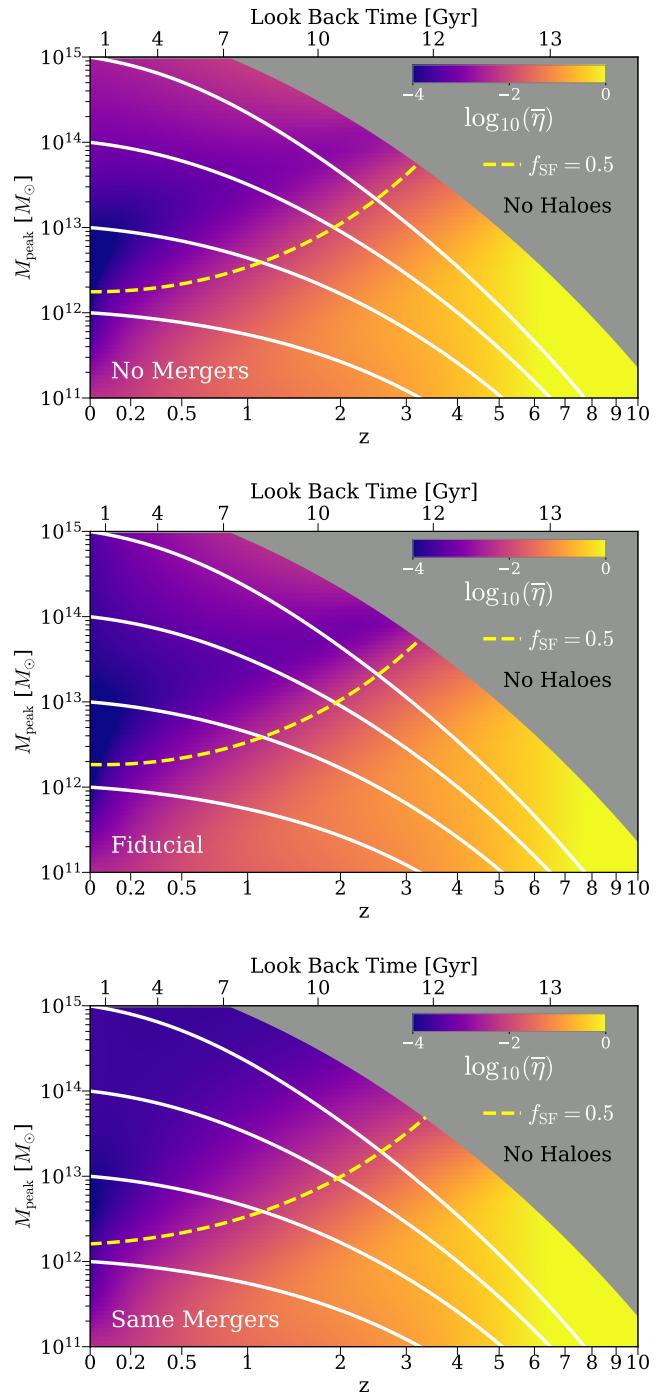


Figure A5. The average Eddington ratios as functions of M_{peak} and z from the “no mergers” model (top panel), the fiducial model (middle panel), and the “same mergers” model (bottom panel). See Appendix A3.2. The yellow dashed line shows the halo mass at which the galaxy star-forming fraction f_{SF} is 0.5 as a function of z . The white solid lines are the average mass growth curves of haloes with $M_{\text{peak}} = 10^{12}, 10^{13}, 10^{14}$, 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 labeled as “No Haloes.” All the data used to make this plot can be found [here](#).

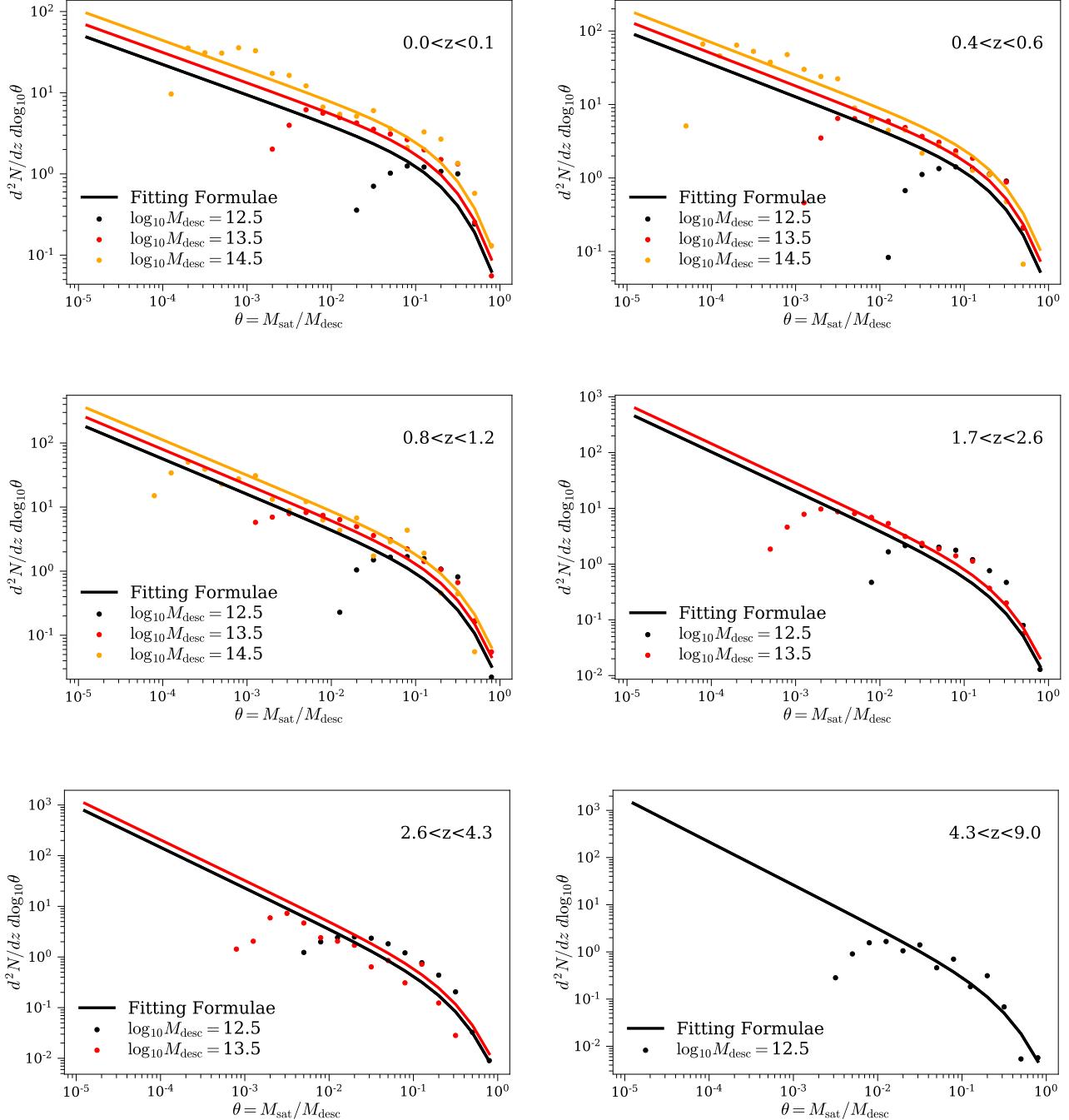


Figure B1. The rate of satellite galaxy disruption in host haloes in the UNIVERSE MACHINE as a function of z , descendant mass M_{desc} , and satellite-to-descendant mass ratio $\theta = M_{\text{sat}}/M_{\text{desc}}$. The solid symbols are the binned estimates of merger rates, and the solid lines are the fitted results. 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](#).

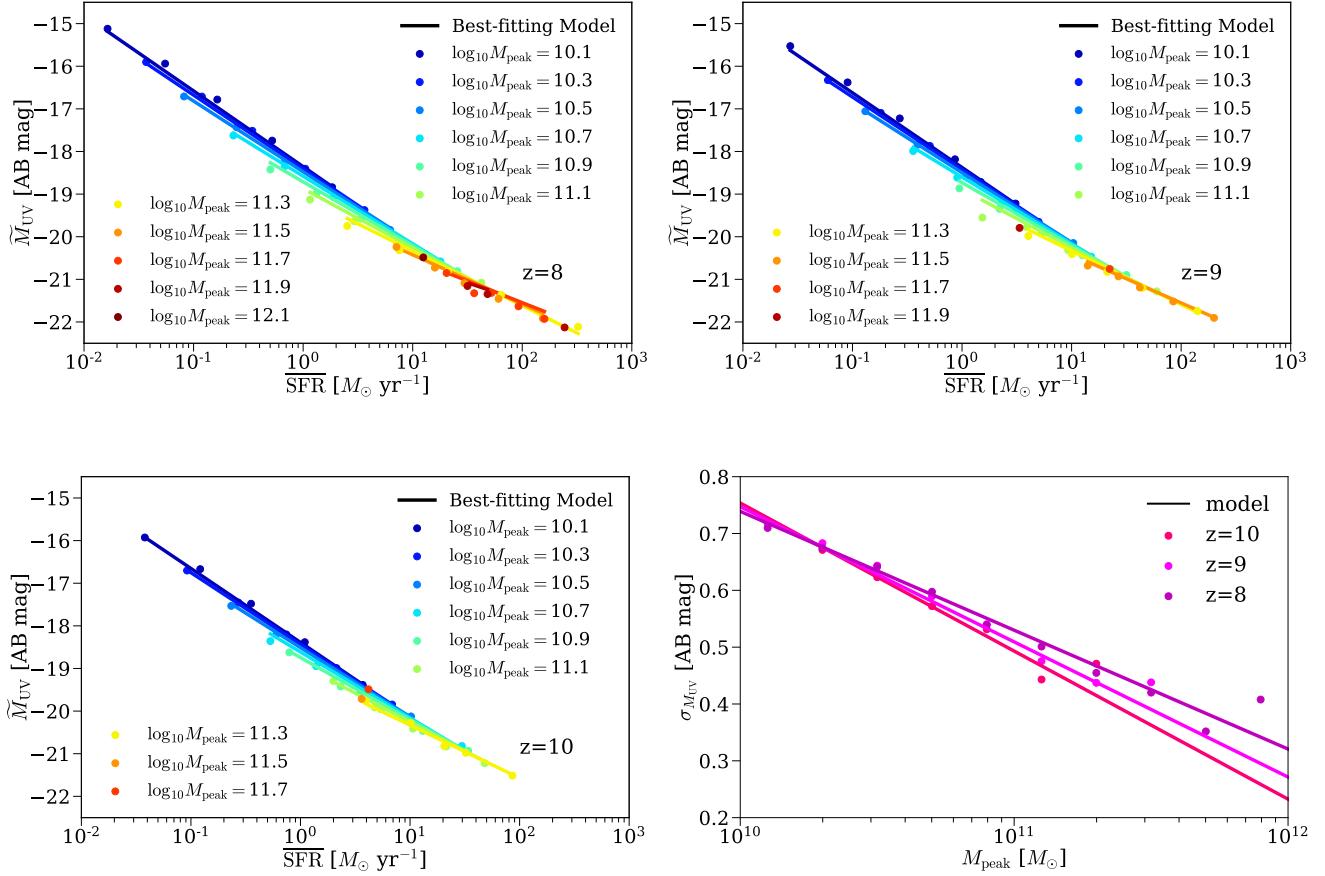


Figure D1. The fits to median UV magnitude, \tilde{M}_{UV} , as a function of M_{peak} , $\overline{\text{SFR}}$, and z , and the corresponding scatter, $\sigma_{M_{\text{UV}}}$, as a function of M_{peak} and z , from the UNIVERSE MACHINE. The filled circles are the data points from the UNIVERSE MACHINE, and the solid lines are the best-fitting models in Eqs. D1–D6. See Appendix D. All the data used to make this plot (including the individual data points and our best-fitting model) can be found [here](#).

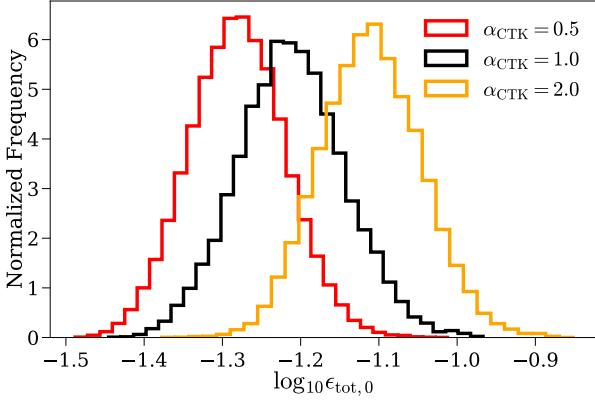


Figure E1. The comparison of SMBH efficiency $\epsilon_{\text{tot},0}$ between models with $\alpha_{\text{CTK}} = 0.5, 1.0$, and 2.0 . See Appendix E1. All the data used to make this plot can be found [here](#).

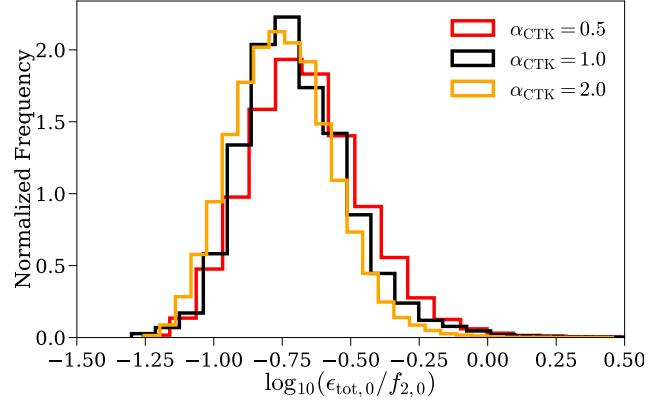


Figure E3. The comparison of $\epsilon_{\text{tot},0}/f_{2,0}$ between models with $\alpha_{\text{CTK}} = 0.5, 1.0$, and 2.0 . See Appendix E1. All the data used to make this plot can be found [here](#).

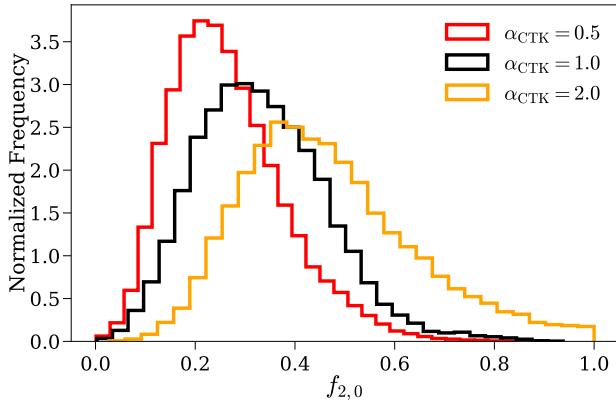


Figure E2. The comparison of SMBH duty cycle $f_{2,0}$ between models with $\alpha_{\text{CTK}} = 0.5, 1.0$, and 2.0 . See Appendix E1. All the data used to make this plot can be found [here](#).

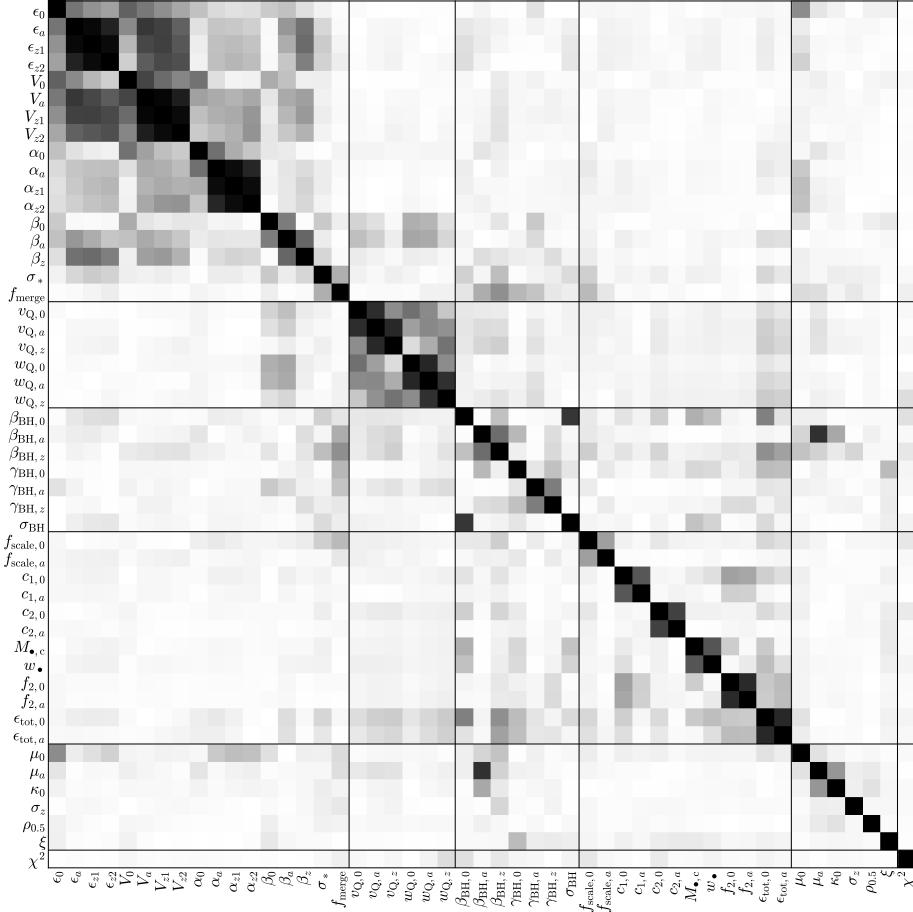


Figure H1. Rank correlation coefficients in the model posterior distribution. Darker shades indicate higher *absolute values* of correlation coefficients (both positive and negative). See Appendix H. All the data used to make this plot can be found [here](#).