MNRAS 000, 1–21 (2025) Preprint 2 September 2025 Compiled using MNRAS LATEX style file v3.2

Tracing individual black hole growth histories and quasar lightcurves in an N-body Universe

Elia Pizzati[1]★, Joseph F. Hennawi[1,2], Joop Schaye[1], et al.
[1] *Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA Leiden, The Netherlands*
[2] *Department of Physics, University of California, Santa Barbara, CA 93106, USA*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present a new model for the evolution of supermassive black holes (SMBHs) and quasars across cosmic time. The framework builds on merger trees from the FLAMINGO large-volume dark-matter-only simulation, linking SMBH growth histories to those of their host halos through parametric prescriptions that capture both average trends and stochastic variability. SMBH accretion is modeled self-consistently and directly drives quasar activity, with the goal of reproducing the observed evolution of the luminous quasar population. Our model is designed to match three key observables: the bolometric quasar luminosity function (QLF), the conditional Eddington ratio distribution function (cERDF) at fixed bolometric luminosity, and the clustering of UV-luminous quasars. Despite residual discrepancies at the faint end of the QLF, in the mean of the cERDF, and in quasar clustering at $z \approx 4$, our model provides a close match to the most robust observational constraints currently available. Additionally, we find that: (a) SMBHs grow primarily through bursts of super-critical accretion ($\dot{M}_{\rm BH} > \dot{M}_{\rm Edd}$); (b) these bursts must be sufficiently long-lived, making the coherence timescale of accretion a key parameter in shaping the SMBH mass distribution; (c) the predicted $M_{\rm BH}$–$M_{\rm halo}$ relation remains approximately constant with redshift, with relatively tight scatter ($\lesssim 0.3$ dex) but extended high-SMBH-mass tails that give rise to luminous quasars; and (d) mergers contribute only marginally to SMBH growth compared to accretion, even under optimistic assumptions about merger timescales and remnant survival. The resulting SMBH growth histories, merger trees, and quasar light curves provide a versatile framework for future comparisons with the expanding range of observational constraints.

Key words: large-scale structure of Universe – quasars: general – quasars: supermassive black holes

1 INTRODUCTION

Understanding the growth and evolution of supermassive black holes (SMBHs) remains a central challenge in astrophysics. These objects are believed to reside at the centers of nearly all massive galaxies (e.g., Magorrian et al. 1998; Ferrarese & Merritt 2000; Kormendy & Ho 2013), where their accretion-powered emission gives rise to Active Galactic Nuclei (AGN) and quasars – some of the most luminous objects in the Universe (Salpeter 1964; Zel'dovich & Novikov 1967; Lynden-Bell 1969). Since the discovery of the first quasar (Schmidt 1963), a robust theoretical framework has emerged: the observed quasar radiation originates from the release of gravitational energy as matter accretes onto the black hole, with a small fraction of this rest-mass energy (known as *radiative efficiency*) converted into radiation, and the rest fueling SMBH growth.

Building on this theoretical framework, a foundational link between quasar activity and black hole growth was established by Soltan (1982). The Soltan argument posits that the redshift evolution of the quasar luminosity function (QLF) directly traces the accretion history of SMBHs. By integrating the observed quasar emission over cosmic time and assuming a value for the radiative efficiency, one can estimate the total black hole mass density accumulated during luminous accretion phases. This insight laid the foundation for a broader class of empirical models for SMBH evolution, which describe the growth of the black hole mass function by solving a continuity equation. These models constrain key physical parameters – such as radiative efficiency, duty cycle, and the Eddington ratio distribution – by requiring consistency between the observed quasar population across cosmic time and the local census of dormant black holes (e.g., Yu & Tremaine 2002; Merloni & Heinz 2008; Shankar et al. 2009; Aversa et al. 2015; Tucci & Volonteri 2017).

A complementary class of empirical models extends the treatment of SMBH evolution by explicitly linking it to galaxy formation, typically by assuming a redshift-dependent relationship between black hole and galaxy properties. These models build on well-established empirical connections between dark matter halos, galaxies, and SMBHs (e.g., Kormendy & Ho 2013; Reines & Volonteri 2015), and often adopt semi-empirical galaxy–halo frameworks (e.g., Behroozi et al. 2013) as a foundation. They then incorporate SMBH growth and quasar activity in a way that is consistent with both galaxy and black hole observables (e.g., Croton 2009; Conroy & White 2013; Caplar et al. 2015). A notable recent example is the TRINITY model (Zhang et al. 2023), which jointly evolves the average properties of halos, galaxies, and SMBHs within a unified, data-constrained

★ pizzati@strw.leidenuniv.nl

framework. By reproducing a wide range of observables – including galaxy stellar mass functions, quasar luminosity functions, and black hole–galaxy scaling relations – these models provide powerful tools to investigate the co-evolution of halos, galaxies, and black holes across cosmic time.

The rationale behind empirical models stems from the recognition that a first-principles treatment of SMBH accretion in a cosmological context remains fundamentally out of reach. While supermassive black holes are routinely included in both semi-analytic models (SAMs) and cosmological hydrodynamical simulations – and AGN feedback is widely acknowledged as a key driver of galaxy evolution (e.g., Somerville & Davé 2015) – current theoretical frameworks are still limited in their ability to model SMBH growth in a self-consistent, physically grounded way. The core difficulty lies in the extreme dynamic range involved: black hole accretion unfolds on scales much smaller than a parsec (pc), yet it is regulated by – and feeds back into – processes acting on kiloparsec to tens-of-megaparsec scales. Bridging these disparate scales in a physically accurate manner remains computationally unfeasible, complicating the development of robust sub-grid (physical) models that can be reliably implemented in large-scale simulations and SAMs.

As a result, while sub-grid prescriptions for star formation processes have reached a relatively mature and consistent formulation, the modeling of black hole seeding, accretion, and feedback remains coarse and exhibits substantial variation across simulation platforms (Habouzit et al. 2021). It is therefore not surprising that even state-of-the-art cosmological simulations – despite their success in reproducing galaxy populations across cosmic epochs (Vogelsberger et al. 2020) – continue to struggle with matching SMBH and AGN observables. Most are calibrated to reproduce local SMBH-galaxy scaling relations (e.g., Di Matteo et al. 2005; Booth & Schaye 2009), yet yield widely divergent predictions for black hole growth and evolution at earlier times (Habouzit et al. 2021, 2022; Porras-Valverde et al. 2025).

To address these shortcomings, significant effort in recent years has gone into refining the treatment of AGN in cosmological simulations, with the goal of more accurately connecting black hole growth to the broader context of galaxy formation and evolution. Advances have been made across multiple fronts, including improved models for black hole seeding (e.g., Bhowmick et al. 2024), more accurate prescriptions for black hole dynamics (e.g., Genina et al. 2024), revised accretion models that go beyond the classical Bondi-Hoyle approach (e.g., Koudmani et al. 2024; Weinberger et al. 2025), and increasingly sophisticated implementations of AGN feedback (e.g., Huško et al. 2024). Nevertheless, fundamental uncertainties remain – particularly in the high-redshift regime – underscoring the continued need for flexible, data-driven models. In this landscape, empirical approaches serve as a valuable counterpart to simulations, extracting physical constraints directly from observations and offering an efficient means to explore parameter space. Ultimately, the combination of improved simulations and empirically anchored models offers a promising path toward unraveling the complex history of SMBH growth across cosmic time.

In parallel with theoretical progress, the observational frontier has been advancing rapidly – particularly at high redshifts. Over the past few decades, large-scale spectroscopic surveys have revealed luminous quasars powered by SMBHs with masses $\gtrsim 10^9\,\mathrm{M}_\odot$ at $z \gtrsim 6$, during the epoch of reionization (Fan et al. 2006; Mazzucchelli et al. 2017; Farina et al. 2022; Fan et al. 2023), and even out to $z \approx 7.5$, just 700 million years after the Big Bang (Bañados et al. 2018; Yang et al. 2020; Wang et al. 2021). These discoveries pose a significant challenge to conventional models of SMBH formation and growth. If black hole seeds originate from Population III stellar remnants with initial masses of $\sim 100\,\mathrm{M}_\odot$ (e.g., Heger et al. 2003), and accrete at the Eddington limit, there is simply not enough time to reach $\gtrsim 10^9\,\mathrm{M}_\odot$ by $z \sim 7$.

This tension has motivated the development of alternative scenarios for early black hole growth. Proposed pathways include the formation of massive seeds through the direct collapse of pristine gas clouds (e.g., Bromm & Loeb 2003; Volonteri et al. 2008; Latif & Ferrara 2016), runaway stellar mergers in dense nuclear star clusters (e.g., Omukai et al. 2008; Devecchi & Volonteri 2009), and sustained phases of super-Eddington accretion (e.g., Volonteri et al. 2015; Lupi et al. 2016; Inayoshi et al. 2016). Despite their theoretical appeal, however, current observations are insufficient to distinguish between these competing models. The problem is fundamentally degenerate: widely different combinations of initial seed masses, accretion rates, duty cycles, and merger histories can be fine-tuned to reproduce the observed SMBH population. Disentangling these scenarios and uncovering the early growth history of SMBHs requires new and more detailed observations.

Interestingly, new observational probes are now beginning to emerge. With the advent of JWST, AGN candidates hosting moderately massive black holes ($\sim 10^6$–$10^8\,\mathrm{M}_\odot$) have been identified at redshifts as high as $z \approx 8$–10 (e.g., Maiolino et al. 2024; Kokorev et al. 2023; Larson et al. 2023; Bogdán et al. 2023). Although the physical nature of many of these sources remains uncertain, ongoing and upcoming wide-field surveys with Euclid and the Roman Space Telescope are expected to deliver statistically robust samples of luminous quasars at the highest redshifts (e.g., Yang et al. in prep.). At the same time, different quasar observables beyond SMBH mass estimates – such as the luminosity function (Schindler et al. 2023; Matsuoka et al. 2023), Eddington ratio distributions (Wu et al. 2022), clustering (Arita et al. 2023; Eilers et al. 2024), and proximity zone sizes (Eilers et al. 2017, 2020; Ďurovčíková et al. 2024) – are now being extended to earlier epochs.

Emerging trends from these early data are already beginning to challenge traditional views of SMBH growth in the high-z Universe. In particular, recent clustering measurements indicate that luminous quasars at $z \approx 6-7$ exhibit surprisingly low duty cycles (Eilers et al. 2024; Pizzati et al. 2024b; Huang et al. in prep.), suggesting that only a small fraction of SMBHs are actively accreting at any given time. These findings are consistent with independent constraints on quasar lifetimes derived from proximity zone sizes and damping wings in quasar spectra (Davies et al. 2019; Ďurovčíková et al. 2024), as well as from the spatial extent of the Lyα nebulae powered by quasar radiation (Ďurovčíková et al. 2025). Taken together, these results are difficult to reconcile with simple scenarios of continuous, Eddington-limited growth. Instead, they point toward a more nuanced evolutionary picture, in which SMBH accretion is stochastic and episodic, luminous phases are short-lived, and multiple growth pathways – including obscured or radiatively inefficient accretion – may contribute to the assembly of the most massive black holes in the early Universe.

At the same time, the analyses presented in Pizzati et al. (2024a,b) embed the quasar population within a cosmological context, leveraging luminosity functions and clustering measurements at different redshifts – including the latest constraints at $z \gtrsim 6$. Using a consistent and homogeneous empirical framework, these studies uncover a puzzling trend: quasar properties, particularly their clustering, appear to evolve rapidly with redshift. This behavior is largely driven by the exceptionally strong clustering signal reported by Shen et al. (2007) at $z \approx 4$ using Sloan Digital Sky Survey (SDSS) data – a result that still awaits independent confirmation, but it nonetheless

underscores our incomplete understanding of quasar evolution beyond cosmic noon. A key limitation of the Pizzati et al. (2024a,b) studies, however, is that each redshift is modeled in isolation, without tracing any evolutionary connection across epochs. Constructing a coherent, time-resolved model of SMBH growth is essential for interpreting these redshift trends within a unified framework.

Motivated by these considerations – and by the growing body of high-redshift observational constraints – we introduce a new empirical framework for modeling the evolution of SMBHs and quasars. Our model, BAQARO (Black Hole Accretion and Quasar Activity in a Realistic Observational framework), is designed to capture the early buildup of SMBH mass and the emergence of luminous quasars from cosmic dawn to cosmic noon. Quasar activity and SMBH growth are treated self-consistently, incorporating constraints not only from quasar luminosity functions but also from clustering measurements as well as from the distribution of SMBH masses and Eddington ratios at different epochs. This integrated approach allows us to explore a wide range of physical growth scenarios within a unified, observationally anchored framework.

The model is built on the merger trees and halo catalogs from the dark-matter-only (DMO) version of the FLAMINGO cosmological simulation suite (Schaye et al. 2023; Kugel et al. 2023). Specifically, we use the $(2.8\,\mathrm{cGpc})^3$ volume run, which offers the statistical power needed to sample the rare, luminous ($L_{\rm bol} \gtrsim 10^{47}\,\mathrm{erg\,s^{-1}}$) quasar population out to the highest redshifts. By leveraging the output of large N-body simulations, we are able to trace SMBH accretion histories and quasar light curves along individual halo growth trajectories, naturally capturing the diversity and stochasticity of black hole evolutionary pathways. This is essential to recover the most massive SMBHs ($\gtrsim 10^9\,\mathrm{M_\odot}$) observed at early times, which are extreme outliers in the distribution of SMBH growth histories. Furthermore, modeling quasar clustering directly from the simulated large-scale structure eliminates the need for linear bias or halo model prescriptions, which are known to perform poorly for the halo mass and redshift regimes relevant to bright quasars (e.g., Mead & Verde 2021).

The paper is structured as follows. In Section 2, we introduce the key components of the BAQARO model and the data it aims to reproduce. In Section 3, we present the main results of our analysis, comparing our fiducial model with all observational constraints. Section 3.3 studies the implications of our model for the growth of SMBHs at early cosmic times and for the scaling relations between quasar/SMBH and halo properties. We summarize our findings and discuss our results from a broader perspective in Section 4.

2 METHODS

At its core, BAQARO combines DMO cosmological simulations with a phenomenological prescription for black hole seeding and growth. We use the merger trees extracted from the FLAMINGO large-volume simulation to trace the assembly histories of dark matter halos, within which we model the evolution of SMBHs. Black holes are seeded in early halos and subsequently grow through a combination of gas accretion and black hole mergers. The accretion rate, together with an assumed radiative efficiency, determines the bolometric luminosity of each SMBH, allowing us to predict quasar light curves along with individual merger histories.

Our primary goal is to reproduce the statistical properties of the bright quasar population from cosmic dawn to cosmic noon. To this end, we calibrate our model to match three key observables across redshift: (i) the quasar luminosity function (QLF), (ii) the conditional Eddington ratio distribution function (cERDF), and (iii) the large-scale clustering of quasars. These constraints jointly inform the underlying growth histories of SMBHs and the physical conditions that shape quasar activity.

In the remainder of this section, we describe each component of the model in detail, beginning with the underlying simulation and halo merger trees, and proceeding through the prescriptions for seeding, merger, accretion, and quasar luminosity modeling.

2.1 Extracting halo mass histories and merger trees from the FLAMINGO simulation

To model the evolution of SMBHs and quasars in a cosmological context, we require a realistic description of the growth and assembly histories of dark matter halos across cosmic time. This is crucial because quasars are rare, highly biased tracers that inhabit the most massive and rapidly evolving structures in the Universe – environments whose complexity and stochasticity cannot be fully captured by analytic models of merger trees or average mass accretion histories (e.g., extended Press–Schechter; Lacey & Cole 1993). In particular, analytic approaches struggle to reproduce the nonlinear structure formation, mergers, and diverse growth trajectories of massive halos, especially at high redshift. Large cosmological N-body simulations, by contrast, can track halo growth and merger histories in detail, providing the physically grounded framework needed to model the evolution of luminous quasars in the Universe.

We obtain these halo growth histories from the DMO version of the FLAMINGO cosmological simulations (Schaye et al. 2023; Kugel et al. 2023), which combine the resolution and large volume necessary to capture the environments in which luminous quasars form and evolve. FLAMINGO is a suite of state-of-the-art simulations run with the SWIFT code (Schaller et al. 2024), which couples an N-body gravity solver with smooth particle hydrodynamics (SPH). Gravitational interactions are computed using the Fast Multipole Method (Greengard & Rokhlin 1987). The simulations adopt the "3×2pt + all" cosmological parameters from Abbott et al. (2022): $\Omega_{\rm m} = 0.306$, $\Omega_{\rm b} = 0.0486$, $\sigma_8 = 0.807$, $H_0 = 68.1\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, and $n_{\rm s} = 0.967$, with a total neutrino mass of 0.06 eV. Massive neutrinos are included via the δf method of Elbers et al. (2021). Initial conditions are generated with multi-fluid third-order Lagrangian perturbation theory (3LPT), using partially fixed phases to reduce cosmic variance (Angulo & Pontzen 2016): the amplitudes of modes with $(kL)^2 < 1025$ are set to match the mean variance, where k is the wavenumber and L the box size.

In this work, we employ the DMO FLAMINGO run with a comoving box size of $L = 2.8\,\mathrm{cGpc}$, comprising 5040^3 cold dark matter (CDM) particles and 2800^3 neutrino particles. This corresponds to a CDM particle mass of $M_{\rm CDM} = 6.72{\times}10^9\,\mathrm{M_\odot}$, which – although relatively low in resolution – is sufficient to resolve the massive haloes expected to host luminous quasars. In future developments of our model, we plan to address this limitation by exploiting the newly developed FLAMINGO-10k simulation (Schaller et al., in prep.; Pizzati et al. 2024b), which contains eight times more particles and will enable us to trace SMBH growth starting from significantly lower-mass progenitors.

2.1.1 *Subhalo masses and specific halo accretion rates*

Our first step is to construct a comprehensive halo catalog across all simulation snapshots of interest. Specifically, we consider the 39 snapshots spanning from $z = 15$ – the highest redshift available in

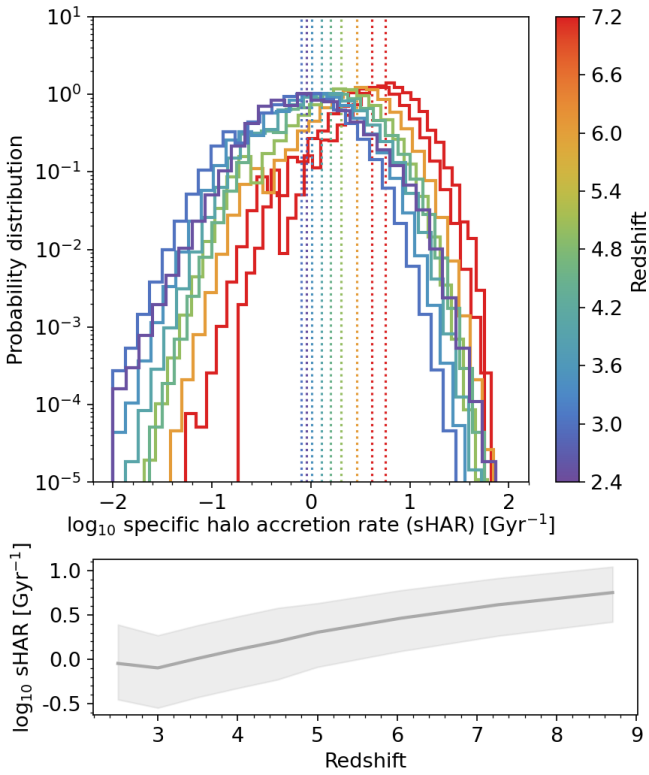

Figure 1. *Top*: Probability distribution of the specific halo accretion rate (sHAR) measured between two consecutive snapshots (Eq. 1). The corresponding redshifts are shown with different colors. The median values of the distributions are highlighted with dotted vertical lines. *Bottom*: Evolution of the median (solid line) and 16th-84th percentiles (shaded region) of the specific halo accretion rate as a function of redshift. The median specific halo accretion rate decreases by almost one order of magnitude between $z \approx 8.5$ and $z \approx 2.5$.

the simulation – down to $z = 2$. This lower redshift limit, which currently bounds our model to cosmic noon, is chosen to reduce computational cost; it will be extended in future iterations of the model. The snapshots are approximately evenly spaced in time, with a mean spacing of $\sim$ 80 Myr. The time intervals in the redshift range considered vary from a minimum of 40 Myr to a maximum of 180 Myr.

We use the FLAMINGO halo catalogs generated with the HBT-HERONS code (Forouhar Moreno et al. 2025), an upgraded implementation of the Hierarchical Bound-Tracing (HBT) algorithm originally developed by Han et al. (2012, 2018). HBT-HERONS identifies subhalos and follows their evolution across time by accounting for key physical processes such as mergers, accretion, and tidal stripping. To achieve this, it tags particles associated with a subhalo at a given snapshot based on their dynamical history, and then propagates these associations forward in time. In later snapshots, particles that originated from the same progenitor subhalo are regrouped to identify descendant subhalo candidates. This approach ensures temporal consistency and enables accurate reconstruction of merger trees for substructures in simulations. Notably, the resulting merger trees exhibit significantly fewer pathological failures – such as mass swapping or unphysical transients – than those created with other tree-building algorithms (Chandro-Gómez et al. 2025). These features make HBT-HERONS particularly well suited to our study, as capturing the hierarchical evolution of SMBHs depends critically on the fidelity of halo merger trees.

We adopt the bound mass definition for measuring subhalo masses: for each subhalo, the mass is computed by summing the gravitationally bound particles. However, satellite subhalos are often subject to strong tidal stripping, which can significantly reduce their instantaneous bound mass and thereby obscure their past gravitational influence. For this reason, we use the peak bound mass, $M_{\rm peak}$, defined as the maximum bound mass a subhalo has reached over the course of its assembly history. Conveniently, HBT-HERONS provides this quantity for each object by default, and we adopt it as our fiducial subhalo mass definition – i.e., we take $M_{\rm h} \equiv M_{\rm peak}$. Using this convention, we construct subhalo mass histories by following each object from the snapshot where it first appears through to the final snapshot included in our model, currently set at $z = 2$.

Although HBT-HERONS is capable of identifying subhaloes down to a minimum of 20 particles, we apply a stricter resolution cut to ensure the robustness of our results. Specifically, we exclude any halo that never exceeds 40 bound particles at any point between its formation and $z = 2$. This threshold removes transient or poorly resolved subhaloes that could otherwise introduce noise or instability, while retaining the well-resolved halos that meaningfully contribute to the buildup of supermassive black holes and the quasar population.

In addition to tracking subhalo masses, we also measure halo accretion rates, defined as the mass growth of subhaloes per unit time. These are computed directly from the merger trees by taking the difference in halo mass between two consecutive snapshots and dividing by the time interval separating them. In FLAMINGO, the typical snapshot spacing is $\sim$ 80 Myr (see above), which corresponds to a substantial fraction of the dynamical time for the halos of interest at the redshifts we study. This snapshot cadence allows us to resolve halo growth on timescales that are well-matched to the key cosmological processes governing halo and galaxy evolution – and, ultimately, the long-term fueling of SMBHs.

While the time spacing between snapshots in FLAMINGO is approximately constant, it is not strictly uniform across all redshifts. As a result, the accretion rates we compute are averaged over different time intervals at different epochs. In this initial analysis, we do not correct for this variation. However, because these differences are modest and the overall cadence remains physically meaningful, we do not expect this to introduce significant biases in our results. Nonetheless, we plan to implement redshift-dependent correction factors in future iterations to more rigorously account for this effect.

We emphasize that the computed accretion rates reflect the total mass growth of halos, including both smooth accretion and mergers. We do not attempt to separate these contributions, as both cosmological inflows and merger events are expected to play important roles in triggering black hole growth and quasar activity. For our purposes – linking SMBH accretion to host halo evolution – this combined accretion measure provides a physically motivated and practical proxy.

The top panel of Fig. 1 shows the distribution of specific accretion rates, $s\dot{M}_{\rm acc}$, for all halos in the simulation across a range of redshifts – corresponding to a subset of all the available snapshots. The specific accretion rate between two consecutive snapshots i and $i+1$ is defined as:

$$s\dot{M}_{\rm acc,i} = \frac{M_{{\rm h},i+1} - M_{{\rm h},i}}{(t_{i+1} - t_i)\, M_{{\rm h},i}}, \tag{1}$$

where $M_{{\rm h},i}$ and t_i are the halo mass and cosmic time at snapshot i, respectively. This quantity captures the relative growth rate of halos and serves as the foundation for our SMBH accretion prescriptions (Sec. 2.2).

The bottom panel of Fig. 1 illustrates the redshift evolution of halo growth by showing the median and scatter (standard deviation) of the specific accretion rate distribution as a function of redshift. As expected, typical accretion rates are higher at earlier cosmic times, reflecting the accelerated pace of structure formation in the high-redshift Universe (e.g., McBride et al. 2009).

2.1.2 *Construction of the merger tree catalogs*

Thanks to the structure of HBT-HERONS, merger trees are naturally constructed by following the evolution of subhalos over time through the tracker particle method described above. To build our merger tree catalog, we extract descendant information for each subhalo flagged as "dead" by the halo finder. This status is assigned when a subhalo is either gravitationally disrupted or sinks toward the center of a larger halo, meeting the merging criterion defined in Forouhar Moreno et al. (2025).

For subhaloes that have sunk, we directly identify the halo they merge into – referred to as the sink halo in the code. For those that are disrupted but not sunk, we search for a descendant subhalo that shares the majority of the tracker particles previously bound to the disrupted object. If such a match is found, it is designated as the descendant. If no suitable descendant can be identified based on particle overlap, we fall back to the HBT parent–child hierarchy, assuming the subhalo merges with its immediate parent as defined by the algorithm.

In a very small fraction of cases ($\lesssim 0.1\%$), no parent halo can be identified. These cases typically involve low-mass or field haloes that become tidally disrupted or were only transiently detected due to noise or numerical artifacts. We classify such objects as "lost", and we do not attempt to track the evolution of their associated SMBHs beyond the point of disruption.

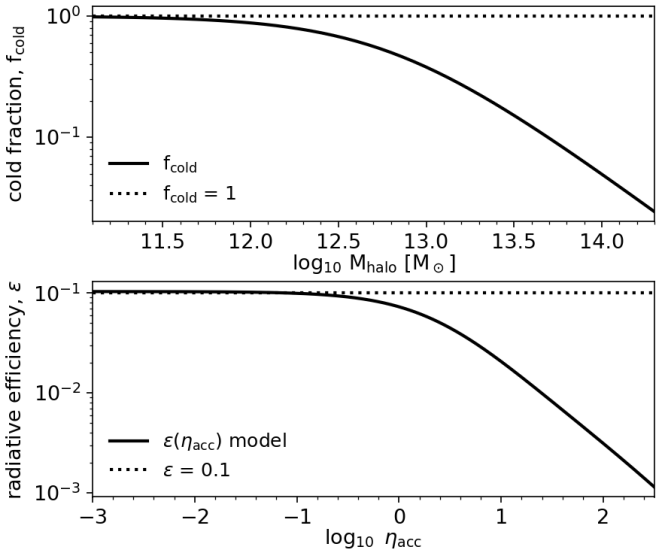

Figure 2. *Top*: Fraction of cold gas accreted onto halos, $f_{\rm cold}$, as a function of halo mass, $M_{\rm halo}$. The cold fraction accounts for the suppression of cold inflows in massive halos, where virial shock heating raises the gas temperature above the cooling threshold. We adopt the parametrization of Correa et al. (2018, see Eq. 5), assuming a redshift-independent $f_{\rm cold}$ since their model shows only weak evolution beyond cosmic noon. *Bottom*: Radiative efficiency, ϵ, as a function of the specific black hole accretion rate, $\eta_{\rm acc}$. We adopt the parametrization of Madau et al. (2014), fixing the spin parameter to $a = 0.67$, which yields a sub-Eddington efficiency of $\epsilon_0 \approx 0.1$. As $\eta_{\rm acc}$ approaches and exceeds the Eddington limit, ϵ decreases due to the transition to the slim-disk regime, where photon trapping and advective processes reduce the radiative output of the accretion flow (e.g., Sądowski et al. 2014).

2.2 Modeling SMBH and quasar evolution

Our model for the growth and radiative output of SMBHs is built around three key components: initialization, accretion (and the associated quasar emission), and mergers. Each of these processes is detailed separately in the subsections that follow.

2.2.1 *Black hole initialization*

For the initialization of black holes, we adopt a simple prescription: a black hole is assigned to the center of each subhalo at the snapshot where the halo first appears in the merger tree. The black hole is initialized with a fixed mass, $M_{\rm start}$. This mass should not be interpreted as a physical seed mass in the early Universe, but rather as an empirical value that marks the beginning of the SMBH growth track within our model. A physical treatment of the seed mass regime would require either simulations with much higher mass resolution, capable of resolving halos down to $\sim 10^6 - 10^7\ {\rm M}_\odot$, or an analytical framework extending SMBH growth histories down to $10^2 - 10^4\ {\rm M}_\odot$.

Using a uniform value of $M_{\rm start}$ for all subhalos is, of course, a strong simplification. In reality, SMBH masses are expected to vary with host halo mass – which is similar for newly formed halos but not identical – as well as with the formation environment and underlying seeding channel (e.g., Li et al. 2021; Jeon et al. 2025). While intrinsic scatter in $M_{\rm start}$ could be easily incorporated into our framework to reflect these diverse formation pathways, its impact on our predictions is largely degenerate with scatter in the accretion rates. In practice, the quasar luminosities and final black hole masses are determined primarily by the integrated accretion history rather than the precise initial mass.

Nevertheless, the interplay between seeding and accretion remains an important open question for SMBH evolution models. In future work, we plan to explore alternative initialization prescriptions and to quantify the extent to which different assumptions about $M_{\rm start}$ can be disentangled from variations in accretion.

2.2.2 *Black hole mergers*

Black hole mergers are implemented in our model using a straightforward, mass-conserving prescription: when two host halos merge, we assume their central SMBHs merge instantaneously, and the remnant is assigned a mass equal to the sum of the progenitor masses. This simplifying assumption neglects the complex dynamical processes that, in reality, delay SMBH coalescence after the host halos (or galaxies) merge.

In hierarchical structure formation, the two SMBHs first sink toward the common centre via dynamical friction against the surrounding dark matter, gas, and stars (Begelman et al. 1980; Mayer et al. 2007). At kiloparsec to parsec scales, the pair forms a bound binary that hardens further through stellar scattering (e.g., Milosavljević & Merritt 2001) and/or interaction with circumbinary gas discs (e.g., Dotti et al. 2012). Only once gravitational wave (GW) emission dominates the energy loss does the binary inspiral and coalesce. This multi-stage process can introduce delays ranging from a few hundred Myr to several Gyr between the halo merger and SMBH coalescence, depending on the host properties, gas content, and redshift.

The coalescence itself can impart a GW recoil velocity to the

Table 1. Free parameters of the model, together with their values for the fiducial run analyzed in this work. A brief description of each parameter, along with the corresponding equation, is also provided.

Parameter	Value (fiducial model)	Description	Equation
$M_{\rm start}$	$5 \times 10^6\,{\rm M}_\odot$	SMBH mass initialized in newly-formed halos	-
$\tau_{\rm coherence}$	10 Myr	Coherence timescale of SMBH accretion	Eq. 12
$\eta_{\rm av,0}$	-1.12	Average sBHAR for $s\dot{M}_{\rm cold,acc} = 1\,{\rm Gyr}^{-1}$	Eq. 7
$\eta_{\rm av,evol}$	0.95	Power-law index of the average sBHAR – $s\dot{M}_{\rm cold,acc}$ relation	Eq. 7
σ_0	0.52	Scatter in the sBHAR for $s\dot{M}_{\rm cold,acc} = 1\,{\rm Gyr}^{-1}$	Eq. 8
$\sigma_{\rm evol}$	-0.17	Power-law index of the sBHAR scatter – $s\dot{M}_{\rm cold,acc}$ relation	Eq. 8

remnant SMBH due to the asymmetric emission of gravitational waves (Bekenstein 1973). Numerical relativity simulations show that these kicks can reach up to several thousand km/s for particular mass ratios and spin configurations (e.g., Herrmann et al. 2007). In massive galaxies, such velocities may cause the SMBH to oscillate about the galactic centre for hundreds of Myr, while in lower-mass systems they can exceed the escape speed, ejecting the SMBH entirely from its host (e.g., Blecha & Loeb 2008).

While these processes can be incorporated into physical models via phenomenological prescriptions (e.g., Volonteri & Rees 2006; Tanaka & Haiman 2009; Kelley et al. 2017), here we adopt the most conservative choice and neglect them. This effectively assumes zero delay between the subhalo merger and SMBH coalescence, and ignores the possibility of displacement or ejection. As such, our treatment yields an upper limit to the contribution of mergers to SMBH mass growth. As shown in Sec. 3.3.3, even under this optimistic assumption, mergers contribute only a subdominant fraction of SMBH growth across almost all redshifts of interest. The details of the merger prescriptions, therefore, have little influence on our main results, although they could become important in future work where we plan to study the SMBH merger rates and the occupation fraction of SMBHs in galaxy populations.

2.2.3 *Black hole accretion and quasar radiation*

In contrast to seeding and mergers, modeling black hole accretion – and the associated quasar activity – requires more detailed physical prescriptions. Accretion is the dominant growth channel for SMBHs over most of cosmic history and directly governs their observable luminosity output. Our approach is to connect the specific black hole accretion rate (sBHAR) to the specific halo accretion rate of the host subhalo, supplemented by a stochastic component that accounts for the scatter seen in observationally inferred accretion rates (e.g., Alexander et al. 2025).

We describe the probability distribution of the sBHAR as a conditional function, $P(\eta_{\rm acc} \mid s\dot{M}_{\rm cold,acc})$, where $\eta_{\rm acc}$ is the black hole accretion rate normalized to the Eddington rate, and $s\dot{M}_{\rm cold,acc}$ is the specific cold gas accretion rate onto the halo. The black hole specific accretion rate, $\eta_{\rm acc}$, is given by

$$\eta_{\rm acc} = \frac{\dot{M}_{\rm BH,acc}}{\dot{M}_{\rm Edd}}, \tag{2}$$

where $\dot{M}_{\rm BH,acc}$ is the SMBH mass accretion rate and $\dot{M}_{\rm Edd}$ is the Eddington accretion rate. The latter depends on the adopted radiative efficiency $\epsilon_0 = 0.1$ via $\dot{M}_{\rm Edd} = L_{\rm Edd}/\epsilon_0 c^2$, with the Eddington luminosity scaling linearly with black hole mass:

$$L_{\rm Edd} = \frac{4\pi G M_{\rm BH} m_p c}{\sigma_T} \approx 1.3 \times 10^{38} \left(\frac{M_{\rm BH}}{{\rm M}_\odot}\right) {\rm erg\,s^{-1}}, \tag{3}$$

where G is the gravitational constant, m_p is the proton mass, c is the speed of light, σ_T is the Thomson cross-section, and $M_{\rm BH}$ is the black hole mass.

The specific cold gas accretion rate onto halos, $s\dot{M}_{\rm cold,acc}$, is obtained by applying a cold fraction, $f_{\rm cold}$, to the total specific halo accretion rate, $s\dot{M}_{\rm acc}$ (Fig. 2, top panel):

$$s\dot{M}_{\rm cold,acc} = f_{\rm cold}\, s\dot{M}_{\rm acc}. \tag{4}$$

The cold fraction encapsulates the physical suppression of cold inflows in massive halos, where virial shock heating raises the gas temperature above the cooling threshold (Dekel & Birnboim 2006). In contrast, low-mass halos – particularly at high redshift – can sustain efficient cold gas accretion through narrow filaments of the cosmic web.

We model $f_{\rm cold}$ as a function of halo mass, $M_{\rm h}$, following the phenomenological fit proposed by Correa et al. (2018). Their model, calibrated on the EAGLE simulation (Schaye et al. 2015), reproduces the transition between cold-mode accretion in low-mass halos and hot-mode accretion in massive halos. While in general $f_{\rm cold}$ depends on both halo mass and redshift, the parametrization of Correa et al. (2018) shows only weak redshift evolution beyond cosmic noon. We therefore adopt a fixed functional form across cosmic time:

$$f_{\rm cold}(M_{\rm h}) = 1 - \frac{1}{1 + \left(\frac{M_{200}}{M_{1/2}}\right)^a}, \tag{5}$$

with $a = -1.07$ and $\log_{10}(M_{1/2}/{\rm M}_\odot) = 12.8$ fixed following Correa et al. (2018). In this model, $M_{1/2}$ marks the characteristic halo mass where half of the inflowing gas is in the cold phase, while a controls the steepness of the transition between cold and hot accretion regimes. As shown in the top panel of Fig. 2, the cold fraction declines sharply above $M_{\rm h} \sim 10^{12.5}\,{\rm M}_\odot$, reflecting the increasing dominance of virial shock heating in massive halos.

Because $s\dot{M}_{\rm cold,acc}$ is the primary driver of SMBH fueling in our framework, this scaling provides a direct link between large-scale halo growth and the small-scale accretion processes that power quasars. In this model, the black hole accretion rate depends solely on the host halo mass – through its cold gas fraction – and on its total accretion rate. We do not impose any explicit redshift dependence, based on the assumption that the fundamental physical mechanisms governing SMBH growth are not directly dictated by cosmic time. Instead, redshift dependence emerges naturally through the evolution of the specific halo accretion rate itself (Fig. 1, bottom panel). As

discussed in Sec. 3, this physically motivated, minimal prescription captures the key trends in SMBH and quasar evolution that our model is designed to reproduce.

The conditional accretion rate distribution, $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$, can in principle take a variety of functional forms. We experimented with several parametrizations, including log-normal distributions, Schechter functions, and broken power laws. For the purposes of this work, we adopt a log-normal form, which we find to provide an adequate fit to the data while remaining mathematically simple. The conditional sBHAR distribution is thus written as

$$P(\log_{10}\eta_{\rm acc}|s\dot{M}_{\rm cold,acc}) = \\ = \frac{1}{\sqrt{2\pi}\sigma(s\dot{M}_{\rm cold,acc})}\exp\left(-\frac{\log_{10}^2\left(\eta_{\rm acc}/\eta_{\rm av}(s\dot{M}_{\rm cold,acc})\right)}{2\sigma^2(s\dot{M}_{\rm cold,acc})}\right) \quad (6)$$

where $\eta_{\rm av}$ and σ are the mean and log-normal scatter (expressed in dex) of the distribution, respectively.

In our implementation, the mean and scatter are each parametrized as a power-law function of the specific cold gas accretion rate:

$$\eta_{\rm av}(s\dot{M}_{\rm cold,acc}) = \eta_{\rm av,0}\left(\frac{s\dot{M}_{\rm cold,acc}}{1{\rm Gyr}^{-1}}\right)^{\eta_{\rm av,evol}} \quad (7)$$

$$\sigma(s\dot{M}_{\rm cold,acc}) = \sigma_0\left(\frac{s\dot{M}_{\rm cold,acc}}{1{\rm Gyr}^{-1}}\right)^{\sigma_{\rm evol}} \quad (8)$$

This choice keeps the model both flexible and interpretable: η_0 and σ_0 set the normalization of the distribution at a fiducial accretion rate of $1\,{\rm Gyr}^{-1}$, while $\eta_{\rm evol}$ and $\sigma_{\rm evol}$ control how the mean and scatter respond to changes in cold gas supply.

Although more elaborate functional forms are possible – including broken power laws or redshift-dependent terms – we find that this simple, four-parameter power-law scaling yields a satisfactory match to the observational constraints considered in this work. As shown later, it captures both the central trend and the dispersion of the sBHAR distribution across the relevant range of halo accretion rates.

Once the sBHAR distribution is specified, we can compute SMBH mass growth by accretion between two consecutive simulation snapshots. Assuming that $\eta_{\rm acc}$ remains constant between snapshots i and $i+1$, the black hole mass evolves according to the exponential growth expected for Eddington-limited accretion:

$$M_{\rm BH}(t_{i+1}) = M_{\rm BH}(t_i)\,\exp\left[\frac{t_{i+1}-t_i}{t_{\rm acc}(\eta_{\rm acc})}\right], \quad (9)$$

where $t_{\rm acc}(\eta_{\rm acc})$ is the Salpeter timescale corresponding to the chosen accretion rate.

The Salpeter timescale quantifies the e-folding time for black hole mass growth at a given $\eta_{\rm acc}$ and radiative efficiency ϵ:

$$t_{\rm acc}(\eta_{\rm acc}) = \frac{\epsilon}{(1-\epsilon)\eta_{\rm acc}}\frac{\sigma_T\,c}{4\pi G m_p} \approx \frac{4.5\times10^7\ {\rm yr}}{\eta_{\rm acc}}\left(\frac{\epsilon}{0.1}\right)\left(\frac{0.9}{1-\epsilon}\right), \quad (10)$$

The numerical approximation corresponds to the canonical Salpeter time for $\epsilon = 0.1$ and $\eta_{\rm acc} = 1$.

In general, the radiative efficiency ϵ is not a fixed quantity, but depends on the accretion rate, $\eta_{\rm acc}$. Both analytic arguments and numerical simulations indicate the existence of distinct accretion regimes with different radiative properties (e.g., Shakura & Sunyaev 1973; Abramowicz et al. 1988; Sądowski et al. 2014). At sub-Eddington rates ($\eta_{\rm acc} \lesssim 1$), accretion flows are typically radiatively efficient and well described by the geometrically thin, optically thick Shakura–Sunyaev disk model (Shakura & Sunyaev 1973). In this regime, ϵ is approximately constant – typically $\epsilon \sim 0.06$–0.3 depending on black hole spin – reflecting the high efficiency with which gravitational binding energy is converted into radiation (Thorne 1974).

As the accretion rate approaches and exceeds the Eddington limit ($\eta_{\rm acc} \gtrsim 1$), however, the efficiency can drop sharply. In this regime, accretion is often described by slim-disk models (Abramowicz et al. 1988; Sądowski et al. 2014) in which photon trapping becomes significant: radiation generated in the inner disk is advected inward with the gas rather than escaping. Combined with powerful radiation-driven outflows, this effect leads to a "saturated" luminosity that increases only logarithmically with $\eta_{\rm acc}$ (Ohsuga et al. 2005; Jiang et al. 2014). Consequently, the SMBH can experience rapid mass growth while radiating only modestly above the Eddington luminosity.

In this work, the adopted functional form of $\epsilon(\eta_{\rm acc})$ is shown in Fig. 2 (bottom panel). It reproduces the two main regimes outlined above: (i) a constant radiative efficiency at low and moderate accretion rates, in agreement with thin-disk theory, and (ii) a capped luminosity at super-Eddington rates, consistent with slim-disk prescriptions. We follow the parametrization of Madau et al. (2014), which empirically fits the results of general relativistic radiation-hydrodynamic simulations (Sądowski et al. 2014) as a function of black hole spin. Since our model does not track spin evolution, we adopt the curve corresponding to a fixed dimensionless spin parameter $a = 0.67$, yielding a constant sub-Eddington efficiency of $\epsilon_0 \approx 0.1$. This choice provides a smooth and physically motivated transition between the sub- and super-Eddington regimes, while preserving the correct asymptotic limits in both.

Because the time interval between consecutive simulation snapshots is typically much longer than the characteristic variability timescales of SMBH accretion, assuming a fixed value of $\eta_{\rm acc}$ across an entire snapshot interval would be a poor approximation. Instead, we decide to "subcycle" the integration of each SMBH's mass history by introducing a shorter timescale, $\tau_{\rm coherence}$, which we interpret as the coherence timescale of the accretion process. Over each interval of length $\tau_{\rm coherence}$, the accretion rate is held constant; at the end of the interval, we draw a new, independent value of $\eta_{\rm acc}$ from the conditional sBHAR distribution.

In reality, SMBH accretion is a stochastic process spanning a wide hierarchy of variability timescales, from days to hundreds of million years. By adopting $\tau_{\rm coherence}$ as an effective coherence timescale, we approximate this stochasticity in a computationally tractable way, while preserving the statistical properties of the underlying accretion distribution. Further discussion on the role of $\tau_{\rm coherence}$ in the growth of SMBHs can be found in Sec. 3.3.2.

Between two consecutive simulation snapshots, i and $i+1$, we therefore draw

$$N = \left\lfloor\frac{t_{i+1}-t_i}{\tau_{\rm coherence}}\right\rfloor \quad (11)$$

independent values $\eta_{{\rm acc},j}$, and compute the SMBH mass at t_{i+1} as:

$$M_{\rm BH}(t_{i+1}) = M_{\rm BH}(t_i)\,\exp\left[\tau_{\rm coherence}\sum_{j=1}^{N} t_{\rm acc}^{-1}(\eta_{{\rm acc},j})\right], \quad (12)$$

where $t_{\rm acc}(\eta_{{\rm acc},j})$ is the Salpeter timescale corresponding to the j-th sampled accretion rate.

At the same time, the sampled accretion rates determine the radiative output of quasars. In practice, we assign the bolometric luminosity of each SMBH at snapshot i using the most recent sampled accretion rate, $\eta_{{\rm acc},i}$, via:

$$L_{\rm bol}(t_i) = \epsilon(\eta_{{\rm acc},i})\,\eta_{{\rm acc},i}\,\dot{M}_{\rm Edd}\,c^2. \quad (13)$$

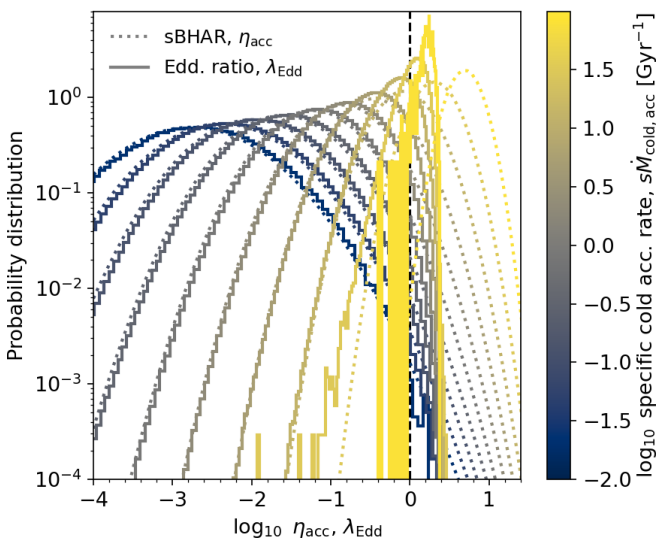

Figure 3. Probability distribution functions of the specific black hole accretion rate, $\eta_{\rm acc}$ (solid lines), and the Eddington ratio, $\lambda_{\rm Edd}$ (dotted lines), shown in bins of the specific cold accretion rate onto the halo. The $\eta_{\rm acc}$ distributions are drawn from the theoretical conditional probability $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ (Eq. 6), using the fiducial parameter values listed in Tab. 1. The $\lambda_{\rm Edd}$ distributions, in contrast, are measured directly from the model output. At sub-Eddington rates, they follow the same trend as $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ (with minor mismatches due to the non-uniform distribution of $s\dot{M}_{\rm cold,acc}$ within each bin), but they deviate markedly as $\lambda_{\rm Edd}$ approaches unity (dashed vertical line), reflecting the drop in radiative efficiency in this regime.

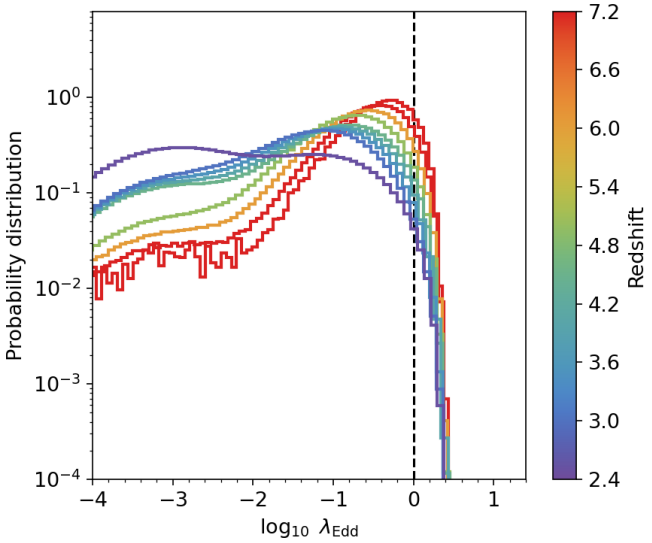

Figure 4. Probability distribution functions of the Eddington ratio, $\lambda_{\rm Edd}$, at different redshifts (solid colored lines) for the fiducial model described in Tab. 1. At high redshift, the majority of SMBHs accrete efficiently with $\lambda_{\rm Edd} \gtrsim 0.1$, whereas at later times accretion becomes progressively less efficient and the distribution shifts toward lower values. The dashed vertical line indicates the Eddington limit, $\lambda_{\rm Edd} = 1$.

This framework yields, for each subhalo in the simulation, a self-consistent prediction for the SMBH mass, bolometric luminosity, and Eddington ratio at every snapshot. These can then be directly compared to observational constraints across cosmic time, enabling a joint test of both SMBH growth and quasar demographics in our model.

2.3 Overview of the observational constraints

We consider three key observational constraints, all targeting the luminous quasar population ($L_{\rm bol} \gtrsim 10^{45}$ erg s^{-1}): (i) the bolometric quasar luminosity function (QLF), (ii) the large-scale clustering of quasars, and (iii) the conditional Eddington ratio distribution function (cERDF) at fixed quasar luminosity.

Our first constraint is the bolometric QLF, which provides the most direct observable for comparison with theoretical models of quasar evolution. Unlike single-band surveys – particularly rest-frame UV selections, which systematically miss obscured quasars – the bolometric QLF combines multi-wavelength AGN datasets to reconstruct the full quasar population. Modern compilations draw on X-ray, mid-infrared, UV-optical, and radio observations (e.g., Hopkins et al. 2007; Shen et al. 2020). X-ray data are especially valuable, as measurements of hydrogen column densities enable population-level obscuration corrections and yield intrinsic luminosities even for heavily absorbed sources (Ueda et al. 2014). By synthesizing these corrections across multiple bands, bolometric QLFs provide the most complete available census of SMBH accretion and thus a robust benchmark for testing models. In particular, using bolometric QLFs allows us to bypass the need for an explicit obscuration prescription in our framework. Likewise, the other two observables we consider – the clustering of quasars and the cERDF – remain unaffected by obscuration, provided that, at fixed luminosity, obscured and unobscured quasars represent a random subsample of the overall population.

Here, we adopt the bolometric QLF of Shen et al. (2020), which provides the most recent and comprehensive compilation of multi-wavelength data spanning $z \approx 0-6$. Since our focus is on luminous quasars ($L_{\rm bol} \gtrsim 10^{45}$ erg s^{-1}), the strongest constraints come from obscuration-corrected UV-optical samples such as those presented by Kulkarni et al. (2019), supplemented at high redshift by dedicated surveys (e.g., Matsuoka et al. 2018; Wang et al. 2019).

Although bolometric QLF reconstructions provide a substantially improved census of AGN activity and serve as a key benchmark for our model, significant uncertainties remain – particularly at high redshift, where both the fraction and physical nature of obscured quasars are still debated. In practice, X-ray-based obscuration corrections are possible only up to $z \lesssim 3$–4 and rely on uncertain extrapolations beyond this regime. Meanwhile, recent observations (e.g., Vito et al. 2018; Circosta et al. 2019; D'Amato et al. 2020; Gilli et al. 2022) and cosmological simulations (e.g., Ni et al. 2020; Vito et al. 2022; Bennett et al. 2024) suggest that quasars in the early Universe may be embedded in dense gas environments that drive high obscuration fractions, implying a rapid evolution of obscuration properties at $z \gtrsim 4$. If so, current bolometric QLF estimates at these redshifts likely underestimate the true space density of quasars, making them a conservative lower limit for comparison with theoretical models.

For the clustering of quasars, we rely on measurements of the two-point auto-correlation function from large spectroscopic surveys, which provide the most direct probe of quasar environments on cosmological scales. Numerous studies have characterized quasar clustering across a broad redshift range (e.g., Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007; da Ângela et al. 2008; Ross et al. 2009; White et al. 2012; Eftekharzadeh et al. 2015), consistently finding that luminous quasars typically reside in dark matter halos of mass $\sim 10^{12}$–10^{13} M$_\odot$. In this work, which focuses on cosmic noon and earlier epochs, we adopt three

key datasets: (i) the high-precision clustering constraints from the BOSS survey at cosmic noon ($z \approx 2.5$, Eftekharzadeh et al. 2015); (ii) the strong clustering of quasars at $z \approx 4$ reported by Shen et al. (2007) using SDSS measurements; and (iii) the recent quasar–galaxy cross-correlation measurements at $z \approx 6$ from the EIGER JWST survey (Eilers et al. 2024; Pizzati et al. 2024b). We do not include the auto-correlation function of faint quasars at $z \approx 6$ presented by Arita et al. (2023) because, as shown in Pizzati et al. (2024a, see their Appendix D), these data are not sufficiently constraining.

Each of the clustering measurements we consider applies a luminosity threshold, selecting quasars brighter than a given $L_{\rm bol}$. Since quasar clustering may depend on luminosity, it is essential to adopt consistent thresholds when comparing our model predictions to the data. Accordingly, we impose the same luminosity cuts as in the observations, denoted $L_{\rm bol,thr}$. For the Eftekharzadeh et al. (2015) and Shen et al. (2007) measurements, we follow the thresholds used in the analysis of Pizzati et al. (2024a, see their Sec. 3.1): these are $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 46.1$ and $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 46.7$, respectively. For the EIGER measurements of Eilers et al. (2024), we adopt the same threshold as in Pizzati et al. (2024b): $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 47.1$.

Finally, the cERDF is defined as the probability distribution of Eddington ratios at fixed bolometric luminosity, $P(\lambda_{\rm Edd} \,|\, L_{\rm bol})$. We use this quantity because it provides direct constraints on SMBH properties through broad-line measurements, while being robust to survey incompleteness. At fixed luminosity, the distribution of black hole mass estimates – and thus of $\lambda_{\rm Edd}$ – is determined primarily by the widths of broad emission lines in quasar spectra. Since these line widths are largely unaffected by survey flux limits, the cERDF is considerably less sensitive to selection effects than other diagnostics such as the total ERDF or the black hole mass function (BHMF).

We estimate the cEDRF using the SDSS quasar compilation in Wu & Shen (2022), which covers the redshift range $z \approx 0-6$. To extend the high-redshift coverage, we also incorporate the compilation of Fan et al. (2023), which includes all quasars known at $z > 5.9$ at the time of publication. In constructing the cERDF, we restrict the sample to sources with reliable SMBH mass estimates, and compute the Eddington ratio of each quasar as $\lambda_{\rm Edd} = L_{\rm bol}/L_{\rm Edd}$. Bolometric luminosities are derived from UV/optical magnitudes using the bolometric correction of Richards et al. (2006).

2.4 Fiducial model and parameter inference

The ultimate goal of our framework is to perform parameter inference and assess the predictive power of current quasar observables in constraining models of SMBH evolution. We plan to do so by writing a joint likelihood function for our model parameters, $\Theta : (M_{\rm start}, \tau_{\rm coherence}, \eta_{\rm av,0}, \eta_{\rm av,evol}, \sigma_0, \sigma_{\rm evol})$. The likelihood will incorporate the independent constraints coming from the three observables described in Sec. 2.3 – the QLF, the quasar auto-/cross-correlation functions ("corr"), and the cERDF:

$$\mathcal{L}^{(\rm total)} = \mathcal{L}^{(\rm QLF)}\, \mathcal{L}^{(\rm corr)}\, \mathcal{L}^{(\rm cERDF)} \tag{14}$$

The first two likelihood terms have the same expression:

$$\mathcal{L}^{(k)}(\mathbf{d}^{(k)}\,|\Theta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{15}$$

with $k \in \{\rm QLF, corr\}$, $\mathbf{d}^{(k)}$ being for the set of n data points with means $\mathbf{y}$ and covariance Σ coming from observations, and $\boldsymbol{\mu}$ the set of values predicted by our models.

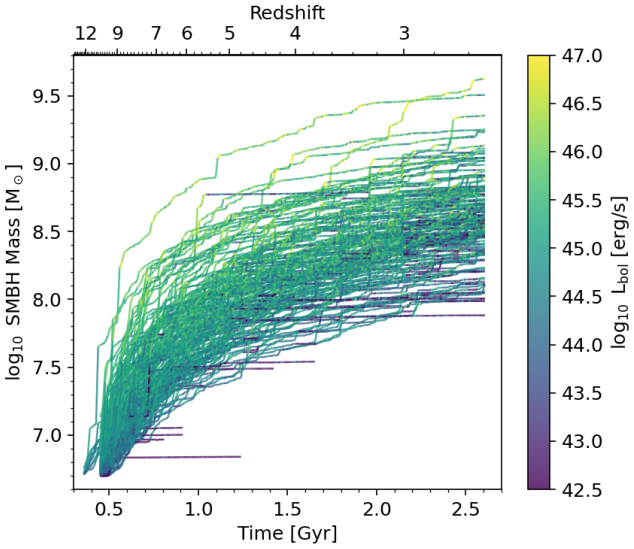

Figure 5. Evolution of black hole mass as a function of cosmic time (bottom axis) and redshift (top axis) for a subsample of 150 objects selected from the first 1000 halos to form in the simulation. Each track is color-coded by the instantaneous bolometric luminosity of the corresponding quasar. Tracks that terminate abruptly correspond to halos that either merge with more massive systems or are lost in the merger tree (see Sec. 2.1.2).

For the third likelihood term, we directly compare the two-dimensional distribution in the bolometric luminosity-Eddington ratio plane predicted by our model with that measured from observations. Let $\mathbf{d}^{(\rm cERDF)} = \{(\lambda_{{\rm Edd},i}, L_{{\rm bol},i})\}$ denote the set of Eddington ratios and bolometric luminosities measured observationally, and let $P(\lambda_{\rm Edd}, L_{\rm bol})$ represent the corresponding probability distribution predicted by the model. The likelihood can then be written as:

$$\mathcal{L}^{(\rm cERDF)}(\mathbf{d}^{(\rm cERDF)}\,|\Theta) = \prod_i P(\lambda_{{\rm Edd},i}, L_{{\rm bol},i}) \tag{16}$$

While this inference roadmap is straightforward in principle, it is challenging to implement in practice. Performing inference requires evaluating the model many thousands of times across parameter space, which is computationally prohibitive: depending on the configuration, a single run of our model can take from several minutes to one hour. This makes direct inference with methods such as Markov Chain Montecarlo (MCMC) unfeasible.

To overcome this limitation, our long-term strategy is to construct a statistical emulator for the model observables, capable of reproducing the output of the full model at negligible computational cost. Such an emulator will enable efficient MCMC exploration of parameter space and a rigorous inference of posterior distributions. Development of this emulator is ongoing and will be included in future updates of this manuscript.

In the present study, we instead adopt a simpler approach: we fix the model parameters to fiducial values, chosen through a combination of trial-and-error exploration and preliminary experiments with the emulator. A systematic parameter inference study, together with an optimized parametrization of the accretion and merger processes, is deferred to forthcoming work.

The fiducial values used for all model parameters are listed in Table 1. The seed black hole mass is set to $M_{\rm start} = 5 \times 10^6\,{\rm M}_\odot$: given the resolution of the simulation, this corresponds to assigning each newly formed halo a black hole initially $\sim 5 \times 10^4$ times less massive than its host, broadly consistent with local SMBH–halo

scaling relations. The coherence timescale of accretion, $\tau_{\rm coherence}$, is set to 10 Myr. This value ensures that individual high-accretion episodes are sufficiently long-lived to contribute meaningfully to SMBH growth, while remaining short compared to cosmological timescales. A more detailed discussion on $\tau_{\rm coherence}$ and our prior knowledge on the duration of SMBH accretion episodes is deferred to Sec. 3.3.2.

The ERDF parameters are set to the following fiducial values: $\eta_{\rm av,0} = -1.12, \eta_{\rm av,evol} = 0.95, \sigma_0 = 0.52, \sigma_{\rm evol} = -0.17$. This parametrization produces the conditional sBHAR distributions shown as dotted lines in Fig. 3. At high cold halo accretion rates ($\log_{10} s\dot{M}_{\rm cold,acc}/{\rm Gyr}^{-1} \gtrsim 0.5$), the sBHAR distribution becomes relatively narrow and centered at values close to or above the Eddington limit. As cold accretion onto halos declines, the distribution broadens and peaks at much lower sBHAR values.

The corresponding conditional distributions of the observed Eddington ratio, $\lambda_{\rm Edd} = L_{\rm bol}/L_{\rm Edd}$, are shown with solid lines. Importantly, $\lambda_{\rm Edd}$ coincides with $\eta_{\rm acc}$ only if the radiative efficiency is constant. In our model, where $\epsilon(\eta_{\rm acc})$ varies with accretion rate, $\lambda_{\rm Edd}$ no longer traces the intrinsic growth rate directly. This is evident in Fig. 3: while the $\lambda_{\rm Edd}$ distributions follow the sBHAR distributions at sub-Eddington rates (where ϵ is roughly constant), they remain sharply peaked near the Eddington limit even when the intrinsic sBHAR far exceeds unity. Physically, this reflects the saturation of radiative output in the super-Eddington regime, where SMBHs can accrete mass at highly efficient rates without producing proportionally higher luminosities. This effect has been invoked to explain the rapid assembly of massive black holes in the early Universe (e.g., Madau et al. 2014; Volonteri et al. 2015), and plays a central role also in our framework (see Sec. 3.1).

Figure 4 shows the distributions of observed Eddington ratios, binned by redshift. The model predicts a pronounced redshift evolution in SMBH accretion properties, driven primarily by the changing halo accretion rates over cosmic time (Fig. 1, bottom). At the epoch of reionization ($z \gtrsim 6$), the majority of SMBHs are actively accreting at $\lambda_{\rm Edd} \gtrsim 0.1$, consistent with rapid and efficient growth. By cosmic noon ($z \sim 2$), however, only a small fraction of SMBHs remain in the high-accretion regime, reflecting the global decline in gas accretion rates and SMBH activity. We note that very low Eddington ratios ($\lambda_{\rm Edd} \lesssim 0.01$–$0.1$) are only useful for theoretical modeling as they are effectively unobservable in current quasar surveys. Consequently, the fraction of SMBHs accreting below this threshold defines an effective duty cycle of quasar activity in our framework. This duty cycle evolves rapidly with redshift, in agreement with expectations from measurements of the QLF and quasar clustering (e.g., Martini & Weinberg 2001; Haiman & Hui 2001; Pizzati et al. 2024b).

3 RESULTS

We now turn to the results of the fiducial run introduced above. We begin by examining how SMBHs assemble their mass in our framework, focusing on the predicted accretion histories and comparing their radiative output to quasar observables. We then explore the broader implications of these results for the growth of SMBHs across cosmic history, with particular attention to the connection between black holes and the properties of their host halos.

3.1 The buildup of supermassive black holes across cosmic history

Figure 5 illustrates the mass assembly of SMBHs for a subsample of 150 objects, selected from the first 1000 halos formed in the simulation (i.e., halos that form at $z > 10$). The growth of these early black holes is initially rapid: within the first billion years, SMBH masses increase from the seed value $M_{\rm start}$ up to $M_{\rm BH} \approx 10^9\,{\rm M}_\odot$ by $z \approx 6$. This phase of accelerated growth coincides with sustained episodes of luminous quasar activity, with bolometric luminosities reaching $L_{\rm bol} \gtrsim 10^{47}\,{\rm erg\,s^{-1}}$. At lower redshifts, however, the buildup of SMBHs slows significantly. The vigorous accretion that characterizes the high-z regime – and enables the emergence of the first bright quasars – gradually gives way to a more quiescent growth pattern. By cosmic noon, quasars are powered less by rapid accretion and more by the sheer mass of their SMBHs: even relatively modest accretion rates can produce large luminosities once $M_{\rm BH}$ exceeds 10^8–$10^9\,{\rm M}_\odot$.

Overall, the bulk of the early-forming population grows from $M_{\rm start}$ at cosmic dawn to $M_{\rm BH} \approx 10^8$–$10^9\,{\rm M}_\odot$ by $z \sim 2$. Only a small subset of outliers – i.e., SMBHs experiencing unusually efficient or repeated accretion episodes – are able to reach the extreme SMBH masses associated with the brightest quasars observed across cosmic time. Tracking these rare growth histories, rather than focusing solely on population averages, is therefore essential for understanding the origin of luminous quasars.

These trends can be examined in more detail by following individual SMBH and quasar luminosity histories, as shown in Fig. 6. Here we select six halos that formed in the second snapshot of the simulation ($z = 12.26$) and plot their SMBH mass evolution (red) alongside the corresponding host halo mass (green, scaled down by 10^5). The associated quasar luminosity histories are shown in blue. Red and blue curves intersect when quasars radiate at the Eddington limit, while blue lines above the red indicate super-Eddington radiative phases. For reference, the light gray curves represent idealized cases of continuous SMBH growth from $M_{\rm start}$ at the Eddington limit and at $0.1\,\dot{M}_{\rm Edd}$. The smaller panels further show the accretion histories in terms of the sBHAR $\eta_{\rm acc}$ (purple), compared against the population median and 16th-84th percentiles (gray dotted line and shaded region) across cosmic time.

Across all examples, the same qualitative pattern emerges: rapid SMBH growth at $z \gtrsim 6$, followed by a marked slowdown at later times. This behavior is driven directly by the evolving distribution of sBHARs. At early times, SMBHs accrete at rates close to the Eddington limit; by lower redshifts, the distribution broadens and shifts toward low values ($\eta_{\rm acc} \lesssim 1\%$), consistent with the decline in cosmic gas supply. In most cases, SMBH growth closely tracks that of the host halo (green lines). However, stochastic variations in the accretion rate lead to significant departures in some cases.

Crucially, short bursts of super-critical accretion ($\eta_{\rm acc} > 1$) play a decisive role in driving SMBHs to higher masses on short timescales. For instance, the bottom-left panel of Fig. 6 shows an SMBH reaching $\sim 10^9\,{\rm M}_\odot$ at $z \approx 6$ through repeated episodes of super-critical accretion. Importantly, because the radiative efficiency ϵ drops steeply in the super-critical regime (Fig.2, bottom), very high accretion rates do not translate into equally high radiative output. Even when $\eta_{\rm acc} \gg 1$, the corresponding Eddington ratio saturates only modestly above unity (Fig. 3). This feature allows SMBHs in our model to gain mass rapidly at high redshift through brief high-accretion bursts ($1 \lesssim \eta_{\rm acc} \lesssim 10$), while remaining consistent with the empirical fact that strongly super-Eddington quasars are not observed at any epoch. In Sec. 3.2, we quantify this comparison by confronting our predicted Eddington ratio distributions with observations.

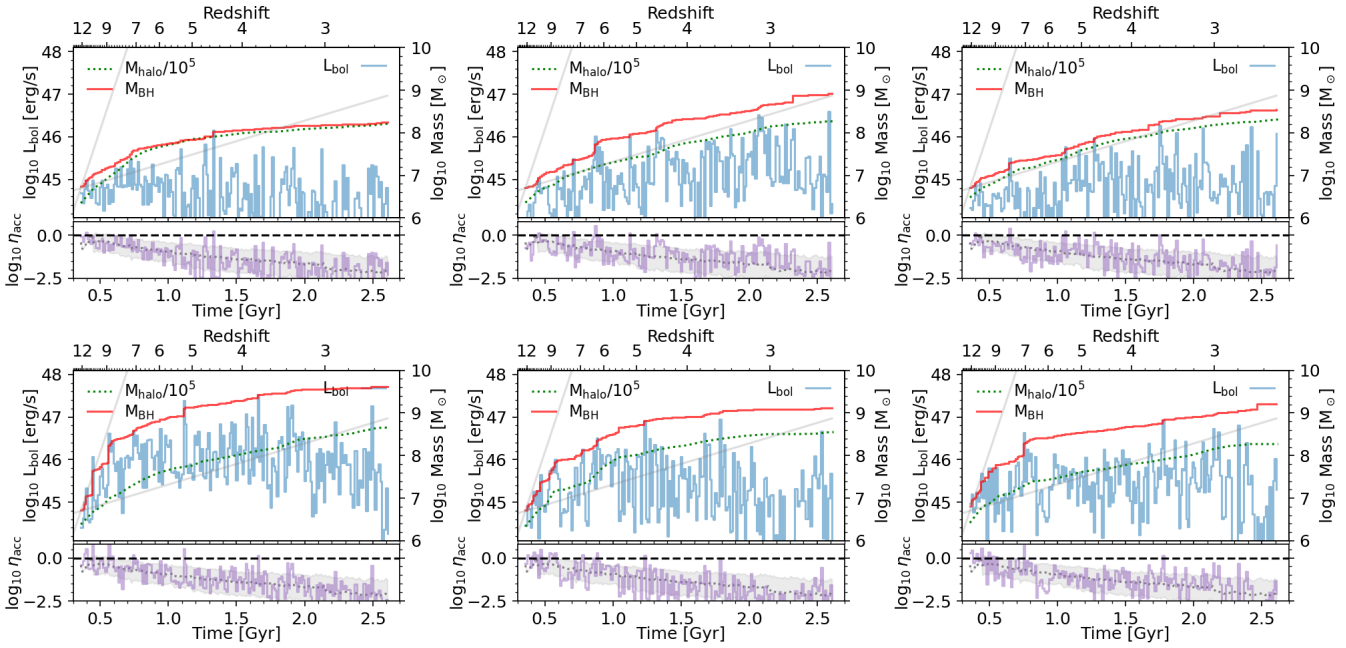

Figure 6. Individual accretion histories of SMBHs and their quasar light curves as a function of cosmic time (bottom axis) and redshift (top axis). We select six halos that form in the second snapshot of the simulation ($z = 12.26$). The SMBH mass is shown in red, alongside the host halo mass scaled down by 10^5 (green dotted). For reference, the steep and shallow gray lines indicate idealized SMBH growth tracks at constant accretion rates of $\eta_{acc} = 1$ (Eddington rate) and $\eta_{acc} = 0.1$, respectively. The quasar bolometric luminosity is plotted in blue; the blue and red curves intersect when quasars radiate at the Eddington limit. The bottom panels show the evolution of the specific black hole accretion rate, η_{acc}, for each object (purple), compared to the population median (dotted gray line) and 16th–84th percentile range (shaded gray region).

Interestingly, even with such strong accretion bursts, SMBH growth never systematically exceeds the analytic Eddington-limited growth curve (light gray). This highlights a central result of our model: sustained, uninterrupted Eddington-limited accretion is not a viable growth pathway. Instead, SMBHs grow through a stochastic sequence of accretion episodes – alternating between super-critical bursts and long periods of less efficient, sub-Eddington accretion. This produces growth tracks that may cluster around, but rarely exceed, the simple Eddington-limited scenario, while still enabling a subset of black holes to assemble the extreme masses required to power luminous quasars at all epochs.

3.2 Comparison with quasar observables

In Fig. 7, we compare the predictions of our fiducial model (solid lines) to the bolometric QLF of Shen et al. (2020, see Sec. 2.3). Each panel shows a different redshift slice from $z = 2$ to $z = 6$, while the bottom-right panel combines all redshifts to highlight the overall evolutionary trend.

We find that our model reproduces the observed QLF very well for luminous quasars ($L_{bol} \gtrsim 10^{45.5}$ erg s^{-1}), which constitute the primary targets of UV-optical surveys. The large cosmological volume of FLAMINGO allows us to follow the quasar population up to extreme luminosities of $L_{bol} \gtrsim 10^{48}$ erg s^{-1}, and down to space densities as low as $n \approx 10^{-9}$ cMpc^{-3}. This statistical power makes it possible to robustly probe the rarest, most luminous quasars, and we find that the bright-end slope of the QLF is accurately reproduced across all redshifts considered.

At the faint end, however, discrepancies arise. Observationally, the QLF flattens significantly toward low luminosities, whereas our model predicts a steeper continuation and thus an excess population of faint quasars/AGN, particularly at high redshift. This tension is not unique to our framework: many theoretical models have long struggled to reproduce the faint-end behavior of the QLF from first principles (e.g., Degraf et al. 2010).

It is important to emphasize that the faint end of the QLF is itself highly uncertain observationally. In UV-optical surveys, the QLF can be probed down to $L_{bol} \approx 10^{45}$ erg s^{-1}, but at these luminosities quasars become increasingly difficult to distinguish from UV-bright galaxies, and completeness corrections are non-trivial. X-ray surveys extend coverage to lower luminosities, $L_{bol} \sim 10^{43}$–10^{44} erg s^{-1}, but rely heavily on uncertain photometric redshifts and rarely extend beyond $z \gtrsim 3$. Moreover, the level of obscuration at low luminosities and high redshifts is still poorly constrained: if the obscured fraction is higher than currently assumed (e.g., Ueda et al. 2014), present-day bolometric QLF estimates may underestimate the true space density of faint quasars.

The advent of JWST has opened a new observational window onto the faint end of the quasar population. Early studies identifying broad-line AGN through rest-frame optical diagnostics suggest the presence of a more numerous faint population than previously inferred, with bolometric luminosities of $L_{bol} \sim 10^{43} - 10^{45}$ erg s^{-1} and black hole masses of $M_{BH} \sim 10^6 - 10^8$ M$_\odot$ (e.g., Harikane et al. 2023; Maiolino et al. 2024; Juodžbalis et al. 2025). Strikingly, our model naturally predicts such a population: low-mass SMBHs that were largely invisible to pre-JWST surveys make a substantial contribution to the faint end of the QLF. This is illustrated in Fig. 7 by comparing the full QLF (solid lines) with that obtained by restricting to SMBHs with $M_{BH} > 10^8$ M$_\odot$ (dashed lines). The strong divergence of the two curves at $L_{bol} \lesssim 10^{45}$ erg s^{-1} highlights the predicted dominance of small SMBHs in this regime – consistent with the emerging JWST results. However, current JWST-based estimates remain highly

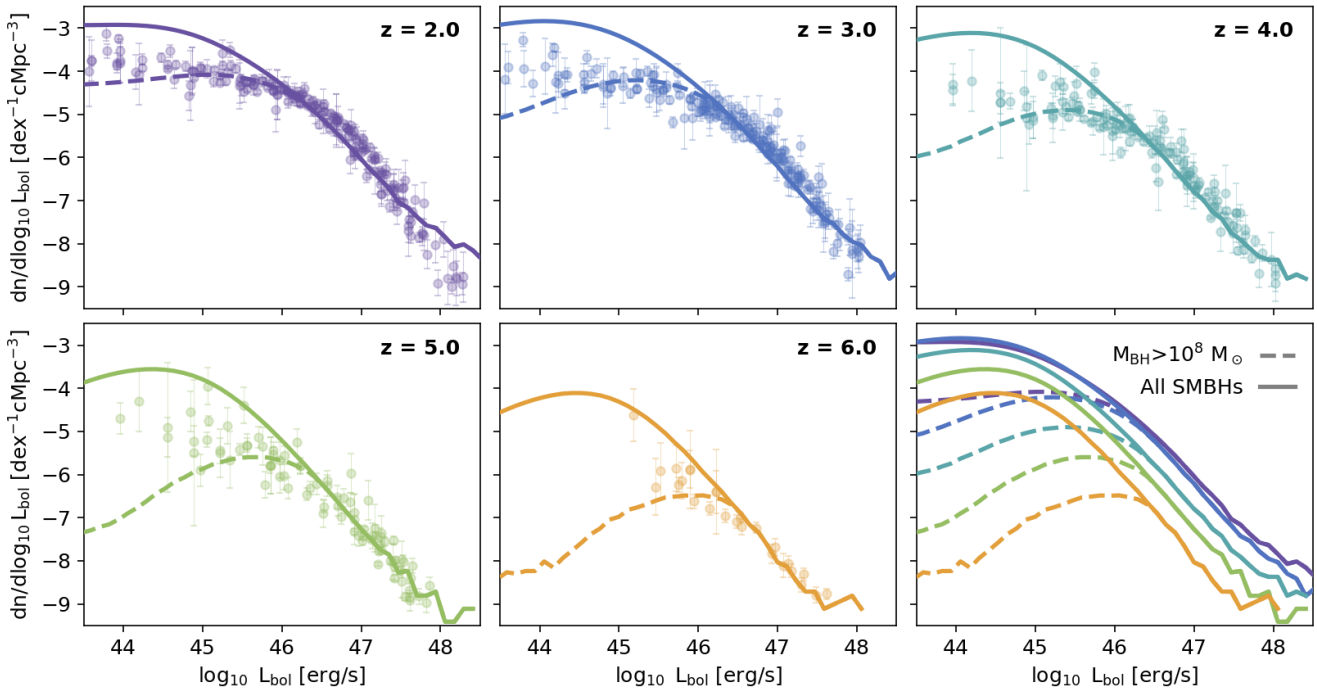

Figure 7. Comparison between the bolometric quasar luminosity function (QLF) predicted by our model and the compilation of Shen et al. (2020). Each panel corresponds to a different redshift, from $z = 2.0$ to $z = 6.0$, while the bottom-right panel combines all redshifts to illustrate evolutionary trends. Model predictions for the full SMBH population are shown as solid colored lines, and those restricted to $M_{BH} > 10^8$ $M_\odot$ as dashed lines. Observational data – drawn from X-ray, UV–optical, infrared, and radio surveys, and corrected for obscuration using wavelength-dependent bolometric corrections – are shown as colored points with error bars.

uncertain, and the physical nature and demographics of these faint sources are still the subject of active debate.

If the faint QLF measurements in Fig. 7 are correct, then the faint-end excess predicted by our model reflects a genuine shortcoming of the framework. In this scenario, the model reproduces the QLF data across luminosities reasonably well only when restricted to $M_{BH} > 10^8$ $M_\odot$ SMBHs (dashed lines). However, the additional population of lower-mass black holes accreting at or above the Eddington limit reaches luminosities comparable to faint quasars, thereby altering the predicted shape of the QLF. If such a population does not exist in reality, this would imply that SMBHs in our model are, on average, accreting – and thus radiating – more efficiently than is observed. This discrepancy could signal missing physical processes that regulate accretion in low-mass systems, such as stronger radiative or mechanical feedback, limited gas supply due to inefficient inflows, or environmental effects that suppress sustained high-Eddington accretion. More complex parametrizations are needed to account for these effects in the context of our framework.

A discrepancy between our model predictions for SMBH accretion and observations becomes indeed apparent when examining the cERDF in Fig. 8, where we show the distribution of Eddington ratios in bins of bolometric luminosity. In practice, the two-dimensional distribution $P(\lambda_{Edd}, L_{bol})$ is projected into a set of one-dimensional distributions by slicing along narrow L_{bol} bins. Comparing the observed distributions (histograms) with those predicted by the model (thin solid lines) reveals a systematic offset: the model consistently produces Eddington ratio distributions skewed toward higher values relative to the data. The difference is modest – the peaks of the two distributions typically agree within 1σ – but its persistence across all redshifts and luminosities suggests that it reflects a genuine limitation of the model rather than statistical noise.

This systematic bias toward higher λ_{Edd} is also what drives the large population of lower-mass SMBHs in our model to reach bolometric luminosities comparable to those of faint quasars. Reducing the average accretion rates would suppress the number of such faint sources and improve agreement with observations, but it would also prevent SMBHs from growing rapidly enough to reach the extreme masses required to power the brightest quasars. This underscores a fundamental challenge: reconciling the high accretion rates seemingly necessary to assemble billion-solar-mass black holes at early times with the empirical evidence that most observed quasars radiate at Eddington ratios around or below unity. Although our prescription for radiative efficiency in the super-critical regime (Madau et al. 2014) suppresses the luminosities of rapidly accreting quasars, limiting them to radiate only modestly above the Eddington limit (Fig. 3), we conclude that this effect alone does not fully reconcile the tension between modest observed Eddington ratios and the growth rates required for SMBH assembly. Resolving this discrepancy will require further work to test whether refined parameter choices, alternative radiative efficiency prescriptions, or more flexible accretion models can provide a better match.

Despite this offset, our fiducial model successfully reproduces the main behavior of the cERDF. In particular, it captures the observed trend that the cERDF peaks at higher λ_{Edd} with both increasing redshift and increasing bolometric luminosity. This agreement suggests that while the model slightly overestimates accretion efficiencies, it nonetheless recovers the key qualitative features of SMBH growth

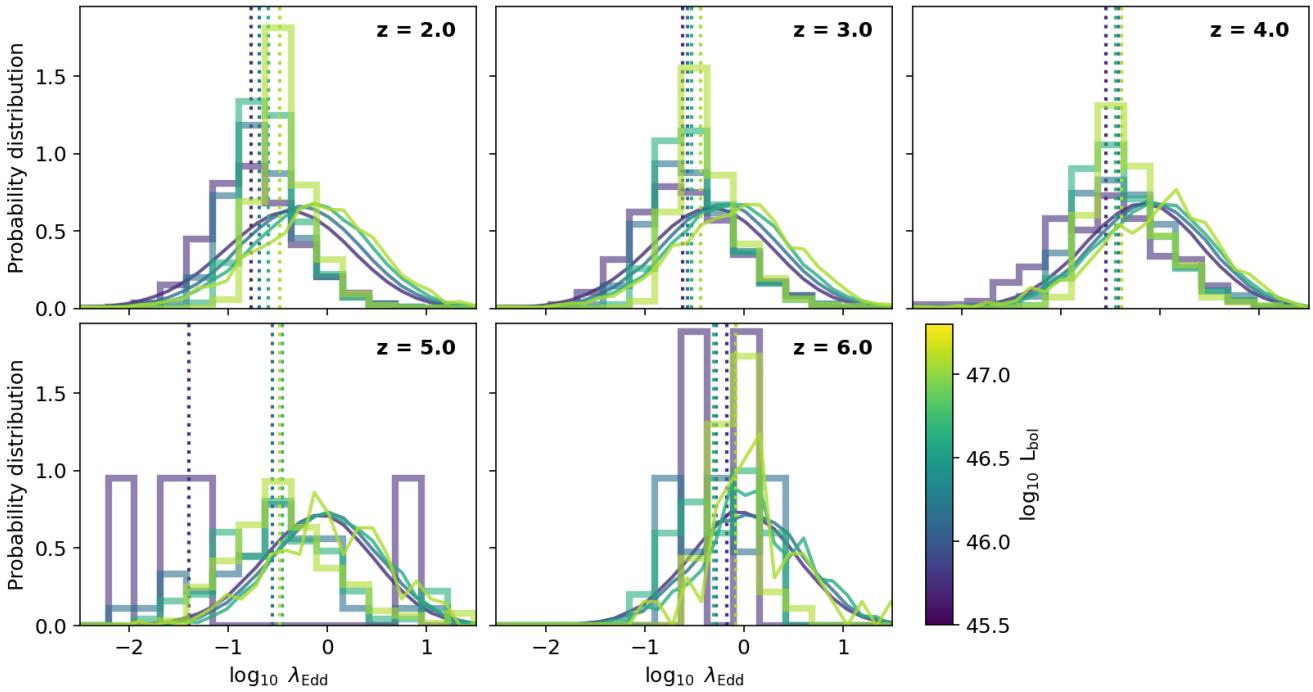

Figure 8. Comparison between the binned conditional Eddington ratio distribution functions (cERDFs) predicted by our model and those derived from SMBH mass and quasar luminosity compilations of Wu & Shen (2022, $z \approx 2$–5) and Fan et al. (2023, $z \gtrsim 6$). Observed distributions are shown as histograms, color-coded by bolometric luminosity, with dotted vertical lines marking their medians. Predictions from the fiducial model are shown as thin solid lines. While the model broadly reproduces the observed trends with both redshift and luminosity, it systematically predicts higher Eddington ratios than observed.

across cosmic time and luminosity, making it a solid foundation for future refinements.

The last observable we consider is the clustering of quasars. As discussed in Sec. 2.3, we include measurements of the quasar auto-correlation function at $z \approx 2.5$ and $z \approx 4$, as well as the quasar–galaxy cross-correlation function at $z \approx 6$. In principle, the auto-correlation could be computed directly from the quasars in our model above a chosen luminosity threshold, $L_{\rm bol,thr}$. However, this approach is not feasible for the quasar-galaxy cross-correlation, since the FLAMINGO run used here does not resolve the majority of galaxy-hosting halos[1]. To ensure consistency across all redshifts, we instead adopt the framework developed by Pizzati et al. (2024a,b), and compute clustering predictions from the quasar host mass function (QHMF)[2].

The QHMF is defined as the halo mass distribution of quasars brighter than $L_{\rm bol,thr}$. Given a QHMF, the clustering can be predicted under the assumption that halo mass alone determines the bias – i.e., neglecting any assembly bias contributions (e.g., Wechsler et al. 2006). This assumption is reasonable in our context, as assembly bias is expected to play a minor role, and current quasar clustering measurements are not yet precise enough to be dominated by such effects (e.g., Bonoli et al. 2010). Using the halo correlation fitting framework of Pizzati et al. (2024a), we compute the quasar auto-correlation functions at $z \approx 2.5$ and $z \approx 4$ given the QHMFs at the respective redshifts (see their Eqs. 7–8). For the quasar–galaxy cross-correlation at $z \approx 6$, we combine the QHMF with a galaxy host mass function (GHMF) following Eqs. 3–5 in Pizzati et al. (2024b). For the GHMF, based on [O III] -emitting galaxies, we adopt the measurements of Eilers et al. (2024), and approximate it with a simple cutoff form in which the HMF is set to zero below $\log_{10} M_{\rm min,gal}/{\rm M}_\odot = 10.56$.

Figure 9 shows the QHMF (solid lines) from our fiducial run, using a uniform luminosity threshold of $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 46.5$ at all redshifts. For comparison, the halo mass function (dashed line) is also shown, illustrating the fraction of halos active as quasars at a given mass. This fraction can be interpreted as the quasar duty cycle (e.g., Martini & Weinberg 2001; Haiman & Hui 2001; Pizzati et al. 2024a). At all redshifts, the QHMF peaks in the range $M_{\rm halo} \sim 10^{12}$–$10^{13}\,{\rm M}_\odot$, in good agreement with general observational trends. As redshift decreases, the QHMF peak shifts to progressively higher halo masses, and the distribution broadens, reflecting the growing diversity of quasar host environments.

Finally, Fig. 10 compares the clustering predicted by our model with the observational measurements described in Sec. 2.3. To ensure consistency, we compute clustering from the QHMFs using the same luminosity thresholds as the data at each redshift. This differs slightly from the QHMFs shown in Fig. 9 (that are obtained with a uniform luminosity threshold), but in practice the impact is modest: in our framework, quasar clustering depends only weakly on luminosity, in agreement with observations that find a mild or negligible luminosity dependence (e.g., Adelberger & Steidel 2005; Porciani et al. 2004; Shen et al. 2009; Eftekharzadeh et al. 2015).

[1] By extending our framework to the larger FLAMINGO-10k simulation (Schaller et al., in prep.; Pizzati et al. 2024b), we will be able to compute the quasar-galaxy cross-correlation function directly from the simulation outputs. Work on this extension is currently underway.

[2] Pizzati et al. (2024a,b) also rely on the FLAMINGO suite of cosmological simulations. In their approach, the simulations are used to calibrate a fitting model that predicts the clustering of any halo population (both auto- and cross-correlations) directly from its mass distribution.

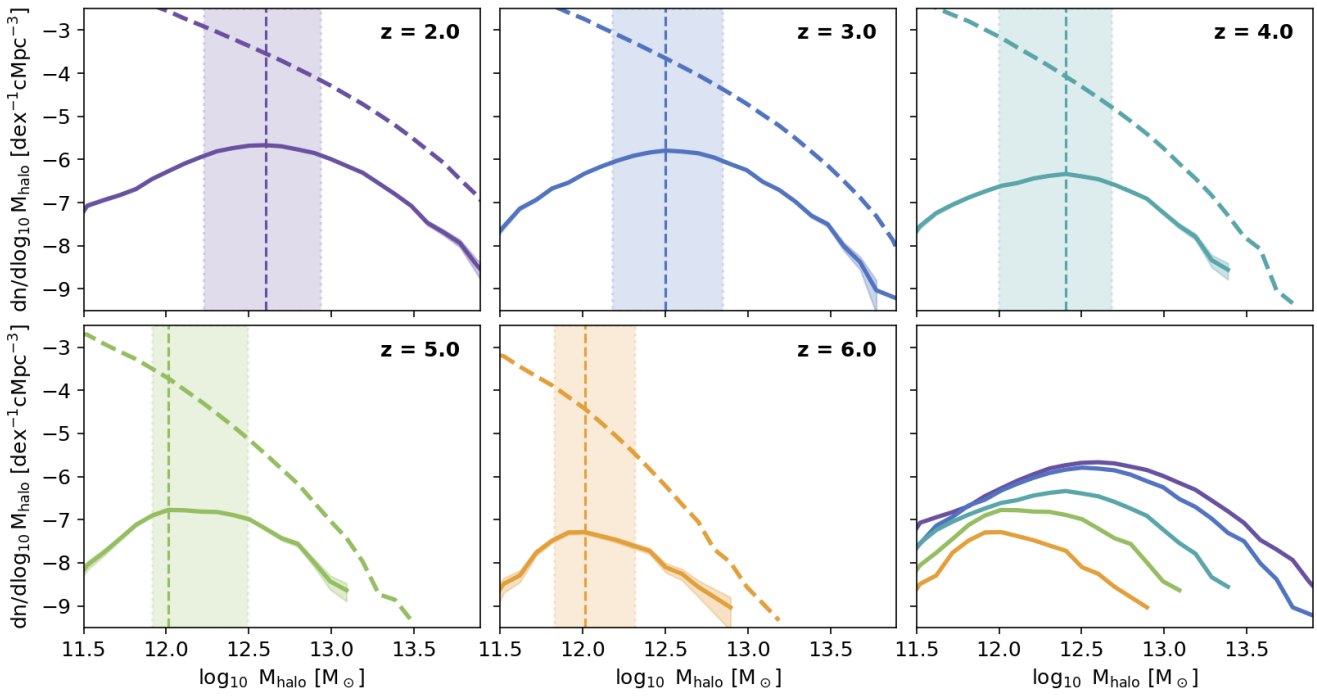

Figure 9. Quasar host mass functions (QHMFs) predicted by our fiducial model at different redshifts (solid colored lines). All QHMFs are computed using a uniform bolometric luminosity threshold of $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 46.5$. The bottom-right panel combines all redshifts to highlight evolutionary trends. For each distribution, vertical dashed lines mark the median, with shaded bands indicating the 16th–84th percentile range. The shaded envelopes around the solid lines represent Poisson uncertainties on the QHMF measurements. For comparison, the corresponding halo mass functions (HMFs) at each redshift are shown with dashed lines.

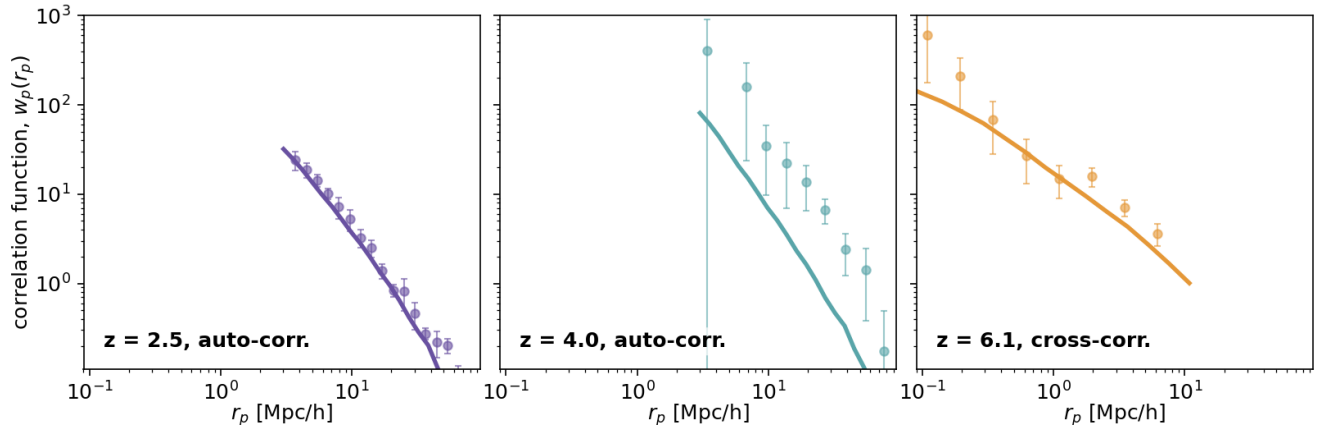

Figure 10. Comparison between quasar clustering predicted by our model and observational measurements at three redshifts: $z \approx 2.5$ (quasar auto-correlation function from Eftekharzadeh et al. 2015), $z \approx 4$ (quasar auto-correlation function from Shen et al. 2007), and $z \approx 6$ (quasar–galaxy cross-correlation function from Eilers et al. 2024). Model predictions are derived from the quasar host mass functions (QHMFs), adopting luminosity thresholds matched to each dataset, as described in Sec. 2.3.

At $z \approx 2.5$, our fiducial run reproduces very well the tight clustering constraints from the BOSS survey (Eftekharzadeh et al. 2015). Indeed, Pizzati et al. (2024a) derived the QHMF at this redshift by jointly fitting the QLF and clustering, finding a broad distribution peaking at $M_{\rm halo} \approx 10^{12.5}\,{\rm M}_\odot$, in excellent agreement with our model predictions.

At $z \approx 4$, by contrast, our model underpredicts the remarkably strong clustering measured by Shen et al. (2007) from SDSS quasars. Those data imply a very rapid evolution of quasar bias with redshift – a result that has long posed challenges for models of quasar and SMBH evolution (e.g., White et al. 2008; Shankar et al. 2010). More recent measurements at comparable (He et al. 2018; Timlin et al. 2018) and higher redshifts (Eilers et al. 2024; Pizzati et al. 2024b) indicate weaker clustering, casting doubt on the extreme values implied by Shen et al. (2007). It is therefore not surprising that our model – like many other empirical models and SAMs (e.g., Conroy & White 2013; Fanidakis et al. 2013) – predicts significantly lower clustering at $z \approx 4$. Matching the Shen et al. (2007) result would

require an extremely narrow QHMF, with virtually all luminous quasars confined to the most massive halos (Pizzati et al. 2024a). Such a scenario is difficult to reconcile with the intrinsic stochasticity that, in our model, drives extreme SMBH growth at early cosmic times (Sec. 3.3.1). The presence of stochasticity inevitably lowers the clustering, as it implies that even lower-mass halos can host massive, highly accreting SMBHs. This tension highlights the need for further work to establish whether the $z \approx 4$ clustering measurements can be reconciled with the broader set of quasar and SMBH constraints, and calls for new clustering analyses at similar redshifts – such as those anticipated from DESI (Yang et al. 2023) – which will be essential to determine whether the strong $z \approx 4$ signal reflects genuine quasar physics or arises from observational systematics.

At $z \approx 6$, the agreement improves again. The EIGER survey (Eilers et al. 2024) measures a clustering signal somewhat stronger than our prediction, consistent with a QHMF peaking near $\approx 10^{12.5}\,\mathrm{M}_\odot$ (Pizzati et al. 2024b). Our model predicts a slightly lower peak halo mass and weaker clustering, but the discrepancy is modest, and a refined choice of model parameters would likely bring the results into even closer agreement. Moreover, the EIGER result is based on only five quasar fields and is therefore highly sensitive to cosmic variance. In future iterations, we will compare against forthcoming results from the JWST ASPIRE survey (Wang et al. 2023, Wang et al. in prep.), which will provide clustering measurements from a much larger sample of 25 quasars. These data, which are largely consistent with EIGER but significantly more robust against cosmic variance (Huang et al. in prep.), will offer a more stringent benchmark for constraining our model at early cosmic times.

3.3 Implications for SMBH growth and scaling relations

The central question we set out to address in this work is straightforward: can a simple, physically-motivated model for black hole formation and evolution reproduce the diverse properties of bright quasars observed across cosmic time? While additional work is required to perform a full inference analysis and refine the model parameters, the results presented in Sec. 3.2 demonstrate that the answer is encouragingly positive. Our framework successfully captures the key observational benchmarks – the quasar luminosity function, the conditional Eddington ratio distribution, and quasar clustering – for a wide redshift range.

Building on this result, we now turn to the broader implications of the model. In particular, we examine how our framework informs the scaling relations between SMBHs and their host halos, and what it reveals about the physical processes that govern SMBH growth across cosmic history. By connecting the global quasar population to the detailed assembly histories of halos, our model provides a natural way to probe both the average evolutionary pathways and the stochastic variability that drive the emergence of the most massive black holes.

3.3.1 The black hole mass-halo mass relation across cosmic history

Figure 11 shows the black hole mass–halo mass ($M_{\rm BH}$–$M_{\rm halo}$) relation predicted by our fiducial model. Each panel displays the full distribution of SMBHs at a given redshift as a two-dimensional histogram in the $M_{\rm BH}$–$M_{\rm halo}$ plane (with logarithmic color scaling to highlight the tails), while points and error bars denote the median and 16th–84th percentiles of $M_{\rm BH}$ in bins of halo mass. For reference, a linear relation with normalization $M_{\rm BH}/M_{\rm halo} = 10^{-5}$ is shown as a dashed line.

It is important to stress that the relation shown here cannot be directly compared to observations. In reality, SMBH mass measurements are subject to systematic uncertainties of order ~ 0.5 dex, and only a biased subset of the population is accessible – either luminous quasars radiating above survey thresholds or, in the nearby Universe, the most massive black holes detectable through dynamical methods. By contrast, our model includes the entire SMBH population, independent of observability, and assumes perfect knowledge of their masses.

With these caveats in mind, the relation exhibits a clear, nearly linear trend with relatively small scatter ($\lesssim 0.3$ dex). The very tight distribution at low SMBH and halo masses primarily reflects our assumption of a fixed seeding mass, $M_{\rm start}$. At larger masses the scatter increases moderately, but the relation remains well-defined, a direct consequence of our prescription that ties SMBH growth to halo accretion. This coupling is also evident in Fig. 6, where the growth of individual SMBHs broadly parallels the assembly of their host halos.

The bottom-right panel of Fig. 11 highlights the redshift evolution of the median relation and its scatter. Overall, the evolution is modest: at $M_{\rm halo} \lesssim 10^{13}\,\mathrm{M}_\odot$, both the normalization and slope of the relation increase toward lower redshifts. At the high-mass end, by contrast, the relation shows a clear flattening, which becomes more apparent once massive halos emerge in significant numbers near cosmic noon. This flattening is a direct consequence of our cold-gas accretion prescription: once halos exceed $\sim 10^{12.5-13}\,\mathrm{M}_\odot$, the smaller cold gas accretion rates limit the ability of SMBHs to grow in lockstep with their hosts. As a result, the most massive halos host SMBHs that grow more slowly relative to halo mass assembly. This behavior mirrors the turnover observed in the stellar-to-halo mass relation (e.g., Behroozi et al. 2019), and underscores a common physical picture in which cooling inefficiencies in massive halos suppress baryonic growth across both galaxies and SMBHs.

Despite the relatively small scatter, most SMBHs in our model remain below $M_{\rm BH} \sim 10^{9},\mathrm{M}_\odot$ across all redshifts and halo mass bins, with the median relation at high redshift reaching only $M_{\rm BH} \sim 10^{8},\mathrm{M}_\odot$ even in the most massive halos. The billion-solar-mass SMBHs powering luminous quasars at early times are therefore not typical products of the mean relation, but instead arise as stochastic outliers in the accretion history distribution. Indeed, the rare objects with $M_{\rm BH} \gtrsim 10^{9}\,\mathrm{M}_\odot$ are found in a wide range of halo masses, indicating that their rapid growth is driven more by fluctuations in accretion than by steady halo mass assembly. This finding reinforces the importance of tracing individual SMBH growth trajectories – rather than relying solely on population averages – to capture the formation pathways of luminous quasars.

This intrinsic stochasticity also explains why our model struggles to reproduce the strong clustering signal measured at $z \approx 4$ by Shen et al. (2007, Sec. 3.2). Matching such strong clustering would require the massive SMBHs powering quasars to reside exclusively in the most massive halos – contrary to the broad range of environments predicted here. One could, in principle, reduce stochasticity by shifting the median relation upward (i.e., assuming more efficient accretion on average), or by allowing SMBHs in massive halos to continue accreting by relaxing the cold-gas suppression. However, both approaches would likely lead to an overproduction of extremely massive black holes at later times, in conflict with constraints in the local Universe. While low-redshift data are not explicitly included here, future work will explore whether low-z constraints, such as the local $M_{\rm BH}$–$M_{\rm halo}$ relation (e.g., Ferrarese & Merritt 2000), can help anchor the high-redshift regime and clarify which pathways of early

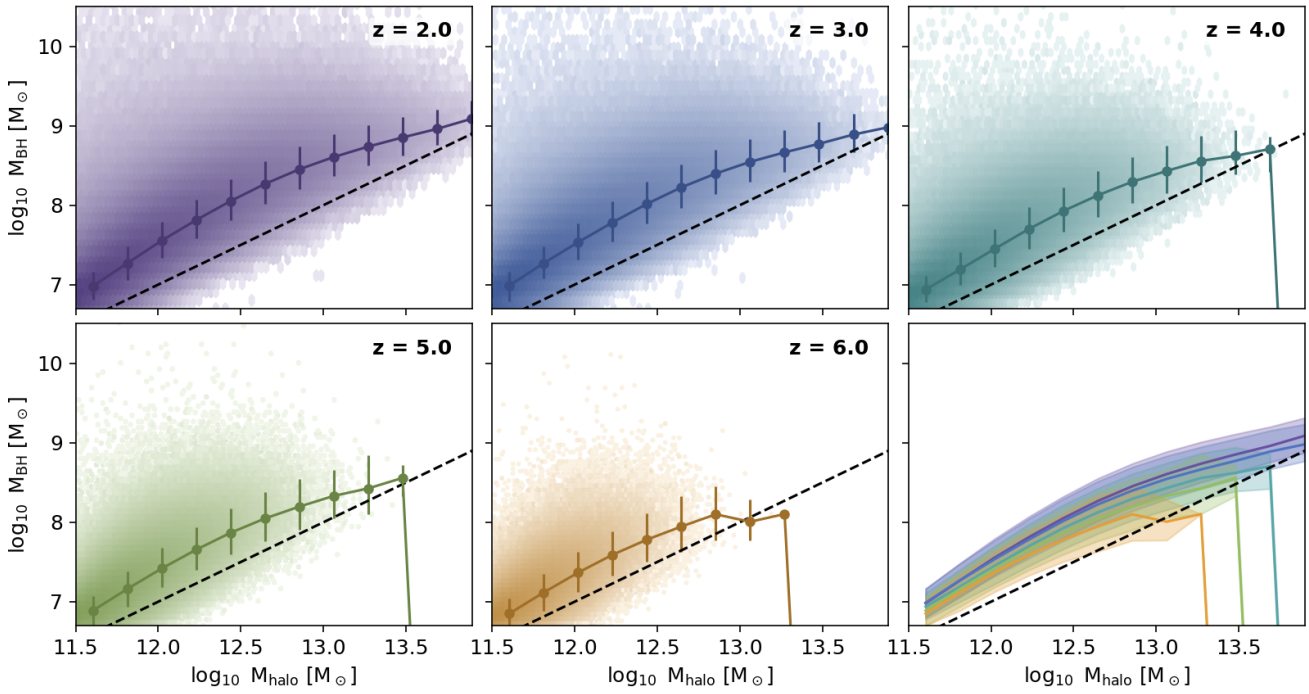

Figure 11. Black hole mass–halo mass relation predicted by our fiducial model at different redshifts. The two-dimensional distribution of SMBHs in the $M_{\rm BH}$–$M_{\rm halo}$ plane is shown with a logarithmic color scale. Median values of $M_{\rm BH}$ in bins of halo mass are plotted as points, with error bars indicating the 16th–84th percentile range. For reference, the dashed line in each panel marks a linear relation with normalization $M_{\rm BH}/M_{\rm halo} = 10^{-5}$. The bottom-right panel combines all redshifts, showing median relations (solid lines) and their scatter (shaded regions) to highlight evolutionary trends.

SMBH growth remain consistent with observational constraints at later cosmic time.

3.3.2 The coherence timescale of the accretion process

A key driver of stochasticity in our SMBH growth model is the coherence timescale of the accretion process, $\tau_{\rm coherence}$. This often-overlooked parameter sets the degree of temporal correlation in SMBH accretion. While the intrinsic shape of the accretion rate (sBHAR) distribution, $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$, specifies only the zeroth moment of the accretion stochastic process, $\tau_{\rm coherence}$ encodes its higher-order temporal structure, determining how fluctuations are sampled and accumulated over time.

If $\tau_{\rm coherence}$ is large, accretion bursts persist for extended periods; over the interval between two snapshots, the accretion history is then determined by only a few draws from the $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ distribution, yielding substantial object-to-object scatter in final SMBH masses. Conversely, a very small $\tau_{\rm coherence}$ yields many (approximately) independent draws in a fixed interval, so individual histories converge toward the mean behavior of the distribution. As Eq. 12 makes explicit, SMBH growth is governed by the sample mean of the $t_{\rm acc}^{-1}(\eta_{\rm acc})$ distribution. If $N \simeq \Delta t/\tau_{\rm coherence}$ is the effective number of independent draws over a snapshot interval Δt, then the standard deviation of the sample mean distribution scales as $N^{-1/2} \propto (\tau_{\rm coherence}/\Delta t)^{1/2}$ – directly linking larger $\tau_{\rm coherence}$ to larger variance in SMBH growth.

Consequently, $\tau_{\rm coherence}$ has clear population-level implications. A very short coherence timescale drives SMBHs with similar seed masses and formation times to follow nearly identical, smooth growth tracks, set primarily by their average halo accretion rate. In contrast, a longer $\tau_{\rm coherence}$ induces genuine diversity in growth paths even at fixed halo accretion rate, generating intrinsic mass scatter in addition to the stochasticity already encoded in the distribution $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$. The latter primarily governs the short-term variability seen in individual quasar light curves.

These effects are illustrated in Fig. 12, which shows the black hole mass function (BHMF) for the entire SMBH population at different redshifts. We compare two cases: our fiducial model with $\tau_{\rm coherence}$ = 10 Myr and an alternative run with $\tau_{\rm coherence}$ = 1 Myr, keeping all other parameters fixed (Tab. 1). The contrast is striking: the 1 Myr run produces a much narrower BHMF, making it substantially more difficult to grow the most massive SMBHs observed at all redshifts. As a consequence, the individual growth histories shown in Figs. 5 and 6 would appear far more uniform for $\tau_{\rm coherence}$ = 1 Myr, with significantly reduced diversity in accretion trajectories.

The $M_{\rm BH}$–$M_{\rm halo}$ relation discussed in Sec. 3.3.1 is likewise strongly influenced by $\tau_{\rm coherence}$. A shorter coherence timescale greatly reduces the scatter in this relation – forcing SMBH growth tracks to closely follow those of their host halos – and suppresses the stochastic outliers that, in our model, give rise to the brightest quasars across cosmic time. Following Eq. 12, and consistent with the central limit theorem, the distribution of SMBH masses at fixed halo mass approaches a narrow log-normal as $\tau_{\rm coherence}$ decreases. Conversely, a longer coherence timescale preserves extended high-mass tails in the distribution, enabling a subset of SMBHs to reach extreme masses and power the billion-solar-mass quasars observed in the early Universe.

Despite its importance, the coherence timescale of accretion is inevitably degenerate with other parameters that regulate SMBH growth. For instance, a shorter $\tau_{\rm coherence}$ could, in principle, be

offset by increasing the scatter in the accretion rate distribution $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ – though this freedom is limited, since the distribution is already anchored to the observed shape of the QLF – or by introducing additional variance through seeding or merger prescriptions. What makes the accretion timescale especially compelling, however, is that it can also be constrained through completely independent methods that probe quasar lifetimes and duty cycles (e.g., Martini 2004). For example, proximity-zone measurements in quasar spectra suggest that quasars must typically have been actively accreting for 10^4–10^7 years to produce the observed ionization structures around them (e.g., Eilers et al. 2017), setting a firm lower limit on the accretion timescale. Meanwhile, clustering-based duty cycle estimates provide complementary constraints by measuring how long quasars, on average, remain above a given luminosity threshold (Martini & Weinberg 2001; Haiman & Hui 2001). Taken together, these independent probes elevate $\tau_{\rm coherence}$ from a tunable modeling parameter to a physically interpretable quantity with broad observational implications.

Indeed, from a physical standpoint, $\tau_{\rm coherence}$ can be interpreted as the characteristic timescale of the processes that regulate quasar activity. These processes remain poorly constrained: it is still unclear whether most variability arises from rapid, small-scale fluctuations in accretion flows, or from longer-term, secular changes associated with galaxy and halo evolution, with short-term variability contributing only secondarily (e.g., Alexander et al. 2025). A more general framework than that developed here could, in principle, capture the full hierarchy of variability timescales by parametrizing the stochastic accretion process in terms of, e.g., its power spectral density, thereby quantifying the relative importance of different physical mechanisms. While developing such a framework lies beyond the scope of this work, it represents a promising avenue for future research. Ultimately, by combining the full suite of constraints – proximity zones and clustering-based estimates of quasar lifetimes and duty cycles, instantaneous accretion traced by the QLF, and long-term SMBH growth inferred from the cERDF and local SMBH mass measurements – it may become possible to phenomenologically uncover the processes that govern SMBH evolution across cosmic time.

3.3.3 *The relative role of mergers and accretion*

In Fig. 13, we examine the relative importance of mergers and gas accretion in driving SMBH growth. The solid lines show the BHMF from our fiducial run, where both accretion and mergers are included. The dashed lines represent a run where mergers are switched off by simply discarding merged SMBHs (i.e., black holes that are "sinked" or disrupted, see Sec. 2.1.2) without adding their mass to the remnant. The comparison reveals that mergers contribute only minimally across the entire redshift and mass range probed. At high redshift, the solid and dashed curves are indistinguishable, while at $z \lesssim 4$ a slight difference emerges around $M_{\rm BH} \approx 10^{8.5}$–$10^{9.5}\,{\rm M}_\odot$. These SMBHs likely reside in massive halos where cold-gas accretion has been suppressed; in such cases, mergers provide the only significant growth channel, producing the small offset. However, this difference is marginal and unlikely to affect any of the quasar observables considered here. This conclusion is broadly consistent with previous studies, which suggest that mergers become important only for the most massive SMBHs whose gas accretion has already been quenched (e.g., Shankar et al. 2009; Volonteri 2012; Pacucci & Loeb 2020). Extending our model to lower redshifts will allow us to probe this regime in more detail.

The dotted lines in Fig. 13 illustrate the BHMF when accretion is completely switched off and SMBHs grow only through mergers. In this scenario, SMBH masses never exceed $\sim 10^8\,{\rm M}_\odot$, even by cosmic noon. The BHMF evolves from being entirely seed-dominated at high redshift to gradually incorporating more growing SMBHs, reflecting the slow accumulation of mass as black holes merge hierarchically in the ΛCDM paradigm.

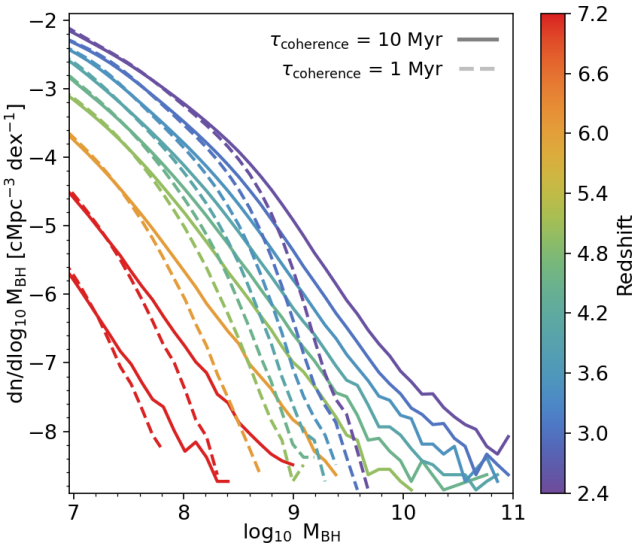

Figure 12. Black hole mass function (BHMF) for all SMBHs in the simulation. Solid lines show the fiducial run (Tab. 1; $\tau_{\rm coherence}$ = 10 Myr) at different redshifts (colors), while dashed lines show the results with a shorter coherence timescale of $\tau_{\rm coherence}$ = 1 Myr, keeping all other parameters fixed. The choice of $\tau_{\rm coherence}$ has a strong impact on the high-mass tail of the BHMF, which corresponds to the most massive SMBHs powering bright quasars across cosmic history.

Although mergers do not play a dominant role in shaping SMBH growth or quasar observables in our current framework, quantifying their contribution is nonetheless crucial for an independent and complementary test of SMBH assembly. Gravitational-wave observations, in particular, are sensitive almost exclusively to mergers and provide a window into a regime that is otherwise invisible to traditional electromagnetic probes. Pulsar timing arrays (PTAs) have already begun to place constraints on the gravitational-wave background generated by SMBH binaries (e.g., Agazie et al. 2023), offering indirect evidence for the demographics of massive black hole pairs at $z \lesssim 2$. The upcoming LISA mission, on the other hand, will directly detect individual SMBH merger events over a wide range of redshifts and masses, reaching into the early Universe and probing the very systems responsible for seeding and assembling today's SMBH population (e.g., Amaro-Seoane et al. 2023).

Because our framework is explicitly constructed from cosmological merger trees, it is particularly well suited to generate detailed predictions for the merger rates, mass ratios, and redshift distribution of SMBH binaries. Even if mergers are subdominant for quasar fueling, their gravitational-wave signatures could provide the cleanest observational handle on SMBH assembly histories. In this sense, gravitational-wave observatories will not only test the merger-driven growth channel but also offer an entirely orthogonal way to validate models like ours. In future work, we plan to extend our model in this direction, leveraging its merger-based nature to make concrete predictions for the SMBH merger landscape in the upcoming era of PTAs and LISA.

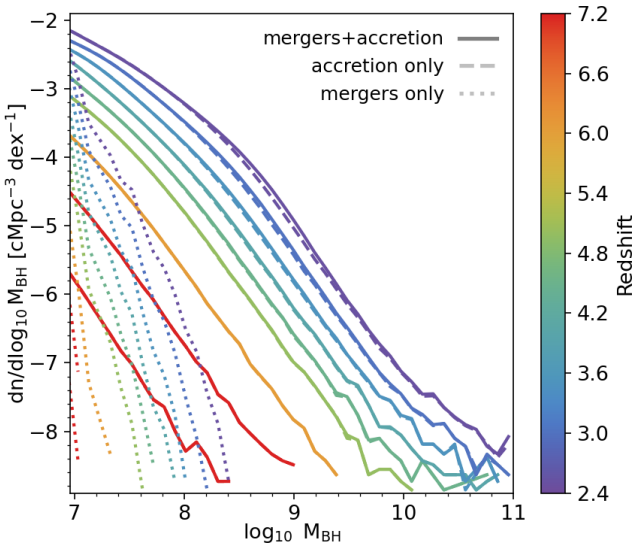

Figure 13. Black hole mass function (BHMF) for all SMBHs in the simulation at different redshifts (colors), illustrating the relative contributions of mergers and accretion to SMBH growth in our fiducial model. Solid lines show the full BHMF including both accretion and mergers. Dashed lines show the BHMF when only accretion is included (merging SMBHs are removed without contributing mass), demonstrating that mergers have only a marginal effect on the overall distribution. When accretion is switched off, mergers alone fail to produce sufficiently massive SMBHs to match observations (dotted lines).

4 DISCUSSION AND SUMMARY

In this work we introduced BAQARO (Black hole Accretion and Quasar Activity in a Realistic Observational framework), a new empirical model for the cosmological evolution of supermassive black holes (SMBHs) and quasars from cosmic dawn to cosmic noon. The framework is built on subhalo merger trees from the N-body version of the FLAMINGO large-volume simulation (Schaye et al. 2023; Kugel et al. 2023), and links SMBH growth to halo assembly through a compact set of parametric prescriptions designed to capture both average evolutionary trends and stochastic variability. A key design choice is the absence of explicit redshift dependence: cosmic evolution enters naturally through the changing specific halo accretion rate (sHAR), allowing the same physical mapping to be applied seamlessly from the epoch of reionization to cosmic noon. The model produces full SMBH mass growth histories, quasar light curves, SMBH merger trees, and host-halo statistics, providing a versatile platform for direct comparison with a wide range of observational constraints.

The model incorporates three main ingredients – seeding, accretion, and mergers – implemented as follows. (i) Seeding. Because subhalos in our merger trees are only resolved once they reach relatively large masses, we initialize each newly formed halo with a fixed "seed" black hole of mass $M_{\rm start}$ (Sec. 2.2.1). This empirical initialization acts as a proxy for unresolved early growth and establishes the baseline of the $M_{\rm BH}$-$M_{\rm halo}$ relation. In this work we adopt a single fiducial value of $M_{\rm start}$, deferring exploration of a distribution of seed masses to future extensions of the model.

(ii) Accretion. We tie the specific black hole accretion rate (sBHAR) to the specific cold halo accretion rate, $s\dot{M}_{\rm cold,acc} = f_{\rm cold}(M_{\rm h})\, s\dot{M}_{\rm acc}$, measured between two consecutive snapshots. The cold fraction $f_{\rm cold}(M_{\rm h})$, taken from Correa et al. (2018), accounts for the suppression of cold inflows in massive halos due to virial shock heating, while allowing efficient accretion in low-mass halos at high redshift. Conditional on $s\dot{M}_{\rm cold,acc}$, we draw the Eddington-normalized accretion rate, $\eta_{\rm acc} = \dot{M}_{\rm BH}/\dot{M}_{\rm Edd}$, from a log-normal distribution whose mean and scatter scale as power laws of $s\dot{M}_{\rm cold,acc}$ (Sec. 2.2.3). The radiative efficiency $\epsilon(\eta_{\rm acc})$ is prescribed following slim-disk models (Sądowski et al. 2014; Madau et al. 2014), transitioning from a thin-disk plateau at sub-Eddington rates to a saturated luminosity at super-Eddington rates. This ensures that the bolometric luminosity, $L_{\rm bol} = \epsilon\, \eta_{\rm acc}\, \dot{M}_{\rm Edd} c^2$, remains physically consistent across regimes. To model stochastic variability, SMBH masses are advanced by sub-cycling each snapshot into intervals of a coherence timescale, $\tau_{\rm coherence}$: $\eta_{\rm acc}$ is held constant over $\tau_{\rm coherence}$ and redrawn thereafter. This single parameter controls how strongly growth histories "average out" versus retain long-lived bursts.

(iii) Mergers. When subhalos merge, their central SMBHs are assumed to coalesce following a simplified and optimistic prescription: the remnant SMBH has a mass equal to the sum of the progenitors, and no black hole is ejected from the host subhalo as a result of gravitational recoil (Sec. 2.2.2). This treatment is similar to that adopted in many large-scale cosmological hydrodynamical simulations (e.g., Habouzit et al. 2021). We also perform control experiments in which we suppress the mass contribution from mergers, or conversely suppress accretion, to isolate their relative roles.

By construction, BAQARO is anchored to three observational diagnostics that probe complementary aspects of quasar physics (Sec. 2.3): (a) the bolometric quasar luminosity function (QLF), which traces the global abundance of quasars as a function of luminosity; (b) the conditional Eddington-ratio distribution function (cERDF), $P(\lambda_{\rm Edd}|L_{\rm bol})$, which leverages broad-line SMBH mass estimates to probe instantaneous fueling at fixed luminosity; and (c) the large-scale clustering of UV-luminous quasars, which constrains typical host halo masses and duty cycles. In practice, we compare our predictions with the bolometric QLF compilation of Shen et al. (2020), cERDF measurements derived from SDSS and high-redshift samples (Wu & Shen 2022; Fan et al. 2023), and the quasar auto-correlation functions from BOSS and SDSS (Eftekharzadeh et al. 2015; Shen et al. 2007) as well as the recent JWST constraints on the high-z quasar-galaxy cross-correlation function (Eilers et al. 2024).

In the analysis presented here, we have focused on the results of a single fiducial run, with the free parameters of the model fixed to the values listed in Tab. 1. This calibration was chosen to approximately reproduce the main quasar observables while enabling us to explore the qualitative implications of the framework. In forthcoming work, we will move beyond this fiducial calibration and perform a full Bayesian inference of the model parameters. This will be made possible by developing an emulator trained on the model outputs that can approximate the predicted observables at negligible computational cost (Sec. 2.4). The emulator will enable Markov Chain Monte Carlo (MCMC) exploration of the parameter space, allowing us to rigorously quantify parameter degeneracies, assess the constraining power of each observable, and obtain posterior distributions jointly constrained by the QLF, cERDF, and clustering. Such an inference pipeline will sharpen the predictive power of the model, provide robust uncertainty estimates, and establish a systematic connection between phenomenological modeling and observational data.

Our results show that the fiducial model provides a satisfactory match to the bright end of the bolometric QLF ($L_{\rm bol} \gtrsim 10^{45} - 10^{46}\,{\rm erg\,s^{-1}}$), the main evolutionary trends of the cERDF, and the clustering of quasars at $z \approx 2.5$ and $z \approx 6$. Nonetheless, several tensions remain: the model overpredicts the abundance of faint quasars, particularly at high redshift; it yields Eddington ratios that are systematically biased toward slightly higher values than those

observed across all redshifts and luminosities; and it underestimates the clustering amplitude at $z \approx 4$, failing to match the strong signal reported by Shen et al. (2007). These discrepancies may point to missing physics in our prescriptions – for example, more complex parametrizations of how accretion is regulated in low-mass SMBHs, or refined treatments of radiative efficiency and luminosity output across different accretion regimes.

At the same time, however, some of these relevant observational constraints remain highly uncertain. The faint end of the QLF is difficult to measure due to incompleteness, obscuration, and contamination from star-forming galaxies. Similarly, the extreme clustering amplitude at $z \approx 4$ is debated, with more recent studies reporting weaker signals (e.g., He et al. 2018). Addressing these issues thus requires advances on both the modeling and observational fronts. Forthcoming wide-field surveys such as DESI, Euclid, and Roman, combined with deep AGN samples from JWST, will provide a far more complete view of quasar demographics and environments, offering critical tests for models like BAQARO.

In addition to confronting the model with key observables, we analyzed the internal assembly of SMBHs in BAQARO and its connection to the growth of their host halos. This leads to several conclusions:

- Accretion dominates SMBH growth. In our model, SMBHs grow predominantly through bursts of near- or super-critical accretion, whereas mergers contribute only a minor fraction of the overall mass budget (Fig. 13). Even under optimistic assumptions about merger timescales and remnant survival, the impact of mergers remains marginal. They can provide modest mass boosts for very massive systems ($M_{\rm BH} \sim 10^9\,{\rm M}_\odot$) in gas-poor halos at late times, but they are incapable of producing the billion-solar-mass SMBHs observed as luminous quasars at all redshifts. This result reinforces a broad consensus from both analytical arguments and cosmological simulations that sustained accretion, rather than mergers, is the dominant channel of SMBH assembly across cosmic history (e.g., Shankar et al. 2009; Volonteri et al. 2016).
- Accreting black holes undergo rapid mass assembly at $z \gtrsim 6$, followed by a marked decline in growth rates toward cosmic noon (Fig. 5). This overall trend reflects the evolution of halo accretion histories, but stochasticity plays a decisive role in shaping the distribution of SMBH masses. In particular, some black holes become extreme outliers, building up significantly more mass than average through short-lived episodes of very high, but radiatively inefficient, accretion (Fig. 6). Such bursts allow rare SMBHs to reach $\sim 10^9\,{\rm M}_\odot$ by $z \approx 6$, consistent with previous models of rapid early SMBH growth (e.g., Madau et al. 2014; Volonteri et al. 2015; Lupi et al. 2016). Crucially, these events are not directly tied to halo mass assembly, but instead emerge from stochastic fluctuations in the accretion rate distribution. Tracking these rare, burst-driven growth histories – rather than focusing solely on population averages – is therefore essential for explaining the emergence of the most massive quasars in the early Universe.
- As a consequence of this SMBH-halo co-evolution, the predicted $M_{\rm BH}$–$M_{\rm halo}$ relation in BAQARO is approximately linear and nearly constant with redshift, with an intrinsic scatter of $\lesssim 0.3$ dex (Fig. 11). This tight correlation reflects the average tendency of SMBH growth to follow halo accretion. However, consistent with the stochastic growth episodes discussed above, the most massive SMBHs powering luminous quasars at all cosmic times do not emerge from the mean relation. Instead, they appear as rare outliers, produced by bursts of unusually efficient accretion rather than steady halo assembly. These stochastic extremes are essential for explaining the billion-solar-mass SMBHs observed at high redshift, but they also complicate simple interpretations of quasar environments that rely on a deterministic $M_{\rm BH}$-$M_{\rm halo}$ mapping.
- A key driver of this stochasticity in SMBH evolution is the coherence timescale of the accretion process, $\tau_{\rm coherence}$. Short values of $\tau_{\rm coherence}$ lead to SMBH mass functions that are narrow, with little diversity in individual growth histories, as accretion fluctuations are averaged out. In contrast, longer coherence timescales preserve broad high-mass tails in the distribution, enabling rare SMBHs to reach $M_{\rm BH} \gtrsim 10^9\,{\rm M}_\odot$ already at early cosmic times (Fig. 12). This makes $\tau_{\rm coherence}$ a critical parameter for determining whether the model can produce the most massive quasars seen at $z \gtrsim 6$. Independent constraints from proximity-zone measurements and clustering-based duty cycle estimates suggest timescales for SMBH accretion and quasar activity of $\sim 10^4$–10^7 yr (e.g., Eilers et al. 2017, 2024; Pizzati et al. 2024b), placing $\tau_{\rm coherence}$ in a regime where it directly connects phenomenological modeling with observables. As such, it provides a promising avenue to tie SMBH accretion physics to measurable quantities, and to test whether the observed diversity in quasar activity is consistent with burst-driven growth.

The version of BAQARO presented here is a preliminary implementation of the framework. Several avenues for further development are already clear. First, a full parameter inference must be carried out. This will enable us to quantify parameter degeneracies, identify which observables drive the strongest constraints, and obtain robust posteriors for SMBH growth prescriptions. Such an inference pipeline will significantly strengthen the predictive power of the model.

Second, the redshift range of the model must be extended. Our current analysis is restricted to $2 \lesssim z \lesssim 15$, mainly to reduce computational costs. Extending it down to $z = 0$ will allow direct tests against the local scaling relations, the observed black hole mass function, and the full history of quasar downsizing. This step will also make it possible to connect high-z accretion-driven growth to the observed demographics of SMBHs in the nearby Universe.

Third, the treatment of the low-mass and seeding regime needs to be improved. At present, the limited resolution of the FLAMINGO simulation prevents us from resolving the formation and early growth of low-mass SMBHs. We plan to tackle this in two complementary ways: (i) by rebuilding the model on the larger, higher-resolution FLAMINGO-10k run (Schaller et al. in prep.), and (ii) by incorporating analytical prescriptions to capture the unresolved early phases of SMBH seeding and growth. Together, these approaches will allow us to explore the critical regime of $M_{\rm BH} \sim 10^4$–$10^7\,{\rm M}_\odot$, which remains one of the most uncertain aspects of SMBH evolution.

With these developments, we hope to address several key questions that remain open in our present analysis. For example: can the strong clustering signal reported by Shen et al. (2007) at $z \approx 4$ be ruled out as an observational systematic, or is there a physically consistent way to connect it with the rapid SMBH buildup at higher redshift and the subsequent evolution to lower redshift? More broadly, how much can we learn from clustering-based estimates of the quasar duty cycle across cosmic time? And how can these constraints be tied to lifetime estimates from quasar proximity zones and damping-wing analyses (e.g., Eilers et al. 2017; Ďurovčíková et al. 2024), and ultimately to the coherence timescale of the accretion process? Because BAQARO resolves individual SMBH accretion histories, it naturally predicts the fraction of time black holes spend above a given luminosity threshold. This definition of duty cycle can be compared directly with clustering-based estimates and with lifetime measurements from proximity zones, providing a coherent test of

whether short-lived, bursty accretion episodes are compatible with the observed demographics of quasars.

Another major uncertainty concerns the role of super-critical accretion in the assembly of early SMBHs. Our results suggest that bursts of highly efficient accretion are essential for producing billion-solar-mass quasars by $z \gtrsim 6$, with short-lived episodes of ~1–10 Myr compatible with lifetime and duty cycle constraints at $z \approx 6$ (Pizzati et al. 2024b). Yet it remains unclear how robust this channel is compared to alternative pathways, and to what degree super-critical accretion can complement or replace heavy seeding scenarios. In the current version of the model, SMBHs are initialized with a fixed seed mass, meaning that the degeneracy between seed properties and subsequent accretion histories remains unresolved. Breaking this degeneracy will be critical for distinguishing between different theories of early black hole formation.

Encouragingly, the next generation of observational constraints will provide powerful tests of these ideas. Ongoing and upcoming surveys with JWST, Euclid, and Roman will directly probe the abundance of 10^6–10^9 $M_\odot$ SMBHs at $z \gtrsim 7$, offering new leverage on the high-redshift early accretion regime. At the same time, a complementary window is opening through gravitational-wave astronomy. Extending the model to the local Universe will allow us to connect with the recent evidence for a nano-Hz gravitational-wave background from pulsar timing arrays – signals that are expected to become even more constraining in the near future (e.g., Agazie et al. 2023). Moreover, the model is well-suited to make forecasts for LISA, which will detect individual SMBH mergers across cosmic time and a broad mass range (e.g., Amaro-Seoane et al. 2023). These gravitational-wave observations will provide an entirely orthogonal test of SMBH assembly, probing the merger-driven channel that is otherwise invisible in traditional electromagnetic surveys.

In summary, our analysis shows that a simple, observationally anchored framework can account for the main demographics of luminous quasars across cosmic time while naturally incorporating the stochasticity required to produce the most extreme SMBHs. Looking ahead, the combination of large-volume simulations, flexible empirical prescriptions, and the rapidly expanding suite of multi-wavelength and multi-messenger observations will enable BAQARO and analogous models to refine our understanding of how the Universe assembled its first quasars and, ultimately, the billion-solar-mass black holes that continue to shape galaxy evolution to the present day.

ACKNOWLEDGEMENTS

EP is grateful to Victor Forouhar Moreno and Rob McGibbon for help with the HBT-HERONS catalogs. We are grateful to the FLAMINGO team for making their dark matter only simulations available. We acknowledge helpful conversations with the ENIGMA group at UC Santa Barbara and Leiden University. JFH and EP acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 885301). This work is partly supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860744 (BiD4BESt). This work used the DiRAC Memory Intensive service (Cosma8) at the University of Durham, which is part of the STFC DiRAC HPC Facility (`www.dirac.ac.uk`). Access to DiRAC resources was granted through a Director's Discretionary Time allocation in 2023/24, under the auspices of the UKRI-funded DiRAC Federation Project. The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

DATA AVAILABILITY

The derived data generated in this research will be shared on reasonable requests to the corresponding author.

REFERENCES

Abbott T. M. C., et al., 2022, Phys. Rev. D, 105, 023520
Abramowicz M. A., Czerny B., Lasota J. P., Szuszkiewicz E., 1988, ApJ, 332, 646
Adelberger K. L., Steidel C. C., 2005, ApJ, 627, L1
Agazie G., et al., 2023, ApJ, 951, L8
Alexander D. M., et al., 2025, arXiv e-prints, p. arXiv:2506.19166
Amaro-Seoane P., et al., 2023, Living Reviews in Relativity, 26, 2
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Arita J., et al., 2023, ApJ, 954, 210
Aversa R., Lapi A., de Zotti G., Shankar F., Danese L., 2015, ApJ, 810, 74
Bañados E., et al., 2018, Nature, 553, 473
Begelman M. C., Blandford R. D., Rees M. J., 1980, Nature, 287, 307
Behroozi P. S., Wechsler R. H., Conroy C., 2013, ApJ, 770, 57
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143
Bekenstein J. D., 1973, ApJ, 183, 657
Bennett J. S., Sijacki D., Costa T., Laporte N., Witten C., 2024, MNRAS, 527, 1033
Bhowmick A. K., Blecha L., Torrey P., Weinberger R., Kelley L. Z., Vogelsberger M., Hernquist L., Somerville R. S., 2024, MNRAS, 529, 3768
Blecha L., Loeb A., 2008, MNRAS, 390, 1311
Bogdán Á., et al., 2023, Nature Astronomy,
Bonoli S., Shankar F., White S. D. M., Springel V., Wyithe J. S. B., 2010, MNRAS, 404, 399
Booth C. M., Schaye J., 2009, MNRAS, 398, 53
Bromm V., Loeb A., 2003, The Astrophysical Journal, 596, 34
Caplar N., Lilly S. J., Trakhtenbrot B., 2015, ApJ, 811, 148
Chandro-Gómez Á., et al., 2025, MNRAS, 539, 776
Circosta C., et al., 2019, A&A, 623, A172
Conroy C., White M., 2013, ApJ, 762, 70
Correa C. A., Schaye J., Wyithe J. S. B., Duffy A. R., Theuns T., Crain R. A., Bower R. G., 2018, MNRAS, 473, 538
Croom S. M., et al., 2005, MNRAS, 356, 415
Croton D. J., 2009, MNRAS, 394, 1109
D'Amato Q., et al., 2020, A&A, 636, A37
Davies F. B., Hennawi J. F., Eilers A.-C., 2019, ApJ, 884, L19
Degraf C., Di Matteo T., Springel V., 2010, MNRAS, 402, 1927
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Devecchi B., Volonteri M., 2009, ApJ, 694, 302
Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604
Dotti M., Sesana A., Decarli R., 2012, Advances in Astronomy, 2012, 940568
Eftekharzadeh S., et al., 2015, MNRAS, 453, 2779
Eilers A.-C., Davies F. B., Hennawi J. F., Prochaska J. X., Lukić Z., Mazzucchelli C., 2017, ApJ, 840, 24
Eilers A.-C., et al., 2020, ApJ, 900, 37
Eilers A.-C., et al., 2024, arXiv e-prints, p. arXiv:2403.07986
Elbers W., Frenk C. S., Jenkins A., Li B., Pascoli S., 2021, MNRAS, 507, 2614
Fan X., et al., 2006, AJ, 132, 117
Fan X., Bañados E., Simcoe R. A., 2023, ARA&A, 61, 373
Fanidakis N., Macciò A. V., Baugh C. M., Lacey C. G., Frenk C. S., 2013, MNRAS, 436, 315
Farina E. P., et al., 2022, ApJ, 941, 106
Ferrarese L., Merritt D., 2000, ApJ, 539, L9

Forouhar Moreno V. J., Helly J., McGibbon R., Schaye J., Schaller M., Han J., Kugel R., 2025, arXiv e-prints, p. arXiv:2502.06932
Genina A., Springel V., Rantala A., 2024, MNRAS, 534, 957
Gilli R., et al., 2022, A&A, 666, A17
Greengard L., Rokhlin V., 1987, Journal of Computational Physics, 73, 325
Habouzit M., et al., 2021, MNRAS, 503, 1940
Habouzit M., et al., 2022, MNRAS, 509, 3015
Haiman Z., Hui L., 2001, ApJ, 547, 27
Han J., Jing Y. P., Wang H., Wang W., 2012, MNRAS, 427, 2437
Han J., Cole S., Frenk C. S., Benitez-Llambay A., Helly J., 2018, MNRAS, 474, 604
Harikane Y., et al., 2023, ApJ, 959, 39
He W., et al., 2018, PASJ, 70, S33
Heger A., Fryer C. L., Woosley S. E., Langer N., Hartmann D. H., 2003, The Astrophysical Journal, 591, 288
Herrmann F., Hinder I., Shoemaker D., Laguna P., Matzner R. A., 2007, ApJ, 661, 430
Hopkins P. F., Richards G. T., Hernquist L., 2007, ApJ, 654, 731
Huško F., Lacey C. G., Schaye J., Nobels F. S. J., Schaller M., 2024, MNRAS, 527, 5988
Inayoshi K., Haiman Z., Ostriker J. P., 2016, MNRAS, 459, 3738
Jeon J., Liu B., Taylor A. J., Kokorev V., Chisholm J., Kocevski D. D., Finkelstein S. L., Bromm V., 2025, ApJ, 988, 110
Jiang Y.-F., Stone J. M., Davis S. W., 2014, ApJ, 796, 106
Juodžbalis I., et al., 2025, arXiv e-prints, p. arXiv:2504.03551
Kelley L. Z., Blecha L., Hernquist L., 2017, MNRAS, 464, 3131
Kokorev V., et al., 2023, ApJ, 957, L7
Kormendy J., Ho L. C., 2013, ARA&A, 51, 511
Koudmani S., Somerville R. S., Sijacki D., Bourne M. A., Jiang Y.-F., Profit K., 2024, MNRAS, 532, 60
Kugel R., et al., 2023, MNRAS, 526, 6103
Kulkarni G., Worseck G., Hennawi J. F., 2019, MNRAS, 488, 1035
Lacey C., Cole S., 1993, MNRAS, 262, 627
Larson R. L., et al., 2023, ApJ, 953, L29
Latif M. A., Ferrara A., 2016, Publ. Astron. Soc. Australia, 33, e051
Li W., Inayoshi K., Qiu Y., 2021, ApJ, 917, 60
Lupi A., Haardt F., Dotti M., Fiacconi D., Mayer L., Madau P., 2016, MNRAS, 456, 2993
Lynden-Bell D., 1969, Nature, 223, 690
Madau P., Haardt F., Dotti M., 2014, ApJ, 784, L38
Magorrian J., et al., 1998, AJ, 115, 2285
Maiolino R., et al., 2024, A&A, 691, A145
Martini P., 2004, in Ho L. C., ed., Coevolution of Black Holes and Galaxies. p. 169 (arXiv:astro-ph/0304009), doi:10.48550/arXiv.astro-ph/0304009
Martini P., Weinberg D. H., 2001, ApJ, 547, 12
Matsuoka Y., et al., 2018, ApJ, 869, 150
Matsuoka Y., et al., 2023, ApJ, 949, L42
Mayer L., Kazantzidis S., Madau P., Colpi M., Quinn T., Wadsley J., 2007, Science, 316, 1874
Mazzucchelli C., et al., 2017, ApJ, 849, 91
McBride J., Fakhouri O., Ma C.-P., 2009, MNRAS, 398, 1858
Mead A. J., Verde L., 2021, MNRAS, 503, 3095
Merloni A., Heinz S., 2008, MNRAS, 388, 1011
Milosavljević M., Merritt D., 2001, ApJ, 563, 34
Ni Y., Di Matteo T., Gilli R., Croft R. A. C., Feng Y., Norman C., 2020, MNRAS, 495, 2135
Ohsuga K., Mori M., Nakamoto T., Mineshige S., 2005, ApJ, 628, 368
Omukai K., Schneider R., Haiman Z., 2008, The Astrophysical Journal, 686, 801
Pacucci F., Loeb A., 2020, ApJ, 895, 95
Pizzati E., Hennawi J. F., Schaye J., Schaller M., 2024a, MNRAS, 528, 4466
Pizzati E., et al., 2024b, MNRAS, 534, 3155
Porciani C., Norberg P., 2006, MNRAS, 371, 1824
Porciani C., Magliocchetti M., Norberg P., 2004, MNRAS, 355, 1010
Porras-Valverde A. J., Ricarte A., Natarajan P., Somerville R. S., Gabrielpillai A., Yung L. Y. A., 2025, arXiv e-prints, p. arXiv:2504.11566
Reines A. E., Volonteri M., 2015, ApJ, 813, 82
Richards G. T., et al., 2006, AJ, 131, 2766
Ross N. P., et al., 2009, ApJ, 697, 1634
Salpeter E. E., 1964, ApJ, 140, 796
Schaller M., et al., 2024, MNRAS, 530, 2378
Schaye J., et al., 2015, MNRAS, 446, 521
Schaye J., et al., 2023, MNRAS, 526, 4978
Schindler J.-T., et al., 2023, ApJ, 943, 67
Schmidt M., 1963, Nature, 197, 1040
Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337
Shankar F., Weinberg D. H., Miralda-Escudé J., 2009, ApJ, 690, 20
Shankar F., Crocce M., Miralda-Escudé J., Fosalba P., Weinberg D. H., 2010, ApJ, 718, 231
Shen Y., et al., 2007, AJ, 133, 2222
Shen Y., et al., 2009, ApJ, 697, 1656
Shen X., Hopkins P. F., Faucher-Giguère C.-A., Alexander D. M., Richards G. T., Ross N. P., Hickox R. C., 2020, MNRAS, 495, 3252
Sądowski A., Narayan R., McKinney J. C., Tchekhovskoy A., 2014, MNRAS, 439, 503
Soltan A., 1982, MNRAS, 200, 115
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Tanaka T., Haiman Z., 2009, ApJ, 696, 1798
Thorne K. S., 1974, ApJ, 191, 507
Timlin J. D., et al., 2018, ApJ, 859, 20
Tucci M., Volonteri M., 2017, A&A, 600, A64
Ueda Y., Akiyama M., Hasinger G., Miyaji T., Watson M. G., 2014, ApJ, 786, 104
Vito F., et al., 2018, Monthly Notices of the Royal Astronomical Society, 473, 2378
Vito F., Di Mascia F., Gallerani S., Zana T., Ferrara A., Carniani S., Gilli R., 2022, MNRAS, 514, 1672
Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020, Nature Reviews Physics, 2, 42
Volonteri M., 2012, Science, 337, 544
Volonteri M., Rees M. J., 2006, ApJ, 650, 669
Volonteri M., Lodato G., Natarajan P., 2008, MNRAS, 383, 1079
Volonteri M., Silk J., Dubus G., 2015, ApJ, 804, 148
Volonteri M., Dubois Y., Pichon C., Devriendt J., 2016, MNRAS, 460, 2979
Wang F., et al., 2019, ApJ, 884, 30
Wang F., et al., 2021, ApJ, 907, L1
Wang F., et al., 2023, ApJ, 951, L4
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
Weinberger R., Bhowmick A., Blecha L., Bryan G., Buchner J., Hernquist L., Hlavacek-Larrondo J., Springel V., 2025, arXiv e-prints, p. arXiv:2502.13241
White M., Martini P., Cohn J. D., 2008, MNRAS, 390, 1179
White M., et al., 2012, MNRAS, 424, 933
Wu Q., Shen Y., 2022, ApJS, 263, 42
Wu J., et al., 2022, MNRAS, 517, 2659
Yang J., et al., 2020, ApJ, 897, L14
Yang J., et al., 2023, ApJS, 269, 27
Yu Q., Tremaine S., 2002, MNRAS, 335, 965
Zel'dovich Y. B., Novikov I. D., 1967, Soviet Ast., 10, 602
Zhang H., Behroozi P., Volonteri M., Silk J., Fan X., Hopkins P. F., Yang J., Aird J., 2023, MNRAS, 518, 2123
da Ângela J., et al., 2008, MNRAS, 383, 565
Ďurovčíková D., et al., 2024, ApJ, 969, 162
Ďurovčíková D., et al., 2025, arXiv e-prints, p. arXiv:2505.00080

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