The Astrophysical Journal, 952:146 (25pp), 2023 August 1 https://doi.org/10.3847/1538-4357/acd9ce
© 2023. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

QUOTAS: A New Research Platform for the Data-driven Discovery of Black Holes

Priyamvada Natarajan[1,2,3], Kwok Sun Tang[4], Robert McGibbon[5], Sadegh Khochfar[5], Brian Nord[6,7,8], Steinn Sigurdsson[9], Joe Tricot[10], Nico Cappelluti[11], Daniel George[10], and Jack Hidary[10]
[1] Department of Astronomy, Yale University, 52 Hillhouse Avenue, New Haven, CT 06511, USA
[2] Department of Physics, Yale University, P.O. Box 208121, New Haven, CT 06520, USA
[3] Black Hole Initiative, Harvard University, 20 Garden Street, Cambridge MA 02138, USA
[4] Department of Astronomy, University of Illinois Urbana-Champaign, USA
[5] Institute for Astronomy, University of Edinburgh, Royal Observatory, UK
[6] Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA
[7] Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
[8] Department of Astronomy and Astrophysics, University of Chicago, IL 60637, USA
[9] Department of Astronomy & Astrophysics and Institute for Gravitation and the Cosmos, Pennsylvania State University, 525 Davey Lab, University Park, PA 16802, USA
[10] Sandbox@Alphabet, Mountain View, CA 94043, USA
[11] Department of Physics, University of Miami, Coral Gables, FL, USA
Received 2021 June 15; revised 2023 April 13; accepted 2023 April 14; published 2023 July 25

Abstract

We present QUOTAS, a novel research platform for the data-driven investigation of supermassive black hole (SMBH) populations. While SMBH data—observations and simulations—have grown in complexity and abundance, our computational environments and tools have not matured commensurately to exhaust opportunities for discovery. To explore the BH, host galaxy, and parent dark matter halo connection—in this pilot version—we assemble and colocate the high-redshift, $z > 3$ quasar population alongside simulated data at the same cosmic epochs. As a first demonstration of the utility of QUOTAS, we investigate correlations between observed Sloan Digital Sky Survey (SDSS) quasars and their hosts with those derived from simulations. Leveraging machine-learning algorithms (ML), to expand simulation volumes, we show that halo properties extracted from smaller dark-matter-only simulation boxes successfully replicate halo populations in larger boxes. Next, using the Illustris-TNG300 simulation that includes baryonic physics as the training set, we populate the larger LEGACY Expanse dark-matter-only box with quasars, and show that observed SDSS quasar occupation statistics are accurately replicated. First science results from QUOTAS comparing colocated observational and ML-trained simulated data at $z3$ are presented. QUOTAS demonstrates the power of ML, in analyzing and exploring large data sets, while also offering a unique opportunity to interrogate theoretical assumptions that underpin accretion and feedback models. QUOTAS and all related materials are publicly available at the Google Kaggle platform. (The full data set—observational data and simulation data—are available at: https://www.kaggle.com/ and the codes are available at: https://www.kaggle.com/datasets/quotasplatform/quotas)

Unified Astronomy Thesaurus concepts: Quasars (1319)

1. Introduction and Motivation

Observations suggest that most, if not all, galaxies in the universe host central supermassive black holes (SMBHs). Theoretical modeling at present successfully embeds and integrates the formation, growth, evolution, and impact of black holes (BHs) on their environments—into the larger framework of structure formation in the standard cosmological model—revealing a profound underlying BH–galaxy–halo connection. Given our empirical access and focus, much attention has been paid to the BH–galaxy coupling, its origin, and evolution. More exhaustive exploration of the underlying correlation with the associated dark matter halo is warranted given that structure formation in the standard paradigm is driven by dark matter. The key motivation for QUOTAS rests on garnering deeper insights into this trium-variate connection by colocating and analyzing observational data and simulated data. Additionally, we derive, in particular, new ways of interrogating our input assumptions regarding BH growth that are currently implemented in cosmological simulations.

The coupling between SMBHs and their host galaxies presents as empirical correlations between the BH mass and properties of their host galactic stellar populations (see Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000; Tremaine et al. 2002; Schawinski et al. 2006; Gültekin et al. 2009; Kormendy & Ho 2013; McConnell & Ma 2013; Heckman & Best 2014). How and when these are established are currently open questions, and two possibilities are under active consideration (Alexander & Hickox 2012; Natarajan 2014). Correlations are suspected to either arise from the initial conditions that govern the formation of BH seeds (Agarwal et al. 2013; Woods et al. 2019), or from the coupled evolution of galaxies and their central SMBHs, which emerges over cosmic time (Haehnelt et al. 1998; Natarajan et al. 2019) in the standard cold dark matter driven structure formation model. Discriminating between these alternatives for the origin of observed correlations and discerning the causal mechanisms that underlie them are active research areas. To address this fundamental question, comprehensive data of SMBH populations across multiple cosmic epochs is required, including at very high redshifts, where observations currently peter out.

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

However, with the spate of new observational facilities and instruments expected to come online shortly, vast amounts of data probing the earliest cosmic epochs will soon be available. Combining the inhomogeneous data from these various independent instruments and surveys will be an exceedingly complex task. Here, we construct the QUOTAS platform as an ongoing, actively updated effort and present the pilot version here as a template for continuing consolidation studies.

A broad data landscape—including numerical simulations and multiwavelength, multiepoch observations—has been critical for building models that have helped us understand the symbiotic growth history of BHs and their host galaxies (Haehnelt et al. 1998; Alexander & Hickox 2012; Volonteri 2012; Natarajan 2014; Storchi-Bergmann & Schnorr-Müller 2019). Simulating BH growth in the full cosmological context is numerically challenging due to the active interplay of disparate physical scales—from the megaparsec scale gas flows in the cosmic web that feed down to parsec-scale gas accretion onto the accreting SMBH. Despite these hurdles, there has been enormous progress with the inclusion of physical processes that cannot be implemented in an ab initio fashion in computations via the use of so-called subgrid models (see Springel et al. 2005; Dubois et al. 2012; Sijacki et al. 2015; Gaspari et al. 2020). Statistical studies of SMBH populations that leverage observational and simulated data with a new suite of tools is needed to enable breakthroughs in the fundamental understanding of the demographics and the cosmic accretion history of BHs. The Hubble Space Telescope (HST) and surveys like the ground-based Sloan Digital Sky Survey (SDSS) have acquired data on more than half a million quasars. Their archives contain not only raw data (e.g., images and spectra), but also high-level, value-added, and science-ready products—often in the form of relational databases. HST data are archived in the Mikulski Archive for Space Telescope (MAST) archive, and SDSS data is publicly available on several online platforms—e.g., CyVerse[12] and the NOAO Data Lab[13]—that include additional tools to investigate data in situ. In the near future, multiple new facilities, e.g., including the Vera Rubin Observatory (LSST) and the Nancy Grace Roman Telescope (NGRST)—will all contribute significant amounts of data to the already large banks available for the investigation of SMBHs and their galaxy hosts. Meanwhile, on the numerical side, next-generation peta-scale simulations are also currently being developed to complement this observational deluge. These two kinds of data are not integrated and used effectively in concert at present, and our work illustrates one way forward to fill this important gap with QUOTAS.

To optimally mine these data, we propose development of customized, sophisticated methodologies and analysis tools that leverage machine-learning (ML) algorithms in combination with our current domain knowledge. Flexible ML methodologies are the natural set of tools that extract maximal information from these varied and rich data. ML is already being vigorously applied in astronomy, for example, the Event Horizon Telescope (EHT) collaboration combined data from multiple radio observatories and relied extensively on ML methods to generate the silhouette of the SMBH at the center of M87 (Event Horizon Telescope Collaboration et al. 2019a, 2019b). ML has demonstrated successes in a growing number of other astrophysics applications (Ntampaka et al. 2019)—including time domain astronomy (Mahabal et al. 2005) image classification using a convolutional neural network to morphologically classify and extract properties from large observed samples of nearby galaxies (Ghosh et al. 2020), and the use of generative adversarial networks to recover features in astrophysical images of galaxies beyond the de-convolution limit (Schawinski et al. 2017). In simulations, trained ML algorithms have been used to successfully mimic a full hydrodynamic cosmological simulation to study the galaxy–halo connection (Kamdar et al. 2016), to obtain new insights into cosmological structure formation (Lucie-Smith et al. 2020), including predictions of the evolution of angular momentum (Cadiou et al. 2021) and characterizing galaxy formation itself (McGibbon & Khochfar 2022; Villaescusa-Navarro et al. 2022). The power and prospects of insights gleaned in astronomy using ML for data-driven discovery have been amply demonstrated (Djorgovski et al. 2016; Bernardini et al. 2020; Moster et al. 2021; Ntampaka et al. 2020; Dai & Seljak 2021).

At the highest redshifts, only the most-luminous quasars are detected. Therefore, we have an incomplete and biased census of the accreting BH population. Our goal is to develop tools that will help uncover the more characteristic, average quasars at these early epochs, beyond the most extreme objects, to derive a comprehensive view the quasar population and disentangle seeding, accretion, and dynamical processes that shape properties of the quasar population. We demonstrate that building the BH–galaxy–halo connection with QUOTAS will permit prediction and understanding of the properties of the more ubiquitous and more representative BH population at high redshifts. Two integrated quantities, the black hole mass function (BHMF) and the quasar luminosity function (QLF) computed directly from observational data are typically used to constrain theoretical models of BH growth evolution. In what follows, we present the design and operation of the QUOTAS framework, built and tailored specifically to address the BH–galaxy–halo connection. QUOTAS is an ongoing endeavor, and we are continually collating, updating, and ingesting new quasar data as it becomes available. The data from multiple observational surveys is currently published, and is hence available in QUOTAS. For all surveys listed in Table 2, we have survey selection functions, survey depths, and survey areas from the painstaking compilation provided in Kulkarni et al. (2019a). Knowledge of selection functions is critical to homogenizing the database and to derive quasar population properties like the BHMF and QLF for study and analysis. Though we have collated all current $z > 3$ observed quasars in QUOTAS, not all of them have complete information; for some, only photometric information is available, and for others additional spectroscopic information is also available. In this paper, we describe the current full landscape of BH data, outlining observations collected from various astronomical surveys and simulations. The database design and the considerations for QUOTAS are discussed, as are illustrations of ease-of-use cases with concrete examples and visualizations. The concrete examples discussed here are focused on the SDSS survey data as this is the largest, full homogenized data set we have, and they are able to compare our results with previously published work and the sample size permits comparison with simulations. For $z > 3$ SDSS quasars, we combine the observational data with corresponding

[12] https://www.cyverse.org/
[13] https://datalab.noao.edu/

simulation slices from the cosmological dark-matter-only LEGACY suite (Meriot et al. 2022) to demonstrate the power that ML offers in interrogating our input model assumptions while tackling the BH growth problem. We follow conventional notation, wherein accreting (actively evolving/growing) BHs detected in optical wavelengths are referred to as "quasars" (with luminosities $\sim 10^{15}\,L_{\odot}$) while those detected in X-ray observations are denoted as active galactic nuclei (AGNs; with lower luminosities ranging from $10^{10}\,L_{\odot}$ to $10^{15}\,L_{\odot}$). AGNs are rare at $z \geqslant 3$, our chosen epochs for study in this pilot project, and the majority have identifiable optical quasar counterparts. When referring to multiwavelength data for accreting SMBHs, we use the term "AGNs," and the term "quasars" when referring to optical data.

ML is yet to be deployed for the comprehensive and systematic study of SMBH populations, though a recent attempt was made in deriving BH masses from spectral data (Yao-Yu Lin et al. 2020). As show in this paper, the BH growth problem offers an important case study to develop and hone new ML algorithms, despite the known challenges for ML techniques, namely, extrapolative prediction (Lucie-Smith et al. 2020), and uncertainty quantification. Deep-learning-based models for instance, lack native uncertainty quantification, which degrades their ability to make predictions beyond the data that they have seen and trained on (Nord et al. 2019; Hortúa et al. 2020; Lucie-Smith et al. 2020).

In this paper, we present the highlights of the QUOTAS pilot project including early science results. In Section 2, we briefly introduce accreting black holes and describe the status of our current understanding of the coevolution of SMBHs and their host galaxies in Section 3. In Section 4, we outline the full current data landscape for quasars. The design and development of the QUOTAS database that collates observational and simulated data in a user-friendly format that is able to be queried is presented in Section 5. The illustration of ease of use case along with sample queries and an example of the workflow is presented in Section 6. The first results from this pilot study, where we focus on the $z > 3$ high-redshift SDSS quasar population deploying ML tools to compare simulated and observational data are presented in Section 7. In this section, upon combining with simulated data, we show how our underlying BH accretion model assumptions stand to be questioned with these new ML tools. We also demonstrate the power of ML to help formulate optimal survey strategies for uncovering the lower-luminosity quasar population. Finally, we conclude in Section 8 with a brief discussion of our longer-term goals: to continually update observational data, and standardize it with appropriate selection functions and to develop a suite of tailored ML analysis tools that will help make predictions to uncover the more characteristic, representative SMBH population, beyond just the most extreme objects in order to fully explore and understand the underlying BH–host galaxy–parent dark matter halo connection.

2. The Properties of Accreting Black Holes

In this section, we briefly outline the current understanding of accreting BHs and their properties to provide a broader context for our project. This discussion includes a brief description of the taxonomy, and a concise summary of current physical models for quasars. We discuss the essential theoretical and observational characteristics of accreting SMBHs, which inform the presently known correlations between galaxy properties and BH mass, and the population statistics relevant for the data-driven investigations explored in this work.

Less than a parsec across, the nucleus of nearly every galaxy appears to harbor a central SMBH (Genzel 2014). A galactic nucleus is considered active when a hot and geometrically thin disk of accreting gas surrounds the central SMBH. While incomplete, there exist descriptions of unified models for AGNs that describe their dynamics and account for their properties across a large range of physical scales around the growing BH (for example, the proposal to unify all radio-loud AGNs is presented in Urry & Padovani (1995). Per these unified schemes that attempt to couple the structure with phenomenology across scales—on small scales—ultraviolet radiation that is emitted by the accretion disk illuminates small, closely orbiting gas clouds and on—larger scales—the disk is inferred to be surrounded by a puffed-up toroidal region that consists of dusty molecular gas. Accretion activity is believed to be episodic, lasting $\lesssim 10^{8}$ yr, during the lifecycle of a galaxy, during which it can produce a tremendous amount of energy, often outshining the entire stellar content of the galaxy by several orders of magnitude. Integrating the various physical scales and the coupling of astrophysical processes therein has been the aspiration of multiple modeling efforts to date. We describe the observational signatures of black hole activity that arise on full the range of scales across wavelengths that are now integrated in a convenient form in QUOTAS.

The typical spectral energy distribution (SED) that models the energy output of an accreting BH comprises weighted emission, which is produced by astrophysical processes occurring in distinct physical regions and hence scales of the object—the accretion disk, the hot corona, and the surrounding dusty torus. The observed SED of an accreting BH spans all the way from the radio to very-high-energy gamma-rays (10^{9}–10^{30} Hz) and consists of a superposition of multiple components: a tall bump at high frequencies, a blackbody component, and a plethora of emission and absorption lines. Details of emission mechanisms and the corresponding scales on which they operate can be found in these two comprehensive reviews: Peterson (2014) and Netzer (2015). Observed sources are organized primarily into two classes—Type Is and Type IIs—based on the width of emission lines in their spectra, a feature which tends to correlate with the viewing angle to the central BH. Type I AGNs are characterized by the presence of relatively broad (1000–20,000 km s^{-1}) emission lines, which are produced by a population of the closely orbiting clouds that move in Keplerian orbits. Both these clouds and the accretion disk itself reside within the broad-line region (BLR), which can most readily be detected when the surrounding torus is viewed face-on. Type II AGNs meanwhile are characterized by the presence of narrow spectral emission lines (300–1000 km s^{-1}), which originate from the more quiescent clouds that reside farther away from the accretion disk in the so-called narrow-line region (NLR). When the galaxy is viewed edge-on, the obscuring dusty torus prohibits a direct view into the accretion disk, but permits a view of the toroidal region. Observed Type I AGNs are more luminous than Type II AGNs and therefore tend to comprise the bulk of observationally detected high-redshift sources. All optical quasars are Type I AGNs, and the thus-far X-ray detected Type II AGNs at $z \geqslant 3$ also have optical counterparts. A population of heavily obscured Type II AGNs that are as yet undetected are expected to exist across cosmic epochs, and the high-redshift end population of these sources stands to be uncovered soon by the James Webb

Space Telescope (JWST). For a more detailed description and illustration of these unified models, again we refer the reader to two comprehensive reviews and the references therein (Peterson 2014; Netzer 2015). Currently available photometric and spectral data for $z > 3$ sources are included in QUOTAS.

3. Current Understanding of BH–Galaxy–Halo Coevolution

All black holes—ranging from stellar mass to SMBHs—can be characterized by three fundamental properties: mass, spin, and charge. This simplicity and small number of parameters needed to characterize BHs makes them attractive objects for systematic study. For astrophysical BHs, the subject of this work, charge is irrelevant, as they are expected to be charge-neutral (Znajek 1977). Conceptual models of BH growth are therefore focused on their mass assembly history and their spin evolution. We briefly describe the currently observed correlations that link central BHs to their host galaxy properties. The formation of galaxies is intricately coupled to the underlying properties of their parent dark matter halos; therefore, the existence of a galaxy–halo connection is well established in cosmology (Wechsler & Tinker 2018). Here, our aim is to build new synthetic models that include BHs into this picture explicitly and extend this conceptual frame to fully understand the BH–galaxy–halo connection by combining observational data with simulated data using ML tools.

Observations in the local universe reveal the existence of a correlation between the central SMBH and the stellar mass, luminosity, and velocity dispersion of stars in the inner regions (bulges) of their host galaxies (Magorrian et al. 1998; Ferrarese & Merritt 2000; Tremaine et al. 2002; Häring & Rix 2004; Gültekin et al. 2009; Kormendy & Ho 2013; McConnell & Ma 2013; Woo et al. 2015; Bentz & Manne-Nicholas 2018; Sexton et al. 2019; Ding et al. 2020). These empirical correlations appear to hold for local, inactive, dormant SMBHs over approximately 5 orders of magnitude in BH mass (from 10^5–$10^{10}\,M_\odot$). At face value these correlations suggest that the assembly and growth of BHs is likely linked to the evolution of their host galaxies, as was first explored in Haehnelt et al. (1998). These correlations, albeit detected at these late cosmic times, could also reflect initial conditions, encoding, for instance, the native properties of the sites where initial BH seeds form efficiently. The origin of these correlations and whether they reflect deeper causal physical processes is a subject of active current study. For instance, it is unclear if these correlations are set up initially at the earliest epochs where galaxies are seeded with the first BHs. For example, one proposed channel for the formation of massive initial BH seeds from the direct collapse of pristine gas, naturally accounts for the origin of these correlations as an imprint of the initial physical seeding conditions (Lodato & Natarajan 2006, 2007; Agarwal et al. 2014, 2019; Wise et al. 2019). Heavy seeds have been invoked to account for the formation and assembly of the rare, most-massive BHs at early cosmic epochs. By contrast, light initial BH seeds appear to be sufficient to account for the more modest mass SMBHs with masses in the range of 10^6–$10^8\,M_\odot$, detected in the nearby universe, like the one harbored at the center of our own galaxy, the Milky Way. Alternatively, these measured local correlations may simply reflect the accumulated history of the stochastic assembly process of galaxies that occurs via accretion of mass and repeated merging (Peng 2007; Hirschmann et al. 2010; Jahnke & Macciò 2011; Ricarte & Natarajan 2018a; Ricarte et al. 2019) over cosmic time.

Understanding if these local correlations reflect deeper physical causes, and whether astrophysical processes that self-regulate BH growth and star formation activity in a galactic nucleus operate efficiently and in tandem, is a question of fundamental importance. Whether such a correlation holds at any earlier times for accreting BHs is not observationally known at present (Treister et al. 2011). More observational data are needed to probe if these correlations were imprinted during early cosmic epochs (Kormendy & Ho 2013; Natarajan 2014; Woods et al. 2019) and if so, how and why they persist to late cosmic times. The differences between competing models of BH seeding and growth are naturally more pronounced at early cosmic times (Agarwal et al. 2012; Ricarte & Natarajan 2018b) closer to the seeding era; however, observational constraints at these epochs are currently sparser. Detecting growing SMBHs at these early epochs has been a technical challenge, as distant accreting sources are significantly fainter. At high redshifts, we preferentially detect only the rare, highest-luminosity quasars, and the models have only an unrepresentative sample of quasars to anchor and for guidance to predict properties of the more ubiquitous, and as yet undetected, lower-luminosity quasar population. Therefore, to develop a more complete theoretical understanding, we need comprehensive sampling of the entire range of luminosities and hence BH masses powering quasars and their associated hosts. Additionally, we are yet to have detectors with the appropriate spectral sensitivity to capture emission from the early dusty universe, a situation that has just started to alter with new data trickling in from JWST (Eilers et al. 2023). Our pilot project work presented here targets early epochs and aims to provide a framework for making predictions for the properties of accreting black holes that have lower luminosities than the currently detected population of high-redshift quasars. Our goal is to showcase QUOTAS by coupling observations in the database with simulations and developing a set of customized ML tools that will permit probing the BHMF and QLF in deeper ways. QUOTAS is a live database that we plan to continually update as more quasar survey data and individual follow-up of sources become available. In this first phase, we use the collated SDSS data and coeval simulation slices to demonstrate the power of colocating these data sets.

In future work, we plan to address and fill in a critical gap using trained ML to predict the abundance and luminosities of quasars that are 1–2 orders of magnitude lower in luminosity than the brightest quasars currently observed.

4. The Current Data Landscape for Quasars

In this section, we first discuss multiwavelength observational data, including extant data archive facilities, their content, and how they are currently configured for access and use before outlining the specific sources that we use to construct QUOTAS. We reiterate that QUOTAS represents an active, ongoing effort. Both space-based and ground-based observatories have been essential for developing a multiwavelength picture of central SMBHs. For example, data from HST, Chandra, XMM, Herschel, and the Spitzer Space Telescope spanning X-ray to infrared wavelengths, and their cross-correlations, have revealed important ways in which BH accretion is related to key host galaxy properties like stellar mass and the star formation rate (Chen et al. 2013). Most space missions and ground-based observational

Table 1
Compiled List of Quasars Including Additional Properties of the SMBHs Included in Our Database

z Range	Num. Quasars	Luminosity Estimator	Vel. Disp. Estim.	Reference
3.01–4.33	43	L_λ(1350), L_λ(3000), L_λ(5100), L(Hβ), L(Mg II),	Hβ, Mg II, C IV	Shen et al. (2019a)
3.00–5.46	7955	L_λ(1350), L_λ(3000), L_λ(5100), L(Hβ), L(Mg II)	Hβ, Mg II, C IV	Shen et al. (2011)
3.00–4.89	26571	L_λ(1350), L_λ(3000), L_λ(5100)	Mg II, C IV	Kozłowski (2017)
3.22–3.76	24	L_λ(3000), L_λ(5100)	Hβ, Mg II	Zuo et al. (2015)
4.52–6.41	21	L_λ(3000)	Mg II	De Rosa et al. (2011)
4.66–4.92	40	L_λ(1450), L_λ(3000)	Mg II	Trakhtenbrot et al. (2011)
5.64–6.42	50	L_λ(1350), L_λ(1700), L_λ(3000)	Mg II, C IV	Shen et al. (2019b)
5.77–6.31	6	L_λ(1350), L_λ(3000)	Mg II, C IV	Jiang et al. (2007)
5.82–6.28	5	L_λ(1350), L_λ(3000)	Mg II, C IV	Kurk et al. (2007)
5.90	1	L_λ(3000)	Mg II	Eilers et al. (2018)
5.98–6.44	9	L_λ(3000)	Mg II	Willott et al. (2010a)
5.98–6.9	38	L_λ(3000)	Mg II, C IV	Farina et al. (2022)
5.98–6.9	69	L(Lyα)	Lyα	Matsuoka et al. (2022)
6.13	1	L_λ(3000)	Mg II	Shao et al. (2017)
6.48–7.08	11	L_λ(3000)	Mg II	Mazzucchelli et al. (2017)
6.60–7.10	4	L_λ(3000)	Mg II, C IV	De Rosa et al. (2014)
7.07	1	L_λ(1350)	Mg II	Matsuoka et al. (2019)

Note. This table comprises all reported spectroscopic data from quasar samples (surveys listed in Table 2) as well as individual quasars.

surveys have been efficiently cataloging their own data to permit public access after the lapse of the data proprietary period. One important challenge in dealing with quasar data pertains to how candidates are traditionally selected from surveys, using color–color diagnostics via the drop-out technique, which was in fact originally developed to detect high-redshift quasars in Warren et al. (1987) that exploits the spectral shape of their SEDs per templates, successfully adapted later to detect high-redshift galaxies (Steidel et al. 1999). Therefore, at progressively higher redshifts, only the most-luminous quasars are detected, making it difficult to discern properties of unobserved fainter sources. This is precisely where we see the power of trained ML techniques that will extract and utilize the characteristic properties of the detected sample to predict the properties of hitherto undetected fainter sources based on sources currently seen and a set of underlying modeling assumptions. In fact, ML studies of these combined data, as we demonstrate in Section 7, in turn also permit interrogation of our underlying modeling assumptions that are incorporated into current simulations regarding BH growth. This finding shows that ML offers a novel and potent arbiter between various model assumptions as well.

HST has acquired data on a large number of quasars, and their galaxy hosts, over more than three decades of operation. This data is efficiently archived in and distributed from MAST, hosted and managed by the Space Telescope Science Institute.[14] In addition to the images—raw, cleaned, and processed—MAST provides high-level science-ready data products. In particular, MAST houses the AGN SED ATLAS (Brown et al. 2019), which contains the full SEDs of 41 AGNs at present, derived from multiwavelength photometry and archival spectroscopy, which combines information from eight MAST-supported projects (HST, SWIFT-UVOT, GALEX, PanSTARRS, IUE, FUSE, HUT, and WUPPE) plus an additional nine other missions/observatories for $z > 3$ quasars. Beyond this Atlas, MAST also contains 80 SEDs of accreting sources, composites that are produced by mixing the SEDs of the central regions of active sources with their host galaxy SEDs. This high-level science products database contains a total of 121 AGNs whose SEDs span a wavelength range from ~0.09–30.0 μm for these $z > 3$ sources. For some sources, additionally, there is even broader wavelength coverage extending into the X-ray, far-infrared, and radio.[15] The archival data is stored in relational database tables, which are accessed through two user interfaces—a Java-based application ("StarView") and a web-based application: neither of these require the user to understand the database architecture or how to use Structured Query Language (SQL) to browse and extract data.

The ground-based SDSS has mapped a third of the sky and provides the most detailed three-dimensional data for more than 3 million unique astronomical objects.[16] SDSS contains more than 0.5 million quasars detected out to $z \sim 7$ and comprises a homogeneous sample of quasars with optical spectra. Each round of the SDSS quasar surveys was motivated by a different science goal. For example, the SDSS DR7 quasar catalog consists of 105,783 spectroscopically confirmed quasars from the SDSS-I/II survey whose aim was to study the QLF and clustering properties (Abazajian et al. 2009). With every data release, the number of detected quasars has grown dramatically. The latest SDSS DR14, not only increased the number of quasars by a factor of 5 compared to the SDSS DR7, it also went 1.5 mag fainter, enabling the probing of quasar properties over a much larger luminosity range (Abolfathi et al. 2018). Though the SDSS catalogs contain the X-ray, UV, optical, IR, and radio imaging properties of the quasars wherever available, they lack spectral information for the majority of sources. Spectra have been typically obtained from detailed, independent follow-up studies, as seen from the references listed in our Tables 1 and 2. The detection of quasars and their redshift estimates are done primarily using the color–color selection of drop-outs (Richards et al. 2002). The SDSS archive also offers a range of sophisticated science products that are available as Value Added Catalogs (VACs).[17] The fully standardized and homogenized data from SDSS serve as our current benchmark for the rest of the observational data that we have collated. For

[14] https://archive.stsci.edu/hst/search_retrieve.html

[15] https://archive.stsci.edu/hlsp/index.html

[16] https://www.sdss.org/dr16/data_access/

[17] https://www.sdss.org/dr14/data_access/value-added-catalogs/

Table 2
List of Quasar Survey Samples for Which We Have Photometric Data That Are Included in Our Database QUOTAS and for Which We also Have Selection Functions, Survey Area, Survey Depths, and Completeness Estimates from Kulkarni et al. (2019a)

z range	Survey	Reference	Number of Quasars	Area (deg^2)
2.2–3.5	BOSS DR9	Ross et al. (2013)	23301	2236.0
3.7–4.7	SDSS DR7	Schneider et al. (2010)	1785	6248.0
3.6–5.2	NDWFS	Glikman et al. (2011)	12	1.71
3.8–5.3	DLS	Glikman et al. (2011)	12	2.05
4.7–5.4	SDSS+WISE	Yang et al. (2016)	99	14555.0
4.7–5.5	SDSS DR7	McGreer et al. (2013)	103	6248.0
4.7–5.5	—Stripe 82	McGreer et al. (2013)	59	235.0
5.7–6.5	SDSS Main	Jiang et al. (2016)	24	11240.0
5.7–6.5	—Overlap	Jiang et al. (2016)	10	4223.0
5.7–6.5	—Stripe 82	Jiang et al. (2016)	13	277.0
5.8–6.6	CFHQS Deep	Willott et al. (2010b)	1	4.47
5.8–6.6	—Very Wide	Willott et al. (2010b)	16	494.0
5.8–6.5	Subaru High-z Quasar	Kashikawa et al. (2015)	2	6.5
4.0–6.5	CANDELS GOODS-S	Giallongo et al. (2015)	19	0.047
6.5–7.4	UKIDSS	Mortlock et al. (2011)	1	3370.0
6.5–7.4	UKIDSS	Venemans et al. (2015)	1	3370.0
6.5–7.4	ALLWISE+UKIDSS+DECaLS	Bañados et al. (2018)	1	2500.0

all of the surveys lists in Table 2—the Quasar Photometry Table—we have in hand the selection functions, survey area, and survey depth—complete survey specifications from the homogenizing work presented in Kulkarni et al. (2019a), which are needed for the construction of the QLF and the BHMF. We note, however, that the BHMFs computed from observations and presented in this work for comparison with simulations are all derived from SDSS DR7. All of the currently available spectral data in QUOTAS are listed in Table 1, and the collated attributes are listed in the schematic in Figure 2. The publicly available resource that we utilize to collect data for QUOTAS is the NASA-IPAC Extra-galactic Database (NED; 2019).[18] This database contains basic information—e.g., coordinates and magnitude—on all observed astronomical objects and references to all published papers that describe the sources, and this includes reports of detailed follow-up work. Unlike other archives like MAST, SDSS, and SDSS VAC, the NED repository does not contain customized value-added properties. However, it contains the most comprehensive compilation of publication references for surveys and follow-up studies and is consistently updated. It is this feature of NED—the comprehensive bibliographic data—that we deploy for QUOTAS. In this first phase of our work, we derive the BHMF of collated SDSS quasars from the database as a function of cosmic time in redshift slices $z > 3$.

Due to the plethora of new observatories and instruments on the ground and in space due to come on line soon, we are entering an era of an unprecedented deluge of multiwavelength and multimessenger data on active and dormant SMBHs. From the ground, the Vera Rubin Observatory's projected output catalog is expected to provide 15 petabytes of data over 10 yr of operation (Jurić et al. 2017). The targeted Rubin AGN Survey, in particular, is expected to uncover $N \sim 10^7$ optical quasars (LSST Science Collaboration et al. 2009) from the near to the distant universe. JWST[19] has instruments with sensitivity in the mid-, near-, and far-infrared wavelengths (ranging from ~1–30 μm), and has started opening up an entirely new window into the early universe, revealing populations of early BHs and the first galaxies (Eilers et al. 2023). With accumulating JWST data in the coming years, earlier theoretical predictions stand to be tested (Pacucci et al. 2016; Natarajan et al. 2017; Barrow et al. 2018). The planned Laser Interferometer Space Antenna mission[20] to be launched in the early 2030s also offers exciting prospects for detecting gravitational waves from binary SMBH mergers that will most likely also be accompanied by multiwavelength electromagnetic counterparts (Kelley et al. 2018; Colpi et al. 2019; Taylor et al. 2019; Katz et al. 2020). Other relevant facilities currently providing new data on SMBHs include the Canadian Hydrogen Intensity Mapping Experiment;[21] the soon to be launched NGRST;[22] the Euclid Space Mission[23]; the Square Kilometer Array;[24] the Next Generation CMB Experiment Stage 4;[25] and the Advanced Telescope for High-ENergy Astrophysics.[26]

4.1. Simulations of Galaxy Assembly and Black Hole Growth

Used together, observations and numerical simulations have enabled detailed studies of many aspects of BH physics, including the accretion process, feedback from growing BHs, the evolution of the BH population and how BHs fit into the larger picture of structure formation in the universe (Springel et al. 2005). The power of simulations lies in the density of information that can be mined from them for all constituent particles—dark matter, baryons, and BH sink particles that enable connecting physical scales and regimes across cosmic epochs. Simulations also place BH growth in its full astrophysical context against the backdrop of dark-matter-driven galaxy formation in our cold dark matter dominated universe. Simulations have shown that the energy released

[18] https://ned.ipac.caltech.edu
[19] https://www.jwst.nasa.gov/
[20] https://www.elisascience.org/
[21] https://chime-experiment.ca/
[22] https://roman.gsfc.nasa.gov/
[23] https://sci.esa.int/web/euclid
[24] https://www.skatelescope.org/
[25] https://cmb-s4.org/
[26] https://sci.esa.int/web/athena

from the accretion process has the capacity to impact a wide range of spatial scales—from the smallest scales where general relativity and magnetohydrodynamic processes dominate, to the largest scales wherein outflows driven by accreting black holes and their feedback are relevant (Beskin & Kuznetsova 2000; Komissarov 2001; Gammie et al. 2003; Komissarov 2005; McKinney 2005; Johnson et al. 2011; Harrison et al. 2018). This coupling of scales might well hold the key to understanding the observed correlations between BHs, their host galaxies, and their parent dark matter halos. Due to the complexity of the involved physics and numerical cost of ab initio simulations of BH formation and subsequent evolution, simulators have adopted a strategy in which they focus on specific aspects of the problem. For instance, in a given simulation suite, a subset or all of the relevant physics involved, such as gravity, hydrodynamics (e.g., Beckmann et al. 2018), magnetohydrodynamics, and radiation transfer (McKinney 2005; Hawley et al. 2007; Tchekhovskoy et al. 2011; Liska et al. 2018), are followed and tracked. In addition to this, they need to account for the initial conditions for forming/seeding BHs and the level of detail with which the environment around them is rendered. Combining information from different simulations has allowed us at present to build a consistent picture spanning a large range of length scales. Take, for example, the role of jets driven by BHs; small-scale general relativistic magnetohydrodynamic simulations are needed to investigate the connection of the BH spin to the power of radio-emitting jets (Tchekhovskoy et al. 2010). Complementary simulations by Gaibler et al. (2011) bridge the gap to larger spatial scales and look at the impact of jets on galactic disks in hydrodynamical simulations. This permits assessing the impact of jets on the provision of fuel for the growing BH and the global star formation in the host galaxy. The latter simulations use idealized initial conditions for their setup and neglect the impact of feedback on the gas in the immediate vicinity of the central BH. Lower-resolution cosmological zoom simulations on the other hand are able to resolve individual objects such as the dark matter halos that host BHs with spatial resolutions of up to tens of parsecs (Debuhr et al. 2011), and at the same time they track the cosmological environment and its coupling with the BH (Beckmann et al. 2019). These simulations not only track BH relevant physics but also follow the evolution of the host galaxy in terms of its star formation, metal enrichment, and feedback from exploding supernovae (Beckmann et al. 2019; Wise et al. 2019; Latif & Khochfar 2020).

The ever-increasing number and diversity of observed BHs and AGNs requires numerical simulations that span a large parameter space of initial conditions for the formation of BHs and their associated cosmic environment. This kind of coupling with the larger-scale environment goes beyond what current zoom simulations can computationally provide. As quasars and AGNs are extremely rare objects, rare even compared to galaxies in the cold dark matter structured universe, simulations in which their formation and growth are tracked self-consistently over cosmic time are extremely computationally challenging. To alleviate this fundamental mismatch, trade-offs are often made between the simulation volume and resolution. For example, large-scale cosmological simulations of cosmic volumes of $\sim(100\ \mathrm{Mpc})^3$ with coarser spatial resolution of $\sim$kpc have been performed (Schaye et al. 2015). These simulations typically include subgrid models for the baryonic physics i.e., to model star formation, cooling of gas, chemical networks, stellar population evolution, supernovae feedback, and metal enrichment tracked via radiative transfer, but these versions can only include limited BH physics (Di Matteo et al. 2005; Robertson et al. 2006; Sijacki et al. 2007; Debuhr et al. 2011; Vogelsberger et al. 2014; Schaye et al. 2015; Sijacki et al. 2015; Habouzit et al. 2017; Weinberger et al. 2018). To match the statistical properties of AGNs detected in large observational surveys, even larger-scale simulations with volume $>(500\ \mathrm{Mpc})^3$ are needed (Bower et al. 2006). Given current computational limits, such simulations have necessarily been dark matter only and are able to only incorporate simplified, semianalytic models for the physics of BHs and galaxies that are added in by-hand post hoc. QUOTAS offers a novel way to test the association between observed BHs with their putative host galaxies and dark matter halos assumed in semianalytic models as well. The use of coeval simulation slices corresponding to redshifts where we have observational data permits unique tests of model assumptions including feedback from accreting BHs that appears to not only shape galaxies but also regulate the BH growth histories in quasars.

We demonstrate with examples the kind of new association studies that QUOTAS allows us to pursue. The first is via detailed tracking of the potential assembly history of the dark matter halo and the central BH that hosts an observed $z = 3$ quasar, shown in Figure 5. The second illustration involves using ML tools and their application to expand available simulation volumes—to use the simulated data from a smaller-volume cosmological simulation—the LEGACY mean box ($\sim 110\ \mathrm{Mpc}^3$)—to derive the full set of characteristic properties of halos that in turn allows us to expand the simulation box size while retaining all of the statistical properties of halos including their correlations and clustering. We demonstrate the power of ML by comparing the normalizing flow model derived characteristic properties of dark matter halos from the LEGACY mean box with the statistics from the larger-volume LEGACY $(1\ \mathrm{Gpc})^3$ Expanse box. Finally, showcasing the combination of simulated and observational data at just one epoch $z \sim 3$, with the constructed BHMF, we show how ML tools permit examining and interrogating assumptions that are integral to modeling the growth of SMBHs. In future work, we plan to develop additional ML tools in the next stage of our project that will allow us to further mimic larger simulated volumes that correspond to observational survey volumes and predict the putative locations for fainter as yet undetected quasars. Here we present the first key step in building up and exploring the BH–galaxy–halo connection that underpins the formation and evolution of all structures including the nonbaryonic (dark matter), baryonic (galaxies), and BHs.

Over the last two decades with the advent of powerful computational architectures, the size of data sets and complexity of numerical simulations has significantly increased. Typical simulated data sets are of Terabyte size and contain $N > 1024^3$ computational elements that represent gas, stars, black holes, and dark matter. For each of these elements, the equation of motion is solved taking into account the full physical information available in the simulation. In addition to containing phase-space information on momentum and spatial coordinates, computational elements also carry further information to compute the evolution of the system in time such as, e.g., the spin and mass of black holes, chemical composition of gas and stars, the present mass function of stars that is associated with an individual star particle, and internal energy

and density of gas. In addition, it has become standard practice to add evolutionary information to elements that are needed for subgrid computations during the simulation or post-processing of the data. These include the birth age of a star particle to track the evolutionary state of the underlying stellar population that it represents or the maximum temperature an individual gas element achieves during its evolution, to name just a few. It is evident that this list could be arbitrarily expanded, putting pressure on resulting data volumes. On top of this, simulation data sets include frequent snapshots with the full information of the state of all constituents—dark matter, stars, gas, and BHs—and their evolution through time. Further dramatic increase in data volumes is expected over the next decade with new peta-scale codes that will increase the number of computational elements significantly. To deal with such large data volumes, the optimal strategy is to focus on only a subset of the information. In our case, we are concerned with BH formation and evolution, which take place in gravitationally bound dark matter halos. Focusing only on material bound to dark matter halos reduces the data volume to roughly ∼20%–30%. One standard way to identify dark matter halos is based on linking dark matter particles by assuming a fixed linking length that is a fraction $b = 0.2$ times the mean interparticle separation (Davis et al. 1985). These so-called friends-of-friends halos contain further substructure in form of subhalos that may contain galaxies and BHs. Several different approaches have been utilized in the literature to identify subhalos based on their being gravitationally self-bound (Springel et al. 2001; Behroozi et al. 2013). Halo catalogs have become the standard way of providing the information from cosmological simulations to the community. They store all of the relevant information on collapsed halos: global properties of dark matter halos (e.g., total mass and angular momentum), properties of their constituent galaxies (e.g., star formation rate and metallicity), gas (temperature, pressure, and entropy), and BHs (e.g., mass, accretion rate, and spin). This set is complemented by information on all of the gas, stars, BHs, and dark matter in the region defined by the halo. Halo catalogs are easily ingested in SQL databases (e.g., Millennium simulation; Springel et al. 2006;[27] Illustris simulation; Nelson et al. 2015;[28] or the EAGLE project; Schaye et al. 2015[29]), which allow easy and fast access for population studies. A complementary approach is to use a web-based application programming interface, which allows for in-depth analyses of data beyond halo catalogs without the need of storing a local copy. Within QUOTAS we include simulation data using a combination of these approaches to maximize the ease of application for ML techniques.

5. QUOTAS: Database Design and Development

With QUOTAS, our goal has been to create an optical quasar repository of $z > 3$ quasars that includes a combination of raw and high-level products; covering a broad range of cosmic epochs, and colocating observational and simulated data of corresponding epochs (see Tables 1 and 2 for published references for the data). In this pilot project, focused on investigating the high-redshift quasar population, we curate data for these $z > 3$ sources in three distinct tables: the Quasar Spectra Table that contains BH mass estimates, spectroscopic data like line-widths used to derive independent BH masses and errors when available, and quantities like the accretion luminosities that also depend on the BH mass (source list with references listed in Table 1). The Quasar Photometry Table that contains all of the photometric data measured across bands (source list with references listed in Table 2) and the Halo Properties Table that includes the properties of the putative parent dark matter halos for the corresponding redshift slices spanning $3 < z < 7$. The distribution of the entire collated quasar population including all sources listed in Table 1 on the sky is shown in Figure 1. Figure 2 depicts a color-coded schematic of these data structures outlining the stored attributes: blue boxes denote the primary data source—the NED; the red Quasar Spectra Table (Table 1) contains the properties of individual objects collected from NED and multiple additional published sources, and the orange Quasar Photometry Table includes data from NED and surveys including their selection functions, survey depth, and survey areas (Table 2). We use the selection functions, survey depth, areas, and other details that have been computed to homogenize data from Kulkarni et al. (2019a). They present selection functions for the surveys listed in Table 2, and these are included in the selection function data structure in QUOTAS, as shown in the schematic of Figure 2. For science results showcased here in this first paper, that include the computation of the BHMF we use only the SDSS data. In QUOTAS, data attributes like the photometry, redshift, luminosity, etc. downloaded from NED are first flattened and combined into the respective attribute tables whose formats follow the column naming convention as described in(Shen et al. 2011).

5.1. Collating the Observational Data for QUOTAS

Our primary source data set is NED, which contains sources retrieved from several independent optical surveys, principally the SDSS. The following procedure was used to construct our database: first, we used the Astropy-affiliated package `astroquery` on NED to search and collect all known quasars at $z \geqslant 3$, and in the next step, we used the `By Parameters` search query to obtain basic information—object name, R.A., decl., redshift, photometry, spectra, images, and associated references—for each extracted object. From painstaking manual follow-up of published NED sources, we scoured tables in papers and the accompanying Vizier catalogs site[30] for raw and derived properties of individual quasars (Kulkarni et al. 2019b). We collected object names, positions, redshifts, luminosities, masses, line-widths, Eddington ratios, and references wherever available.

5.2. Measured and Derived Properties of BHs Powering Quasars

Mass and spin are fundamental properties of black holes. While mass estimates are available for several thousand sources at the present time (Kelly & Shen 2013; Peterson 2014; Vestergaard 2019), spin measurements are available only for a handful of sources (Nandra et al. 2006; Reynolds 2020). As the observed correlations are between SMBH mass and host galaxy properties, in this work, we restrict our investigation to BH masses. By construction the database structure of QUOTAS is

[27] https://www.g-vo.org
[28] https://www.illustris-project.org
[29] http://icc.dur.ac.uk/Eagle/database.php
[30] https://vizier.u-strasbg.fr/

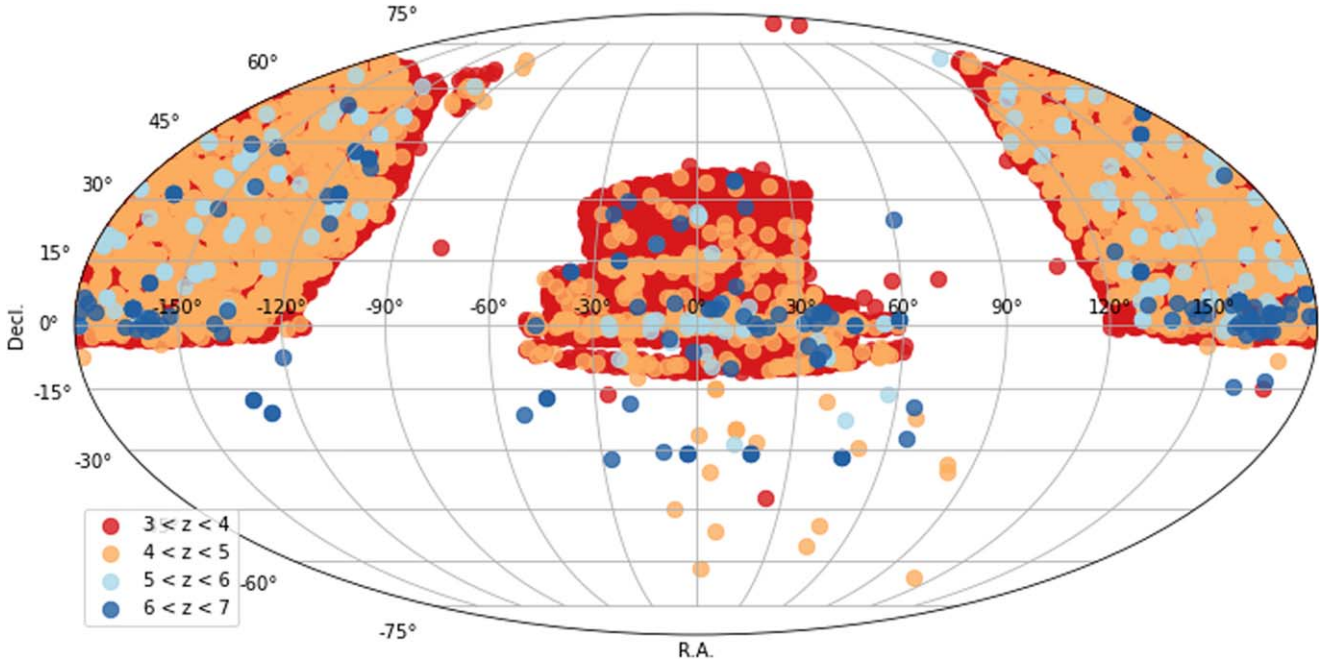

Figure 1. The distribution of all quasars collated in QUOTAS on the sky that are listed in Table 2. Sources are color-coded according to their redshifts: $3 < z < 4$—red; $4 < z < 5$—orange; $5 < z < 6$—cyan; and $6 < z < 7$—blue.

kept inherently flexible to accommodate data on BH spins as and when they become available, at which time spins can be seamlessly incorporated to the other stored source attributes. QUOTAS includes BH masses derived from multiple techniques where available. Many independent methods have been used to derive BH masses—including from the mapping of the orbits of individual stars, which is possible only for the Milky Way (Genzel et al. 1997; Ghez et al. 1998)—to modeling the orbits of bulge stars from imaging and spectroscopy as performed for nearby galaxies (Tremaine et al. 1994) and using measurements of the speed of rotating gas using water mega-masers as tracers of the mass (Miyoshi et al. 1995). It has been demonstrated with local data that the luminosity of an accreting SMBH can be used to infer its BH mass across a range of redshifts.

One widely adopted method for SMBH mass determination, available for many sources in QUOTAS assumes that the BLR is virialized and that the motion of the emitting clouds reflects the gravitational potential of the central BH (Blandford & McKee 1982; Peterson 1993). Under this assumption, the black hole mass $M_\bullet$ can be estimated via

$$M_\bullet = f\frac{V_{\rm vir}^2 R_{\rm BLR}}{G}, \tag{1}$$

where $V_{\rm vir}$ is the virial velocity, $R_{\rm BLR}$ is the size of the BLR, G is Newton's gravitational constant, and f is the virial coefficient that accounts for the geometry and kinematics of the material around the BLR (Shen 2013). The virial velocity can be estimated using the velocity dispersion derived from the width of observed BLR emission lines. The time-lagged broad-line response to variations in the continuum flux enables the measurement of the light travel time from the central ionizing source to the broad-line regions, which can be used to estimate $R_{\rm BLR}$. Acquiring these time lags from reverberation data is challenging, as it requires a long observational baseline, monitoring an accreting BH for 6 months to a year (Peterson et al. 2004; Grier et al. 2017). In QUOTAS, we include BLR determined masses and errors therein where available. There are many interesting questions relevant to observed variable quasars, like the sources referred to as "changing look" quasars. However for these, uniform sampled variability data are not currently available, though deep-learning methods have been recently used to fill in data gaps (Tachibana et al. 2020). Data on time variability of quasars is not included in this current pilot version of QUOTAS.

An alternative method for SMBH mass measurements, that is not predicated on the assumption of virialization of the BLR, is one in which line and continuum luminosities from the X-ray, ultraviolet, infrared, and optical wavelengths can instead be used to estimate the BLR size. For this purpose, continuum luminosities $L_{\rm cont}$ are often preferred over line luminosities, because they tend to yield a tighter correlation with the size of the BLR (Kaspi et al. 2000, 2005; Bentz et al. 2009). Reverberation mapping has revealed a tight correlation between the size of BLR and the continuum luminosity. This empirical scaling relationship is then used to derive black hole mass. Therefore, the continuum luminosity L, combined with the widths of broad emission lines (Δv), can be used to derive estimates of the mass of the black hole:

$$\log\left(\frac{M_\bullet}{M_\odot}\right) = a + b\log\left(\frac{L}{L'}\right) + c\log\left(\frac{\Delta v}{\Delta v'}\right), \tag{2}$$

where $L' = 10^{44}$ erg s^{-1}, $\Delta v' = 1$ km s^{-1}. In general, the values of pre-factors a and b depend on the choice of the luminosity and velocity dispersion estimators, and if a virialized BLR is assumed, then the coefficient of the line width is typically taken to be $c = 2$ (Shen & Liu 2012). Various cross-calibrations for this relation have been proposed (Vestergaard & Peterson 2006; Vestergaard & Osmer 2009; Shen & Liu 2012; Trakhtenbrot & Netzer 2012). Additional

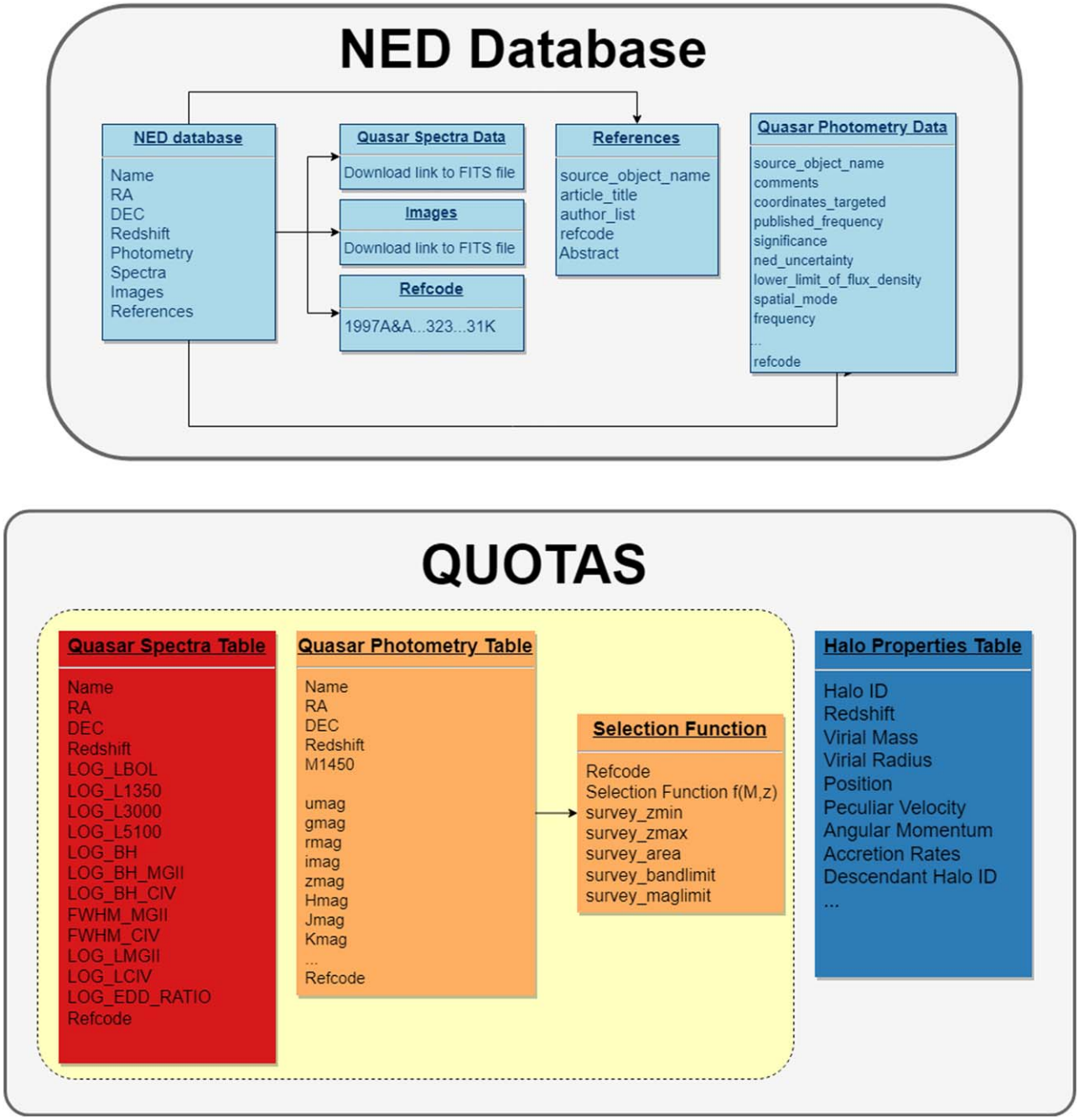

Figure 2. Schematic of the data structures and derived custom data tables that comprise QUOTAS. The information sourced from NED is shown in blue; the collated data in Quasar Spectra Table is shown in red; the Quasar Photometry data table is shown in orange; and the data table for Halo Properties derived from the simulated LEGACY catalogs is shown in blue. The stored attributes in each of the data tables is listed.

calibrations are needed to reconcile with estimates from other methods (Coatman et al. 2017) and for the determination of robust error bars. For example, the masses derived using the continuum luminosity are often calibrated with BH mass measurements obtained from reverberation mapping. Additional BH mass estimates are often made using individual lines like the Carbon-IV lines (Vestergaard & Peterson 2006). In QUOTAS we provide all available BH mass estimates as shown in the schematic (Figure 2) into the Quasar Spectra Table, computing them when relevant measured quantities are available. We also keep track of the error bars in the mass estimates when available. We focus on SMBH masses as the BHMF is a key component of theoretical models to study BH growth that occurs via both merging and accretion. Apart from the study of individual objects, population studies of quasars can reveal aspects of their evolution across multiple cosmic epochs and are crucial for building evolutionary models. Large quasar samples derived from ambitious astronomical surveys have enabled the study of the cosmological evolution of the QLF. A critical quantity for theoretical modeling, the QLF, encapsulates the growth history of quasars through accretion. The importance of the QLF was demonstrated in early modeling work, and it continues to be vigorously studied in the field (for example, see early work by Small &

Blandford 1992; Haehnelt et al. 1998; Hasinger 2008). In studies using the QLF, the growth via mergers is additionally accounted for with the modeling assumption that mergers themselves trigger additional accretion episodes. The QUOTAS database by design permits easy computation of both the QLF and BHMF as demonstrated in the workflow examples provided below.

The QLF is one of the key inputs to current conceptual models of BH growth (see reviews that outline the methodology by Volonteri 2012; Natarajan 2014, and references therein). The specifications of observational surveys need to be taken into account to determine the QLF. Observationally determined QLFs are then compared with those predicted by theoretical models of BH growth. Calibrated with data, the honed models are extrapolated and then used to predict QLFs down to fainter luminosities compared to current observed limits and out to larger redshifts than current detections. The most recent predicted QLFs extrapolated out to $z = 9$ derived from the combination of observational data and modeling can be found in Ricarte & Natarajan 2018b). The QLF provides a census of the number of sources at a given redshift and absolute magnitude M and is not easy to estimate. It is computed as follows:

$$\phi(M, z) \approx \sum_{j=1}^{N} \frac{1}{V_a^j} (\Delta M)^{-1}, \tag{3}$$

where ΔM is magnitude range $\{M, M + dM\}$, and V_a^j is the effective volume for a source j. This is usually estimated with the binned-volume method (Avni & Bahcall 1980), where the effective volume is estimated using:

$$V_a^j = \int_M^{M+\Delta M} \int_\Omega \int_{z_{\min}}^{z_{\max}} dM dz f(M, z) \frac{dV}{dz d\Omega}, \tag{4}$$

where $f(M, z)$ is the survey sample completeness, $z_{\min}$ and $z_{\max}$ define the redshift bin over which the QLF is averaged, $\frac{dV}{dz d\Omega}$ is the comoving volume element, and $d\Omega$ is the survey area. The minimum $z_{\min}$ is determined by the specifications of the survey. The maximum redshift $z_{\max}$ depends on individual sources and is defined by the maximum redshift out to which the source would remain detectable given the survey detection limit, as specified by the selection function of the survey. QUOTAS collates data for computing the QLF by merging observations taken by several independent surveys while keeping track of the individual selection functions of the surveys—these are stored in the Quasar Photometry Table (see the schematic in Figure 2).

The BHMF is a critical diagnostic that permits understanding of the mass assembly history of black holes over cosmic time. We briefly outline its computation from quasar survey data, and in this work, focus exclusively on deriving the BHMF from SDSS data to illustrate. The first step in computing the BHMF involves evaluating the detectable volume of quasars in a given survey, using its design specifications and the computed QLF as noted above. Using for instance, the SDSS DR3 data, we calculate the volume $V_{\rm bin}$, the effective volume within which a quasar of a given luminosity would be detectable. This is done with the data from our photometry data table. $V_{\rm bin}$ is defined as

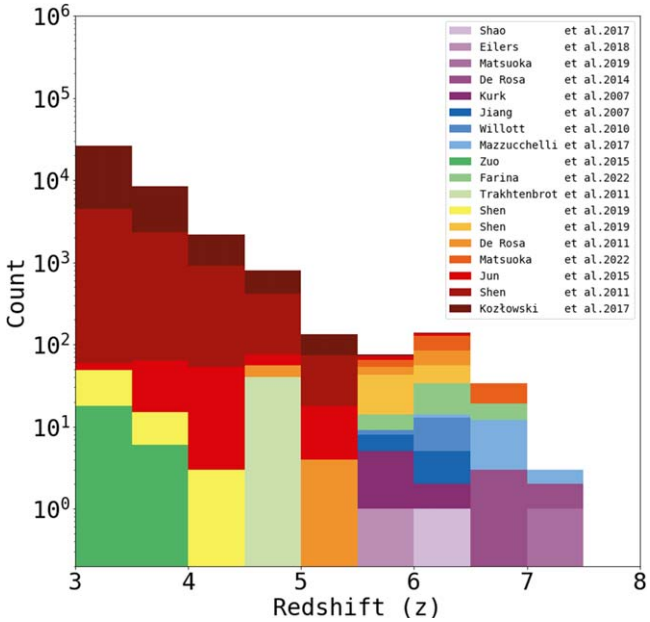

Figure 3. BH/quasar population demographics—the number of quasars plotted as a function of redshift collected from all surveys listed in Table 1 (referenced in the inset) that are in the QUOTAS database.

follows:

$$V_{\rm bin} = \int_{M_{\min}}^{M_{\max}} dM \int_{z_{\min}}^{z_{\max}} dz f(M, z) \frac{dV}{dz}, \tag{5}$$

where $f(M, z)$ is the survey selection function, for quasars as a function of magnitude M and redshift z; $M_{\min}$ and $z_{\min}$ are the magnitude and redshift of the observed quasar, respectively; $M_{\max}$ and $z_{\max}$ are the magnitude of the survey depth limit and the maximum detectable redshift given the luminosity of the quasar, respectively; and dV/dz is the redshift evolution of the volume element in the LCDM universe, which in turn is a function of the luminosity distance $d_L(z)$; $\Omega_{\rm m}$ and Ω_Λ. For the next step, we look up the corresponding BH mass for the quasar from the spectroscopic data table and then bin in logarithmic mass intervals. Finally, we combine the volume calculation to compute the volume density of quasars that falls within these mass bins. For each quasar in QUOTAS, its computed value of $V_{\rm bin}$ is also stored using the $f(M, z)$ for the survey in which it is detected. Every survey has a selection function, and as noted previously, we use the compilation from Kulkarni et al. (2019a) in QUOTAS. In this work, we use the selection function for SDSS quasars and compute the BHMF for SDSS quasars, which is shown in Figures 4 and 12. As part of our ongoing, continual assembly of QUOTAS, we are updating additional attributes to our tables as they become available. Our database is currently fully updated for SDSS quasars.

Instead of archiving only the reported SMBH mass, in QUOTAS all of the parameters used to derive the mass, like the line-widths, and line luminosities are also tabulated when available. The redshift distribution of quasar counts in the database is plotted in Figure 3. We also include serendipitously discovered high-redshift quasars into QUOTAS. We have carefully assigned unique object identifiers and recorded relevant information like the survey flux limits, redshift limits, sky area, and magnitude limit for sources in the database whenever available, as described above. To date, QUOTAS contains the

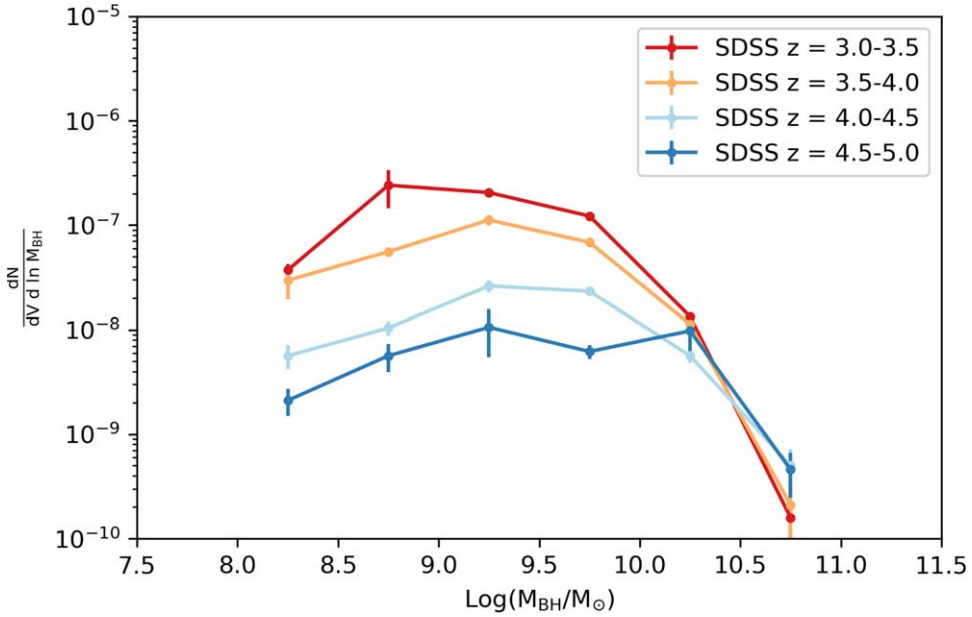

Figure 4. BHMF for SDSS computed from QUOTAS: the BHMF is an important quantity for calibrating/constraining theoretical models of black hole mass assembly and growth, and it can be derived in one step from the collated tables in QUOTAS when all relevant attributes are available to estimate quantities in Equations (3)–(5). The BHMFs queried from QUOTAS sourced from the SDSS survey (SDSS DR7 for the $z < 5$ quasars) are plotted.

most comprehensive and growing compilation of optical quasars and contains in excess of 30000 unique objects. Quasars in Table 2 in our database contain sufficient parameters to compute their BHMFs and the QLFs, the two key quantities needed for comparison with models of the assembly history of black holes. In this work, we present comparison of the observationally determined BHMF with that derived from the corresponding simulated slice in QUOTAS for SDSS quasars, as this is the largest uniform sample that permits straightforward comparison with simulations (see Section 7.3 and Figure 12).

5.3. Collation of Simulation Data for QUOTAS

We ingest simulated data from the LEGACY suite[31] runs, a set of large-scale cosmological dark-matter-only simulations (Meriot et al. 2022) into QUOTAS. We outline the key properties of the runs from the LEGACY suite used in this pilot project work in Table 3. The LEGACY simulations consist of a parent run with a $(2.3\,\mathrm{Gpc})^3$ volume of 2048^3 particles at a mass resolution of $M_{\rm DM} = 5.43 \times 10^{10} M_\odot$ (referred to in Table 3 as the parent run) accompanied by a set of higher-resolution zoomed-in simulations with a volume of $(119\,\mathrm{Mpc})^3$ and mass resolution of $M_{\rm DM} = 1.32 \times 10^7 M_\odot$, which corresponds to an effective particle number count of $32{,}768^3$ (labeled as the mean run in Table 3) and another further zoomed-in simulation with $V = 1\,\mathrm{Gpc}^3$ and $M_{\rm DM} = 6.78 \times 10^9 M_\odot$ to capture sufficiently large-scale modes (Klypin & Prada 2019; labeled as the "expanse run" in Table 3). The initial conditions of the simulation are created using the public code MUSIC (Hahn & Abel 2011), adopting the following values for cosmological parameters based on the results from the WMAP satellite (Hinshaw et al. 2013): $\Omega_\Lambda = 0.715$, $\Omega_M = 0.285$, $\Omega_{\rm b} = 0.045$, $h = 0.695$, and $\sigma_8 = 0.828$. For each of the simulations (parent and zoom-ins), gravitationally bound structures and substructures were determined using the halo finder ROCKSTAR (Behroozi et al. 2013). Particles and halo/subhalo properties are stored in 300 snapshots from $z = 20$, covering the redshift range of interest in this study, $z \geqslant 3$, though slices are available down to $z = 0$.

Table 3
List of LEGACY Suite Runs Used in This Work

Run Name	Box Size (Mpc/h)	DM Particle Mass in $M_\odot$	Effective Number of Particles	Comment
Parent	1600	5.43×10^{10}	2048^3	Full Box
Mean	83	1.32×10^7	32768^3	Zoom-in Box
Expanse	703.1	1.32×10^7	4096^3	Zoom-in Box

Note. Halo catalogs from the parent, mean, and expanse runs are included in the halo properties table in QUOTAS.

The halo catalogs provide detailed information—81 individual attributes for each identified halo—and all of these characteristic parameters are stored in QUOTAS. We also store detailed evolutionary information for each halo in finely spaced time steps for each snapshot, which is used to generate merging histories for individual halos. This data structure allows us to uniquely identify for each halo its progenitors and ancestors. This information provides further leverage in establishing links to evolutionary models of BH growth providing the BH–halo connection piece. QUOTAS permits association of the host dark matter halo hierarchy for each observed quasar, an essential link for our goal of probing the underlying BH–galaxy–halo connection. Assuming for instance, that the most-massive simulated dark matter halo at $z = 3$ hosts the brightest observed quasar at $z = 3$, from the data in QUOTAS we can trace its detailed merger history as shown in Figure 5. In the case shown, the host dark matter halo has a mass of $9.3 \times 10^{12}\ M_\odot$ at $z = 3$.

The QUOTAS platform and developed tools are publicly available for open access at the Google Kaggle platform. The

[31] Details of the LEGACY suite runs are available at: https://sites.google.com/site/tmoxwiki/legacysuite/infolegacy.

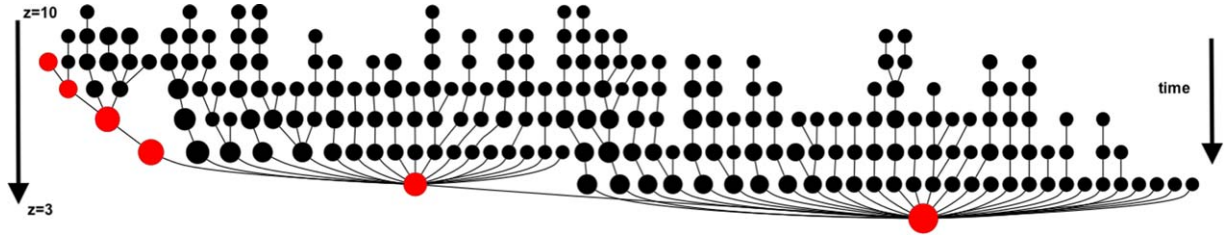

Figure 5. Assembly history studies permitted by QUOTAS. The merger and assembly history of a dark matter halo that likely hosts the brightest observed quasar at $z = 3$ tracked through time generated from QUOTAS is illustrated here. The size of the circle is proportional to the logarithm of the dark matter halo mass. Red circles mark the most-massive progenitor of a halo at each time step. In this plot, only the progenitor halos with mass $\geqslant 5 \times 10^9\ M_\odot$ and only time steps with $\Delta z \sim 1$ are shown.

complete observational data and simulation data are available at https://www.kaggle.com/. The codes are available at DOI: 10.34740/kaggle/ds/3165166 (Natarajan et al. 2023), and examples of use are available at the same location in the data-exploration-with-quasarnet section.

6. Demonstration of Ease of Use and Workflow: Querying QUOTAS

To facilitate querying the database, searches can be performed with SQL-like `BigQuery`on `Google Public Dataset`on Kaggle.[32] With the data hosted on the public Google Kaggle Platform, researchers can directly access and interact with the data through a `Colab`notebook available at DOI: 10.34740/kaggle/ds/3165166 that we provide and perform analyses directly on the data set. The subset of photometry data for instance, can be accessed for all quasars with the following query:

```
SELECT
qtable.*,
ptable.*
FROM `black hole-database.ml_test.qso` as
qtable
JOIN `black hole-database.ml_test.photome-
try` as ptable
ON qtable.object_name = ptable.source_
object_name
```

The Quasar Spectra Table and the Quasar Photometry Table are stored similarly; therefore, simply replacing "qtable" with "qlftable" will yield the existing photometry data for all quasars across surveys.

```
SELECT
qlftable.*,
ptable.*
FROM `black hole-database.ml_test.qso` as
qlftable
JOIN `black hole-database.ml_test.photome-
try` as ptable
ON qlftable.object_name = ptable.source_
object_name
```

The Quasar Photometry Table includes the selection function for every source from Kulkarni et al. (2019a), and with it we can calculate the effective volume for each source and therefore generate QLFs in a straightforward fashion.[33] Next, we illustrate ease of use and utility of QUOTAS and describe the computation of the BHMF. To derive the BHMF, we use the measurements of BH masses and the flux limits for every observed source compiled in QUOTAS, to calculate the spatial density of objects. In quasar surveys, every detected quasar can be assigned a detection probability and a detection volume in a given redshift bin, in accordance with the survey detection limit and its completeness. These quasar properties are combined with the derived spatial density to compute the BHMF (Willott et al. 2010a; Jiang et al. 2016). This provides a swift estimate of the BHMF with a single line query to QUOTAS. We illustrate the ease of use by plotting the resulting estimate of the BHMF in Figure 3, which is volume complete. Kelly et al. (2010) derived the BHMF from the SDSS DR3 data release; here we demonstrate the ease-of-use to generate the updated BHMF with the full current available SDSS DR7 quasar data. The plotted Figure 3 clearly reveals that there appears to be an upper limit for BH masses across redshift bins. Interestingly, such a limit is theoretically predicted and the rendering of data compiled in QUOTAS aligns with modeling expectations, as per Equation (7) in Natarajan & Treister (2009).

7. First Science Results from QUOTAS

We present early science results from (i) mining the observational data stored in QUOTAS reporting on findings of $z > 3$ SDSS quasar properties and correlations with their host properties; (ii) illustrating the use and power of ML to expand dark-matter-only simulation volumes that faithfully reproduce the underlying statistical properties of halos; (iii) showing how ML can be used to "populate" these larger dark-matter-only simulation volumes after training on full hydrodynamic simulations that include baryonic physics; and (iv) comparing SDSS quasar properties with the ML derived properties from the "populated" simulation. Finally, we demonstrate the power of ML to interrogate key assumptions of BH growth, in this case, the accretion paradigm, in the subgrid models that are currently implemented in simulations. We focus primarily on $z \sim 3$ as the benchmark redshift to explore and compare properties of observed SDSS quasars ($3 < z < 3.5$) and the corresponding $z = 3.25$ LEGACY simulation output stored in QUOTAS.

For this work, the homogenized, largest observational data set that is available is from the SDSS. For the SDSS, we have access to all survey properties including the selection function and survey specification, and we have in hand all attributes that permit easy calculation of the QLF and the BHMF in QUOTAS. For data collated from other surveys, listed in Table 2, selection functions are curated from Kulkarni et al. (2019a). In this first paper we showcase results for the BHMF

[32] DOI: 10.34740/kaggle/ds/3165166

[33] For the plots shown here, the following values of standard cosmological parameters were adopted: $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, and $H_0 = 70$ km s^{-1} Mpc^{-1}.

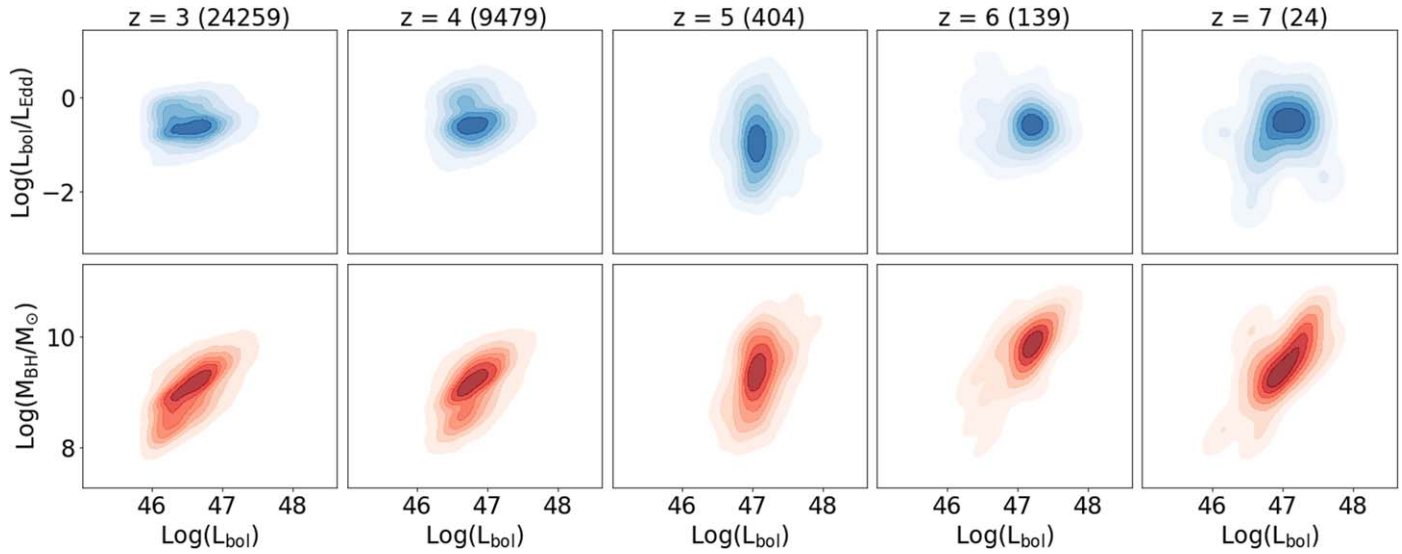

Figure 6. Key derived statistical properties of the population of observed $z > 3$ quasars relevant to modeling their growth history over cosmic time—BH mass, bolometric luminosity, and Eddington ratio in five redshift slices for sources in Table 1. The number in parentheses denotes the number of observed quasars in the redshift bin. These data were drawn from QUOTAS collated in the Quasar Spectra Table and the Quasar Photometry Table per the schematic shown in Figure 2. Each column corresponds to redshift bins ranging from: $3 < z < 3.5$; $3.5 < z < 4.5$; $4.5 < z < 5.5$; $5.5 < z < 6.5$; and $z > 6.5$, respectively.

computed for SDSS quasars. With all of the information on selection function and completeness from SDSS and the large number of sources, we are able to mimic the relevant cuts in our simulated LEGACY slice that is populated with quasars at $z = 3$.

7.1. Exploration of SDSS Quasar Properties and Correlations

To start with, we demonstrate exploration of the observational data stored in the Quasar Spectra Table and Quasar Photometry Table convolved with the corresponding selection functions in QUOTAS. In Figure 6, we plot derived BH properties: BH mass and the Eddington ratio for the SDSS quasar population from a sample query to QUOTAS in five redshift bins ranging from $3 < z < 7$. New insights from this exercise include a noticeable precipitous decline in the number of quasars with redshift that constrains the general shape of the QLF and its slope as a function of redshift after taking into account the effect of incompleteness and our preferential detection of the brightest quasars with increasing redshift. In contrast to what we find at these higher redshifts, at lower redshifts ($z < 2$ for instance), although a similar decline in the abundance of quasars is detected, this does not reflect our inability to detect sources that exist but rather signals the true deficit of actively accreting BHs at these late times when galactic nuclei are gas-poor and much of the gas in galaxies has already been consumed by star formation. Bright as the currently detected quasars are, we note from the top panel of Figure 6 that the majority of the detected quasar population is still accreting at sub-Eddington rates. We caution that the dusty super-Eddington sources that are yet to be uncovered at these early epochs are missing from our current optical data sets, and therefore also in QUOTAS. We reiterate that QUOTAS includes all X-ray detected Type I AGNs, since they all have optical counterparts. Given the current limits of X-ray instruments on Chandra and XMM-Newton, accreting sources peter out at $z > 3$ where the Type II AGN inventory is expected to pick up. This is where we expect JWST and NGRST to start filling in the landscape in the very near future.

As a function of redshift, as expected, our current observational surveys detect progressively brighter quasars. Reading off the peaks of the contour plots in Figure 6, we see that the bolometric luminosity of the average detected quasar at $z = 3$ is $3.6 \times 10^{46}\,\mathrm{erg\,s^{-1}}$, while by $z = 4$ it is an order of magnitude higher at $2.8 \times 10^{47}\,\mathrm{erg\,s^{-1}}$. The typical measured BH mass powering the corresponding average detected quasar exceeds $10^9\,M_\odot$ at all redshifts beyond $z > 3$. This demonstrates clearly that we sample different subportions of the QLF and BHMF as a function of redshift, and are preferentially accessing the most extreme objects at earlier epochs. This biased observational view severely limits our ability to model the growth history of the overall BH population. And it is this limitation that we would like to circumvent with ML techniques that will permit us to predict and hence detect the lower luminosity, more representative quasar population at every epoch. This will finally yield a more complete picture of BH growth and galaxy assembly.

As noted earlier, correlations between the properties of central SMBHs and their host galaxy luminosity, stellar mass and bulge velocity dispersion are well documented for the dormant, local SMBH population. QUOTAS offers a unique opportunity to explore and discover new correlations between active SMBHs and their host galaxies. Turning now to Figure 7, we demonstrate once again the ease of visualizing data and the ability to explore correlations between multiple measured quantities from QUOTAS. Several new science results and insights are obtained from just this single analysis. Take for instance, the anticorrelation seen between the FWHM of C IV lines versus Eddington ratio plotted in the second panel on the bottom row. In a recent paper, Marziani et al. (2019) discussed that a correction needs to be applied to the FWHM of C IV derived BH masses that depends on the Eddington ratio for a lower-redshift population of quasars $0 < z < 3$. Our plot from QUOTAS clearly demonstrates and supports that such a correction may be required and suggests that the Marziani et al. (2019) finding also extends to higher-redshift quasars at $z > 3$ studied here. With these set of first results, we illustrate the power of QUOTAS to reveal correlations and their evolution

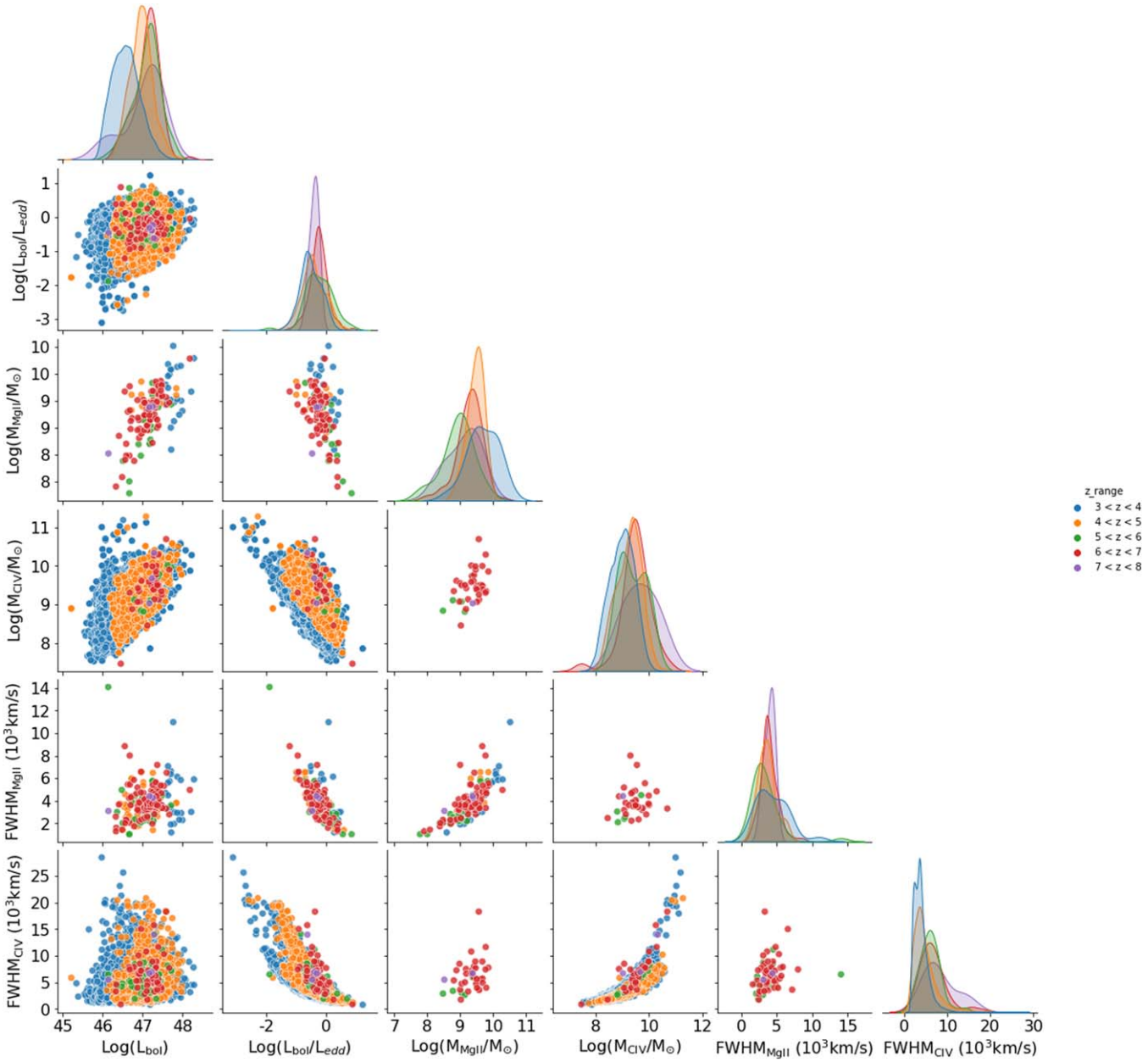

Figure 7. Preliminary exploration of correlations using QUOTAS from information in Table 1: correlations between the bolometric luminosity; the Eddington ratio; BH mass determined from MgII line-widths; BH mass determined from C IV line-widths; the FWHM of Mg II lines and the FWHM of C IV lines queried from the Quasar Spectra Table. The data are color-coded according to redshift as follows: $3 < z < 4$ (orange); $4 < z < 5$ (blue); $5 < z < 6$ (cyan); and $6 < z < 7$ (pink). Along the diagonal, we plot the kernel density estimations of the distributions for each of these parameters. The off-diagonal panels show the cross-correlation between different parameters.

Table 4
Examples of Dark Matter Halo Properties Stored in QUOTAS

Name	Units	Description
z	⋯	redshift
$M_{\rm vir}$	$M_\odot$	virial mass
$r_{\rm vir}$	kpc	virial radius
$x,\ y,\ z$	Mpc	position
$v_x,\ v_y,\ v_z$	km s^{-1}	peculiar velocity
$J_x,\ J_y,\ J_z$	$M_\odot$ Mpc km s^{-1}	angular momentum
$\dot{M}_{\rm acc}$	$M_\odot$ yr^{-1}	accretion rate onto halo
$I_{\rm desc}$	⋯	descendant halo ID
$T/\lvert U\rvert$	⋯	virial parameter
$x_i,\ y_i,\ z_i$	Mpc	position of each particle
$v_{x,i},\ v_{y,i},\ v_{z,i}$	km s^{-1}	peculiar velocity of each particle

with redshift as seen in the diagonal kernel density plots as a function of cosmic epoch.

7.2. Tailored ML Tools to Expand Simulation Volumes

We show how the explicit association of observed quasars at $z > 3$ with their putative parent dark matter halo sites and their history shown in Figure 5 can be used for improved modeling and providing insights into the BH–galaxy–halo connection. However, as noted previously there are several important limitations that prevent direct comparison of observational data with simulations. We start by deploying ML to circumvent the limitations arising from small simulation volumes and resolution due to the prohibitive computational cost of including

baryonic physics in larger box simulations. It is numerically inexpensive to scale up and perform large volume N-body dark-matter-only simulations, and generate halo catalogs from volumes that would permit detailed comparison to the comparable volumes probed by observational surveys. This is where we use ML to train on the smaller-volume full hydrodynamical simulation that includes baryonic physics, in this case the Illustris-TNG300[34] simulation to then "populate" a larger-volume dark-matter-only simulation with SMBHs from the LEGACY suite stored in QUOTAS. As a first step in this process, we perform an important task, i.e., use ML to extract the key, characteristic statistical properties of the dark matter halos from a subset, the training set, of dark matter halos, in the LEGACY mean box (see Table 4) and then see if we can successfully recover the overall statistical properties—beyond the training set—in the larger-volume (lower-mass resolution) LEGACY $(1\,\mathrm{Gpc})^3$ Expanse box.

Since our goal is to explore the BH–galaxy–halo connection, we ultimately need to relate simulated dark matter halos, the putative sites for BH formation and growth, to the population of observed quasars at every epoch. To conduct the association studies between observed quasars and their potential parent dark matter halos, we require simulated volumes to match the large observational survey volumes. Directly simulating corresponding volumes with the full baryonic physics implemented is not computationally possible currently. Given that quasars are rare objects, in order to circumvent the restrictions of the simulated box size for our planned association studies, we present the application of the ML normalizing flow algorithm to expand our sample of simulated halo catalogs.

Therefore, we first utilize ML methods to generate mock simulated surveys that are better volume matched to the data. To do so, we first need to characterize ensemble properties of the simulated halo population to mimic the halo population in a significantly larger-volume box that successfully reproduces its statistical properties. The underlying statistical properties of the halo population of the simulated LEGACY mean run at $z = 3.25$ stored in QUOTAS are derived using the class of generative ML models referred to as normalizing flows (Rezende & Mohamed 2016; Kobyzev et al. 2019; Papamakarios et al. 2019). Normalizing flow provides a nonparametric fit to the joint data distribution and offers a powerful way to generate new halos by sampling latents. The observed variables X, in our case, the properties of each simulated halo such as its mass, angular momentum (modeled via its spin parameter), mass accretion redshift (redshift at which half the mass of the halo is assembled), etc. are mapped to a latent variable $z \sim p_z(z)$ of the same dimension drawn from a known distribution (usually Gaussian) through an invertible transformation parameterized by a neural network, i.e., $z = f_\theta(X)$. The marginal likelihood of $p(X)$ can be obtained with the changes of variables given by:

$$p_X(X;\ \theta) = p_z\left|\det\frac{\partial z}{\partial x}\right| = p_z\left(f_\theta^{-1}(x)\right)\left|\det\left(\frac{\partial f_\theta^{-1}(x)}{\partial x}\right)\right|. \quad (6)$$

The model is trained by maximizing the log likelihood of the estimate under the observed sample distribution X. Once trained, the model can be used to generate as yet unseen data by sampling the latent (z') space from the base distribution and transforming it through the invertible transform $X' = f_\theta^{-1}(z')$. Although, several other kinds of transformations have been proposed (Dinh et al. 2017; Kingma et al. 2017; Papamakarios et al. 2018), we use a version of normalizing flows as they are exactly invertible and are particularly well attuned to deriving the posterior probability distributions from which larger-volume statistics for the halo population can be generated for our future association studies. In other astronomical applications, normalizing flows have been shown to successfully constrain the posterior estimate of the distance estimates from Gaia data (Cranmer et al. 2019), performing likelihood free Bayesian inference on cosmological parameters (Alsing et al. 2019) and in constraining the re-ionization history of the universe (Hortúa et al. 2020). We adopt the Real Non-Volume Preserving model (Dinh et al. 2017) to fit our simulated halo properties tables. The results of this crucial first test step in the combined exploitation of simulated and observed data are shown in Figure 8. The properties of the generated host halo population (orange) from sampling the trained normalizing flow model on the LEGACY mean runs are distributed very similarly to those of true halo population (blue) stored in the Halo Properties Table from the larger-volume zoom-in, the LEGACY $(1\,\mathrm{Gpc})^3$ Expanse cosmological simulation at $z = 3.25$ in QUOTAS. This coherence in recovered halo properties demonstrates that the trained model can be used for downstream ML tasks to explore the quasar–host halo connection exploiting the large observational data sets currently in hand and expected deluge in the near future.

As seen in Figure 8, the generated host halo population reproduces all properties of interest of the true halo population from the LEGACY $(1\,\mathrm{Gpc})^3$ simulation slice at $z = 3.25$ in QUOTAS extremely well. The use of normalizing flows therefore permits us to train on a smaller-volume simulation (the $z = 3.25$ slice from the LEGACY mean run) and accurately extend our simulated catalog to correctly generate the rarer dark matter halos that would only appear in significantly larger-volume simulations (in this case, the $z = 3.25$ slice of the LEGACY $(1\,\mathrm{Gpc})^3$ Expanse run). ML permits us to extend and expand our reach to capture the parent dark matter hosts of the rare, bright observed high-redshift quasars that would only appear in significantly larger-volume simulations. In future work, we plan to use this trained model to generate halo catalogs in volumes that correspond more closely to those of observational surveys for further detailed association studies.

7.3. Application of ML Tools to Probe the BH–Galaxy–Halo Connection

In the next step, we employ an ML algorithm to populate the dark-matter-only Legacy simulation volume with accreting BHs. To do this we train a model to learn the relationship between BHs and their host halos using a full physics simulation. The publicly available IllustrisTNG hydrodynamical simulations (Nelson et al. 2019) incorporate baryonic processes by way of subgrid model recipes for galaxy formation and evolution in which BHs grow via accretion and mergers (Weinberger et al. 2017). Full information about black holes and their associated parent halos is therefore available to train an ML model. As this is a supervised learning problem with tabular data, we use a random forest algorithm (Breiman 2001).

[34] Details of the Illustris-TNG300 set of simulations can be found at: https://www.tng-project.org/.

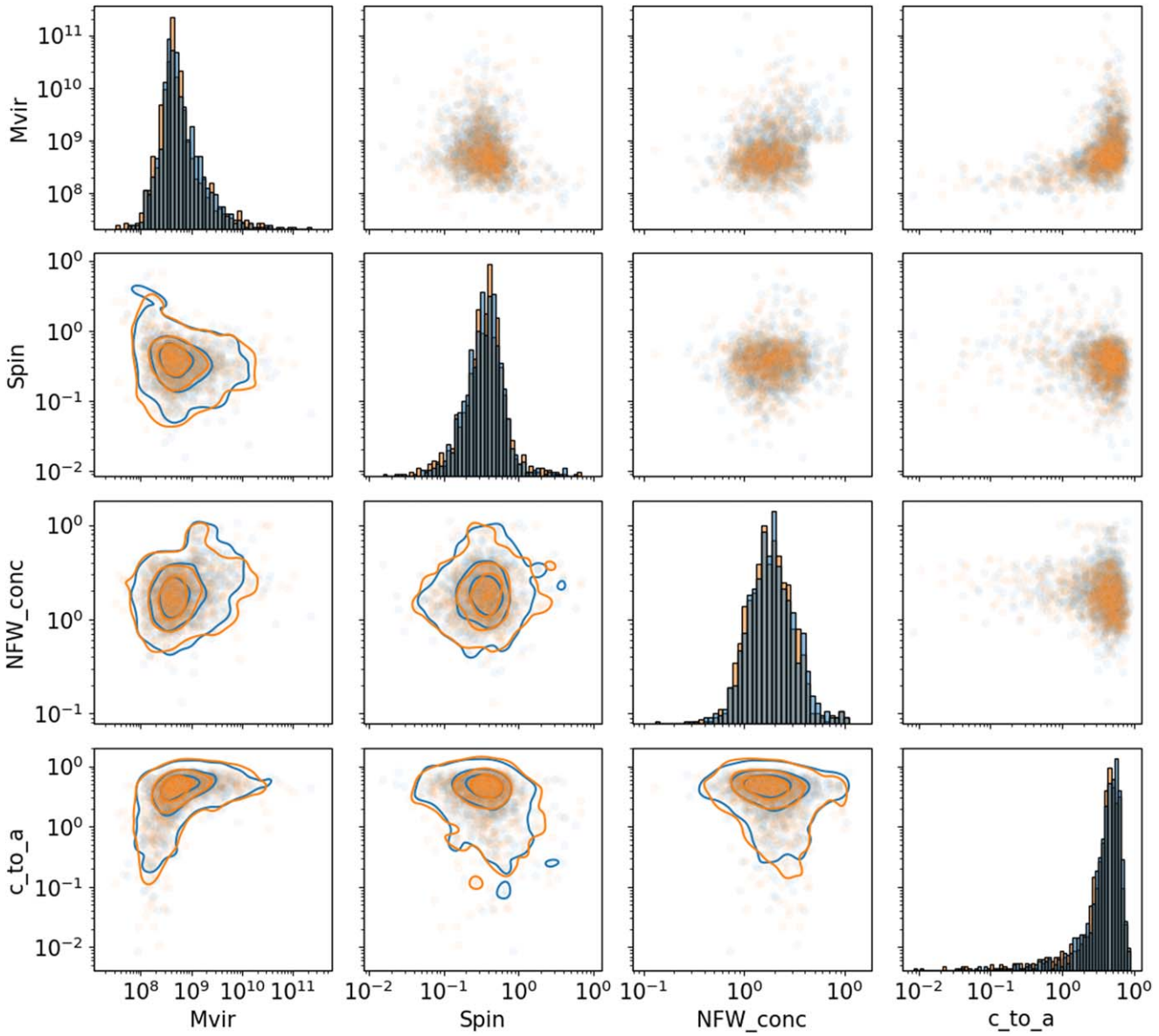

Figure 8. Derived underlying statistical properties of dark matter halos at redshift $z = 3$ using the normalizing flow ML algorithm. These prior distributions for halo properties are derived from the Halo Properties table in QUOTAS. The distributions of "true" halo properties are colored in blue; the orange points are generated by sampling with the trained normalizing flow model. Here we select the halo mass, halo spin, halo NFW concentration, and halo ellipticity as key characteristic properties to train the normalizing flow model. This is the first of a set of many tailored ML tools that will be used to fill in the key missing link to probe the black hole–host galaxy–parent dark matter halo connection.

This approach uses an ensemble of decision trees to make a prediction. Each decision tree is constructed top-down from a root node. At each node, the data is split into two bins based on the values of its input parameters. The splits are chosen such that the weighted average of the mean squared error (MSE) of the two bins is minimized. This partitioning of the data results in each leaf node at the bottom of the tree containing a small subset of the data, where almost all members of the subset have a similar output value. Predictions from decision trees are based on the assumption that test data points will have a similar output value to the other members of the leaf node it is placed into. A random forest is made up of a number of decision trees. There is a bootstrapping procedure such that each decision tree within the forest is trained on a randomly generated subset of the training data. Further randomness is added in that for each split only a subset of input features can be used. The prediction from a random forest is the average prediction of its component decision trees. A major advantage of random forests is that they are significantly less prone to overfitting data compared with a single decision tree. This results from the randomness added when training the individual decision trees.

Specifically we choose the extremely randomized tree (ERT) ensemble algorithm, which has been shown in several previous works (e.g., Kamdar et al. 2016; Moews et al. 2021) to be suitable for this task. The ERT algorithm adds in additional randomization by computing a random split for each feature at each node, rather than the optimal split. We follow the approach in McGibbon & Khochfar (2022), who used halo properties from a wide range of redshifts as model inputs, and showed how this can provide information about the times when galaxy properties are determined. As we are predicting properties at a high redshift, and so do not have a large

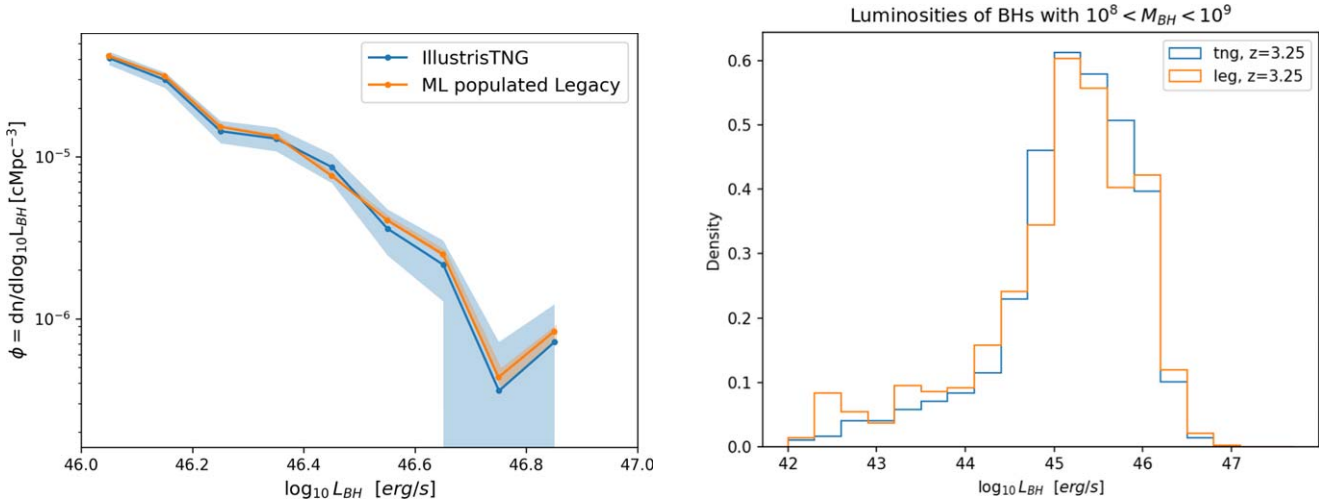

Figure 9. Left panel: comparison of the QLF derived for sources with bolometric luminosity $L > 10^{45}$ erg s^{-1} from the Illustris-TNG300 trained ML used to populate the larger LEGACY 1 Gpc Expanse run box from the $z = 3.25$ slice. Right panel: comparison of the census of accreting SMBHs—the number density of quasars as a function of bolometric luminosity in the training set (the Illustris-TNG300 box) and the larger-volume LEGACY 1 Gpc Expanse box at $z = 3.25$. There is excellent agreement showing that the ML algorithm (ERT random forest) deployed here accurately reproduces the number statistics of SMBHs in the larger simulation box slice.

number of prior snapshots, we adopt the base model from McGibbon & Khochfar (2022). We train the model using the IllustrisTNG300 simulation volume, with the BH mass and accretion rate as our target variables. We then apply the trained model to the dark-matter-only Legacy (1 Gpc)^3 Expanse simulation. This gives us a catalog with approximately 40x the number of black holes that are present in the largest IllustrisTNG simulation.

To test the robustness of this ML procedure for our specific task, in Figure 9, in the left panel, we plot the QLF derived from the training set—the Illustris-TNG300 box—with that from the ML populated LEGACY (1 Gpc)^3 Expanse box at $z = 3.25$. We note that the QLF is in excellent agreement for the full sample and the subsample, demonstrating that the deployed ML algorithm successfully enables us to expand simulation volumes while accurately capturing the occupation fraction of accreting BHs. In the right panel of Figure 9, we compare and find excellent agreement in the number density of accreting SMBHs in the Illustris-TNG300 slice and the larger-volume LEGACY (1 Gpc)^3 Expanse box. Further, in Figure 10, we dissect the accreting SMBH population in these two simulations in three BH mass bins and once again find excellent agreement. Notable in these histograms is the bimodal luminosity distribution that starts to emerge at higher BH masses (as seen in the middle and right panels of the plot).

Confident that the extended random forest ML algorithm has statistically populated the LEGACY (1 Gpc)^3 Expanse box in concordance with the Illustris-TNG300 simulation, we delve deeper into studying BH–parent dark matter halo connection. Therefore, next, we proceed to compare the BHMF derived from observations—SDSS quasars in the redshift bin $3 < z < 3.5$ in this instance—from QUOTAS with that derived from the ML populated LEGACY (1 Gpc)^3 Expanse box. However, before doing so, we need to pay attention to the fact that the optically bright SDSS quasars collated in QUOTAS are a subset (Type Is) of the full population of accreting black holes (AGNs), and therefore we need to appropriately scale the accreting black hole population to determine the fraction of Type I AGNs in the simulation by classifying them. We partition the simulated population into Type I and Type II AGNs using an independent empirical finding by Ananna et al. (2019). First, detailed multiwavelength studies of AGNs find that by $z \sim 3$, close to 90% of the sources are obscured Type II AGNs (not optically bright quasars), and their fraction decreases slightly and monotonically as function of X-ray luminosity in the range $L_X \sim 10^{42}$–10^{46} erg s^{-1} (for the latest census, see Ananna et al. 2019 and references therein). Adopting this empirically derived fraction from Ananna et al. (2019), with an X-ray to bolometric luminosity correction factor, we scale the number of accreting BHs populated in the LEGACY (1 Gpc)^3 Expanse box to mimic the SDSS quasars (Type I AGNs) at this epoch. With this selection implemented, we plot the NFW concentration parameters for the dark matter halos that host quasars (the Type I AGNs) in the LEGACY (1 Gpc)^3 Expanse slice at $z = 3.25$ in Figure 11. Combining AGN abundance with clustering measurements from large X-ray surveys, Powell et al. (2018, 2020) reported that Type I AGNs occupy dark matter halos with concentration parameters $< c_{\rm NFW} < 10.14$ and Type II AGNs are found in $10.14 < c_{\rm NFW} < 100$. Plotting the concentration parameter distribution as a function of bolometric accreting luminosity for sources in the LEGACY (1 Gpc)^3 Expanse box, we find that our selection of Type Is in the box is in excellent agreement with empirical findings as shown in Figure 11. We note that all of our sources selected as quasars are indeed consistent with the empirically determined classification criteria derived by Powell et al. (2018), as all selected sources lie leftward of the white dashed line that marks $c_{\rm NFW} < 10.14$) in the Expanse box.

In order to compute the BHMF, we need the selection function and survey specifications including depth and completeness, which are available for SDSS in QUOTAS. In Figure 10, we compare the BHMF of SDSS quasars and the equivalent simulated population from QUOTAS. The mass functions match extremely well at the turnover point, which corresponds to a bolometric luminosity cut of $L \sim 10^{46.5}$ erg s^{-1}. In Figure 12, the dashed vertical line marks the location of BHs that would be accreting at the Eddington limit. Therefore, we see clearly that every accreting BH with mass $M_{\rm bh} > 10^{8.4} M_\odot$ in the simulation is accreting at sub-Eddington rates. At BH masses greater than $10^9\, M_\odot$, neither the

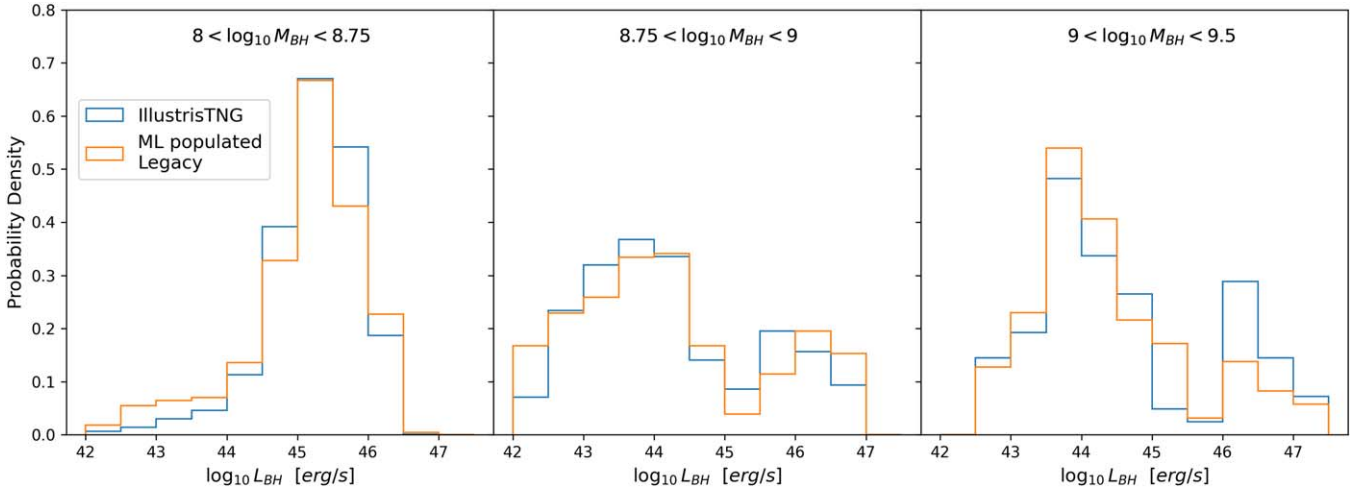

Figure 10. Inventory of quasar population at $z \sim 3$: sources with bolometric luminosity $L > 10^{42}$ erg s^{-1} from the Illustris-TNG300 trained ML used to populate the larger LEGACY 1 Gpc Expanse run box from the $z = 3.25$ slice in three mass bins.

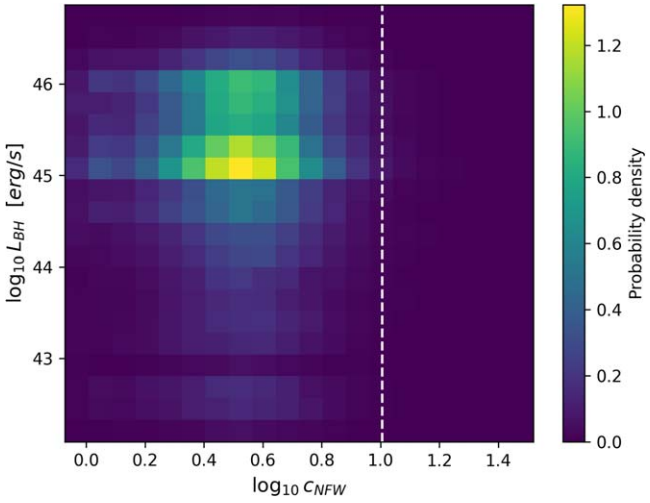

Figure 11. Robustness of the selection of accreting quasars in the ML populated LEGACY 1 Gpc Expanse box from the $z = 3.25$ slice as Type Is. To test if the selected quasars inhabit the right parent dark matter halos, here we plot the distribution of the NFW fit concentration parameters for their parent halos and show that our selected sample is in excellent agreement with the observational determination combining clustering measurements for AGNs reported in Powell et al. (2018, 2020).

amplitude nor the slope match well, as the BH population is underestimated in simulations. This indicates, as expected, that the simulation under-produces the rarest, most-massive black holes, which likely signals incompleteness arising from the smaller volume compared to the SDSS survey volume footprint. What is however completely unexpected is that even at lower masses, $M_{\rm bh} < 10^{8.5} M_{\odot}$, BHs that are detected by the SDSS $z \sim 3$ are missing from the simulations. At the high-mass end of the BHMF, this deficit of accreting SMBHs in the simulations could be partially attributed to a mismatch in sampling volumes between the SDSS and the simulation box, or as potentially arising as a limitation of our training set. However, the mismatch at lower masses, namely the lack of accreting BHs in mass range $M_{\rm bh} < 10^{8.5} M_{\odot}$ in simulations, is glaring and strongly suggests that the adopted subgrid accretion and feedback prescriptions in simulations are suspect. Note that accretion in our training set Illustris-TNG300 simulations is capped at the Eddington limit, so unsurprisingly no super-Eddington sources are predicted in these simulations, and therefore none are found the LEGACY (1 Gpc)^3 Expanse box either. Our comparison clearly reveals that lower-mass black holes in the simulation box are not luminous enough, as they are not accreting at high enough rates to survive the luminosity cut. This suggests that the subgrid model for accretion adopted in Illustris-TNG300 does not accurately capture the accretion in observed quasars. The high accretion rate tail of the distribution of accretion rates appears to be missing in the simulations. Note that even though the LEGACY (1 Gpc)^3 Expanse box given its larger volume, encapsulates a larger range and diversity of formation and assembly histories for SMBHs, it replicates the issue with accretion rates found in the Illustris-TNG300. Our work suggests that the subgrid, multimode BH feedback operating in a thermal "quasar" mode at high accretion states, and a kinetic "wind" mode at low accretion states adopted in Illustris-TNG300, is over-efficient and appears to choke accretion onto BHs prematurely. Note that current state-of-the-art simulations like the Illustris-TNG300 successfully reproduce observed properties like the BHMF at $z = 0$ and the integrated mass density in BHs over cosmic time at epochs $z < 5$. Therefore, these simulations with unique recipes for feedback from growing BHs, model for gas accretion onto them and prescriptions for star formation reproduce integrated quantities well. However, as we show above, QUOTAS permits a unique diagnosis of the mass assembly history over time, the differential mass accumulation in the BH population by focusing on a slice at $z \sim 3$. The mismatch we find between SDSS quasars and their simulated counterparts points to the fact that current subgrid models of accretion and feedback do not reproduce the mass buildup over time accurately. More specifically, the probability distribution of accretion rates onto black holes with masses between 10^8 and $10^{8.75}$ $M_{\odot}$ does not match observations, and this is not just a cosmic variance issue.

In recent work, Habouzit et al. (2022) showed that nearly none of the currently deployed subgrid models for AGN feedback across simulation suites adequately capture the underlying physics and match the observational data at $z > 5$. Here, we show that the problem starts to appear earlier, at even more modest redshifts, of $z \sim 3$. This growing divergence shows that the undetected lower-luminosity quasar population

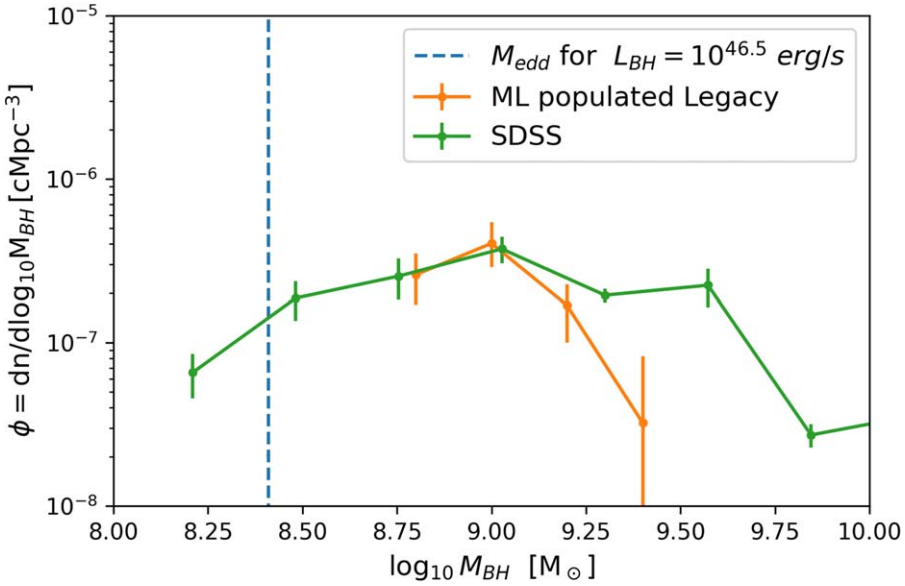

Figure 12. Comparison of the BHMF from $z = 3$–3.5 SDSS quasars and the ML populated LEGACY 1 Gpc box ($z = 3.25$ slice). The blue dashed line marks the BH mass that corresponds to the imposed luminosity cut of $L_{\rm bol} > 10^{46.5}$ erg s^{-1}. This cut implies that all BHs included in the census of the BHMF are accreting at sub-Eddington luminosities.

at modest to high redshift needs to be critically filled in to fully understand BH growth and evolution. Analyzing and comparing results on BH growth and assembly across redshift in several large-scale independent cosmological simulations like the Illustris-TNG100, Illustris-TNG300, Horizon-AGN, EAGLE, and SIMBA suites, Habouzit et al. (2022) showed that while all of them predict the same BHMF and relation between BH mass and host galaxy stellar mass M_* at $z = 0$ in agreement with local observations, their predictions disagree at higher redshifts. For instance, there is much disagreement on whether BHs at $z > 4$ are overmassive or undermassive at fixed host galaxy stellar mass with respect to the $z = 0$ $M - {\rm bh} - M_*$ relation. And as can be seen in Figure 5 of their paper, none of the simulations reproduce the observed accretion luminosities of observed SDSS quasars at $z \geqslant 5.8$. Only the Illustris-TNG300 and SIMBA simulations produce enough SMBHs with $M_{\rm bh} > 10^8\,M_\odot$; however, they both clearly underpredict the luminosities of these sources at $z \sim 6$. Aside from the disagreements between various independent simulations arising from the range of feedback prescriptions and subgrid models they each adopt, here we show explicitly that even at $z \sim 3$ closer to epochs when simulations reproduce local observations, and where the BHMF is observationally well sampled, the Illustris-TNG300 simulation, for instance, does not adequately capture the growth rate of SMBHs that current SDSS observations provide. That is, the mass accretion history of the average SMBH that would lead to the production of a quasar with a bolometric luminosity in the range of $L_{\rm bol} \sim 10^{45-46}$ erg s^{-1}, even at modest redshifts of $z \sim 3$, is not currently reproduced by simulations. The AGN feedback and BH growth prescriptions adopted in simulations do not get the growth buildup over time correct, and analysis with QUOTAS reveals that this is already evident by $z \sim 3$. This strongly suggests that reexamination of our current subgrid theoretical models is warranted, specifically, the models adopted for gas accretion and feedback. It is evident that BH gas feeding rates are insufficient, and therefore, less-aggressive feedback models that do not scupper the accretion flow are needed in order to facilitate uninterrupted gas supply into the centers of galactic nuclei.

In addition to permitting the analysis of large, complex data sets, ML techniques as we illustrate above also help us interrogate key theoretical model assumptions. We have presented a specific instance where observational data from QUOTAS simultaneously leveraged with ML populated LEGACY simulations has afforded deeper diagnosis of the subgrid model for BH accretion and its implementation in simulations. Our key science result from this first deployment of the pilot version of QUOTAS has revealed that two key coupled theoretical assumptions, namely the accretion rate and efficiency of AGN feedback, need to be revisited, and the current subgrid models deployed in simulations are too efficient in driving feedback and hence end up suppressing the growth of low-mass black holes. Essentially, new ML techniques applied simultaneously to analyze the colocated observational and simulation data in QUOTAS have clearly demonstrated that SMBHs in simulations even at modest redshifts of $z \sim 3$ fail to accrete, shine, and grow at high enough rates to match observations of quasars at these cosmic epochs.

While our work has revealed that BH populations in simulations simply do not grow enough, QUOTAS offers a deeper look that will enable us to probe if the association between SMBHs and their corresponding dark matter halos can accurately reproduce the observed clustering properties of the quasar population. This can be tested easily by comparing clustering of observed quasars with the simulated population in the LEGACY $(1\,{\rm Gpc})^3$ box. Comparison of the clustering properties via the two-point correlation function will help investigate the robustness of the association between SMBHs and their parent halos in the SDSS data and the LEGACY $(1\,{\rm Gpc})^3$ Expanse box. In Figure 13, we show that the clustering properties of optical SDSS quasars in this redshift range, $z \sim 3$, are well reproduced by the accreting SMBH population at the corresponding redshift slice ($z = 3.25$) in the LEGACY $(1\,{\rm Gpc})^3$ Expanse box. This clearly suggests that simulations accurately capture the statistics of the association

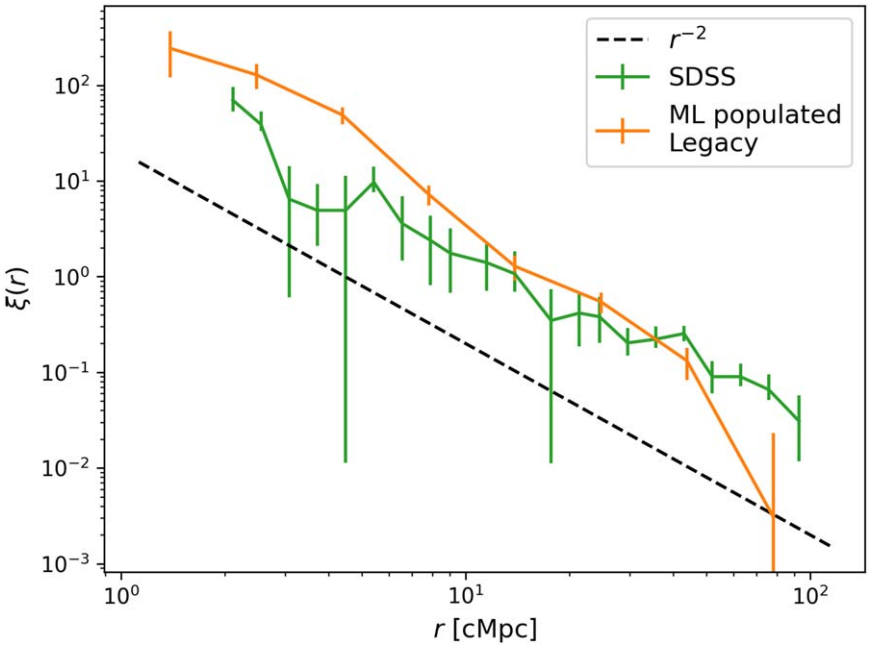

Figure 13. Comparison of the clustering of quasars from SDSS and the ML populated LEGACY 1 Gpc Expanse box at $z \sim 3$. The SDSS quasars plotted here are also part of the BOSS survey, with clustering measurements reported in Eftekharzadeh et al. (2015).

of SMBHs with their parent dark matter halos inferred from clustering studies of SDSS quasars. Therefore, simulated quasars occupy the right parent dark matter halos as inferred from observations, and where they fail is in how they grow. The mass assembly history over time, which is represented by the accretion rate and its evolution for the SMBH population, is suspect in simulations. Comparison of the predictions of simulations by Habouzit et al. (2022) reveals that not only do they progressively diverge from each other at higher redshifts, but that they also progressively deviate from observations. We have demonstrated the power of QUOTAS to produce new science insights on our fundamental input theoretical model assumptions, the implementation of accretion physics in the adopted in subgrid models. This too with the analysis of the observed and simulated quasar population at just one epoch $z \sim 3$. The prospects and science returns from the comprehensive future analysis and application of the tests explored here to even earlier epochs with the publicly available user-friendly research platform QUOTAS look to be extremely promising.

7.4. Probing the Environments of the Brightest Quasars

Finally, one of the fundamental science questions that motivated us to create QUOTAS was the ability to devise future survey strategies leveraging ML, simulations, and observational data to uncover the lower-luminosity, more characteristic, accreting BH population in the modest- and high-redshift universe. We have established above that one key feature of SDSS quasars that our ML-trained simulated quasars accurately reproduce is the occupation statistics, namely, that we get the association between quasars and their parent dark matter halos right as evidenced by (i) the agreement of the correlation function determined from the two samples (as shown in Figure 13) and (ii) the NFW concentration parameter distribution of the parent halos in the LEGACY $(1\,\mathrm{Gpc})^3$ Expanse box (as shown in Figure 11). Therefore, we are well placed to infer an observational survey strategy that will help with uncovering the fainter quasar population. To do this, leveraging the established accuracy of the occupation statistics, we construct a novel neighborhood statistic and evaluate its robustness using QUOTAS. The proposed statistic is derived by enumerating the BHMF found in the neighborhood of a bright quasar powered by a $10^9\,M_\odot$ SMBH in annular rings of varying radius, 8 and 16 comoving Mpc, respectively. As we change the radial region under consideration, the mass function of BHs within that annular radius changes. In Figure 14, we plot this neighborhood occupation statistic for quasars in the ML populated LEGACY $(1\,\mathrm{Gpc})^3$ Expanse box. Included in this plot is the mean neighborhood occupation statistic, shown with a solid line, derived from averaging over a much larger annulus with radius $r > 40$ comoving Mpc. We clearly see that there is a preferential excess of lower-luminosity sources well above the mean value of the neighborhood statistic in the vicinity of a bright quasar. A strong excess of sources is detected even for sources with $L_{\mathrm{bol}} > 10^{46}\,\mathrm{erg\,s^{-1}}$ (just a factor of 3 lower in luminosity), compared to the cut at $L_{\mathrm{bol}} > 10^{46.5}\,\mathrm{erg\,s^{-1}}$. This suggests that the optimal observational strategy to enlarge the high-redshift quasar sample, and probe beyond just the brightest quasars, would be to survey regions around these most-luminous currently detected sources.

8. Conclusions and Future Directions

In this paper, we present the new research platform QUOTAS that for the first time comprises colocated observational and simulated data for the study of the growth and assembly history of SMBHs over cosmic time. QUOTAS is an ongoing project to collect and homogenize observational quasar data and their simulated counterparts. Leveraging the existing normalizing flow and random forest ML algorithms, we provide a template for how we can expand simulated samples of quasars to better match observational survey volumes that permit detailed studies of quasar properties and their environments. In this paper, outlining the pilot phase of

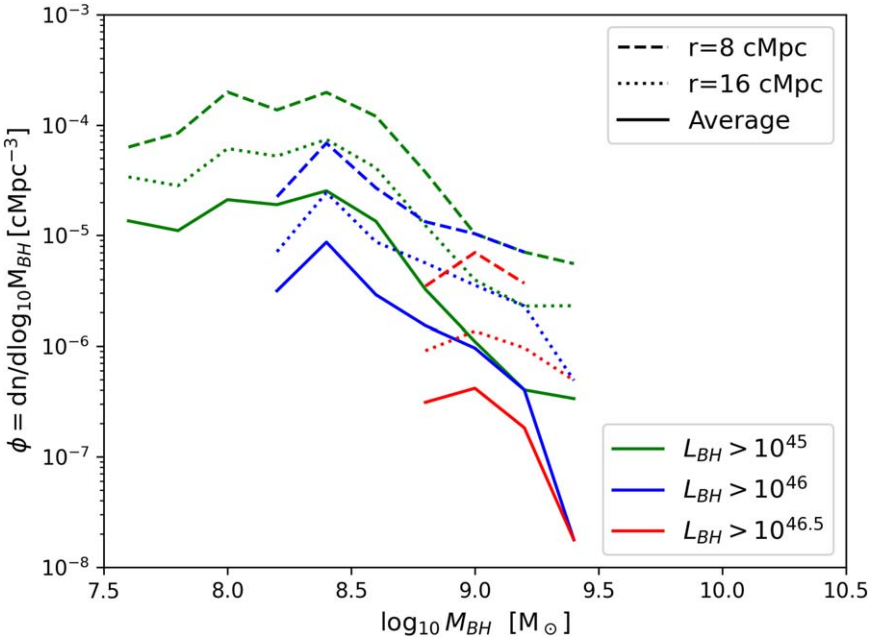

Figure 14. New neighborhood statistic at $z \sim 3$—comparison of the BHMF of the accreting SMBH population that lies within 8, 16, and 40 comoving Mpc radius annuli around $10^9\,M_\odot$ quasars in the LEGACY 1 Gpc Expanse box and SDSS quasars computed with QUOTAS.

our project, we describe the successful compilation of the new database and provide a proof of concept demonstration of ease of use and first science results from querying QUOTAS exploring the SDSS data. With this pilot project, we demonstrate that the profound connection between black holes, their host galaxies, and parent dark matter halos in the context of the standard model for structure formation in the universe can be and needs to be more deeply investigated in order to fully understand how black hole growth unfolds in the larger cosmological context.

We present a collated database of high-redshift $z > 3$ optical quasars in a format ready for application of ML techniques, and have made it available publicly on the Google Kaggle platform. Collecting data from many disparate observational surveys, we make available detailed information and properties where available for individual objects tailored to address the problem of the mass assembly history of the earliest black holes. While full exploration of the BH–galaxy–halo connection with further customized ML tools is the longer-term goal of this live database that we plan to continually update, here we present the QUOTAS database and some early results. We anticipate that QUOTAS will be actively used by the astronomy community to investigate quasar populations and BH growth models. We report on the critical first steps that leverage the QUOTAS platform to construct key quantities of theoretical interest—the BHMF and the QLF—by querying the collated data. Picking $z \sim 3$ as the benchmark redshift, we then compare the Illustris-TNG300 trained ML algorithm used to populated dark-matter-only LEGACY $(1\,\mathrm{Gpc})^3$ Expanse simulation slice with observed SDSS quasars. We find that the occupation fraction of BHs, and the clustering properties of the associated parent dark matter halos, is accurately reproduced. We demonstrate a use case with the BHMF derived from SDSS data and ML-trained extended LEGACY simulations. Comparing the BHMFs, we find that the simulations underpredict the accretion luminosities for BH masses below $M_{\rm bh} < 10^{8.75}\,M_\odot$. This deficit is also clearly reflected in the comparison of the census of the luminosity of accreting BHs as a function of BH mass. Essentially, analysis with QUOTAS reveals that the subgrid models deployed to model the physics of accretion in the Illustris-TNG300 suite are unable to match the observed mass assembly history of BH populations over time. The accretion luminosities are too low, suggesting that AGN feedback is likely too aggressive and causes BHs to stunt their own growth prematurely. Detection of the lower-luminosity quasar population is needed to anchor models and sharpen our understanding of BH growth over cosmic time; therefore, we deployed ML once again to help devise optimal observational survey strategies to uncover this population. ML tools we have shown permit us to question some of the basic assumptions embedded in our theoretical modeling of accretion processes in simulations. So despite simulations matching $z = 0$ observations of the properties of BHs and their hosts, by modest redshifts of just $z \sim 3$, we find that the predictions of simulations are out of line with observed properties of SDSS quasar populations. In this work, we demonstrate how tailored ML tools can be leveraged to visualize, analyze, and compare observational and simulated data of the quasar population.

This successful combination of simulation and observational data sets in a single platform by epoch, we show, allows seamless application of ML techniques. In follow-up work, we plan to expand the data sets with the inclusion of spectral data; addition of the high-z AGN sources that will be detected in upcoming infrared surveys by JWST; and expand with populating host galaxy properties in the LEGACY $(1\,\mathrm{Gpc})^3$ Expanse simulation to study correlations between the BH and the host galaxy, namely, correlations like $M_{\rm bh} - M_*$ and beyond. The ultimate goal of this project is to make full use of the powerful capabilities of ML to analyze these rich, complex multidimensional data sets together. The power of colocated simulated and observational data sets in QUOTAS coupled with the expected improvements in computational infrastructure will allow us to, before long, conduct mock surveys in ML-aided expanded simulation volumes that mimic the exact

footprint of observational surveys. Adding in more attributes for halo properties in the near future will allow for the generation of expansive mock catalogs corresponding to galaxy surveys of varying depths and volumes. This will allow us to connect across data types and test against more detailed model predictions beyond the BHMF and the QLF, for instance, the occupation fraction, and generate maps of larger-scale density of environments around SMBH hosts. In this work, with the introduction of a novel neighborhood statistic, we propose that one reliable detection strategy to uncover the lower-luminosity quasar population is to perform targeted, deeper surveys around currently detected luminous SDSS quasars. Our ML propelled work has provided this new insight that has the potential to open up discovery space for the lower-luminosity quasar population that will significantly constrain our theoretical models. Our future plan is to further leverage these current tools, build upon them using ML to predict the detailed properties of the host galaxies of the quasars, and utilize them to devise additional detection strategies for the as yet undetected lower-luminosity quasars at $3 < z < 7$ to fill in the current gaps in our knowledge of this population. Information on lower-luminosity quasars is critically needed to hone our understanding of accreting black holes and their mass assembly history over cosmic time. In future work, we intend to explore the interplay between observations and simulations and generate predictions for the more representative undetected quasar populations and their properties using ML algorithms that will enable deeper exploration of the BH–galaxy–halo connection.

Acknowledgments

P.N. gratefully acknowledges support from the Black Hole Initiative (BHI) by grants from the Gordon and Betty Moore Foundation and the John Templeton Foundation. P.N. thanks her colleagues at the BHI and members of the next generation Event Horizon Telescope (ngEHT) Science Working Group on Black Holes and their Cosmic Context for many useful conversations on evolving black hole populations and is grateful to Alphabet-X for technical support and computational resources for this project. P.N. acknowledges conversations on databases with Sanjay Sarma and Brian Subirana during the early stages of this project. P.N. and K.S.T. thank Rick Ebert at the Infra-Red Processing and Analysis Center (IPAC) at the California Institute of Technology for his help with accessing the NED database. K.S.T. thanks Frank Wang at Google for his help with the Google Cloud Platform. S.K. acknowledges use of the ARCHER UK National Super-computing Service (http://www.archer.ac.uk) for running the LEGACY simulation. B.N. acknowledges support from the Fermi National Accelerator Laboratory, managed and operated by Fermi Research Alliance, LLC under contract No. DE-AC02-07CH11359 with the U.S. Department of Energy. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes. S.S. acknowledges the Aspen Center for Physics where parts of this work were done, which is supported by National Science Foundation grant PHY-1607611.

Data Availability

All data and associated codes for QUOTAS are publicly available on the Google Kaggle platform for community use. The complete observational data and simulation data sets are available at: https://www.kaggle.com/ and the developed related codes are all available at: https://www.kaggle.com/data sets/quasarnet/quasarnet/code and examples of use are available at: https://www.kaggle.com/code/quasarnet/bhmf-quasarnet; https://www.kaggle.com/code/quasarnet/data-exploration-with-quasarnet.

ORCID iDs

Priyamvada Natarajan https://orcid.org/0000-0002-5554-8896
Sadegh Khochfar https://orcid.org/0000-0001-5956-7843
Brian Nord https://orcid.org/0000-0001-6706-8972
Steinn Sigurdsson https://orcid.org/0000-0002-8187-1144
Nico Cappelluti https://orcid.org/0000-0002-1697-186X

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42
Agarwal, B., Cullen, F., Khochfar, S., Ceverino, D., & Klessen, R. S. 2019, MNRAS, 488, 3268
Agarwal, B., Dalla Vecchia, C., Johnson, J. L., Khochfar, S., & Paardekooper, J.-P. 2014, MNRAS, 443, 648
Agarwal, B., Davis, A. J., Khochfar, S., Natarajan, P., & Dunlop, J. S. 2013, MNRAS, 432, 3438
Agarwal, B., Khochfar, S., Johnson, J. L., et al. 2012, MNRAS, 425, 2854
Alexander, D. M., & Hickox, R. C. 2012, NewAR, 56, 93
Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440
Ananna, T. T., Treister, E., Urry, C. M., et al. 2019, ApJ, 871, 240
Avni, Y., & Bahcall, J. N. 1980, ApJ, 235, 694
Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018, Natur, 553, 473
Barrow, K. S. S., Wise, J. H., Aykutalp, A., et al. 2018, MNRAS, 474, 2617
Beckmann, R. S., Devriendt, J., & Slyz, A. 2019, MNRAS, 483, 3488
Beckmann, R. S., Slyz, A., & Devriendt, J. 2018, MNRAS, 478, 995
Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, ApJ, 762, 109
Bentz, M. C., & Manne-Nicholas, E. 2018, ApJ, 864, 146
Bentz, M. C., Peterson, B. M., Netzer, H., Pogge, R. W., & Vestergaard, M. 2009, ApJ, 697, 160
Bernardini, M., Mayer, L., Reed, D., & Feldmann, R. 2020, MNRAS, 496, 5116
Beskin, V. S., & Kuznetsova, I. V. 2000, NCimB, 115, 795
Blandford, R. D., & McKee, C. F. 1982, ApJ, 255, 419
Bower, R. G., Benson, A. J., Malbon, R., et al. 2006, MNRAS, 370, 645
Breiman, L. 2001, Mach. Learn., 45, 5
Brown, M. J. I., Duncan, K. J., Landt, H., et al. 2019, MNRAS, 489, 3351
Cadiou, C., Pontzen, A., & Peiris, H. V. 2021, MNRAS, 502, 5480
Chen, C.-T. J., Hickox, R. C., Alberts, S., et al. 2013, ApJ, 773, 3
Coatman, L., Hewett, P. C., Banerji, M., et al. 2017, MNRAS, 465, 2120
Colpi, M., Holley-Bockelmann, K., Bogdanovic, T., et al. 2019, arXiv:1903.06867
Cranmer, M. D., Galvez, R., Anderson, L., Spergel, D. N., & Ho, S. 2019, arXiv:1908.08045
Dai, B., & Seljak, U. 2021, PNAS, 118, e2020324118
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
De Rosa, G., Decarli, R., Walter, F., et al. 2011, ApJ, 739, 56
De Rosa, G., Venemans, B. P., Decarli, R., et al. 2014, ApJ, 790, 145
Debuhr, J., Quataert, E., & Ma, C.-P. 2011, MNRAS, 412, 1341
Di Matteo, T., Springel, V., & Hernquist, L. 2005, Natur, 433, 604
Ding, X., Silverman, J., Treu, T., et al. 2020, ApJ, 888, 37
Dinh, L., Sohl-Dickstein, J., & Bengio, S. 2017, arXiv:1605.08803
Djorgovski, S. G., Graham, M. J., Donalek, C., et al. 2016, arXiv:1601.04385
Dubois, Y., Devriendt, J., Slyz, A., & Teyssier, R. 2012, MNRAS, 420, 2662
Eftekharzadeh, S., Myers, A. D., White, M., et al. 2015, MNRAS, 453, 2779
Eilers, A.-C., Hennawi, J. F., & Davies, F. B. 2018, ApJ, 867, 30
Eilers, A.-C., Simcoe, R. A., Yue, M., et al. 2023, ApJ, 950, 68
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019a, ApJL, 875, L1

Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019b, ApJL, 875, L3
Farina, E. P., Schindler, J.-T., Walter, F., et al. 2022, ApJ, 941, 106
Ferrarese, L., & Merritt, D. 2000, ApJL, 539, L9
Gaibler, V., Khochfar, S., & Krause, M. 2011, MNRAS, 411, 155
Gammie, C. F., McKinney, J. C., & Tóth, G. 2003, ApJ, 589, 444
Gaspari, M., Tombesi, F., & Cappi, M. 2020, NatAs, 4, 10
Gebhardt, K., Bender, R., Bower, G., et al. 2000, ApJL, 539, L13
Genzel, R. 2014, arXiv:1410.8717
Genzel, R., Eckart, A., Ott, T., & Eisenhauer, F. 1997, MNRAS, 291, 219
Ghez, A. M., Klein, B. L., Morris, M., & Becklin, E. E. 1998, ApJ, 509, 678
Ghosh, A., Urry, C. M., Wang, Z., et al. 2020, ApJ, 895, 112
Giallongo, E., Grazian, A., Fiore, F., et al. 2015, A&A, 578, A83
Glikman, E., Djorgovski, S. G., Stern, D., et al. 2011, ApJL, 728, L26
Grier, C. J., Pancoast, A., Barth, A. J., et al. 2017, ApJ, 849, 146
Gültekin, K., Richstone, D. O., Gebhardt, K., et al. 2009, ApJ, 698, 198
Habouzit, M., Onoue, M., Bañados, E., et al. 2022, MNRAS, 511, 3751
Habouzit, M., Volonteri, M., & Dubois, Y. 2017, MNRAS, 468, 3935
Haehnelt, M. G., Natarajan, P., & Rees, M. J. 1998, MNRAS, 300, 817
Hahn, O., & Abel, T. 2011, MNRAS, 415, 2101
Häring, N., & Rix, H.-W. 2004, ApJL, 604, L89
Harrison, C. M., Costa, T., Tadhunter, C. N., et al. 2018, NatAs, 2, 198
Hasinger, G. 2008, A&A, 490, 905
Hawley, J. F., Beckwith, K., & Krolik, J. H. 2007, Ap&SS, 311, 117
Heckman, T. M., & Best, P. N. 2014, ARA&A, 52, 589
Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19
Hirschmann, M., Khochfar, S., Burkert, A., et al. 2010, MNRAS, 407, 1016
Hortúa, H. J., Malagò, L., & Volpi, R. 2020, MLS&T, 1, 035014
Hortúa, H. J., Volpi, R., Marinelli, D., & Malagò, L. 2020, PhRvD, 102, 103509
Jahnke, K., & Macciò, A. V. 2011, ApJ, 734, 92
Jiang, L., Fan, X., Vestergaard, M., et al. 2007, AJ, 134, 1150
Jiang, L., McGreer, I. D., Fan, X., et al. 2016, ApJ, 833, 222
Johnson, J. L., Khochfar, S., Greif, T. H., & Durier, F. 2011, MNRAS, 410, 919
Jurić, M., Kantor, J., Lim, K. T., et al. 2017, in ASP Conf. Ser. 512, Astronomical Data Analysis Software and Systems XXV, ed. N. P. F. Lorente, K. Shortridge, & R. Wayth (San Francisco, CA: ASP), 279
Kamdar, H. M., Turk, M. J., & Brunner, R. J. 2016, MNRAS, 457, 1162
Kashikawa, N., Ishizaki, Y., Willott, C. J., et al. 2015, ApJ, 798, 28
Kaspi, S., Maoz, D., Netzer, H., et al. 2005, ApJ, 629, 61
Kaspi, S., Smith, P. S., Netzer, H., et al. 2000, ApJ, 533, 631
Katz, M. L., Kelley, L. Z., Dosopoulou, F., et al. 2020, MNRAS, 491, 2301
Kelley, L. Z., Blecha, L., Hernquist, L., Sesana, A., & Taylor, S. R. 2018, MNRAS, 477, 964
Kelly, B. C., & Shen, Y. 2013, ApJ, 764, 45
Kelly, B. C., Vestergaard, M., Fan, X., et al. 2010, ApJ, 719, 1315
Kingma, D. P., Salimans, T., Jozefowicz, R., et al. 2017, arXiv:1606.04934
Klypin, A., & Prada, F. 2019, MNRAS, 489, 1684
Kobyzev, I., Prince, S. J. D., & Brubaker, M. A. 2019, arXiv:1908.09257
Komissarov, S. 2005, in KITP Conference: Physics of Astrophysical Outflows and Accretion Disks (Santa Barbara, CA: Kavli Inst. for Theoretical Physics), 15
Komissarov, S. S. 2001, MNRAS, 326, L41
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Kozłowski, S. 2017, ApJS, 228, 9
Kulkarni, G., Worseck, G., & Hennawi, J. F. 2019a, MNRAS, 488, 1035
Kulkarni, G., Worseck, G., & Hennawi, J. F. 2019b, MNRAS, 488, 1035
Kurk, J. D., Walter, F., Fan, X., et al. 2007, ApJ, 669, 32
Latif, M. A., & Khochfar, S. 2020, MNRAS, 497, 3761
Liska, M., Hesp, C., Tchekhovskoy, A., et al. 2018, MNRAS, 474, L81
Lodato, G., & Natarajan, P. 2006, MNRAS, 371, 1813
Lodato, G., & Natarajan, P. 2007, MNRAS, 377, L64
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Lucie-Smith, L., Peiris, H. V., Pontzen, A., Nord, B., & Thiyagalingam, J. 2020, arXiv:2011.10577
Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, AJ, 115, 2285
Mahabal, A. A., Djorgovski, S. G., Graham, M. J., et al. 2005, in ASP Conf. Ser. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert (San Francisco, CA: ASP), 604
Marziani, P., del Olmo, A., Martínez-Carballo, M. A., et al. 2019, A&A, 627, A88
Matsuoka, Y., Iwasawa, K., Onoue, M., et al. 2022, ApJS, 259, 18
Matsuoka, Y., Onoue, M., Kashikawa, N., et al. 2019, ApJL, 872, L2
Mazzucchelli, C., Bañados, E., Venemans, B. P., et al. 2017, ApJ, 849, 91
McConnell, N. J., & Ma, C.-P. 2013, ApJ, 764, 184
McGibbon, R., & Khochfar, S. 2022, MNRAS, 513, 5423
McGreer, I. D., Jiang, L., Fan, X., et al. 2013, ApJ, 768, 105
McKinney, J. C. 2005, ApJL, 630, L5
Meriot, R., Khochfar, S., Oñorbe, J., & Smith, B. 2022, MNRAS, 512, 27
Miyoshi, M., Moran, J., Herrnstein, J., et al. 1995, Natur, 373, 127
Moews, B., Davé, R., Mitra, S., Hassan, S., & Cui, W. 2021, MNRAS, 504, 4024
Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Natur, 474, 616
Moster, B. P., Naab, T., Lindström, M., & O'Leary, J. A. 2021, MNRAS, 507, 2115
Nandra, K., O'Neill, P. M., George, I. M., Reeves, J. N., & Turner, T. J. 2006, AN, 327, 1039
NASA/IPAC Extragalactic Database (NED) 2019, NASA/IPAC Extragalactic Database (NED), IPAC, doi:10.26132/NED1
Natarajan, P. 2014, GReGr, 46, 1702
Natarajan, P., Pacucci, F., Ferrara, A., et al. 2017, ApJ, 838, 117
Natarajan, P., Ricarte, A., Baldassare, V., et al. 2019, BAAS, 51, 73
Natarajan, P., Tang, K. S., McGibbon, R., & Khochfar, S. 2023, QUOTAS, Kaggle, doi:10.34740/KAGGLE/DS/3165166
Natarajan, P., & Treister, E. 2009, MNRAS, 393, 838
Nelson, D., Pillepich, A., Genel, S., et al. 2015, A&C, 13, 12
Nelson, D., Springel, V., Pillepich, A., et al. 2019, ComAC, 6, 2
Netzer, H. 2015, ARA&A, 53, 365
Nord, B., Connolly, A. J., Kinney, J., et al. 2019, BAAS, 51, 224
Ntampaka, M., Avestruz, C., Boada, S., et al. 2019, BAAS, 51, 14
Ntampaka, M., Eisenstein, D. J., Yuan, S., & Garrison, L. H. 2020, ApJ, 889, 151
Pacucci, F., Ferrara, A., Grazian, A., et al. 2016, MNRAS, 459, 1432
Papamakarios, G., Nalisnick, E., Jimenez Rezende, D., Mohamed, S., & Lakshminarayanan, B. 2019, arXiv:1912.02762
Papamakarios, G., Pavlakou, T., & Murray, I. 2018, arXiv:1705.07057
Peng, C. Y. 2007, ApJ, 671, 1098
Peterson, B. M. 1993, PASP, 105, 247
Peterson, B. M. 2014, SSRv, 183, 253
Peterson, B. M., Ferrarese, L., Gilbert, K. M., et al. 2004, ApJ, 613, 682
Powell, M. C., Cappelluti, N., Urry, C. M., et al. 2018, ApJ, 858, 110
Powell, M. C., Urry, C. M., Cappelluti, N., et al. 2020, ApJ, 891, 41
Reynolds, C. S. 2020, ARA&A, 59, 117
Rezende, D. J., & Mohamed, S. 2016, arXiv:1505.05770
Ricarte, A., & Natarajan, P. 2018a, MNRAS, 474, 1995
Ricarte, A., & Natarajan, P. 2018b, MNRAS, 481, 3278
Ricarte, A., Tremmel, M., Natarajan, P., & Quinn, T. 2019, MNRAS, 489, 802
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Robertson, B., Bullock, J. S., Cox, T. J., et al. 2006, ApJ, 645, 986
Ross, N. P., McGreer, I. D., White, M., et al. 2013, ApJ, 773, 14
Schawinski, K., Khochfar, S., Kaviraj, S., et al. 2006, Natur, 442, 888
Schawinski, K., Zhang, C., Zhang, H., Fowler, L., & Santhanam, G. K. 2017, MNRAS, 467, L110
Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521
Schneider, D. P., Richards, G. T., Hall, P. B., et al. 2010, AJ, 139, 2360
Sexton, R. O., Canalizo, G., Hiner, K. D., et al. 2019, ApJ, 878, 101
Shao, Y., Wang, R., Jones, G. C., et al. 2017, ApJ, 845, 138
Shen, Y. 2013, BASI, 41, 61
Shen, Y., Hall, P. B., Horne, K., et al. 2019a, ApJS, 241, 34
Shen, Y., & Liu, X. 2012, ApJ, 753, 125
Shen, Y., Richards, G. T., Strauss, M. A., et al. 2011, ApJS, 194, 45
Shen, Y., Wu, J., Jiang, L., et al. 2019b, ApJ, 873, 35
Sijacki, D., Springel, V., Di Matteo, T., & Hernquist, L. 2007, MNRAS, 380, 877
Sijacki, D., Vogelsberger, M., Genel, S., et al. 2015, MNRAS, 452, 575
Small, T. A., & Blandford, R. D. 1992, MNRAS, 259, 725
Springel, V., Di Matteo, T., & Hernquist, L. 2005, MNRAS, 361, 776
Springel, V., Frenk, C. S., & White, S. D. M. 2006, Natur, 440, 1137
Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726
Steidel, C. C., Adelberger, K. L., Giavalisco, M., Dickinson, M., & Pettini, M. 1999, ApJ, 519, 1
Storchi-Bergmann, T., & Schnorr-Müller, A. 2019, NatAs, 3, 48
Tachibana, Y., Graham, M. J., Kawai, N., et al. 2020, ApJ, 903, 54
Taylor, S., Burke-Spolaor, S., Baker, P. T., et al. 2019, BAAS, 51, 336
Tchekhovskoy, A., Narayan, R., & McKinney, J. C. 2010, ApJ, 711, 50
Tchekhovskoy, A., Narayan, R., & McKinney, J. C. 2011, MNRAS, 418, L79
Trakhtenbrot, B., & Netzer, H. 2012, MNRAS, 427, 3081
Trakhtenbrot, B., Netzer, H., Lira, P., & Shemmer, O. 2011, ApJ, 730, 7

Treister, E., Schawinski, K., Volonteri, M., Natarajan, P., & Gawiser, E. 2011, Natur, 474, 356
Tremaine, S., Gebhardt, K., Bender, R., et al. 2002, ApJ, 574, 740
Tremaine, S., Richstone, D. O., Byun, Y.-I., et al. 1994, AJ, 107, 634
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
Venemans, B. P., Bañados, E., Decarli, R., et al. 2015, ApJL, 801, L11
Vestergaard, M. 2019, NatAs, 3, 11
Vestergaard, M., & Osmer, P. S. 2009, ApJ, 699, 800
Vestergaard, M., & Peterson, B. M. 2006, ApJ, 641, 689
Villaescusa-Navarro, F., Ding, J., Genel, S., et al. 2022, ApJ, 929, 132
Vogelsberger, M., Genel, S., Springel, V., et al. 2014, MNRAS, 444, 1518
Volonteri, M. 2012, Sci, 337, 544
Warren, S. J., Hewett, P. C., Irwin, M. J., et al. 1987, Natur, 325, 131
Wechsler, R. H., & Tinker, J. L. 2018, ARA&A, 56, 435
Weinberger, R., Springel, V., Hernquist, L., et al. 2017, MNRAS, 465, 3291
Weinberger, R., Springel, V., Pakmor, R., et al. 2018, MNRAS, 479, 4056
Willott, C. J., Albert, L., Arzoumanian, D., et al. 2010a, AJ, 140, 546
Willott, C. J., Delorme, P., Reylé, C., et al. 2010b, AJ, 139, 906
Wise, J. H., Regan, J. A., O'Shea, B. W., et al. 2019, Natur, 566, 85
Woo, J.-H., Yoon, Y., Park, S., Park, D., & Kim, S. C. 2015, ApJ, 801, 38
Woods, T. E., Agarwal, B., Bromm, V., et al. 2019, PASA, 36, e027
Yang, J., Wang, F., Wu, X.-B., et al. 2016, ApJ, 829, 33
Yao-Yu Lin, J., Pandya, S., Pratap, D., Liu, X., & Carrasco Kind, M. 2020, arXiv:2011.15095
Znajek, R. L. 1977, MNRAS, 179, 457
Zuo, W., Wu, X.-B., Fan, X., et al. 2015, ApJ, 799, 189