Identifying RR Lyrae Variables and Globular Clusters in the Milky Way and Simulating Supermassive Black Hole Early Evolution

Kuan-Wei Huang

Department of Physics
Carnegie Mellon University

Dissertation Committee:
Sergey E. Koposov
Matthew G. Walker
Carl Rodriguez
Rachel Bezanson

January 1, 2022

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Abstract

This thesis presents four research projects about substructures of the Milky Way from large-sky surveys and supermassive black holes at high redshift in cosmological simulations.

First, I reported the searching result for globular clusters (GCs) associated with 55 Milky Way dwarf galaxies using *Gaia* DR2. Eleven candidates were identified and all were either known GCs or galaxies, yet only the six Fornax GCs were associated with the dwarf. The completeness of the GC search was above 90% for most dwarf galaxies and the 90% credible intervals on the GC specific frequency S_N were $12 < S_N < 47$ for Fornax, $S_N < 20$ for the dwarfs with $-12 < M_V < -10$, $S_N < 30$ for the dwarfs with $-10 < M_V < -7$, and $S_N < 90$ for the dwarfs with $M_V > -7$. Based on S_N, I obtained that the probability of galaxies fainter than $M_V = -9$ to host GCs was lower than 0.1.

Second, I presented a RR Lyrae (RRL) catalog based on the combination of ZTF DR3 and *Gaia* EDR3. Covering the sky of declination $\geq -28°$, the catalog contained 71,755 RRLs with period and light curve parameter measurements, with completeness of 0.92 and purity of 0.92 compared to the SOS *Gaia* DR2 RRLs. Compared with several other RRL catalogs covering the Northern sky, this catalog had more RRLs around the Galactic halo and was more complete at low Galactic latitude areas. Analyzing the spatial distribution of RRL in the catalog revealed the known major over-densities of the Galactic halo, such as the Virgo over-density and the Hercules-Aquila Cloud, with some evidence of an association between the two.

Third, I examined the early growth of supermassive black holes (SMBHs) with different BH seeding scenarios, by conducting multiple constrained cosmological simulations. Among the simulations, only the ones with a low-tidal field and high-density peak in the initial conditions induced the fastest BH growth required to explain the $z > 6$ quasars. In the simulations with different BH seed masses of 5×10^3, 5×10^4, and $5 \times 10^5\, h^{-1} M_\odot$, the SMBH masses converged to $\sim 10^9\, M_\odot$ except the one with the smallest seed. The vast BHs in the small-seed scenario merged frequently during the early phases of SMBH growth, which provided an exciting prospect for discriminating BH formation mechanisms at high-z.

Fourth, I established scaling relations between SMBH mass ($M_\bullet$) and host galaxy properties (stellar mass $M_\star$ and velocity dispersion σ), using the BlueTides simulation. The relations at $z = 8$ were: $\log_{10}(M_\bullet) = 8.25 + 1.10 \log_{10}(M_\star / 10^{11} M_\odot)$ and $\log_{10}(M_\bullet) = 8.35 + 5.31 \log_{10}(\sigma / 200 \mathrm{kms}^{-1})$, both consistent with current local measurements. However, the intrinsic scatter in the $M_\bullet - \sigma$ relation was larger than the one inferred from observations and that of the $M_\bullet - M_\star$ relation. I found that the scatter of $M_\bullet - \sigma$ was significantly reduced when galaxies with high gas fractions were excluded such that the sample was comparable to low-z galaxies. The excluded systems had extremely large star formation rates and BH accretion rates, indicating that these fast-growing systems were still moving toward the relations at high redshift.

Acknowledgements

In completing this dissertation and the Physics Ph.D. journey at Carnegie Mellon, I am grateful for all the assistance and support I have received on the way.

First and foremost, I thank my advisor Sergey Koposov for being supportive and patient when advising me during the years. I appreciate his patient when he explained some concepts or points if I could not get the gist in the first place. I am also thankful for the freedom he provided, such as working time and place or choosing research projects. In short, I am grateful to have Sergey as my advisor, and I enjoy the time working with him.

Next, I would like to thank Matthew Walker for taking over the role of academic advisor after Sergey left Carnegie Mellon. I felt stressed at the moment, but thanks to Matthew, I could continue on the work and smoothly finish the degree. I am grateful for all the paperwork he has done during the transition.

Besides Sergey and Matthew, I would like to thank the other two committee members, Rachel Bezanson and Carl Rodriguez. It is my pleasure to have you on my committee, and I appreciate all the advice and suggestions for the dissertation.

I thank all the collaborators for the papers and projects included in the dissertation. First, I appreciate Tiziana Di Matteo's efforts and advice for my early work at Carnegie Mellon. Next, I thank Yu Feng and Ananth Tenneti for all the help on the technical issues. Last but not least, I am thankful for the discussion with Chung-Pei Ma.

I am grateful to have Aklant Bhowmick and Yueying Ni as my partners in Wean 8309 for such a long time. All the laughter and fun moments with you overtook all the stress and pressure on the Ph.D. journey. I am thankful to have you two as collaborators, officemates, and friends.

I am thankful to have Manfred Paulini as the graduate associate in the department of Physics, who has been extremely helpful and supportive on many essentials. I am grateful to have Manfred Paulini and Scott Dodelson to talk to during a hard time.

I thank the company of my fellow physicists who joined the department together in 2016 and experienced the journey together. Hüsni Almoubayyed, Sayan Mandal, Ana Paula Vizcaya, Madeline Sauleda, Daniel Darvish, Yizhou He, Jeffrey Patrick, Muhammed Fethullah Ergüder, Andrew Smith, Olga Navros, Ying Liu.

I am glad to share most of my time in the department with all the people I have met. Hsiu-Hsien Lin, Evan Tucker, Zhonghao Luo, Hongyu Zhu, Mao-Sheng Liu, Abel Sun, Hung-Jin Huang, Chien-Hao Lin, Bai-Cian Ke, Alexander Moskowitz, Michael Andrews, Tassia Ferreira, Larisa Thorne, Michael Sinko, Dacen Waters, Francois Lanusse, Matthew Ho, Shashin Pavaskar, I-Hsuan Kao, Nianyi Chen, Andresa Rodrigues de Campos, Tianqing Zhang, Zhaozhou An, Yingzhang Chen, Samuel Foley, Nathaniel Dene Hoffman, Yesukhei Jagvaral, Christopher Kervick, Kuldeep Sharma, Beka Modrekiladze, Ryan Muzzio.

I am glad to have friends outside of the department who made my time here in Pittsburgh.

Yi-Chung Lin, Michael Liang, Shih-En Wei, Ching-Yi Lin, Janet Tseng, Marcos Mazari Armida, Won Eui Hong.

To some of my Physics friends from Taiwan whom I have met in different states during the years of my Ph.D., I wish you all the best. Joshua Yao-Yu Lin, Huan-Kuang Wu, Wen-Chen Lin, Po-Wen Chang, Hsiao-Yi Chen.

To my family. Thank you for everything.

Contents

List of Tables

List of Figures

1 Introduction

Looking at the dark sky at night, we can see various astronomical objects catching our attention, such as stars, galaxies, and even galaxy clusters. Why these structures are there and how they form are the questions that many structure formation theories in cosmology are trying to answer. A widely accepted model for understanding structure formation and evolution of the Universe is the modern ΛCDM model. This "dark energy + cold dark matter" model has achieved great success in depicting the Universe and predicting many cosmological structures that we have observed.

The basic concepts of cold dark matter for galaxy formation have been around for decades (White & Rees, 1978; Fall & Efstathiou, 1980). The increasing amount of evidence that all galaxies are baryon condensates at the bottom of massive dark halos (White & Rees, 1978) has motivated the creation of cold dark matter (Zwicky, 1933; Rubin et al., 1978), despite the unknown nature of the matter constituting dark halos (e.g., Bergström, 2000; Bertone et al., 2005). In the ΛCDM paradigm, baryonic structure formation began at the recombination era at redshift $\sim$ 1000 when the ionized hydrogen recombined after the Big Bang (Dicke et al., 1965; Peebles, 1968; Blumenthal et al., 1984). During this era, baryons, which had been almost uniformly distributed and in quasi-equilibrium with radiation, started to feel no radiation pressure and fell into the potential wells that had been already formed by dark matter (White & Rees, 1978). As the cooling of baryons continued, first stars proceeded to form and kept forming at redshift $\sim 20-50$ in dark matter halos with masses of $\sim 10^5 - 10^6 M_\odot$ (Abel et al., 2002), starting to ionize and chemically enrich the surrounding interstellar and intergalactic medium. The increasing amount of first stars resulted in overlaps of first galaxies and finally made the Universe ionized again at redshift $\sim 7-11$ (Gnedin & Ostriker, 1997; Barkana & Loeb, 2001; Fan et al., 2006). Since then, subsequent galaxy formation continued with many complicated processes, including star formation (White & Rees, 1978), gas cooling under UV ionizing background (Quinn et al., 1996; Thoul & Weinberg, 1996), stellar feedback (Dekel & Silk, 1986; Martin, 1999), ANG feedback (Springel et al., 2005), and satellite accretion (Hernquist & Mihos, 1995; Abadi et al., 2003b).

Until today, the ΛCDM model has successfully explained the large-scale structure of the Universe (Percival et al., 2001; Cole et al., 2005; Eisenstein et al., 2005) and the Cosmic Microwave Background (CMB) (Penzias & Wilson, 1965; Dunkley et al., 2009; Planck Collaboration et al., 2016). Its postulations, that dark matter halos are the breeding grounds to form galaxies (White & Rees, 1978; Blumenthal et al., 1984) and that the Universe is composed of $\sim$ 25% baryonic matter and $\sim$ 75% dark matter (Planck Collaboration et al., 2016), are hence widely accepted in modern cosmology.

However, there are still several unresolved problems concerning the structural properties of halos on galactic scales and some discrepancies between the ΛCDM prediction and

observation. The missing satellites problem is one of the discrepancies, where ΛCDM predicted two orders of magnitude more than the number of known Milky Way satellites (Kauffmann et al., 1993; Klypin et al., 1999; Moore et al., 1999; Diemand et al., 2008; Springel et al., 2008). For the missing satellites problem, there have been both theoretical attempts to reduce the number of predicted substructures (Bode et al., 2001; Narayanan et al., 2000; Zentner & Bullock, 2003) and observational attempts to discover possibly lurking substructures through large-sky surveys. The observational searches for substructures using these surveys have achieved tremendous success in finding substructures of globular clusters, dwarf galaxies, and stellar streams (e.g., Ibata et al., 2001; Odenkirchen et al., 2001; Ferguson et al., 2002; Majewski et al., 2003; Grillmair & Dionatos, 2006; Belokurov et al., 2006, 2007a, 2009; Koposov et al., 2015; Torrealba et al., 2016b,a; Homma et al., 2016; Koposov et al., 2017; Massari & Helmi, 2018; Torrealba et al., 2018; Homma et al., 2018; Torrealba et al., 2018, 2019b). Albeit this slightly eased the missing satellites problem, the problem still has not been fully resolved yet and remains challenging for the standard ΛCDM paradigm.

The Milky Way Galaxy is the perfect place to test ΛCDM predictions at small scales, so the search and analysis for substructures around or near the Galaxy are crucial for the following reasons. First, it hints us the history of the accretion process and the formation of the Galaxy (Searle & Zinn, 1978; Helmi et al., 1999; Ibata et al., 2001; Helmi, 2004; Fellhauer et al., 2006; Koch et al., 2006; Koposov et al., 2010). Second, it provides a chance to see the lowest-luminosity galaxies formed in the early Universe and further open a gate to study the formation process at high redshifts (Ricotti & Gnedin, 2005; Koposov et al., 2009). Third, it is helpful to understand the dynamics of stellar structures and trace the Galactic potential (e.g., Fellhauer et al., 2006; Bell et al., 2008; Grillmair, 2009; Klement et al., 2009; Koposov et al., 2010). Therefore in Chapters 2 and 3, I present our published works about globular clusters and RR Lyrae variable stars in the Galaxy.

Globular clusters are some of the oldest luminous observable objects with ages comparable to the age of the Universe (VandenBerg et al., 2013). These compact and bright star clusters typically have masses of $10^4 - 10^6 M_\odot$, luminosities of $M_V = -5$ to -10, and sizes of a few parsecs (Harris, 1991; Brodie & Strader, 2006). Globular clusters might have played an essential role in the early formation of galaxies, and they could have been the potential drivers of cosmic reionization (Boylan-Kolchin, 2018) despite the issues with the escape fraction of ionizing radiation (Howard et al., 2018b; He et al., 2020). However, the formation of globular clusters themselves remains an open question in astrophysics (some recent literature that discusses the formation of globular clusters, e.g., Howard et al., 2018a; Reina-Campos et al., 2019; Choksi & Gnedin, 2019; El-Badry et al., 2019; Ma et al., 2020). For detailed reviews of globular clusters, we refer readers to Gratton et al. (2004), Brodie & Strader (2006), and Gratton et al. (2019).

In the Milky Way, the number of known globular clusters has increased to about 150 (Harris, 1996, 2010) since Abraham Ihle discovered the first one in 1665. Some of the

globular clusters that are concentrated around the Galactic Center are believed to have been formed *in-situ* (Forbes et al., 1997; Harris et al., 1999). Other ones in the outskirts are believed to have been accreted together with their parent dwarf galaxies (e.g., Searle & Zinn, 1978; Mackey & Gilmore, 2004; Beasley et al., 2018; Kruijssen et al., 2019), which were destroyed by tides. In particular, some of the globular clusters can still be found within the Milky Way satellites themselves, offering a window on the formation of globular clusters in dwarf galaxies. The three most luminous Milky Way satellites, the Large Magellanic Cloud, the Small Magellanic Cloud, and the Sagittarius dwarf spheroidal galaxy, have large globular cluster populations (Mackey & Gilmore, 2003a,b,c; McLaughlin & van der Marel, 2005). Specifically, the clusters associated with the Sagittarius dwarf are spread out along the stellar stream (Lynden-Bell & Lynden-Bell, 1995; Bellazzini et al., 2003; Luque et al., 2017; Vasiliev, 2019). The Small Magellanic Cloud has a large population of star clusters in general, but few are classically old globular clusters. The only other two Milky Way satellite galaxies known to possess globular clusters are the Fornax dwarf and the Eridanus 2. The Fornax dwarf spheroidal galaxy, the fourth most luminous Milky Way satellite, has six globular clusters, and the ultra-faint system Eridanus 2 contains a faint cluster (Koposov et al., 2015; Crnojević et al., 2016).

The fact that some globular clusters in the Milky Way still have been found until recently (Koposov et al., 2015; Koposov et al., 2017; Wang et al., 2019) motivates us to further search for possibly missing ones. In **Chapter 2**, we reproduce the published work of Huang & Koposov (2021c), where we search for globular clusters associated with 55 Milky Way dwarf galaxies. Intuitively, faint globular clusters within dwarf galaxies are more likely to have been missed, especially when located within luminous dwarf galaxies where the ground-based data can be crowded, e.g., Fornax 6. Instead of looking for this kind of object by chance, we apply the systemic overdensity searching algorithm to the areas around the Milky Way satellite galaxies within the distance of 450 kpc except for the three most luminous ones: the Large Magellanic Cloud, the Small Magellanic Cloud, and the Sagittarius dwarf. That is, we target the areas where globular clusters are likely to lurk from previous inspections of deep imaging to look for overdensities in dense dwarfs. For each targeted area, we investigate the stellar distribution in the *Gaia* data to detect possible globular cluster candidates, because the angular resolution of *Gaia* exceeds most ground-based surveys. This search provides a chance for us to detect previously missed objects that are not well resolved or missed by ground-based searches; for instance, Koposov et al. (2017) has found star clusters in *Gaia* that were missed by previous searches.

Searching for further faint substructures such as globular clusters and dwarf galaxies in and around the Milky Way has been essential to understanding the Galactic structure and the Galaxy formation. RR Lyrae variable stars are another powerful tool to search and locate the Milky Way substructures since almost every Milky Way dwarf satellite galaxy has at least one RR Lyrae star (Sesar et al., 2014; Baker & Willman, 2015). It opens a gate to locate the

Milky Way dwarf satellites by using distant RR Lyrae stars, even for ultra-faint objects like Antlia 2 (Torrealba et al., 2019b).

RR Lyrae stars are pulsating variables with periodic light curves of a period ranging from 0.2 to 0.9 days (Smith, 1995), found primarily in the horizontal branches of old stellar systems (age > 10 Gyr). These old, metal-poor ([Fe/H] < −0.5), bright (M_V = 0.59 at [Fe/H] = −1.5; Cacciari & Clementini (2003)) variable stars follow a well-understood period-luminosity-metallicity (PLZ) relation (e.g., Cáceres & Catelan, 2008; Marconi, 2012). This relation makes RR Lyraes excellent distance indicators for old, low-metallicity stellar populations in the outer halo of the Milky Way (e.g., Catelan et al., 2004; Vivas et al., 2004; Cáceres & Catelan, 2008; Sesar et al., 2010; Stetson et al., 2014; Fiorentino et al., 2015). Besides, RR Lyraes are sufficiently luminous to be detected at large distances so that they can be the tracer of the halo substructures with a good spatial resolution (e.g., Vivas & Zinn, 2006; Sesar et al., 2010; Sesar et al., 2014; Baker & Willman, 2015; Torrealba et al., 2015; Martínez-Vázquez et al., 2019).

Being beneficial to many Galactic studies, there have been several RRL catalogs classified from existing surveys over the years, e.g. SDSS Stripe 82 (Sesar et al., 2010), CRTS (Drake et al., 2014), PS1 (Sesar et al., 2017), nTransits:2+ *Gaia* DR2 (Holl et al., 2018), SOS *Gaia* DR2 (Clementini et al., 2019a), ZTF DR2 (Chen et al., 2020), and DES Y6 (Stringer et al., 2021). The quality of the catalogs has progressed from being either deep with limited sky coverage (e.g. the SDSS Stripe 82 catalog) or wide-coverage but not as deep (e.g. the CRTS catalog) to having decent depth and wide sky coverage at the same time (e.g. the PS1 catalog), pushing the Galactic studies furthermore. However, catalogs with decent depth and coverage usually suffer from incompleteness and contamination due to the low number of epochs in the light curves. This motivates us to identify a RRL catalog from the ZTF survey thanks to its uniformly high number of observation epochs of light curves across the Northern sky while having decent depth. Another challenge of the catalogs is to cover the Galactic plane; the PS1, *Gaia* DR2, and ZTF DR2 catalogs do cover this area though the *Gaia* catalog suffers the completeness issue here. In **Chapter 3**, we reproduce our published work of Huang & Koposov (2021b) to demonstrate our RR Lyrae catalog from the joint set of the *Gaia* early third data release (*Gaia* EDR3; Gaia Collaboration et al., 2020) and the third data release of the Zwicky Transient Facility (ZTF DR3; Masci et al., 2019).

Another challenging astronomical object in the Milky Way is the supermassive black hole in the Galatic Center (e.g., Gravity Collaboration et al., 2018). How it forms and how it interacts with the Milky Way are still unclear. Almost every massive galaxy has a supermassive black hole more massive than a million solar masses lurking at its center. The formation of supermassive black holes and the interaction between black holes and their host galaxies remain mysterious in the standard paradigm of structure formation. Therefore in Chapters 4 and 5, I present our published works about supermassive black holes in cosmological hydrodynamic simulations.

Over the last three decades, scaling relations between the mass of supermassive black holes and several stellar properties of their host galaxies such as bulge stellar mass and bulge velocity dispersion (Magorrian et al., 1998; Häring & Rix, 2004; Gebhardt et al., 2000; Tremaine et al., 2002; Gültekin et al., 2009; Kormendy & Ho, 2013; McConnell & Ma, 2013; Reines & Volonteri, 2015) have been locally discovered and measured for galaxies with supermassive black holes and active galactic nuclei (AGN) from $z = 0$ and up to $z \sim 2$. Two important follow-up questions are when the scaling relations are established and if they still persist at high redshifts when the first supermassive black holes form ($z > 6$). To understand this, galaxies with AGN play a key role in observations (Bennert et al., 2010; Merloni et al., 2010; Kormendy & Ho, 2013). For example, a strong direct constraint on the high-redshift evolution of supermassive black holes comes from the luminous quasars at $z \sim 6$ in SDSS (Fan et al., 2006; Jiang et al., 2009; Mortlock et al., 2011). However, it is still not established as to whether these objects follow the local supermassive black hole-galaxy relations and whether there is a redshift evolution, due to the systematic uncertainties (Woo et al., 2006) and selection effects (Lauer et al., 2007; Treu et al., 2007; Schulze & Wisotzki, 2011, 2014).

These extremely rare and luminous supermassive black holes are known to exist in the early universe, even up to $z \sim 7.5$, including the current record holder at $z = 7.54$ (Bañados et al., 2018), the one at $z = 7.09$ (Mortlock et al., 2011; Wu et al., 2015), and the one discovered in the Sloan Digital Sky Survey at $z \sim 6$ (Fan et al., 2006; Jiang et al., 2009). The presence of luminous quasars observed within the first billion years of the Universe highlights that the supermassive black hole seeds for the supermassive black hole population must have assembled at the cosmic dawn, concurrently with the time when the first stars or galaxies form. However, the precise supermassive black hole seed formation mechanism remains unknown, nor is it clear if there is only one seed formation channel at play over the entire supermassive black hole seed mass spectrum of models. Current scenarios suggest that seed supermassive black holes may be remnants of the first generation of stars (PopIII) (e.g. Madau & Rees, 2001; Abel et al., 2002; Johnson & Bromm, 2007), direct gas collapse within the first massive halos (e.g. Lodato & Natarajan, 2006; Begelman et al., 2006; Regan & Haehnelt, 2009; Ferrara et al., 2014; Latif et al., 2013), or runaway collapse of dense nuclear star clusters (e.g. Begelman & Rees, 1978; Devecchi & Volonteri, 2009; Yajima & Khochfar, 2016; Katz et al., 2015).

To push the limit of our understanding of structure formation from the local Universe to high redshifts, cosmological hydrodynamic simulation is a powerful tool that we can utilize. For example, several cosmological simulations that model the formation, growth of supermassive black holes and their host galaxies have successfully reproduced the scaling relations at low-z, including the *Illustris* simulation (Vogelsberger et al., 2014; Sijacki et al., 2015), the Magneticum Pathfinder SPH simulation (Steinborn et al., 2015), the Evolution and Assembly of GaLaxies and their Environment (*EAGLE*) suite of SPH simulation (Schaye et al., 2015), and the *MassiveBlackII (MBII)* simulation (Khandai et al., 2015; DeGraf et al.,

2015). In **Chapter 4**, we reproduce the published work of Huang et al. (2020), where we propose a new method of constrained cosmological simulation to study different supermassive black holes seeding scenarios. In **Chapter 5**, we reproduce the published work of Huang et al. (2018) to extend the study of the scaling relations to $z = 8$ to 10, using the high-resolution large-volume cosmological hydrodynamic simulation, BLUETIDES.

In short, this thesis is composed of four research projects to study multiple astronomical objects that are relevant to the structure formation, using both large-sky observational surveys and large-volume cosmological hydrodynamic simulations. Chapter 2 is a reproduce of the published work of Huang & Koposov (2021c), where we report the result of searching for globular clusters around 55 Milky Way satellite dwarf galaxies within the distance of 450 kpc from the Galactic Center. Chapter 3 is a reproduce of the published work of Huang & Koposov (2021b), where we classify a RR Lyrae catalog from the joint set of *Gaia* EDR3 and ZTF DR3. Chapter 4 is a reproduce of the published work of Huang et al. (2020), where we examine the early growth of supermassive black holes with different black hole seeding scenarios by constrained cosmological hydrodynamic simulations. Chapter 5 is a reproduce of the published work of Huang et al. (2018), where we establish the scaling relations between the mass of supermassive black holes and host galaxy properties, stellar mass and velocity dispersion, at high redshifts using BLUETIDES. In Chapter 6, we conclude the thesis and discuss future outlooks of the field.

2 Search for globular clusters associated with the Milky Way dwarf galaxies using *Gaia* DR2

Kuan-Wei Huang[1] and Sergey E. Koposov[2,1,3]

[1] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[3] Institute of Astronomy, Madingley Rd, Cambridge, CB3 0HA

Abstract

We report the result of searching for globular clusters (GCs) around 55 Milky Way satellite dwarf galaxies within the distance of 450 kpc from the Galactic Center except for the Large and Small Magellanic Clouds and the Sagittarius dwarf. For each dwarf, we analyze the stellar distribution of sources in *Gaia* DR2, selected by magnitude, proper motion, and source morphology. Using the kernel density estimation of stellar number counts, we identify eleven possible GC candidates. Crossed-matched with existing imaging data, all eleven objects are known either GCs or galaxies and only Fornax GC 1 – 6 among them are associated with the targeted dwarf galaxy. Using simulated GCs, we calculate the GC detection limit $M_{\rm V}^{\rm lim}$ that spans the range from $M_{\rm V}^{\rm lim} \sim -7$ for distant dwarfs to $M_{\rm V}^{\rm lim} \sim 0$ for nearby systems. Assuming a Gaussian GC luminosity function, we compute that the completeness of the GC search is above 90 percent for most dwarf galaxies. We construct the 90 percent credible intervals/upper limits on the GC specific frequency $S_{\rm N}$ of the MW dwarf galaxies: $12 < S_{\rm N} < 47$ for Fornax, $S_{\rm N} < 20$ for the dwarfs with $-12 < M_{\rm V} < -10$, $S_{\rm N} < 30$ for the dwarfs with $-10 < M_{\rm V} < -7$, and $S_{\rm N} < 90$ for the dwarfs with $M_{\rm V} > -7$. Based on $S_{\rm N}$, we derive the probability of galaxies hosting GCs given their luminosity, finding that the probability of galaxies fainter than $M_{\rm V} = -9$ to host GCs is lower than 0.1.

2.1 Introduction

Globular clusters (GCs) are some of the oldest luminous observable objects with ages comparable to the age of the Universe (VandenBerg et al., 2013). Characterized by being compact

and bright, GCs typically have masses of $10^4 - 10^6 M_\odot$, luminosities of $M_V = -5$ to -10, and sizes of a few parsecs (Harris, 1991; Brodie & Strader, 2006). GCs might have played an important role in the early formation of galaxies, and they could have been the potential drivers of cosmic reionization (Boylan-Kolchin, 2018) despite the issues with the escape fraction of ionizing radiation (Howard et al., 2018b; He et al., 2020). However, the formation of GCs themselves remains an open question in astrophysics (some recent literature that discuss the formation of GCs, e.g. Howard et al., 2018a; Reina-Campos et al., 2019; Choksi & Gnedin, 2019; El-Badry et al., 2019; Ma et al., 2020). For detailed reviews of GCs, we refer readers to Gratton et al. (2004), Brodie & Strader (2006), and Gratton et al. (2019).

In the Milky Way (MW), the number of known GCs has increased to around 150 (Harris, 1996, 2010) since the first one was discovered in 1665 by Abraham Ihle. While some of these GCs that are more concentrated around the Galactic Center are believed to have been formed *in-situ* (Forbes et al., 1997; Harris et al., 1999), the ones in the outskirts are believed to have been accreted together with their parent dwarf galaxies (e.g. Searle & Zinn, 1978; Mackey & Gilmore, 2004; Beasley et al., 2018; Kruijssen et al., 2019), which were destroyed by tides. Some of the GCs however can still be found within the MW satellites themselves, offering a window on the formation of GCs in dwarf galaxies. The three most luminous MW satellites, the Large and Small Magellanic Clouds (LMC and SMC) and the Sagittarius dwarf spheroidal galaxy, have large populations of GCs (Mackey & Gilmore, 2003a,b,c; McLaughlin & van der Marel, 2005). In particular, the clusters of the Sagittarius dwarf are spread out along the stellar stream (Lynden-Bell & Lynden-Bell, 1995; Bellazzini et al., 2003; Luque et al., 2017; Vasiliev, 2019), and the SMC has a large population of star clusters in general but few of them are classically old GCs. The only other two MW satellite galaxies known to possess GCs are the Fornax dwarf spheroidal galaxy which is the fourth most luminous MW satellite with six GCs, and the Eridanus 2, an ultra-faint system containing a faint cluster (Koposov et al., 2015; Crnojević et al., 2016).

The fact that some GCs in the MW still have been found until recently (Koposov et al., 2015; Koposov et al., 2017; Wang et al., 2019) motivates us to further search for possibly missing ones. Intuitively, faint GCs within dwarf galaxies are more likely to have been missed, especially when located within luminous dwarf galaxies where the ground-based data can be crowded e.g. Fornax 6. Instead of looking for this kind of objects by chance, we apply the systemic overdensity searching algorithm (which will be explained in Section 2.2) to the areas around the MW satellite galaxies within the distance of 450 kpc except for the three most luminous ones: the LMC, the SMC, and the Sagittarius dwarf. That is, we target the areas where GCs are likely to lurk from previous inspections of deep imaging to look for overdensities in dense dwarfs.

Focusing on a small area of the sky, a targeted search is less computationally expensive so that it can afford a lower detection threshold. For each targeted area, we investigate the stellar distribution in the *Gaia* data to detect possible GC candidates (see Section 2.2.1 for

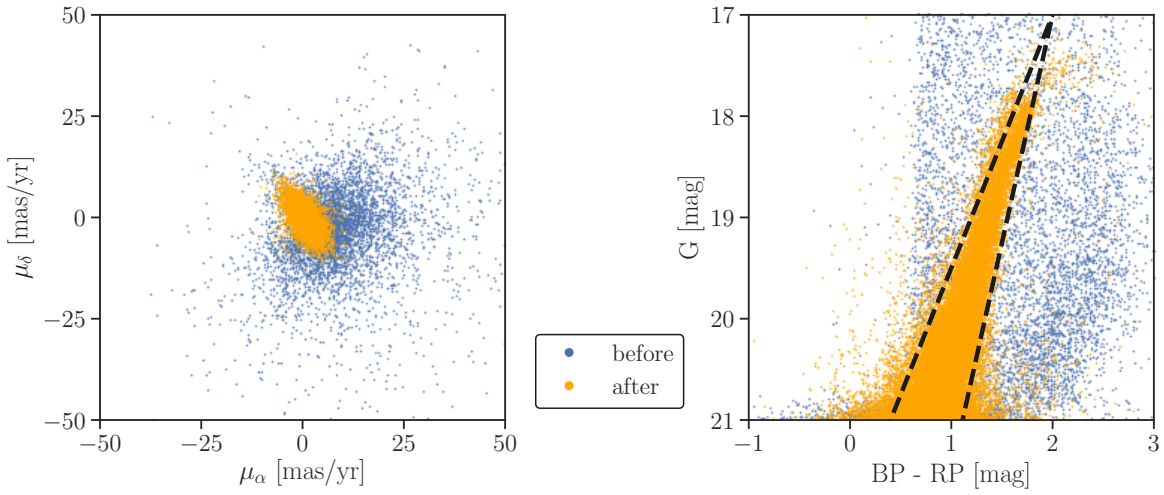

Figure 2.1: The *Gaia* sources around the Fornax dwarf before (blue) and after (orange) the proper motion selection defined in Equation 2.3. **Left**: the distribution in proper motion space. **Right**: the color-magnitude diagram. The black dashed lines define a lasso to roughly distinguish possible member stars in the red-giant branch of Fornax.

more detail about *Gaia* and the dataset). Thanks to the high angular resolution that exceeds most ground-based surveys, *Gaia* allows us to detect previously missed objects that are not well resolved or missed by ground-based searches. For instance, Koposov et al. (2017) has found star clusters in *Gaia* that were missed by previous searches.

We organize the paper as follows. In Section 2.2, we explain the methodology with more detail about the *Gaia* data, sample selection, and kernel density estimation procedure. In Section 2.3, we demonstrate the main results of the detection. In Section 2.4, we discuss the limit and completeness of the detection, the inferred specific frequency of GCs, and the derived probability of dwarfs to host GCs based on our findings. In Section 2.5, we conclude the paper.

2.2 Methodology

2.2.1 *Gaia* DR2 and data selection

The space-based astrometric mission *Gaia* was launched by the European Space Agency in 2013 and started the whole-sky survey in 2014 (Gaia Collaboration et al., 2016). Released in 2018, the second *Gaia* data release (*Gaia* DR2) contains the data collected during the first 22 months of the mission (Gaia Collaboration et al., 2018a) and has approximately 1.7 billion sources with 1.3 billion parallaxes and proper motions. *Gaia* DR2, therefore, provides high-resolution stellar distribution in the MW for us to look for possibly missing GCs around

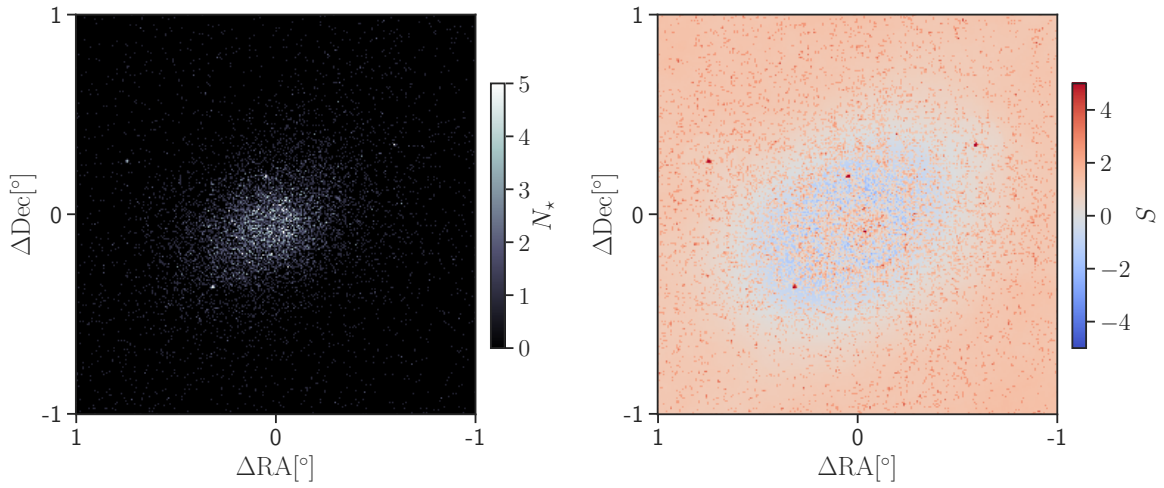

Figure 2.2: **Left**: the two-dimensional histogram of *Gaia* DR2 sources selected using Equations 2.1, 2.2, and 2.3 around the Fornax dwarf. **Right**: the over-density significance (S) map according to Equation 2.8.

the MW dwarf galaxies. The overall scientific validation of the data is described in Arenou et al. (2018).

The entire analysis of this paper utilizes the `GAIA_SOURCE` catalog of *Gaia* DR2 (ESA & DPAC, 2019), particularly the position `ra` and `dec` (α and δ), the proper motion (PM) `pmra` and `pmdec` (μ_α and μ_δ), the G-band magnitude `phot_g_mean_mag` (G), and the value of the `astrometric_excess_noise` parameter (ϵ). Gaia Collaboration et al. (2018a) contains the detail on the contents and the properties of this catalog. We use this dataset to identify stellar density peaks as possible candidates of GCs around in the vicinity and inside nearby dwarf galaxies.

Throughout the whole paper, we apply two main selection cuts on the *Gaia* catalog. The first selection is

$$17 < \mathrm{G} < 21. \tag{2.1}$$

The faint-magnitude cut $\mathrm{G} < 21$ approximately corresponds to the faint-end limit of *Gaia* DR2; Gaia Collaboration et al. (2018a) reported that only 4 percent of the sources are fainter than $\mathrm{G} = 21$ and those sources lack PMs and parameters. There are two reasons for the bright-magnitude cut $\mathrm{G} > 17$. The first reason to get rid of the bright stars is that the foreground contamination dominates at bright magnitudes. Conversely, the expected rapid rise of the stellar luminosity function for the majority of GCs and dwarf galaxies at reasonable distances from the Sun at $\mathrm{G} > 17$ results in the majority of stars being fainter than $\mathrm{G} = 17$. The other reason is that most bright GCs with large numbers of $\mathrm{G} < 17$ stars would have likely been detected already. The second selection criterion is

$$\ln \epsilon < 1.5 + 0.3 \max\{\mathrm{G} - 18, 0\}. \tag{2.2}$$

This cut is used to reject potentially extended sources (see Koposov et al., 2017; Wang et al., 2019, for more detail).

Another optional selection that we use to further clean the source list is based on the PM, with the goal of removing sources whose PMs are different from the mean PM of a given targeted dwarf galaxy, as these sources are less likely to be member stars of the given dwarf. For each targeted dwarf, we exclude stars with PMs (μ_α, μ_δ) differing from a systemic PM of the dwarf ($\mu_\alpha^{\rm dwarf}$, $\mu_\delta^{\rm dwarf}$) by more than three times the PM uncertainty (σ_{μ_α}, σ_{μ_δ}). That is, only the stars satisfying

$$\sqrt{(\mu_\alpha - \mu_\alpha^{\rm dwarf})^2 + (\mu_\delta - \mu_\delta^{\rm dwarf})^2} < 3\sqrt{\sigma_{\mu_\alpha}^2 + \sigma_{\mu_\delta}^2} \tag{2.3}$$

survive after the PM selection.

For example, Figure 2.1 shows the *Gaia* sources around the Fornax dwarf before and after the PM selection in Equation 2.3. The source distribution in PM space in the left panel shows that there are many foreground sources with PMs that are 10 – 100 mas yr^{-1} different from the PM of the dwarf. This PM selection is thus applied to remove this kind of contamination; the sources colored in orange survive after the selection. It is worth noting that the PM uncertainty of the studied dwarfs is around the order of 10^3 – 10^5 km s^{-1} (see Table 2.2) which is much larger than the typical velocity dispersion of dwarf galaxies around the order of 10 km s^{-1} (Walker et al., 2007) or 0.02 mas yr^{-1} if at 100 kpc, so the survived sources under this PM selection still have a fairly large range of internal space velocity. To investigate the PM selection for the stars that are more likely to be member stars of the Fornax dwarf, we draw a lasso with the black dashed lines to roughly distinguish the member stars in the red-giant branch of Fornax from the other stars in the color-magnitude diagram in the right panel. For the stars that are likely to be member stars inside of the lasso, 91 percent of the sources survive after the PM selection, whereas most of the sources outside of the lasso are excluded. Moreover, in the left panel of Figure 2.2, the stellar distribution after the PM selection retains the shape of the Fornax dwarf.

2.2.2 Kernel density estimation

Convolving the spatial distribution of the data with various kernels is a common approach to identify the excess number of stars associated with a satellite or clusters in imaging data. The density is calculated by convolving all the data points interpreted as delta functions with different kernels, e.g. a moving average in Walsh et al. (2009), two circular indicator functions in Torrealba et al. (2019a) and Gaussian kernels in Koposov et al. (2008a), Koposov et al. (2008b), and Drlica-Wagner et al. (2015).

To identify star clusters in dwarf galaxies, we use the kernel density estimation on the stellar distribution, while assuming the Poisson distribution of stellar number counts.

1. We obtain the distribution of stars

$$\Sigma(x, y) = \sum_i \delta(x - x_i, y - y_i) \tag{2.4}$$

where (x_i, y_i) is the position of the i^{th} star on the local coordinates which takes care of the projection effect.[1]

2. Using the circular indicator function with a given radius R defined as $\mathbb{1}(x, y; R) = \begin{cases} 1 \text{ if } x^2 + y^2 \leq R^2 \\ 0 \text{ otherwise} \end{cases}$, we define the inner kernel $K_{\text{in}}(x, y; \sigma_1) = \mathbb{1}(x, y; \sigma_1)$, where σ_1 corresponds to the scale of GCs which is 3, 5, or 10 pc. We then convolve $\Sigma(x, y)$ with $K_{\text{in}}(x, y; \sigma_1)$ to estimate the number density of stars on the scale of σ_1 as

$$\Sigma_{\text{in}}(x, y) = \Sigma(x, y) * K_{\text{in}}(x, y; \sigma_1). \tag{2.5}$$

3. Defining the outer kernel $K_{\text{out}}(x, y; \sigma_1, \sigma_2) = \mathbb{1}(x, y; \sigma_2) - \mathbb{1}(x, y; 2\sigma_1)$, we convolve $\Sigma(x, y)$ with $K_{\text{out}}(x, y; \sigma_1, \sigma_2)$ as

$$\Sigma_{\text{out}}(x, y) = \Sigma(x, y) * K_{\text{out}}(x, y; \sigma_1, \sigma_2) \tag{2.6}$$

to estimate the number density of stars on the annular area of radius between $2\sigma_1$ and σ_2, where $\sigma_2 > 2\sigma_1$ and σ_2 corresponds to either the angular scale of parent dwarf galaxy or a fixed angular scale of 0.5° (see more detail in the next paragraph).

4. We estimate the expected background number density within the inner kernel from $\Sigma_{\text{out}}(x, y)$ through the ratio of the inner and outer areas

$$\Sigma_{\text{bg}}(x, y) = \frac{\sigma_1^2}{\sigma_2^2 - (2\sigma_1)^2} \Sigma_{\text{out}}(x, y). \tag{2.7}$$

5. We convert the tail probability of Poisson into the z-score of the standard normal distribution to evaluate the significance as

$$S(x, y) = F^{-1}_{\text{N}(0,1)} \left(F_{\text{Poi}\left(\Sigma_{\text{bg}}(x,y)\right)} \left(\Sigma_{\text{in}}(x, y) \right) \right), \tag{2.8}$$

where F is the cumulative distribution function.

As an example, Figure 2.2 shows the original two-dimensional histogram of the sources around the Fornax dwarf in the left panel and the significance map of that stellar distribution in the right panel. According to the significance map, we identify positive detection with significance higher than a certain significance threshold. For nearby pixels with significance

[1] In the algorithm, we always divide a targeted area into small patches with a side of 0.5°. For each patch centered at (α_0, δ_0), we define the local coordinates (x, y) with the origin of $(x_0, y_0) = (\alpha_0, \delta_0)$. Since the patch is very small, we approximate the projection effect as $x \approx (\alpha - \alpha_0) \cos \delta_0$ and $y = \delta - \delta_0$.

Table 2.1: The nine known GCs and the two known galaxies found in our detection and their detected positions (α and δ), significance values (S), and inner kernel sizes (σ_1). Only Fornax 1 – 6 are actual clusters belonging to their parent dwarf galaxy.

Objects	α [°]	δ [°]	S	σ_1 [pc]
Fornax 1	40.5871	-34.1016	8.5	10
Fornax 2	39.6842	-34.8092	8.2	10
Fornax 3	39.9502	-34.2593	7.4	10
Fornax 4	40.0343	-34.5375	5.4	10
Fornax 5	39.2570	-34.1845	7.5	10
Fornax 6	40.0298	-34.4204	5.2	10
Palomar 3	151.3788	0.0731	7.3	10
Messier 75	301.5206	-21.9233	37.6	10
NGC 5466	211.3615	28.5321	37.7	10
Leo I	152.1122	12.3001	37.2	10
Sextans A	151.3799	0.0714	6.3	10

higher than the threshold, we merge them as one single positive detection if the radial distance between the pixels is shorter than the size of the inner kernel. We assign the maximum significance on the merged pixels as the detected significance and use the center of mass coordinates of the merged pixels as the detected position.

The main reason for σ_2 in step (iii) corresponding to either the angular scale of parent dwarf galaxy or the fixed angular scale of 0.5° is that the kernel density estimates are biased in crowded areas, which may lead to missing objects around big dwarfs. Given a dwarf with a half-light radius of r_h, σ_2 is chosen to be $0.5r_h$ or 0.5° for pixels inside ($r < r_h$) or outside ($r > r_h$) of the dwarf respectively, where r is the distance from the position of any pixel to the center of the dwarf. The latter large σ_2 of 0.5° is to take care of the sparse outskirts of the dwarf. Besides, when dealing with the pixels outside of the dwarf, we exclude the effect of the pixels inside of the dwarf ($r_h < r < r_h + 0.5°$) because the relatively high number density of stars in the dwarf will lead to over-estimate of $\Sigma_{\rm out}(x, y)$ which will suppress the background estimate too much later.

2.3 Results

The objective of the paper is to search for possibly missing GCs around the MW satellites by identifying stellar over-densities with the searching algorithm described in Section 2.2.2. The list of dwarf galaxies considered in this paper was created by selecting dwarf galaxies within the distance of 450 kpc from the Galactic Center with the exception of the LMC, the SMC, and the Sagittarius dwarf. The dwarf list in Table 2.2 summarizes all 55 targeted

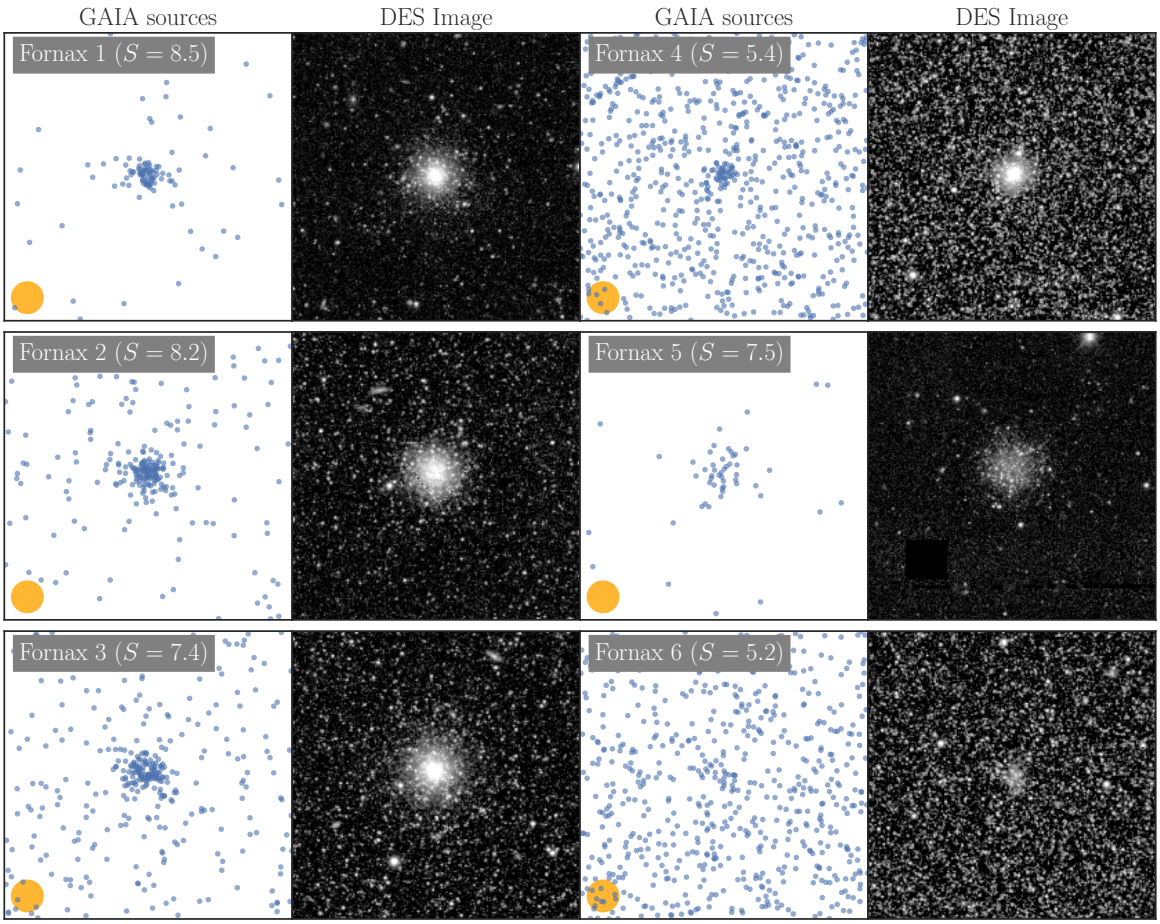

Figure 2.3: The source maps and the images of the six GCs of Fornax. **Left** and **Middle-right** panels: the stellar distributions of *Gaia* sources centered at each over-density satisfying the detection criteria. The legends show the names of GCs and their significance values S. The yellow circles illustrate the inner kernel size of 10 pc. The dimension of each panel is 100×100 pc^2. **Middle-left** and **Right** panels: the corresponding images from DES DR1 made with the HiPS.

dwarfs investigated in the paper and their properties. The reason to exclude the three most massive satellites of the MW is that their relatively large sizes will lead to a huge portion of the sky to be searched, which conflicts with our goal of conducting a targeted search. In the construction of the dwarf galaxy list, we use the data from the McConnachie (2012) compilation and include some of the recent discoveries: Antlia 2 (Torrealba et al., 2019b), Aquarius 2 (Torrealba et al., 2016b), Bootes 3 (Massari & Helmi, 2018), Carina 2 (Torrealba et al., 2018), Carina 3 (Torrealba et al., 2018), Cetus 3 (Homma et al., 2018), Crater 2 (Torrealba et al., 2016a), and Virgo I (Homma et al., 2016).

For each targeted dwarf, we search the area within the radius of $\min\{8^\circ, R_{\rm vir}\}$, where $R_{\rm vir}$ is the virial radius of a $10^9 M_\odot$ halo (Walker et al., 2007) (at the distance of 100 kpc this corresponds to 10°). We choose the inner kernel sizes of $\sigma_1 = 3$, 5, and 10 pc which covers the range of physical sizes of a typical GC (Brodie & Strader, 2006). We run the searching algorithm for each inner kernel size on the *Gaia* sources after the selections of Equation 2.3 if the dwarf has known measured PM (see Table 2.2), Equation 2.1, and Equation 2.2.

To balance the completeness of search with the number of false positives, we define two thresholds for identifying possible candidates: a significance threshold $S > 5$ and the limit of the number of stars inside the inner kernel $\Sigma_{\rm in} > 10$. For $S > 5$, as the z-score of the standard normal distribution, its false alarm probability is of the order of 10^{-7}.[2] Assuming a targeted dwarf at the distance of 100 kpc with a searching radius of 8°, the total number of spatial pixels is around the order of 10^8.[3] With the false alarm probability $\sim 10^{-7}$ on the targeted area of $\sim 10^8$ pixels, the number of expected false positives is around the order of 10. Moreover, we apply the other threshold, $\Sigma_{\rm in} > 10$, to prevent a large number of false positives for the pixels with very low background number density. For example in Figure 2.2, it is noticeable that the significance can easily be large in the area with very sparse stellar density even if only a handful of stars are detected in the inner kernel. These pixels typically have $\Sigma_{\rm out} < 1$ where the significance estimator breaks down due to the very low rate parameter of Poisson. Hence by applying $\Sigma_{\rm in} > 10$, we effectively increase the threshold on S for pixels with $\Sigma_{\rm out} < 1$, e.g. the threshold is $S = 5.6$ for $\Sigma_{\rm out} = 1$ and $S = 8.9$ for $\Sigma_{\rm out} = 0.1$. This avoids the detection of false-positive peaks due to Poisson noise in the $\Sigma_{\rm bg}$ estimates, binary stars, or unresolved galaxies in *Gaia* that are expected to show more clustering than stars. Particularly for binary star systems or unresolved galaxies, the pairs of them are much more likely to occur because they are more correlated; thus they are likely to reach 5 significance and cause false positives.

After running the searching algorithm on all 55 targeted dwarfs, we identify eleven stellar over-density candidates, based on the highest detected significance of each candidate if it is detected multiple times with different searching parameters. Cross-matched with the SIMBAD

[2] $\int_5^\infty \frac{1}{\sqrt{2\pi}} e^{-0.5z^2} dz \sim 10^{-7}$

[3] The searching radius of 8° corresponds to $\sim 10^4$ pc at the distance of 100 kpc so the searching area is $\sim 10^8$ pc^2. With the spatial resolution ~ 1 pc^2, the total number of pixels is then $\sim 10^8$.

database (Wenger et al., 2000), all eleven candidates are known objects. Nine of them are known GCs: Fornax GC 1 – 5 (Shapley, 1938; Hodge, 1961), Fornax GC 6 (Shapley, 1939; Verner et al., 1981; Demers et al., 1994; Stetson et al., 1998; Wang et al., 2019), Messier 75 (Shapley & Sawyer, 1927), NGC 5466 (Shapley & Sawyer, 1927), and Palomar 3 (Wilson, 1955). The other two of them are known galaxies: the Leo I dwarf spheroidal galaxy (Harrington & Wilson, 1950) and the Sextans A dwarf irregular galaxy (Zwicky, 1942). We remark that Leo I is found when searching for overdensities near Segue I, as they are close to each other in the sky. Table 2.1 summarizes the eleven known objects and their detected positions (RA and Dec), significance values (S), and inner kernel sizes (σ_1). Figure 2.3 shows the stellar distribution of *Gaia* sources for the six GCs of Fornax and the corresponding images from DES DR1 (Abbott et al., 2018) made with the HiPS (Hierarchical Progressive Surveys, Fernique et al., 2015). The yellow circles show the inner kernel of 10 pc (note that it happens to be that all the significance values with 10 pc are greater than with 3 or 5 pc in our detection of the nine GCs). Most of those known GCs are detected with the strong significance of $S > 7$ except for Fornax GC 6 with $S = 5.2$, which emphasizes that our algorithm can detect GCs from the regions of high stellar density such as Fornax GC 6. Figure 2.3 further indicates that the significance values are reasonable: bright GCs located at low-density areas (e.g. Fornax GC 1 and 2) have high significance ($S > 8$) and faint GCs located at high-density areas (e.g. Fornax GC 6) have low significance ($S \sim 5$). However, we are aware of missing the ultra-faint GC in the Eridanus 2 in our detection, which we will further discuss later in Section 2.4.2.

2.4 Discussion

2.4.1 Detection limit in V-band magnitude

In this section, we will demonstrate how we carry out the detection limit in V-band magnitude $M_V^{\rm lim}$ of the search for each targeted dwarf, which indicates that GCs brighter than $M_V^{\rm lim}$ are detectable in our search. To do so, we generate 1000 mock GCs with luminosity in the range of $-10 < M_V < 0$ assuming the age = 12 Gyr and [Fe/H] = -2 of the stellar populations. Sampling the stars of each GC population according to the log-normal initial mass function in Chabrier (2005), we interpolate the isochrone based on the PARSEC isochrone (Bressan et al., 2012), then utilizing the isochrones of all the mock GC stellar populations to carry out the detection limit for each targeted dwarf as follows.

Given a targeted dwarf, to compute the detection limit $M_V^{\rm lim}$, we first calculate the number of observable stars of each mock GC satisfying the G-band selection by counting the number of stars within $17 < G < 21$ according to its isochrone at the distance of the dwarf. Based on the number of observable stars, we compute the number of stars of each GC within the inner

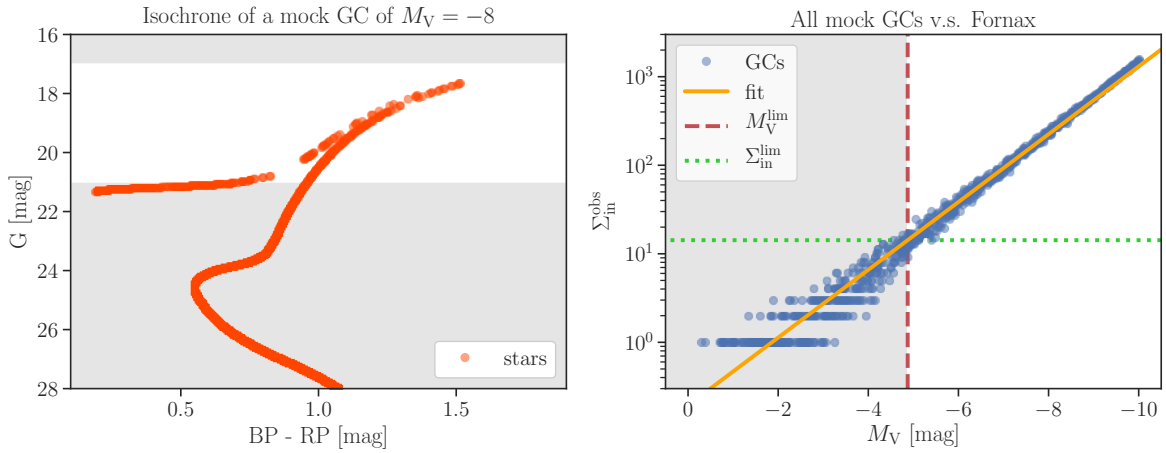

Figure 2.4: **Left**: the isochrone of a single mock GC of $M_V = -8$ at the distance of the Fornax dwarf spheroidal. The stars in the white area are observable within our *Gaia* G-band cut. **Right**: the numbers of observable stars Σ_{in}^{obs} versus M_V of all 1000 mock GCs for Fornax. The green dashed line shows the threshold number of stars Σ_{in}^{lim} to reach 5 significance according to the maximum background estimate of Fornax. The yellow line is the linear best fit and the red dashed line is the detection limit M_V^{lim} derived based on the best fit and Σ_{in}^{lim}. The GCs in the white area are detectable.

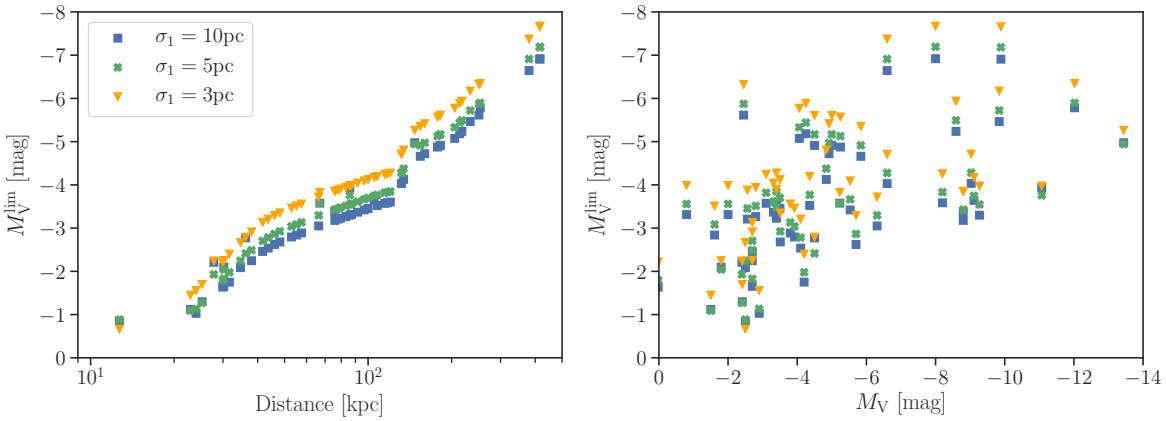

Figure 2.5: The detection limit M_V^{lim} of all targeted dwarfs with the different inner kernels σ_1 = 3, 5, and 10 pc. **Left**: M_V^{lim} versus the distance of the dwarfs. **Right**: M_V^{lim} versus the M_V of the dwarfs.

kernel size σ_1 as

$$\Sigma_{\rm in}^{\rm obs} = f\,(\sigma_1;\, r_{\rm h} = 3{\rm pc}) \times (\text{total number of observable stars})\,, \tag{2.9}$$

where $f\,(\sigma_1;\, r_{\rm h}) = \frac{\sigma_1^2}{\sigma_1^2+r_{\rm h}^2}$ is the fraction of the number of stars within the radius of σ_1 according to the Plummer model of 2D surface density profile of a GC with a half-light radius $r_{\rm h} = 3$pc (Plummer, 1911). With $\Sigma_{\rm in}^{\rm obs}$ of all the mock GCs at hand, we then use a linear best fit to describe the relation between $\log_{10}\left(\Sigma_{\rm in}^{\rm obs}\right)$ and $M_{\rm V}$ of the GCs. According to the maximum background estimate of the given dwarf, we know the threshold number of stars $\Sigma_{\rm in}^{\rm lim}$ to be observed to reach 5 significance. By comparing $\Sigma_{\rm in}^{\rm lim}$ to the best fit, we can obtain the detection limit $M_{\rm V}^{\rm lim}$ for the given targeted dwarf.

We take the Fornax dwarf as an example of the procedure of injection of mock GCs. In the left panel of Figure 2.4, we show the isochrone of a single mock GC of $M_{\rm V} = -8$ at the distance of Fornax and the stars in the white area are observable within our *Gaia* G-band cut. By counting the number of stars satisfying $17 < G < 21$ corrected by the fraction of stars located within the inner kernels, we know the number of observable stars $\Sigma_{\rm in}^{\rm obs}$ for the given mock GC. Applying the calculation of $\Sigma_{\rm in}^{\rm obs}$ for each mock GC, we show the relation between $\Sigma_{\rm in}^{\rm obs}$ and $M_{\rm V}$ for all the mock GCs in the right panel of Figure 2.4. The green dashed line shows the threshold number of stars $\Sigma_{\rm in}^{\rm lim}$ to reach $S = 5$ according to the maximum background estimate of Fornax; that is, the GCs above the green dashed line are expected to be detectable. Fitting the relation between $\log_{10}\left(\Sigma_{\rm in}^{\rm obs}\right)$ and $M_{\rm V}$ with a linear best fit as shown in the yellow line, we solve the detection limit $M_{\rm V}^{\rm lim}$ by finding the value of $M_{\rm V}$ satisfying the fit at the value of $\Sigma_{\rm in}^{\rm lim}$ (the green dashed line). The red dashed line indicates the derived $M_{\rm V}^{\rm lim}$ and the GCs brighter than $M_{\rm V}^{\rm lim}$ in the white area are thus detectable in our search. It is worth noting that the *Gaia* magnitude limit is brighter than G = 21 in some areas of the sky, which will decrease $\Sigma_{\rm in}^{\rm obs}$ if it happens in our targeted area, resulting in a brighter $M_{\rm V}^{\rm lim}$.

Repeating the same calculation of $M_{\rm V}^{\rm lim}$ for all the targeted dwarfs, we obtain the detection limits of the dwarfs and show the comparison of the derived $M_{\rm V}^{\rm lim}$ to the distances and the luminosities of the dwarfs in Figure 2.5. In the left panel, there is an obvious trend that the $M_{\rm V}^{\rm lim}$ are fainter for the dwarfs that are closer because the injected $\Sigma_{\rm in}^{\rm obs}$ of the GCs for these dwarfs with small distance modulus is typically larger than that of the dwarfs with large distance modulus. On the other hand in the right panel, the relation between $M_{\rm V}^{\rm lim}$ and $M_{\rm V}$ of the dwarfs is more scattered yet there is a slight trend of fainter $M_{\rm V}^{\rm lim}$ for the fainter dwarfs. This is likely because the faint dwarfs, compared to the bright ones, tend to have less-crowded stellar distributions and hence lower thresholds $\Sigma_{\rm in}^{\rm lim}$ to reach 5 significance. To sum up, the faint $M_{\rm V}^{\rm lim}$ for the close dwarfs or the faint dwarfs is reasonable because the ability of dwarfs to hide GCs from our detection is intuitively weaker for the dwarfs that are closer or fainter. It is also worth noting that most of the time $M_{\rm V}^{\rm lim}$ with $\sigma_1 = 10$ pc is the faintest, $M_{\rm V}^{\rm lim}$ with $\sigma_1 = 5$ pc is the intermediate, and $M_{\rm V}^{\rm lim}$ with $\sigma_1 = 3$ pc is the brightest mainly because the fractions of stars observed within the inner kernels are around 0.9, 0.7 and 0.5 for $\sigma_1 = 10$, 5,

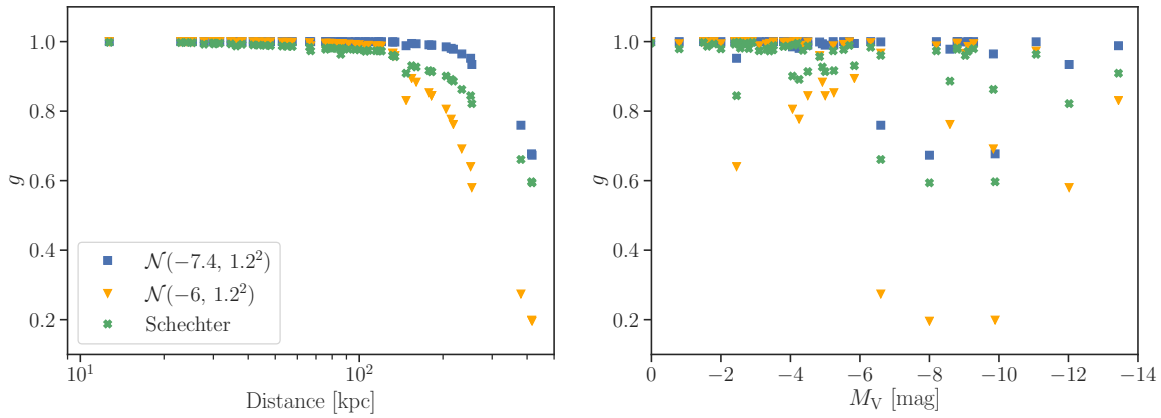

Figure 2.6: The completeness g of the GC search for all targeted dwarfs with three GCLFs: the Gaussian of $\mathcal{N}(-7.4,\ 1.2^2)$ and $\mathcal{N}(-6,\ 1.2^2)$ and the evolved Schechter in Jordán et al. (2007). **Left**: Completeness versus the distance of the dwarfs. **Right**: Completeness versus dwarf galaxy luminosity.

and 3 pc respectively according to the Plummer model. That is, the low $\Sigma_{\rm in}^{\rm obs}$ due to the small fraction for small σ_1 makes the faint GCs less likely to meet 5 significance, thus resulting in a bright $M_{\rm V}^{\rm lim}$.

2.4.2 Completeness of the search

With the limiting magnitudes of GC detection at hand, we can calculate the completeness of the search according to the typical GC luminosity function (GCLF). In this section, we will calculate the completeness factor g with three different GCLFs: (a) the typical MW GCLF in Harris (2001): a Gaussian distribution with a peak at $M_{\rm V} = -7.4$ and a standard deviation of 1.2, $\mathcal{N}(-7.4,\ 1.2^2)$, (b) the evolved Schechter function in Jordán et al. (2007) with a peak at $M_{\rm V} \sim -7.4$, and (c) a presumed Gaussian distribution with a peak at $M_{\rm V} = -6$ and a standard deviation of 1.2, $\mathcal{N}(-6,\ 1.2^2)$. We calculate g by evaluating the cumulative distribution functions of those GCLFs at $M_{\rm V}^{\rm lim}$ based on the search with $\sigma_1 = 10$ pc thanks to its better detecting sensitivity compared to $\sigma_1 = 3$ and 5 pc (all the detected objects with the highest significance are detected with $\sigma_1 = 10$ pc in Section 2.3).

We begin with the GCLF in (a); in the MW, the GCLF is approximately a Gaussian distribution of $\mathcal{N}(-7.4,\ 1.2^2)$ (Harris, 2001). With this MW GCLF, we compute the completeness factor g and show them in the blue points in Figure 2.6. The completeness of the search is higher than 90 percent for most of the dwarfs and around 70 percent for the lowest three, Eridanus 2, Leo T, and Phoenix. This high completeness is a consequence of $M_{\rm V}^{\rm lim} > -7$ for all the dwarfs; that is, the detection limits are fainter than the peak magnitude of the MW GCLF. Besides, as a result of the trend of brighter $M_{\rm V}^{\rm lim}$ for the farther targeted dwarfs in the left panel of Figure 2.5, the completeness gets lower for the dwarfs that are more distant.

In addition, Villegas et al. (2010) described that the dispersion of GCLF can be as small as 0.5 for small dwarfs. Calculating the completeness with this GCLF, we find that the result is almost the same as that of the MW GCLF.

Compared to the Gaussian MW GCLF peaking at $M_V = -7.4$, the evolved Schechter function with a similar peak magnitude proposed in Jordán et al. (2007) can describe the GCLF well too, particularly taking good care of the low-mass faint GCs. We compute the completeness factor g with this GCLF as shown in the green points in Figure 2.6, finding that the difference in g with this GCLF from the traditional Gaussian is less than 5 – 10 percent lower. The reason for the larger difference ($\sim$ 10 percent) in g of the two GCLFs for the targeted dwarfs that are more distant than 100 kpc is that the probability density of the evolved Schechter function is higher than that of the Gaussian MW GCLF in the faint end. Thus as these dwarfs have brighter $M_V^{\rm lim}$ than the close dwarfs, their cumulative distribution functions at $M_V^{\rm lim}$ of the evolved Schechter GCLF are lower than that of the Gaussian MW GCLF. On the other hand, for the dwarfs that are closer than 100 kpc, $M_V^{\rm lim}$ is much fainter than the peaks of the two GCLFs so the corresponding g approaches 1 for both GCLFs.

So far, we have assumed the GC population for all the dwarfs follows the GCLFs based on the results from bright galaxies, the Gaussian in Harris (2001) and the evolved Schechter in Jordán et al. (2007). These two GCLFs have similar peaks but different shapes: the evolved Schechter one extends more toward the faint end to account for faint GCs (see the black curves in Figure 2.8). However, these GCLFs might not hold in the faint host galaxies such as the faint satellites of the MW since there has been no reason for them being universal. Especially some of the dwarfs investigated in the paper are even fainter than the peak magnitude of these GCLFs, whether such systems may host GCs that are brighter than the dwarfs themselves is unclear, and is probably unlikely. Despite the lack of robust constraints on this, van den Bergh (2006) has pointed out that the peak of GCLF can be at $M_V = -5$ for faint galaxies. Moreover, the peak magnitude of GCLFs for different galaxies can vary in the range of $-7 < M_V < -5$ (see Richtler, 2003, Table 1 in particular). Therefore, we look at the known GC populations of the MW, NGC 6822, Sagittarius, Fornax, and Eridanus 2 in Appendix 2.A and decide to consider the peak of GCLF at $M_V = -6$ based on Figure 2.8 to calculate the completeness again. The orange points in Figure 2.6 show the completeness g computed with the GCLF $\mathcal{N}(-6,\ 1.2^2)$. As this GCLF peaks at the fainter magnitude than the other two GCLFs, g hardly changes for close dwarfs with much fainter $M_V^{\rm lim}$ than the peak of GCLF at $M_V = -6$ whereas g drops for the ones that are more distant than 100 kpc with small $M_V^{\rm lim}$, e.g. $g = 20 - 30$ percent for Eridanus 2, Leo T, and Phoenix.

Section 2.3 has mentioned that the ultra-faint GC with the luminosity of $M_V = -3.5$ (Koposov et al., 2015; Crnojević et al., 2016) in the Eridanus 2 is missing in our detection. This is mainly because the luminosity of this GC is much fainter than the detection limit $M_V^{\rm lim} \sim -6.5$ for the Eridanus 2 in the search. Hosting the ultra-faint GC of $M_V = -3.5$ and having the luminosity of $M_V = -6.6$ close to the peak magnitude of the MW GCLF, the

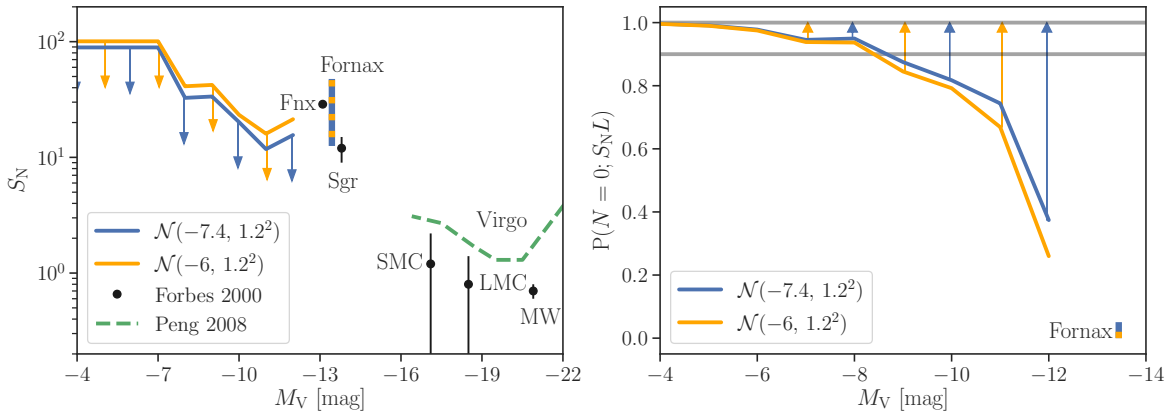

Figure 2.7: **Left**: The 90 percent credible intervals on S_N versus M_V of the dwarfs with two different GCLFs: double-sided intervals for Fornax and one-sided upper bounds for the others. The black data points are S_N of the MW, LMC, SMC, Sagittarius (Sgr), and Fornax (Fnx) in Forbes et al. (2000). The green dashed curve is the mean trend curve of the S_N for 100 galaxies in the Virgo Cluster in Peng et al. (2008). **Right**: The probability of hosting no GC for a galaxy with luminosity L and specific frequency S_N, $P(N = 0; S_N L)$. The 90 percent credible intervals on S_N is used to derive the range of $P(N = 0; S_N L)$. The two greys lines indicate $P(N = 0; S_N L) = 0.9$ and $P(N = 0; S_N L) = 1$.

Eridanus 2 is likely to have a GCLF peaking at a fainter magnitude than $M_V = -7.4$. As shown in Figure 2.6, the completeness g for the Eridanus 2 is 75 percent with the Gaussian MW GCLF and 65 percent with the evolved Schechter GCLF. When we shift the peak of GCLF to $M_V = -6$, the completeness factor drops to only 30 percent for the Eridanus 2, which further explains the existence of the ultra-faint GC in the Eridanus 2 while it is missing in our search.

2.4.3 Specific frequency of the globular clusters

The specific frequency of GCs is a common quantity to indicate the richness of GC system for a galaxy, first formulated as $S_N = N_{gc} \times 10^{0.4\,(M_{V,gal}+15)}$ where N_{gc} is the total number of GCs in a host galaxy and $M_{V,gal}$ is the absolute magnitude of the host galaxy (Harris & van den Bergh, 1981). With $L_{gal} \equiv 10^{-0.4\,(M_{V,gal}+15)}$ defined as the galactic V-band luminosity normalized to $M_V = -15$, $S_N = N_{gc}/L_{gal}$ then indicates the number of GCs per unit normalized luminosity. When the galaxy luminosity and the number of clusters are large, simply taking a ratio between the number and luminosity makes sense; however, a more statistical approach is required for dwarf galaxies.

Here, we define S_N as the specific frequency for a group of galaxies. In that case, the

observed number of clusters for each galaxy in a group will be Poisson distributed:

$$N_{\rm gc} \sim \text{Poisson}\left(S_{\rm N} L_{\rm gal}\right) \tag{2.10}$$

where $L_{\rm gal}$ is the luminosity of the galaxy and $N_{\rm gc}$ is the random variable describing the number of clusters in this galaxy. Assuming that our samples of GCs are incomplete with different completeness correction g for each dwarf, we can update the model to include incompleteness as

$$N_{\rm gc} \sim \text{Poisson}\left(S_{\rm N} g L_{\rm gal}\right). \tag{2.11}$$

Among the nine objects that we identify in our search in Section 2.3, only the six GCs found around the Fornax dwarf are associated with the parent dwarf galaxy. That is, the dwarfs targeted in the paper except for Fornax have no associated GCs detected around them. Due to the lack of associated GCs and the fact that most of the dwarfs are much fainter than Fornax, the formal $S_{\rm N}$ is hence expected to be zero with large upper bounds. To properly take into account the non-detections and to still be able to constrain the specific frequency of the dwarf population, we assume that $S_{\rm N}$ is constant for the dwarfs with similar luminosities and will provide upper bounds on $S_{\rm N}$ for the dwarf population as a whole.

Assuming that we look at m dwarfs as a group at once, we know the luminosity L_i and the completeness g_i for the $i^{\rm th}$ dwarf, where $1 \leq i \leq m$. The total expected number of observed GCs in this group of m dwarfs is the sum of the expected number of GCs in each dwarf. Defining $L \equiv \sum_{i=1}^{m} L_i g_i$ and with the constant specific frequency $S_{\rm N}$ shared among the m dwarfs, we can write down the total expected number of GCs as

$$\sum_{i=1}^{m} S_{\rm N} g_i L_i = S_{\rm N} \sum_{i=1}^{m} g_i L_i \equiv S_{\rm N} L. \tag{2.12}$$

Together with the definition of the total number of observed GCs of the m dwarfs as $N = \sum_{i=1}^{m} N_i$ where N_i is the number of observed GCs of the $i^{\rm th}$ dwarf from our detection. We model N similarly to Equation 2.11 as $N \sim \text{Poisson}\,(S_{\rm N} L)$ and therefore the likelihood function $P(N \mid S_{\rm N}) \propto S_{\rm N}^{N} e^{-S_{\rm N} L}$. Using the Jeffreys prior $S_{\rm N}^{-1/2}$ as the distribution of the parameter $S_{\rm N}$, we have the posterior distribution

$$P(S_{\rm N} \mid N) \propto P(S_{\rm N}) P(N \mid S_{\rm N}) \propto S_{\rm N}^{N-\frac{1}{2}} e^{-S_{\rm N} L}. \tag{2.13}$$

This is a Gamma distribution; that is, $S_{\rm N} \sim \text{Gamma}\left(N + \frac{1}{2}, L\right)$.

With the posterior in Equation 2.13, we construct the 90 percent credible intervals on the parameter $S_{\rm N}$ with the Gaussian MW GCLF for the dwarfs as shown in the blue curve in the left panel of Figure 2.7. Also, we show the $S_{\rm N}$ of the MW and its four most luminous satellites (LMC, SMC, Sagittarius, and Fornax) based on Forbes et al. (2000) and the mean trend curve of $S_{\rm N}$ of 100 galaxies in the Virgo Cluster from Peng et al. (2008). Separating the Fornax dwarf from the others due to its richness of GCs, we first calculate its double-sided credible

interval on the specific frequency of $12 < S_N < 47$. For the other dwarfs with no discovered GCs, we bin the ones brighter than $M_V = -7$ with a window width of 2 mag and look at the others all at once, where the value of $M_V = -7$ is chosen as it is close to the peak magnitude of the GCLF. For the dwarfs in each bin, we obtain the one-sided credible intervals as the upper bounds of the specific frequency: $S_N < 20$ for the dwarfs with $-12 < M_V < -10$, $S_N < 30$ for the dwarfs with $-10 < M_V < -7$, and $S_N < 90$ for the dwarfs with $M_V > -7$. Similarly, we also construct the credible intervals on S_N with the evolved Schechter GCLF for the dwarfs, finding a similar result as with the Gaussian MW GCLF. The difference in S_N with the two GCLFs is less than 5 – 10 percent so we only show the one with $\mathcal{N}(-7.4,\ 1.2^2)$ in Figure 2.7.

The reason for grouping the dwarfs fainter than $M_V = -7$ is that they are in general faint so the expected number of GCs is much smaller than one, which makes them not very informative. Besides, the posterior becomes more prior-dependent for the fainter dwarfs as well. Thus, finding no GCs for the dwarfs in the brighter M_V bins constrains the upper bounds stronger than in the fainter bins. Especially at $M_V < -10$, the relatively low upper bounds indicate that the Fornax dwarf has a relatively higher S_N than the other dwarfs, especially than the ones with $M_V < -10$.

As mentioned in Section 2.4.2, the completeness will drop if the GCLF peaks at a fainter M_V than the typical peak magnitude at $M_V = -7.4$, which would effectively increase the upper bounds on S_N because the dropping completeness decreases the L. We, therefore, calculate the credible intervals on S_N again for the dwarfs with the GCLF $\mathcal{N}(-6,\ 1.2^2)$ as the orange curve shows in the left panel of Figure 2.7, finding that the upper bounds on S_N with this shifted GCLF (the orange curve) are higher than that with the MW GCLF (the blue curve) as expected. This effect is also expected to influence the upper limits more for the fainter dwarfs since the GCLFs are expected to shift more if the host galaxies are fainter; however, the upper limit is already more prior-dependent and less informative on the faint end so this upper limit increasing effect is less influential.

Besides S_N, the probability of a galaxy with luminosity L and S_N to host N GCs, $P(N; S_N L)$, is also interesting. With the 90 percent credible intervals on S_N, we show the range of $P(N = 0; S_N L)$ for a galaxy with L based on the model $N \sim \text{Poisson}\,(S_N L)$ in the right panel of Figure 2.7, which indicates the probability of a galaxy to host no GCs. Except for Fornax, the upper limits of S_N result in the lower limits of $P(N = 0; S_N L)$. Based on $P(N = 0; S_N L)$, galaxies fainter than $M_V = -9$ have $P(N = 0; S_N L) > 0.9$, which means the probability of these galaxies to have at least one GC is lower than 10 percent. Our finding of $P(N = 0; S_N L) > 0.9$ for galaxies with $M_V > -9$ is in agreement with the claims of the lowest galaxy mass of $\sim 10^5 M_\odot$ or luminosity $M_V \sim -9$ to host at least one GC from Georgiev et al. (2010) and Forbes et al. (2018). This may further explain the observation that galaxies less massive than $10^6 M_\odot$ tend not to have nuclei (Sánchez-Janssen et al., 2019) if we assume that the nuclei originate from GCs sunk by dynamical friction to the center. Given our

constraints on the specific frequency, Eridanus 2 with $M_V \sim -7$ has $P(N = 0; S_N L) \sim 0.95$, which highlights that the GC inside Eridanus 2 is indeed an outlier.

2.5 Conclusions

We have reported the results of the search for possibly hiding GCs around 55 dwarf galaxies within the distance of 450 kpc from the Galactic Center excluding the LMC, SMC, and Sagittarius. This was a targeted search around the dwarfs so we excluded those three satellites to avoid a huge portion of the sky to be searched due to their relatively large sizes. For each targeted dwarf galaxy, we have investigated the stellar distribution of the sources in *Gaia* DR2, selected with the magnitude, proper motion, and stellar morphology cuts.

Using the kernel density estimation and the Poisson statistics of stellar number counts, we have identified eleven stellar density peaks of above 5 significance as possible GC candidates in the targeted area. Cross-matching the eleven possible candidates with the SIMBAD database and existing imaging data, we have found that all of them are known objects: Fornax GC 1 – 6, Messier 75, NGC 5466, Palomar 3, Leo I and Sextans A. Only the six GCs of Fornax are associated with the parent dwarf galaxy.

We have calculated the GC detection limit in M_V for each dwarf using 1000 simulated GCs, finding that $M_V^{\rm lim} > -7$ for all the dwarfs. According to the $M_V^{\rm lim}$ of the dwarfs, we have then calculated the completeness of detection with the Gaussian MW GCLF $\mathcal{N}(-7.4,\ 1.2^2)$, the evolved Schechter GCLF peaking at $M_V^{\rm lim} \sim -7.4$, and the assumed Gaussian GCLF $\mathcal{N}(-6,\ 1.2^2)$. With the Gaussian MW GCLF and the evolved Schechter GCLF, the completeness of the detection for most of the dwarfs was higher than 90 percent and even that of the lowest three, Eridanus 2, Leo T, and Phoenix, was around 70 percent. With the assumed Gaussian GCLF, the completeness of our search was lower for the dwarfs that are more distant than 100 kpc, such as the Eridanus 2, Leo T, and Phoenix where it reached 20 – 30 percent. Using the completeness, we have constructed the 90 percent credible intervals on the GC specific frequency S_N of the MW dwarf galaxies. The Fornax dwarf had the credible interval on the specific frequency of $12 < s < 47$, the dwarfs with $-12 < M_V < -10$ had $S_N < 20$, the dwarfs with $-10 < M_V < -7$ had $S_N < 30$, and dwarfs with $M_V > -7$ had non-informative $S_N < 90$. Based on these credible intervals on S_N, we have derived the probability of galaxies to host GCs given their luminosity, finding that the probability of galaxies fainter than $M_V = -9$ to possess GCs is lower than 10 percent.

Acknowledgements

We acknowledge the support by NSF grants AST-1813881, AST-1909584, and Heising-Simons Foundation grant 2018-1030. This paper has made use of the Whole Sky Database (wsdb) created by Sergey Koposov and maintained at the Institute of Astronomy, Cambridge

Table 2.2: The list of properties of the studied dwarf galaxies: the positions (α and δ), the heliocentric distance ($D_\odot$), the V-band magnitude (M_V), the proper motions (μ_α and μ_δ), the reference (ref.), and the 3 $\sigma_\mu = 3\sqrt{\sigma^2_{\mu_\alpha} + \sigma^2_{\mu_\delta}}$ PM uncertainty converted to km s^{-1} at the distance of the dwarf.

dwarf	α [°]	δ [°]	$D_\odot$ [kpc]	M_V [mag]	ref.[a]	μ_α [mas yr^{-1}]	μ_δ [mas yr^{-1}]	ref.[a]	3 σ_μ [km s^{-1}]
Antlia 2	143.89	-36.77	132.0	-9.0	T19b	-0.095 ± 0.018	0.058 ± 0.024	T19b	2e+04
Aquarius 2	338.48	-9.33	107.9	-4.4	T16b	-0.252 ± 0.526	0.011 ± 0.448	Fritz et al. (2018)	3e+05
Bootes I	210.02	14.50	66.4	-6.3	M12	-0.459 ± 0.041	-1.064 ± 0.029	Gaia Collab. 18b.	2e+04
Bootes II	209.50	12.85	41.7	-2.7	M12	-2.686 ± 0.389	-0.530 ± 0.287	Fritz et al. (2018)	9e+04
Bootes III	209.25	26.80	46.0	-5.7	MH18	-1.210 ± 0.130	-0.920 ± 0.170	MH18	5e+04
Canes Venatici I	202.01	33.56	217.8	-8.6	M12	-0.159 ± 0.094	-0.067 ± 0.054	Fritz et al. (2018)	1e+05
Canes Venatici II	194.29	34.32	100.0	-4.9	M12	-0.342 ± 0.232	-0.473 ± 0.169	Fritz et al. (2018)	1e+05
Carina	100.40	-50.97	105.2	-9.1	M12	0.495 ± 0.015	0.143 ± 0.014	Gaia Collab. 18b.	1e+04
Carina 2	114.11	-58.00	36.2	-4.5	T18	1.810 ± 0.080	0.140 ± 0.080	MH18	2e+04
Carina 3	114.63	-57.90	27.8	-2.4	T18	3.035 ± 0.120	1.558 ± 0.136	Simon (2018)	2e+04
Cetus II	19.47	-17.42	29.9	0.0	M12				
Cetus III	31.33	-4.27	251.0	-2.4	Homma18				
Columba I	82.86	-28.03	182.0	-4.5	M12	-0.020 ± 0.240	-0.040 ± 0.300	Pace & Li (2019)	3e+05
Coma Berenices	186.75	23.90	43.7	-4.1	M12	0.471 ± 0.108	-1.716 ± 0.104	Fritz et al. (2018)	3e+04
Crater 2	177.31	-18.41	117.5	-8.2	T16a	-0.184 ± 0.061	-0.106 ± 0.031	Fritz et al. (2018)	4e+04
Draco	260.05	57.92	75.9	-8.8	M12	-0.019 ± 0.009	-0.145 ± 0.010	Gaia Collab. 18b.	5e+03
Draco II	238.20	64.57	24.0	-2.9	M12	1.170 ± 0.297	0.871 ± 0.303	Simon (2018)	5e+04
Eridanus 2	56.09	-43.53	380.2	-6.6	M12	0.160 ± 0.240	0.150 ± 0.260	Pace & Li (2019)	6e+05
Eridanus 3	35.69	-52.28	87.1	-2.0	M12				
Fornax	40.00	-34.45	147.2	-13.4	M12	0.376 ± 0.003	-0.413 ± 0.003	Gaia Collab. 18b.	3e+03
Grus I	344.18	-50.16	120.2	-3.4	M12	-0.250 ± 0.160	-0.470 ± 0.230	Pace & Li (2019)	2e+05
Grus II	331.02	-46.44	53.0	-3.9	M12	0.430 ± 0.090	-1.450 ± 0.110	Pace & Li (2019)	3e+04
Hercules	247.76	12.79	131.8	-6.6	M12	-0.297 ± 0.118	-0.329 ± 0.094	Fritz et al. (2018)	9e+04
Horologium I	43.88	-54.12	79.4	-3.4	M12	0.950 ± 0.070	-0.550 ± 0.060	Pace & Li (2019)	3e+04
Horologium II	49.13	-50.02	78.0	-2.6	M12				
Hydra II	185.43	-31.99	134.3	-4.8	M12	-0.416 ± 0.519	0.134 ± 0.422	Fritz et al. (2018)	4e+05
Indus I	317.20	-51.17	100.0	-3.5	M12				
Indus II	309.72	-46.16	213.8	-4.3	M12				
Leo I	152.12	12.31	253.5	-12.0	M12	-0.097 ± 0.056	-0.091 ± 0.047	Gaia Collab. 18b.	9e+04
Leo II	168.37	22.15	233.4	-9.8	M12	-0.064 ± 0.057	-0.210 ± 0.054	Gaia Collab. 18b.	8e+04
Leo IV	173.24	-0.53	154.2	-5.8	M12	-0.590 ± 0.531	-0.449 ± 0.358	Fritz et al. (2018)	5e+05
Leo V	172.79	2.22	177.8	-5.3	M12	-0.097 ± 0.557	-0.628 ± 0.302	Fritz et al. (2018)	5e+05
Leo T	143.72	17.05	416.9	-8.0	M12				
Pegasus 3	336.09	5.42	205.1	-4.1	M12				
Phoenix	27.78	-44.44	415.0	-9.9	M12	0.079 ± 0.099	-0.049 ± 0.120	Fritz et al. (2018)	3e+05
Phoenix 2	355.00	-54.41	83.2	-2.8	M12	0.490 ± 0.110	-1.030 ± 0.120	Pace & Li (2019)	6e+04
Pictoris I	70.95	-50.28	114.8	-3.1	M12				
Pisces II	344.63	5.95	182.0	-5.0	M12	-0.108 ± 0.645	-0.586 ± 0.498	Fritz et al. (2018)	7e+05
Reticulum II	53.93	-54.05	30.2	-2.7	M12	2.340 ± 0.120	-1.310 ± 0.130	MH18	2e+04
Reticulum III	56.36	-60.45	91.6	-3.3	M12	-1.020 ± 0.320	-1.230 ± 0.400	Pace & Li (2019)	2e+05
Sagittarius II	298.17	-22.07	67.0	-5.2	M12	-1.180 ± 0.140	-1.140 ± 0.110	MH18	6e+04
Sculptor	15.04	-33.71	85.9	-11.1	M12	0.082 ± 0.005	-0.131 ± 0.004	Gaia Collab. 18b.	3e+03
Segue I	151.77	16.08	22.9	-1.5	M12	-1.697 ± 0.195	-3.501 ± 0.175	Fritz et al. (2018)	3e+04
Segue II	34.82	20.18	34.7	-2.5	M12	1.270 ± 0.110	-0.100 ± 0.150	MH18	3e+04
Sextans I	153.26	-1.61	85.9	-9.3	M12	-0.496 ± 0.025	0.077 ± 0.020	Gaia Collab. 18b.	1e+04
Triangulum II	33.32	36.18	30.2	-1.8	M12	0.651 ± 0.193	0.592 ± 0.164	Simon (2018)	4e+04
Tucana II	342.98	-58.57	57.5	-3.8	M12	0.910 ± 0.060	-1.160 ± 0.080	Pace & Li (2019)	3e+04
Tucana III	359.15	-59.60	25.2	-2.4	M12	-0.030 ± 0.040	-1.650 ± 0.040	Pace & Li (2019)	7e+03
Tucana IV	0.73	-60.85	48.1	-3.5	M12	0.630 ± 0.250	-1.710 ± 0.200	Pace & Li (2019)	7e+04
Tucana V	354.35	-63.27	55.2	-1.6	M12				
Ursa Major I	158.72	51.92	96.8	-5.5	M12	-0.659 ± 0.093	-0.635 ± 0.131	Simon (2018)	7e+04
Ursa Major II	132.88	63.13	31.6	-4.2	M12	1.661 ± 0.053	-1.870 ± 0.065	Simon (2018)	1e+04
Ursa Minor	227.29	67.22	75.9	-8.8	M12	-0.182 ± 0.010	0.074 ± 0.008	Gaia Collab. 18b.	4e+03
Virgo I	180.04	-0.68	87.0	-0.8	Homma16				
Willman I	162.34	51.05	38.0	-2.7	M12	0.199 ± 0.187	-1.342 ± 0.366	Fritz et al. (2018)	7e+04

[a] Some of the citations are abbreviated: Gaia Collab. 18b. is for Gaia Collaboration et al. (2018b); MH18 is for Massari & Helmi (2018); M12 is for McConnachie (2012); T18 is for Torrealba et al. (2018); T19b is for Torrealba et al. (2019b); T16b is for Torrealba et al. (2016b); T16a is for Torrealba et al. (2016a); Homma18 is for Homma et al. (2018); Hargis16 is for Hargis et al. (2016); Homma16 is for Homma et al. (2016).

with financial support from the Science & Technology Facilities Council (STFC) and the European Research Council (ERC). This software has made use of the Q3C software (Koposov & Bartunov, 2006).

This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are being processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the *Gaia* MultiLateral Agreement (MLA). The *Gaia* mission website is https://www.cosmos.esa.int/gaia. The *Gaia* archive website is https://archives.esac.esa.int/gaia.

Software: NUMPY (van der Walt et al., 2011), SCIPY (Jones et al., 2001), PANDAS (Mckinney, 2010), MATPLOTLIB (Hunter, 2007), SEABORN (Waskom et al., 2016), ASTROPY (Astropy Collaboration et al., 2013), IMF (Ginsburg et al., 2020), SQLUTILPY (Koposov, 2018).

2.A GC luminosity functions

In Section 2.4, we adopt the Gaussian MW GCLF in Harris (2001) and the evolved Schechter GCLF in Jordán et al. (2007) for all the dwarfs to carry out the completeness factor and the specific frequency. However, the GCLF may shift toward the faint end for faint dwarfs, e.g. Richtler (2003); van den Bergh (2006). To investigate this, we show the GCLFs in the histogram with Gaussian probability density distributions of the MW ($M_V \sim -21$), NGC 6822 ($M_V \sim -16$), Sagittarius ($M_V \sim -14$), and Fornax ($M_V \sim -13$) with the solid curves in Figure 2.8. Besides, we also show the evolved Schechter GCLF with the black dashed curve and the ultra-faint GC of the Eridanus 2 ($M_V \sim -7$) with the red dashed line. We collect the GC lists for these galaxies according to Harris (2010), Veljanoski et al. (2015), Koposov et al. (2015), Vasiliev (2019), or the SIMBAD database. Based on the Gaussian distributions of the GCLFs and the existence of Eridanus 2 GC, there is a possible shift of the GCLF peak toward the faint luminosity for faint galaxies, e.g. the peaks of the dwarf galaxies are closer to $M_V \sim -6$ as opposed to the peak of the GC distribution in the MW at $M_V = -7.4$.

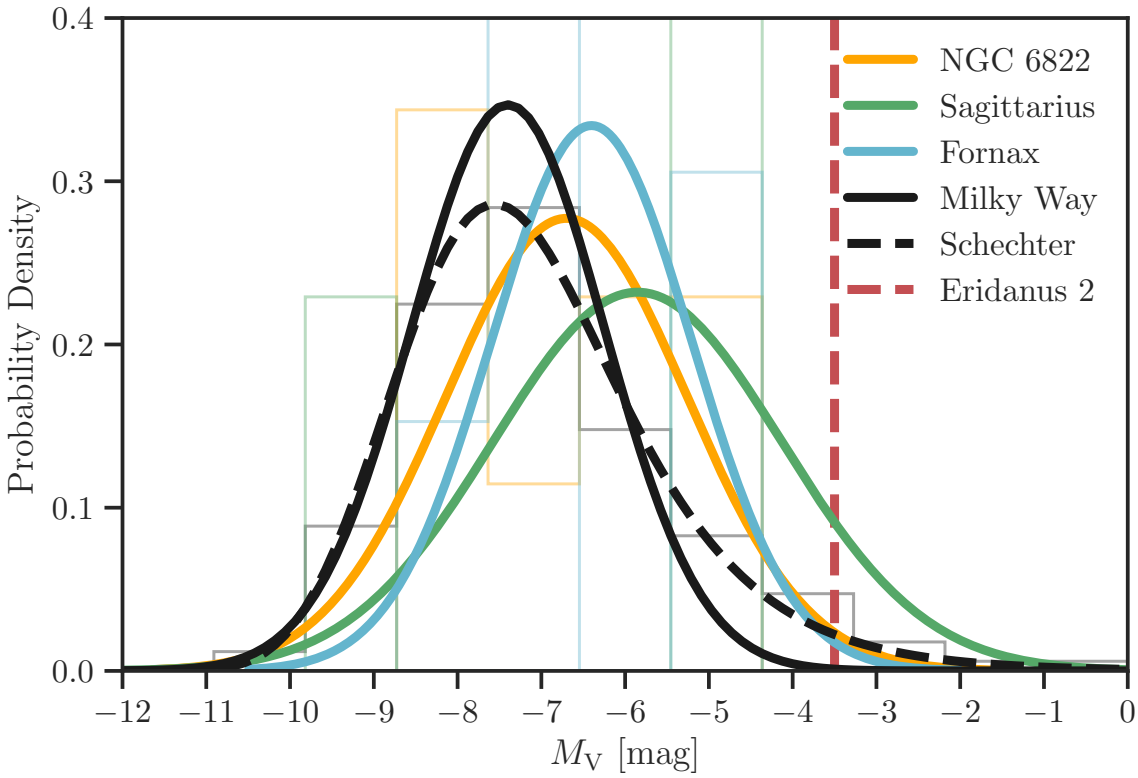

Figure 2.8: GCLFs of the Milky Way, NGC 6822, Sagittarius, and Fornax. For each galaxy, the solid curve is the Gaussian fit to the histogram of the probability density of the number of GCs in each magnitude bin. The black dashed curve is the evolved Schechter function in Jordán et al. (2007). The red dashed line indicates the ultra-faint GC of the Eridanus 2.

3 Identifying RR Lyrae in the ZTF DR3 dataset

Kuan-Wei Huang[1] and Sergey E. Koposov[2,3,1]

[1]McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[3] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

Abstract

We present a RR Lyrae (RRL) catalogue based on the combination of the third data release of the Zwicky Transient Facility (ZTF DR3) and *Gaia* EDR3. We use a multi-step classification pipeline relying on the Fourier decomposition fitting to the multi-band ZTF light curves and random forest classification. The resulting catalogue contains 71,755 RRLs with period and light curve parameter measurements and has completeness of 0.92 and purity of 0.92 with respect to the SOS *Gaia* DR2 RRLs. The catalogue covers the Northern sky with declination $\geq -28°$, its completeness is $\gtrsim 0.8$ for heliocentric distance ≤ 80 kpc, and the most distant RRL at 132 kpc. Compared with several other RRL catalogues covering the Northern sky, our catalogue has more RRLs around the Galactic halo and is more complete at low Galactic latitude areas. Analysing the spatial distribution of RRL in the catalogue reveals the previously known major over-densities of the Galactic halo, such as the Virgo over-density and the Hercules-Aquila Cloud, with some evidence of an association between the two. We also analyse the Oosterhoff fraction differences throughout the halo, comparing it with the density distribution, finding increasing Oosterhoff I fraction at the elliptical radii between 16 and 32 kpc and some evidence of different Oosterhoff fractions across various halo substructures.

3.1 Introduction

RR Lyrae (RRL) stars are pulsating variables with periodic light curves of a period ranging from 0.2 to 0.9 days (Smith, 1995), found primarily in the horizontal branches of old stellar systems (age $>$ 10 Gyr). These old, metal-poor ([Fe/H] < -0.5), bright ($M_V = 0.59$ at [Fe/H] $= -1.5$; Cacciari & Clementini (2003)) variable stars follow a well-understood

period-luminosity-metallicity (PLZ) relation (e.g. Cáceres & Catelan, 2008; Marconi, 2012). This relation makes RRLs excellent distance indicators for old, low-metallicity stellar populations in the outer halo of the Milky Way (e.g. Catelan et al., 2004; Vivas et al., 2004; Cáceres & Catelan, 2008; Sesar et al., 2010; Stetson et al., 2014; Fiorentino et al., 2015). Besides, RRLs are sufficiently luminous to be detected at large distances so that they can be the tracer of the halo substructures with a good spatial resolution (e.g. Vivas & Zinn, 2006; Sesar et al., 2010; Sesar et al., 2014; Baker & Willman, 2015; Torrealba et al., 2015; Martínez-Vázquez et al., 2019). Proposed by Sesar et al. (2014) (see also Baker & Willman, 2015), the fact that almost every Milky Way dwarf satellite galaxy has at least one RRL star opens up a gate of locating the Milky Way dwarf satellites even for the ones that are very faint by using distant RRL stars, for example, Antlia 2 (Torrealba et al., 2019b).

Being beneficial to many Galactic studies, there have been several RRL catalogues classified from existing surveys over the years, e.g. SDSS Stripe 82 (Sesar et al., 2010), CRTS (Drake et al., 2014), PS1 (Sesar et al., 2017), nTransits:2+ *Gaia* DR2 (Holl et al., 2018), SOS *Gaia* DR2 (Clementini et al., 2019a), ZTF DR2 (Chen et al., 2020), and DES Y6 (Stringer et al., 2021). The quality of the catalogues has progressed from being either deep with limited sky coverage (e.g. the SDSS Stripe 82 catalogue) or wide-coverage but not as deep (e.g. the CRTS catalogue) to having decent depth and wide sky coverage at the same time (e.g. the PS1 catalogue), pushing the Galactic studies furthermore. However, large-coverage and deep surveys usually suffer from significant incompleteness and contamination due to the low number of epochs in the light curves. This motivates us to identify a RRL catalogue from the ZTF survey thanks to its uniformly high number of observation epochs of light curves across the Northern sky while having decent depth. Another challenge of the catalogues is to cover the Galactic plane; the PS1, *Gaia* DR2, and ZTF DR2 catalogues do cover this area though the *Gaia* catalogue suffers the completeness issue here. The PS1 data suffer the issues of sparse temporal coverage, cadence, and asynchronous multi-band observations where they overcame them by the multi-stage classification in Hernitschek et al. (2016). Compared to the ZTF DR2 catalogue (Chen et al., 2020), the more recent data release used in this work provides more observation epochs which is beneficial for the light curve fitting to achieve more accurate period measurement. Also in this work for the period determination, we used all the bands simultaneously during the light curve fitting stage.

In this paper, we utilize the joint set of the *Gaia* early third data release (*Gaia* EDR3; Gaia Collaboration et al., 2020) and the third data release of the Zwicky Transient Facility (ZTF DR3; Masci et al., 2019) to classify RRL stars in the Northern sky. Thanks to the high angular resolution of *Gaia* and the fast cadence of ZTF observations, the sources in the joint set thus have high spatial resolution and multi-band light curves with large observation epochs. Assisted with the Specific Objects Study (SOS) *Gaia* DR2 RRL catalogue as the label, we process the dataset following the pipeline we come up with, which includes data labelling, feature building, and classifier training, to obtain the predicted RRL catalogue. In

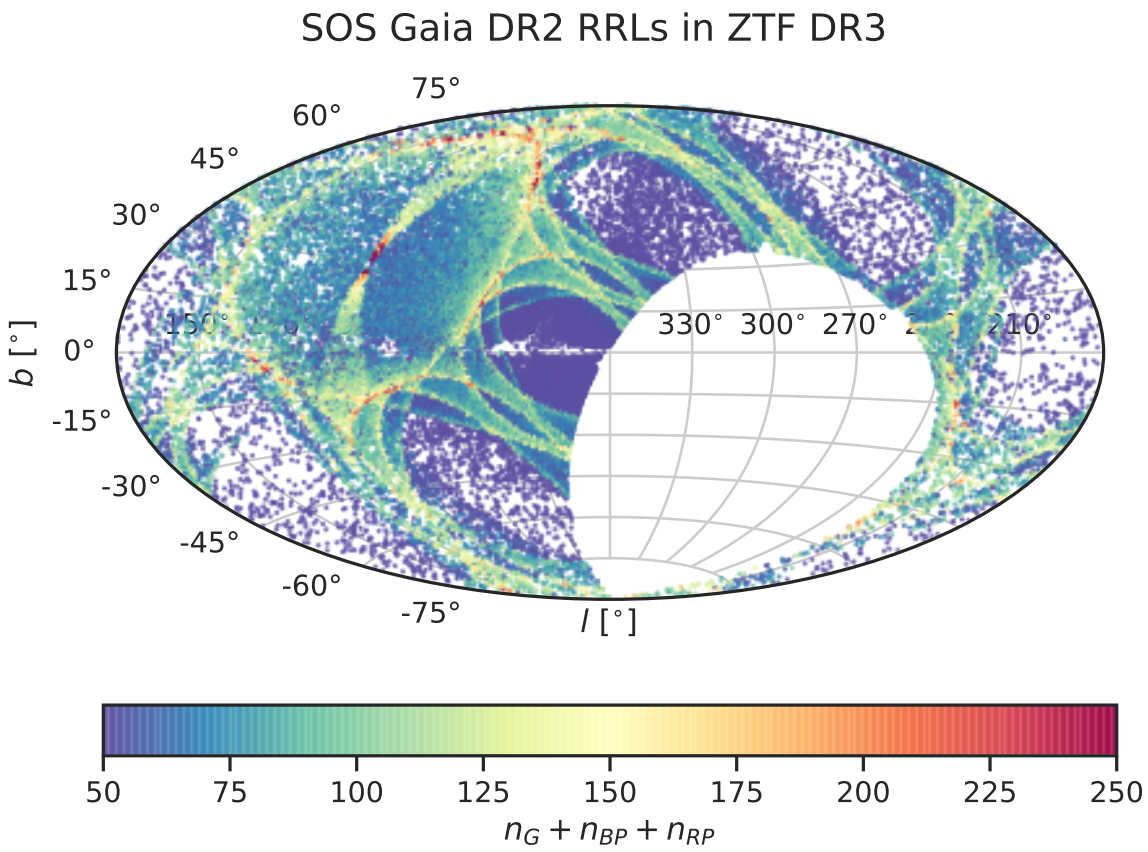

Figure 3.1: The spatial distribution of 48,365 SOS *Gaia* DR2 RRLs with detected ZTF DR3 light curves by the closest separation within one arcsec on the sky, colour-coded by the total number of *Gaia* epochs.

Section 3.2, we describe the datasets above in more detail. In Section 3.3, we explain the pipeline step by step. In Section 3.4, we demonstrate the classification results and present the predicted RRL catalogue. In Section 3.5, we conclude the paper.

3.2 Datasets

To identify RRLs in the Northern sky, we utilize three datasets in this work: ZTF DR3, *Gaia* EDR3, and the SOS *Gaia* DR2 RRL catalogue. The joint set of ZTF DR3 and *Gaia* EDR3 is the main dataset and the SOS *Gaia* DR2 RRL catalogue serves as the label for training models.

ZTF DR3 (Bellm et al., 2019): As a time-domain survey using the 48-inch Schmidt telescope equipped with a 47 squared degree camera at Palomar Observatory, ZTF started scanning the entire Northern sky in March 2018, covering the area of $\sim 3\pi$ steradians. In

the Northern sky of declination $> -31°$, ZTF has conducted two surveys: the Galactic Plane Survey with a one-day cadence of all visible fields at $|b| < 7°$ and the Northern Sky Survey with a three-day cadence at all fields with centres at $|b| > 7°$. Released in June 2020, ZTF DR3 contains the data collected during the first 21.4 months of the survey and has approximately 2.5 billion light curves constructed from the single-exposure extractions, with limiting magnitudes at about $g = 20.8$, $r = 20.6$, and $i = 19.9$, and the angular resolution of about one arcsec.

***Gaia* EDR3** (Gaia Collaboration et al., 2020): The space-based astrometric mission *Gaia* was launched by the European Space Agency in 2013 and started the whole-sky survey in 2014 (Gaia Collaboration et al., 2016). Released in December 2020, *Gaia* EDR3 contains the data collected during the first 34 months of the mission and has approximately 1.8 billion sources with 1.5 billion parallaxes and proper motions, down to the magnitude limit of G = 20.7. The angular separation limit, below which two sources are considered duplicates, has been lowered to 180 mas in EDR3, while it was 400 mas in DR2.

The SOS *Gaia* DR2 RRL catalogue: Using the Specific Objects Study (SOS) pipeline, Clementini et al. (2019a) presented 140,784 RRL stars in *Gaia* DR2 using the *Gaia* multi-band time-series photometry of all-sky candidate variables. We note that there are two RRL catalogues from Gaia DR2, the SOS catalogue and the nTransits:2+ catalogue (Holl et al., 2018), which is expected to be of lower quality due to a significantly smaller number of epochs per source.

To start the data preparation, we first create the joint dataset of ZTF DR3 and *Gaia* EDR3 by cross-matching the closest sources from the two surveys with an angular separation smaller than one arcsec. The resulting dataset contains 675,640,523 sources in the Northern sky down to the magnitude of about 20.5. The sources in the dataset thus are clearly identified but not mismatched single sources because *Gaia* has a higher angular resolution than ZTF. Each source in the joint set thus not only has the astrometric and photometric measurements from *Gaia* but also has the light curves in the gri bands from ZTF which in particular are essential for the classification pipeline explained in the following paragraphs. We note that we lose about 800 million sources from the original 1,471,263,267 sources in the ZTF DR3 dataset by this cross-match mainly because ZTF is slightly deeper than *Gaia* in some regions, despite the similar limiting magnitudes of the two surveys. However, the majority of the missing objects are very faint with magnitudes > 21 and have extremely large photometric errors.

Besides *Gaia* EDR3 and ZTF DR3, we use the SOS *Gaia* DR2 RRLs as the label for the binary classification task; we label each source in the joint dataset as true if it is classified as a RRL in the SOS *Gaia* DR2 RRL catalogue and as false otherwise. Amongst the 140,784 RRLs in the SOS *Gaia* DR2 RRL catalogue, 48,365 RRLs have ZTF light curves when cross-matched by the closest separation within one arcsec. In Figure 3.1, we show the distribution of these 48,365 *Gaia* RRLs in the Galactic coordinate colour-coded by the total number of

Gaia epochs, where n_G, n_{BP}, and n_{RP} are `num_clean_epochs_g`, `num_clean_epochs_bp`, and `num_clean_epochs_rp` respectively. Figure 3.1 illustrates the incompleteness issue that the SOS *Gaia* DR2 RRL catalogue suffers in the low-epoch areas due to the scanning trajectory of *Gaia*, which we will take into account during the classification pipeline.

3.3 The classification pipeline

With the dataset of 600 million sources in the joint set of *Gaia* EDR3 and ZTF DR3 and the label of the SOS *Gaia* DR2 RRLs, we then proceed to the supervised classification of RR Lyrae candidates through the multi-step process summarized below and described in detail in later sections.

The initial variability selection: To make the period fitting process computationally feasible, in Section 3.3.1, we first reduce the size of the dataset to 155,095,514 sources by applying an initial variability selection based on the residuals of constant flux fits to the ZTF light curves.

The broad selection of RRL candidates: Since the computational cost of the full Fourier period fitting for 155 million sources is still prohibitive, in Section 3.3.2, we perform a further filtering step by doing a discretised single sinusoidal fit to characterize the periodic variability of the sources. Together with the results from the previous step, we further rule out the unlikely variable sources using a random forest classifier and end up with 3,041,677 sources.

The final classification of RRLs: In Section 3.3.3, we build features for the dataset of 3 million sources using the parameters obtained by fitting truncated Fourier Series to each light curve in multiple bands. Then we train another random forest classifier to predict the probability of a source being a RRL and generate a catalogue of 71,755 RRLs.

Since we employ the ZTF light curves for every step, we here lay out the data we use before diving into the detail of the classification process. For each band $k = g, r, i$ in ZTF, n_k is the number of ZTF detection with `catflags` < 32768, which flags bad or generally unusable observation epochs (Masci et al., 2019). For the i-th detection for $i \in \{1, 2, ..., n_k\}$, $t_{k,i}$ is the observed time `mjd_k`, $m_{k,i}$ is the observed magnitude `mag_k`, and $\sigma_{k,i}$ is the uncertainty of the observed magnitude `magerr_k`.

3.3.1 The initial variability selection

We start to process the 600 million sources in the joint set of *Gaia* EDR3 and ZTF DR3 by two selections to make the size of the dataset feasible for variable light curve fittings in the following steps. The first selection is

$$n_k \geq 10 \tag{3.1}$$

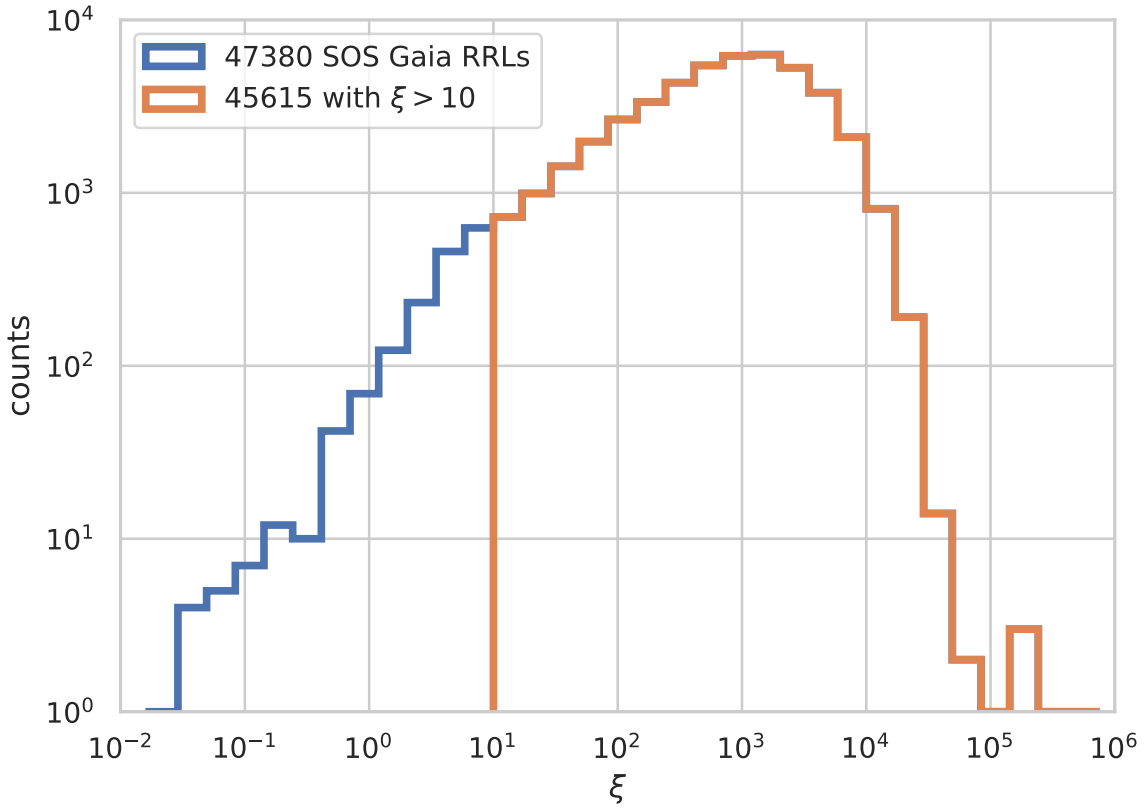

Figure 3.2: The distribution of SOS *Gaia* RRLs in terms of ξ defined in Equation 3.4. The blue and orange histograms are before and after the selection of Equation 3.5 respectively.

for any ZTF band $k = g, r, i$. The reason is to keep the sources with at least 10 light curve data points in any given band such that the single sinusoidal fitting and the truncated Fourier fitting in the following steps are reasonable. After the selection of Equation 3.1, 47,380 out of the total 48,365 SOS *Gaia* RRLs in ZTF DR3 survive.

The second selection is based on the variability inferred by the residuals of constant light curve fits. The constant light curve model for band k is defined as

$$m_{\mathrm{C}_k}(t) = C_k. \tag{3.2}$$

The estimator of the parameter C_k is the mean of the observed light curve data points; $C_k = \frac{1}{n_k}\sum_{i=1}^{n_k} m_{k,i}$. For each light curve, we evaluate the sum of squared residuals as

$$\chi^2_{\mathrm{C}_k} = \sum_{i=1}^{n_k}\left(\frac{m_{k,i} - m_{\mathrm{C}_k}\left(t_{k,i}\right)}{\sigma_{k,i}}\right)^2. \tag{3.3}$$

Using the g and r band statistics, we characterize the significance of variability as a scalar quantity

$$\xi = \frac{\chi^2_{\mathrm{C}_g} + \chi^2_{\mathrm{C}_r} + \nu}{\sqrt{2\nu}} \tag{3.4}$$

where $\nu = n_g + n_r - 2$ is the degrees of freedom, similar to Equation 1 in Hernitschek et al. (2016). We exclude the i band because $\sim 96\%$ of the ZTF sources have < 10 epochs in their i-band light curves. The blue histogram in Figure 3.2 shows the distribution of the 47,380 SOS *Gaia* RRLs in terms of ξ. To keep as many SOS *Gaia* RRLs as possible while shrinking the size of the overall dataset as small as possible, we decide to have the cut of

$$\xi > 10 \tag{3.5}$$

as the second selection. As shown in the orange histogram in Figure 3.2, this selection keeps 45,615 from the 47,380 SOS *Gaia* RRLs.

After the selections of Equation 3.1 and Equation 3.5, 155,095,514 out of the 600 million sources in the joint set of *Gaia* EDR3 and ZTF DR3 survive, entering the next step in the following section. The completeness of the SOS *Gaia* RRLs after the two selections of Equation 3.1 and Equation 3.5 is 0.94.

3.3.2 The broad selection of RRL candidates

Because it is still too computationally expensive to perform higher-order Fourier fitting of all 155 million sources selected in the previous step, we need an extra step to further select a smaller subset of sources. Utilizing two simple and computationally feasible models of the multiple-band ZTF light curves described in Section 3.3.2, we obtain features to characterize the periodicity and variability of the sources and train the random forest classifier I to broadly select the possible RRL candidates in Section 3.3.2.

Table 3.1: The features we use to train the random forest classifier I. The total ZTF epoch $n_{\rm tot} = n_g + n_r + n_i$. $\bar{k}$ and $\tilde{k}$ are the mean and median of the k-band magnitude with $k = g$ and r. $Q_j(k)$ is the jth quartile of the k-band magnitude with $k = g$ and r.

symbol	explanation	range
$\log_{10} n_{\rm tot}$	log of total ZTF epochs	
$(\bar{g} - \bar{r})_0$	$\bar{g} - \bar{r} - E(B-V)$	
$(\tilde{g} - \tilde{r})_0$	$\tilde{g} - \tilde{r} - E(B-V)$	
ρ_{gr}	correlation of g and r light curves	$[-1, 1]$
ρ_{gg}	auto-correlation of g light curves	$[-1, 1]$
ρ_{rr}	auto-correlation of r light curves	$[-1, 1]$
$Q_{12}(g)$	$Q_1(g) - Q_2(g)$	
$Q_{12}(r)$	$Q_1(r) - Q_2(r)$	
$Q_{32}(g)$	$Q_3(g) - Q_2(g)$	
$Q_{32}(r)$	$Q_3(r) - Q_2(r)$	
$\delta\chi^2_{\rm S,C}$	normalized delta chi-square in Equation 3.8	
$P_{\rm sin}$	best fitting period from single sinusoidal fit	$[0.1, 30]$

Constant and single sinusoidal light curve fitting

The first of the two simple and computationally feasible models is the constant light curve fit mentioned in the previous section. The other model is a single discretized sinusoidal light curve formulated as

$$m_{{\rm S}_{k,i}} = A_k \cos^* \left(\frac{2\pi}{P} t_{k,i} + \phi_k \right) + B_k \tag{3.6}$$

where $\cos^*$ is the discretized cosine and the parameters A_k and B_k are the amplitudes, ϕ_k is the phase, and P is the period. For each band k, the sum of squared residuals for the single sinusoidal model is defined as

$$\chi^2_{{\rm S}_k} = \sum_{i=1}^{n_k} \left(\frac{m_{k,i} - m_{{\rm S}_{k,i}}}{\sigma_{k,i}} \right)^2 . \tag{3.7}$$

Fitting the period ranging between 0.1 and 30 days for the light curve in each band with more than 10 ZTF detections with `catflags` < 32768, we have the best fits with the residual sums of squares in the multiple bands for each source for each trial period. Then we pick the fit with the best period $P_{\rm S}$ that minimizes the total residual sum of squares $\chi^2_{\rm S}$ in the multiple bands as the best fit of the single discretized sinusoidal light curve. This fitting process for the 155 million sources took about 300k CPU hours to complete (one month on machines of 420 cores of Intel Haswell E5-2695 v3 CPUs).

From the fits of the two models, we select a set of features summarized in Table 3.1 for the broad selection in the following step. The selected features are the total number of epochs,

the de-reddened colour index $g - r$ based on the mean and the median observed light curves, the difference of the magnitudes between quartiles, the correlations, the best period of the single sinusoidal fit, and the difference of the residual sum of squares $\delta\chi^2_{\rm S,C}$ defined as

$$\delta\chi^2_{\rm S,C} = \frac{\chi^2_{\rm C} - \chi^2_{\rm S}}{\sqrt{2\chi^2_{\rm C}}}, \tag{3.8}$$

where the total residual sums of squares $\chi^2_{\rm C} = \chi^2_{{\rm C}_g} + \chi^2_{{\rm C}_r} + \chi^2_{{\rm C}_i}$ and $\chi^2_{\rm S} = \chi^2_{{\rm S}_g} + \chi^2_{{\rm S}_r} + \chi^2_{{\rm S}_i}$ according to Equation 3.3 and Equation 3.7 respectively, the term of $\sqrt{2\chi^2_{\rm C}}$ is the approximate uncertainty from the variance of the chi-square distribution. Ideally, given a number of epochs, a source with a higher $\delta\chi^2_{\rm S,C}$ is more periodically variable than a source with a lower $\delta\chi^2_{\rm S,C}$.

Random forest classification I

With the features listed in Table 3.1 and the label from the SOS *Gaia* DR2 RRLs, we train a 10-fold cross-validation random forest classifier on the 155 million sources to identify periodic variable sources that are likely to be RRLs by predicting the probability of a source being a possible RRL candidate. Utilizing the random forest classifier in Scikit-Learn (Pedregosa et al., 2011), we employ the default parameters from the module but customize the objective function to be the cross-entropy function and the weights to be adjusted inversely proportional to class frequencies in the input data. For details of random forests and the module, we refer readers to Breiman (2001) and Pedregosa et al. (2011). The 10-fold cross-validation is done by randomly shuffling the 155 million entries and partitioning them into 10 subsets. For each subset, we train a classifier using the other nine subsets as the training set and use the classifier to compute the predicted probability for the subset. Repeating this for all 10 subsets, we accomplish the cross-validation prediction for all the 155 million sources.

Based on the cross-validation prediction, we show the completeness versus the number of selected sources with different probability thresholds between 0 and 0.1 in Figure 3.3. Limited by our computational resources, we can only afford to fit at most roughly 3 million sources with higher-order Fourier Series in the next step, so we decide to use the probability threshold of 0.01 for the selection. With the probability larger than 0.01, there are 3,041,677 selected sources, whose completeness is 0.95 and purity is 0.014. This dataset of 3 million sources then enters the final step of the pipeline described in the following sections.

3.3.3 The final RRL classification step

Using the 3 million sources selected previously as the dataset, we are ready to process the last step in the pipeline to identify RRLs. We first fit each ZTF light curve with the third order of Fourier Series to find the best period and select a set of features that characterizes

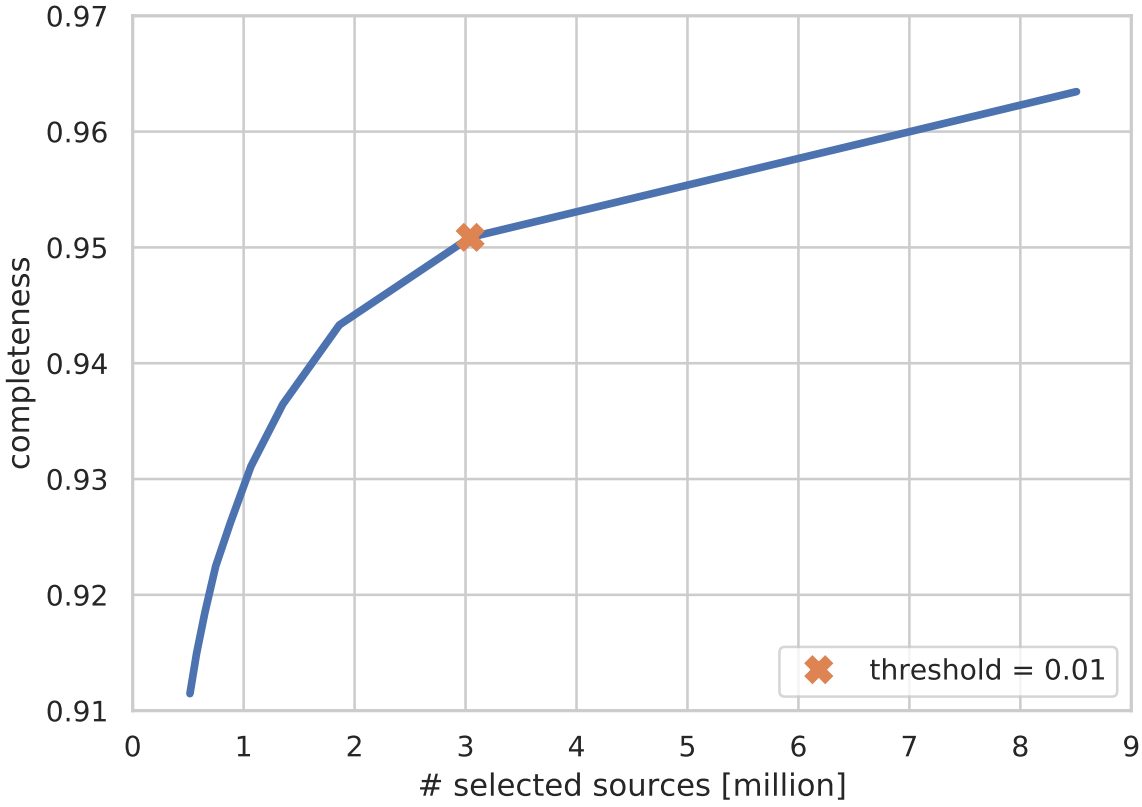

Figure 3.3: The relation between the completeness and the number of selected sources according to different probability thresholds ranging between 0 and 1. The orange mark shows the threshold of 0.01 which is the one we use for the random forest classification I in our pipeline.

Table 3.2: The features of the training set we use for the random forest classifier. Note that k denotes g or r bands.

symbol	explanation	range
$P_{\rm best}$	best fitting period	$[0.1, 1]$
$(g-r)_0$	$A_{g,0} - A_{r,0} - E(B-V)$	
$\ln A_{k,1}$	log of the first Fourier amplitude	
$\ln A_{k,2}$	log of the second Fourier amplitude	
$\ln A_{k,3}$	log of the third Fourier amplitude	
$\phi_{k,21}$	the second relative phase	$[-\pi, \pi]$
$\phi_{k,31}$	the third relative phase	$[-\pi, \pi]$
$\delta\chi^2_{\rm F,C}$	normalized delta chi square in Equation 3.12	

the shape of light curves in Section 3.3.3. With the selected feature set, we train the random forest classifier II to predict the probability of each source being a RRL in Section 3.3.3.

Fourier Series fitting

We model each ZTF light curve in band k using the third order of the Fourier Series as

$$m_{{\rm F}_k}(t) = A_{k,0} + \sum_{j=1}^{3} A_{k,j} \cos(j\omega t + \phi_{k,j}) \tag{3.9}$$

with the parameters of the angular frequency $\omega = \frac{2\pi}{P}$, the period P, the Fourier amplitudes $A_{k,0}, A_{k,j}$ and phases $\phi_{k,j}$ for $j = \{1, 2, 3\}$. We note that for the objects with a large number of light curve points, the accurate description of the light curve might require more high-order Fourier terms than three. To fit a light curve using the model if there is more than 10 detection with `catflags` < 32768 for the light curve, we use a uniform grid in $\frac{1}{P}$ with 10^5 points of the period between 0.1 and 1 days. Given a period, we fit each light curve using the model with the lowest residual sum of squares computed as

$$\chi^2_{{\rm F}_k} = \sum_{i=1}^{n_k} \left(\frac{m_{k,i} - m_{{\rm F}_k}\left(t_{k,i}\right)}{\sigma_{k,i}} \right)^2 . \tag{3.10}$$

For each trial period we perform fits to data in every band and then sum their resulting chi-squares as $\chi^2_{\rm F} = \chi^2_{{\rm F}_g} + \chi^2_{{\rm F}_r} + \chi^2_{{\rm F}_i}$ to be the indicator for determining the best fitting result, that is, the fit with the best period $P_{\rm best}$ that minimizes $\chi^2_{\rm F}$. We note that practically we fit light curves using the model (Eq. 3.9) for each period by doing linear regression with respect to the 1, $[\sin(j\omega t), \cos(j\omega t)]$ for $j = \{1, 2, 3\}$, which can be done with one single matrix operation. This Fourier fitting process for the 3 million sources took about 600k CPU hours to complete (two months on machines of 420 cores of Intel Haswell E5-2695 v3 CPUs).

With the fitted parameters $\left(P_{\rm best},\ A_{k,0},\ A_{k,j},\ \phi_{k,j}\right)$ for $j = \{1, 2, 3\}$, to choose features for the classifier, we aim to use the parameters that characterize the shape of light curves because of the unique shape of RRL light curves. The terms of the zeroth amplitude $A_{k,0}$ and the first phase $\phi_{k,1}$ are essentially the mean magnitude and the phase shift respectively for the light curve so they contribute no meaningful information about the shape of light curves. Thus we exclude them. Because $\phi_{k,1}$ does affect the other phase terms, we rewrite Equation 3.9 in the form of

$$\begin{aligned} m_k(t) = A_{k,0} &+ A_{k,1}\cos(\omega\tau_k) \\ &+ A_{k,2}\cos(2\omega\tau_k + \phi_{k,21}) + A_{k,3}\cos(3\omega\tau_k + \phi_{k,31}) \end{aligned} \tag{3.11}$$

to take care of the time shift caused by $\phi_{k,1}$, where $\tau_k = t + \frac{\phi_{k,1}}{\omega}$, $\phi_{k,21} = \phi_{k,2} - 2\phi_{k,1}$, and $\phi_{k,31} = \phi_{k,3} - 3\phi_{k,1}$. Unlike $\phi_{k,1}$, these relative phases $\phi_{k,21}$ and $\phi_{k,31}$ do characterize the shape of light curves so we include them in the feature set. It is worth noting that there is a correlation between metallicity and $\phi_{k,31}$ (Simon & Clement, 1993; Jurcsik & Kovacs, 1996; Sandage, 2004; Sesar et al., 2010).

Besides the shape of light curves, the difference in the goodness of the Fourier fit and that of the constant light curve fit is essential to the classification because it indicates the goodness of the two competing models. Similar to Equation 3.8 in Section 3.3.2, we define the normalized delta chi-square as

$$\delta\chi^2_{\rm F,C} = \frac{\chi^2_{\rm C} - \chi^2_{\rm F}}{\sqrt{2\chi^2_{\rm C}}} \tag{3.12}$$

and include it in the feature set, where $\chi^2_{\rm F}$ and $\chi^2_{\rm C}$ are the residual sums of squares of the best Fourier fit and that of the constant fit. For example, for the light curve of a true RRL, the Fourier light curve tends to fit it better than the constant light curve does, resulting in low $\chi^2_{\rm F}$, high $\chi^2_{\rm C}$, and thus a large value of $\delta\chi^2_{\rm F,C}$.

To sum up, the features that we decide to use for the final classifier are the best fitting period $P_{\rm best}$, the de-reddened colour index $g - r$, the amplitudes $A_{k,j}$ for $j = \{1, 2, 3\}$ and the relative phases $\phi_{k,21}$ and $\phi_{k,31}$ in the g and r bands, and $\delta\chi^2_{\rm F,C}$, summarized in Table 3.2. With these features and the label from the SOS *Gaia* DR2 RRL catalogue, we have prepared all the ingredients for the final classification of the RRL stars among the dataset of 3 million sources.

Random forest classifier II

To carry out the last step of the binary classification task, we again utilize the random forest classifier in Scikit-Learn (Pedregosa et al., 2011) and describe the detail of the process below. First, we partition the dataset of 3 million sources into two subsets, the high-quality set and the low-quality set, due to the incompleteness of the SOS *Gaia* DR2 RRLs in the low

galactic latitude areas and the low *Gaia* epoch areas. Based on HEALPix (Górski et al., 2005) pixels with `nside` = 128, if a source is at the pixel with the galactic latitude at $|b| > 10°$ and at the pixel with a number of *Gaia* epochs larger than the global mean of 250, we assign the source to the high-quality set, otherwise, it goes into the low-quality set. The reasoning for this partition is to only train models in the following steps using the high-quality set because the incompleteness of the SOS *Gaia* DR2 RRLs on the HEALPix pixels that do not satisfy the above criteria is expected to cause some miss-labelled samples in the low-quality set.

For the high-quality set of 1,273,760 sources, we randomly shuffle the rows and partition the set into 10 subsets such that we can perform a 10-fold cross-validation prediction by training 10 classifiers. That is, we train a random forest classifier using all the sources that are not in the k^{th} subset as the training data to predict the probability of each source in the k^{th} subset for $k = \{1, 2, ..., 10\}$ to be a RRL. For the low-quality set of 1,767,917 sources, we use the entire high-quality set as the training data to train a random forest classifier and predict the probability of each source in the low-quality set to be a RRL. For each random forest classifier, the classifier parameters are the same as the ones used for the random forest classifier I in Section 3.3.2. In the end, we concatenate both sets back together to a single set and thus have the predicted probability for each of the 3 million sources being a RRL from the result of the final random forest classification.

Determination of the probability threshold

Using the predicted probability for each source being a RRL in the dataset of 3 million sources from Section 3.3.3, we investigate the completeness and purity for different probability thresholds to determine the threshold for our RRL catalogue. Given a probability threshold, to compute the completeness and purity of the predicted RRLs, we compare our predicted RRLs to the SOS *Gaia* DR2 RRL samples in the high galactic latitude areas with $|b| > 10°$ and in the high *Gaia* epoch areas with the number of *Gaia* epochs > 250 as the high-quality set explained in Section 3.3.3. The reason for applying these two conditions to the calculation of completeness and purity is that the SOS *Gaia* DR2 RRL samples in these areas are supposed to be more complete compared to the other areas. We show the completeness and purity of the predicted RRLs for different probability thresholds in Figure 3.4, choosing the probability threshold of 0.15 which maximizes the F_1 score[1] defined as

$$F_1 = 2 \cdot \frac{\text{completeness} \cdot \text{purity}}{\text{completeness} + \text{purity}} \tag{3.13}$$

as the orange cross mark shows. The probability threshold of 0.15 results in a RRL catalogue of 71,755 predicted RRLs with 0.92 purity and 0.92 completeness, which contains 39,502 out of the original labels of 48,365 SOS *Gaia* DR2 RRLs.

[1]We note that completeness and purity are the synonyms of recall and precision respectively.

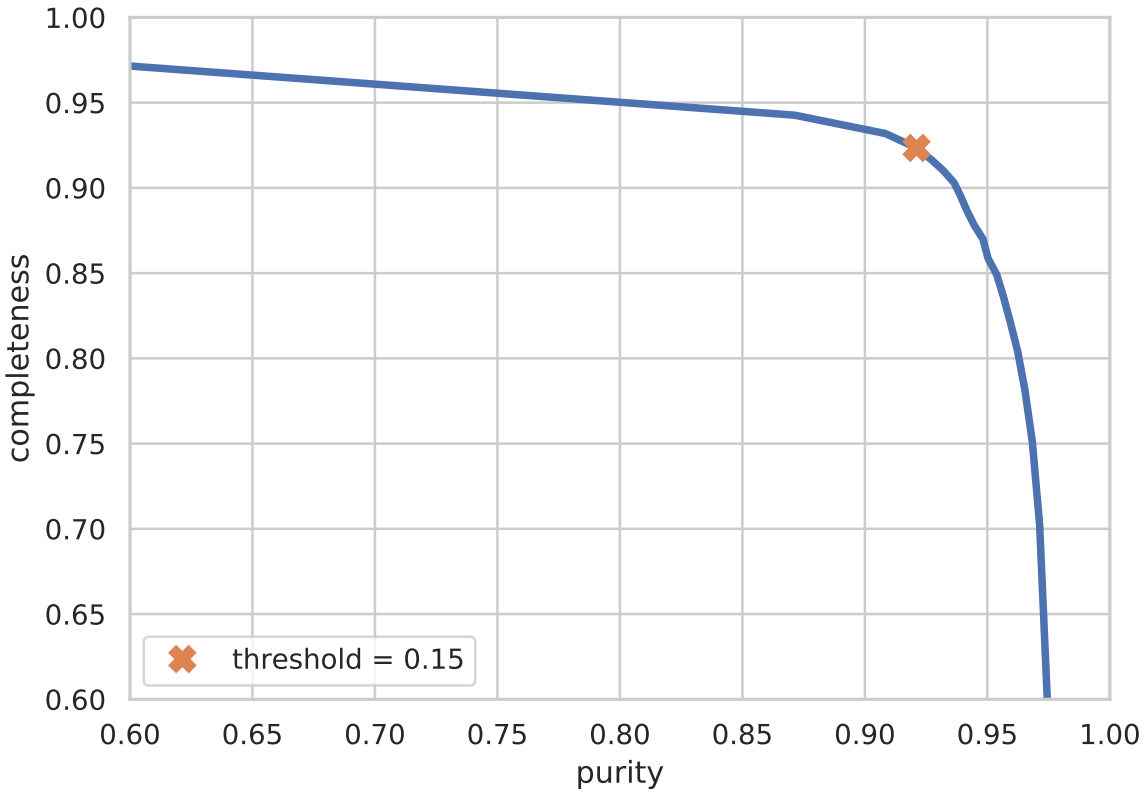

Figure 3.4: Completeness versus purity of the predicted RRL catalogue with probability thresholds between 0 and 1. The orange mark shows the probability threshold of 0.15.

Table 3.3: The description of our catalogue of 71,755 RRLs. Note that $k = g, r, i$ band in ZTF in the description.

column	description
`objid`	ZTF DR3 `objid`
`source_id`	*Gaia* EDR3 `source_id`
`ra`	right ascension [deg]
`dec`	declination [deg]
`prob_rrl`	predicted probability for being a RRL
`best_period`	best fitting period [day]
`amp_1_`k	$A_{k,1}$, first Fourier amplitudes [mag]
`amp_2_`k	$A_{k,2}$, second Fourier amplitudes [mag]
`amp_3_`k	$A_{k,3}$, third Fourier amplitudes [mag]
`phi_1_`k	$A_{k,1}$, first Fourier phases [rad]
`phi_2_`k	$A_{k,1}$, second Fourier phases [rad]
`phi_3_`k	$A_{k,1}$, third Fourier phases [rad]
`mean_`k	$A_{k,0}$, mean k-band magnitude [mag]
`ngooddet_`k	number of ZTF epochs
`phot_g_mean_mag`	*Gaia* EDR3 mean G magnitude [mag]
`ebv`	$E(B-V)$ [mag]
`distance`	heliocentric distance [pc]

3.4 The RRL catalogue

3.4.1 Overview of the catalogue

In this section, we give an overview of the RRL catalogue produced by the pipeline described in Section 3.3. Covering the Northern sky, this catalogue containing 71,755 RRLs in the joint set of *Gaia* EDR3 and ZTF DR3 will be the main RRL catalogue of the paper. A detailed description of the catalogue contents is provided in Table 3.3. The catalogue is released in electronic form with the paper at DOI 10.5281/zenodo.5774017 (Huang & Koposov, 2021a) with a short snippet of the table provided in Table 3.4.

To evaluate the heliocentric distances in the catalogue, we first derive the absolute magnitudes of the RRLs according to the PS1 period-luminosity relations in Sesar et al. (2017)

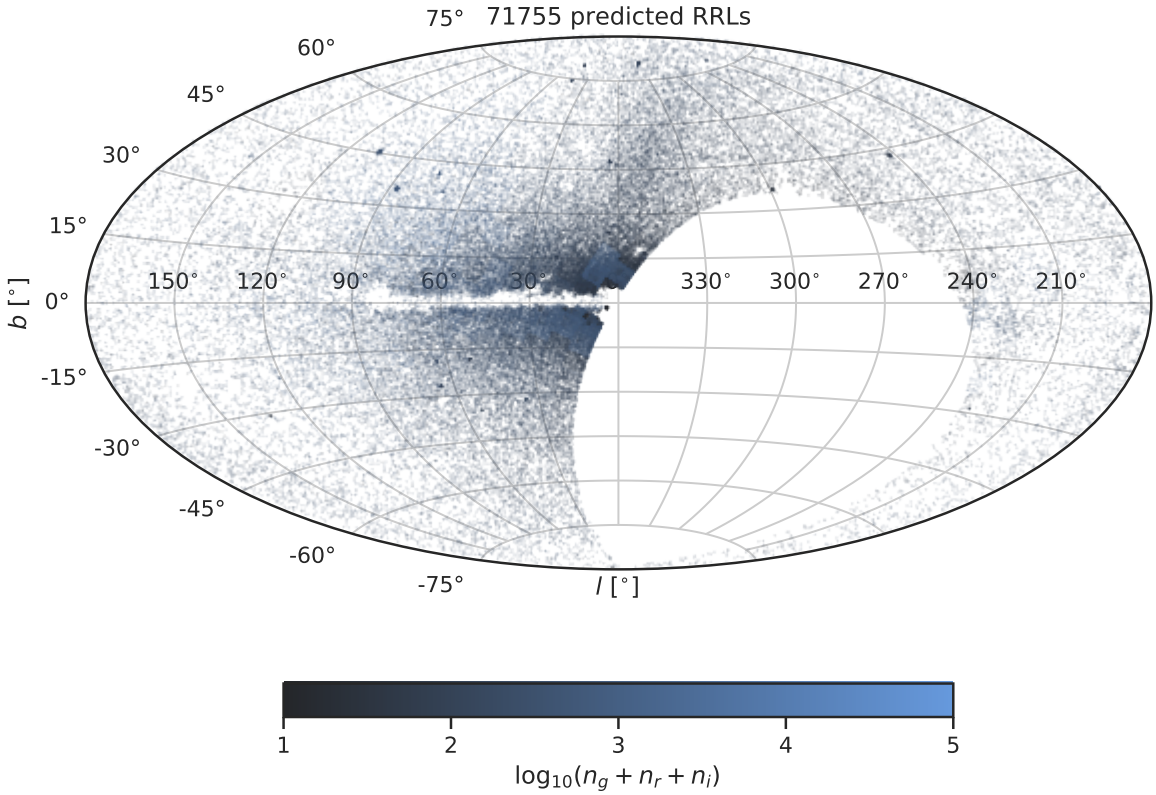

Figure 3.5: The distribution of our 71,755 RRLs in the Galactic coordinates, color-coded by the total number of ZTf observation epochs in the *gri* bands. There are some visible stripes associated with the ZTF fields along declination.

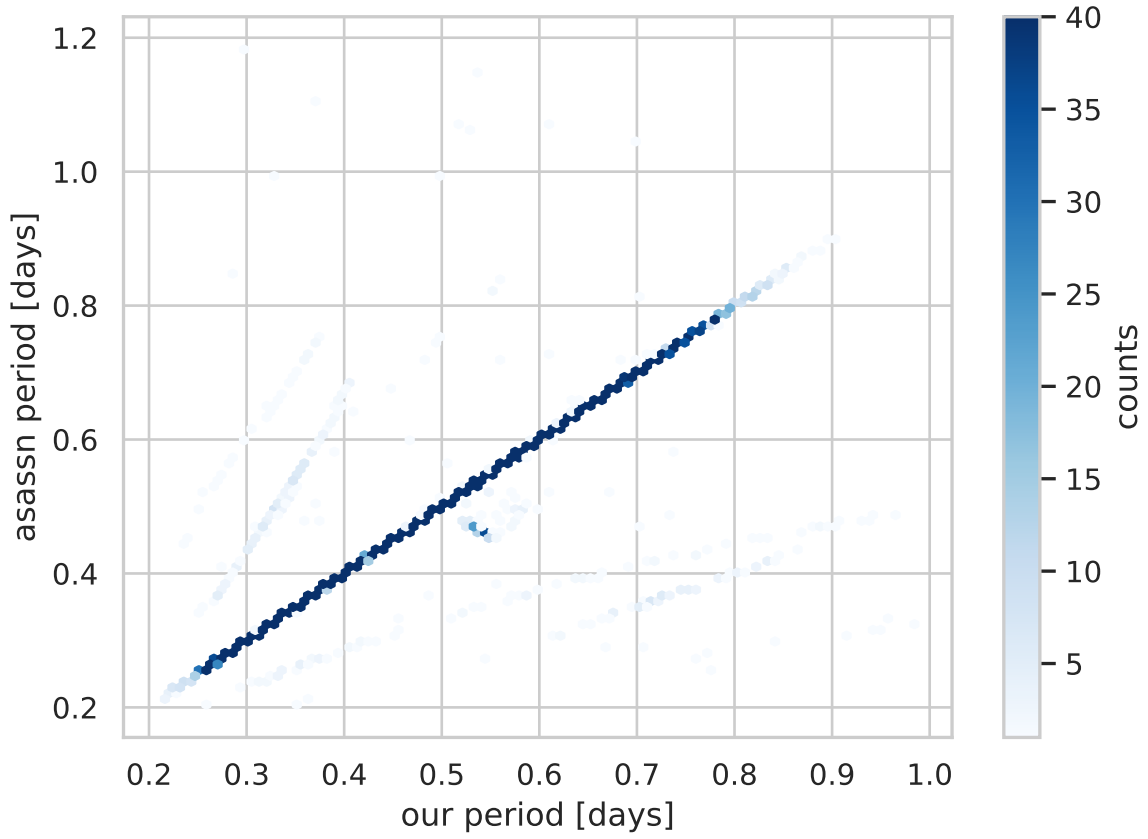

Figure 3.6: Our best fitting period P_{best} versus the period provided by the ASAS-SN catalogue (Jayasinghe et al., 2020) for the common 18,854 RRLs in both datasets.

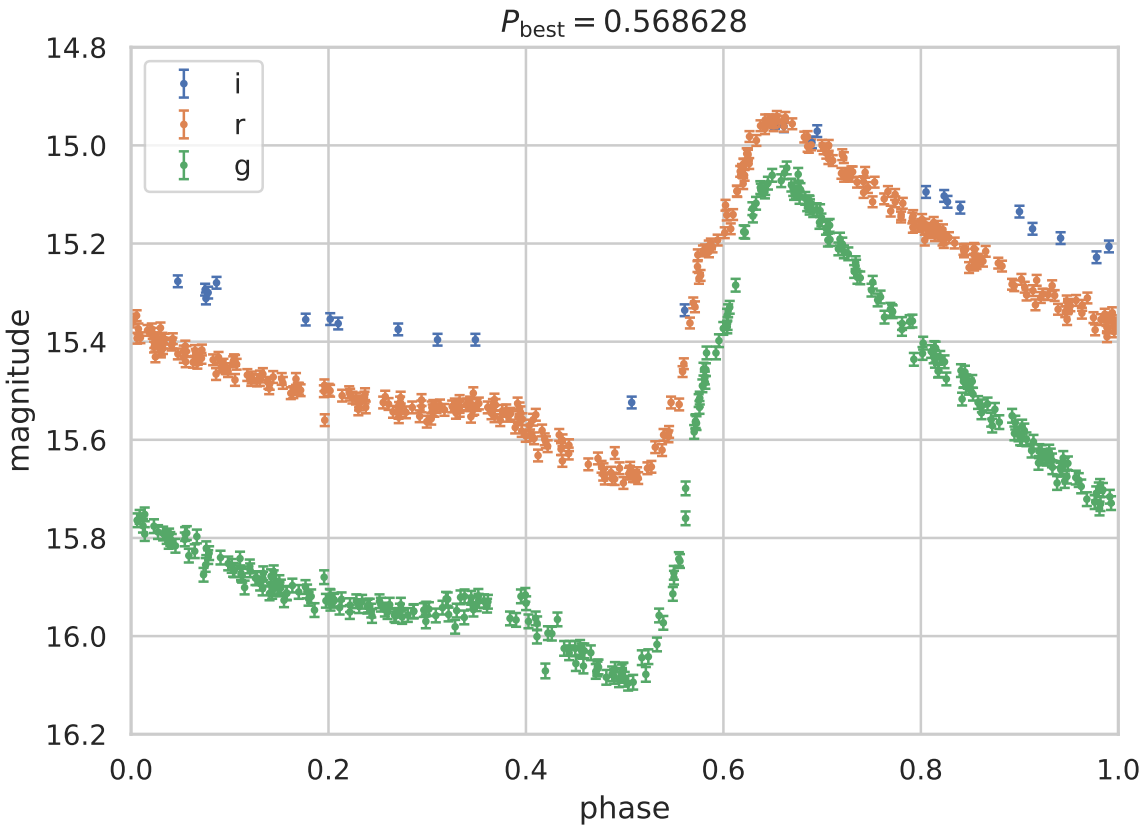

Figure 3.7: An example of RRL ZTF light curves folded by its best period P_{best}, whose *Gaia* EDR3 `source_id` = 2294134898301488640.

Table 3.4: A snippet of the machine-readable table for the RRL catalogue (split into three parts below due to space limitation). The detailed description of the columns is in Table 3.3.

objid	source_id	ra [deg]	dec [deg]	prob_rrl	best_period [day]	ebv [mag]	distance [pc]
245101100001850	2323207596351730304	4.34881	-26.732536	0.95	0.621282	0.017800	35202.500000
245101200001823	2323151181956812672	3.44762	-26.736970	0.89	0.363568	0.022338	20908.599609

phot_g_mean_mag [mag]	ngooddet_r	ngooddet_g	ngooddet_i	mean_r [mag]	mean_g [mag]	mean_i [mag]
18.278099	76	74	0	18.263773	18.448895	
17.573000	81	80	0	17.580330	17.662470	

amp_1_r [mag]	amp_1_g [mag]	amp_1_i [mag]	amp_2_r [mag]	amp_2_g [mag]	amp_2_i [mag]	amp_3_r [mag]	amp_3_g [mag]	amp_3_i [mag]
0.310712	0.429619		0.136991	0.186742		0.077740	0.111422	
0.214430	0.310684		0.025369	0.049442		0.029519	0.016883	

phi_1_r [rad]	phi_1_g [rad]	phi_1_i [rad]	phi_2_r [rad]	phi_2_g [rad]	phi_2_i [rad]	phi_3_r [rad]	phi_3_g [rad]	phi_3_i [rad]
-0.420907	-0.554002		1.119307	1.172803		2.935552	2.603256	
2.610607	2.676576		0.784964	0.691132		-1.650670	-1.853666	

assuming a halo metallicity of $[\mathrm{Fe/H}] = -1.5$ (Ivezić et al., 2008)

$$\begin{aligned} M_g &= -1.7 \log_{10}\left(\frac{P_{\mathrm{best}}}{0.6}\right) + 0.69 \\ M_r &= -1.6 \log_{10}\left(\frac{P_{\mathrm{best}}}{0.6}\right) + 0.51 \\ M_i &= -1.77 \log_{10}\left(\frac{P_{\mathrm{best}}}{0.6}\right) + 0.46. \end{aligned} \tag{3.14}$$

Together with the mean ZTF magnitudes as the zeroth-order fitted Fourier amplitude $A_{k,0}$ for $k = g, r, i$ corrected by the extinction in Schlafly & Finkbeiner (2011), we evaluate the distance moduli μ_k as

$$\begin{aligned} \mu_g &= A_{g,0} - 3.17E(B-V) - M_g \\ \mu_r &= A_{r,0} - 2.27E(B-V) - M_r \\ \mu_i &= A_{i,0} - 1.68E(B-V) - M_i \end{aligned} \tag{3.15}$$

and then derive the heliocentric distance by averaging the distance moduli.

As a first look at the catalogue, we show the sky distribution of the 71,755 predicted RRLs in the Galactic coordinates in Figure 3.5, observing the Galactic halo and the Sagittarius Stream despite the lack of coverage of the Southern sky. Compared to the SOS *Gaia* DR2 RRLs which serves as the label in our classification pipeline, there are several facts about

our RRL catalogue which are worth noting. Our RRL catalogue contains more sources than the 48,365 SOS *Gaia* RRL samples in the Northern sky coverage with the completeness of 0.92 and purity of 0.92 globally. Colour-coded by the total ZTF observation epochs, Figure 3.5 shows that our RRL catalogue is more complete in the areas where *Gaia* suffers incompleteness due to its scanning trajectory as the patches with fewer RRLs shown in Figure 3.1, and in the low galactic latitude areas, e.g. $3^\circ < |b| < 10^\circ$.

To show the robustness of our fitting period, we compare our best-fitting periods to the periods provided in the ASAS-SN catalogue (Jayasinghe et al., 2020), for the 18,854 RRLs that are in both catalogues by matching the *Gaia* EDR3 `source_id` provided in both catalogues. The reason to choose the ASAS-SN catalogue to compare with is due to its high number of epochs (each ASAS-SN field in the V-band has roughly 100 – 600 epochs Jayasinghe et al., 2018) and thus its reliable period determination. Figure 3.6 shows the alignment on the one-to-one line on the period plane and indicates the goodness of our period fitting result. We note that 97% of the 18,854 RRLs have a period percentage difference smaller than 0.1%, though several objects suffer the aliasing period issue during the Fourier fitting process (Lomb, 1976; Scargle, 1982; VanderPlas, 2018). Moreover in Figure 3.7, we display an example of ZTF light curves from our RRL catalogue in the *gri* bands, folded by its best-fitting period $P_{\rm best}$. Demonstrating a typical shape of a folded RRL light curve, this furthermore shows the robustness of our Fourier Series fitting described in Section 3.3.3 and the resulting period in the catalogue.

To further investigate the RRL catalogue, we will look into the completeness of the catalogue in Section 3.4.2, compare the catalogue with other existing catalogues in Section 3.4.3, and study the Galactic halo profile in Section 3.4.4

3.4.2 Completeness of the catalogue

As mentioned in Section 3.3.3, our RRL catalogue has overall completeness of 0.92 compared to the SOS *Gaia* DR2 RRLs grouped by the HEALPix pixels with `nside` = 128 with the number of *Gaia* epochs > 250 globally. In this section, we will look into the completeness in more detail and we begin by investigating the completeness as a function of heliocentric distance in Figure 3.8. Using the SOS *Gaia* DR2 RRLs grouped by the same HEALPix pixels to compute the completeness in heliocentric distance bins, we find that the completeness is higher than ~ 0.8 at the regions with distance smaller than 80 kpc, is roughly 0.5 at 100 kpc, and drops drastically to 0 at 130 kpc. We note that the most distant RRL in our catalogue is at a distance of 132 kpc. Thanks to the deeper RRL catalogue from DES Y6 with the most distant RRL at $\sim$ 300 kpc (Stringer et al., 2021), we cross-match the closest RRL within one arcsec at the areas above -20° declination and evaluate the completeness, finding that the completeness is consistent with the one compared to the SOS *Gaia* RRLs for distance smaller than 80 kpc. However, at distance larger than 100 kpc, the completeness drastically drops to 0.2 and then 0.

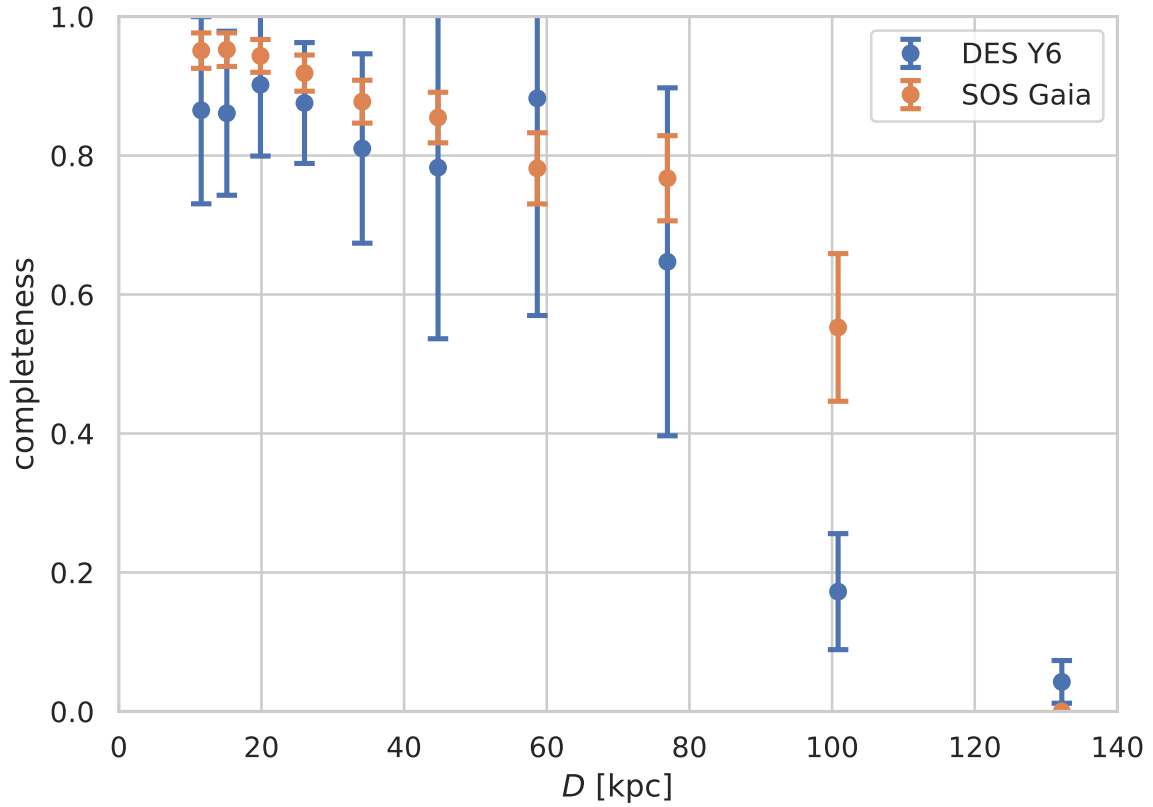

Figure 3.8: The completeness of our RRL catalogue as a function of heliocentric distance compared to the SOS *Gaia* DR2 RRL catalogue and the DES Y6 RRL catalogue.

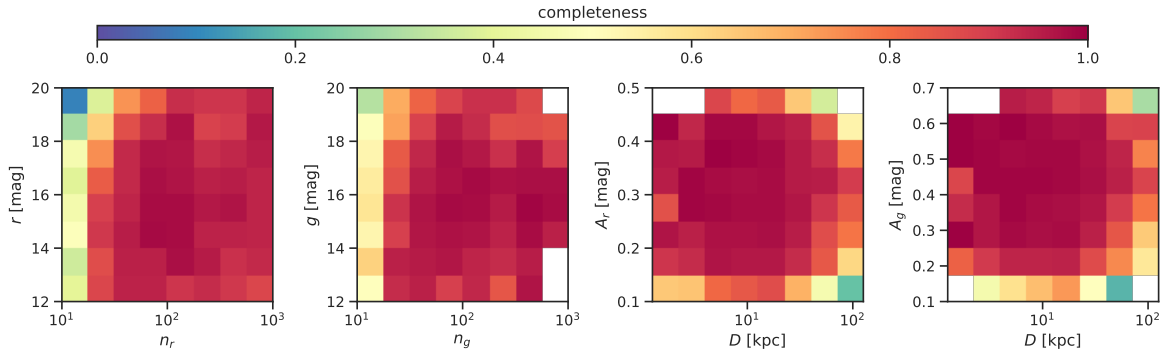

Figure 3.9: **Left** and **middle-left**: the completeness as functions of the mean magnitudes r and g and the ZTF numbers of epochs in r and g bands n_r and n_g. **Middle-right** and **right**: the completeness as functions of the amplitudes A_r and A_g and the heliocentric distance D.

Table 3.5: The completeness of our catalogue compared to some external RRL catalogues. For each catalogue, we apply the selections of declination $> -20°$, $|b| > 10°$, and magnitude between 15 and 20. After the selections, N is the number of RRLs in the external catalogues, and N_x is the number of RRLs from each external catalogue that have a match in our catalogue.

catalogue	N	N_x	N_x/N	reference(s)
ZTF DR2	28883	27993	0.96	Chen et al. (2020)
DES Y6	769	665	0.86	Stringer et al. (2021)
ASAS-SN	12765	10704	0.83	Jayasinghe et al. (2018)
PS1	32045	25862	0.80	Sesar et al. (2017)
SOS	30002	24090	0.80	Clementini et al. (2019b)
OGLE	701	567	0.80	Soszyński et al. (2019)
CRTS	6917	5473	0.79	Drake et al. (2014)

We move on to investigate the influence of several quantities on the completeness of our RRL catalogue, including the distance, the amplitudes, the magnitudes, and the numbers of epochs, again utilizing the SOS *Gaia* DR2 RRLs on the HEALPix pixels with `nside` = 128 with the number of *Gaia* epochs > 250. In the left and the middle-left panels of Figure 3.9, we show the completeness as a function of r and n_r and that of g and n_g respectively, where r and g are the mean magnitudes corrected by the extinction and n_r and n_g are the numbers of ZTF detection with `catflags` < 32768 in r and g bands. We find that the completeness is lower when there is less detection for a source given a magnitude and when the luminosity is fainter given a number of detection. The middle-right and the right panels show the completeness as a function of the r-band amplitude A_r and the heliocentric distance D and that of g-band amplitude A_g and D respectively. The amplitudes A_r and A_g defined from the combination of multiple Fourier terms

$$\begin{aligned} A_r &= \sqrt{A_{r,1}^2 + A_{r,2}^2 + A_{r,3}^2} \\ A_g &= \sqrt{A_{g,1}^2 + A_{g,2}^2 + A_{g,3}^2} \end{aligned} \tag{3.16}$$

are the best fitting amplitudes from the third-order Fourier Series in the g and r bands. We find that the completeness gradually decreases as the distance increases given an amplitude, meaning that our catalogue is less complete at more distant regions, which is consistent with Figure 3.8. When given a distance, the completeness drops faster at the small-amplitude ends than at the large-amplitude end.

3.4.3 Comparison with other catalogues

We start this section by comparing our RRL catalogue to several recent RRL catalogues covering the entire Northern sky. Figure 3.10 shows the RRL distributions of different

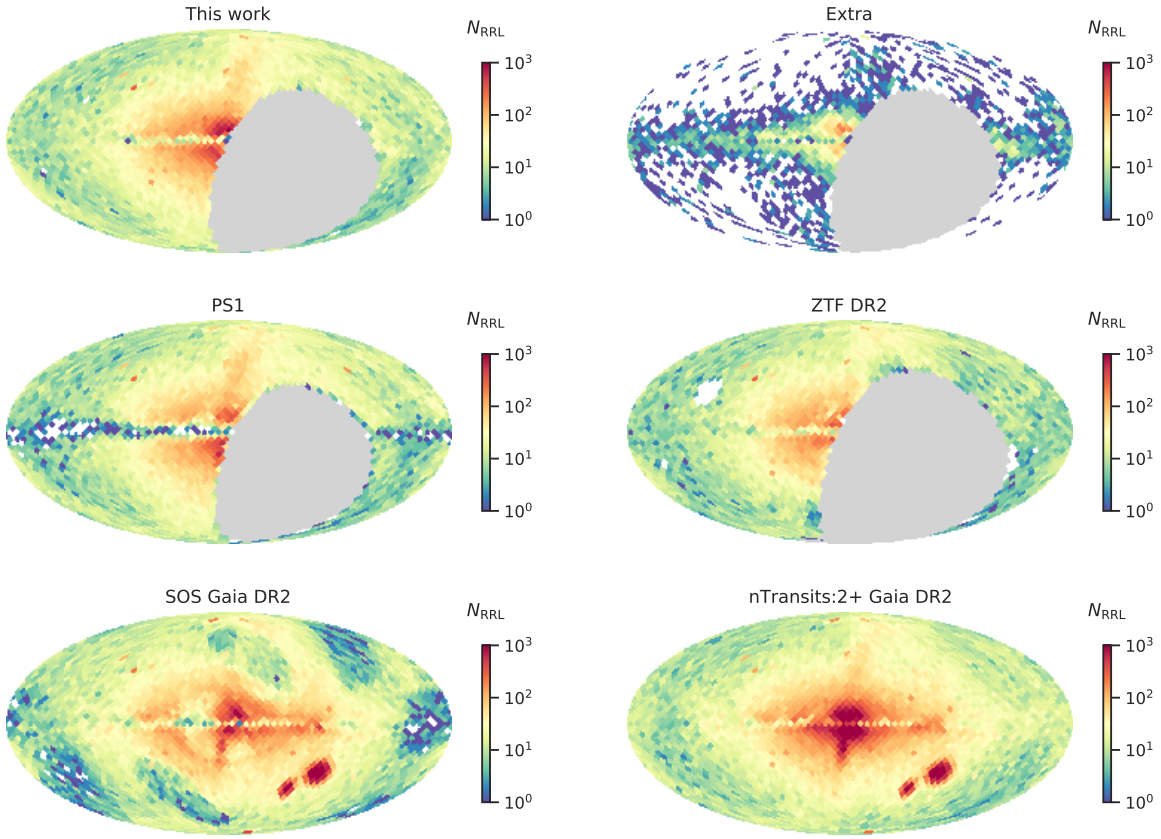

Figure 3.10: The RRL distributions in the Galactic coordinate of the catalogue from this work, the PS1 catalogue (Sesar et al., 2017), the ZTF DR2 catalogue (Chen et al., 2020), the SOS *Gaia* DR2 catalogue (Clementini et al., 2019a), and the nTransits:2+ *Gaia* DR2 catalogue (Holl et al., 2018), colour-coded by the number of RRLs on each grid $N_{\rm RRL}$. The top-right panel illustrates the extra RRLs from this work that are not in any external catalogues mentioned in Section 3.4.3.

catalogues in the Galactic coordinate, colour-coded by the number of RRLs $N_{\rm RRL}$ on each HEALPix pixel of `nside` = 16. The catalogues plotted are the RRL catalogue from this work, the PS1 catalogue (Sesar et al., 2017), the ZTF DR2 catalogue (Chen et al., 2020), the SOS *Gaia* DR2 catalogue (Clementini et al., 2019a), and the nTransits:2+ *Gaia* DR2 catalogue (Holl et al., 2018). In particular, we apply the score thresholds of 0.8 and 0.55 for types ab and c RRLs according to Sesar et al. (2017) when utilizing the PS1 catalogue.

Overall, our catalogue, the PS1 catalogue, and the nTransits:2+ *Gaia* DR2 catalogue illustrate the Galactic halo and the Sagittarius Stream better than the ZTF DR2 catalogue and the SOS *Gaia* DR2 catalogue do. Even though the nTransits:2+ *Gaia* DR2 catalogue covers the whole sky, it is generally more contaminated than the SOS RRL catalogue (see Holl et al., 2018; Clementini et al., 2019a, for more detail) and it does not provide periods and light curve fits for the RRL samples. Compared to the SOS *Gaia* DR2 catalogue which serves as our training label, our RRL catalogue outperforms at the incomplete areas caused by the *Gaia* scanning trajectory and has more RRLs in the Northern sky coverage. Compared to the PS1 catalogue of 61,144 RRLs, our catalogue has more RRL samples, especially around the Galactic halo, and covers the low galactic latitude areas better. Compared to the ZTF DR2 catalogue of 46,358 RRLs, our catalogue has more RRLs globally, especially near the Galactic halo and the Sagittarius Stream, and tends to have more numbers of observed epochs due to the usage of ZTF DR3.

Besides the above five catalogues covering the entire Northern sky, we also compare our catalogue to other existing RRL catalogues, including the DES Y6 catalogue (Stringer et al., 2021), the CRTS catalogue (Drake et al., 2014), the ASAS-SN catalogue (Jayasinghe et al., 2018), the OGLE catalogue (Soszyński et al., 2019), and the NSVS catalogue (Wils et al., 2006). For the comparison, we apply three selections on every catalogue, the selection of declination $> -20^\circ$ due to the sky coverage of the ZTF survey, the selection of $|b| > 10^\circ$ to exclude the region near the Galactic disc, and the selection of magnitude between 15 and 20 based on our RRL magnitude distribution because the depth of the catalogues varies. After the selections, we count the number of RRLs in each catalogue N, amongst them we count the number of RRLs from each catalogue that have a match in our table as $N_{\rm x}$, and from that we calculate the overall completeness of our catalogue as $N_{\rm x}/N$. The cross-matching is done by selecting the closest objects based on the angular separation within 1 arcsec for most of the catalogues, except for the CRTS and ASAS-SN catalogues. When cross-matching the CRTS catalogue to our catalogue, we use the angular separation of 2.5 arcsec as it is the pixel size for CRTS (Drake et al., 2009). For the ASAS-SN catalogue, it has already provided the *Gaia* EDR3 `source_id`, which our catalogue also provides, so we directly utilize the `source_id` to cross-match the two catalogues.

The results of the comparison are summarized in Table 3.5. We note that there are only 8 stars left in the NSVS catalogue after the selections, so we exclude NSVS from the table. Our catalogue achieves high completeness of 96% compared to the ZTF DR2 catalogue, which is

expected to be the highest as these two catalogues are based on the same survey but different data releases. For all the other catalogues, DES Y6, ASAS-SN, PS1, *Gaia* SOS, OGLE, and CRTS, our catalogue has the completeness $\gtrsim 80\%$. We note that our catalogue is possibly less complete for distant RRLs, for small-amplitude RRLs, for type c RRLs, or for the RRLs located on the field boundary regions.

We end the section by identifying the extra RRLs from our catalogue when cross-matched with all the external RRL catalogues mentioned in the section. In total, we have 6547 extra RRLs, and we visualize them in the top-right panel in Figure 3.10. This panel indicates the extra RRLs in our catalogue concentrate around the Galactic halo and near the Galactic disk. When making this panel, we mask out 844 RRLs with the period within 0.5 ± 0.01 days because they are most likely contaminated objects due to the aliasing period issue.

3.4.4 The Galactic halo profile

Knowing that our catalogue contains more RRLs around the Galactic halo and near the low Galactic latitude areas compared to the other catalogues in Section 3.4.3, we study the Galactic halo profile using our RRL catalogue in the Galactocentric coordinate in this section. Focusing on the Galactic halo profile, we mask out the RRLs in the Milky Way dwarf galaxies and globular clusters with declination above -28° due to the coverage of ZTF and with heliocentric distance smaller than 100 kpc due to the completeness of our catalogue. This criterion includes 90 globular clusters from Harris (1996, 2010) and 17 dwarf galaxies of Bootes I and II, Cetus II, Coma Berenices, Draco, Draco II, Sagittarius II, Segue I and II, Sextans I, Triangulum II, Ursa Major I and II, Ursa Minor, and Willman I from McConnachie (2012), and Bootes III (Massari & Helmi, 2018) and Virgo I (Homma et al., 2016). After the selection, there are 70950 RRLs for the study of the Galactic halo profile in this section.

We briefly lay out the Galactocentric coordinate adopted in this section. The right-handed Cartesian coordinate (X, Y, Z) is computed by the Galactic longitude, Galactic latitude, and heliocentric distance (l, b, D) as

$$\begin{aligned} X &= D \cos l \cos b - R_\odot \\ Y &= D \sin l \cos b \\ Z &= D \sin b \end{aligned} \tag{3.17}$$

where $R_\odot = 8$ kpc is the distance between the Galactic Centre and the Sun. This coordinate is centred at the Galactic Centre with the Galactic disk on the (X, Y) plane, the Z-axis pointing to the north Galactic pole, and the X-axis pointing from the Sun at $X = -8$ kpc to the Galactic Centre at $X = 0$ kpc. We define the cylindrical radius R_{XY} and the elliptical radius r_e as

$$\begin{aligned} R_{XY} &= \sqrt{X^2 + Y^2} \\ r_\mathrm{e} &= \sqrt{X^2 + Y^2 + \left(\frac{Z}{q}\right)^2} \end{aligned} \tag{3.18}$$

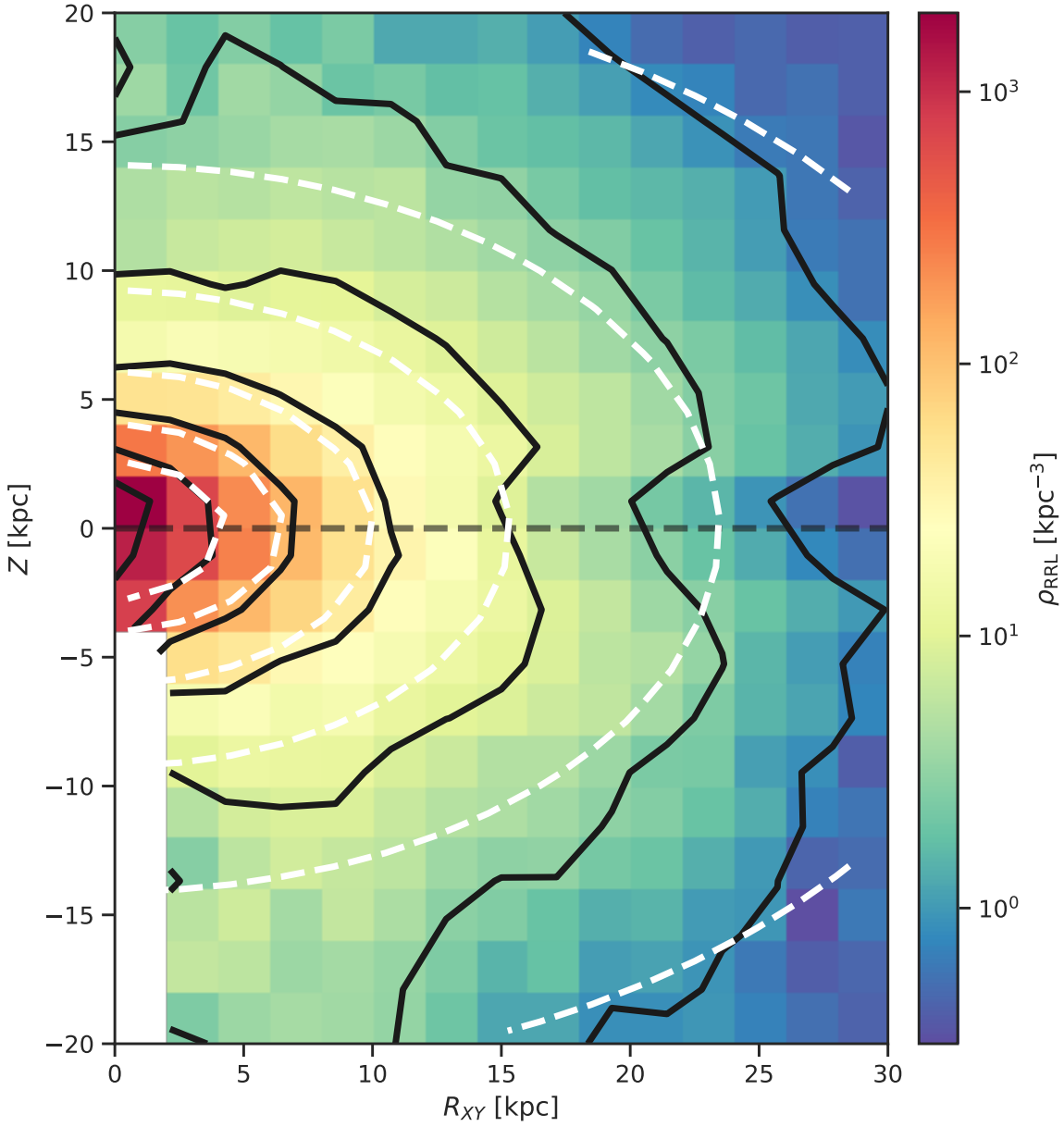

Figure 3.11: The 2D histogram of the RRL distribution in the cylindrical galactocentric coordinates ($R_{XY} - Z$) colour-coded by the RRL number density $\rho_{\rm RRL}$ on each grid. The black curves are the contours of $\rho_{\rm RRL} = 10^0, 10^{0.5}, 10^1, 10^{1.5}, 10^2, 10^{2.5}, 10^3$ kpc^{-3}. The white elliptical contours are the single power law density profile with $q = 0.6$ and power of -2.7 from Iorio et al. (2018).

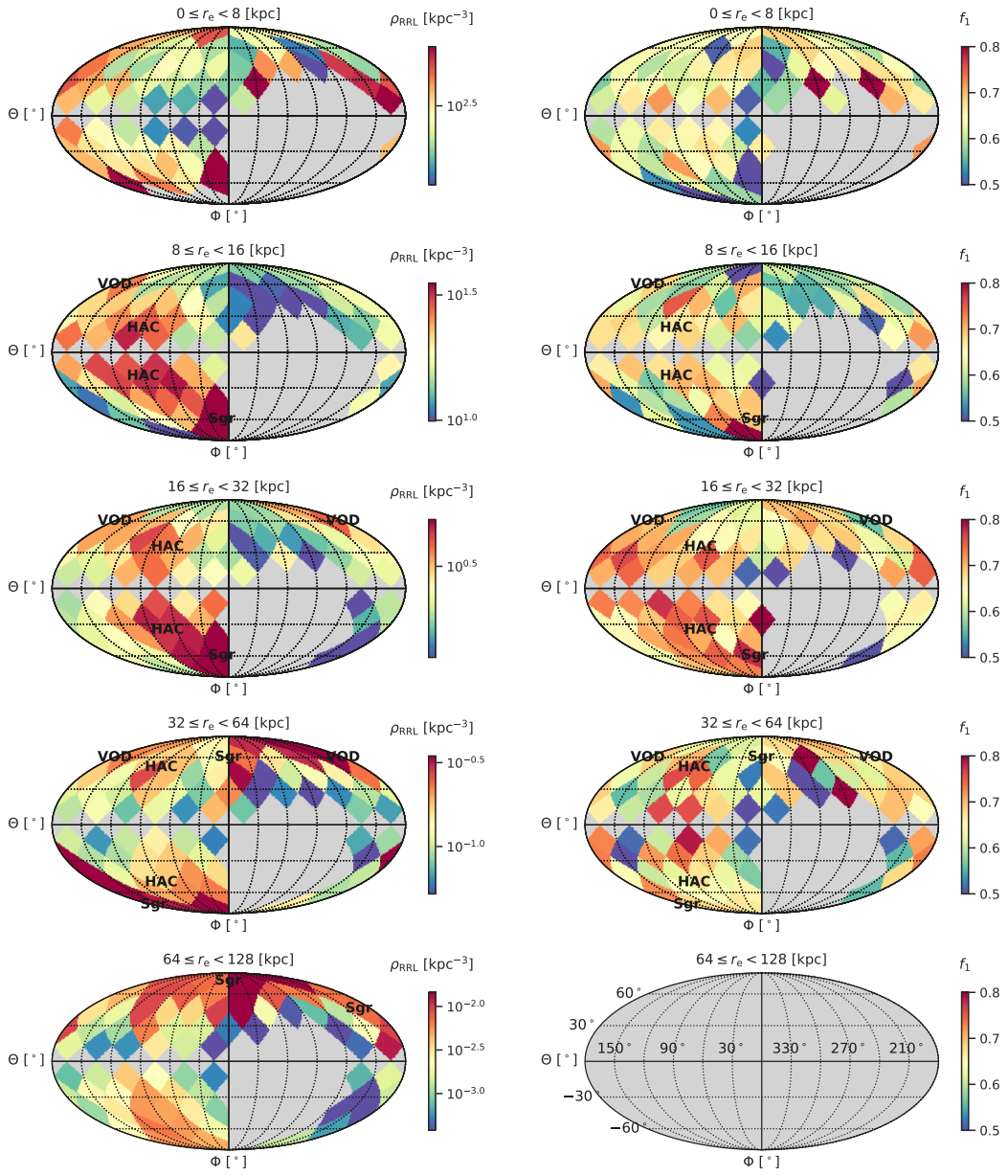

Figure 3.12: **Left column**: The RRL number density $\rho_{\rm RRL}$ on spheroidal shells of different elliptical radii $r_{\rm e}$ in the coordinate of the Galactocentric longitude Φ and latitude Θ. **Right column**: The Oosterhoff type I fraction f_1 on each spheroidal shell in Φ and Θ. For the grids on each panel, the edges from left to right are $\Phi = 180°, 150°, 120°, 90°, 60°, 30°, 0°, 330°, 300°, 270°, 240°, 210°, 180°$ and from top to bottom are $\Theta = 90°, 60°, 30°, 0°, -30°, -60°, -90°$. The annotations HAC, VOD, and Sgr are the Hercules-Aquila Cloud, the Virgo over-density, and the Sagittarius Stream respectively.

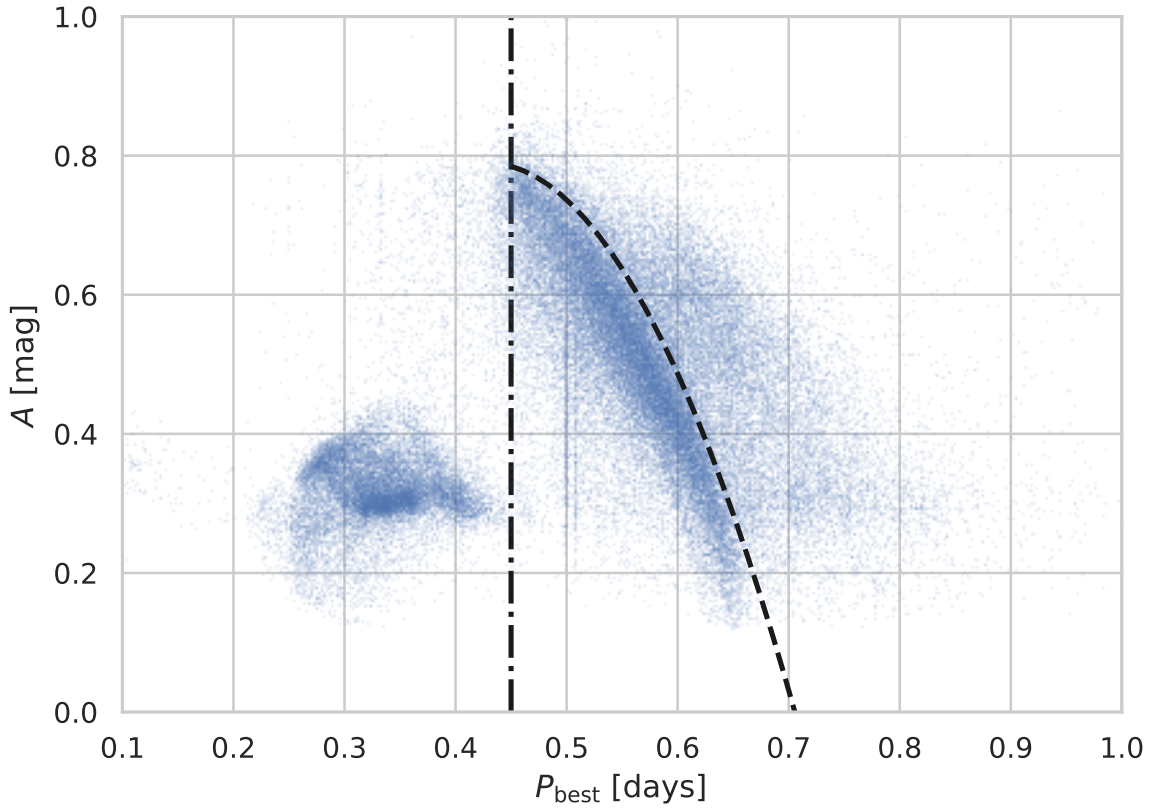

Figure 3.13: The distribution of detected RRLs on the period-amplitude diagram, where $P_{\rm best}$ is the period and $A = \sqrt{A_r^2 + A_g^2}$ is the total amplitude of the best fit in g and r bands. The dash-dotted line of $P_{\rm best} = 0.45$ days is the boundary to roughly separate RRab and RRc stars. The dashed curve is the boundary we adopt to separate Oosterhoff I and II for RRab stars in Equation 3.20.

where the flattening $q \sim 0.6$ for the spheroidal stratification according to literature about the Galactic density profile fitting (e.g. Iorio et al., 2018). We further define the Galactocentric longitude Φ and latitude Θ as

$$\begin{aligned} \Theta &= \arctan \frac{Z}{R_{XY}} \\ \Phi &= \arctan \frac{Y}{X}. \end{aligned} \tag{3.19}$$

To study the density profile of the Galactic halo from different perspectives in the Galactocentric coordinate, we need to evaluate the RRL number density $\rho_{\rm RRL}$ based on our RRL catalogue. The calculation of $\rho_{\rm RRL}$ is the number of RRLs per volume, where we take into account two factors when evaluating the volume: the ZTF coverage of declination $> -28°$ and the completeness as a function of ZTF epoch and magnitude. Given a grid at (X, Y, Z), we compute its declination and count the grid if it is above $-28°$. Based on the position of the grid, we calculate the mean r-band epoch utilizing HEALPix with `nside` = 8 for all ZTF sources. Besides, knowing the heliocentric distance of the grid and using the r-band absolute magnitude of $M_r = 0.54$ mag which maximizes the histogram of M_r of all 71,755 RRLs, we evaluate the r-band magnitude of RRLs for the grid. With the mean r-band epoch and the r-band magnitude of RRLs for the given grid, we compute the completeness on the grid by interpolating the value from the completeness matrix shown in the left panel of Figure 3.9.

To visualize the spheroidal stratification of the Galactic halo density profile and to look for Galactic disk RRLs, we show the RRL number density $\rho_{\rm RRL}$ around the Galactic halo on the $R_{XY} - Z$ plane in Figure 3.11, assuming the density profile is cylindrically symmetric. The black contours of $\rho_{\rm RRL} = 10^0, 10^{0.5}, 10^1, 10^{1.5}, 10^2, 10^{2.5}, 10^3$ kpc^{-3} verify the roughly spheroidal density profile with the flattening $q \sim 0.6$ for $r_{\rm e}$ in Equation 3.18, as indicated by the white dashed elliptical contours. The change of the exponent if modelled by the power-law models, indicated by the distance of any two neighbored contours getting larger as the radius increasing, is consistent with the findings of the single power-law in Iorio et al. (2018). We note that some recent works have found a break in the radial profile of the halo at the Galactocentric distances of 25-30 kpc (e.g. Medina et al., 2018; Stringer et al., 2021). Despite having more RRLs near the disk compared to other catalogues as discussed in Section 3.4.3, our catalogue still lacks some RRLs at the regions near the disk with roughly $|Z| < 2$ kpc, which can be seen at the regions with roughly $|b| < 3°$ in Figure 3.5 as well.

As the Galactic halo stellar density profile is potentially triaxial (Iorio et al., 2018), to study the substructure in the Galactic halo, we also look at the RRL density distribution in the coordinate of Galactocentric longitude Φ and latitude Θ defined in Equation 3.19. The left panels in Figure 3.12 illustrate the RRL number density $\rho_{\rm RRL}$ on the spheroidal shells of different elliptical radii $0 < r_{\rm e} < 128$ kpc with the flattening $q = 0.6$, each of which demonstrates the density on the sky view with $\Phi = 180°$ pointing to the Sun and Φ increasing towards the left in the figure. We observe and annotate some known over-densities

of the Galactic halo, including the Sagittarius Stream (Hernitschek et al., 2017), the Virgo over-density (Vivas et al., 2001; Newberg et al., 2002; Duffau et al., 2006; Jurić et al., 2008; Bonaca et al., 2012), and the Hercules-Aquila Cloud (Belokurov et al., 2007b; Simion et al., 2014, 2018). An interesting point from the panels of $16 < r_e < 64$ kpc is that the Northern part of the Hercules-Aquila Cloud is very close to the Virgo over-density, where the possible association of the two over-densities has been discussed in recent literature (e.g. Li et al., 2016; Simion et al., 2019; Balbinot & Helmi, 2021), as well as the Eridanus–Phoenix over-density which however is not in ZTF coverage. It is worth noting that there are over-densities in the Northern and the Southern hemispheres with Φ roughly from 30° to 120° in the outer halo in the bottom panel of $64 < r_e < 128$ kpc, where the south one may be the local wake and the north one may be the collective halo response due to the dynamical reaction of the Galactic halo to the Large Magellanic Cloud (e.g. Garavito-Camargo et al., 2019; Erkal et al., 2020; Conroy et al., 2021).

Apart from the density profile, the composition of RRLs, particularly for the observed over-densities mentioned above, is interesting to study because it is likely related to their birth environment (van Albada & Baker, 1973; Lee & Carney, 1999; Sandage, 2004). The period-amplitude diagram is typical to study the composition of RRLs and to verify the quality of a RRL catalogue, so we show the distribution of our RRLs in Figure 3.13, where the amplitude $A = \sqrt{A_r^2 + A_g^2}$ with A_r and A_g defined in Equation 3.16, and $P_{\rm best}$ is the best fitting period. We note that the location of a star in this diagram can be affected by the presence of the Blazhko effect (Blažko, 1907) or by the period aliasing during the Fourier fitting stage (Lomb, 1976; Scargle, 1982; VanderPlas, 2018). There are two main clusters of the RRL type ab and c (RRab and RRc) roughly separated by the black dash-dotted line of $P_{\rm best} = 0.45$ days; the RRab cluster is to the right whereas the RRc cluster is to the left. It is worth noting that during the classification process, we never separate the two types of RRLs yet the classifier can still identify both of them. There are vertical patterns of RRLs at $P_{\rm best} = 0.33$ and 0.51 days, which are very likely caused by the aliasing period issue when fitting the light curves. Also we note that the RRc stars might be contaminated by binary stars of the W Ursae Majoris type due to their sinusoidal light curves and period ranging between 0.25 and 0.6 days (Rucinski, 1998), which would be hard to distinguish with on our classification pipeline.

Looking closely at each cluster, we can see the Oosterhoff dichotomy (Oosterhoff, 1939; Catelan, 2009), the more populated Oosterhoff I (OoI) and the less populated Oosterhoff II (OoII) that is shifted to longer periods given an amplitude. For the stars of the RRab cluster in our catalogue (whose periods are greater than 0.45 days), we compute the number counts on every grid of the period-amplitude plane, and utilize the grids with maximum number counts in each amplitude bin to fit a relation of $A = -10.26P_{\rm best}^2 + 8.27P_{\rm best} - 0.88$ to describe the distribution of OoI stars on the period-amplitude plane in Figure 3.13. Then we shift the

curve by 0.025 days in the direction of period to roughly separate the OoI and OoII stars as

$$A = -10.26\,(P_{\rm best} - 0.025)^2 + 8.27\,(P_{\rm best} - 0.025) - 0.88 \tag{3.20}$$

which is shown by the black dashed curve in Figure 3.13. With the OoI RRLs to the left of the boundary and the OoII RRLs to the right of the boundary, we define the OoI fraction as $f_1 = N_1/(N_1 + N_2)$ where N_1 and N_2 are the numbers of OoI and OoII RRLs, finding the overall $f_1 = 0.65$.

According to the explained separation of OoI and OoII above, we are able to study the variation of OoI fraction f_1 across the Galactic halo, together with the RRL density distribution. The right panels of Figure 3.12 show f_1 on shells of different elliptical radii $r_{\rm e}$ in the coordinate of the Galactocentric longitude Φ and latitude Θ for the RRab stars in our catalogue. We note that for $64 < r_{\rm e} < 128$ kpc in the bottom panel, f_1 is so noisy that we grey it out. Overall, f_1 is higher at the radii between $16 < r_{\rm e} < 64$ kpc, especially between $16 < r_{\rm e} < 32$ kpc which is roughly consistent with the finding in Figure 2 in Belokurov et al. (2018). We observe that f_1 seems particularly anisotropic for $16 < r_{\rm e} < 32$ kpc. When looking at the locations of individual over-densities such as the Hercules-Aquila Cloud, the Virgo over-density and the Sagittarius Stream on the left panels, we observe no particular high or low f_1 corresponding to these over-densities in the right panels with the exception of somewhat higher f_1 for the Hercules-Aquila Cloud in the $16 < r_{\rm e} < 32$ kpc distance range.

Another interesting point is the slightly higher f_1 around the solar neighbourhood $(\Phi, \Theta) = (180^\circ, 0^\circ)$ for $r_{\rm e} \sim 8$ kpc, which might be the Splash stars dubbed in Belokurov et al. (2020), yet f_1 around the disk for $16 < r_{\rm e} < 32$ kpc is way higher than f_1 in the solar neighbourhood with Φ between 120° to 210° and Θ between -30° to 30°.

3.5 Conclusions

In this work, we have presented the RRL catalogue constructed from the combination of ZTF DR3 with *Gaia* EDR3, where *Gaia* provides accurate positions and proper motions on the whole sky and ZTF provides the vast amount of light curves with large epochs in multi-bands in the Northern sky. Starting from the source list in the join set of *Gaia* EDR3 and ZTF DR3 and the label of the SOS *Gaia* DR2 RRLs, we have processed them through the classification pipeline, that included the light curve fitting by a constant, single sinusoidal, third-order Fourier model in multiple bands, and two random forest classification steps to predict the probability for each source being a RRL.

Generating the RRL catalogue based on the predicted probability, we have obtained a catalogue that consists of 71,755 objects predicted to be RR Lyrae with at least 92% purity and 92% completeness compared to the SOS *Gaia* DR2 RRLs in the high galactic latitude areas with a high number of Gaia observations. The completeness of the RRL catalogue is generally higher than 80% at the heliocentric distances closer than 80 kpc but drops drastically

to 0 after 100 kpc. The catalogue covers the Northern sky above -28° in declination and the most distant RRL in it is at the heliocentric distance of 132 kpc. Compared with other RRL catalogues covering the Northern sky, the RRL catalogue of this work has more RRLs in the Galactic halo and is more complete at low Galactic latitude areas.

Using the new constructed RRL catalogue to analyze the Galactic halo density distribution, we observe the broadly ellipsoidal stellar distribution with flattening around 0.6 and power-law density profile with three known major over-densities of the halo substructure dominating: the Virgo over-density, the Hercules-Aquila Cloud, and the Sagittarius Stream. We do not observe a significant population associated with the Galactic disk (Iorio & Belokurov, 2021). The RRL density distribution seems to demonstrate the connection between the Virgo over-density and the Hercules-Aquila Cloud, supporting the possible association of several over-densities such as Hercules-Aquila, Virgo, Eridanus–Phoenix and their link to the Gaia-Encelladus-Sausage merger (i.e. Simion et al., 2019). Besides, the RRL over-density in the Northern hemispheres is in broad agreement with the effect of the dynamical response of the Galactic halo to the Large Magellanic Cloud (i.e. Conroy et al., 2021). We also analyse the Oosterhoff fraction differences across the halo, comparing it to the density distribution. We observe a higher fraction at the radii between $16 < r_e < 32$ kpc with some anisotropy across the sky, but no clear association of this with known major over-densities.

Acknowledgements

SK was previously supported by NSF grants AST-1813881, AST-1909584, and Heising-Simons Foundation grant 2018-1030. This paper has made use of the Whole Sky Database (wsdb) created by Sergey Koposov and maintained at the Institute of Astronomy, Cambridge with financial support from the Science & Technology Facilities Council (STFC) and the European Research Council (ERC). This paper has made use of the Q3C software (Koposov & Bartunov, 2006).

This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are being processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the *Gaia* MultiLateral Agreement (MLA). The *Gaia* mission website is https://www.cosmos.esa.int/gaia. The *Gaia* archive website is https://archives.esac.esa.int/gaia.

Software: PYTHON (Van Rossum & Drake, 2009), NUMPY (van der Walt et al., 2011), SCIPY (Jones et al., 2001), PANDAS (Mckinney, 2010), MATPLOTLIB (Hunter, 2007), SEABORN (Waskom et al., 2016), ASTROPY (Astropy Collaboration et al., 2013), SQLUTILPY (Koposov, 2018), HEALPY (Górski et al., 2005; Zonca et al., 2019).

4 The early growth of supermassive black holes in cosmological hydrodynamic simulations with constrained Gaussian realizations

Kuan-Wei Huang[1], Yueying Ni[1], Yu Feng[2], and Tiziana Di Matteo[1,3]

[1] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2] Berkeley Center for Cosmological Physics, University of California at Berkeley, Berkeley, CA, 94720, USA
[3] School of Physics, The University of Melbourne, VIC 3010, Australia

Abstract

The paper examines the early growth of supermassive black holes (SMBHs) in cosmological hydrodynamic simulations with different BH seeding scenarios. Employing the constrained Gaussian realization, we reconstruct the initial conditions in the large-volume BlueTides simulation and run them to $z = 6$ to cross-validate that the method reproduces the first quasars and their environments. Our constrained simulations in a volume of $(15\,h^{-1}\mathrm{Mpc})^3$ successfully recover the evolution of large-scale structure and the stellar and BH masses in the vicinity of a $\sim 10^{12}\,M_\odot$ halo which we identified in BlueTides at $z \sim 7$ hosting a $\sim 10^9\,M_\odot$ SMBH. Among our constrained simulations, only the ones with a low-tidal field and high-density peak in the initial conditions induce the fastest BH growth required to explain the $z > 6$ quasars. We run two sets of simulations with different BH seed masses of 5×10^3, 5×10^4, and $5\times10^5\,h^{-1}M_\odot$, (a) with the same ratio of halo to BH seed mass and (b) with the same halo threshold mass. At $z = 6$, all the SMBHs converge in mass to $\sim 10^9\,M_\odot$ except for the one with the smallest seed in (b) undergoing critical BH growth and reaching $10^8 - 10^9\,M_\odot$, albeit with most of the growth in (b) delayed compared to set (a). The finding of eight BH mergers in the small-seed scenario (four with masses $10^4 - 10^6\,M_\odot$ at $z > 12$), six in the intermediate-seed scenario, and zero in the large-seed scenario suggests that the vast BHs in the small-seed scenario merge frequently during the early phases of the growth of SMBHs. The increased BH merger rate for the low-mass BH seed and halo threshold scenario provides an exciting prospect for discriminating BH formation mechanisms with the advent of multi-messenger astrophysics and next-generation gravitational wave facilities.

4.1 Introduction

The formation of the first supermassive black holes (SMBHs) remains challenging in our standard paradigm of structure formation. SMBHs, as massive as those in galaxies today, are known to exist in the early universe, even up to $z \sim 7.5$. Luminous, extremely rare, quasars at $z \sim 6$ were initially discovered in the Sloan Digital Sky Survey (Fan et al., 2006; Jiang et al., 2009) and, until recently, the highest redshift quasar known (Wu et al., 2015) at $z = 7.09$ (Mortlock et al., 2011) has been surpassed by the discovery of a bright quasar at $z = 7.54$ (Bañados et al., 2018), which is currently the record holder for known high redshift quasars. A further sample of two $z > 7$ has also been recently discovered (Yang et al., 2019). The presence of luminous quasars observed within the first billion years of the Universe highlights that the BH seeds for the SMBH population must have assembled at the cosmic dawn, concurrently with the time of the formation of the first stars or galaxies.

However, the precise SMBH seed formation mechanism remains unknown, nor is it clear if there is only one seed formation channel at play over the entire SMBH seed mass spectrum of models. The current scenarios suggest that seed BHs are (a) remnants of the first generation of stars (PopIII) (e.g. Madau & Rees, 2001; Abel et al., 2002; Johnson & Bromm, 2007) or (b) direct gas collapse within the first massive halos (e.g. Lodato & Natarajan, 2006; Begelman et al., 2006; Regan & Haehnelt, 2009; Ferrara et al., 2014; Latif et al., 2013) or (c) runaway collapse of dense nuclear star clusters (e.g. Begelman & Rees, 1978; Devecchi & Volonteri, 2009; Yajima & Khochfar, 2016; Katz et al., 2015). The seed BHs then range in mass from a few hundred for (a) to $10^5 M_\odot$ for (b) and (c).

In large-volume cosmological simulations, a common and widely used sub-grid model for SMBHs and active galactic nuclei (AGN) feedback has been proposed in Di Matteo et al. (2005). Since the SMBH seed formation process is not resolved by cosmological simulations (see Regan & Haehnelt, 2009, for a review), it is assumed that every halo above a certain threshold mass hosts a central BH seed. Halos are selected for seeding by regularly running the 'Friends-of-Friends' (FoF) halo finder on the dark matter distribution. The BH seed mass ($M_\bullet^{\rm seed}$) and the threshold halo mass ($M_{\rm fof}^{\rm seed}$) are the parameters in simulations. Although this is an ad-hoc seeding procedure, the initial seed BH mass subsequently grows in these simulations via mergers and accretion. Many simulations have adopted this or a similar scenario and gotten good agreements with observations such as the MassiveBlack simulation (Di Matteo et al., 2012), the Illustris simulation (Vogelsberger et al., 2014), the Evolution and Assembly of GaLaxies and their Environment (EAGLE) suite of SPH simulation (Schaye et al., 2015), the MassiveBlack II simulation (Khandai et al., 2015), and the BlueTides simulation (Feng et al., 2016a). Some recent studies have implemented different, physically motivated approaches where the BH seeding is based on gas properties such as Bellovary et al. (2011); Habouzit et al. (2017); Tremmel et al. (2017). However, it is worth noting that these models were adopted in much smaller volume simulations than,

for example, MassiveBlack II or BlueTides. From those, it is not possible to validate the basic statistical properties of the BH population (e.g. luminosity functions or mass functions) against currently observed samples.

Taking BlueTides, a large-volume and high-resolution cosmological hydrodynamic simulation with 2×7040^3 particles in a box of $400\,h^{-1}$Mpc on a side, as an example, SMBHs are modeled as follows. For each FoF halo with a mass above $M_{\rm fof}^{\rm seed} = 5 \times 10^{10}\,h^{-1}M_\odot$, a SMBH is seeded with an initial seed mass $M_\bullet^{\rm seed} = 5 \times 10^5\,h^{-1}M_\odot$ at the position of the local minimum potential if there is no SMBH in that halo. After being seeded, gas accretion proceeds according to Bondi (1952) while the BH accretion rate is limited to two times the Eddington rate. When SMBHs are accreting, we assume that some fraction of the radiated luminosity can couple thermally and isotropically to surrounding gas in the form of feedback energy (Springel, 2005; Di Matteo et al., 2005).

Adopting this SMBH model and appropriate sub-grid physics for the galaxy formation modeling, the BlueTides simulation has predicted various quantities in good agreements with current observational constraints in the high-z universe such as UV luminosity functions (Feng et al., 2016a; Waters et al., 2016a,b; Wilkins et al., 2017), the first galaxies and the most massive quasars (Feng et al., 2015; Di Matteo et al., 2017; Tenneti et al., 2018), the Lyman continuum photon production efficiency (Wilkins et al., 2016, 2017), galaxy stellar mass functions (Wilkins et al., 2018), angular clustering amplitude (Bhowmick et al., 2017), BH-galaxy scaling relations (Huang et al., 2018), and gas outflows from the $z = 7.54$ quasar (Ni et al., 2018). Important for our work here, BlueTides, with its large volume and appropriate resolution, is currently the only cosmological hydrodynamic simulation that makes direct contact with the rare, first quasar population at $z > 7$.

However, an essential question for the SMBH sub-grid model is how different parameters (e.g. $M_\bullet^{\rm seed}$ or $M_{\rm fof}^{\rm seed}$) may affect the growth of SMBHs and the local environment in cosmological simulations. Changing the BH seed mass and re-running such a large-volume simulation multiple times is completely prohibitive because it is computationally expensive even on the largest current national facilities. To reduce the demand on computational resources, a common method is to run a "zoom-in re-simulation" with a higher resolution or different physical parameters from a certain region selected from a large-volume lower-resolution simulation. This allows people to focus on a specific environment numerically and has been applied to study SMBHs in simulations for various purposes (e.g. Li et al., 2007; Sijacki et al., 2009; Hopkins & Quataert, 2010; Bournaud et al., 2011; Romano-Diaz et al., 2011; Bellovary et al., 2013; Dubois et al., 2013; Anglés-Alcázar et al., 2014; Costa et al., 2014; Feng et al., 2014).

In this paper, we combine the technique of constrained Gaussian realization and cosmological hydrodynamic simulations to reduce the demand on computational resources. Hoffman & Ribak (1991) first introduced an optimal solution to the problem of the construction of constrained realizations of Gaussian fields by demonstrating how the algorithm

generates constrained fields with a simple single-density peak. Later on, van de Weygaert & Bertschinger (1996) in addition developed an algorithm to set up initial Gaussian random density and velocity fields containing multiple constraints of arbitrary amplitudes and positions. Integrating the algorithm to cosmological hydrodynamic simulations has arisen in the past few years to explore dark matter halos and galaxy formation (Roth et al., 2016; Porciani, 2016; Pontzen et al., 2017).

With the constrained Gaussian realization, we can constrain the initial density field by adding a desirable height of a density peak when generating the initial condition such that a more massive halo can still form in a relatively small box compared to those large-volume ($\sim$ Gpc per side) cosmological simulations (with uniform/unconstrained initial condition) . For instance, we can grow a halo with a mass $\sim 10^{12}\,M_{\odot}$ in a box of $15\,h^{-1}$Mpc on a side at $z = 8$, whose mass is similar to the one hosting the most massive BH in BlueTides simulation ($400\,h^{-1}$Mpc on a side) under the same resolution. This reduces the computational demand by a factor of $(400/15)^3 \sim 20000$. This approach is a more general way to study the growth of SMBH compared to the zoom-in re-simulation method because the goal of the latter is to exactly study a particular object/region (for example, a particular halo). However, our approach is aiming to study characteristic environments by creating one or more different realizations but with similar properties such as halo mass or tidal field to the object/region of our interest (which we extract, for comparison, in the uniform large volume simulations with the exact physics).

As we shall further demonstrate, besides the density constraint, we need another condition related to the ICs that induces the fastest growth for SMBHs in cosmological simulations. This is expected, as the observed population of quasar-like SMBHs at high redshifts is even rarer than the massive halos. For example, there is only one SMBH with a mass above $10^8\,M_{\odot}$ in a halo with a mass $\sim 10^{12}\,M_{\odot}$ while there are > 50 halos more massive than that halo at $z = 8$. An environmental property, also related to the ICs that has been found to be relevant to induce fast BH growth is the local tidal field strength (Di Matteo et al., 2017). In particular, using the large volume BlueTides simulations, Di Matteo et al. (2017) has shown that isolated regions of low tidal fields are key to the fast growth of the firs SMBHs. As a consequence, we also choose the realization with a lower tidal field around the local environment where the halo forms, which indeed helps more massive SMBHs grow in simulations (see Section 4.2.4).

After significantly decreasing the demand on computational resources with the constrained Gaussian realizations and a lower tidal field realization, we are finally able to examine how sensitive the SMBH growth is to the BH seed mass in the sub-grid model by running multiple cosmological simulations. According to the different hypotheses of the BH formation scenario, the BH seed mass has been suggested to range from 10^2 to $10^6\,h^{-1}M_{\odot}$ (Haehnelt & Rees, 1993; Loeb & Rasio, 1994; Eisenstein & Loeb, 1995; Bromm & Loeb, 2003; Koushiappas et al., 2004; Begelman et al., 2006; Lodato & Natarajan, 2006; Zhang

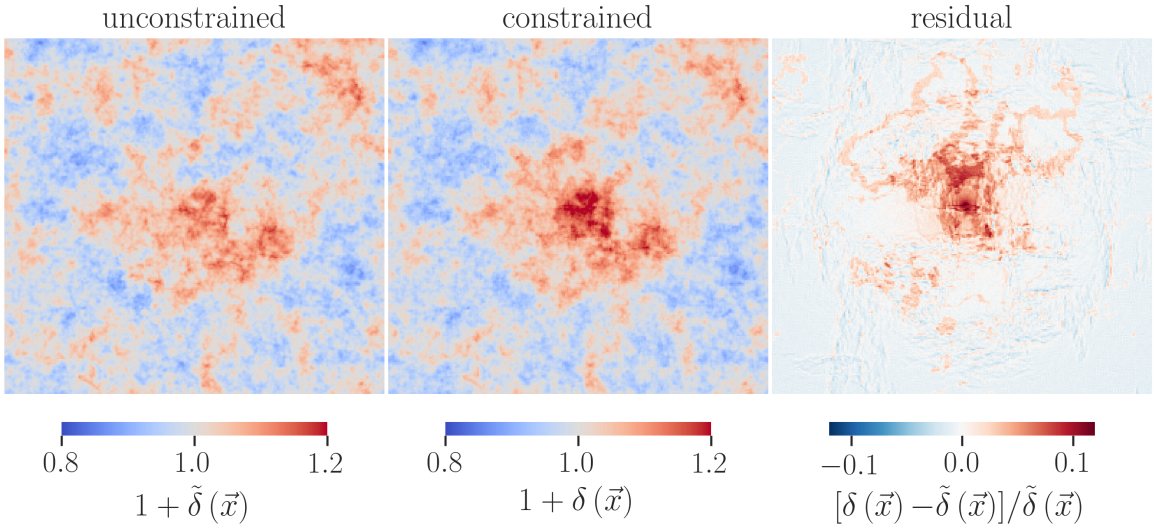

Figure 4.1: Slices of density fields of the initial conditions with the same realization number. **Left**: without any constraints. **Middle**: with a constrained density peak at the densest region of the original field (also the center of the panel). The boxes are 15 h^{-1}Mpc per side with a thickness of 5 h^{-1}Mpc. **Right**: the residual of the constrained and unconstrained density fields.

et al., 2008; Volonteri, 2010; Latif et al., 2013; Schleicher et al., 2013; Ferrara et al., 2014). Therefore we focus on the three different seed masses: 5×10^3, 5×10^4, and 5×10^5 $h^{-1}M_\odot$, with the same ratio of halo to BH seed mass and with the same halo threshold mass in this paper.

We organize the paper as the following. Section 4.2 describes the constrained Gaussian realizations, compares the constrained and unconstrained initial conditions and simulations, and discusses the effect of different tidal fields on the local environment of SMBHs. Section 4.3 demonstrates the results of the early growth of SMBHs and their hosts in the simulations with different BH seeding scenarios. Section 4.4 concludes the paper.

4.2 methodology

4.2.1 Constrained initial conditions

The first quasars are extremely rare and hence the massive halos hosting these first SMBHs also have to be commensurately rare. Traditionally one needs extremely large-volume simulations to simulate such objects. For example, there is only one SMBH with a mass above 10^8 $M_\odot$ in a halo with a mass of $\sim 10^{12}$ $M_\odot$ at $z = 8$ in the BlueTides simulation with 400 h^{-1}Mpc on a side of the cube. Using simulations of such scale to study the effect on BH growth due

to subgrid prescriptions is prohibitively expensive. In this work, we work with much smaller simulation boxes ($15\,h^{-1}$Mpc), but we add constraints to the initial condition (linear field) to ensure the existence of extreme density peaks. With high-density peaks in the linear field, we can guarantee the formation of massive halos at later times even in a smaller simulation box. These halos are then used to study the early growth of SMBHs in the massive halos in cosmological hydrodynamic simulations with different BH seeding parameters.

According to the linear growth theory, to generate a $10^{12}\,M_{\odot}$ halo at z=8, we need a $5\,\sigma$ peak in the underlying over-density field. To achieve that, we use FastPM, a particle mesh based quasi N-body solver (Feng et al., 2016b), to generate initial conditions with constrained Gaussian realizations for the simulations in the paper for the first time. Thanks to the contribution of Aslanyan et al. (2016), FastPM is capable of producing constrained Gaussian density fields to the initial condition, which we will describe the basic idea and the implementation in the following paragraphs.

First introduced by Hoffman & Ribak (1991), the constrained realization technique was then explained in a lot more detail in van de Weygaert & Bertschinger (1996) which we refer readers to. The goal is to construct a field $f(\mathbf{x})$ subject to a set of M constraints:

$$\Gamma = \{C_i \equiv C_i[f; \mathbf{x}_i] = c_i;\ i = 1, ..., M\} \tag{4.1}$$

where the constraint C_i can be viewed as a functional of $f(\mathbf{x})$ field (here in our specific case, the overdensity field) to have the specific value c_i at the position $\mathbf{x}_i$.

To obtain a field $f(\mathbf{x})$ satisfying the constraint Γ, one can start with a random, unconstrained Gaussian realization $\tilde{f}(\mathbf{x})$ and impose on that an "ensemble mean field" $\bar{f}(\mathbf{x})$ corresponding to the desired constraint Γ. More specifically, the ensemble mean field $\bar{f}(\mathbf{x})$ can be written in the form of:

$$\bar{f}(\mathbf{x}) = \langle f(\mathbf{x}) \,|\, \Gamma \rangle = \xi_i(\mathbf{x})\, \xi^{-1}_{ij}\, c_j \tag{4.2}$$

where $\xi_i(\mathbf{x}) = \langle f(\mathbf{x})\, C_i \rangle$ is the cross-correlation between the $f(\mathbf{x})$ field and the i^{th} constraint C_i, and ξ^{-1}_{ij} is the (i, j) element of the inverse of the constraint covariance matrix $\langle C_i\, C_j \rangle$. Note that the summation over repeated indices is used. The ensemble mean field $\bar{f}(\mathbf{x})$ can be interpreted as the "most likely" field subject to the set of constraints Γ.

We further introduce the "residual field" $F(\mathbf{x}) \equiv f(\mathbf{x}) - \bar{f}(\mathbf{x})$ as the difference between an arbitrary Gaussian realization $f(\mathbf{x})$ satisfying the constraint set Γ and the ensemble mean field $\bar{f}(\mathbf{x})$. The crucial idea of the constrained realization construction method is based on the fact that, the complete probability distribution $\mathscr{P}[F \,|\, \Gamma]$ of the residual field $F(\mathbf{x})$ is independent of numerical values c_i of the constraints in Γ (c.f. Hoffman & Ribak, 1991; van de Weygaert & Bertschinger, 1996, for the detailed derivations). That is, for any Γ_p and Γ_q where $p \neq q$,

$$\mathscr{P}[F \,|\, \Gamma_p] = \mathscr{P}[F \,|\, \Gamma_q] \tag{4.3}$$

Therefore, we can construct the desired realization under a constraint set Γ by properly sampling the residual field $F(\mathbf{x})$ from a random, unconstrained realization $\tilde{f}(\mathbf{x})$ and then add that $F(\mathbf{x})$ to the ensemble field $\bar{f}(\mathbf{x})$ corresponding to Γ. The formalism can be written as

$$\begin{aligned} f(\mathbf{x}) &= F(\mathbf{x}) + \bar{f}(\mathbf{x}) \\ &= \left(\tilde{f}(\mathbf{x}) - \xi_i(\mathbf{x})\, \xi_{ij}^{-1}\, \tilde{c}_j\right) + \xi_i(\mathbf{x})\, \xi_{ij}^{-1}\, c_j \\ &= \tilde{f}(\mathbf{x}) + \xi_i(\mathbf{x})\, \xi_{ij}^{-1}\, \left(c_j - \tilde{c}_j\right) \end{aligned} \tag{4.4}$$

In other words, we treat the original $\tilde{f}(\mathbf{x})$ as a field subject to a constraint set $\tilde{\Gamma}$ with $\tilde{c}_j = C_j[\tilde{f}; \mathbf{x}_j]$ where $\tilde{c}_j$ is the original value of the unconstrained field. We then have the ensemble mean field corresponding to $\tilde{\Gamma}$ as $\bar{\tilde{f}}(\mathbf{x}) = \xi_i(\mathbf{x})\, \xi_{ij}^{-1}\, \tilde{c}_j$. Getting the residual field $F(\mathbf{x})$ from a random unconstrained realization by $\tilde{f}(\mathbf{x}) - \bar{\tilde{f}}(\mathbf{x})$, we then add $F(\mathbf{x})$ to $\bar{f}(\mathbf{x})$ to obtain the field $f(\mathbf{x})$ satisfying the constraint Γ. It is well established in van de Weygaert & Bertschinger (1996) that the $f(\mathbf{x})$ field constructed in this way is a properly sampled realization subject to the desired constraint Γ. This is what we implemented in our code.

We note that, however, one limitation in our implementation is that the super-sampling variance (DC mode) is missing. Super-sampling variance is the effect of the coupling to modes at scales larger than the box size (Li et al., 2014). In our simulation, we assume that the overdensity of the whole simulation box to be zero, i.e., the DC mode is zero. The DC mode can be incorporated by the so-called separate universe technique that absorbs the overdensity of the simulation volume into a modified cosmology. (see, e.g., Sirko, 2005; Gnedin et al., 2011; Wagner et al., 2015; Li et al., 2014, 2018, for more details). However, we estimate that with our simulation box size ($15\,h^{-1}$Mpc per side) at these high redshifts ($z > 6$), this effect accounts only about 10 percent.

Here we demonstrate the constrained realization generated via FastPM with a single $5\,\sigma$ density peak. Figure 4.1 shows examples of the density field with and without a constrained density peak and its associated residual map in a dense region at the center in the domain with a box size of $15\,h^{-1}$Mpc. As expected, we find that the density increases in the region where we put the constraint without changing the overall pattern of the density field. According to the residual map, 0.04 percent of the pixels have greater than 10 percent residuals; 2 percent of the pixels exceed 5 percent residuals; none of the pixels have residuals less than -5 percent.

4.2.2 Simulation setup

We use the massively parallel cosmological smoothed particle hydrodynamic (SPH) simulation software, MP-Gadget (Feng et al., 2016a), to run all the simulations in this paper. Its hydrodynamics solver adopts the new pressure-entropy formulation of SPH (Hopkins, 2013). The main sub-grid models in MP-Gadget are

- star formation based on a multiphase star formation model (Springel & Hernquist, 2003) with modifications following Vogelsberger et al. (2013),

Table 4.1: Parameters adopted in our simulations.

h	0.697	$\Omega_{\rm matter}$	0.2814	$M_{\rm DM}$	$1.7 \times 10^7\, M_\odot$
σ_8	0.820	$\Omega_{\rm baryon}$	0.0464	$M_{\rm gas}$	$3.4 \times 10^6\, M_\odot$
n_s	0.971	Ω_Λ	0.7186	$M_\star$	$8.4 \times 10^5\, M_\odot$
ϵ	$1.5\, h^{-1}$kpc	$N_{\rm particle}$	2×264^3	$L_{\rm box}$	$15\, h^{-1}$Mpc

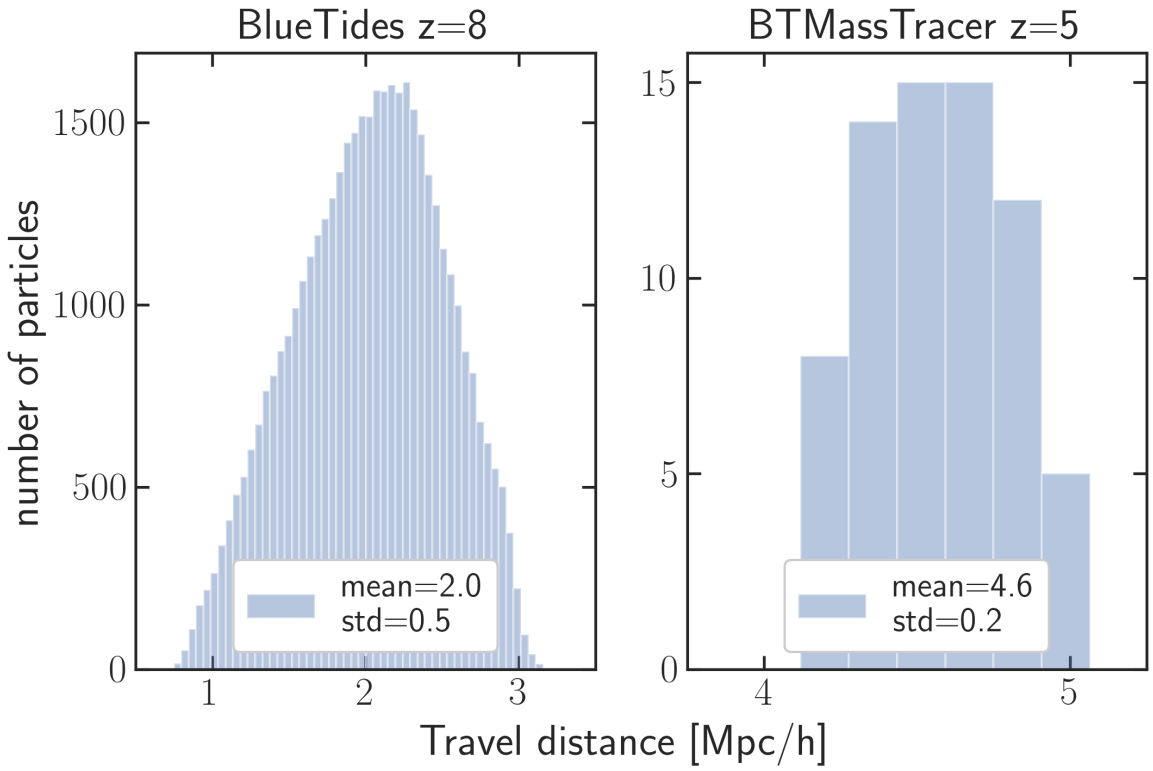

Figure 4.2: Histogram of the travel distance of all particles in the halo hosting the most massive BH in BlueTides from $z = 99$ to $z = 8$ and in the same halo in BTMassTracer from $z = 99$ to $z = 5$.

- gas cooling through radiative processes (Katz et al., 1996) and metal cooling (Vogelsberger et al., 2014),
- formation of molecular hydrogen and its effects on star formation (Krumholz & Gnedin, 2011),
- type II supernovae wind feedback (Nelson et al., 2015),
- SMBH growth and AGN feedback (Di Matteo et al., 2005).

All the new constrained simulations in the paper are run with periodic boundary conditions from $z = 99$ to $z = 6$ (as we are mostly interested in the seed mass and the early growth of SMBHs). Each simulation contains 2×264^3 particles in a cube with the box size $L_{\rm box} = 15\,h^{-1}$Mpc (the choice of $L_{\rm box}$ size will be further discussed later in detail). We adopt the cosmological parameters based on the Nine-Year Wilkinson Microwave Anisotropy Probe Observations (Hinshaw et al., 2013). All the simulations have the same cosmology and resolution as in BlueTides, so that we can always use the direct large volume simulations to assess the validity of the new simulations. Table 4.1 summarizes all the basic parameters of our new runs. Note that a star particle has a mass of $M_\star = \frac{1}{4} M_{\rm gas}$ and that the gravitational smoothing length ϵ is the same for all kinds of particles.

Constraining a high-density peak in the initial density field to get a massive halo allows us to study rare objects in a small simulation box rather than in large-volume cosmological hydrodynamic simulations with the box size of a few hundred Mpc on the side. In order to set a box size for our constrained runs, we need to make sure that the growth history of the halo needs to be well converged. In particular, we would like to make sure that all the particles that make into the halo and all the way into the central BH are captured. In particular, here, we want to track the growth of SMBHs at the center and their host galaxies, so we look into how far the particles in the halo hosting the most massive BH in BlueTides have traveled from $z = 99$ to $z = 8$ in Figure 4.2. The mean of the travel distance is $2\,h^{-1}$Mpc with a standard deviation of $0.5\,h^{-1}$Mpc, indicating that at least a box size of $L_{\rm box} \geq 2\,h^{-1}$Mpc is necessary to contain the halo up to $z = 8$. As we will run our new simulations beyond $z = 8$ here we also make use of a dark matter only realization of BlueTides, the BTMassTracer (Tenneti et al., 2018). We track all the dark matter particles in the halo down to $z = 5$. We show that the particles typically travel a mean of $4.6\,h^{-1}$Mpc and a standard deviation of $0.2\,h^{-1}$Mpc. This implies an absolute minimum box size of $L_{\rm box} \sim 5\,h^{-1}$Mpc. To be rather conservative and make sure we have the appropriate growth history of the halo and its black holes, we choose a size of $15\,h^{-1}$Mpc and stick to this for all the simulations in the paper.

4.2.3 Constrained versus unconstrained simulations

To illustrate the basic features of the constrained simulations, we first run the constrained and unconstrained initial conditions in Figure 4.1 down to $z = 6$ while keeping all the other simulation parameters. Figure 4.3 shows the density fields of the constrained and

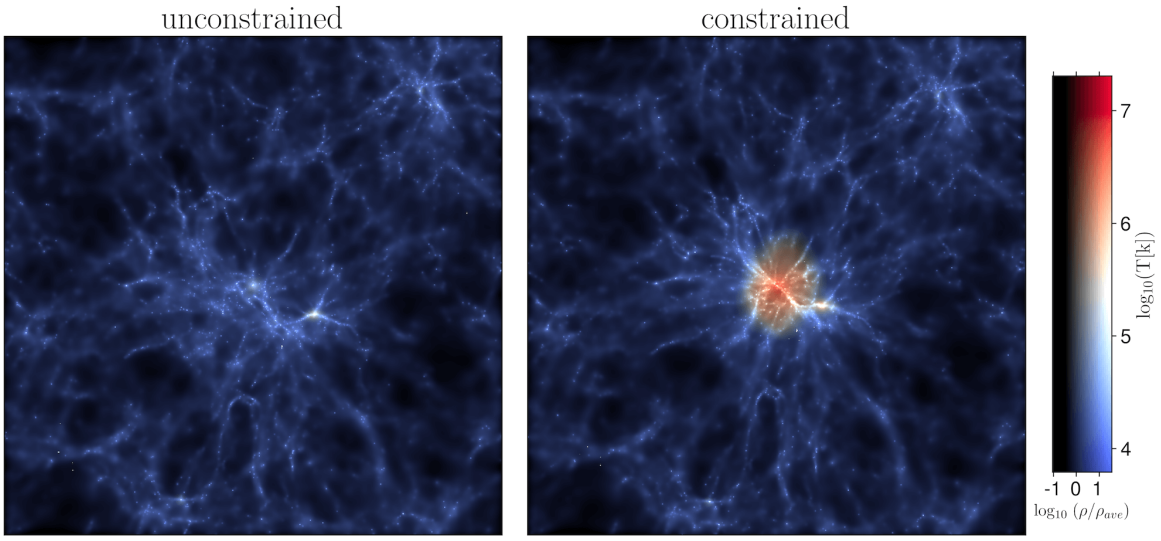

Figure 4.3: The slices of gas density fields of the unconstrained (left) and constrained (right) simulations at $z = 6$. The gas density field is color-coded by temperature as well. The boxes are 15 h^{-1}Mpc per side with a thickness of 5 h^{-1}Mpc.

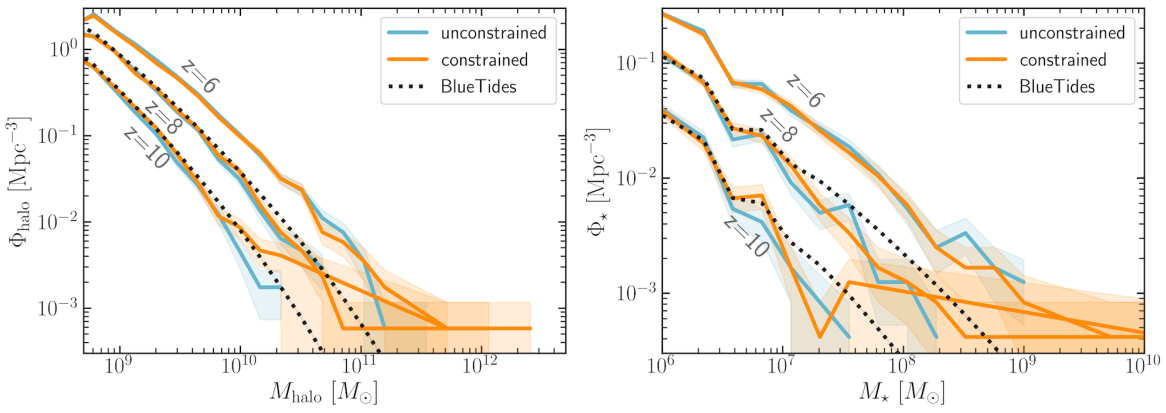

Figure 4.4: Mass functions in the constrained and unconstrained simulations at $z = 6$, 8, and 10 in comparison with BLUETIDES. **Left**: halo mass functions $\Phi_{\rm halo}$. **Right**: galaxy stellar mass functions $\Phi_\star$.

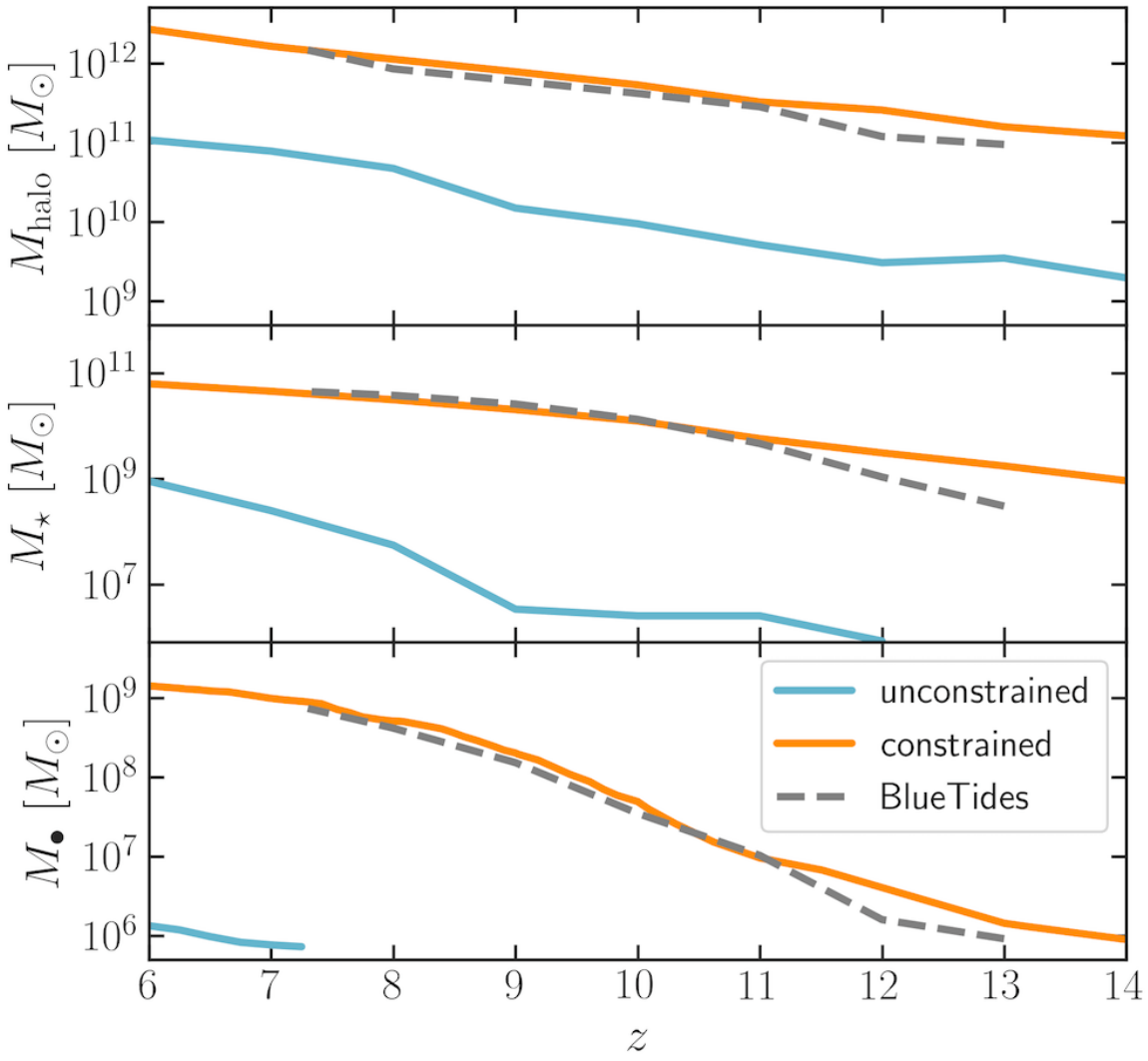

Figure 4.5: The growth history of the host halo and galaxy ($M_{\rm halo}$ and $M_\star$) and the most massive BHs ($M_\bullet$) in the constrained and unconstrained simulations in comparison with BlueTides.

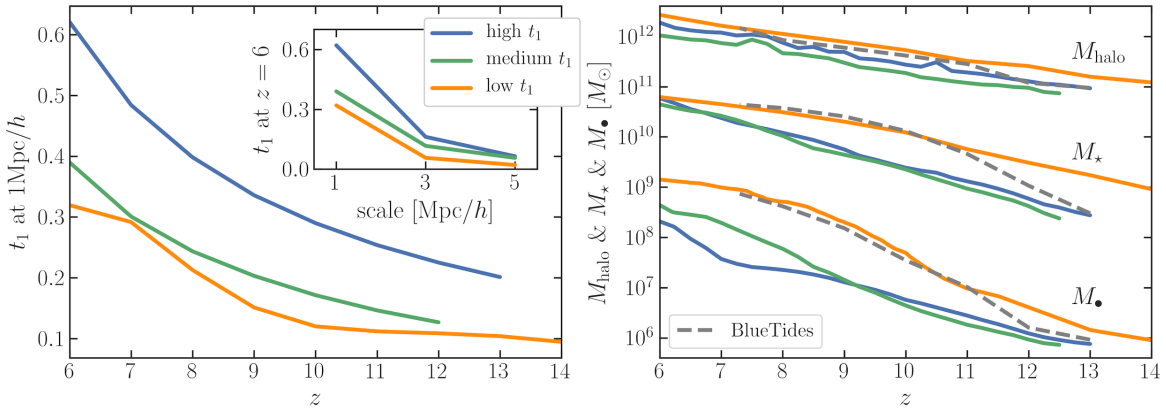

Figure 4.6: **Left**: tidal field strength t_1 measured at the position of the most massive BHs in the three simulations. The inner panel shows t_1 measured at different scales at $z = 6$ and the outer panel shows the evolution of t_1 measured at 1 h^{-1}Mpc. **Right**: the growth history of the host halo and galaxy ($M_{\rm halo}$ and $M_\star$) and the most massive BHs ($M_\bullet$) in the three simulations. The grey dashed curves are the quantities of the most massive BH in BlueTides.

unconstrained simulations at $z = 6$ color-coded by temperature as well. As expected, the density around where we put the constrained peak in the constrained simulation is higher than the unconstrained one while the overall structure maintains. So is the temperature. Figure 4.4 shows the halo and stellar mass functions ($\Phi_{\rm halo}$ and $\Phi_\star$) at $z = 6$, 8, and 10, compared with BlueTides. In particular, there is one halo in the very massive end of these functions in the constrained simulation due to the constrained high-density peak. Aside from the massive objects, the consistency of both mass functions with each other and with the ones in BlueTides indicates that the constrained simulation appropriately captures the growth of halo and stellar mass function statistically.

We then investigate the growth history of the most massive BHs ($M_\bullet$) and their hosts (halo mass $M_{\rm halo}$ and stellar mass $M_\star$) in the two simulations compared with that of the BlueTides simulation in Figure 4.5. With a proper density peak, the growth history of the three masses in the constrained simulation converges to the ones in BlueTides (note that a total convergence is not expected as this a new constrained simulation but not a zoom-in simulation). On the other hand, the halo mass of the unconstrained simulation at $z = 6$ is an order of magnitude less massive than the one in the constrained simulation; the stellar mass is around two orders of magnitude less massive; the BH mass is three orders of magnitude less massive.

4.2.4 Tidal fields of the SMBHs

While a highly biased region (as in our constrained simulations) is a necessary condition for growing a massive BH, it is not sufficient. For example in the BlueTides simulation, only

one of the 50 most massive halos of mass similar or greater than the one hosting the most massive BH has a BH more massive than $10^8\,M_\odot$. As we shall further show, not all of the constrained Gaussian realizations that can grow massive halos guarantee to grow SMBHs in them as well. Directly related to the density field in the initial conditions, the local tidal field has been identified as the environmental property that is the most strongly correlated to the growth of the first quasars in Di Matteo et al. (2017). In these findings, the extreme early growth depends on the early interplay of high gas densities and the tidal field that shapes the mode of accretion in those halos.

The tidal field is characterized by the three eigenvalues (t_1, t_2, t_3) of the local tidal tensor $T_{ij} \equiv S_{ij} - \frac{1}{3}\sum_i S_{ii}$, where the strain tensor is the second derivative of the potential, $S_{ij} \equiv \nabla_i \nabla_j \phi$. According to Dalal et al. (2008), S_{ij} is calculated in Fourier space as $\hat{S}_{ij} = \frac{k^2}{k_i k_j}\hat{\delta}$. The three eigenvalues are by definition $t_1 > t_2 > t_3$ and satisfy $t_1 + t_2 + t_3 = 0$ so that t_1 is always positive and t_3 is negative. Thus, the tidal field stretches material along with $\mathbf{t_1}$ and compresses material along with $\mathbf{t_3}$, where ($\mathbf{t_1}$, $\mathbf{t_2}$, $\mathbf{t_3}$) are the corresponding eigenvectors. To use t_1 as the indicator of the local tidal field strength following the standard usage, we calculate t_1 numerically using NBODYKIT (Hand & Feng, 2015). We read all the particle data from a snapshot into a mesh object weighted by the particle mass to get $\hat{\delta}$; transform them to Fourier space; apply a kernel of $\frac{k^2}{k_i k_j}$ to get $\hat{S}_{ij}$; transform them back to the real space and evaluate T_{ij} at the position of the SMBH.

To further evaluate the role of the tidal field in the growth of the first massive SMBHs, we generate a number of constrained realizations; select the ones with the minimum, intermediate, and maximum t_1 around the density peak as the initial conditions; run them from $z = 99$ to $z = 6$. The left panel of Figure 4.6 shows t_1 at the position of the most massive BHs in the three simulations. t_1 measured at $z = 6$ on different scales of $1 - 5\,h^{-1}$Mpc in the inset suggests that the simulation with a lower or higher t_1 is always lower or higher across the scales. The evolution of t_1 measured at the scale of $1\,h^{-1}$Mpc in the main panel shows that a lower or higher t_1 environment tends to maintain a lower or higher t_1 as time goes.

The right panel of Figure 4.6 shows the growth history of the most massive BHs and their hosts in the three simulations. Several interesting results we find include that, the masses of the halos $M_{\rm halo}$ always differ by a factor less than 10 among the three simulations; the stellar mass $M_\star$ of the low t_1 simulation is around an order of magnitude higher than the others at an earlier stage; the BH masses $M_\bullet$ differ by a factor of 10 – 100. The three simulations suggest that the tidal field has a larger impact on the growth of SMBHs: a SMBH can grow more or less massive when it is in a lower or higher t_1 surrounding environment. Besides, the growth history of the most massive BH and its host galaxy and halo in the low t_1 simulation also converges better to the BLUETIDES simulation.

The fact that a lower tidal field environment helps a more massive growth of the SMBHs in our simulations strengthens the findings in Di Matteo et al. (2017) that the local tidal field is strongly correlated to the growth of the first quasars. Moreover, we utilize the constrained

Table 4.2: The sets, names, the BH seed mass $M_\bullet^{\rm seed}$, and threshold halo mass $M_{\rm fof}^{\rm seed}$ in the simulations.

Set	Name	$M_\bullet^{\rm seed}$ $[h^{-1}M_\odot]$	$M_{\rm fof}^{\rm seed}$ $[h^{-1}M_\odot]$
A [a,e]	B3H8	5×10^3	5×10^8
A	B4H9	5×10^4	5×10^9
A, B	B5H10 [c,d]	5×10^5	5×10^{10}
B	B4H10	5×10^4	5×10^{10}
B [b]	B3H10	5×10^3	5×10^{10}

[a] Set A contains the simulations with different $M_\bullet^{\rm seed}$ and $M_{\rm fof}^{\rm seed}$.
[b] Set B contains the simulations with different $M_\bullet^{\rm seed}$ only.
[c] B5H10 is the same simulation as the constrained simulation and the low-t_1 simulation in Section 4.2.
[d] B5H10 has the same seeding parameters as that of BlueTides.
[e] Feng et al. (2014) has examined the exact pairs of $M_\bullet^{\rm seed}$ and $M_{\rm fof}^{\rm seed}$ in Set A using zoom-in simulations from the MassiveBlack simulation.

Table 4.3: The numbers of BHs in the simulations at different redshifts.

	B3H8	B4H9	B5H10	B4H10	B3H10
$z=10$	1678	36	1	1	1
$z=8$	4994	135	2	2	2
$z=6$	12210	488	11	11	11

realization that provides the lowest t_1 environment as the initial condition for the study of the early growth of SMBHs with different BH seeding parameters in the following sections.

4.3 Results: different SMBH seeding scenarios

The main objective of this work is to study the effect of different BH seeding parameters on the early growth of SMBHs using a set of constrained cosmological simulations that statistically reproduce the environments for early growth in the large-volume BlueTides simulation. Here, we conduct two sets of new constrained simulations and investigate the growth history of SMBHs starting from different BH seed masses. In particular, we use three different BH seed masses $M_\bullet^{\rm seed} = 5\times10^3$, 5×10^4, and $5\times10^5\ h^{-1}M_\odot$ and perform two sets of simulations.

- In **Set A**: we lower the halo mass threshold to $M_{\rm fof}^{\rm seed} = 5\times10^8$, 5×10^9, and $5\times10^{10}\ h^{-1}M_\odot$ commensurate with keeping the ratio of $M_\bullet^{\rm seed}/M_{\rm fof}^{\rm seed}$ constant. This is motivated by the physical models for BH seed formation implying that smaller BHs

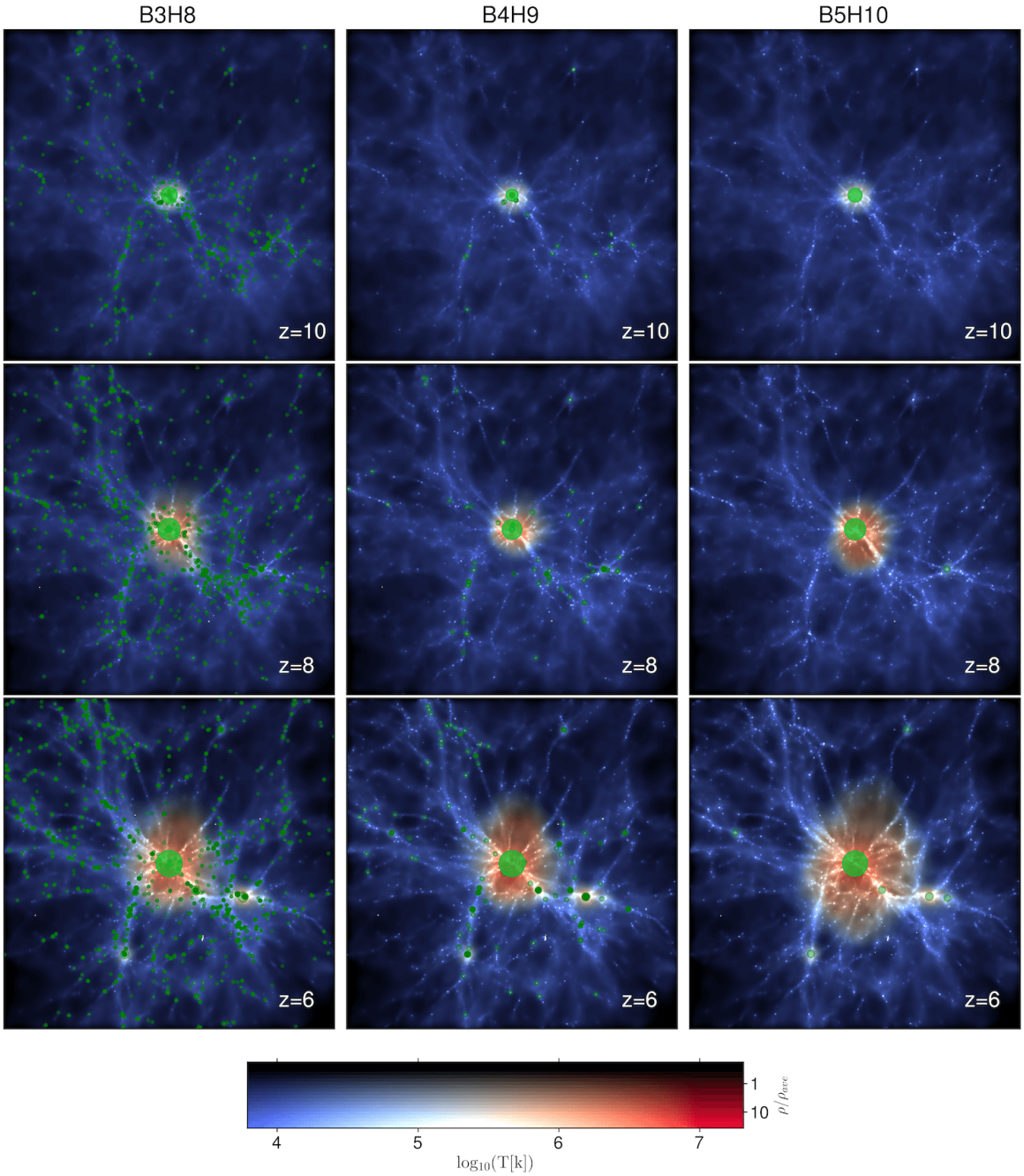

Figure 4.7: The gas density fields of the simulations B3H8, B4H9, and B5H10 (from the left to the right) at $z = 10$, 8, and 6 (from the top to the bottom). Each of them is centered at the most massive BH with a zoomed-in cube of $6\,h^{-1}$Mpc per side. The gas density fields are color-coded by temperature (blue to red indicating cold to hot respectively, as shown by the color bar at bottom). The green marks show the BHs and are sized according to their masses.

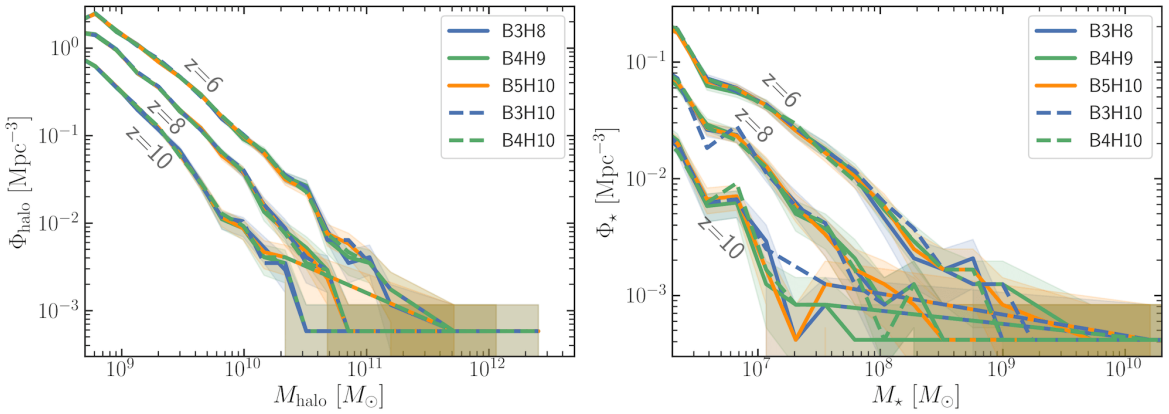

Figure 4.8: Mass functions of the simulations at $z = 6$, 8, and 10. **Left**: halo mass functions $\Phi_{\rm halo}$. **Right**: galaxy stellar mass functions $\Phi_\star$.

may form earlier in the first, molecular cooling halos which have smaller mass (see, e.g., Johnson & Bromm, 2007). Hence in this set, BH seeds with smaller masses (than the canonical $5 \times 10^5\, h^{-1} M_\odot$) are seeded at earlier times.

- In **Set B**: we fix the threshold halo mass at $M_{\rm fof}^{\rm seed} = 5 \times 10^{10}\, h^{-1} M_\odot$ and study the effect of changing the BH seed mass in a given fixed halo mass. All the BH seeds in this set are seeded at the same time and in the same halos but simply with different BH seed masses.

Table 4.2 summarizes the sets of simulations, the adopted naming and their respective BH and halo seeding parameters. We emphasize that all of the simulations have the same constrained initial condition. Also, in particular, B5H10 has the same BH seeding parameters as that of BlueTides with the canonical/reference choice of BH seed mass and halo mass.

To illustrate the results of our simulations, we start by showing the environments of the BHs at $z = 6$, 8, and 10 in Figure 4.7. In particular, we show the projected gas density field color-coded by the gas temperature in each of the simulations. Each panel is $6\, h^{-1}$Mpc per side with the most massive BH residing at the center. The green circles mark out all the BHs, where the size of the circles scales with the BH masses. The relatively hot region of gas around the BH results mostly from the effects of AGN feedback. Here we specifically show the results of the simulations in **Set A** (B3H8, B4H9, and B5H10). We note that the density field/environments of B3H10 and B4H10 in **Set B** will be similar to those shown for B5H10 except for the gas temperatures in the region around the central BH (which would typically be less affected by AGN feedback and associated heating, as we will discuss later).

Figure 4.7 highlights two major points. First, the effects of BH feedback, represented by the heated gas phase (reddish colors) are more prominent as the seed BH grows and as in the case for a single, larger BH seed. Second, the BH populations in the simulations B3H8, B4H9, and B5H10 are different. This is due to the adopted values for the threshold halo mass

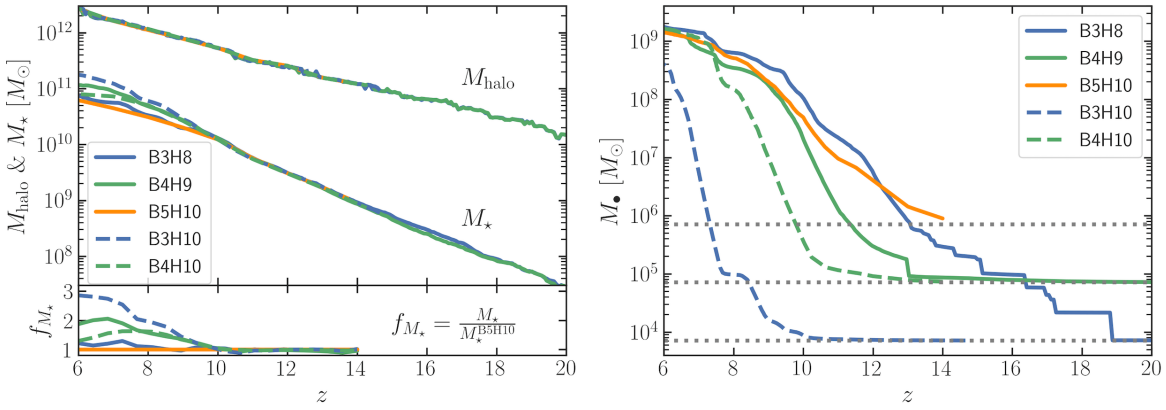

Figure 4.9: **Left-top**: the growth history of the host halo and galaxy ($M_{\rm halo}$ and $M_\star$) of the most massive BHs in the simulations. **Left-bottom**: the stellar mass ratio $f_{M_\star}$ between $M_\star$ in each simulation and in B5H10. **Right**: the growth history of the most massive BH ($M_\bullet$) in each simulation. The horizontal grey dotted lines show the BH seed masses.

in each of the respective simulations. In particular, the lower the threshold halo mass is, the more BH seeds are placed in a simulation, resulting in more BHs in B3H8 than in B5H10. Table 4.3 summarizes the numbers of BHs seeded and growing in each of the simulations.

Figure 4.8 shows the halo mass functions $\Phi_{\rm halo}$ and the stellar mass functions $\Phi_\star$ in all our simulations (both **Set A** and **Set B**). The consistency of the mass functions among the simulations suggests that the choice of BH seeding parameters does not affect the global halo and galaxy population in significant ways. In particular, the lower-mass end is virtually unaffected in both halo and stellar mass functions. The high-mass end of the stellar mass function shows some differences. As we shall show later, this is a reflection of the different SFR histories (for $z < 10$ in the different seed models which are modulated by different amounts of AGN feedback in different BH seed models.

In the following sections, we aim to explore in more detail the growth histories of most massive BHs and their hosts in the simulations including their masses and mass assembly rates. In particular, in Section 4.3.1 we will show results for the simulations with different BH seed masses and different halo thresholds, B3H8, B4H9, and B5H10 while in Section 4.3.2 the simulations with different BH seed masses at fixed halo mass, B3H10, B4H10, and B5H10.

4.3.1 Set A: different BH seed masses and halo mass thresholds

Here we describe the results from the simulations B3H8, B4H9, and B5H10. Those are the ones in which halo thresholds for BH seeding are adjusted such that the ratio $M_\bullet^{\rm seed}/M_{\rm fof}^{\rm seed}$ is fixed. The left-top panel of Figure 4.9 shows the growth history of host halo and galaxy ($M_{\rm halo}$ and $M_\star$) for the most massive BHs in the simulations. The halos show the same

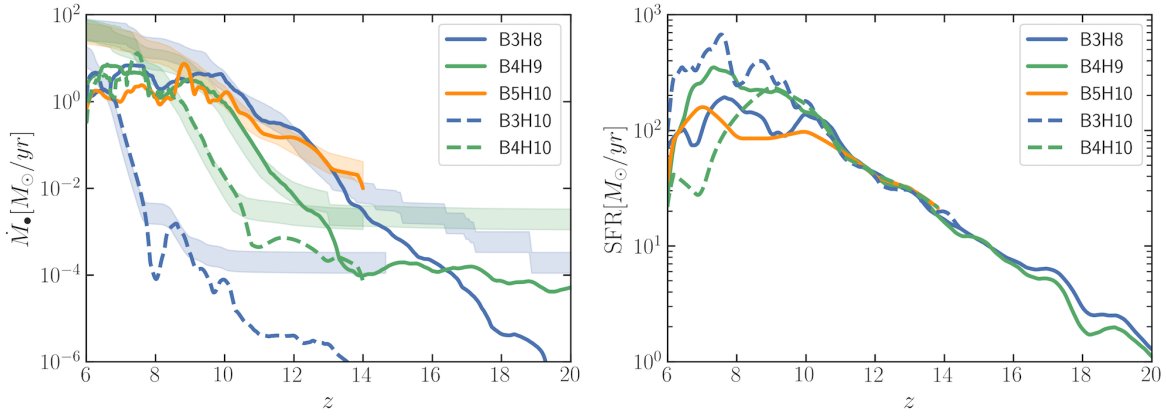

Figure 4.10: **Left**: the BH accretion rate ($\dot{M}_\bullet$) of the most massive BHs in the simulations. The shady regions show the Eddington rates of the BHs. **Right**: the star formation rate (SFR) of the most massive BHs in the simulations.

growth history among the simulations and their masses at $z = 6$ are $3 \times 10^{12}\, h^{-1} M_\odot$. The galaxy stellar mass also have a similar growth history except for B4H9 in which the host stellar mass ends up being larger than the others after $z = 8$ but less than a factor of two different at $z = 6$ according to the stellar mass ratio in the left-bottom panel. The stellar mass ratio $f_{M_\star} = \frac{M_\star}{M_\star^{\rm B5H10}}$ compares the galaxy mass in each of the simulation to $M_\star$ in B5H10. At $z = 6$, $M_\star = 10^{11}\, h^{-1} M_\odot$ for B3H8 and B5H10 and $M_\star = 6 \times 10^{10}\, h^{-1} M_\odot$ for B4H9. Our simulations suggest that the choice of BH seeding parameters does not affect the growth of the hosts for more than a factor of two.

Of our interest, the solid curves in the right panel of Figure 4.9 show the growth history of the most massive BHs in the simulations B3H8, B4H9, and B5H10. As apparent in the figure, the BHs are seeded with different masses according to $M_\bullet^{\rm seed}$ and at different times according to $M_{\rm fof}^{\rm seed}$; that is, a smaller BH seed emerges in a lower-mass halo at an earlier time than a corresponding higher-mass one. For example, when a BH seed of $5 \times 10^3\, h^{-1} M_\odot$ is placed in halos of mass $5 \times 10^8\, h^{-1} M_\odot$ it can be seeded at $z > 20$. As halos of this mass are not so rare at high redshifts, a lot of BH seeds emerge rather than just one seed does as in the case of $M_\bullet^{\rm seed} = 5 \times 10^5\, h^{-1} M_\odot$. The most massive BHs in the simulations start to converge in mass at $z \sim 8$ and reach a mass of $2 \times 10^9\, h^{-1} M_\odot$ by $z = 6$ even though the seed population and the total number of BHs are different. This suggests that the choice of different pairs of BH seed mass and threshold halo mass does not significantly affect the growth of the most massive BHs, at least in the expected range of $M_\bullet^{\rm seed} = 10^3 - 10^6\, M_\odot$. However, the early growth of SMBHs in the three simulations can be faster or slower at $z > 10$; that is, the small and large BH seeds start more massively while the intermediate one remains at its seed mass the longest but catch up drastically once it starts growing. Moreover, the discrete jumps in the growth history of B3H8 at early times indicate a lot of mergers occur even at such high

redshifts compared to the others.

Figure 4.10 shows the evolution of BH accretion rates $\dot{M}_\bullet$ of the most massive BHs in the simulations. The solid curves show the $\dot{M}_\bullet$ and the shaded bands indicate the regime between the Eddington rate and two times the Eddington rate as the upper limit for $\dot{M}_\bullet$ in the simulations. We find that $\dot{M}_\bullet$ has the same overall evolution: starting with an initial low accretion phase, followed by a close to exponential Eddington growth, and ending with a final quenched feedback-regulated phase. It is noticeable that $\dot{M}_\bullet$ gradually converges at $z > 10$, the time when the three SMBHs enter the feedback-regulated phase where their growth saturates and $M_\bullet$ starts converging. However, in the early phases during $10 < z < 13$, the BH accretion rates have different trajectories as the BHs experiencing exponential growth but constrained by the upper limit in the simulations.

The right panel of Figure 4.10 shows the star formation history of the host galaxies for the simulations. The evolution of the star formation rates (SFRs) appears similar at $z > 10$ but diverges in the later phase because the AGN feedback starts to regulate the star formation rate by coupling significant energy to the star forming gas. Therefore, $\dot{M}_\bullet$ and SFR start to couple with each other after a significant BH growth phase ($z < 10$) though the physical scales of the two quantities are quite different (SFR is in the galactic scale of tens of h^{-1}kpc, whereas $\dot{M}_\bullet$ is determined by local gas density at the scale of h^{-1}kpc).

As a further comparison, Feng et al. (2014) has examined the exact pairs of $M_\bullet^{\rm seed}$ and $M_{\rm fof}^{\rm seed}$ in **Set A** (see Table 4.2) using zoom-in simulations. With the zoom-in technique, they re-simulated a high-redshift ($z > 5.5$) halo hosting a $10^9\,M_\odot$ BH from the $\sim$ Gpc volume, MassiveBlack cosmological hydrodynamic simulation. They reported that regardless of the BH seed mass, the BH masses converged to $M_\bullet = 10^9\,M_\odot$ at $z = 6$, the BHs underwent a similar history of BH accretion, and the evolution of SFRs was the same at the earlier times before AGN feedback starts to regulate the SFR. It is interesting that both their findings and ours are fully consistent with completely different methods, the zoomed-in simulation and the constrained simulation. This strengthens the conclusion that the choice of BH seed mass - when keeping the ratio $M_\bullet^{\rm seed}/M_{\rm fof}^{\rm seed}$ fixed - does not affect the growth of SMBHs significantly.

4.3.2 Set B: different BH seed masses at fixed host halo mass

We then move on to investigate the growth history of host halo and galaxy ($M_{\rm halo}$ and $M_\star$) of the most massive BHs in the simulations B3H10, B4H10, and B5H10 in Figure 4.9. The halos show the same growth history among the simulations and their masses at $z = 6$ are $3 \times 10^{12}\,h^{-1}M_\odot$. On the other hand, the galaxies seem to have a very similar growth history at $z > 10$ but then their masses start to differ. According to the stellar mass ratio $f_{M_\star}$, the galaxy in B3H10 is three times more massive than the one in B5H10 at $z = 6$. This can be inferred through the evolution of SFR in Figure 4.10; that is, the galaxy in B3H10 undergoes a rather bursty star formation history at $z < 10$ with variations in SFR up to an order of

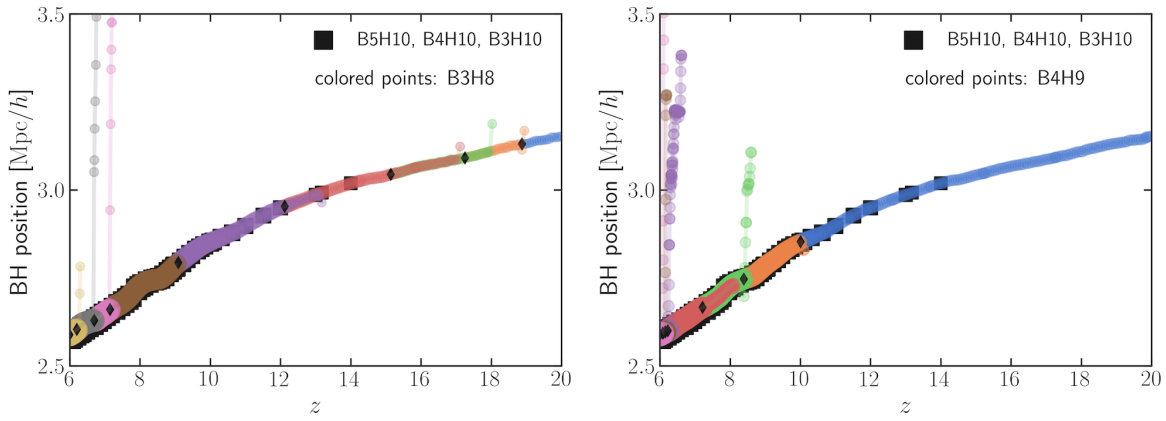

Figure 4.11: Positions of the most massive BHs in simulations B3H8 (left) and B4H9 (right) compared with B5H10 (and the others). The black diamonds mark the mergers the two most massive BHs experience. The size and color of the data points illustrate the mass of BHs and the ID of BH particles.

magnitude compared to the others. The reason is that the BH mass in B3H10 is smaller than the others and therefore the corresponding AGN feedback is not strong enough to bring sufficient suppression on the local star formation at $6 < z < 10$. Our simulations suggest that the choice of BH seed mass does not affect the growth of halo by $z = 6$; does not affect the growth of galaxy by $z = 10$; do affect the growth of galaxy after $z = 10$ but less than a factor of three by $z = 6$. There is another noticeable trend that a galaxy grows more when the BH seed mass is smaller.

The growth of the most massive BHs in Figure 4.9 starts with different masses and then converges to $\sim 10^9\, M_\odot$ by $z = 6$ except for the one in B3H10, which is less than an order of magnitude difference. The evolution of BH accretion rates of the BHs in Figure 4.10 shows that the BH in B3H10 is still experiencing Eddington exponential growth at $z < 8$. This indicating that the BH in B3H10 is still catching up in mass and will probably converge to the others at later times. Our simulation results hint that the growth of the most massive BHs will converge at the later times regardless of the choice of the SMBH seeding parameters in cosmological simulations.

4.3.3 BH-BH mergers

Figure 4.7 and Table 4.3 indicate that there is a major difference in the BH populations of the simulations B3H8, B4H9 and B5H10. Particularly in B3H8, there is a vast BH population in the environment of the most massive BH since we also adjust the minimum halo, implying that BH mergers are more likely to happen in the early times. The step-like feature in the BH mass assembly history in Figure 4.9 then infers that mergers occur. These pieces of evidence motivate the following investigation of BH mergers in the simulations. In MP-GADGET, a BH-

BH merger occurs when the distance between two BHs is smaller than their SPH smoothing kernel and the relative velocity of the two BHs is smaller than $\frac{1}{2}\,c_{\rm s}$, where $c_{\rm s}$ is the local sound speed of the gas. Convection is also applied to the BH dynamics by repositioning the BH particle to the local minimum potential at every time step.

To investigate the BH merger history of the simulations B3H8, B4H9, and B5H10 (as well as B3H10 and B4H10), we plot the merger tree of BH projected positions in Figure 4.11. The left and right panels show the position of the most massive BHs in B3H8 and B4H9 respectively compared with B5H10. The size of the data points scales with the BH masses; the data points are color-coded by the ID of BH particles; the black diamonds mark where and when the BH mergers occur. By $z = 6$, there are eight and six BH mergers that happen in B3H8 and B4H9 respectively whereas there is no merger in B5H10 despite mergers likely to occur below this redshift as the closest BH below 100 kpc distance from the most massive one. Besides, B3H10 and B4H10 have no merger as well as B5H10 since they contain the same number and position of BHs but with smaller BH masses.

At $z > 12$, interestingly, four mergers happen in B3H8 whereas none in the others. This explains why the SMBH in B3H8 grows faster than the others during the earlier phase and further implies that mergers dominate the early growth of SMBHs in small BH seeding scenarios. Our simulations suggest that if less-massive BH seeds are more common, SMBHs can still grow via mergers at the early times even though they may be expected to grow slower. In other words, a different BH merger history results in a different growth of the SMBHs particularly at the early times. Despite the fact that the high halo occupation fraction of SMBHs in cosmological simulations will increase the number of BH-BH mergers since these simulations are implemented with simple merger models without considering the BH dynamics that could make BH-BH mergers more difficult between low-mass BHs, it is still interesting that different seeding scenarios are expected to produce different BH populations and associated merger rates that can discriminate the different scenarios at early times while the final BH mass converges to a similar value.

In contrast, at $6 < z < 10$, SMBHs seeded with a mass of $10^3 - 10^4\,M_\odot$ undergo a few BH mergers whereas the one in the largest seed models ($\sim 10^5\,M_\odot$) does not experience any merger until $z < 6$. These different predictions of the merger history for the first massive BHs constitute an interesting prospect for constraining BH seed masses or models for the first quasars that will become within reach with the planned *LISA* mission (Amaro-Seoane et al., 2017).

4.4 Conclusion

In this paper, we have investigated new constrained cosmological simulations designed to reproduce the environments and large-scale structures relevant for the growth of the first quasars at $z \geq 6$. In particular, we have focused on the effects of different choices of BH

seeding scenarios (different parameters in the SMBH sub-grid model) on the growth of SMBHs at the early times. Employing the technique of constrained Gaussian realizations (Hoffman & Ribak, 1991; van de Weygaert & Bertschinger, 1996), we have reconstructed the initial conditions to reproduce the large-scale structure and the local environment of the most massive BH in the BlueTides simulation. BlueTides has been the only cosmological hydrodynamic simulation that directly predicted the rare-observed first quasars (Di Matteo et al., 2017; Ni et al., 2018; Tenneti et al., 2018) thanks to its sufficiently high resolution and large volume. The first quasars are extremely rare such that there were only four SMBHs with mass $\sim 10^9\,M_\odot$ by $z = 7$ in BlueTides with $L_{\rm box} = 400\,h^{-1}$Mpc.

We have compared the new constrained simulations with the BlueTides simulation to validate this method by running the constrained initial conditions forward in time until $z = 6$. Our new simulations in boxes of $15\,h^{-1}$Mpc on a side have successfully recovered the evolution of the large-scale structure, the mass functions, as well as the growth history of the most massive BHs and their hosts at the high redshifts of interests. At $z = 8$, the most massive BH and its hosts had a halo mass of $\sim 10^{12}\,M_\odot$; a stellar mass of $\sim 4\times10^{10}\,M_\odot$; a BH mass of $\sim 4\times10^{8}\,M_\odot$. This is consistent with BlueTides within a factor of 1.5 in mass while keeping the resolution. More importantly, the demand on computational resources has decreased significantly by a factor of $(400/15)^3 \sim 20000$.

By running a set of different realizations such that each of them has a different local tidal field, we have further shown that a low-tidal field environment is crucial for the growth of the earliest and most massive SMBHs. This is consistent with the finding from BlueTides in Di Matteo et al. (2017). For our highest tidal field realization, the mass of the most massive BH was only $\sim 2\times10^7\,M_\odot$ at $z = 7$ which was two orders of magnitude lower than the SMBH in the lowest tidal field realization. Among the simulations, the SMBH in the lowest tidal field environment had a mass an order of magnitude more than the one in the highest tidal field environment at $z = 6$.

After selecting the initial conditions that best recovered the original quasar environment in BlueTides, we have run other simulations to investigate the effects of the choice of BH seeding parameters on the growth of these first massive objects. In BlueTides simulation, the BH seed mass has been chosen to be $5\times10^5\,h^{-1}M_\odot$, which is at the high end of the predicted mass for SMBH seeds in theories. With the same constrained initial conditions, we have conducted two sets of simulations with different SMBH seed masses $M_\bullet^{\rm seed} = 5\times10^3$, 5×10^4, and $5\times10^5\,h^{-1}M_\odot$. Set A with different threshold halo masses such that the ratio $M_\bullet^{\rm seed}/M_{\rm fof}^{\rm seed}$ is fixed, while set B has a fixed halo threshold. Our simulations have suggested that the final mass of the SMBH is insensitive to the initial seed mass regardless of the choice of BH seeding parameters; the mass of SMBH in our constrained simulations has converged to $\sim 10^9\,M_\odot$ at $z = 6$. In the early times at $z > 10$, the growth of SMBHs varies among the simulations with different seeding scenarios; less massive seed models tend to grow slower initially unless they are seeded in more common but less massive halos so that they can merge

frequently. A significant fraction of the early growth occurs in this mode in a low mass seed scenario in set A, effectively allowing the SMBH growth to catch up with that of a more massive seed. There were four SMBH mergers at $z \gtrsim 12$ for the most massive SMBH with the lowest seed mass while no mergers happened for the other two runs, suggesting that the smallest seed grows faster at earlier times when seeded in less massive halos.

The significant differences in the early merger rates provide an interesting discriminating feature for small versus large BH seed models at the early time. The space-based gravitational wave telescope *LISA* will open up new investigations into the dynamical processes involving SMBHs and new exciting prospects for tracing the origin, merger history of SMBHs across cosmic ages.

Acknowledgements

We acknowledge the funding from NSF ACI-1614853, NSF AST-1517593, NSF AST-1616168, NASA ATP 80NSSC18K1015 and NASA ATP 17-0123. The BLUETIDES simulation is run on the BlueWaters facility at the National Center for Supercomputing Applications. Some of the simulations in this work are carried out on the Stampede2 supercomputing cluster. The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu.

This work has made use of the following software as well: NUMPY (van der Walt et al., 2011), SCIPY (Jones et al., 2001), MATPLOTLIB (Hunter, 2007), SEABORN (Waskom et al., 2016), GAEPSI2 (Feng, 2018), NBODYKIT (Hand & Feng, 2015).

5 BlueTides simulation: establishing black hole-galaxy relations at high-redshift

Kuan-Wei Huang[1], Tiziana Di Matteo,[1] Aklant K. Bhowmick,[1] Yu Feng[2], and Chung-Pei Ma[3]

[1] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2] Berkeley Center for Cosmological Physics, University of California at Berkeley, Berkeley, CA, 94720, USA
[3] Department of Astronomy, University of California at Berkeley, Berkeley, CA, 94720, USA

Abstract

The scaling relations between the mass of supermassive black holes ($M_\bullet$) and host galaxy properties (stellar mass, $M_\star$, and velocity dispersion, σ), provide a link between the growth of black holes (BHs) and that of their hosts. Here we investigate if and how the BH-galaxy relations are established in the high-z universe using BlueTides, a high-resolution large volume cosmological hydrodynamic simulation. We find the $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations at $z = 8$: $\log_{10}(M_\bullet) = 8.25 + 1.10\ \log_{10}(M_\star/10^{11}M_\odot)$ and $\log_{10}(M_\bullet) = 8.35 + 5.31\ \log_{10}(\sigma/200\mathrm{kms}^{-1})$ at $z = 8$, both fully consistent with the local measurements. The slope of the $M_\bullet - \sigma$ relation is slightly steeper for high star formation rate and $M_\star$ galaxies while it remains unchanged as a function of Eddington accretion rate onto the BH. The intrinsic scatter in $M_\bullet - \sigma$ relation in all cases ($\epsilon \sim 0.4$) is larger at these redshifts than inferred from observations and larger than in $M_\bullet - M_\star$ relation ($\epsilon \sim 0.14$). We find the gas-to-stellar ratio $f = M_{\rm gas}/M_\star$ in the host (which can be very high at these redshifts) to have the most significant impact setting the intrinsic scatter of $M_\bullet - \sigma$. The scatter is significantly reduced when galaxies with high gas fractions ($\epsilon = 0.28$ as $f < 10$) are excluded (making the sample more comparable to low-z galaxies); these systems have the largest star formation rates and black hole accretion rates, indicating that these fast-growing systems are still moving toward the relation at these high redshifts. Examining the evolution (from $z = 10$ to 8) of high mass black holes in $M_\bullet - \sigma$ plane confirms this trend.

5.1 Introduction

Over the last three decades, scaling relations between mass of supermassive black holes (SMBHs) and several stellar properties of their host galaxies such as bulge stellar mass and bulge velocity dispersion (Magorrian et al., 1998; Häring & Rix, 2004; Gebhardt et al., 2000; Tremaine et al., 2002; Gültekin et al., 2009; Kormendy & Ho, 2013; McConnell & Ma, 2013; Reines & Volonteri, 2015) have been discovered and measured for galaxies with black holes (BHs) and active galactic nuclei (AGN) from $z = 0$ and up to $z \sim 2$ (using different techniques).

Many theoretical models have been developed to understand the origin of these relations. Several cosmological simulations that follow the formation, growth of BHs and their host galaxies have successfully reproduced the scaling relations at low-z; these include recent simulations such as the *Illustris* simulation (Vogelsberger et al., 2014; Sijacki et al., 2015), the Magneticum Pathfinder SPH simulation (Steinborn et al., 2015), the Evolution and Assembly of GaLaxies and their Environment (*EAGLE*) suite of SPH simulation (Schaye et al., 2015), and the *MassiveBlackII (MBII)* simulation (Khandai et al., 2015; DeGraf et al., 2015). Thus, the scaling relations from observations and simulations agree with each other at low-z, linking the growth of SMBHs to the growth of their hosts via AGN feedback.

A popular way to interpret the scaling relations is by invoking AGN feedback. Many models (and simulations) show that the SMBHs regulate their own growth with their hosts by coupling a fraction of their released energy back to the surrounding gas (Di Matteo et al., 2005). The BHs grow only until sufficient energy is released to unbind the gas from the local galaxy potential (Silk & Rees, 1998; King, 2003; Springel, 2005; Bower et al., 2006; Croton et al., 2006; Di Matteo et al., 2008; Ciotti et al., 2009; Fanidakis et al., 2011). However, there are also models which have been proposed to explain the scaling relations without invoking the foregoing coupled feedback mechanism. For instance, it has also been shown that dry mergers can potentially drive BHs and their hosts towards a mean relation (Peng, 2007; Hirschmann et al., 2010; Jahnke & Macciò, 2011) and that BH growth regulated by gravitational torques can also explain the relations (Anglés-Alcázar et al., 2015, 2017). Regardless, studying the scaling relations in both observation and simulation is essential for understanding the coupled growth of galaxies and BHs across cosmic history.

An important related question is when the scaling relations are established, and if they still persist at higher redshifts when the first massive BHs form ($z > 6$). To understand this, galaxies with AGN play a key role in observations (Bennert et al., 2010; Merloni et al., 2010; Kormendy & Ho, 2013). A strong direct constraint on the high-redshift evolution of SMBHs comes from the luminous quasars at $z \sim 6$ in SDSS (Fan et al., 2006; Jiang et al., 2009; Mortlock et al., 2011). Most recently, the earliest quasar is discovered at $z = 7.5$ (Bañados et al., 2017) in ALLWISE, UKIDSS, and DECaLS. However, it is still not established as to whether these objects follow the local BH-galaxy relations and whether there is a redshift

evolution, because of the systematic uncertainties (Woo et al., 2006) and selection effects (Lauer et al., 2007; Treu et al., 2007; Schulze & Wisotzki, 2011, 2014).

At high-z, AGN is our only proxy for studying the BH mass assembly. The luminosity functions (LFs) of AGN however remain uncertain. For example, BH mass function at $z = 6$ has been inferred from optical AGN LFs in Willott et al. (2010). On the other hand, several works (Wang et al., 2010; Volonteri & Stark, 2011; Fiore et al., 2012; Volonteri & Reines, 2016) have argued that there are large populations of obscured, accreting BHs at high-z. For instance, the BH mass density at $z = 6$ from X-ray observations of AGN has been shown to be greater than that inferred from optical quasars by an order of magnitude or more (Treister et al., 2011; Willott, 2011). A luminosity dependent correction for the obscured fraction is proposed in Ueda et al. (2014) and the obscured fraction tends to increase with redshift up to $z \sim 4$ (Merloni et al., 2014; Vito et al., 2014; Buchner et al., 2015; Vito et al., 2018). Regardless of the exact amount of the obscured AGNs, it is certain that a fraction of AGNs is obscured, and therefore missed by observations. As a result, quantities such as the BH mass function, BH mass density, and BH accretion rate density are still uncertain at high-z.

Here, we use the BlueTides simulation (Feng et al., 2015) to make predictions for both the global BH mass properties and the scaling relations ($M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations) from $z = 8$ to $z = 10$. BlueTides is a large-scale and high-resolution cosmological hydrodynamic simulation with 2×7040^3 particles in a box of $400h^{-1}$Mpc on a side, which includes improved prescriptions for star formation, BH accretion, and associated feedback processes. With such high resolution and large volume, we are able to study the scaling relations and the global properties of BH mass at high-z for the first time. So far, various quantities measured in BlueTides have been shown to be in good agreement with all current observational constraints in the high-z universe such as UV luminosity functions (Feng et al., 2016a; Waters et al., 2016a,b; Wilkins et al., 2017), the first galaxies and the most massive quasars (Feng et al., 2015; Di Matteo et al., 2017; Tenneti et al., 2018), the Lyman continuum photon production efficiency (Wilkins et al., 2016, 2017), galaxy stellar mass functions (Wilkins et al., 2018), and angular clustering amplitude (Bhowmick et al., 2017).

The paper is organized as follows. In Section 5.2, we briefly describe the BlueTides simulation and several physics implementations. In Section 5.3, we report BH mass properties: mass function, mass density, and accretion rate. In Section 5.4, we demonstrate the scaling relations between $M_\bullet$ and $M_\star$, and σ. In Section 5.5, we study the selection effects for several galaxy properties on the scaling relations. In Section 5.6, we investigate the assembly history of how BHs evolove on $M_\bullet - M_\star$ and $M_\bullet - \sigma$ planes. In Section 5.7, we summarize the conclusion of the paper.

Table 5.1: Numerical parameters for the BlueTides simulation.

h	0.697	Boxsize	$400h^{-1}$Mpc
Ω_Λ	0.7186	$N_{\rm particle}$	2×7040^3
$\Omega_{\rm matter}$	0.2814	$M_{\rm DM}$	$1.2 \times 10^7 h^{-1} M_\odot$
$\Omega_{\rm baryon}$	0.0464	$M_{\rm gas}$	$2.36 \times 10^6 h^{-1} M_\odot$
σ_8	0.820	ϵ	$1.5h^{-1}$kpc
$n_{\rm s}$	0.971	$M_{\bullet \rm seed}$	$5 \times 10^5 h^{-1} M_\odot$

5.2 Methods

5.2.1 BlueTides hydrodynamic simulation

The BlueTides simulation has been carried out using the Smoothed Particle Hydrodynamics code MP-Gadget on the Blue Waters system at the National Center for Supercomputing Applications. The hydrodynamics solver in MP-Gadget adopts the new pressure-entropy formulation of smoothed particle hydrodynamics (Hopkins, 2013). This formulation avoids non-physical surface tensions across density discontinuities. BlueTides contains 2×7040^3 particles in a cube of $400h^{-1}$Mpc on a side with a gravitational smoothing length $\epsilon = 1.5h^{-1}$kpc. The dark matter and gas particles masses are $M_{\rm DM} = 1.2 \times 10^7 h^{-1} M_\odot$ and $M_{\rm gas} = 2.36 \times 10^6 h^{-1} M_\odot$, respectively. The cosmological parameters used were based on the Wilkinson Microwave Anisotropy Probe nine years data (Hinshaw et al., 2013) (see Table 5.1 for a brief summary of the parameters). With an unprecedented volume and resolution, BlueTides runs from $z = 99$ to $z = 8$. BlueTides contains approximately 200 million star-forming galaxies, 160000 of which have stellar mass $> 10^8 M_\odot$, and 50 thouthand BHs, 14000 of which have BH mass $> 10^6 M_\odot$ (the most massive BH's mass $\sim 4 \times 10^8 M_\odot$). A full description of BlueTides simulation can be found in Feng et al. (2016a).

5.2.2 Sub-grid physics and BH model

A number of physical processes are modeled via sub-grid prescriptions for galaxy formation in BlueTides. Below we list the main features of the sub-grid models:

- Star formation based on a multiphase star formation model (Springel & Hernquist, 2003) with modifications following Vogelsberger et al. (2013).
- Gas cooling through radiative processes (Katz et al., 1996) and metal cooling (Vogelsberger et al., 2014).
- Formation of molecular hydrogen and its effects on star formation (Krumholz & Gnedin, 2011).
- Type II supernovae wind feedback (the model used in *Illustris* (Nelson et al., 2015)).

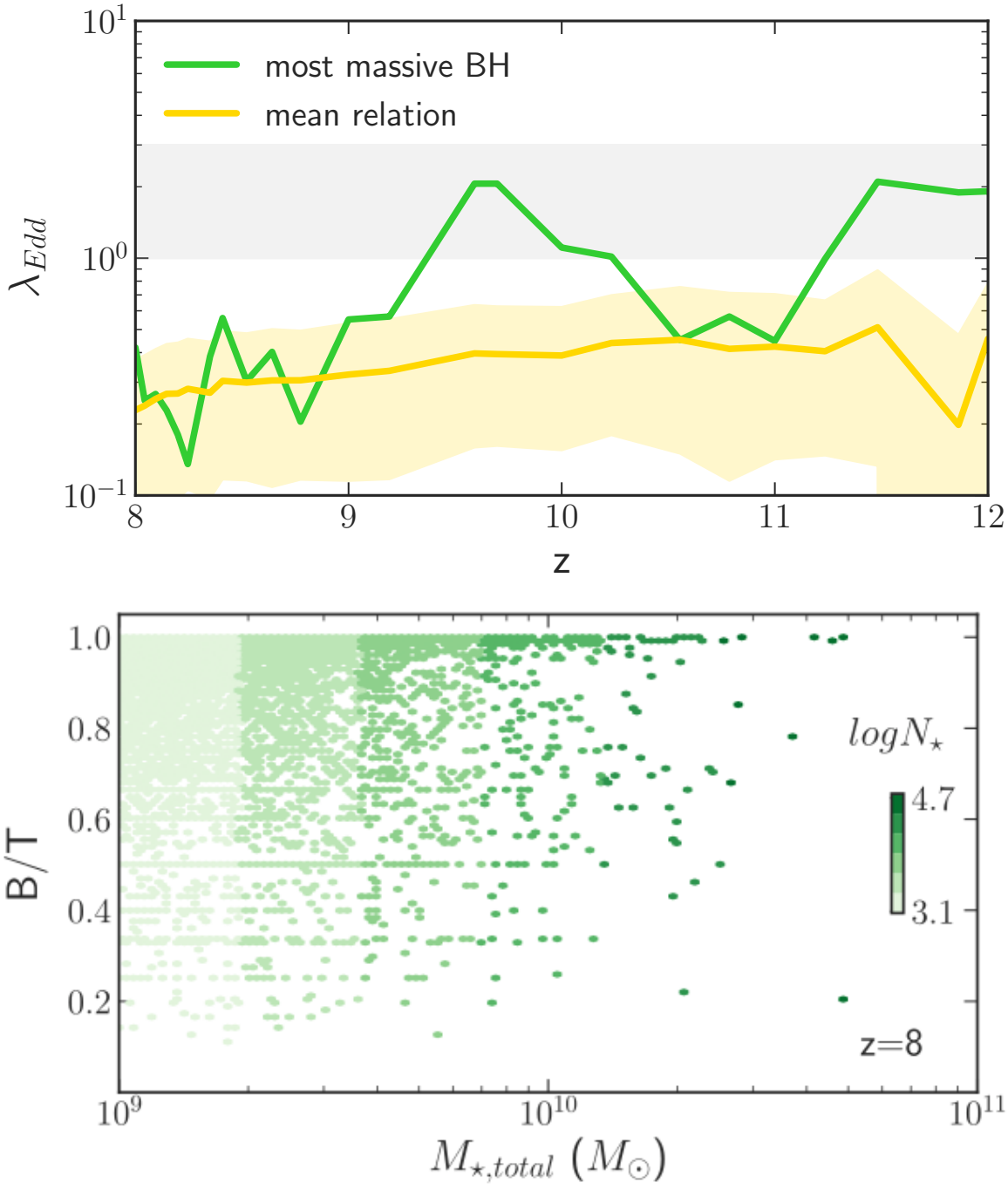

Figure 5.1: Top panel: the evolution of $\lambda_{\rm Edd}$ in BLUETIDES: green curve for the most massive BH, yellow curve for the mean relation shaded with one standard deviation, and grey shade for $1 < \lambda_{\rm Edd} < 3$. Bottom panel: the relation between B/T and total stellar mass ($M_\star$) color coded according to number of star particles for each galaxy at $z = 8$ in BLUETIDES.

- A model of 'patchy' reionization (Battaglia et al., 2013) yielding a mean reionization redshift $z \sim 10$ (Hinshaw et al., 2013), and incorporating the UV background estimated by Faucher-Giguère et al. (2009);
- Black growth and AGN feedback. BHs grow in mass by gas accretion and by merging with other BHs.

We model BH growth and AGN feedback in the same way as in the *MassiveBlack I & II* simulations, using the SMBH model developed in Di Matteo et al. (2005) with modifications consistent with *Illustris*. BHs are seeded with an initial seed mass of $M_{\bullet \rm seed} = 5 \times 10^5 h^{-1} M_\odot$ (commensurate with the resolution of the simulation) in halos more massive than $5 \times 10^{10} h^{-1} M_\odot$ while their feedback energy is deposited in a sphere of twice the radius of the SPH smoothing kernel of the black hole. Gas accretion proceeds via $\dot{M}_\bullet = \frac{4\pi\alpha G^2 M_\bullet^2 \rho}{(c_s^2+v^2)^{-3/2}}$ according to Hoyle & Lyttleton (1939); Bondi & Hoyle (1944); Bondi (1952), where ρ and c_s are the density and sound speed of the gas respectively, α is a dimensionless parameter, and v is the velocity of the BH relative to the gas. We allow for super-Eddington accretion but limit the accretion rate to three times of the Eddington rate: $\dot{M}_{\rm Edd} = \frac{4\pi G M_\bullet m_{\rm p}}{\eta \sigma_{\rm T} c}$, where $m_{\rm p}$ is the proton mass, $\sigma_{\rm T}$ is the Thomson cross-section, and η is the radiative efficiency. The Eddington ratio is defined as $\lambda_{\rm Edd} = \frac{\dot{M}_\bullet}{\dot{M}_{\rm Edd}}$. In the top panel of Figure 5.1, we show the evolution of $\lambda_{\rm Edd}$ in BlueTides. Although we allow super-Eddington accretion to $\lambda_{\rm Edd} < 3$ (grey area), the BHs do not grow that fast: on average, the BHs accrete with $\lambda_{\rm Edd} < 1$ (yellow curve) and even the most massive one (green curve) spends 39% and 7% of time growing with $\lambda_{\rm Edd} > 1$ and $\lambda_{\rm Edd} > 2$. A similar plot of a few most massive BHs in BlueTides can be found in Figure 5 of Di Matteo et al. (2017), it shows that there are only 4 BHs that grow (for not a large fraction of the time) at the critical rate. This is consistent with the incidence of 'first quasar', only roughly 1 object per Gpc^3. Thus, all the interesting super-Eddington models (Madau et al., 2014; Lupi et al., 2016; Pezzulli et al., 2016, 2017; Jiang et al., 2017) do not appear to be relevant for the overall BH population at $z > 8$ in BlueTides. For this reason, we fix radiative efficiency at the average value of 0.1 according to Shakura & Sunyaev (1973) throughout the simulation. Note also that this is consistent with all other cosmological simulations work that follow BH growth across cosmic history and in particular with our *MassiveBlack II* simulation (DeGraf et al., 2015), which was run to $z = 0$ shows direct agreement with $M_\bullet - \sigma$ relation local measurements. Also, BH is assumed to radiate with a bolometric luminosity ($L_{\rm B}$) proportional to the accretion rate ($\dot{M}_\bullet$) by $L_{\rm B} = \eta \dot{M}_\bullet c^2$.

5.2.3 Kinematic decomposition

As σ or $M_\star$ in observational studies of $M_\bullet - \sigma$ or $M_\bullet - M_\star$ relations are often measured from the bulge component of galaxies, we perform a kinematic decomposition for the stellar particles of the galaxies in BlueTides as in Feng et al. (2015). This allows us to determine which stars are on planar circular orbits and which are associated with a bulge, in each

galaxy (Vogelsberger et al., 2014; Tenneti et al., 2016), providing kinematically classified disks and bulges, and a disk to total (D/T) ratio for our galaxies. We perform this analysis following Abadi et al. (2003a): a circularity parameter is defined for every star particle as $\kappa = j_z/j\,(E)$, where j_z is the specific angular momentum around a selected z-axis and $j\,(E)$ is the possible maximum specific angular momentum of the star with the specific binding energy E. The star particle with $\kappa > 0.7$ is identified as a disk component according to Vogelsberger et al. (2014) and Tenneti et al. (2016). Thus, the D/T ratio for the stellar component of each galaxy is obtained, allowing us to calculate the bulge stellar mass and the bulge velocity dispersion for our galaxies. In the bottom panel of Figure 5.1, we show the relation between B/T = 1 − D/T (bulge to total ratio) and total stellar mass ($M_\star$) color coded according to number of star particles for each galaxy. According to the standard assumption D/T < 0.3 is considered a bulge dominated galaxies (Feng et al., 2015; Tenneti et al., 2016). For galaxies with $M_\star > 10^9 M_\odot$, the number of star particles is higher than 1000. We require this minimum number of star particles to have a reliable kinematic decomposion. For this reason, for the rest of the analysis we will only consider objects with $M_\star > 10^9 M_\odot$.

5.3 The global property of BH mass

We begin by investigating the global properties of BH mass ($M_\bullet$) at $z = 8 \sim 12$ in BlueTides. We choose a BH population with $M_\bullet > 1.5 \times 10^6 M_\odot$ which is roughly twice the BH seed mass ($M_{\bullet\mathrm{seed}} = 7.2 \times 10^5 M_\odot$), in order to minimize any possible influence of the seeding prescription on our analysis.

5.3.1 BH mass function and bolometric luminosity

We first look at the bolometric luminosity (L_B) of BH population in BlueTides at $z = 8$ in the left panel in Figure 5.2. The brown dashed and dotted lines are X-ray luminosity $L_\mathrm{X} = 10^{42.5}$ and 10^{43}erg/s, which are calculated by the bolometric correction in Marconi et al. (2004). These two values will be used as thresholds when studying other global properties of BH mass in this section. Statistically, there are more than 76 and 16 percent of our BHs with $L_\mathrm{X} > 10^{42.5}$ and 10^{43}erg/s respectively. This indicates that the global quantities of BH mass is sensitive to L_X when measured from X-ray survey in observation. In addition, L_B with the mean Eddington ratio in our BH population ($\overline{\lambda}_\mathrm{Edd} = 0.3$) is shown (the green solid line).

The right panel in Figure 5.2 shows BH mass functions (BHMFs) in BlueTides from $z = 8$ to $z = 12$ (the solid curves), as well as the ones with thresholds $L_\mathrm{X} > 10^{42.5}$ and 10^{43}erg/s (the dashed and dotted curves respectively). We also show the BHMFs inferred from optical quasars at $z = 6$ in Willott et al. (2010) (W10 hereafter; the purple dotted curve) and the theoretical prediction combined with observed Lyman break galaxy population (Stark et al., 2009; Volonteri & Stark, 2011) (the red squares). The slope of BHMFs in BlueTides

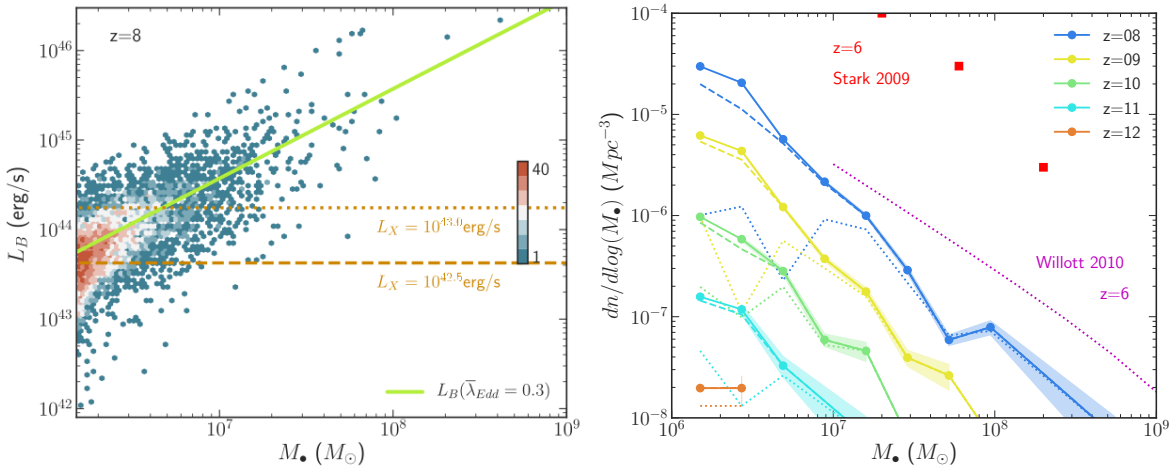

Figure 5.2: Left panel: the relation between BH bolometric luminosity ($L_{\rm B}$) and BH mass ($M_\bullet$) at $z = 8$ in BLUETIDES color coded according to the number of galaxies. The green line shows $L_{\rm B}$ with the mean Eddington ratio of our all BH population ($L_{\bar{\lambda}_{\rm Edd}=0.3}$). The brown dashed and dotted lines show the X-ray luminosity $L_X = 10^{42.5}$ and 10^{43}erg/s respectively according to the bolometric correction from Marconi et al. (2004). Right panel: BH mass functions in BLUETIDES at $z = 8 \sim 12$ (the solid curves). The dashed and dotted curves show the BH mass functions with thresholds of $L_{\rm X} = 10^{42.5}$ and 10^{43}erg/s respectively at the corresponding redshift. Also, the results at $z = 6$ in Willott et al. (2010) and Volonteri & Stark (2011) are shown.

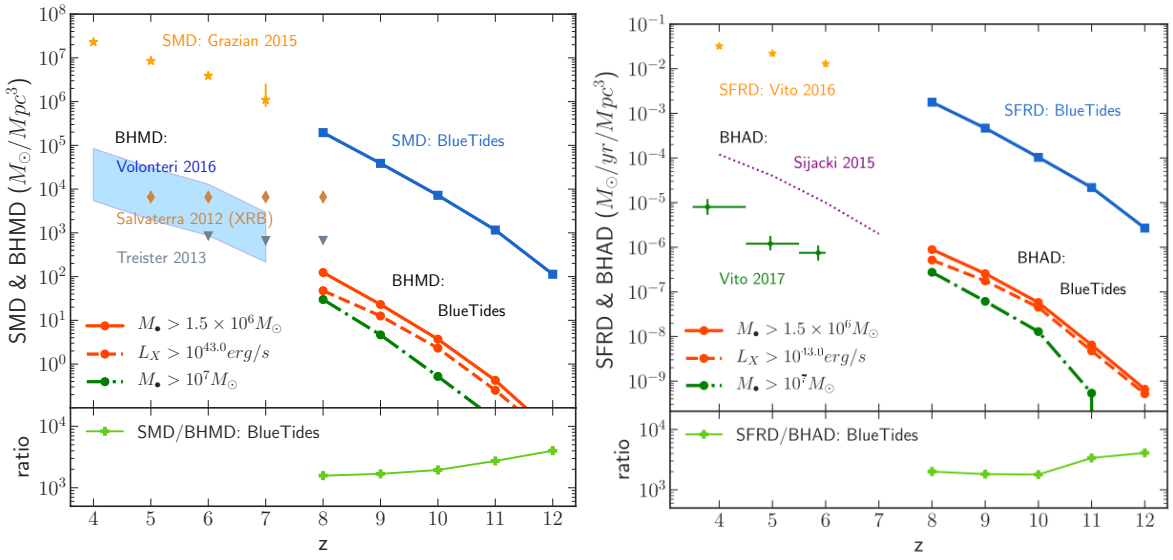

Figure 5.3: Left panel: the stellar mass and BH mass density (SMD and BHMD) inBLUETIDESand their ratio SMD/BHMD (the green solid curve). For SMD, galaxies with $M_\star > 10^8 M_\odot$ are selected in simulation (the blue solid curve) and observation (Grazian et al. (2015); the orange stars). For BHMD, galaxies with $M_\bullet > 1.5 \times 10^6 M_\odot$ (double of $M_{\bullet\mathrm{seed}}$), with $M_\bullet > 10^7 M_\odot$, and with $L_X > 10^{43}$erg/s are shown (the red solid, olive dash-dotted, and red dash curves respectively). The blue shaded area is the result in Volonteri & Reines (2016) and the brown diamonds and gray triangles are current upper limits from X-ray observations (Salvaterra et al., 2012; Treister et al., 2013). Right panel: the SFR and BH accretion rate density (SFRD and BHAD) in BLUETIDES and their ratio SFRD/BHAD (the green curve). The same thresholds are used as the left panel. Observational results in Vito et al. (2016) and Vito et al. (2018) for SFRD and BHAD are shown (the orange stars and green dots respectively), as well as the simulation prediction from Sijacki et al. (2015) (purple dotted curve; thresholds: $M_{\bullet\mathrm{seed}} > 10^5 h^{-1} M_\odot$).

are generally steeper than the one in W10 but similar to theoretical predictions (see Volonteri & Stark (2011) for more detail and comparison), particularly at the low mass end (which is currently unconstrained at these redshifts). In addition, the normalization of the BHMFs in BLUETIDES suggests that there is a larger BH population than those which have been observed from optical quasars, consistent with the claim that there is a large population of obscured accreting BHs at high-z.

5.3.2 BH mass density and stellar mass density

The left panel in Figure 5.3 shows BH and galaxy stellar mass density in BLUETIDES from $z = 8$ to $z = 12$ with observations from $z = 4$ to $z = 7$. For stellar mass density (SMD),

BlueTides (the solid blue curve) agrees with the trend from Grazian et al. (2015) (the orange stars). Galaxies with $M_\star > 10^8 M_\odot$ are selected for both cases. For BH mass density (BHMD), we report the results with two different $M_\bullet$ thresholds: $M_\bullet > 1.5 \times 10^6 M_\odot$ (double of $M_{\bullet\mathrm{seed}}$ in BlueTides) and $M_\bullet > 10^7 M_\odot$) and with $L_X > 10^{43}$erg/s (the red solid, olive dash-dotted, and red dash curves respectively) to show the influence from $M_\bullet$ and L_X thresholds. Current upper limits from X-ray observation are also presented: Salvaterra et al. (2012) (cosmic X-ray background (XRB)) and Treister et al. (2013) (the brown diamonds and gray triangles respectively). BHMD in BlueTides complies with those upper limits (at least at $z = 8$), pointing that our BH mass function is just steeper than the one in W10 but still within the upper limits from X-ray observation. In addition, results in Volonteri & Reines (2016) (the blue shade) are also included to support our BHMF, arguing that the integrated BH density depends on the $M_\bullet - M_\star$ relation.

It has been discussed that the stellar mass density exceeds the BH mass density roughly by a factor of 10^3 for low-z. To understand the ratio of these two quantities at higher redshifts, we show the ratio by normalizing SMD to BHMD in the left panel in Figure 5.3 (the green solid curve). Overall, SMD grows more rapidly than the BHMD at early times. Parameterizing the ratio by an evolutionary factor $(1+z)^\alpha$, we find that SMD/BHMD $= 1.3(1+z)^{3.1}$.

5.3.3 BH accretion rate density and SFR density

After BHMD and SMD, we investigate their assembly rate: the BH accretion rate density and the SFR density. The right panel in Figure 5.3 shows the BH accretion rate density (BHAD) again with $M_\bullet$ and L_X thresholds $M_\bullet > 1.5 \times 10^6 M_\odot$, $M_\bullet > 10^7 M_\odot$, and $L_X > 10^{43}$erg/s (the red solid, olive dash-dotted, and red dash curves respectively) and the SFR density (SFRD; the blue solid curve) in BlueTides from $z = 8$ to $z = 12$. For SFRD, we show observational result in Vito et al. (2016) (the orange stars), and on the other hand for BHAD, we not only show results from observation (Vito et al. (2018); the green dots) but also from simulation (Sijacki et al. (2015); the dotted purple curve) because it has been noticed that the prediction from simulation tends to be higher than current observation by more than an order of magnitude. It is possibly due to the difficulty of observing AGNs from deep X-ray surveys.

Similar to Section 5.3.2, we compare these two quantities by their ratio via normalizing SFRD to BHAD (the green curve). The ratio of the SFRD and the BHAD increases as z increases and the order of which (ranging from 10^3 to 10^4) is close to the order of the ratio of the BHMD and the SMD in our simulation. Again, we fit the ratio as a function of $(1+z)^\alpha$: SFRD/BHARD $= 3.4(1+z)^{2.7}$.

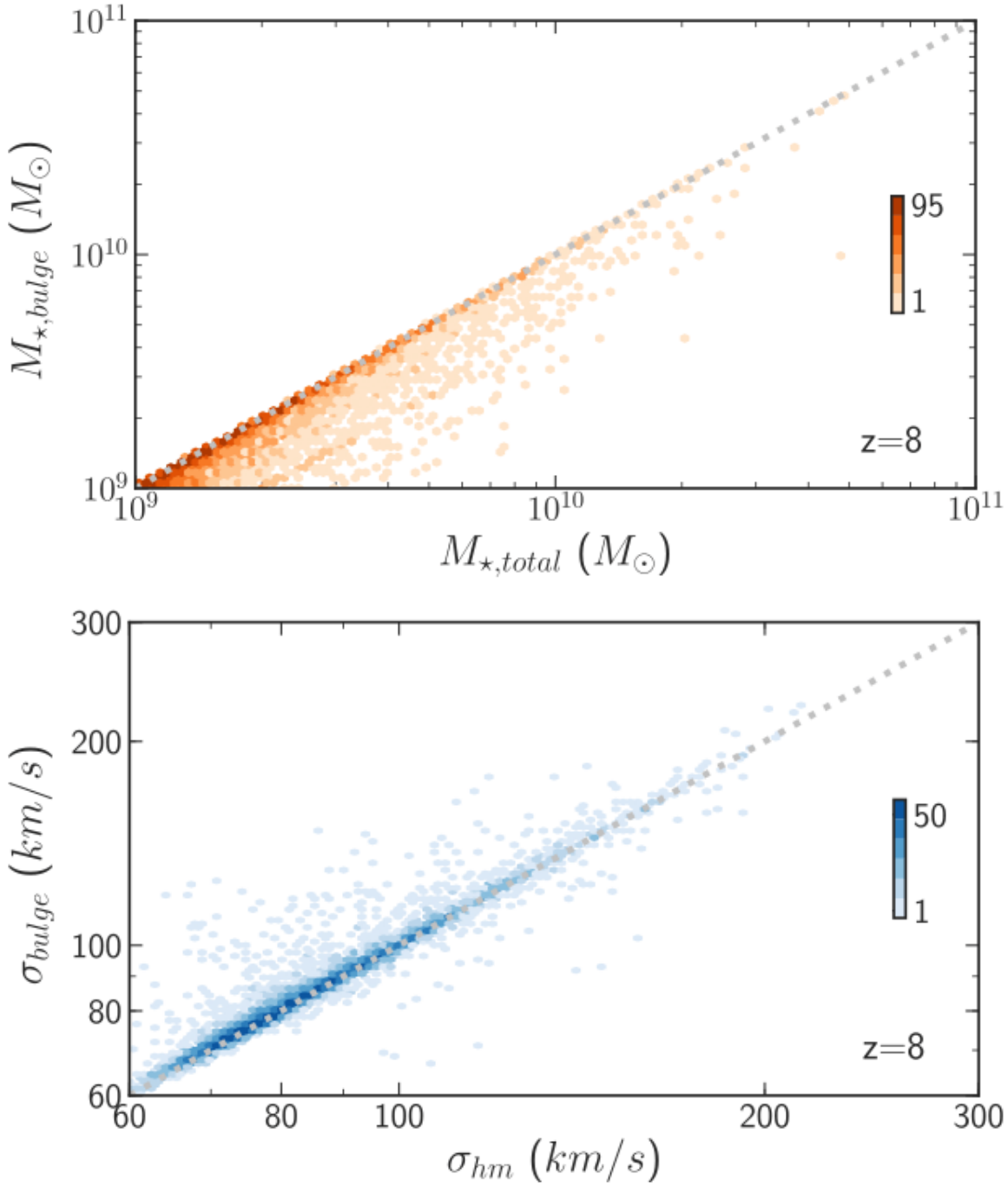

Figure 5.4: The top and bottom panels show $M_{\star,\mathrm{bulge}}$ versus $M_{\star,\mathrm{total}}$ and σ_{bulge} versus σ_{hm} respectively, color coded by the number of galaxies at $z = 8$ in BlueTides.

Table 5.2: The fitting coefficients α and β (normalization and slope) of equation (5.1), the total number of data points N, and the standard deviation of residuals ϵ of the scaling relations at each redshift.

		$M_\bullet - M_{\star,\mathrm{total}}$			$M_\bullet - \sigma_{\mathrm{hm}}$		
z	N	α	β	ϵ	α	β	ϵ
8	8131	$8.25_{\pm 0.03}$	$1.10_{\pm 0.01}$	0.14	$8.35_{\pm 0.08}$	$5.31_{\pm 0.04}$	0.36
9	1567	$8.44_{\pm 0.08}$	$1.19_{\pm 0.01}$	0.14	$8.50_{\pm 0.23}$	$5.95_{\pm 0.12}$	0.40
10	269	$8.76_{\pm 0.20}$	$1.35_{\pm 0.02}$	0.13	$8.49_{\pm 0.54}$	$6.06_{\pm 0.28}$	0.40

		$M_\bullet - M_{\star,\mathrm{bulge}}$			$M_\bullet - \sigma_{\mathrm{bulge}}$		
z	N	α	β	ϵ	α	β	ϵ
8	8131	$8.43_{\pm 0.06}$	$1.16_{\pm 0.01}$	0.15	$8.60_{\pm 0.17}$	$6.15_{\pm 0.09}$	0.42
9	1567	$8.61_{\pm 0.14}$	$1.24_{\pm 0.02}$	0.15	$8.63_{\pm 0.47}$	$6.56_{\pm 0.24}$	0.46
10	269	$8.98_{\pm 0.37}$	$1.43_{\pm 0.04}$	0.14	$8.73_{\pm 1.24}$	$6.95_{\pm 0.63}$	0.46

5.4 The Scaling Relations

5.4.1 Measuring $M_\star$ and σ

The scaling relations between BH mass and their host galaxy properties have been measured by total stellar mass ($M_{\star,\mathrm{total}}$) or bulge stellar mass ($M_{\star,\mathrm{bulge}}$) for $M_\bullet - M_\star$ relation, and by velocity dispersion of bulge stars for $M_\bullet - \sigma$ relation. In simulations, total or half stellar mass is available as a proxy for $M_{\star,\mathrm{bulge}}$. A proxy for σ often used is the velocity dispersion within half-light (or mass) radius (as for example in Sijacki et al. (2015) and DeGraf et al. (2015)). With the dynamical disk-bulge decomposition (see Section 5.2.3) for the stellar components of galaxies in BlueTides, we directly have $M_{\star,\mathrm{bulge}}$ and σ_{bulge} in our galaxies.

The top panel in Figure 5.4 shows the comparison between the $M_{\star,\mathrm{bulge}}$ and $M_{\star,\mathrm{total}}$ color coded according to the number of galaxies. We find that about 82% of objects are bulge-dominated with D/T < 0.3. The bottom panel in Figure 5.4 shows the comparison between σ_{bulge} and σ_{hm} color coded according to the number of galaxies. Again, there is certainly a strong correlation between the two but an increased scatter above the one-to-one relation for a small number of objects, for which we have larger values of σ_{bulge} for a given σ_{hm}. In particular, we find that over 94% and 97% of our galaxies have the difference between σ_{bulge} and σ_{hm} less than 10% and 20% respectively. We shall see in the next section how the detailed dynamical decomposition and the resulting σ_{bulge} and $M_{\star,\mathrm{bulge}}$ impact the scaling relations.

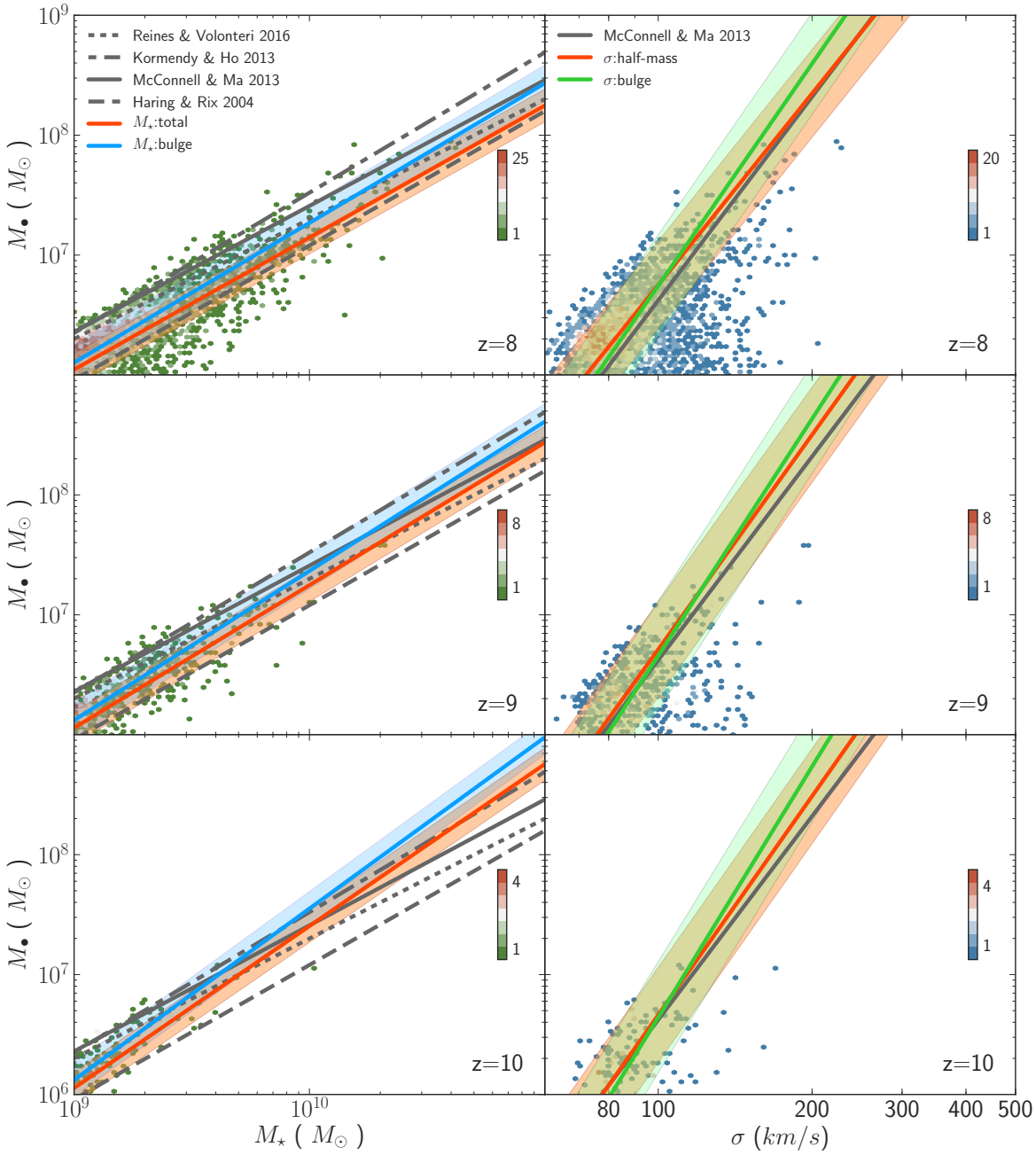

Figure 5.5: The scaling relations at $z = 8$, 9, and 10 in BLUETIDES color coded according to the number of galaxies. Left panels: the $M_\bullet - M_\star$ relations with $M_{\star,\mathrm{bulge}}$ for the data points. The red and blue lines show the best-fitting relation using $M_{\star,\mathrm{total}}$ and $M_{\star,\mathrm{bulge}}$ respectively while the gray lines show the observations (Häring & Rix, 2004; McConnell & Ma, 2013; Kormendy & Ho, 2013; Volonteri & Reines, 2016). Right panels: the $M_\bullet - \sigma$ relations with σ_{bulge} for the data points. The red and green lines show the best-fitting relation with σ_{hm} and σ_{bulge} respectively while the gray lines show the observations in McConnell & Ma (2013). The shaded area shows the standard deviation of residuals.

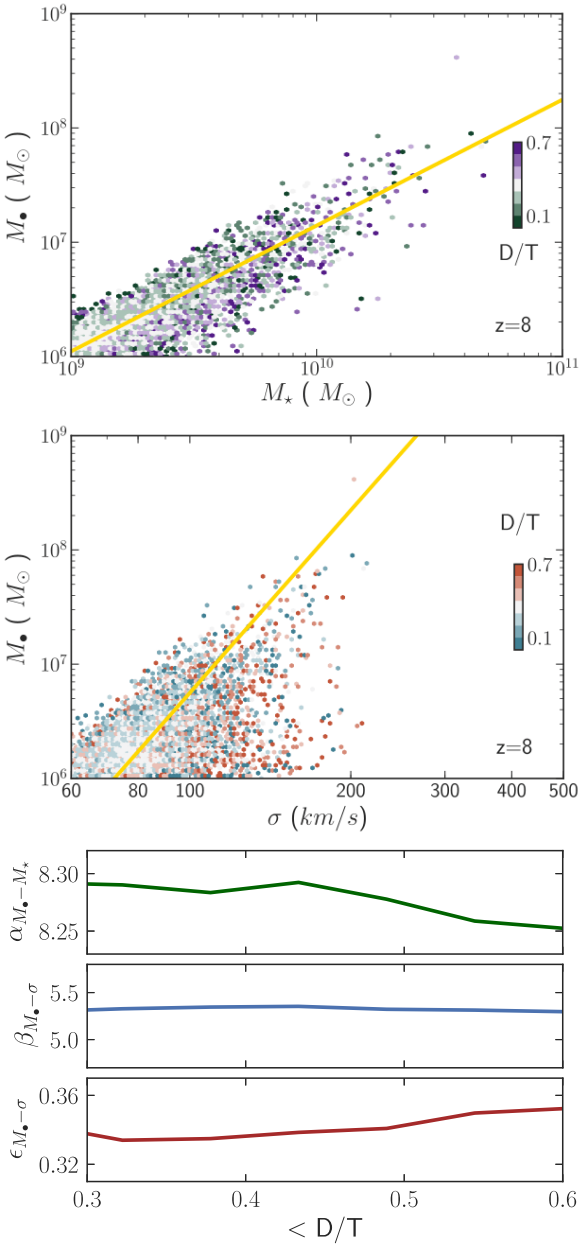

Figure 5.6: Top and middle panels: the $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations color coded according to D/T at $z = 8$ in BlueTides. The yellow lines show the overall fits as the ones in Figure 5.5. Bottom panel: α of $M_\bullet - M_\star$ relation and β and ϵ of $M_\bullet - \sigma$ relation as functions of different limiting D/T.

5.4.2 $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relation

Figure 5.5 shows the $M_\bullet - M_\star$ relation and the $M_\bullet - \sigma$ relation at $z = 8$, 9, and 10 in BlueTides, color coded according to the number of galaxies. Note that the data points are shown with $M_{\star,\rm bulge}$ and $\sigma_{\rm bulge}$. We plot and fit only galaxies with $M_\star > 10^9 M_\odot$ where a sufficient number of star particles are available to carry out the dynamical decomposition reliably. Both scaling relations in BlueTides are best-fitted by power laws as

$$\log_{10}(M_\bullet) = \alpha + \beta \ \log_{10}(X)\,, \tag{5.1}$$

where $M_\bullet$ is in units of $M_\odot$, and X is $M_\star/10^{11}M_\odot$ or σ/kms^{-1}. The fitting coefficients (normalization α and slope β) are summarized in Table 5.2, including the total number of data points N and the standard deviation of the residuals ϵ.

The left panels in Figure 5.5 show the $M_\bullet - M_\star$ relations. The red and blue lines show the best-fitting relation with $M_{\star,\rm total}$ and $M_{\star,\rm bulge}$ respectively while the gray lines show the observations: Häring & Rix (2004), McConnell & Ma (2013) and Kormendy & Ho (2013) with bulge stellar mass and elliptical samples while Volonteri & Reines (2016) with total stellar mass. Our simulation provides the $M_\bullet - M_\star$ relation in the form of $\log_{10}(M_\bullet) = 8.25 + 1.10\ \log_{10}(M_\star/10^{11}M_\odot)$ with $M_{\star,\rm total}$ at $z = 8$, suggesting that the slopes are consistent with the observations but the normalizations are lower than most observations except for the one in Häring & Rix (2004). Both α and β with $M_{\star,\rm total}$ (the red lines) are lower than the ones with $M_{\star,\rm bulge}$ (the blue lines) across all three redshifts, and both get steeper as z is higher. The standard deviation of the residuals (ϵ) is shown as the shaded area.

The right three panels in Figure 5.5 show the $M_\bullet - \sigma$ relations. The red and green lines show the best-fitting relation with $\sigma_{\rm hm}$ and $\sigma_{\rm bulge}$ respectively while the gray lines show the observations in McConnell & Ma (2013). Our simulation provides the $M_\bullet - \sigma$ relation with $\sigma_{\rm hm}$ as $\log_{10}(M_\bullet) = 8.35 + 5.31\ \log_{10}(\sigma/200\mathrm{kms}^{-1})$ at $z = 8$, which is consistent with the results of McConnell & Ma (2013). We note that both α and β using $\sigma_{\rm hm}$ (the red lines) are lower by $\sim 3\%$ and $\sim 10\%$, respectively than the ones with $\sigma_{\rm bulge}$ (the blue lines) across all three redshifts, and both get steeper with increasing z . Moreover, $M_\bullet - \sigma$ relations with $\sigma_{\rm bulge}$ are higher than local measurements. ϵ is shown as the shaded area and, more importantly, $M_\bullet - \sigma$ relation shows a larger scatter than the $M_\bullet - M_\star$ relation ($\epsilon \sim 0.4$ and $\epsilon \sim 0.1$ respectively) in our simulations. We will examine this in Section 5.5.

For most observational results, the scaling relations are established with bulge-dominated galaxies. Here, we report how the relations change in our simulation with bulge-dominated galaxies (D/T < 0.3). In the top two panels in Figure 5.6, we show both relations color coded according to the D/T ratio. In the bottom panels, we show α of $M_\bullet - M_\star$ relations and β and ϵ of $M_\bullet - \sigma$ relations as functions of limiting D/T. We find that the relations hardly change even with different D/T, even for the bulge-dominated regime D/T < 0.3.

5.5 The slope and scatter in the scaling relations

We have shown that the high-z relation is consistent with the locally measured ones (both in slope and normalization). However here we wish to investigate possible selection effects and/or physical parameters that may affect the slope and scatter in the scaling relations (Figure 5.5 and Table 5.2). In particular, we have found that there is a more significant scatter in the $M_\bullet - \sigma$ relation (larger than in local measurements) than the $M_\bullet - M_\star$ relation. The relatively large scatter in the $M_\bullet - \sigma$ relation appears to be due to a significant amount of objects that lie below the main relation: galaxies with relative high σ compared to their relative low $M_\bullet$. Note that $M_\star$ denotes $M_{\star,\mathrm{total}}$ and σ denotes σ_{hm} hereafter unless stated otherwise.

5.5.1 SFR, λ_{Edd}, and $M_\star$ dependence

Here, we examine the dependency of the scaling relations at $z = 8$ on galaxy properties. The top and middle panels in Figure 5.7 show the $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations for threshold samples selected for (from the left to right) SFR (a proxy for the galaxy luminosity), stellar mass ($M_\star$), and BH accretion rate, in units of Eddington (Eddington ratio λ_{Edd}; a proxy for the AGN luminosity), while the yellow line is the best fit from all galaxies. We wish to examine if the above selection thresholds may have an influence on the slopes, normalization and scatter in the relations.

The bottom panels of Figure 5.7 show the variation in normalization α of the $M_\bullet - M_\star$ relation and in the slope β and scatter ϵ of the $M_\bullet - \sigma$ relation as functions of thresholds of SFR, λ_{Edd}, and $M_\star$. We find that the normalization (α) of $M_\bullet - M_\star$ relation increases slightly for higher thresholds of SFR and $M_\star$ (while it is not sensitive to the Eddington rate/AGN luminosity). The slope (β) of $M_\bullet - \sigma$ increases for higher thresholds of SFR and $M_\star$, indicating that selection effects at these high redshifts (which typically bias samples toward high SFR and $M_\star$ objects) are likely to play an important role and lead to biased interpretations of evolutionary effects in these relations when compared to those seen at low-z. The scatter also tends to increase particularly when high SFR or high $M_\star$ samples are selected (which populate the $\sigma \geq 200$ km/s range).

5.5.2 The gas fraction: $f = M_{gas}/M_\star$

In Section 5.5.1, we have examined the dependence of the scaling relation on the various galaxy or AGN parameters; none of them helps to explain the relatively large intrinsic scatter in the $M_\bullet - \sigma$ relation. Here, we test for the effects due to the large range and high value of gas fraction in the high-z galaxies and how that may affect the $M_\bullet - \sigma$ relation.

We use the gas-to-stellar ratio (f) to measure the gas fraction in our galaxies in BlueTides. Figure 5.8 shows the $M_\bullet - \sigma$ relation at $z = 8$ color coded according to f. We find that

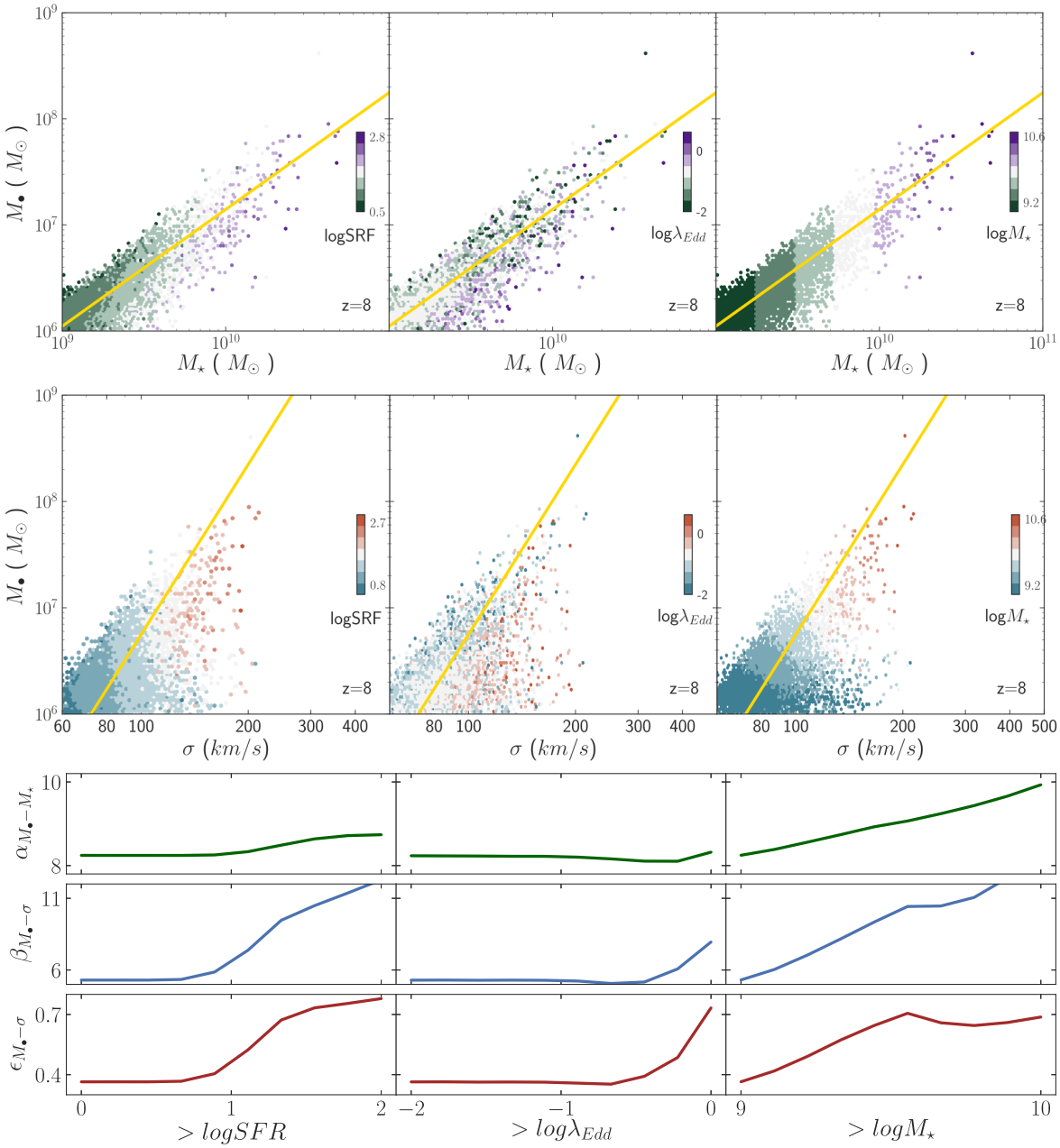

Figure 5.7: Top and middle panels: the $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations color coded according to SFR, $\lambda_{\rm Edd}$, and $M_\star$ (from left to right respectively) at $z = 8$ in BLUETIDES. The yellow lines show the overall fits as the ones in Figure 5.5. Bottom panel: α of $M_\bullet - M_\star$ relation and β and ϵ of $M_\bullet - \sigma$ relation as functions of different thresholds of SFR, $\lambda_{\rm Edd}$, and $M_\star$ from left to right respectively.

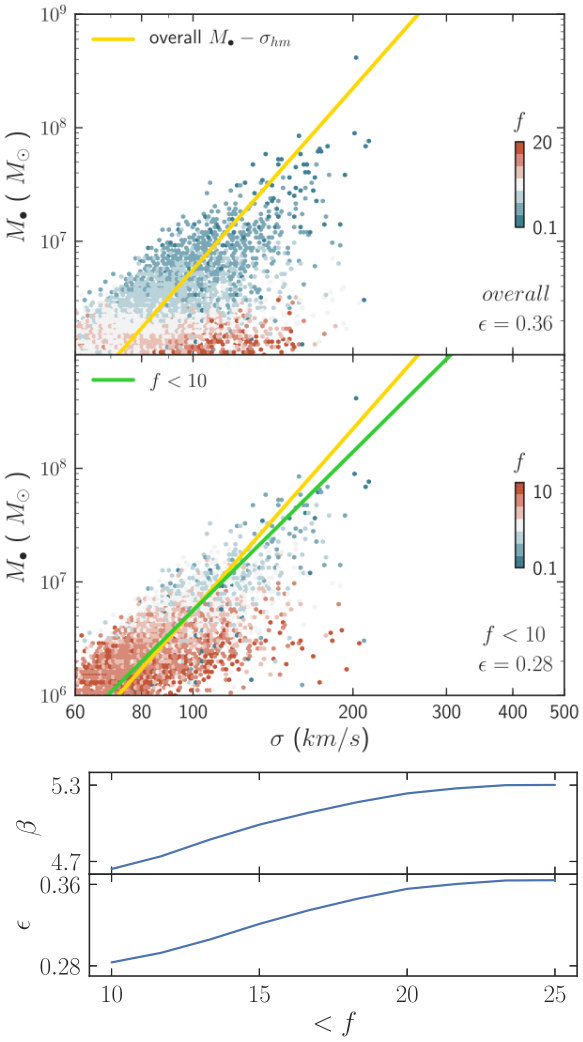

Figure 5.8: Top and middle panels: the $M_\bullet - \sigma$ relation at $z = 8$ color coded according to the gas-to-stellar ratio (f). The yellow and green lines are the best fits with overall galaxies and galaxies with $f < 10$ respectively. Bottom panel: β and ϵ of $M_\bullet - \sigma$ relation at $z = 8$ as functions of limiting f.

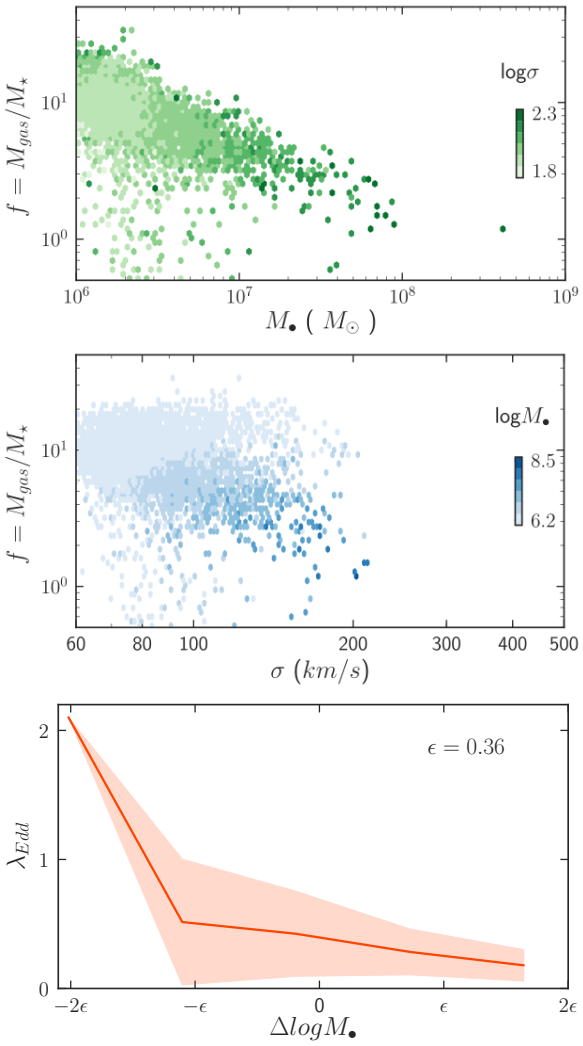

Figure 5.9: Top and middle panels: the relations between $f = M_{\rm gas}/M_\star$, $M_\bullet$, and σ color coded according to σ and $M_\bullet$. Bottom panel: the relation between $\lambda_{\rm Edd}$ and BHs according to their relative position to the main fit on the $M_\bullet - \sigma$ plane, where ϵ is the standard deviation.

the gas fraction in galaxies f has a significant impact on the scatter of the $M_\bullet - \sigma$ relation. In particular, the scatter is significantly reduced for galaxies with a smaller value of f. We find that ϵ decreases significantly from 0.36 (all galaxies) to 0.28 (with $f < 10$) while the slope of the relation (β) decreases from 5.31 (all galaxies) to 4.64 (with $f < 10$). We further illustrate this trend of decreasing of ϵ and β with different limiting f, in the bottom panel in Figure 5.8. Lower gas fractions are indeed more representative of local galaxies, which have been used to measure the $M_\bullet - \sigma$.

The decrease of both ϵ and β with the lower limiting f implies that objects with higher σ but lower $M_\bullet$ are those that have higher f. To look into the relations between f, $M_\bullet$, and σ, we show the top and middle panels of Figure 5.9, color coded according to σ and $M_\bullet$ respectively. The top panel indicates a trend between a decreasing f at increasing $M_\bullet$. Gas fractions of galaxies decrease as BH masses are high/reach the relation, indicating that BH growth is quenched/self-regulated due to AGN feedback. The galaxies with higher σ but lower $M_\bullet$ are indeed those that have larger f (the light blue area in the middle panel), which results in that the $M_\bullet - \sigma$ relation is more scattered if there is no limiting f applied. The bottom panel of Figure 5.9 shows the relation between $\lambda_{\rm Edd}$ and BHs according to their relative position to the main fit on the $M_\bullet - \sigma$ plane. We can see those objects with higher σ but lower $M_\bullet$ tend to have higher $\lambda_{\rm Edd}$ ($\lambda_{\rm Edd}$ is higher if the object is more below to the main fit). These are relatively sizeable galaxies where significant BH growth is still occurring and objects are still moving toward the relation (feedback has not yet saturated the BH growth; see also Figure 5.10) Such an actively growing BH population is indeed rare among the sample of quiescent local BHs.

5.6 The assembly history

To further investigate how the $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations are established we trace the evolution of ~ 200 black holes (from $z \sim 10$ to $z \sim 8$) and their hosts. In Figure 5.10, we show sample tracks on the $M_\bullet - M_\star$ and $M_\bullet - \sigma$ plane from $z = 10$ to $z = 8$. The solid lines shown in orange and blue show the average track in the evolution for higher mass (with $M_\bullet \gtrsim 5 \times 10^7 M_\odot$) and lower mass (with $M_\bullet \lesssim 5 \times 10^6 M_\odot$) respectively. The tracks in the $M_\bullet - \sigma$ plane suggest a slightly steeper growth in the low mass galaxies compared to the high mass galaxies. This likely indicates a fast growth of the black hole mass at approximately fixed values of σ up to the point where the galaxies reach the average relation.

To characterize the overall evolution in these planes, we parameterize the growth of $M_\bullet$, $M_\star$, and σ as

$$\begin{aligned} M_\bullet(z) &\propto (1+z)^{\gamma_\bullet} \\ M_\star(z) &\propto (1+z)^{\gamma_\star} \\ \sigma(z) &\propto (1+z)^{\gamma_\sigma}. \end{aligned} \tag{5.2}$$

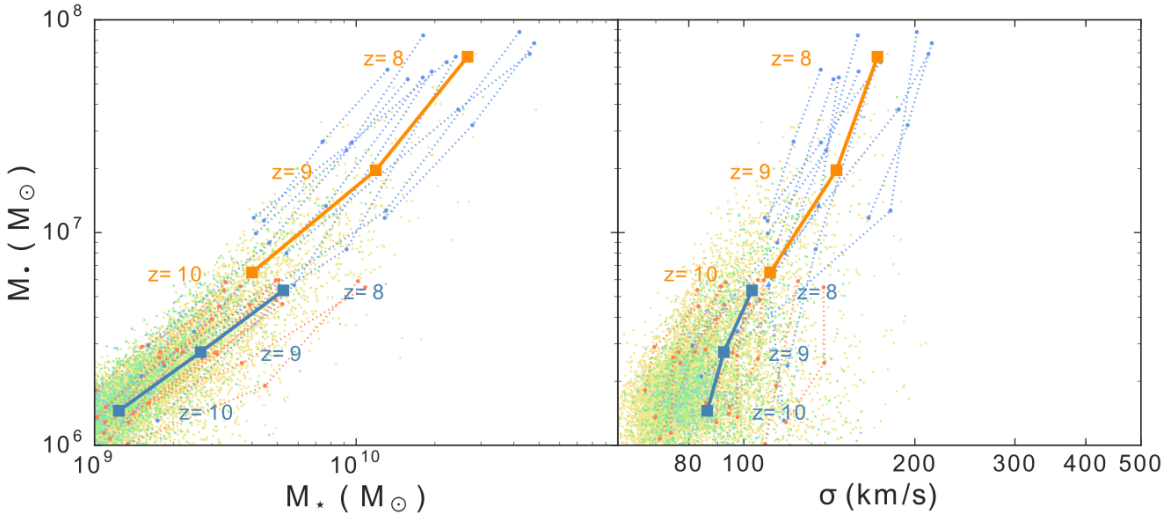

Figure 5.10: The left and the right panels show the growth history on $M_\bullet - M_\star$ and $M_\bullet - \sigma$ planes of our galaxies from $z = 8$ to $z = 10$ respectively. The blue and orange dashed curves show the evolution of ~ 10 galaxies with $M_\bullet > 5 \times 10^7 M_\odot$ and another ~ 10 galaxies with $M_\bullet \sim 5 \times 10^6 M_\odot$. The blue and orange thick curves demonstrate the average growth history for either groups.

where the exponents are $\gamma_\bullet = -9.1 \pm 1.0$, $\gamma_\star = -8.1 \pm 2.2$, and $\gamma_\sigma = -1.6 \pm 0.8$, (the error bars are standard deviation errors). Note that $\gamma_\bullet/\gamma_\star \sim 1.1$ and $\gamma_\bullet/\gamma_\sigma \sim 5.7$, which is consistent with the slope of the scaling relation shown in Table 5.2 (if we use Eqs. (5.2) and eliminate $(1 + z)$, then $\gamma_\bullet/\gamma_\star \sim \beta_{\mathrm{M}_\bullet - \mathrm{M}_\star}$ and $\gamma_\bullet/\gamma_\sigma \sim \beta_{\mathrm{M}_\bullet - \sigma}$). This suggests that, on average, the redshift evolution of these black holes traces the overall scaling relation (at a given redshift), with commensurate growth in black hole mass and stellar mass.

We now take a step further and look into two distinct mass regimes of the above sample of 200 BHs: 1) A low mass range of BHs with $M_\bullet < 5 \times 10^6 M_\odot$, and 2) A high mass range of BHs with $M_\bullet > 1 \times 10^7 M_\odot$. We note that, as discussed in the previous section, the lower mass subsample has more scatter in the $M - \sigma$ relation (see Section 5.4). We find that $\gamma_\bullet = -5.7 \pm 3.5$, $\gamma_\sigma = -0.72 \pm 0.49$ and $\gamma_\bullet = -10 \pm 2.6$, $\gamma_\sigma = -1.8 \pm 0.5$ for low mass and high mass subsamples respectively; therefore, the higher mass BHs have steeper increase in both $M_\bullet$ and σ, as is clearly illustrated in Figure 5.10. Interestingly, we also find $\gamma_\bullet/\gamma_\sigma \sim 7.9$, which is higher than the slope of the overall fit of the $M - \sigma$ relation in Section 5.4, suggesting that indeed these lower mass BHs tend to grow their black holes faster than their velocity dispersion and saturate their growth as they move closer to the mean relation. This behavior can then potentially lead to a decrease in the scatter in the relation at the low redshifts, where most black holes and galaxies have low gas fractions and have quenched their growth and star formation (particularly in the bulge dominated samples where the relations are typically measured). For the high mass sample, $\gamma_\bullet/\gamma_\sigma \sim 5.55$ which is consistent with the slope of

the overall fit of the $M - \sigma$ relation, so that high mass objects continue to move along the relation.

5.7 Conclusions

We investigate the global properties of supermassive black holes at high redshifts ($z \sim 8$, 9, 10), which include scaling relations w.r.t properties of their host galaxies (σ, $M_\star$) and their redshift evolution using the BlueTides simulation. The bolometric luminosity of BHs span more than two orders of magnitude around a mean of 0.3 of the Eddington luminosity. The BH mass functions in our simulation tend to have steeper slopes compared to the one inferred at $z = 6$ measured from optical quasars. While this may be due to obscuration, we find that it is consistent with the large range of luminosities spanned and the flux cuts implied by observations. We have also shown that the BH mass density and BH accretion rate are broadly consistent with current observational constraints at the highest redshifts ($z \sim 7$).

The scaling relations, $M_\bullet - M_\star$, $M_\bullet - \sigma$ predicted by BlueTides reveal that correlations between the growth of black holes and their host galaxies persist at high-z ($z = 8$ to $z = 10$), with the slopes and normalizations consistent with published relations at low-z. For the scatter, we find that the $M_\bullet - \sigma$ relation has a significantly higher scatter compared to current measurements as well as the $M_\bullet - M_\star$ relation. We further show that this large scatter can be primarily attributed to the gas-to-stellar ratio ($f = M_{\rm gas}/M_\star$), wherein we observe a significant decrease in the scatter ($\epsilon = 0.36$ to $\epsilon = 0.28$) upon exclusion of galaxies with $f = M_{\rm gas}/M_\star > 10$. Such high gas fraction systems have the largest star formation rates and black hole accretion rates indicating that these systems have not yet converged to the relation. We also find that the assembly history of the evolution of BHs on $M_\bullet - M_\star$ and $M_\bullet - \sigma$ planes is, on an average, consistent with the corresponding scaling relations; in other words, the average trajectory of the evolution of BHs traces the mean scaling relations well.

Note that the deviation between BH-galaxy relation at high-z and the local relation is important for studying how the BH seeds grow. BlueTides suggests that the relations are generally consistent with the ones of local measurements (at $z = 8$ particularly) but also that the scatter of the $M_\bullet - \sigma$ relation has been larger than one from the local relation. It implies that even the overall relation is consistent with local measurements, gas-rich systems at such high-z have started influencing the BH-relations. Since BlueTides has run to z=8 only, we would not have much validation for any other model. Therefore, we adopt models that have been used in other simulations (e.g. *Illustris*, *EAGLE*, and *MassiveBlack II*), which have successfully reproduced the evolution of the AGN luminosity functions, mass functions, and local Mbh-sigma relation. A currently proceeding work is to re-simulate some subparts of the volume in BlueTides to test different BH models (AGN feedback and BH seeding and accretion models) to see if/how different models will affect the BH growth.

Acknowledgements

We thank A. Tenneti, C. DeGraf, and M. Volonteri for discussions on D/T ratio, the fitting method for scaling relation, and BH mass density respectively. We acknowledge funding from NSF ACI-1614853, NSF AST- 1517593, NSF AST-1616168 and the BlueWaters PAID program. The BLUETIDES simulation was run on the Blue Waters facility at the National Center for Supercomputing Applications.

6 Conclusion

To push the boundaries of our knowledge on structure formation in astrophysics, studying astronomical objects is essential. For astronomical objects at different scales, there are various methods adequate to each of them. In the local Universe, observation with large-sky surveys is one of the most powerful ways to investigate the Milky Way as the Galaxy provides rich information and clues for astrophysicists. At high redshift, cosmological simulations allow us to go further in time and space, despite the lack of observations. This thesis consists of four different projects studying multiple astronomical objects from the local Universe to high redshift. Two focus on globular clusters and RR Lyrae stars in and around the Milky Way from large-sky surveys. The other two focus on supermassive black holes at high redshift in cosmological simulations.

In Chapter 2, I implemented an automatic searching algorithm to systemically search for globular clusters in *Gaia* DR2, targeting the areas associated with the 55 dwarf galaxies in and around the Milky Way. Eleven possible candidates were identified through the targeted search. Crossed-matched with existing imaging data, all eleven objects are known globular clusters or galaxies, and only Fornax globular clusters 1 – 6 among them are associated with the targeted dwarf galaxy. Despite no newly found objects, the result still provided constraints on the specific frequency of globular clusters. Assuming a Gaussian luminosity function for globular clusters, I computed that the completeness of the globular cluster search was above 90 percent for most dwarf galaxies, given the detection limit in magnitude obtained from a set of simulated globular clusters. The resulting 90 percent credible intervals/upper limits on the globular cluster specific frequency suggested that the probability of galaxies fainter than $M_V = -9$ to host globular clusters is lower than 0.1.

In Chapter 3, I presented a RR Lyrae catalog based on the combination of ZTF DR3 and *Gaia* EDR3, using a multi-step classification pipeline relying on the Fourier decomposition fitting to the multi-band ZTF light curves and random forest classification. The resulting catalog contained 71,755 RR Lyraes with period and light curve parameter measurements, covering the Northern sky with declination $\geq -28°$. Based on the catalog, the Galactic halo density distribution suggested the broadly ellipsoidal stellar distribution with flattening around 0.6 and power-law density profile with three known major over-densities of the halo substructure: the Virgo over-density, the Hercules-Aquila Cloud, and the Sagittarius Stream. The RR Lyrae density distribution demonstrated a possible connection between the Virgo over-density and the Hercules-Aquila Cloud, supporting the possible association of several over-densities such as Hercules-Aquila, Virgo, Eridanus–Phoenix and their link to the Gaia-Encelladus-Sausage merger (i.e., Simion et al., 2019). Besides, the RR Lyrae over-density in the Northern hemispheres was in broad agreement with the effect of the dynamical response of the Galactic halo to the Large Magellanic Cloud (i.e., Conroy et al., 2021).

In Chapter 4, I investigated the early growth of supermassive black holes using constrained cosmological simulations with multiple black hole seeding scenarios. With a box size of only 15 h^{-1}Mpc on each side, the constrained simulations successfully reconstructed the initial conditions in the BlueTides simulation for the first quasars and their environments. The results of realizations with different local tidal fields suggested that a low-tidal field environment was crucial for the growth of the earliest and most massive black holes. With the same constrained initial conditions of a low tidal field, I conducted multiple simulations with different black hole seed masses of 5×10^3, 5×10^4, and 5×10^5 $h^{-1} M_\odot$. The resulting black hole masses in our constrained simulations converged to $\sim 10^9$ $M_\odot$ at $z = 6$, suggesting that the final supermassive black hole mass was insensitive to the initial seed mass regardless of the choice of black hole seeding parameters. However, in the early times at $z > 10$, the supermassive black hole growth varied among the constrained simulations with different seeding scenarios. Black holes in less massive seed models tended to grow slower initially unless they were seeded in more common but less massive halos so that they could merge frequently. The significant difference in the early merger rates provided an interesting discriminating feature for black hole models with small or large seed mass at high redshift.

In Chapter 5, I investigated the scaling relations between supermassive black holes and their host galaxy properties together with their redshift evolution at high redshift using the BlueTides simulation. The $M_\bullet - M_\star$ and $M_\bullet - \sigma$ relations in BlueTides revealed that the correlations between the growth of black holes and their host galaxies persisted at $z = 8$ to $z = 10$, with the slopes and normalizations consistent with published relations at low-z. However, the $M_\bullet - \sigma$ relation had a significantly larger scatter than current measurements and the $M_\bullet - M_\star$ relation, primarily attributed to a high gas-to-stellar ratio of host galaxies. Such high gas fraction systems had huge star formation rates and black hole accretion rates, indicating that these systems have not yet converged to the relation. According to the assembly history of the black hole evolution on $M_\bullet - M_\star$ and $M_\bullet - \sigma$ planes, the average trajectory of the black hole evolution traced the mean scaling relations well. BlueTides suggested that the overall relationship was consistent with current local measurements and that gas-rich systems at such high-z had started influencing the relations.

Overall the thesis has provided the studies of globular clusters and RR Lyrae stars in and around the Milky Way using large-sky surveys and supermassive black holes at high redshift using cosmological simulations. Understanding the Galaxy using large-sky surveys is essential to develop, test, or constrain modern theories about structure formation at small scales. Beyond the current observation limit at high redshift, cosmological simulation provides a gate to explore the formation and evolution of astronomical objects. The projects in this thesis can be expanded and continued in the future. Below I propose some future projects which extend the ideas or results summarized in this thesis.

One follow-up project comes from the reusability for different surveys of the searching algorithm in Chapter 2. It is worth executing the algorithm again to search for globular

clusters around the Milky Way dwarf galaxies but using *Gaia* EDR3, which has higher resolution and more faint sources than *Gaia* DR2 used in Chapter 2. The searching result can push the detection limit to the fainter magnitude and thus result in more complete searches compared to what we have in Chapter 2. There is a chance to find unknown globular clusters associated with the dwarfs or better constrain the globular cluster specific frequency of the dwarfs. Either finding more globular clusters or constraining the number of globular clusters better can push the current limit to the modern paradigm of structure formation on small scales.

The RR Lyrae catalog presented in Chapter 3 has several possible future projects, as RR Lyrae stars are great tracers for substructures and usually come with decent distance measurements. Not only the 3D spatial coordinates but also the 2D *Gaia* proper motions are available in the catalog. First, we can search for local overdensities of the RR Lyrae spatial distribution to find possibly missing substructures in the Northern sky, such as dwarf galaxies, star clusters, or streams. The potential of discovering unknown substructures with distance measurements is beneficial to understanding the Milky Way structure. Another project associated with the catalog is the detailed study of the Milky Way halo structure using the 3D density distribution of the RR Lyrae stars. The discussion of the RR Lyrae density distribution in Chapter 3 is to demonstrate the RR Lyrae catalog, which can be seen as a preliminary result to more detailed and quantitative studies, such as density fitting for halo triaxiality.

Besides the RR Lyrae catalog itself, the classification pipeline used in Chapter 3 can be applied to other time-series survey data with light curve measurements as for now or in the future to classify other RR Lyrae catalogs, which is a potential future project. Another future project is to improve the pipeline itself, as it currently contains multiple stages whose light curve fitting step is relatively computationally expensive. The random forest classifier relies on the resulting parameters from the light curve fitting, so the fitting step is inevitable. A possible way to avoid light curve fitting is to make the classification process from multi-stage to end-to-end by using sequential deep learning models to take in measured time-series light curve data as input directly and have binary predictions as output. This end-to-end classification may be more efficient for the entire classification process than the multi-stage pipeline by avoiding light curve fitting. Furthermore, the trained model can generalize to other time-series surveys as long as they contain light curve measurements that can be used as input directly.

Bibliography

Abadi M. G., Navarro J. F., Steinmetz M., Eke V. R., 2003a, ApJ, 591, 499

Abadi M. G., Navarro J. F., Steinmetz M., Eke V. R., 2003b, ApJ, 597, 21

Abbott T. M. C., et al., 2018, ApJS, 239, 18

Abel T., Bryan G. L., Norman M. L., 2002, Science, 295, 93

Amaro-Seoane P., et al., 2017, arXiv e-prints, p. arXiv:1702.00786

Anglés-Alcázar D., Davé R., Özel F., Oppenheimer B. D., 2014, ApJ, 782, 84

Anglés-Alcázar D., Özel F., Davé R., Katz N., Kollmeier J. A., Oppenheimer B. D., 2015, ApJ, 800, 127

Anglés-Alcázar D., Davé R., Faucher-Giguère C.-A., Özel F., Hopkins P. F., 2017, MNRAS, 464, 2840

Arenou F., et al., 2018, A&A, 616, A17

Aslanyan G., Feng Y., White M., 2016, Constrained Gaussian relealization in FastPM, `https://github.com/rainwoodman/fastpm`

Astropy Collaboration et al., 2013, A&A, 558, A33

Bañados E., et al., 2017, preprint, (`arXiv:1712.01860`)

Bañados E., et al., 2018, Nature, 553, 473

Baker M., Willman B., 2015, The Astronomical Journal, 150, 160

Balbinot E., Helmi A., 2021, arXiv e-prints, p. arXiv:2104.09794

Barkana R., Loeb A., 2001, Phys.Rep., 349, 125

Battaglia N., Trac H., Cen R., Loeb A., 2013, ApJ, 776, 81

Beasley M. A., Trujillo I., Leaman R., Montes M., 2018, Nature, 555, 483

Begelman M. C., Rees M. J., 1978, MNRAS, 185, 847

Begelman M. C., Volonteri M., Rees M. J., 2006, MNRAS, 370, 289

Bell E. F., et al., 2008, ApJ, 680, 295

Bellazzini M., Ferraro F. R., Ibata R., 2003, AJ, 125, 188

Bellm E. C., et al., 2019, PASP, 131, 018002

Bellovary J., Volonteri M., Governato F., Shen S., Quinn T., Wadsley J., 2011, The Astrophysical Journal, 742, 13

Bellovary J., Brooks A., Volonteri M., Governato F., Quinn T., Wadsley J., 2013, ApJ, 779, 136

Belokurov V., et al., 2006, ApJ, 642, L137

Belokurov V., et al., 2007a, ApJ, 654, 897

Belokurov V., et al., 2007b, ApJ, 657, L89

Belokurov V., et al., 2009, MNRAS, 397, 1748

Belokurov V., Deason A. J., Koposov S. E., Catelan M., Erkal D., Drake A. J., Evans N. W., 2018, MNRAS, 477, 1472

Belokurov V., Sanders J. L., Fattahi A., Smith M. C., Deason A. J., Evans N. W., Grand R. J. J., 2020, MNRAS, 494, 3880

Bennert V. N., Treu T., Woo J.-H., Malkan M. A., Le Bris A., Auger M. W., Gallagher S., Blandford R. D., 2010, ApJ, 708, 1507

Bergström L., 2000, Reports on Progress in Physics, 63, 793

Bertone G., Hooper D., Silk J., 2005, Phys.Rep., 405, 279

Bhowmick A. K., Di Matteo T., Feng Y., Lanusse F., 2017, preprint, (arXiv:1707.02312)

Blažko S., 1907, Astronomische Nachrichten, 175, 325

Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517

Bode P., Ostriker J. P., Turok N., 2001, ApJ, 556, 93

Bonaca A., et al., 2012, AJ, 143, 105

Bondi H., 1952, MNRAS, 112, 195

Bondi H., Hoyle F., 1944, MNRAS, 104, 273

Bournaud F., Dekel A., Teyssier R., Cacciato M., Daddi E., Juneau S., Shankar F., 2011, ApJ, 741, L33

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645

Boylan-Kolchin M., 2018, MNRAS, 479, 332

Breiman L., 2001, in Machine Learning. pp 5–32

Bressan A., Marigo P., Girardi L., Salasnich B., Dal Cero C., Rubele S., Nanni A., 2012, MNRAS, 427, 127

Brodie J. P., Strader J., 2006, Annual Review of Astronomy and Astrophysics, 44, 193

Bromm V., Loeb A., 2003, ApJ, 596, 34

Buchner J., et al., 2015, ApJ, 802, 89

Cacciari C., Clementini G., 2003, Globular Cluster Distances from RR Lyrae Stars. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 105–122, doi:10.1007/978-3-540-39882-0_6, https://doi.org/10.1007/978-3-540-39882-0_6

Cáceres C., Catelan M., 2008, The Astrophysical Journal Supplement Series, 179, 242

Catelan M., 2009, Ap&SS, 320, 261

Catelan M., Pritzl B. J., Smith H. A., 2004, The Astrophysical Journal Supplement Series, 154, 633

Chabrier G., 2005, in Corbelli E., Palla F., Zinnecker H., eds, The Initial Mass Function 50 Years Later. Springer Netherlands, Dordrecht, pp 41–50

Chen X., Wang S., Deng L., de Grijs R., Yang M., Tian H., 2020, ApJS, 249, 18

Choksi N., Gnedin O. Y., 2019, MNRAS, 486, 331

Ciotti L., Ostriker J. P., Proga D., 2009, ApJ, 699, 89

Clementini G., et al., 2019a, A&A, 622, A60

Clementini G., et al., 2019b, A&A, 622, A60

Cole S., et al., 2005, MNRAS, 362, 505

Conroy C., Naidu R. P., Garavito-Camargo N., Besla G., Zaritsky D., Bonaca A., Johnson B. D., 2021, Nature, 592, 534

Costa T., Sijacki D., Trenti M., Haehnelt M. G., 2014, MNRAS, 439, 2146

Crnojević D., Sand D. J., Zaritsky D., Spekkens K., Willman B., Hargis J. R., 2016, ApJ, 824, L14

Croton D. J., et al., 2006, MNRAS, 367, 864

Dalal N., White M., Bond J. R., Shirokov A., 2008, ApJ, 687, 12

DeGraf C., Di Matteo T., Treu T., Feng Y., Woo J.-H., Park D., 2015, MNRAS, 454, 913

Dekel A., Silk J., 1986, ApJ, 303, 39

Demers S., Irwin M. J., Kunkel W. E., 1994, AJ, 108, 1648

Devecchi B., Volonteri M., 2009, ApJ, 694, 302

Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604

Di Matteo T., Colberg J., Springel V., Hernquist L., Sijacki D., 2008, ApJ, 676, 33

Di Matteo T., Khandai N., DeGraf C., Feng Y., Croft R. A. C., Lopez J., Springel V., 2012, ApJ, 745, L29

Di Matteo T., Croft R. A. C., Feng Y., Waters D., Wilkins S., 2017, MNRAS, 467, 4243

Dicke R. H., Peebles P. J. E., Roll P. G., Wilkinson D. T., 1965, ApJ, 142, 414

Diemand J., Kuhlen M., Madau P., Zemp M., Moore B., Potter D., Stadel J., 2008, Nature, 454, 735

Drake A. J., et al., 2009, ApJ, 696, 870

Drake A. J., et al., 2014, ApJS, 213, 9

Drlica-Wagner A., et al., 2015, ApJ, 813, 109

Dubois Y., Pichon C., Devriendt J., Silk J., Haehnelt M., Kimm T., Slyz A., 2013, MNRAS, 428, 2885

Duffau S., Zinn R., Vivas A. K., Carraro G., Méndez R. A., Winnick R., Gallart C., 2006, ApJ, 636, L97

Dunkley J., et al., 2009, ApJS, 180, 306

ESA DPAC 2019, Documentation release 1.2, https://gea.esac.esa.int/archive/documentation/GDR2/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html

Eisenstein D. J., Loeb A., 1995, ApJ, 443, 11

Eisenstein D. J., et al., 2005, ApJ, 633, 560

El-Badry K., Quataert E., Weisz D. R., Choksi N., Boylan-Kolchin M., 2019, MNRAS, 482, 4528

Erkal D., Belokurov V. A., Parkin D. L., 2020, MNRAS, 498, 5574

Fall S. M., Efstathiou G., 1980, MNRAS, 193, 189

Fan X., et al., 2006, AJ, 132, 117

Fanidakis N., Baugh C. M., Benson A. J., Bower R. G., Cole S., Done C., Frenk C. S., 2011, MNRAS, 410, 53

Faucher-Giguère C.-A., Lidz A., Zaldarriaga M., Hernquist L., 2009, ApJ, 703, 1416

Fellhauer M., et al., 2006, ApJ, 651, 167

Feng Y., 2018, gaepsi2, https://github.com/rainwoodman/gaepsi2

Feng Y., Di Matteo T., Croft R., Khandai N., 2014, MNRAS, 440, 1865

Feng Y., Di Matteo T., Croft R., Tenneti A., Bird S., Battaglia N., Wilkins S., 2015, ApJ, 808, L17

Feng Y., Di-Matteo T., Croft R. A., Bird S., Battaglia N., Wilkins S., 2016a, MNRAS, 455, 2778

Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016b, MNRAS, 463, 2273

Ferguson A. M. N., Irwin M. J., Ibata R. A., Lewis G. F., Tanvir N. R., 2002, AJ, 124, 1452

Fernique P., et al., 2015, A&A, 578, A114

Ferrara A., Salvadori S., Yue B., Schleicher D., 2014, MNRAS, 443, 2410

Fiore F., et al., 2012, A&A, 537, A16

Fiorentino G., et al., 2015, ApJ, 798, L12

Forbes D. A., Brodie J. P., Grillmair C. J., 1997, AJ, 113, 1652

Forbes D. A., Masters K. L., Minniti D., Barmby P., 2000, A&A, 358, 471

Forbes D. A., Read J. I., Gieles M., Collins M. L. M., 2018, MNRAS, 481, 5592

Fritz T. K., Battaglia G., Pawlowski M. S., Kallivayalil N., van der Marel R., Sohn S. T., Brook C., Besla G., 2018, A&A, 619, A103

Gaia Collaboration et al., 2016, A&A, 595, A1

Gaia Collaboration et al., 2018a, A&A, 616, A1

Gaia Collaboration et al., 2018b, A&A, 616, A12

Gaia Collaboration Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Biermann M., 2020, arXiv e-prints, p. arXiv:2012.01533

Garavito-Camargo N., Besla G., Laporte C. F. P., Johnston K. V., Gómez F. A., Watkins L. L., 2019, ApJ, 884, 51

Gebhardt K., et al., 2000, ApJ, 539, L13

Georgiev I. Y., Puzia T. H., Goudfrooij P., Hilker M., 2010, MNRAS, 406, 1967

Ginsburg A., Koposov S., Christian T., 2020, IMF: Simple tools to work with the Initial Mass Function

Gnedin O. Y., Ostriker J. P., 1997, ApJ, 474, 223

Gnedin N. Y., Kravtsov A. V., Rudd D. H., 2011, ApJS, 194, 46

Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759

Gratton R., Sneden C., Carretta E., 2004, ARA&A, 42, 385

Gratton R., Bragaglia A., Carretta E., D'Orazi V., Lucatello S., Sollima A., 2019, A&A Rev., 27, 8

Gravity Collaboration et al., 2018, A&A, 618, L10

Grazian A., et al., 2015, A&A, 575, A96

Grillmair C. J., 2009, ApJ, 693, 1118

Grillmair C. J., Dionatos O., 2006, ApJ, 643, L17

Gültekin K., et al., 2009, ApJ, 698, 198

Habouzit M., Volonteri M., Dubois Y., 2017, MNRAS, 468, 3935

Haehnelt M. G., Rees M. J., 1993, MNRAS, 263, 168

Hand N., Feng Y., 2015, nbodykit, https://nbodykit.readthedocs.io/en/latest/

Hargis J. R., et al., 2016, The Astrophysical Journal, 818, 39

Häring N., Rix H.-W., 2004, ApJ, 604, L89

Harrington R. G., Wilson A. G., 1950, PASP, 62, 118

Harris W. E., 1991, ARA&A, 29, 543

Harris W. E., 1996, AJ, 112, 1487

Harris W. E., 2001, in Labhardt L., Binggeli B., eds, Saas-Fee Advanced Course 28: Star Clusters. p. 223

Harris W. E., 2010, arXiv e-prints, p. arXiv:1012.3224

Harris W. E., van den Bergh S., 1981, AJ, 86, 1627

Harris G. L. H., Harris W. E., Poole G. B., 1999, AJ, 117, 855

He C.-C., Ricotti M., Geen S., 2020, MNRAS, 492, 4858

Helmi A., 2004, ApJ, 610, L97

Helmi A., White S. D. M., de Zeeuw P. T., Zhao H., 1999, Nature, 402, 53

Hernitschek N., et al., 2016, ApJ, 817, 73

Hernitschek N., et al., 2017, ApJ, 850, 96

Hernquist L., Mihos J. C., 1995, ApJ, 448, 41

Hinshaw G., et al., 2013, ApJS, 208, 19

Hirschmann M., Khochfar S., Burkert A., Naab T., Genel S., Somerville R. S., 2010, MNRAS, 407, 1016

Hodge P. W., 1961, AJ, 66, 83

Hoffman Y., Ribak E., 1991, ApJ, 380, L5

Holl B., et al., 2018, A&A, 618, A30

Homma D., et al., 2016, The Astrophysical Journal, 832, 21

Homma D., et al., 2018, PASJ, 70, S18

Hopkins P. F., 2013, MNRAS, 428, 2840

Hopkins P. F., Quataert E., 2010, MNRAS, 407, 1529

Howard C. S., Pudritz R. E., Harris W. E., 2018a, Nature Astronomy, 2, 725

Howard C. S., Pudritz R. E., Harris W. E., Klessen R. S., 2018b, MNRAS, 475, 3121

Hoyle F., Lyttleton R. A., 1939, Proceedings of the Cambridge Philosophical Society, 35, 405

Huang K.-W., Koposov S. E., 2021a, The RR Lyrae variable catalog of ZTF DR3, doi:10.5281/zenodo.5774018, https://doi.org/10.5281/zenodo.5774018

Huang K.-W., Koposov S. E., 2021b, arXiv e-prints, p. arXiv:2112.06017

Huang K.-W., Koposov S. E., 2021c, MNRAS, 500, 986

Huang K.-W., Di Matteo T., Bhowmick A. K., Feng Y., Ma C.-P., 2018, MNRAS, 478, 5063

Huang K.-W., Ni Y., Feng Y., Di Matteo T., 2020, MNRAS, 496, 1

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90

Ibata R., Lewis G. F., Irwin M., Totten E., Quinn T., 2001, ApJ, 551, 294

Iorio G., Belokurov V., 2021, MNRAS, 502, 5686

Iorio G., Belokurov V., Erkal D., Koposov S. E., Nipoti C., Fraternali F., 2018, MNRAS, 474, 2142

Ivezić Ž., et al., 2008, ApJ, 684, 287

Jahnke K., Macciò A. V., 2011, ApJ, 734, 92

Jayasinghe T., et al., 2018, Research Notes of the American Astronomical Society, 2, 18

Jayasinghe T., et al., 2020, VizieR Online Data Catalog, p. II/366

Jiang L., et al., 2009, AJ, 138, 305

Jiang Y.-F., Stone J., Davis S. W., 2017, preprint, (arXiv:1709.02845)

Johnson J. L., Bromm V., 2007, MNRAS, 374, 1557

Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/

Jordán A., et al., 2007, ApJS, 171, 101

Jurcsik J., Kovacs G., 1996, A&A, 312, 111

Jurić M., et al., 2008, ApJ, 673, 864

Katz N., Weinberg D. H., Hernquist L., 1996, ApJS, 105, 19

Katz H., Sijacki D., Haehnelt M. G., 2015, MNRAS, 451, 2352

Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201

Khandai N., Di Matteo T., Croft R., Wilkins S., Feng Y., Tucker E., DeGraf C., Liu M.-S., 2015, MNRAS, 450, 1349

King A., 2003, ApJ, 596, L27

Klement R., et al., 2009, ApJ, 698, 865

Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, ApJ, 522, 82

Koch A., Grebel E. K., Wyse R. F. G., Kleyna J. T., Wilkinson M. I., Harbeck D. R., Gilmore G. F., Evans N. W., 2006, AJ, 131, 895

Koposov S., 2018, Sqlutilpy module to access SQL databases, https://github.com/segasai/sqlutilpy

Koposov S., Bartunov O., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, Astronomical Society of the Pacific Conference Series Vol. 351, Astronomical Data Analysis Software and Systems XV. p. 735

Koposov S. E., Glushkova E. V., Zolotukhin I. Y., 2008a, A&A, 486, 771

Koposov S., et al., 2008b, ApJ, 686, 279

Koposov S. E., Yoo J., Rix H.-W., Weinberg D. H., Macciò A. V., Escudé J. M., 2009, ApJ, 696, 2179

Koposov S. E., Rix H.-W., Hogg D. W., 2010, ApJ, 712, 260

Koposov S. E., Belokurov V., Torrealba G., Evans N. W., 2015, The Astrophysical Journal, 805, 130

Koposov S. E., Belokurov V., Torrealba G., 2017, MNRAS, 470, 2702

Kormendy J., Ho L. C., 2013, ARA&A, 51, 511

Koushiappas S. M., Bullock J. S., Dekel A., 2004, MNRAS, 354, 292

Kruijssen J. M. D., Pfeffer J. L., Reina-Campos M., Crain R. A., Bastian N., 2019, MNRAS, 486, 3180

Krumholz M. R., Gnedin N. Y., 2011, ApJ, 729, 36

Latif M. A., Schleicher D. R. G., Schmidt W., Niemeyer J. C., 2013, MNRAS, 436, 2989

Lauer T. R., Tremaine S., Richstone D., Faber S. M., 2007, ApJ, 670, 249

Lee J.-W., Carney B. W., 1999, AJ, 118, 1373

Li Y., et al., 2007, ApJ, 665, 187

Li Y., Hu W., Takada M., 2014, Phys.Rev.D, 89, 083519

Li T. S., et al., 2016, ApJ, 817, 135

Li Y., Schmittfull M., Seljak U., 2018, J. Cosmology Astropart. Phys., 2018, 022

Lodato G., Natarajan P., 2006, MNRAS, 371, 1813

Loeb A., Rasio F. A., 1994, ApJ, 432, 52

Lomb N. R., 1976, Ap&SS, 39, 447

Lupi A., Haardt F., Dotti M., Fiacconi D., Mayer L., Madau P., 2016, MNRAS, 456, 2993

Luque E., et al., 2017, MNRAS, 468, 97

Lynden-Bell D., Lynden-Bell R. M., 1995, MNRAS, 275, 429

Ma X., et al., 2020, MNRAS, 493, 4315

Mackey A. D., Gilmore G. F., 2003a, MNRAS, 338, 85

Mackey A. D., Gilmore G. F., 2003b, MNRAS, 338, 120

Mackey A. D., Gilmore G. F., 2003c, MNRAS, 340, 175

Mackey A. D., Gilmore G. F., 2004, MNRAS, 355, 504

Madau P., Rees M. J., 2001, ApJ, 551, L27

Madau P., Haardt F., Dotti M., 2014, ApJ, 784, L38

Magorrian J., et al., 1998, AJ, 115, 2285

Majewski S. R., Skrutskie M. F., Weinberg M. D., Ostheimer J. C., 2003, ApJ, 599, 1082

Marconi M., 2012, Memorie della Societa Astronomica Italiana Supplementi, 19, 138

Marconi A., Risaliti G., Gilli R., Hunt L. K., Maiolino R., Salvati M., 2004, MNRAS, 351, 169

Martin C. L., 1999, ApJ, 513, 156

Martínez-Vázquez C. E., et al., 2019, Monthly Notices of the Royal Astronomical Society, 490, 2183

Masci F. J., et al., 2019, PASP, 131, 018003

Massari D., Helmi A., 2018, A&A, 620, A155

McConnachie A. W., 2012, AJ, 144, 4

McConnell N. J., Ma C.-P., 2013, ApJ, 764, 184

McLaughlin D. E., van der Marel R. P., 2005, ApJS, 161, 304

Mckinney W., 2010, Data Structures for Statistical Computing in Python, doi:10.25080/Majora-92bf1922-00a

Medina G. E., et al., 2018, ApJ, 855, 43

Merloni A., et al., 2010, ApJ, 708, 137

Merloni A., et al., 2014, MNRAS, 437, 3550

Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, ApJ, 524, L19

Mortlock D. J., et al., 2011, Nature, 474, 616

Narayanan V. K., Spergel D. N., Davé R., Ma C.-P., 2000, ApJ, 543, L103

Nelson D., et al., 2015, Astronomy and Computing, 13, 12

Newberg H. J., et al., 2002, ApJ, 569, 245

Ni Y., Di Matteo T., Feng Y., Croft R. A. C., Tenneti A., 2018, preprint, (arXiv:1806.00184)

Odenkirchen M., et al., 2001, ApJ, 548, L165

Oosterhoff P. T., 1939, The Observatory, 62, 104

Pace A. B., Li T. S., 2019, ApJ, 875, 77

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peebles P. J. E., 1968, ApJ, 153, 1

Peng C. Y., 2007, ApJ, 671, 1098

Peng E. W., et al., 2008, ApJ, 681, 197

Penzias A. A., Wilson R. W., 1965, ApJ, 142, 419

Percival W. J., et al., 2001, MNRAS, 327, 1297

Pezzulli E., Valiante R., Schneider R., 2016, MNRAS, 458, 3047

Pezzulli E., Volonteri M., Schneider R., Valiante R., 2017, MNRAS, 471, 589

Planck Collaboration et al., 2016, A&A, 594, A13

Plummer H. C., 1911, MNRAS, 71, 460

Pontzen A., Tremmel M., Roth N., Peiris H. V., Saintonge A., Volonteri M., Quinn T., Governato F., 2017, MNRAS, 465, 547

Porciani C., 2016, MNRAS, 463, 4068

Quinn T., Katz N., Efstathiou G., 1996, MNRAS, 278, L49

Regan J. A., Haehnelt M. G., 2009, MNRAS, 396, 343

Reina-Campos M., Kruijssen J. M. D., Pfeffer J. L., Bastian N., Crain R. A., 2019, MNRAS, 486, 5838

Reines A. E., Volonteri M., 2015, ApJ, 813, 82

Richtler T., 2003, The Globular Cluster Luminosity Function: New Progress in Understanding an Old Distance Indicator. pp 281–305, doi:10.1007/978-3-540-39882-0_15

Ricotti M., Gnedin N. Y., 2005, ApJ, 629, 259

Romano-Diaz E., Shlosman I., Trenti M., Hoffman Y., 2011, ApJ, 736, 66

Roth N., Pontzen A., Peiris H. V., 2016, MNRAS, 455, 974

Rubin V. C., Ford W. K. J., Thonnard N., 1978, ApJ, 225, L107

Rucinski S. M., 1998, AJ, 115, 1135

Salvaterra R., Haardt F., Volonteri M., Moretti A., 2012, A&A, 545, L6

Sánchez-Janssen R., et al., 2019, ApJ, 878, 18

Sandage A., 2004, AJ, 128, 858

Scargle J. D., 1982, ApJ, 263, 835

Schaye J., et al., 2015, MNRAS, 446, 521

Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103

Schleicher D. R. G., Palla F., Ferrara A., Galli D., Latif M., 2013, A&A, 558, A59

Schulze A., Wisotzki L., 2011, A&A, 535, A87

Schulze A., Wisotzki L., 2014, MNRAS, 438, 3422

Searle L., Zinn R., 1978, ApJ, 225, 357

Sesar B., et al., 2010, ApJ, 708, 717

Sesar B., et al., 2014, The Astrophysical Journal, 793, 135

Sesar B., et al., 2017, AJ, 153, 204

Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337

Shapley H., 1938, Nature, 142, 715

Shapley H., 1939, Proceedings of the National Academy of Science, 25, 565

Shapley H., Sawyer H. B., 1927, Harvard College Observatory Bulletin, 849, 11

Sijacki D., Springel V., Haehnelt M. G., 2009, MNRAS, 400, 100

Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575

Silk J., Rees M. J., 1998, A&A, 331, L1

Simion I. T., Belokurov V., Irwin M., Koposov S. E., 2014, MNRAS, 440, 161

Simion I. T., Belokurov V., Koposov S. E., Sheffield A., Johnston K. V., 2018, MNRAS, 476, 3913

Simion I. T., Belokurov V., Koposov S. E., 2019, MNRAS, 482, 921

Simon J. D., 2018, ApJ, 863, 89

Simon N. R., Clement C. M., 1993, ApJ, 410, 526

Sirko E., 2005, ApJ, 634, 728

Smith H. A., 1995, Cambridge Astrophysics Series, 27

Soszyński I., et al., 2019, Acta Astron., 69, 321

Springel V., 2005, MNRAS, 364, 1105

Springel V., Hernquist L., 2003, MNRAS, 339, 289

Springel V., Di Matteo T., Hernquist L., 2005, MNRAS, 361, 776

Springel V., et al., 2008, MNRAS, 391, 1685

Stark D. P., Ellis R. S., Bunker A., Bundy K., Targett T., Benson A., Lacy M., 2009, ApJ, 697, 1493

Steinborn L. K., Dolag K., Hirschmann M., Prieto M. A., Remus R.-S., 2015, MNRAS, 448, 1504

Stetson P. B., Hesser J. E., Smecker-Hane T. A., 1998, PASP, 110, 533

Stetson P. B., Fiorentino G., Bono G., Bernard E. J., Monelli M., Iannicola G., Gallart C., Ferraro I., 2014, PASP, 126, 616

Stringer K. M., et al., 2021, ApJ, 911, 109

Tenneti A., Mandelbaum R., Di Matteo T., 2016, MNRAS, 462, 2668

Tenneti A., Di Matteo T., Croft R., Garcia T., Feng Y., 2018, MNRAS, 474, 597

Thoul A. A., Weinberg D. H., 1996, ApJ, 465, 608

Torrealba G., et al., 2015, MNRAS, 446, 2251

Torrealba G., Koposov S. E., Belokurov V., Irwin M., 2016a, MNRAS, 459, 2370

Torrealba G., et al., 2016b, MNRAS, 463, 712

Torrealba G., et al., 2018, MNRAS, 475, 5085

Torrealba G., Belokurov V., Koposov S. E., 2019a, MNRAS, 484, 2181

Torrealba G., et al., 2019b, MNRAS, 488, 2743

Treister E., Schawinski K., Volonteri M., Natarajan P., Gawiser E., 2011, Nature, 474, 356

Treister E., Schawinski K., Volonteri M., Natarajan P., 2013, ApJ, 778, 130

Tremaine S., et al., 2002, ApJ, 574, 740

Tremmel M., Karcher M., Governato F., Volonteri M., Quinn T. R., Pontzen A., Anderson L., Bellovary J., 2017, MNRAS, 470, 1121

Treu T., Woo J.-H., Malkan M. A., Blandford R. D., 2007, ApJ, 667, 117

Ueda Y., Akiyama M., Hasinger G., Miyaji T., Watson M. G., 2014, ApJ, 786, 104

Van Rossum G., Drake F. L., 2009, Python 3 Reference Manual. CreateSpace, Scotts Valley, CA

VandenBerg D. A., Brogaard K., Leaman R., Casagrand e L., 2013, ApJ, 775, 134

VanderPlas J. T., 2018, ApJS, 236, 16

Vasiliev E., 2019, MNRAS, 484, 2832

Veljanoski J., et al., 2015, MNRAS, 452, 320

Verner G., Demers S., Hardy E., Kunkel W. E., 1981, AJ, 86, 357

Villegas D., et al., 2010, ApJ, 717, 603

Vito F., et al., 2014, MNRAS, 441, 1059

Vito F., et al., 2016, MNRAS, 463, 348

Vito F., et al., 2018, MNRAS, 473, 2378

Vivas A. K., Zinn R., 2006, The Astronomical Journal, 132, 714

Vivas A. K., et al., 2001, ApJ, 554, L33

Vivas A. K., et al., 2004, AJ, 127, 1158

Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2013, MNRAS, 436, 3031

Vogelsberger M., et al., 2014, MNRAS, 444, 1518

Volonteri M., 2010, A&A Rev., 18, 279

Volonteri M., Reines A. E., 2016, ApJ, 820, L6

Volonteri M., Stark D. P., 2011, MNRAS, 417, 2085

Wagner C., Schmidt F., Chiang C. T., Komatsu E., 2015, MNRAS, 448, L11

Walker M. G., Mateo M., Olszewski E. W., Gnedin O. Y., Wang X., Sen B., Woodroofe M., 2007, ApJ, 667, L53

Walsh S. M., Willman B., Jerjen H., 2009, AJ, 137, 450

Wang R., et al., 2010, ApJ, 714, 699

Wang M. Y., et al., 2019, ApJ, 875, L13

Waskom M., et al., 2016, seaborn: v0.7.0 (January 2016), doi:10.5281/zenodo.45133, http://dx.doi.org/10.5281/zenodo.45133

Waters D., Wilkins S. M., Di Matteo T., Feng Y., Croft R., Nagai D., 2016a, MNRAS, 461, L51

Waters D., Di Matteo T., Feng Y., Wilkins S. M., Croft R. A. C., 2016b, MNRAS, 463, 3520

Wenger M., et al., 2000, A&AS, 143, 9

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Wilkins S. M., Feng Y., Di-Matteo T., Croft R., Stanway E. R., Bouwens R. J., Thomas P., 2016, MNRAS, 458, L6

Wilkins S. M., Feng Y., Di Matteo T., Croft R., Lovell C. C., Waters D., 2017, MNRAS, 469, 2517

Wilkins S. M., Feng Y., Di Matteo T., Croft R., Lovell C. C., Thomas P., 2018, MNRAS, 473, 5363

Willott C. J., 2011, ApJ, 742, L8

Willott C. J., et al., 2010, AJ, 140, 546

Wils P., Lloyd C., Bernhard K., 2006, MNRAS, 368, 1757

Wilson A. G., 1955, PASP, 67, 27

Woo J.-H., Treu T., Malkan M. A., Blandford R. D., 2006, ApJ, 645, 900

Wu X.-B., et al., 2015, Nature, 518, 512

Yajima H., Khochfar S., 2016, MNRAS, 457, 2423

Yang J., et al., 2019, AJ, 157, 236

Zentner A. R., Bullock J. S., 2003, ApJ, 598, 49

Zhang W., Woosley S. E., Heger A., 2008, ApJ, 679, 639

Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, Journal of Open Source Software, 4, 1298

Zwicky F., 1933, Helvetica Physica Acta, 6, 110

Zwicky F., 1942, Phys. Rev., 61, 489

van Albada T. S., Baker N., 1973, ApJ, 185, 477

van de Weygaert R., Bertschinger E., 1996, MNRAS, 281, 84

van den Bergh S., 2006, AJ, 131, 304

van der Walt S., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13, 22