Theses and Dissertations

December 2022

Chasing Transients: Constructing Local Galaxy Catalogs for Electromagnetic Follow-Up of Gravitational Wave Events

Chaoran Zhang
University of Wisconsin-Milwaukee

CHASING TRANSIENTS: CONSTRUCTING LOCAL GALAXY CATALOGS FOR ELECTROMAGNETIC FOLLOW-UP OF GRAVITATIONAL WAVE EVENTS

by

Chaoran Zhang

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Physics

at

The University of Wisconsin-Milwaukee

December 2022

ABSTRACT

CHASING TRANSIENTS: CONSTRUCTING LOCAL GALAXY CATALOGS FOR
ELECTROMAGNETIC FOLLOW-UP OF GRAVITATIONAL WAVE EVENTS

by

Chaoran Zhang

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Patrick Brady & David Kaplan, PhD

Gravitational waves (GWs) provide a new window for observing the universe which is not possible using traditional electromagnetic (EM) wave astronomy. The coalescence of compact object binaries, such as black holes (BHs) and neutron stars (NSs) generates "loud" GW signals that are detectable by the LIGO-Virgo-KAGRA (LVK) GW Observatory. If the binary contains at least one NS, there is a possibility that an observable EM counterpart will be launched during and/or after the merger. The first joint detection of GW radiation (GW170817) and its EM counterpart (AT 2017gfo) greatly extended our understanding of the universe in many fields, such as the birth of heavy elements and the independent measurements of Hubble constant; it also announced the era of multi-messenger astronomy (MMA). As the early EM emission in optical and infrared, known as the kilonova (KN) fades rapidly in hours to days, prompt follow-up of the counterpart is essential. However, it is challenging due to the large localizations of the GW events and numerous distant false positives enclosed. Since GW170817, unprecedented EM follow-up efforts have been made during LVK's latest third observing run (O3), but no EM counterparts were identified. In this dissertation, I present the details of my work with the Global Relay of Observatories Watching Transients Happen (GROWTH) collaboration in where I helped improve the efficiency of EM follow-up to GW events by constructing galaxy catalogs in the local universe.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GW   Gravitational wave

EM   Electromagnetic

BH   Black Hole

NS   Neutron Star

MMA   Multi-Messenger Astronomy

KN   Kilonova

GR   General Relativity

BBH   Binary Black Hole

BNS   Binary Neutron Star

NSBH   Neutron Star - Black Hole

KNe   Kilonovae

IR   Infrared

GRB   Gamma-ray Burst

sGRB   Short Gamma-ray Burst

ISM   Interstellar Medium

UV   Ultraviolet

NUV   Near-ultraviolet

FUV   Far-ultraviolet

AGN   Active Galactic Nucleus

Y_e Electron Fraction

ML   Machine Learning

RF   Random Forest

Photo-z Photometric Redshifts

FOV   Field of View

PSF   Point Spread Function

S/G   Star/Galaxy Separation

SFR   Star Formation Rate

TPR   True Positive Rate

FPR   False Positive Rate

ACKNOWLEDGMENTS

I started to pursue my PhD in physics as a bachelor of engineering. The lack of solid background in physic made my PhD journey more difficult than average, but I never regret for starting this journey. I would like to thank Paul Lyman and Daniel Agterberg for teaching me undergraduate level quantum mechanics and electrodynamics, which served as concrete foundations for my graduate level courses. Many thanks to other UWM physics faculty for their excellent coursework, such as electrodynamics by Alan Wiseman, theory of relativity and gravitational-wave by Jolien Creighton, high-energy/relativistic astrophysics by David Kaplan/John Friedman and a lot more.

I am grateful to my doctoral advisors — David Kaplan and Patrick Brady — for their constant guidance, support and understanding. I will never forget the numerous moments I walked into David's office for questions. I will never forget the help he generously offered when I met difficulties. He taught me how to figure out the essence of a question; I always found a question much simpler after discussing with him. Thank you

CHAPTER 1

Introduction

Albert Einstein published his relativistic theory of gravity–general relativity (GR; Einstein, 1916) in 1916. In this theory, gravity is considered the curvature of space-time, light is also bent by the curved space-time. The theory predicted a wave-like perturbation in the space-time metric traveling at the speed of light caused by the acceleration of massive objects, named gravitational waves (GWs). The strength of GWs are so weak when they propagate to the earth that they were never directly detected by scientists until the second decade of this century. The first indirect evidence of GWs was observed in 1974 from a binary pulsar system PSR1913+16 (Hulse & Taylor, 1975); during years of observations, the orbital period of the binary decayed at a rate precisely matching that predicted by GR if they were emitting GWs (Taylor & Weisberg, 1989).

The ground-based GW detectors that directly measured the GWs for the first time, the Laser Interferometer Gravitational-wave Observatory (LIGO; Abramovici et al., 1992) was initially constructed in the 1990s, and began its operation in the mid 2000s. In parallel, a similar GW detector with slightly shorter arms, the Virgo observatory (Acernese et al., 2006) was built in Italy. LIGO and Virgo evolved into the Advanced LIGO (aLIGO; Aasi et al., 2015) and Advanced Virgo (AdV; Acernese et al., 2014) later by upgrading their sensitivity with almost an order of magnitude improvement in 2015 and 2017 respectively. These GW observatories are designed to detect GWs at the high frequency end of the GW spectrum, which come from the coalescence of compact object binaries like stellar/intermediate mass black holes (BHs) and neutron stars (NSs). Figure 1.1 shows the full GW spectrum and sources at different frequencies, LIGO/Virgo listens to the colliding binaries that emit high frequency GWs.
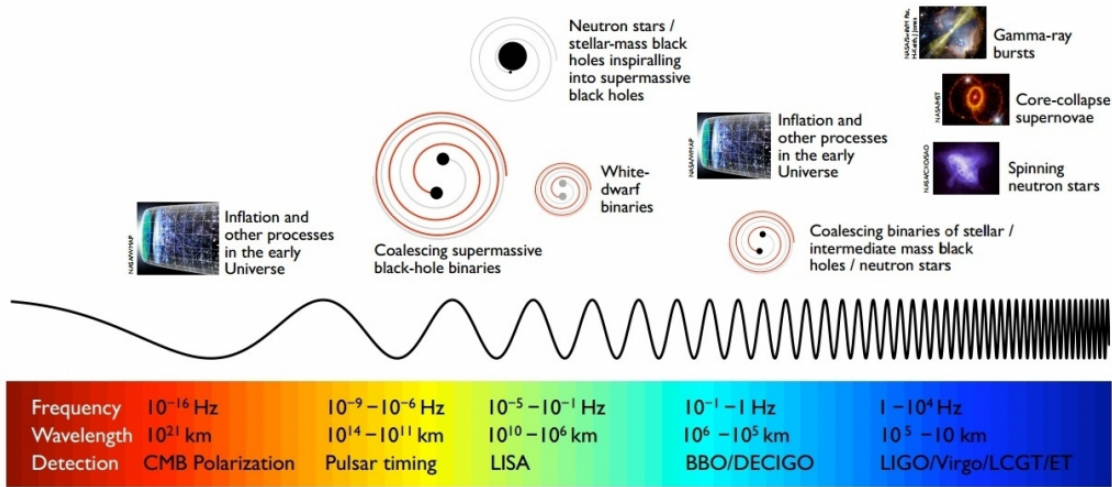
Figure 1.1: The GW spectrum and potential sources. Image credit: LIGO-India.

The basic mechanism of the GW observatories is identical to Michelson interferometers (Aasi et al., 2015). The L-shaped detectors consist of two orthogonal arms with length (L), in which the laser is emitted and reflected. At the end of each arm, a mirror is suspended and used as a test mass. If gravitational waves pass the arms, they produce differential change in the length of the arms which will also change the travel paths of the beams in each arm, thus altering the interference pattern of the beams. By monitoring the patterns, scientists can derive the amplitude and frequency of the GWs. LIGO has two detector sites with 4 kilometer (km) long arms in Hanford, Washington and Livingston, Louisiana in the United States. Virgo has a detector with ~ 3 km arms near Pisa in Italy. They combine as a global network for searching GWs. The three detectors are shown in Figure 1.2. The Kamioka Gravitational Wave Detector (KAGRA; Kagra Collaboration et al., 2019) is another GW detector built in Japan, which is expected to join the GW network in the next observing run.
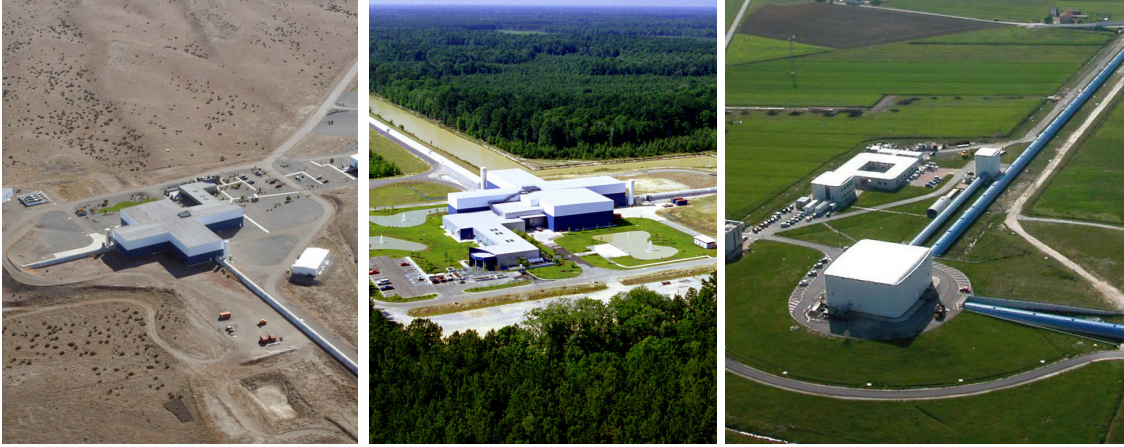
Figure 1.2: Views of the gravitational wave observatories LIGO-Hanford (left), LIGO-Livingston (center) and Virgo (right). Image credits: LIGO Caltech.

1.1 GRAVITATIONAL WAVES: RIPPLES IN SPACE-TIME TO BE DETECTED

Gravitational waves (GWs) produce periodic perturbations in the space-time which travel at the speed of light. GWs are transverse waves whose oscillations are orthogonal to the direction of propagation. In weak-gravity fields, GWs can be linearly approximated by small first order perturbations on the flat space-time metric, written as,

$$g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}, \tag{1.1}$$

where $\eta_{\alpha\beta} = \mathrm{diag}(-c^2, +1, +1, +1)$ is the flat space-time metric, and $h_{\alpha\beta}$ are small perturbations to the flat space-time metric produced by GWs. For simplicity, consider a plane GW propagating along the z-axis. The GWs have two independent polarizations: the plus (+) polarization and the cross ($\times$) polarization. The form of a plane GW propagating in the positive z-direction in the transverse-traceless gauge is:

$$h_{\alpha\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{1.2}$$

where h_+ and $h_\times$ denote the plus and cross polarization respectively. The line element of the space-time metric $g_{\alpha\beta}$ is thus:

$$ds^2 = -c^2 dt^2 + (1 + h_+)dx^2 + (1 - h_+)dy^2 + 2h_\times dx dy + dz^2. \tag{1.3}$$

To see the effect of GWs passing through, imagine two test masses in free fall on the xy-plane, separated by distance L_* in the unperturbed flat space-time. One test mass at the origin $(0, 0, 0)$ and the other at $(L_*, 0, 0)$. For simplicity, I consider only the plus polarization, the cross polarization is nothing different but rotated by $45°$. As proved in Chapter 16 of Hartle (2021), the coordinate positions of the test masses remain unchanged as the GW passes to first order in the amplitude of the wave. However, the distance between the two test masses changes with time even if their coordinate separation does not. The distance $L(t)$ along the x-axis can be computed by:

$$L(t) = \int_0^{L_*} [1 + h_{xx}(t)]^{1/2}\, dx \approx L_*[1 + \frac{1}{2}h_{xx}(t)] = L_*[1 + \frac{1}{2}h_+(t)]. \tag{1.4}$$

Subtract $L(t)$ by L_*, we get the change in distance, $\delta L(t) = \frac{1}{2}h_+(t)L_*$. Thus the fractional strain produced by the GW is:

$$\frac{\delta L(t)}{L_*} = \frac{1}{2}h_+(t). \tag{1.5}$$

The GW interferometers set up two perpendicular arms to construct a Michelson in-

terferometer. A schematic diagram of the LIGO interferometer is shown in Figure 1.3. An incident beam of light from a laser is split into two beams by the beam splitter, the beams then travel along the perpendicular arms and are reflected by the mirrors suspended at the end which serve as test masses. The beams from the two arms meet again at the beam splitter to create an interference pattern at the photodetector.
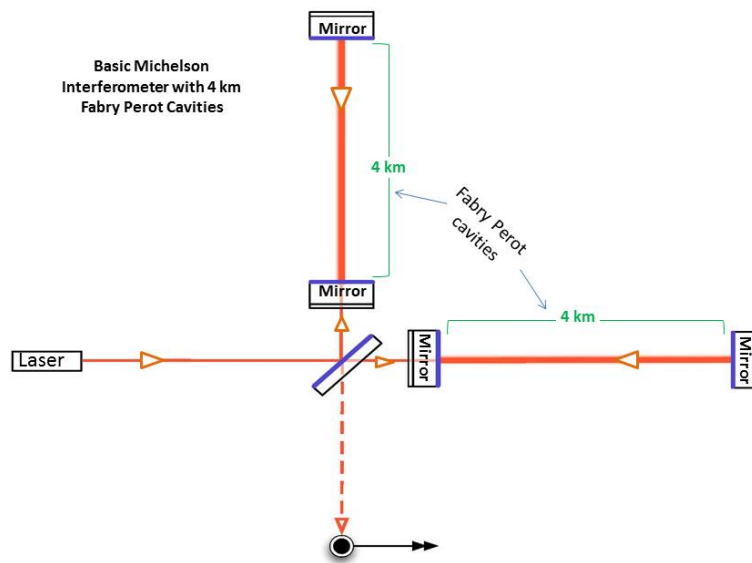
Figure 1.3: Schematic diagram of the LIGO interferometer. Image credits: Caltech/MIT/LIGO Lab.

In Figure 1.3, assume the horizontal arm is along the x-axis, and the vertical arm is along the y-axis, the lengths of the arms are $L_{(x)}(t)$ and $L_{(y)}(t)$. The interference pattern is determined by:

$$\Delta L(t) = L_{(x)}(t) - L_{(y)}(t) = h_+(t)L_*. \tag{1.6}$$

The beams will produce a constructive pattern if the path difference is equal to an integer number of the laser wavelengths, and a destructive pattern if the path difference is equal

to an odd number of half-wavelengths. LIGO monitors the interference pattern to track the strain and hence the GW signal,

$$\frac{\Delta L(t)}{L_*} = h_+(t). \tag{1.7}$$

The longer arm length is, the higher sensitivity that the interferometer can achieve. Since the physical length of the arms are limited by the facilities, LIGO built a Fabry Perot cavity in each arm to bounce the laser about 300 times before it being merged with the beam from the other arm. This configuration greatly increased the effective arm length from $4\,\mathrm{km}$ to $1200\,\mathrm{km}$, which significantly improved the sensitivity.

The initial LIGO reached a GW strain sensitivity of $\sim 10^{-21}$ corresponding to a distance range of $\sim 30\,\mathrm{Mpc}$ for an optimally oriented binary neutron star (BNS) system comprised of a pair of $1.4 M_\odot$ stars (Abbott et al., 2009). The initial Virgo could detect signals from the same system at a closer distance of $\sim 8\,\mathrm{Mpc}$. Advanced LIGO and Advanced Virgo upgraded their detectors in 2015 and 2017, with a factor of ~ 10 improvement in overall strain sensitivity. The upgrade of aLIGO directly led to the first discovery of GWs from a binary black hole (BBH) merger in the same year.

1.2 DETECTIONS OF GRAVITATIONAL WAVES: NEW WINDOW TO THE UNIVERSE

Advanced LIGO started its first observing run (O1) after the upgrade in September 2015. On September 14, 2015 at 09:50:45 UTC, the aLIGO detectors observed GWs for the first time in history from the coalescence of a binary black hole (BBH) system at $\sim 400\,\mathrm{Mpc}$ (GW150914; Abbott et al., 2016b). This discovery opened a new window of observing the universe, and officially launched a new era of gravitational-wave astronomy. Two more BBH mergers, GW151012 and GW151226 were detected during O1 (Abbott et al., 2016a)

by aLIGO, while AdV was in commissioning period.

The second observing run (O2) started in November 2016, and AdV joined the observation in August 2017, a month before O2 reached its end. Almost immediately after AdV joined, on August 17, 2017, the LIGO-Virgo GW detector network discovered the first GW signal from a BNS merger, GW170817 (Abbott et al., 2017a) in our nearby universe at $\sim 40\,$Mpc. This is a particularly interesting and important event: electromagnetic (EM) counterparts were observed from gamma-rays, optical/infrared (IR) to radio/X-ray across the entire EM spectrum by global efforts spanning the astronomy community (Abbott et al., 2017c). The joint detection of GW170817 and its EM counterparts established another milestone, which announced the beginning of multi-messenger astronomy (MMA). O2 observed 7 more GW events from BBH mergers (Abbott et al., 2019).

The LIGO-Virgo third observing run (O3) operated from April 2019 to March 2020, resulting in the detection of 74 compact binary mergers (Abbott et al., 2021a; The LIGO Scientific Collaboration et al., 2021). Most of the observed events are BBH mergers with the expectation of no EM counterparts. O3 detected the second BNS coalescence at a distance of $\sim 159\,$Mpc (Abbott et al., 2020c) and the first confident GW signals from neutron star - black hole (NSBH) binaries (Abbott et al., 2021b). No counterparts were detected in the EM follow-ups to these NS mergers either.

1.3 MULTI-MESSENGER ASTRONOMY AND ELECTROMAGNETIC FOLLOW-UP

The compact binary mergers with a neutron star as one of the components are hypothesized to launch high-energy EM radiations across large range of wavelengths. Within seconds after the merger, a short duration (milliseconds to seconds) gamma-ray burst (GRB) is powered by a collimated ultra-relativistic jet created during the remnant being rapidly accreted onto the central compact object (Berger, 2011; Troja et al., 2016; Metzger

& Berger, 2012; Fernández & Metzger, 2016). Due to relativistic beaming, the on-axis short GRB (sGRB) emission is restricted to narrow angles; however, the weaker off-axis sGRB that is less beamed can be observed from wider directions (Metzger, 2019b; Burns, 2020). An isotropic thermal transient in optical and IR lasting days to weeks is later powered by the radioactive decay of r-process enhanced material, known as the kilonova (KN; Li & Paczyński, 1998; Metzger et al., 2010; Roberts et al., 2011; Tanaka & Hotokezaka, 2013; Grossman et al., 2014; Metzger, 2019a; Fernández & Metzger, 2016). As the matter ejected during the merger interacts with the interstellar medium (ISM) and decelerates, radio emission is produced by synchrotron radiation (Nakar & Piran, 2011; Piran et al., 2013; Hotokezaka & Piran, 2015; Hotokezaka et al., 2016; Fernández & Metzger, 2016). An X-ray afterglow is another possible emission produced by multiple mechanisms (Zhang, 2013; Sun et al., 2017; Kisaka et al., 2015; Gao et al., 2013; Fernández & Metzger, 2016). Figure 1.4 shows the predicted evolution of a NS merger as it undergoes different phases; a subset of the physical phenomena and their associated EM emissions are plotted as a function of time.

All types of emission described above were observed during the EM follow-up campaign of GW170817. Figure 1.5 shows the observed EM counterparts of GW170817 and their origins. Combining the EM observations with the GW observations, exciting scientific breakthroughs were made in many fields, such as tracing the progenitors of sGRBs (Abbott et al., 2017d) and understanding the nucleosynthesis of heavy elements (Kasliwal et al., 2017). GW170817 and its EM follow-up were of great success as a commencement of the MMA era, however, it is so far the only GW event whose EM counterpart was successfully identified. Benefiting from the public GW alerts in O3, the EM follow-up attempts became more active than ever. Such massive efforts all over the globe led to zero detection of EM counterparts, this fact reminds us how lucky we were in 2017, and at the
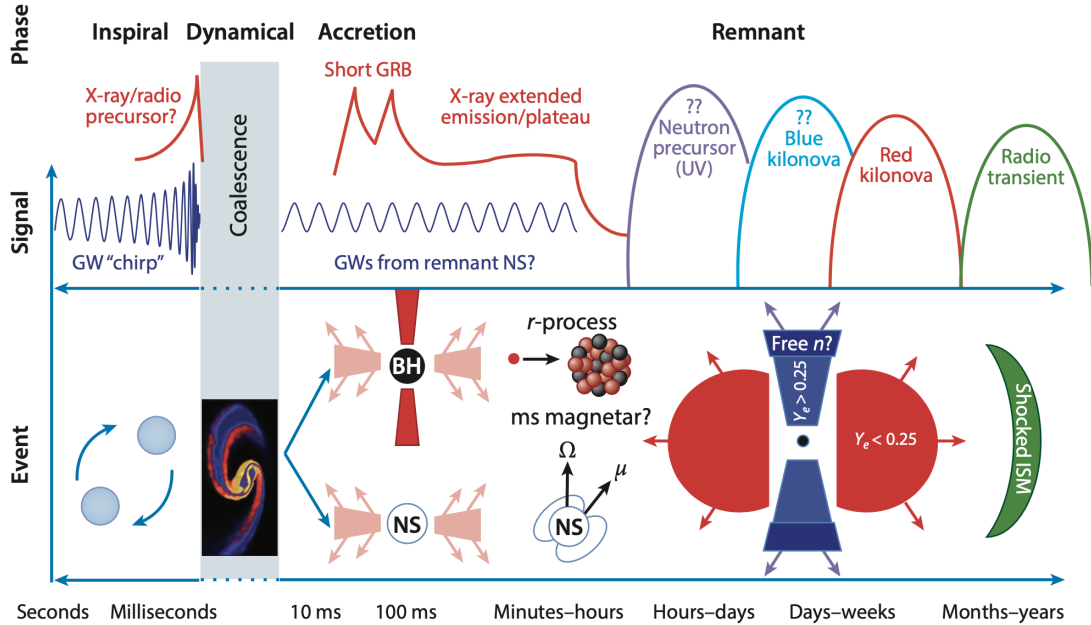
Figure 1.4: Predicted timeline of a binary neutron star (BNS) merger, showing the associated observational signatures and underlying physical phenomena. Image credits: Fernández & Metzger (2016) by Annual Reviews.

same time, how challenging EM follow-up could be.

Once a NS merger happened, the sGRB is expected to launch within seconds after the GWs and only last a timescale of milliseconds to seconds. Because the sGRB can only be detected by the passive GRB detectors, it is very difficult to localise the source promptly. Moreover, since the sGRB is beamed, and the observed luminosity drops quickly with increasing observing angle, there are chances that the sGRB will be missed with large observing angle. Thus the EM hunters choose to trace kilonovae (KNe) for the following reasons:

1. KNe launch shortly (hours) after the merger, they are early emissions that can be identified before the rise of other afterglows.

2. KNe are nearly isotropic, so their visibility is not limited by the direction of the jets.

9
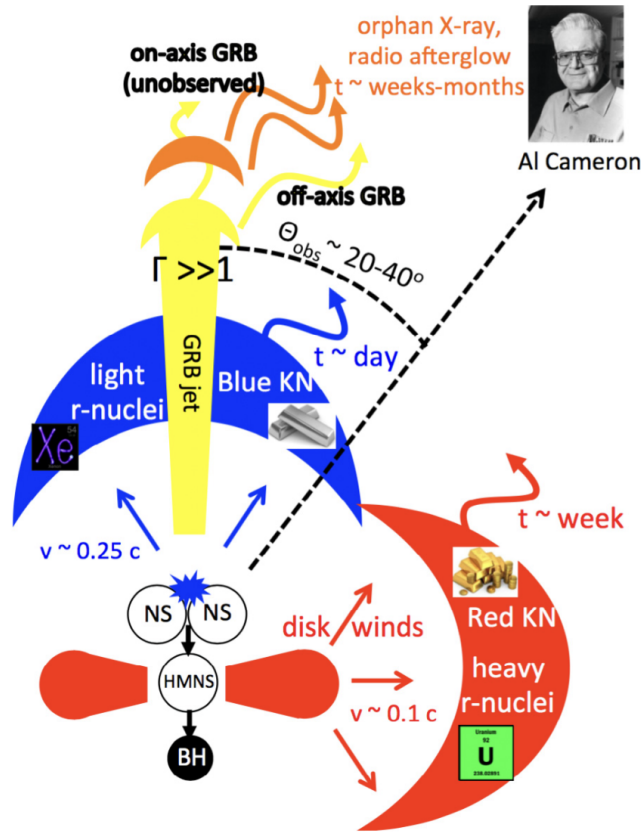
Figure 1.5: Schematic diagram showing the observed EM counterparts of GW170817, as viewed by an observer from the binary inclination angle $\theta_{obs} \approx 20°$-$40°$. Image credits: Metzger (2019b) by Annals of Physics.

3. KNe shine in optical and near-IR. There are large numbers of ground-based telescopes targeting these wavelengths can join the search by pointing the GW skymap in an active fashion.

4. The KN emissions themselves are valuable because of the abundant science can be derived from their photons, we want to observe them as early as possible.

To search for the EM counterparts, the astronomy community works closely with the LIGO-Virgo-KAGRA (LVK) GW collaboration. When LVK detects a compact binary coalescence, they will send a public alert within minutes to the astronomy community with

preliminary information of the event including the localization skymap, the source classes and the possible distance range of the event. If the binary contains at least one NS, the EM hunters will target the regions in the skymap depending on circumstances. The sky localization of the LVK GW events can range from $10\,\mathrm{deg}^2$ to more than $1000\,\mathrm{deg}^2$ (Abbott et al., 2020a); in O3 there were multiple events with skymaps larger than $2000\,\mathrm{deg}^2$. The large areas make the regions hard to cover with typical telescope fields-of-view in a reasonably short time (e.g., DECam, LCO, Gemini; Herner et al. 2020; Arcavi et al. 2017; Eikenberry et al. 2012). Moreover, the sky is noisy, there will be numerous background transients like supernovae and AGN flares (Metzger et al., 2013; Metzger & Berger, 2012; Nissanke et al., 2013) enclosed which make the identification more difficult. KNe rise and decay rapidly (Metzger & Berger, 2012; Metzger, 2019a; Abbott et al., 2017c; Kasliwal et al., 2017; Arcavi, 2018; Cowperthwaite et al., 2017). The optical component starts to fade in hours to days, thus the the candidate transients need to be quickly selected and the false positives need to be quickly eliminated to capture the emission from KNe in time. Since the GW detectors are only sensitive to the nearby universe, most of the galaxies which have been observed in the sky are not reachable by LVK. So when searching for the EM counterparts, only a tiny fraction of galaxies are potential host galaxies. If we were able to select only these galaxies in the sensitive volume beforehand along with their astrophysical properties, targeted searches prioritizing to the most probable host galaxy candidates become possible. Such searching strategies can narrow down the search area by a factor of 100 (Gehrels et al., 2016; Nissanke et al., 2013). This is the topic of this dissertation. I will discuss how can we select these galaxies to improve the efficiency of EM follow-ups.

1.4 DISSERTATION OUTLINE

The multi-messenger astronomy era has begun. The detection of GW events is becoming a routine. We have detected nearly 100 compact binaries so far. With a new round of upgrades to the LVK detectors, the detection frequency is expected to be higher in the upcoming fourth observing run (O4) and later. GW170817 revealed the power of combining GW and EM observations, the successful follow-ups of EM counterparts are more important than ever. However, due to the blockers in multiple aspects, the prompt identification of the host galaxies remains challenging. In this dissertation, I present my work with the Global Relay of Observatories Watching Transients Happen (GROWTH) collaboration to make EM follow-up efforts more promising and efficient.

The organization of the dissertation is as follows. In Chapter 2, I give details about selecting the galaxies in the nearby universe that the GW detectors are sensitive to and deriving their astrophysical properties. I constrain the redshift ranges of the galaxies using machine learning (ML) techniques and discuss how they can be utilized to optimize EM follow-ups. These galaxies are recorded in a galaxy catalog with positions, photometry, redshifts and other information. In Chapter 3, I conclude by discussing the possible usage, limitations, and future work related to the catalog and EM follow-ups.

CHAPTER 2

Census of the Local Universe (CLU) II: Identifying Nearby Galaxies Using Machine Learning

2.1   INTRODUCTION

On September 14, 2015, the Advanced Laser Interferometer Gravitational-Wave Observatory (aLIGO) detected gravitational waves (GWs) for the first time from a binary black hole (BBH) system merging (Abbott et al., 2016b), which opened a new window of observing the universe. Two years later, on August 17, 2017, the LIGO-Virgo detector network observed the first gravitational wave signal from a binary neutron star (BNS) merger, GW170817 (Abbott et al., 2017a) during its second observing run (O2). Theories predict that BNS mergers and some neutron star-black hole (NSBH) mergers will emit electromagnetic (EM) waves as well as GWs, with possibilities including: a prompt short-duration gamma-ray burst (sGRB; Berger, 2011; Troja et al., 2016; Metzger & Berger, 2012; Fernández & Metzger, 2016); early optical and infrared emission from neutron-rich ejecta known as a kilonova (Li & Paczyński, 1998; Metzger et al., 2010; Roberts et al., 2011; Tanaka & Hotokezaka, 2013; Grossman et al., 2014; Metzger, 2019a; Fernández & Metzger, 2016); delayed radio emission (Nakar & Piran, 2011; Piran et al., 2013; Hotokezaka & Piran, 2015; Hotokezaka et al., 2016; Fernández & Metzger, 2016) and others (e.g., X-ray; Zhang, 2013; Sun et al., 2017; Kisaka et al., 2015; Gao et al., 2013; Fernández & Metzger, 2016). A global effort was made to search and follow-up the EM counterpart of GW170817 across the electromagnetic spectrum (Abbott et al., 2017c). The first joint detection of gravitational and electromagnetic radiation led to great scientific breakthroughs in many fields, such as the progenitors of short Gamma-ray Bursts (sGRBs; Abbott et al., 2017d),

our understanding of the nucleosynthesis of heavy elements (Kasliwal et al., 2017), and an independent measurement of Hubble constant (Abbott et al., 2017b).

In LIGO-Virgo's latest observing run (O3), 74 candidate GW events were detected with 39 of them from the first half of O3 (O3a; Abbott et al., 2021a) and 35 from the second half (O3b; The LIGO Scientific Collaboration et al., 2021). The vast majority of these events are BBH mergers. In O3a, the detectors observed the second GW event consistent with a BNS coalescence (Abbott et al., 2020c), and for the first time, we discovered binary systems with significantly asymmetric mass ratios (Abbott et al., 2020b,d). The O3b observed no BNS mergers, but delivered the first confident observations of NSBH binaries (Abbott et al., 2021b). Unfortunately, no EM counterparts were detected during O3. A lot more BNS/NSBH mergers (Abbott et al., 2020a) and potentially their EM counterparts are expected to be observed in the future, and such multi-messenger astronomy (MMA) studies will significantly improve our understanding of the universe.

However, identifying the EM counterparts and hence the host galaxies of GWs is challenging due to the poor localization of the GW events. The sky localization of the LIGO-Virgo-KAGRA (LVK) GW events can range from less than $10 \deg^2$ to more than $1000 \deg^2$ (Abbott et al., 2020a), which may be significantly larger than the field-of-view (FOV) of many EM follow-up instruments (e.g., DECam, LCO, Gemini; Herner et al. 2020; Arcavi et al. 2017; Eikenberry et al. 2012). Large numbers of exposures and long integration times are required to fully cover the sky map. Moreover, the large numbers of false positive transients (e.g., background supernovae, M dwarf flares, AGN flares; see Metzger et al. 2013; Metzger & Berger 2012; Nissanke et al. 2013) enclosed make the identification even harder. Given the rapid evolution (hours to days) of the EM counterparts in optical (Metzger & Berger, 2012; Metzger, 2019a; Abbott et al., 2017c; Kasliwal et al., 2017; Arcavi, 2018; Cowperthwaite et al., 2017), the candidate transients need to be quickly se-

lected and the false positives need to be quickly eliminated so that deeper photometry or spectroscopy could be performed before the kilonova fades.

As most galaxies in the sky map are distant, only a tiny fraction of them are within the distance that LVK is sensitive to. Targeted search of likely host galaxies can potentially reduce the number of pointings by a factor of 100 (Gehrels et al., 2016). By simply restricting distances of the candidate galaxies, the reduction in false positives can be as high as a factor of $\sim 10^3$ without tiling strategies (Nissanke et al., 2013). In addition, multi-telescope networks can take advantage of galaxy catalogs to deploy sophisticated tiling strategies, which will be able to image a higher fraction of an event's probability area (Coughlin et al., 2019). Galaxy catalogs can also help prioritize follow-up efforts, e.g., target massive galaxies with higher priority (Ducoin et al., 2020). Hanna et al. (2014) discussed how the utility of a galaxy catalog can improve the probability of successfully imaging the host galaxy by prioritizing the pointings. To achieve the goals above, a complete and accurate catalog for galaxies in the local universe is required to maximize the scientific returns from GW detections.

Previous efforts have been made towards such spectroscopic galaxy catalog for EM follow-up (Kopparapu et al., 2008; White et al., 2011), but both catalogs were limited to $100\,\mathrm{Mpc}$ while the distant BNS mergers that could be detected by LIGO-Virgo in O3 were as far as $\sim 200\,\mathrm{Mpc}$. Other spectroscopic galaxy surveys have added more secure distance for new galaxies (e.g., SDSS, 6dFRGS, ALFALFA; Alam et al. 2015; Jones et al. 2009; Haynes et al. 2011), but still do not fill the vacancy in the local universe. The Galaxy List for the Advanced Detector Era (GLADE; Dálya et al., 2018, 2022) greatly extends the volume to well above $200\,\mathrm{Mpc}$ and achieves high completeness by joining multiple surveys and catalogs, but a large fraction of its distance information are photometric redshifts (photo-z). The photo-z at low redshifts is likely overestimated (see Sec. 2.6 for detailed

discussion), causing a loss of galaxies identified in the targeted volume. There are other efforts trying to compile galaxy catalogs (e.g., HECATE; Kovlakas et al. 2021), and to develop searching and ranking tools (e.g., Salmon et al. 2020, 2021) for EM follow-up. However, they do not add previously unknown distance information to our knowledge.

The Census of the Local Universe (CLU) catalog (Cook et al., 2019) aims to provide the most complete list of galaxies with distance information in the LVK sensitivity volume by: 1) carefully compiling galaxies with known distances and redshifts from existing galaxy databases, referred as CLU-compiled; 2) estimating distance constraints by finding Hα emission-line galaxies from redshift 0 to 0.047 ($\sim 200\,\mathrm{Mpc}$), referred as CLU-Hα; and 3) combining photometry from narrow-band (CLU-Hα) and broad-band surveys across large range of EM wavelengths to establish a machine learning (ML) model for constraining redshift ranges for the galaxies with no previous distances, referred as CLU-ML. In this paper, we introduce the CLU-ML catalog. This work is not the first effort towards a complete galaxy catalog with distance in large areas using photometry, given the expense of spectroscopy. We discuss more backgrounds and the uniqueness of CLU-ML in Sec. 2.6.

The CLU-ML catalog covers $\sim 3\pi\,\mathrm{sr}$ of the sky (north of $\delta = -30°$). It will provide distance constraints to the galaxies in its footprints and identify galaxies in the $\sim 200\,\mathrm{Mpc}$ local universe. Aside from GW host galaxies, CLU can also be used to search for extreme emission-line galaxies, such as blue compact dwarfs (BCDs; Kunth & Östlin, 2000; Cairós et al., 2010) in the local universe and more distant green peas (Cardamone et al., 2009): extreme emission-line galaxies at intermediate redshifts ($0.11 \leq z \leq 0.4$) whose strong [O III] emissions-lines give them green colors. CLU, with its large sky coverage and unique data-set will provide a rich sample of these extreme galaxies, which will lead to better statistics of evolutionary trends and help test star formation theories.

This chapter is organized as follows. In Section 2.2, we summarize the data used then introduce our source catalog and training set. In Section 2.3, we present our algorithms and build the ML model. In Section 2.4, we evaluate the performance of our method with different metrics. In Section 2.5, we explore the behavior of our method with a specific classification threshold, and compare it with the GLADE catalog. In Section 2.6, we discuss the motivation behind our method, the use of the output catalog and the possible future work. Finally, we draw our conclusions in Section 2.7.

2.2 CONSTRUCTION OF CLU-ML DATA

Here we describe the construction of the CLU-ML data-set, which combines archival broad-band optical and infrared (IR) photometry with CLU-Hα narrow-band imaging to identify galaxies within 200 Mpc. We also describe a list of sources that have measured spectroscopic redshifts as "ground truth", which we use both as a training set for our machine learning model and to evaluate its performance.

2.2.1 External Surveys and Catalogs

In addition to our proprietary Hα data, there are two external surveys and catalogs spanning optical to mid-infrared wavelengths that we use. These are briefly described below.

2.2.1.1 Pan-STARRS1

The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al. 2016) PS1 survey imaged the northern sky ($\delta > -30°$) in five broad-band optical filters (g, r, i, z, y). We downloaded the photometric data for all PS1 data release 1 (DR1) detections in the stacked images from the PS1 StackObjectView table (Flewelling et al.,

[2020](#)) through the Mikulski Archive for Space Telescopes (MAST)[1]. There are a total of 3.48 billion sources including duplicates. Because the goal of this work is to identify galaxies in the local volume of our universe, we use a PS1 point source catalog (PS1-PSC; Tachibana & Miller 2018) to separate stars and galaxies and to eliminate transients and spurious sources. The PS1-PSC catalog uses a machine learning model to classify a subset of the 3.48 billion sources from PS1 as either resolved extended objects or unresolved point sources. The sources in PS1-PSC are required to pass two criteria that validate the sources as real (not artifacts), leaving ∼1.5 billion unique sources in the catalog. PS1-PSC provides a probabilistic ranking for those sources with a score of 0 corresponding to extended objects and 1 corresponding to point sources. The classifications are based on morphological properties measured from the PS1 stacked images, and evaluations of the results show that this method delivers better star/galaxy separation than other techniques such as a cut on the difference between the i-band point spread function (PSF) magnitude and Kron (Kron, 1980) magnitude discussed in Farrow et al. (2014) and elsewhere. We discuss this catalog in more details in Sec. 2.2.3 and Sec. 2.3.2.

After de-duplication and cross-matching to PS1-PSC, the PS1 table works as our primary source detection table and provides the primary optical broad-band photometry.

2.2.1.2 WISE

The *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al., 2010) mapped the entire sky in four mid-infrared (mid-IR) bands with central wavelengths at $3.4, 4.6, 12$ and $22\,\mu$m (W1, W2, W3, and W4 respectively), with angular resolution of $\approx 6''$ for W1–W3 and $12''$ for W4. We use the Source Catalog from the AllWISE Data Release (Cutri et al., 2021), which contains information for over 747 million objects.

---

[1]https://archive.stsci.edu.

2.2.1.3 GALEX?

The *Galaxy Evolution Explorer* (*GALEX*; Bianchi, 2009, 2014; Bianchi et al., 2017) imaged the sky in the near-ultraviolet (NUV; $\lambda = 1770\text{-}2730\,\text{Å}$) and far-ultraviolet (FUV; $\lambda = 1350\text{-}1780\,\text{Å}$) bands. Eventually *GALEX* observed 77% of the sky at various depths in at least one band.

We initially considered utilizing UV photometry in our model, but gave up because most galaxies in PS1 do not have associated UV measurements in *GALEX*.

2.2.2 CLU-Hα Survey

The Census of the Local Universe (CLU) emission-line (Hα) galaxy survey (Cook et al., 2019) is a narrow-band survey that covers the northern sky ($\delta > -20°$, although the two reddest filters avoid the Galactic plane, $|b| < 3°$). The survey aims to constrain galaxy distances out to 200 Mpc ($z = 0.047$) by imaging $\sim 3\pi\,\text{sr}$ of the sky with four contiguous narrow-band filters, Hα1, Hα2, Hα3, Hα4, as part of the Intermediate Palomar Transient Factory (iPTF; Law et al., 2009). The first filter (Hα1) is centered near the $z = 0$ Hα emission line ($\lambda = 6563\,\text{Å}$), with subsequent filters covering redder (and hence more redshifted) wavelengths, with the last filter (Hα4) covering Hα out to $z = 0.047$ (i.e., $\sim$200 Mpc). CLU-Hα uses standard narrow-band color selection criteria (Bunker et al., 1995) to search for nearby galaxies, where the distance is constrained by the filter that contains the excess flux.

To help fill in chip gaps and compensate for a missing CCD, at least 3 dithered 60-s images were taken in each filter for the survey, although some fields had considerably more (see Cook et al. 2019 for details). Previous work focused on source detection and association in each individual image. Here, to provide maximum depth we co-added the multiple exposures in each filter using `swarp` (Bertin et al., 2002). First we used the

individual photometry to determine photometric zero-point variations between each observation, and then we weighted each image by the inverse of the flux density variance (i.e., inverse variance weighting) before co-adding.

To illustrate the motivation of the usage of all the surveys above spanning a wide range of wavelengths and how CLU-Hα constrains the distances of galaxies, we plot the spectra of three different types of galaxies and the band-passes of all the filters in Figure 2.1. We plot the simulated spectra [2] of an early-type galaxy, a late-type galaxy and an intermediate galaxy which are all redshifted to a redshift of 0.02, on top of the transmissions of the filters. The late-type galaxy has emission lines at various wavelengths, which can be utilized to trace the redshift (see insert of Figure 2.1). Although for photometric surveys we do not have spectroscopic data to match emission lines in general, we can identify the existence of the Hα emission line in the narrow-band Hα filters for the nearby galaxies to infer their possible redshift ranges. If the Hα emission line is present in one of the four Hα filters, a flux excess and hence a color excess will be observed compared to the adjacent narrow-band or continuum filters (Cook et al., 2019). The other types of galaxies do not have obvious emission lines, but their spectra still show certain features across a wide wavelength range such as a UV continuum from young stars or a mid-IR excess from warm dust, thus their photometry in multiple broad-band filters contains implicit information about their redshifts. Our CLU-ML method combines the photometry in broad-band and narrow-band filters, and uses ML techniques to constrain the redshift of these galaxies. A summary of all the surveys above is in Table 2.1.

To take full advantage of the deep PS1 optical survey for source discovery, and to measure Hα photometry for as many sources as possible, we used forced photometry technique for Hα measurements instead of independent source finding, more details are

---

[2]We use the `fsps` spectrum templates from the `EAZY photometric redshift code` github repository: https://github.com/gbrammer/eazy-photoz/tree/master/templates/fsps_full.
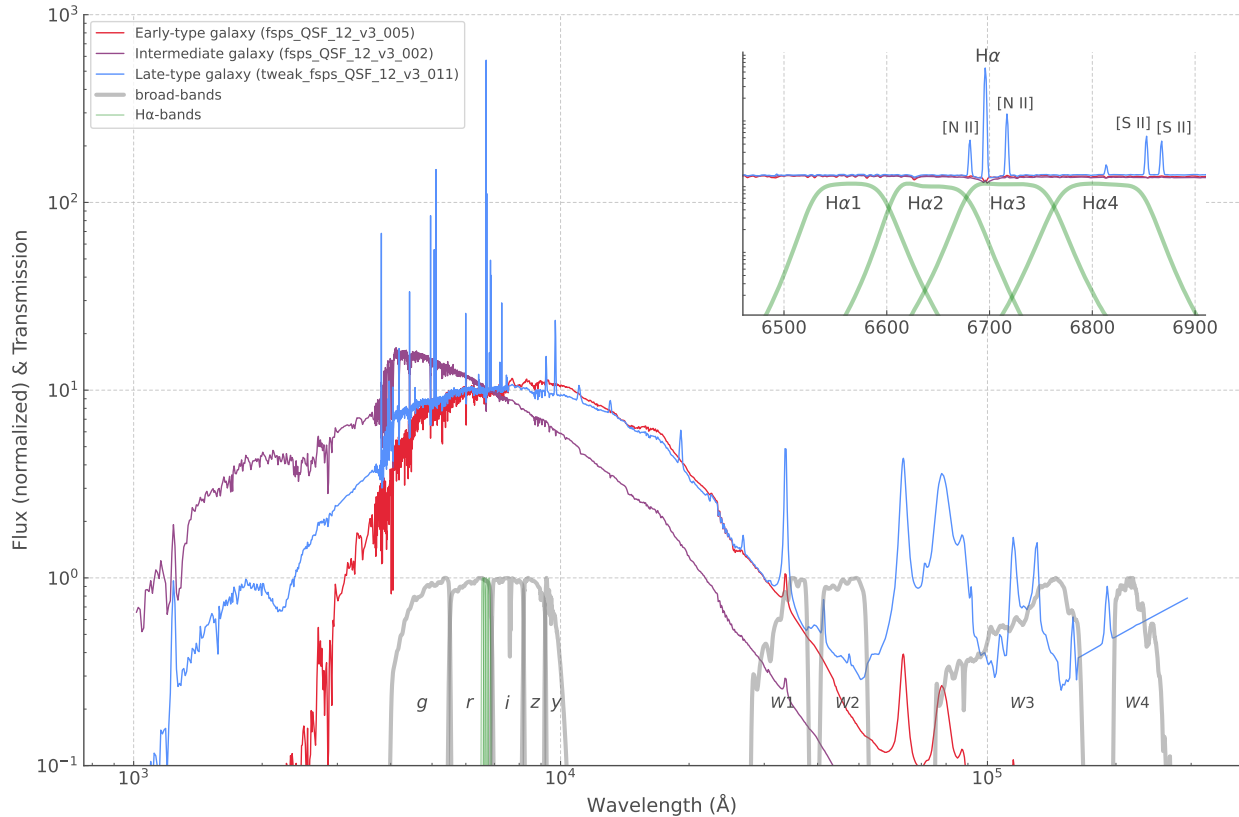
Figure 2.1: Illustration of CLU-ML methodology. We show spectra of three different types of galaxies overlaid on the band-passes of all filters used here. The spectra are redshifted by 0.02, an amount in the middle of our desired range. The gray curves represent the transmissions of the broad-band filters from the optical to the mid-IR, while the green curves represent the transmissions of the four narrow-band Hα filters. The red, blue and purple lines show the spectra of an early-type galaxy with no emission lines, dust, or UV emission; a late-type galaxy with significant nebular emission lines and dust emission; and an intermediate galaxy with no emission lines but with significant UV emission from a young stellar continuum. All spectra are normalized at $6600\,\text{Å}$. The inset shows a zoom around the Hα filters. In the case of the late-type galaxy, the redshifted Hα emission line in one of the Hα filters would cause a color excess, which CLU-Hα uses to constrain the distances of galaxies, but note that other emission lines ([N II], [S II]) can complicate the classification (Metzger et al., 2013).

discussed in Sec. 2.2.3. Since the PS1 optical images are much deeper than the stacked Hα

images, the forced photometry measurements are limited by the depth of the Hα images.

Most of the faint sources do not have strong enough fluxes for significant Hα detections,

<div align="center">Summmury of Surveys</div>

Survey Name	Source Number (#)	Filter Name	Filter λ (Å)	Filter $\Delta\lambda$ (Å)	Limiting Magnitude (mag)	Resolution (″)	References
Pan-STARRS	3.48×10^9	g	4881.5	1256.3	23.3	1.3	1, 3
		r	6198.4	1404.4	23.2	1.2	
		i	7549.3	1296.7	23.1	1.1	
		z	8701.4	1034.3	22.3	1.0	
		y	9509.8	628.2	21.4	1.0	
WISE	7.48×10^8	$W1$	34655.2	6357.9	19.2	6.1	1, 4
		$W2$	46443.0	11073.2	18.8	6.4	
		$W3$	132156.4	62758.0	16.4	6.5	
		$W4$	222228.8	47397.3	14.5	12.0	
CLU-Hα	N/A (forced photometry)	Hα1	6584.2	76.1	19.1	2.0	2
		Hα2	6663.7	77.9	19.2	2.0	
		Hα3	6730.9	90.1	19.3	2.0	
		Hα4	6822.1	92.1	19.4	2.0	

Table 2.1: The properties of all the surveys utilized, where the columns present the survey name, number of sources contained, filter name, central wavelength, FWHM, 5σ limiting AB magnitude, angular resolution, and references, from left-to-right.
References. (1) Rodrigo et al. 2012, Rodrigo & Solano 2020; (2) Cook et al. 2019; (3) Chambers et al. 2016, Flewelling et al. 2020; (4) Wright et al. 2010, Cutri et al. 2012, Cutri et al. 2021.

thus their magnitude measurements are assigned as 3σ upper limits. We plot the Hα magnitude and the significance of Hα flux as a function of r-band magnitude in Figure 2.2. From the r-Hα scatter plot panel (left), it can be easily seen that the Hα magnitudes of the faint sources ($r \gtrsim 20$ mag) no longer scale linearly with the r-band magnitudes. This is due to these sources approaching the limiting depth of the Hα images, thus their magnitudes are largely upper limits, shown as the flat tail. The right panel displays the cumulative fraction of sources that have significant ($> 3\sigma$) Hα detection as a function of r-band magnitude, where the fraction drops rapidly as the r magnitude approaches and passes 20 mag. For sources fainter than 22 mag, nearly none of them have significant Hα fluxes.

2.2.3   Source Catalog

With all the data from the different surveys available, we must first combine them into a single multi-wavelength source catalog. As discussed above we use PS1 as our primary
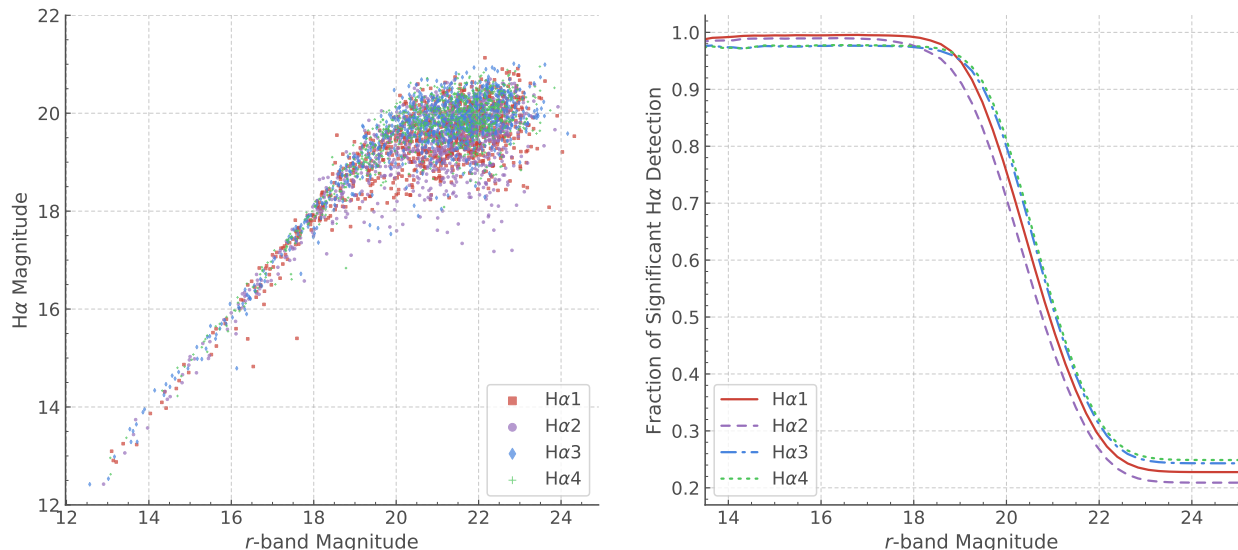
Figure 2.2: Illustration of the depth of Hα images. The left panel shows the forced photometry Hα magnitudes of 1000 sources versus their r-band magnitudes in the four Hα filters respectively. The Hα filter 1–4 are represented by red square, purple dot, blue diamond and green plus. The faint sources tend to measure upper limits in the Hα images, which is shown by the flat tail above $r \approx 20$ mag. The right panel plots the cumulative fraction of sources that have significant ($> 3\sigma$) Hα detection as a function of r-band magnitude. The four Hα filters are colored in red, purple, blue and green respectively.

catalog, as it has deeper images and higher angular resolution than the CLU-Hα data. We augment it with *WISE* data to fill out the spectral energy distribution at mid-IR, and then add CLU-Hα data. This catalog includes both galaxies and stars: while we select only those sources classified by the PS1-PSC, we do not yet limit those classifications.

We start with the PS1 `StackObjectView` table, which contains all sources detected from their stacked images. However, a large fraction of them are duplicates or spurious. We can remove the duplicates and some of the poor detections by matching the unique `objID` for multiple entries and requiring that a `primaryDetection == 1` entry exists. The PS1-PSC catalog applies two more criteria to remove most of the spurious detections as well as transients:

1. Only sources from the `StackObjectView` table with `nDetections` $\geq$ 3 are se-

lected.

2. Non-unique sources (i.e., if a single `objID` corresponds to multiple rows with `primaryDetection == 1`) are excluded.

For the remaining sources PS1-PSC keeps only the `primaryDetection == 1` row as the final source to remove duplicates. We apply all the criteria by:

1. Keeping only one row with `primaryDetection == 1` for every single `objID` in the `StackObjectView` table, as indicated.

2. Cross-matching the table against the PS1-PSC catalog with the unique `objID`, selecting only the intersection and adding the PS1-PSC probabilistic classification for each surviving source.

The steps taken result in a catalog with unique entries which is a subset of the PS1 `StackObjectView` table that contains likely real stars and galaxies, and includes their star/galaxy probabilistic classifications.

Then we cross-match the *WISE* catalog to the PS1 sources above. For every PS1 source, we search its associations in *WISE* within a $5.0''$ radius, associate the nearest source found and update its photometry to the source catalog.

Finally, we add CLU-Hα photometry to this catalog. In contrast to Cook et al. (2019), we did not perform independent source finding and association on the CLU-Hα data. Rather, we use the positions and aperture sizes from the PS1 `StackObjectView` catalog to perform forced photometry in the stacked narrow-band images described above. We use both fixed aperture sizes with radii of $1.0''$,$2.0''$,$2.5''$,$4.0''$,$5.0''$, as well as the r-band Kron aperture, to match the reported PS1 photometry. This is repeated in all 4 CLU-Hα filters.

For comparison with Cook et al. (2019) we also computed the "color sigma" (hereafter Csig) Σ_{ab}, which is a metric from standard narrow-band color selection methods that indicates the significance of a color excess in filter a compared to filter b (also see Bunker et al. 1995):

$$\Sigma_{ab} \equiv \frac{c_a - c_b}{\delta_{ab}}, \tag{2.1}$$

where c_a is the detected counts in filter a and:

$$\delta_{ab} = \sqrt{\pi r^2 (\sigma_a^2 + \sigma_b^2)}, \tag{2.2}$$

is the combined uncertainty between filters a and b, r is the aperture radius, and σ_a is the individual uncertainty in the counts of filter a.

The significance of the narrow-band colors (Σ) represents the number of standard deviations a given color is above that of random scatter for a source with zero color, and can be used to identify emission lines for a given object. The δ term defined in Eqn.(2.2) describes the total color error expected given the standard deviation of sky counts in each image, and grows exponentially large at fainter source magnitudes while approaches zero color at the brightest magnitudes (see Figure 5 of Cook et al., 2019). However, this curve does not account for steep continuum sources or other measurement errors, as can be seen in the scatter of bright sources (typically 0.03 mag). Consequently, we define and apply a second Σ cut based on the standard deviation of bright continuum sources. Thus, the final Σ value of a given object will depend on its brightness, where Σ for faint and bright objects is based on the scatter in the sky fluctuations and the standard deviation in colors for bright stars, respectively. This piece-wise selection has been used in many previous narrow-band studies (e.g., Bunker et al., 1995; Sobral et al., 2009; Stroe & Sobral, 2015; Khostovan et al., 2020).

25

The final source catalog contains ~ 1.5 billion sources and occupies ~ 4 TB space.

2.2.4 Ground Truth

To get a clean training set for our machine learning algorithm, and to understand the performances of the algorithms, we gather a set of galaxies with known spectroscopic redshifts in a sub-region of sky from the Galaxy And Mass Assembly survey (GAMA; Liske et al., 2015; Baldry et al., 2018) as our "ground truth" of the local universe. The GAMA survey used the AAOmega multi-object spectrograph (Saunders et al., 2004; Smith et al., 2004; Sharp et al., 2006) on the Anglo-Australian Telescope (AAT) to observe the spectra of $\sim 300,000$ galaxies down to $r = 19.8$ mag over $\sim 286\,\mathrm{deg}^2$. This was then combined with previous spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS; York et al., 2000; Eisenstein et al., 2011; Alam et al., 2015), the 2dF Galaxy Redshift Survey (2dFGRS; Colless et al., 2001, 2003) and the Millennium Galaxy Catalogue (MGC; Driver et al., 2005) to build a large and complete spectroscopic dataset. Various other target selection criteria such as the z-band and K-band constraints were applied (Baldry et al., 2010) in addition to the r-band magnitude limit to acquire spectra for more objects. The *Herschel* Astrophysical Terahertz Large Area Survey (H-ATLAS; Eales et al., 2010) made extraordinary contribution to the GAMA survey by providing a large number of selected survey targets.

GAMA splits their sky coverage into 5 regions, G02, G09, G12, G15 and G23, each of which is $\sim 60\,\mathrm{deg}^2$. Regions G09, G12 and G15 are three equatorial regions with the highest completeness (98.5%; Baldry et al. 2018) among the five, and the data for these three regions are released in their data release 3 (DR3). Thus we use these three regions as our "ground truth".

To construct the ground truth catalog, we first downloaded the spectroscopic redshifts for all the galaxies in the three equatorial regions released in GAMA DR3, then applied a

few filters suggested by the GAMA team to keep only the galaxies with securely-derived redshifts, resulting in $118,438$ galaxies in total. As described in the previous section, we choose PS1 as our origin for optical photometry and source detection, so we cross-matched GAMA galaxies against PS1 to fetch optical photometry for these sources. Because PS1 is a significantly deeper survey than GAMA, in principle every single galaxy in GAMA should have a detection in PS1. Thus we decided to match the nearest neighbor from PS1 as the association. To find the best angular separation limit, we did our analysis and diagnostics on a $12 \deg^2$ subset in the G09 field. We searched for nearest neighbors with a sequence of angular separation thresholds and compared the numbers of associations (see Figure 2.3). When the threshold is greater than $4.0''$, 5267 out of 5269 galaxies are matched and the number remains constant until the threshold reaches $8.0''$. We eliminated $8.0''$ because after visual inspection, we found the extra one galaxy was a wrong match. The two GAMA galaxies not matched to PS1 were too close to a large bright source so were not detected by PS1. To be conservative, we decided to use $7.0''$ as our angular separation threshold for association.

After this matching, the distribution of separations of the GAMA positions from the PS1 positions follows a χ distribution with two degrees of freedom (i.e., a Rayleigh distribution), as expected. We also visually inspected ~ 1000 pairs of matched galaxies by plotting the GAMA positions and matching radius on the PS1 stacked images. From this we find very few incorrect associations, with a contamination of at most 0.365%. Overall, our simple cross-match threshold is able to provide us a fairly clean ground truth sample with an average $0.15''$ separation between the two surveys.

We applied the above matching threshold to all $118,438$ galaxies in the three GAMA equatorial regions and found matches for $118,432$ of them in PS1. Only 6 GAMA galaxies did not have an association. In addition to the optical photometry from PS1, we followed
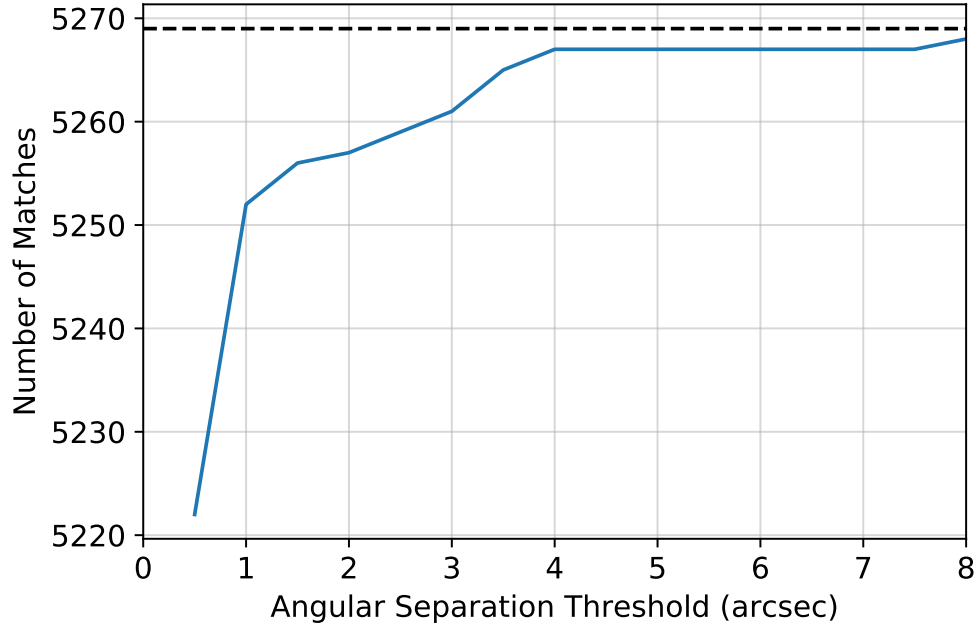
Figure 2.3: Number of associations between a subset of GAMA G09 galaxies and PS1 for different angular separation thresholds (blue line). There are 5269 GAMA galaxies total (black dashed line). Once the threshold is greater than $4.0''$, the number of matches reaches 5267 and remains constant until $8.0''$.

the procedures in Section 2.2.3 to incorporate additional mid-IR data from *WISE*, as well as forced Hα photometry from CLU-Hα.

While validating the Hα forced photometry, we discovered that in the Hα2 stacked images, a few areas located in the GAMA G09 region are masked due to bad image quality and hence do not have Hα2 measurements. Including these objects in our model could introduce a bias, as the number of masked objects in one filter significantly exceeds those in the other filters, is not common. Therefore we decided to remove those areas from our ground truth catalog, which reduced its size to $111,212$ galaxies.

2.3 MACHINE LEARNING CLASSIFICATION ALGORITHM

In this section we discuss the machine learning algorithm we use for constraining galaxy redshift ranges. We use a supervised machine learning algorithm, the Random Forest (RF; Breiman, 2001) classifier. To train the ML model we identify a training set, which contains objects with features and known labels based on external information. We can then train the RF classifier and use the model it builds to predict a redshift range for a given source.

As we are looking to classify galaxies only, we use a cut on the star/galaxy probabilistic ranking to limit our predictions on galaxies.

2.3.1 Discussion on Limitations of Brightness

In Sec. 2.2.3, we discussed construction of our source catalog. As this catalog is based on PS1 detections, its photometric depth is that of PS1, i.e., $r = 23.2$ mag for stacked images.

The goal of our work is to identify galaxies in the local universe (< 200 Mpc; $z < 0.047$). If galaxies had a fixed luminosity, because farther sources look fainter, there would be an apparent magnitude where above that there would be no local galaxies. Even in a more realistic scenario that the luminosities of galaxies follow some distribution, there will still be an apparent magnitude limit above which the sample is dominated by more distant galaxies. We do not have abundant redshift measurements for galaxies in the faint regime, so we are interested in how faint can most local galaxies reach.

To determine this limit we cannot use our GAMA-based ground truth, as it is only complete to $r = 19.8$ mag, compared to $r = 23.2$ mag for PS1. Instead we use the zCOS-MOS spectroscopic redshift survey (Lilly et al., 2007), in particular its zCOSMOS-bright survey which focuses on $z < 1$ galaxies. This survey selects targets with apparent AB magnitudes 15.0 mag $< I < 22.5$ mag and classified as galaxies to construct the input catalog (Koekemoer et al., 2007; Leauthaud et al., 2007).

We downloaded the $20,689$ objects across the whole $1.7\,\mathrm{deg}^2$ COSMOS field from the zCOSMOS-bright target catalog. The catalog contains position, redshift, and confidence class information for every object: the confidence class indicates the confidence level of a measured spectroscopic redshift. Among the $20,689$ objects, $1,539$ of them do not have redshifts, which leaves $19,150$ objects with redshifts. While not every redshift measurement has the same level of reliability, we follow the instruction in Lilly et al. (2007) and limit our sample to only recommended confidence classes (very secure and completely secure redshifts with a few special scenarios) and obtain a total number of $17,358$ objects. Finally, in the $17,358$ objects, 708 are identified as stars at zero redshift, leaving $16,650$ galaxies with secure redshifts. To acquire r-band photometry of the galaxies, we did a positional cross-match between the zCOSMOS-bright samples and PS1 with a $5''$ separation threshold, leading to $16,521$ objects (and we performed the same quality checks described in Sec. 2.2.4 which suggested high reliability). Of those, $14,228$ of them have a measured r-band Kron magnitude in PS1, which are what we use for our evaluation sample. The ~ 2000 objects do not have Kron magnitude are faint in r-band.

Among these objects, only 36 of them are galaxies with $z < 0.047$, 12 of those have r_{Kron} between $21\,\mathrm{mag}$ and $22\,\mathrm{mag}$, the rest are brighter than $21\,\mathrm{mag}$. The faintest one has $r_{\mathrm{Kron}} = 21.95\,\mathrm{mag}$. We show the cumulative completeness of local ($z < 0.047$) and non-local ($z > 0.047$) galaxies in Figure 2.4. The cumulative completeness of the local galaxies increases rapidly and reaches 100% at $r = 22\,\mathrm{mag}$. The cumulative completeness of the non-local galaxies rises more slowly and achieves only $\sim 50\%$ at the same magnitude, it reaches 100% at $r = 24\,\mathrm{mag}$. This suggests as the apparent magnitude reaches fainter regime, the number of local galaxies decreases much faster than the non-local ones, and almost no local galaxies lie in the fainter than $22\,\mathrm{mag}$ regime.

Because the zCOSMOS field is small and the size of the evaluation sample — espe-
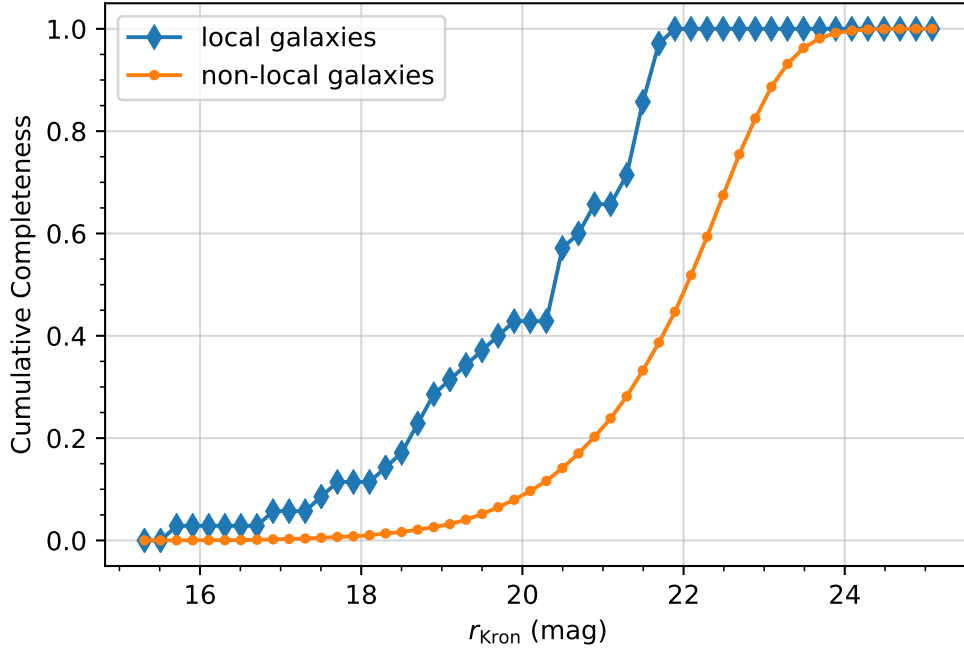
Figure 2.4: Cumulative completeness of the zCOSMOS-bright galaxies in the local ($z < 0.047$) and non-local ($z > 0.047$) volumes.

cially the number of local galaxies — is tiny, there is chance that in our sample some faint ($r > 22\,\text{mag}$) local galaxies are missed. With reasonable assumptions, we can estimate the probability of missing galaxies. We start with the $\Omega = 1.7\,\text{deg}^2$ zCOSMOS field where 67% of the objects have been observed spectroscopically (the sampling rate) and $\sim 80\%$ of those objects have successful redshift measurements for low redshifts (the redshift success rate). Given that we observed 0 galaxy with $r > 22\,\text{mag}$ in this field, we can estimate that fewer than 3 galaxies would have $r > 22\,\text{mag}$, or a fraction $< 3/14228 = 2.1 \times 10^{-4}$ at 95% confidence.

In addition to the r-band magnitude, our study in Sec. 2.2.2 shows, the forced photometry measurement is greatly limited by the depth of the Hα images. We likely measure Hα photometry in poor quality for faint objects thus make them harder to classify using a model trained with brighter objects (recall the vast majority of ground truth objects are

brighter than $20\,\mathrm{mag}$), which further implies we should focus our model on the relatively brighter galaxies.

2.3.2 Star/Galaxy Separation

As mentioned in Sec. 2.2.1.1, we use the PS1-PSC catalog to separate stars and galaxies. For our purpose, we want to eliminate as many stars as possible and keep all the galaxies for prediction. As no algorithm gives a perfect result, a trade off is inevitable.

The PS1-PSC catalog provides a star/galaxy probabilistic ranking instead of a single binary classification for the sources it contains, giving us the freedom to adjust the threshold for dividing the 2 types of sources. In Tachibana & Miller (2018), the authors report a table (Table 3) that presents the accuracies, false positive and true positive rates (FPR and TPR) of different separations when various probability thresholds are applied. For a binary classification, candidates are classified as either positives or negatives. Predictions can be categorized to four categories: true positive (TP), where the candidate is in the desired category ("positive") and the classifier correctly identified it as such; false positive (FP), where the candidate is not in the desired category but the classifier misidentified it; true negative (TN), where the candidate is correctly identified as not in the desired category; and false negative (FN), where the candidate is misclassified as not in the desired category. We define the TPR and FPR as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.3}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2.4}$$

In this context, positives denote point sources (stars) while negatives denote resolved

sources (galaxies), hence the TPR here corresponds to the fraction of stars that are correctly identified and FPR means the fraction of galaxies that are misclassified as stars. The authors demonstrate their thresholds by discussing the classifications in 3 different magnitude regimes: 1) rKronMag < 20; 2) rKronMag < 21; 3) all sources with no magnitude limits; with the accuracy increasing for brighter sources.

We use the thresholds in the rKronMag < 21 regime as our baseline because:

1. As discussed in Sec. 2.3.1, most local galaxies are likely in the rKronMag < 21 regime.

2. In the rKronMag < 21 regime and especially the rKronMag < 20 regime, the point sources dominate the distribution; in the fainter regime, the extended sources dominate.

So we want to achieve great separation in the rKronMag < 21 regime, and if so, the worse classification in fainter regime will have less impact on the overall performance due to lower proportion of stars in that regime given the fact that the model looses accuracy at selecting stars faster than galaxies in faint regime. After evaluating the performance of this classifier, we chose a threshold $\mathrm{ML\ score} = 0.406$ as our criterion to separate stars and galaxies in our source catalog. This threshold results in $\mathrm{TPR} = 0.980$ and $\mathrm{FPR} = 0.02$ for sources with rKronMag < 21. It balances the stars and the galaxies by eliminating most of the stars to reduce the contaminants while not losing many galaxies.

2.3.3 Machine Learning Model

In this section we discuss our machine learning model, including details such as design, feature choice, hyper-parameters and the final training set.

We use the RF method to build our model. RF is a supervised ensemble machine learning method that operates by combining the outputs of multiple decision trees trained with

the training set to predict a final classification or regression value for previously unseen targets. The use of random decision trees reduces overfitting that appears commonly in a decision tree method (Quinlan, 2014). RF makes use of the "bagging" technique (Breiman, 1996) to decrease the variance of a model while not increasing the bias, where only bootstraped samples of the training set are used to construct each tree in the forest. When each tree is constructed, only a random subset of all the features provided by the training set are selected as potential splitting criteria at the nodes of the tree. This process is applied to reduce the correlation of the trees in a forest. Once the forest is established, RF offers final predictions for a new candidate by averaging the predictions among the individual trees. Overall, RF is an accurate, low-bias, low-variance machine learning algorithm and has become popular in the astronomy community due to its natural suitability to astronomical data sets (Ivezić et al., 2014). We utilize the `Python scikit-learn` package (Pedregosa et al., 2011) for the RF algorithm in this study.

2.3.3.1 *Training and Test Sets*

As described in Sec. 2.2.4, we construct a ground truth catalog for model training and evaluation. Eventually the entire ground truth catalog will be used as the final training set to build the best ML model, but for evaluation purpose, we extract a subset of the ground truth catalog to serve as the test set. Therefore, the training set used for analysis in this paper is also a subset of the ground truth catalog. Note that the training set plus the test set do not necessarily equal the ground truth catalog, as some data may not be utilized depends on the goal of the analysis.

2.3.3.2 *Classifier Design*

We initially came up with 3 designs of the classifier, discussed below.

The first design we established is a straight-forward, standard RF classifier, hereafter, the "simple model". In this model, we divide all the galaxy candidates into 5 classes according to the associated redshift ranges of the $4\,\mathrm{H}\alpha$ filters, where the non-local galaxies ($z > 0.0471$) are assigned to class-0. For the galaxies in the local volume, the classes are: class-1 for candidates with $0 < z < 0.0094$ (Hα1), class-2 for candidates with $0.0094 < z < 0.0213$ (Hα2), class-3 for candidates with $0.0213 < z < 0.0325$ (Hα3), and class-4 for candidates with $0.0325 < z < 0.0471$ (Hα4). We train the model with a set of features we selected and construct the model based on the data (see details in Sec. 2.3.3.3). We then let the machine predict the probabilities (scores) for a test candidate to belong to each class.

For our final analysis, simply taking the probabilities of each class separately would not fully represent the data since class-0 corresponds to the non-local volume which has the largest physical volume by far. To partially mitigate this, and because our top-level question is whether or not a galaxy is nearby (and hence appropriate for EM followup), we sum the scores of class-1 to 4 as the overall probability of being local and compare it with the score of class-0 to discriminate between local objects and non-local objects. To account for different prior volumes and to fully utilize our model we consider the local vs. non-local probability as a score where we can set a threshold as for any other classifier, balancing completeness with contamination.

The second classifier ("2-step classifier") adds another step to help deal with the very uneven volumes for our training set. To do so, 1) we first figure out if a candidate is in the local volume or not; 2) if the candidate is in the local volume, we estimate its most possible filter (redshift range). Specifically, the first step divides the candidates coarsely between $z < 0.1$ and $z > 0.1$ regions. A threshold of $z = 0.1$ is chosen as the boundary because it is small enough to exclude most of the non-local candidates outside our narrow-band selection ($z > 0.0471$), but also provides sufficient tolerance to the local candidates when

dealing with uncertain predictions: we can adjust the threshold to guarantee that a very high fraction of real local candidates are in the $z < 0.1$ class. We train the RF classifier with samples that span the entire redshift range of the ground truth catalog, assign class-0 (same as the simple model) to candidates predicted as $z > 0.1$, then take the $z < 0.1$ class to the next step for further classification. The second step focuses on a finer classification for the nearby galaxies. Among all the candidates in this class, most of them are indeed closer than $z = 0.1$, but there is a "tail" passes over the boundary. To handle this, we introduce a tolerance Δz, defined as the amount of redshift that the nearest 70% of the objects that are classified in the $z < 0.1$ class but whose real redshifts are greater than 0.1 can exceed the $z = 0.1$ boundary. For example, if the most distant 30% of the objects in the tail ($z > 0.1$, across the boundary) have $z > 0.15$, we define $\Delta z = 0.15 - 0.1 = 0.05$. In the second step of this model, we train the classifier with samples that have redshifts $z < 0.1 + \Delta z$ with the complete five labels, then predict the scores for candidates in the $z < 0.1$ class with the same 5 sub-classes as above. Finally, we combine the results in the two steps as the final classification.

The third design is an extension of the "2-step classifier", but replacing the RF classifiers by RF regressors. We take the same two step strategy, but in every step we predict for redshift directly using a trained regressor instead of just determining a class. This is essentially a version of a photo-z technique, with more features spanning all wavelengths; eventually the predicted redshifts are translated to classes-0 to 4 to match the desired result. Unfortunately, the classification accuracy of this model is not comparable to the previous two models due to the large errors of photo-z at low redshifts (we will discuss more later in Sec. 2.6), which makes this regressor model not desirable.

We applied various different thresholds to the simple model and the 2-step classifier model to optimize their classification results and compare their performances. However,

with the same features and settings, the two models behave nearly identically. Therefore for simplicity, we have decided to use the simple model as our basic structure.

2.3.3.3   *Feature Selection*

When using machine learning, appropriate features lead to the success of a model (Liu & Motoda, 1998), thus extracting relevant and efficient features from the data is crucial. We start with the multi-band photometry from the optical and mid-IR surveys discussed above. We use the Kron, PSF and aperture-based (Chambers et al., 2016) magnitudes in $grizy$-bands from PS1, and also use them to construct optical colors for each candidate. We add the PSF major/minor axis full width at half maximum (FWHM) and the difference between Kron and PSF magnitudes in the optical bands from PS1 (Farrow et al. 2014 uses $\mathrm{Kron} - \mathrm{PSF}$ magnitudes to separate stars and galaxies) to include shape information. As described in Sec. 2.2.3, we measure magnitudes in the 4 Hα bands with 6 aperture sizes: Kron, $1.0''$, $2.0''$, $2.5''$, $4.0''$, $5.0''$. For each aperture, we construct the corresponding colors and compute the Csig values. We add the mid-IR colors from *WISE*. Finally, we use the PS1 and *WISE* photometry to construct a few optical-IR colors. All photometry are corrected for Galactic extinction using the SFD dust reddening map (Schlegel et al., 1998) and the Fitzpatrick (1999) dust extinction function (Fitzpatrick, 1999).

Of these features, many of the colors and Csigs are computed for multiple apertures and thus are correlated. It may also be that not every feature has a strong contribution to the model. Correlations between features could possibly reduce the accuracy of the model, and additionally, correlated features may split the importance to each other and lead to underestimates of the feature importance. Although RF methods are relatively insensitive to weak and correlated features (e.g., Richards et al. 2012), we can reduce the correlation of the features, and investigate if removing some of the features will lead to

better model performance.

We evaluated the correlation by the Pearson product-moment correlation coefficients (PPMCC) and visual inspection of the resulting plots. We find the optical/Hα colors and the Hα Csigs have high correlations between different apertures, so we only keep the Kron aperture for these features. A list of all the features used is in Table 2.2.

To investigate the combination of features that achieves best performance, we apply a procedure similar to the one employed in Miller et al. (2017). We build a series of RF models whereby we iteratively add one feature at a time in the order of the most important RF feature to the least important. The feature importance is measured by the built-in RF significance function. For each combination, we assess the accuracy, completeness, contamination and the TPR at a fixed $\mathrm{FPR} = 0.03$ of the model via a five-fold cross validation (CV) run on the training set[3], and this procedure is repeated three times to minimize the scatter; we take the average as the final performance estimate. Because our top-level target is to identify the local galaxies, we estimate the binary local/non-local classification, where the local candidates are positives and the non-local candidates are negatives. The completeness and contamination are defined as:

$$\mathrm{Completeness} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.5}$$

$$\mathrm{Contamination} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.6}$$

The completeness measures the fraction of all the local galaxies that are correctly selected by our classifier, while the contamination measures the fraction of the identified local candidates that are misclassified. For the series of models, the accuracies, defined as the frac-

---

[3]We randomly take 80% of the ground truth samples as training set and leave 20% for evaluation, which we then repeat 5 times.

Importance Ranking	Features	Description
1	rKronMag − wise1	Color between r-band Kron magnitude and *WISE* $W1$-band magnitude
2	rKronMag − wise2	Color between r-band Kron magnitude and *WISE* $W2$-band magnitude
3	iPSFMag − iKronMag	PSF magnitude − Kron magnitude in i-band
4	zPSFMag − zKronMag	PSF magnitude − Kron magnitude in z-band
5	**gKronMag**	g-band Kron magnitude
6	yPSFMag − yKronMag	PSF magnitude − Kron magnitude in y-band
7	rKronMag − iKronMag	Kron color between r-band and i-band
8	gKronMag − iKronMag	Kron color between g-band and i-band
9	**csig3_HaKron**	Csig of Hα3-band for fluxes in a Kron aperture
10	rPSFMag − rKronMag	PSF magnitude − Kron magnitude in r-band
11	rKronMag − wise3	Color between r-band Kron magnitude and *WISE* $W3$-band magnitude
12	**csig4_HaKron**	Csig of Hα4-band for fluxes in a Kron aperture
13	**rKronMag**	r-band Kron magnitude
14	gPSFMag − gKronMag	PSF magnitude − Kron magnitude in g-band
15	gKronMag − zKronMag	Kron color between g-band and z-band
16	**iKronMag**	i-band Kron magnitude
17	wise1 − wise2	Color between *WISE* $W1$-band and $W2$-band
18	**maxcsig_HaKron**	Maximum Csig among the 4 Hα-bands
19	**yKronMag**	y-band Kron magnitude
20	rKronMag − zKronMag	Kron color between r-band and z-band
21	HaKron_f3_mag − HaKron_f4_mag	Kron color between Hα3-band and Hα4-band
22	**zKronMag**	z-band Kron magnitude
23	gKronMag − rKronMag	Kron color between g-band and r-band
24	**ipsfMajorFWHM**	i-band major-axis FWHM
25	yKronMag − gKronMag	Kron color between y-band and g-band
26	**ipsfMinorFWHM**	i-band minor-axis FWHM
27	wise4 − wise1	Color between *WISE* $W4$-band and $W1$-band
28	**rpsfMinorFWHM**	r-band minor-axis FWHM
29	**gpsfMajorFWHM**	g-band major-axis FWHM
30	**gpsfMinorFWHM**	g-band minor-axis FWHM
31	**rpsfMajorFWHM**	r-band major-axis FWHM
32	iKronMag − zKronMag	Kron color between i-band and z-band
33	wise1 − wise3	Color between *WISE* $W1$-band and $W3$-band
34	HaKron_f4_mag − HaKron_f1_mag	Kron color between Hα4-band and Hα1-band
35	rKronMag − wise4	Color between r-band Kron magnitude and *WISE* $W4$-band magnitude
36	wise2 − wise3	Color between *WISE* $W2$-band and $W3$-band
37	**ypsfMinorFWHM**	y-band minor-axis FWHM
38	HaKron_f1_mag − HaKron_f3_mag	Kron color between Hα1-band and Hα3-band
39	**zpsfMinorFWHM**	z-band minor-axis FWHM
40	wise2 − wise4	Color between *WISE* $W2$-band and $W4$-band
41	rKronMag − yKronMag	Kron color between r-band and y-band
42	**ypsfMajorFWHM**	y-band major-axis FWHM
43	**zpsfMajorFWHM**	z-band major-axis FWHM
44	wise3 − wise4	Color between *WISE* $W3$-band and $W4$-band
45	iKronMag − yKronMag	Kron color between i-band and y-band
46	HaKron_f2_mag − HaKron_f3_mag	Kron color between Hα2-band and Hα3-band
47	HaKron_f2_mag − HaKron_f4_mag	Kron color between Hα2-band and Hα4-band
48	zKronMag − yKronMag	Kron color between z-band and y-band
49	**csig2_HaKron**	Csig of Hα2-band for fluxes in a Kron aperture
50	HaKron_f1_mag − HaKron_f2_mag	Kron color between Hα1-band and Hα2-band
51	**csig1_HaKron**	Csig of Hα1-band for fluxes in a Kron aperture

Table 2.2: The features that are selected to build the RF model ranked by their feature importance.

tion of objects classified correctly, are consistently high, with only two features needed to reach a classification accuracy within 1% of the maximum. We attribute this to the fact that the non-local galaxies completely dominate the samples (see Figure 2.7), which makes the accuracy almost independent of the local galaxies. This is why we introduce completeness, contamination and TPR in the evaluation, which focus on the local samples. The results of the performance analysis for different feature combinations is shown in Figure 2.5. Accuracy, completeness and contamination are evaluated at the default splitting $\text{threshold} = 0.5$, the TPRs are evaluated at thresholds that return fixed $\text{FPR} = 0.03$.
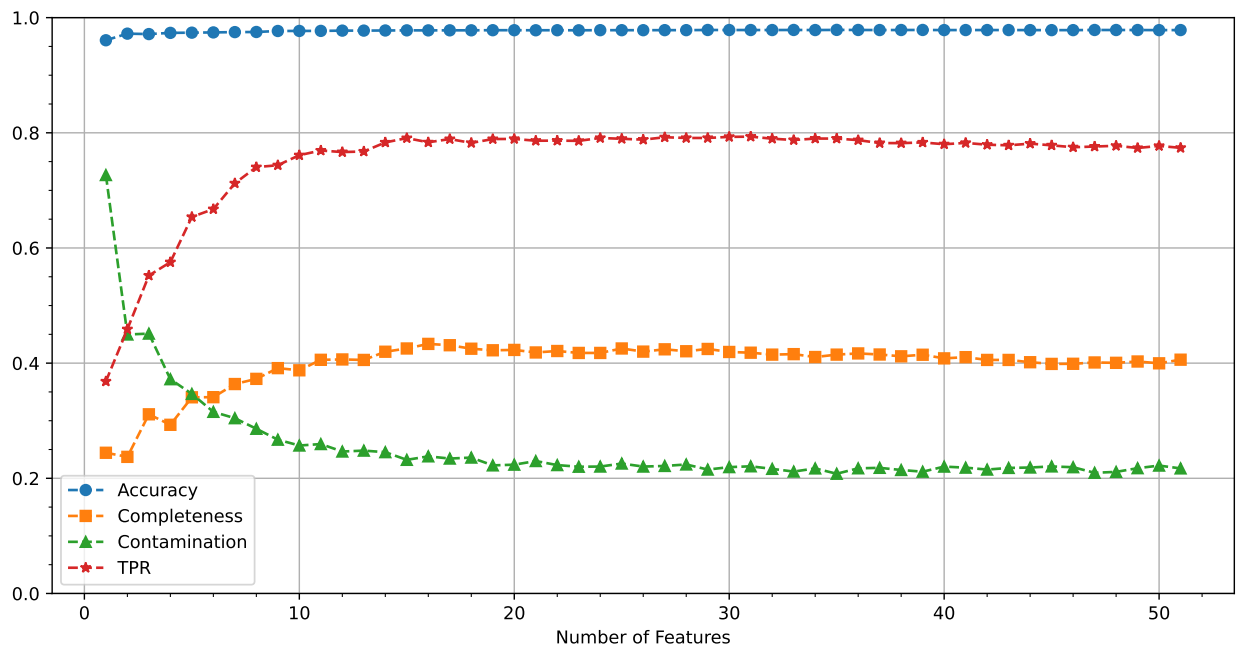
Figure 2.5: Accuracy, completeness, contamination and TPRs at fixed $\text{FPR} = 0.03$ for the model series with different number of features in the model. The blue circles show the overall accuracy of the model with certain number of features; the orange squares show the completeness of the local galaxies; the green triangles show the contamination over the selected local candidates; and the red stars show the TPRs given the fixed FPR. The accuracy is constantly high due to the dominating amount of non-local galaxies, the completeness rapidly reaches 43% and then slowly decreases while more features are added into the model, the contamination also quickly reaches a fairly low level and then gradually declines along with completeness. The TPR rises quickly and reaches the maximum 79% with 31 features, then slowly declines with more features.

The completeness grows rapidly and reaches about 40% when the 9 most important features are used. It keeps increasing gradually as a few more important features are added, then slowly decreases when the number of features gets larger. Contamination, in contrast, decreases monotonically, it quickly reaches a fairly low level and then gradually declines further as the number of features increases. It is hard to define a combination of features as having "the best performance" as many combinations perform very similarly. Considering the extremely small ratio between the numbers of local to non-local galaxies (the number of local galaxies is tiny compared to the non-local ones), the small variations we observe in contamination do not change the number of false positives (i.e., non-local galaxies in the predicted local candidates set) by a significant amount, especially when the completeness only varies slightly likewise. For this reason, we do not regard contamination as a major criterion for selecting the best features. As one major goal of the CLU catalog is to provide the host galaxies for following up the EM counterparts of the GW events, we would rather seek for a higher completeness by tolerating a few more misclassified non-local candidates than take the risk of missing the correct host galaxy. Hence we evaluate the TPRs at a reasonable fixed FPR, try to maximize the amount of local galaxies when the model selects the same amount of false positives. The model achieves the maximum $\mathrm{TPR} = 0.79$ with 31 features. We will further tune the settings of the model in Sec. 2.3.3.5, and select optimal thresholds in Sec. 2.4, the metrics in this section are computed from a simple model with default settings and threshold whose performance is purely determined by the features.

The model is not sensitive to the number of features fed when it is greater than 10, but more features can still improve. With the first 31 most important features, the model recovers the most local galaxies, thus we decided to use these 31 features for constructing the final model, the features can be found in Table 2.2.

2.3.3.4   Imbalanced Data

The local volume that we probe with our Hα filters (out to $200\,\mathrm{Mpc}$) occupies only a tiny portion of the entire volume observed by modern surveys. This leads to a result that the number of galaxies outside the local volume greatly surpasses the number inside. In our training set, the size of these non-local galaxies is more than 30 times larger, leaving us extremely imbalanced classes.

Imbalanced data can introduce challenges to machine learning classification tasks (Kaur et al., 2019) as the classifiers tend to bias towards the majority class: they spend more effort on getting the majority class right. If the majority class and the minority classes are not well separated in the feature space, the classification for the minority classes could become particularly bad. When the minority classes are the targets of interest, such bias can cause the classifier to deviate from the desired result; e.g., in this project, simply throwing every candidate into the majority class would achieve an accuracy as high as 97%. Thus, further investigation of the impact introduced by the imbalanced data is necessary.

By investigating the predicted scores (probability of being a local galaxy) of the local and non-local galaxy candidates, we find the imbalance in the training data does make the classifier less accurate in selecting local galaxies. Ideally, we want the scores of the local candidates to be 1 and the scores of the non-local candidates to be 0. While the vast majority of the non-local candidates have scores close to 0, the scores of the local candidates span a large range (0–0.95) almost uniformly. This indicates there is not a clear boundary between the local candidates and the non-local candidates in scores.

Re-sampling techniques are commonly used in scenarios where imbalanced data occurs (Kaur et al., 2019). They can affect the predicted scores and potentially improve the classification results, but it is always a trade-off between better scores for local candidates

and better scores for non-local candidates. We examined three popular re-sampling techniques: 1) Random over-sampling; 2) Random under-sampling; 3) Tomek links method (Ivan, 1976). The random over-sampling method returned the best results, and slightly improved the TPRs at high FPRs, with the cost of having lower TPRs at low FPRs. But overall, these re-sampling techniques did not help us achieve better classification results due to not adding new information to the model: they may improve the scores for one class, but do not separate the two classes better.

An alternative solution to the imbalanced data problem is cost-sensitive learning (Kaur et al., 2019). The basic idea behind cost-sensitive learning is to impose additional cost on the model when making incorrect classifications on the minority classes, which therefore forces the model to pay more attention to them. One can either allocate higher weights to the minority samples or use algorithms designed to focus on the misclassified data during training processes by their frameworks such as boosting algorithms. We examined both scenarios:

1. Classes reweight

   To balance the importance of the minority and majority classes, we assigned weights to each class inversely proportional to their sample frequencies. We tested two different strategies: 1) assign unique weights for all the tree classifiers in the random forest; 2) assign weights in the tree basis where the weights are computed at each tree according to the class distribution of the subsamples used in that tree.

   The results showed the two strategies gave almost identical performance, both very similar to the original model, but the original model returned a modestly purer set at low completeness.

2. Boosting algorithms

Boosting algorithms stack weak classifiers iteratively to build a final strong classi-
fier. When each new weak classifier is added, it focuses more on the examples that
were misclassified by the previous weak learners. By nature, boosting algorithms
focus on hard examples, in the situation of imbalanced data, a boosting algorithm
is driven by the minority classes. We explored the performance of two boosting
algorithms: 1) Adaptive Boosting (AdaBoost; Freund & Schapire, 1997) classifier;
2) Gradient boosting (Mason et al., 1999). AdaBoost built a model comparable to
the original one but did not outperform, gradient boosting returned slightly worse
results.

By examining all the solutions above, we conclude the various techniques and algo-
rithms have little impact on the final classification results, the performance is intrinsically
limited by the input data available. We eventually decide to retain the original model due
to its higher completeness at low FPRs.

2.3.3.5 *Hyper-parameters*

After determining the model design and features, we explored the hyper-parameters of
the RF algorithm to seek any further classification improvements. In RF algorithms, there
are two important hyper-parameters that often effect the performance of a model promi-
nently (Scornet, Erwan, 2017): 1) total number of trees used in the forest N_{tree}; 2) the
size of a random subset of full features that is chosen as potential splitting criterion at
each node of the tree, f_{try}. To optimize the model, we applied a grid search over the two
hyper-parameters: N_{tree} spans 50 to 1000 in steps of 50; and f_{try} varies from 1 to 31. At
each node on the grid, we perform a five-fold CV on the training set to reduce the random
scattering of the results.

The final classification results are not strongly sensitive to the choice of these hyper-

parameters. The optimized model achieves highest TPR at a fixed FPR equal to 0.03; the corresponding hyper-parameters are $N_{\mathrm{tree}} = 850$ and $f_{\mathrm{try}} = 5$.

2.4 EVALUATION

In this section we explore the behavior of the final model and compare its performance to previously-studied methods of identified local galaxies from Hα images from Cook et al. (2019). We also discuss the astrophysical properties and their completeness for the recovered local galaxies on the basis of our knowledge of the universe, try to approach the reality as much as possible. Finally, we address how limitations in the star-galaxy separation will influence the identification of local galaxies.

2.4.1 Test Set for Evaluations

To evaluate the performance of the optimized model we again use the GAMA DR3 dataset. Those data achieve high completeness for galaxies with $r < 19.8$ mag in the G15 region, and $r < 19.0$ mag in the G09 and G12 regions. There are some galaxies fainter than $r = 19.8$ mag in all regions but the sample is not well defined, so we exclude them from use in the test sets. To make sure we included sufficient fainter sources for testing, we took 80% of the sources in G09, G12 and G15 as the training set, then used the remaining 20% sources in G15 as the test set.

2.4.2 Model Evaluation in Counts

We assess the performance of the optimized model via a five-fold CV run on the ground truth samples, and we repeat this procedure three times to estimate the scatter. For every run we train the RF model with the 31 selected features and the optimized hyper-parameters from Sec. 2.3.3, using the training set discussed above, and then take test with

the G15 test set.

Once trained, we evaluate the classification results using the Receiver Operating Characteristic (ROC) metric: we plot the TPR against the FPR at different threshold settings. Because our main purpose is to identify local galaxies, the ROC curves here only consider the classification results in binary classes: the non-local candidates (negatives) and the local candidates (positives); i.e., class-0 is considered negative, class-1 to class-4 are all considered positive. The threshold varies from 1 to 0, where smaller thresholds tend to assign more candidates to the local class. We plot the ROC curves for each CV run and their averaged results in Figure 2.6. The figure contains ROC curves for the test set ($r < 19.8\,\mathrm{mag}$) and for two brighter subsets ($r < 19.0\,\mathrm{mag}$, $r < 18.0\,\mathrm{mag}$). For comparison, we also plot the evaluation of the predicted results using the method from Cook et al. (2019) for the $r < 19.8\,\mathrm{mag}$ test set in the same figure. In Cook et al. (2019) the authors select the local galaxies by inspecting the flux excess in the narrow-band Hα filters. If an emission-line galaxy is in the local universe, its Hα emission will cause a flux excess in one of the four filters compared to its adjacent continuum filter, quantified by Csig. A galaxy is selected if its maximum Csig is statistically significant, and the redshift range is determined by the filter where the maximum Csig is observed. We apply two different Csig thresholds in our comparisons: a conservative $\mathrm{Csig} > 5.0$ cut and a more aggressive $\mathrm{Csig} > 2.5$ cut.

Overall our classifier model performs very similarly across the three sets, with the $r < 19.8\,\mathrm{mag}$ test set slightly surpassing the others. We infer that this is due to the large number of samples between $19.0\,\mathrm{mag}$ and $19.8\,\mathrm{mag}$, which results in a better prediction over that interval. Figure 2.6 reveals promising results: with $\mathrm{FPR} = 0.01$ we achieve $\mathrm{TPR} = 0.615$; at $\mathrm{FPR} = 0.03$, TPR is greater than 0.80; if we allow for a higher $\mathrm{FPR} = 0.06$, our model can recover 90% of the local galaxies. Moreover, the similarity among the per-

formances for the three sets implies, although the sample sizes for the objects in different brightness regimes are fairly non-uniform in the training set and do not necessarily map the real universe, our model is not sensitive to the candidates' brightness. Considering that in the training set, a decent amount (14.5%) of training samples lie in the $r > 19.8$ mag regime, it is highly possible our model can successfully pick those fainter local galaxies as well.

In contrast, the Csig methods achieve excellent FPRs (especially the $\mathrm{Csig} > 5.0$ cut which has negligible FPR), but the TPRs for both cuts are low. At the same FPRs, our ML method always returns significantly higher TPRs, which outperforms the Csig method. This big performance difference is likely due to the Csig method's dependence on the existence of a moderate-to-strong $H\alpha$ emission line, where our ML model does not have such limits. Hence the ML model is more sensitive to the low $H\alpha$ emission passive galaxies.

2.4.3 Model Evaluation with Astrophysical Properties

While number counts for the classification results give a good sense of the model performance, what is really of interest in this project is the chance that our CLU catalog finds the galaxies which host compact binary object merger events that are detectable by the GW detectors. Therefore we introduce multiple astrophysical properties into the model evaluation.

Studies (Cao et al., 2018; Mapelli et al., 2018; Artale et al., 2019, 2020b,a) suggest that BBH, BNS and NSBH merger rates have strong correlations with the host galaxy's stellar mass and star formation rate (SFR). The B-band luminosity is often taken as another proxy for the compact object merger rate (Nissanke et al., 2013; Gehrels et al., 2016) because it can be used to trace the population of young stars and thus extrapolate the star
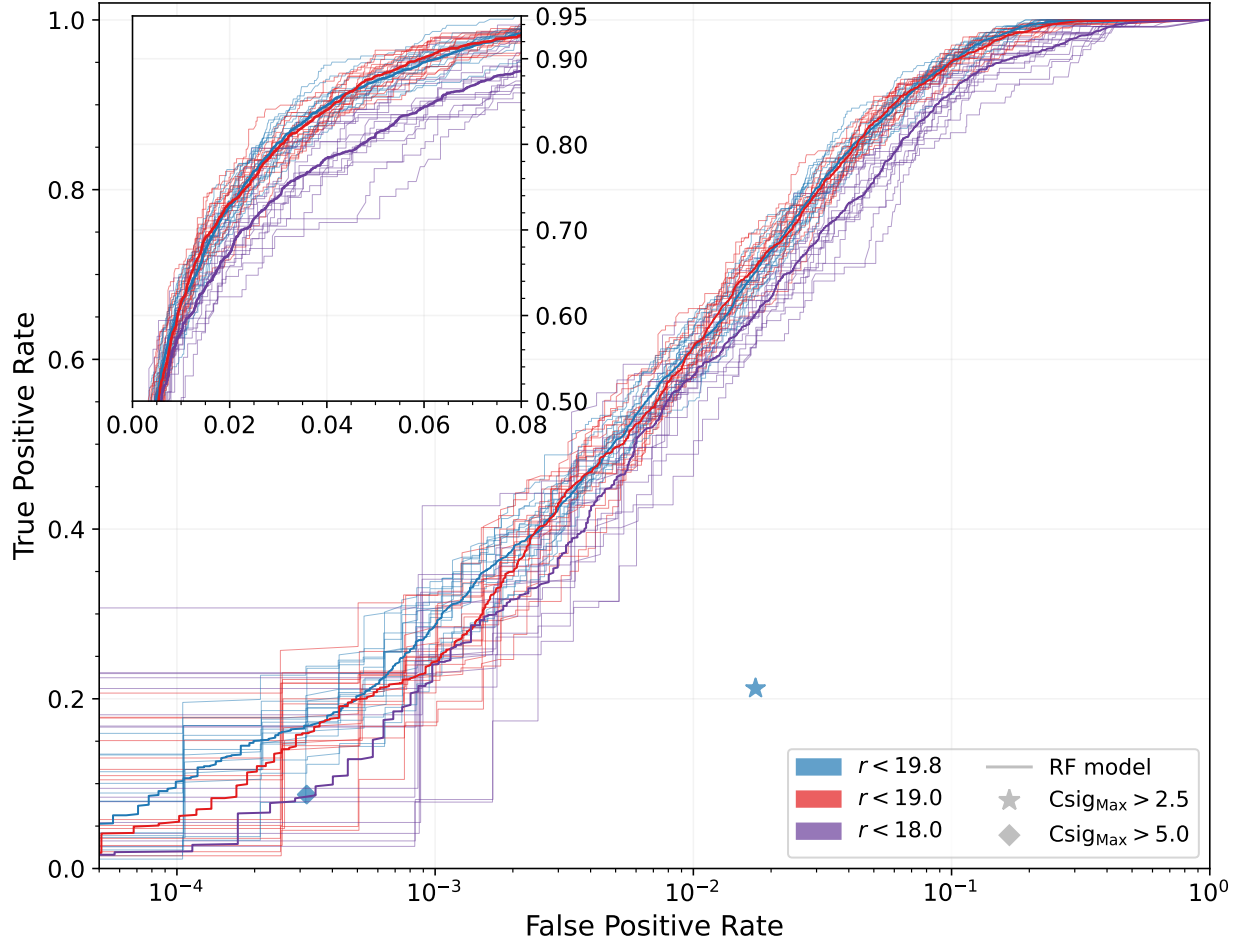
Figure 2.6: ROC curves illustrating the performances of our ML model for three different test sets, weighted by number count. The thick, solid blue, red, and purple lines show the ROC curves for the $r < 19.8$ mag, the $r < 19.0$ mag and the $r < 18.0$ mag sets, respectively. The light, thin lines show the ROC curves for the individual CV folds. The inset on the upper-left shows a zoom-in around $\mathrm{FPR} = 0.04$, where the TPR spans 0.5 to 0.95 given various FPRs. The performances of the Csig methods using the maximum-$\mathrm{Csig} > 2.5$ cut and the maximum-$\mathrm{Csig} > 5.0$ cut are represented as a star and a diamond, the results are evaluated under the $r < 19.8$ mag test set.

formation (Phinney, 1991). We add these three astrophysical properties as weights in our evaluations. Because our optical data PS1 does not have a B-band filter, we use the PS1 g-band magnitudes to derive the B-band luminosities. For stellar mass and SFR, we adopt the same definitions as in Cook et al. (2019).

We estimate the stellar mass via a constant mass-to-light ratio using the *WISE* $3.4\,\mu\mathrm{m}$ fluxes (Jarrett et al., 2013; Norris et al., 2014):

$$\Upsilon_\star^{3.4\mu\mathrm{m}} = 0.5 \; M_\odot/L_{\odot,3.4\mu\mathrm{m}}, \tag{2.7}$$

where $M_\odot/L_{\odot,3.4\mu\mathrm{m}}$ is the mass-to-light ratio in units of solar masses per the solar luminosity in the *WISE* $3.4\,\mu\mathrm{m}$ $W1$ filter bandpass ($m_{\odot,3.4\mu\mathrm{m}} = 3.24\,\mathrm{mag}$; $L_{\odot,3.4\mu\mathrm{m}} = 1.58 \times 10^{32}\,\mathrm{erg\,s^{-1}}$; Jarrett et al., 2013).

SFRs are estimated from the tracer Hα luminosity plus the *WISE* $22\,\mu\mathrm{m}$ luminosity which accounts for internal dust extinction (Calzetti et al., 2010; Murphy et al., 2011):

$$\left(\frac{\mathrm{SFR}_{\mathrm{H}\alpha,\mathrm{corr}}}{M_\odot\,\mathrm{yr}^{-1}}\right) = 5.37 \times 10^{-42} \left(\frac{\nu L_{\mathrm{H}\alpha} + 0.031\,\nu L_{22\mu\mathrm{m}}}{\mathrm{erg\,s}^{-1}}\right), \tag{2.8}$$

where νL are the observed monochromatic luminosities of both Hα and the *WISE* $W4$ band at $22\,\mu\mathrm{m}$.

These definitions can be difficult to apply in practice because of the significant redshifts of some of our test sources (e.g., 18% are above 0.3, and 3.1% are above 0.4). A simple resolution to this issue is to limit our evaluation to a finite volume that the emission stays in the $W1$, $W3$, and B bands. To keep the B-band emission in the PS1 g-band, we need a limit of $z \approx 0.23$. For the same reason, the *WISE* bandpass parameters (Jarrett et al., 2011) for $W1$ and $W4$ bands suggest an upper limit of $z \approx 0.2$. We use the GAMA line fluxes (Gordon et al., 2017) from their DR3 for Hα, which are taken from spectra that cover a wavelength range of 3750—$8850\,$Å (Hopkins et al., 2013) and can theoretically identify Hα emission lines up to $z \approx 0.35$. For objects with no available line flux, we use our forced-photometry measurements to derive the Hα luminosities for sources with $z < 0.047$ following Cook et al. (2019).

We plot the redshift distribution for all the galaxies in our CV test sets in Figure 2.7, with each galaxy labeled as local/non-local by the classifier, using classification thresholds resulting in TPRs of 80% and 90% in number. The reason for choosing 80% and 90% is that we want to inspect how the classifier will select candidates as locals with high TPRs. In our analysis we find the high-TPR thresholds tend to identify more higher-redshift galaxies as local ones; not only does the total number of false positives increase, but the redshifts of those false positives tend to shift towards the higher end as well. Both panels of Figure 2.7 illustrate two features:

1. With high TPRs, a fair number of non-local galaxies are mistakenly classified as "local".

2. The non-local galaxies with relatively low redshifts (closer to $z = 0.047$) are more likely to be classified as "local" objects.

With $\mathrm{TPR} = 80\%$, few candidates above $z = 0.2$ are classified as "local"; for $\mathrm{TPR} = 90\%$, a few more candidates in that regime are thrown into the "local" group, but the vast majority of the false positives are still in the $z < 0.2$ volume. This means that if we evaluate our ML model in the $z < 0.2$ volume, all actual local galaxies are included, plus we are able to capture almost all false positives. Combined with the upper limits suggested by the photometric bands, we decided to evaluate the weighted performances of our model in the $z < 0.2$ volume.

Adding the astrophysical properties as weights, we plot the weighted ROC curves for our test candidates in the $z < 0.2$ volume in Figure 2.8, alongside the results using the Csig methods. Limiting the evaluation to the $z < 0.2$ volume does not change the value of TPR for a given threshold, since all the real local galaxies are already in this volume. But such a constraint will overestimate the FPR because the vast majority of false positives happen to be $z < 0.2$ candidates for any reasonable threshold, but the true negatives above $z = 0.2$
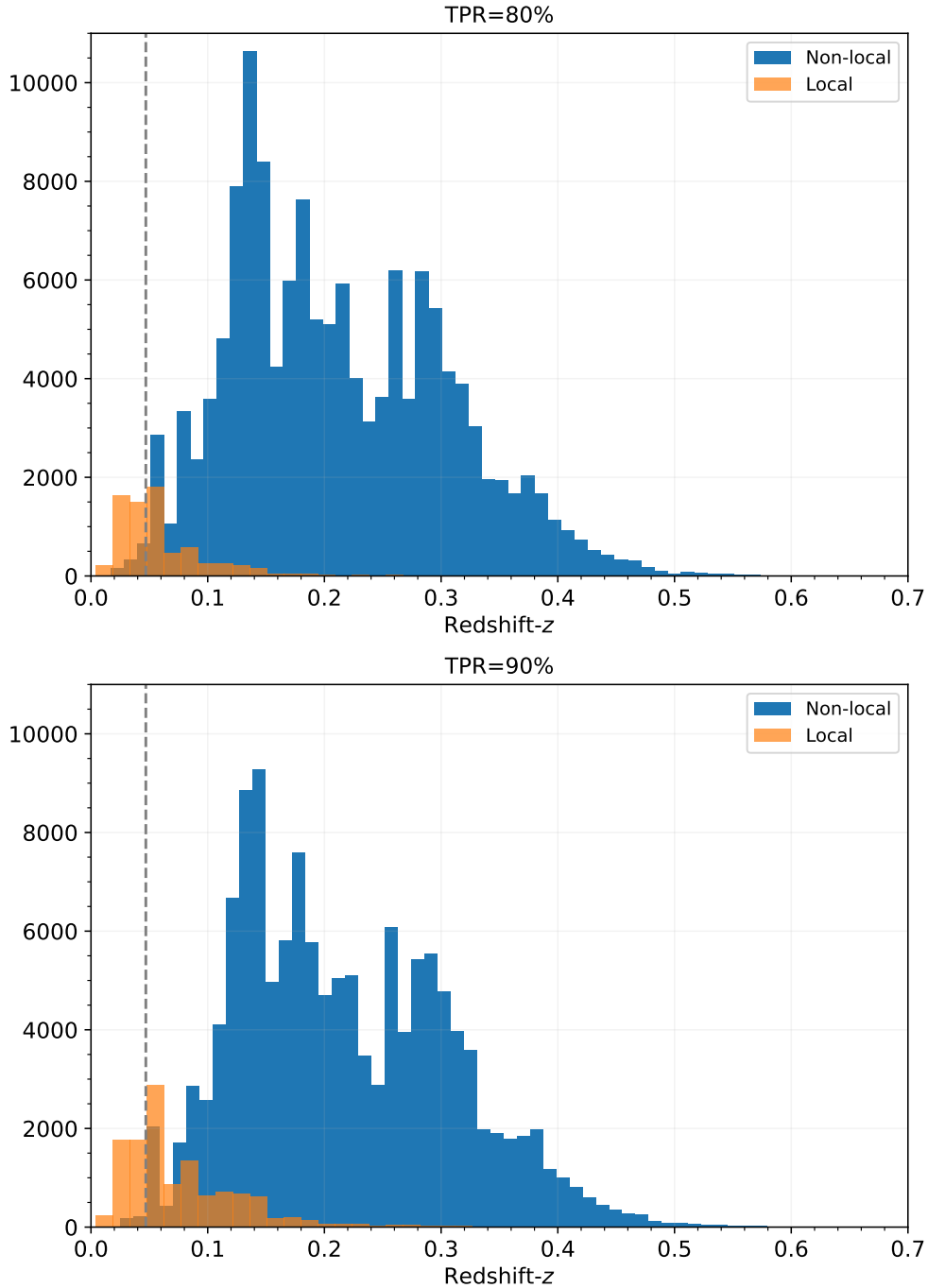
Figure 2.7: Redshift distributions for test galaxies in all 15 CV runs. The top panel shows the redshift distribution of those galaxies with assigned labels by the threshold gives $\mathrm{TPR} = 80\%$. Galaxies labeled as "local" are plotted in orange, the ones labeled "non-local" are colored in blue. The grey dashed line indicates $z = 0.047$, the boundary of the local universe. The bottom panel shows the same distribution for $\mathrm{TPR} = 90\%$.

51

are not included in the calculations (see Figure 2.7). This effect can be easily seen from the orange ROC curve in Figure 2.8. The orange curve is weighted by number counts, like the blue curve in Figure 2.6. For the same TPR (i.e., same threshold), the orange curve constantly has a larger FPR than the blue curve which represents the results of the full test set without redshift constraints.
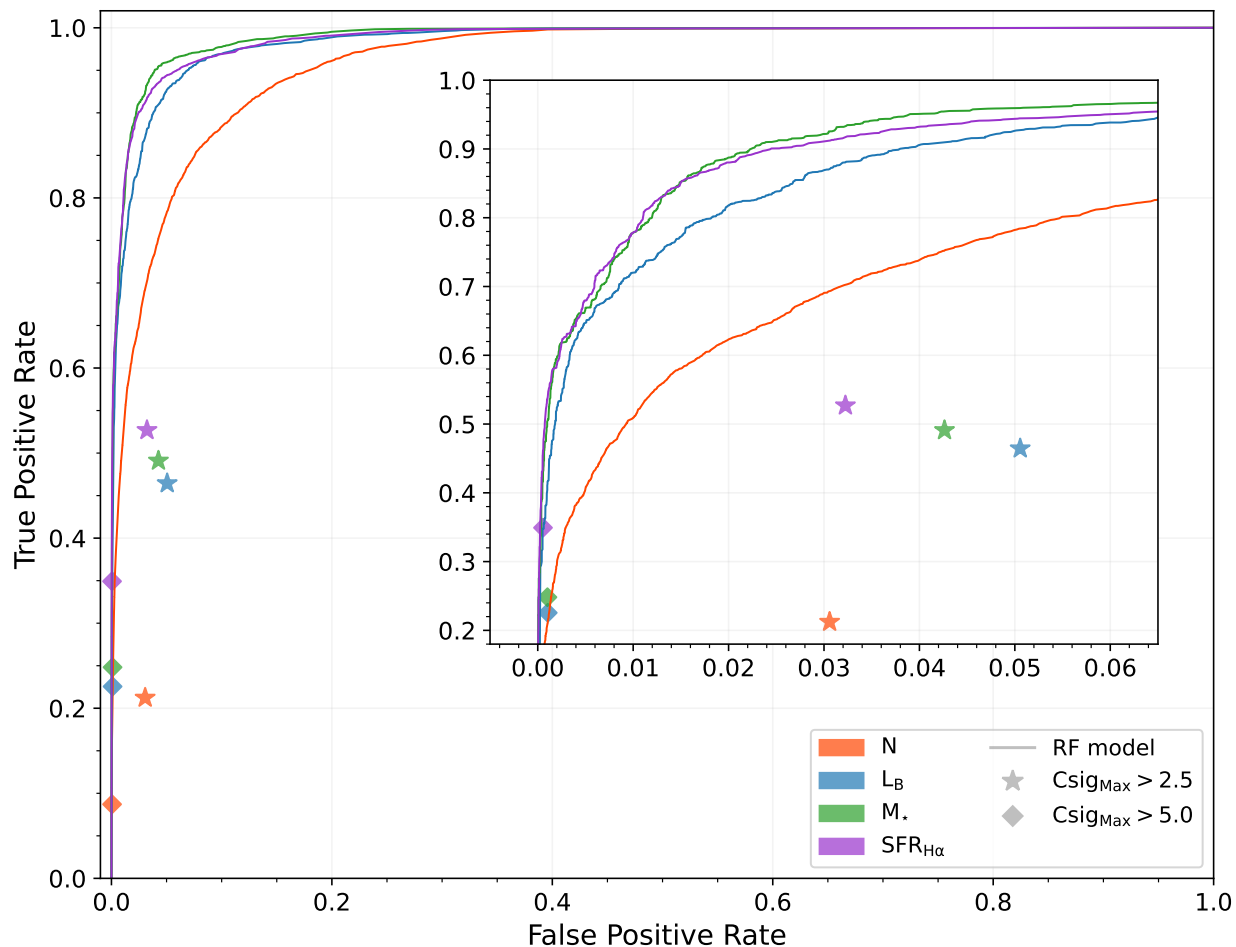
Figure 2.8: ROC curves for our ML model and results from the Csig methods weighted by astrophysical properties, evaluated among the test candidates ($r < 19.8\,\mathrm{mag}$) in the $z < 0.2$ volume. The ML ROC curves are plotted in solid lines, the orange, blue, green and purple colors represent number, B-band luminosity, stellar mass and star formation rate. The results using the Csig methods are plotted as stars and diamonds with the same color scheme. The inset on the upper-right shows a zoom-in around $\mathrm{FPR} = 0.03$.

Although the FPRs are overestimated, we find our ML model performs well at finding targets: it can reach very high completeness in B-band luminosity, stellar mass and SFR with low FPRs. The same behavior is found for the Csig methods. Overall, the performance for these astrophysical metrics significantly surpasses the ones for simple number counts. The Csig methods again achieve approximately 10%–20% TPRs with low FPRs in number, but the TPRs with same thresholds increase to nearly 25%–50% when weighted by B-band luminosity and stellar mass; if weighted by SFR, the TPRs can even reach 35% and 53%. This is not surprising, as the Csig methods rely on Hα emission lines for selection, and that emission is itself a proxy for SFR. However, neither Csig threshold is able to select the majority of the galaxies (by astrophysical property or number) in the local volume. In contrast, our ML model can achieve high TPRs with appropriate thresholds in all categories, while not introducing unacceptable FPRs. At the same FPRs, our ML model always accomplishes higher TPRs than the Csig methods. For a FPR of 1% derived in the $z < 0.2$ volume, the ML model recovers 51% of galaxies in the local universe and they constitute more than 77% of the total mass. With a FPR equal to 3% that is close to the Csig > 2.5 cut, we recover more than 69% of the targets and 92% of stellar mass. We emphasize again these FPRs are overestimated. Their small values suggest that true negatives completely dominate the denominator in Eqn.(2.4); if the sum of true negatives doubles, the FPR will decrease by roughly a factor of 2. The number of test galaxies in the $z > 0.2$ volume is approximately equal to the amount of true negatives in the $z < 0.2$ volume, thus the 1% and 3% FPRs, if derived from the full test set with $r < 19.8\,\mathrm{mag}$, will translate to around 0.5% and 1.5% respectively; they are exactly the FPRs for TPRs equal to 50% and 68% in Figure 2.6.

2.4.4 Impact of Incomplete Star/Galaxy Separation

As described in Sec. 2.2.3 and Sec. 2.3.2, we use the PS1-PSC catalog to separate stars and galaxies in our source catalog. Because no ML algorithm gives perfect results, some real galaxies will be classified as stars by our chosen classification threshold, thus causing a loss of completeness in the local universe. Similarly, some stars will be mistakenly selected as galaxy candidates, potentially adding contaminants. In this section, we estimate the impact introduced by this imperfect star/galaxy separation (S/G) in our test region G15, limiting the depth to $r < 19.8$ mag.

We evaluate the impact by exploring the change of completeness and contamination of the classification results when the S/G selection is applied on the test set. After the S/G step, the test set is a mix of galaxies and stars, with the galaxy sample size slightly smaller than the original. Stars which passed the S/G selection are regarded negatives like non-local galaxies. We use contamination as a metric instead of FPR because we care more about the newly introduced contaminants in our local galaxy catalog than the tiny changes in FPR caused by the small number of stars given the huge amount of non-local galaxies. We check the changes at various completeness levels (i.e., different thresholds). With a S/G filter, the completeness weighted in number and astrophysical properties drop by 4.2% at maximum with respect to their original values without a S/G filter. The decrease is smaller at lower completeness; the median fractional drop is 2.5%. The situation for contamination is more complex. While a S/G filter rejects some of the true local galaxies (decreasing TP) and adds some stars (increasing FP), it also prevents a certain fraction of non-local galaxies from being selected which reduces the FP.

Given the fact stars have negligible astrophysical properties compared to the galaxies, contamination weighted in the properties can actually become smaller with S/G. We find the contamination with the properties remains almost identical until the completeness in-

creases to 40%, then we see a quick drop of contamination due to some non-local galaxies are filtered out by S/G. Because the number of false positives was originally small, this drop is significant and can be as high as 24.2% at completeness equal to $\sim 60\%$; however, the absolute difference in contamination is still small due to the low contamination at this completeness. When the completeness gets larger, the drop of contamination becomes more modest quickly, the median decrease of contamination weighted in astrophysical properties is 2.1%. On the contrary, the contamination in number greatly grows because of the stars added by S/G, especially at low completeness levels where only very few non-local galaxies are selected by our classifier, so that a small addition of stars can easily raise the contamination by multiple times even though the absolute difference is not large. As the completeness level rises up, the fractional growth of contamination drops rapidly and becomes low at the high end of completeness. Overall, the median increase of contamination in number is 87.9% respect to its original value; the difference is less than 26% with completeness greater than 80%.

We emphasize that the "stars" in this evaluation are not all real stars. We consider an object in our candidate set a star if it meets both conditions:

1. The object passes the PS1-PSC threshold cut for galaxies, i.e., this object is identified as a galaxy in our S/G algorithm.

2. The object is in our source catalog but not in the GAMA catalog; i.e., this object is a star, or it is a galaxy but had bad spectroscopy so was not included by GAMA, or the galaxy was not observed due to the limited completeness of GAMA.

With every 1000 GAMA galaxies, the S/G filter will leak less than 59 such "stars" into the candidate set. Because the GAMA survey is not 100% complete to their targeted depth, some of the "stars" are actually galaxies missed by GAMA. We estimate roughly one third of the "stars" are such missed galaxies, and about 3% of those galaxies will be local. Since

they are regarded as false positives if selected by our classifier, while actually being true positives, they will artificially increase the calculated contamination. Thus we conclude that the actual contamination is lower than what we estimated above. We assess the median fractional decrease of contamination with the correction by such missed galaxies is about 1.2%; at very low completeness levels ($\sim 10\%$), this decrease can be nearly 10%.

2.4.5 Contaminants

The evaluations so far show that our ML model can recover a large fraction of the local galaxies with low FPRs. However, local galaxies only occupy a small fraction among all the observed galaxies spanning a wide range of redshifts, thus even a low FPR can result in an amount of false positives that is comparable to the true positives in our classified "local" galaxy set. For example, a 74.2% completeness corresponds to a FPR of 2.2%, but among all the selected positives, half of them are false, leads to a contamination rate of 50%. The stars introduced by the S/G step contribute even more contaminants as described in Sec. 2.4.4: for the same 74.2% completeness, contamination rises up to 68% with the stars. It is inevitable that high completeness comes with high contamination. When using this catalog, high contamination may have big impact for some research scenarios, but users are free to adjust their own thresholds for lower contamination and lower completeness; moreover, users can apply more aggressive S/G thresholds than ours to reduce the number of stars in the contaminants, which will lower the contamination significantly at low completeness levels.

2.5 PERFORMANCE AND COMPARISON

In this section we explore a series of appropriate classification thresholds for identifying the local galaxies, and estimate the corresponding FPR, TPR and accuracy. We report

Classification Thresholds for the CLU-ML Catalog

Evaluation Metric	Threshold	0.453	0.341	0.232	0.166	0.126	0.099	0.078	0.033
Number (N)	FPR	0.005	0.01	0.02	0.03	0.04	0.05	0.06	0.1
	TPR	0.505	0.615	0.728	0.802	0.846	0.874	0.896	0.950
	Accuracy	98.1%	97.9%	97.3%	96.5%	95.7%	94.8%	93.9%	90.3%
B-band Luminosity (L_B)	FPR	0.011	0.020	0.037	0.053	0.069	0.085	0.102	0.165
	TPR	0.737	0.820	0.898	0.931	0.950	0.962	0.970	0.983
	Accuracy	98.5%	97.7%	96.2%	94.7%	93.1%	91.6%	89.9%	84.0%
Stellar Mass (M_*)	FPR	0.006	0.010	0.017	0.025	0.034	0.042	0.051	0.089
	TPR	0.680	0.776	0.874	0.913	0.936	0.952	0.960	0.974
	Accuracy	99.2%	98.9%	98.2%	97.4%	96.6%	95.8%	94.9%	91.4%
Star Formation Rate ($SFR_{H\alpha}$)	FPR	0.004	0.008	0.015	0.025	0.035	0.045	0.056	0.106
	TPR	0.653	0.755	0.855	0.901	0.923	0.938	0.948	0.970
	Accuracy	99.3%	99.0%	98.4%	97.5%	96.5%	95.5%	94.5%	89.7%

Table 2.3: Appropriate thresholds for identifying the local galaxies in the CLU-ML catalog. All the evaluations are based on the $r < 19.8$ mag test samples. The evaluations with B-band Luminosity, Stellar Mass and Star Formation Rate are limited in the $z < 0.2$ volume due to the limitations of these metrics discussed in Sec. 2.4.3; these FPRs are overestimated by roughly a factor of 2, and the accuracy are underestimated.

the appropriate thresholds and their performance with multiple metrics in Table 2.3. We then demonstrate the behavior of our model in detail with a reasonable threshold, and compare its performance against the GLADE catalog (Dálya et al., 2018, 2022), one of the most commonly used catalogs in EM follow-up.

2.5.1 A Worked Example

To better illustrate what our catalog can achieve, we choose a classification threshold that people will likely use in EM follow-up and evaluate the results it leads to. We use 0.167 as the threshold which results in an overall completeness of 80% for the test. This test uses purely galaxies in our test set, with no S/G filtering.

We use this threshold to separate local and non-local galaxies, then plot the cumulative completeness as a function of redshift and r-band magnitude in Figure 2.9. The cumulative completeness is defined as the fraction of galaxies that are recovered in terms

of number counts or astrophysical properties within a certain distance. The left panel shows the completeness versus redshift, where the completeness is evaluated with multiple weights: number counts, B-band luminosity, stellar mass and star formation rate. At the very low redshifts, our ML algorithm can recover almost all the local galaxies. As the distance increases, the completeness gradually decreases until reaching 80% at $z = 0.047$ if weighted in number count. When weighted by the astrophysical properties, the cumulative completeness is noticeably higher and reaches more than 90% at the same distance. This suggests with the chosen threshold, our model can identify the absolute majority of the local galaxies, and thus maximize the probability of finding the host galaxies for EM follow-up. On the right panel, we plot the completeness with respect to the r-band magnitudes of the local galaxies. We are able to recover almost all the local galaxies in the very bright regime and 80% for all objects brighter than $r = 19.8$; the proportions recovered in luminosity, mass and SFR are even higher.

We do the same test for the contamination and plot the results in Figure 2.10. Unlike completeness, because our model predicts redshift constraints instead of the exact redshifts, we can only evaluate the contamination in the four redshift ranges (class-1 to class-4) corresponding to the Hα filters. The contamination here is defined as the fraction of objects that are not in the local volume among all the candidates predicted to a specific redshift range, i.e., only non-local galaxies (class-0) are considered contaminants, if a galaxies in class-2 is predicted to class-3, it does not contribute to the contamination. We will also discuss the mis-classification between the local classes shortly. In Figure 2.10, the left panel shows the contamination in the four redshift ranges (classes), the contamination in class-1 is as low as 28.8%; in the other classes, the contamination is higher, around 40%–60%. The contamination for the astrophysical properties is constantly higher still as model tends to select larger non-local galaxies. The right panel reveals how the cumu-

58
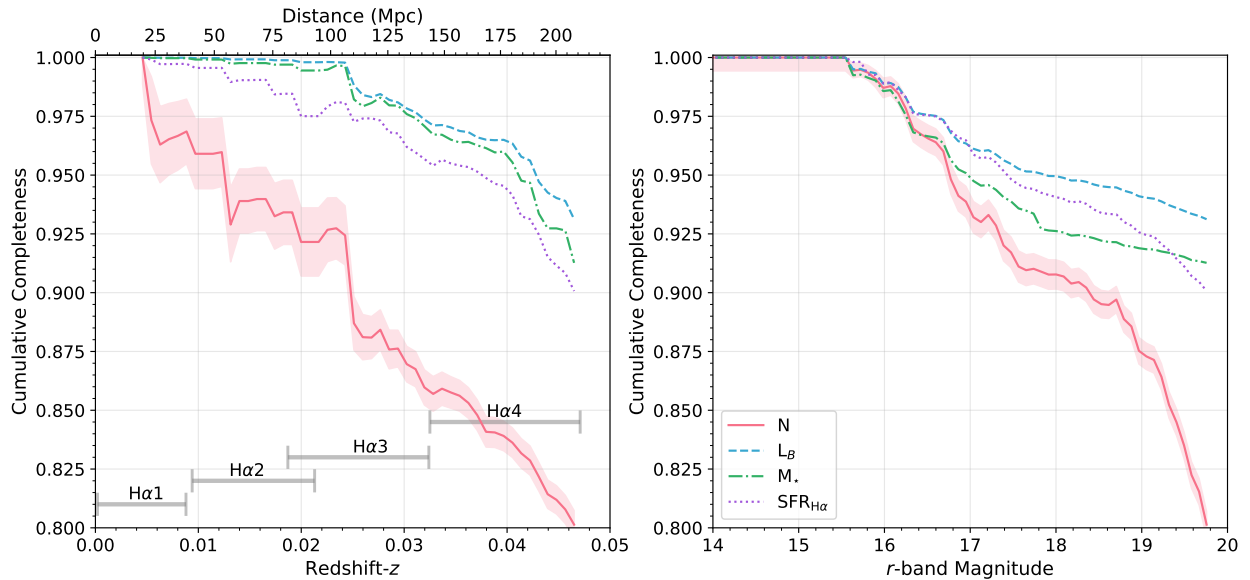
Figure 2.9: Cumulative completeness as a function of redshift (left panel) and r-band magnitude (right panel). The red solid curve represents the completeness weighted in number counts, the shaded areas represent the errors. The dashed curves colored in blue, green and purple represent the completeness weighted in B-band luminosity, stellar mass and star formation rate. We label the luminosity distances associated with the redshifts. We denote the redshift ranges for the four Hα filters by grey error bars. The errors for completeness equal to 100% on the right panel at small r-band magnitude is the root mean square (RMS) error across all magnitudes.

lative contamination changes with r-band magnitude. The cumulative contamination is defined as the fraction that contaminants occupy among all the selected local candidates down to a certain magnitude. The contamination is low in the bright regime, and gradually rises to 56.1% at 19.8 mag, because the model picks more massive non-local galaxies, this value is higher for luminosity, mass and SFR.

As mentioned above, although our model identifies most of the local galaxies, many of them are not necessarily assigned to the correct redshift range/class. The recovered local galaxies may be assigned to any of the four local classes, e.g., more than 84% of the local galaxies in class-2 are recovered, but the majority of those are classified as class-3. To explore how the mis-classification behaves across the classes, we plot a confusion
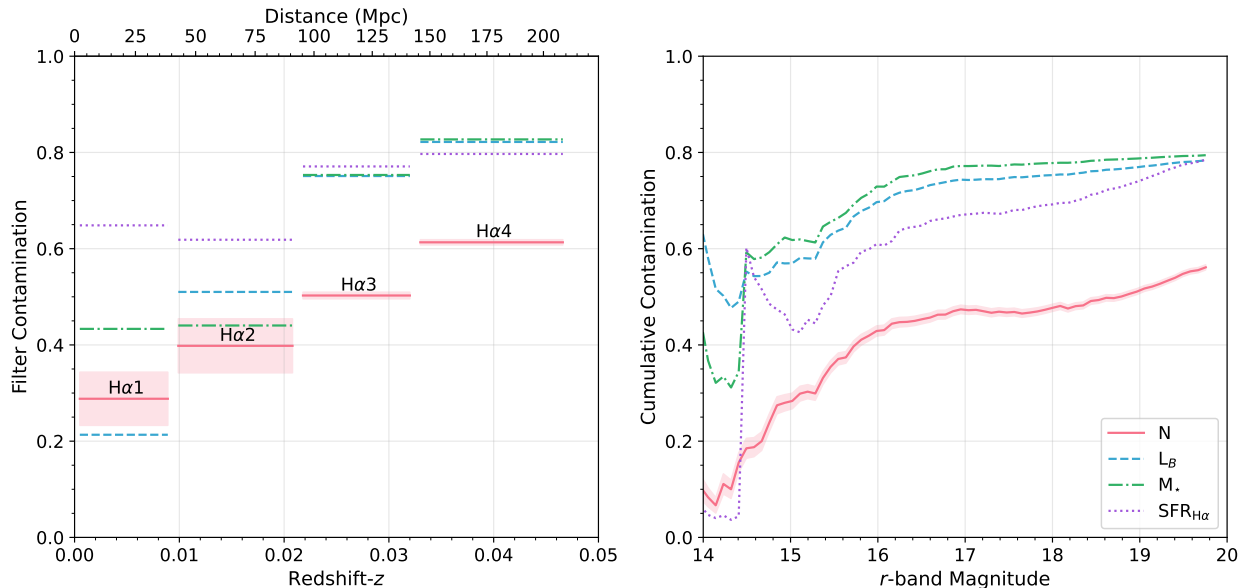
Figure 2.10: Contamination in the four local redshift ranges (Hα filters; left panel) and the cumulative contamination as a function of r-band magnitude (right panel). The red solid lines represent the contamination weighted in number counts, the shaded areas represent the errors. The dashed lines colored in blue, green and purple represent the contamination weighted in B-band luminosity, stellar mass and star formation rate. We label the luminosity distances associated with the redshifts.

matrix in Figure 2.11. With this specific threshold, the overall classification accuracy is 95.8%; the ML model is able to identify 97% of the non-local galaxies correctly; for the 3% wrong predictions, the majority are classified as class-3 and class-4 as they are closer to the local volume boundary. In the four local classes, we lose local galaxies mostly in class-3 and class-4, again because they are closer to the boundary. For the local objects that are successfully assigned to one of the four local classes, predictions are greatly biased to class-3 and class-4 due to their significantly larger volumes and thus dramatically larger training samples. About half of the objects in class-3 and class-4 can be selected to the right sub-local classes, but a large portion of the local galaxies as class-1 and class-2 will be mis-classified as class-3. Therefore, we conclude that our algorithm is very useful at distinguishing between the local and non-local galaxies, but that the predicted sub-classes

for the true local galaxies are highly unreliable.
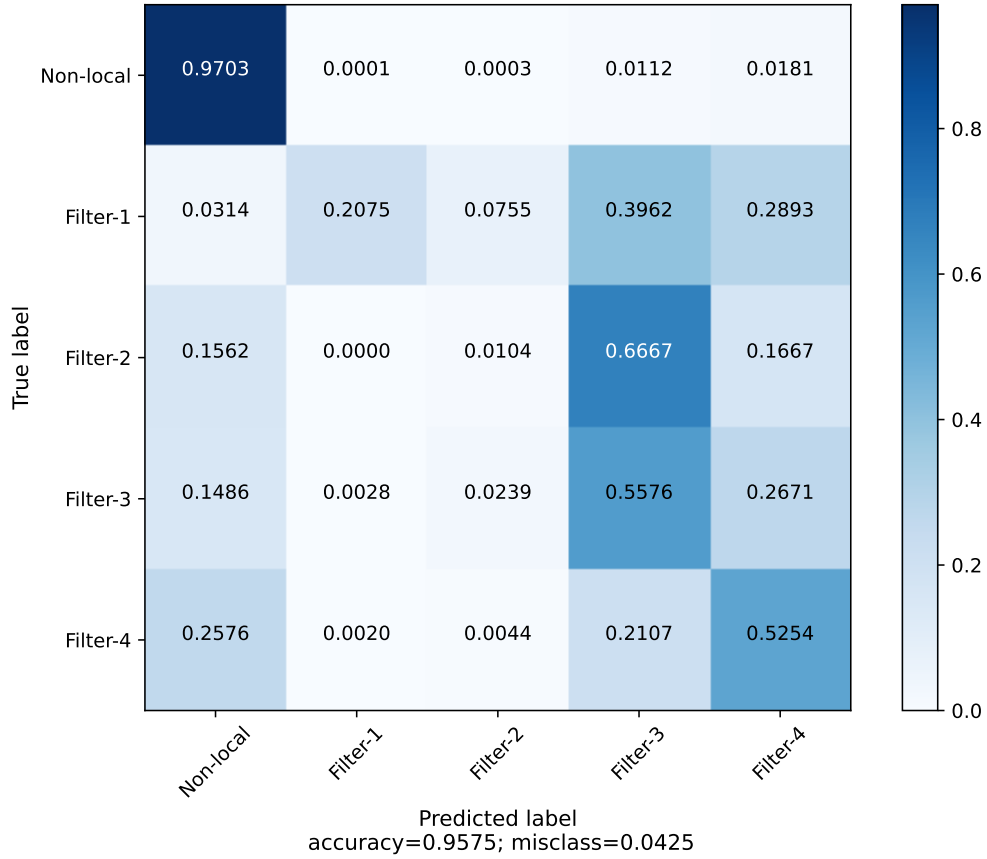
Figure 2.11: Confusion matrix of the CLU-ML algorithm that illustrates mis-classification. The true classes are labeled on the vertical axis, the predicted classes are on the horizontal axis. The label "Non-local" denotes class-0, and the "Filters" denote class-1 to class-4. The number in each box indicates the fraction of objects in this true class that are selected to the corresponding predicted class by the machine; the diagonal implies the prediction equal to its true class. The overall accuracy of the classification with this threshold is 95.8%.

2.5.2 Comparison to GLADE

In this section we compare the completeness of CLU-ML to the GLADE catalog (Dálya et al., 2018). GLADE aims to support the EM follow-ups of GW events by cross-matching and combining five separate astronomical catalogs. The GLADE catalog contains ~ 3.26

million objects including galaxies, quasars and globular clusters. Because the objects in GLADE come from multiple surveys with different sensitivities and uneven distributions toward different directions on sky, the object number density of GLADE varies significantly with the direction as well (see figure 1 in Dálya et al. 2018).

The number densities of GLADE objects in some of the densest directions can be orders of magnitude higher than the average. The densest regions in GLADE correspond to the deep and sensitive HyperLEDA catalog (Makarov et al., 2014). HyperLEDA contains over 3 million objects with highly reliable redshifts and redshift-independent distance measurements. GLADE kept ~ 2.6 million galaxies and quasars in their own catalog. The distance measurements of the GLADE objects in the HyperLEDA regions are dominated by spectroscopic redshifts, which are very accurate and do not suffer from the problems caused by photometric redshift estimates. One of the three densest GLADE patches located in the HyperLEDA regions entirely overlaps with the GAMA G15 equatorial region, thus we compare the results of our CLU-ML algorithm with GLADE in a sub-region of GAMA G15.

The GLADE team recently upgraded the GLADE catalog by adding another WISExSCOSPZ galaxy catalog (Bilicki et al., 2016) and updating the SDSS quasar catalog to the DR16 (Lyke et al., 2020) version. The new catalog is named GLADE+ (Dálya et al., 2022). The WISExSCOSPZ catalog greatly increased the completeness of the original GLADE above 200 Mpc and makes the number density of objects more even across the sky, but the three densest HyperLEDA patches are still notable. We compared CLU-ML to both GLADE and GLADE+, but since the difference between GLADE and GLADE+ in the 200 Mpc local volume is negligible, we only present the comparison with GLADE+ below.

Because the GAMA catalog is nearly 100% complete in its G15 region down to $r =$

19.8 mag, we use half of G15 as the ground truth to test the completeness of CLU-ML and GLADE+. We compare our CLU-ML algorithm with GLADE+ down to the GAMA magnitude limit and plot the results in Figure 2.12. The red curves show the cumulative completeness of CLU-ML weighted by number count and B-band luminosity. As with Figure 2.9, the completeness starts very high and then gradually decreases to 80% at $\sim$ 200 Mpc. The luminosity is more complete, at over 90%. The brown curves represent the completeness of GLADE+. When weighted by luminosity, it is slightly lower to although comparable with CLU-ML. However, the number of GLADE+ galaxies is constantly less than CLU-ML. Note that we test this in a region where GLADE+ has the highest number density of different areas across the sky. For the majority of the sky, GLADE+ will have much lower completeness, while CLU-ML will have roughly even completeness across its footprint. Furthermore, out of the the HyperLEDA regions, GLADE+ will suffer from the increasing fraction of photo-z measurements, which suppresses its completeness even more.

2.6 DISCUSSION

This ML algorithm is applied to all the candidates in our source catalog, where the training set are the entire ground truth samples. Astronomers can utilize our catalog to optimize their pointing strategies, remove the false positive transients, and prioritize the rankings of targets for the EM follow-up of the GW events. When used in EM follow-up, we want this catalog to be more complete with the cost of more contaminants, rather than taking the risk of missing the host galaxy. Thus more aggressive thresholds are preferred. Although such thresholds lead to higher FPRs so adds more contaminants to the classified "local" galaxy set, the absolute majority of negatives are removed. For instance, as evaluated in Sec. 2.4.5, a threshold with 74.2% completeness results in a contamination
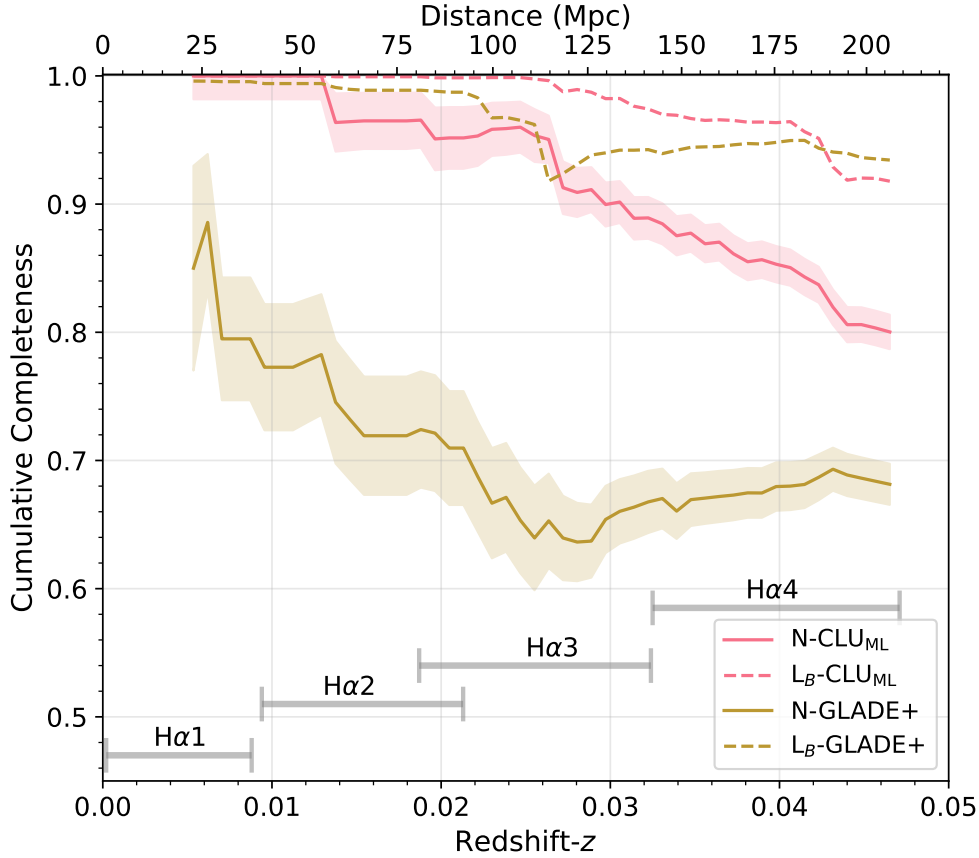
Figure 2.12: Comparison of the cumulative completeness of the CLU-ML algorithm and GLADE+ in over half of the GAMA G15 region (one of the densest regions for GLADE+) for $r < 19.8$ mag objects. The completeness for CLU-ML is plotted as red curves, with the solid curve representing the completeness in number count and the dashed curve represents that in B-band luminosity; the shaded areas represent the errors. The same results for GLADE+ are plotted as brown curves. We label the luminosity distances associated with the redshifts. We denote the redshift ranges for the four Hα filters by grey error bars.

of 50% (68% with imperfect S/G), but 97.8% of the non-local galaxies are filtered out by our model. Prioritizing the selected likely-local galaxies in the search will improve the searching efficiency dramatically.

In addition to identifying the host galaxies for EM follow-ups, the CLU catalog will also contribute to finding galaxies with extreme emission-lines. Dwarf galaxies with extreme emission-line properties at low-to-intermediate redshifts ($z < 0.6$) resemble the

low mass, low metallicity galaxies in the early universe that likely reionized the universe (Jaskot & Oey, 2013; Finkelstein et al., 2019; Guseva et al., 2020). Due to the difficulty of observing those distant galaxies in the early universe with current facilities, understanding the properties of their low-z analogs provides an alternative approach that will give us insights into the conditions of the first galaxies and the reionization of the universe. CLU is an ideal resource catalog for searching these emission-line galaxies. CLU will be able to find the BCDs (Kunth & Östlin, 2000; Cairós et al., 2010) out to $z = 0.047$ via their Hα emission, and discover the green peas (Cardamone et al., 2009) at $0.30 \lesssim z \lesssim 0.37$ via their strong [O III]$\lambda 5007$ emission line. Using our preliminary source catalogs (Cook et al., 2019), we have identified $3,400$ BCD and green pea candidates, where 80% have no previous redshift information, we expect to find many more with CLU-ML.

Massive efforts have been made across the community towards complete galaxy catalogs with distance in the past decades for the important roles such catalogs play in various fields, the most common method used is the photometric redshift (photo-z) due to the expense of spectroscopy. Traditionally scientists use two main techniques to perform photo-z: template fitting and machine learning algorithms. The template fitting methods (e.g., Arnouts et al. 1999; Benítez 2000; Brammer et al. 2008) establish templates that model the physical processes governing the light emission of the objects to obtain their redshifts. The machine learning methods, on the other hand, do not try to reproduce the physical processes, but are data-driven. Numerous classical (e.g., Csabai et al. 2007; Carliles et al. 2010) and deep learning (e.g., Collister & Lahav 2004; Pasquet et al. 2019; Bilicki et al. 2014; Christodoulou et al. 2012; Beck et al. 2021) algorithms have been used to derive the photometric redshifts. The template fitting methods and the machine learning methods both result in decent photometric redshifts, with the latter preferred nowadays due to their overall better precision. The photo-z estimation methods usually yield unbi-

ased predicted redshifts (z_{phot}) in a sense that the mean true redshift (z_{spec}) at given z_{phot} is equal to z_{phot}. However, photo-z estimations tend to systematically overestimate z_{phot} at low z_{spec} and underestimate z_{phot} at high z_{spec}; see figure 6 & 7 in Bilicki et al. (2014), figure 8 in Pasquet et al. (2019), figure 3 in Beck et al. (2016) and figure 4 in Tarrío & Zarattini (2020). Moreover, the large scatter of photo-z is a serious issue at low redshift, where the standard deviation of $\delta z = z_{\mathrm{phot}} - z_{\mathrm{spec}}$ is comparable to z_{spec} at low redshift. These defects are intrinsic issues to photo-z, which make the photo-z results unreliable in the local volume. For this reason, photo-z is not a qualified solution to our goal. This makes us investigate a redshift-binning classification, which leads to the solution in this paper. With the redshift-bins as predicted results, we do not suffer the systematical bias so badly and can adjust the classification threshold accordingly as intended.

We are now in between LVK's third (O3) and fourth observing (O4) run, LVK is upgrading their GW detectors to achieve higher sensitivity in the next observing run. According to the prospects from LVK (Abbott et al., 2020a), the expected median luminosity distance of the detectable events for BNS and NSBH in O4 are $170\,\mathrm{Mpc}$ and $330\,\mathrm{Mpc}$ respectively with all four detectors, we should be prepared to react to more distant mergers that possibly locate beyond the $200\,\mathrm{Mpc}$ local volume. The predicted sensitive volume (the space-time volume surveyed per unit detector time) for the fifth observing (O5) run is expected to extend significantly compared to O4, by then, our current catalog limited to $200\,\mathrm{Mpc}$ will no longer be sufficient for searching the EM follow-ups and their host galaxies. Here we discuss the possibility of extending our catalog to a larger volume. In this work we limit our prediction volume to $200\,\mathrm{Mpc}$ for the greatest utilization of our unique Hα data. However, the framework we built does not have such constraint intrinsically, its limit is only constrained by the training data we provide. We currently have four positive classes corresponding to the redshift-bins of the four Hα bands, if we extend

the positive classes to higher redshifts (e.g., add a fifth positive class for $0.0471 < z < 0.1$), a new model classifying galaxies in a new larger "local volume" can be easily established. When above $z = 0.0471$, the Hα data will no longer provide much useful information, thus we lose accuracy, but the broad-band data is still powerful and can be used for predictions alone. We test a simple circumstance where we target the galaxies in the volume up to a luminosity distance equal to $\sim 465\,\mathrm{Mpc}$ ($z = 0.1$), we add a fifth positive class for $0.0471 < z < 0.1$ and train the RF classifier with the new labeling. The input training data is unchanged except re-labeling the objects between $z = 0.0471$ and $z = 0.1$ as in class-5, we keep all the classifier settings unchanged. Applying this simple modification, our new model can achieve a completeness almost equal to 72% at the same FPR where the $200\,\mathrm{Mpc}$ model results in 80% of completeness. Even though the FPRs for the two models are the same, because the new model has a larger "local volume", the negatives in the "non-local volume" are less, thus the number of contaminants for the new model is smaller, and when combining it with the much larger number of positives, the corresponding contamination will be reduced a lot. We emphasize that this naive new model is not optimized, with more careful studies, it should be able to perform even better. Overall, in the much larger volume, our framework can achieve a promising completeness with reasonable amount of contaminants, its potential for the future is encouraging.

As the demand of complete galaxy catalogs keeps increasing for the studies of cosmology, galaxies and so on, a number of new surveys have been designed targeting the spectra of galaxies. Some of such surveys cover large areas of the sky deeply, observing millions of galaxies, their spectroscopic/spectral redshifts will provide accurate distance information in substantially larger volumes. The Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al., 2016) is conducting a large redshift survey over more than $\frac{1}{3}$ of the sky reaching to redshift 3.5, where about 10 million bright ($r < 19.5$) galax-

ies will be observed with spectra in the redshift range $0.05 < z < 0.4$. Those galaxies are ideal candidates for EM follow-up, however, DESI does not target galaxies closer than $z = 0.05$ ($\sim 200\,\mathrm{Mpc}$) and covers significantly smaller area than CLU. DESI started its 5-year survey in 2021 May, with annual data releases expected in the near future.

The *Spectro-Photometer for the History of the Universe, Epoch of Reionization and Ices Explorer* (*SPHEREx*; Doré et al., 2018) mission is a 2-year all-sky spectral survey in optical and near-infrared. *SPHEREx* will survey hundreds of millions of galaxies and provide more than 120 million high-quality redshifts, those spectral redshifts will greatly enhance the completeness of the host galaxies. *SPHEREx* is sensitive down to 19–$20\,\mathrm{mag}$ in near-infrared, with a planned launch date of June 2024. While DESI and *SPHEREx* will greatly improve our efficiency of EM follow-up searches, they are both limited by their depths. CLU can constrain redshifts for galaxies as faint as PS1 reaches ($\sim 23\,\mathrm{mag}$).

At other wavelengths, radio surveys of neutral hydrogen may be complementary. For instance, the Widefield ASKAP L-band Legacy All-sky Blind surveY (WALLABY; Koribalski et al., 2020) will start later this year. WALLABY will survey three-quarters of the sky ($-90° < \delta < 30°$) in two years for neutral hydrogen (H I) emission up to $z \leq 0.26$, and is expected to detect half a million galaxies with a mean redshift of $z \sim 0.05$. WALLABY will be effective at identifying local galaxies, especially for the southern sky. Metzger et al. (2013) estimates WALLABY could achieve a SFR completeness of about 93% and 44% with respect to total stellar mass out to a distance of $200\,\mathrm{Mpc}$; though less complete, WALLABY will be able to add more galaxies even further. The spectral surveys will push the edge of the host galaxy searching to larger volume and higher completion, but CLU retains its unique value at low redshifts.

Cross-matching CLU against such spectroscopic/spectral surveys can help identify the mis-classified non-local galaxies, which will further improve the purity of the CLU

local galaxies. Moreover, these spectroscopic/spectral galaxies can be used as new training samples to improve our model, and as we described above, to expand our distance limit to catch up the horizons of the future observing runs. The GAMA survey recently released their latest data release 4 (GAMA DR4; Driver et al., 2022), DR4 provides more spectroscopic data than DR3 including the objects between $r = 19.0$ mag to $r = 19.8$ mag in G09 and G12; the new data can be added to the training set as well to further improve the CLU-ML model.

2.7 CONCLUSION

In this paper, we presented the methodology used in creating the CLU-ML catalog, and evaluated the completeness and the accuracy of the catalog. CLU-ML builds a RF classifier using photometry across a large range of EM wavelengths to constrain redshift ranges for the galaxies that have no previous distance, and identify the galaxies in the local universe ($z < 0.047$; ~ 200 Mpc) for EM follow-up of gravitational wave events. CLU-ML uses photometry spanning optical to IR, combining with Hα narrow-band imaging to construct the model. CLU-ML is trained using the GAMA survey spectroscopic redshifts. We utilized the PS1 survey for the detection of astrometric sources in the northern sky (~ 1.5 billion), and used the PS1-PSC catalog to exclude stars.

In a test region, the CLU-ML model achieved 61.5% completeness with FPR $= 0.01$; 80.2% completeness with FPR $= 0.03$; and 90% completeness with FPR $= 0.06$ for galaxies brighter than $r = 19.8$ mag in the local universe. As the compact binary merger rate is correlated with the astrophysical properties of the host galaxy, we also evaluated the completeness of our model weighted by B-band luminosity, SFR and stellar mass. Due to the difficulty of deriving the astrophysical properties for the high-z galaxies, this evaluation was limited to the $z < 0.2$ volume. In the $z < 0.2$ volume, our model recovered more

than 77% of the total mass with $\mathrm{FPR} = 0.01$; 92% of stellar mass with $\mathrm{FPR} = 0.03$ in the local universe. The model has comparable performance with B-band luminosity and SFR. Note that roughly half of the galaxies in our test set are more distant than $z = 0.2$, thus the FPRs were greatly overestimated; we assess the actual FPRs are $\sim 0.5\%$ and $\sim 1.5\%$ respectively. We showed that the CLU-ML method outperformed the Csig method applied in the CLU-Hα survey by achieving higher completeness. This is due to the profit of ML and extra photometric data. We found that most of the misclassification happened near the local volume boundary ($z = 0.047$). Finally, we compared the CLU-ML method with the GLADE+ catalog by specifying a threshold which led to a completeness of 80% for the method, and further explored the behavior of our model. The CLU-ML method recovered slightly more B-band luminosity and modestly higher fraction of local galaxies in one of the densest patches in the GLADE+ coverage. Although the CLU-ML method could produce a local galaxy catalog with very high completeness, we warn the users that the constrained redshift ranges in the local universe are not reliable due to biasing towards class-3 and class-4. We listed a series of suggested thresholds along with their performance in Table 2.3.

The CLU-ML catalog is a unique complement to our current knowledge of the galaxies nearby, and will serve as an important tool in the EM follow-up campaigns during the upcoming LIGO-Virgo-KAGRA observing runs. Moreover, the CLU catalog is an ideal resource for searching extreme emission-line galaxies such as BCDs and green peas. The framework of CLU-ML can be easily extended to larger volume to suite the more distant horizons of the future observing runs. With a simple test model whose boundary is extended to $z = 0.1$ ($\sim 465\,\mathrm{Mpc}$), we estimated that it was able to identify 72% of the galaxies in the targeted volume with the same FPR where the $200\,\mathrm{Mpc}$ model recovered 80% of the local galaxies. As a few ongoing and upcoming large spectroscopic galaxy surveys (e.g.,

70

DESI, *SPHEREx*) will measure precise redshifts for millions of galaxies, CLU will be able to use the new distance information to evolve the model and extend its boundary.

CHAPTER 3

Conclusions and Future Directions

Gravitational waves (GWs) provide a new way of observing the universe, making the behavior of compact objects in strong-gravity environments clearer than ever. The direct detection of GWs not only proves the validity of general relativity once again, but also makes active localization of the merging binaries possible. Scientists have been hypothesizing that the merger of such compact object binaries power high-energy astrophysical phenomena like GRBs and kilonovae (Eichler et al., 1989; Li & Paczyński, 1998) for a long time, and observations with electromagnetic waves (EMs) have provided indirect evidences supporting these hypothesises. However, it was the joint observation of GWs and EMs from the same source that confirmed the neutron star (NS) mergers are indeed the progenitors of these violent EM phenomena.

The targeted follow up towards the afterglows of the mergers (EM follow-up) searches the transients in the probability area given by GWs, with the purpose of identifying the EM counterparts promptly. As the successful EM follow-up to GW170817 has shown the great scientific outputs that can be derived from joint observations, we want to identify the EM counterparts accurately and early, so longer and deeper observations can be performed. It is particularly important for kilonovae because of their rapid evolution in hours to days (Metzger & Berger, 2012; Metzger, 2019a). Such a goal is not easily accomplishable due to the large areas to search and numerous false positive transients. GW170817 remains the only neutron star merger whose identified EM counterpart is widely accepted. In Chapter 2, I discussed this problem in detail, and claimed a complete galaxy catalog in the local universe can improve the efficiency of EM follow-ups thus increase the probability of detecting the EM counterparts. I described the efforts

of constructing such a catalog, the census of the local universe (CLU), and evaluated its completeness, the results were promising. At this moment, the fourth observing run (O4) of LVK is around the corner in Spring 2023, and CLU is a perfect complement to existing galaxy catalogs that can contribute its unique value in O4 and upcoming runs.

3.1   USAGE AND LIMITATIONS

The CLU-ML catalog presented in Chapter 2 can be utilized in different circumstances, where the main idea is to reduce the time for spotting the host galaxy. People can either target the likely host galaxies with a telescope to reduce the pointings; or accelerate the process of eliminating false positives by associating them to distant galaxies that will never be seen by the GW detectors. Sophisticated tiling strategies can also be applied with CLU-ML to achieve a higher fraction of the probability area in limited time, or to prioritize towards more massive/greater SFR galaxies that have higher chance of hosting NS mergers. Various possibilities can be attempted, the different approaches can even be combined to generate the optimal solution.

Despite all these potentials, the CLU-ML catalog has limitations that cannot be ignored. The current version of CLU-ML selects local galaxies out to $200\,\mathrm{Mpc}$. The LVK detectors undergo upgrades between observing runs. The expected median luminosity distance for detectable BNS mergers is $170\,\mathrm{Mpc}$ in O4 (Abbott et al., 2020a), and this number is likely much greater ($\sim 300\,\mathrm{Mpc}$; Abbott et al., 2020a) in the fifth observing run (O5). By then, many BNS signals may exceed CLU-ML's $200\,\mathrm{Mpc}$ distance limit. The NSBH signals may come from even further distances. To catch up the evolution of the GW detectors, CLU-ML needs to extend its boundary as well; we discussed the details in Section 2.6. CLU-ML is trained using the GAMA spectra. GAMA is highly complete down to $r = 19.8\,\mathrm{mag}$, but its completeness drops quickly at fainter magnitudes. This

being said, CLU-ML does not have a lot training samples in the faint regime, which may cause lower accuracy there. With appropriate thresholds, CLU-ML can easily eliminate more than 97% of non-local galaxies while recovering 80% of local galaxies which contribute more than 90% of the total mass and SFR. But due to the extremely asymmetric ratio of local and non-local galaxies, as many as about half of the selected local galaxies can be false positives. This caveat has to be kept in mind when using CLU-ML. Another deficiency of the catalog is the constrained redshifts for the local galaxies are not very reliable as we discussed in Sec. 2.5.1.

3.2   FUTURE WORK

3.2.1   Finishing Implementation

We have built the model for the CLU-ML catalog, and have constructed the source catalog which contains ~ 1.5 billion objects. But the implementation of the classification method on all the sources is not completed due to the large amount of data and the difficulties we encountered when using databases for managing the data. We will finish the classification, and then release the catalog soon.

3.2.2   Expanding Boundary and Training Samples

As mentioned in Sec. 3.1, CLU-ML is limited by its relatively close boundary and this needs to be extended. The framework of CLU-ML is designed to be flexible. Its boundary is not tied to any specific distance. In the current stage, the boundary is set to $200\,\mathrm{Mpc}$ for the best utilization of the CLU-Hα data, but the model can be easily modified to cover a larger volume with promising performance as discussed in Sec. 2.6. Several large spectroscopic galaxy surveys like DESI (DESI Collaboration et al., 2016) and *SPHEREx* (Doré et al., 2018) have the expect to release their data in the next few years, we can add their

spectra to our training set, then re-train and expand the model. Faint galaxies from deep surveys like the zCOSMOS spectroscopic redshift survey (Lilly et al., 2007) are also desired to improve the accuracy of the model in that regime.

3.2.3 Deep Learning

At this stage, CLU-ML is built using a classical ML algorithm, the random forest (RF) classifier. One possible way to improve the model is to try deep learning methods. Although RF naturally suits astronomical data-sets well and it already produced promising results, deep learning algorithms usually are able to achieve better performance. The deep learning methods have a unique advantage: while the features of the classical algorithms need to be defined by humans, the neural networks of deep learning can automatically learn features from the input data. We constructed a number of features in CLU-ML including colors, shape and other types of information, they are certainly important and useful, but these features may be limited by human power. On the other hand, the neural networks will be able to discover the potentially useful features that missed by human selection. Deep learning methods can possibly help us achieve higher accuracy, so that the contamination with high completeness can be reduced and the constraining on the redshifts of local galaxies can become more reliable.

3.2.4 Telescopes

From the perspective of the EM observers, there are also many improvements can be possibly made. As many times the GW probability areas are large, it is hard for any telescope to cover the entire region in one night, not to mention the GW signals do not always arrive early in a night. They sometimes come in the daytime. But fortunately, there are always telescopes in night on the other side of the earth. Thus more observatories all

over the world should collaborate more closely, to build a relay of observations that cover more areas faster. Groups with full communication can even observe different regions simultaneously using their own telescopes to maximize the area coverage per unit time. The GROWTH collaboration is a great example of this.

Finally, as the fluxes of the galaxies decrease quadratically with distance, the distant kilonovae can be more than an order of magnitude fainter than that of GW170817. The kilonova from GW170817 about $40\,\mathrm{Mpc}$ away peaked at $\sim 17\text{--}18\,\mathrm{mag}$ in r-band (Cowperthwaite et al., 2017), for a kilonova from a similar BNS merger at $400\,\mathrm{Mpc}$, the flux is 100 times fainter, translating to the apparent magnitude is $\sim 5\,\mathrm{mag}$ fainter. So the peaked brightness for such an event is roughly $22\text{--}23\,\mathrm{mag}$, beyond the depth of a lot of follow-up telescopes, and this is especially challenging for spectroscopy. More next generation large telescopes with fast-response systems like the Large Synoptic Survey Telescope (LSST; Ivezić et al., 2019) or even deeper ones are needed.

BIBLIOGRAPHY

Aasi J., et al., 2015, Classical and Quantum Gravity, 32, 074001

Abbott B. P., et al., 2009, Reports on Progress in Physics, 72, 076901

Abbott B. P., et al., 2016a, Phys. Rev. X, 6, 041015

Abbott B. P., et al., 2016b, Phys. Rev. Lett., 116, 061102

Abbott B. P., et al., 2017a, Phys. Rev. Lett., 119, 161101

Abbott B. P., et al., 2017b, Nature, 551, 85

Abbott B. P., et al., 2017c, ApJ, 848, L12

Abbott B. P., et al., 2017d, ApJ, 848, L13

Abbott B. P., et al., 2019, Phys. Rev. X, 9, 031040

Abbott B. P., et al., 2020a, Living Reviews in Relativity, 23, 3

Abbott R., et al., 2020b, Phys. Rev. D, 102, 043015

Abbott B. P., et al., 2020c, ApJ, 892, L3

Abbott R., et al., 2020d, ApJ, 896, L44

Abbott R., et al., 2021a, Physical Review X, 11, 021053

Abbott R., et al., 2021b, ApJ, 915, L5

Abramovici A., et al., 1992, Science, 256, 325

Acernese F., et al., 2006, Classical and Quantum Gravity, 23, S635

Acernese F., et al., 2014, Classical and Quantum Gravity, 32, 024001

Alam S., et al., 2015, ApJS, 219, 12

Arcavi I., 2018, ApJ, 855, L23

Arcavi I., et al., 2017, ApJ, 848, L33

Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540

Artale M. C., Mapelli M., Giacobbo N., Sabha N. B., Spera M., Santoliquido F., Bressan A., 2019, MNRAS, 487, 1675

Artale M. C., Mapelli M., Bouffanais Y., Giacobbo N., Pasquato M., Spera M., 2020a, MNRAS, 491, 3419

Artale M. C., Bouffanais Y., Mapelli M., Giacobbo N., Sabha N. B., Santoliquido F., Pasquato M., Spera M., 2020b, MNRAS, 495, 1841

Baldry I. K., et al., 2010, MNRAS, 404, 86

Baldry I. K., et al., 2018, MNRAS, 474, 3875

Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, MNRAS, 460, 1371

Beck R., Szapudi I., Flewelling H., Holmberg C., Magnier E., Chambers K. C., 2021, MNRAS, 500, 1633

Benítez N., 2000, ApJ, 536, 571

Berger E., 2011, New A Rev., 55, 1

Bertin E., Mellier Y., Radovich M., Missonnier G., Didelon P., Morin B., 2002, in Bohlender D. A., Durand D., Handley T. H., eds, Astronomical Society of the Pacific Conference Series Vol. 281, Astronomical Data Analysis Software and Systems XI. p. 228

Bianchi L., 2009, Ap&SS, 320, 11

Bianchi L., 2014, Ap&SS, 354, 103

Bianchi L., Shiao B., Thilker D., 2017, ApJS, 230, 24

Bilicki M., Jarrett T. H., Peacock J. A., Cluver M. E., Steward L., 2014, ApJS, 210, 9

Bilicki M., et al., 2016, ApJS, 225, 5

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

Breiman L., 1996, Machine learning, 24, 123

Breiman L., 2001, Machine Learning, 45, 5

Bunker A. J., Warren S. J., Hewett P. C., Clements D. L., 1995, MNRAS, 273, 513

Burns E., 2020, Living Reviews in Relativity, 23, 4

Cairós L. M., Caon N., Zurita C., Kehrig C., Roth M., Weilbacher P., 2010, A&A, 520, A90

Calzetti D., et al., 2010, ApJ, 714, 1256

78

Cao L., Lu Y., Zhao Y., 2018, MNRAS, 474, 4997

Cardamone C., et al., 2009, MNRAS, 399, 1191

Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511

Chambers K. C., et al., 2016, arXiv e-prints, p. arXiv:1612.05560

Christodoulou L., et al., 2012, MNRAS, 425, 1527

Colless M., et al., 2001, MNRAS, 328, 1039

Colless M., et al., 2003, arXiv e-prints, pp astro–ph/0306581

Collister A. A., Lahav O., 2004, PASP, 116, 345

Cook D. O., et al., 2019, ApJ, 880, 7

Coughlin M. W., et al., 2019, MNRAS, 489, 5775

Cowperthwaite P. S., et al., 2017, ApJ, 848, L17

Csabai I., Dobos L., Trencséni M., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astronomische Nachrichten, 328, 852

Cutri R. M., et al., 2012, Explanatory Supplement to the WISE All-Sky Data Release Products, Explanatory Supplement to the WISE All-Sky Data Release Products

Cutri R. M., et al., 2021, VizieR Online Data Catalog, p. II/328

DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036

Dálya G., et al., 2018, MNRAS, 479, 2374

Dálya G., et al., 2022, MNRAS, 514, 1403

Doré O., et al., 2018, arXiv e-prints, p. arXiv:1805.05489

Driver S. P., Liske J., Cross N. J. G., De Propris R., Allen P. D., 2005, MNRAS, 360, 81

Driver S. P., et al., 2022, MNRAS, 513, 439

Ducoin J. G., Corre D., Leroy N., Le Floch E., 2020, MNRAS, 492, 4768

Eales S., et al., 2010, PASP, 122, 499

Eichler D., Livio M., Piran T., Schramm D. N., 1989, Nature, 340, 126

Eikenberry S., et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. p. 84460I, doi:10.1117/12.925679

Einstein A., 1916, Annalen der Physik, 354, 769

Eisenstein D. J., et al., 2011, AJ, 142, 72

Farrow D. J., et al., 2014, MNRAS, 437, 748

Fernández R., Metzger B. D., 2016, Annual Review of Nuclear and Particle Science, 66, 23

Finkelstein S. L., et al., 2019, ApJ, 879, 36

Fitzpatrick E. L., 1999, PASP, 111, 63

Flewelling H. A., et al., 2020, ApJS, 251, 7

Freund Y., Schapire R. E., 1997, Journal of computer and system sciences, 55, 119

Gao H., Ding X., Wu X.-F., Zhang B., Dai Z.-G., 2013, ApJ, 771, 86

Gehrels N., Cannizzo J. K., Kanner J., Kasliwal M. M., Nissanke S., Singer L. P., 2016, ApJ, 820, 136

Gordon Y. A., et al., 2017, MNRAS, 465, 2671

Grossman D., Korobkin O., Rosswog S., Piran T., 2014, MNRAS, 439, 757

Guseva N. G., et al., 2020, MNRAS, 497, 4293

Hanna C., Mandel I., Vousden W., 2014, ApJ, 784, 8

Hartle J. B., 2021, Gravity: An Introduction to Einstein's General Relativity. Cambridge University Press, doi:10.1017/9781009042604

Haynes M. P., et al., 2011, AJ, 142, 170

Herner K., et al., 2020, Astronomy and Computing, 33, 100425

Hopkins A. M., et al., 2013, MNRAS, 430, 2047

Hotokezaka K., Piran T., 2015, MNRAS, 450, 1430

Hotokezaka K., Nissanke S., Hallinan G., Lazio T. J. W., Nakar E., Piran T., 2016, ApJ, 831, 190

Hulse R. A., Taylor J. H., 1975, ApJ, 195, L51

Ivan T., 1976, I.E.E.E. TRANS. SYST. MAN CYBERN.; U.S.A.; DA. 1976; VOL. 6; NO 6; PP. 448-452; BIBL. 3 REF.

Ivezić Ž., Connelly A. J., Vand erPlas J. T., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy

Ivezić Ž., et al., 2019, ApJ, 873, 111

Jarrett T. H., et al., 2011, ApJ, 735, 112

Jarrett T. H., et al., 2013, AJ, 145, 6

Jaskot A. E., Oey M. S., 2013, ApJ, 766, 91

Jones D. H., et al., 2009, MNRAS, 399, 683

Kagra Collaboration et al., 2019, Nature Astronomy, 3, 35

Kasliwal M. M., et al., 2017, Science, 358, 1559

Kaur H., Pannu H. S., Malhi A. K., 2019, ACM Comput. Surv., 52

Khostovan A. A., et al., 2020, MNRAS, 493, 3966

Kisaka S., Ioka K., Nakamura T., 2015, ApJ, 809, L8

Koekemoer A. M., et al., 2007, ApJS, 172, 196

Kopparapu R. K., Hanna C., Kalogera V., O'Shaughnessy R., González G., Brady P. R., Fairhurst S., 2008, ApJ, 675, 1459

Koribalski B. S., et al., 2020, Ap&SS, 365, 118

Kovlakas K., et al., 2021, MNRAS, 506, 1896

Kron R. G., 1980, ApJS, 43, 305

Kunth D., Östlin G., 2000, A&A Rev., 10, 1

Law N. M., et al., 2009, PASP, 121, 1395

Leauthaud A., et al., 2007, ApJS, 172, 219

Li L.-X., Paczyński B., 1998, ApJ, 507, L59

Lilly S. J., et al., 2007, ApJS, 172, 70

Liske J., et al., 2015, MNRAS, 452, 2087

Liu H., Motoda H., 1998, Feature extraction, construction and selection: A data mining perspective. Vol. 453, Springer Science & Business Media

Lyke B. W., et al., 2020, ApJS, 250, 8

Makarov D., Prugniel P., Terekhova N., Courtois H., Vauglin I., 2014, A&A, 570, A13

Mapelli M., Giacobbo N., Toffano M., Ripamonti E., Bressan A., Spera M., Branchesi M., 2018, MNRAS, 481, 5324

Mason L., Baxter J., Bartlett P., Frean M., 1999.

Metzger B. D., 2019a, Living Reviews in Relativity, 23, 1

Metzger B. D., 2019b, Annals of Physics, 410, 167923

Metzger B. D., Berger E., 2012, ApJ, 746, 48

Metzger B. D., et al., 2010, MNRAS, 406, 2650

Metzger B. D., Kaplan D. L., Berger E., 2013, ApJ, 764, 149

Miller A. A., Kulkarni M. K., Cao Y., Laher R. R., Masci F. J., Surace J. A., 2017, AJ, 153, 73

Murphy E. J., et al., 2011, ApJ, 737, 67

Nakar E., Piran T., 2011, Nature, 478, 82

Nissanke S., Kasliwal M., Georgieva A., 2013, ApJ, 767, 124

Norris M. A., Meidt S., Van de Ven G., Schinnerer E., Groves B., Querejeta M., 2014, ApJ, 797, 55

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26

Pedregosa F., et al., 2011, the Journal of machine Learning research, 12, 2825

Phinney E. S., 1991, ApJ, 380, L17

Piran T., Nakar E., Rosswog S., 2013, MNRAS, 430, 2121

Quinlan J. R., 2014, C4. 5: programs for machine learning. Elsevier

Richards J. W., Starr D. L., Miller A. A., Bloom J. S., Butler N. R., Brink H., Crellin-Quick A., 2012, ApJS, 203, 32

Roberts L. F., Kasen D., Lee W. H., Ramirez-Ruiz E., 2011, ApJ, 736, L21

Rodrigo C., Solano E., 2020, in XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society. p. 182

Rodrigo C., Solano E., Bayo A., 2012, SVO Filter Profile Service Version 1.0, IVOA Working Draft 15 October 2012, doi:10.5479/ADS/bib/2012ivoa.rept.1015R

Salmon L., Hanlon L., Jeffrey R. M., Martin-Carrillo A., 2020, A&A, 634, A32

Salmon L., Hanlon L., Jeffrey R. M., Martin-Carrillo A., 2021, in Revista Mexicana de Astronomia y Astrofisica Conference Series. pp 67–74, doi:10.22201/ia.14052059p.2021.53.17

Saunders W., et al., 2004, in Moorwood A. F. M., Iye M., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 5492, Ground-based Instrumentation for Astronomy. pp 389–400, doi:10.1117/12.550871

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Scornet, Erwan 2017, ESAIM: Procs, 60, 144

Sharp R., et al., 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 62690G (arXiv:astro-ph/0606137), doi:10.1117/12.671022

Smith G. A., et al., 2004, in Moorwood A. F. M., Iye M., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 5492, Ground-based Instrumentation for Astronomy. pp 410–420, doi:10.1117/12.551013

Sobral D., et al., 2009, Monthly Notices of the Royal Astronomical Society, 398, 75

Stroe A., Sobral D., 2015, MNRAS, 453, 242

Sun H., Zhang B., Gao H., 2017, ApJ, 835, 7

Tachibana Y., Miller A. A., 2018, PASP, 130, 128001

Tanaka M., Hotokezaka K., 2013, ApJ, 775, 113

Tarrío P., Zarattini S., 2020, A&A, 642, A102

Taylor J. H., Weisberg J. M., 1989, ApJ, 345, 434

The LIGO Scientific Collaboration et al., 2021, arXiv e-prints, p. arXiv:2111.03606

Troja E., et al., 2016, ApJ, 827, 102

White D. J., Daw E. J., Dhillon V. S., 2011, Classical and Quantum Gravity, 28, 085016

Wright E. L., et al., 2010, AJ, 140, 1868

York D. G., et al., 2000, AJ, 120, 1579

Zhang B., 2013, ApJ, 763, L22