Multi-messenger astronomy in the era of gravitational wave detections

by

Thomas L. Killestein

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Physics

June 2023

"A still more glorious dawn awaits, not a sunrise, but a galaxy-rise. A morning filled with 400 billion suns. The rising of the Milky Way."

Carl Sagan

Contents

List of Tables

List of Figures

Acknowledgments

When I first started my PhD, I was told by someone I don't remember that the PhD experience is more like a marathon than a sprint. As I sit here making the final edits prior to submission, reflecting on the past 3.5 years, the other side of a global pandemic, I never could have anticipated how true those words would be. It is important that I thank everyone who supported me along this journey, now that I am arriving at the destination.

Above all, I'd like to thank my supervisors for the kindness, guidance, and patience they have shown me over the past years. To Prof. Danny Steeghs, for constantly pushing me to be the best scientist I can, encouraging me to be independent and creative, and supporting me to strike out in new directions. To Dr. Joe Lyman, for being a constant mentor, collaborator, and source of support throughout some of the hardest times, for teaching me that good code is self-documenting, and for always having insightful thoughts and feedback on whatever I'm working on. Without both of you, this thesis would simply be a hollow collection of pages, and the PhD would not have been as enjoyable and rewarding as it was.

I'd also like to acknowledge Warwick Astro past and present, who have made this journey far more enjoyable and provided so much fun and joy over the years. I wish I could name all of you, but given the size of the group now I have to be mindful of the word limit! Special thanks go to Cat, for being a constant source of good humour and support even throughout the shared pain of thesis writing, to my office mates the Bens and Ginger for bearing with me while I was in thesis mode.

Finally, I am truly grateful to my parents, and my sister, for their constant love and support throughout this long journey. Without your fostering of my curiosity, patience, and kindness, I would not be the person I am today, and likely would not be here writing this text.

Despite this thesis and my studies focusing on the skies above, perhaps the biggest revelation of my PhD is that what matters most is happening down here on Earth, and I am eternally grateful to all who helped make these past years as extraordinary as they have been, whether explicitly noted here or not. The Universe is a far richer place for you, and I will carry fond memories of these times forward with me, wherever that may lead.

Declarations

This thesis is submitted to the University of Warwick Doctoral College in satisfaction of the requirements of the degree of Doctor of Philosophy, and I declare that all work presented herein is my own. The contents of this thesis have not been submitted for any other degree at any university.
Multiple chapters of this thesis are based on peer-reviewed publications accepted to journals:

- Chapter 3 is based on a MNRAS publication, entitled *Transient-optimized real-bogus classification with Bayesian convolutional neural networks - sifting the GOTO candidate stream* (Killestein et al., 2021).
- Chapter 4 is based on the MNRAS publication *Precision Ephemerides for Gravitational-wave Searches - IV: Corrected and refined ephemeris for Scorpius X-1* (Killestein et al., 2023)

Specific details of co-author contributions to these works are given at the beginnning of the corresponding chapter for clarity. Content from other chapters will form the basis of further peer-reviewed publications in future, under preparation at the time of writing this thesis.

Abstract

From the first detection of gravitational waves, to the discovery of the optical counterpart to a binary neutron star merger, gravitational-wave multi-messenger astronomy has been a powerful driving force in the development of new large-scale optical sky surveys, techniques, and methods, seeking to exploit the powerful synergies this new way of looking at the Universe has unlocked. This thesis is a compilation of original work from across time-domain astronomy – with a common thread of applying statistical methods to large datasets to extract new conclusions.

Chapters 3 and 6 fuse deep learning and databases to build high-performance, uncertainty-aware source classification algorithms for large-scale optical sky surveys, breaking new ground in integrating contextual information directly into deep-learned classifiers. Chapter 4 constructs a Bayesian inference pipeline for homogeneous reprocessing of over 20 years of high-resolution spectra of the principal continuous-wave source and cornerstone LMXB, Sco X-1 – delivering the most precise ephemerides for the system thus far to enable high-sensitivity searches for gravitational waves. Chapter 5 presents a search for short- timescale variability in supernova light curves, with the aim of providing novel constraints on the structuring and density of the circumstellar medium in these systems. Although null results were obtained, the data constrain the amplitude of and rate of occurrence of previously-observed fluctuations, and allow us to develop the techniques necessary to extend this study to a larger sample in future.

Chapter 1

Introduction

"I don't believe in astrology; I'm a Sagittarian and we're sceptical."

— Arthur C. Clarke

1.1 The advent of time-domain astrophysics

Since the dawn of time, humanity has been captivated by the dynamic Universe we find ourselves in. The periodicity of the seasons, the re-emergence of familiar constellations from below the horizon, and the slow dynamical dance of the planets across the night sky have formed part of our human culture for many thousands of years – dictating when to sow our crops, guiding epic voyages across vast oceans, and forming the inspiration for ancient myths and legends that have transcended the generations through oral tradition.

This neat, ordered picture of the Cosmos has proven incompatible with reality however, with the night sky presenting countless irregular and chaotic phenomena, many observed with the naked eye throughout antiquity. The arrival of comets from the outer Solar system, fleeting flashes of light from the sky, and stars changing brightness erratically all called the 'static' perspective into question in ancient times, revealing a Universe more complex than could be comprehended at the time, and causing untold problems for the reigning philosophical perspectives of the day. Among the more telling hints of this dissonance is in the translated Arabic name of the naked-eye eclipsing

binary star Algol (β Persei) – the head of the Ghoul.

These 'transients' were among the first hints of the extraordinary complexity of the time-varying sky, that we now know to be populated at all time, length, and energy scales – from short and low-luminosity stellar flares, to the explosive deaths of massive stars as supernovae (SNe) that outshine their entire host galaxy for weeks, to energetic gamma-ray bursts (GRBs) in the distant Universe illuminating the large-scale structure of the Universe. Although ultimatedly non-exhaustive owing to the sheer scale of the field, a brief history of time-domain astrophysics follows, focusing largely on our knowledge of explosive transients.

The earliest verifiable records of astrophysical transients in human history come from compendia published in ancient China. Chinese astronomers noted 'guest stars', bright new astrophysical sources that appeared on short timescales, reached naked-eye brightness, and faded over timescales of months. From the sparse historical records that survive to the present day, and association to supernova remnants identified in all-sky surveys, these are now inferred to be associated with galactic supernovae. One particularly notable 'guest star' (Ye, 500) remained visible to the naked eye for 18 months after its' sudden appearance in the Southern Gate (南門) – the modern Western constellation of Circinus. It would be over 1800 years before this event was tied to the diffuse emission nebula RCW 86 as our earliest recorded SN, SN 185: with X-ray studies yielding a consistent age, and elemental abundances pointing towards a Type Ia supernova (SN Ia) origin (Williams et al., 2011).

Although numerous other naked eye supernovae were reported throughout antiquity (Tycho's Supernova in 1572 ; Kepler's Supernova in 1604), systematic study of such objects (and indeed, understanding their true astrophysical significance) was not possible until the early 20th century with the availability of photographic plates. With a means to now objectively capture the night sky, and make precise and repeatable measurements, more systematic efforts could truly begin.

1.1.1 The dawn of supernova science

The first major supernova surveys began from the Palomar Observatory in the late 1930s, led by Walter Baade and Fritz Zwicky, imaging the sky using the Palomar 18-inch Schmidt telescope. Supernovae were identified using a blink comparator, which flipped between two aligned images of the same patch of sky, revealing sources that had changed in brightness between the two observations. The first discovery from this survey was made in March 1937, 18 months after the start of the survey. This foundational work continued until Zwicky's death in 1974, having discovered 120 supernovae over the space of $\sim$40 years, providing a rich sample to begin learning more about these explosive transients.

The theoretical question of what these 'super-novae' could be was addressed in two seminal papers, Baade & Zwicky (1934a,b). These visionary papers illuminate many aspects of supernovae that are taken for granted today – that supernovae begin life as ordinary stars and these events constitute an explosive end to their lives, that these events must be in a different luminosity class to any common novae observed thus far (owing to their association with distant galaxies), and the first tentative rate estimates for both our own galaxy and the Local Volume. One particularly prescient remark was that supernovae may represent the transition of an ordinary star into a dense, compact neutron star, largely supported by neutron degeneracy pressure. The energetics of this situation suggested production of high-energy cosmic rays, with this later being suggested to originate from particles being accelerated in shocks (Fermi, 1949). It would be over 70 years later that gamma-ray observations with *Fermi*confirmed supernova remnants as among the cosmic ray sources, a testament to this early theoretical work. Both papers underscored the need for more follow-up observations: both high-quality multi-colour light curves and spectroscopy. These techniques remain the cornerstone in our understanding of supernovae, even today.

The first high-quality spectroscopic studies of supernovae began in 1941, with Minkowski (Minkowski, 1941) obtaining the first optical spectra of Zwicky's discoveries. These were classified into two broad groups: Type I and II, depending whether they showed signatures of hydrogen – the progenitor to the supernova classification scheme

in use today. By the 1970s (Oke & Searle, 1974), sufficient numbers been classified such that 'peculiar' SNe could be identified, that did not neatly fit into the Type I/II dichotomy. A more in-depth description of supernovae and other explosive transients is deferred to Section 1.2.2, but it suffices now to say this provided an early hint of the extraordinary complexity present in the transient zoo – and there is still no fully satisfactory classification scheme at the time of writing.

1.1.2 SN 1987A – the birth of multi-messenger astronomy

One specific supernova, SN 1987A (Arnett et al., 1989), merits special discussion: as the main instigator of modern multi-messenger astrophysics, the closest supernova in modern astronomy, and as a cornerstone in our understanding of core-collapse supernovae (CCSNe) at all observable wavelengths (Wooden et al., 1993; Chevalier & Dwarkadas, 1995; Matsuura et al., 2022; Larsson et al., 2023), at late times (Woosley et al., 1989; Jerkstrand et al., 2011; Arendt et al., 2020), and with resolved ejecta (Fryxell et al., 1991; Mueller et al., 1991; Wang et al., 2002). Discovered while still rising, SN1987A (Kunkel et al., 1987) occurred in the Large Magellanic Cloud (50 kpc; Pietrzyński et al. 2013), a satelite galaxy of the Milky Way. A supernova at this distance enabled two key landmark discoveries. Deep pre-imaging of the Large Magellanic Cloud (LMC) enabled the identification of the progenitor star for the first time. Sanduleak -69 202, a blue supergiant star in the Tarantula Nebula, was identified on photographic plates (Sanduleak, 1970) as astrometrically consistent (West et al., 1987) with the explosion site. The observation of the progenitor served as direct confirmation of the supernova-massive star connection, yet also raised additional questions: CCSNe were thought to be primarily from larger red supergiants, with signficant implications for the progenitor system (Podsiadlowski, 1992). Later work has revealed that SN 1987A is not a typical CCSN: evolving far more slowly, and with a lower overall luminosity as a result of the different progenitor. A small but growing class of '1987A-like' supernovae has emerged, showing comparable behaviour and consistently pointing towards a blue supergiant (BSG) origin. The detection of 31 neutrinos (Bionta et al., 1987; Hirata et al., 1987) from the direction of the LMC further cemented this event as the first extragalactic multi-

messenger event. The flux of neutrinos provided crucial data with which to test existing supernova explosion models, as well as placed an upper limit on the mass of the neutrino (Burrows & Lattimer, 1987). For the first time, a test of supernova physics with both light and neutrinos was possible, enabling powerful tests of the theorised mechanisms, nuclear reactions, and energy transport occurring at the heart of the supernova explosion.

Astronomers and particle physicists eagerly await the next CCSN in the local Universe (Adams et al., 2013), with new detectors primed and ready to detect $\sim$thousands of neutrinos, make detailed inferences about the explosion mechanisms (Migenda, 2020) and deep core physics of supernovae, and provide some of the most constraining tests yet of fundamental physics. Rates remain uncertain, ($\sim$1/century in the Milky Way; Tammann et al. 1994; Reed 2005), and there is a non-zero chance such an event will occur in the Zone of Avoidance created by the dust extinction from our own Milky Way. Nevertheless, such an event is sure to bring about a significant paradigm shift in our understanding of the physics of supernova explosions, in a way that observing more distant extragalactic supernovae simply cannot: we must be ready.

1.2 The Transient Zoo

With the step changes in survey capability over the past decades, we have discovered significant numbers of supernovae and other extragalactic transients, and begun to unveil a rich taxonomy of different types of transient, each with unique progenitor systems, pathways to explosion, and observable properties. In the following sections, I summarise some of the key properties of supernovae and other transients.

1.2.1 The landscape of transients

To contextualise the diversity of transients we see, it is useful to consider the key observables of transient events that connect directly to our theoretical models: the luminosity and characteristic timescale (Kulkarni, 2012). These two observables constrain directly the underlying physics of the transient (Villar et al., 2017) – with radioactively- powered

transients for example having tightly-constrained timescales owing to the fixed half-life of radioactive isotopes (with variance emerging from ejecta velocity.) Similarly, the luminosity of SNe Ia is tightly constrained due to the (approximately) fixed progenitor mass through the Chandrasekhar limit, yielding a comparable nickel mass for each transient. Nickel-powered transients show a characteristic late-time decay slope (see Figure 1.2), fixed primarily by the half-life of the isotopes involved. Figure 1.1 shows the luminosity-timescale paraameter space of transients, using data from the ZTF Bright Transient Survey (Perley et al., 2020) sample split into broad phenomenological classes (see Section 1.2.2). As noted above, thermonuclear transients show a very narrow distribution of timescales and peak absolute magnitudes, whereas the more complex core-collapse-like transients show a broader range owing to their broader ranges of progenitors. Significant diversity in the light curves of specific transient classes is challenging, although exciting: this means the properties of the explosion are strongly governed by parameters of the system, and thus successful modelling can yield insights into the nature of these transients, and constraints on their properties – *if* we can understand the underlying physics.

A key open question however surrounds the 'filling' of this luminosity-timescale parameter space – where do the limits of this parameter space lie? Is this plot fully filled with transients of all types, or are there regions 'forbidden' by explosion physics or stellar evolution? Conversely, pushing out into regions unexplored by current surveys may yield novel new physics, insights into stellar evolution, and previously unforeseen phenomena. With the higher cadences and greater photometric depths of current-generation transient surveys, we are beginning to fill out regions of this luminosity-timescale parameter space that were previously challenging – populated by only a few serendipitous detections. We are successfully expanding this space to both shorter timescales and fainter luminosities via new surveys, finding new types of transient in seemingly all directions we explore. Nevertheless, we are also unveiling ever larger populations of familiar transients – supernovae, which I discuss in the following Section.

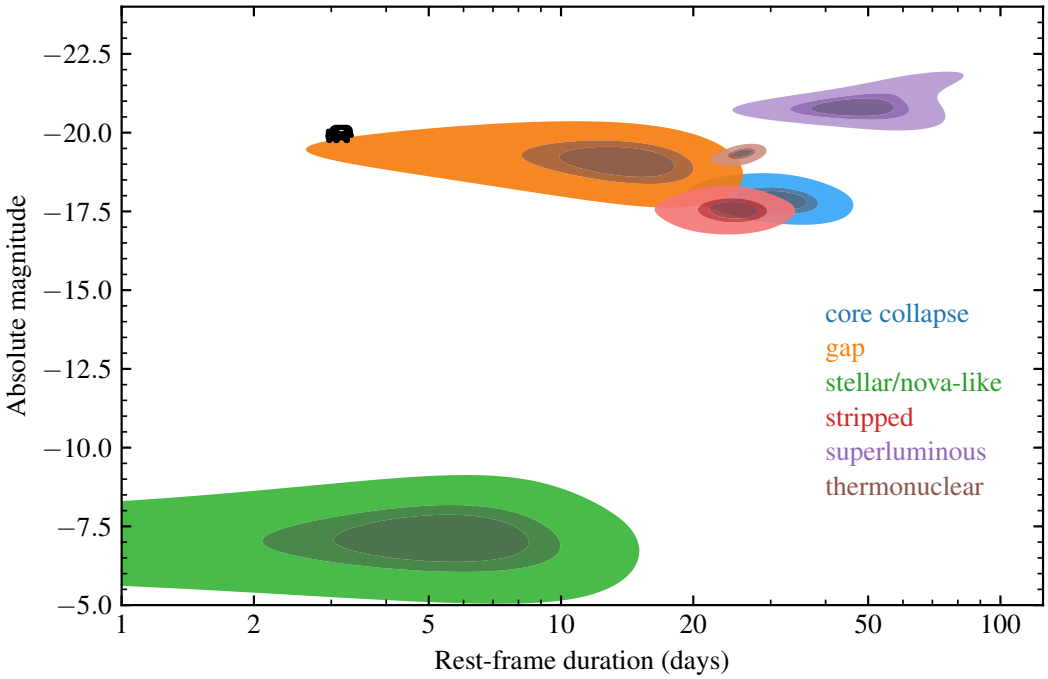

Figure 1.1: Luminosity-timescale plot for supernovae and other associated transients, constructed based on kernel-density estimates of the ZTF-BTS (Perley et al., 2020) sample, aggregated by broad transient class. The 68th, 90th, and 95th percentiles are plotted as contours for each class. The extraordinary fast blue optical transient AT 2018cow is represented with the cow-shaped marker, laying at the extreme end of the gap transient distribution in terms of luminosity and timescale.

1.2.2 Supernovae

Supernovae (SNe) are the violent endpoints of stellar evolution, creating some of the brightest transient events observed and outshining their entire host galaxies for weeks at a time. This extreme brightness has underpinned use of supernovae as cosmological probes (Schmidt et al., 1998; Riess et al., 1998; Perlmutter et al., 1999) due to the great distances they are visible at. Only around 1% of stars will experience this catastrophic fate, with the vast majority ending life as white dwarfs, destined to slowly cool for billions of years unhindered (Althaus et al., 2010). Despite this comparatively small fraction, SNe have a profound influence in the evolution of the Universe across a range of spatial scales. SN explosions are responsible for seeding the interstellar medium (ISM) with heavy elements, with a large amount of these being synthesised via silicon burning (Woosley et al., 1973) in the intense pressure created by the infalling outer layers of the star onto the stiff core. Through this interaction with the ISM, SNe also directly influence the evolution of their galaxies and modulate the star formation rate, via coupling and feedback (Hopkins et al., 2018; Smith et al., 2018). The dual is also true, with the environments in which supernovae explode encoding information about their progenitors (e.g. Modjaz et al. 2008; Rigault et al. 2013; Anderson et al. 2015). As briefly introduced in Section 1.1.1, SNe are separated into classes based on their spectroscopic properties – with the top-level split being between Type I (hydrogen-poor) and II (hydrogen-rich) (Minkowski, 1941). This classic dichotomy has persisted to the present day, with numerous extensions to cover the various supernovae we have uncovered.

SNe Ia are the result of a detonation and deflagration of a CO white dwarf that exceeds the Chandrasekhar limit (1.4 $M_{\odot}$), where neutron degeneracy pressure can fully support the degenerate star against gravitational collapse. When the white dwarf gets sufficiently close to this limit, the increased pressure permits a deflagration front of carbon-oxygen burning to advance through the star, synthesising significant yields of radioactive elements – with heavy elements like nickel (the main source of luminosity) and iron, with the characteristic Si features of SNe Ia being synthesised closer to the surface of the dying white dwarf. The star is destroyed in the process leaving no rem-

nant behind, as sufficient energy is released to fully unbind the white dwarf, launching ejecta at significant velocity. Owing to the precise constraint on the mass at which SN Ia explosions are triggered and the largely homogeneous composition of WDs, the energy budget (and thus luminosity) of these events is tightly bounded – making these transients crucial standard candles for cosmological studies. White dwarfs massive enough to detonate in a SN Ia explosion are typically the result of significant accretion from a main-sequence companion star (single-degenerate origin), or the merger of two white dwarfs (double-degenerate origin), with the progenitor system embedding on the properties of the explosion. In contrast, core-collapse (Type II) supernovae (CCSNe) are a more diverse group of transients, stemming from the death of massive $\gtrsim 5-8\ M_{\odot}$ stars. As these massive stars fuse successively heavier and heavier elements in shells surrounding the core as they approach the end of the their lives, the energy yield from fusion decreases – causing the star to expand to greater radii and cool, whilst the core also contracts inwards. Fusion ceases at Fe/Ni and the core can no longer generate sufficient pressure The intense luminosity of supernovae (both Type I and II) comes primarily from the radioactive decay of ^{56}Ni (Arnett, 1982), synthesised in the explosion – with the half-life of ^{56}Ni setting the characteristic timescales for the decay of luminosity in SN Ia and other hydrogen-poor transients. Hydrogen-rich supernovae from massive star progenitors are more complex (e.g. Heger et al. 2003; Woosley & Janka 2005), partially due to the presence of hydrogen. After the photosphere has cooled sufficiently (4000-6000K), hydrogen begins to recombine providing an additional source of energy at later times. This yields a long (~ 100d) plateau in the light curve in the case of SNe IIP (plateau), where abundant hydrogen is available. In SNe IIL (linear), there is no significant hydrogen envelope, and so this effect is negligible and the light curve drops at approximately the rate of radioactive decay instead. Core-collapse supernovae also find utility in cosmological studies, not as standard candles via their (highly variable, see e.g. Richardson et al. 2014) peak luminosities, but via the 'expanding photosphere' method (Kirshner & Kwan, 1974; Eastman et al., 1996) – linking the expansion velocity (measured from spectral lines) and angular size of the photosphere (from photometry, assuming a blackbody emitter) to obtain a purely geometric distance measurement to

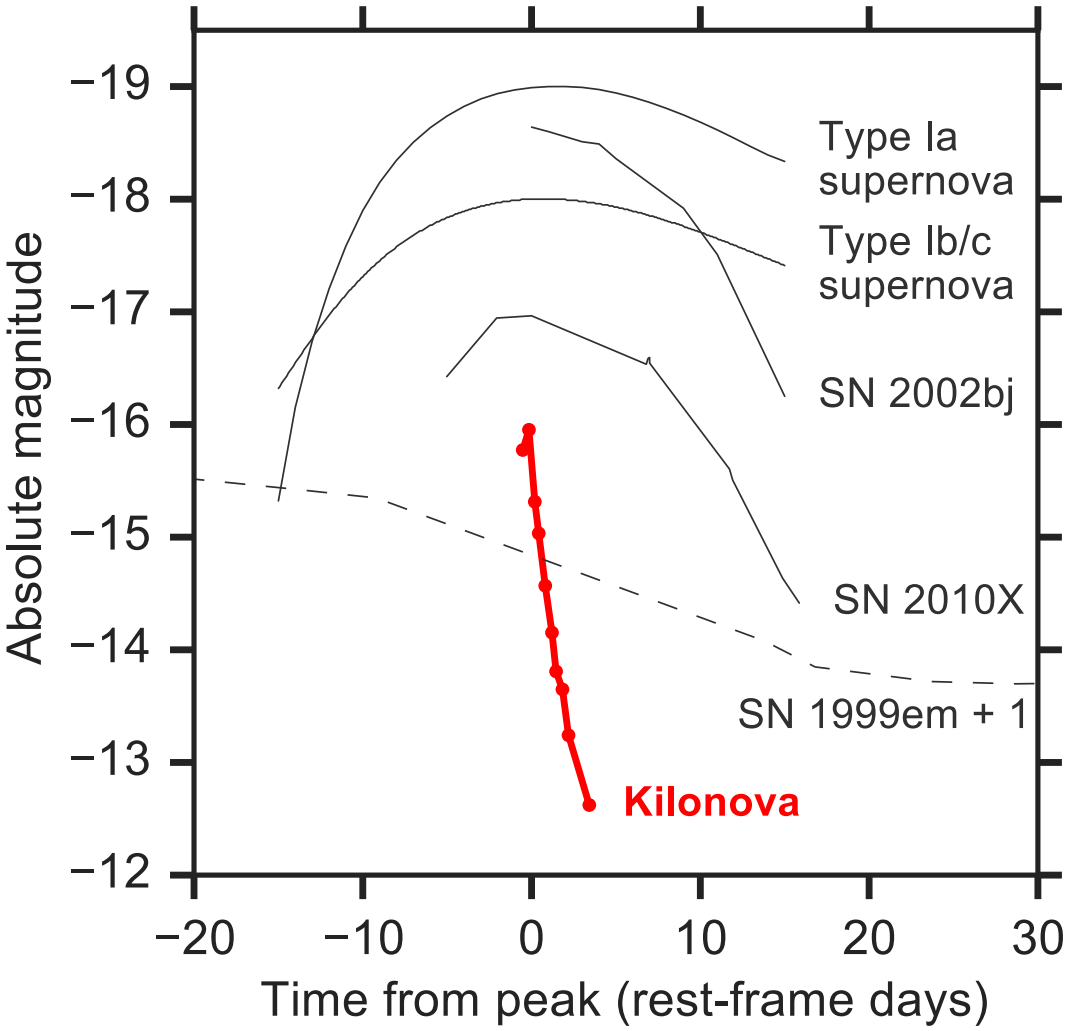

Figure 1.2: Typical light curves for thermonuclear, core-collapse, and stripped supernovae plotted to show their typical timescales (reproduced from Arcavi et al. 2017).

the transient.

Lacking both strong hydrogen emission lines, and the characteristic Si bump, stripped envelope supernovae represent an extreme endpoint of stellar evolution (Lyman et al., 2016). The progenitors of these transients have had their outer layers removed either by intense stellar winds (Smith, 2014), or by binary interaction (e.g. Eldridge et al. 2008. The degree of stripping dictates the overall evolution and observational parameters of these transients, with mildly-stripped supernovae (with a remaining He layer) known as SNe Ib, and supernovae fully stripped of H and He known as SNe Ic – determined primarily by their spectral appearance. These events are nominally core -collapses of massive stars, although are given the SN I label as they are hydrogen-poor. SNe IIb (e.g. SN 1993J Woosley et al. 1994) represent a transitional phase between SNe II and SNe Ib, with some remaining hydrogen still visible in the spectrum, that disappears giving way to strong helium emission.

One key complicating factor is the presence of 'interaction' across a wide range of supernovae (Fraser, 2020). In interacting supernovae, the supernova blast wave ploughs into a surrounding dense circumstellar medium (CSM), driving additional emission from heating of this material. This manifests observationally (Ofek et al., 2014a) as strong, blue emission at early times close-in to the supernova (where the CSM is densest, e.g. Chevalier 1982), with strong, narrow line emission from metals. Depending on the extent and density of the CSM, these emission features can either disappear days after explosion (flash features, e.g. Khazov et al. 2016; Bruch et al. 2022) or persist for weeks afterwards in the case of dense CSM, often denoted by appending an **n** to the classification to denote **n**arrow lines . Unlike stripped envelope supernovae, the mass must remain around the SN progenitor, likely favouring stellar winds or eruptive mass loss (Ofek et al., 2014b) as a mechanism (despite being challenging for thermonuclear explosions like SNe Ia). The most common of these interacting transients are SNe IIn (e.g. Schlegel 1990; Kiewe et al. 2012), with the CSM arising from eruptive mass loss (Chugai et al., 2004) (see Section 5.1) – although populations of more rare Ibn, Icn, and Ia-CSM supernovae have been noted. Signatures of interaction, and searching for them, lie at the heart of Chapter 5 – providing a novel route to test predic-

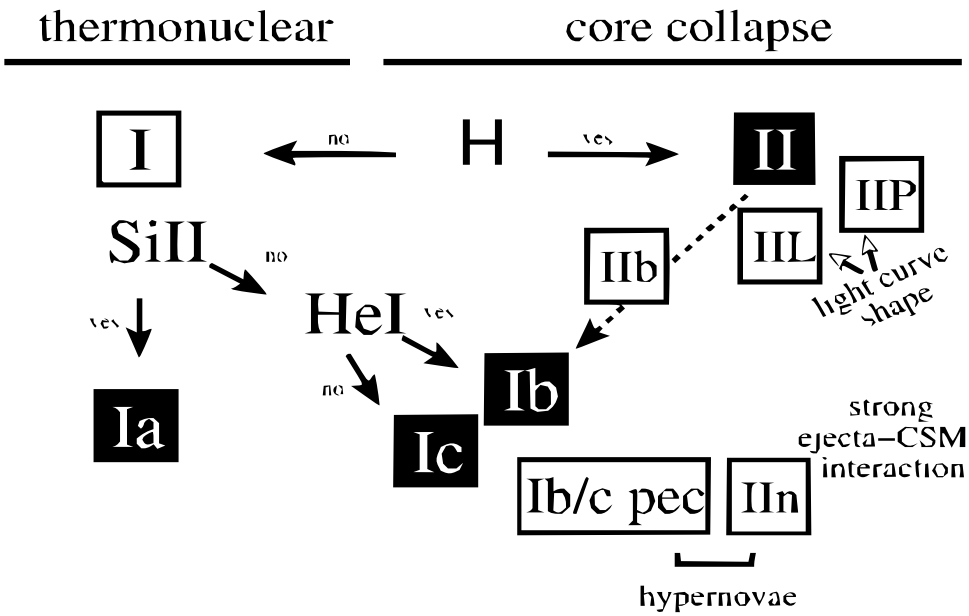

Figure 1.3: Schematic diagram illustrating the various supernova classes, with a particular focus on the often hierarchical nature of the classification schemes. Reproduced from Turatto (2003)

tions of stellar mass loss, and constrain these parameters to improve our evolutionary models going forward.

Figure 1.3 attempts to illustrate the existing classification scheme of supernovae graphically, and emphasises the mismatch between the classification scheme and e.g. progenitor pathways/mechanisms. Although replacement schemes have been proposed (e.g. Gal-Yam 2017), none have succeeded in gaining traction in the community. This is largely driven by the staggering diversity of SN spectra, as well as the difficulties in assigning quantitative 'grades' using data of inhomogeneous quality. As we discover and characterise more transients , we consistently find new objects that defy classification within the framework of whichever labels are currently in use – either being classified as 'peculiar' variants of existing types, or branching out into new classes of their own.

The growing population of fast blue optical transients (FBOTs) (Drout et al., 2014; Pursiainen et al., 2018; Ho et al., 2021) is a key example of this – as survey cadences increased to sample large areas at $\sim$day cadence, a population of tran-

sients with evolution timescales comparable to this emerged. Although the first transients emerged from local, small-area surveys, an increasing number began to emerge serendipitously from wide-field sky surveys (although largely post-peak). The hallmarks of this class of transients are: short time spent above half maximum light $\lessapprox$12d (Drout et al., 2014), and blue optical colours throughout rise/peak. Many FBOTs also show radio/X-ray emission, indicative of shock interaction with a dense circumstellar medium. Among the most remarkable of these events was AT 2018cow (Prentice et al., 2018), a bright, nearby FBOT with a high luminosity and rapid rise time (3.5d) that is the hallmark FBOT owing to the rich dataset we gathered on this object. Since then, many more 'fast transients' (Ho et al., 2021) have been discovered, but only 5 of these are 'Cow-like' in terms of luminosity, indicating again rich diversity in this region of transient parameter space. Many of the sub-luminous FBOTs are thought to be interaction-dominated IIb/Ib/Ic explosions, with the origins of the most luminous members of the class still largely unclear (e.g. Perley et al. 2019; Leung et al. 2020; Lyutikov 2022) owing to the intrinsic difficulty of gathering datasets of the quality required to distinguish between scenarios, especially at early time. Particularly intriguing is the observation of very short-timescale flaring in the 'Cow-like' explosions, up to a factor 100 over timespans of minutes (Ho et al., 2022a). Although further discussion is left to Section 5.1, we note here that this remarkable behaviour is further evidence for compact objects being responsible for many of the properties of these transients – the causal length-scale of this phenomena must be smaller than 10^{12} cm, incompatible with the photospheric radius/shock-front radius at the times observed. Engine-driven variability of a nascent magnetar, or modulation of accretion onto a compact object is compatible both with this timescale, as well as the significant X-ray (e.g. Matthews et al. 2022) and radio emission of these events generated through non-thermal emission.

Closer to home, increasing numbers of intermediate luminosity optical transients (ILOTs) in local galaxies are being discovered through dedicated local- Universe surveys. As their name suggests, these transients characteristically have luminosities between classical novae and supernovae, and have long evolution timescales of $\sim$ months – predominantly red due to the cool, dusty components they eject. ILOTs encapsulate

a diverse range of origins, including luminous red novae (LRNe; Kulkarni et al. 2007; Williams et al. 2015; Cai et al. 2022) from stellar mergers, failed supernovae (Gerke et al., 2015), and supernova 'impostors' (Pessi et al., 2022). From a theoretical perspective we expect many more poorly-characterised transient sub-types to inhabit this region of luminosity-timescale space: the explosive deaths of cataclysmic variable stars (Metzger & Fernández, 2021) and planetary engulfments (De et al., 2023). As always, more observations are required to place these events within context, and further probe the diversity of individual events within our broad classifications.

At the most extreme luminosities, and shortest timescales, lie the explosive GRB, among the most energetic explosions in time-domain astrophysics. Whilst in-depth discussion of GRBs is left to the extensive literature on the subject (see e.g. Paradijs et al. 2000), these events divide into short and long GRBs based on their duration, the T_{90} value, over which 90% of total counts are detected. Short GRBs (sGRBs) have T_{90} values less than 2 seconds, and are believed to originate predominantly from the mergers of binary neutron stars. The electromagnetic counterparts of short gamma-ray bursts receive extensive discussion in Section 1.3.2. Long GRBs originate from the deaths of the most massive stars, powered by the 'collapsar' mechanism – meta-stable accretion disks formed from the outer layers of the star re-accreting down onto the nascent compact remnant of the star. Relativistic jets are launched off of the compact object, which interact with the dense material surrounding the collapsar, and create a gamma-ray burst when they 'punch' out of the material. To both form the accretion disk required and leave a massive-enough central remnant to power the emission, the progenitors of GRBs must be extremely massive ($\gtrsim 40 M_{\odot}$). After the main gamma-ray burst, a luminous supernova can emerge on timescales of $\sim$ days, typically of type Ib/Ic, but with *significant* ejecta velocity from the violent GRB event. Observations of GRB-SNe (e.g. SN 1998bw, Woosley et al. 1999), their host environments, and the lack of H or He in their spectra further cement the collapsar origin of these events.

The section above has provided a broad, non-exhaustive overview primarily focused on our knowledge of transients from the perspective of electromagnetic observations. Recent advancements have enabled us to move beyond this, and obtain

unprecedented new insights into the side of transients we cannot 'see' directly.

1.3 The promise of multi-messenger astrophysics

Multi-messenger astronomy is a powerful emerging paradigm in the study of transient events. Different 'messengers' carry information about different physical processes, potentially occurring at different mass/length/energy scales than classical electromagnetic waves. A key example is in supernova explosions: at early times, the stellar interior/ejecta is optically thick to the oncoming supernova shockwave. The shockwave travels at local sound speed, and must reach the edge of the star/where the CSM becomes optically thin before being visible (shock breakout; Waxman & Katz 2017) – taking hours to days dependent on stellar radius and density. Unlike the electromagnetic emission however, neutrinos can traverse the dense regions relatively unattenuated at the speed of light, emerging well before the shock breakout occurs. Neutrinos not only carry information from the very heart of the supernova explosion, but can also provide early warning of supernova explosions.

Multi-messenger astronomy has only recently come to prominence as a field of research, although multi-messenger observations themselves go back many decades. Our own Sun was the first multi-messenger source, with observations of solar energetic particles (Forbush, 1946) being made and associated to the Sun. These observations probe particle acceleration in flares on the Sun's surface, and have yielded constraints on the elemental abundances of our Sun (Breneman & Stone, 1985). The landmark extragalactic multi-messenger event, SN 1987A Section 1.1.2, remains our cornerstone in the understanding of core-collapse supernovae, and is our principal constraint on the innermost workings of supernova explosions, despite only 31 neutrinos being observed.

A recent significant breakthrough in the domain of multi-messenger astrophysics was the discovery of neutrinos (IceCube Collaboration et al., 2018a) and very high energy (VHE) emission (Ansoldi et al., 2018) temporally coincident with an optical flaring episode in the blazar TXS 0506+05 (IceCube Collaboration et al., 2018b; Keivani et al., 2018). This event cemented the previously-theorised role of active galaxies as nature's

most powerful particle accelerators, and provided a wealth of new observational data to constrain jet models (Cerruti et al., 2019). This picture has been borne out by more recent detections of neutrino excesses from another active galaxy, NGC 1068 (IceCube Collaboration et al., 2022). Further, there is emerging evidence that tidal disruption events (TDEs; van Velzen et al. 2021) are another significant ($\gtrsim 10\%$ of total flux) source of astrophysical neutrinos (Stein et al., 2021, 2023), with some models also predicting observable neutrino emission from kilonovae (Kimura et al., 2018).

Most recently, the advent of gravitational wave (GW) multi-messenger astrophysics with the discovery and in-depth study of GW 170817 (Abbott et al., 2017e) has proven transformative in our understanding of explosive transients, the neutron star (NS) equation of state (EoS) equation of state, and relativistic astrophysics in the extreme – underpinned by the most precise metrology in the physical sciences thus far, distributed across multiple continents. I devote the next few sections to discussing this profound new approach in depth.

1.3.1 Gravitational-wave astrophysics

The existence of GWs were theorised prior to Einstein's general relativity by Poincare and Heaviside as analogs to electromagnetic radiation, however these theories lacked mathematical grounding until the advent of general relativity (GR) by Einstein (Einstein, 1916). The Einstein field equations of GR permit simple plane wave-like solutions in the weak-field regime. These solutions are the 'gravitational waves': transverse, traceless perturbations of space- time – that is they only cause perturbations perpendicular to the direction of propagation, and cause only 'shearing' forces. These perturbations propagate at the speed of light as ripples on the fabric of space-time, and are generated in principle via motion of mass. More specifically, GW sources require non-axisymmetric motion of mass – or more rigorously, the second derivative of the quadrupole moment of the system's stress-energy tensor must be non- zero. This is mathematically encapsulated in the 'quadrupole formula',

$$\bar{h}_{ij}(t,r) = \frac{2G}{c^4 r}\ddot{I}_{ij}\left(t - \frac{r}{c}\right)$$

, where I_{ij} is the stress- energy tensor (with $_{ij}$ representing spatial indices following Einstein sum convention), and h_{ij} is the metric perturbation caused by the gravitational wave itself, sometimes called the 'strain'. Two points are worth remarking upon here. The principal observable h scales as $\frac{1}{r}$, in contrast to the typical $\frac{1}{r^2}$ scaling seen in electromagnetic (EM) astronomy. The amplitude $|h_{ij}|$ scales approximately as I/t^2 – meaning the amplitude of GWs can be made greater by moving more mass-energy, or simply by moving it more quickly. These properties directly inform what we expect to be the most prominent sources of gravitational-wave radiation: the orbital decay of compact binary star systems, the spindown of rapidly-rotating non- axisymmetric sources. Some other astrophysical events like supernovae have predicted GW signatures, although these give off more stochastic 'bursts'.

Despite this, GWs are incredibly feeble, with typical predicted strains $\lesssim 10^{-20}$, requiring metrology below the scale of individual nucleons to detect. Claimed direct detections of GW radiation were made as early as the 1960s (Weber, 1967, 1969) using a 'Weber bar' antenna: this was composed of two aluminium cylinders with resonant frequencies close to the expected frequency of gravitational waves. Although these claims were ultimately not reproducible by numerous independent attempts (e.g. Levine & Garwin 1973), the ideas and metrology developed by Weber laid the groundwork for many modern breakthroughs.

The first experimental evidence for gravitational-wave emission being a real and measurable effect came from radio timing measurements of the binary pulsar PSR J1915+1606 – known now as the Hulse-Taylor binary after the two discoverers. The system is comprised of two neutron stars – one pulsar and one in relative quiescence. The $\approx$ 59ms pulse period of the pulsar, although challenging to measure, provided an exquisitely accurate clock with which to trace the orbit. Through measuring the pulse arrival times over a period of 14 years, it was revealed that the orbital period of the pulsar was decaying (Hulse & Taylor, 1975), indicating a loss of angular momentum from the system. The measured rate of decay ($\dot{\nu} = 8.62713(8) \times 10^{-18}$ s/s; Taylor & Weisberg 1989) shows remarkable consistency with the predicted period derivative expected from gravitational-wave radiation (Peters & Mathews, 1963), providing both a

pioneering test of general relativity and the first (albeit indirect) measurement of GWs.

All current gravitational-wave detectors follow a design inherited from the Michelson interferometer (Michelson & Morley, 1887), by sending a beam of light down two perpendicular arms using a beam-splitter, then combining the two reflected beams and measuring the interference pattern created. Any differences in the path length in each arm will manifest in subtle changes in the interference fringes. This technique enables measurement of changes in the arm lengths significantly smaller than the wavelength of the light used. This is illustrated in Figure 1.4. The principal observable GW detectors measure is the strain time-series $h(t)$ - the ratio of the difference in length between the two perpendicular arms divided by arm length as a function of time. The 'shearing' behaviour of GWs causes one arm to contract (or expand) relative to the other. Naturally, these detectors have a characteristic sensitivity pattern for detection of GW sources, with 'blind spots' where the GW shear axis is perpendicular to the plane of the detector. Localisation requires multiple detectors ideally, with differences in the time of arrival helping to constrain sources to annuli on the sky. The simple descriptions above abstracts away decades of instrumentation development, computational advances, and work – the step change comes not only from significant scale changes ($\sim$cm to $\sim$km) but a continuous heritage of technological advancement (e.g. Lück et al. 2006).

The first direct detection of gravitational waves came in 2015 with GW 150914 (Abbott et al., 2016a,d), a binary black hole (BBH) merger at a distance of around 500 Mpc – generating a peak strain of $h \approx 10^{-21}$, or a deviation in arm length approximately $1/1000^{\text{th}}$ the width of a nucleon. Modelling with numerical relativity suggested this event was caused by the merger of two black holes (36 and 29 $M_\odot$ respectively), with $3M_\odot$ of mass-energy being radiated away in gravitational waves. Alongside validating another direct prediction of general relativity and showing an excellent fit to predictions from numerical relativity, this event further constrained the parameter space for post-Newtonian theories of gravity (Abbott et al., 2016c). This event heralded the beginning of a new era of gravitational-wave astrophysics in its' own right, but also re- invigorated prospects of gravitational-wave multi-messenger astronomy.

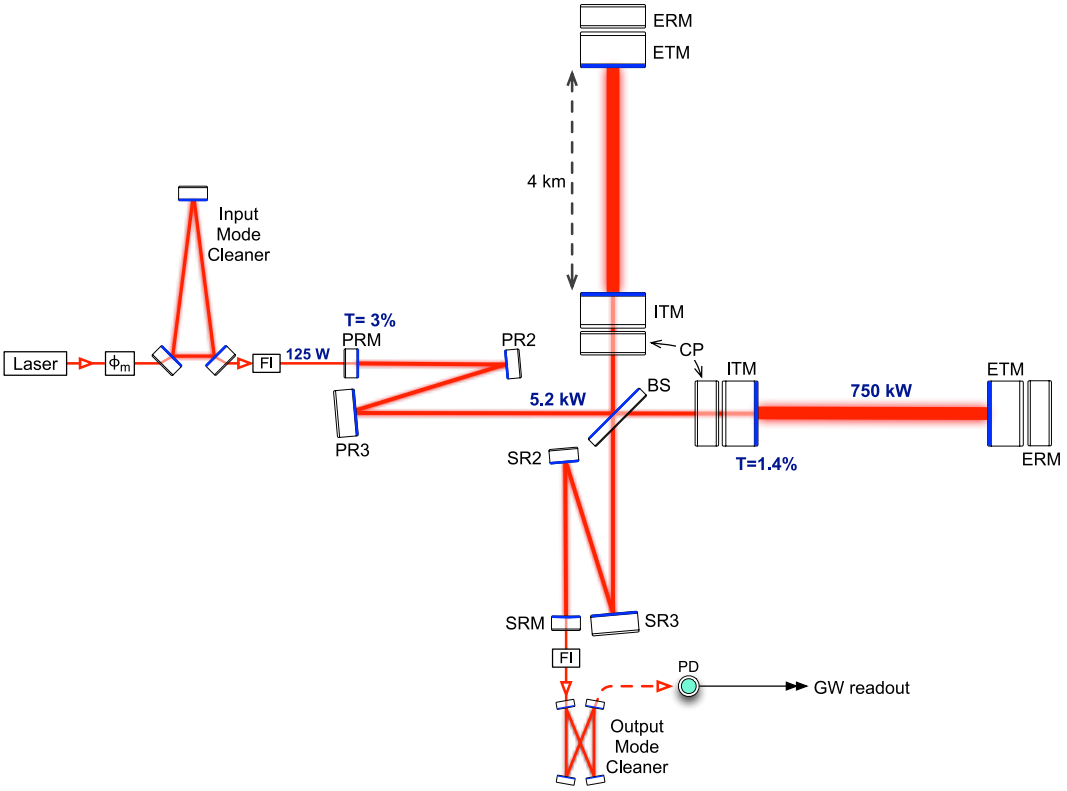

Figure 1.4: Schematic diagram of Advanced LIGO (reproduced from LIGO Scientific Collaboration et al. 2015), showing the Michelson interferometer-based design in more detail. Laser light is split along two perpendicular arms, and recombined to measure phase differences indicating changes in the arm lengths.

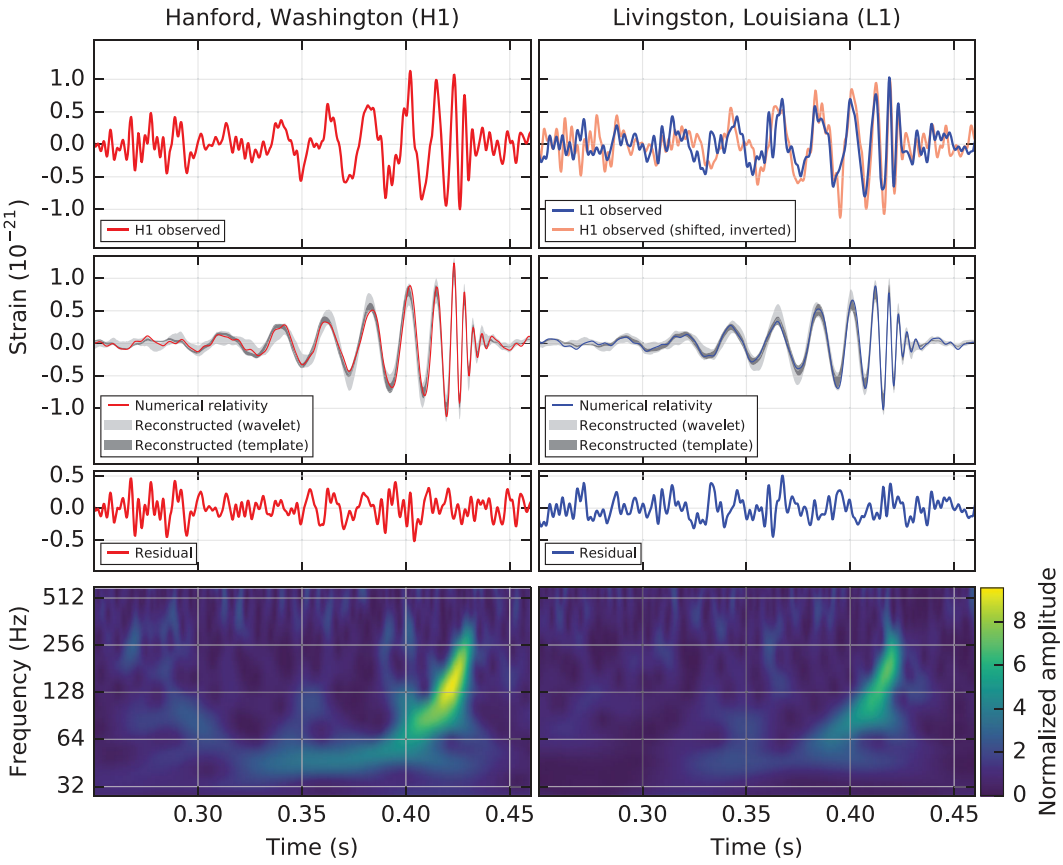

Figure 1.5: The 'chirp' waveform of GW150914 (reproduced from Abbott et al. 2016a), as detected by the LIGO Livingston and Hanford detectors in 2015. The 'chirp' increases in frequency as the binary orbit decays to shorter periods, and the amplitude dramatically increases towards coalescence as it too depends on frequency.

1.3.2 Electromagnetic counterparts to GW-driven mergers

With the unambiguous[1] detection of a compact binary coalescence, focus shifted in the astronomical community to identifying potential EM counterparts to the events being unveiled successfully through gravitational-wave detections. Given the significant amounts of mass being shifted around, it seemed natural that such events would be accompanied by significant, energetic signatures where baryonic matter was involved. Even prior to the detection of GWs, EM counterparts to binary neutron star (BNS) mergers had been theorised to exist (Li & Paczyński, 1998; Kulkarni, 2005; Metzger et al., 2010) – termed 'macro-novae' or 'kilo-novae'. These transients were theorised to occupy a region of luminosity space between classical novae, and SNe – powered by the radioactive decay of freshly-synthesised r-process elements. The synthesis of lanthanides on $\sim$day timescales drives changes in opacity, yielding a rapidly-reddening, fast-evolving transient. Although some transients discovered through high-energy triggers (e.g. GRB 130603B; Tanvir et al. 2013) bore many of the hallmarks of a kilonova, none could be directly linked to a compact binary coalescence. Similar events are predicted for NSBH mergers (Tanaka et al., 2014; Kawaguchi et al., 2016; Barbieri et al., 2019; Gompertz et al., 2023), with lower luminosities (around 2–10 magnitudes less) given the overall smaller mass of ejecta present.

Given that no baryonic matter is present, BBH mergers are expected to be largely EM-dark, as no normal matter is involved in typical isolated mergers of black holes. Some more exotic progenitors (Liebling & Palenzuela, 2016; Loeb, 2016; de Mink & King, 2017) or environments have the potential to trigger short-timescale, high-energy events from more exotic progenitors. Some have claimed EM counterparts to BBH mergers within active galactic nucleus (AGN) (Graham et al., 2020, 2023) – with the merger kicking the resultant BH through the dense accretion disk of the AGN, creating a short-lived optical flare. Accounting for the intrinsic variability of AGN (which already show regular flares) however remains challenging

With the prospects of resolving long-standing uncertainties about the birth sites of r-process elements, and the progenitors of short GRBs, GW triggers were intensively

[1] Some groups (Creswell et al., 2017) disagreed with this initial detection.

followed up by the EM community – with the hope of finding kilonovae within the localisation region of GW-detected binary neutron star mergers . It would not be long before the community were granted their wish, with GW170817.

1.3.3 GW170817 and AT2017gfo

GW 170817/GRB 170817A/AT 2017gfo was perhaps the single most significant target in multi-messenger astronomy thus far, with over 1,637 peer-reviewed publications referencing the event at the time of writing. The rich, multi-facility, multi-wavelength dataset on this event is among the richest in observational astronomy, crucial to the interpretation and understanding we have gleaned from it. Although a fully comprehensive account is too large to fit within this thesis, I outline some of the key events, outcomes, and findings from this extraordinary cosmic explosion.

GRB 170817A (Goldstein et al., 2017; Abbott et al., 2017f) was the first reported trigger of this event, with a clear detection from the *Fermi* GBM instrument. 1.7s before this high-energy trigger, the LIGO Livingston, LIGO Hanford, and Virgo detectors observed a compact binary coalescence, given the name GW 170817 (Abbott et al., 2017c). Parameter estimation routines confirmed this as a merger of two neutron stars of masses 1.16 and 1.60 $M_{\odot}$respectively (Abbott et al., 2019a), with a joint sky localisation of 31 square degrees, and median luminosity distance of 40 Mpc. This is illustrated in Figure 1.6 The time delay of 1.7 seconds between the two triggers provided strong constraints on the speed of propagation of gravitational waves – predicted to be the speed of light, and empirically verified to 1 part in 10^{16}. Although this messenger arrived before, latencies in the detection pipeline meant it was only reported after the fact, when the significance of the association became clear. An extensive multi-messenger campaign ensued to search for potential optical counterparts in the joint localisation area between GW 170817 and GRB 170817A (Abbott et al., 2017e). 10.9 hours post-trigger, a bright transient (known as AT2017gfo, see Coulter et al. 2017b) in the nearby galaxy NGC 4993 (Hjorth et al., 2017) was discovered (Coulter et al., 2017a), with multiple independent discoveries (Soares-Santos et al., 2017; Arcavi et al., 2017; Lipunov et al., 2017) confirming the discovery. Later spectroscopic confirmation (Lyman et al.,

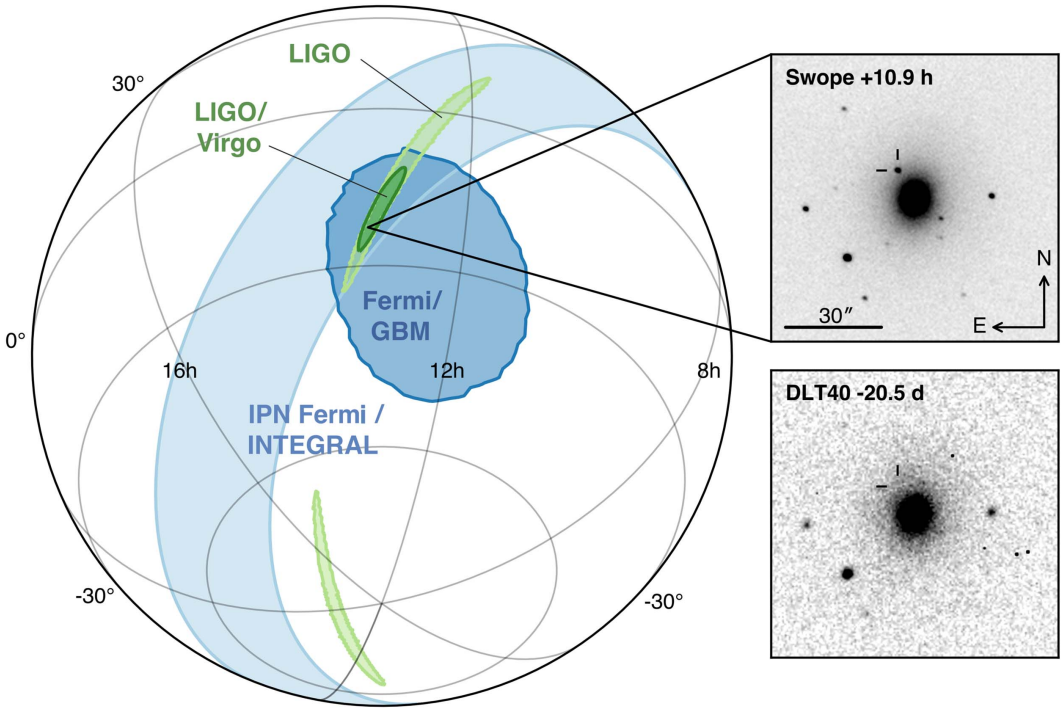

Figure 1.6: Graphical summary of GW170817, reproduced from Abbott et al. (2017e). The joint localisation between GW and EM triggers is illustrated

2017; Shappee et al., 2017) confirmed that this was a transient entirely unlike any supernovae seen prior – with a featureless spectrum with a clear deficit of flux in the bluer wavelengths, a cool black-body. Intensive follow-up was triggered on facilities around the world, and in space, providing among the richest datasets in astronomy.

From this singular event, many long-standing questions and theories have been verified empirically. The optical afterglow counterpart cemented the association between GW 170817 and GRB 170817A, providing observational confirmation of the theorised link between compact binary mergers and short gamma-ray bursts (e.g. Eichler et al. 1989; Narayan et al. 1992). Modelling the rapidly-reddening observed spectral energy distribution (Drout et al., 2017), paired with spectroscopic observations (Smartt et al., 2017), provided strong observational evidence for the synthesis of r-process elements. Based on the inferred yields of lanthanides from these observations, it is likely that binary neutron star mergers are the dominant r-process sites in the Universe (Tanvir et al., 2017).

Late-time observations continue to the present day across all bands, as the kilonova has faded, and the jet launched by the GRB collides with the ISM dominates the emission (Lyman et al., 2018). GRB 170817A is an off-axis GRB, and shows evidence for a structured jet (Alexander et al., 2018; Troja et al., 2019) with a simple power-law spectrum. Long-term X-ray emission (Troja et al., 2017; Margutti et al., 2017; Piro et al., 2019; Troja et al., 2020) continues from synchotron emission from the shocked ISM in front of the jet. This has proven brighter than expected at very late times (Hajela et al., 2022), likely due to fallback accretion (Rossi & Begelman, 2009; Metzger & Fernández, 2021; Ishizaki et al., 2021).

Despite everything we have learned from this event, and the validation of our theoretical predictions of kilonova observables, many uncertainties lie ahead: how typical was AT 2017gfo of the wider population of kilonovae? Out to what distances can we effectively conduct follow-up of typical kilonovae? We expect significant diversity in the light curves and properties of kilonovae (Gompertz et al., 2018) from theoretical models (Bulla, 2019), with the viewing angle, lanthanide fraction, and ejecta massses driving significant changes in the observational appearance. Paired with numerical relativity modelling to map between merger parameters and kilonova properties, this provides a potential route to inferring binary properties from high-quality light curve data (Coughlin et al., 2019; Nicholl et al., 2021), with potential synergistic constraints possible from a joint GW-EM detection. Further, the recent detection of a kilonova from a long gamma-ray burst (lGRB) (Rastinejad et al., 2022) suggests that some lGRBs are the result of compact binary mergers, further complicating the picture. This particular event showed (temporally) extended emission (Norris & Bonnell, 2006) in gamma rays, in the context of the compact binary merger powered by a high rate of fallback accretion onto the remnant. Although the LVK detectors were offline at the time of this event, at 350 Mpc it is unlikely to have been detected (e.g. Buikema et al. 2020), making the follow-up of high-energy triggers an important complement to pure gravitational-wave triggers for this type of object. Resolving the broad uncertainties remaining will require the discovery and follow-up of more optical counterparts to gravitational-wave events, a uniquely challenging task owing to the nature of gravitational-wave detectors.

1.4 The challenges of multi-messenger astrophysics

Despite the great promise of GW-EM multi-messenger astrophysics, significant challenges remain even now. The sky localisations of gravitational wave events can be large, especially in the case of detection in a single detector where the localisation is largely driven by individual detector response patterns, rather than time-of- arrival differences between detectors in the network. Whilst in the case of GW 170817 we were lucky to obtain a separate high-energy trigger, we have no reason to expect that every BNS merger observable by LVK will be accompanied with a gamma-ray burst. Off-axis, lower fluence bursts like that of GRB 170817A would simply not be visible at greater distances. The expanded detector horizon resulting from increased sensitivity also means more BNS events will be observed, and at typically greater distances (Petrov et al., 2022), further complicating follow-up efforts and reducing the prospects of joint high-energy triggers. Large localisations are especially challenging for surveys not dedicated to GW-EM follow-up, where interrupting the overall survey cadence and strategy for triggers is not possible/heavily constrained.

A further issue is the volumes of unrelated transients (termed 'contaminants') unearthed by multi-messenger searches (e.g. Smartt et al. 2016). Given the markedly higher volumetric rates of regular supernovae, stellar variability, and other phenomena, identifying intrinsically rare GW-related transients is challenging. Recent searches have unveiled families of transients with similar evolutionary timescales and observed properties to kilonovae (KNe) at early times (Agudo et al., 2022), further complicating the process of triaging candidates in the error region. Recent observational (Gompertz et al., 2018) and theoretical (Kawaguchi et al., 2020; Barnes et al., 2021; Bulla, 2023) efforts further reveal significant diversity in the photometric and spectroscopic signatures of kilonovae themselves – resulting from varying systemic parameters. Although such diversity in observational properties may make initial identification more challenging, it also directly encodes information about the kilonova that may be retrievable with accurate models (Bulla, 2019; Nicholl et al., 2021).

Putting aside the challenges of multi-messenger follow-up efforts: upcoming

observing runs will require not only wide fields of view to cover the vast localisations, but the depth to push to more distant targets – all while operating at high enough cadence to identify counterparts before they fade to luminosities incompatible with intensive follow-up (e.g. Chase et al. 2022). Such a balance is ideally met by purpose-built all-sky surveys that are dedicated predominantly to GW-EM follow-up.

1.5 Eyes on the skies: the modern landscape of large scale sky surveys

With the wide-scale adoption of the charge-coupled device (CCD), and rapidly growing computational power (Moore, 1965) available to astronomers, the process of surveying the sky could now be digitised. CCDs improved significantly over the photographic plates used previously, with higher sensitivity, a more linear response to light, and the ability to perform processing on the raw pixel data. This last point is particularly important: with human eyes on photographic plates being the limiting factor previously, image processing and automated routines could reduce this burden and enable surveys to extend to greater fields of regard.

The first digital transient sky surveys such as the Lick Observatory Supernova Search (LOSS; Li et al. 2000) and ESSENCE (Miknaitis et al., 2007) discovered as many supernovae as Zwicky had in 40 years in just a few years, marking a step change in capability. Surveys such as OGLE (Udalski et al., 2015) also began large-scale time-domain surveys of the Magellanic Clouds and Galactic Plane, delivering vast catalogs of stellar variability and microlensing events. Difference imaging (see Section 2.4 for a more in-depth mathematical discussion) proved transformative in facilitating searches at larger scale, by making automated measurements of candidate transient sources possible. Through matching point spread functions (PSFs) and background fluxes between a 'science' image and a suitable 'template' or reference image of the same patch of sky, the flux of non-varying sources can be subtracted away, leaving (ideally) residual PSF-like sources on a clean background that reveal sources that have changed in brightness. Traditional source detection algorithms and photometry routines (e.g. Bertin

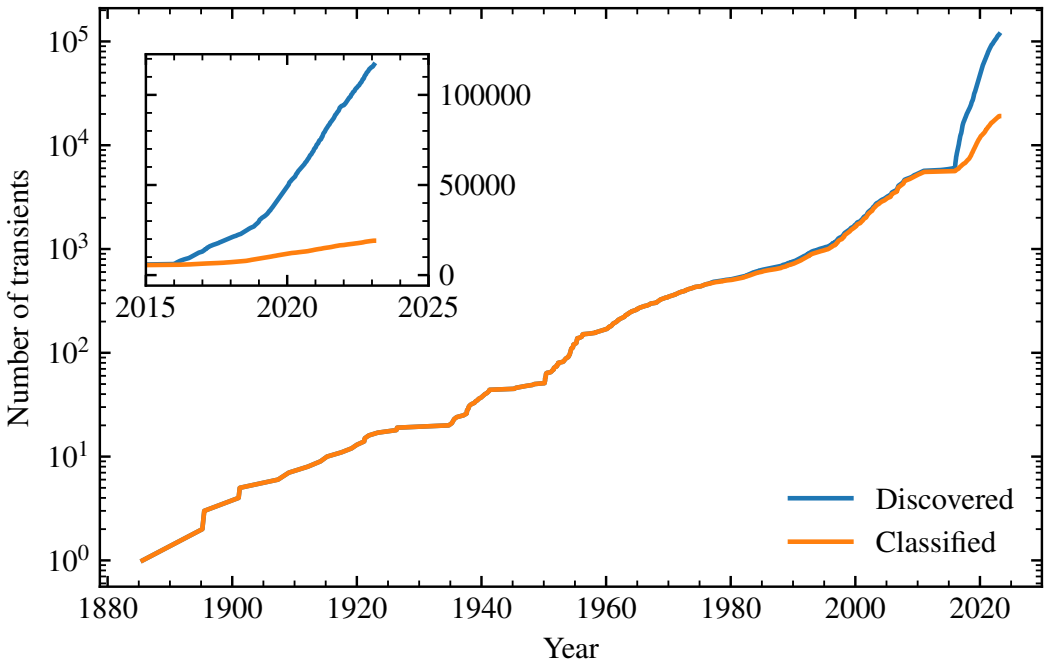

Figure 1.7: Supernova discoveries and classifications plotted with time, as reported to the Transient Name Server. The inset panel shows the last $\sim$ 10 years on a linear scale, to highlight the rapidly-widening gap between transient discoveries and successful classifications

& Arnouts 1996) can then be applied to this 'difference' image to identify candidate transients and make measurements of their brightness. Whilst initially limited in capability (e.g. Alard & Lupton 1999), this technique was rapidly scaled up to wide field of view instruments, enabling deep searches for transients over 10s of square degrees. This technique now underpins the vast majority of modern transient surveys. These surveys are now unveiling significant yields of astrophysical transients, and observing at the necessary cadence to begin to discover transients that evolve on timescales of $\sim$ days. The evolution of technologies and techniques used in wide-field transient surveys have driven an explosion in the number of transients discovered over the past 10 years. Figure 1.7 illustrates the progress of the field over the past century, showing exponential growth over time. The inset shows the extraordinary growth in the numbers of discovered transients over the past 10 years, but also highlights one of the central issues of the field – the ever-widening gap between the number of transient discoveries, and

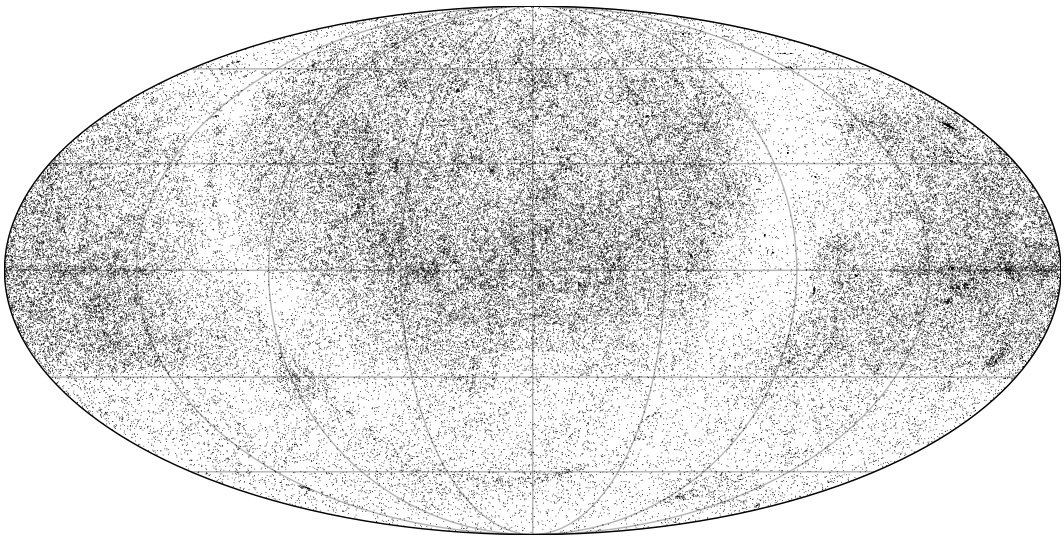

Figure 1.8: Mollweide projection of all extragalactic transients discovered up to the time of writing. There is a notable deficit of transients within the galactic plane, owing to the strong dust extinction from our own Milky Way, and less transients in the Southern Hemisphere owing to fewer transient surveys operating below declinations of -30°. Some 'clumps' are also visible, corresponding to 'deep' sky surveys

the number of transient classifications. This shows no sign of abating, as spectroscopic classifications require both larger telescopes and longer integrations to achieve suitable signal to noise than simple imaging of transients.

The on-sky distribution of extragalactic transient discoveries is illustrated in Figure 1.8, showing that we are finding objects all across the sky. There is a notable deficit of discoveries in the Galactic plane, where foreground extinction limits the effective depth of surveys, as well as in the Southern sky, where fewer discovery facilities exist.

Two major paradigms for large-scale optical sky surveys have emerged (here named 'monolithic' and 'modular'), driven by strategic tradeoffs between depth, field of view (grasp), and cadence. 'Monolithic' surveys tend to be wide-field instruments installed on a single, large telescope. The large aperture allows deep imaging in comparatively short exposure times (e.g. ZTF's 1.2m aperture enables imaging down to $g' = 21.5$ in 30 second exposures). Some key examples of this type of survey include ZTF (Bellm et al., 2019), PanSTARRS (Chambers et al., 2016a), and SkyMapper (Keller et al., 2007). The upcoming Legacy Survey of Space and Time (LSST; Ivezić et al.

2019) delivered by the Vera C. Rubin Observatory also falls into this category, although on a thus unprecedented scale. In contrast, 'modular' surveys increase sky coverage by having multiple telescopes (potentially at multiple sites). Some examples of these survey designs include ASAS-SN (Shappee et al., 2014), ATLAS (Tonry et al., 2018), BlackGEM (Bloemen et al., 2016), CRTS (Drake et al., 2012), and GOTO (see Section 1.6, Steeghs et al. 2021) Although limited to somewhat shallower depths than 'monolithic' surveys, the possibility of operating in parallel significantly increases the instantaneous field of view, and the additional scheduling flexibility and resilience to potential bad weather provide significant advantages. Collectively, these surveys are responsible for driving the significant increase in the discovery of extragalactic transients over the past 10 years (see Figure 1.7).

In contrast to the numerous wide-field efforts, a number of surveys have been launched focusing on surveying smaller areas of sky, at higher cadences and depths using large-aperture telescopes such as PanSTARRS and the Victor Blanco 4m (home to the DECAM instrument). Some prominent programs include the Young Supernova Experiment (YSE; Jones et al. 2021), Deeper Wider Faster (DWF; Andreoni & Cooke 2019), DECAMERON, and DESIRT (Palmese et al., 2022). More limited surveys have also been performed with space-based observatories with similar goals, such as the HST-based See-Change program (Hayden et al., 2021) and the Spitzer SPIRITS survey (Kasliwal et al., 2017). These deeper, more focused surveys have played a pivotal role in unveiling populations of fast transients at high redshifts (Pursiainen et al., 2018), providing robust and well-characterised samples of SNe Ia at cosmological distances (Scolnic et al., 2018; Brout et al., 2019; Hayden et al., 2021), and probing the minute-timescale variability of the extragalactic sky (Andreoni & Cooke, 2019).

Whilst the majority of transient discoveries have been driven by a limited group of professional sky surveys, it is important to acknowledge the efforts of hard-working amateur astronomers in discovering many transients, and in some cases providing classifications and early-time data that have proven crucial to informing professional studies. Data-mining efforts are also unveiling populations of historical supernovae

from historical data that were previously missed (e.g. PALEO[2]), and discoveries and follow-up (Bersten et al., 2018) of bright transients being performed by a small but dedicated group of amateur observers. Indeed, during the writing of this thesis the nearest (5.4 Mpc) core-collapse supernova in a decade, SN 2023ifx, was discovered by Koichi Itagaki, a prolific amateur discoverer of nearby supernovae (Itagaki, 2023).

We leave more in-depth discussions of future prospects and upcoming surveys in time-domain astrophysics to Chapter 7, but remark here that the discovery capabilities of time-domain astrophysics are evolving rapidly, populated with a diverse set of discovery instruments with complementary capabilities. The demands of gravitational-wave follow-up are markedly different to the requirements for other transient survey programs, requiring a tailor-made approach in both hardware (to survey rapidly) and survey strategy (requiring rapid response to incoming triggers). These needs are ideally met by purpose-built sky surveys, which I discuss in the following Section.

1.6 The Gravitational-wave Optical Transient Observer (GOTO)

Built specifically with the follow-up of poorly-localised optical counterparts (see White 2014), the Gravitational-wave Optical Transient Observer (GOTO) is a wide-field, modular optical sky survey with a flexible design. Beyond the content of this thesis, a full description of the prototype phase is given in Steeghs et al. (2021), with some key scientific 'highlights' from the Laser Interferometer Gravitational Observatory (LIGO) O3a observing run published in Gompertz et al. (2020). Below, I summarise some of the key ideas, methods, and features of the design-specification GOTO systems as they are at the time of writing, to provide broader context to Chapters 3 and 6.

1.6.1 Hardware

The GOTO project has two antipodal installations, one at Roque de los Muchachos Observatory, La Palma, and the other at Siding Spring Observatory, Australia. Each 'node' has two telescopes, each comprised of 8 co-mounted $40\,\mathrm{cm}$ f/2 astrographs, having a

[2] http://scan.sai.msu.ru/~denis/Paleo/paleo-sky-map.html

45 square degree field of view each. Each astrograph is paired with a 50 megapixel CCD, sampling the sky at 1.4 arcsecond/pixel resolution. Typical survey images with the final design-specification hardware reach a 5-sigma limiting magnitude of$L = 20.3$ in a single 4x60s pointing, with typical survey speeds of ~ 600 square degrees per hour, per mount, per site. This capacity is ideally suited to rapidly tiling the ~ 1000s square degree localisations of typical GW events, uniquely also positioned to survey some of the more poorly localised events, and among the transient surveys with the best prospects of detecting a kilonova (Chase et al., 2022). Figure 1.9 depicts the full design specification GOTO hardware, across both hemispheres. GOTO is fully instrumented for autonomous operations (see Dyer 2020), with observing orchestrated by a distributed, just-in-time scheduler that allocates pointings to each mount independently, and performed by a fault-tolerant control system for each mount (Dyer et al., 2018). This is designed from the ground up to be responsive to time-sensitive alerts in low-latency. The 'sentinel' listens to the NASA Gamma-ray Coordinates Network (GCN; Barthelmy et al. 1998) for incoming alerts from a variety of trigger sources, including gravitational-wave, gamma-ray, and neutrino facilities. Upon receipt of an alert, the skymap is passed to an in-house processing utility (`goto-tile`) which generates an optimal tiling strategy for the GOTO network, and submits pointings to the scheduler for observation. GOTO responds in real time to incoming events, being primarily limited by sky visibility rather than response time itself (e.g. the events with 5 minutes response time in Gompertz et al. 2020). Even while in the prototype phase (2018–2021), based on targets of opportunity, GOTO contributed to various studies on variable stars (Duffy et al., 2021), transient follow-up (Gompertz et al., 2020; Ackley et al., 2020; Mong et al., 2021), and minor planets (Borisov et al., 2018).

1.6.2 Software

Real-time response to gravitational-wave events requires real-time data reduction, classification, and candidate vetting. The main GOTO science pipeline (`kadmilos`; Lyman et al., in prep.) performs calibration and reduction of the raw data gathered by GOTO in real time, with $\sim$10 minutes from close of shutter to transient candidates. The pipeline

Figure 1.9: The GOTO network, as of April 2023. Image credit: M. Dyer / GOTO Collaboration

automatically creates optimal stacks of biases, flats, and darks from nightly calibration observations, calibrates incoming raw data with these, performs astrometry and photometry, and combines multiple (typically 4) dithered observations from a single visit to improve depth and image quality. Difference imaging (see Section 2.4) is performed on incoming stacks of science images using the HOTPANTS (Becker, 2015) algorithm to remove sources of static flux, with source extraction being applied to identify transient candidates. Templates are provided as part of a dedicated reference survey, with an extra $\sim$magnitude of depth over regular all-sky imaging. Difference imaging is a notably 'noisy' process however, generating significant volumes of false-positive candidates that must be sifted through to identify genuine astrophysical sources.

GOTO makes extensive use of machine learning source classification techniques to filter the significant incoming data volumes to levels where humans are able to keep pace and follow up promising candidates. I reserve further discussion of this to a dedicated Chapter (Chapter 3), although here remark that application of these techniques can reduce the number of candidates by two orders of magnitude. Candidates that pass this initial machine-learning filter are propagated to the GOTO Marshall (Lyman et al., in prep.), a web-based platform for humans to validate candidates, assess their significance in the context of external GW triggers, and trigger further follow-up as required.

A key part of this vetting process is considering contextual information: pulling in data from existing catalogs and images to provide a richer picture of the surroundings of a candidate. For example, historical data may reveal variability at the explosion site from many years prior, suggesting a variable star or AGN is responsible. Looking at nearby (in sky separation) galaxies can give hints as to the host of a given candidate, with associated distance estimates giving bounds on the absolute magnitude/luminosity. For a multi- messenger trigger, this can rule out potential transients by comparing to the (e.g.) GW-estimated distance. The concept of contextual classification, and automating this process to minimise human effort, is discussed further in Chapter 6.

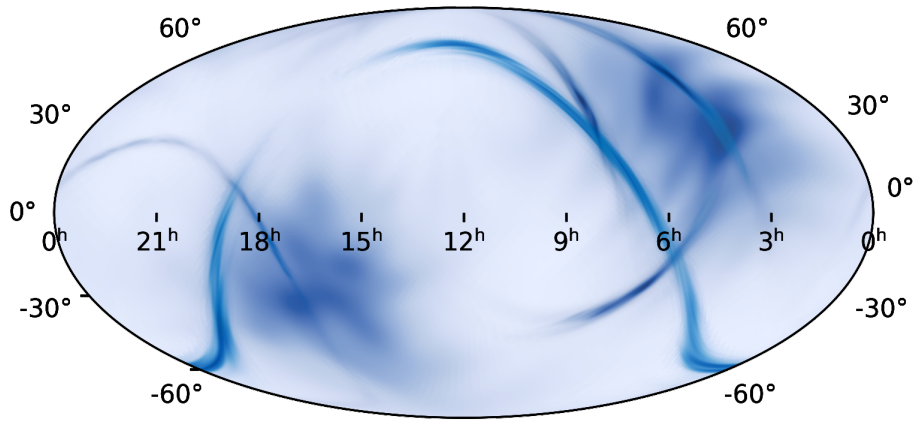

Figure 1.10: The first GW events of the LVK O4 observing run, from the first month of observing. There were 6 BBH mergers, and one 1 black-hole neutron star/BNS merger in total. Figure based on all publicly-available data from GraceDB, and plotted with the `ligo.skymap` Python package.

1.6.3 The first month of O4

Owing to logistical delays, the LIGO-Virgo-KAGRA O4 observing run began during the writing period for this thesis, on 24th May 2023, and is anticipated to continue for 18 months. I here discuss some of the early outcomes from this pivotal phase of GW-EM astronomy. 1 month into this period, a total of 6 statistically significant events in 3 weeks have occurred, a large step change in the number of events being detected. Alongside this, 38 low significance events were also reported. The high-significance events are plotted in Figure 1.10. The participation of the Virgo detector in the run has been delayed until later in the year, owing to significant technical issues that require the detector to be brought out of deep vacuum for investigation. This means that only two detectors (excluding KAGRA, whose sensitivity is limited) are actively observing the extragalactic sky, meaning that most localisations so far have been poor in nature – displaying the characteristic arcs associated with two- detector coincidences. This

is a vastly different landscape to what was expected, but one that is ideally suited for wide-field surveys to explore and make significant contributions to.

1.7 Thesis Outline

With the advent of gravitational-wave multi-messenger astrophysics, and the optical sky survey capability to follow up the challenging localisations of compact binary mergers, we are uniquely placed to be able to resolve some of the key scientific questions of multi-messenger astrophysics – the birth sites of heaviest elements produced in the Universe, the progenitors of the most luminous explosions, and the underlying (and emergent) properties of compact binary coalescences.

Significant challenges remain however – both in terms of the data volumes generated as part of these searches, and the numbers of unrelated transients we discover routinely. Triaging this data, and identifying promising counterparts with minimal human effort is a crucial step to being able to perform the science we want to do – and the current methods are not well-optimised for the new scales we are looking at. This thesis explores both the observational and computational aspects of time-domain astrophysics, and the cross- cutting synergies between them, aiming to build new techniques that work well at the scales involved, and drive the study of transients in new directions to gain novel insights that would not otherwise be possible.

Chapter 2 introduces the overarching techniques that power modern scientific discovery in transient astrophysics, and underpins the latter chapters. Chapter 3 discusses the development of the real-bogus classifier employed in the live GOTO pipeline, `gotorb`. Chapter 4 presents the most precise ephemeris yet for the candidate continuous-wave source, Scorpius X-1 – delivering the best constraints yet on the orbital motion of the donor star, and facilitating high-precision searches for the upcoming (at time of writing) LVK 'O4' observing run. Chapter 5 presents a novel time-domain study of bright, nearby supernovae at high cadence, searching for signatures of inhomogeneities in the circumstellar material surrounding supernovae. This study uses SALT and LT data to place some of the strongest constraints yet on such variability, and pre-

empts on-going studies led by myself using NTT/ULTRACAM. Chapter 6 presents my recent work on developing contextual classification for GOTO and other time-domain surveys, making use of extensive literature catalogs to provide high-accuracy predictions on potential hosts for transients, as well as delivering rich contextual information to aid human classification. In the concluding Chapter 7, I give an overall summary of the thesis contents, and outline the bright future of time-domain astrophysics with the advent of upcoming large-scale sky surveys, whilst also noting some critical problems that the community must overcome to make the most of the opportunities afforded by an order-of- magnitude increase in data volumes. This thesis is composed of original published work from the PhD.

Chapter 2

Methodologies

As the data volumes generated by astrophysics have grown, so too have the methods that we use to process and analyse our gathered data. This chapter is intended to provide an overview of some of the key methods employed in the following chapters to provide some context for their use.

2.1 Machine learning in astronomy

Machine learning (ML) is a family of statistical techniques that employ algorithms that 'learn' to predict quantities from data without explicit instruction, or assuming some underlying model. In particular, the more data that is used to train a ML model, the better the performance on a given task. These powerful, non-parametric methods have been at the heart of many modern breakthroughs, attaining human-level performance on computer vision and speech recognition tasks, and are now ubiquitous in modern society (for better or worse).

ML workflows are now beginning to receive mainstream acceptance in the scientific community as an essential and necessary tool in processing the vast datasets in play in modern scientific experiments. ML is even being used in the optimal design, execution, and automation of experiments, with entire lab setups being orchestrated by 'robot scientists' (King et al., 2009) in search of new phenomena. Astrophysics is no exception to this. In contrast to computer science, where datasets are complex,

messy, and inhomogeneous (e.g. Gao et al. 2020; Schuhmann et al. 2022), astronomical datasets are clearly defined in terms of modality, selection criteria, and quality. Problems in astrophysics are more likely to be data quality-limited than architecture limited: with classification performance being set by label noise, instead of the capacity of the model architecture. Astronomical datasets do not carry many of the privacy/ethical issues associated with the large-scale ensemble datasets deployed in training current state-of- the-art image, text, and audio-based algorithms, making them a more benign playground for the development of new algorithms and methodologies. To this end, the astronomical commmunity has made extensive use of, and embraced, machine learning approaches since the early 1990s (e.g. Sandler et al. 1991; Odewahn et al. 1992). Providing a full summary of all uses of machine learning in astronomy is beyond the scope of this thesis, and would almost certainly be out of date by time of publication given the extreme pace the field is advancing at – spurred on by the latest techniques from the computer science literature.

Machine learning lies at the heart of parameter estimation in modern astronomy, delivering photometric redshifts (e.g. Collister & Lahav 2004; Carrasco Kind & Brunner 2013; Beck et al. 2016; Duncan 2022), star-galaxy separation scores (e.g. Fadely et al. 2012; Soumagnac et al. 2015; Miller et al. 2017; Beck et al. 2021), morphology (e.g. Dieleman et al. 2015; Cheng et al. 2020; Walmsley et al. 2020), source properties (e.g. Ness et al. 2015; Huppenkothen et al. 2017), and more across all sub-fields of astrophysics. The huge scale of modern astronomical catalogs ($\sim$ billions of sources) is ideally suited to exploration with ML, driving data-driven discovery with clustering algorithms (e.g. Rubin & Gal-Yam 2016; Chattopadhyay & Maitra 2017; Hunt & Reffert 2023), anomaly detection (e.g. Storey-Fisher et al. 2020; Ishida et al. 2021; Lochner & Bassett 2021; Malanchev et al. 2021), and other unsupervised learning techniques (e.g. George et al. 2018; Hocking et al. 2018; Shih et al. 2022). ML has found an important role in reducing the computational time of parameter inference with machine-learned surrogates for computationally-expensive processes (e.g. Himes et al. 2022; Mould et al. 2022; Spurio Mancini et al. 2022) and amortising posterior sampling (e.g. Dax et al. 2021; Zhang et al. 2021; Vasist et al. 2023) Even theoretical astrophysics

has not been spared the influence of machine learning, with neural networks inferring conservation laws (Cranmer et al., 2020; Liu & Tegmark, 2021), predicting planetary dynamics (Cranmer et al., 2021), and accelerating N- body simulations (Emsenhuber et al., 2020). In the following section, I will provide an overview of the central concepts of machine learning, to provide context for my use of it in the upcoming chapters, and tie these back to the astronomical literature.

2.2 Overarching concepts

All machine learning, whether supervised or unsupervised, online or offline, has three core steps. These are:

- Evaluating a model on some target datapoints to generate some 'predictions'
- Computing a metric of choice based on the target datapoints and the model outputs
- Adjusting model parameters to optimise the metric of choice

These three steps apply from the simplest linear models, up to the largest language models in existence – with the complexity arising primarily from our choice of model and the nature of our metric. The steps above are written in a deliberately abstract manner to encapsulate the breadth of possibilities for each of these steps.

2.2.1 Machine learning models

For all the complexities, machine learning models essentially have one goal – predicting the values of an arbitrary function. More rigorously, the models in machine learning must map potentially high-dimensional input spaces to potentially high-dimensional output spaces. Such spaces are not well-represented by the (often) parametric models we deal with in astrophysics, and so machine learning works in a different domain of complex, high-parameter count models. Although a full summary of *all* machine learning models is beyond the scope of this section, I focus in on two of the most popular classes

of models – decision trees and neural networks, as the two that find most usage in the Chapters that follow.

Tree-based models were some of the first 'machine learning' models – emerging from the usage of decision trees to model human decision making. Starting at the top, each internal node branches based on a simple condition on a given feature. The tree is traversed, evaluating the condition at each internal node, until a leaf node is reached and a classification is made. Optimally building the tree structure has been the subject of extensive (Quinlan, 1986, 1992; Breiman et al., 1984) work, but each seeks to build the tree by choosing the attribute that best splits the dataset (in terms of entropy/other quantities) at each candidate node. This is naturally a greedy approach, but given the vast potential state space this is necessary. Unlike many other ML models, decision trees are inherently interpretable, as one can simply traverse the tree, evaluating the single criterion at each branch until a leaf node (classification) is reached.

With rapidly growing computational power, ensemble learning became possible – providing predictions based on the aggregated output of a number of smaller, weaker learners. This aggregation may be as simple as a (weighted) vote between all members of the ensemble, or as complex as 'stacking' classifiers together by feeding the outputs of the weak learners into another classifier. Two main approaches to improving the performance of simple decision trees emerged from this: Random forests (Breiman, 2001) make use of 'bagging', or bootstrapping – by fitting individual decision trees to randomised subsets of both the training dataset and features within the training dataset. This simultaneously improves predictive performance, and improves generalisation to unseen data. Boosted decision trees (Schapire, 1990; Freund & Schapire, 1995) iteratively learn a dataset by fitting successive decision trees – with the next tree trained on the residuals of the previous to 'correct' the predictive errors. This 'boosting' rapidly improves the performance of the overall classifier, minimising predictive bias and variance, at the cost of being prone to overfitting. RF and BDT algorithms are commonplace in ML, especially in commercial/enterprise settings as they require very little optimisation of hyperparameters (see Section 2.2.5). Tree-based classification also finds broad utility in astronomy, as it works well on tabular data where features have been pre-extracted –

features here being numerical data like fluxes, line widths, distances, and other parameters inferred by some model/routine, as well as categorical data (which other, more complex techniques cannot process natively).

In parallel, inspired by the biological structures present in the brain, artificial neural networks (ANNs; McCulloch & Pitts 1943) emerged, with the Rosenblatt perceptron (Rosenblatt, 1958) being among the first physical realisations [1] of this new architecture. The building blocks of ANNs are neurons, simplistic predictors that take some input, multiply by a tunable weight, and add a bias value. As described above, the neuron is essentially linear regression, a statistical methodology used for hundreds of years. What transforms a linear predictor into a neuron is the addition of non-linearity via an 'activation function' – a function that rectifies the output of the neuron by mapping it to some (generally) constrained space. The additional expressiveness provided by this non-linear transformation makes single-layer perceptrons powerful general models. The (comparatively) simple architecture, combined with analytic derivatives, makes ANNs well-suited to optimisation by stochastic gradient descent. Some families of models (e.g. neural networks with non-linear activations) are provably 'universal function approximators (Cybenko, 1989; Hornik et al., 1989): that is, they can reliably approximate **any** function to an arbitrary degree of accuracy, given sufficient numbers of neurons (width). Any is emphasised here, as this 'any' is a far looser condition than approximations with e.g. Taylor series (requiring the function to be infinitely differentiable to attain arbitrary precision) or Fourier series (requirement to be piecewise smooth). This is also somewhat of a curse, as the weak constraints on the 'manifold' geometries make it trivial to fit directly to noise in the data ('overfitting', see Section 2.2.4).

2.2.2 Optimisers

For a a smooth, differentiable function $F(\theta)$, we can find the extremal value by iteratively taking steps in the direction of steepest gradient – the method of gradient descent. Mathematically, given a smooth , at least once differentiable function, $F(\theta)$, and a set of initial starting parameters θ_0, the values θ that minimise $F(\theta)$ can be found with

[1] Implemented entirely in electronic hardware, rather than as a computer program

iterative application of the update rule

$$\theta_{i+1} = \theta_i - \eta \nabla F(\theta_i)$$

, where η is the step size or 'learning rate'. The value η must be chosen carefully: too small, and many steps will be required to converge to a local minimum, too large, and the algorithm will be numerically unstable, overshooting minima and failing to converge. It is also possible to set a per-parameter learning rate η_i to mitigate non-uniform curvature in each model parameter, although it is hard to know *a priori* which learning rates are appropriate.

For general-purpose optimisation, gradient descent is largely replaced in practice with approaches with provably better convergence, that adaptively tune η based on information about the curvature of the minimisation space. The ubiquitous Levenberg-Marquadt (Levenberg, 1944) method is a key example of this, using the Jacobian matrix $\mathbf{J}$ of residuals to compute the optimal step size that guarantees stable, optimal convergence for well-behaved (twice- differentiable) functions. Gradient-free optimisers also exist, trading the requirement to compute gradients for more function evaluations – the Nelder-Mead method (Nelder & Mead, 1965) being among the most popular. These methods struggle to perform well in high-dimensional parameter spaces however, owing to the exponentially-increasing parameter space they must search. Gradient information allows a 'directed' optimisation, leading to more rapid convergence. The requirement of gradient/Jacobian evaluations makes the application of these methods challenging in large- model, large-data scenarios however, scaling $O(N^3)$ with N parameters in the worst case (Gauss-Newton). In these scenarios, valuating $F(\theta)$ is prohibitively expensive, and even with modifications (e.g. Liu & Nocedal 1989) too costly for large models on large datasets. Approximate methods must be applied to subsets of the dataset. Stochastic gradient descent (Robbins & Monro, 1951) uses 'batches' (subsamples) of the dataset to compute an approximation to the true gradient. This estimate will be inherently noisy owing to the sampling error (dictated by the size of batch), but much quicker to evaluate in comparison to the full gradient – effectively

trading convergence rate for speed of evaluation.

SGD and derivative algorithms have a distinct advantage over ‘standard‘ gradient descent in high-dimensional optimisation problems: resilliency against convergence to local minima. This is particularly pronounced when training multi-million parameter neural networks (see Section 2.2.4), with complex and non-convex loss landscapes. The issue of step size is more pronounced here, as adaptive step size methods must infer the step size based on noisy gradient estimates, which lead to inherently unstable step size estimates. The complex, high-order loss landscapes may also require learning rate to be tuned per parameter for efficient optimisation. A whole family of techniques have emerged using moving averages of the gradient estimate to address this, among the simplest being RMSprop (Tieleman et al., 2012). This method maintains a running average of the squared gradients, with a chosen decay factor $\beta \ll 1$ dictating the relative importance of historic versus current squared gradients

$$E\left[g^2\right] = (1-\beta)g_i^2 + \beta\, g_{i-1}^2$$

where g_i is the gradient vector at the i^{th} step. The learning rate is then scaled through by this running average, giving the update rule,

$$\theta_{i+1} = \theta_i - \frac{\eta \nabla F(\theta_i)}{\sqrt{E\left[g^2\right] + \varepsilon}}$$

with the ε term ensuring that the denominator does not vanish, causing large steps to be taken spuriously. Algorithms such as these are at the heart of training the largest machine learning models in existence, with even marginal improvements in the convergence rate of algorithms translating to potential time savings of days in these peta-scale scenarios (Coleman et al., 2018). With efficient approaches to optimising potentially complex, high- dimensional models in hand, the final remaining element of machine learning is specifiying the function we’re trying to extremise.

2.2.3 Loss functions

What is $F(\theta)$ in the context of machine learning – how to we optimise our models to maximise their predictive power, given a training dataset to learn from? We seek a function that measures the degree of disagreement between our observed data and the predictions of our model – specifically so that we can minimise this and in turn maximise the fidelity of the model. This is known as a loss (or cost) function, often denoted $\mathscr{L}(\theta)$, which should decrease monotonically in the quality of fit. From a theoretical perspective, one can consider loss functions as distance metrics between the predictions of the model and the dataset observed, and minimisation of this distance is equivalent to minimising the predictive error. The chi-squared χ^2 goodness of fit used throughout the physical sciences is a valid loss function - albeit one with some convenient statistically-motivated properties for parameter inference.

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

This almost directly maps to the common L2 mean-squared error loss function for regression problems in machine learning - only with the residuals unscaled by the error, and a prefactor of $1/N$ for numeric stability [2].

$$\mathscr{L} = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

where all symbols retain their original meaning. We can modify the loss function to change what we want to prioritise in the model. For example, the L2 loss is sensitive to the presence of outliers in the data, with the quadratic dependence on the residuals leading to them dominating the overall loss (and thus gradients) disproportionately. A simple modification is to take the absolute values of the residuals, rather than squaring them, known as L1 loss – this change minimises the dominance of outliers, whilst still having a well-behaved derivative (everywhere apart from zero). The Huber loss (Huber, 1964) is a smooth interpolant between these two cases, and is widely used in robust

[2] Machine learning operations are often done at reduced numerical precision for speed and compatibility with hardware accelerators

regression contexts.

For classification, there are more optimal choices than simple power-residual forms – given that valid probabilities are bounded between zero and one, and labels are binary (at least per class). The categorical cross-entropy is given by

$$CE = -\frac{1}{e}\sum_{c,e} y_c \log p_c$$

where y_i is the class label for a given example, p_i the model's predicted probability, and e enumerates chosen examples in the batch – with the mean over examples being taken. This has the benefit of being positive-definite (as $\log p_c$ is always less than zero), and has convenient numerical properties when paired with the common sigmoid activation function. One issue with this loss function (indeed all) is its' unsuitability for imbalanced datasets – that is where the numbers of classes are not comparable. Weights can be introduced into both classification and regression losses to address this.

To tie together all the overarching concepts in the previous subsections, let us mathematically construct the simplest possible neural network for classification – with multiple inputs to a single layer of neurons, and an identity activation function. This construction is equivalent to (multi-)linear regression: with the weight and bias terms mapping directly to the 'slope' and 'intercept' of the underlying linear model. The single neuron model can be written (using the Einstein summation convention):

$$y_i = w_j x_{ij} + b$$

where i indexes each item of data, and j indexes the input dimensions of the data, w_j is the weight vector, and b is the bias term. In order to train this model, we need to choose a loss function – for convenience of mathematics I opt for the mean- squared error ('L2' loss).

$$\mathscr{L} = \frac{1}{N}\sum_i (y_i - f(x_i))^2$$

Using the chain rule, we can compute the gradient of this loss function with respect to

$f(x_i)$ as

$$\frac{\partial \mathscr{L}}{\partial f(x_i)} = -\frac{2}{N}\sum_i (y_i - f(x_i))$$

$$f(x_i) = w_j x_{ij} + b$$

.

$$f(x_i)$$

is kept as an arbitrary function here in what follows, to emphasise the generality of this procedure for more complex models. Given this derivative, we need only apply the chain rule again to obtain derivatives of the loss with respect to our parameters w_j and b. Dropping the x_i from f for clarity:

$$\frac{\partial \mathscr{L}}{\partial w_j} = \frac{\partial \mathscr{L}}{\partial f}\frac{\partial f}{\partial w_j}, \frac{\partial \mathscr{L}}{\partial b} = \frac{\partial \mathscr{L}}{\partial f}\frac{\partial f}{\partial b}$$

Therefore, computing the relevant derivatives (assuming the simple linear model for convenience), we obtain

$$\frac{\partial \mathscr{L}}{\partial w_j} = -\frac{2}{N}\sum_i (y_i - w_j x_{ij} + b)\, x_j$$

$$\frac{\partial \mathscr{L}}{\partial b} = -\frac{2}{N}\sum_i (y_i - w_j x_{ij} + b)$$

as the gradients of the loss function with respect to our weights and biases. Extension to non-linear activations is straight-forward, but more involved, requiring another application of the chain rule to account for the derivative of the activation function. The true power of neurons comes when we knit them together into neural networks – where outputs of neurons are passed as the inputs to a new set of neurons. This powerful approach underpins all of modern machine learning, and I discuss it in more depth in the next chapter, along with the rich toolkit of methods that have arisen to tame the complexity of these models.

2.2.4 Deep learning

Deep learning (DL; LeCun et al. 2015) is a subset of machine learning that seeks to represent increasingly complex datasets through the use of deep neural network architectures. Through stacking multiple layers of learnable blocks (either of dense neurons, or arbitrarily more complex operations), DL algorithms can extract higher-order features composed of the outputs of previous layers, and represent more complex decision boundaries than otherwise possible. Deep learning approaches generally lack the requirement for human-selected features, instead learning an optimal feature set directly from the data itself. As the network is trained by selecting parameters that minimise the loss, the network will move to identify the most salient aspects of the data that enable the best classification. It is this behaviour that is at the heart of deep learning's power. Whilst at first glance computing the derivatives of deep neural networks to be able to optimise them with SGD may seem complex, the only non-linear components are the activation functions, which are typically chosen to have well-behaved, analytic derivatives. The derivative of the loss function with respect to the model parameters can therefore be computed with simple application of the chain rule, with the derivatives with respect to successive layers using the results of the previous layer. This process is known as 'backpropagation', and directly facilitates the training of large ($\sim 100,000+$ parameter) neural networks. Modern automatic differentiation frameworks (e.g. The Theano Development Team et al. 2016) are trivially able to model this with a functional 'graph' representation, with derivatives computed by traversing the graph and successively applying chain rule to the function.

As the data modalities processed by ML have evolved, so too have the techniques to process them. We can exploit the unique properties of different types of data to choose optimal extractors, and minimise the computational cost involved. Perhaps the most salient example of this is images. The information contained in images is highly correlated: the information encoded is shared across multiple pixels, often spatially co-incident or proximal. We as humans know this to be the case – lines, shadows, patterns, and symmetries are implicit in our understanding of the world, but these are not natural features of a pixel-by-pixel understanding of images. Convolutional neural

networks (CNNs; LeCun et al. 1995; Krizhevsky et al. 2017) evolved out of a need to efficiently process images – with naive fully-connected neural networks the number of neurons required to process an image scale quadratically with the side length of the image. The principal building block of a CNN is the convolutional filter, in which a learned kernel is convolved with the input image(s) to obtain feature maps – outputs of typically reduced size compared to the input images, encoding concise local information. The learned kernel is shared between all spatial positions, dramatically reducing the model complexity and number of parameters. Discrete convolution has a well-defined derivative (discrete correlation), which can be trivially computed by re-indexing the kernel, making this a very efficient computational primitive on which to build networks when paired with chain rule for backpropagation. The convolution operation itself is one of the most optimised in ML (e.g. Chetlur et al. 2014). The true power of CNNs emerges when feature maps are fed as inputs into other convolutional filters, allowing the composition of higher-order filters with stronger representative power.

As noted above, one requirement to attain maximal performance deep learning-based workflows is a large quantity of labelled training data. With too little data, a model with large representative power will tend to 'overfit' – memorising the input training data, and interpolating exactly between input datapoints, fitting the noise (see Figure 2.1). An overfitted model will show poor predictive performance when applied to unseen data as a result. In the opposite case, models that 'underfit' either lack sufficient expressiveness to capture the dataset, or have not been trained for long enough. Both of these conditions can be diagnosed by testing a model in training on a 'held-out' validation set, which the model has not been trained on. A model beginning to overfit will show decreasing training loss but a stall in the validation loss. The best performance on unseen is obtained precisely at this point , and so training should stop here. If the validation loss begins to increase beyond this point, the model is overfitted.

In some cases, gathering additional data may be difficult, or impossible. Rather than opt for a simpler model with worse performance, there are techniques to improve the diversity of the training dataset. Augmentation is a strategy to increase the effective data size by applying plausible transformations to your input data. Strategies differ

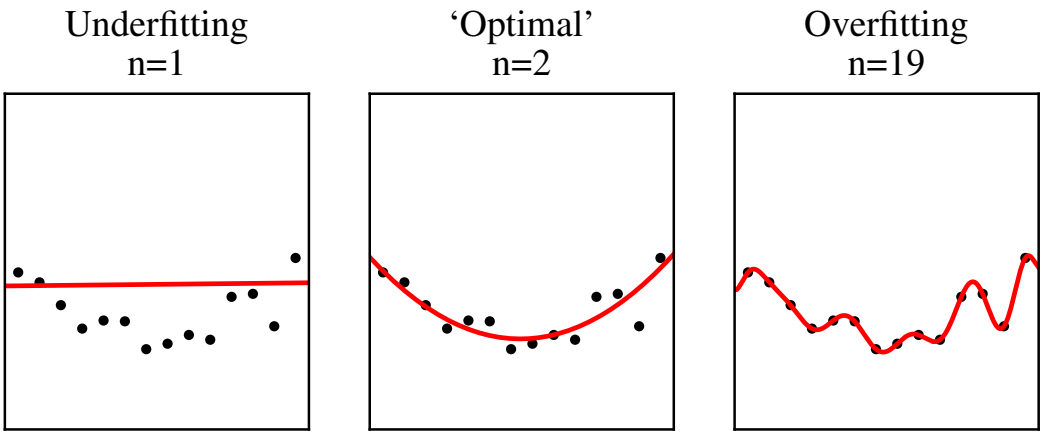

Figure 2.1: Overfitting, the 'optimal' fit, and underfitting, illustrated on a simple toy dataset generated from a quadratic model with additive Gaussian noise, with Chebyshev polynomials of varying degree fitted. In the case of underfitting, the model does not fully capture the data. In the case of overfitting , the model fully captures the dataset, including the noise – and thus would note generalise well to predicting an equivalent dataset with a different realisation of additive Gaussian noise. The 'optimal' model here captures the curvature of the dataset without interpolating noise.

based on data modality, but for images this some effective transformations are flips, rotations, contrast adjustments, hue changes, and crops. As a base example (also employed in Chapter 3, consider applying vertical/horizontal reflections and 90 degree rotations to an image. The symmetry group representing these transformations is known as the dihedral group D_4, and represents the set of all transformations applicable to a square tile that preserve the square footprint. Figure 2.2 shows the Cayley diagram for this transformation, and shows that by applying this transformation we can multiply by a factor 8 the input data size. Naturally this is an upper bound, as it is not guaranteed that each orientation is exactly as informative as each – but by including these transformations we can improve the model's resilliency to subjects in different orientations. This is particularly important for many problems in astronomy – a key example arises from galaxy morphology studies. We as humans intuitively understand that a spiral galaxy is a spiral galaxy, no matter which orientation (position angle) it has on the sky – but this is not inherent in a convolutional neural network's understanding of galaxy morphology and so makes the task significantly harder. Augmentation, or architectures that are invariant to rotation (e.g. Dieleman et al. 2015) are critical especially in the case of

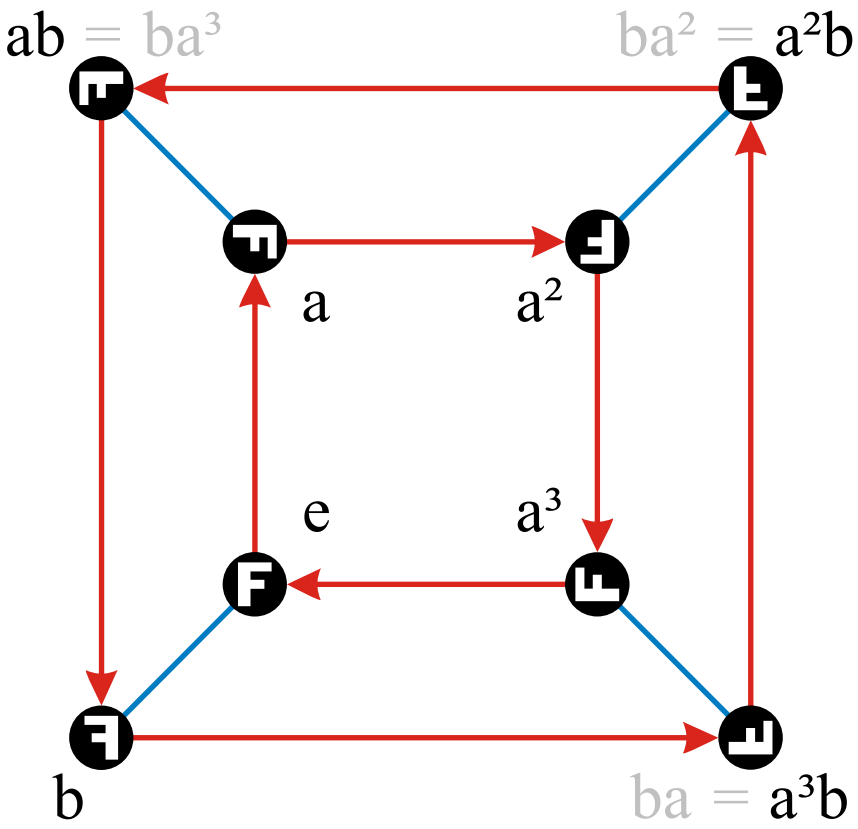

Figure 2.2: Cayley graph illustrating the various members of the D_4 symmetry group. The generator a denotes 90 degree rotations, whereas the generator b denotes reflection about the horizontal axis. Reproduced from Piesk (2016)

unsupervised learning/clustering.

2.2.5 Hyperparameter optimisation

Aside from the model parameters themselves, performance is a function of the specific *configuration* of the model itself. Items such as the number of hidden layers in a simple neural network, number of trees in a random forest, or number of convolutional layers per filter dramatically affect the classification performance of a given algorithm. These 'hyperparameters' should be optimised for maximal performance, however this is a non-trivial task. Compared to traditional optimisation, evaluating the performance of model architectures may take minutes-hours of real-world time given the length and computational cost of the training process. There may be additional complications arising from non-convex geometry and potential noise in the evaluation output arising from the stochasticity of the training process. A broad family of techniques have emerged to evaluate complex, high-dimensional, and expensive objective functions.

A typical approach is to conduct a 'grid search' – that is to evaluate candidate solution points sampled from the solution space on a discrete grid. This suffers from the curse of dimensionality however, with an exhaustive search becoming exponentially more computationally expensive with increasing number of parameters. Grid search is largely a legacy approach inherited from low-dimensional problems, and unless specific conditions are met (small number of function evaluations,requirement of covering the entire parameter space evenly) there are more performant choices. A random search (sampling uniformly over the solution space) is provably more effective (Bergstra & Bengio, 2012) at delivering optimal parameters than a grid search. Geometrically two points share the same parameter value in a random search, unlike in a grid search where by nature the 'grid' is defined by fixing one parameter.

However, as the name suggests, random search randomly explores the parameter space, with no guarantees on convergence We ideally want to incorporate knowledge of previous evaluations into where we choose our next solution to evaluate. Bayesian optimisation (Snoek et al., 2012) is a black-box surrogate optimisation algorithm that is especially well-suited to this task. Although many distinct variants and

derivatives exist, the general concept remains the same – we seek to optimise an unknown function $f(\theta)$. Given an arbitrary function $f(\theta)$, Bayesian optimisation follows three steps:

Evaluate Evaluate a given set of parameters θ – $f(\theta)$ may be computationally (or otherwise) expensive to evaluate.

Condition Condition the surrogate model on the evaluated parameters.

Predict Use a minimisation algorithm (e.g. Levenberq-Marquadt) to find the maximum of a given 'acquisition function', which yields the next set of trial points to evaluate.

This loop repeats until an appropriate solution is found, or the resource budget for optimisation is exhausted. The key idea of Bayesian optimisation is to perform optimisation (often requiring expensive Jacobian/Hessian evaluations) on a cheaper, better behaved model than the true model $f(\theta)$, and improve this model as optimisation proceeds to explore the parameter space whilst also selecting salient points. The surrogate model is generally a Gaussian process (GP; Rasmussen & Williams 2003), due to the ability to impose relevant priors on the length scale , and the ability to sample not only the mean of the function but it's covariance at trial points. The choice of acquisition function is also important, and multiple prescriptions exist for this – a natural one that lends itself well to the use of a GP is the upper confidence bound (UCB, Auer 2003), where the function to minimise is the GP function mean + some number of standard deviations. If the cost function $f(\theta)$ arises from an iterative optimisation process (e.g. training of a neural network), a further technique can be applied in conjunction with Bayesian optimisation to prune 'bad' parameter combinations before fully training them if they do not exceed some fraction of other trial points in quality, known as successive halving (Jamieson & Talwalkar, 2015). Naturally, no one metric can encompass what makes a given result 'optimal', thus multi-objective optimisation is an emerging field of study – where trial points aim to optimise multiple metrics. It is useful to introduce the concept of 'Pareto' optimality to choose between solutions in this context – a solution is Pareto optimal if a given metric cannot be improved without making at least one other worse. Plotting the set of Pareto-optimal parameters yields the Pareto 'frontier', or the set of

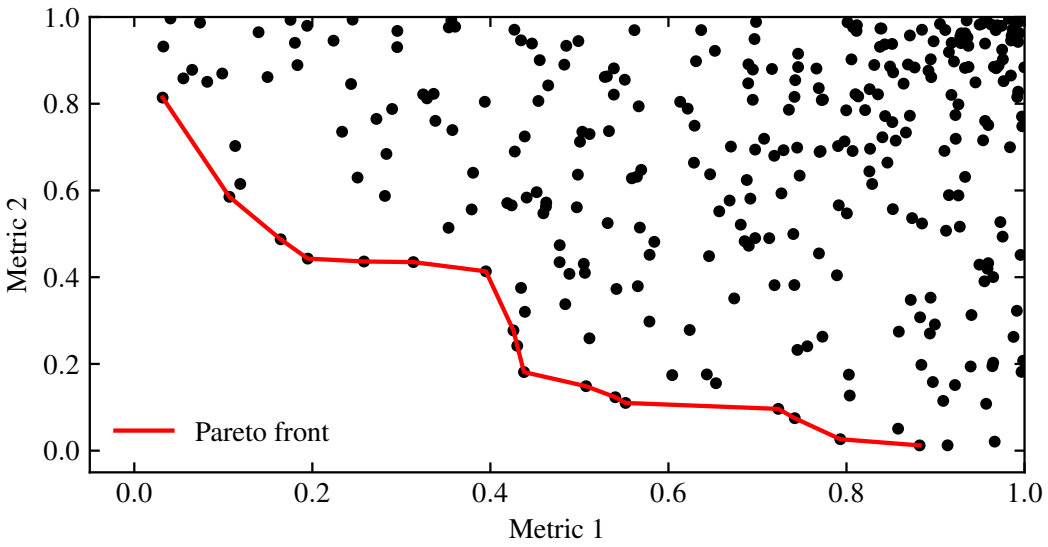

Figure 2.3: Example Pareto frontier from an optimisation process, plotting two unspecified metrics (where lower values are better) against each other to illustrate the Pareto-optimal solution set.

non-dominated candidate solutions for a given optimisation problem. By moving along the Pareto frontier, algorithm designers can make optimal tradeoffs between e.g. performance and overall runtime, rather than needing to consider all sampled solutions. This is also known as the 'skyline' in other contexts (e.g. Börzsönyi et al. 2001). Figure 2.3 illustrates the solution set yielded by a test optimisation, evaluated against two unnamed metrics where lower values correspond to better solutions. I will revisit the algorithms introduced in this section in Section 2.4.3, where they are applied directly to optimisation of parameters for difference imaging algorithms, with surprising results. Although dominant in machine learning, these optimisation methods show significant promise for some computational workflows in astrophysics where gradient descent (and more intelligent derivatives) struggle. Bayesian optimisation and related methods are already being applied to other domains of physics, where experimental setups (an 'expensive' process to evaluate) are being optimised using these techniques.

2.3 Bayesian methods

Probability lies at the heart of science, and astronomy – from quantifying the significance of detections, to rate estimates of astrophysical events and obtaining the credible intervals of parameters of interest. Science traditionally relied heavily on frequentist (classical) statistics to make headway. As our demands of our data grow more sophisticated, and the exquisite precision with which we measure quantities begins to bump up against irreducible sources of systematic error, our analyses must keep pace – incorporating existing information (priors), mitigating the effect of variables we cannot control (marginalisation), and weighing the relative likelihoods of potentially nested hypotheses (model selection). Bayesian statistics is a vital tool and facilitator for these more nuanced analyses, providing a robust framework to work within. There are also some problems that simply cannot be resolved with a frequentist outlook (e.g. the 'lighthouse problem ', see Gull 1988), owing to pathological distributions and ill-posed problems. The following subsections provide a brief overview of the heart of Bayesian methods, to provide introductions to some of the methods used in the latter chapters (see Chapter 4).

2.3.1 Bayes' theorem, maximum likelihood, and modelling

In a Bayesian outlook, probability encodes a degree of belief about the state of the world – whether it be the probability of an event occurring within a given timescale, or our belief about the value of a parameter. Our belief about the world can change upon receiving new information, and the heart of Bayesian statistics is Bayes' theorem, which provides a way to reason given conditional probabilities, and update our beliefs in line with new evidence.

Bayes' theorem takes a simple form,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where $P(\theta)$ is the prior, encoding our prior belief about the values of the parameters. $P(D|\theta)$ is the likelihood function, encoding the probability of realising a given set of ob-

servations, assuming some fixed parameters. The model that best explains the data has the *maximum likelihood* (or with prior constraints the maximum *a posteriori* probability), and we can define credible intervals for parameters of interest by bracketing the posterior probability (the highest posterior density intervals (HPDIs). $P(D)$ is the Bayesian evidence or marginal likelihood, which can be thought of as the average likelihood over the prior parameter space. This is a complex and often intractable quantity to compute, but for most purposes can simply be taken as a normalising constant. Bayes' theorem combines the above probabilities to infer the *posterior* probability $P(\theta|D)$ - that is the probability of obtaining given parameter values, given the observed data (and priors):

In contrast to frequentist statistics, where we often ask 'what is the probability of our model being correct, given our observed data', Bayesian statistics asks the subtly different question 'what is the probability of observing our data, given some underlying model (parameters)'. We do not think of our best estimate of the model parameters as one that minimises some goodness of fit (e.g. χ^2), but instead those that maximise the likelihood of observing our data given the parameters. This approach is incredibly powerful, as creating a 'generative' model forces us to consider all effects at play in the problem, and allows us to explicitly model e.g. intrinsic noise in a system, on top of our measurement noise. We can also model systematic errors in our measurement process, and 'marginalise' over them, in the process converting them to statistical errors and mitigating their impact.

Model comparison is a subtle endeavour both in frequentist and Bayesian statistics, however a natural route to proceed is by computing the ratio of posterior probabilities between the two models of interest (potentially multiplied by the ratio of the priors if they differ). This yields a Bayes factor, which quantifies the degree of evidence for one model being favoured over the other.

2.3.2 Sampling and marginalisation

Given the apparatus provided by Bayes' theorem, we can now evaluate the posterior probability of parameters θ_i given the evidence/data D. For low-dimensional θ_i, we can

simply evaluate combinations of parameters via brute-force, over a grid/random sampling of points and obtain satisfactory results. For values of n $\gtrsim$ a few however, this becomes computationally inefficient owing to the exponential increase in the volume of parameter space to evaluate, the 'curse of dimensionality'. Direct sampling is often computationally infeasible or challenging . Rejection sampling approaches are a crucial tool in the toolbox of Bayesian inference, enabling sampling from high-dimensional posterior parameter spaces in a robust way, which in turn underpins estimation of model parameters and their respective credible intervals. We also want to 'marginalise' out irrelevant parameters by summing or integrating over them to obtain the posterior distributions of the quantities we are interested in. The chief family of algorithms for this (although others exist) are Markov Chain Monte Carlo (MCMC) algorithms, in particular the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is among the most popular rejection-sampling algorithms, in part owing to it's simplicity. We seek to sample from a target distribution (in the case of Bayesian inference, the posterior distribution), by constructing a Markov Chain that draws samples from a distribution that asymptotically approaches our target distribution. We initialise our chain at a given set of parameters (ideally close to maximum likelihood), then draw samples from a 'proposal distribution' or 'transition kernel', which yields the next state we move to. This is the property that makes the Markov Chain Markovian – our next state depends only on the current state. The proposal distribution can be any distribution, although most typically is chosen to be Gaussian due to it's property of symmetry and the ease of sampling. More formally, given target distribution $\pi(\theta)$, a starting location θ_0 and a proposal distribution $Q(\theta'|\theta)$, MH sampling proceeds as follows. Over time, our sampling distribution will approach the target distribution, and we will efficiently draw samples from the target distribution – in the case of Bayesian inference our posterior. The principal tunables are the acceptance probability α, and the exact form of the proposal distribution $Q(\theta'|\theta)$. Algorithm 1 below provides a pseudocode for computing each MH step.

Algorithm 1: Metropolis-Hastings algorithm

$i \leftarrow 0$;
Set $\theta_i \leftarrow \theta_0$;
for $i \leftarrow 1$ **to** N **do**
  Sample $\theta' \sim Q(\theta'|\theta_i)$;
  Calculate acceptance ratio $\alpha = \min\left(1, \frac{\pi(\theta')Q(\theta_i|\theta')}{\pi(\theta_i)Q(\theta'|\theta_i)}\right)$;
  Generate uniform random number $u \sim \text{Uniform}(0,1)$;
  if $u \leq \alpha$ **then**
    $\theta_{i+1} \leftarrow \theta'$;
  else
    $\theta_{i+1} \leftarrow \theta_i$;
return Samples $\theta_i \ldots$;

It is **crucial** to note: Markov Chain Monte Carlo algorithms are not for 'fitting', nor do they directly maximise the likelihood. The Metropolis-Hastings algorithm draws samples with density *proportional to* the posterior probability, which may then be used to estimate the maximum *a posteriori* parameters and associated credible intervals. It is always advisable to initialise chains around a maximum likelihood estimate of the parameters to aid in 'convergence' - that is convergence of the sampling distribution to the true underlying distribution.

Adaptive samplers

One drawback of Metropolis-Hastings is the requirement to tune the parameters of the proposal distribution to obtain a good acceptance fraction. Depending on the posterior geometry, if the proposal distribution is poorly matched the Markov Chain will not fully explore the posterior space, and achieve a low acceptance fraction. This is particularly the case for correlated dimensions, where the sampling efficiency of classic MH is dramatically reduced as it is forced to take smaller jumps. Naturally, a number of 'adaptive' approaches have emerged as a result to achieve higher acceptance fractions and more efficient sampling – making use of ensembles (Foreman-Mackey et al., 2013), more powerful moves (Turner et al., 2013), and other Markov-like approaches (Neal, 2000).

One remaining issue lies in the correlation between steps in the Markov Chain,

even with adaptive methods. Small steps that remain near each other are correlated are less informative than big steps that explore the posterior, and the reduced variance of correlated steps weakens parameter estimates and statistical inferences. The key metric of interest is the 'effective sample size' (ESS; Gelman et al. 2013), which measures the number of effectively independent samples, taking into account the correlation length of the chain. The estimation error on a given quantity scales as $\frac{1}{\sqrt{N_{eff}}}$, as correlation between samples effectively reduces the information content of each step. Although it is always possible to increase ESS by simply drawing more samples, particularly for high-dimensional problems the computational cost of this may be infeasible, not to mention that issues like poorly-behaved posterior geometry cannot simply be overcome by 'sampling harder'. New methods that minimise inter- and intra-chain correlations are therefore crucial for efficient posterior exploration, with significant efforts having been invested in 'hybrid Monte Carlo' schemes, that keep the core 'accept-reject' behaviour of Metropolis-Hastings, yet use more advanced schemes for the proposal distribution, to suggest moves of higher quality (lower autocorrelation and higher acceptance probability). Hamiltonian Monte Carlo is among the most common of these, leveraging the gradients of function parameters with respect to the target distribution, and using Hamiltonian mechanics to make moves that explore well the posterior distribution. Naturally, the requirement of gradients makes this more challenging to implement than the naive Metropolis-Hastings algorithm. With the emergence of mature linear algebra frameworks (e.g. Bradbury et al. 2018) that support automatic differentiation, this 'barrier to entry' has been effectively lifted, and Hamiltonian Monte Carlo methods have found increasing utility in astronomical inference tasks. Further, general purpose 'probabilistic programming languages' (e .g. `pyro`; Bingham et al. 2019) enable the specification of complex probabilistic models in a principled and intuitive way, and are a powerful tool for constructing Bayesian models that fully capture the behaviour and parameters of the system of interest. Such tools minimise the need for more complex ad-hoc models or implementation of likelihood functions directly, minimising the risk of errors.

A full description of Hamiltonian Monte Carlo (HMC)), and the deep geometrical/dynamical connections underpinning it, is well beyond the scope of this thesis,

with the reader being referred to Betancourt (2017) for a comprehensive exploration. Section 4.3.2 of this thesis makes use of Hamiltonian Monte Carlo to sample from a non-trivial posterior probability encoding both the model parameters of a radial velocity fit, and a series of systematic 'nuisance' parameters that are later marginalised over. Although the posterior geometry of this problem was largely Gaussian, the significantly improved sampling speed and stability bodes well for reformulation of this problem into a Bayesian hierarchical model in future. HMC lies at the heart of many complex and computationally-intensive inference tasks in modern astronomy, from exoplanetary modelling (Foreman-Mackey et al., 2021), large-scale cosmological inference (Heavens, 2009; Jasche & Kitaura, 2010), and in powering data-rich hierarchical inference (Sanders et al., 2015; Shabram et al., 2020; Shen et al., 2022) tasks.

2.4 Difference imaging

As noted in Chapter 1, the technique of difference imaging lies at the heart of modern time-domain astrophysics, underpinning the discovery capabilities of large-scale sky surveys. This section provides an overview of the algorithms involved from a computational and mathematical perspective.

The central aim of difference imaging is to subtract reference flux from a new 'science' image, to search for sources that may have varied in brightness between the two epochs. Difference imaging does not rely on pre-existing knowledge of a variable sources' location, and provides significant increases in recovery of variable sources in crowded field (see e.g. Wozniak 2000; Alcock et al. 2000). Assuming pre-aligned science and reference images, three steps are required for successful difference imaging:

- Matching of PSFs between science and reference images.
- Removal of differential background between science and reference images.
- Photometric calibration of difference image.

Of these steps, matching PSF between science and reference images is the most challenging, and the one that the majority of this Section will be devoted to discussing.

For an 'ideal' image with zero noise or background contribution, the PSF-matching kernel follows from the convolution theorem Ciardullo et al. (1990); Tomaney & Crotts (1996) as the (inverse Fourier transform of) ratio of the Fourier transforms of the science and reference PSFs. Such a 'Fourier division' is poorly conditioned however, suffering both from the finite precision of floating-point arithmetic, and the non-zero Poisson noise of astrophysical observations. Strictly, if the 'science' PSF has smaller spatial scale than the 'template' PSF, this operation amounts to a deconvolution, which has the potential to significantly increase the noise of the science image and introduce spurious artifacts. Approaches need to be appropriately regularised to limit the effects of this, as well as to make better use of image pixels to avoid overfitting. Early approaches directly approximated the PSF as Gaussians, making use of the property that the Fourier transform of a Gaussian is another Gaussian to avoid the poor numerical conditioning of direct inversion – although this assumption is far too restrictive for most true instrumental PSFs. Although approaches such as Wiener filtering have seen adoption for generic PSF-filtering tasks (e.g. Boucaud et al. 2016), the often limited quality with which the PSF can be determined poses issues. We would ideally be able to compute the PSF-matching kernel with zero knowledge of the actual PSFs, as we predominantly care about the quality of the match itself. An appropriate solution to the problem in this form naturally arises via the method of least squares.

2.4.1 Difference imaging as a linear least-squares problem

To mathematically formalise the concepts introduced previously, in this subsection I present the key equations of difference imaging introduced by Alard & Lupton (1998). We adopt the notation of Bramich (2008) for clarity going forward, although note this is fully equivalent to the original A&L prescription.

In difference imaging, we seek to match a reference image R_{ij} to a science image I_{ij} using a combination of spatial convolution with a kernel K and differential background B_{ij}, where i and j are indexers for the spatial coordinates of the input

image. This forward model is written

$$M_{ij} = (K \otimes R)_{ij} + B_{ij}$$

. We seek the values of K and B_{ij} that minimise the residual sum of squares between the science image I_{ij} and our matched template R_{ij}, that is:

$$\chi^2 = \sum_{ij} \left(\frac{I_{ij} - M_{ij}}{\sigma_{ij}} \right)^2$$

, where σ_{ij} are the per-pixel uncertainties. As discrete convolution ($\otimes$) is a linear operator, we can write our kernel as a sum of basis components and factor out the coefficients:

$$M_{ij} = \sum_n (a_n k_n \otimes R)_{ij} + b_n B_{n,ij}$$

.

$$M_{ij} = B_{ij} + \sum_n a_n (k_n \otimes R) + b_n B_{n,ij}$$

, where a_n and b_n are the coefficients for each background/kernel component, and k_n and B_n are the kernel/background basis vectors. This makes the matched reference image a linear sum of kernel-convolved reference images, which is a linear model by definition – and thus solvable via least-squares. Spatially-varying kernels can be accommodated within this framework by multiplying basis components by polynomials in image space, such as

$$M_{ij} = \sum_{nm} a_{nm} \phi_{mij} (k_n \otimes R) + b_{nm} \phi_{mij} B_{n,ij}$$

where ϕ_{mij} contains m spatially-varying basis functions. In the simplest case this may correspond to a series of polynomial terms in the x, y coordinates of each pixel

The choice of basis components is naturally crucial for a proper matching of PSFs and backgrounds, and as a result many different choices have emerged. A basis set must be expressive enough to capture the potentially complex shapes of PSF-matching kernels, yet be computationally tractable for large kernels. The ‘canonical’

choice of basis vectors was introduced in Alard & Lupton (1998), as a sum of Gaussian basis functions (GFB) of differing full-width half maxima, each multiplied by a modulating polynomial. The widths of the basis functions should be tuned relative to the PSF size such that they can capture the relevant scales of the problem. Bramich (2008) introduced the 'delta-function' basis (DFB), which spans the set of basis functions with delta functions localised each pixel - with the basis coefficients a_n mapping directly to the kernel intensity values. Although far more expressive than other more constrained basis sets (e.g. GFBs), this flexibility can lead to numerical artifacting if not regularised somehow. Multiple prescriptions for this have arisen (Becker et al., 2012; Bramich et al., 2016), penalising overly complex kernels through adding a curvature-like term to the normal equations similar in concept to Tikhonov regularisation. Some other good candidates for basis functions include Gauss-Hermite polynomials, Cartesian/polar shapelets (Refregier, 2003; Massey & Refregier, 2005), and mixed-resolution delta-function bases (Bramich et al., 2013).

The quantity $\sum_{nij} a_n k_{nij}$ gives the 'photometric' normalisation factor, which gives the difference in flux scaling between the science and reference image. This should be close to one ideally, and gives the factor by which the photometric calibration of the reference should be scaled to match the science image, for difference image photometry. This too may be spatially varying, particularly important in the case of wide-field imaging, where airmass, seeing, and transparency may vary across the frame.

Implementation of the above least-squares methods on real imaging data has many subtleties that must be taken into account. However, in comparison to Fourier space methods, least squares operates on the pixel level, making weighting, masking, and other similar operations more simple than would be in frequency space. The core methods of difference imaging have remained largely unchanged since Alard & Lupton (1999), instead with focus moving to better treatment of data artifacts (via masking), better selection of optimal stamps to determine the kernel and any spatial variation (via sigma clipping), and propagation of uncertainties for optimal source extraction. Multiple implementations of the core Alard and Lupton algorithm exist, implementing various levels of pre-processing . The original, `ISIS` (Alard & Lupton, 1999), is still regularly

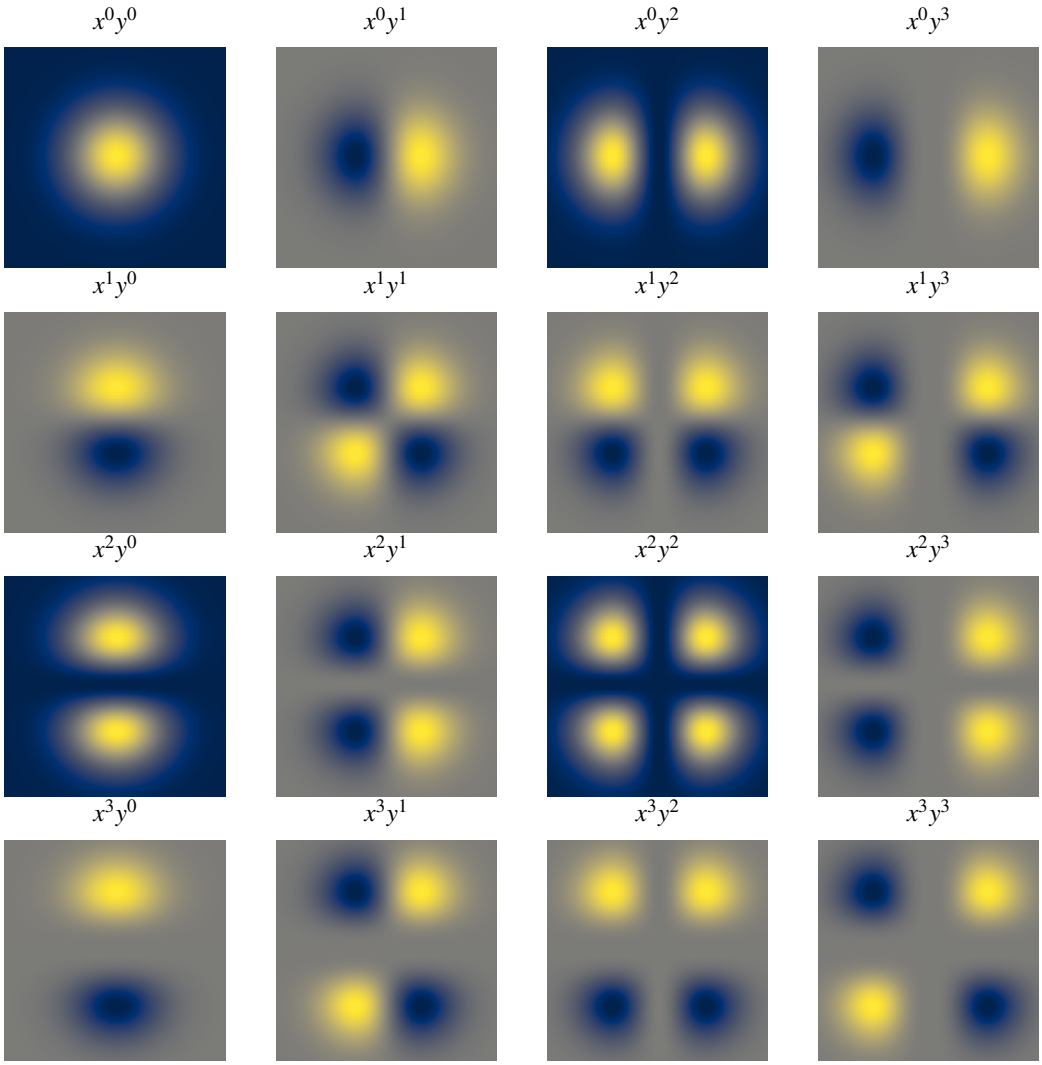

Figure 2.4: Illustration of the Gaussian basis function set, with up to third order terms in x and y. Typically three of these of differing widths are combined to provide representative power for a range of PSFs.

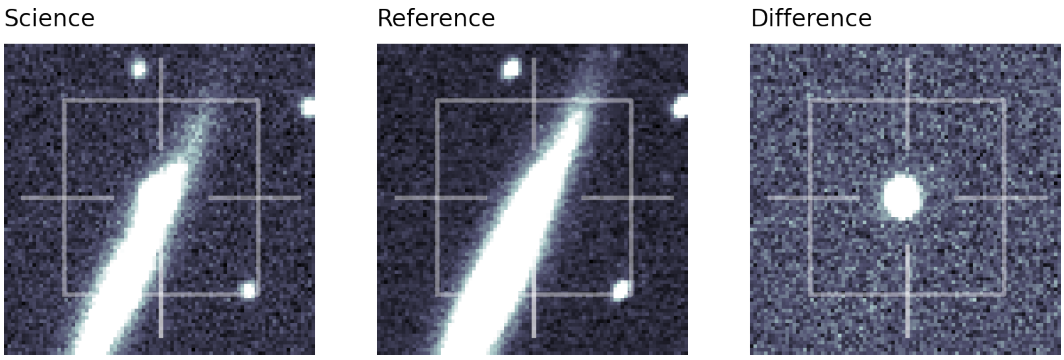

Figure 2.5: GOTO difference image of the nearby SN Ia, SN 2021hiz. The source is visible in the science image, but embedded in the bright disk of the host galaxy. Subtraction of the reference image yields a clear PSF-shaped residual at the location of the SN.

used today. More recent implementations such as HOTPANTS (Becker, 2015) are better optimised and built for wide- field surveys with non-trivial spatial dependencies, varying noise levels, and correct noise scaling. An example difference image is shown in Figure 2.5, of the nearby SN Ia SN 2021hiz.

Some modern approaches have returned to the 'frequency-based' approaches, leveraging the availability of highly- optimised Fast Fourier Transform libraries (e.g. Frigo 1999). The Zackay-Ofek-Gal Yam (ZOGY; Zackay et al. 2016) is among the most popular of these approaches, delivering a provably 'optimal' matching of PSF and image subtraction under Gaussian noise. Most recently, Hu et al. (2022) provided a hybrid approach, performing a least-squares fit in Fourier space that is equivalent to a least-squares fit in real space under Parseval's theorem, in the 'saccadic Fast Fourier Transform' (SFFT) code, designed for rapid execution on graphics processing units (GPUs). Both methods require explicit knowledge of, or empirical construction of, point spread functions for both the science and reference image, which is a challenging and computationally-intensive task in it's own right.

Despite a diverse set of algorithms, the various failure modes of difference imaging are common across all algorithms. Mismatched PSFs lead to ringing artifacts around bright sources, where the radial extents of the science and reference PSFs are not perfectly matched. Misalignments between science and reference images yields 'dipole' artifacts, with positive and negative flux residuals in close proximity around

sources. These propagate through source detection/photometry into false detections, thus significant work continues on improving difference imaging algorithms, especially in the shadow of upcoming next-generation surveys like the Vera Rubin Legacy Survey of Space and Time.

2.4.2 Challenges

PSF variations add additional complexity to the solution of the best-fit kernel and background parameters, requiring the incorporation of spatial terms and thus increasing the size of the least-squares matrix to be solved. For modern wide-field instrumentation this is a necessity, however. One simple approach is to split the image into 'regions' – smaller cutouts of the image that tile it exactly. This allows the representation of arbitrarily-complex spatial variations (stricly, piecewise-polynomial) and enables trivial parallelisation. There are two significant problems with this however: existing codes do not enforce continuity of PSF or background across region boundaries, leading to significant discontinuities which may cause issues with source extraction in latter pipeline steps. Sub-division into regions also reduces the number of stars per region, providing poorer constraints on the kernel than otherwise.

One significant remaining challenge for all approaches is the treatment of correlated noise in images. Correlated noise results from spatially/temporally-close pixels showing non-zero covariances in their intensities, thus resulting in correlated residuals upon subtraction of a model. Mathematically, the covariance matrix of measurements contains off-diagonal terms (implying co-variance between measurements). This is illustrated in Figure 2.6. Correlated noise can arise from a variety of sources , however for difference imaging it principally arises from interpolation artifacts, background subtraction, or from the kernel itself, which naturally spreads the (approximately) Gaussian noise of the image out on spatial scales comparable to the kernel size. The method of least-squares is the best linear unbiased estimator (BLUE) for linear regression only when the core assumptions are met, that is residuals are normally distributed, with significant deviations from optimality in the presence of correlated residuals – affecting kernel/background determination on correlated images. The Poissonian statistics im-

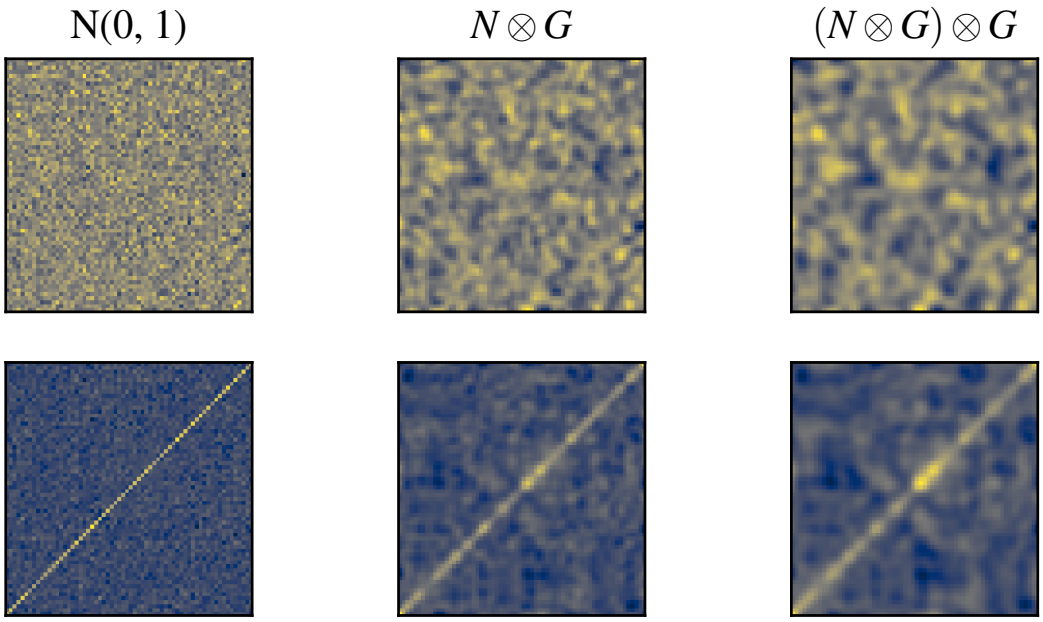

Figure 2.6: Illustration of white and correlated noise, generated via convolution. The top panels show the progressive correlation of initially white, Gaussian noise as it is convolved with a Gaussian with σ=1 pixel. The bottom panels show the covariance matrix for each image – starting with a pure diagonal covariance matrix for white noise, and widening to include nearby off-diagonal terms in the correlated noise images.

plicit in many source detection routines do not account for inter-pixel correlation in the image, leading to the detection of spurious (often real-looking) sources. The correlations lead to an underestimation of the variance, in turn leading to a lower 'true' source detection threshold than specified. Whilst a simple ad-hoc increase in the source detection threshold will reject these sources, different difference images will have differing degrees of correlation, making this non-ideal. Modelling the correlation itself seems like a natural next step, following work on modelling covariances in time-series using Gaussian processes (GPs; Rasmussen & Williams 2006). However, applying this in 2D is more challenging owing to the complexity scaling of GPs (naively $O(N^3)$, dominated by the inversion of the covariance matrix), and many of the optimisations (e.g. Foreman-Mackey 2018) applied to 1D (or N-D independent) datasets to reduce the computational complexity are not directly applicable. Whilst ZOGY provides a theoretical guarantee of white noise in the difference images, in practice implementation details and deviations from the model assumptions on Gaussianity limit this. There is no robust

extension[3] to ZOGY that permits non-stationary kernels and backgrounds, limiting applicability to wide-field imaging. Recent work has proposed an extension to the original Alard and Lupton pixel-space difference imaging, involving an 'afterburner' decorrelation kernel derived from the PSF-matching kernel to whiten the image noise (Reiss & Lupton, 2016). This is a promising first step to resolving the issues of correlated noise in difference images, yet requires further work to extend to spatially-varying kernels, and spatially-varying degrees of correlation in a robust way.

All difference imaging algorithms, whether explicit (the kernel basis/spatial dependencies in Alard and Lupton-like algorithms) or implicit (pre-processing steps required for ZOGY/SFFT) have configurable parameters – these are largely left unchanged from the defaults suggested by a given algorithm's author owing to the complexity of tuning these per- instrument. In the following sub-section, I introduce a metric-driven Bayesian optimisation framework for performing principled parameter tuning of difference imaging algorithms, inspired by 'hyperparameter tuning' techniques from the computer science literature.

2.4.3 Data-driven optimisation of difference imaging parameters

> **Note:** This chapter is based on an in-preparation manuscript. Full performance evaluation, and release of accompanying code is deferred to the full publication.

As with any algorithm, parameter tuning in difference imaging is important to extract the maximum performance possible. A similar problem arises in machine learning, with the 'hyperparameters' of a particular machine learning model being tunables that can yield significant improvements in classification performance (see e.g. Section 3.3.1) There is significant overlap between these tasks: both processes are computationally expensive to evaluate, have high- dimensional non-trivial solution spaces, and may be subject to stochastic 'noise' in their evaluations. As a proof-of-concept for developing more general-purpose optimisation routines in future, I present here some early work

[3] Sub-division of the image into 'regions' does not enforce 'smooth' (in the mathematical sense) variation of PSF or background, thus can cause severe artifacts, spurious edge effects

Table 2.1: Difference imaging parameter bounds, and their accompanying distributions

Parameter	Lower bound	Upper Bound	Default	Optimised
g_0 (times seeing)	0.5	1.5	0.5	0.684
g_1 (times seeing)	g_0	2.0	1.0	1.305
g_2 (times seeing)	g_1	3.5	2.0	2.900
Kernel size	11	19	-	13
Kernel substamp size	16	23	-	17

on applying Bayesian optimisation to optimising the (hyperparameters) of the difference imaging algorithm HOTPANTS, with a view to minimising the number of spurious sources created as part of the process.

The parameters we choose to optimise, and their bounds are given in Table 2.1 – focusing primarily on the kernel. Note that we choose to specify the widths of each Gaussian basis function scaled according to seeing (estimated as the median across the frame). This ensures that the PSF-matching kernel remains an appropriate size across a range of observing conditions . We also enforce an ordering on the Gaussian basis function components according to intuition – the zeroth one that is modulated by a high-order polynomial should have the smallest width, to better fit the complex core features, with the lower order ones requiring a wider width to fit the wings of the differential PSF.

We choose to optimise a single event follow-up as a proof of concept, taken from a *Fermi* GRB follow-up campaign. The images reach typical depths of 20th magnitude, and have a range of seeing values. As a proxy metric for the quality of a difference image, we choose to optimise the number of difference image detections divided by the number of science image detections (f_{diff}). This should be close to zero in each image (excluding transient detections), so minimisation of this minimises the number of spurious difference image detections. Scaling through by the number of science image detections ensures that each image is correctly weighted: we naturally expect more residuals in a dense stellar field, and thus we should not inadvertantly upweight these images. We use the `optuna` (Akiba et al., 2019) package, using the Tree-Parzen estimator algorithm (Bergstra & Bengio, 2012) to successively minimise f_{diff}, pre-conditioning with 5 trial solutions uniformly sampled from the parameter space, and performing a fur-

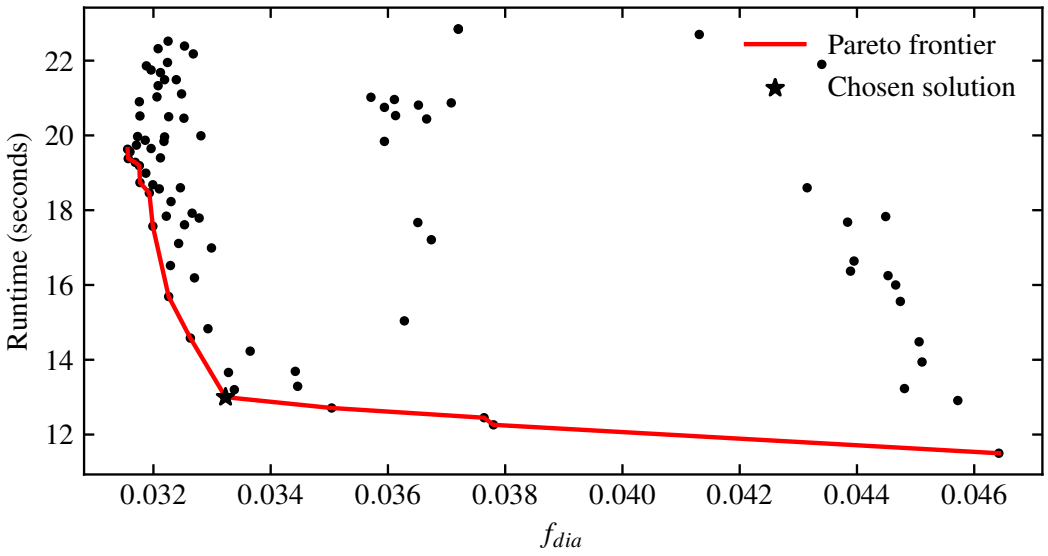

Figure 2.7: Plot showing the performance of a range of parameter solutions found via Bayesian optimisation, as a function of the runtime and figure of merit. The Pareto frontier is overplotted to show the optimal solution set.

ther 75 acquisitions using the upper confidence bound acquisition function. The resultant family of solutions is plotted in Figure 2.7 as a function of f_{diff} and the runtime of the algorithm. The optimal parameters (see Table 2.1) we obtain for the GOTO test dataset are somewhat different from the defaults suggested – favouring a smaller high-order component, and broader low-order components than default, and favouring a smaller kernel and stamp size than suggested by default. Although we only use a small subset of images, it is clear we can achieve significant improvements over the 'canonical' function values through data-driven optimisation, especially in the case of wide-field imaging surveys where spatial and PSF variations cause issues. This approach enables us to choose a tradeoff between runtime and image quality that suits the chosen use case we are optimising for. A crucial next step is extending this to a larger set of images to minimise sampling variance in the metrics of choice, as well as expanding the parameter space probed to remain open to possible unforeseen combinations. Owing to the compute time required for such searches, node-level parallelisation is an important implementation detail required for scaling up.

There is no reason the data-driven optimisation methods outlined in this sub-

section apply only to difference imaging . The (hyper)parameters of any algorithm are in principle optimisable in this way, provided a suitable metric/metrics exists and a large enough volume of test data exists such that any inter-sample variance in statistics can be minimised. In particular, algorithms such as source detection/segmentation, background extraction, and image alignment are ripe for improvements via these techniques. Given the relative agnosticity of Bayesian optimisation to what is being optimised, one could even optimise all stages of e.g a difference imaging pipeline simultaneously, to exploit the dependence of sequential steps on each other, to enable the maximal performance. It is important to acknowledge there are issues with this approach – in particular the potential for mis-specification of metrics to skew results, however metric-driven optimisation shows significant promise in extracting maximal performance from astronomical pipelines. We are in the infancy of such techniques, and work from the machine learning community on hyperparameter optimisation will only further improve the viability of high-dimensional exotic optimisation.

2.4.4 Future directions

Taking into account the challenges listed in Section 2.4.2, significant scope exists for the improving the efficacy of existing difference imaging algorithms, with the potential to reduce the considerable false-positive rate such algorithms carry due to the artifacts they cause – whilst also bringing depth improvements for wide-field surveys.

A particularly promising avenue for improving execution time is by leveraging the highly optimised convolution and image processing routines delivered by modern deep learning frameworks, in tandem with the tried-and-tested approaches of the past. This makes offloading highly-parallelisable steps of difference imaging (namely the convolutions and matrix multiplications) to the GPU and just-in-time compilation of key serial-only branches possible, with potentially significant performance gains available. Whilst GPU implementations of these algorithms exist already, they are implemented in domain-specific languages (e.g. CUDA C++) that make them inaccessible to the average astronomer. In contrast, deep learning frameworks provide bindings from high-level languages, which makes development , debugging, and optimisation more intuitive.

Some early work (Hitchcock et al., 2021) makes use of this approach, using stochastic gradient descent (see Section 2.2) to solve for the optimal kernel and background parameters numerically. The convergence properties of this approach are not ideal for deployment into large-scale automated discovery pipelines, however this work underscores the significant performance gains possible from adopting new architectures for the difference imaging problem, and extension to more complex scenarios where a linearised treatment of the problem is simply not possible.

Chapter 3

Transient-optimised real-bogus classification

Note

This chapter is taken from the 2021 MNRAS article, *Transient-optimized real-bogus classification with Bayesian convolutional neural networks - sifting the GOTO candidate stream* (Killestein et al., 2021), and based on the Authors' Accepted Manuscript. The majority of work contributing to this publication was performed by myself, with some additional contributions and guidance from co-authors.

Abstract

Large-scale sky surveys have played a transformative role in our understanding of astrophysical transients, only made possible by increasingly powerful machine learning-based filtering to accurately sift through the vast quantities of incoming data generated. In this paper, we present a new real-bogus classifier based on a Bayesian convolutional neural network that provides nuanced, uncertainty-aware classification of transient candidates in difference imaging, and demonstrate its application to the datastream from the GOTO wide-field optical survey. Not only are candidates assigned a well-calibrated probability of being real, but also an associated confidence that can be used to priori-

tise human vetting efforts and inform future model optimisation via active learning. To fully realise the potential of this architecture, we present a fully-automated training set generation method which requires no human labelling, incorporating a novel data-driven augmentation method to significantly improve the recovery of faint and nuclear transient sources. We achieve competitive classification accuracy (FPR and FNR both below 1%) compared against classifiers trained with fully human-labelled datasets, whilst being significantly quicker and less labour-intensive to build. This data-driven approach is uniquely scalable to the upcoming challenges and data needs of next-generation transient surveys. We make our data generation and model training codes available to the community.

3.1 Introduction

Transient astronomy seeks to identify new or variable objects in the night sky, and characterise them to learn about the underlying mechanisms that power them and govern their evolution. This variability can occur on timescales of milliseconds to years, and at luminosities ranging from stellar flares to luminous supernovae that outshine their host galaxy (Kulkarni, 2012; Villar et al., 2017). Through observations of optical transient sources we have obtained evidence of the explosive origins of heavy elements (e.g. Abbott et al. 2017e; Pian et al. 2017), traced the accelerating expansion of our Universe across cosmic time (e.g. Perlmutter et al. 1999), and located the faint counterparts of some of the most distant and energetic astrophysical events known: gamma-ray bursts (e.g. Tanvir et al. 2009). Requiring multiple observations of the same sky area to detect variability, transient surveys naturally generate vast quantities of data that require processing, filtering, and classification – this has driven the development of increasingly powerful techniques bolstered by machine learning to meet the demands of these projects.

Many of the earliest prototypical transient surveys began as galaxy-targeted searches, performed with small field-of-view instruments. In the early stages of these surveys candidate identification was performed manually, with humans 'blinking' im-

ages to look for varying sources. This process is time-consuming and error-prone, and represented a bottleneck in the survey dataflow which heavily limited the sky coverage of these surveys. The first 'modern' transient surveys (e.g. LOSS; Filippenko et al. 2001) used early forms of difference imaging to detect candidates in the survey data, automating the candidate detection process and enabling both faster response times and greater sky coverage. LOSS proved extremely successful, discovering over 700 supernovae in the first decade of operation, providing a homogeneous sample that has proven useful in constraining supernova rates for the local Universe (Leaman et al., 2011; Li et al., 2011).

Difference imaging has since emerged as the dominant method for the identification of new sources in optical survey data. With this method, an input image has a historic reference image subtracted to remove static, unvarying sources. Transient sources in this difference image appear as residual flux, which can be detected and measured photometrically using standard techniques. Various algorithms have been proposed for optical image subtraction, either attempting to match the point spread function (PSF) and spatially-varying background between an input and reference image (Alard & Lupton, 1998; Becker, 2015), or accounting for the mismatch statistically (Zackay et al., 2016) to enable clean subtraction. Difference imaging also provides an effective way to robustly discover and measure variable sources in crowded fields (Wozniak, 2000).

Driven by both improvements in technology (large-format CCDs, wide-field telescopes) and difference imaging algorithms, large-scale synoptic sky surveys came to the fore. In this mode, significant areas of sky can be covered each night to a useful depth and candidate transient sources automatically flagged. This has driven an exponential growth in discoveries of transients, with over 18,000 discovered in 2019 alone[1]. Wide-field surveys such as the Zwicky Transient Facility (ZTF; Bellm et al. 2019), PanSTARRS1 (PS1; Chambers et al. 2016a), the Asteroid Terrestrial-impact Last Alert System (ATLAS; Tonry et al. 2018), and the All Sky Automated Survey for SuperNovae (ASAS-SN; Shappee et al. 2014) have proven to be transformative, collectively discovering hundreds of new transients per night.

[1] https://wis-tns.org/

With the ability to repeatedly and rapidly tile large areas of sky in order to search for new and varying sources, the follow-up of optical counterparts to poorly localised external triggers became possible, in the process ushering in the age of multi-messenger astronomy. An early example was detection of optical counterparts to *Fermi* gamma-ray bursts by the Palomar Transient Factory (PTF; Law et al. 2009). Typical localisation regions from the *Fermi* GBM instrument (Meegan et al., 2009) were of order 100 square degrees at this time, representing a significant challenge to successfully locate comparatively faint ($r \sim 17 - 19$) GRB afterglows. Of the 35 high-energy triggers responded to, 8 were located in the optical (Singer et al., 2015), demonstrating the emerging effectiveness of synoptic sky surveys for this work.

Another recent highlight has been the detection of an optical counterpart to a TeV-scale astrophysical neutrino detected by the IceCUBE facility (Aartsen et al., 2017). Recent and historical wide-field optical observations of the localisation area combined with high-energy constraints from *Fermi* enabled the identification of a flaring blazar, believed to be responsible for the alert (IceCube-170922A; IceCube Collaboration et al. 2018b) . This rapidly increasing survey capability has culminated recently in the landmark discovery of a multi-messenger counterpart to the gravitational wave (GW) event GW170817 (Abbott et al., 2017c,e).

3.1.1 Real-bogus classification

For many years, the rate of difference image detections generated per night by sky surveys has significantly exceeded the capacity of teams of humans to manually vet and investigate each one. This has motivated the development of algorithmic filtering on new sources, to reject the most obvious false positives and reduce the incoming datastream to something tractable by human vetting. With the growing scale and depth of modern sky surveys, simple static cuts on source parameters cannot keep pace with the rate of candidates, with high false positive rates leading to substantial contamination by artifacts. This situation has motivated the development of machine learning (ML) and deep learning (DL) classifiers, which can extract subtle relationships/connections between the input data/features and perform more effective filtering of candidates. The

dominant paradigm for this task has so far been the real-bogus formalism (e.g. Bloom et al., 2012), which formulates this filtering as a binary classification problem. Genuine astrophysical transients are designated 'real' (score 1), whereas detector artefacts, subtraction residuals and other distractors are labelled as 'bogus' (score 0). A machine learning classifier can then be trained using these labels with an appropriate set of inputs to make predictions about the nature of a previously-unseen (by the classifier) source within an image.

This real-bogus classification is only one step in a transient detection pipeline. Having established the candidates appearing as astrophysically real sources, further filtering is required to determine if they are scientifically interesting, or distractors – the definition of "interesting" is naturally governed by the science goals of the survey. This process draws in contextual information from existing catalogues, historical evolution, and more fine-grained classification routines. The last step before triggering follow-up and further study (at least currently) is human inspection of the remaining candidates. No single filtering step is 100% efficient in removing false positives/low significance detections, thus human vetting is required to identify promising candidates and screen out any bogus detections that have made it this far. Real-bogus classification is the most crucial step, reducing the volume of candidates that later steps must process and the amount of bogus candidates that humans must eventually sift through to find interesting objects – a balance between sensitivity (to avoid missing detections irretrievably) and specificity (avoiding floods of low-quality candidates) must be reached.

Real-bogus classification is a well-studied problem, beginning with early transient surveys (Romano et al., 2006; Bailey et al., 2007), and evolving both in complexity and performance with the increasing demands placed on it by larger and deeper sky surveys such as PTF (Brink et al., 2013), PanSTARRS1 (Chambers et al., 2016a), and the Dark Energy Survey (Goldstein et al., 2015). Early classifiers were generally built on decision tree-based predictors such as random forests (Breiman, 2001), using a feature vector as input. Feature vectors comprise extracted information about a given candidate, and often include broad image-level statistics/descriptions designed to maximally separate real and bogus detections in the feature space. Examples include the

source full-width half maximum computed from the 2D profile, noise levels, and negative pixel counts. More elaborate features can be composed via linear combinations of these quantities, which may exploit correlations and symmetries. Another method of deriving features is to compute compressed numerical representations of the source via Zernicke/shapelet decomposition (Ackley et al., 2019).

However, feature selection can represent a bottleneck to increasing performance. Features are typically selected by humans to encode the salient details of a given detection, attempting to find a compromise between classification accuracy and speed of evaluation. This introduces the possibility of missing salient features entirely, or choosing a sub-optimal combination of them.

Directly using pixel intensities as a feature representation avoids choosing features entirely, instead training on flattened and normalised input images (Wright et al., 2015; Mong et al., 2020), these have demonstrated improved accuracy over fixed-feature classifiers. However, this approach quickly (quadratically) becomes inefficient for large inputs. Using a smaller input size means information on the surrounding area of each detection is unavailable, limiting the visible context and affecting classification accuracy as a result.

Recently, convolutional neural networks (CNN; LeCun et al. 1995) have led to a paradigm shift in the field of computer vision and machine learning, which has been transformative in the way we process, analyse, and classify image data across all disciplines. CNNs use learnable convolutional filters known as kernels to replace feature selection. These filters are cross-correlated with the input images to generate 'feature maps', effectively compact feature representations. Through the training process, the filter parameters are optimised to extract the most salient details of the inputs, which can then be fed into fully-connected layers to perform classification or regression. In this way, the model can select its own feature representations, avoiding the bottleneck of human selection. Multiple layers can be combined to achieve greater representational power, known as deep learning (LeCun et al., 2015). Recent work using CNNs has demonstrated state-of-the-art performance at real-bogus classification (Gieseke et al., 2017; Cabrera-Vives et al., 2017; Duev et al., 2019; Turpin et al., 2020). CNNs

are also efficiently parallelisable making them suitable for high-volume data processing tasks. Whilst providing substantial accuracy improvements over previous techniques, deep learning is particularly reliant upon large and high quality training sets to minimise overfitting, arising from the high number of model parameters. Although augmentation and regularisation techniques can minimise this risk, they are no substitute for a larger dataset. The performance of any classifier is ultimately limited by the error rate on the training labels, so it is important to also ensure the dataset is accurately labelled. Making a large, pure, and diverse training set can be among the most challenging parts of developing a machine learning algorithm, and significant effort has been focused on this area in recent years.

Traditionally the 'gold-standard' for machine learning datasets across computer science and astronomy has been human-labelled data, as this represents the ground truth for any supervised learning task. Use of citizen science has proven to be particularly effective, leveraging large numbers of participants and ensembling their individual classifications to provide higher accuracy training sets for machine learning through collaborative schemes such as Zooniverse (Lintott et al., 2008; Mahabal et al., 2019). However, even in large teams, human labelling of large-scale datasets is time-consuming and inefficient requiring hundreds–thousands of hours spent collectively to build a dataset of a suitable size and purity. Specifically for real-bogus classification, there are also issues with completeness and accuracy for human labelling of very faint transients close to the detection limit. These faint transients are where a classifier has potential to be the most helpful, so if the training set is fundamentally biased in this regime, any classifier predictions will be similarly limited. To go beyond human-level performance, we cannot solely rely on human labelling, additional information is required. One specific aspect of astronomical datasets that can be leveraged to address both issues discussed above is the availability of a diverse range of contextual data about a given source. Sizeable catalogues of known variable stars, galaxies, high energy sources, asteroids, and many other astronomical objects are freely available and can be queried directly to identify and provide a more complete picture of the nature of a given source.

Significant effort is being invested in data processing techniques for transient astronomy in anticipation of the Vera C. Rubin Observatory (Ivezić et al., 2019), due to begin survey operations in 2024. Via the Legacy Survey of Space and Time (LSST), the entire southern sky will be surveyed down to a nominal depth of $g' \sim 24.5$ in 5 colours at high cadence, providing an unprecedented discovery engine for transients to depths previously unprobed at this scale. The dataflow from this project is expected to be a factor 10 greater than current transient surveys, and promises to be transformative in the fields of supernova cosmology, detection of potentially hazardous near-Earth asteroids, and mapping the Milky Way in unprecedented detail. The main high-cadence deep sky survey promises to provide a significant increase in the number of genuine transients we detect, but also a significant increase in the number of bogus detections assuming there are not similarly large improvements in the capability of machine learning-based filtering techniques. Development of higher-performance classifiers is crucial to fully exploit this stream, but also more granular classification involving contextual data (as recently demonstrated by Carrasco-Davis et al. 2020) to ensure that novel and scientifically important candidates are identified promptly enough to be propagated to teams of humans and followed up.

A related goal of increasing importance in the big data age of the Rubin Observatory and similar projects is that of quantifying uncertainty – being able to identify detections that the classifier is confident are real, and providing a classifier a way to indicate uncertainty on more tenuous examples. This objective goes beyond the simple value of the real-bogus score, and can then be used to find the optimal edge cases to feed to human labellers, allowing new data to be continually integrated to improve performance and keep the classifier's knowledge current and applicable to a continuously evolving set of instrumental parameters. Current generation transient surveys provide a crucial proving ground for development of these new techniques.

3.1.2 The Gravitational-Wave Optical Transient Observer (GOTO)

The Gravitational-Wave Optical Transient Observer (Steeghs et al., 2021) is a wide-field optical array, designed specifically to rapidly survey large areas of sky in search of the

weak kilonovae and afterglows associated with gravitational wave counterparts. The work we present in this paper was conducted during the GOTO prototype stage , using data taken with a single 'node' of telescopes situated at the Roque de los Muchachos observatory on La Palma . Each node comprises 8 co-mounted fast astrograph OTAs (optical tube assemblies) combining to give a ~ 40 square degree field of view in a single pointing. The GOTO prototype performs surveys using a custom wide L band filter (approximately equivalent to $g' + r'$) down to $L \approx 20$, providing an effective combination of fast and deep survey capability uniquely suited to tackling the challenging large error boxes associated with gravitational wave detections. As demonstrated in Gompertz et al. (2020), the prototype GOTO installation is capable of conducting sensitive searches for the optical counterparts of nearby binary neutron star mergers, even with weak localisations of $\sim$1000 square degrees. When not responding to GW events, GOTO performs an all-sky survey utilising difference imaging to search for other interesting transient sources. Although the GOTO prototype datastream will be the primary data source used to investigate the performance of the machine learning techniques developed in this paper, the methods are inherently scalable and will also be deployed for the future GOTO datastream from 4 nodes spread over two sites. For now, we concentrate on a calendar year of prototype operations (spanning 01-01-2019 – 01-01-2020) – which represents a significant dataset, comprising 44,789 difference images in total.

Raw images are reduced with the GOTO pipeline (Steeghs et al., 2021). Here we provide a very brief overview of the process for context, and delegate more in-depth discussion to the specific upcoming pipeline papers. The typical survey strategy for GOTO is three exposures per pointing, which undergo standard bias, dark and flat correction, and then are median-combined to reject artifacts and improve depth. Throughout this paper we refer to this median-combined stack of subframes as a 'science image'. Each combined image is matched to a reference template, which passes basic quality checks, and aligned using the SPALIPY [2] code. Image subtraction is performed on the aligned science and reference images with the HOTPANTS algorithm (Becker,

[2] https://github.com/Lyalpha/spalipy

2015) to generate a difference image. To locate residual sources in the difference image, source extraction is performed using SExtractor (Bertin & Arnouts, 1996). Detections in the difference image are referred to as 'candidates' through the remainder of this paper. For each candidate, a set of small stamps are cut out from the main science, template and difference images and this forms the input to the GOTO real-bogus classifier. This process and proposed improvements are discussed in more detail in Section 3.2.1. From here, candidates that pass a cut on real-bogus score (using a preliminary classifier) are ingested into the GOTO Marshall – a central website for GOTO collaborators to vet, search and follow-up candidates (Lyman et al., in prep.).

In line with the principal science goals of the GOTO project, the real-bogus classifier discussed in this work is constructed specifically to maximise the recovery rate of extragalactic transients and other explosive events such as cataclysmic variable outbursts. Small-scale stellar variability can be easily detected via difference imaging, but is better studied through the aggregated source light curves. An operational requirement for the current version of this classifier is the ability to perform consistently across multiple different hardware configurations. During classifier development, the GOTO prototype used two different types of optical tube design, each with varying optical characteristics that led to different point spread functions, distortion patterns, and background levels/patterns. Due to limited data availability, training a classifier for each individual OTA (or group of OTAs of the same type) was not viable. This requirement adds an additional operational challenge over survey programs such as the Zwicky Transient Facility (ZTF, Bellm et al., 2019) and PanSTARRS1 (PS1, Chambers et al., 2016a), which use a static, single-telescope design. If acceptable results can be achieved with this heterogeneous hardware configuration, then further performance gains can be expected when the design GOTO hardware configuration is deployed. This will use telescopes of consistent design and improved optical quality meaning less model capacity needs to be directed towards making the classification performance stable and across a diverse ensemble of optical distortions.

In this paper, we propose an automated training set generation procedure that enables large, minimally contaminated, and diverse datasets to be produced in less

time than human labelling and at larger scales. This procedure also introduces a data-driven augmentation scheme to generate synthetic training data that can be used to significantly improve the performance of any classifier on extragalactic transients of all types, but with particular effectiveness for nuclear transients. Using this improved training data, we apply Bayesian convolutional neural networks (BCNNs) to astronomical real-bogus classification for the first time, providing uncertainty-aware predictions that measure classifier confidence, in addition to the typical real-bogus score. This opens up promising future directions for more complex classification tasks, as well as optimally utilising the predictions of human labellers. We emphasise that although this classifier is discussed in the context of GOTO and our associated science needs, the techniques discussed are fully general and could be applied to general real-bogus classification at other projects easily. Our code, GOTORB, is made freely available online [3] with this in mind.

3.2 Training set generation and augmentation

The 'real' content of our training set is composed of minor planets, similar to Smith et al. (2020). Assuming the sky motion is large (but not so large that the source is trailed) these objects are typically detected in the science image but not the template image, which provides a clean subtraction residual resembling an explosive transient. Due to the large pixels of the GOTO detectors and short exposure times of each sub-image, very few asteroids move sufficiently quickly to trail. We estimate that sky motions of 1 arcsec per minute or greater will lead to trailing.

There are significant numbers of asteroids detectable down to $L \sim 20.5$ with GOTO, and the sky motion ensures that a diverse range of image configurations are sampled. With the large ~ 40 square degree field of view provided by GOTO, an whole-sky average of 4.6 asteroids per pointing are obtained, with this number significantly increasing towards the ecliptic plane. Using ephemerides provided by the `astorb` database (Moskovitz et al., 2019), based on observations reported to the Minor

[3] https://github.com/GOTO-OBS/gotorb

Planet Center[4], difference image detections can be robustly cross-matched to minor planets in the field. This provides a significant pool of high-confidence, unique, and diverse difference image detections from which to build a clean training set.

We use the online SkyBoT cone search (Berthier et al., 2006, 2016) to retrieve the positions and magnitudes of all minor planets within the field of view of each GOTO image, then cross-match this table with all valid difference image detections using a 1 arcsec threshold value to identify the asteroids present in the image. The ephemerides provided are of sufficient quality that this is adequate to match even faint ($L \sim 20$) asteroids. To avoid spurious cross-matches, only asteroids brighter than the 5-sigma limiting magnitude of the image are considered. An alternative offline cone search is made accessible via the PYMPC package[5] Python package, which the code can fall back on if SkyBoT is unavailable. Using minor planets, the training set can reliably be extended to fainter magnitudes, where the performance of human vetters begins to significantly decrease. Figure 3.1 illustrates the magnitude distribution of minor planets used to construct the training set.

To create the bogus content of our training set, we randomly sample detections in the difference image following Brink et al. (2013). Bogus detections overwhelmingly ($\gtrsim 99\%$) outnumber real detections in each difference image, so it is justified to sample in this way. One significant source of contamination taking this approach is variable stars, therefore we remove all known variable stars from the random bogus component by cross-matching against the ATLAS Variable Star Catalogue (Heinze et al., 2018) with a 5 arcsec radius. These variable star detections can constitute 2–4% of the entire bogus dataset. Of the detections removed by this step, a small fraction of these will be high-amplitude variable stars which have a strong subtraction residual in a given night's data, and thus represent real sources lost. Automating the correct labelling of these sources using light curve information is feasible, but would add significant complexity and more potential failure modes, so we instead opt to remove the variable stars entirely and simply add more verifiably 'real' detections in their place in the form of more minor planets. Inevitably, some small fraction of uncatalogued variable stars will be missed

[4] https://www.minorplanetcenter.net/
[5] https://pypi.org/project/pympc/

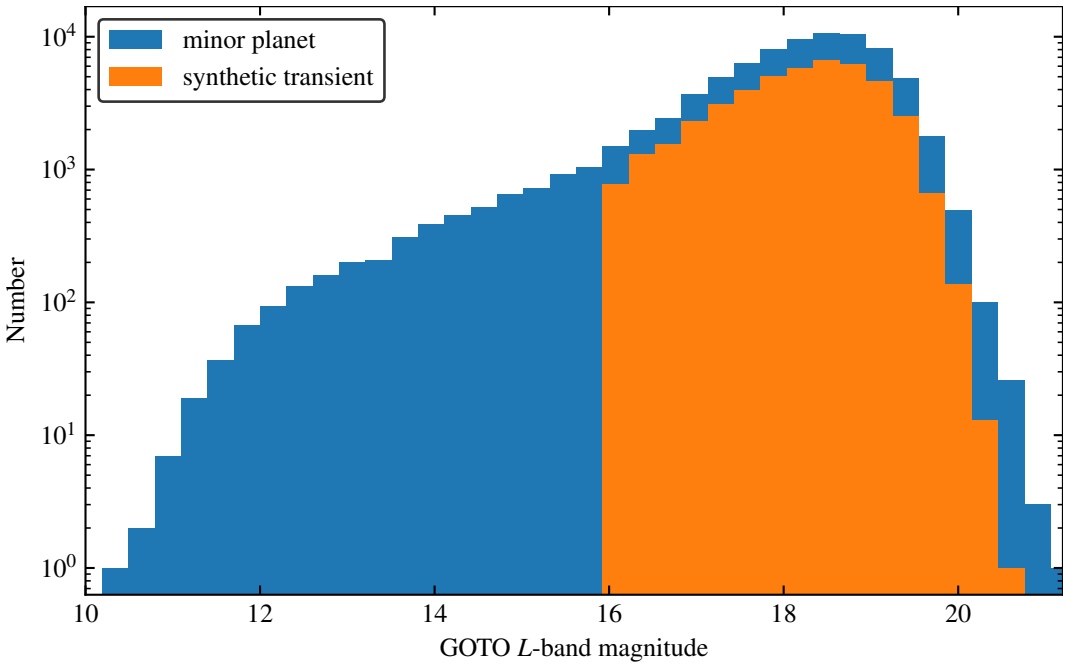

Figure 3.1: Magnitude distribution of the minor planets (MP) used to build our training set. Bright-end number densities are dominated by the true magnitude distribution of the minor planets, where the faint-end density is constrained by the GOTO limiting magnitude. The magnitude distribution of synthetic transients (SYN) is a sub-sample of the minor planet magnitude distribution, except with a cut at $L \sim$16, to avoid unrealistically bright objects.

with this procedure, and we develop tools to identify them retrospectively after model training in Section 3.3.3.

To improve the classifier's resistance to specific challenging subtypes of data poorly represented in our algorithmically generated training set, we inject human-labelled detections into the dataset. More specifically, candidates from the GOTO Marshall (discussed in full in Lyman et al., in prep.) are included, which were misidentified by the classifier in the pipeline at the time as real and later labelled as bogus by human vetters. The previous classifier was a rapidly-deployed prototype CNN similar in design to that presented here, trained on a smaller dataset of minor planets and random bogus detections. These detections are included to allow the classifier to screen out artifacts missed by the prototype image processing pipeline, including satellite trails and highly wind-shaken PSFs. This artifically increases the diversity of the bogus component of the training set, as these edge-case detections would rarely be selected by naive random sampling and so be poorly represented within the model. Although these detections represent a small fraction of the overall training set ($\sim$ 5%), they provide a marked improvement in performance in the real-world deployment of the classifier, including marginal gains on more typical detections.

3.2.1 Data extraction and format

For each detection identified for inclusion in our training/validation/test sets, a series of stamps are cut out from the larger GOTO image centred on the difference image residual. In common with previous CNN-based classifiers, we use small cutouts of the median-stacked science and template images, as well as the resultant difference image after image subtraction. The size of these stamps is an important model hyperparameter, which we explore in more detail in Section 3.3.1. A example of the model inputs for a synthetic source are illustrated in Figure 3.2.

An important addition to our network's inputs compared to previous work is a peak-to-peak (`p2p`) layer. This is included to characterise variability across the individual images that make up a median stacked science image, and is calculated as the peak-to-peak (maximum value - minimum value) variation of each pixel computed across all

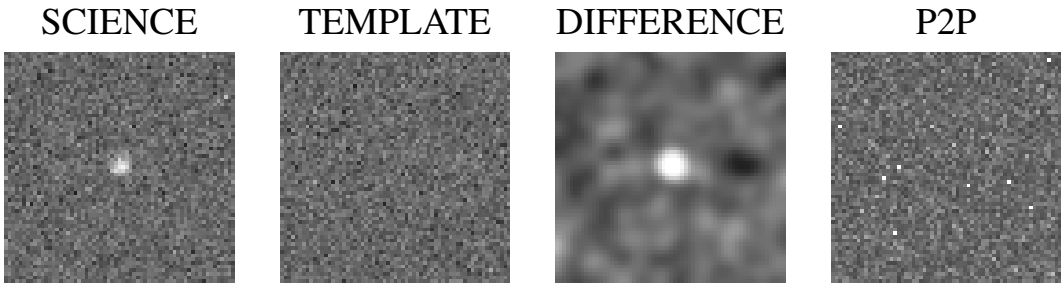

Figure 3.2: Example data format for a set of idealised synthetic images of a single Gaussian source newly appearing in the science image. We apply a naive convolution of science image with template PSF and vice versa in producing the difference image for visualisation purposes. From left to right: science median, template median, difference image, pixel-wise peak-to-peak variation across contributing images to science median. Cutouts are 55x55 pixels square, corresponding to a side length of 1.1 arcminutes.

individual images that composed the median stack. To ensure consistent alignment across all individual stamps and remove any jitter, we cut out the region based on the RA/Dec coordinates of the source detection in the median stack. This additional provides an effective discriminator for spurious transient events such as cosmic ray hits and satellite trails. If sufficiently bright, these are not removed by the simple median stacking in the current pipeline due to the small number of sub-frames used. This is particularly problematic for cosmic ray hits which are convolved with a Gaussian kernel for image subtraction, and appear PSF-like in the difference image. This can create convincing artifacts which are difficult to identify without access to the individual image level information. In testing, this reduced the false positive rate on the test set by $\sim$ 0.2%. Although this is not a sizeable improvement when evaluated on the full dataset, cosmic ray hits constitute a very small percentage of overall detections. Testing instead on a human-labelled set of bogus detections which were initially scored as real by the existing deployed classifier (without a $\mathrm{p2p}$ layer), there is a 2–3% decrease in false positive rate.

For all of the above steps, stamps extending beyond the edge of the detector have missing areas filled in with a constant intensity level of 10^{-6}, to distinguish them quantitatively from masked (i.e. saturated) pixels which are assigned a value of zero in the difference image by the pipeline. The specific intensity level chosen for this off-setting is not important, and we choose our value to be well above machine precision

(significant enough to influence the gradients) but well below the typical background level. To ensure that the classifier remains numerically stable in later training steps, each stack of stamps undergoes layer-wise L2 normalisation to reduce the input's magnitude. Each stamp has the mean subtracted and is then divided through by the L2 ($\sqrt{\vec{x} \cdot \vec{x}}$) norm.

3.2.2 Synthetic transients

Although asteroids provide a convenient source of PSF-like residuals to train on, it is important to note that they cannot fully replicate genuine transients. Asteroids are markedly simpler to learn and discriminate for a classifier since they lack the complex background of a host galaxy. The main goal of this classifier is to detect extragalactic transients, so adapting the training set to maximise performance on these objects is important. An ideal approach would be to add a large number of genuine transients into the training set. However, GOTO has not been on-sky long enough to collect a suitably large set of these detections, and we only build the training set from the previous year of data. Even assuming every supernova over the past year is robustly detected in our data this will still yield a number of transients that is significantly less than the target size of our training set. This would create a severely imbalanced dataset, which could in principle be used but with reduced classification performance. Using spectroscopically confirmed transients may also inject an element of observational bias into our training set, as events that have favourable properties for spectroscopy (in nearby galaxies, offset from their host, bright) are preferentially selected (Bloom et al., 2012) to be followed up. Instead we reserve a set of real, spectroscopically confirmed transients GOTO has detected (~ 900 as of August 2020) for benchmarking purposes, as they represent a valuable insight into real-world performance and can be used to directly evaluate the effectiveness of any transient augmentation scheme we employ, as in Section 3.4.2.

PSF injection has been used heavily in prior work to generate synthetic detections for testing recovery rates and simulating the feasibility of observations. This process can be computationally intensive, involving construction of an effective PSF (ePSF) from combining multiple isolated sources or fitting an approximating function

(e.g. a Gaussian) to sources in the image. The ePSF model can then be scaled and injected into to the image to simulate a new source. By injecting sources in close proximity to galaxies in individual images then propagating this through the data reduction pipeline, synthetic transients could be generated in a realistic way. However, the fast optical design of GOTO makes this a complex task, as the PSF varies as a function of source position on the detector. Sources in the corners of an image display mild coma, which, combined with wind-shake and other optical distortion, can lead to unusual PSFs that are not accurately reproduced by the mean PSF. In principle this could be accounted for by computing PSFs for sub-regions of a given image or assuming some spatially-varying kernel to fit for, but this would add sizeable overheads to the injection process and will always be an approximation.

Recent new techniques such as generative adversarial networks (GANs, Goodfellow et al. 2014) have shown promise in generating novel training examples that can be used to address class imbalances/scarcity in training sets (Mariani et al., 2018), and have recently started to be applied to astrophysical problems (Yip et al., 2019). However these networks are computationally expensive, complex to train and understand the outputs of, and don't fully remove the need for large datasets. A robust human-interpretable method for generating synthetic examples is a better approach for the noisy, diverse datasets used in real-bogus classification.

We propose a novel technique for synthesising realistic transients that can be used to significantly improve transient-specific performance when compared to a pure minor planet training set, without requiring PSF injection or other CPU-intensive approaches. For each minor planet detected in an image, the GLADE galaxy catalogue (Dálya et al., 2018) is queried for nearby galaxies within a set angular distance of 10 arcminutes, chosen such that the PSF of sources within this region are consistent. Pre-built indices are used via CATSHTM (Soumagnac & Ofek, 2018) to accelerate querying GLADE. The algorithm chooses the brightest galaxy (minimum B band magnitude) within range, then generates a cutout stamp with with a randomly chosen x, y offset relative to the galaxy centre. For the implementation within this work, the x, y pixel offsets are drawn from a uniform distribution $U(-7.7)$ chosen to fully cover the range

of offsets for nearby galaxies. Sources that are completely detached from any host galaxy are better represented by the minor planet component of the training set. This ensures that a diverse range of transient configurations (nuclear, offset, orphaned) are sampled. The minor planet and galaxy stamp are then directly summed to produce the synthetic transient. For the purposes of real-bogus classification, accurately matching the measured transient host-offset distribution is not crucial. The host offset distribution contains implicit and difficult to quantify biases resulting from the specific selection functions of the transient surveys that populate it – it does not reflect accurately the underlying distribution of astrophysical transients. By choosing from a uniform distribution, we instead aim to attain consistent performance across a wide range of host offsets that overlap with the range inferred from the transient host offset distribution.

The original individual images for each component are retrieved to correctly compute the peak-to-peak variation of the combined stamp. Model inputs are pre-processed and undergo L2 normalisation (as discussed in Section 3.2.1) prior to training and inference, so additional background flux introduced by this method does not affect the model inputs. The noise characteristic of this combined stamp is not straightforward to compute due to the highly correlated noise present in the difference image and varying intensity levels, and could be higher or lower depending on the specific stamps – with the straightforward Gaussian case providing a $\sqrt{2}$ reduction in noise. This is likely not problematic for the classifier, providing a form of regularisation that could improve generalisation accuracy. We also assume that the spatial gradients in background across both stamps are $\sim$ constant, as the stamp scale is far smaller than the overall frame scale – naturally this breaks down in the presence of nebulosity/galaxy light but this represents a overwhelmingly small fraction of the sky. We also reject all minor planets with $L < 16$, as these are significantly brighter than the selected host galaxy so are better represented by the pure minor planet candidates. This also cuts down significantly on saturated detections of dubious quality. This choice has no detrimental effect on bright-end performance, as discussed in Section 3.4. A random sample of synthetic transients generated with this approach is shown in Figure 3.3. Our method bears some similarity in retrospect to the approach of (Cabrera-Vives et al., 2017), who

added stamps from the science image into difference images to simulate detections in 'random' locations. Our approach uses confirmed difference image detections of MPs and puts them in more purposeful locations, whilst preserving the noise characteristics of the difference stamp.

This approach has strong advantages over simply injecting transients into galaxies. By selecting only galaxies close to each minor planet, the PSF is preserved and is consistent, regardless of how distorted it may be. Injection-based methods require estimation/assumption of the image PSF, which is typically a parameterised function determined by fitting isolated sources. Given the variation in PSF across images and across individual unit telescopes, this would be a computationally intensive task, and would likely lead to poorer results compared to using minor planets. However, using only these synthetic transients introduces unintended behaviour in the trained model that significantly degrades classification performance if not remedied. Since every synthetic transient in the training set is associated with a host galaxy by design, the model will over time learn to associate all detections with galaxies as being real as there is no loss penalty for doing so. To resolve this, we also inject galaxy residuals as bogus detections, randomly sampling from the remaining GLADE catalog matches at a 1:1 transient:galaxy residual ratio. This way, the model learns that the salient features of these detections are not the galaxy, but the PSF-like detection embedded in them.

3.2.3 Training set construction

Using the techniques developed in the sections above, we build our training set with GOTO prototype data from 01-01-2019 to 01-01-2020. This ensures that our performance generalises well across a range of possible conditions – with PSF shape and limiting magnitude being the most important properties that benefit from this randomisation. A breakdown of training set proportions and properties is given in Table 3.1.

Our code is fully parallelised at image level, meaning that a full training set of $\sim$400,000 items can be constructed in under 24h on a 32-core machine. Training sets can also be easily accumulated on multiple machines and then combined thanks to the use of the HDF5 file format. The main bottlenecks of training set generation are

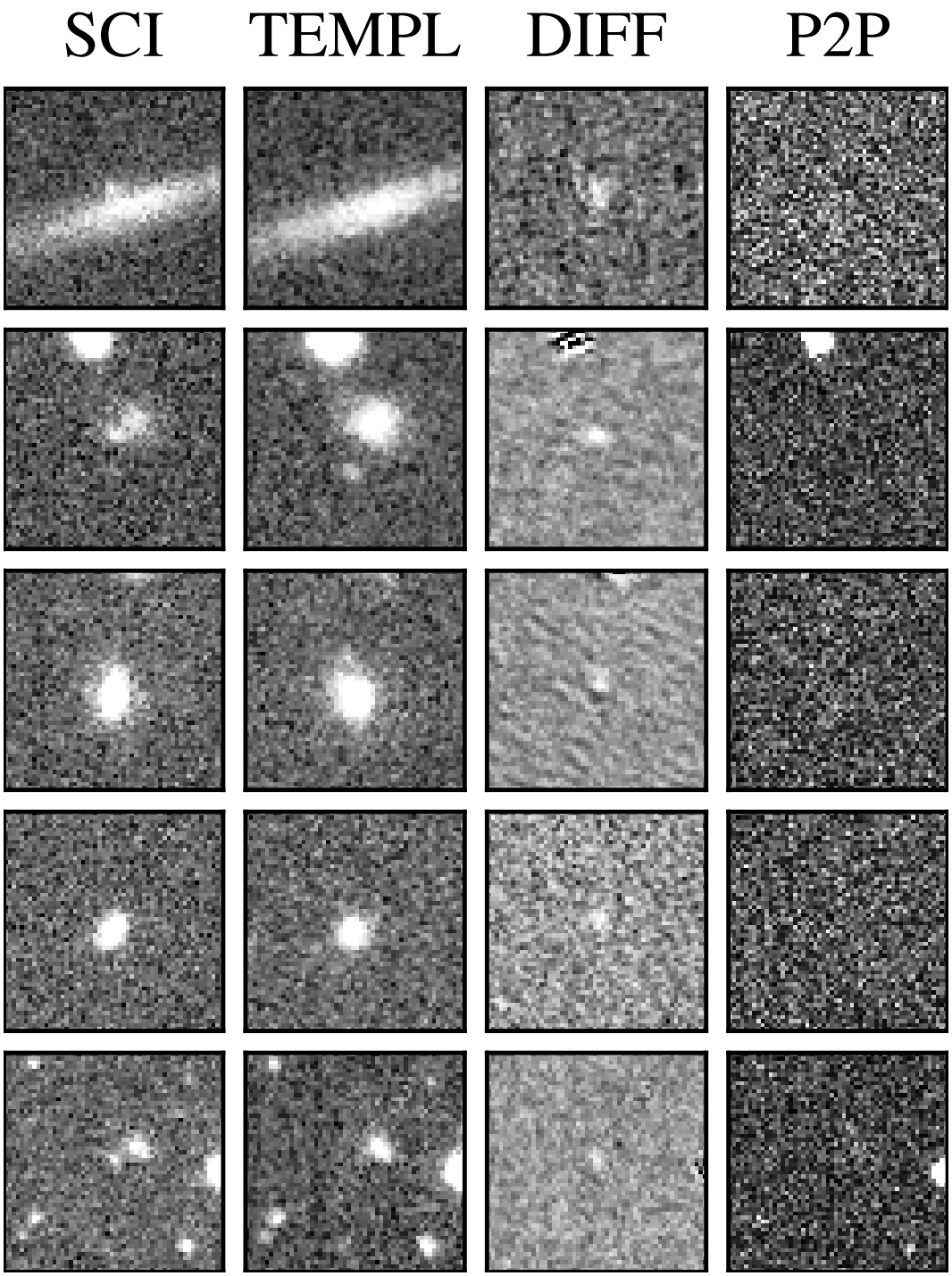

Figure 3.3: Randomly selected sample of synthetic transients generated with our algorithm, displayed in the same format as in Figure 3.2. Significant variations in the PSF are visible due to sampling directly from the image, improving classifier resilience.

Metalabel	Train	Test	
Minor planet	72992	8133	
Synthetic transient	40192	4521	
Random bogus	177556	19645	
Galaxy residual	28040	3190	
Marshall bogus	24577	2662	
Total	343357	38151	**381508**

Table 3.1: Breakdown of the composition of our dataset, partitioned according to training and test sets. The validation dataset is not shown, but is composed of 10% of the training dataset, chosen randomly at training time.

IO-related – loading in image data to prepare the stamps, and querying the GLADE catalogue and SkyBoT cone search.

3.3 Classifier architecture

As a starting point, we follow the BRAAI classifier of Duev et al. (2019) in using a downsized version of the VGG16 CNN architecture of Simonyan & Zisserman (2014). This network architecture has proven to be very capable across a variety of machine learning tasks, and is a relatively simple architecture to implement and tweak. This architecture uses conv-conv-pool blocks as the primary component – two convolutions are applied in sequence to extract both simple and compound features, then the resultant feature map is reduced in size by a factor 2 by 'pooling', taking the maximum value of each 2x2 group of pixels. This architecture also uses small kernels (3x3) for performance. These structures are illustrated in Figure 3.4. We use the configuration as presented in Duev et al. (2019) for development, but later conduct a large-scale hyperparameter search to fine-tune the performance to our specific dataset (Section 3.3.1). The primary inputs to the classifier are small cut-outs of the science, template, difference, and `p2p` images as discussed in Section 3.2.1 which we refer to as stamps.

The sample weights for real and bogus examples are adjusted to account for the class imbalance in our dataset, set to the reciprocal of the number of examples with each label. Class weights are not adjusted on a per-batch basis, as our training set is only mildly imbalanced. For regularisation, we apply a penalty to the loss based on

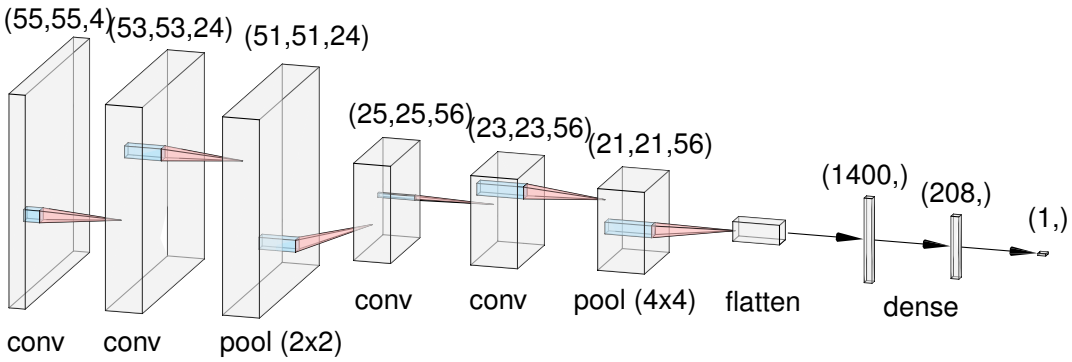

Figure 3.4: Block schematic of the optimal neural network architecture found by hyper-parameter optimisation in Section 3.3.1. Each block here represents a 3D image tensor, either as input to the network, or the product of a convolution operation generating an 'activation map'. Classification is performed using the scalar output of the neural network. Directly above each 3D tensor block the dimensions in pixels are shown, along with the operation that generates the next block below it represented by the coloured arrow. Not illustrated for clarity here are the dropout masks applied between each layer and the activation layers. Base figure produced with NNSVG (LeNail, 2019).

the L2 norm of each weight matrix. This penalises exploding gradients and promotes stability in the training phase. L1 regularisation was trialled but did not produce significantly better results. We also use spatial dropout (Tompson et al., 2015) between all convolutions which provides some regularisation, but primarily is used for the purposes of uncertainty estimation (see Section 3.3.3) – a small dropout probability of ~ 0.01 is found to be optimal from work in Section 3.3.1. Due to the significant training set size and our use of augmentation, very little regularisation is needed for a model of this (comparatively) low complexity.

To further increase the effective size of our training set we randomly augment training examples with horizontal and vertical flips, which provide a factor 4 increase in effective training set size over unaugmented stamps. We also trialled the usage of 90 degree rotations following (Dieleman et al., 2015), which do not require interpolations and thus do not introduce spurious artifacts that could add additional learning complexity. In constrast to other works (Cabrera-Vives et al., 2017; Reyes et al., 2018), we find consistent performance (over multiple training runs) with simple reflections – potentially having already reached the saturation region of the learning curve.

Our model is implemented with the KERAS framework (Chollet et al., 2015),

running with an optimised build of the TENSORFLOW backend (Abadi et al., 2015). For parameter optimisation we use the ADAM optimiser of Kingma & Ba (2014), which provides reliable convergence, and use the binary cross-entropy as the loss function. To avoid overfitting, we utilise an early stopping criterion conditioned on the validation dataset loss — if there has been no decrease in validation loss within 10 epochs, the model training is terminated. We perform model training and inferencing on CPU only, to mirror the deployment architecture used in the main GOTO pipeline. Using a single 32-core compute node, training the finalised model to early-stopping at $\sim$170 epochs takes around 10 hours. Inferencing is significantly quicker, with an average throughput of 7,500 candidates per second with no model ensembling performed. Our model training code is freely available via the `gotorb` Python package [6], which includes the full range of tunable parameters and model optimisations we implement.

3.3.1 Tuning of hyperparameters/training set composition

To achieve the maximum performance possible with a given neural network, we conduct a search over the model hyperparameters to assess which combinations lead to the best classification accuracy and model throughput. Initially the ROC-AUC score (Fawcett, 2006) was used as the metric to optimise as in many cases this is a more indicative performance metric than others, however this did not translate directly to improvements in classification performance. We conjecture this may be due to the score-invariant nature of the ROC-AUC statistic – it only captures the probability that a randomly selected real example will rank higher than a randomly selected bogus example, which is independent of the specific real-bogus threshold chosen. We instead opt to use the accuracy score, as this directly maps to the quantity we want to maximise in our model.

Data-based hyperparameters (ratio of real-to-bogus examples, stamp size, data augmentation) are optimised iteratively by hand due to computational constraints. An approximate real-bogus ratio between 1:2 to 1:3 was found to be optimal, with greater values giving better bogus performance at the cost of recovery of real detections – we opt for 1:2 in the final dataset. The overall dataset size was found to be the biggest

[6] https://github.com/GOTO-OBS/gotorb

determinant of classification accuracy, with larger datasets showing improved performance – although this increase was subject to diminishing returns with larger and larger datasets. We chose a training set of O(4×10^5) examples, as this was roughly the largest dataset we could fit into RAM on training nodes – naturally this could be increased further by reading data from disk on demand, but given CPUs were used for training there was a need to minimise input pipeline latencies as much as possible to compensate. Model performance was found to be relatively insensitive to the ratio of synthetic transients to minor planets, as long as there were at least 10,000 of both in the training set. Using a dataset where 100% of the real content came from minor planets led to a $\sim 5\%$ drop in the recovery rate of transients on the test set (see Fig. 3.11), whereas a 100% synthetic transient dataset led to a detrimental 15% decrease in the recovery rate of minor planets, and a 5% drop on the transient test set. This surprising result implies that combining both minor planets and synthetic transients has a synergistic effect, with the combination providing better performance overall. The specific composition of the final dataset is listed in Table 3.1, we found a roughly 2:1 minor planet:synthetic transient ratio to provide the correct balance between overall test set performance and sensitivity to astrophysical transients.

A key parameter explored as part of this study is the input stamp size. Larger stamps take longer to generate and more time to perform inference on, so identifying the minimum stamp size possible without affecting performance is crucial. In Figure 3.5 we show the results of training identical models on an identical 330k-example dataset, with varying stamp size between 21 and 63 pixels. We find that there is no significant increase in performance for our training dataset beyond a stamp size of 55 pixels. The upper limit of this search was set by available RAM, and took 118 hours of compute time to complete. When scaled through by the ratio of the GOTO/ZTF plate scales (1.4x), our best value of 55 pixels appears remarkably consistent with the 63 pixel stamps that Duev et al. (2019) found optimal for their network. This is an interesting result, and could imply that the angular scale is actually the more relevant parameter – this might represent some characteristic length scale that encodes the optimal amount of information about the candidate and surrounding context without including too much

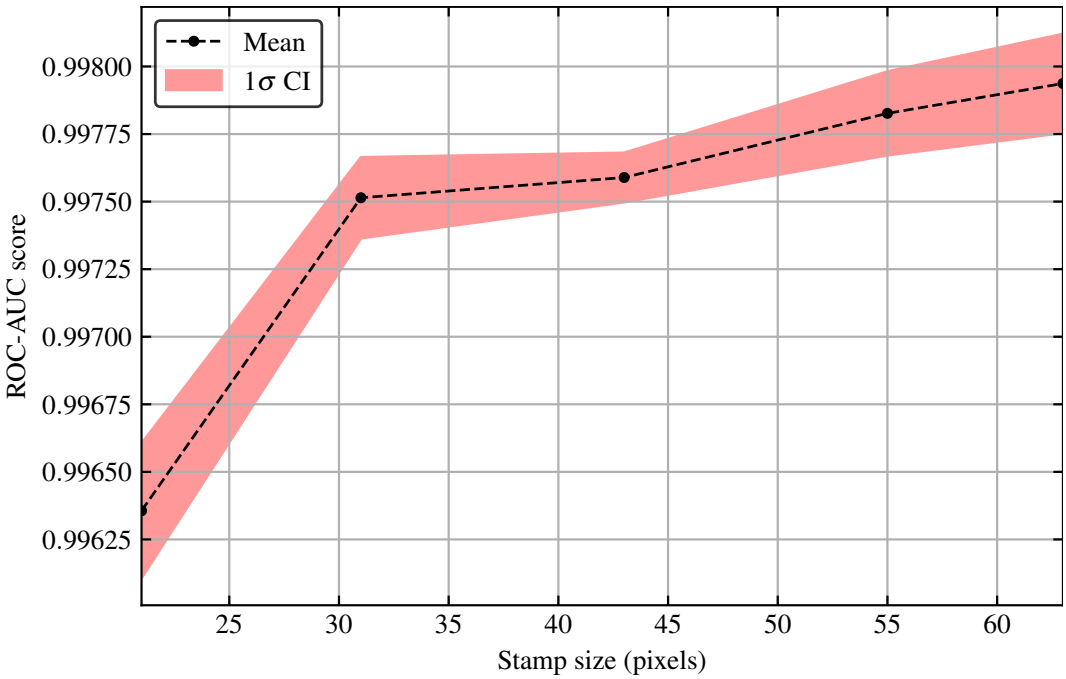

Figure 3.5: Classifier performance on the test set of a 330,000 example training set as a function of input stamp size. Each point is the average of 3 independent training runs on the same input training set, with the shaded region representing the 1σ confidence interval.

irrelevant data.

Network hyperparameters are optimised using the `Hyperband` algorithm (Li et al., 2017) as implemented in the Keras-Tuner package (O'Malley et al., 2019). This algorithm implements a random search, with intelligent allocation of computational resources by partially training brackets of candidate models and only selecting the best fraction of each bracket to continue training. In testing, this consistently outperformed both naive random search and Bayesian optimisation in terms of final performance. Table 3.2 illustrates the region of (hyper)parameter space we choose to conduct our search over. The upper limits for the neuron/filter parameters are set by purely computational constraints – networks above this threshold take too long to evaluate and train, and so are excluded. We also set an upper limit of 500,000 on the number of model parameters to avoid overly complex models and promote small but efficient architectures. Based on initial experimentation, we require the number of convolutional filters in the second block must be greater than or equal to the number in the first block. This ensures that the largest (and most computationally expensive) convolution operations are performed on tensors that have been max-pooled and thus are smaller, reducing execution time. To maximise performance across all possible deployment architectures, the number of convolutional filters and fully-connected layer neurons are constrained to be a multiple of 8. This is one of the requirements for fully leveraging optimised GPU libraries (such as cuDNN, Chetlur et al. 2014), and also enables use of specialised hardware accelerators such as tensor cores in the future. Conveniently, this discretisation also makes the hyperparameter space more tractable to explore.

This search took around 1 month to complete on a single 32-core compute node, and sampled 828 unique parameter configurations. The three top-scoring models were then retrained from random initialisation through to early stopping to validate their performance, and confirm that the hyperparameter combination led to stable and consistent results. The top three scoring models achieved accuracies on the hyperparameter validation set of 98.88, 98.64 and 98.54% respectively. Some of the candidate models had to be pruned from the list due to excessive overfitting. The best model was then selected based on the minimum test set error. Our final model achieved a test set

continuous Hyperparameter	Min	Max	Prior	Selected
Block 1 filters (N_1)	8	32	linear	24
Block 2 filters (N_2)	N_1	64	linear	56
N_{fc}	64	512	linear	208
Dropout rate	10^{-2}	0.5	log	5.2×10^{-2}
Learning rate	10^{-5}	10^{-2}	log	6×10^{-5}
Regulariser penalty	10^{-8}	10^{-2}	log	2.0×10^{-8}

discrete Hyperparameter	Choice	Selected
Kernel initialiser	He, Glorot	Glorot
Kernel regulariser	L1, L2	L2
Activation function	ReLU, LeakyReLU, ELU	LeakyReLU

Table 3.2: Hyperparameter space over which the optimisation search was conducted, split by numerical and categorical variables. The final adopted values are given in the rightmost column.

class-balanced accuracy of 98.72 ± 0.02% (F1 score 0.9826 ± 0.0003), with the selected hyperparameters listed in Table 3.2. This outperforms the version human-tuned by the authors through iterative improvement by 0.6%, and trains to convergence in around half the number of epochs. We adopt this model architecture going forward, and characterise the overall performance in greater detail in Section 3.4. For this final model, the theoretical maximum ROC-AUC is obtained when the real-bogus threshold is set to 0.4, although in live deployment we opt for a conservative value of 0.8 to minimise contamination.

3.3.2 Quantifying classification uncertainty

Uncertainty estimation in neural networks is an open problem, but is of critical importance for a range of applications. Traditional deterministic neural networks output a single score per class between 0 and 1. This single value would be sufficient to provide a measure of confidence, if properly calibrated. However, neural networks are often regarded as providing over-confident predictions in general, and, worse, providing misidentifications at high confidence. Giving neural networks the ability to make nuanced predictions and account for their own uncertainty in decision making is a po-

tentially powerful improvement, that we discuss in more detail over the next sections.

It is important to be specific and distinguish between epistemic (systematic) and aleatoric (random) uncertainty for the purposes of our classification problem (Kendall & Gal, 2017). Aleatoric uncertainty is captured by the classifier's score value, and originates from noise in the input data. More crucial for our application is quantifying the epistemic uncertainty – that is the uncertainty in our choice of neural network's model weights. This epistemic source of error is directly quantifiable through Bayesian neural networks, and in later sections this is the error, confidence, or uncertainty we refer to and attempt to quantify. In the Bayesian framework, this can be achieved by casting model parameters as probability distributions, and using the mechanics of Bayesian statistics to marginalise the neural network output over these distributions, in the process finding the score posterior. In this way, the uncertainty inherent in model selection can be quantified. There are various approximate and exact approaches to achieve this which we outline below.

Dropout (Srivastava et al., 2014) provides a useful form of regularisation in neural networks. At each training step, a fraction p (a tunable hyperparameter) of neuron weights are randomly set to zero, decreasing the effective number of parameters of the model. In this way, overfitting can be prevented and generalisation accuracy can be increased. In traditional neural networks, dropout is not active at inference time so that all neurons are used for making predictions. However, Gal & Ghahramani (2015a) demonstrate the profound result that training and evaluating neural networks with dropout is equivalent to performing the approximate Bayesian inference discussed above, with multiple evaluations being equivalent to Monte Carlo integration of the posterior distribution. This is directly applicable to convolutional neural networks, via the Monte Carlo dropout technique (Gal & Ghahramani 2015b; referred to as MCDropout for brevity from now on).

Alternative approaches to uncertainty estimation exist (Bayes by Backprop, Blundell et al. 2015), which instead directly performs the approximate Bayesian inference by instead casting neuron weights as distributions with associated hyperparameters, then updating these according to the backpropagated gradients (like deterministic NNs). In

this work, we opt to use MCDropout for computational efficiency and for maximal compatibility with existing network architectures and software. No changes to the training loop are required, and only a simple wrapper is required at inference to perform multiple predictions with dropout enabled. The only significant additional computational cost for a Bayesian neural network using the MCDropout technique over a deterministic CNN is at inference time, as multiple samples need to be drawn to approximate the posterior. This performance overhead can be mitigated with suitable batching of the dataset. The ability of neural networks to learn complex, non-linear representations in high-dimensional vector spaces is well-known and utilised throughout machine learning. However, estimation of the uncertainty of products of neural networks is often a barrier to their implementation in scientific applications, where well-grounded determination of errors is important. MCDropout provides a principled way to introduce this.

Although a comparatively new technique, Bayesian neural networks show emerging promise across a variety of astronomical classification and regression tasks – including supernova light curve classification (Möller & de Boissière, 2020), efficient learning of galaxy morphology (Walmsley et al., 2020), and age estimation of stars for galactic archaeology (Ciucă et al., 2020).

There is disagreement in the literature on the precise nature of a Bayesian neural network and how to implement it 'properly', from approximate variational inference as used here, to applying some variant of the Markov Chain Monte Carlo sampler over the weight and bias parameters of the neural network. However, what is relevant for the implementation in this work is that examples the classifier is unconfident about are assigned lower confidence scores than obviously real/bogus detections. More complex tests, such as confirming that the classifier's confidence matches the actual confidence of the dataset/some human-derived uncertainty score are beyond the scope of the introductory work presented here.

Whilst these posterior predictions are informative to human vetters, converting them to a single informative summary parameter that captures the overall uncertainty is more useful for integration into pipelines and enabling coarse filtering of candidates. To convert the posterior distributions to meaningful information about the confidence

of a given prediction, we utilise the information entropy $\mathbb{H}$. For a binary classification problem, the generic entropy formula can be reduced to:

$$\mathbb{H}(p) = -p\log_2 p - (1-p)\log_2 1 - p$$

where p is the probability of a given detection being real (the real-bogus score). The entropy is maximised for $p = 0.5$, where the probability of being real vs. bogus is equal, or the classifier prediction carries no useful information. We define the classifier confidence $\mathbb{C}$ in terms of the average entropy of the posterior distribution samples, scaling to confidences in the range $[0, 1]$ with the relation

$$\mathbb{C} = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbb{H}_i$$

where N is the number of posterior samples and $\mathbb{H}_i$ is the binary entropy of the i^{th} posterior sample. This metric is equivalent the second term of the BALD acquisition function of Houlsby et al. (2011), and is chosen as it is pre-normalised to $[0, 1]$ unlike standard deviation or similar metrics. Naturally the uncertainties we derive here are correlated with the actual output score, but the multiple samples provide sufficient dispersion that this metric is useful to assess model confidence. In future implementations, these raw posterior samples (or some approximating distribution parameters to reduce data needs) could be fed directly into downstream, more specialised classification tools to enable them to make use of the real-bogus classifier's probabilistic predictions in their own score/posterior.

3.3.3 Using the uncertainty in classifier predictions

One immediate advantage of Bayesian neural networks over deterministic neural networks is the ability to improve classification performance through model ensembling. Figure 3.6 illustrates the gain in accuracy observed by averaging the predictions of our BNN, as a function of the number of posterior samples. Although small, this is a definite improvement over single-evaluation predictions, and is likely constrained by our

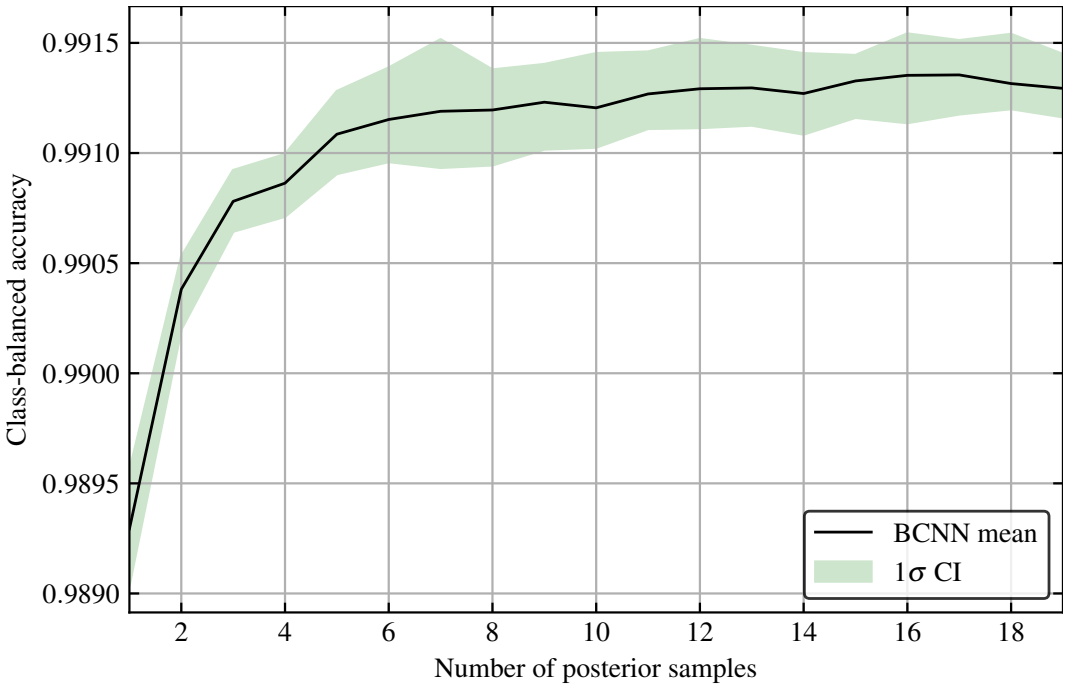

Figure 3.6: Classification accuracy on the test set from Section 3.2.3 as a function of the number of posterior samples averaged. Each point is the average of 10 model runs, with the shaded area corresponding to the 2σ confidence interval. The BCNN quickly recovers the performance of a deterministic CNN within statistical uncertainty (99.18 ± 0.03% accuracy, F1: 0.9877) and provides additional information in the form of confidence. No significant improvement in classification accuracy is obtained beyond 10 samples, remaining consistent out to 50 samples.

dataset. For the majority of positive and negative examples the model is highly confident about the assigned RB score, so averaging over the posteriors does not improve them significantly. This increase in performance is likely to be greater on more complex (multi-class) classification problems, or scenarios where significantly less training data is available.

Posteriors and/or associated confidence scores can be added to any downstream candidate evaluation tools, providing an additional metric to inform decisions. Objects with both high score and high confidence are highly likely to be genuine, so can be prioritised in human vetting of candidates. This means more time can be spent looking at more marginal candidates, and obvious detections can quickly be identified. Confidence provides a complementary metric to the pure real-bogus score that can

help alleviate some of the issues with the poor dynamic range observed in the classifier outputs at low/high scores. Classification is still performed on the consensus real-bogus score derived from the posterior, with the confidence score intended to aid human decision making. In Figure 3.7, we illustrate some example candidates, their associated real-bogus score, and the score posterior.

Classifier confidence is also a useful tool for the training and development process, providing deeper insight into the functioning of the classifier and the associated training set. Predictive uncertainty provides a useful heuristic to clean datasets of mislabelled data. Misclassified detections that the classifier returns a high confidence for are very likely to be mislabelled, as the confidence score is partially based on seeing large numbers of similar detections in the training set. These frames can be actively prioritised in any human relabelling efforts, or fixed cuts on the confidence can be utilised to perform this in a semi-automated way. This 'optimal relabelling' scheme provides a method for human vetters and machine learning models to collaboratively and iteratively refine noisy labels. Our label noise is introduced as humans are imperfect judges of real/bogus, and interpret the vetting rubric in different ways leading to inconsistencies which can harm model performance.

We demonstrate the effectiveness of this procedure on the training set built in this work by training the model first on the uncleaned dataset, then attempt to relabel the misclassified detections in the training and test set ordered by decreasing confidence. This amounts to a substantial task of 3580 stamps, which would take a prohibitively long time to relabel by hand, notwithstanding the possibility of human bias in the relabelling. We instead here propose a heuristic re-labelling scheme based on the BALD score of Houlsby et al. (2011) that leverages the simplistic nature of binary classification.

The model is first trained on the 'unclean' dataset generated with the approaches in Section 3.2.3, then the BALD score is evaluated over the misidentifications in the test and training sets. From here, a new set of labels is derived by flipping the labels of those examples that have a BALD score less than (thus confidence higher than) the median – effectively accepting the prediction of the classifier over the human vetter. This approach is naturally capable of flipping the labels of accurately labelled stamps

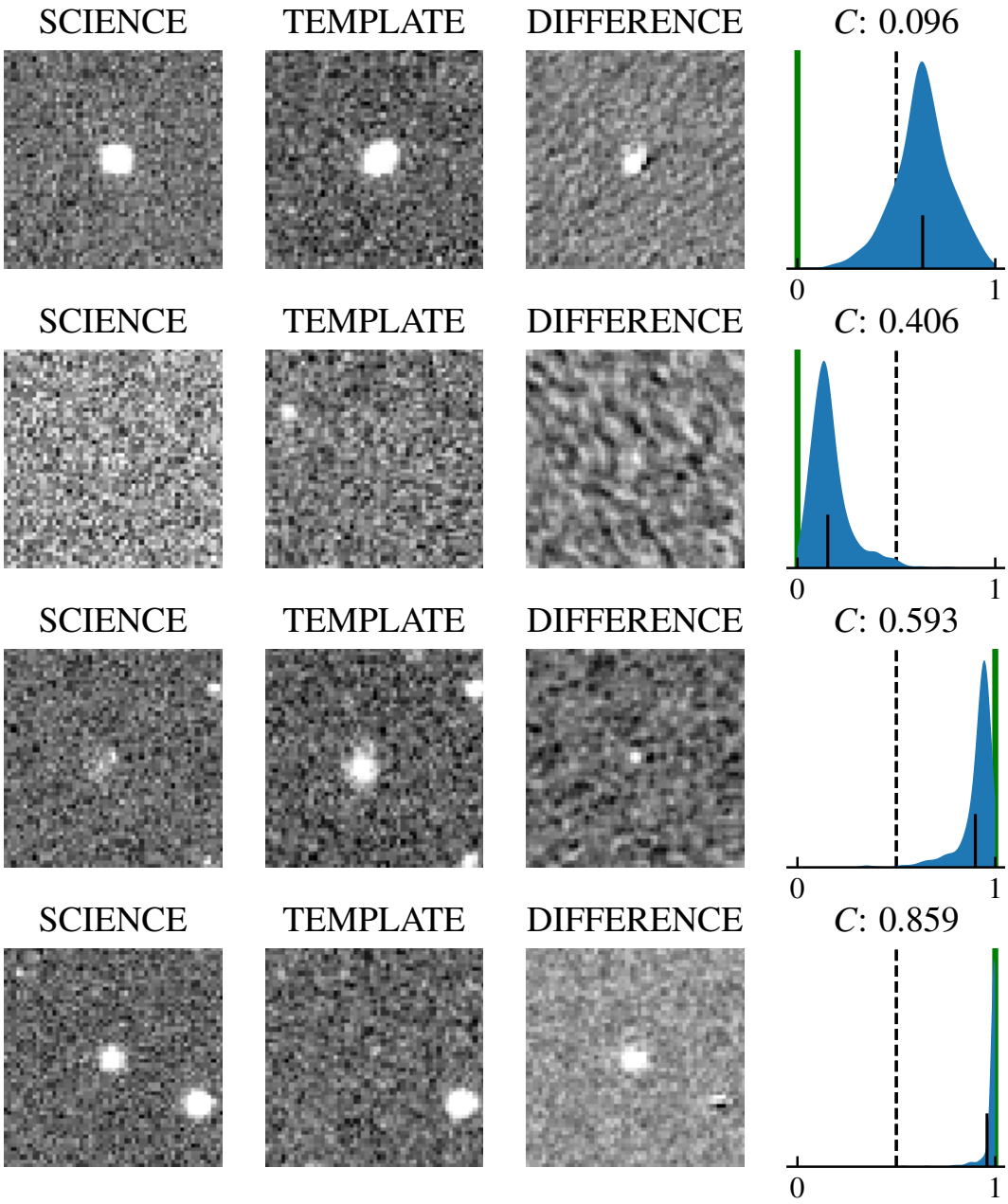

Figure 3.7: A selection of example posteriors, taken from real GOTO data. The majority of predictions are highly confident, so we select examples of increasing confidence score ($\mathbb{C}$) to display here. Plotted here is a Gaussian kernel-density estimate constructed from 500 posterior samples. The green line indicates the correct label for each candidate, with the black line indicating the mean of the distribution. The dashed line indicates $P_{\text{real}} = 0.5$

.

incorrectly, but by imposing this cut in classifier confidence it ensures that the majority of relabelled stamps each round correspond to regions of classifier parameter space that are well-covered by the training set and so are classified at high confidence. This method effectively trades active human labelling time for passive background computational time, and can be applied iteratively as suggested above to progressively improve the quality of the dataset labelling. We manually checked a subset of the sources selected to be re-labelled to verify these were sensible and indeed found they were mislabelled detections that had leaked through the quality cuts we applied. After 1 round of the heuristic relabelling routine outlined above, the class-balanced accuracy achieved on the classifier test set improved markedly from 98.72 ± 0.02 to $99.12 \pm 0.01\%$ (F1 score: 0.9826 ± 0.0003 to 0.9877 ± 0.0002), demonstrating the efficacy of this approach. We adopt this cleaned dataset for the following sections.

When visualised in an intuitive way, this confidence score can provide insights into the specific families of detection that the classifier is uncertain about. A natural approach to combine this with is to examine the latent space of the neural network. The first convolutional stage of our network can be thought of as a feature extractor, with the resultant feature vector encoding high-level information about the morphological characteristics of our dataset, providing insight about potential groupings of detection types through clusterings in this space. To explore the latent space within our model, we apply t-stochastic neighbour embedding (t-SNE, Maaten & Hinton 2008) to the output vector of the layer prior to the fully-connected classification layer to reduce the dimensionality and identify clusterings of common data points. The combined process projects an 800-dimensional vector space down to (in our case) a 2D plane. In this space, points with similar latent vectors appear close to each other, thus providing a clustering of the latent space which can be used to visualise the internals of the neural network. This is a purely diagnostic clustering for visualisation purposes, as t-SNE does not preserve global distances, nor does it provide a bidirectional mapping from the compressed latent space to the full latent vector space. Figure 3.8 illustrates this technique applied to the test set, coloured by both detection sub-class (left) and classifier confidence (right).

A useful insight this compressed space provides is the ability to identify clusters

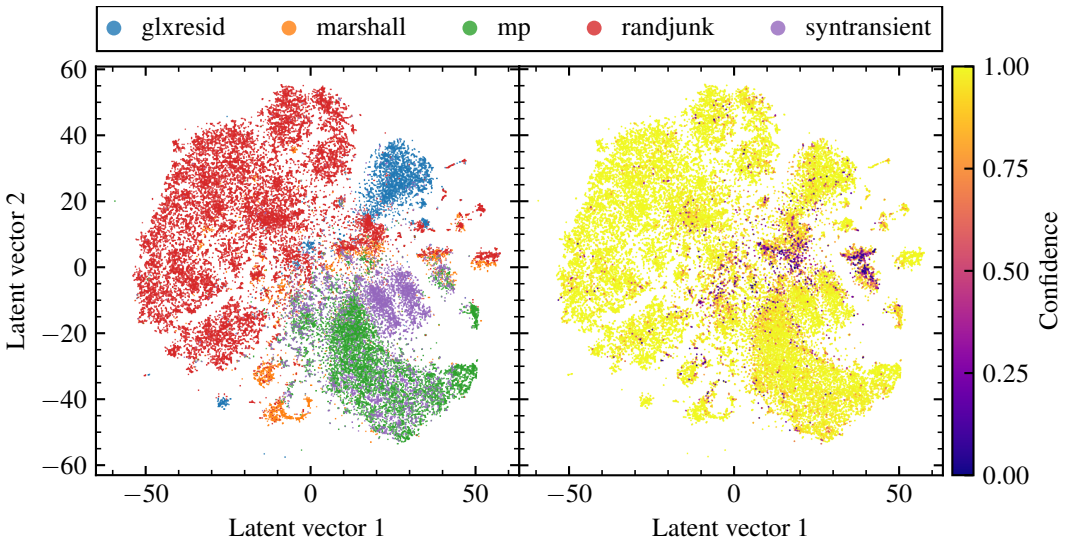

Figure 3.8: Example class-clustering (left) and confidence (right) maps generated from the classifier's test set. Each colour in the left panel represents a specific sub-class of detections, where colour on the right represents classifier confidence. The top legend gives the classes corresponding to each colour in the left panel. Regions of low confidence in the right panel tend to correspond to cluster boundaries in the left, where there is more uncertainty about which class each example belongs to.

of low-confidence points. This immediately reveals types of detection where the classifier may be uncertain, due to intrinsic difficulty of classification (sources close to the detection limit, nuclear transients, unusual PSFs), or scarcity of training data in general. The fact that there are clear divisions between the coloured sub-classes in the left panel of Figure 3.8 implies that the classifier has learned something about the intrinsic morphology of the detections beyond simple real-bogus division. Neither the classifier nor the clusterer receive these higher-level metalabels, so the clear partitions between the subclasses is purely a result of the internal representations learned.

For more complex datasets where the labelling budget for training examples is limited, Bayesian neural networks enable active learning – a process where the model identifies input data from a large unlabelled pool that would provide the greatest gain in information to it, using the uncertainty. This has been applied to convolutional neural networks with great success (Gal et al., 2017), and is likely a useful tool for fine-tuning existing training sets in light of new data. We trialled Bayesian active learning as a tool to build the training set presented in this work using the BALD score (Houlsby et al.,

2011) as our acquisition function, although it showed no significant improvement over a random selection from the unlabelled pool. This is likely due to the formulation of our classification problem – using only binary labels, and our data being dominated by large numbers of high-confidence real and bogus examples – only rare examples which add little to the overall classification accuracy are acquired. The additional complexity introduced by a multi-class labelling scheme along with the greater entropy provided by having multiple output neurons would likely yield better results.

3.4 Evaluation of classifier performance

Machine learning algorithms acquire inherent and often subtle biases based on the training set used in their construction. Given the automated nature of our data set generation, it is particularly important to verify that performance is consistent across a range of parameters of interest, such as transient magnitude. Some care is required in choosing the test set for evaluating classifier performance in a real-world setting, as the training set has been augmented with both human-labelled data and fully synthetic data. Although a low FPR/FNR on the validation and test data is encouraging as it is artificially made more difficult for the classifier to learn, it is not directly representative of the performance we should expect in deployment as a non-negligible component of it is synthetic. Performance characterisation should be reinforced with extensive testing on representative samples of GOTO data. A particular focus is to confirm that the synthetic augmentation scheme we implement leads to genuine improvements in the classifier's recovery rate of real transient detections. We also emphasise that in following sections, we effectively test the performance of the real-bogus classifier in isolation – the 'real-world' detection efficiency is the product of the efficiency of multiple pipeline stages, most crucially image subtraction and source extraction. Exploring the impact of these steps is beyond the scope of this paper, and thus are left to future work.

In the following sections, we use 'accuracy' to refer to the class-balanced accuracy, as it is more appropriate for our mildly imbalanced classification task. We also quote results based on the mean scores of 10 posterior samples (motivated by

the saturation observed in Figure 3.6) since individual evaluations of a Bayesian neural network using MCDropout are based on weaker classifiers due to the presence of dropout. Typical uncertainties (estimated as the standard deviation) on the metrics below are $< 0.05\%$, largely arising from the small number of examples around the decision boundary – where uncertainties exceed this they are given explicitly.

3.4.1 Performance on the test set

To provide a more granular view of the classifier performance, we further split the test set into two groups for the purposes of evaluation. The first comprises of only the minor planet and random bogus detections. We also test a synthetic transient/galaxy residual test set, to verify that the classifier can genuinely discriminate between galaxies and galaxies with transients. This also reveals any strong performance differences between the two main positive classes, which could skew metrics evaluated on the whole dataset. For both test sets, the human-inspected Marshall data is deliberately excluded, since it is significantly more challenging for the classifier than normal detections and does not accurately reflect the true data distribution.

The best-scoring classifier after hyperparameter optimisation shows excellent performance, attaining balanced accuracies of 99.49% (F1: 0.9935) and 99.19% (F1: 0.9925) on the minor planet and synthetic transient test datasets respectively. Figure 3.9 illustrates the false positive and negative rates for the classifier on both the minor planet and transient datasets, as a function of the real-bogus threshold chosen. There is a clear difference in false negative rate between the minor planet and transient datasets, reflecting the increased difficulty associated with the complex host morphology associated with the transient examples. The classifier displays a notable skew in the FPR/FNR equality point towards lower values. This is a result of the Marshall injections in the training set, which are made more difficult to learn than the random bogus detections due to being misclassified by the previous classifier. This does not affect classification accuracy, and could be fixed by applying a power transform to the classifier output if required, conditioned on the validation set.

Given the spatially-variable optical characteristics present in the GOTO proto-

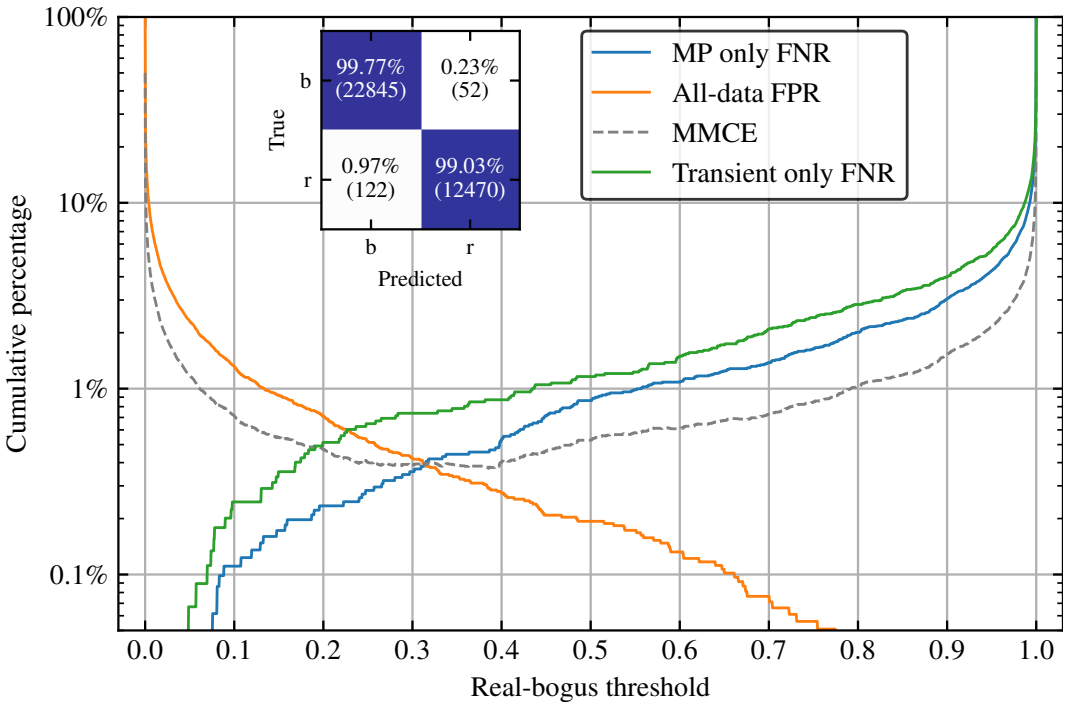

Figure 3.9: False positive/negative rate evaluated on the test set, excluding Marshall examples. Performance metrics are split based on minor planet and synthetic transients. The grey dashed line (MMCE) represents the full-dataset mean misclassification error, which is below 1% between real-bogus scores of 0.1 – 0.6. Inset: confusion matrix, evaluated on the full test set. There is a slight difference in the false negative rates achieved between the minor planets and synthetic transients, reflecting the increased difficulty posed by complex host morphology and subtraction residuals.

type, it is important to confirm that our classifier provides good performance across the full detector – and not simply in the centre where distortion is minimal. In Figure 3.10 we plot the class-balanced accuracy score as a function of radial position on the detector, using a series of radial bins chosen to equalise source density. These radial bins are scaled through by the maximum value (corresponding to the image corner) to provide a scale-free measurement of detector position. Class-balanced accuracy is used here as the real-bogus fraction varies as a function of detector position, and care must be taken to account for this. We find a consistent performance of $\sim$99% out to a fractional radial distance of 0.7, with a slight drop of 1% out at the far edge of the image. This is primarily due to the severe distortion found in the image corners of the GOTO prototype optical tubes, which produces very challenging detections (abnormal PSFs, strong vignetting) both for source extraction and real-bogus classification. Some contribution to this performance decrease is likely from good quality sources close to the edge of the image or close to the edge of the science-template overlap. Estimating reliably these sources and their contribution to the numbers in each bin is a complex task. Suffering only a 1% decrease in performance in these extremely challenging conditions demonstrates the overall robustness of the classifier. With the significantly improved optical quality of the GOTO design specification OTAs, we anticipate that future versions of our classifier trained on data from the upgraded system will display a constant (within statistical error) classification accuracy as a function of detector position.

3.4.2 Performance on spectroscopically confirmed transients

To provide the most accurate assessment of transient-specific classifier performance and further confirm that our algorithmically-generated training set generalises well, we assemble a test set of genuine astrophysical transients. This set was found by cross-matching a list of all spectroscopically confirmed supernovae reported to the Transient Name Server (TNS) since January 2019 with the GOTO master candidate table. Those with an associated GOTO candidate within 3 arcsec, with TNS discovery magnitude greater than the GOTO source magnitude, and only found in GOTO data after the formal TNS discovery date are accepted. With these cuts, purity is favoured over com-

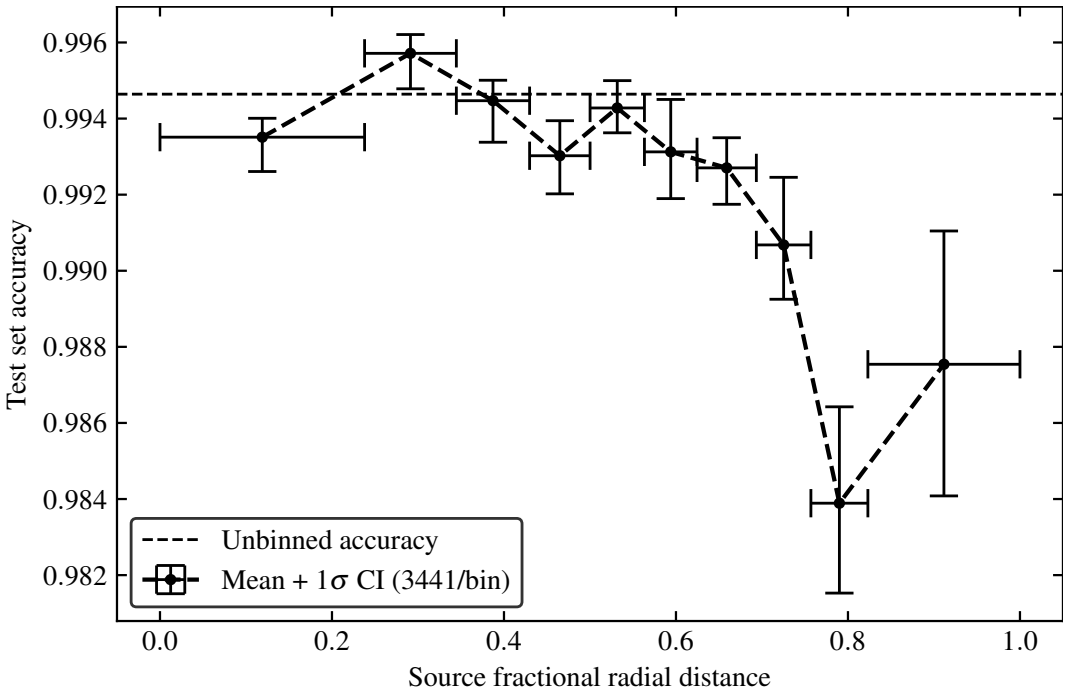

Figure 3.10: Class-balanced accuracy evaluated on the test set as a function of detector position. We use a series of concentric radial bins, chosen to contain equal numbers of sources for uniform statistics. We scale the radius through by the detector size to give a relative picture of performance. The drop in performance at large radial distances is primarily caused by the extreme optical distortion present in the early GOTO prototype, and only a minor drop of 1% in accuracy in these challenging conditions demonstrates the very robust performance of our classifier. With the design-specification GOTO optics, we anticipate this curve will be level within error.

pleteness, a deliberate choice to ensure that the test set is as clean of false positives as possible. This yields 877 known transients recovered in the GOTO prototype data. The whole-sample recovery rate is $97.2 \pm 0.3\%$, consistent with the performance achieved on the synthetic transients. This is a strong indicator that our generation algorithm for synthetic transients produces convincing detections which are useful for learning to detect genuine transients. Uncertainties on the TNS-derived set are larger than for our synthetic datasets due to both the smaller sample size and the increased complexity of the real dataset.

To confirm that consistent performance across a wide range of magnitudes is attained, the recovery rate is evaluated across a series of magnitude bins. Figure 3.11 illustrates the transient recovery rate as a function of GOTO L band magnitude. We find that the classifier maintains excellent performance across the full magnitude range of detections accessible to GOTO, even towards fainter magnitudes. Our galaxy augmentation scheme provides up to a 30% improvement in recovery rate at magnitudes fainter than $L \sim$19.5 over a pure minor planet training set. This marked improvement at the faint end of our detection range is powerful, as the expected number of other transients increases as a function magnitude, meaning this improvement in recovery rate will yield a corresponding increase in the total number of transients recovered by GOTO. Of particular relevance for GOTO, we expect the majority of kilonovae within the current GW detection volume to also occupy this magnitude range, increasing significantly our recovery rate of these faint transients in particular.

Our augmentation scheme also provides a significant improvement for sensitivity to nuclear transients, considered to be a more difficult transient morphology to detect. Motivated by the typical RMS astrometric noise level of GOTO images, we adopt a fixed threshold of 0.5 arcsec to distinguish between nuclear and offset transients. We find a 13 ± 5 % increase in the recovery rate of nuclear transients using the transient-optimised classifier compared to a pure minor planet classifier, on a sample of 15 confirmed detections. This is a direct result of the host offset distribution chosen for the augmentation scheme, which permits full freedom to generate close-in nuclear configurations. The main obstacle to improving this further is the inherent quality of the

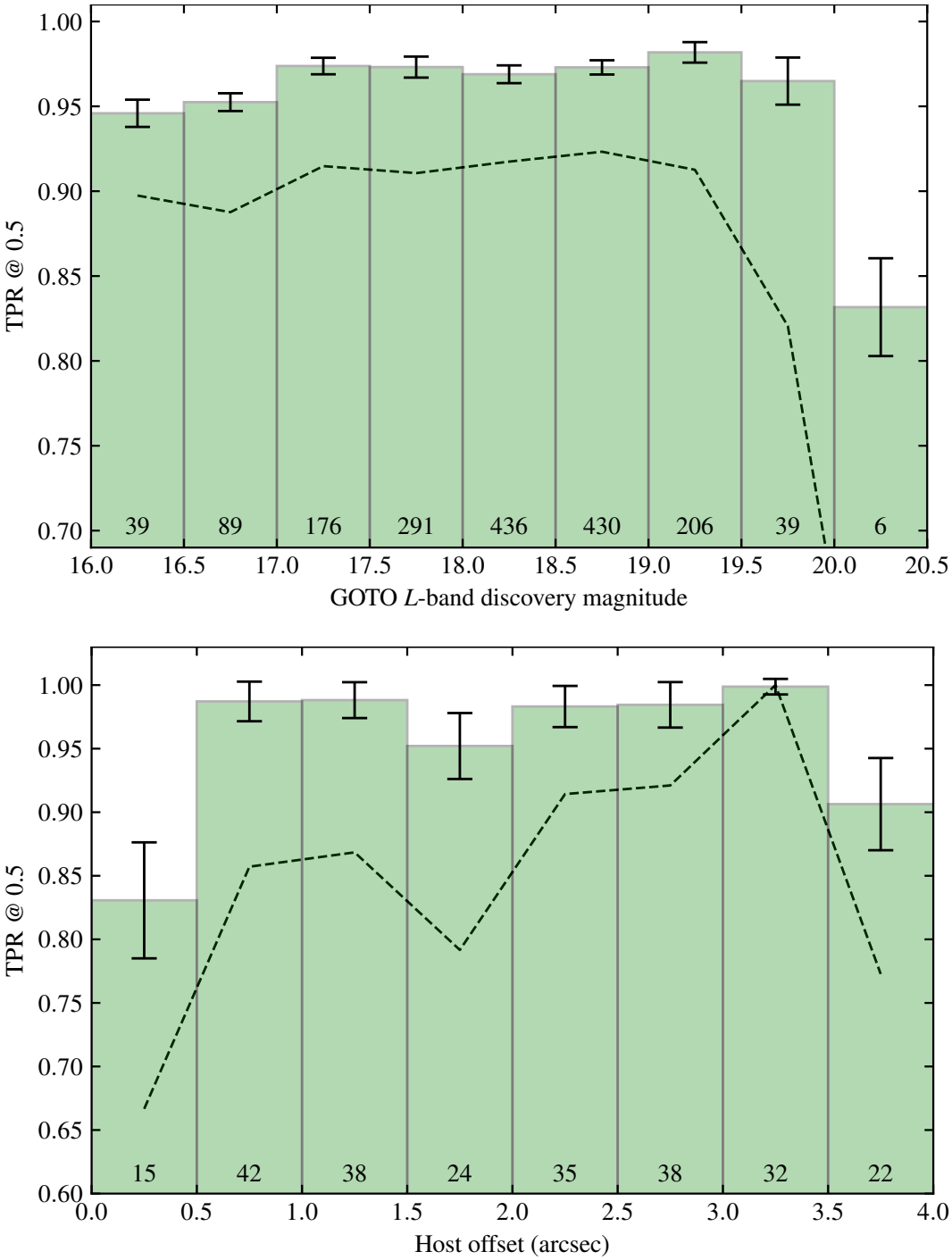

Figure 3.11: *Top panel:* Recovery rate (TPR) as a function of GOTO discovery magnitude, at a fixed real-bogus threshold of 0.5. The dashed line indicates the performance of a classifier with a similarly sized training set, but with only minor planet detections. Error bars are derived directly from the classifier score posteriors. The number of detections per bin is written below each bar. The sharp drop-off in the number of detections beyond $L \sim 19.5$ is associated with the median 5-sigma limiting magnitude of the GOTO prototype, thus expected. *Bottom panel:* Recovery rate of transients that can be reliably associated with a host galaxy (as cross-matched with WISExSuperCosmos, Bilicki et al. 2016) as a function of host offset. As above, a similarly-sized minor planet-based classifier is plotted for comparison. There is a marked improvement in the recovery rate for very small host offsets, particularly for nuclear transients.

galaxy subtraction residuals, which limits our bright-end performance.

3.4.3 Further characterisation

Although the main transient sources of interest for GOTO will overwhelmingly be fainter than the saturation level ($L \sim 15$), there are still important secondary science Galactic targets as well as rare transients occurring in nearby Local Group galaxies (e.g. SN2014J; Fossey et al., 2014) which have the potential to brighten beyond the well-sampled regions of our training set. To simulate these bright transients, GOTO detections of the first 100 minor planets are used. These have magnitudes from $L \sim$10–14, and have well-constrained orbits. Using the SKYFIELD code (Rhodes, 2019), we generate nightly ephemerides for each minor planet, and locate all difference image detections associated with each object. This yields a benchmark set of around 200 bright asteroid detections. Of the 207 detections, 99.5% are recovered, showing good consistency with the recovery rate on the fainter minor planets in the classifier test set. Of those minor planets with $L \lesssim 10$, 100% are recovered, although small-number statistics limits the usefulness of this metric. This bright-end testing demonstrates the excellent dynamic range of the classifier, showing high (>90%) recovery rates from $10^{\text{th}} - 20^{\text{th}}$ magnitude.

Through the host offset distribution choice we make, we expect to generate a reasonable number of transients at zero offset, so this region of parameter space should not be empty in the training set. To test the performance in this regime we repeated the procedure outlined in Section 2.2, except with the host offsetting routine disabled to generate synthetic detections overlapping the galaxy nucleus only. This generated 5,100 synthetic nuclear transients, with a magnitude distribution consistent with that in Figure 3.1. Testing our model against this dataset (with the negative examples being galaxy residuals as in Section 3.2.3, we obtain a 97.5% accuracy, with a recovery rate (TPR) of $\sim$ 96%. These scores are lower than the full-dataset scores, reflecting the increased difficulty of classification in this regime. The average prediction confidence on the real component of this set is 0.9390, which is less than the average prediction confidence on the real members of the test set is 0.9626, reinforcing that

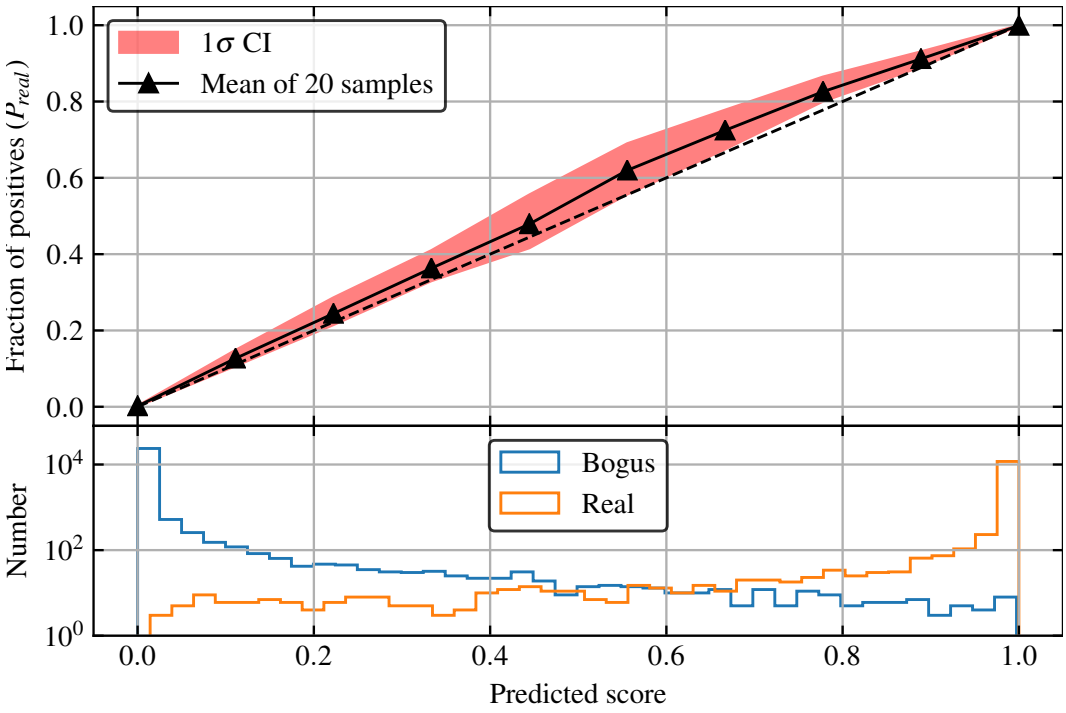

Figure 3.12: Top panel: classifier calibration curve, illustrating how well the classifier's output score corresponds to probability. The mean of 20 samples and the 1σ confidence interval are plotted to show that individual draws from the posterior remain well-calibrated. Bottom panel: Score distribution for both real and bogus examples – with the relative scarcity of examples with $0.2 < RB < 0.8$ accounting for the greater uncertainty in calibration.

these detections are more difficult than the 'average' real detection.

Another important factor to consider with any classifier is how closely the output correlates with probability – known as calibration. Although this does not necessarily impact on the classification performance, having scores that accurately reflect the probabilities of being real/bogus is important for human use of classification outputs and is important for performing inference using classifier scores. In Figure 3.12, we illustrate the calibration of this classifier by plotting as a function of classifier score the fraction of real detections at a given score. Our uncalibrated classifier shows excellent calibration, and does not show the characteristic sigmoidal calibration curve of other uncalibrated classifiers such as random forests (Niculescu-Mizil & Caruana, 2005). Calibration becomes increasingly important if different machine learning models are chained together, with downstream classifiers using the posterior probabilities of the main real-bogus clas-

sifier. With our high degree of calibration, we are justified to use our RB score as a proxy for P_{real} (the probability a given source is real) in such implementations.

One significant benefit of using a Bayesian neural network is a built-in indicator of out-of-distribution data – that is data poorly represented by or unseen in the training set. For input data that is completely different to the training set, the classifier will return a low confidence score which can then be used to remove/deprioritise the candidate in downstream applications. This confidence can also be used to optimise candidate vetting efforts, with the highest-confidence candidates being a natural choice to prioritise over lower-confidence, lower quality detections.

In principle, the task-specific knowledge encoded in our trained network weights can be used to accelerate the training of similar real-bogus classifiers through transfer learning, and in principle increase generalisation (Yosinski et al., 2014). This requires that the same data input structure is used and there are no changes to model hyperparameters. However, we caution that training in this way is susceptible to local minima and does not offer the opportunity to change the model hyperparameters that training from scratch does – in Section 3.3.1 we have demonstrated the sizeable performance improvements doing a full hyperparameter search can yield, and so encourage this.

The techniques and framework we implement in this paper are naturally extensible to more challenging astronomical classification tasks such as those outlined at the end of Section 3.1.1. A key focus is more fine-grained classification – being able to distinguish variable stars, supernovae, nuclear transients and other astrophysical objects of interest in an automated (and crucially, accurate) way. Figure 3.8 already hints at this being a fruitful approach, as we see evidence of morphological differentiation in both the positive and negative classes through the emergence of smaller sub-clusters. Similarly, leveraging the wealth of contextual information available from astrophysical surveys in a principled, informative, and efficient way within the framework of deep learning poses an open challenge, with potentially significant gains possible. We aim to address these challenges, among others, with development of future generations of the classifier we implement here.

3.5 Conclusions

We demonstrate a data-driven approach to generating large, low-contamination training sets, which along with our novel augmentation scheme can be used to train high-performance, transient-optimised real-bogus classifiers. By combining real PSFs from minor planets with galaxies, we generate realistic synthetic transients that provide a measurable improvement in the recovery of genuine astrophysical transients. This technique is computationally lightweight, easily implemented, and directly applicable to a variety of both current and future transient survey streams/datasets.

We also demonstrate the efficacy of Bayesian neural networks for the first time in real-bogus classification, and demonstrate the unique insights that confidence estimation can bring to the real-bogus problem. Being able to assign epistemic confidences to classifier predictions in addition to the more typical real-bogus score provides another parameter for human vetters further downstream to use in identifying promising candidate detections – this can potentially be used in future to further automate decision making in the context of follow-up and reporting. Techniques such as this that minimise human involvement in data-gathering and labelling will become increasingly important in the new 'big-data' era of astronomy that large-scale projects such as the Rubin Observatory and SKA will bring about.

Our classifier demonstrates excellent performance across a wide magnitude range, with a missed detection rate of 0.5% at a fixed 1% false positive rate, and up to 30% improvement in recovery rate of astrophysical transients in the challenging faint end. This has the potential to markedly increase the number of faint transients GOTO can discover, and significantly improves the prospects for detecting the kilonova afterglows of gravitational-wave driven mergers GOTO was designed to find. We anticipate that improvements to the quality and stability of GOTO's hardware and dataflow will bring significant performance gains for the real-bogus classifier presented here.

GOTO is due to undergo significant expansion over the coming years, with a final configuration of 4 installations spread across a northern (La Palma) and southern (Siding Spring) site providing a high-cadence datastream covering almost the whole

sky down to 20th magnitude every 2–3 days. The tools developed in this work have generated a classifier that is capable of handling and sifting the accompanying volume of candidate transient detections with robust accuracy and high sensitivity.

Acknowledgements

We thank the anonymous referee for their insightful comments which helped improve the quality of this manuscript. The Gravitational-wave Optical Transient Observer (GOTO) project acknowledges the support of the Monash-Warwick Alliance; Warwick University; Monash University; Sheffield University; the University of Leicester; Armagh Observatory & Planetarium; the National Astronomical Research Institute of Thailand (NARIT); the University of Turku; the University of Manchester; the University of Portsmouth; the Instituto de Astrofísica de Canarias (IAC) and the Science and Technology Facilities Council (STFC). DS, KU, BG and JDL acknowledge support from the STFC via grants ST/T007184/1, ST/T003103/1 and ST/P000495/1. JDL acknowledges support from a UK Research and Innovation Fellowship (MR/T020784/1). RPB, MRK and DMS acknowledge support from the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715051; Spiders). POB and RS acknowledge support from the STFC.

This research made use of Astropy,[7] a community-developed core Python package for Astronomy (Astropy Collaboration et al., 2013, 2018), and scikit-learn (Pedregosa et al., 2011). The resources to support `astorb.dat` were originally provided by NASA grant NAG5-4741 (PI E. Bowell) and the Lowell Observatory endowment, and more recently by NASA PDART grant NNX16AG52G (PI N. Moskovitz). This research has made use of IMCCE's SkyBoT VO tool. This research has made use of data and/or services provided by the International Astronomical Union's Minor Planet Center.

[7] `http://www.astropy.org`

Data Availability

The GOTORB code is made freely available at https://github.com/GOTO-OBS/gotorb, along with validation examples for testing. Accompanying observational data used in this work will be made available via upcoming GOTO public data releases.

Postscript

The GOTORB classifier continues to run in the live GOTO pipeline, filtering over 100,000 candidates on a typical night, and delivering candidates to the GOTO Marshall for human inspection. GOTO underwent significant hardware changes after the publication of this manuscript, with optical tube assemblies of a different design. The optical performance was vastly improved by this change, although this meant that the large training datasets used for the real-bogus classifier no longer fully represented the real world dataset. Despite this , the model shows excellent performance (in recovery of minor planets) on the new datastream. The false positive rate however is elevated, owing to new changes to the pipeline that allow us to go deeper than prior, meaning we are now routinely sampling detections at the faintest end of the distribution in Figure 3.1. We are currently waiting to gather enough data to retrain the classfier and fully expect this will resolve issues with the concept drift we are seeing currently. The dataset we constructed as part of this work forms part of Chapter 6, as the multi-class CNN discussed there, and we intend to make all future datasets in this multi-purpose fashion to facilitate further exploration.

The next-generation real-bogus classifier is simultaneously in development, moving to a new machine learning framework (JAX; Bradbury et al. 2018) and architecture of rotationally equivariant convolutions (e.g. Dieleman et al. 2015). Given the higher quality pipeline now and better quality human labelling, we can begin to explore more powerful architectures, and ongoing work aims to tie the uncertainties predicted by the Bayesian neural network architecture to human confidence.

Chapter 4

A precision ephemeris for the continuous-wave source Scorpius X-1

Note

This chapter is taken from the 2023 MNRAS article, *Precision Ephemerides for Gravitational-wave Searches – IV: Corrected and refined ephemeris for Scorpius X-1* Killestein et al. (2023). I led the re-analysis, validation, and writing of the manuscript. This work continues and extends the previous analyses presented in Galloway et al. (2014) and Wang et al. (2018)

Abstract

Low-mass X-ray binaries have long been theorised as potential sources of continuous gravitational-wave radiation, yet there is no observational evidence from recent LIGO/Virgo observing runs. Even for the theoretically 'loudest' source, Sco X-1, the upper limit on gravitational-wave strain has been pushed ever lower. Such searches require precise measurements of the source properties for sufficient sensitivity and computational feasibility. Collating over 20 years of high-quality spectroscopic obser-

vations of the system, we present a precise and comprehensive ephemeris for Sco X-1 through radial velocity measurements, performing a full homogeneous reanalysis of all relevant datasets and correcting previous analyses. Our Bayesian approach accounts for observational systematics and maximises not only precision, but also the fidelity of uncertainty estimates — crucial for informing principled continuous-wave searches. Our extensive dataset and analysis also enables us to construct the highest signal-to-noise, highest resolution phase-averaged spectrum of a low-mass X-ray binary to date. Doppler tomography reveals intriguing transient structures present in the accretion disk and flow driven by modulation of the accretion rate, necessitating further characterisation of the system at high temporal and spectral resolution. Our ephemeris corrects and supersedes previous ephemerides, and provides a factor three reduction in the number of templates in the search space, facilitating precision searches for continuous gravitational-wave emission from Sco X-1 throughout the upcoming LIGO/Virgo/KAGRA O4 observing run and beyond.

4.1 Introduction

With the completion of their third observing run, the catalogue of gravitational-wave (GW) signals detected by the Advanced Laser Interferometer Gravitational Wave Observatory (LIGO; LIGO Scientific Collaboration et al. 2015) and Advanced Virgo (Acernese et al., 2015) is growing rapidly (The LIGO Scientific Collaboration et al., 2021). These sources include the mergers of binary black holes (BHs) (Abbott et al., 2016b), a binary neutron star (NS) (Abbott et al., 2017c,e), and NS–BH binaries (Abbott et al., 2021b). However, compact-object mergers are not the only sources expected to produced detectable GW emission. Unlike these transient signals, continuous GW (CW) sources present persistent quasi-monochromatic emission. This as-yet undetected type of gravitational radiation is emitted by rapidly rotating asymmetric NSs, whether found as isolated sources or within stellar binaries; see, e.g., Lasky (2015); Sieniawska & Bejger (2019); Piccinni (2022) for recent reviews.

Numerous mechanisms have been suggested to induce the time-varying quadrupole

required for GW emission from spinning NSs, whose angular momentum loss would limit the spin rate and account for the observation that NS spins are measured at $\lesssim 700$ Hz (Hartman et al., 2003; Hessels et al., 2006; Patruno et al., 2017), below estimated breakup frequencies (Chakrabarty et al., 2003). Low-mass X-ray binaries (LMXBs) with NS primaries have received particular attention as target sources (Abbott et al., 2007; Whelan et al., 2015; Aasi et al., 2015; Abbott et al., 2017a,a, 2019c; Middleton et al., 2020; Zhang et al., 2021; Abbott et al., 2022b). In this scenario, a torque balance is achieved between the spin-up due to accretion from a stellar companion and spin-down due to GW emission (Papaloizou & Pringle, 1978; Wagoner, 1984; Bildsten, 1998; Andersson et al., 1999). Such a rotational equilibrium leads to a characteristic strain that increases with increasing X-ray flux of the source (a proxy for the accretion rate; see Bildsten 1998). The most promising candidates are thus those that are most bright in X-rays.

The prototypical LMXB, Sco X-1 (Giacconi et al., 1962; Sandage et al., 1966; Shklovsky, 1967), is composed of an accreting NS primary and donor star, and has been intensively studied since its discovery as among the closest X-ray binaries known (Gottlieb et al., 1975; Bradshaw et al., 1999; Fomalont et al., 2001). It shows strong emission across the electromagnetic (EM) spectrum, from radio (Fomalont et al., 1983) to gamma-rays (Brazier et al., 1990), powered by a near-Eddington accretion rate from the donor star. Intriguing null detections of very high energy (VHE, TeV) emission (Aleksić et al., 2011) suggest that the high-energy emission mechanism in Sco X-1 is markedly different to other systems. The donor star in the system remains enigmatic; the high accretion luminosity shrouds any stellar absorption features present in the near-infrared (Mata Sanchez et al., 2015), but these observations combined with dynamical constraints suggest a donor mass of $0.28 < M_2 < 0.7 M_\odot$ (stellar type K4IV or later). As the strongest source of X-rays on the sky, Sco X-1 has been used as a 'lighthouse' to study intervening interstellar material (García et al., 2011), magnetic fields (Titarchuk et al., 2001), and even the Martian atmosphere (Rahmati et al., 2020). Very recently, polarised X-ray emission was detected by the *PolarLight* mission at keV energies (Long et al., 2022), providing a strong constraint on the system geometry and

suggesting that the X-ray emission arises primarily from a compact, optically thin corona near the disc transition layer.

Despite being a cornerstone of our understanding of compact binary systems across the Universe, Sco X-1 still remains enigmatic, showing great complexity and variability. These complexities are simply not revealed in more distant systems where lower signal-to-noise ratio (SNR) limits their visibility. Putting this aside, Sco X-1 is the most luminous extra-solar X-ray source, which combined with its relative proximity (2.3 ± 0.1 kpc, Lindegren et al. 2021), implies it should be among the loudest CW sources detectable by current GW detectors (Watts et al., 2008) and has made it the target of a great number of search efforts.

No concrete detection of CW emission has yet been made of what we a priori expect is the strongest GW source, which remains puzzling. However, improvements in detector sensitivity have led to correspondingly stronger upper limits on the GW strain from Sco X-1, (Abbott et al., 2007; Aasi et al., 2015; Abbott et al., 2017a; Meadors et al., 2017; Abbott et al., 2017d, 2019c; Zhang et al., 2021; Abbott et al., 2022b), with the most recent analyses reaching $\lesssim 4 \times 10^{-26}$ (in the most sensitive frequency bands and assuming knowledge of the inclination angle; Abbott et al. 2022c). This level of strain begins to push below the torque-balance limit (Bildsten, 1998), where the expected GW emission balances the spin-up torque from donor accretion. One complicating factor in the case of Sco X-1 is the absence of a measured spin period, unlike many other accreting NS systems (Abbott et al., 2022a). Sco X-1 has also been the target of directional, stochastic searches (Abbott et al., 2017b, 2019b, 2021a), leading to further sensitive (yet also null) results.

As one of the primary uncertainties in any directed search, extensive X-ray timing observations have been performed in search of the spin period of the (assumed) NS primary, as revealed by X-ray pulsations or bursts (e.g. Galloway et al. 2010). The most recent placed an upper limit of 0.034% (90% confidence) on any putative X-ray variability using Rossi X-ray Timing Explorer (RXTE) data (Galaudage et al., 2022). It remains unclear whether Sco X-1 is intrinsically or intermittently variable in the X-ray, and whether any pulsed X-ray emission is being scattered away by surrounding mate-

rial.

Another crucial component to facilitate CW searches are precise orbital constraints for the LMXB systems of interest (Watts et al., 2008). In order to combine long sections of data coherently, the motion of both the Earth relative to the Solar System and the orbital motion of the NS in the binary must be accounted for. It is computationally infeasible to marginalise over the vast, high-dimensional parameter space this presents, and therefore, it is critical to place constraints on this orbital motion to reduce the search dimensionality and increase sensitivity (Dhurandhar & Vecchio, 2001; Messenger et al., 2015; Leaci & Prix, 2015). This presents a signficant observational challenge however; in a high accretion rate system like Sco X-1, the Balmer features are blurred significantly, limiting precision and making them unsuitable for precision radial velocity (RV) measurements. The disc structure is also dynamic and chaotic, leading to time- *and* phase-dependent changes in the emission line geometry.

Fortuitously, the irradiated donor star provides an alternative observational probe via narrow emission lines generated by Bowen fluorescence, which provide a remarkably precise probe of orbital motion, as first demonstrated by Steeghs & Casares (2002). The NS irradiates the face of the donor star with a strong X-ray flux, ionising He II which then de-excites, triggering a cascade of emission from C/N/O atomic orbitals (Kastner & Bhatia, 1996). The (comparatively) compact emission geometry generates narrow emission lines (with little intrinsic broadening), which precisely trace the heated face of the donor in contrast to disc emission, which has contributions from either side of the disc and thus experiences significant Doppler broadening and complex geometric distortions from specific regions, such as the hot-spot/bulge or the gas stream. This tracer of orbital motion can be used to estimate the orbital parameters of the NS with RV measurements. Beyond Sco X-1, this technique has found broad applicability to LMXBs in general (e.g., Casares et al. 2003; Cornelisse et al. 2007; Brauer et al. 2018). To this end, Sco X-1 has benefited from an extensive spectroscopic monitoring campaign since 1999 as a part of the Precision Ephemerides for Gravitational-wave Searches (PEGS) project. This has lead to incremental improvements in the ephemeris accuracy (Galloway et al., 2014; Wang et al., 2018), with coverage continuing up to the present day

through an extensive all-weather campaign with the Very Large Telescope (VLT) Ultraviolet and Visual Echelle Spectrograph (UVES; Dekker et al. 2000) providing over 200 high-quality spectra.

In support of ongoing CW searches, in this paper, we present the most up-to-date ephemeris for Sco X-1 in the literature. We leverage 20 years of spectroscopic coverage with a homogeneous approach that accounts for potential systematic errors — with a view to maximising not only the precision of the ephemeris, but also the quality of uncertainty estimates — to correct and improve upon previous analyses. This is crucial when constraining the parameter space for CW searches; over-optimistic predictions lead to missing out valid areas of parameter space, whereas over-estimated errors greatly increase the computational burden of such searches. In Section 4.2, we describe the observational data used. In Section 4.3, we describe the method to infer the orbital properties of Sco X-1 via RV measurements and present our updated — and corrected — ephemeris. In Section 4.4, we further explore the uncertain variability in Sco X-1. We present our conclusions in Section 4.5

4.2 Data reduction

4.2.1 Spectroscopic observations

As part of the comprehensive analysis performed in this paper, we collate all spectroscopy presented in the previous PEGS project papers (Galloway et al. 2014; Wang et al. 2018) and combine this with more recent VLT/UVES data, extending the overall observational baseline to 22 years (1999-2021). We obtained (or retrieved) 264 spectra with the UVES (Dekker et al., 2000) instrument mounted on the 8.2m VLT Unit Telescope 2. All observations were obtained in service mode, across variable observing conditions but with a typical SNR in the range 50–100. We focus here on the spectra taken with the blue arm using the `CD#2` dispersive element, achieving a typical resolution $\Delta\lambda/\lambda \sim 40,000$ per spectral element with the 1" slit. Data were reduced using the v5.10.13 ESO UVES pipeline. We also include all relevant legacy datasets presented in the previous PEGS releases — two runs of time-resolved spectroscopy taken with the

Group	Instrument	Program ID	No. spectra
1	VLT/UVES	077.D-0384(A)	5
1	VLT/UVES	087.D-0278(A)	52
1	VLT/UVES	089.D-0272(A)	57
1	VLT/UVES	098.D-0688(A)	32
1	VLT/UVES	599.D-0353(A)	29
1	VLT/UVES	599.D-0353(B)	89
2	WHT/ISIS	W/1999A/CAT42	137
3	WHT/ISIS	W/2011A/P23	157
			558

Table 4.1: Summary table containing all observations included in this ephemeris version, along with instrument and program IDs. 'Group' here refers to the fitting group introduced in Section 4.3.2 — datasets in the same group are fitted with common error scaling and velocity offset parameters.

ISIS double-beam spectrograph on the William Herschel Telescope (WHT) in 1999 and 2011, respectively. These observations were taken with the R400B grating and reduced using the `molly` software (Marsh, 2019). All spectroscopic datasets are enumerated in Table 4.1. A pictorial summary of the VLT/UVES data is given in Figure 4.1, where we plot the phase-folded spectrogram of the Bowen line region implied by the binary constraints made in Section 4.3.2.

For all datasets, we recompute both the barycentric time and velocity corrections at mid-exposure to reduce any scatter introduced by the different conversion routines used in the data processing pipelines. We use the `astropy.time` module and adopt the JPL DE405 ephemerides as our reference. To provide both consistency with previous ephemerides and an authoritative value, we give ephemeris times in both UTC Barycentric Julian Date (BJD) and GPS seconds.

4.2.2 Measuring Bowen line velocities

We obtain our Bowen radial velocities by fitting a constrained line model similar to the one used in Wang et al. (2018) to the reduced spectra. An example spectrum with the model is given in Figure 4.2. The narrow Bowen components are modelled with Gaussians of fixed width ($50\,\mathrm{km/s}$) and variable amplitudes, offset from their rest frame with a common velocity. In the region of interest, we identify five narrow-line N III/C III/O II

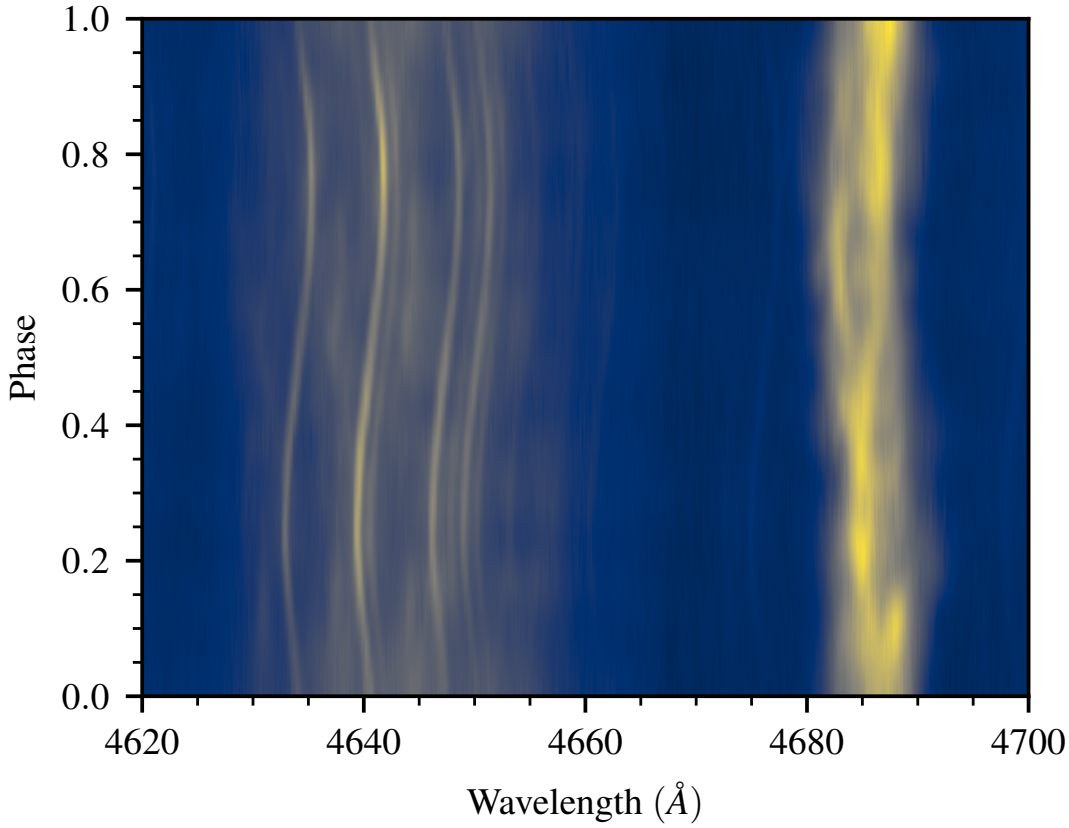

Figure 4.1: Trailed spectrogram of the VLT/UVES data presented in this work, folded by the best-fit ephemeris derived in Section 4.3 and corrected to the binary rest frame. Centred on 4640 Å is the Bowen line region of interest, and 4686 Å is the He II line. We restrict our analyses in this paper to this specific region of the spectrum, although other weaker Bowen lines are present in the spectrum.

components that optimally constrain the Bowen RVs: 4634.13 Å, 4640.64 Å, 4647.42 Å, 4650.25 Å, and 4643.37 Å respectively. Some of these narrow lines are not visible at specific phases due to system geometry. To accommodate this, we constrain all line amplitudes to be ≥ 0, such that these emission lines cannot be forced by continuum modulation to unphysical non-negative fluxes when they are not present. The broad-line component is modelled with a Gaussian of fixed width $1250\,\mathrm{km/s}$. The amplitude and centroid of this component are fitted for to remove additional correlations with the centres of the narrow lines. The line widths above were measured by Gaussian fits to the individual components, and are kept fixed to limit the number of free parameters. Empirically, the narrow-line components do not change width significantly as a function

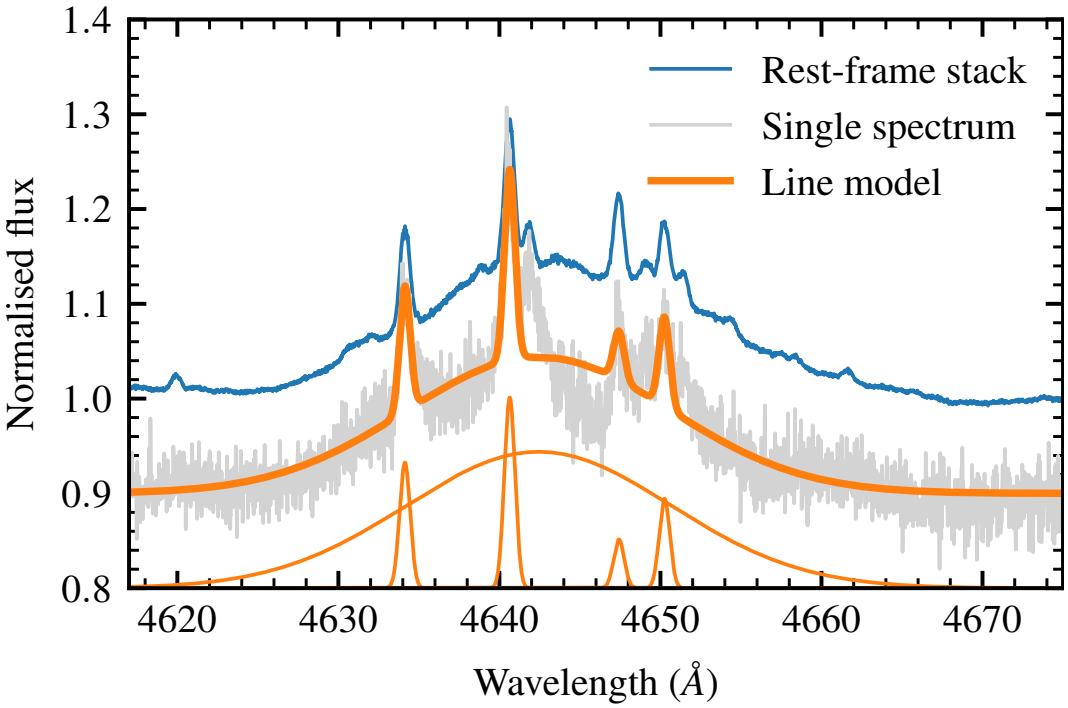

Figure 4.2: Example plot of the Bowen region in one UVES spectrum after continuum subtraction, with the Bowen line model discussed in Section 4.2.2 overplotted along with the individual components. Note that one of the fitted line components is not visible in this spectrum so has zero amplitude. We take care to avoid including the bright HeII 4686 Å line. Individual spectra typically have SNR $\sim$ 50.

of orbital phase, and any model–data mismatch is unlikely to cause large deviations in the fitted velocities as the emission-line peaks are typically clear and share a common velocity.

The final line model has eight free parameters that are well constrained by the dense spectral sampling (≈ 0.03 Å/pix) of the UVES data. Our specific line model mitigates correlations between parameters, such that uncertainties in the other fitted parameters do not skew the Bowen line velocities. Through experimentation, the UVES data could potentially benefit from the addition of more lines, but this leads to issues in the lower-quality WHT data with deviant line amplitudes so we opt for the above model for all datasets.

As a pre-processing step, we continuum-subtract the spectra by fitting a third-order Chebyshev polynomial to 10 Å regions at the red and blue ends of the Bowen line

region ($4605\,$Å to $4675\,$Å). We fit the line model using nonlinear least squares, as implemented in the `least_squares` routine in the `scipy` package (Virtanen et al., 2020). The original ephemeris of Wang et al. (2018) is used to set the initial Bowen line velocity, and other parameters are initialised with sensible defaults from the highest SNR spectrum. The RVs are then computed from the mutual Doppler shift of the measured emission-line locations with respect to their known rest-frame values. Uncertainties are rescaled to enforce a unity reduced chi-squared test statistic. We add a constant value of $0.5\,\mathrm{km/s}$ in quadrature with the error values to ensure uncertainties are not underestimated due to poor conditioning of the least-squares fit, and to include uncertainties in the absolute velocity calibration of UVES. At this stage of reanalysis we are more concerned with correct *relative* error scaling between velocity measurements, as we rescale the errors between each dataset in latter analysis steps.

4.3 Binary ephemeris

4.3.1 Corrections to previous Sco X-1 ephemerides

As part of our homogeneous reanalysis, we identified calibration errors that affect the timing parameters of both previously published ephemerides for the Sco X-1 system (Galloway et al., 2014; Wang et al., 2018). These were not readily apparent in previous work as the statistical uncertainties on the WHT-derived datasets masked their effect on the ephemeris. With the inclusion of four times as many VLT spectra, these discrepancies become significant and thus must be corrected. For full transparency, we elaborate in greater detail on the specific errors and their effect on the ephemeris:

- The subset of VLT data used in PEGS I and PEGS III were inadvertently not corrected to the mid-exposure time, nor adjusted to the Solar System heliocentre. As a result, the values of T_0 presented in these works bear systematic errors of $\sim$ 600s.

- Covariances between orbital parameters are underestimated in Wang et al. (2018) due to being taken from the initial least-squares fit to the data (using a two-point

finite-difference Jacobian approximation), rather than being computed from the samples from which the ephemeris is derived. This changes the covariance by two orders of magnitude.

Discrepancies were identified by cross-checking the reduced data against the raw frames located at the European Southern Observatory (ESO) Archive, in particular the metadata stored in the file headers. Through a careful reanalysis of the datasets, we confirm that addressing these issues brings the data used in both G14 and W18 into strong agreement, and jointly agree with the VLT data presented in this paper (accounting for deviations in the period due to dependence on low-resolution WHT data). We plot marginal posteriors for each ephemeris in Figure 4.3 to illustrate how the ephemerides change with the corrections applied above. These issues underscore the importance of both robust archiving of astronomical data, such that raw frames are easily retrievable even decades later, and the importance of storing metadata alongside these frames and documenting reduction steps; without this, these timing issues could not have been easily diagnosed. With this additional scrutiny, we are now confident any calibration artifacts are accounted for in our reanalysis – our previously-published ephemerides are considered obsoleted by the ephemeris presented in this paper, and should not be used in CW searches going forward.

4.3.2 Bayesian modelling of the Keplerian orbit

Moving forward with the reanalysis, we apply here the velocity corrections derived during the previous steps to bring all spectra into the Solar System barycentric frame, and convert observation times to BJD.

We derive our ephemeris using a standard Keplerian orbit model (assuming zero eccentricity), using for the Bowen line RVs

$$v(t) = K \sin\left(\frac{2\pi(t - T_0)}{P}\right) + \gamma,$$

where K is the velocity semi-amplitude, T_0 is the reference epoch, P is the period, and γ is the systemic velocity. To ensure good initialisation prior to sampling, we first do

Parameter	Value	Units
T_0	2456723.3272 ± 0.0004	(BJD UTC)
	1078170682 ± 33	(GPS seconds)
$T_{\mathrm{asc,ns}}$	1078153676 ± 33	(GPS seconds)
P	0.7873139 ± 0.0000002	(days)
	68023.92 ± 0.02	(seconds)
K	76.8 ± 0.2	(km/s)
γ	-113.8 ± 0.2	(km/s)
e	≤ 0.0132	

Table 4.2: Tabulated posterior distribution summary statistics for our ephemeris. Values correspond to the posterior median, while uncertainties represent the marginal 1σ confidence intervals. Explicitly, the value of T_0 refers to the inferior conjunction of the companion star, and $T_{\mathrm{asc,ns}}$ refers to the time of ascending node crossing for the NS, moving away from the observer. The top rows assume zero eccentricity. Our upper limit on the eccentricity is given in the bottom row, as the 90th percentile of the marginal eccentricity distribution. Accompanying this table is a corner plot of posterior samples in Figure 4.5.

a simple fit to the above model with nonlinear least-squares. To minimise covariance between P and T_0 in our ephemeris, we use a simple bracketing line search to find the time of conjunction $T_0' = T_0 + nP$ that minimises $\mathrm{cov}(P, T_0)$ for an integer number of orbital cycles n. This 'seed' ephemeris is then used to initialise parameters for the more complex modelling that follows, reducing burn-in time during sampling and ensuring our final ephemeris provides the minimal covariance.

To marginalise over systematics, we include two nuisance parameters per dataset: a constant offset term, δ_i, to correct for differences in absolute wavelength calibration between the datasets, and an error scaling term, ε_i, intended to correct for underestimated uncertainties. We assume a Gaussian likelihood on the RVs μ and their uncertainties σ measured at time t, given by

$$\log \mathscr{L}\left(\mu, \sigma, t | T_0, P, K, \gamma, \delta_i, \varepsilon_{i\_} \propto \sum_i \left(\frac{\mu - v(t) - \delta_i}{\varepsilon_i \sigma}\right)^2,\right.$$

where i represents the index of each individual dataset, δ_i represents the per-dataset velocity offset, ε_i represents the per-dataset error scaling. The VLT datasets have the most stable and accurate long-term calibration, and so the δ_i value for this dataset is

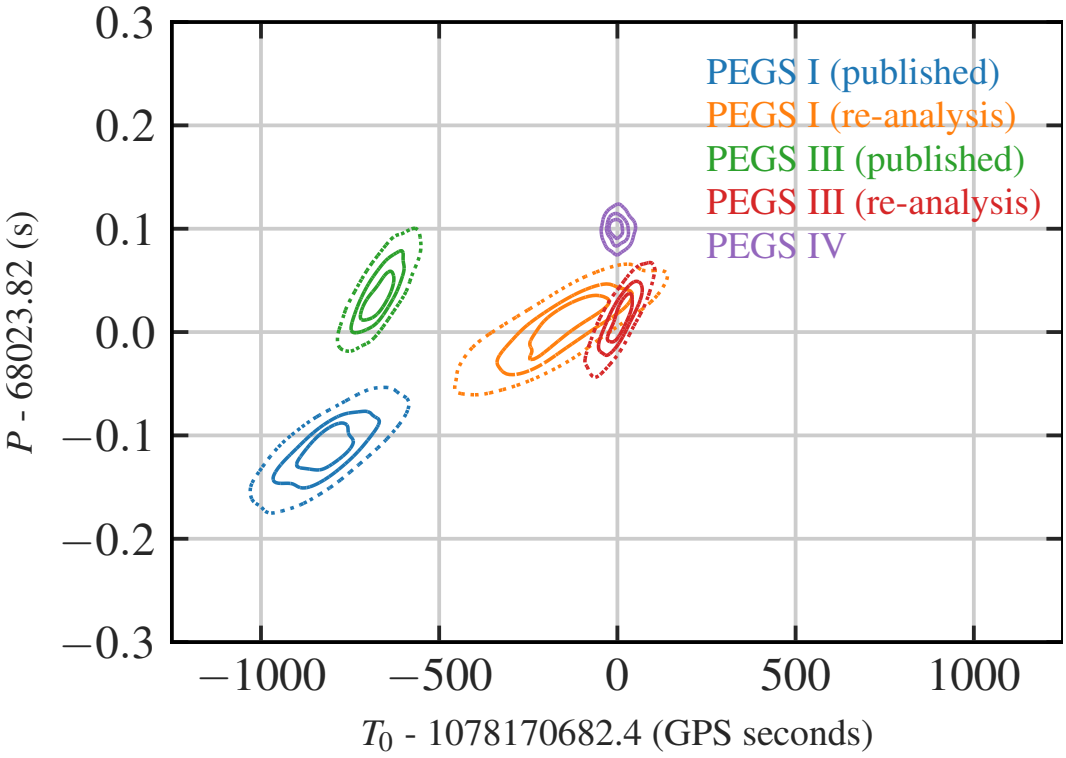

Figure 4.3: Marginal two-dimensional posteriors of T_0 and P for all previous ephemerides, propagated forwards to the epoch of our ephemeris (see Table4.2). The contours show the 1-, 2-, and 3-σ confidence intervals respectively.

fixed to zero. It should be noted that this offset term effectively also absorbs long-term velocity variations in the system — e.g., due to perturbations from a distant companion. Over the observational baseline, these would be linear in order. Regardless, the focus of this work is to provide the most precise timing of the system possible, and we have no reason to expect significant secular motion. To avoid being biased by previous ephemerides calculated from these datasets, we assume uninformative uniform priors on all variables. To mitigate complications associated with multimodality arising from an unconstrained T_0, the prior range of T_0 is centred on the seed value and bounded on either side within the seed period.

We generate samples from the posterior distributions with Hamiltonian Monte Carlo (HMC) sampling (Duane et al., 1987; Betancourt, 2017) and the No U-Turn Sampler (NUTS; Hoffman & Gelman 2011), using `JAX` (Bradbury et al., 2018) and `NumPyro` (Phan et al., 2019; Bingham et al., 2019). The number of burn-in and sampling steps are tuned

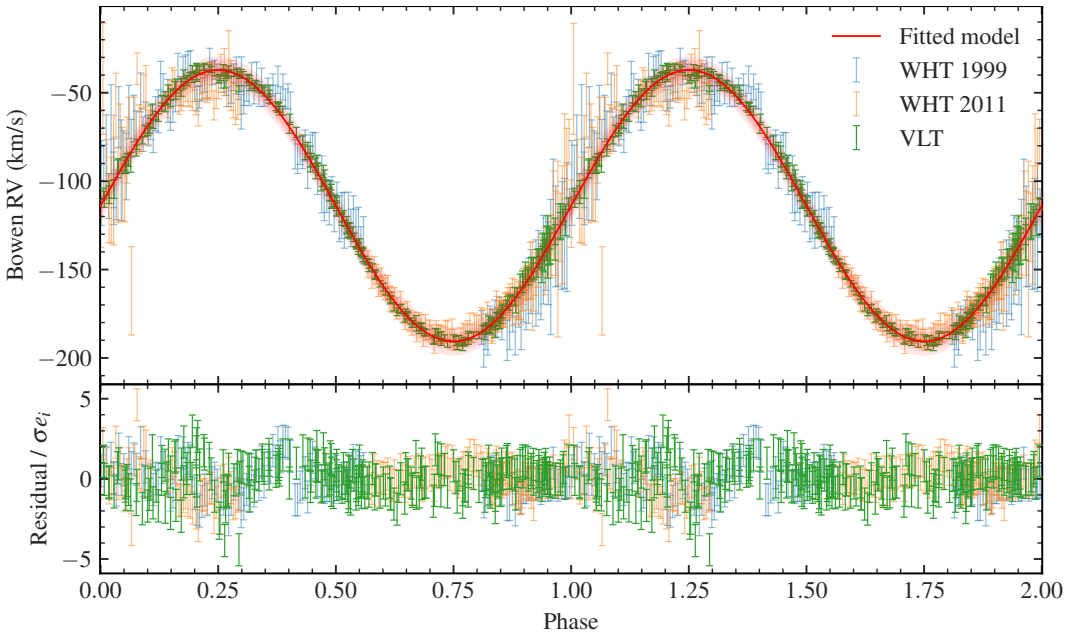

Figure 4.4: Top panel: RV curve for Sco X-1. We show the 1σ confidence intervals of the RVs measured from the 1999 WHT (blue), 2011 WHT (orange), and VLT (green) spectroscopic datasets, derived from the Bowen line model. We also plot the median posterior predictive RV curve implied by our ephemeris constraints (red) and the corresponding 1σ, 2σ, and 3σ confidence regions (darker to lighter red shaded regions, respectively). Note the significantly reduced intrinsic scatter of the higher resolution VLT data with respect to the older WHT data. Bottom panel: Normalized residuals between the observational RV measurements with the median RV curve, scaled by the respective error factors e_i

to ensure convergence via the Gelman-Rubin split-$\hat{r}$ convergence diagnostic (Gelman & Rubin, 1992). We sample four chains in parallel for 4000 steps in total, discarding the first 2000 samples of each chain as burn-in. This procedure takes just two minutes to complete on commodity hardware, running a single chain per core.

As a final step to further minimise $\mathrm{cov}(P, T_0)$, we repeat the process of shifting T_0 by an integer number of periods, this time using the HMC samples to provide a more robust estimate of the covariances. We find a shift of -106 periods minimises this, and thus we adopt this value as T_0 going forward.

Our updated ephemeris is presented in Table 4.2. Compared to (our corrected version of) the ephemeris from PEGS III, we achieve an factor 2.9 improvement in the uncertainty on the time of conjunction when propagated to the start of O4. In the top

panel of Figure 4.3, we compare the two-dimensional marginal distribution of (T_0, P) for our updated ephemeris with that for each of the previous (corrected and uncorrected) ephemerides, propagated to the epoch of the new ephemeris. Our corrections are clear in the now consistent values of the reference epoch, T_0. There are some deviations in the period P, however this is expected due to the inclusion of more VLT data (with better resolution of the spectral lines) at later epochs.

In Figure 4.3, we propagate the uncertainty on the PEGS IV T_0 to past and future epochs. This again demonstrates the consistency of our corrected and new ephemerides with each other. Of course, the uncertainty in the propagated time of inferior conjunction for the neutron star $T_{ns,asc}$ grows in time, due to the posterior uncertainty in the ephemeris, illustrated by Figure 4.6. However, our updated measurements reduce this rate of growth into the future observing runs, O4 and O5, of the LIGO/Virgo/KAGRA GW detectors, as is crucial for increasing the sensitivity of searches for CWs from Sco X-1. In Figure 4.4, we present the RV curve of Sco X-1 implied by our ephemeris inference. The high-resolution VLT spectra result in an excellent match with the fitted RV curve, with robust uncertainty quantification carried through our Bayesian measurements of the ephemeris. Finally, Figure 4.5 presents a corner plot of our samples, along with diagnostic statistics to illustrate the proper exploration of the parameter space.

For completeness, we present the ephemeris posterior and additional diagnostic quantities in Figure 4.5. To enable the principled calculation of search boundaries based on the full posterior distributions, we make high-quality samples available alongside this paper in the Supplementary Material section.

Examining the posteriors in Figure 4.5, our marginal distributions are largely Gaussian, and inter-parameter correlations are very weak. Our two-step line-search algorithm has successfully reduced the covariance between T_0 and P significantly to an (absolute) value of $3.6 \times 10^{-15}\,\mathrm{d}^2$ – multiple orders of magnitude improvement over the value quoted in Galloway et al. (2014). It is further reassuring to see that both $\delta_{1999WHT}$ and $\delta_{2011WHT}$, the systemic offset parameters for each dataset, are consistent with zero at the 1σ confidence level, verifying the validity of the absolute wavelength calibration of

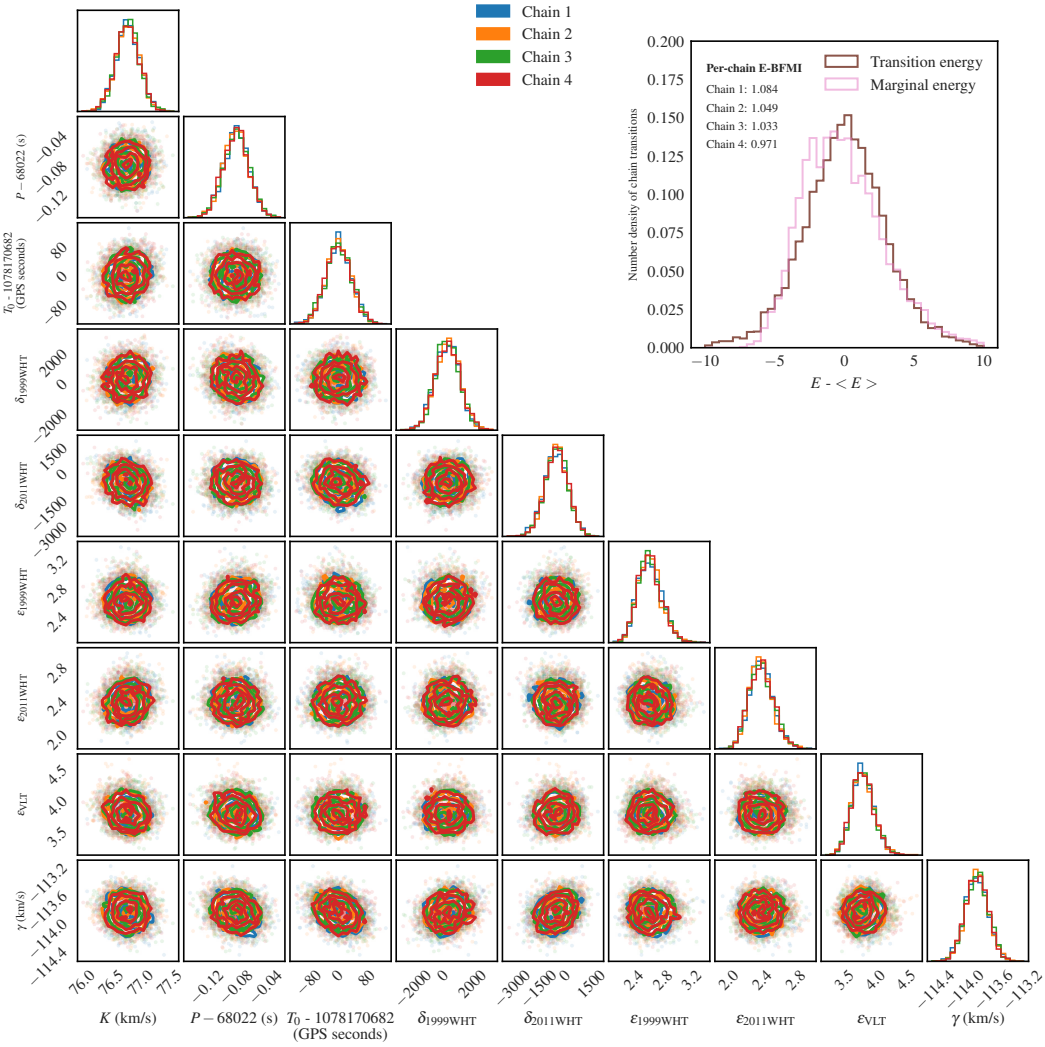

Figure 4.5: Corner plot of posterior samples for the $e = 0$ ephemeris, coloured according to each independent chain to confirm randomly-initialised chains converge on the same posterior mode. Inset: Histogram showing the distributions of marginal energies against the transition energies for each step taken by the NUTS sampler - the close match between these two distributions implies full and efficient (low autocorrelation) exploration of the parameter space. The estimated Bayesian fraction of missing information is close to 1 for all chains, providing a quantitative verification of this also. These diagnostics are unique to Hamiltonian Monte Carlo and provide an orthogonal check to the usual split-$\hat{r}$ diagnostics employed in MCMC that we use above. Refer to Betancourt (2016) for a full theoretical explanation, and further details.

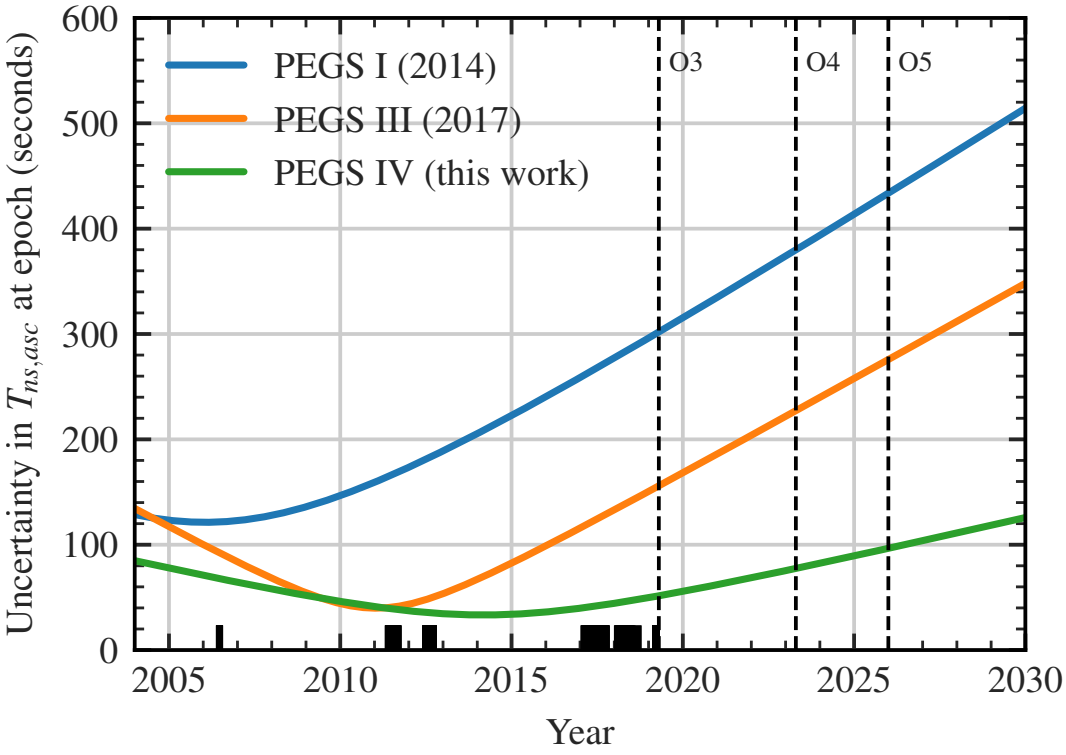

Figure 4.6: Plot of uncertainty in T_0, propagated to past and future epochs for all updated ephemerides presented in this work. The black bars overplotted denote the times of individual spectra contributing to the ephemeris. Our ephemeris provides a factor 2 improvement in uncertainty for the O3 observing run, growing out to later times.

both datasets – nevertheless their inclusion is important to account for a potentially non-trivial systematic effect on the ephemeris, and to provide correct uncertainty estimates on each parameter.

Our ephemeris provides an excellent (~ 30s) precision on the time of conjunction T_0, providing a significant ($\sim$ factor 3) reduction in the required template search space (following Galloway et al. 2014). At these levels of precision, even subtle effects begin to have marked impacts on the ephemeris. It is therefore crucial to acknowledge that there are additional sources of statistical and systematic uncertainty on timings that may hamper efforts to push the ephemeris to even greater precision.

Sco X-1 shows optical variability on the level of $\sim$ 0.5 mag, with typical variability timescales of ~ 1 hour. As a result, we observe a flux-weighted average radial velocity over the length of our exposure, adding additional uncertainties to our velocity estimates. This is likely a small effect at the sub-$\mathrm{km/s}$ level owing to our comparatively short exposure times ($\lesssim 5\%$ flux variability over one $\sim$ 700s UVES integration), however may begin to be important when searching for deviations from Keplerian radial velocities — we discuss this further in Section 4.3.3. There are also potential uncertainties arising from the timestamps applied to each spectrum acquired – it is challenging to quantify the temporal accuracy as this is often poorly documented, and rarely validated experimentally. This is particularly true in the case of 'historic' datasets from many decades ago. In the ideal scenario, timestamps would be derived from GPS time (e.g. see Dhillon et al. 2007), or derived from system time that is updated regularly via Network Time Protocol (NTP), each of which can keep clock errors minimal. This also implies potential per-observatory time offsets, which may manifest in non-trivial ways. These effects are currently dwarfed by the statistical error present on our ephemeris, and are only likely to become problematic given a significantly increased volume of data – which will reduce the statistical errors on the ephemeris. We nevertheless caution that the 'true' uncertainties on our ephemeris may be larger than implied by our posterior samples, and encourage conservative allocation of CW parameter spaces to accommodate this. Future work will look to incorporate some of these effects into our Bayesian framework, as well as further extending the observational baseline to provide

the volume of data required to model them.

4.3.3 Eccentricity constraints

We now consider potential deviations from the standard circular orbit model discussed above. As remarked upon in Wang et al. (2018), obtaining direct constraints on the orbital eccentricity of Sco X-1 is complicated by the Roche lobe geometry of the companion star. As it fills its Roche lobe, emission occurs from an extended surface of the donor, which continuously changes aspect ratio with orbital phase. This naturally leads to deviations from the pure sinusoidal RV curve expected from the base Keplerian orbit model, and adds additional periodic modulation on the $\sim 1\,\mathrm{km/s}$ level. This creates an apparent eccentricity dominating over — and potentially degenerate with — any true orbital eccentricity, which makes it difficult to disentangle the effects of each. As with other LMXBs, there are strong physical reasons to expect the orbital eccentricity e of the system to be close to $e = 0$, due to tidal dissipation occurring during the main sequence lifetime of the NS progenitor (see, e.g., Tassoul & Tassoul 1992). As a further complication, the vanishing phase space (Lucy & Sweeney, 1971) as $e \to 0$ induces a bias towards non-zero orbital eccentricities, and therefore, any detection of $e > 0$ should be interpreted with caution mathematically also.

Bearing this in mind, we repeat the analysis of Section 4.3, but fit for a nonzero eccentricity, using a more general form of the Keplerian orbital model described previously. We adopt the common parameterisation $(\sqrt{e}\cos\omega, \sqrt{e}\sin\omega)$ to sample the orbital eccentricty e and the argument of periapsis ω. This decorrelates the posterior at low eccentricities and make sampling easier, assuming uniform priors on both of these parameters in the range (-1, 1). We use the `JAX`-based `kepler` solver from the `exoplanet` package (Foreman-Mackey et al., 2021) for compatibility with the NUTS algorithm.

We empirically find an upper limit on the orbital eccentricity of Sco X-1 of 0.0132 (0.0161) at 90% (99%) confidence level. The commonly-employed significance test of Lucy & Sweeney (1971) is of limited utility here as it considers solely orbital eccentricity, and (falsely) suggests we should adopt an elliptical orbit as a result. We expect any non-zero eccentricity originates entirely from the Roche geometry, bearing in mind

the tidal dissipation seen in similar LMXBs. This provides a conservative constraint of parameter space for potential CW searches; it is difficult to move beyond this without more intensive modelling efforts, involving full modelling of the binary system in both light curves and RVs. Some early progress is being made (see, e.g., Cherepashchuk et al. 2021), although it is important to note that light curve modelling carries potential systematics that must be carefully accounted for when combined with other independent constraints. Even with the high-quality UVES data here, the expected RV modulation from the distorted secondary (see Figure 6 of Wang et al. 2018, of order $100\,\mathrm{m/s}$) is comparable to the statistical uncertainty, making the prospect of direct measurement unlikely — the greatest modulation occurs around phase zero, where the Bowen emission lines are weakest. Deviations from a pure sinusoid are directly informative on the inclination i (Masuda & Hirano, 2021), making further reduction of RV uncertainties an important goal going forward.

4.4 Binary properties

As the prototypical LMXB, Sco X-1 has been the focus of intense study since its discovery in the early 1960s (Giacconi et al., 1962; Chodil et al., 1965). However, the high (super-Eddington) accretion rate and its effect on the Balmer lines means there is still uncertainty surrounding the structure of the accretion disc in Sco X-1 — whether the observed form is stable over long timescales, and how this correlates with the various X-ray and optical states of the system (as identified in Scaringi et al. 2015). The long-term, high-resolution dataset collected as part of the PEGS program provides the means to probe secular evolution of the disc, as well as search for period derivatives and other phenomena not detectable with the typical dense but short-baseline observational sampling presented in previous work, although one pays a penalty in the resolution of fine structure owing to the dynamic and changing disc state. We expand on this in the subsections below.

4.4.1 High-resolution, high-SNR line atlas for Sco X-1

With over 200 high-resolution spectra of Sco X-1, and a high-precision ephemeris, we construct a 'line atlas' by combining the spectra in the donor-star rest frame. By doing this, weak spectral lines not visible in the individual spectra can be detected, and the quality of all lines improved. The variability in the broad H/He line profiles leads to some smearing of these lines. However weaker, narrow lines are not affected as heavily. A significant number of VLT/UVES spectra taken as part of PEGS also have a simultaneous observations with the red arm, which, although of limited utility for constraining the binary motion due to the lack of strong and narrow spectral features, encodes information about the cooler material present in the Sco X-1 system. We include this in our line atlas primarily for completeness, although caution that telluric subtraction was not performed as part of data reduction, and so the red arm suffers from residual telluric contamination. This is somewhat suppressed by our shift-adding scheme.

We use the measured Bowen line velocities to shift all spectra into alignment, and normalise out the continuum flux by fitting Chebyshev polynomials (7th order for the UVES Blue arm, 5th for UVES Red arm CCD 1, 1st for UVES Red arm CCD 2), rejecting outlier points to avoid fitting spectral features. The heavy telluric contamination redwards of 9000 Å necessitates the use of a low-order polynomial to avoid erratic ringing effects. Spectra are then reinterpolated onto a common wavelength grid, and median-combined with a sigma-clipping outlier rejection scheme to remove any cosmic ray hits or defects. This process is performed separately for the blue arm and the red arm, and we take care to treat the two red CCDs separately to avoid discontinuities. Note that the continuum normalisation at the far-red end of the spectrum is hampered by the strong telluric bands, so we opt for a simple linear continuum only.

Our finalised line atlas is presented across Figure 4.7 (CD#2, blue) and Figure 4.8 (CD#4, red) — we achieve a median SNR of ~ 1000 in the blue arm, marginally above the expected $\sqrt{N}$ scaling owing to our use of sigma-clipping, but likely affected by line variability.

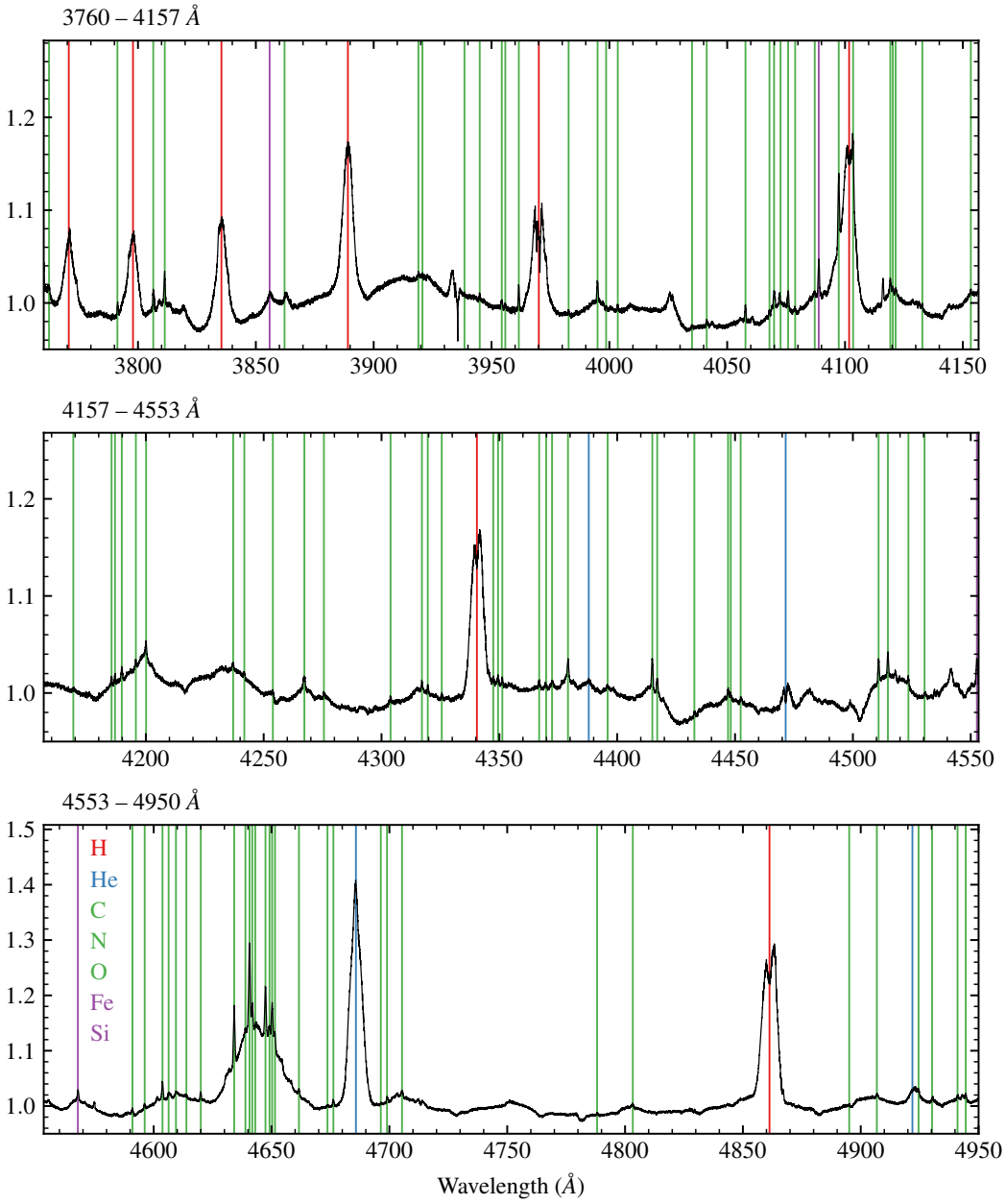

Figure 4.7: Co-addition of all VLT/UVES spectra in the binary rest-frame, to produce a line atlas that reveals the presence of many low-strength emission lines. Spectra are broken into chunks of 400 Å to aid visualisation. We annotate potential lines of interest to the LMXB community.

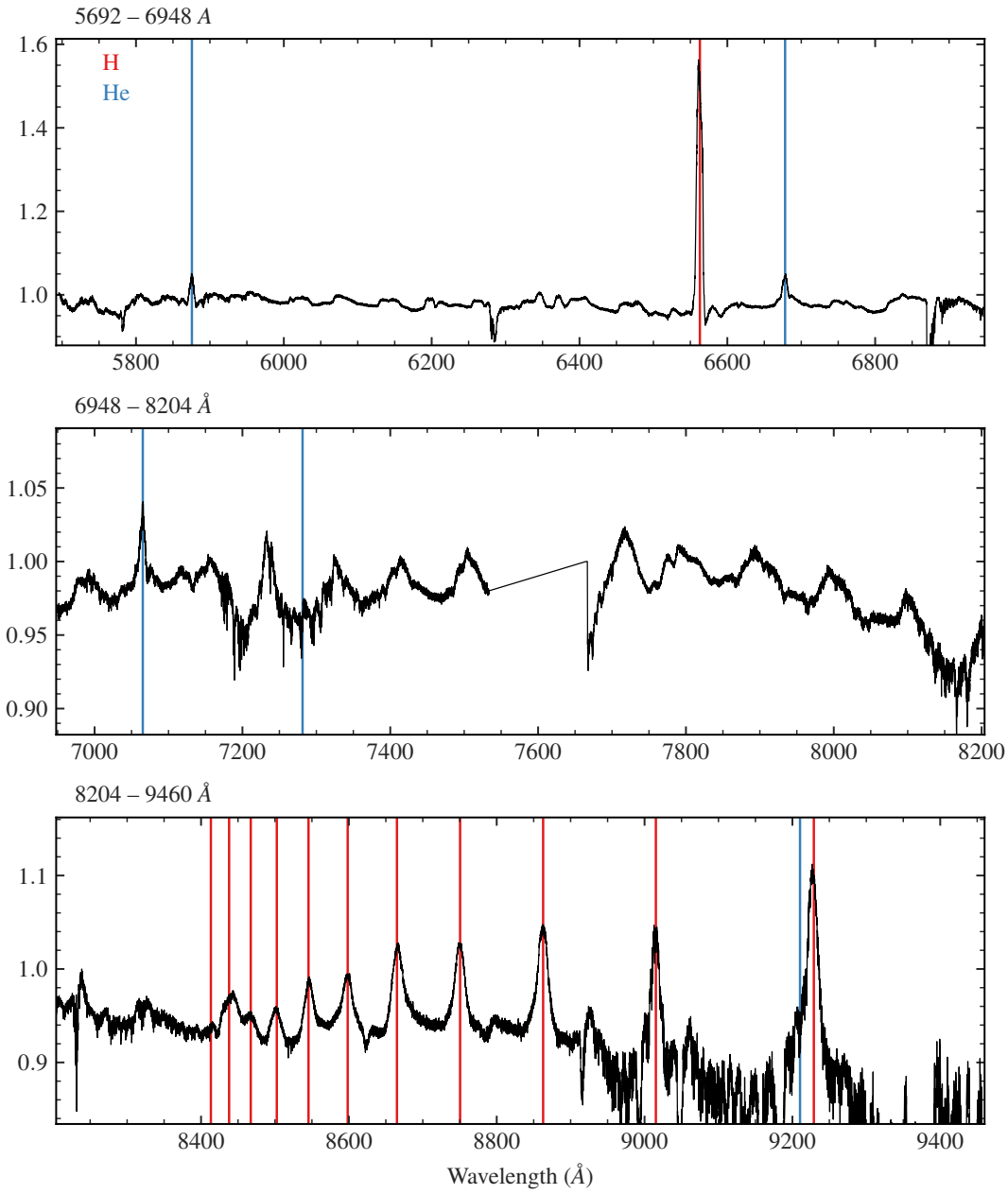

Figure 4.8: Line atlas for Sco X-1 in the binary rest frame continued from Figure 4.7 – this figure shows the UVES red arm spectra, covering 5692 Å to 9460 Å. Note that no telluric correction has been applied, and we instead use the binary motion to median the weaker features out. Some contamination persists in the red, hence we only attempt to identify the strong spectral lines of H and He, with the weaker C/N/O lines obscured by atmospheric absorption. A full treatment of telluric absorption is required to identify all faint lines present, but this is beyond the scope of this paper. UVES has a chip gap covering approximately 7540 Å to 7660 Å in this configuration. Spectra are broken into chunks of 1200 Å to aid visualisation.

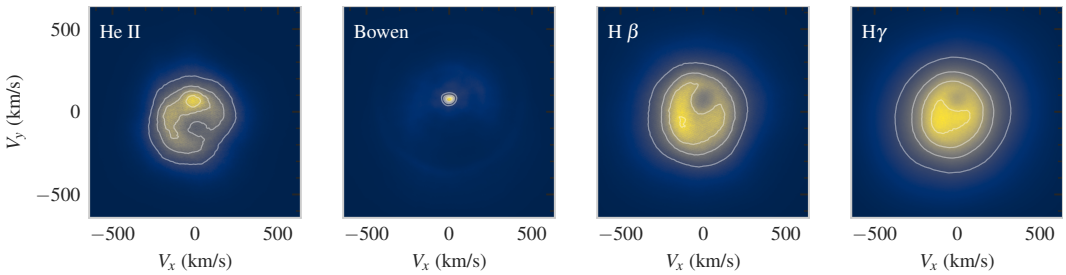

Figure 4.9: Doppler tomograms for prominent spectral features of interest in the UVES blue arm. Contours of constant intensity are overplotted to aid visualisation. Note that the Bowen image is dominated by the strong donor star features, with only tenuous hints of the sparse disc visible.

4.4.2 Doppler tomography

We applied Doppler tomography (Marsh & Horne, 1988; Marsh, 2005) to our phase-resolved UVES spectra to probe the structure of the accretion disc surrounding Sco X-1. Our spectra are pre-processed to remove continuum flux using the same procedure as Section 4.4.1 and normalised in flux over the line region of interest to mitigate the impact of flux variability on the final Doppler map. We use the Python bindings of the `DOPPLER`[1] code to produce the Doppler maps. The period and systemic velocity are fixed according to the median of our ephemeris posteriors (see Table 4.2), and we compute the Doppler map over a square grid $500\,\mathrm{km/s}$ in side length, with a per-pixel resolution of $1.5\,\mathrm{km/s}$ in the x-y disc plane velocities v_x and v_y. We neglect any velocities in the z plane for computational reasons, but note that our spectra are unlikely to be constraining along this axis (see Marsh 2022). Doppler maps are constructed for the Hβ, Hγ, He II and Bowen lines individually. These are shown in Figure 4.9, alongside the target χ^2 value used for the maximum-entropy optimisation procedure. Despite the marked resolution improvement afforded by VLT/UVES over previous datasets, the output Doppler tomograms are of similar quality to the previous maps presented based solely on WHT data in Wang et al. (2018). We speculate that this likely arises to long-term seasonal variations in the structure of the accretion disc around Sco X-1, and explore this possibility further in the following section.

[1] `https://github.com/trmrsh/trm-doppler`

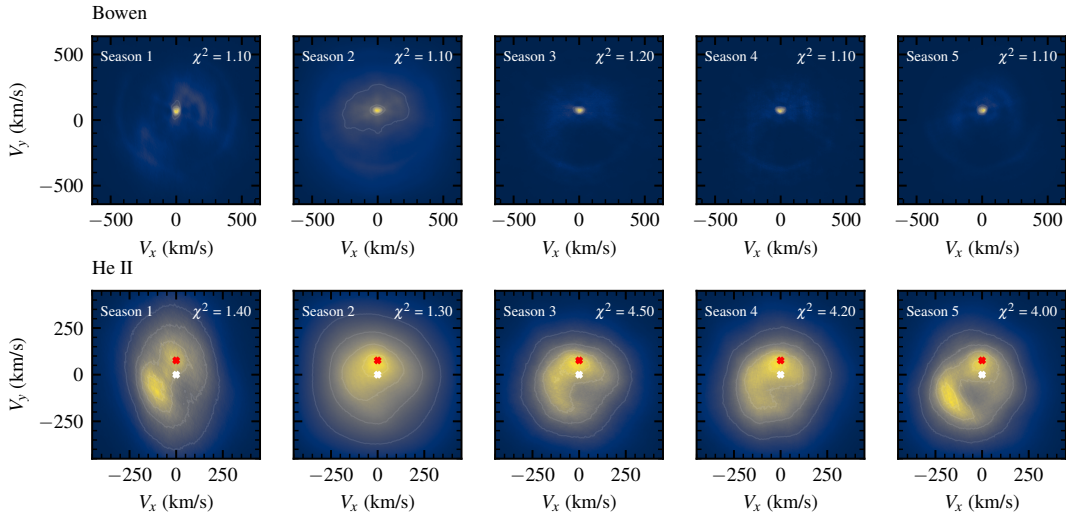

Figure 4.10: Doppler tomograms for each observing 'season', using the method defined in Section 4.4.3. We compute season-averaged tomograms for the Bowen lines (top panel) and the He II lines(bottom panel), to search for long-term secular variations in the disc structure. The Bowen Doppler maps are largely dominated by the strong donor star signature, but some structure is visible. The red and white markers indicate the positions of the donor star, and the system centre-of-mass respectively.

4.4.3 Secular variability in the He II disc?

To explore the possibility of seasonal or secular variability in the disc structure of Sco X-1, we compute He II Doppler maps on subsets of the full UVES dataset presented in this paper. We manually group contiguous subsets of our data into five 'seasons' – selected roughly corresponding to the ESO observing semesters. We then compute individual Doppler maps for each season, to avoid averaging over any long-term variability. The target χ^2 is adjusted per-season to avoid over-resolving disc features. The resulting Doppler maps are shown in Figure 4.10.

We focus our attention on the He II Doppler maps, as these show a clearly-resolved disc structure. Some tenuous variability is present in the Bowen disc also, although this is hard to quantify or comment on phenomenologically, given that the donor star feature (at $V_x = 0$, $V_y = 76.4$ km/s) dominates the reconstruction. In the He II maps, we see a strong asymmetric emission component present in the accretion disc. Across multiple observing seasons, we see evidence for evolution in intensity of this 'rim'-like structure — with minimal intensity in Season 2 and maximal intensity in Season

5. We interpret this feature as an extended region of gaseous material suspended above the disc and corotating with it — potentially a signature of stream–disc overflow occurring in the system (see, e.g., Kunze et al. 2001). Given the high accretion rate of Sco X-1, this is not surprising, as the accretion process is stochastic and variable. As our Doppler maps are averaged over many orbital periods, this is likely to be a longer-term secular variation rather than an effect correlated with the disc state. This may potentially coincide with periods of enhanced accretion onto the NS, as implied by short- ($\sim$ weeks–months) timescale fluctuations in the X-ray luminosity seen in MAXI light curves (Hynes et al., 2016). These disc states are well-known (e.g., Scaringi et al. 2015), but we do not resolve these with our sparsely-sampled VLT/UVES dataset.

We encourage additional observations and characterisation of these secular variations in the disc structure, preferentially via intensive spectroscopic observations over blocks of a few nights, every few months. This ensures each block has adequate phase coverage, and minimises the blurring of structure caused by a more coarse observing program such as our VLT/UVES data. Simultaneous X-ray coverage is crucial to correlate structural changes with potential modulation of accretion rate, and to understand the disc state. As the prototypical LMXB, targeted studies of this phenomenon in Sco X-1 have the potential to provide insight into the longer-term dynamics of accretion across the domain of X-ray binaries and other high accretion rate compact binaries.

4.4.4 Behaviour of the Balmer lines

Complex absorption features are present in the Balmer lines — as reported by Scaringi et al. 2015 — reminiscent of P Cygni-style outflow profiles with red/blue absorption wings superimposed on a broad central emission feature. These occur antiphase with the overall optical brightness of the system (and with the Balmer line intensity itself), showing maximum absorption when the Balmer lines are strongest and no absorption when the Balmer lines are barely visible. Both red and blue components are visible, showing considerable variations in absorption depth. Given the strong time variability demonstrated in previous sections, it is challenging to probe this with our VLT/UVES dataset. Nevertheless, we present some broad overview properties below in the hope

of encouraging further characterisation. Such profiles have also been noted previously in NIR spectra of the system on the Brγ line, e.g. see Bandyopadhyay et al. (1999).

We focus on the Hα line as this shows the strongest absorption (see Figure 4.11) and thus provides the most robust velocity measurements, although this phenomenon is also visible in $H_{\beta,\gamma}$. We apply a local continuum normalisation to bring all spectra onto a common baseline, then fit a Chebyshev polynomial to the line wing to provide a smooth approximation to the spectrum on this specific region. We find the minimum value of this polynomial and use bootstrap iterations to measure the uncertainty on our minimum wavelength value. To reject spectra where no absorption is present and thus ensure our velocity measurements are robust, we only consider spectra from this point on where there is $\geq$ 3% absorption in this red wing.

Under the assumption that this is an outflow, the 'apparent' velocity varies between $100\,\mathrm{km/s}$ to $1000\,\mathrm{km/s}$, with variable absorption depth as remarked upon above, and appears to show some correlation with orbital phase. However, such a phase-dependent morphology could equally be replicated with the addition of a broad absorption component superimposed on the emission component (see Figure 4.11). To illustrate this, we fit a toy model to one of our spectra showing high absorption, composed of a narrow emission line centred on Hα, and a broad (at least 2x wider than the emission component) absorption line with a free mean value. We note that the true structure of the Balmer lines is markedly more complex than this, and this simplified model is intended to be illustrative, rather than fully reproducing all features of the data.

Owing to degeneracies between the strength of the $H\alpha$ line and this component it is not possible to constrain the amplitude of such absorption, but one solution suggests a width of around $8\times$ that of $H\alpha$. This could be created by optically-thick absorption of the inner disk surrounding the NS in Sco X-1, potentially probing some NS-driven outflow that may act as a crude tracer of the true NS radial velocity amplitude. This component appears to have some systemic velocity offset compared to the Balmer line in emission. It relies largely on serendipity to observe the disc in the right state to make these measurements, and would benefit strongly from simultaneous X-ray observations, to this end. Disentangling the origin is not possible with the sparse sampling

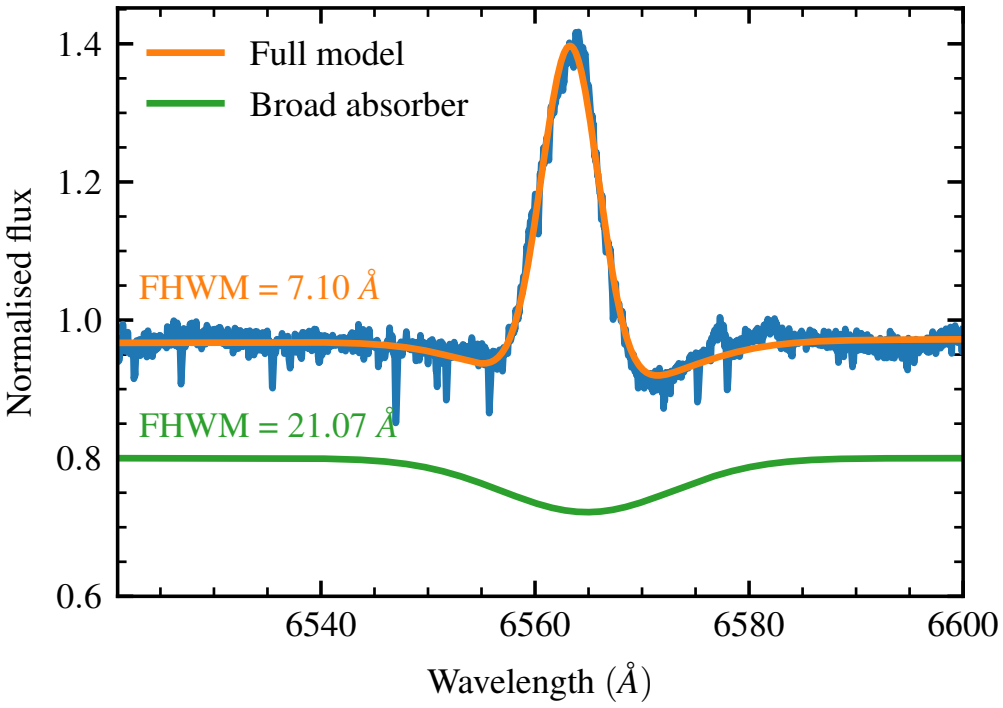

Figure 4.11: Spectrum of Sco X-1 showing strong absorption wings on the Hα line. Overplotted is a two- Gaussian model that reproduces the line shape shown, with one positive component representing the broad Hα line, and one constrained-negative component reproducing the broad absorption wings present.

of the VLT/UVES data, and we therefore advocate for additional intensive monitoring to investigate the nature of this absorption.

4.4.5 Constraints on other system parameters?

Despite the marked improvement in timing uncertainties offered by our reanalysis and extensive dataset, it is difficult to further constrain the system's orbital parameters, which are important in CW searches. The uncertainties on the other system parameters of interest are dominated by the poor constraints on the mass ratio q and the unknown neutron star radial velocity K_1, that are difficult to further improve on. The additional VLT/UVES data provide no stronger constraint on K_1 owing to the variability we discuss in Section 4.4.3 blurring the Doppler tomograms. Similarly, we cannot tighten the lower bound on q as the lines are already fully resolved by our spectra. We recreated the Monte-Carlo analysis of Wang et al. (2018), casting the problem instead as a Bayesian forward- modelling approach (using the HMC methods discussed in Section 4.3), but this failed to provide any stronger constraints, owing to the above considerations.

Further observational work is required to better constrain these orbital parameters, yet the complexity of Sco X-1 makes this challenging — detection of donor star features in the near-infrared is hampered by the strong accretion flux (Mata Sanchez et al., 2015), and light-curve modelling is made more difficult by the distinct optical high and low states in the system. Only by jointly considering all available constraints, and pushing the capabilities of current-generation instruments, can we begin to more robustly constrain the orbital parameters of Sco-X 1 and further reduce the parameter space for CW searches going forwards.

4.5 Conclusions

As the prototypical LMXB, Sco X-1 continues to remain central to current searches for continuous GW sources. Tensions between the predicted GW emission and the GW strain limits obtained from previous observing runs are beginning to emerge, as specific sub-bands now reach below the torque–balance between accretion spin-up and GW

emission spin-down. However, this is largely contingent on the inclination constraint obtained from Fomalont et al. (2001), and further data is required to reach constraining upper limits whilst simultaneously marginalising over the unknown inclination; it may be the case, e.g., that the radio lobes do not trace the orbital inclination.

Although any GW emission thus far has eluded detection, the upcoming LIGO/Virgo/KAGRA O4 observing run promises to further improve instrument sensitivities. Furthermore, searches with increased sensitivity (e.g., Mukherjee et al. 2022) may yield more stringent constraints over previous results. With upcoming third-generation GW detectors such as the Einstein Telescope (Maggiore et al., 2020) and Cosmic Explorer (Reitze et al., 2019) — bringing orders of magnitude increase in strain sensitivity and delivering high SNR GW detections that will unveil populations of compact objects currently out of reach for the current ground-based detectors (e.g., Cieślar et al. 2021) — likely including Sco X-1, alongside a wealth of other science outcomes (Kalogera et al., 2021).

From a compact object perspective, further studies of Sco X-1 focused on the transient structure of the disc and the nature of the companion star is crucial in underpinning our understanding of more distant LMXBs, and may yield even stronger constraints that may synergistically further constrain the parameter space for GW searches. It is clear many processes remain poorly understood.

Leveraging the improved search sensitivity afforded by the enhanced detectors and computational methods discussed above is contingent on a precise — and, critically, up-to-date — ephemeris for the donor star in Sco X-1. The ephemeris presented in this work will enable precision searches for CW emission from Sco X-1, which will further constrain any emission down to the torque–balance limit across the entire band of search frequencies — both in the upcoming LIGO-Virgo-Kagra O4 observing run and beyond. Sparsely-sampled, high-resolution observations of Sco X-1 over the coming years can efficiently keep this ephemeris current for the foreseeable future, underpinned and facilitated by the extensive and corrected VLT/UVES constraints that we present here.

Data Availability

All reduced data (excluding those under a proprietary period) are publicly available via the ESO Archive and ING Archive respectively. All associated data and reduction codes will be made publicly available after the release of this manuscript at `https://github.com/tkillestein/pegs_plus`. We intend to regularly publish updates to this ephemeris in light of new data, with the version of record hosted at Zenodo. (DOI: 10.5281/zenodo.7635465) This paper makes use of the following software packages: `numpy` (Harris et al., 2020), `scipy` (Virtanen et al., 2020), `astropy` (Astropy Collaboration et al., 2013, 2018), `matplotlib` (Hunter, 2007), `jax` (Bradbury et al., 2018), `numpyro` (Phan et al., 2019; Bingham et al., 2019), `corner` (Foreman-Mackey, 2016), `pandas` (McKinney, 2010), `cividis` colour map (Nuñez et al., 2018)

Acknowledgements

We thank the anonymous referee for their careful review of our manuscript. Based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programmes 077.D-0384(A), 087.D-0278(A), 089.D-0272(A), 098.D-0688(A), 599.D-0353(A), and 599.D-0353(B). and on observations made with the William Herschel Telescope operated on the island of La Palma by the Isaac Newton Group of Telescopes in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias under programmes W1999A/CAT42 and W/2011A/P23 respectively. This paper makes use of data obtained from the Isaac Newton Group Archive which is maintained as part of the CASU Astronomical Data Centre at the Institute of Astronomy, Cambridge.

T.L.K. gratefully acknowledges support fom the UK Science and Technology Facilities Council (STFC, grant number ST/T506503/1). M.M. is supported by European Union's H2020 ERC Starting Grant No. 945155–GWmining and Cariplo Foundation Grant No. 2021–0555. D.S. acknowledges support from the UK Science and Technology Facilities Council (STFC, grant numbers ST/T007184/1, ST/T003103/1 and ST/T000406/1). J.C. acknowledges support by the Spanish Ministry of Science under

grant PID2020-120323GB-100. J.T.W. acknowledges support from NSF grants PHY-1806824 and PHY-2110460.

Postscript

After the publication of this manuscript, an updated search based on the corrected ephemeris was conducted (see Whelan et al. 2023). Although no concrete detection of continuous waves was made, the ephemeris enabled a search of similar sensitivity to the prior LVK searches, but with $\sim$3 times less computation time. The wider GW community remain hopeful for a positive detection in the LIGO-Virgo-KAGRA O4 observing run, 18 months in length, and with a factor 2 greater sensitivity than the prior O3 run. Additional monitoring observations of the Sco X-1 system are also being undertaken with the High Resolution Spectrograph (HRS) on the South African Large Telescope (SALT; Buckley et al. 2006) at a weekly cadence as part of a 4-semester long-term proposal I am leading, to provide further phase constraints for the LVK O4 observing run. Figure 4.12 shows the most recent spectra obtained at the time of writing. The SALT HRS instrument has a comparable ($R \approx 40,000$) spectral resolution to UVES in our chosen configuration, and we achieve comparable signal-to-noise with each visit (900s integration time), even during bright time. Based on Monte Carlo simulations, this $\sim$ weekly sampling will deliver $\sim$ 30s ephemeris accuracy across O4, uniquely enabling the most sensitive CW searches thus far for Sco X-1, and reducing reliance on the poorer-quality WHT data. One particular region of phase space of interest is the slight bump in the residuals of Figure 4.4 around $\phi = 0.3$, which may be real and tied to properties of the donor. Only further modelling can help constrain these donor properties, along with further observations which we are currently pursuing.

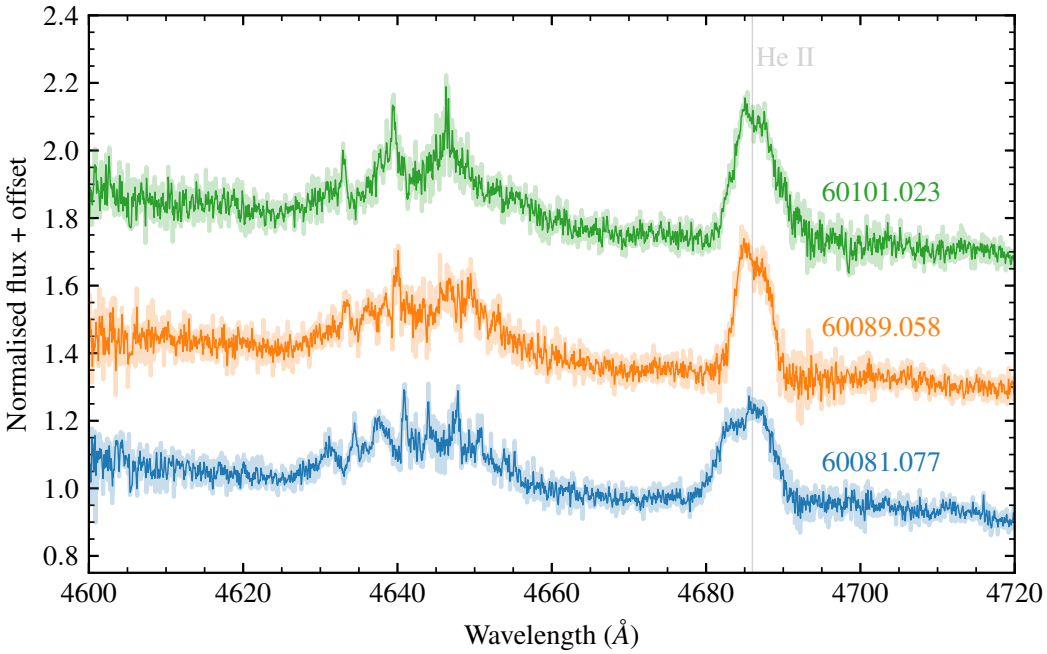

Figure 4.12: Recent spectra obtained of Sco X-1 with SALT/HRS, plotted centred on the Bowen blend and He II lines of interest. Wavelengths are given in the binary rest frame, with systemic velocity and heliocentric velocity corrected for, and times are given in MJD (UTC). Spectra are plotted with a 3-pixel boxcar smoothing for visualisation, with the unsmoothed spectra plotted behind.

Chapter 5

Towards a systematic census of short-timescale variability in supernova light curves

5.1 Introduction

Supernovae do not explode in a perfect vacuum, and interaction with the surrounding circumstellar medium has long been known to play an important role in observed explosions across the transient zoo. Circumstellar material is a powerful probe of the poorly-understood latter phases of massive star evolution. We know that massive stars shed significant amounts of mass prior to supernova from observations of local red supergiants (de Jager et al., 1988; Beasor & Davies, 2018) – as their outer envelopes become less dense and expand, stellar winds are able to drive significant amounts of envelope material off, creating dense circumstellar material. It is thus far unclear whether the dominant mechanism for this occurs on short timescales, in eruptive outbursts (e.g. Davidson & Humphreys 1997), or on longer timescales via a 'superwind' (Heger et al., 1997). Regardless of how the material arrives at large radial extent from the progenitor, the shock front from the supernova traverses this CSM, leading to strong interaction signatures in the light curve (Chevalier & Irwin, 2011) and spectrum that can persist for

weeks after explosion (dependent on the CSM density and radial distribution). CSM interaction leads to faster rising, bluer, more luminous light curves than expected from simplistic radioactive decay models. Some level of CSM interaction is required to reproduce the observed light curves of RSG-progenitor SNe II (Morozova et al., 2017). We also see with some regularity (e.g. Bruch et al. 2022) flash-ionised features present in early spectra of CCSNe, further indicative of the dense CSM from pre-explosion mass loss – with observations constraining mass-loss rates and providing spectral diagnostics of abundances (Ofek et al., 2010). These lines are narrow as they arise from more localised regions in space (i.e. the CSM) and show lower velocity dispersion than the main supernova lines themselves. Further, sufficiently dense CSM can produce detectable non-thermal (synchotron self-absorption) emission, visible in radio and X-ray bands (Chevalier, 1982). Under the assumptions of equipartition, it is possible to directly infer the density of the CSM as a function of radial extent – typically within the context of a power law. Radio emission has been observed extensively in CCSNe (Weiler et al., 2002; Bietenholz et al., 2021), with only null detections in SNe Ia (Panagia et al., 2006; Chomiuk et al., 2016) – as expected from the progenitor channels for these events. Superluminous supernovae (SLSNe, see (Moriya et al., 2018)) have shown 'bumpy' light curves, with at least some evidence pointing towards modulations of density in the 'lumpy' circumstellar material being key to this observed behaviour (Hosseinzadeh et al., 2021).

Although traditionally associated with CCSNe (with pre-explosion mass-loss from a giant progenitor providing the CSM structure), spectroscopic evidence of CSM interaction has been detected in a very small subset of discovered SN Ia (see e.g . Fox et al. 2015). These transients, known as Ia-CSM supernovae, exhibit the same narrow emission lines as found in SNe IIn, superimposed on more typical Ia spectra. The origins of this interaction continue to elude explanation, with suggestions of planetary nebulae or winds of evolved binary companions providing the material (Uno et al., 2023; Sharma et al., 2023). With the explosion in the numbers of transients being discovered by current-generation sky surveys, we are routinely sampling a growing subclass of rarer transients where interaction is important for explaining the observed properties.

For a more comprehensive view of interaction across all families of extragalactic transients, see Fraser (2020).

One regime of study for supernovae that remains poorly probed is their light-curve behaviour on sub-day timescales ($\sim$ minutes – hours). This time interval is largely inaccessible with the typical $\sim$ nightly cadence of current transient surveys, has not been studied extensively from the perspective of theoretical modelling, and suffers from a general dearth of high-quality, homogeneous data suitable for placing strong limits. Some of the earliest high-cadence light curves of explosive transients have been obtained serendipitously through space- based exoplanet surveys. Although the focus of these missions is predominantly high-cadence ($\sim$ minutes) light curves of bright ($V \lessapprox 12$) stars in search of exoplanet transits (e.g. Charbonneau et al. 2000), through stacking of multiple full-frame exposures into and difference imaging it is possible to construct high signal-to- noise light curves of fainter extragalactic transients. Unhindered by the rotation of the Earth, these surveys observe fields in long ($\sim$ 30 day) stares providing uninterrupted coverage, and in turn unprecedented opportunities to sample the very-early temporal evolution of transients.

The *Kepler* (Borucki et al., 2010) and Transiting Exoplanet Survey Satellite (TESS; Ricker et al. 2015) missions have been the primary sources of early-time high-cadence photometry thus far. A growing number of transients have serendipitously exploded within the field of view of these surveys, and have provided stacked $\sim$hourly photometry of these targets for many weeks. Although illuminating, such studies suffer from systematics from difference imaging, limited by the large PSF and plate scale of *Kepler* and *TESS*, and are intrinsically limited to the brightest transients owing to the small aperture sizes involved. Nevertheless, the repurposing of these missions towards transient science has proven to be fruitful for our understanding of the very early time light curves of a diverse range of supernovae – probing short-timescale variability on typical hours–day long timescales. Vallely et al. (2021) studied the very early time light-curves of a core-collapse supernovae serendipitously observed by the *TESS* satellite, achieving precise measurements of the rise times and inferring the presence of shock breakout. More recent work (Wang et al., 2023a,b) using *TESS* has unveiled rapid evo-

lution at early times in stripped envelope supernovae associated with shock breakout from the progenitor (Waxman & Katz, 2017). Similarly, work using the *Kepler* and K2 missions have shown hints of short-timescale early-time bumps in the light-curve, a potential signature of the interaction of the supernova blast wave with a nearby companion star (Shappee et al., 2019; Dimitriadis et al., 2019; Li et al., 2019). Very few studies have thus far searched for high-cadence optical variability using larger telescopes – enabling fainter targets in more distant hosts to be characterised, without the significant caveats involved with stacking shallower data. In the following sections, I summarise some of the key works.

Fast blue optical transients (see Section 1.2) have proven enigmatic: showing behaviours completely incompatible with standard supernovae, unclear progenitor systems, and relative rarity in comparison to other transient families. The recent discovery of optical flaring in AT 2022tsd (the 'Tasmanian Devil'; Ho et al. 2022b) has renewed interest in the prospect of short-timescale variability in transients. Whilst the source of this variability remains unclear (alongside the progenitors, emission mechanisms, and other key parameters) of FBOTs, using arguments of causality, the variability must be occurring on scales of 10^{12}cm $(\sim c\delta\, t)$. There is an emerging evidence for FBOTs being powered by highly-collimated outflows launched from nascent compact objects –making engine-driven variability a natural candidate for this. Quite how the engine experiences factor $\sim$100 changes in luminosity on such short timescales remains unclear. It should be noted that short-timescale variability is commonplace in more energetic transients like GRBs (MacLachlan et al., 2013) (in the high-energy (X-ray – γ- ray) bands), so it is perhaps not surprising that transients bearing many of the same hallmarks (relativistic outflows, extreme luminosity, stripped progenitors) also show similar behaviours.

5.1.1 The curious case of SN 2014J

Putting aside exotic transients, one normal supernovae has been claimed to show such short-timescale variability. Observations of the nearby Type Ia SN 2014J (Bonanos & Boumis, 2016) at high cadence revealed tentative evidence of short-timescale oscillations in the light curve, on a typical scale of around 50 mmag. These oscillations were

consistent across both the Johnson B and V bandpasses, and were seen consistently across 4 nights of observations, disfavouring any photometric errors or systematic effects. This remains an unexpected and puzzling observation – as a SN Ia there is no central engine to provide engine-driven variability and light curves are driven by largely-predictable ^{56}Ni decay at early times, strongly suggesting the source of variability must be extrinsic to the remnant itself. Furthermore, deep non-detections in the radio (Pérez-Torres et al., 2014) and X-ray (Margutti et al., 2014) close to the time of these optical observations are inconsistent with the presence of a dense CSM around the SN. Considering the causal timescale $c\Delta t$ of such variability, the driving mechanism must operate on scales of $\sim 10^{12}$ cm, far smaller than the spatial extent of the supernova ejecta (estimated as $v_{ej}\Delta t_{expl}$). As the brightest nearby supernova in a decade, SN 2014J received extensive spectrophotometric coverage (e.g. Cox et al. 2014; Goobar et al. 2014; Patat et al. 2014) – none of these spectra bear the emission lines (e.g. He II) traditionally seen in transients where CSM interaction is important. Inspired by this event, further observations were presented of 5 supernovae in Paraskeva et al. (2020), which showed no evidence of short-timescale variability, although this study primarily observed fainter targets so is less sensitive than the SN 2014J observations.

Assuming that the variability seen by Bonanos & Boumis (2016) is related to the circumstellar medium, detections of photometric signatures of circumstellar interaction could provide a method to probe the density structure of material surrounding transients, complementary to the traditional spectroscopic approaches in providing independent constraints. Regardless of the mechanism powering these observed fluctuations, it is crucial to verify their existence independently. One dataset is not sufficient evidence, and it is still unclear how this may manifest in other (non- thermonuclear) types of transients. With the recent discovery of short-timescale variability in FBOTs, there is now additional motivation to establish rates of occurrence, and further provide strong upper limits across a diverse range of transient types. The question remains: is this behaviour not being observed in regular supernovae simply because we aren't looking, or is this phenomenon a clue to the origins of FBOTs?

In this chapter, I present the results of a pilot program using the South African

Large Telescope (SALT; Buckley et al. 2006) and Liverpool Telescope (LT; Steele et al. 2004) building on existing literature studies – by searching for excess short- timescale variability using high-cadence, single-colour photometry obtained of two bright core-collapse supernovae in nearby ($\lesssim 100$ Mpc) galaxies. Using these datasets, we place stringent upper limits on the scale of potential variability through the light curves obtained, and develop the techniques necessary to extend this to a larger sample currently being gathered.

5.2 Targets and observations

SN2021acya (Tonry et al., 2021) was discovered by ATLAS, and later classified by the ePESSTO+ collaboration (Ragosta et al., 2021) as a SN IIn on account of it's strong, narrow metal emission lines. Although distant at 290 Mpc, the strong He II signature, broad $H\alpha$ lines, and bright absolute magnitude made this a priority target for high-cadence follow-up. This target received continued follow-up and was later re-classified as SLSN IIn, making it an even more compelling target for study. SLSNe II are well-explained via circumstellar interaction, with the augmented luminosity arising from a greater CSM mass (e.g. Moriya et al. 2018) – making the prospects of interaction-induced variability signatures more promising in this target.

SN2022mm was discovered by ATLAS (Smith et al., 2022; Tonry et al., 2022) in a nearby [] galaxy and showed a rapid rise of ≥ 0.6 mag/day. Of central interest was the strong explosion constraint, with a non-detection just 24h before. We proactively triggered LT/RISE on this transient prior to it being spectroscopically classifed to obtain a high-cadence early-time light curve. On the same night, this transient was classified by the ePESSTO+ Collaboration (Reguitti et al., 2022) as a normal Type II supernova, showing very weak $H\alpha$ emission, and later revealed emerging P-Cygni features.

These two targets provide a direct comparison (albeit at different epochs) – a transient likely to host dense circumstellar material (in the case of SN 2021acya), and a transient with very weak/no interaction (in the case of SN 2022mm). If interaction were to be a dominant mechanism in producing high-cadence variability, we would expect

Table 5.1: Summary of high-cadence observations

Target	Instrument	Phase (discovery)	Run length (h)	Cadence (s)
SN 2021acya	SALT/SALTICAM	+41.5	1	2
SN 2021acya	SALT/SALTICAM	+84.4	1	10
SN 2022mm	LT/RISE	+1.4	1	30
SN 2022mm	LT/RISE	+25.4	1	45

to see it more strongly in SN 2021acya, with SN 2022mm showing no signatures of interaction. The details of observations undertaken are summarised in Table 5.1, along with filters used, cadence, and instrument used.

5.2.1 SALT/SALTICAM observations

We obtained 2 epochs of high-speed photometry on SN 2021acya using the SALTICAM instrument (O'Donoghue et al., 2006) on the South African Large Telescope (SALT; Buckley et al. 2006). In the first epoch, SALTICAM's frame transfer mode was employed to minimise dead time and maximise cadence, yielding a usable (vignetted) field of view of 4'$\times$8' using the Sloan i' band filter. In the second epoch, we obtained full-frame images using the Sloan g' filter at 10 second cadence. Raw frames were processed using the SALT science pipeline (Crawford et al., 2010), with additional corrections performed by the authors to account for additional instrumental systematics. The most important of these (removing pickup noise) is summarised in Section 5.2.2. We also apply a two-step flatfielding procedure using calibration screen exposures as recommended to correct for both the time-varying vignetting pattern induced by SALT's tracker-based optical design, and the pixel-to-pixel variations associated with the detector. For each science frame, we compute a 'low-frequency' flatfield by estimating the spatially-varying background across the illuminated portion of the frame with `sep`. For the 'high-frequency' pixel-to-pixel variation, we divide out the smooth variations in illumination from the calibration screen images with the same procedure as the science frames, then compute a median stack of the normalised calibration screen images. The final science 'flat field' is computed as the low-frequency illumination correction multiplied by the 'high-frequency' calibration frame, and is markedly more effective at

correcting pixel defects.

5.2.2 Pickup noise correction step

The data taken in the Sloan i' filter suffer from variable pickup noise, which adds additional uncertainty on the background flux measurements, especially paired with the ~ 20 pixel drift across the window of the observation. Although the photometric precision we obtain with these data are already more than adequate for our studies, failing to correct for this adds significant correlated noise which propagates into our light curves. We apply a simple empirical model to correct for this noise component – this provides a model-independent correction and significantly improves correlated noise in the images (and final light-curves). Parametric models fail to fully capture the complex morphology of this pattern, which seems to have prominent sub-harmonics between the main peaks. The pattern is also poorly-localised and time-variable in Fourier space limiting the utility of Fourier- based approaches (e.g. Brault & White 1971)

For each raw frame, we rotate the image to a series of trial angles θ_i, then average along image columns, creating an sigma-clipped median profile along the given θ_i. We sweep through values of θ_i, seeking to maximise the chi-squared value to maximise the structure present in this estimate of the background. This is equivalent to accumulating signal along the lines of the pickup noise pattern. An example of this process is shown in Figure 5.1, along with some examples of the inferred pattern. To mitigate issues with evaluating a grid of rotation angles, we interpolate the 5 highest χ^2 values with a quadratic, and take the analytic maximum of this polynomial as the 'optimal' value θ_{opt}. Once this is found, we fit a high-order Chebyshev polynomial to the column-averaged signal and subtract it, to remove any row-wise variation as a result of not having flat-fielded yet. This yields a good estimate of the pickup noise, which we then rotate back into alignment and subtract off. This pickup noise is additive (i.e. as a result of the detector readout process),therefore we subtract it before applying any further corrections. This process must be repeated for each individual amplifier, as we observe each has a slightly different noise pattern. The improvements afforded by each step are shown in Figure 5.2. This correction step has the potential to significantly im-

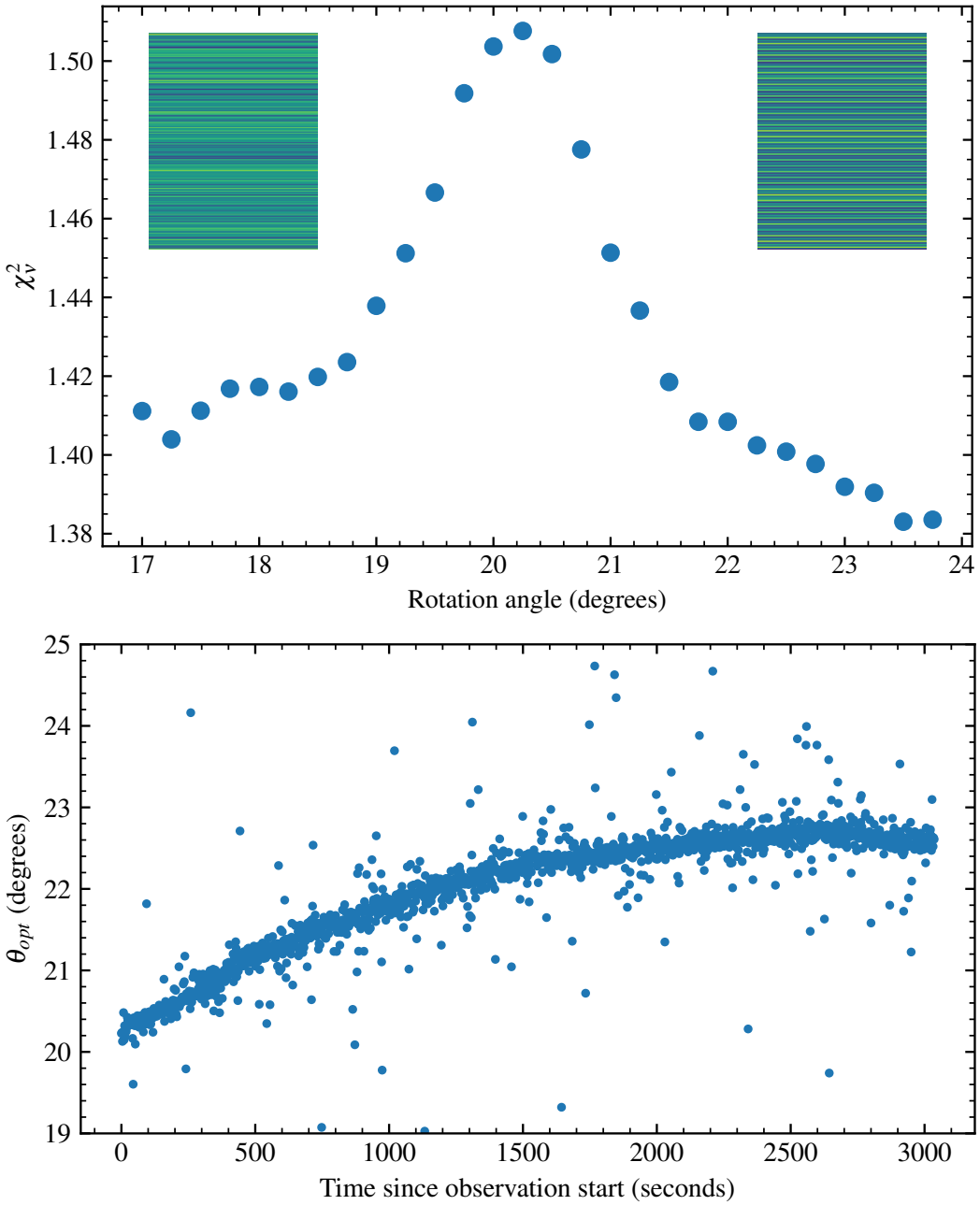

Figure 5.1: Top: Reduced chi-squared of the pickup noise estimator as a function of rotation angle. The two inset images show the inferred pickup noise pattern prior to rotation back into the detector frame at $\theta = 18^\circ$ (left) and the optimal $\theta = 20.25^\circ$ (right). The structure in the angle with the maximum reduced chi-squared value is markedly more defined. **Bottom:** evolution of the optimal rotation angle θ_{opt} throughout the observational series. The smooth variation is unexpected: with the most plausible explanation being of a cable changing orientation as the telescope tracks.

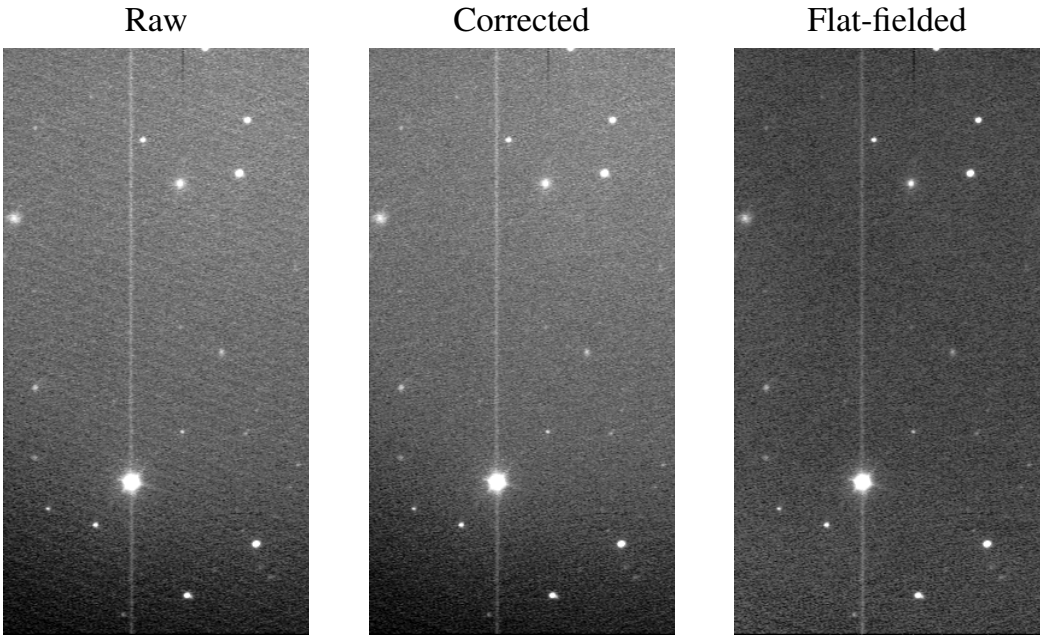

Figure 5.2: SALTICAM data in raw, pickup noise corrected, and pickup noise corrected and flat-fielded form. All frames are plotted with the same scaling to emphasise the improvements in image quality resulting from these corrective steps. We plot only a single amplifier here (256 pixels across) for more compact visualisation.

prove many archival SALTICAM datasets, therefore the method is presented here in the hope it is useful to the wider community. Accurate flatfielding is currently a significant limiting factor in the usage of SALTICAM imaging in general, not just for differential photometry as we perform here. The resultant light curves are shown in Figure 5.3.

5.2.3 Liverpool Telescope/RISE observations

We also obtained 2 epochs of 30s/45s cadence photometry using RISE (Steele et al., 2008) on the Liverpool Telescope (Steele et al., 2004). Observations were taken with permanently-mounted $V + R$ band filter chosen for maximal throughput , with 1x1 binning in frame-transfer mode. Data are reduced with the standard Liverpool Telescope reduction pipeline utilising library calibration products,and were retrieved directly from the LT archive. We use the pipeline products as-is in our analyses, and absorb any additional systematic effects arising from this choice into our detrending model. Our second epoch required a longer integration time than the first. The resultant light curves (after processing) are plotted in Figure 5.4

5.2.4 Common photometry pipeline

To minimise systematic errors and obtain consistent photometry, we use a common set of techniques between the SALT and LT datasets. Both sensors have unilluminated patches present that may cause issues if not dealt with – to mitigate this we build a coverage mask for each image. We first threshold the image based on an empirical sampling of the residual illumination present in these regions, then apply a binary dilation of 30 pixels to ensure we also mask the 'edges', regions where the illumination changes sharply and may cause issues. Aperture photometry with `sep` (Barbary, 2016) is used to obtain the raw fluxes of each source included in our ensemble, and we optimise the aperture size using a curve-of-growth approach by choosing the aperture size that maximises the signal-to-noise of each source. Sky background is subtracted with a sky annulus of inner radius 10 and outer radius 14 pixels. We also obtain a per-frame estimate of the seeing by computing the half-flux radius of the supernova to include in latter detrending algorithms, as the amount of flux caught by each aperture

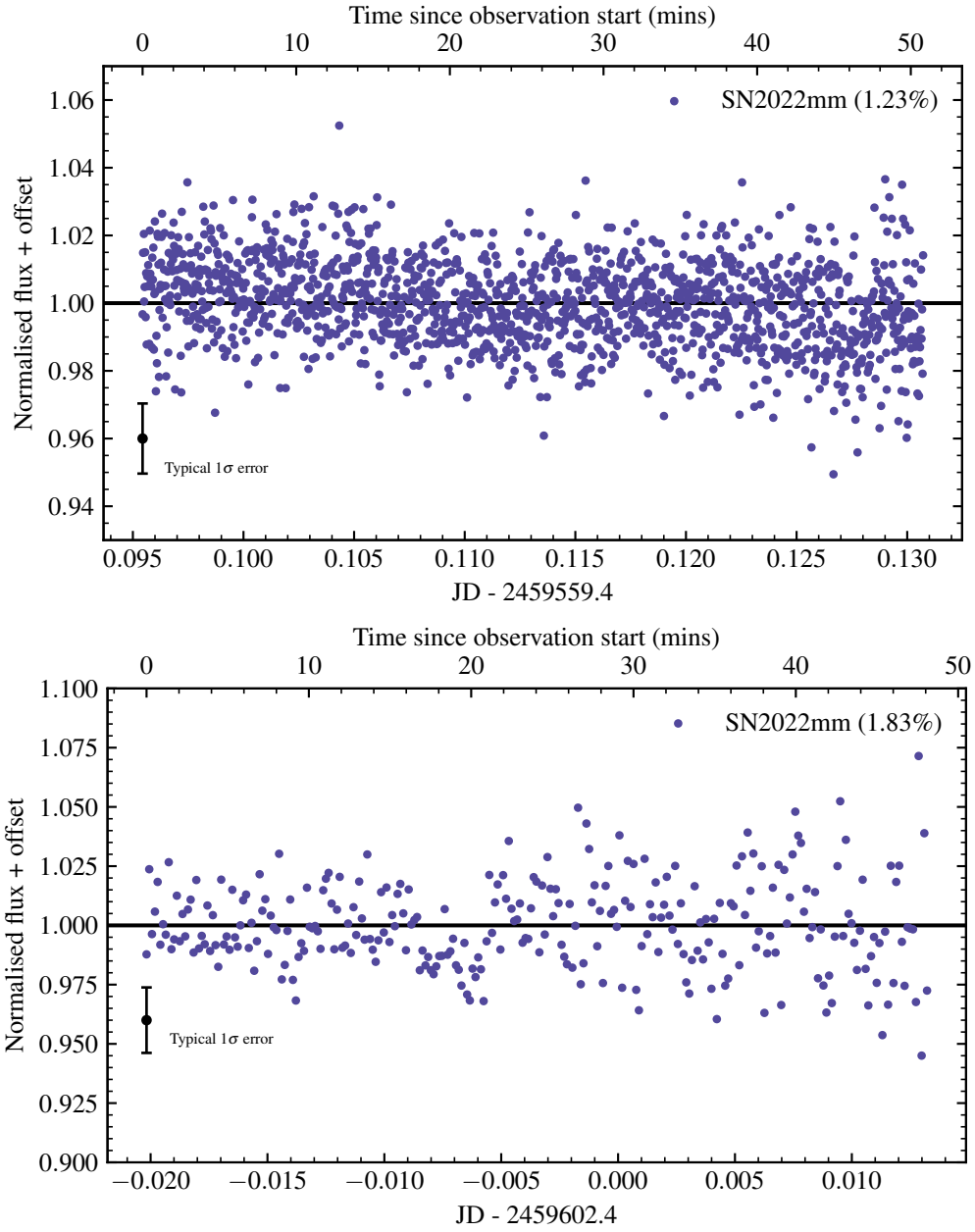

Figure 5.3: SALTICAM light curves of SN 2021acya

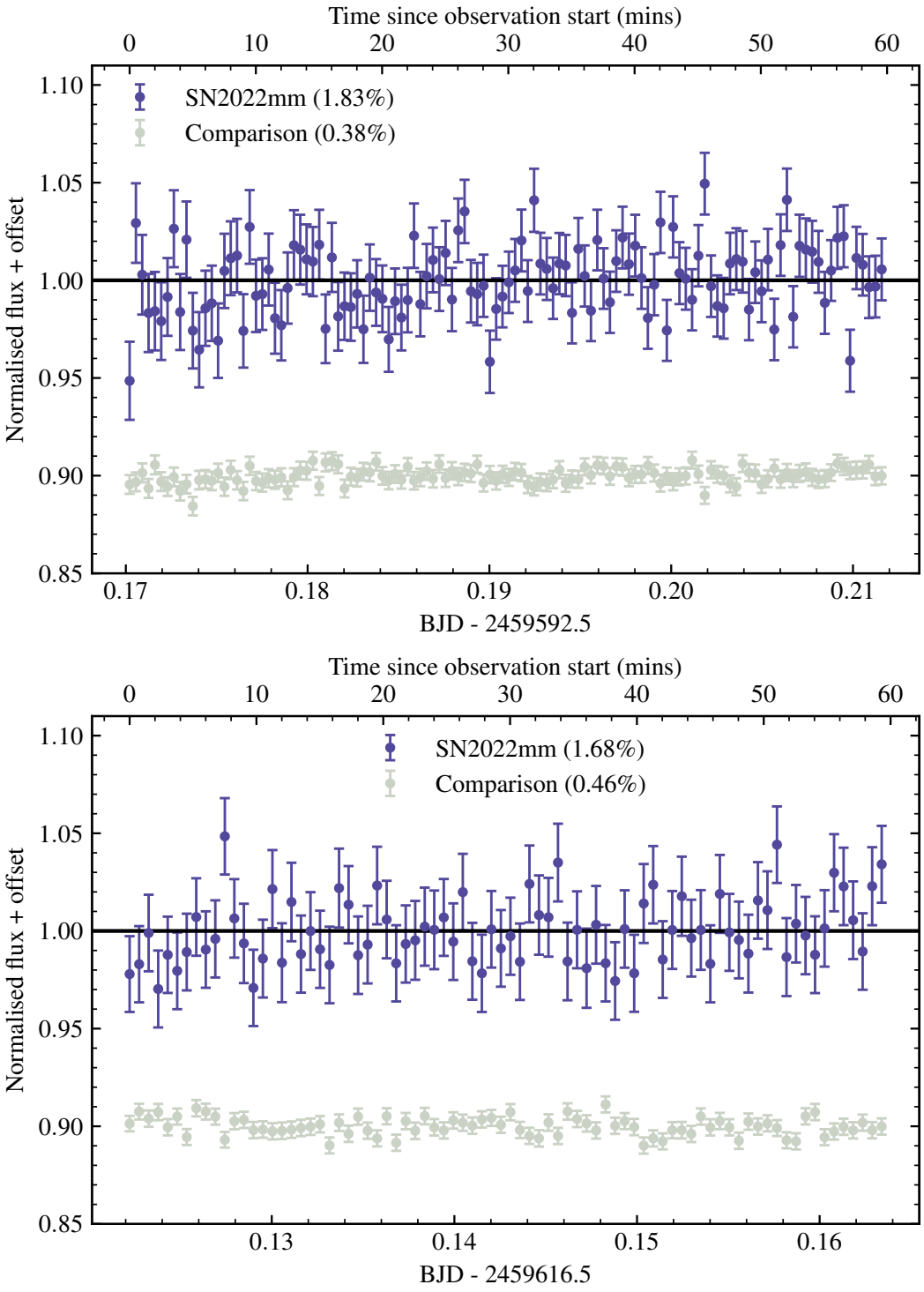

Figure 5.4: Detrended LT light curves of SN 2022mm

will change if the seeing does. The first image in each set is solved astrometrically with `astrometry.net` (Lang et al., 2010), then track arbitrary drifts in source position across the detector by centroiding at the location of each aperture. This reliably re-positions the aperture in each frame and allows measurement of the x-y positions of each source for diagnostic purposes/detrending.

5.2.5 Detrending instrumental systematics

Despite the use of differential photometry, some of our light curves suffer from residual systematic effects. We apply two different approaches to this, depending on the number of stars available. For the Liverpool Telescope data with a wide field of view, we apply the SYSREM (Tamuz et al., 2005; Mazeh et al., 2007) algorithm to the raw fluxes of all ensemble stars. This approach iteratively removes correlated trends between stars in the ensemble, and corrects for effects with polynomial dependence on 'airmass' and 'colour', removing common-mode systematics. Our SALT data contains less suitable stars to compare to, so we use simple differential photometry to generate our light curves. Despite our flat-fielding efforts, stars distant from the supernova on the detector have significant systematic shifts in flux.

For both datasets, we also apply detrending with a basic linear model using instrumental vectors. This approach can successfully deal with linear-order effects in state vectors without overfitting and removing potential variability. We specifically use $x, y, FWHM$ for the target and comparison stars, and avoid using any flux-based vectors as these are co-linear with the final flux. We solve the following matrix equation for each light-curve, and subtract the result from the differential light curve.

$$Ax = B$$

where A is the column vector $[1, x_i, y_i, FWHM_i]$ corresponding to the target state vectors and B is the vector of differential fluxes f_i. We perform the fit using zero-centred normalised fluxes for numerical stability. For the SALT/SALTICAM dataset in particular (which has a small usable field of view and thus few nearby comparison stars), this ap-

proach bears a significantly lower risk of overfitting than PCA-like approaches that are commonplace in wide-field variability surveys (Kovács et al., 2005; Waldmann, 2014; del Ser et al., 2018). Our light curves are then normalised to the median flux for ease of interpretation, and clearer visualisation.

5.3 Light curve analysis

With calibrated light curves, we can now begin to assess the significance of any potential variability present, and place upper limits on detectable variability given the noise present. Temporal correlation and quasi-periodic oscillations (QPOs) may also be instructive – with a robust velocity measurement from spectra these can be converted into length scales in the ejecta, revealing potential mechanisms at play. Simultaneously however, we must be cautious, as instrumental systematics often show short-timescale temporal correlations (red noise) and aliasing on the observation frequency. This section presents some preliminary analyses. We compute the Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982; Zechmeister & Kürster, 2009) for all light curves, to probe for periodic signals in the data – the resultant periodograms are shown in Figure 5.5. No significant variability aside from instrumental aliases are detected in either supernova. There is a slight excess of power visible in the Epoch 1 SN 2021acya light curve that exceeds our chosen false-alarm level of 5%. However, this lies close to half of the total timespan of our data, making it likely this is a spurious peak caused by remaining low-order trends in our data – potentially associated with the rising flux of this object. As a notional measure of the dispersion of the light curve, we quote the inter-quartile range (IQR) and A_{90} (the difference between the 5th and 95th percentile) as robust estimators. These values are tabulated in Table 5.2 for both targets. The dip in the second epoch of SALTICAM is not thought to be real, and remains as an artifact of the flat-fielding procedure. All light curves are qualitatively compatible with non-variability at our photometric precision, with the SALTICAM data providing some of the strongest constraints. More sophisticated analyses are needed to directly place constraints on any short-timescale or correlated variability, but this is statistically chal-

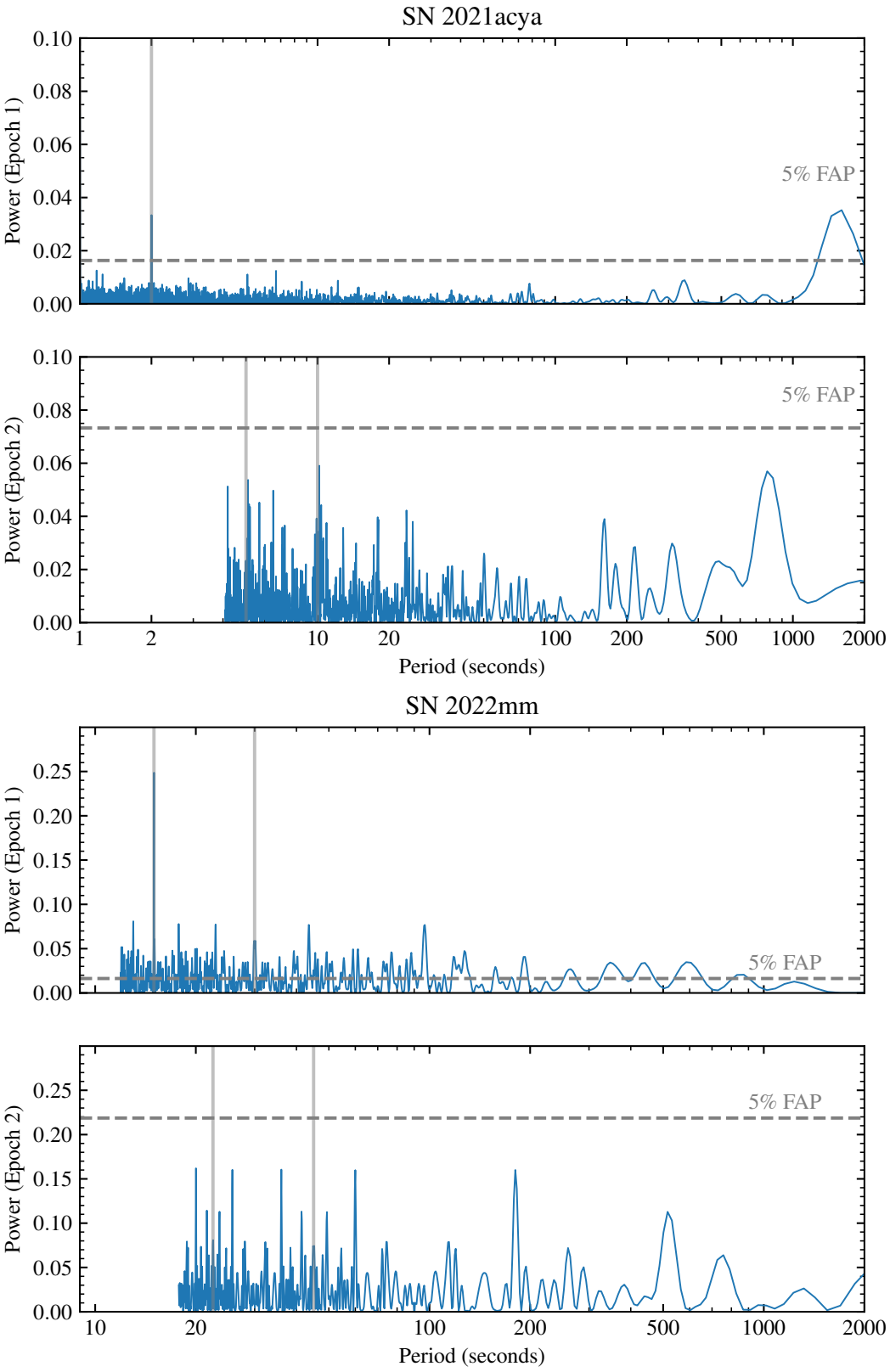

Figure 5.5: Power spectra of the light curves of SN 2021acya and SN 2022mm. The grey vertical lines correspond to aliases of one and two times the instrumental cadence,where we expect spurious peaks. The horizontal dashed line corresponds to the 5% false alarm level, as estimated by the procedure in Baluev (2008)

Table 5.2: Table of upper limits on short-timescale supernova variability

Target	Instrument	Phase (discovery)	IQR (%)	A_{90} (%)
SN 2021acya	SALT/SALTICAM	+41.5	1.65	2.97
SN 2021acya	SALT/SALTICAM	+84.4	2.34	4.90
SN 2022mm	LT/RISE	+1.4	2.47	4.51
SN 2022mm	LT/RISE	+25.4	2.14	4.05

lenging, and likely requires better characterisation of the photometric noise than performed here. Non-parametric approaches such as Gaussian processes (see Gibson et al. 2012) may have greater success here in extracting variability in the presence of correlated noise, although such modelling efforts are left to future work as care is required to avoid overfitting and removing real astrophysical variability. Techniques heavily utilised in the study of active galactic nuclei (e.g. Kozłowski 2016) may also be of use in characterising correlated variability, although we note that these are typically used on longer-timescale data. We leave implementation of these to future work, where a larger sample of data will be available to investigate the nuances of these approaches.

5.4 Conclusions

Although both the SALT/SALTICAM and LT/RISE datasets present null detections of short-timescale variability, there is significant red noise present in our data that limits our sensitivity to variability on timescales of $\sim$ minutes. Nevertheless, the data obtained rule out variability at the levels presented in Bonanos & Boumis (2016) in both SN 2021acya and SN 2022mm. The modest sample of supernovae presented here is not enough to draw a firm conclusion as to the existence of short-timescale variability, but taken in conjunction with the literature results of Paraskeva et al. (2020) begins to disfavour the presence of such phenomena (at least of the persistent kind seen in Bonanos & Boumis 2016). Studying this behaviour across a diverse range of supernova subtypes, each with different explosion mechanisms and environments, will help confirm the presence of variability, constrain the source mechanism, and enable robust measurements of environmental parameters. Even null detections are directly informative, providing the data required for some population-level estimates of (non-)variability.

Further, systematic observations of fast blue optical transients are necessary to begin to assess the prevalence of short-timescale variability going forward.

The numbers of transients being generated by modern transient surveys provides a rich sample to extend this work to a larger population of bright, local transients, to provide further number statistics to constrain rates. Using the ZTF Bright Transient Survey (ZTF-BTS; Perley et al. 2020) sample, roughly 40 transients per semester per sky have a peak magnitude less than $r = 16.5$ – this threshold is chosen to facilitate high signal-to-noise light curves on 2m-class facilities. With the pilot study presented here, we now have the methodologies required to test for the prevalence of short-timescale variability at scale.

Postscript

At the time of writing, 9 further transients have been observed as part of a successful ESO program in P110 using ULTRACAM (Dhillon et al., 2007), a high-speed simultaneous tri-colour imager mounted on the ESO 3.6m New Technology Telescope. ULTRACAM delivers zero-readout-time imaging in u', g' and r/i' bands over a ~ 4 arcminute field of view, with exquisite calibration and levels of systematics. It is perfectly suited to the kinds of study discussed in this Chapter and has a rich heritage of galactic time-domain studies (e.g. Jeffery et al. 2004; Brinkworth et al. 2006; Paice et al. 2018). With P111 upcoming at the time of writing, we hope to extend this study to a further 6 transients using a 3 day time allocation. We also obtained a further 3 triggers from the Liverpool Telescope on young core-collapse supernovae. Analyses of these datasets are ongoing, although conditions were sub-optimal for these observations. Analysis of this dataset is ongoing, with focus being placed on developing the robust statistical methodologies needed to perform a full injection-recovery simulation for each target.

Chapter 6

Building the next generation of context-aware source classification algorithms

Note

This chapter is based on the in-preparation manuscript *moriarty: a data-driven, open source contextual classifier and knowledge store for time-domain astrophysics*. Detailed characterisation, data-driven optimisation, and implementation of machine-learned transient classification is left to this manuscript, where here we focus on the construction of the databases and the underlying algorithms.

6.1 Moving beyond real-bogus classification - multi-class contextual classifications

With the successful implementation of deep-learned real-bogus classifiers (Gieseke et al., 2017; Duev et al., 2019; Killestein et al., 2021), human effort has now shifted from sifting vast through large quantities of false positives to identifying the most interesting 'real' transients to follow up.

Emerging approaches make use of multi-label classification, aiming to predict the class of a given candidate from a coarse, high-level taxonomy of sources. Such a classification task is inherently hierarchical – a SN Ia for example belongs to the family of SNe, which is in turn a galaxy-associated transient, which is in turn a real source. The more granular a classification scheme imposed, the harder a given model will be to train – both from a machine learning (projecting to a high-dimensional output space) and a data perspective (by taking smaller and smaller subsets of the dataset). The potential for 'label noise' also increases significantly on intrinsically rarer targets. For this reason, existing efforts (Carrasco-Davis et al., 2020; Duev & van der Walt, 2021) have focused on classifying into coarse, high-level taxonomies using either ensembles of binary classifiers, or low- dimensional output multi-class classifiers. In many ways, this is a more subtle classification task than real-bogus classification – for example, misclassifying a supernova as a nuclear transient is less problematic than misclassifying the same supernova as a variable star from the perspective of performing follow-up observations. Similarly, for some rare sub-classes of transient (e.g. kilonovae) we have only one, a data volume simply not adequate for training classification algorithms on. Simulated datasets (e.g. PLAsTiCC; Malz et al. 2019) can mitigate the significant class imbalances present – although of course come with the caveat that simulated data may not accurately represent the real world, with errors and subtle systematic effects that are impossible to model *a priori*, and a strong reliance on the correctness of the models being used to derive the synthetic data.

An important distinction is between 'ontological' (actual) and 'phenomenological' (apparent) classifications for a given object (e.g. van Roestel et al. 2021). The apparent properties of candidates (is near a galaxy) may not map directly onto their true nature (is a supernova) – and indeed may be degenerate, with multiple different 'ontological' classes being plausible conditioned on the 'phenomenological' behaviour. Splitting these two classes is complex, and as a result we classify with 'phenomenological' classes for the remainder of this Chapter – the difference between ontological and phenomenological classes is more important in the study of Galactic variable sources than for extragalactic transients. Whether a galaxy-associated transient is a

core-collapse or thermonuclear supernova is largely only something we can tell with follow-up spectroscopy, and with existing solutions these subtleties are largely out of reach. Some early works are beginning to be able to make inferences on the ontological class of transients through contextual information (Gagliano et al., 2021), but should be interpreted with caution owing to the limited sample sizes available.

The end goal of multi-class contextual classification is full automation of transient follow-up – with discovery and vetting of candidates largely automated via high-performance real-bogus classifiers, a multi-class contextual agent can sift the stream of candidates for those that meet specific triggering criteria, and automatically submit observing blocks to appropriate robotic telescope facilities. The issue of remote triggering itself is largely solved, with multiple facilities now offering programmatic (Hessman, 2006) submission of targets to the telescope queue, and robust automated data reduction (e.g. Smith et al. 2016; McCully et al. 2018). The critical 'missing link' that needs to be addressed is accurate context-aware classification of candidates – the following sections detail the issues complicating this, current solutions, and the next-generation contextual classifier I am leading development of: which aims to address many of the existing issues.

6.2 Image and catalog-based approaches

6.2.1 Image-level classification

By definition, an image of a given transient candidate is always available at point of discovery, encoding context on what surrounds the transient. Image-level classification is naturally limited however, by the quality of the input images. The discovery/reference image of a transient may not be deep enough to detect the transient host, or suffer from poor image quality. This can be alleviated with survey data, however this is not always available across the entire sky /region of interest. Images provide purely phenomenological classifications – without distance estimates, all 'alignments' can only be apparent. Nevertheless, faint hosts may be visible in imaging but missed by source extraction, and could be recovered by an image-level approach. Perhaps the biggest issue with this

approach is that Images are also an inefficient representation for contextual information, with the vast majority of pixels in a given image containing only sky background. Representing a large region of sky (to enclose potential associated host galaxies in the local Universe, for example) requires either a prohibitively large, full-resolution stamp (with the storage cost scaling quadratically with stamp size), or a low resolution downsampled stamp bearing interpolation artifacts. Multi-scale stamps (Reyes-Jainaga et al., 2023) containing multiple 'channels' at varying resolution and field-of-view have been proposed as a potential solution to this, although as of yet have not seen widescale adoption. As a test, we trained a multi-class CNN on the training dataset from `gotorb` (see Chapter 3), using the meta-labels from the training set generation code to create 'nuclear transient' (NT), 'orphan' (OR), 'supernova' (SN), and 'variable star' (VS) positive classes, along with the 'bogus' (BS) negative class. OR, SN, and NT are all extragalactic classes, with the distinction being the presence (and distance) of a host. NT-class objects are within astrometric uncertainty of the host galaxy's nucleus, whereas SNe are near to a galaxy and OR have no host galaxy. The architecture and hyperparameters were kept entirely the same,with the exception of setting the output layer to have 5 neurons (for the 5 classes). Figure 6.1 presents the results from the held out test set. It is clear that images are only partially useful in providing contextual classification, in part due to their limited context – with significant confusion between the OR and SN classes, and NT and VS classes owing to their visual similarity.

6.2.2 Catalog-level classification

Astronomical catalogs are a rich source of contextual information for a given part of the sky. Indeed, as humans vetting transients, we often consult these resources in deciding if a given transient is worthy of further follow-up. In a real sense, the 'hard work' is done for us in using these data – the salient information about detections has already been extracted in the form of source photometry/distances/other properties, and contained in a tabular form well-suited to analysis with existing machine learning approaches. Despite the distilled information present in astronomical catalogs, many challenges exist for their usage in machine learning applications.

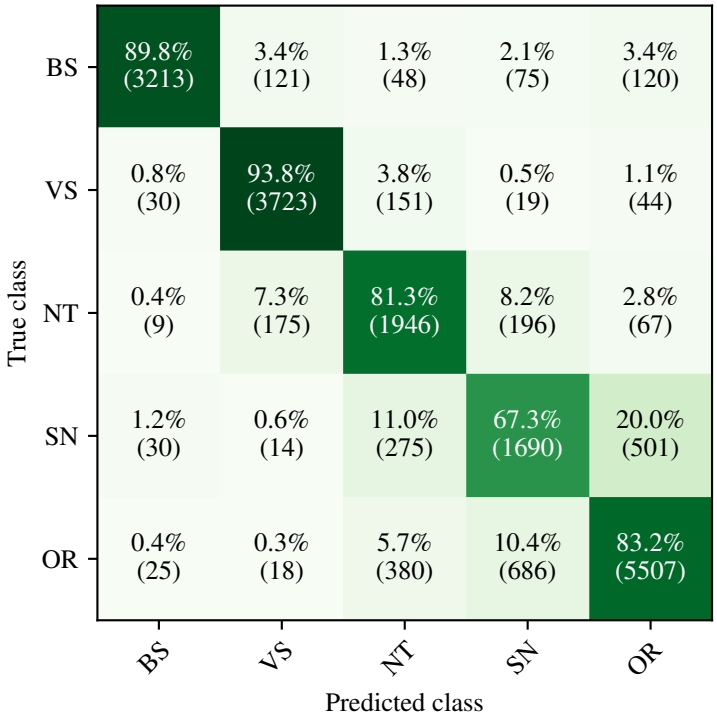

Figure 6.1: Confusion matrix based on the test set for the image-only classifier. A significant proportion of supernova-like transients are mis-classified as orphan sources, owing to the host galaxy not being present in the image.

Incomplete Almost all astronomical surveys are magnitude-limited, displaying a sharp drop in detection efficiency around some characteristic magnitude. One particularly important example to consider is in local Universe galaxy catalogs, with completeness[1] rapidly dropping beyond ~ 40 Mpc (Dálya et al., 2018) in the most up-to-date catalogs. This arises both from photometric non-detection, but also the limited capability to follow up and obtain redshifts for faint local galaxy candidates.

Inhomogeneous Catalogs do not always have uniform coverage of the sky - whether observed at all, or varying depth. The PanSTARRS1 survey, for example, is conducted from a single site, and has negligible coverage below declinations of -30°. Further, many 'catalogs' themselves are composed of data from multiple other sources, each with their own selection biases, deficiencies, and uncertainties.

Incorrect Many source disposition catalogs carry inherent misclassifications: owing to both the limited amount of data they make use of, but also the intrinsic similarity in observables between different source classes. Variable star catalogs quite often contain active galactic nuclei, and quasar catalogs host many hot white dwarfs.

6.2.3 An optimal fusion?

Ideally, any algorithm should be able to combine both image-level and catalog-level contextual information in a principled way – being robust to the potential issues associated with catalog information, whilst also integrating inferences about context from the discovery image. As a prototype, we take the multi-class convolutional neural network (CNN) trained in Section 6.2.1, and feed the output probability for each class alongside the nearest-neighbour contextual source distances into a random forest classifier (Breiman, 2001) implemented in the `sklearn` (Pedregosa et al., 2011) package. The resultant model is a so-called 'meta-classifier' – combining predictions from multiple independently-learned models. Training the model naïvely as-is yields two major issues:

[1] This is further complicated by a lack of consensus on how best to determine the completeness of a given catalog (Kulkarni et al., 2018).

- With no contextual information, the model fails catastrophically, yielding NaN results for all output classes.
- Model performance is drastically reduced as the tree classifier does not optimise well for the missing contextual information.

To mitigate these issues and make a more robust overall classifier, we apply some novel data augmentation. We know that our contextual features may suffer from the issues discussed in Section 6.2.2, so we apply 'contextual censorship' – that is randomly removing contextual information from training items at train time. As a proof of concept, this censorship is applied with a fixed uniform probability of 50% for all entries. This can be thought of as simulating 'missing' sources, as a result of catalog incompletenesses.

Furthermore, by tuning the censorship probability between 0% and 100%, we can directly control the relative importances of the image-level context and catalog-level context. Figure 6.2 illustrates this by plotting the random forest feature importances for P_{SN} (the CNN-predicted probability of the example being a supernova), and the nearest-neighbour GLADE source distance as a function of the censorship probability they were trained at – as a surrogate for image-level and catalog-level contexts. All models were trained on and evaluated on the same splits of the dataset between runs, making changes in feature importance a pure result of the varying censorship. The two features reach equivalence in importance at around 0.2 (20%) censorship, making this a natural choice for the probability at train time. Conversely, if image-derived information is trusted more, a higher value of censorship could be chosen. For evaluation we choose 50% censorship, as we want the image-level CNN to be slightly dominant owing to our incomplete context. Figure 6.3 illustrates the confusion matrix on the held-out test set. By accounting for the lossy nature of the contextual cross-matches in the training process, we are not only able to balance the relative importances of image and catalog-level context, but create a robust model that smoothly interpolates between zero context and full context. This is a desirable property given the large regions of sky that context may be unavailable over, owing to the limited coverage of some surveys. A natural next step is to tune the censorship of individual catalogs as a function of their limiting

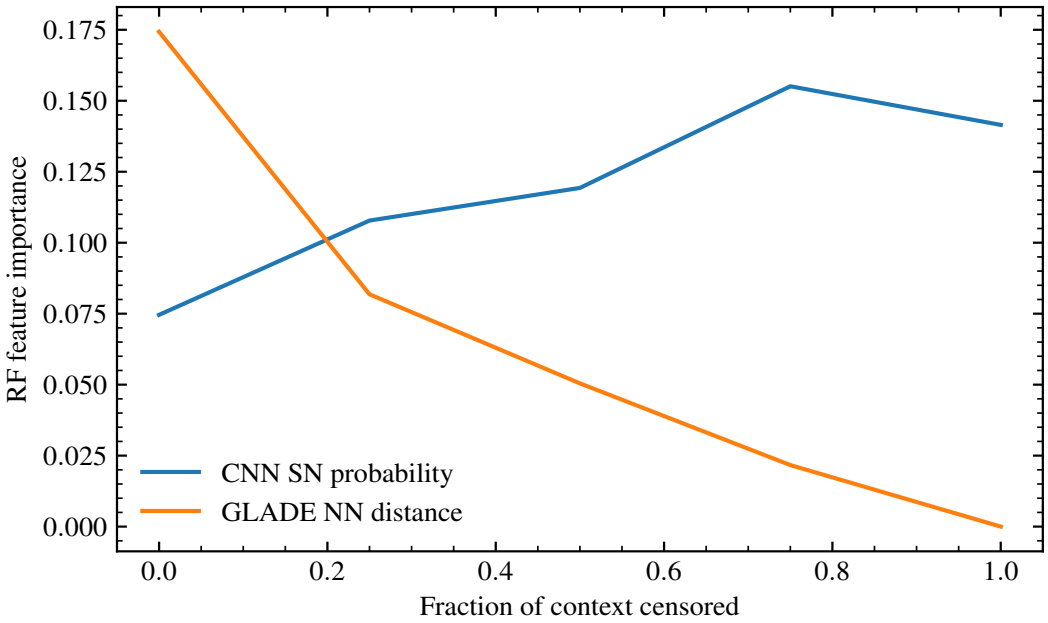

Figure 6.2: Varying contextual censorship and the result on relative importances of image and catalog-level contexts

magnitudes – for example by assuming a probability of rejection proportional to:

$$P_{\text{reject}} = \frac{1}{1+\exp\left(-\sigma_m\left(m-m_{lim}\right)\right)}$$

where m is the detection magnitude, σ_m is a characteristic width over which detection efficiency drops off, and m_{lim} is a characteristic limiting magnitude (here where 50% of sources are recovered). Accounting for mis-identifications, or catastrophic redshift failures (Bernstein & Huterer, 2010) in catalogs with photometric redshifts is also possible with this approach, but is left to future work as this requires more careful treatment to avoid biasing the training set. With a principled way to control the trade-off between image-level and catalog-level context, it now falls to individually optimising classifiers and algorithms operating on each data modality. The next section details `moriarty`, a state-of-the-art contextual agent designed to form the contextual knowledge store for future multi-class classification algorithms.

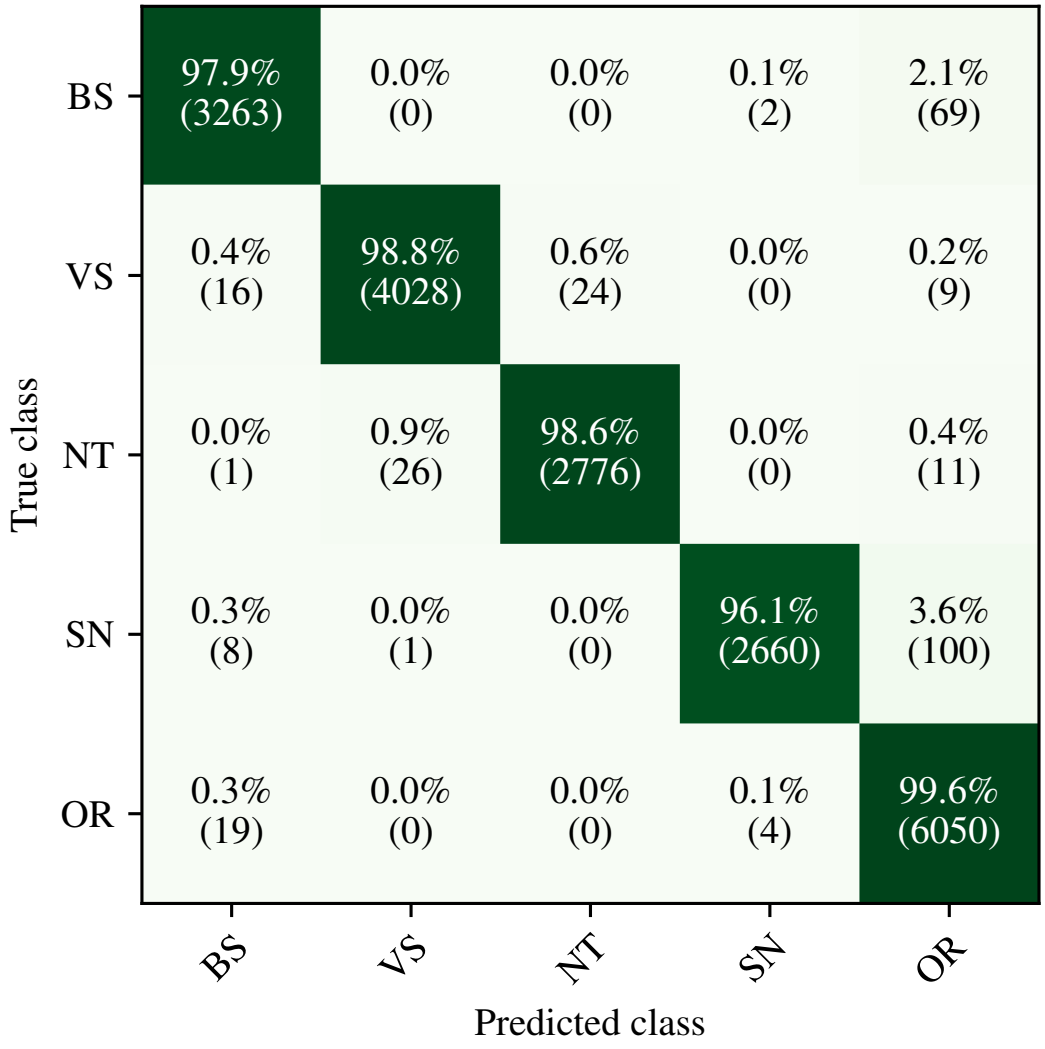

Figure 6.3: Confusion matrix based on the test set for the combined image and context classifier. We observe excellent performance across the classes, with around 3% of supernovae being misclassified as orphan transients.

6.3 `moriarty`: a data-driven, open source contextual classifier and knowledge store for time- domain astrophysics

`moriarty`[2] is named in homage to the current state-of-the-art in contextual classification, the `sherlock`[3] (Smith et al., 2020) algorithm. The frontend code is written entirely in Python, with the backend being implemented as a PostgreSQL database cluster. Special attention is given to interoperability of all catalogs – GOTO is distributed across two hemispheres, unlike some of the catalogs we rely on, and so care is needed to fully make use of all information available in a principled and predictable way. It is planned in future to make `moriarty` available to the community, both as a hosted API where users can submit requests online, and also as a 'build-your-own' solution – so that all can make use of advanced contextual classifiers. The following subsections outline some of the key implementation details, with some rationale behind the specific design choices made.

6.3.1 Catalog selection

Providing a comprehensive contextual coverage of a given patch of sky requires a rich selection of astronomical catalogs to combine and infer information from. To this end we construct a database cluster of 15 catalogs, listed in Section 6.3.1. A full description of the contents of each can be found in the accompanying reference, but at a bare minimum we ingest the coordinates, magnitude, and corresponding filter of each source from each catalog into a main `context_source` base table. At the time of writing, the only remaining catalogs to add to this list are *Gaia* DR3 (Gaia Collaboration et al., 2022), DESI Imaging Legacy Surveys (Dey et al., 2019), and SkyMapper DR3 (Keller et al., 2007). The latter two are important for deep coverage of the Southern Hemisphere, filling in for Pan-STARRS both in terms of deep photometry and photometric redshifts.

[2] The cunning and formidable nemesis of Sherlock Holmes
[3] `https://github.com/thespacedoctor/sherlock`

Name	Disposition	Number of rows	Reference
Pan-STARRS + PS1-PSC + PS1-STRM	-	2,070,490,112	Chambers et al. (2016b) Tachibana & Miller (2018); Beck et al. (2021)
GLADE+ (galaxies)	galaxy	22,431,348	Dálya et al. (2022)
WISE AGN Catalog	agn	4,543,530	Assef et al. (2018)
AAVSO Variable Star Index	varstar	2,235,893	Watson et al. (2006)
Million Quasars Catalog v7.2	agn	1,461,834	Flesch (2021)
ZTF Variable Star Catalog	varstar	781,602	Chen et al. (2020)
GLADE+ (AGN)	agn	750,410	Dálya et al. (2022)
ATLAS Variable Star Catalog	varstar	427,196	Heinze et al. (2018)
ASAS-SN Variable Star Catalog X	varstar	378,861	Christy et al. (2023)
Zwicky Cluster Catalog	area	9134	Zwicky et al. (1961)
LASR AGN Catalog	agn	4309	Asmus et al. (2020)
Abell Cluster Catalog	area	2712	Abell et al. (1989)
LMXBCAT	varstar	347	Avakyan et al. (2023)
HMXBCAT	varstar	169	Neumann et al. (2023)

Table 6.1: All contextual catalogs currently available in the context database for `moriarty`, listed with their approximate number of rows and reference.

To provide the deepest source context possible, we create a large derived catalog from multiple different Pan-STARRS data products. We discard most of the PS1 source-level metadata for the purposes of this classifier, to create a lightweight and fast photometry table tailored exactly to our needs. The Pan-STARRS1 Point Source Catalog (PS1-PSC; Tachibana & Miller 2018) provides a 'star-galaxy score' between 0 and 1 for 1.5B sources based on a random forest model trained on PS1 photometry. The Pan-STARRS1 Source Types and Redshifts with Machine learning (PS1-STRM; Beck et al. 2021) catalog provides source dispositions and photometric redshifts for the majority of sources detected by Pan-STARRS thus far. Both of these high-level science products are locally crossmatched against the entire Pan-STARRS1 `StackObjectView` photometry table, containing over 2 billion sources. For sources with both PS1-STRM and PS1-PSC entries, we prefer PS1-STRM-provided probabilities, bagging the galaxy and AGN probabilities into one to maximise diagnostic power. We adopt the fiducial $p_{galaxy} \geq 0.65$ figure as suggested in Beck et al. (2021) to select galaxies in practice, with sources less than this being marked 'unclear'.

To provide a unified set of attributes, we give sources from galaxy catalogs a star-galaxy score of zero, and sources from known stellar catalogs a star-galaxy score of one. This has a well-justified probabilistic interpretation for our cross-matching (see Section 6.3.4), and we set a flag on the source to indicate this is a synthetic star-galaxy score.

6.3.2 Database design

As new catalogs are released regularly, the database is designed from the ground up to be modular and support the ingestion (and removal) of catalogs as required, with a consistent schema and query structure. This further extends to the priorities, dispositions and astrometric uncertainties associated with each catalog. Extensive use of table inheritance provides a unified query structure for tables with somewhat disparate attributes. All sources are ingested into a main `context_source` table, with `galaxy`, `varstar`, and `star_galaxy_score` child tables providing additional attributes. The high-level structure of the database is specified using configurable 'views' to present

a tidy and efficient table structure to the user. For fast spatial queries, we make use of the `q3c` PostgreSQL extensions Koposov & Bartunov (2006), by defining spatial indices over each table. Database query times scale with the region of sky searched, as expected, but for a 10 arcminute search radius, queries take $\lesssim$ 400ms on commodity hardware – with use of composite indices mitigating overheads from the multiple joins involved.

6.3.3 Source aggregation and voting

To mitigate the previously-discussed issues with catalogs, we aggregate synonymous sources and employ a voting scheme to robustly determine their properties. We use a joint astrometric uncertainty for our association criteria, computed by adding the astrometric uncertainties of each source in quadrature. Sources within a separation of 3 times the joint astrometric uncertainty are regarded as synonymous, and are passed into the voting algorithm for de-duplication. For each set of synonymous sources the following algorithm is applied:

- Sort the sources in order of priority, and then by the quality of redshift estimate.
- Back-fill this source table with the first non-null value (i.e. the highest quality estimate) as an aggregation
- Determine the overall source disposition via a weighted voting scheme.

Future versions will migrate to a fully probabilistic approach to aggregating synonymous sources, but for now simple cross-matching suffices owing to all catalogs having similar astrometric uncertainties. Future planned inclusions of radio/X-ray source catalogs

6.3.4 Robust galaxy associations

One significant issue with existing catalog-level host association approaches is the use of fixed association radii (known as a 'cone search'). Whilst fast, this approach does not take into account galaxies as extended objects: bright, local galaxies are extended on the sky, and thus transients can be produced and found far from the photo- centre of

the object. Setting the cone search radius too large will lead to spurious associations with implausible hosts, whereas setting it too small will exclude many matches in the local Universe from nearby galaxies. It is clear that the association distance scale must vary with the size/mass/brightness of the putative host to avoid these issues. We seek an approach that sits somewhere between cone searches and advanced image-level approaches such as those in Gagliano et al. (2021) in the runtime-accuracy space.

With this in mind, we construct a probabilistic catalog-level approach, making use of context like potential host magnitudes in our decision making. As a starting point we adopt the $\mathrm{p_{chance}}$ measure (e.g. Bloom et al. 2002) commonly employed in the study of host galaxies of gamma-ray bursts:

$$\mathrm{p_{chance}} = 1 - \exp\left(-\pi\theta^2\rho(m)\right)$$

where θ is the angular distance between transient and putative host, and $\rho(m)$ is the areal density of sources with $m \leq m_{host}$: that is galaxies brighter than the putative host. For ease of computation and to avoid rounding/truncation errors, we work internally with the logarithm of these probabilities instead:

$$\log\left(\mathrm{p_{chance}}\right) = \left(-\pi\theta^2\rho(m)\right)$$

Not all source catalogs in our ensemble carry a B-band magnitude for each source, therefore we convert source magnitudes to a 'pseudo-B' magnitude using a set of synthetic colours calculated from the galaxy template SEDs in Brown et al. (2014). With $\rho(B)$ being defined as a rank statistic, $\mathrm{p_{chance}}$ is robust to errors in the assumption of average colour. We deliberately exclude AGN from this association process, which show markedly different (~ 2 mag) colours dependent on type. We leave the inclusion of this class of object into our galaxy association algorithm to future work, noting that the majority of AGN noted in our chosen catalogs are compact on the sky, and thus any transients occuring within them will be flagged as 'synonymous'.

With the inclusion of PanSTARRS sources in the table, the formula for $\mathrm{p_{chance}}$ must be modified to take into account the probabilistic source classifications provided.

PS1-STRM provides $\mathrm{p_{star}}$ as an entry, and the `sgscore` in PS1-PSC is effectively $(1 - \mathrm{p_{star}})$ (putting aside issues of probabilistic calibration in each). Noting that $\mathrm{p_{chance}}$ and $\mathrm{p_{star}}$ are statistically independent, we can simply multiply together to obtain the joint probability of the transient being associated with the source, and the source being a galaxy:

$$\mathrm{p_{assoc}} = \exp\left(-\pi\theta^2\rho(m)\right)(1 - \mathrm{p_{star}})$$

Even outside the natural probabilistic interpretation, $(1 - \mathrm{p_{star}})$ can be interpreted as a penalty term that downweights potential stellar associations. For sources directly matched to a galaxy catalog (or equivalently a stellar catalog), we can fix p_{star} to zero or one respectively, such that this term only operates on star-galaxy scored entities. This is a simplification, as naturally these catalogs have mis-classifications also – nevertheless this extension enables us to make use of some of the deepest source context available. This approach also works for Southern Hemisphere surveys like SkyMapper and the Legacy Surveys. Figure 6.4 illustrates a typical match resulting from `moriarty` for the calcium-rich gap transient SN 2003H listed in the Dong et al. (2022) sample.

6.3.5 Areal associations

Although the 'source level' context is the more important aspect of the algorithm, associations with galaxy cluster groups can be highly diagnostic for specific classes of transient, and facilitate large-scale searches – for example finding lensed transients (Goobar et al., 2022; Ryczanowski et al., 2023) of key scientific interest, as well as informing the search for sub-luminous transients in the local Universe (Kulkarni et al., 2007; Cai et al., 2021). To enable these key science goals, we build in an 'areal association' routine into the contextual classifier.

Clusters from the Abell et al. (1989) and Zwicky et al. (1961) galaxy cluster catalogs are included in the database by default. These two catalogs, although based on photographic plate observations, provide a relatively comprehensive census of rich, low- redshift clusters with which to make associations. Both catalogs provide an estimated on-sky 'cluster radius' (in arcseconds) with which we define their extent in the

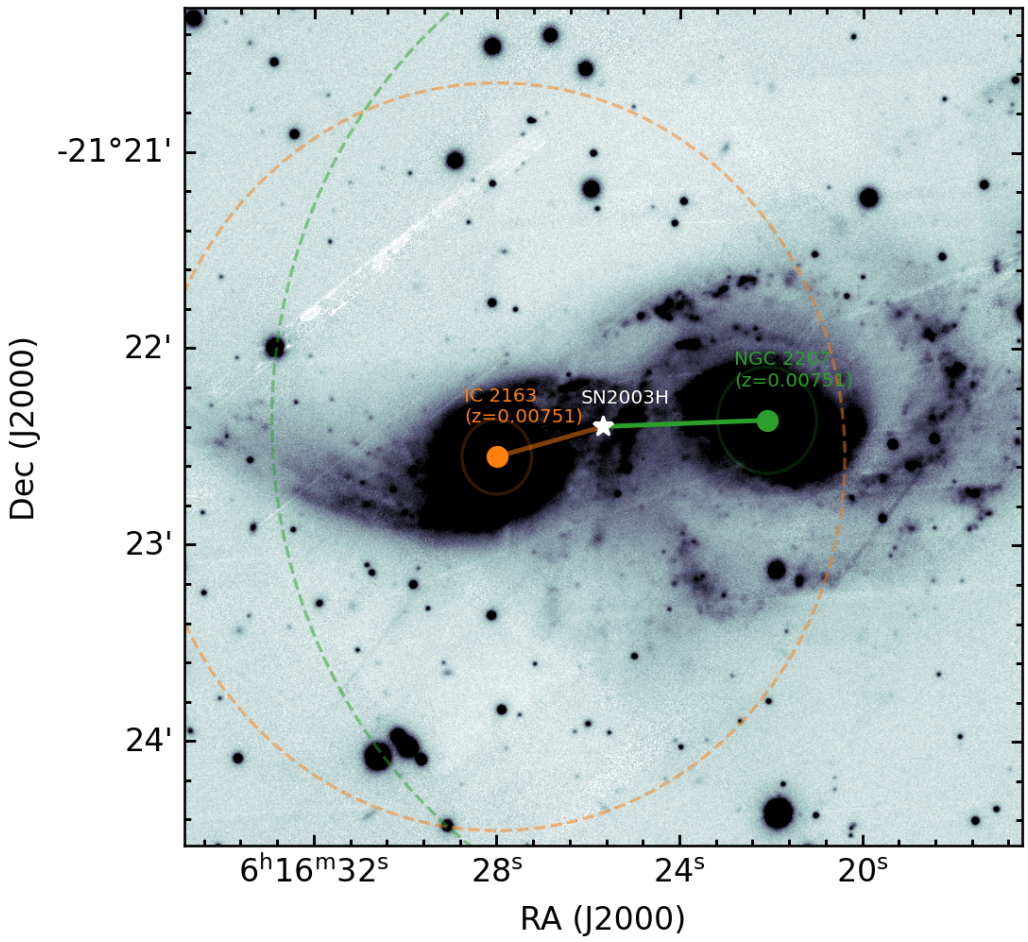

Figure 6.4: Example of the rich contextual output from `moriarty` for the gap transient SN 2003H. Both hosts are associated with the transient, but the correct host has marginally higher p_{assoc} and so is favoured.

database, and the Abell catalog also provides a cluster redshift estimate. Notably, neither catalog includes the Virgo Cluster, the closest galaxy cluster to us ($\sim$ 20 Mpc) as it extends over multiple photographic plates and thus was not recorded. We manually include this as a high-priority entry, and rename a subset of clusters to correspond to their common name (e.g. *Abell 1060* $\rightarrow$ *Coma Cluster*) for ease of use.

As a deliberate choice to be conservative about associations, as well as to enable a unified query structure for more complex regions (e.g. survey footprints, arbitrary sky regions), we assume each cluster's area covers a box with side lengths equal to 2 times the cluster radius. These polygons are then queried using the `q3c_polyquery` routine. Typical queries of the entire database table for a single source take O(100ms), which can be executed in parallel with other operations required for the contextual classification. A natural next step is to extend this framework to include deeper and more complete galaxy cluster catalogs (e.g. Wen & Han 2022), although this is left to future work as the majority of transients that GOTO will discover are at low redshift and so are unlikely to benefit from this. There are also significant completeness issues with galaxy cluster catalogs, owing to the relative errors on redshift – future catalogs backed by large-scale spectroscopic surveys may mitigate this to some extent.

6.3.6 Example outputs

The rich context generated by the algorithms discussed above is propagated through into a human-readable format, presented to the end user as a contextual string. The two examples below are taken from real transients discovered by GOTO, and focus primarily on extragalactic transients.

> GOTO23fi is likely associated with the B=15.16 mag galaxy HyperLEDA 30974 in the gladeplus_galaxies, ps1_stackobjectview_minimal catalogs (8.44″ away). Probability of connection 99.960% – host is at 217.2 Mpc (z=0.0473+/-0.0014), implying a transient absolute magnitude of -18.60 and sky-projected offset of 8.89 kpc.

> GOTO23fg is likely associated with the B=15.84 mag galaxy HyperLEDA

> 139567 in the gladeplus_galaxies, ps1_stackobjectview_minimal catalogs (12.40″ away). Probability of connection 99.915% – host is at 276.9 Mpc (z=0.0598+/-0.0150), implying a transient absolute magnitude of -18.68 and sky-projected offset of 16.65 kpc. This transient is spatially coincident with the clusters ACO 1238 (231.0 Mpc), ZwCl 4067

Future work will tie the full contextual information into an easy-to-browse webpage format, plotting associated sources on the sky and enabling access to the full catalog information behind each source identified by `moriarty`.

6.3.7 Performance verification

As an end-to-end test of the performance of `moriarty`, we evaluate it's performance on the Zwicky Transient Facility Bright Transient Survey (ZTF-BTS) sample (Perley et al., 2020) – specifically all Type Ia supernovae. This numbers 3601 transients in total, 3244 of which are matched to a galaxy with known redshift and redshift error. If `moriarty` returns a host compatible with the quoted redshift at 3 sigma confidence, that is

$$\left| \frac{z_{moriarty} - z_{bts}}{\sigma_z} \right| < 3$$

we mark this as a successful host identification. By this criterion, we identify the correct host with 90% accuracy. This is shown in Figure 6.5. We note that a number of the supernovae in the sample have had redshifts derived via template matching (e.g. Blondin & Tonry 2007), rather than via host galaxy lines or manual association by human eye, and thus have their own (significant) intrinsic error that must be accounted for. There is no flag present in the ZTF-BTS sample to identify these cases, and inferring based on the last significant digit is likely to be 'unfair' to high-redshift transients. Extreme outliers in this normalised residual space are likely a result of cases where we have assigned a host with an extremely low error as a result of a robust spectroscopic redshift. One immediate outcome of applying `moriarty` to this sub-sample is the projected offset distribution of SNe Ia from their identified hosts. We define a 'gold' sample, where the normalised redshift error is less than one, to cut down on the number of misidentifica-

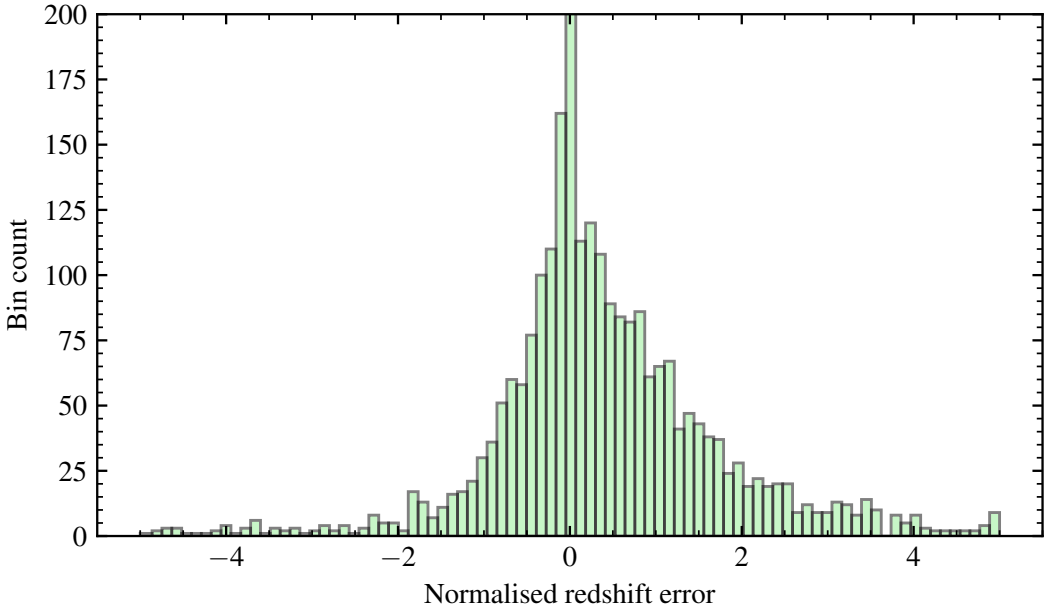

Figure 6.5: Histogram of normalised redshift error for the full ZTF-BTS Ia sample with good spectroscopic matches in `moriarty`. Histogram binning is computed using the Knuth (2006) rule, and we deliberately truncate the bin at zero error, as this corresponds to the human-identified host associations in ZTF-BTS. We also compute the histogram over the range (-5, 5) to remove extreme outliers (see text) and instead focus on the core of the distribution.

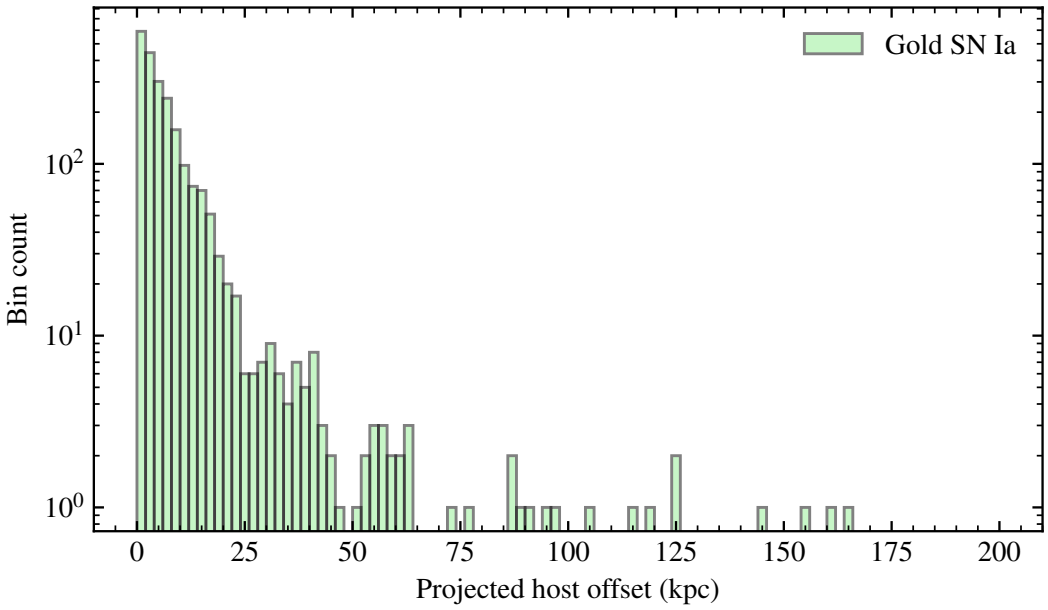

Figure 6.6: Sky-projected host offset distribution of Type Ia supernovae, using the distances derived through the `moriarty` algorithm. At large separations (~ 90 kpc) the matches are likely dominated by misidentifications, rather than real hosts.

tions, and plot the resultant distribution in Figure 6.6. We know that some specific types of transients are more likely found at larger physical separations from their hosts (e.g. Kasliwal et al. 2012), and thus `moriarty` is a unique tool for targeting our follow-up efforts – by algorithmically identifying these transients at discovery time, we can in future find more of them at earlier times.

We also plot a Hubble diagram using both `moriarty`-derived redshifts, and the spectroscopic redshifts obtained as part of ZTF-BTS. There is larger dispersion for the contextual redshifts, as expected, but it is reassuring to see that they track the spectroscopic redshifts in bulk as expected. Whilst the contextual classifier is not intended for cosmology, it is important that these redshifts are accurate – as derived properties depend on them. All results presented here are prior to fine-tuning of the algorithm using methods from Section 2.2.5, which can be expected to increase our figures of merit significantly. Around 5% of the transients in the ZTF-BTS Ia sample are identified with a host, but lack a redshift estimate for the object so derived properties cannot be computed. The availability of upcoming large-scale spectroscopic catalogs promises

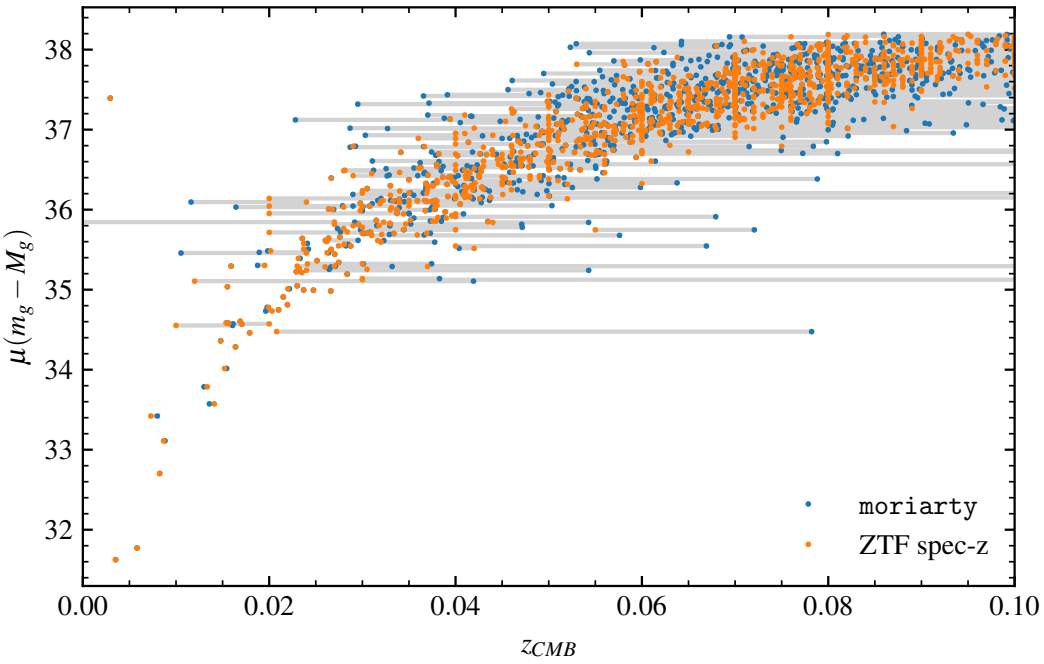

Figure 6.7: Hubble diagram constructed with the ZTF-BTS 'gold sample' using g' band photometry only. The grey lines connect the redshift estimates for each individual event, to reveal significant outliers.

to be useful for identifications like this, providing redshifts for many faint point sources. The remaining 5% are not successfully matched to a given source in the `moriarty` catalogues. This may be due to missing sources from catalogs, the transient being located at an extreme offset and so failing the p_{assoc} cut , or the transient may be intrinsically hostless (down to the depth of PS1, our faintest catalog). This 5% is the focus of remaining optimisation efforts. It is important to acknowledge such algorithms require extensive testing under real-world conditions to provide empirical validation of their performance and reveal intrinsically rare edge cases – however, the high recovery rate on a known subset of astrophysical transients shows great promise. may make use of the algorithm.

6.4 Future prospects

The `moriarty` classifier is now running on the live GOTO transient discovery stream, annotating every target to be inspected by humans. The finalised version of the code will be made publicly available in future via an easily-queryable REST API so that the community. Robust contextual classification is both a requirement and significant challenge for the Vera C Rubin Observatory's upcoming Legacy Survey of Space and Time (LSST) - given the depth of single-visit exposures, very few areas of the sky will have deep enough coverage from other surveys to provide contextual information about the host of a potential transient, or the faint underlying source associated with a variable star. Deep stacks will not be available until later in the survey's lifetime, therefore much of the context will need to be inferred from the reference template. At the time of writing, the alert system will generate $5" \times 5"$ cutouts (Ivezić et al., 2019), which are not large enough to capture nearby sources/potential hosts in the local Universe. Overcoming these inherent issues will require a fusion of image-level and catalog- level information as discussed in this chapter, and it is hoped some of the methods introduced here can help address some of the challenges inherent.

Chapter 7

Conclusions and future prospects

In this concluding chapter, I summarise the content of this thesis and highlight the most important findings, with an eye to upcoming future developments in time-domain astrophysics and future work on each of the projects I have presented.

7.1 Summary of the thesis

The landscape of time-domain astrophysics has changed significantly over the past 10 years, moving from small, single telescope surveys of the local Universe, to vast, distributed deep optical sky surveys that cover the entire night sky every night – driven by the need to rapidly survey large areas of sky in response to poorly-localised external triggers, searching for rapidly decaying sources associated with cataclysmic cosmic events. Developing the methodologies and algorithms required to rapidly sift the data volumes associated with such large-scale searches has been (and remains) a central challenge in the time-domain astrophysics community, along with capitalising on the discoveries made to advance our understanding of the transient parameter space and the diverse range of objects that inhabit it. This thesis has straddled the border between these observational and computational domains, leveraging the connection between the two in the form of domain-specific knowledge, and application of robust statistical methodologies, to make novel contributions in both.

The challenge of real-bogus classification continues to remain a signficant bot-

tleneck in the identification of transients, especially in the context of modular surveys like GOTO where the PSF, detector quality, and other parameters may be subject to variance across the individual telescopes. Chapter 3 presented an algorithmic approach to generating the vast volumes of training data necessary to train a deep-learned source classifier using minor planets already in the data, as well as a novel data augmentation scheme to improve the recovery of faint and nuclear transients associated with host galaxies. Scaling the datasets involved to even larger sizes, and adapting the techniques to work on multi-class classification tasks (see Section 6.2.1) remains an important focus of future work – with the number of false positives significantly reduced, we are likely to instead be swamped with real (but ultimately uninteresting) detections instead. There is a strong case for introducing datasets generated by citizen science and large-scale human vetting efforts to fine-tune performance, and performing active learning leveraging the Bayesian neural network architectures constructed as part of this work to optimise this allocation process.

Although the predominant focus of this thesis was on transient gravitational-wave sources and what we can learn about them via optical follow-up, Chapter 4 demonstrates the power of EM observations for the discovery of continous- wave sources. Through a homogeneous reprocessing of over 20 years of high-resolution optical spectroscopy, we derive the most precise ephemeris to date for the Sco X-1 system, the continuous-wave source with the strongest promise of being detected by LIGO/Virgo. Whilst also correcting previous errors, through robust Bayesian modelling we minimise potential systematics arising from calibration uncertainties in the data, and employ Hamiltonian Monte Carlo for efficient sampling that explores well the posterior landscape. We also produce some of the highest signal-to-noise spectral atlases of any low-mass X- ray binary, revealing many potential lines of interest to the compact object community for further follow-up. As part of future work, I have secured 2 years of monitoring of Sco X-1 to continue to provide high-quality ephemerides in support of the upcoming LIGO/Virgo/KAGRA O4 observing run – with the hope of the first detection of continuous waves. The modelling built in this chapter will also be applied to other LMXB sources of interest to the continuous-wave community.

Chapter 5 presented the results of a sensitive search for short-timescale variability in two core-collapse supernovae, conducted with Liverpool Telescope and South African Large Telescope observations. As part of this I obtained the highest cadence light curves of any supernova to date in the literature, and probed for potential variability at sub-minute timescales for the first time. We successfully rule out the presence of $\sim 1\%$ scale variability in both of the targets studied, and pave the way for a future (ongoing) large, homogeneous census of supernova variability at short timescales across the transient zoo. The techniques and methodology presented in this Chapter will be crucial for investigating the recent emergence of flaring in transients of the FBOT family in future, and this is the focus of future work.

In Chapter 6, I introduced the catalog engine for `moriarty`, a data-driven contextual classifier that aggregates and combines a wide range of astronomical datasets, with the goal of providing accurate source classifications and contextual information for transient discoveries based solely on their sky location. As part of this, I implement a novel extension of the p_{chance} algorithm for host association that shows marked improvements over a naive cross-match, whilst accommodating uncertainties in cross-matching and enabling the recovery of transients at extreme physical separations from their hosts. The final algorithm, paired with deep-learned image-level classification introduced in this Chapter will be released as part of an in-preparation manuscript – with an early alpha version of `moriarty` being tested in the GOTO Marshall. Future work hopes to incorporate the contextual information from `moriarty` into automated decision- making frameworks for astronomical follow-up: with the end goal being fully autonomous spectroscopic classification on a subset of 'gold-standard' sources that are discovered by GOTO, using RTML and similar frameworks to send the observing blocks to facilities like Liverpool Telescope with zero human intervention.

7.2 Looking forward: the future of time-domain astrophysics

At the time of submission, the LIGO-Virgo-KAGRA O4 observing run is 1 month underway, generating significant numbers of gravitational wave alerts already at the time of

writing. Lowered signifance thresholds, in addition to the enhanced detector sensitivity are contributing to this – with prioritisation, ranking, and significance assessment of events becoming a challenge due to the $\sim 3/d$ rate. Instrumentation issues at Virgo mean that it is currently not participating, expected to join late in Summer 2023 at the time of writing. The lack of a third sensitive detector is currently causing issues, with the quality of sky localisations being significantly worse than anticipated. Although problematic for galaxy-targeted searches, wide- field surveys such as GOTO are ideally suited to this challenging task. The current range of KAGRA does not reach outside our own Milky Way's satellite galaxies, although towards the latter end of the O4 run this detector may gain diagnostic power in the case of Local Group events, ($\lesssim 20$ Mpc). This observing run is 50% longer than the previous O3 observing run, which combined with the $\sim 2\times$ greater detector sensitivity promises to deliver a significant yield of EM-bright gravitational wave triggers to follow up. The time-domain community stands ready to respond to these events, with both large-scale sky surveys and dedicated follow-up programs standing by for the next kilonova (or something more interesting!) The relevant lessons have been learned from the detection of GW 170817, and with significant progress in both modelling and instrumentation the community is well-poised to make the most of both the GW and EM information provided by these events – making these months a truly exciting time to be in the field.

GOTO has completed expansion to the second antipodal site in recent months, with now 32 telescopes across 4 mounts in two hemispheres – providing truly unprecedented follow-up capability. With an instantaneous field of view of 80 degrees from each site, and near-continuous observing capability (barring dusk and dawn at each site), survey operations can cover the largest localisation regions in ~ 1 night, with well-localised events receiving multiple repeat observations. This is crucial for both validating transient candidates, but also obtaining photometric evolution to rule out contaminants in any searches. We have been active since the very beginning of the O4 observing run, following up significant gravitational-wave events and reporting transients from the GW error boxes. All the machinery is in place for the next nearby kilonova, and with 17 months remaining of this observing run, we are ready.

Looking ahead beyond the O4 run, the recent approval of LIGO-India is a significant development in the GW-EM landscape of the coming decade. A five detector network promises to significantly improve the localisation quality of the highest signal-to-noise events, a crucial limiting factor for GW-EM searches, as well as dramatically improving parameter estimation and signal-to-noise for less massive coalescences. Continuous-wave searches will also benefit from the additional detector, pushing the GW strain limits into further tension with our existing models of neutron star physics, or perhaps yielding a significant detection of the elusive gravitational waves from galactic compact binaries. This expanded network promises to deliver more opportunities for multi-messenger synergies, ripe for upcoming electromagnetic surveys to make significant impacts.

Building on the already explosive rise of the past decade, the future of time-domain astrophysics looks incredibly promising – with a new generation of upcoming modular sky surveys are currently in development. Some examples include wFAST (Nir et al., 2021), LAST (Ben-Ami et al., 2023), and the Argus Optical Array (Law et al., 2022), each surveying the entire visible sky at high cadence. Existing surveys are not standing still, expanding to new nodes (like GOTO- South) and migrating to new, more sensitive hardware. Technologies like complementary metal-oxide semiconductor (CMOS) are rapidly becoming popular in professional astronomy (e.g. Alarcon et al. 2023) due to their low noise, high quantum efficiency, and ultra-high-cadence capabilities owing to an electronic shutter. Whether this marks a paradigm shift for wide-field surveys remains to be seen, but continually evolving instrumentation provides new opportunities for novel study– by pushing coverage of transient parameter space to shorter timescales. With a wealth of optical surveys on sky, combining data from multiple surveys is likely to yield exquisitely-sampled light curves of a range of transient phenomena even from survey observations alone, further expanding the possibilities for large-scale population studies. Naturally, bridging the gap from single-object to population studies is a challenge, likely requiring principled use of Bayesian statistics to draw appropriate and robust conclusions from challenging datasets.

The prospect of dedicated near-infrared all-sky survey instruments in the near

future, such as DREAMS (Soon et al., 2020, 2022) and WINTER (Lourie et al., 2020; Frostig et al., 2022), is exciting both from the perspective of late-time recovery of kilonova counterparts from the next-generation GW detector network, but also from shining new light on dust-obscured transients and probing the poorly understood near- infrared evolution of supernovae and other explosive transients. Ultraviolet missions such as ULTRASAT (Ben-Ami et al., 2022) and UVEX (Kulkarni et al., 2021) promise to deliver unprecedented early-time NUV/FUV observations of a range of transients, probing poorly-understood phenomena such as shock breakout and prompt emission and in turn delivering insights into massive star evolution, neutron star physics, and more. The successful fusion of ultraviolet, optical, and near-infrared data at all-sky scale is a crucial step to beginning robust population-level studies. This level of in-depth characterisation is only achievable with 'pointed' observations currently, making its' extension to populations of $\sim$1000s of bright transients a year an exciting step change in capability. Whether the larger politics of some of these missions will conspire to limit the correlation of these datastreams remains to be seen, however I hope that the profound scientific gains possible will be incentive enough to forge new collaborations between previously isolated groups.

The next generation of spectroscopic follow-up facilities are due to be fully commissioned in the mid-2020s: with both multi-object (DESI, 4MOST, and WEAVE) (DESI Collaboration et al., 2016; de Jong et al., 2012; Dalton et al., 2012) and single-object (SoXS, NTE, NRT) (Schipani et al., 2016; Fynbo, 2022; Copperwheat et al., 2015) facilities devoting significant shares of their time towards the classification and follow-up of transients, with some of the above effectively being earmarked for transients. A particularly exciting prospect the community is building towards (e.g. Nordin et al. 2019) is 'fully autonomous' follow-up using robotic facilities, underpinned by techniques such as those developed in Chapter 6. Removing the human element of triggering confers two crucial advantages:

- Minimisation of latency: follow-up can commence in near-real time, providing access to both the infant stages of regular supernovae, and timely observations of fast blue optical transients.

- Minimisation of biases: defining precise, algorithmic constraints on when to trigger observations minimises human bias (i.e. on bright/nearby/interestingly-named transients), and so can probe novel regions of parameter space.

Naturally, this is a challenge: telescope time is expensive (both in financial costs and time), and ensuring that time is not wasted on false positives is crucial for the longevity of any program doing fully-autonomous observations. As elaborated on in Section 6.4, significant challenges await in the sheer scale of the LSST alert stream: with the anticipated *10 million* alerts per night (Hambleton et al., 2022), even with current state-of-the-art real-bogus performances of 99.5%, thousands of false-positives will be generated. Further, the majority of transients discovered via the LSST will be too faint to receive effective follow-up – further straining the finite spectroscopic capability available to the community. This emphasises the importance of usage of contextual classification to draw probabilistic inferences about the type of a given transient from existing datasets available, and the usage of light curve information. It is clear that despite the promises of Vera Rubin Observatory, 'local Universe' transient surveys still have a critical role to play in time-domain astrophysics going forward – it is only with a better in-depth understanding of seemingly ordinary supernovae and their peculiarities that we can begin to make the most of the unprecedented samples that deeper surveys can deliver.

Regardless, with a diverse range of new facilities on the horizon with dramatically improved survey capabilities, the trend of exponential growth in the discovery of transients (Figure 1.7) shows little sign of slowing down. If we can keep pace with this deluge, multi-messenger time-domain astrophysics promises to continue to lead the way towards understanding the cataclysmic fates of stars and stellar remnants, and their role in shaping the Cosmos we observe today. As we push back the boundaries of transient parameter space one cosmic explosion at a time, it is truly unclear what new things we will find[1] – it is precisely here, at the edge of humanity's knowledge, where the joy of discovery lies.

[1] In all likelihood, more Type Ia supernovae.

Appendix A

Data, funding, and acknowledgements

The work presented in this thesis was funded via an Science and Technology Facilities Council (STFC) doctoral studentship, no. ST/T506503/1, and makes use of the facilities funded via STFC grants no. ST/T007184/1 and ST/T003103/1. I gratefully acknowledge the financial support from the Visitor and Mobility program of the Finnish Centre for Astronomy with ESO (FINCA), funded by the Academy of Finland grant nr. 306531, for enabling a fruitful research visit during the PhD.

Chapter 5

This project was enabled via several generous awards of telescope time from various time allocation committees:

- Observations in this paper were obtained under the Reactive Time proposal `PQ21B03` (PI Killestein). The Liverpool Telescope is operated on the island of La Palma by Liverpool John Moores University in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias with financial support from the UK Science and Technology Facilities Council.
- Observations reported in this paper were obtained with the Southern African

Large Telescope (SALT) under proposal 2021-2-DDT-002 (PI Killestein). We thank the SALT team for their excellent support in undertaking these observations.

- I thank Vik Dhillon for helpful discussions surrounding pickup noise in the SALT data.

Chapter 6

I thank Bernie Hsiao at MAST for help in downloading and ingesting the PS1 DR2 StackObjectView database table.

Appendix B

List of original publications

Whelan, J. T., Tenorio, R., Wofford, J. K., et al. (2023) *Search for Gravitational Waves from Scorpius X-1 in LIGO O3 Data with Corrected Orbital Ephemeris*, ApJ, **949**, 117

Deckers, M., Graur, O., Maguire, K., et al. (2023) *Photometric study of the late-time near-infrared plateau in Type Ia supernovae*, MNRAS, **521**, 4414-4430

Killestein, T. L., Mould, M., Steeghs, D., et al. (2023) *Precision Ephemerides for Gravitational-wave Searches - IV. Corrected and refined ephemeris for Scorpius X-1*, MNRAS, **520**, 5317-5330

Wang, Q., Armstrong, P., Zenati, Y., et al. (2023) *Revealing the Progenitor of SN 2021zby through Analysis of the TESS Shock-cooling Light Curve*, ApJL, **943**, L15

Pasham, D. R., Lucchini, M., Laskar, T., et al. (2023) *The Birth of a Relativistic Jet Following the Disruption of a Star by a Cosmological Black Hole*, NatAs, **7**, 88-104

Mong, Y.-L., Ackley, K., Killestein, T. L., et al. (2023) *Self-supervised clustering on image-subtracted data with deep-embedded self-organizing map*, MNRAS, **518**, 752-762

Hosseinzadeh, G., Kilpatrick, C. D., Dong, Y., et al. (2022) *Weak Mass Loss from the Red Supergiant Progenitor of the Type II SN 2021yja*, ApJ, **935**, 31

Steeghs, D., Galloway, D. K., Ackley, K., et al. (2022) *The Gravitational-wave Optical Transient Observer (GOTO): prototype performance and prospects for transient science*, MNRAS, **511**, 2405-2422

Mong, Y.-L., Ackley, K., Galloway, D. K., et al. (2021) *Searching for Fermi GRB optical counterparts with the prototype Gravitational-wave Optical Transient Observer (GOTO)*, MNRAS, **507**, 5463-5476

Froebrich, D., Derezea, E., Scholz, A., et al. (2021) *A survey for variable young stars with small telescopes - IV. Rotation periods of YSOs in IC 5070*, MNRAS, **506**, 5989-6000

Burhanudin, U. F., Maund, J. R., Killestein, T., et al. (2021) *Light-curve classification with recurrent neural*

networks for GOTO: dealing with imbalanced data, MNRAS, **505**, 4345-4361

Makrygianni, L., Mullaney, J., Dhillon, V., et al. (2021) *Processing GOTO survey data with the Rubin Observatory LSST Science Pipelines II: Forced Photometry and lightcurves*, PASA, **38**, e025

Killestein, T. L., Lyman, J., Steeghs, D., et al. (2021) *Transient-optimized real-bogus classification with Bayesian convolutional neural networks - sifting the GOTO candidate stream*, MNRAS, **503**, 4838-4854

Mullaney, J. R., Makrygianni, L., Dhillon, V., et al. (2021) *Processing GOTO data with the Rubin Observatory LSST Science Pipelines I: Production of coadded frames*, PASA, **38**, e004

Mong, Y.-L., Ackley, K., Galloway, D. K., et al. (2020) *Machine learning for transient recognition in difference imaging with minimum sampling effort*, MNRAS, **499**, 6009-6017

Gompertz, B. P., Cutter, R., Steeghs, D., et al. (2020) *Searching for electromagnetic counterparts to gravitational-wave merger events with the prototype Gravitational-Wave Optical Transient Observer (GOTO-4)*, MNRAS, **497**, 726-738

Evitts, J. J., Froebrich, D., Scholz, A., et al. (2020) *A survey for variable young stars with small telescopes: II - mapping a protoplanetary disc with stable structures at 0.15 au*, MNRAS, **493**, 184-198

Bibliography

Aartsen M. G., et al., 2017, Journal of Instrumentation, 12, P03012

Aasi J., et al., 2015, Phys. Rev. D, 91, 062008

Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/

Abbott B., et al., 2007, Phys. Rev. D, 76, 082001

Abbott B. P., et al., 2016a, Phys. Rev. D, 93, 122003

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

Abbott B. P., et al., 2016c, Phys. Rev. Lett., 116, 221101

Abbott B. P., et al., 2016d, Phys. Rev. Lett., 116, 241102

Abbott B. P., et al., 2017a, Phys. Rev. D, 95, 122003

Abbott B. P., et al., 2017b, Phys. Rev. Lett., 118, 121102

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

Abbott B. P., et al., 2017d, ApJ, 847, 47

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

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

Abbott B. P., et al., 2019a, Physical Review X, 9, 011001

Abbott B. P., et al., 2019b, Phys. Rev. D, 100, 062001

Abbott B. P., et al., 2019c, Phys. Rev. D, 100, 122002

Abbott R., et al., 2021a, Phys. Rev. D, 104, 022005

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

Abbott R., et al., 2022a, Phys. Rev. D, 105, 022002

Abbott R., et al., 2022b, Phys. Rev. D, 106, 062002

Abbott R., et al., 2022c, ApJ, 941, L30

Abell G. O., Corwin Harold G. J., Olowin R. P., 1989, ApJS, 70, 1

Acernese F., et al., 2015, Classical and Quantum Gravity, 32, 024001

Ackley K., Eikenberry S. S., Yildirim C., Klimenko S., Garner A., 2019, AJ, 158, 172

Ackley K., et al., 2020, A&A, 643, A113

Adams S. M., Kochanek C. S., Beacom J. F., Vagins M. R., Stanek K. Z., 2013, ApJ, 778, 164

Agudo I., et al., 2022, arXiv e-prints, p. arXiv:2208.09000

Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Alarcon M. R., Licandro J., Serra-Ricart M., Joven E., Gaitan V., de Sousa R., 2023, PASP, 135, 055001

Alard C., Lupton R. H., 1998, ApJ, 503, 325

Alard C., Lupton R., 1999, ISIS: A method for optimal image subtraction, Astrophysics Source Code Library, record ascl:9909.003 (ascl:9909.003)

Alcock C., et al., 2000, ApJ, 541, 734

Aleksić J., et al., 2011, ApJ, 735, L5

Alexander K. D., et al., 2018, ApJ, 863, L18

Althaus L. G., Córsico A. H., Isern J., García-Berro E., 2010, A&A Rev., 18, 471

Anderson J. P., James P. A., Habergham S. M., Galbany L., Kuncarayakti H., 2015, PASA, 32, e019

Andersson N., Kokkotas K. D., Stergioulas N., 1999, ApJ, 516, 307

Andreoni I., Cooke J., 2019, in Griffin R. E., ed., Vol. 339, Southern Horizons in Time-Domain Astronomy. pp 135–138 (arXiv:1802.01100), doi:10.1017/S1743921318002399

Ansoldi S., et al., 2018, ApJ, 863, L10

Arcavi I., et al., 2017, Nature, 551, 64

Arendt R. G., Dwek E., Bouchet P., John Danziger I., Gehrz R. D., Park S., Woodward C. E., 2020, ApJ, 890, 2

Arnett W. D., 1982, ApJ, 253, 785

Arnett W. D., Bahcall J. N., Kirshner R. P., Woosley S. E., 1989, ARA&A, 27, 629

Asmus D., et al., 2020, MNRAS, 494, 1784

Assef R. J., Stern D., Noirot G., Jun H. D., Cutri R. M., Eisenhardt P. R. M., 2018, ApJS, 234, 23

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

Astropy Collaboration et al., 2018, AJ, 156, 123

Auer P., 2003, J. Mach. Learn. Res., 3, 397–422

Avakyan A., Neumann M., Zainab A., Doroshenko V., Wilms J., Santangelo A., 2023, arXiv e-prints, p. arXiv:2303.16168

Baade W., Zwicky F., 1934a, Proceedings of the National Academy of Science, 20, 254

Baade W., Zwicky F., 1934b, Proceedings of the National Academy of Science, 20, 259

Bailey S., Aragon C., Romano R., Thomas R. C., Weaver B. A., Wong D., 2007, ApJ, 665, 1246

Baluev R. V., 2008, MNRAS, 385, 1279

Bandyopadhyay R. M., Shahbaz T., Charles P. A., Naylor T., 1999, MNRAS, 306, 417

Barbary K., 2016, Journal of Open Source Software, 1, 58

Barbieri C., Salafia O. S., Perego A., Colpi M., Ghirlanda G., 2019, A&A, 625, A152

Barnes J., Zhu Y. L., Lund K. A., Sprouse T. M., Vassh N., McLaughlin G. C., Mumpower M. R., Surman R., 2021, ApJ, 918, 44

Barthelmy S. D., et al., 1998, in Meegan C. A., Preece R. D., Koshut T. M., eds, American Institute of Physics Conference Series Vol. 428, Gamma-Ray Bursts, 4th Hunstville Symposium. pp 99–103, doi:10.1063/1.55426

Beasor E. R., Davies B., 2018, MNRAS, 475, 55

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

Becker A., 2015, HOTPANTS: High Order Transform of PSF ANd Template Subtraction (ascl:1504.004)

Becker A. C., Homrighausen D., Connolly A. J., Genovese C. R., Owen R., Bickerton S. J., Lupton R. H., 2012, MNRAS, 425, 1341

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

Ben-Ami S., et al., 2022, in den Herder J.-W. A., Nikzad S., Nakazawa K., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 12181, Space Telescopes and Instrumentation 2022: Ultraviolet to Gamma Ray. p. 1218105 (arXiv:2208.00159), doi:10.1117/12.2629850

Ben-Ami S., et al., 2023, arXiv e-prints, p. arXiv:2304.02719

Bergstra J., Bengio Y., 2012, J. Mach. Learn. Res., 13, 281–305

Bernstein G., Huterer D., 2010, MNRAS, 401, 1399

Bersten M. C., et al., 2018, Nature, 554, 497

Berthier J., Vachier F., Thuillot W., Fernique P., Ochsenbein F., Genova F., Lainey V., Arlot J. E., 2006, SkyBoT, a new VO service to identify Solar System objects. , p. 367

Berthier J., Carry B., Vachier F., Eggl S., Santerne A., 2016, MNRAS, 458, 3394

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Betancourt M., 2016, arXiv e-prints, p. arXiv:1604.00695

Betancourt M., 2017, arXiv e-prints, p. arXiv:1701.02434

Bietenholz M. F., Bartel N., Argo M., Dua R., Ryder S., Soderberg A., 2021, ApJ, 908, 75

Bildsten L., 1998, ApJ, 501, L89

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

Bingham E., et al., 2019, J. Mach. Learn. Res., 20, 28:1

Bionta R. M., et al., 1987, Phys. Rev. Lett., 58, 1494

Bloemen S., et al., 2016, in Hall H. J., Gilmozzi R., Marshall H. K., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9906, Ground-based and Airborne Telescopes VI. p. 990664, doi:10.1117/12.2232522

Blondin S., Tonry J. L., 2007, ApJ, 666, 1024

Bloom J. S., Kulkarni S. R., Djorgovski S. G., 2002, AJ, 123, 1111

Bloom J. S., et al., 2012, PASP, 124, 1175

Blundell C., Cornebise J., Kavukcuoglu K., Wierstra D., 2015, arXiv e-prints, p. arXiv:1505.05424

Bonanos A. Z., Boumis P., 2016, A&A, 585, A19

Borisov G., et al., 2018, MNRAS, 480, L131

Borucki W. J., et al., 2010, Science, 327, 977

Börzsönyi S., Kossmann D., Stocker K., 2001, Proceedings 17th International Conference on Data Engineering, pp 421–430

Boucaud A., Bocchio M., Abergel A., Orieux F., Dole H., Hadj-Youcef M. A., 2016, A&A, 596, A63

Bradbury J., et al., 2018, JAX: composable transformations of Python+NumPy programs, http://github.com/google/jax

Bradshaw C. F., Fomalont E. B., Geldzahler B. J., 1999, ApJ, 512, L121

Bramich D. M., 2008, MNRAS, 386, L77

Bramich D. M., et al., 2013, MNRAS, 428, 2275

Bramich D. M., Horne K., Alsubai K. A., Bachelet E., Mislis D., Parley N., 2016, MNRAS, 457, 542

Brauer K., Vrtilek S. D., Peris C., McCollough M., 2018, MNRAS, 478, 4894

Brault J. W., White O. R., 1971, A&A, 13, 169

Brazier K. T. S., et al., 1990, A&A, 232, 383

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

Breiman L., Friedman J., Stone C., Olshen R., 1984, Classification and Regression Trees. Taylor & Francis

Breneman H. H., Stone E. C., 1985, ApJ, 299, L57

Brink H., Richards J. W., Poznanski D., Bloom J. S., Rice J., Negahban S., Wainwright M., 2013, MNRAS, 435, 1047

Brinkworth C. S., Marsh T. R., Dhillon V. S., Knigge C., 2006, MNRAS, 365, 287

Brout D., et al., 2019, ApJ, 874, 150

Brown M. J. I., et al., 2014, ApJS, 212, 18

Bruch R. J., et al., 2022, arXiv e-prints, p. arXiv:2212.03313

Buckley D. A. H., Swart G. P., Meiring J. G., 2006, in Stepp L. M., ed., Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 6267, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 62670Z, doi:10.1117/12.673750

Buikema A., et al., 2020, Phys. Rev. D, 102, 062003

Bulla M., 2019, MNRAS, 489, 5037

Bulla M., 2023, MNRAS, 520, 2558

Burrows A., Lattimer J. M., 1987, ApJ, 318, L63

Cabrera-Vives G., Reyes I., Förster F., Estévez P. A., Maureira J.-C., 2017, ApJ, 836, 97

Cai Y. Z., et al., 2021, A&A, 654, A157

Cai Y. Z., et al., 2022, A&A, 667, A4

Carrasco-Davis R., et al., 2020, arXiv e-prints, p. arXiv:2008.03309

Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483

Casares J., Steeghs D., Hynes R. I., Charles P. A., O'Brien K., 2003, ApJ, 590, 1041

Cerruti M., Zech A., Boisson C., Emery G., Inoue S., Lenain J. P., 2019, MNRAS, 483, L12

Chakrabarty D., Morgan E. H., Muno M. P., Galloway D. K., Wijnands R., van der Klis M., Markwardt C. B., 2003, Nature, 424, 42

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

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

Charbonneau D., Brown T. M., Latham D. W., Mayor M., 2000, ApJ, 529, L45

Chase E. A., et al., 2022, ApJ, 927, 163

Chattopadhyay S., Maitra R., 2017, MNRAS, 469, 3374

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

Cheng T.-Y., et al., 2020, MNRAS, 493, 4209

Cherepashchuk A. M., Khruzina T. S., Bogomazov A. I., 2021, MNRAS, 508, 1389

Chetlur S., Woolley C., Vandermersch P., Cohen J., Tran J., Catanzaro B., Shelhamer E., 2014, arXiv e-prints, p. arXiv:1410.0759

Chevalier R. A., 1982, ApJ, 258, 790

Chevalier R. A., Dwarkadas V. V., 1995, ApJ, 452, L45

Chevalier R. A., Irwin C. M., 2011, ApJ, 729, L6

Chodil G., Jopson R. C., Mark H., Seward F. D., Swift C. D., 1965, Phys. Rev. Lett., 15, 605

Chollet F., et al., 2015, Keras, `https://keras.io`

Chomiuk L., et al., 2016, ApJ, 821, 119

Christy C. T., et al., 2023, MNRAS, 519, 5271

Chugai N. N., et al., 2004, MNRAS, 352, 1213

Ciardullo R., Tamblyn P., Phillips A. C., 1990, PASP, 102, 1113

Cieślar M., Bulik T., Curyło M., Sieniawska M., Singh N., Bejger M., 2021, A&A, 649, A92

Ciucă I., Kawata D., Miglio A., Davies G. R., Grand R. J. J., 2020, arXiv e-prints, p. arXiv:2003.03316

Coleman C., et al., 2018, arXiv e-prints, p. arXiv:1806.01427

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

Copperwheat C. M., et al., 2015, Experimental Astronomy, 39, 119

Cornelisse R., Steeghs D., Casares J., Charles P. A., Barnes A. D., Hynes R. I., O'Brien K., 2007, MNRAS, 380, 1219

Coughlin M. W., Dietrich T., Margalit B., Metzger B. D., 2019, MNRAS, 489, L91

Coulter D. A., et al., 2017a, Science, 358, 1556

Coulter D. A., et al., 2017b, Transient Name Server Discovery Report, 2017-1030, 1

Cox N. L. J., Davis P., Patat F., Van Winckel H., 2014, The Astronomer's Telegram, 5797, 1

Cranmer M., Sanchez-Gonzalez A., Battaglia P., Xu R., Cranmer K., Spergel D., Ho S., 2020, arXiv e-prints, p. arXiv:2006.11287

Cranmer M., Tamayo D., Rein H., Battaglia P., Hadden S., Armitage P. J., Ho S., Spergel D. N., 2021, Proceedings of the National Academy of Science, 118, e2026053118

Crawford S. M., et al., 2010, in Silva D. R., Peck A. B., Soifer B. T., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 7737, Observatory Operations: Strategies, Processes, and Systems III. p. 773725, doi:10.1117/12.857000

Creswell J., von Hausegger S., Jackson A. D., Liu H., Naselsky P., 2017, J. Cosmology Astropart. Phys., 2017, 013

Cybenko G. V., 1989, Mathematics of Control, Signals and Systems, 2, 303

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

Dalton G., 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. 84460P, doi:10.1117/12.925950

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

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

Davidson K., Humphreys R. M., 1997, ARA&A, 35, 1

Dax M., Green S. R., Gair J., Macke J. H., Buonanno A., Schölkopf B., 2021, Phys. Rev. Lett., 127, 241103

De K., et al., 2023, Nature, 617, 55

Dekker H., D'Odorico S., Kaufer A., Delabre B., Kotzlowski H., 2000, in Iye M., Moorwood A. F., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4008, Optical and IR Telescope Instrumentation and Detectors. pp 534–545, doi:10.1117/12.395512

Dey A., et al., 2019, AJ, 157, 168

Dhillon V. S., et al., 2007, MNRAS, 378, 825

Dhurandhar S. V., Vecchio A., 2001, Phys. Rev. D, 63, 122001

Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441

Dimitriadis G., et al., 2019, ApJ, 870, L1

Dong Y., et al., 2022, ApJ, 927, 199

Drake A. J., et al., 2012, in Griffin E., Hanisch R., Seaman R., eds, Vol. 285, New Horizons in Time Domain Astronomy. pp 306–308 (arXiv:1111.2566), doi:10.1017/S1743921312000889

Drout M. R., et al., 2014, ApJ, 794, 23

Drout M. R., et al., 2017, Science, 358, 1570

Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Physics Letters B, 195, 216

Duev D. A., van der Walt S. J., 2021, arXiv e-prints, p. arXiv:2111.12142

Duev D. A., et al., 2019, MNRAS, 489, 3582

Duffy C., et al., 2021, MNRAS, 502, 4953

Duncan K. J., 2022, MNRAS, 512, 3662

Dyer M. J., 2020, PhD thesis, University of Sheffield, UK

Dyer M. J., Dhillon V. S., Littlefair S., Steeghs D., Ulaczyk K., Chote P., Galloway D., Rol E., 2018, in Observatory Operations: Strategies, Processes, and Systems VII. p. 107040C (arXiv:1807.01614), doi:10.1117/12.2311865

Eastman R. G., Schmidt B. P., Kirshner R., 1996, ApJ, 466, 911

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

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

Eldridge J. J., Izzard R. G., Tout C. A., 2008, MNRAS, 384, 1109

Emsenhuber A., Cambioni S., Asphaug E., Gabriel T. S. J., Schwartz S. R., Furfaro R., 2020, ApJ, 891, 6

Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15

Fawcett T., 2006, Pattern Recognition Letters, 27, 861

Fermi E., 1949, Physical Review, 75, 1169

Filippenko A. V., Li W. D., Treffers R. R., Modjaz M., 2001, in Paczynski B., Chen W.-P., Lemme C., eds, Astronomical Society of the Pacific Conference Series Vol. 246, IAU Colloq. 183: Small Telescope Astronomy on Global Scales. p. 121

Flesch E. W., 2021, VizieR Online Data Catalog, p. VII/290

Fomalont E. B., Geldzahler B. J., Hjellming R. M., Wade C. M., 1983, ApJ, 275, 802

Fomalont E. B., Geldzahler B. J., Bradshaw C. F., 2001, ApJ, 558, 283

Forbush S. E., 1946, Phys. Rev., 70, 771

Foreman-Mackey D., 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey D., 2018, Research Notes of the American Astronomical Society, 2, 31

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Foreman-Mackey D., et al., 2021, The Journal of Open Source Software, 6, 3285

Fossey S. J., Cooke B., Pollack G., Wilde M., Wright T., 2014, Central Bureau Electronic Telegrams, 3792, 1

Fox O. D., et al., 2015, MNRAS, 447, 772

Fraser M., 2020, Royal Society Open Science, 7, 200467

Freund Y., Schapire R. E., 1995, in Vitányi P., ed., Computational Learning Theory. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 23–37

Frigo M., 1999, SIGPLAN Not., 34, 169–180

Frostig D., et al., 2022, ApJ, 926, 152

Fryxell B., Mueller E., Arnett D., 1991, ApJ, 367, 619

Fynbo J., 2022, in NOT - A Telescope For the Future. p. 19, doi:10.5281/zenodo.7219529

Gagliano A., Narayan G., Engel A., Carrasco Kind M., LSST Dark Energy Science Collaboration 2021, ApJ, 908, 170

Gaia Collaboration et al., 2022, arXiv e-prints, p. arXiv:2208.00211

Gal Y., Ghahramani Z., 2015a, arXiv e-prints, p. arXiv:1506.02142

Gal Y., Ghahramani Z., 2015b, arXiv e-prints, p. arXiv:1506.02158

Gal-Yam A., 2017, in Alsabti A. W., Murdin P., eds, , Handbook of Supernovae. p. 195, doi:10.1007/978-3-319-21846-5_35

Gal Y., Islam R., Ghahramani Z., 2017, arXiv e-prints, p. arXiv:1703.02910

Galaudage S., Wette K., Galloway D. K., Messenger C., 2022, MNRAS, 509, 1745

Galloway D. K., Lin J., Chakrabarty D., Hartman J. M., 2010, ApJ, 711, L148

Galloway D. K., Premachandra S., Steeghs D., Marsh T., Casares J., Cornelisse R., 2014, ApJ, 781, 14

Gao L., et al., 2020, arXiv e-prints, p. arXiv:2101.00027

García J., Ramírez J. M., Kallman T. R., Witthoeft M., Bautista M. A., Mendoza C., Palmeri P., Quinet P., 2011, ApJ, 731, L15

Gelman A., Rubin D. B., 1992, Statistical Science, 7, 457

Gelman A., Carlin J., Stern H., Dunson D., Vehtari A., Rubin D., 2013, Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, https://books.google.co.uk/books?id=ZXL6AQAAQBAJ

George D., Shen H., Huerta E. A., 2018, Phys. Rev. D, 97, 101501

Gerke J. R., Kochanek C. S., Stanek K. Z., 2015, MNRAS, 450, 3289

Giacconi R., Gursky H., Paolini F. R., Rossi B. B., 1962, Phys. Rev. Lett., 9, 439

Gibson N. P., Aigrain S., Roberts S., Evans T. M., Osborne M., Pont F., 2012, MNRAS, 419, 2683

Gieseke F., et al., 2017, MNRAS, 472, 3101

Goldstein D. A., et al., 2015, AJ, 150, 82

Goldstein A., et al., 2017, ApJ, 848, L14

Gompertz B. P., et al., 2018, ApJ, 860, 62

Gompertz B. P., et al., 2020, MNRAS, 497, 726

Gompertz B. P., Nicholl M., Smith J. C., Harisankar S., Pratten G., Schmidt P., Smith G. P., 2023, arXiv e-prints, p. arXiv:2305.07582

Goobar A., et al., 2014, ApJ, 784, L12

Goobar A., et al., 2022, arXiv e-prints, p. arXiv:2211.00656

Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, arXiv e-prints, p. arXiv:1406.2661

Gottlieb E. W., Wright E. L., Liller W., 1975, ApJ, 195, L33

Graham M. J., et al., 2020, Phys. Rev. Lett., 124, 251102

Graham M. J., et al., 2023, ApJ, 942, 99

Gull S. F., 1988, Bayesian Inductive Inference and Maximum Entropy. Springer Netherlands, Dordrecht, pp 53–74, doi:10.1007/978-94-009-3049-0_4, https://doi.org/10.1007/978-94-009-3049-0_4

Hajela A., et al., 2022, ApJ, 927, L17

Hambleton K. M., et al., 2022, arXiv e-prints, p. arXiv:2208.04499

Harris C. R., et al., 2020, Nature, 585, 357

Hartman J. M., Chakrabarty D., Galloway D. K., Muno M. P., Savov P., Mendez M., van Straaten S., Di Salvo T., 2003, in AAS/High Energy Astrophysics Division #7. p. 17.38

Hayden B., et al., 2021, ApJ, 912, 87

Heavens A., 2009, arXiv e-prints, p. arXiv:0906.0664

Heger A., Jeannin L., Langer N., Baraffe I., 1997, A&A, 327, 224

Heger A., Fryer C. L., Woosley S. E., Langer N., Hartmann D. H., 2003, ApJ, 591, 288

Heinze A. N., et al., 2018, AJ, 156, 241

Hessels J. W. T., Ransom S. M., Stairs I. H., Freire P. C. C., Kaspi V. M., Camilo F., 2006, Science, 311, 1901

Hessman F. V., 2006, Astronomische Nachrichten, 327, 751

Himes M. D., et al., 2022, , 3, 91

Hirata K., et al., 1987, Phys. Rev. Lett., 58, 1490

Hitchcock J. A., Hundertmark M., Foreman-Mackey D., Bachelet E., Dominik M., Street R., Tsapras Y., 2021, MNRAS, 504, 3561

Hjorth J., et al., 2017, ApJ, 848, L31

Ho A. Y. Q., et al., 2021, arXiv e-prints, p. arXiv:2105.08811

Ho A. Y. Q., Perley D. A., Chen P., Schulze S., Sollerman J., Gal-Yam A., 2022a, Transient Name Server AstroNote, 267, 1

Ho A. Y. Q., Perley D. A., Chen P., Schulze S., Sollerman J., Gal-Yam A., 2022b, Transient Name Server AstroNote, 267, 1

Hocking A., Geach J. E., Sun Y., Davey N., 2018, MNRAS, 473, 1108

Hoffman M. D., Gelman A., 2011, arXiv e-prints, p. arXiv:1111.4246

Hopkins P. F., et al., 2018, MNRAS, 480, 800

Hornik K., Stinchcombe M., White H., 1989, Neural Networks, 2, 359

Hosseinzadeh G., Berger E., Metzger B. D., Gomez S., Nicholl M., Blanchard P., 2021, arXiv e-prints, p. arXiv:2109.09743

Houlsby N., Huszár F., Ghahramani Z., Lengyel M., 2011, arXiv e-prints, p. arXiv:1112.5745

Hu L., Wang L., Chen X., Yang J., 2022, ApJ, 936, 157

Huber P. J., 1964, The Annals of Mathematical Statistics, 35, 73

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

Hunt E. L., Reffert S., 2023, A&A, 673, A114

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

Huppenkothen D., Heil L. M., Hogg D. W., Mueller A., 2017, MNRAS, 466, 2364

Hynes R. I., Schaefer B. E., Baum Z. A., Hsu C.-C., Cherry M. L., Scaringi S., 2016, MNRAS, 459, 3596

IceCube Collaboration et al., 2018a, Science, 361, 147

IceCube Collaboration et al., 2018b, Science, 361, eaat1378

IceCube Collaboration et al., 2022, Science, 378, 538

Ishida E. E. O., et al., 2021, A&A, 650, A195

Ishizaki W., Ioka K., Kiuchi K., 2021, ApJ, 916, L13

Itagaki K., 2023, Transient Name Server Discovery Report, 2023-1158, 1

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

Jamieson K., Talwalkar A., 2015, arXiv e-prints, p. arXiv:1502.07943

Jasche J., Kitaura F. S., 2010, MNRAS, 407, 29

Jeffery C. S., Dhillon V. S., Marsh T. R., Ramachandran B., 2004, MNRAS, 352, 699

Jerkstrand A., Fransson C., Kozma C., 2011, A&A, 530, A45

Jones D. O., et al., 2021, ApJ, 908, 143

Kalogera V., et al., 2021, arXiv e-prints, p. arXiv:2111.06990

Kasliwal M. M., et al., 2012, ApJ, 755, 161

Kasliwal M. M., et al., 2017, ApJ, 839, 88

Kastner S. O., Bhatia A. K., 1996, MNRAS, 279, 1137

Kawaguchi K., Kyutoku K., Shibata M., Tanaka M., 2016, ApJ, 825, 52

Kawaguchi K., Shibata M., Tanaka M., 2020, ApJ, 889, 171

Keivani A., et al., 2018, ApJ, 864, 84

Keller S. C., et al., 2007, PASA, 24, 1

Kendall A., Gal Y., 2017, arXiv e-prints, p. arXiv:1703.04977

Khazov D., et al., 2016, ApJ, 818, 3

Kiewe M., et al., 2012, ApJ, 744, 10

Killestein T. L., et al., 2021, MNRAS, 503, 4838

Killestein T. L., Mould M., Steeghs D., Casares J., Galloway D. K., 2023, MNRAS,

Kimura S. S., Murase K., Bartos I., Ioka K., Heng I. S., Mészáros P., 2018, Phys. Rev. D, 98, 043020

King R. D., et al., 2009, Science, 324, 85

Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980

Kirshner R. P., Kwan J., 1974, ApJ, 193, 27

Knuth K. H., 2006, arXiv e-prints, p. physics/0605197

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

Kovács G., Bakos G., Noyes R. W., 2005, MNRAS, 356, 557

Kozłowski S., 2016, ApJ, 826, 118

Krizhevsky A., Sutskever I., Hinton G. E., 2017, Commun. ACM, 60, 84–90

Kulkarni S. R., 2005, arXiv e-prints, pp astro–ph/0510256

Kulkarni S. R., 2012, arXiv e-prints, p. arXiv:1202.2381

Kulkarni S. R., et al., 2007, Nature, 447, 458

Kulkarni S. R., Perley D. A., Miller A. A., 2018, ApJ, 860, 22

Kulkarni S. R., et al., 2021, arXiv e-prints, p. arXiv:2111.15608

Kunkel W., et al., 1987, IAU Circ., 4316, 1

Kunze S., Speith R., Hessman F. V., 2001, MNRAS, 322, 499

LIGO Scientific Collaboration et al., 2015, Classical and Quantum Gravity, 32, 074001

Lang D., Hogg D. W., Mierle K., Blanton M., Roweis S., 2010, AJ, 139, 1782

Larsson J., et al., 2023, ApJ, 949, L27

Lasky P. D., 2015, PASA, 32, e034

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

Law N. M., et al., 2022, PASP, 134, 035003

LeCun Y., Bengio Y., et al., 1995, The handbook of brain theory and neural networks, 3361, 1995

LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436

LeNail A., 2019, The Journal of Open Source Software, 4, 747

Leaci P., Prix R., 2015, Phys. Rev. D, 91, 102003

Leaman J., Li W., Chornock R., Filippenko A. V., 2011, MNRAS, 412, 1419

Leung S.-C., Blinnikov S., Nomoto K., Baklanov P., Sorokina E., Tolstov A., 2020, ApJ, 903, 66

Levenberg K., 1944, Quarterly of Applied Mathematics, 2, 164

Levine J. L., Garwin R. L., 1973, Phys. Rev. Lett., 31, 173

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

Li W. D., et al., 2000, in Holt S. S., Zhang W. W., eds, American Institute of Physics Conference Series Vol. 522, Cosmic Explosions: Tenth AstroPhysics Conference. pp 103–106 (arXiv:astro-ph/9912336), doi:10.1063/1.1291702

Li W., et al., 2011, MNRAS, 412, 1441

Li L., Jamieson K., DeSalvo G., Rostamizadeh A., Talwalkar A., 2017, The Journal of Machine Learning Research, 18, 6765

Li W., et al., 2019, ApJ, 870, 12

Liebling S. L., Palenzuela C., 2016, Phys. Rev. D, 94, 064046

Lindegren L., et al., 2021, A&A, 649, A2

Lintott C. J., et al., 2008, MNRAS, 389, 1179

Lipunov V. M., et al., 2017, ApJ, 850, L1

Liu D. C., Nocedal J., 1989, Mathematical Programming, 45, 503

Liu Z., Tegmark M., 2021, Phys. Rev. Lett., 126, 180604

Lochner M., Bassett B. A., 2021, Astronomy and Computing, 36, 100481

Loeb A., 2016, ApJ, 819, L21

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

Long X., et al., 2022, ApJ, 924, L13

Lourie N. P., et al., 2020, in Evans C. J., Bryant J. J., Motohara K., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 11447, Ground-based and Airborne Instrumentation for Astronomy VIII. p. 114479K (arXiv:2102.01109), doi:10.1117/12.2561210

Lück H., et al., 2006, Classical and Quantum Gravity, 23, S71

Lucy L. B., Sweeney M. A., 1971, AJ, 76, 544

Lyman J. D., Bersier D., James P. A., Mazzali P. A., Eldridge J. J., Fraser M., Pian E., 2016, MNRAS, 457, 328

Lyman J., et al., 2017, GRB Coordinates Network, 21582, 1

Lyman J. D., et al., 2018, Nature Astronomy, 2, 751

Lyutikov M., 2022, MNRAS, 515, 2293

Maaten L. v. d., Hinton G., 2008, Journal of machine learning research, 9, 2579

MacLachlan G. A., et al., 2013, MNRAS, 432, 857

Maggiore M., et al., 2020, J. Cosmology Astropart. Phys., 2020, 050

Mahabal A., et al., 2019, PASP, 131, 038002

Malanchev K. L., et al., 2021, MNRAS, 502, 5147

Malz A. I., et al., 2019, AJ, 158, 171

Margutti R., Soderberg A., Kamble A., Zauderer A., Milisavljevic D., Parrent J., Chomiuk L., 2014, The Astronomer's Telegram, 5851, 1

Margutti R., et al., 2017, ApJ, 848, L20

Mariani G., Scheidegger F., Istrate R., Bekas C., Malossi C., 2018, arXiv e-prints, p. arXiv:1803.09655

Marsh T. R., 2005, Ap&SS, 296, 403

Marsh T., 2019, molly: 1D astronomical spectra analyzer (ascl:1907.012)

Marsh T. R., 2022, MNRAS, 510, 1340

Marsh T. R., Horne K., 1988, MNRAS, 235, 269

Massey R., Refregier A., 2005, MNRAS, 363, 197

Masuda K., Hirano T., 2021, ApJ, 910, L17

Mata Sanchez D., Munoz-Darias T., Casares J., Steeghs D., Ramos Almeida C., Acosta Pulido J. A., 2015, MNRAS, 449, L1

Matsuura M., et al., 2022, MNRAS, 517, 4327

Matthews D., Brethauer D., Margutti R., Chornock R., Jacobson-Galan W., Laskar T., Migliori G., Alexander K. D., 2022, Transient Name Server AstroNote, 229, 1

Mazeh T., Tamuz O., Zucker S., 2007, in Afonso C., Weldrake D., Henning T., eds, Astronomical Society of the Pacific Conference Series Vol. 366, Transiting Extrapolar Planets Workshop. p. 119 (arXiv:astro-ph/0612418), doi:10.48550/arXiv.astro-ph/0612418

McCulloch W. S., Pitts W., 1943, The Bulletin of Mathematical Biophysics, 5, 115

McCully C., Volgenau N. H., Harbeck D.-R., Lister T. A., Saunders E. S., Turner M. L., Siiverd R. J., Bowman M., 2018, in Guzman J. C., Ibsen J., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 10707, Software and Cyberinfrastructure for Astronomy V. p. 107070K (arXiv:1811.04163), doi:10.1117/12.2314340

McKinney W., 2010, in Stéfan van der Walt Jarrod Millman eds, Proceedings of the 9th Python in Science Conference. pp 56–61, doi:10.25080/Majora-92bf1922-00a

Meadors G. D., Goetz E., Riles K., Creighton T., Robinet F., 2017, Phys. Rev. D, 95, 042005

Meegan C., et al., 2009, ApJ, 702, 791

Messenger C., et al., 2015, Phys. Rev. D, 92, 023006

Metzger B. D., Fernández R., 2021, ApJ, 916, L3

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

Michelson A. A., Morley E. W., 1887, American Journal of Science, s3-34, 333

Middleton H., Clearwater P., Melatos A., Dunn L., 2020, Phys. Rev. D, 102, 023006

Migenda J., 2020, arXiv e-prints, p. arXiv:2002.01649

Miknaitis G., et al., 2007, ApJ, 666, 674

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

Minkowski R., 1941, PASP, 53, 224

Modjaz M., et al., 2008, AJ, 135, 1136

Möller A., de Boissière T., 2020, MNRAS, 491, 4277

Mong Y.-L., et al., 2020, arXiv e-prints, p. arXiv:2008.10178

Mong Y. L., et al., 2021, MNRAS, 507, 5463

Moore G. E., 1965, Electronics, 38

Moriya T. J., Sorokina E. I., Chevalier R. A., 2018, Space Sci. Rev., 214, 59

Morozova V., Piro A. L., Valenti S., 2017, ApJ, 838, 28

Moskovitz N., Schottland R., Burt B., Wasserman L., Mommert M., Bailen M., Grimm S., 2019, in EPSC-DPS Joint Meeting 2019. pp Epsc–dps2019–644

Mould M., Gerosa D., Taylor S. R., 2022, Phys. Rev. D, 106, 103013

Mueller E., Fryxell B., Arnett D., 1991, A&A, 251, 505

Mukherjee A., Prix R., Wette K., 2022, arXiv e-prints, p. arXiv:2207.09326

Narayan R., Paczynski B., Piran T., 1992, ApJ, 395, L83

Neal R. M., 2000, Slice Sampling (arXiv:physics/0009028)

Nelder J. A., Mead R., 1965, The Computer Journal, 7, 308

Ness M., Hogg D. W., Rix H. W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16

Neumann M., Avakyan A., Doroshenko V., Santangelo A., 2023, arXiv e-prints, p. arXiv:2303.16137

Nicholl M., Margalit B., Schmidt P., Smith G. P., Ridley E. J., Nuttall J., 2021, MNRAS, 505, 3016

Niculescu-Mizil A., Caruana R., 2005, in Proceedings of the 22nd international conference on Machine learning. pp 625–632

Nir G., et al., 2021, PASP, 133, 075002

Nordin J., et al., 2019, A&A, 631, A147

Norris J. P., Bonnell J. T., 2006, ApJ, 643, 266

Nuñez J. R., Anderton C. R., Renslow R. S., 2018, PLOS ONE, 13, 1

O'Donoghue D., et al., 2006, MNRAS, 372, 151

O'Malley T., Bursztein E., Long J., Chollet F., Jin H., Invernizzi L., et al., 2019, Keras Tuner, https://github.com/keras-team/keras-tuner

Odewahn S. C., Stockwell E. B., Pennington R. L., Humphreys R. M., Zumach W. A., 1992, AJ, 103, 318

Ofek E. O., et al., 2010, ApJ, 724, 1396

Ofek E. O., et al., 2014a, ApJ, 788, 154

Ofek E. O., et al., 2014b, ApJ, 789, 104

Oke J. B., Searle L., 1974, ARA&A, 12, 315

Paice J. A., Gandhi P., Dhillon V. S., Marsh T. R., Green M., Breedt E., 2018, The Astronomer's Telegram, 12197, 1

Palmese A., et al., 2022, Transient Name Server AstroNote, 107, 1

Panagia N., Van Dyk S. D., Weiler K. W., Sramek R. A., Stockdale C. J., Murata K. P., 2006, ApJ, 646, 369

Papaloizou J., Pringle J. E., 1978, MNRAS, 184, 501

Paradijs J. v., Kouveliotou C., Wijers R. A. M. J., 2000, Annual Review of Astronomy and Astrophysics, 38, 379

Paraskeva E., Bonanos A. Z., Liakos A., Spetsieri Z. T., Maund J. R., 2020, A&A, 643, A35

Patat F., et al., 2014, The Astronomer's Telegram, 5830, 1

Patruno A., Haskell B., Andersson N., 2017, ApJ, 850, 106

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

Pérez-Torres M. A., et al., 2014, ApJ, 792, 38

Perley D. A., et al., 2019, MNRAS, 484, 1031

Perley D. A., et al., 2020, ApJ, 904, 35

Perlmutter S., et al., 1999, ApJ, 517, 565

Pessi T., et al., 2022, ApJ, 928, 138

Peters P. C., Mathews J., 1963, Phys. Rev., 131, 435

Petrov P., et al., 2022, ApJ, 924, 54

Phan D., Pradhan N., Jankowiak M., 2019, arXiv preprint arXiv:1912.11554

Pian E., et al., 2017, Nature, 551, 67

Piccinni O. J., 2022, Galaxies, 10, 72

Piesk T., 2016, Dih 4 Cayley Graph, https://commons.wikimedia.org/wiki/File:Dih_4_Cayley_Graph;_generators_a,_b.svg

Pietrzyński G., et al., 2013, Nature, 495, 76

Piro L., et al., 2019, MNRAS, 483, 1912

Podsiadlowski P., 1992, PASP, 104, 717

Prentice S. J., et al., 2018, ApJ, 865, L3

Pursiainen M., et al., 2018, MNRAS, 481, 894

Quinlan J. R., 1986, Machine Learning, 1, 81

Quinlan J. R., 1992.

Ragosta F., et al., 2021, Transient Name Server AstroNote, 297, 1

Rahmati A., Larson D. E., Cravens T. E., Lillis R. J., Lee C. O., Dunn P. A., 2020, Geophys. Res. Lett., 47, e88927

Rasmussen C. E., Williams C. K. I., 2003, in Adaptive computation and machine learning. https://api.semanticscholar.org/CorpusID:1430472

Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning

Rastinejad J. C., et al., 2022, Nature, 612, 223

Reed B. C., 2005, AJ, 130, 1652

Refregier A., 2003, MNRAS, 338, 35

Reguitti A., et al., 2022, Transient Name Server AstroNote, 14, 1

Reiss D. J., Lupton R. H., 2016, DMTN-021: Implementation of Image Difference Decorrelation, doi:10.5281/zenodo.192833. , https://doi.org/10.5281/zenodo.192833

Reitze D., et al., 2019, in Bulletin of the American Astronomical Society. p. 35 (arXiv:1907.04833)

Reyes-Jainaga I., et al., 2023, arXiv e-prints, p. arXiv:2304.13080

Reyes E., Estévez P. A., Reyes I., Cabrera-Vives G., Huijse P., Carrasco R., Forster F., 2018, in 2018 International Joint Conference on Neural Networks (IJCNN). pp 1–8

Rhodes B., 2019, Skyfield: High precision research-grade positions for planets and Earth satellites generator (ascl:1907.024)

Richardson D., III R. L. J., Wright J., Maddox L., 2014, The Astronomical Journal, 147, 118

Ricker G. R., et al., 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Riess A. G., et al., 1998, AJ, 116, 1009

Rigault M., et al., 2013, A&A, 560, A66

Robbins H., Monro S., 1951, The Annals of Mathematical Statistics, 22, 400

Romano R. A., Aragon C. R., Ding C., 2006, in 2006 5th International Conference on Machine Learning and Applications (ICMLA'06). pp 77–82

Rosenblatt F., 1958, Psychological Review, 65, 386

Rossi E. M., Begelman M. C., 2009, MNRAS, 392, 1451

Rubin A., Gal-Yam A., 2016, ApJ, 828, 111

Ryczanowski D., Smith G. P., Bianconi M., McGee S., Robertson A., Massey R., Jauzac M., 2023, MNRAS, 520, 2547

Sandage A., et al., 1966, ApJ, 146, 316

Sanders N. E., Betancourt M., Soderberg A. M., 2015, ApJ, 800, 36

Sandler D. G., Barrett T. K., Palmer D. A., Fugate R. Q., Wild W. J., 1991, Nature, 351, 300

Sanduleak N., 1970, Contributions from the Cerro Tololo Inter-American Observatory, 89

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

Scaringi S., Maccarone T. J., Hynes R. I., Körding E., Ponti G., Knigge C., Britt C. T., van Winckel H., 2015, MNRAS, 451, 3857

Schapire R. E., 1990, Machine Learning, 5, 197

Schipani P., et al., 2016, in Evans C. J., Simard L., Takami H., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI. p. 990841 (arXiv:1607.03729), doi:10.1117/12.2231866

Schlegel E. M., 1990, MNRAS, 244, 269

Schmidt B. P., et al., 1998, ApJ, 507, 46

Schuhmann C., et al., 2022, arXiv e-prints, p. arXiv:2210.08402

Scolnic D. M., et al., 2018, ApJ, 859, 101

Shabram M. I., et al., 2020, AJ, 160, 16

Shappee B. J., et al., 2014, ApJ, 788, 48

Shappee B. J., et al., 2017, Science, 358, 1574

Shappee B. J., et al., 2019, ApJ, 870, 13

Sharma Y., et al., 2023, arXiv e-prints, p. arXiv:2301.04637

Shen J., et al., 2022, ApJ, 925, 1

Shih D., Buckley M. R., Necib L., Tamanas J., 2022, MNRAS, 509, 5992

Shklovsky I. S., 1967, ApJ, 148, L1

Sieniawska M., Bejger M., 2019, Universe, 5, 217

Simonyan K., Zisserman A., 2014, arXiv e-prints, p. arXiv:1409.1556

Singer L. P., et al., 2015, ApJ, 806, 52

Smartt S. J., et al., 2016, MNRAS, 462, 4094

Smartt S. J., et al., 2017, Nature, 551, 75

Smith N., 2014, ARA&A, 52, 487

Smith R. J., Piascik A. S., Steele I. A., Barnsley R. M., 2016, in Chiozzi G., Guzman J. C., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9913, Software and Cyberinfrastructure for Astronomy IV. p. 991317, doi:10.1117/12.2232771

Smith M. C., Sijacki D., Shen S., 2018, MNRAS, 478, 302

Smith K. W., et al., 2020, PASP, 132, 085002

Smith K. W., et al., 2022, Transient Name Server AstroNote, 13, 1

Snoek J., Larochelle H., Adams R. P., 2012, arXiv e-prints, p. arXiv:1206.2944

Soares-Santos M., et al., 2017, ApJ, 848, L16

Soon J., et al., 2020, in Ellis S. C., d'Orgeville C., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 11203, Advances in Optical Astronomical Instrumentation 2019. p. 1120307, doi:10.1117/12.2539594

Soon J., Galla T., Moore A. M., Travouillon T., 2022, in Marshall H. K., Spyromilio J., Usuda T., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 12182, Ground-based and Airborne Telescopes IX. p. 121822K, doi:10.1117/12.2629072

Soumagnac M. T., Ofek E. O., 2018, PASP, 130, 075002

Soumagnac M. T., et al., 2015, MNRAS, 450, 666

Spurio Mancini A., Piras D., Alsing J., Joachimi B., Hobson M. P., 2022, MNRAS, 511, 1771

Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, The journal of machine learning research, 15, 1929

Steeghs D., Casares J., 2002, ApJ, 568, 273

Steeghs D., et al., 2021, The Gravitational-wave Optical Transient Observer (GOTO): prototype performance and prospects for transient science, in prep.

Steele I. A., et al., 2004, in Oschmann Jacobus M. J., ed., Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 5489, Ground-based Telescopes. pp 679–692, doi:10.1117/12.551456

Steele I. A., Bates S. D., Gibson N., Keenan F., Meaburn J., Mottram C. J., Pollacco D., Todd I., 2008, in McLean I. S., Casali M. M., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 7014, Ground-based and Airborne Instrumentation for Astronomy II. p. 70146J (arXiv:0809.3351), doi:10.1117/12.787889

Stein R., et al., 2021, Nature Astronomy, 5, 510

Stein R., et al., 2023, MNRAS, 521, 5046

Storey-Fisher K., Huertas-Company M., Ramachandra N., Lanusse F., Leauthaud A., Luo Y., Huang S., 2020, arXiv e-prints, p. arXiv:2012.08082

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

Tammann G. A., Loeffler W., Schroeder A., 1994, ApJS, 92, 487

Tamuz O., Mazeh T., Zucker S., 2005, MNRAS, 356, 1466

Tanaka M., Hotokezaka K., Kyutoku K., Wanajo S., Kiuchi K., Sekiguchi Y., Shibata M., 2014, ApJ, 780, 31

Tanvir N. R., et al., 2009, Nature, 461, 1254

Tanvir N. R., Levan A. J., Fruchter A. S., Hjorth J., Hounsell R. A., Wiersema K., Tunnicliffe R. L., 2013, Nature, 500, 547

Tanvir N. R., et al., 2017, ApJ, 848, L27

Tassoul J.-L., Tassoul M., 1992, ApJ, 395, 259

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

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

The Theano Development Team et al., 2016, arXiv e-prints, p. arXiv:1605.02688

Tieleman T., Hinton G., et al., 2012, COURSERA: Neural networks for machine learning, 4, 26

Titarchuk L. G., Bradshaw C. F., Geldzahler B. J., Fomalont E. B., 2001, ApJ, 555, L45

Tomaney A. B., Crotts A. P. S., 1996, AJ, 112, 2872

Tompson J., Goroshin R., Jain A., LeCun Y., Bregler C., 2015, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 648–656

Tonry J. L., et al., 2018, PASP, 130, 064505

Tonry J., et al., 2021, Transient Name Server Discovery Report, 2021-3717, 1

Tonry J., et al., 2022, Transient Name Server Discovery Report, 2022-104, 1

Troja E., et al., 2017, Nature, 551, 71

Troja E., et al., 2019, MNRAS, 489, 1919

Troja E., et al., 2020, MNRAS, 498, 5643

Turatto M., 2003, in Weiler K., ed., , Vol. 598, Supernovae and Gamma-Ray Bursters. pp 21–36, doi:10.1007/3-540-45863-8_3

Turner B. M., Sederberg P. B., Brown S. D., Steyvers M., 2013, Psychological methods, 18 3, 368

Turpin D., et al., 2020, MNRAS

Udalski A., Szymański M. K., Szymański G., 2015, Acta Astron., 65, 1

Uno K., et al., 2023, ApJ, 944, 204

Vallely P. J., Kochanek C. S., Stanek K. Z., Fausnaugh M., Shappee B. J., 2021, MNRAS, 500, 5639

Vasist M., Rozet F., Absil O., Mollière P., Nasedkin E., Louppe G., 2023, A&A, 672, A147

Villar V. A., Berger E., Metzger B. D., Guillochon J., 2017, ApJ, 849, 70

Virtanen P., et al., 2020, Nature Methods, 17, 261

Wagoner R. V., 1984, ApJ, 278, 345

Waldmann I. P., 2014, ApJ, 780, 23

Walmsley M., et al., 2020, MNRAS, 491, 1554

Wang L., et al., 2002, ApJ, 579, 671

Wang L., Steeghs D., Galloway D. K., Marsh T., Casares J., 2018, MNRAS, 478, 5174

Wang Q., et al., 2023a, arXiv e-prints, p. arXiv:2305.03779

Wang Q., et al., 2023b, ApJ, 943, L15

Watson C. L., Henden A. A., Price A., 2006, Society for Astronomical Sciences Annual Symposium, 25, 47

Watts A. L., Krishnan B., Bildsten L., Schutz B. F., 2008, MNRAS, 389, 839

Waxman E., Katz B., 2017, in Alsabti A. W., Murdin P., eds, , Handbook of Supernovae. p. 967, doi:10.1007/978-3-319-21846-5_33

Weber J., 1967, Phys. Rev. Lett., 18, 498

Weber J., 1969, Phys. Rev. Lett., 22, 1320

Weiler K. W., Panagia N., Montes M. J., Sramek R. A., 2002, ARA&A, 40, 387

Wen Z. L., Han J. L., 2022, MNRAS, 513, 3946

West R. M., Lauberts A., Jorgensen H. E., Schuster H. E., 1987, A&A, 177, L1

Whelan J. T., Sundaresan S., Zhang Y., Peiris P., 2015, Phys. Rev. D, 91, 102005

Whelan J. T., et al., 2023, arXiv e-prints, p. arXiv:2302.10338

White D. J., 2014, PhD thesis, University of Sheffield, UK

Williams B. J., et al., 2011, ApJ, 741, 96

Williams S. C., Darnley M. J., Bode M. F., Steele I. A., 2015, ApJ, 805, L18

Wooden D. H., Rank D. M., Bregman J. D., Witteborn F. C., Tielens A. G. G. M., Cohen M., Pinto P. A., Axelrod T. S., 1993, ApJS, 88, 477

Woosley S., Janka T., 2005, Nature Physics, 1, 147

Woosley S. E., Arnett W. D., Clayton D. D., 1973, ApJS, 26, 231

Woosley S. E., Pinto P. A., Hartmann D., 1989, ApJ, 346, 395

Woosley S. E., Eastman R. G., Weaver T. A., Pinto P. A., 1994, ApJ, 429, 300

Woosley S. E., Eastman R. G., Schmidt B. P., 1999, ApJ, 516, 788

Wozniak P. R., 2000, Acta Astron., 50, 421

Wright D. E., et al., 2015, MNRAS, 449, 451

Ye F. c., 500, The Book of the Later Han (後漢書). Vol. 102

Yip K. H., et al., 2019, in AAS/Division for Extreme Solar Systems Abstracts. p. 305.04

Yosinski J., Clune J., Bengio Y., Lipson H., 2014, arXiv e-prints, p. arXiv:1411.1792

Zackay B., Ofek E. O., Gal-Yam A., 2016, ApJ, 830, 27

Zechmeister M., Kürster M., 2009, A&A, 496, 577

Zhang Y., Papa M. A., Krishnan B., Watts A. L., 2021, ApJ, 906, L14

Zwicky F., Herzog E., Wild P., Karpowicz M., Kowal C. T., 1961, Catalogue of galaxies and of clusters of galaxies, Vol. I

de Jager C., Nieuwenhuijzen H., van der Hucht K. A., 1988, A&AS, 72, 259

de Jong R. 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. 84460T (arXiv:1206.6885), doi:10.1117/12.926239

de Mink S. E., King A., 2017, ApJ, 839, L7

del Ser D., Fors O., Núñez J., 2018, A&A, 619, A86

van Roestel J., et al., 2021, AJ, 161, 267

van Velzen S., et al., 2021, ApJ, 908, 4