Painting with starlight : optical techniques for the high-contrast imaging of exoplanets

Wilby, M.J.

Chapter 1

Introduction

"The history of astronomy is a history of receding horizons."

– Edwin Hubble

Prior to the 1990s, the study of planetary systems was exclusively confined to our own solar system. Theories on planet formation and evolution have therefore traditionally focussed on explaining all aspects of this single data point, which in turn have shaped our expectations regarding extra-solar planetary (exoplanetary) systems. A subsequent explosion in the number of detected exoplanets has since revolutionised the field; at the time of writing there are now known to be 3,791 confirmed planetary-mass companions in 2,836 planetary systems[1], with this number increasing on an almost-daily basis. This has generated a number of surprises which have forced us to re-evaluate many established aspects of planetary physics. The most prominent of these is the discovery of hot Jupiters; massive gas-giant planets occupying infernally small orbits well inside that of Mercury in our own solar system, which are not expected to be able to form at these locations.

The detailed study of planetary systems orbiting other stars now promises to provide significant insights into some of the most fundamental philosophical questions of *life, the universe and everything*. How unique is the solar system? How did the Earth form? And perhaps most importantly, are we alone in the universe?[2]

Dedicated exoplanet surveys such as the Kepler mission have recently revealed that planets are in fact abundant in our galactic neighbourhood; somewhere between 50-100% of sun-like stars host some form of planetary system (Winn & Fabrycky 2015), while approximately 10-20% are expected to host one or more rocky planets in the Earth-mass category (Fressin et al. 2013; Petigura et al. 2013). By observing young stellar systems at different stages in the process of formation, we can also build up a timeline of how our own solar system, and in turn the Earth, most likely formed.

The exoplanet community is now nearing the capability not only to detect nearby Earth-like planets, but also to determine the composition and dynamics of their atmospheres. This raises the intriguing possibility of detecting biosignatures, such as high atmospheric oxygen content or even industrial pollutants, which could indicate the presence of extraterrestrial life at varying levels of development.

This thesis is primarily concerned with helping to overcome some of the remaining technological hurdles in exoplanet-related instrumentation, to make these goals a reality within the next decade.

[1] Source: exoplanet.eu/, correct as of the 7th of June, 2018

[2] The answers to which, despite the claims of Adams (1986), are significantly more complex than the number *42*.

1.1 Detecting and characterising exoplanets

1.1.1 Overview of exoplanet detection techniques

The first confirmed detection of an exoplanetary system was made in 1992 by Aleksander Wolszczan and Dale Frail, who identified two planetary-mass bodies around the millisecond pulsar *PSR 1257+12* (Wolszczan & Frail 1992). The existence of these planets was revealed by the reflex motion of the host star orbiting around the common centre of mass of the star-planet system, which induces a measurable periodic delay in the arrival times of radio pulses emitted by the dead star.

Although this technique, dubbed *pulsar timing*, is limited to detecting planets around this exotic class of star, the same principle of reflex orbital motion also induces time-varying Doppler shifts in the light spectrum of all stars which host planetary systems. The *radial velocity (RV)* detection technique exploits this fact to look for the corresponding periodic shifts in wavelength of features in the stellar spectrum over the course of a planet's orbit. The first detection of an exoplanet around a main-sequence star was achieved using this technique just a few years later, with the discovery of 51 Pegasi b: a 0.5 Jupiter-mass planet orbiting a sun-like G2-type star (Mayor & Queloz 1995). Any similarity with the solar system ends here however: this object orbits its host star within just four days at a separation six times smaller than the orbit of Mercury, making it an archetypal member of the hot-Jupiter class of exoplanets. The fact that such an extreme object was one of the first detected exoplanets is however not surprising; the size of the RV signal is directly proportional to the mass of the planet, while a short orbital period helps to generate a statistically significant signal within a relatively short observing window.

Figure 1.1 provides an overview of currently known exoplanets by plotting their orbital semi-major axes[3] against the best estimate of planetary mass, with each point coloured according to the method with which it was originally detected. Three major populations can be seen in this way: the "Jupiters", gas-giants on temperate orbits beyond approximately one astronomical unit (AU); the aforementioned "hot Jupiters" on extremely tight orbits; and "super-Earths" with masses of the order of one hundredth of a Jupiter mass (three Earth masses). Planets in this last class are typically detected with orbital separations similar to Mercury (0.1 AU), although this is primarily due to the sensitivity limits of current surveys (e.g. Howard et al. 2010): planetary population synthesis simulations predict that this distribution should extend smoothly to encompass the parameter space in which the solar system planets exist (Mordasini et al. 2009).

It is apparent that the hot-Jupiter and super-Earth populations feature a large number of detections made using the *transit* method, where photometric monitoring of the star's light curve reveals characteristic, periodically repeating dips due to the planet passing in front of its host star. This technique is also heavily biased towards detecting giant close-in planets: the size of the transit signal scales with planet radius squared, while the likelihood of a planet's orbit being sufficiently well aligned with Earth for it to transit the host star becomes rapidly more improbable with increasing orbital distance. The source of the majority of these transit detections is the highly successful Kepler space observatory mission (Borucki et al. 2010), which provided continual high-precision photometric monitoring of stars in a 116 square degree

[3] Equivalent to the orbital radius for a circular orbit.

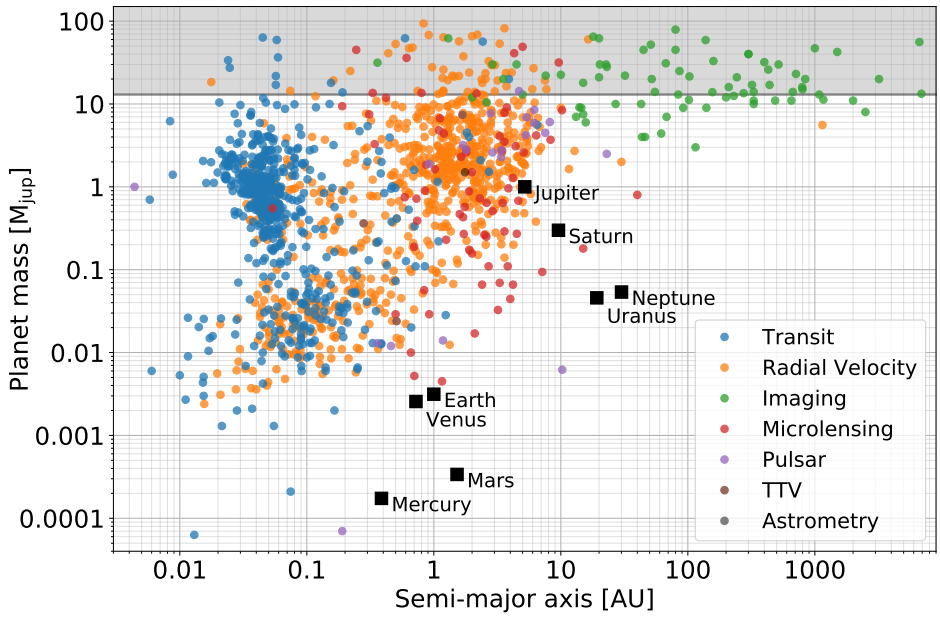

Figure 1.1: Logarithmic plot orbital semi-major axis against mass for all confirmed sub-stellar companions below 100 Jupiter masses (M_{jup}), coloured by discovery method. The horizontal grey line and shaded region denotes the 13 M_{jup} deuterium-burning mass limit, above which these objects begin to more closely resemble brown dwarf "failed stars" rather than planets (Spiegel et al. 2011). The properties of the solar system planets are also plotted for comparison. Data from http://exoplanet.eu/ (May 2018).

field of view in the constellation of Cygnus, between 2009 and 2013. This alone is currently responsible for over $1,500$ confirmed planet detections, with more than $2,000$ candidates still awaiting confirmation.

The transit and RV techniques are are by far the most productive methods to date in terms of number of discoveries, but are by no means the only methods by which exoplanets may be detected. High-precision *astrometry* can be used to look for the spatial rather than spectral modulation of starlight due to reflex orbital motion induced by a planetary system, achieved by imaging the star's changing position in the plane of the sky. Only a handful of planets have so far been detected via this method, however with the recent launch and first two data releases of the all-sky astrometry satellite GAIA (Gaia Collaboration et al. 2016a,b; Molnár et al. 2018), this number is expected to increase significantly in the coming years.

Microlensing events occur when a stellar system containing a planet passes in front of a background source, with both foreground objects acting as gravitational lenses and weakly amplifying its brightness. This method has detected some of the closest known analogues to solar system planets in terms of mass and orbital radius (e.g. Gould et al. 2014; Shvartzvald et al. 2017), but has the major drawback that lensing events are highly improbable, thus providing no opportunity for follow-up or further characterisation of the detected planet.

In systems with at least one known transiting exoplanet, the identification of *transit timing variations* (TTVs) in the signal, where the exact timing of the transit event is seen to vary from one orbit to the next, can be used to infer the existence of additional non-transiting planets in the system. This technique has proved capable of detecting

the gravitational perturbations of additional planets down to 7.5 Earth masses (Masuda 2014) and has also been proposed as a possible technique for detecting super-Galilean exomoons orbiting around giant transiting planets (Kipping 2009a,b).

All of the aforementioned detection methods are inherently indirect techniques, relying on effects induced by the planet on the light of the host star. *Direct imaging* techniques on the other hand attempt to resolve the light of the planet from that of the star, allowing us to study these planetary systems in great detail. As shown in Fig. 1.1, direct imaging probes a very different part of the exoplanet population to other techniques: currently it is only capable of detecting massive giant planets on extremely wide orbits, which are typically young and still hot from their formation and so particularly luminous at near-infrared wavelengths. This limitation is due to the severe technical challenge of teasing out faint planetary signals from a halo of starlight which is naturally many orders of magnitudes brighter. This task, also known as *high-contrast imaging* (HCI), only becomes more challenging when trying to detect signals closer in towards the host star. However, with rapid advances currently being made in instrumentation and data reduction techniques, the field is moving towards the capabilities required to image more conventional planetary systems, with an ultimate goal of directly resolving Earth-like planets around nearby stars.

The work contained in this thesis attempts to address some of the main instrumentation challenges associated with high-contrast imaging; the various benefits, difficulties and methodologies of this technique are therefore presented in significantly more detail in Sec. 1.3.

1.1.2 From detection to characterisation

Due to the limited amount of planetary information provided by most detection techniques, the exoplanet field has until recently broadly focussed on identifying the existence of planetary-mass companions and determining their orbital properties. A particular focus is made on the search for "Earth 2.0"; a planet with a sufficiently similar mass to Earth, orbiting in the habitable zone of a sun-like (F, G or K-type) star where liquid water may exist on the surface. For this reason the super-Earth class of exoplanets is of significant interest: not only are they currently the most similar known objects to Earth in terms of mass, but their nature is largely unknown. Are they rocky and Earth-like, or in fact more like low-density, atmosphere-dominated mini-Neptunes?

This question can already be partially addressed by determining the overall density of these objects, from which it is possible to constrain their bulk composition. This feat can be achieved via complementary transit and RV study, where the transit measurements constrain the planetary radius while RV provides the mass[4]. Fig. 1.2 shows the result of this characterisation for a subset of transiting exoplanets, directly compared to simulated mass-radius relationships for different idealised planetary compositions. The exoplanet population appears to undergo a smooth transition between gaseous (hydrogen-dominated) and solid (rock/iron-dominated) compositions at around ten Earth masses, which is consistent with the trend seen in the solar system. While the deduction of exact planetary composition in this way is inherently degener-

[4]Specifically, RV provides only a lower limit on the planetary mass, $m_{RV} = m_p \sin(i)$; this is degenerate with the (often unknown) system inclination i, since RV only measures the component of orbital motion along our line of sight. For transiting planets however the inclination must by necessity be very close to 90°, which in turn facilitates an accurate measurement of the true planetary mass m_p.

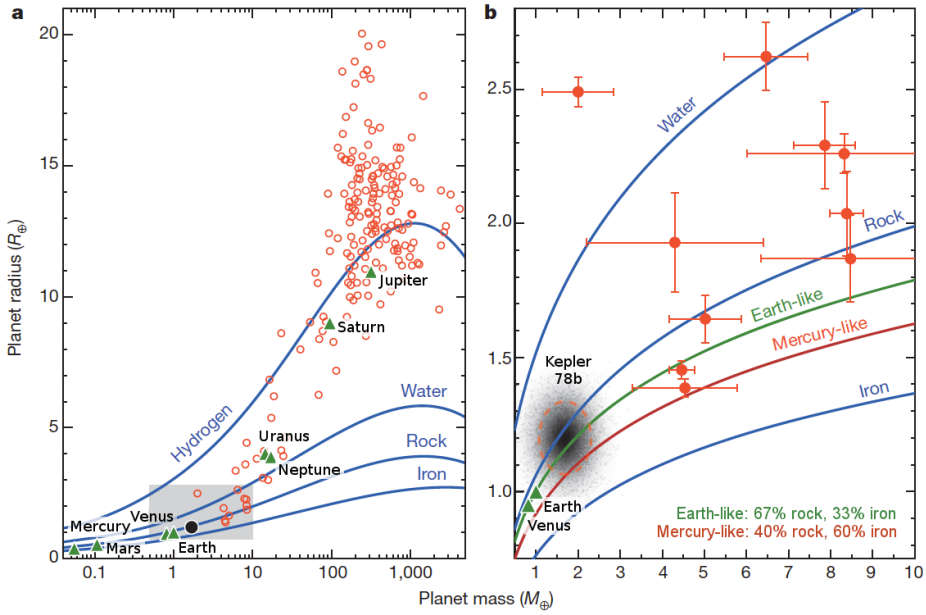

Figure 1.2: Planetary mass and radius (in terms of Earth mass and radii) of a subset of exoplanets which have been jointly characterised by transit and RV study (open circles). The over-plotted blue curves show the expected mass-radius relationships for idealised planetary models consisting entirely of the labelled constituent. The right panel shows a zoom-in on the non-gaseous super-Earth region, including error bars on detections and the then newly-discovered rocky Kepler-78b. Figure credit: Howard et al. (2013)

ate due to the sheer number of possible constituents, it nonetheless provides a useful indication of whether a given planet is likely to be most similar in nature to Earth, Neptune, or Jupiter.

Recent years have also seen the start of the next phase of characterisation of these exoplanetary systems: detailed single-object studies which are capable of determining planetary parameters beyond mass, radius and orbital separation, and hence can provide significant insights into the true nature of these bodies. The most productive way to achieve this is by using *high-resolution spectroscopy* (HRS) instruments such as CRIRES at the VLT (Käufl et al. 2004). Observing planetary light with spectral resolving powers of the order of 100,000 makes it possible to resolve individual spectral lines and molecular feature bands for simple molecules such as oxygen, water and methane in the atmospheres of these planets. Since the spectral "fingerprint" of these molecules can be well characterised via laboratory measurements, it is possible to construct high-resolution templates of their absorption and emission features. By cross-correlating observed spectra with such a template, it is possible to combine the signal from thousands of individual features in the planet spectrum and hence detect the presence of that molecular species in its atmosphere with much higher signal-to-noise than by looking for individual lines in the spectrum itself (e.g. Snellen et al. 2010; Brogi et al. 2012b; de Kok et al. 2013). By using such methods, the sub-field of astrobiology can hope to detect the spectral signatures of molecules such as oxygen and methane in exoplanetary atmospheres: since these molecules are not expected to be present in large quantities due to geological processes alone, they may constitute bio-marker signals which indicate the presence of life. In addition to the composition

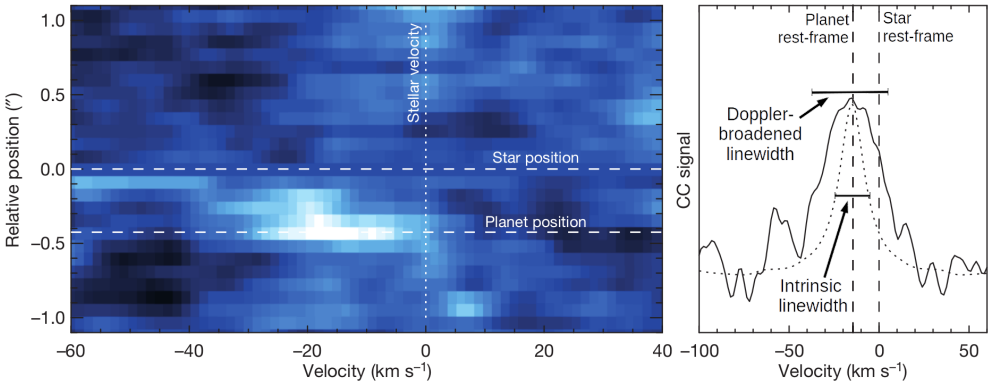

Figure 1.3: Determination of the spin rate of the directly-imaged planet β Pictoris b, using the molecular template cross-correlation method on data from the CRIRES instrument (Snellen et al. 2014). Left: a datacube of the cross-correlation signal from a water (H_2O) plus carbon monoxide (CO) spectral template, as a function of Doppler-shift velocity on the x-axis and angular separation from the star on the y-axis. The planet can be seen as a bright feature at -0.4". Right: A cut through the datacube at the planet's location, showing the average spectral line profile compared to the intrinsic linewidth expected for a non-rotating object. The additional broadening of the line reveals that the planet rotates with a day length of just eight hours, while the average velocity offset from the stellar rest frame measures the orbital velocity of the planet around its host star.

of planetary atmospheres, the identification of telltale atomic and molecular signals can also be used to provide detailed information about other planetary properties, including the atmospheric vertical temperature and pressure structure (Stevenson et al. 2014), and surface gravity (Martín & Zapatero Osorio 2003). It is even possible to detect the signatures of "disintegrating exoplanets", where material is being ejected from the atmosphere or surface of extremely close-orbiting planets due to the impact of powerful stellar winds (Brogi et al. 2012a; Rappaport et al. 2012; Ridden-Harper et al. 2016).

The molecular template cross-correlation technique can also be used to measure the planet's rotation speed and hence determine its day length, by detecting a slight broadening of the combined line profile with respect to the intrinsic width of the spectral features in question. This signal is caused by the rotating planet's limbs moving towards and away from us respectively, causing a corresponding Doppler shift in components of the molecular signal originating in these regions[5]. This measurement has now been achieved for a number of planets including the archetypal directly-imaged planet β Pictoris b (Snellen et al. 2014) shown in Fig. 1.3. This information can provide an insight into how the spin rates of planets change as they age (Schwarz 2017, PhD thesis), which is an important step towards understanding the planet formation process.

High-resolution spectroscopy observations of planetary atmospheres can be obtained in two main ways. The first is to place an optical fibre or spectrograph slit at the location of a directly-imaged planet to isolate its signal, as was done in the case of β Pictoris b. The other main method is the use of *transmission spectroscopy*, where the atmospheric signature of a transiting planet can be measured from the small fraction of starlight which passes through the planet's atmosphere during the transit. The planetary radius appears fractionally larger at wavelengths corresponding to the ab-

[5]This perhaps gives a new, and quite literal, meaning to the quote at the beginning of this introduction.

sorption features of molecular species present in the atmosphere, making the transit signal marginally deeper in these wavelengths. An extension to this is possible for hot Jupiters, where the planetary signal is often bright enough with respect to that of the star to be a detectable component of the total, making it observable even when out of transit. Monitoring the changing spectrum of this planetary component at different points along its orbit allows us to build up a *phase curve* for the planet. Since these planets are most often tidally locked (always pointing the same face towards the star), it is possible to build up a one-dimensional longitude map of the temperature and composition of the unresolved planet's upper atmospheric layers (Knutson et al. 2007; Stevenson et al. 2014). The hot Jupiter population is therefore becoming increasingly well characterised: the challenge for the immediate future is therefore to push towards achieving the same level of detailed characterisation for temperate Earth-sized planets. It is in this effort where an inter-optimised combination of high-resolution spectroscopy with high-contrast imaging techniques (Snellen et al. 2015) stands to make a major impact, by facilitating the detailed characterisation of directly-resolved planets, including those not detectable by transit or RV methods (e.g. Schwarz et al. 2016a; Hoeijmakers et al. 2018).

1.2 Planet formation

1.2.1 The planet-disk connection

Following the description of Williams & Cieza (2011), stars and planetary systems are born from the gravitational collapse and fragmentation of over-dense regions of *giant molecular clouds* (GMCs), triggered by external factors such as gravitational interactions or shock waves from nearby supernovae. Any amount of net angular momentum in the collapsing cloud causes it to spin up as it contracts, resulting in a thick disk of material surrounding a central protostar, which is not yet dense enough to initiate nuclear fusion. Over the course of typically 1 million years, the protostar contracts and accretes material from the surrounding gas- and dust-rich *proto-planetary disk* (PPD) before initiating nuclear fusion and transitioning into either a T-Tauri or Herbig Be/Ae class pre-main-sequence star, depending on its mass (Herbig 1960). This star- and planet-formation process from GMC collapse through to a mature solar system is illustrated in Fig. 1.4, along with examples of each class of object.

As the radiation output of the newly-formed central star intensifies, it begins to clear the gas and small dust grains in the proto-planetary disk from the inside-out through a combination of photo-evaporation and radiation pressure. These so-called *transition disks* are defined by the presence of a central cavity devoid of material, but often also display numerous interesting features in the remaining disk material such as narrow gaps, spirals and local over-densities, which may be a signature of ongoing planet formation.

It is during this transition process that planets are indeed expected to form out of the dust- and gas-rich material of the proto-planetary disk. There are three main competing theories as to how this occurs, and all may be responsible for forming planets under different conditions: no single theory is currently capable of explaining the full population of known exoplanets. The first is *core accretion* (Pollack et al. 1996; Hubickyj et al. 2005), whereby planets form slowly in a bottom-up manner from the collisional accretion of dust grains into rocky *planetesimals* of increasing size. In the

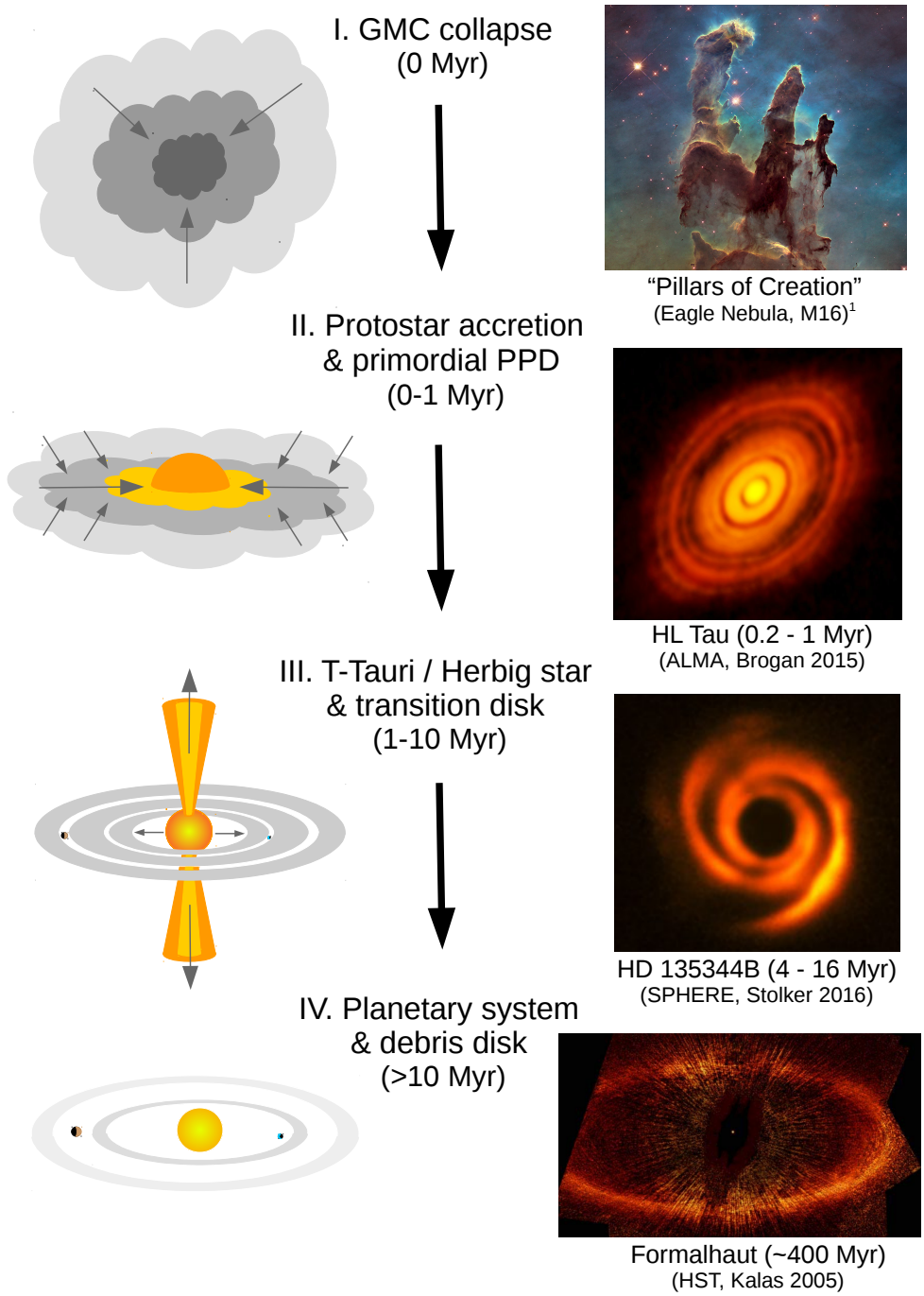

Figure 1.4: Schematic diagram of the four main phases of star and planet formation.
[1]www.nasa.gov/image-feature/the-pillars-of-creation

case of giant planets, these cores eventually grow massive enough to accrete gas from the surrounding disk, thus forming their massive atmospheres. *Gravitational instability* (Boss 2001; Mayer et al. 2002) on the other hand proposes a top-down approach whereby natural turbulence or the rapid cooling of regions in the disk can result in significant gas and dust overdensities, which accumulate matter until gravity takes over: a planet then forms rapidly by runaway accretion on timescales of just a few orbits. Since both gas and dust are accreted at the same time, gas-giant cores then form later by differentiation. In the case of very massive exoplanets or brown dwarf companions, a third method of *binary fragmentation* (Chabrier et al. 2014) is also a possibility; here the initial GMC can fragment as it collapses, forming a secondary protostellar core in the same manner as binary stars form, although this secondary object never becomes massive enough to achieve fusion.

Once the proto-planetary disk has been fully evaporated by the central star after approximately 10 million years (Myr), the formation of gaseous planets must come to an end, since there is no longer any free gas available to accrete and thereby form their atmospheres. Only the newly-formed planetary system and a residual gas-poor *debris disk* (Wyatt 2008) remain, the latter consisting of rocky and icy grains large enough such that the effect of radiation pressure is negligible, plus smaller components from the subsequent collisional fragmentation of proto-planetary bodies. This is the final product of the planet formation process, and the direct equivalent to our solar system including the asteroid and Kuiper belts, and an Oort cloud of material ejected onto extremely eccentric orbits by close gravitational encounters.

1.2.2 Observational challenges to current planet formation theory

The first major issue with the core accretion formation mechanism is that it is difficult to form metre-sized bodies by the process of collisional build-up from mm-scale dust. This is firstly due to the dynamics typically present in these disks; cm-sized bodies tend to approach each other at velocities which will more often shatter the nascent planetesimals under typical conditions, rather than having them stick together (Ros & Johansen 2013). A potentially larger issue is that cm to m-sized bodies begin to decouple from the gas component of the disk, which orbits at sub-Keplerian velocities due to the effect of radiation pressure. These bodies then spiral in towards the central star due to gas drag and disappear on timescales of less than a hundred years (Brauer et al. 2007); far too quickly to assemble planets. It has been proposed that pressure traps in the gas could be able to halt this inward drift and help accumulate planet-forming material (Meheut et al. 2012). The existence of such a pressure trap typically requires a planet to already exist in the system however, which leads to a chicken-and-egg problem. This process of bottom-up planet formation is also comparatively slow in forming gas-giant planets: building up a 10 M_{Earth} planetary core, which is thought to be needed in order to accrete a significant atmosphere, is expected to take a comparable time to the transition disk clearing timescale (Rice & Armitage 2003). This makes it especially challenging to explain the formation of extremely wide-orbit giant planets such as Uranus and Neptune, where collision rates between accreting bodies are low and hence the growth rate is especially slow. Gravitational instability on the other hand is able to form planets significantly faster, but is only thought to be feasible in extremely massive disks such as those around Herbig stars (Mayer et al.

2005): it is therefore unlikely to be responsible for giant planets around solar-type stars.

The discovery of a sizeable population of hot Jupiters was initially surprising, not just because no such object exists in our solar system but because accepted planet formation theories hold that large gas-giant planets can only form further out in the solar system. Bias-corrected statistics from the Kepler mission in fact show that only approximately 0.5% of stars host hot Jupiters, making them relatively rare, however it is still necessary to determine the mechanism by which this population came to exist. It is most likely that some form of migration, whether through slow inward migration due to interaction with the disk (Chambers 2006) or from gravitational encounters between planets (Weidenschilling & Marzari 1996) is responsible for the existence of these objects. This allows them to form in the same manner as conventional gas-giants (although this formation mechanism is itself uncertain) out beyond the snow-line of the proto-planetary disk where it is cold enough for ice species to condense onto dust grains, providing significantly more solid material for the accretion process.

At the opposite extreme, the wide-orbit giant planets to which direct imaging is most sensitive are often massive enough to lie in the grey area in terms of what may be considered a planet. The fusion of deuterium becomes possible inside the cores of objects larger than approximately 13 M_{jup}, beyond which point these bodies become increasingly similar to low-mass stars rather than planets. Studies of these brown dwarf "failed stars" is a diverse field in its own right, and there is much ongoing discussion regarding the exact mass limit cut-offs (Spiegel et al. 2011). To add to this uncertainty, accurately determining the masses of these directly-imaged exoplanets is also a significant challenge: no estimate of this quantity is provided directly by observations, which give only the relative photometry and hence luminosity of the target. Models which constrain the mass-to-light ratio of exoplanets must therefore be used to convert between the two. Poor constraints on the planet formation mechanism and hence the amount of internal heat planets start with, and how fast this is emitted as the planet ages, therefore lead to significant systematic uncertainties in mass estimates when trying to characterise these systems. The two main model classes used to estimate directly-imaged planet masses are referred to as *hot-start* (e.g. Baraffe et al. 2003) and *cold-start* models (Fortney et al. 2008) respectively, which loosely correspond to the gravitational instability and core accretion planet formation methods mentioned earlier; the slow accretion of rocky bodies allows heat to radiate away more effectively during the formation process, resulting in dimmer planets, while gravitational instability forms planets significantly faster, making them hotter for any given age. Understanding planetary formation mechanisms by observing proto-planetary disk environments will therefore ultimately allow for significantly more robust mass estimates of these planets.

1.2.3 Observing proto-planetary disks

Young T-Tauri and Herbig stars are excellent sites for studying the formation of planetary systems, since the proto-planetary disks they host are typically at some stage of the transition process to a mature system. By combining detailed single-object studies with comparative analyses of many disks of different ages, it is possible to begin constraining an overall picture of the planet-formation process.

A good rule of thumb when observing proto-planetary disks is that the observ-

ing wavelength is more or less equal to the size of dust particles that dominate the observed signal. The reasoning behind this is non-trivial, stemming from a combination of the wavelength-dependent photon scattering and emission cross-sections of dust particles and number-density arguments (Draine 2006). Near-infrared (NIR) observations with wavelengths of the order of microns therefore trace the distribution of micron-sized dust, specifically from the scattering of stellar photons by the protoplanetary disk. Due to the high opacity of the disk at these wavelengths, this signal typically corresponds to only the outermost layers of the disk. These "scattered-light" observations can therefore be used to probe the surface features of the disk (such as gaps, spirals and cavities), measure the relative scale height at different radii in the disk, and constrain the dust grain composition by making series of observations at different wavelengths.

By contrast, observations at (sub-)mm wavelengths are primarily sensitive to the thermal emission of mm-sized dust grains. These typically lie close to the mid-plane of the disk, due to dynamical processes which stratify the disk and allow larger grains to settle towards the ecliptic plane. Sub-mm observations using large interferometric arrays, such as the Atacama Large Millimeter Array (ALMA), are also able to measure the presence of certain gas species (such as carbon monoxide, CO) which produce molecular emission features in the sub-mm. These gas species are often used as tracers for more interesting but less visible molecules such as hydrogen, which makes up the majority of the gas mass of the disk (Schwarz et al. 2016b). Complementary NIR scattered-light and sub-mm emission observations of the same target (e.g. de Juan Ovelar et al. 2016; Muro-Arena et al. 2018) can therefore provide a much more complete picture of the disk in question, including constraints on the dust-to-gas ratio, vertical structure and the total disk mass. This is crucial for understanding what material is available for planet formation as a function of distance from the star, and hence what types of object are most likely to form at different locations in the disk.

1.3 Direct imaging of exoplanets

1.3.1 Challenges and rewards

While the premise of direct imaging is a straightforward one, achieving the goal of resolving Earth-like planets around even the nearest stars is anything but simple. This is primarily due to the extreme flux ratio between planetary signals and the host star, which is typically of the order of 10^{-4} to 10^{-6} at near-infrared wavelengths for young, thermally luminous giant planets. This drops to contrasts of 10^{-8} to 10^{-10} at visible wavelengths for mature Jupiter-like and Earth-like planets respectively, whose brightest signal is from reflected starlight. Coupled with the fact that these planets are separated by a fraction of an arcsecond (1/3600th of a degree) from the star in the plane of the sky, any instrument hoping to disentangle the two signals has to make use of advanced starlight rejection techniques and have exceedingly precise control over all of its constituent optical components.

A common analogy for this task is trying to detect a firefly (which emits typically 10^{-6} W of power[6]) fluttering around a lighthouse of approximately 1000 W, which is equivalent to a contrast ratio of the order of 10^{-9}. However, in this particular analogy

[6]As adopted in https://what-if.xkcd.com/151/

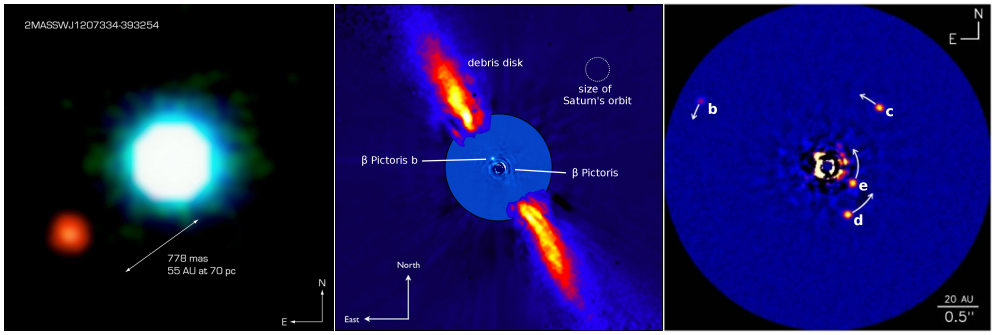

Figure 1.5: Three of the most notable directly-imaged planetary systems. Left: first directly-imaged planetary-mass companion 2M1207 b (Chauvin et al. 2004). Middle: β Pictoris b and debris disk (Lagrange et al. 2010). Right: HR 8799 four-planet system (Marois et al. 2008, 2010b).

we are looking through a telescope at Leiden Observatory at Dover lighthouse 240 km away[7], and trying to detect the firefly when it is only 12 cm away from the lamp, equivalent to an angular separation 100 milli-arcseconds (mas).

If this extreme technical challenge can be overcome however, direct imaging presents major advantages over other techniques for detecting solar system analogues. Aside from the opportunities for detailed atmospheric characterisation already outlined in Sec. 1.1.2, a major advantage is that it does not require planets to complete a significant fraction of an orbit in order to produce a statistically significant detection. This is a major limitation for transit and RV studies, and makes it extremely difficult to detect solar-system like bodies with these methods without running observing campaigns spanning years or even decades. By capturing the entire planetary system in one go, direct imaging is also a versatile tool for studying multi-planet systems and the interplay between nascent proto-planetary disks and their planetary progeny in young, still-forming systems.

The first directly-imaged planetary-mass companion was the 3-10 Jupiter-mass (M_{jup}) 2M1207 b (Chauvin et al. 2004), for which the discovery image is shown in the left-hand panel of Fig 1.5. This object has a contrast ratio with its host of only 10^{-2}: since the host is itself not a main sequence star but a brown dwarf approximately five times more massive than 2M1207 b itself, this system more closely resembles a very low-mass binary than a conventional star-planet system. Images of two other notable directly-resolved planetary systems are also shown in Fig. 1.5; β Pictoris b (Lagrange et al. 2010) has already been introduced in Sec. 1.1.2, and is shown in the middle panel to be surrounded by a significant debris disk. β Pictoris b orbits in a nearly edge-on but non-transiting configuration (Wang et al. 2016), which made its most recent closest approach in 2017. This raised hopes for the observation of circum-planetary material transiting the host star (Stuik et al. 2017), but no significant detections have yet been reported. The final panel shows the multi-planet system HR 8799 (Marois et al. 2008, 2010b), which consists of four giant planets between 4-10 M_{jup} orbiting in a nearly face-on configuration with periods of 45-460 years. During the last decade, repeated observations of these systems has allowed us to follow the orbital motion of these planets[8] and perform detailed characterisation studies (e.g. Marley et al. 2012;

[7]Ignoring for now the fact that obstruction to our line of sight by the Dutch barrier dunes, let alone the Earth's curvature, would in reality make this feat impossible.

[8]Time-lapse videos of which can be found at https://jasonwang.space/orbits.html

Bonnefoy et al. 2016; van Holstein et al. 2017).

1.3.2 Current state of the field

Figure 1.6 provides a quantitative overview of the field of direct imaging, in terms of its angular separation and planet-star contrast capabilities. Here a selection of the most notable directly-imaged planets to date are shown as orange named circles, in addition to reflected-light estimates for the solar system planets when seen in an edge-on configuration from a distance of 10 parsecs, with grey lines denoting how their detectability will vary over the course of an orbit. Overlaid on this are the performance curves of various high-contrast imaging instruments, which in general show a steady improvement between subsequent generations. NIRC2 on the W.M. Keck telescope and VLT-NACO (Lenzen et al. 2003) are examples of "zeroth-generation" instruments which began operation in the early 2000s; they were not originally designed with the direct imaging of exoplanets in mind but have since been extensively used for this task by the astronomical community. The Gemini Planet Imager (GPI, Macintosh et al. 2008) and VLT-SPHERE (Beuzit et al. 2008) came online in 2013-2014 and represent the first (and current) generation of dedicated ground-based high-contrast imagers. These instruments feature extreme adaptive optics systems along with dedicated spectroscopic and polarimetric capabilities designed to significantly enhance the ultimate contrast performance these instruments can obtain. Using advanced data reduction techniques it is now possible to achieve a *5σ contrast*[9] of approximately 10^{-6} at angular separations of 200 mas (Zurlo et al. 2016).

The minimum angular separation these instruments can achieve has however not improved greatly between these two generations. This is primarily due to the performance limits of the adaptive optics and coronagraphic systems they employ, but is also fundamentally constrained by the resolution limit of the 8 m-diameter telescopes on which they are mounted. The resolving power of any telescope is ultimately set by the diffraction limit: the *Rayleigh criterion* states that the minimum angular separation of two point sources at which they may still be resolved is $\theta_{\text{min}} = 1.22\ \lambda/D$, where λ is the observing wavelength and D is the telescope primary mirror diameter. This means that the *inner-working angle* (IWA) of an instrument is ultimately a trade-off between telescope primary mirror diameter D and the observing wavelength λ. Space missions such as the Hubble Space Telescope (HST) and the upcoming James Webb Space Telescope (JWST) are therefore limited in terms of IWA since the largest single-piece mirror which can fit into launch vehicles is currently around 2.5 m; this is a significant disadvantage compared to larger ground-based facilities when it comes to imaging close-in planets.

1.3.3 Outlook for next-generation instrumentation

The next generation of HCI instrumentation for 30-m-class extremely large telescopes (ELTs), such as EPICS (Kasper et al. 2010) for the European ELT and the Planet Formation Imager (PFI, Monnier et al. 2014) for the Thirty-Meter Telescope (TMT) as shown in Fig. 1.6, are due to see first light in the 2020s and early 2030s. This will lead to a significant improvement in both IWA and contrast capabilities, but only if

[9] A common contrast metric in high-contrast imaging, where a point source must be at least five times higher than the noise and/or variability of the residual starlight field in order to count as a detection.

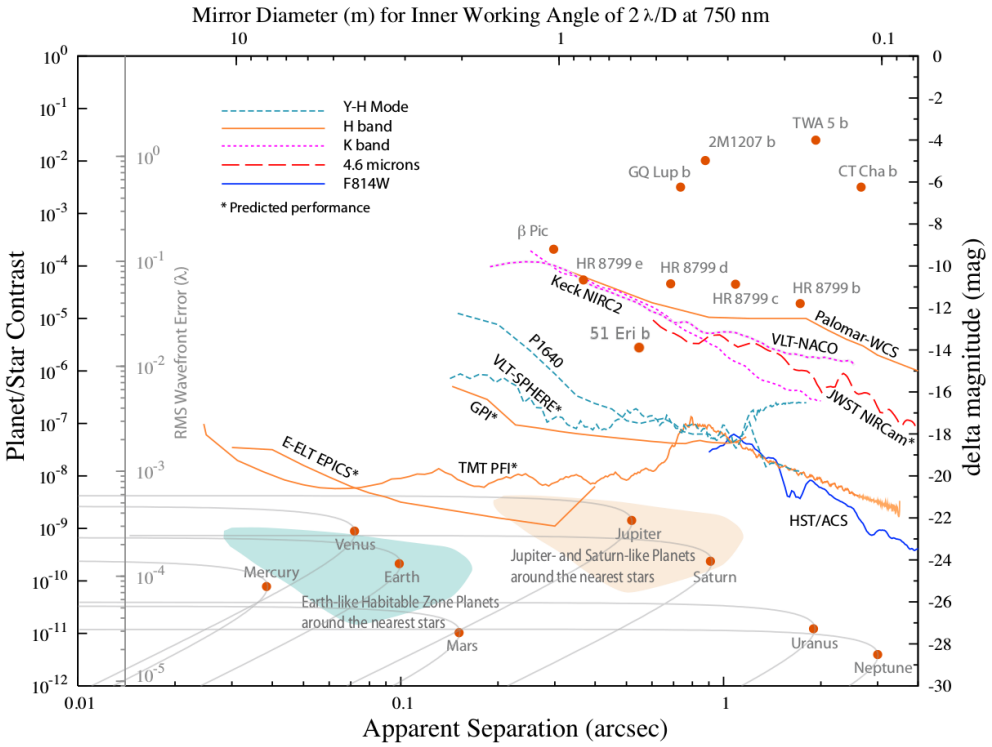

Figure 1.6: Ultimate contrast performance of current and future high-contrast imaging instruments, as a function of apparent (angular) separation from the host star. Also plotted are the solar system planets as seen in reflected light from a distance of 10 parsecs (32 light years), and a selection of the most notable giant exoplanets discovered to-date by direct imaging. Figure credit: Lawson et al. (2012); Mawet et al. (2012), updated with the subsequently discovered 51 Eridani b (Macintosh et al. 2015).

the imaging instruments concerned can actually reach the diffraction limit. In order to achieve the extreme contrast goals, efficient coronagraphs or starlight suppression techniques must first be developed; on current instruments these have functional IWAs that are much larger than the diffraction limit. The contrast of raw images taken with ground-based instruments is also significantly degraded by the distorting effects of atmospheric turbulence and optical instabilities inside the instrument. Extremely high-performance adaptive optics systems will therefore be required, even by the standards of current high-contrast imagers: the difficulty of accurate atmospheric wavefront correction only increases with increasing telescope diameter, and the accuracy of instrumental wavefront control is already the main limiting factor of current HCI instruments. These factors are discussed explicitly in Sec. 1.4.

The ultimate goal of high-contrast imaging is to detect solar system analogues, however Fig 1.7 also illustrates the interesting parameter space which contains a large number of currently known exoplanets which have not yet been directly imaged. This exoplanet population is predicted to form a broad trend with logarithmically increasing contrast requirements as a function of angular separation, since exoplanets on wider orbits will receive significantly less flux from their stars and so are correspondingly fainter when observed in reflected light at visible wavelengths. By comparing this with the predicted contrast performance of upcoming ELT instruments in Fig. 1.6,

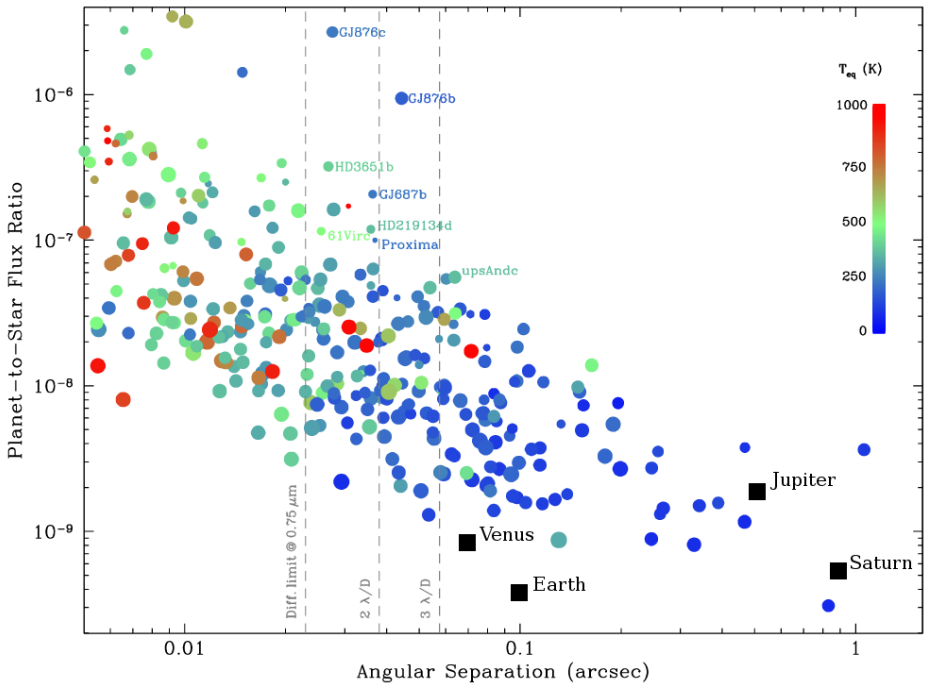

Figure 1.7: Estimated reflected-light contrast ratios of known non-directly-imaged exoplanets at an observing wavelength of 750 nm (R/I-band), calculated from known orbital parameters and estimates of likely planetary albedos, and colour coded by expected surface temperature. Vertical dashed lines indicate the diffraction-limited resolution of the 8-m VLT, plus the two and three diffraction-width cut-offs which are the main performance targets for next-generation coronagraphic imaging systems. Figure credit: Lovis et al. (2017).

we can expect that a large number of these planets should be resolvable within the next two decades. This includes the recently-discovered Proxima Centauri b (Anglada-Escudé et al. 2016), which at its point of greatest elongation sits at an angular separation of 37 mas, with an estimated contrast ratio of 10^{-7}. As the closest exoplanet to Earth, which is also potentially rocky in nature and sits on the inner edge of the nominal liquid-water habitable zone of Proxima Centauri, this object is a target of extreme interest for further characterisation studies. It is worth noting that the spatial resolution of a 30-m-class telescope operating in the near-infrared ($\lambda \approx 3\ \mu$m), such as the upcoming ELT-METIS instrument (Brandl et al. 2014), is roughly equal to that fundamentally achievable by 8-m-class instruments operating at visible wavelengths ($\lambda \approx 750$ nm), such as the current-generation VLT-SPHERE-ZIMPOL subsystem (Thalmann et al. 2008): in principle, both are therefore capable of resolving Proxima Centauri b. This goal has motivated an ongoing upgrade initiative for the SPHERE instrument, which will install cutting-edge coronagraphs, wavefront control techniques and high-resolution spectroscopy capabilities in order to bring the performance of the instrument close to the fundamental limits set by the telescope (Lovis et al. 2017).

1.4 The high-contrast imager's toolbox

The following section provides an overview of the various technologies and techniques which can be used to overcome the twin challenges of contrast and angular separation in current-generation high-contrast imaging instruments.

1.4.1 Coronagraphy

The fundamental purpose of a coronagraph is to act as an angular filter; it should suppress as much of the starlight contribution as possible at the location of the planetary signal, whilst simultaneously not degrading the strength of the planetary signal itself. The first coronagraph was invented by Bernard Lyot (Lyot 1939) in order to study the solar corona by masking out the disk of the sun, and this name has been carried over into the exoplanet field where it is instead used to observe circumstellar material by filtering out the light of the central star. There now exists a whole suite of different coronagraph designs which achieve this filtering in a variety of ways, and are comprehensively summarised by Mawet et al. (2012). These are broadly split into two main families, the difference between which is illustrated in Fig. 1.8. Focal-plane coronagraphs such as the Lyot coronagraph use an opaque mask to block out the starlight in an intermediate image plane, although they also typically use upstream and downstream pupil stops in order to optimally suppress diffraction effects (e.g. Martinez et al. 2007). Pupil-plane coronagraphs, on the other hand, exclusively use apodising optics in the telescope pupil to modify the point-spread function (PSF) of the telescope, creating regions of destructive interference, or dark zones, where the natural diffraction pattern of the telescope is significantly suppressed. This can be done either via amplitude such as with the shaped pupil coronagraph (Kasdin et al. 2004), or phase as with the Apodizing Phase Plate (APP) coronagraph (Kenworthy et al. 2010c). Interferometric nulling techniques such as Baudoz et al. (2000) can also be used for starlight suppression, while sparse aperture masking techniques (e.g. Cheetham et al. 2016) can be used to achieve angular resolutions below the conventional 1.22 λ/D resolution limit, albeit at the cost of contrast performance.

In order to make the most of the telescope's diffraction limit, it is important to develop coronagraphs which have an IWA which is as small as possible; the Apodised-Pupil Lyot Coronagraph (APLCs) is currently the workhorse coronagraph of VLT-SPHERE (Guerri et al. 2011), but is limited to inner-working angles of 3-4 λ/D by the size of the obscuring focal-plane mask. Small IWA focal-plane coronagraphs such as the vector vortex (Mawet et al. 2010) and Phase-Induced Amplitude Apodisation (PIAA, Guyon 2003) coronagraphs have been developed to push this significantly further by using phase rather than amplitude masks in the focal plane: this results in higher off-axis transmission whilst nulling out any on-axis starlight with greater precision. In laboratory environments these have been shown to reach extinction ratios of 10^{-8} to 10^{-10} at angular separations down to 1.5 λ/D, meeting the requirements for imaging solar-system analogues. The fundamental challenge with focal-plane coronagraphs however is that as the IWA becomes smaller, the coronagraph becomes increasingly sensitive to errors in instrumental alignment: any shift to the central star will cause it to move out of the nulling region and re-appear in the image, significantly degrading the contrast performance of the coronagraph. This is most often caused by high-frequency vibrations, due to cooling systems or resonances in the telescope and

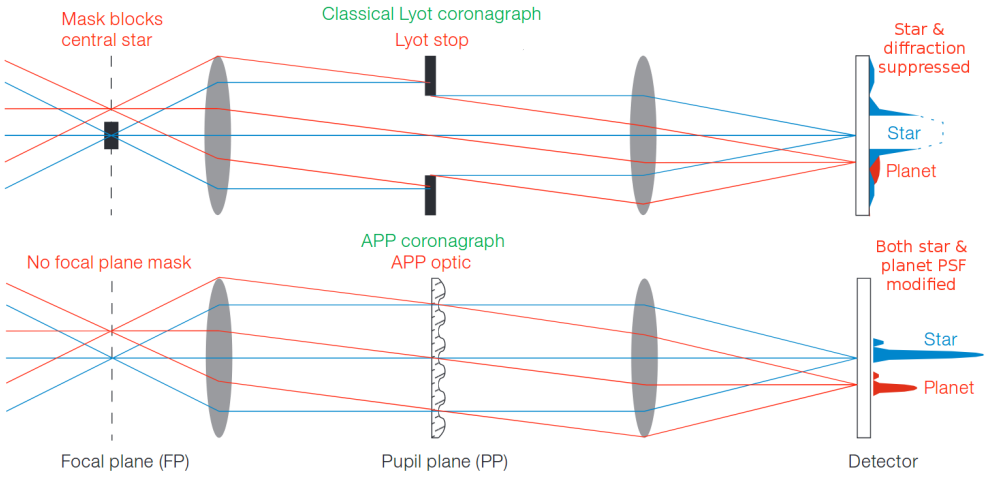

Figure 1.8: Optical layout of two example coronagraphs: Top: the traditional Lyot focal-plane coronagraph, and Bottom: the Apodizing Phase Plate (APP) pupil-plane coronagraph. Here the focal plane (FP) denotes an intermediate image plane, while the pupil plane (PP) corresponds to the re-imaged telescope aperture. In both diagrams blue rays trace the light path of on-axis starlight (which should be rejected), while the red rays illustrate how an off-axis source (i.e. the planet) is transmitted through the system. Figure credit: Kenworthy et al. (2010b).

instrument structure. Extremely fine alignment tolerances as small as a hundredth of a λ/D need to be achieved in order to reproduce laboratory performance at the telescope, meaning that these systems often do not perform at the expected level during on-sky observations.

Pupil-plane coronagraphs such as the APP sidestep this vibration-sensitivity problem, since they work by modifying the PSF of the instrument. This means that every point-source in the image is modified in the same way regardless of its exact position in the image, such that small shifts have minimal impact on the contrast in the dark zone as a whole. Such coronagraphs can also be designed to allow the high-contrast observation of multiple targets simultaneously, enabling planet searches in binary systems. It does however mean that the peak flux of the planetary signal is lowered, since this apodisation process comes at the cost of a lower *Strehl ratio*[10] for all point sources in the image. The vector Apodizing Phase Plate coronagraph (vAPP, Snik et al. 2012) is a powerful extension to the APP concept, and is discussed further in Sec. 1.5.1. vAPP optics have been demonstrated on-sky to achieve contrast levels of better than 10^{-5} over wide areas (Otten et al. 2017), and can be designed to produce contrasts of 10^{-10} in small dark zones of the order of a few square λ/D in size (Keller 2016). This optimisation parameter space is complex, with many trade-offs between contrast, IWA and image Strehl ratio (Por 2017): very different designs are therefore favoured for the detection versus the subsequent characterisation of exoplanets. The APP is straightforward to install and calibrate since it requires only a single pupil-plane optic, which has contributed to it becoming widespread in recent years. APPs are available for science observations with VLT-NACO (Quanz et al. 2010), MagAO (Otten et al. 2017) and the Large Binocular Telescope (LBT, Kenworthy et al. 2010a), with a number of new designs being produced for VLT-ERIS (Boehle et al. 2018) and Subaru-ScExAO (Lozi

[10] A common measure of image quality which is defined by the peak signal of the non-coronagraphic instrument PSF, where a perfect, diffraction-limited image has a Strehl ratio of unity.

et al. 2018), amongst others.

1.4.2 Adaptive optics

Adaptive optics (AO) is an essential component of any ground-based direct imaging instrument (Milli et al. 2016). This is because turbulent layers in the atmosphere heavily distort the phase of incoming light, which breaks up the stellar image into a widely distributed halo of speckles. This phase distortion changes on extremely short timescales on the order of milliseconds, which results in long-exposure images averaging out into a smooth, extended *seeing halo* which has a Strehl ratio of just a few percent and contains very little spatial information about the target. Such a seeing-limited image is shown in the top inset panel of Fig. 1.9.

AO systems aim to correct for this atmospheric wavefront distortion in real time, by using a *wavefront sensor* (WFS) to measure the instantaneous distorting phase aberration. The appropriate correction can then be applied via a *deformable mirror* (DM), which can change its shape in order to induce an equal and opposite phase shift into the light beam, as shown in the main diagram of Fig. 1.9. This AO correction not only increases the spatial resolution of the instrument but also dramatically enhances the signal-to-noise ratio of all point sources in the image, since the majority of the light is re-concentrated into a diffraction-limited core. Deformable mirrors can only provide an approximate correction however; the precision to which they can correct the distorting wavefront is limited by the density of actuators which control the mirror shape. This *fitting error* leads to a specific cut-off in terms of the spatial scales which it can correct, resulting in AO-corrected images displaying a typically circular *control region* inside which the PSF is nearly diffraction-limited, while regions outside see no correction whatsoever. Extreme adaptive optics (XAO) systems are now included in almost every high-contrast imaging instrument, defined by their ability to produce Strehl ratios on the order of 90% at near-infrared wavelengths thanks to their use of high-resolution deformable mirrors containing thousands of individual actuators. An example science-quality XAO-corrected image is shown in the lower inset panel of Fig. 1.9, showing the substantial difference between this and the original seeing-limited image.

Wavefront sensing is most commonly achieved by splitting off a fraction of the light from the main science beam of the instrument into a wavefront-sensing arm, which measures some form of spatial derivative of the incoming phase of the wavefront in a pupil plane of the instrument, such as the local tilt. This is mainly done because it allows the wavefront to be spatially resolved in a way that makes it more straightforward to control the individual actuators of the DM, and does not interfere with science observation. These so-called pupil-plane wavefront sensors have a major disadvantage however, in that they create regions of *non-common path* (NCP) between the wavefront sensor and the science camera; these are highlighted in red in Fig. 1.9. Any deformation to the wavefront which occurs in these regions, for example due to vibrations, internal turbulence or the slow flexing of optics due to temperature changes, will lead to an erroneous measurement of the correction needed to form a perfect image at the science focal plane. As shown in Fig. 1.10, these so-called *non-common path aberrations* (NCPAs), also known as *non-common path errors* (NCPEs), result in a field of *quasi-static speckles* (QSS) which typically evolve on timescales of seconds to hours (Fitzgerald & Graham 2006; Soummer et al. 2007a).

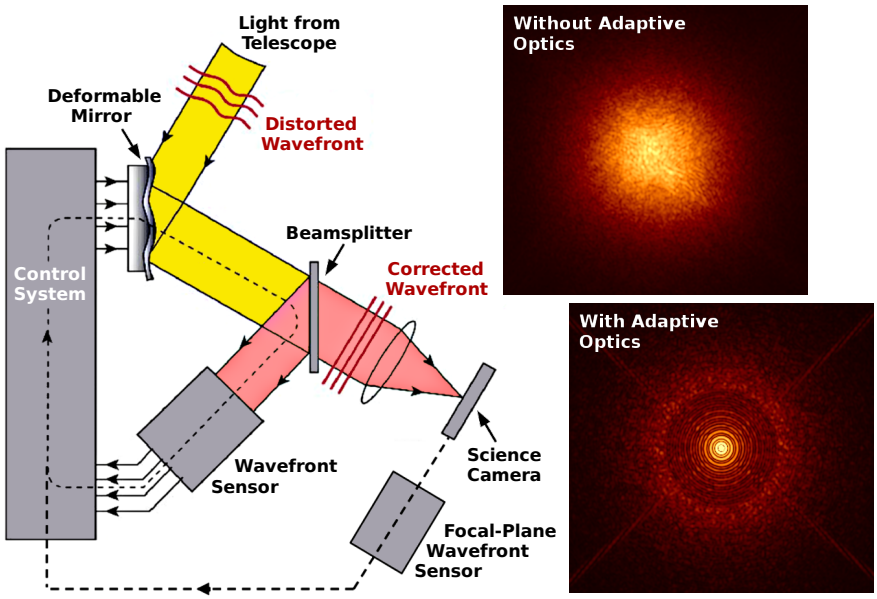

Figure 1.9: Schematic of an adaptive optics (AO) system, with image panels showing the stellar image before (top) and after (bottom) simulated AO correction. The Strehl ratio (SR) is an indicator of image quality, where a perfectly diffraction-limited image has $\mathbf{SR} = 1$. Non-common path (NCP) regions of the main pupil-plane wavefront sensor are shaded in red. A focal-plane wavefront sensor has been included, which operates on science-camera images to provide additional corrections for these NCP regions. Figure adapted from http://lyot.org.

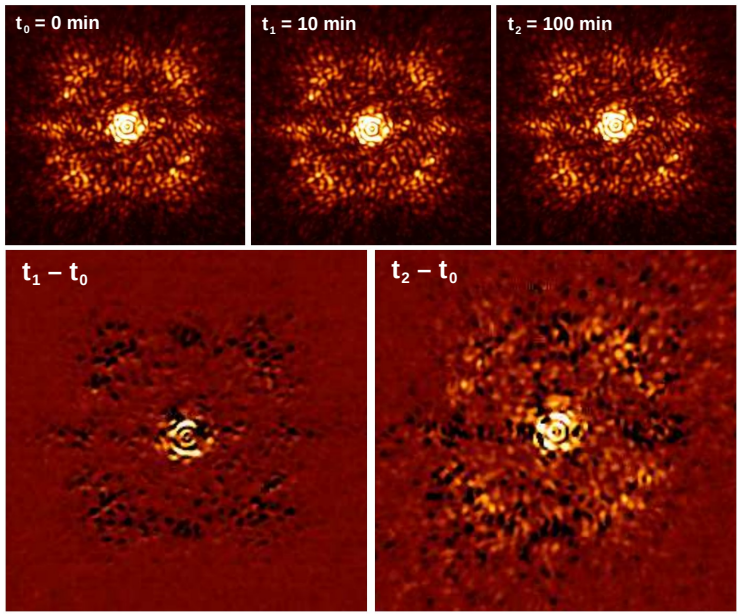

Figure 1.10: Example of a temporally evolving quasi-static speckle field in SPHERE-IRDIS (Dohlen et al. 2008) coronagraphic calibration images, due to non-common path aberrations in the instrument. Lower panels show the subtraction of very similar raw images (upper panels) taken at different times, revealing contrast-limiting residuals of the changing speckle field. Figure credit: Martinez et al. (2013).

The impact of these speckles becomes apparent when subtracting two images taken at different times during an observation window, with the raw image contrast currently limited to 10^{-3} to 10^{-4} at best (Fusco et al. 2016). A typical speckle then only has to change in brightness by 0.01% in order to produce a residual signal of the same amplitude as that expected from Proxima Centauri b.

The most straightforward data reduction technique of *reference differential imaging* (RDI) uses more or less the same approach as Fig. 1.10, by making separate observations of a planet-less star and subtracting this from the data in order to remove residual starlight. The fact that this approach is severely limited by quasi-static speckles due to the time lag between data and reference image acquisition means that NCPAs are currently the main contrast-limiting factor of high-contrast imagers. Maintaining a stable PSF is therefore in most cases more crucial to large, complex ground-based instruments than designing efficient coronagraphs capable of reaching contrasts of 10^{-10} under diffraction-limited conditions. The use of more aggressive coronagraph designs is in general only necessary once the residual QSS noise has been reduced below the photon noise limit of the static speckle background. That said, coronagraphs do help to reduce the phenomenon of *speckle pinning*, where local maxima in the diffraction-limited PSF tend to display worse speckle behaviour than regions where the starlight is suppressed (Aime & Soummer 2004).

The ideal solution to these NCPAs is to use a *focal-plane wavefront sensor*, which analyses the science image in order to determine the optimal correction without the complication of NCP regions. This is shown in Fig. 1.9 as a secondary wavefront sensing loop joined to the science camera, which provides an update to the correction of a primary pupil-plane wavefront sensor. Such a two-sensor setup is optimal because the many existing types of pupil-plane wavefront sensors have already been optimised with fast, high-resolution and high-amplitude atmospheric correction in mind, so can remove the vast majority of wavefront error from the system. Focal-plane correction schemes can then be designed with a much higher sensitivity but lower dynamic range suitable for controlling the residual QSS aberrations (which are typically less than a radian RMS[11] in total), and operate at speeds more in line with the typical coherence timescales of NCPAs and science imaging exposure times.

The development and implementation of focal-plane wavefront sensing techniques is however non-trivial; this task forms a major focus of this thesis and is therefore discussed in significantly more detail in Sec. 1.5 and Chapters 2, 3 & 4 of this thesis.

1.4.3 Image post-processing algorithms

Quasi-static speckles limit the raw contrast of coronagraphic, XAO-corrected images delivered by current HCI instruments to orders of magnitude above the planetary signatures we wish to detect. A wide variety of advanced data reduction techniques have therefore been developed in order to partially overcome this barrier, which are essential in order to reach the final 5σ contrast performance limits described in Sec. 1.3.2 and Fig. 1.6. These techniques operate by exploiting a difference in the behaviour of the planetary signal of interest compared to the halo of unwanted starlight in which it is embedded, which provides a source of diversity by which to distinguish the two components. This can then be used to construct a reference model of the stellar speckle

[11] A root-mean-square error of $1/2\pi$ wavelengths across the full telescope aperture.

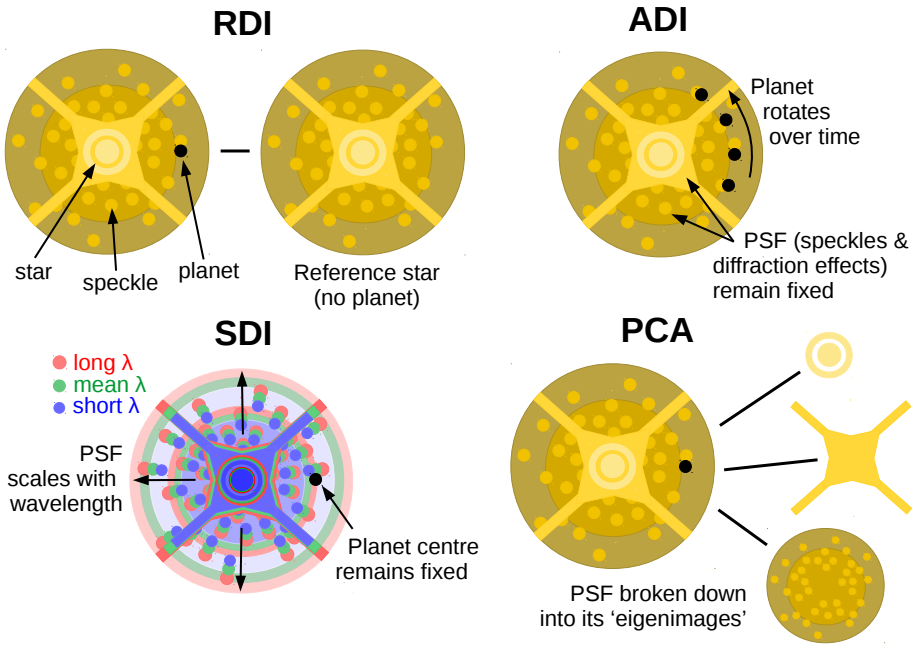

Figure 1.11: Diagrams of the various common post-processing techniques used for contrast enhancement, illustrating the form of diversity between the star and planet signal that each exploits.

field using the scientific dataset itself, thus minimising the time difference between science and reference data, over which the speckle field has a chance to evolve.

The most widely used of these techniques are illustrated in Fig. 1.11. As already mentioned, *reference differential imaging* (RDI) uses a star without any circumstellar material as the PSF reference, but is limited by the time delay (and additional NCPAs generated by changing the telescope pointing) between science and reference image acquisition. *Angular differential imaging* (ADI, Marois et al. 2006) takes advantage of the fact that altitude-azimuth mounted telescopes can be configured to allow the sky to rotate during the observation window, whilst keeping the pupil of the instrument stabilised. This means that any astronomical sources will slowly rotate around the centre of the image over time, while all instrumental speckles remain fixed in place. Many variants of ADI such as LOCI (Lafrenière et al. 2007) then use knowledge of the sky rotation rate to construct an optimal reference image in which the planet is not present, thereby minimising *self-subtraction* of the useful signal. *Principal component analysis* (PCA, Soummer et al. 2012) is an alternative method to LOCI for reducing ADI datasets, where in each frame the stellar image is decomposed into a limited set of basis functions from which a PSF reference image can be directly formed. A form of *Spectral differential imaging* (SDI, Crepp et al. 2011), on the other hand, is possible using low- and medium-spectral resolution integral field unit (IFU) imaging spectrographs, which simultaneously create a low-resolution image of the target at a number of neighbouring wavelengths. Since the diffraction limit scales with wavelength, all instrumental speckles shift radially between the bands while any true companion remains centred on the same location.

These advanced techniques are currently only capable of gaining approximately one to two orders of magnitude in contrast over a basic RDI reduction (e.g. Zurlo et al.

2016), which is primarily limited by the amount of diversity that can be introduced by these imaging modes within the typical coherence timescales of NCPAs. This is especially true at the smallest angular separations, where the amount of sky rotation and spectral dispersion is correspondingly smaller than in the outer regions of the image. ADI algorithms are also carefully optimised to detect point sources with minimal self-subtraction, but by doing so heavily remove or distort the signals of extended objects such as proto-planetary disks, and can also significantly modify the relative photometry of the final image (Milli et al. 2012).

The following subsections therefore present two major ways in which HCI instruments can be augmented, to make them sensitive to intrinsic differences between star and planet light. This is in contrast to the instrumentally-induced diversities which have been presented in this section: the following methods can be used to much more efficiently disentangle and remove unwanted stellar signal, while producing fewer data reduction artefacts.

1.4.4 Polarimetry

Light can be described as a superposition of linearly polarised components, where the electric field of the electromagnetic wave oscillates in only one direction. This fact can be exploited in order to image circumstellar material, since light is linearly polarised by scattering interactions with dust particles in the same way that the blue sky is polarised by Rayleigh scattering. By contrast, un-scattered starlight is un-polarised to an extremely high degree. At its core, the technique of *differential polarimetric imaging* (DPI, Kuhn et al. 2001) uses linear polarisation filters (analysers) to filter out all but a single polarisation state. By subtracting images of the target made with these polarisers in different configurations, an image of the total linearly polarised intensity can be built up which contains no un-scattered signal, providing an efficient means of starlight rejection.

Mathematically, the polarisation state of light can be described using the Stokes vector $\boldsymbol{S} = [I, Q, U, V]$, where I is total intensity, Q is the difference between linear polarisation along the horizontal (positive) and vertical (negative) axes, U is the equivalent linear polarisation vector rotated by 45° with respect to Q, and V is the circular polarisation. Circular polarisation is rarely seen in scattered light but is fundamental to the operation of liquid crystal-based optics, which are introduced in Sec. 1.5.1. Polarimeters are therefore designed to be sensitive to Q and U, from which the fractional degree of linear polarisation (P_I) and angle of linear polarisation (P_θ) of the astronomical signal can be calculated according to

$$P_I = \sqrt{Q^2 + U^2}/I, \tag{1.1}$$

$$P_\theta = \arctan(U/Q)/2. \tag{1.2}$$

The power of this technique when compared to intensity imaging is illustrated by Fig. 1.12, which shows a thin, high-inclination ring structure around the star HR 4796 A (Perrin et al. 2015). Unfortunately the efficiency of polarisation for light scattered by circumstellar disk material is typically observed to be of the order of only 5-10% (e.g. Krist et al. 2010; Mulders et al. 2013): DPI observations therefore throw away a significant fraction of photons from the target of interest, resulting in elevated photon noise. This is however more than outweighed by the gains in contrast that this technique can achieve over intensity imaging at small angular separations, and most

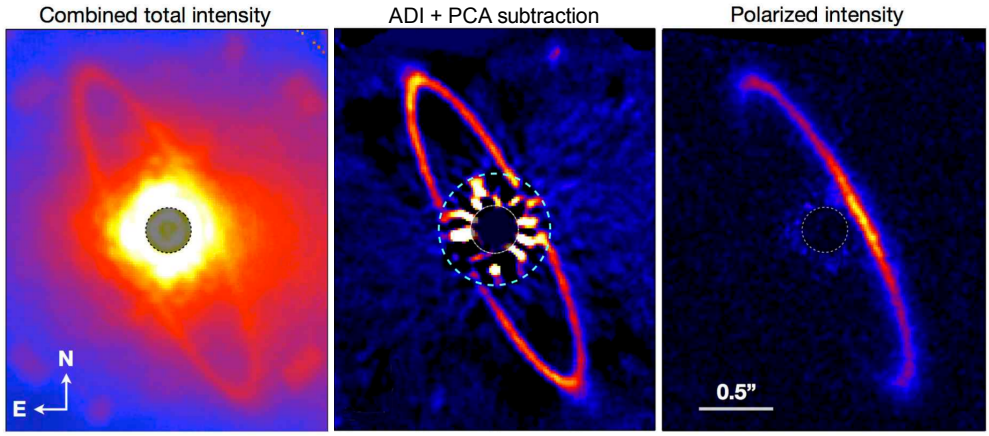

Figure 1.12: First-light images with the Gemini Planet Imager (GPI) of HR 4796 A, a massive A-type star hosting a narrow ring-like debris disk. Polarimetric imaging (right) displays significantly fewer residual stellar artefacts than intensity imaging, shown in the left panel with only minimal processing and in the middle panel following a combined ADI and PCA reduction. Only one side of the disk is strongly polarised due to the angular dependence of polarisation by scattering, revealing that the north-west side of the disk is closest to Earth. Figure credit: Perrin et al. (2015).

importantly results in significantly less distortion of extended objects than equivalent processing with ADI-like techniques.

Self-luminous planets are expected to have only a very small net polarisation signal at NIR infrared wavelengths (Stolker et al. 2017), although this rises significantly for mature planets observed in reflected light (Stam et al. 2004). The degree of linear polarisation of the HR 8799 planets has recently been constrained to less than 1% in the H-band, or a 1σ contrast of almost 10^{-7} (van Holstein et al. 2017), and while the first detection of a strongly polarised, spatially-resolved companion was recently made around CS Cha (Ginski et al. 2018), this is expected to be due to the existence of a circum-planetary disk rather than the companion itself. This technique has nonetheless had great success in imaging proto-planetary disks in scattered light; the SPHERE instrument is extremely productive for this purpose at both visible (e.g. Roelfsema et al. 2016) and near-infrared wavelengths (e.g. Benisty et al. 2015; de Boer et al. 2016; van Boekel et al. 2017). As discussed in Sec. 1.2.3, these scattered-light observations allow the determination of a number of key properties of the surface layers of these disks, including the potential identification of ongoing planet formation (Keppler et al. 2018). The interpretation of polarimetric images is often non-trivial however: reflective surfaces in the instrument often provide a strong source of artificial linear polarisation which hampers the exact determination of P_I and must first be calibrated out. Self-shadowing effects can often make it challenging to determine whether observed features in the disk are due to a dearth of scattering material or to geometric effects, while the strong angular dependence of polarisation by scattering (Pinte et al. 2009) induces a further modulation to the signal in highly inclined disks such as HR 4796 A, which does not reflect the underlying density of material.

1.4.5 Spectroscopy

As already introduced in Sec. 1.1.2, the combination of high-contrast imaging with high-resolution spectroscopy (HRS) is an extremely powerful characterisation approach, since planetary spectra can provide a wealth of information when sampled at a sufficiently high resolution. The significant diversity between planetary and stellar spectra at these resolutions also facilitates a variety of contrast-enhancing techniques.

The simplest method of using spectral information as a contrast-enhancing technique is *simultaneous differential imaging* (SDI), which originally referred to the approach of imaging the target in parallel through two narrowband filters of very slightly different wavelengths (Smith 1987; Biller et al. 2006). These filters are chosen such that one lies on top of a strong emission or absorption feature of the planet which is not present in the star, while the other lies just outside this feature. By subtracting the two images, only sources containing the chosen spectral feature remain and the stellar speckle field should be efficiently removed. This technique was employed in a number of exoplanet surveys (e.g. Biller et al. 2007; Nielsen et al. 2008) which focussed on searching for a 1.62 μm methane absorption feature, since it was predicted from observations of free-floating field brown dwarfs that exoplanet atmospheres should contain a significant methane component. Unfortunately, the dearth of new detections reported by these surveys implies that this prediction of exoplanet and brown-dwarf spectral similarity is not necessarily true.

1.4.5.1 Future contrast-enhancing applications of HRS

In addition to its value as a characterisation tool, the HRS spectral template cross-correlation method can itself be used as a contrast-enhancing technique, capable of gaining an estimated additional factor of 10^{-4} over raw coronagraphic images (Snellen et al. 2015). This is possible since both the differences in spectral features between the star and planet light and the overall Doppler shift from the orbital motion of the planet with respect to the star provide sources of diversity which can be exploited during data reduction. The process of coupling HCI and HRS instrumentation also holds the potential to provide additional starlight nulling. This coupling is typically achieved by feeding light from the HCI focal plane into an optical fibre, through which it can be passed to the spectrograph. While wide-diameter (multi-mode) optical fibres are often used in order to transmit as much light as possible, the newly developed SCAR coronagraph concept (Por & Haffert 2018; Haffert et al. 2018a) shows that single-mode fibres can be used to filter out non-point-source-like light to increase the contrast performance of a fibre-fed integral field unit (IFU).

In order to meet the performance requirements of the next generation of ELT-class exoplanet characterisation instruments, it is necessary to combine as many of the the contrast-enhancing techniques described in this section as is possible. In this way, a fully system-engineered instrument which optimally combines high-contrast imaging with high resolution spectroscopy, and ideally also with polarimetric and IFU capabilities, is the most promising approach to optimise the ultimate contrast performance of instruments such as ELT-EPICS (Kasper et al. 2010).

1.5 This thesis: Focal-plane wavefront sensing techniques

As discussed in Sec. 1.4.2, focal-plane wavefront sensing is the most direct method for eliminating the quasi-static speckles which currently limit high-contrast imaging instruments: it is however a fundamentally non-trivial task when compared to pupil-plane wavefront sensing. When combined with the need to avoid interfering with the science observations, focal-plane sensing with the main science camera constitutes an unattractive solution for most AO applications and is therefore comparatively underdeveloped at the present time.

The two main reasons why most wavefront sensors operate in the pupil plane are that it allows the wavefront to be spatially resolved, and that there are numerous ways to straightforwardly measure some derivative of the phase (e.g. local gradient) for each resolution element. This is illustrated in the top row of Fig. 1.13 for the example of the ubiquitous Shack-Hartmann pupil-plane wavefront sensor (SH-WFS). The SH-WFS uses a lenslet array to create a grid of subsidiary PSFs on the sensor camera, each of which samples only a small region of the pupil. The location of each PSF is linearly related to the local gradient (tip-tilt phase error) of the incoming wavefront in each resolution element, allowing for a direct reconstruction of the full phase map by assuming continuity of the wavefront.

Conversely, the science focal-plane image $p(x, y)$ is formed by the absolute value

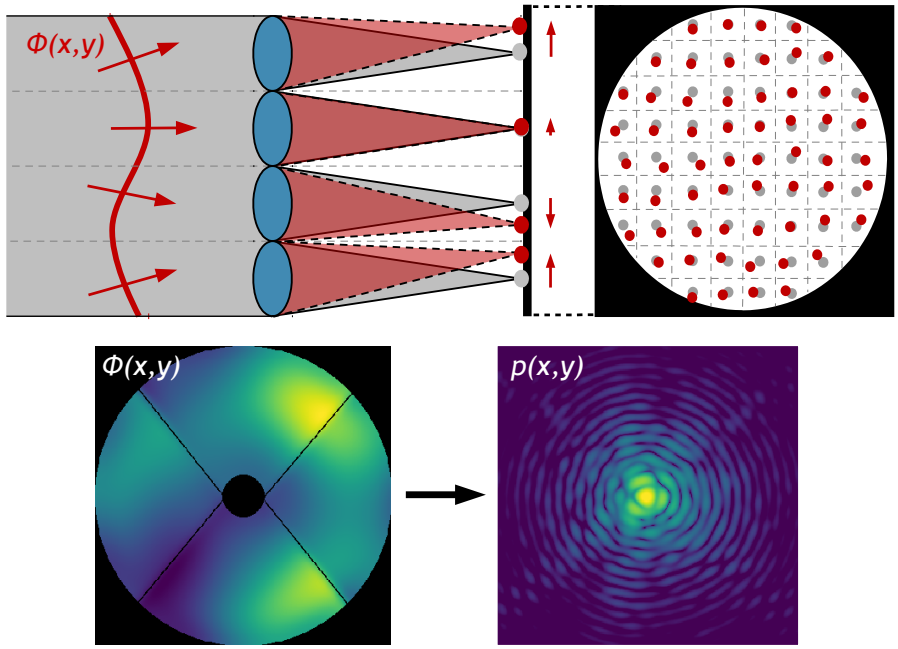

Figure 1.13: Illustration of the information content available to pupil-plane and focal-plane wavefront sensors. Top: Diagram of the Shack-Hartmann pupil-plane wavefront sensor, showing how the position of each subsidiary PSF is sensitive to the local gradient of the aberrating wavefront $\phi(x, y)$ (red spots, with respect to the grey nominal locations). Bottom: An example wavefront phase map $\phi(x, y)$ and the corresponding aberrated focal-plane PSF $p(x, y)$, which by comparison contains no easily retrievable phase information.

squared of the complex Fourier transform of the complex electric field in the pupil plane, expressed as

$$p(x,y) = \left|\mathcal{F}\left[A(x,y)e^{i\phi(x,y)}\right]\right|^2, \tag{1.3}$$

where $A(x,y)$ is the telescope aperture function and $\phi(x,y)$ is the wavefront phase to be sensed. This is illustrated in the bottom row of Fig. 1.13: not only does this mean that the wavefront is no longer spatially resolved, but since imaging cameras are only capable of making intensity measurements there is also a fundamental loss of information about the complex focal-plane electric field, $\mathcal{F}\left[A(x,y)e^{i\phi(x,y)}\right]$. The PSF therefore by default contains no easily retrievable phase information, and in general it is not theoretically possible to reconstruct the full wavefront phase from a single PSF intensity image.

A variety of focal-plane wavefront sensing techniques are now under development which can overcome these challenges, many of which are summarised by Jovanovic et al. (2018). Two specific approaches which are directly relevant to this thesis are briefly outlined in the following subsections.

1.5.1 Holographic wavefront sensing

1.5.1.1 Principles of holography

Since the science image does not contain easily retrievable phase information by default, holographic optics can effectively be used to engineer the PSF of the instrument in such a way as to make direct phase retrieval possible. This works by splitting off small amounts of light from the main science PSF to different locations in the focal plane, which can each be modified with independent wavefront aberrations whilst maintaining a dominant, un-affected central image for science use.

This approach is heavily based on the Holographic Modal Wavefront Sensor (HMWS), originally developed for confocal microscopy (Neil et al. 2000; Booth 2003). The principle behind this is illustrated in Fig. 1.14, where a phase-only holographic optic is used to create secondary copies of the science PSF at different locations in the focal plane. Since these secondary copies can be biased with independent wavefront aberrations, it is possible to encode phase information about a finite set of important aberration modes into the focal plane, in a manner which is easily measurable. The coronagraphic Modal Wavefront Sensor (cMWS) uses this principle in conjunction with the APP coronagraph to provide simultaneous high-contrast imaging and focal-plane wavefront sensing capabilities with the same detector: the development and testing of this technique is the subject of Chapter 2.

These *computer-generated holograms* may be designed via three main encoding methods, which form a direct analogy with the various common types of *diffraction grating* (Brown & Lohmann 1969):

$$H_{cosine}(x,y) = \sum_k a_k \cos\left(O_k(x,y) + R_k(x,y)\right), \tag{1.4}$$

$$H_{binary}(x,y) = \arg\left(H_{cosine}(x,y)\right), \text{and} \tag{1.5}$$

$$H_{blazed}(x,y) = \arg\left(\sum_k a_k e^{i(O_k(x,y)+R_k(x,y))}\right). \tag{1.6}$$

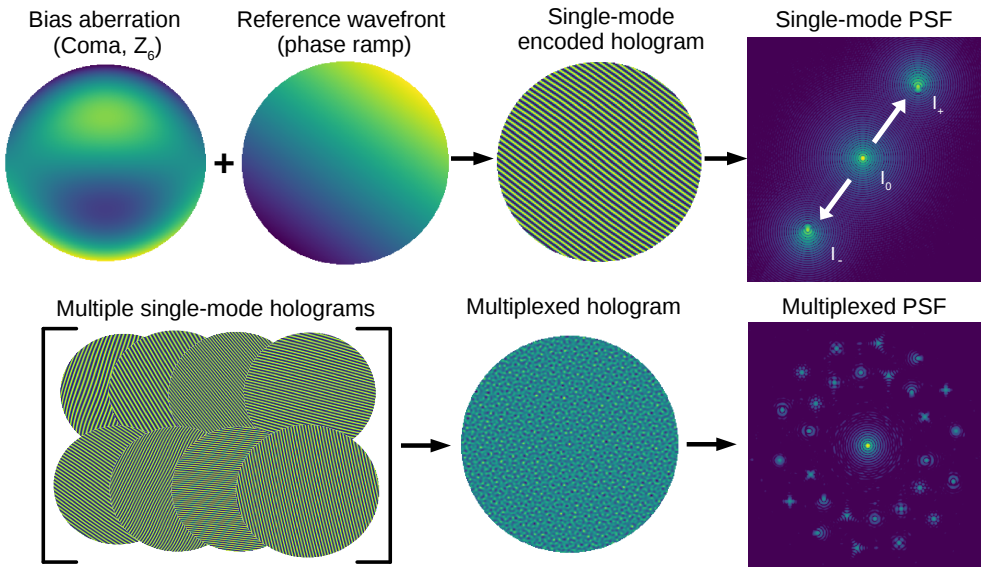

Figure 1.14: Illustration of the design process of a Holographic Modal Wavefront Sensor (HMWS). Top row: A single phase aberration (in this case coma) forms the bias wavefront (O_k), which corresponds to the wavefront mode to be sensed, while the amplitude and direction of the tilted reference wavefront (R_k) defines the location of the biased PSF copies in the focal plane. This is encoded into a phase-only hologram according to one of Eqs. 1.4 - 1.6 (in this case Eq. 1.4), which functions as a form of spatially modulated diffraction grating. The resulting instrument PSF then contains secondary PSF copies in the ±1 diffraction orders of the grating, biased with opposite signs of the wavefront aberration. Bottom row: The summation of multiple single-mode holograms (H_k) allows for many pairs of PSF copies to be created simultaneously, with independent positions, wavefront biases, and relative brightnesses with respect to the central science image.

Here $O_k(x, y)$ is the object (bias) wavefront and $R_k(x, y)$ is the reference (tilted) wavefront as illustrated in Fig. 1.14: this terminology is conventional for traditional forms of optically-exposed hologram (Dong et al. 2012). arg() denotes the complex argument function: multiplexed holograms can be created by summing the individual single-mode components before taking the argument, with differing amplitude a_k controlling the relative brightness of each resulting copy.

These different holographic encodings have various strengths and weaknesses in terms of performance: the cosine hologram creates significantly less speckle noise in the science image plane than either binary or blazed holograms, since it does not suffer from phase information loss (due to taking the complex argument in the latter two cases), however it is not capable of 100% diffraction efficiency and is limited in the number of modes that can be multiplexed (Changhai et al. 2011). This trade-off is explicitly discussed for the case of binary and cosine holograms in Wilby et al. (2016a). The more recent development of optimised cMWS optics based on multiplexed blazed holograms is published in Haffert et al. (2018b), as part of further on-sky cMWS demonstration work with the *Leiden EXoplanet Instrument* (LEXI). These blazed grating holograms provide an excellent trade-off between the binary and cosine approaches, outperforming both in the case of highly multiplexed holograms.

1.5.1.2 The vector-APP coronagraph

The vector-APP (vAPP) coronagraph was developed in Snik et al. (2012) and Otten et al. (2014a,b) as a work-around for some of the limitations of the classical APP coronagraph. The most notable of these is that only one side of the PSF could be made dark at any one time, meaning that for survey work two sets of observations would need to be made for each target, rotating the optic in between. The vAPP is created using a modified polarisation grating to create two PSFs simultaneously, each containing an APP dark zone on the opposite side of the star, as illustrated by Fig. 1.15. This is fundamentally the same principle as the blazed HMWS, simply using a single wavefront aberration which is customised to function as an APP coronagraph rather than as a wavefront sensor: this makes it perfect for incorporation into the cMWS.

Since this work it has now become possible to design annular (360 degree) dark-hole-generating APPs with useful throughput. However, the 180 degree vAPP concept remains useful: it can achieve inner-working angles as small as 1.6 λ/D whilst retaining a useful level of throughput and contrast performance, whereas the design parameter space of 360 degree APPs is significantly more restricted (Por 2017).

1.5.1.3 Liquid crystal manufacturing technologies

The development of the vAPP coronagraph is crucial to the work in this thesis, since many advances in its manufacture have been made in recent years. APP optics were originally produced as diamond-turned glass plates, with the thickness at each location in the pupil inducing the desired phase delay into the wavefront (Codona et al. 2006; Kenworthy et al. 2007): unfortunately this manufacturing technique has relatively limited spatial resolution and is not capable of producing sharp jumps in phase. This rules out a large fraction of the design parameter space of both the vAPP and the cMWS, which typically produce extreme phase patterns which are dominated by discontinuities and high-spatial frequency phase content.

Direct-write liquid crystal technologies (Miskiewicz & Escuti 2014), on the other hand, can be used to create phase optics with spatial resolutions of the order of 1 μm and with artefact-free phase wrapping, allowing almost any phase design to be accurately manufactured. This is achieved by using a scanning UV laser to expose a photo-alignment layer, which in turn controls the orientation of subsequently-deposited liquid crystal layers at each point on the substrate. The resulting optic is effectively a *spatially-modulated half-wave retarder*, which acts on incident circular polarisation, flipping its handedness while inducing a phase shift which is proportional to the angle between the crystal fast axis and the incident electric field. The resulting phase pattern is also achromatic since the phase modulation effect of liquid crystals is produced via so-called *geometric phase*, which is independent of wavelength (Komanduri et al. 2013). For a significantly more detailed description of this manufacturing technique and its implications, see the PhD thesis of Gilles Otten (Otten 2016).

1.5.1.4 Future applications for holographic optics

The success of liquid crystal phase optic manufacturing techniques has led to a recent explosion in the number of vAPP coronagraphs being implemented in high-contrast imaging instruments around the world: many of these are listed in Sec 1.4.1. This also applies to the cMWS, with optics now being implemented in LEXI at the William

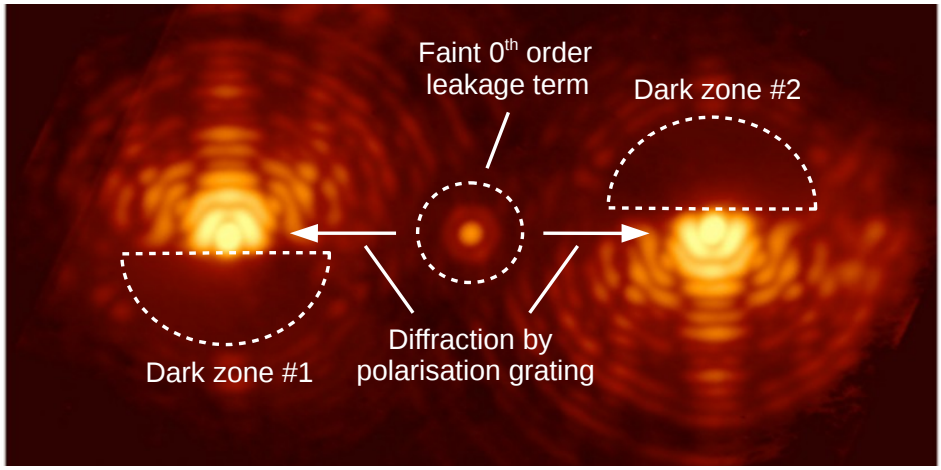

Figure 1.15: On-sky image of a single point source using the vector-APP coronagraph installed in the MagAO instrument at the Magellan Clay Telescope, Chile (Morzinski et al. 2014). Each of the two science PSF copies consists of light of a single circular polarisation, artificially split off from the stellar position by means of a diffraction grating. A small fraction of the stellar light passes through the optic unaltered, forming the zeroth-order leakage term at the original stellar location. Image credit: Gilles Otten.

Herschel Telescope (Haffert et al. 2018a), the HiCIBaS high-altitude balloon mission (Côté et al. 2018), the MagAO-X upgrade for the Magellan Telescope (Miller et al. 2018) and SCExAO at the Subaru Telescope (Doelman et al. 2017).

The development of holographic optics for the cMWS has also already led to a number of spin-off applications. There exists a wide variety of other focal-plane wavefront sensing strategies which, like the original (non-holographic) concept of modal wavefront sensing, are designed to be performed in multiple sequential steps which each degrade the science image. The main potential of the multiplexed holographic approach is that it can transform almost any such temporally-modulated technique into a simultaneous, spatially-modulated equivalent, thereby enabling it to operate in real time and in parallel with science observations. A notable example of this is the electric field conjugation technique (EFC, Give'on et al. 2007) which, like the cMWS, uses phase probes to reconstruct the wavefront. Instead of using a set of low-order modes however, EFC simultaneously adds many high-frequency, low-amplitude aberrations into the dark zone of an APP coronagraph. By making four or more holographic copies of the coronagraphic PSF (Por & Keller 2016) to which these EFC phase probes have been added, the full aberrated electric field and hence the necessary wavefront correction can be retrieved without degrading the science image. An additional high-contrast imaging concept which has been enabled by the development of multiplexed blazed grating holograms is the holographic aperture masking (HAM) technique (Doelman et al. 2018). This is an extension to sparse aperture masking (SAM, Tuthill et al. 2010), which effectively divides up the telescope aperture into multiple independent SAMs which each produce interference fringes on different parts of the image, thereby maximising useful throughput of the system. SAM exploits the fact that interferometric techniques can beat the conventional diffraction limit and so gain significantly in effective inner-working angle performance, while HAM can also be engineered to provide simultaneous coverage of multiple position angles around the star.

In summary, the power of holographic optics to grant unprecedented control over

the instrument PSF structure and hence the information content delivered in science images is now opening up a large window of opportunity for new high-contrast imaging techniques, both for wavefront control and coronagraphy applications.

1.5.2 Phase diversity wavefront sensing

An alternative class of focal-plane wavefront sensing techniques focus on overcoming the over-arching PSF phase degeneracy issue via the use of *phase diversity*. The core principle behind this approach is that, while full phase reconstruction is in general not possible using a single science image, it is still possible to probe the underlying complex focal-plane electric field by deliberately modifying the total phase aberration content of the system in a controlled manner. When using two images which differ only by a known, deliberately injected phase aberration, it becomes possible to reconstruct the full electric field and hence any unknown component of the aberrating wavefront. Fig. 1.16 illustrates how this may be achieved in practice by defocussing the science camera; a commonly-used source of phase diversity since it is typically reproducible and easy to implement.

There are many possible reconstruction algorithms which make use of such phase diversity information (Carrano et al. 1998; Gonsalves 2001; Lamb et al. 2016), some of which are also capable of working with coronagraphic images (Paul et al. 2014b). However, if this compensation scheme is to be operated in real time (as is required for accurate NCPA correction) it is necessary to interleave defocussed images into the acquisition process, thereby degrading the imaging performance of the instrument for 50% of science frames. In most cases this means that the frequency of wave-

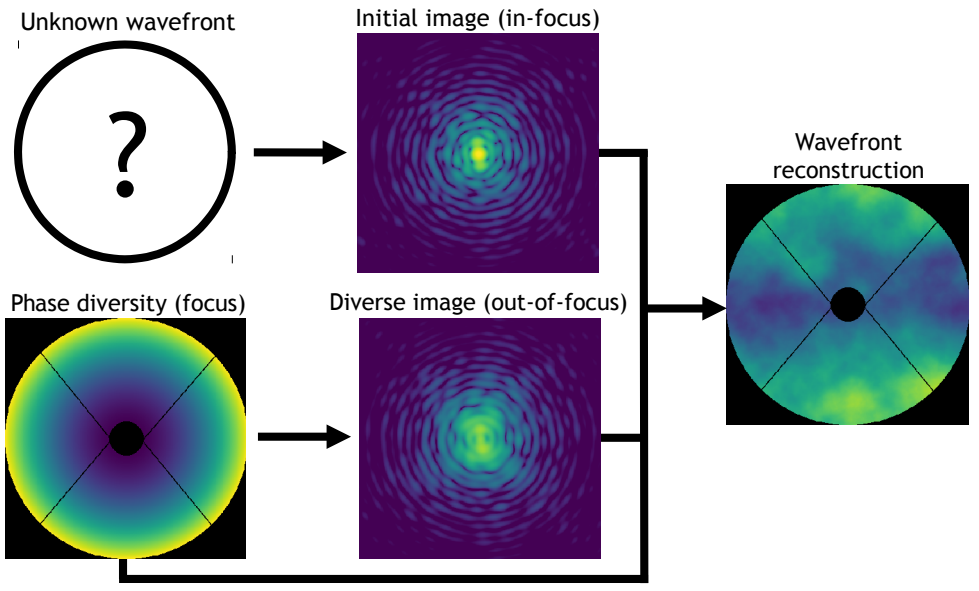

Figure 1.16: Illustration of the "classical" phase diversity wavefront reconstruction technique. A controlled phase aberration, in most cases strong defocus, is added to the system in order to modify the complex electric field in the focal plane in a manner which breaks the phase degeneracy of the PSF. By using the in- and out-of-focus intensity images and the known phase aberration which separates them, it is possible to reconstruct the full electric field in the focal plane and hence the unknown aberrating wavefront.

front correction updates will instead be significantly lowered, in order to maximise the yield of science-quality data: quasi-static wavefront aberrations will then not be fully addressed, allowing contrast-limiting speckle residuals to creep back into the final dataset.

One way around this limitation is to utilise the phase diversity information which is naturally present in any system with active wavefront control: after its first iteration, a focal-plane correction algorithm can in principle use its own previous correction update along with the current and previous image as the necessary input for phase diversity reconstruction, by direct analogy with the original approach in Fig. 1.16. By avoiding the need for destructive phase diversity probes this ensures that the image quality always improves from one iteration to the next, making it possible to operate at appropriately fast correction cadences (typically 10 Hz for NCPA control) without impacting science observations. An example of such a real-time phase diversity routine is the so-called "Fast & Furious" (F&F) sequential phase diversity algorithm, originally developed by Keller et al. (2012) and Korkiakoski et al. (2014). F&F is proposed in Chapters 3 & 4 as a software-only phase reconstruction solution for wavefront control issues seen in the SPHERE current-generation high-contrast imager at the VLT, with wider applications to other current and future instruments.

1.6 Thesis outline

The goal of this thesis is twofold: to address some of the outstanding issues limiting the performance of existing high-contrast imagers, and to develop new technologies which will ultimately facilitate the characterisation of Earth-like exoplanets with ELT-class telescopes. To this end, Chapters 2, 3 and 4 of this work focus on the wavefront control aspects of high-contrast imaging, since this is where the biggest performance gains currently stand to be made. Chapter 5 also illustrates how detailed instrumental characterisation work can be used to optimise the data reduction procedure, and ultimately improve science yields at the smallest angular separations.

Chapter 2: Introducing the coronagraphic Modal Wavefront Sensor

As outlined in Sec. 1.5.1, holographic focal-plane wavefront sensing techniques are a promising approach for the real-time correction of non-common path aberrations (NCPAs), which limit the contrast performance of current high-contrast imaging instruments. This chapter presents the theory, laboratory implementation and first on-sky validation of the *coronagraphic Modal Wavefront Sensor* (cMWS), which is capable of simultaneous wavefront retrieval and coronagraphy using the science focal plane. We show that this sensor is capable of the real-time on-sky recovery of known aberrations for low-order wavefront modes at a cadence of 50 Hz, for a large observing bandwidth of 50 % in the R-band (600 nm). This is important since many focal-plane wavefront retrieval algorithms rely on a quasi-monochromatic image, making the cMWS a versatile focal-plane sensor which is capable of controlling those aberrations which contribute most to the wavefront error budget of NCPAs. The cMWS has since been further validated as part of the *Leiden EXoplanet Instrument* (LEXI, Haffert et al. 2016): more recent on-sky results including on-sky closed-loop operation can be found in Wilby et al. (2016a) and Haffert et al. (2018a).

Chapters 3 & 4: Tackling the low-wind effect at VLT-SPHERE

NCPAs are not always the limiting factor in current high-contrast instruments. In these two chapters a potential real-time wavefront control solution is developed and tested for the so-called *low-wind effect* (LWE, Sauvage et al. 2016), which is seen to consistently degrade the contrast performance of the SPHERE current-generation high-contrast imager under otherwise optimal seeing conditions. Chapter 3 presents closed-loop simulation results for the *"Fast & Furious"* (F&F, Korkiakoski et al. 2014) phase diversity algorithm under SPHERE-like observing conditions, in order to assess the suitability of the control scheme as an immediately implementable, software-only solution for the LWE. Chapter 4 then reports on the subsequent laboratory testing of F&F at the MITHIC high-contrast testbench (Vigan et al. 2016b), located at the Laboratoire d'Astrophysique de Marseille, in order to verify the results of Chapter 3 and compare with other focal-plane wavefront control methods. F&F was found in both cases to be capable of robust real-time wavefront control under LWE-like conditions without degrading the science image, making it a promising algorithm for implementation both on SPHERE and other instruments suffering from similar wavefront control issues.

Chapter 5: Calibrating the SPHERE-IRDIS APLC coronagraph

Optimal reduction techniques are equally crucial when it comes to making the most of the data produced by current high-contrast imaging facilities. This in turn requires the thorough calibration of all instrumentation behind the observations, the results of which are not always easily available to observers. This chapter presents a characterisation effort of the apodised Lyot coronagraph system of the SPHERE-IRDIS near-infrared imager (Dohlen et al. 2008), in order to develop a suitable calibration algorithm for coronagraphic, polarimetric imaging data. This is important since the innermost regions of circumstellar disk observations, which are typically the most valuable to study, are often dominated by artefacts of the imaging system. An improved calibration routine which can correctly restore the relative photometry of these innermost regions is therefore capable of significantly improving the effective inner-working angle of the instrument. This approach is particularly crucial for the identification of central cavities in transitional proto-planetary disks; one of the main science yields of the SPHERE instrument. Calibration observations were made using the minor planet Ceres in order to accurately determine the extinction profile of the coronagraphic system. This was combined with extensive optical modelling in order to fully understand the observed signal, before re-reducing polarimetric observations of TW Hydrae (van Boekel et al. 2017) in order to validate the accuracy of the resulting calibration routine.