Exoplanet Imaging Speckle Subtraction: Current Limitations and a Path Forward

by

Benjamin Lionel Gerard
B.A., University of Colorado, USA, 2014
M.Sc., University of Victoria, 2016

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Physics and Astronomy

Exoplanet Imaging Speckle Subtraction: Current Limitations and a Path Forward

by

Benjamin Lionel Gerard
B.A., University of Colorado, USA, 2014
M.Sc., University of Victoria, 2016

Supervisory Committee

Dr. Christian Marois, Co-Supervisor
Department of Physics and Astronomy

Dr. Jon Willis, Co-Supervisor, Departmental Member
Department of Physics and Astronomy

Dr. Colin Bradley, Outside Member
Department of Mechanical Engineering

Abstract

The direct detection and detailed characterization of exoplanets using extreme adaptive optics (ExAO) is a key science case of both current and future telescopes. However, both quasi-static and residual atmospheric wavefront errors currently limit the sensitivity of this endeavour, generating "speckles" in a coronagraphic image that initially obscure any faint exoplanet(s) from detection.

I first demonstrate the current limits of exoplanet imaging using datasets taken with the Gemini Planet Imager and Subaru Coronagraphic ExAO systems. Even when using advanced post-processing algorithms, speckle evolution over time and wavelength is shown to limit the final contrasts that can be reached with current state-of-the-art instruments. A new approach is thus needed to detect fainter exoplanets below these limits.

I then illustrate a path forward to reach contrasts near the fundamental photon noise limit: fast focal plane wavefront sensing of both quasi-static and atmospheric speckles. My new method, called the Fast Atmospheric Self-coherent camera Technique (FAST), deploys new hardware and software to overcome these limitations. Looking toward the future, the contrast improvements from fast focal plane wavefront sensing techniques such as FAST are expected to play an essential role in the ground-based detection and characterization of lower mass exoplanets.

Contents

List of Tables

List of Figures

Glossary

ADC atmospheric dispersion corrector. 70

ADI Angular Differential Imaging (Marois, 2004; Marois et al., 2006a). 39, 40, 43, 44, 50, 51, 57, 58, 59, 71, 76, 165, 166, 167, 174

AO adaptive optics. 22, 23, 24, 25, 26, 27, 29, 31, 32, 34, 37, 40, 42, 43, 45, 46, 47, 58, 68, 69, 83, 85, 86, 91, 96, 101, 103, 106, 111, 115, 130, 132, 135, 137, 138, 158, 163, 165, 167, 168, 170, 172, 175, 200, 201, 202, 203, 204, 205, 206, 224, 228, 229, 230

APLC apodized Lyot coronagraph. 30, 31, 33, 86, 87, 117, 178, 180

ATS acquisition and tracking sensor. 228, 229, 230

BD brown dwarf. 2, 9, 10, 11, 16, 17

CA Core Accretion (Marley et al., 2007). 7, 8, 15

CDI coherent differential imaging. 44, 45, 46, 47, 48, 49, 76, 81, 87, 95, 101, 102, 103, 111, 115, 135, 143, 144, 145, 160, 165, 168, 169, 170, 171, 172, 173, 174, 179, 180, 181, 182

CHARIS Coronagraphic High Angular Resolution Imaging Spectrograph (Groff et al., 2016, 2017). 51, 57, 58, 60, 65, 69, 70, 71, 72, 74, 177

CREATE Collaborative Research and Training Experience. 135, 154, 200

DH dark hole. 45, 46, 87, 89, 90, 115, 116, 118, 120, 121, 123, 132, 133, 134, 142, 164, 180

DM deformable mirror. 23, 25, 26, 27, 28, 33, 34, 44, 45, 46, 48, 57, 78, 80, 81, 84, 85, 86, 87, 89, 90, 91, 92, 93, 115, 116, 117, 118, 119, 120, 122, 123, 124, 127, 132, 133, 134, 135, 137, 140, 141, 142, 143, 145, 146, 148, 149, 150, 160, 164, 170, 172, 173, 174, 178, 180, 181, 203, 204, 205, 230

DRP data reduction pipeline. 35, 36, 53, 54, 58, 70, 166

ELT extremely large telescope. 18, 38, 76, 161, 162, 163, 182

ETH Züruch Swiss Federal Institute of Technology Zurich. 157, 182

ExAO extreme adaptive optics. 27, 32, 33, 34, 35, 37, 45, 47, 48, 54, 73, 86, 165, 166, 177, 230

FAST Fast Atmospheric SCC Technique. 77, 87, 95, 102, 103, 111, 115, 117, 118, 119, 120, 122, 123, 124, 127, 132, 134, 135, 136, 143, 145, 150, 151, 152, 155, 158, 161, 162, 163, 164, 165, 167, 168, 169, 170, 171, 173, 174, 175, 178, 179, 180, 181, 182, 183

FFT fast Fourier transform. 119

FM forward model. 40, 41, 59, 60, 61

FOV field of view. 20, 27, 35, 36, 39, 57, 71, 99, 171

FPM focal plane mask. 29, 30, 31, 32, 52, 53, 57, 70, 77, 78, 84, 86, 87, 88, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 104, 105, 106, 111, 116, 117, 118, 124, 126, 127, 128, 131, 135, 136, 137, 140, 142, 145, 151, 152, 153, 154, 155, 156, 158, 159, 161, 168, 169, 170, 172, 173, 178, 179, 180, 182

FWHM full width at half maximum. 20, 96, 128, 129, 151

GEO geocentric orbit. 200, 206, 229

GI Gravitational Instability (Boss, 2000). 7, 8, 15

GPI Gemini Planet Imager (Macintosh et al., 2014). 11, 16, 17, 30, 35, 36, 38, 44, 51, 52, 53, 54, 56, 70, 85, 86, 87, 88, 117, 159, 166, 167, 174, 176, 178, 180, 183

HZ habitable zone. 17, 18, 38

IFS integral field spectrograph). 15, 24, 25, 28, 34, 35, 36, 39, 42, 50, 51, 52, 54, 58, 85, 166, 167, 168, 170, 171

IWA inner working angle. 29, 31, 57, 70, 78, 91, 97, 116, 117, 118, 137, 140, 151, 154, 156, 180

LEO low earth orbit. 205, 207, 224, 228

LESIA Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique—Laboratory for Space Studies and Instrumentation in Astrophysics. 154, 155, 182

LLOWFS Lyot-based low order wavefront sensor. 136, 137, 138, 139, 140, 141, 142, 143, 172, 173, 181

LOWFS low order wavefront sensor. 32, 38, 48, 86, 88, 136, 137, 138, 140, 178, 181

MTF modulation transfer function. 81, 85, 97, 99, 100, 105, 122, 126, 131, 136, 137, 152, 157, 158, 160, 169, 170

NCPA non common path aberration. 24, 167

NEW EARTH NRC's Extreme Wavefront lab for Exoplanet Advanced Research Topics at Herzberg. 152, 153, 158, 159, 182

NIR near infrared. 9, 10, 13, 18, 21, 22, 24, 25, 31, 35, 47, 88, 182, 230

NRC-HAA National Research Council of Canada, Herzberg Astronomy and Astrophysics. 158

NSERC Natural Sciences and Engineering Research Council of Canada. 135, 154, 200

OGR optical ground receiver. 201, 203, 205, 206, 209, 210, 211, 212, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229

OTF optical transfer function. 81, 82, 83, 126, 156

PAA point ahead angle. 201, 203, 204, 205, 224, 228, 229

PSD power spectral density. 21, 85, 86, 92, 118, 119, 120, 121, 122, 203, 214, 215, 225

PSF point spread function. 20, 21, 22, 26, 29, 30, 31, 33, 35, 36, 37, 39, 40, 41, 44, 50, 51, 52, 56, 58, 59, 61, 69, 72, 73, 77, 78, 80, 81, 83, 84, 86, 89, 90, 92, 95, 96, 97, 98, 99, 100, 101, 102, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 118, 122, 124, 125, 127, 129, 130, 143, 144, 145, 156, 157, 159, 166, 168, 170, 177, 179, 180, 181

PV peak-to-valley. 137, 138, 140, 181

RDI Reference Differential Imaging. 39, 40, 43, 44, 165

RMS root mean square. 21, 22, 34, 37, 38, 45, 52, 86, 87, 106, 110, 117, 126, 130, 132, 133, 134, 137, 140, 142, 145, 148, 150, 151, 153, 205, 209

RTC real time controller. 23, 25, 26, 28, 167, 230

RV radial velocity. 3, 4, 17, 182

S/N signal-to-noise ratio. 16, 52, 60, 61, 63, 81, 84, 90, 91, 95, 99, 100, 104, 106, 107, 121, 123, 124, 125, 126, 127, 130, 131, 132, 134, 135, 150, 157, 169, 178, 180

SCC Self-Coherent Camera. 46, 76, 77, 78, 80, 82, 84, 87, 88, 89, 90, 91, 94, 95, 97, 98, 99, 101, 102, 103, 104, 108, 110, 113, 115, 116, 117, 118, 120, 125, 126, 127, 128, 129, 131, 132, 133, 135, 136, 137, 138, 139, 140, 141, 142, 143, 145, 150, 152, 155, 156, 157, 160, 168, 169, 170, 171, 172, 173, 178, 179, 180, 181, 182

SCExAO Subaru Coronagraphic Extreme Adaptive Optics (Jovanovic et al., 2015b). 51, 54, 69, 74, 166, 167, 174, 177, 183

SDI Spectral Differential Imaging (Racine et al., 1999; Marois et al., 2000; Sparks & Ford, 2002). 39, 40, 43, 44, 50, 51, 54, 58, 59, 65, 66, 69, 71, 72, 76, 165, 166, 172, 174, 177

SED spectral energy distribution. 13, 16

SHWFS Shack Hartmann wavefront sensor. 25, 27, 34, 52, 88, 136

SLM spatial light modulator. 157

SNPS super-Nyquist power law phase screen. 92, 93, 94

SPHERE Spectro-Polarimetric High-contrast Exoplanet REsearch instrument (Beuzit et al., 2019). 56

SR Strehl ratio. 57, 69, 71, 86, 166, 204, 206

SVD singular value decomposition. 60, 65, 116

T/T tip/tilt. 216, 217, 220, 221, 222, 225, 228, 229

TEPS Technologies for Exo-Planetary Science. 135, 154, 200

TG Tip/tilt+Gaussian. 86, 94, 95, 96, 97, 99, 100, 101, 104, 105, 106, 111, 117, 118, 124, 126, 128, 129, 130, 132, 133, 134, 143, 145, 151, 152, 153, 154, 155, 156, 159, 168, 169, 170, 178, 179, 180

TGV Tip/tilt+Gaussian+Vortex. 79, 86, 124, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 138, 140, 142, 145, 146, 147, 148, 149, 150, 156, 157, 158, 159, 160, 163, 180, 181, 182

THD2 high contrast imaging testbed (Baudoz et al., 2018). 154, 155, 156

TLOCI template locally optimized combination of images (Marois et al., 2014). 51, 59, 60

TTV transit timing variation. 4

WFE wavefront error. 24, 28, 31, 32, 33, 37, 45, 85, 92, 132, 134, 138, 140, 144, 145, 148, 150, 153, 154, 166, 202, 203, 204, 205, 208, 209, 214, 215, 216, 217, 225, 229

WFS wavefront sensor. 23, 24, 25, 26, 27, 28, 31, 32, 47, 48, 84, 115, 119, 120, 124, 130, 131, 132, 135, 138, 141, 145, 146, 147, 148, 149, 150, 166, 167, 172, 173, 178, 181, 204, 224, 228, 229, 230

ZWFS Zernike wavefront sensor. 145, 146, 147, 148, 149, 150, 173, 181

Acknowledgements

First and foremost, for your love, unconditional support, and camaraderie, I'd like to thank my partners, Dr. Charli Sakari and Sir Tuckington Fitzwilliam Sakari 🐱; my family, Dr. Stephen K. Gerard, Dr. Trudy Lionel, and Dr. Juliana Rose Gerard; and my friends, Dr. Jared Keown and (soon to be) Dr. Collin Kielty. You all supported me through this stage of my professional development, and I will be forever grateful to have received this support from you. Thank you!

Next, I'd like to thank my collaborators, both in Victoria, Meudon, and all over the world: Dr. Christian Marois, Dr. Jean-Pierre Véran, Dr. Olivier Lardière, Dr. Raphaël Galicher, the GPIES collaboration, the SCExAO collaboration, the NEW EARTH research, the THD research group, the KAUST computational imaging group, Honeywell/Comdev, and all the others that have supported me throughout my PhD studies. Your hospitality, patience, and innovation made me into the researcher I am today, and supported an inclusive environment for collaboration. I will also be forever grateful for the support I have received from you. Thank you!

I gratefully acknowledge research support of the Natural Sciences and Engineering Council of Canada through the Postgraduate Scholarships-Doctoral award, Discovery Grant, and Technologies for Exo-Planetary Science Collaborative Research and Training Experience programs. I also acknowledge support from the GPIES collaboration throughout my PhD studies. The GPI project has been supported by Gemini Observatory, which is operated by AURA, Inc., under a cooperative agreement with the NSF on behalf of the Gemini partnership: the NSF (USA), the National Research Council (Canada), CONICYT (Chile), the Australian Research Council (Australia), MCTI (Brazil), MINCYT (Argentina), and KASI (S. Korea).

The data in §2.2 was obtained through 2017B open use Canadian and US Gemini exchange time (Gerard, 2017a). The development of SCExAO was supported by the JSPS (Grant-in-Aid for Research #23340051, #26220704 #23103002), the Astrobiology Center of the National Institutes of Natural Sciences, Japan, the Mt Cuba Foundation and the directors contingency fund at Subaru Telescope. CHARIS was built at Princeton University under a Grant-in-Aid for Scientific Research on Innovative Areas from MEXT of the Japanese government (# 23103002). I thank Jason Wang and Jean-Baptiste Ruffio for helpful discussions and suggestions about forward modelling. I acknowledge and support the cultural and spiritual importance of the

summit of Mauna Kea to the Hawaiian native community. I am grateful to share and respect this land with the community.

I thank B. Macintosh, O. Guyon, J. Lozi, P. Pathak, and A. Sahoo for helpful discussions which improved the work in sections 3.5.1 and 3.5.2.

I thank Raphaël Galicher for comments, suggestions, and discussions that have significantly improved the work in §3.5.4. I also thank Pierre Baudoz, Johan Mazoyer, Garima Singh, and J.-P. Véran for helpful discussions and suggestions on this work.

I thank LESIA for the accommodation received at Paris Observatory to complete the laboratory tests presented in §3.5.7.2.

I thank Honeywell/Comdev for the funding and support I received to complete the work presented in appendix A.

Co-Authorship

In addition to the co-authors of the published papers included in this dissertation (Gerard et al. 2018a, Gerard et al. 2019b, Gerard et al. 2018b, Gerard et al. 2019a, and Gerard & Marois 2020) and those listed in the Acknowledgments, the following co-authors contributed to the content of this dissertation:

- whole dissertation - **Christian Marois**, for the advising, input, comments, and suggested edits.
- §2.1 - **GPIES collaboration**. For the design, testing, construction, commissioning, data reduction, and references therein of the Gemini Planet Imager Exoplanet Survey (GPIES), from which data in this section is used for my analysis and pipeline development.
- §2.2 - **SCExAO collaboration**. For the design, testing, construction, commissioning, data reduction, and references therein of the Subaru Coronagraphic Extreme Adaptive Optics (SCExAO) instrument, from which data in this section is used for my analysis and pipeline development.
- §3.5.5 - **Garima Singh, Raphaël Galicher, and Pierre Baudoz**. For the advising, input, and comments used for this section.
- §3.5.6 - **Jean-Pierre Véran, Mamadou N'Diaye**. For the advising, input, and comments used for this section.
- §3.5.7.4 - **Olivier Lardière**, for comments and suggested edits throughout this section.
- Appendix A - **Com Dev/Honeywell**. For the advising, input, and comments from Neil Rowllands, Hugh Podmore, and Alan Scott.
- whole dissertation - **Gaël Chauvin**, for comments and suggested edits.
- whole dissertation - **Charli Sakari**, for comments and suggested edits.
- whole dissertation - **Trudy Lionel**, for suggested edits.

Dedication

To my family and friends, both human and feline, that helped me along the way.

Chapter 1

Introduction

In this dissertation I present my development of new algorithms and technologies that enable studies of extrasolar planets, or exoplanets. Light that is directly detected from exoplanets can be used to understand the formation and evolution of solar systems and, ultimately, the commonality of life in the Universe.

Chapter Outline

§1.1 What is an "exoplanet"?

§1.2 How are exoplanets detected?

§1.3 Why bother directly imaging an exoplanet?

§1.4 How does one directly image an exoplanet?

1.1 What is an Exoplanet?

Discussions of exoplanet detections (§1.2) must begin with a definition of an "exoplanet."

Exoplanet: A planet orbiting an object, other than the Sun, which is capable of Hydrogen fusion.

Thus, we must also first define a "planet."

Planet:

low-mass end: (IAU, 2006)

"A planet is a celestial body that:

(a) is in orbit around the Sun,

(b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and

(c) has cleared the neighbourhood around its orbit."

high-mass end: (Basri & Brown, 2006) A round object, not capable of deuterium fusion, that orbits the Sun.

Relatedly, we must also define "brown dwarf (BD)."

BD: An object capable of deuterium fusion but not Hydrogen fusion.

From Burrows et al. (2001), at solar metallicity the lower mass limit for hydrogen fusion is $M \sim 0.07 - 0.074\ M_\odot \equiv 73 - 78\ M_{\rm Jup}$ and for deuterium fusion is $M \gtrsim 13\ M_{\rm Jup}$, where $1\ M_{\rm Jup}$ is the mass of Jupiter. Note that although there is currently no single accepted definition for an exoplanet, the definition here is consistent with that of the IAU Working Group on Extrasolar Planets.[1]

[1] `http://w.astro.berkeley.edu/~basri/defineplanet/IAU-WGExSP.htm`

1.2 Methods of Exoplanet Detection

Many observational techniques have been developed over the past 40 years, leading to a large increase in the number of exoplanet detections. Many of these exoplanets are illustrated in Figure 1.1, highlighting different detection methods. I will summarize exoplanet detection methods below, utilizing the descriptions from Perryman (2011) and Bozza et al. (2016).

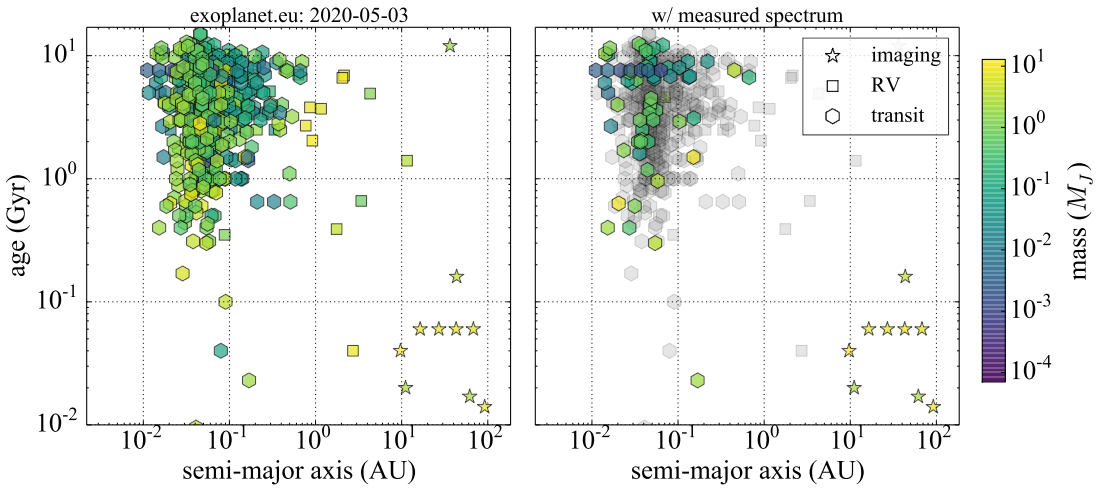


Figure 1.1 Left: All exoplanets from the the exoplanet.eu database (Schneider et al., 2011, updated as of the date listed in the title of the left panel) that have measurements for their host star's age, plotted vs. exoplanet separation (within 100 au), colour coded by exoplanet mass where symbols illustrate different detection methods. Right: all points from the left panel, where exoplanets with remaining coloured points have measured spectra, also determined from the exoplanet.eu database.

The radial velocity (RV), or "velocimetry" method, pioneered by Canadian researchers (Campbell et al., 1984), uses the Doppler effect to detect a periodic differential redshift and blueshift of the host star in an exoplanetary system. Any RV variations infer the presence of one or more exoplanets, which are revolving about the exoplanet-star centre of mass. Such a detection requires high resolution spectroscopy and advanced velocity calibration methods to reach < 1 m/s radial velocity precision for stars later than spectral type $\sim$G0. RV measurements of this motion provide a direct measurement of exoplanet orbital period and an lower limit on exoplanet mass, since for the latter radial velocity variations are not sensitive to any tangential orbital motion. This method led to the first detection of an exoplanet and now the first No-

bel Prize for the exoplanetary field (Mayor & Queloz, 1995). Biases of this method, which are apparent in Figure 1.1, include a limited sensitivity to wider-orbits beyond $\sim$10 au, a general inability to reach sufficient velocity precision on rapidly rotating hot stars because of line broadening and the subsequent relative lack of spectral lines, and systematic uncertainties due to velocity variations from stellar activity such as flares.

The transit method is used to detect exoplanetary systems that are aligned nearly edge-on with respect to the Earth by measuring the periodic "shadow" of an eclipsing exoplanet in the time-domain light curve of the host star. Due to potential systematic errors from false-positive detections (e.g., from slightly grazing eclipsing binaries), candidate exoplanet detections typically also require follow-up RV detections for confirmation. If the stellar radius is known a priori (e.g., from stellar modelling and/or interferometric observations), the period, transit depth, and transit duration can be used to measure the exoplanet radius, which, when combined with exoplanet mass from an RV detection of the same target, can be used to estimate an upper limit of bulk exoplanet density. The transit method is similarly biased to detections at smaller separations $\lesssim$1 au due to limited cadences and is also inherently limited in photometric precision by stellar activity.

The transit method is also supplemented with spectroscopy and transit timing variation (TTV) techniques. If a spectrum is obtained during an exoplanet's transit, the relative transit depth as a function of wavelength provides a spectrum of the exoplanet's atmosphere, typically requiring photometric precision better than at least a few hundred parts per million (ppm, relative to the flux of the host star) to detect atmospheric absorption features of Jupiter-mass transiting exoplanets. Additionally, if multiple exoplanets are present in a single system detected by the transit method (even if only one planet is transiting with respect to Earth) and transits from multiple orbital periods can be obtained, exoplanet masses can be measured directly by the TTV method—a numerical and analytical approach to measuring the effect of planet-planet gravitational interactions.

Exoplanets can also be detected by gravitational "microlensing." When a background star passes in and out of alignment with a foreground star, gravitational lensing can magnify the background star while the foreground star remains undetected; if a planetary system exists around the foreground star (i.e., the "lens"), the the light curve from the microlensing event can be modelled to determine the stellar mass, exoplanet mass, and exoplanet separation in the foreground exoplanetary sys-

tem. However, no microlensing events are shown in Figure 1.1 because (1) light from the foreground planetary system is never detected, and (2) a microlensing event from a single system is not predictable/periodic, preventing a measurement of stellar age.

Although other detection methods have been proposed, they have been unsuccessful thus far in discovering new exoplanets. These alternative methods include high precision stellar astrometry (Unwin et al., 2008), magnetospheric radio emission from exoplanets (Zarka et al., 2015), the existence of extraterrestrial life (Siemion et al., 2015), polarization (Jensen-Clem et al., 2018), and interferometric techniques (Lacour, 2019). With that said, several of these approaches have successfully characterized exoplanets already detected from other methods, including stellar astrometry (e.g., Muterspaugh et al., 2010) and interferometry (e.g., Gravity Collaboration et al., 2019). Furthermore, Gaia data release 4 (2022) is expected to provide many new exoplanet detections via stellar astrometry (Sozzetti, 2017).

Lastly, the imaging method (also known as "exoplanet imaging," "high contrast imaging," and/or "direct imaging") aims to directly detect the light from an exoplanet, producing an image of the exoplanet that is separate from the host star. Direct imaging will be the remaining subject of this dissertation. Figure 1.1 clearly illustrates that exoplanets detected by direct imaging cover a unique region of parameter space that is not accessible by other methods, and that each directly imaged planet can be characterized with spectroscopy, compared to only of a handful of the total exoplanets detected by the transit method via transit spectroscopy. Thus, at separations greater than about 10 au, ages less than about 50 Myrs, and (at the moment) masses greater than about 1 M_{Jup}, exoplanet imaging is uniquely poised to understand the commonality of Jovian exoplanets and their formation and evolution mechanisms and, in the future, will be poised to understand the same for rocky exoplanets and the search for life beyond the Solar System. In §1.3 below I will give a brief overview of such topics addressed by direct imaging.

1.3 Scientific Motivation for Direct Imaging

Direct imaging with current telescopes and instrumentation enables several important scientific questions to be addressed about giant exoplanets, including formation channels (§1.3.1), the composition, physical characteristics, and evolution of exoplanet atmospheres (§1.3.2), occurrence rates (§1.3.3), and connections between exoplanets and debris disks (§1.3.4). Section 1.3.5 also provides a few highlights of individual

analyses of some of the exoplanets that have been directly imaged thus far. Finally, in §1.3.6 I discuss the future goals of the field to detect and characterize habitable, rocky exoplanets.

1.3.1 Giant Exoplanet Formation

The direct imaging method enables measurements of an exoplanet's apparant magnitude at the observed wavelengths of detection/characterization and the projected separation from its host star. Spectral type fitting and/or assumptions of spectral type can then be extrapolated over the full electromagnetic spectrum to provide a bolometric apparent magnitude (e.g., Filippazzo et al. 2015). When combined with a parallax measurement, this bolometric apparent magnitude enables a measurement of exoplanet bolometric luminosity, as shown in Figure 1.2 on the y-axis for all known directly imaged exoplanets with separations less than 100 au (Marois et al., 2008b, 2010b; Lagrange et al., 2010; Rameau et al., 2013; Macintosh et al., 2015; Chauvin et al., 2017; Bowler, 2016, and references therein). As illustrated in Figure 1.2, di-

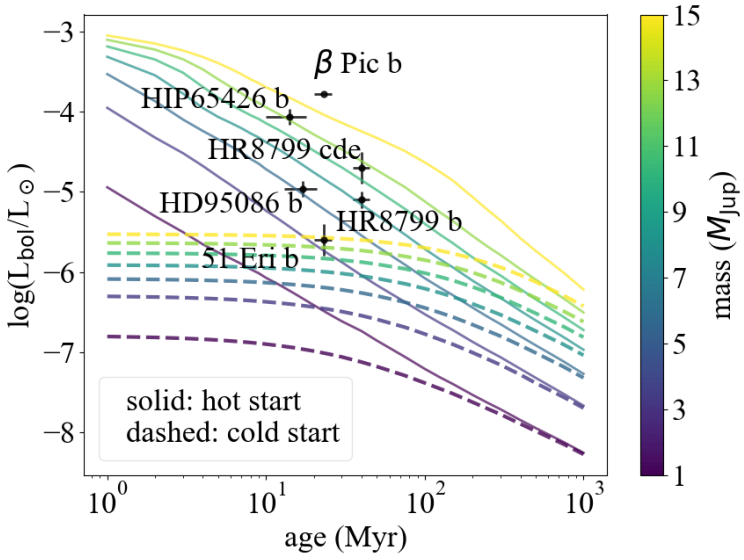


Figure 1.2 Bolometric luminosity as a function of age, shown both for real data (Marois et al., 2008b, 2010b; Lagrange et al., 2010; Rameau et al., 2013; Macintosh et al., 2015; Chauvin et al., 2017; Bowler, 2016, and references therein) and theoretical cooling curves as a function of mass (Spiegel & Burrows, 2012).

rect imaging does not directly measure exoplanet mass; this must instead be done

by combining an exoplanet luminosity measurement with a priori information about the system's age and comparisons with theoretical cooling curves (but see below). As summarized in Bowler (2016), masses can also be measured through dynamical measurements of exoplanet orbital motion as well as numerical simulations of interactions with other (detected or undetected) exoplanets and/or debris disks (see §1.3.4), although these effects are often less constraining and/or require data over long timescales.

The theoretical bolometric cooling curves in Figure 1.2 are obtained by integrating the model spectra from Spiegel & Burrows (2012) at a given, age, mass, and initial entropy over the full available 0.8 - 15μm range (a simple bolometric correction by mass or effective temperature is not realistic in this regime, as we will see in §1.3.2). For a given mass, initial entropy can be thought of as the amount of energy left over from exoplanet formation at the start of the simulation (t =1 Myr). The "hot start" curves have the highest simulated entropy and the "cold start" curves have the lowest. Without yet considering any physical mechanisms that would cause a relatively high or low residual energy of formation, it is important to note that the difference between a hot and cold start exoplanet formation scenario at $t > 1$ Myr is simply described by how these objects cool over time. With this in mind, two theories of giant exoplanet formation currently motivate the hot and cold start paradigms:

1. Gravitational Instability (GI; Boss 2000), involving the hydrodynamical collapse of gas in a protostellar accretion disk forming exoplanets on less than about 100 year orbital timescales.

2. Core Accretion (CA; Marley et al. 2007), involving the initial formation of a rocky core followed by accretion of gas onto the core. Early models from Marley et al. (2007) simulated core formation by the accretion of large planetesimals and predicted a formation timescale of a few Myr; more recent data suggesting that exoplanets may already be fully formed within a few Myrs and indicating a high abundance of smaller grains (e.g., ALMA Partnership et al., 2015) has prompted faster "pebble accretion" models (e.g., Johansen & Lambrechts, 2017) that can form similar mass cores on timescales of a few $\times 10^5$ years.

Although the cold start formation can only be attributed to CA, note that both formation scenarios are degenerate with the hot start paradigm, as identified by Mordasini (2013) (see below for further discussion).

In the CA formation scenario, gas loses energy while accreting onto the exoplanet core (Marley et al., 2007). An accretion shock is the main mechanism for energy loss during this event. This concept was initially proposed to explain protostar collapse by Stahler et al. (1980). In short, the accreting gas is "shocked" by an increase of kinetic energy because of its high incoming speed as it encounters the much heavier mean molecular weight at the core boundary. By the Virial Theorem, this increase of kinetic energy is half of the gravitational energy in the core, setting its central temperature. The gas must then cool to the temperature at the core's surface, and so it radiates away this excess energy generated from the shock. However, more recent work by Mordasini (2013) has shown that core mass governs the initial entropy and luminosity of a cooling curve, with massive cores (100 Earth masses) producing post-formation luminosities that are degenerate (as a function of total exoplanet mass) with those predicted by GI.

No such accretion shock occurs in the hot start formation scenario; exoplanets simply cool over time. Figure 1.2 best illustrates the discrepancy between hot start and cold start cooling curves at ages of less than about 100 Myrs. Thus, any observational constraints on exoplanet formation can only be made at $t \lesssim$100 Myrs (i.e., the models converge for $t \gtrsim$100 Myrs). Figure 1.2 demonstrates that 87.5% of all directly imaged exoplanets with separations less than 100 au are inconsistent with cold start exoplanet formation (i.e., despite all exoplanets remaining consistent with CA, as discussed above). However, with that said, Figure 1.2 also demonstrates a clear observational bias towards brighter exoplanets formed via the "hot start" process. The exoplanet 51 Eri b (Macintosh et al., 2015, see §1.3.5) is the only detection consistent with the Marley et al. (2007) cold start formation models; however, its bolometric luminosity also represents the sensitivity limit for current state-of-the-art exoplanet imaging instruments. Future high contrast imaging instruments and planned upgrades to current instruments will improve this sensitivity limit and will thus considerably decrease the effect of observational bias in understanding how giant exoplanets form.

Note that also as future exoplanet imaging instruments improve detection sensitivities, the overlap of detection parameter space with other methods that directly measure dynamical mass, such as stellar astrometry and/or velocimetry (§1.2), will enable improved calibration to use these hot and/or cold start cooling curves as a proxy for exoplanet mass; some of this work has even already begun (e.g., Snellen & Brown, 2018). Relatedly, newer instruments and surveys are beginning to enable

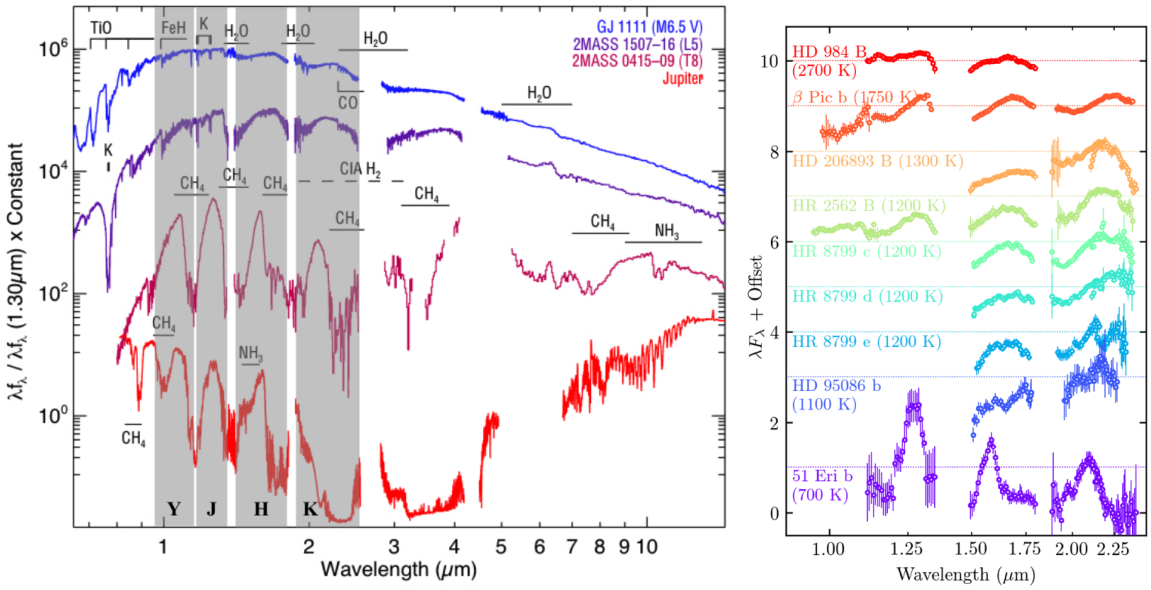
direct imaging of exoplanets at even younger ages within protoplanetary and transition disks (e.g., Keppler et al., 2018; Haffert et al., 2019); as the exoplanet is still forming in these types of systems, the measured emission is governed by the gas accretion rate, no longer directly measuring an isolated exoplanet luminosity. Thus, such calibrations between cooling curves and dynamical mass measurements will become increasingly more essential in the future.

1.3.2 Probing Brown Dwarf and Giant Exoplanet Atmospheres

In this section I will give a brief overview of the physical processes involved in the modelling of exoplanet atmospheres: the L-T transition in BDs vs. giant exoplanets. A detailed review of relevant input physics (including radiative transfer, dynamics and mixing, chemistry, gas opacities, and clouds and condensates) can be found in the review paper by Marley & Robinson (2015) and references therein.

In order to understand giant exoplanet atmospheres we must first consider the atmospheric and evolutionary properties of BDs, whose atmospheres are thought to be analogues to giant exoplanets but for which there is significantly more available data to constrain the models. Spectroscopy of BDs has shown that their spectra evolve over time, changing from spectral type L ($\sim$2000 K) to T ($\sim$1000 K) to Y ($\lesssim$500 K; see Figure 1.3a, which is adapted from Marley & Leggett 2009a). An M star spectrum is also shown for comparison to a young BD L-type spectrum; in both cases, the shape of the blackbody function is apparent. In T- and Y-type BDs, however, molecular absorption leads to significant deviations from the blackbody continuum. The cause of this change at the L to T (or L-T) transition is thought to be due to the changing conditions in the BD atmospheres, particularly the dissipation of clouds and changing chemistries, as summarized in Burrows et al. (2001) and Saumon & Marley (2008).

The atmospheres of L-type objects are thought to be dominated by H_2O and CO. Although some of these absorption features are present for the L-type spectrum in Figure 1.3 (a), L-type BDs also have clouds. In solid form, the condensates in L-type objects reach particle sizes that are greater than the wavelength of near infrared (NIR) light, thereby acting as a "grey" opacity source at these wavelengths (i.e., the particles, which are optically thick, absorb light from the emergent spectrum and reradiate it as a blackbody). These clouds must be at a relatively high elevation from the exoplanet's surface in order to obscure any other absorption bands that would otherwise dominate the spectrum. Time series data also supports the presence of



(a) (b)

Figure 1.3 (a) An illustration of spectral type evolution in free-floating BDs, adapted from Marley & Leggett (2009b). (b) A spectral library of bound BDs and exoplanets detected in the Gemini Planet Imager Exoplanet Survey (Rameau et al., in prep., from R. De Rosa, private communication). The Y, J, H, and K bands highlighted in panel a span the full wavelength range of coverage in panel b.

clouds; e.g., Apai et al. (2017) observe photometric variations in late L-type BDs that are consistent with a banded cloud structure, similar to the morphology of Neptune.

As a BD cools at any temperature, molecules will eventually condense into solids (i.e., separate from the cloud particles discussed above) and "rain out" of the atmosphere, no longer acting as an opacity source. For example, TiO and VO have well-defined absorption bands in M stars (Fig. 1.3a) that are not present in early L-type objects because these molecules condense around $T_{\rm eff} \lesssim$2000 K. Similarly, CO condenses around $T_{\rm eff} \lesssim$1500 K and forms CH_4 via the reaction: $CO+3H_2 \rightarrow CH_4+H_2O$ (Kirkpatrick, 2005, also increasing H2O absorption). Many other condensation reactions occur throughout the L-T transition around 1500 K, which are described by Kirkpatrick (2005) as a"bizarre and intricate dance." As a result of changing chemistry and the loss of clouds, BD spectra at the L-T transition begin to deviate significantly from the Planck function by as much as two to five orders of magnitude. NIR low resolution T-type spectra are typically characterized by strong methane absorption in the J and H bands (Figure 1.3a) as well as an overall blue colour relative to L-type spectra. The exact process of cloud removal during the L-T transition is

still an active area of research but in general requires that the condensates forming the clouds physically move into the optically thick region of the atmosphere at these temperatures.

Figure 1.3 (b) shows spectra of the directly imaged bound BDs and exoplanets from the Gemini Planet Imager (GPI; Macintosh et al. 2014) Exoplanet Survey (GPIES; Nielsen et al. 2019), enabling a comparison to free-floating BDs. A similar evolution of spectral type can be seen as a function of temperature. In a notable difference from BDs, the L-T transition in giant exoplanets appears to occur at colder temperatures, below at least ∼1100 K (compared to ∼1500 K for BDs). Although the generally redder colours of L-type giant exoplanets compared to field BDs have suggested that this L-T transition does indeed follow a different pathway for bound companions (e.g., Filippazzo et al., 2015), the data is far from conclusive at this point. 51 Eri b is thus far the only young (<100 Myrs), directly imaged T-type exoplanet on Solar System scales (<100 au); more data is needed to robustly sample this transition pathway in giant exoplanets, both for L- and T-type objects. Current data also further complicates this process in that all companions in Fig. 1.3b other than 51 Eri b are inconsistent with low entropy/cold start cooling curves; if there is a separate (currently undetected) population of cold start exoplanets, their evolutionary mechanisms should be considered independently from the hot start population, further motivating the need for more data and more sensitive instruments.

1.3.3 Occurrence Rate of Wide-Orbit Giant Exoplanets

Despite the relatively low number of exoplanet detections through direct imaging to date, an important result that has come out of large high contrast imaging surveys has been constraints on the mass-orbital separation distribution of exoplanets spanning various stellar host masses and system ages. In general, although wide-orbit population synthesis models of giant exoplanets have been proposed (e.g., Ida & Lin, 2004), the current lack of detections from direct imaging data prevents any reasonable predictive power. For example, these models require understanding the relative impacts of a cold start vs. hot start formation; as we saw in §1.3.1 these theories are still maturing on individual targets and will require more data to understand the dominant effects of exoplanet formation in a statistical context. Accordingly, most work thus far on this topic has been empirical, motivated and influenced by results from large observational surveys. However, with that said, these empirical results have placed

important constraints on terrestrial exoplanet formation and the predicted habitability of earth-like exoplanets; e.g., dynamical simulations from Raymond (2006) suggest that in systems with giant exoplanets, the formation of a terrestrial exoplanet in the habitable zone (see §1.3.6) is less likely when the giant exoplanet is within 2.5 au, and that water delivery to a habitable, terrestrial exoplanet is only possible in significant amounts when the giant exoplanet is beyond 3.5 au.

Nielsen et al. (2008) carried out one of the first studies on the occurrence rate of exoplanets constrained by direct imaging surveys. In short, an occurrence rate analysis involves (1) running Monte Carlo simulations to sample the full range of physical orbital parameters within a range of exoplanet masses and separations for a given set of exoplanetary targets (corresponding to targets observed in a survey), (2) comparing these simulations with corresponding instrument sensitivity as a function of angular separation, and (3) using this comparison to understand the commonality of exoplanets at a given mass and separation, for a given instrument sensitivity. Using data from a sample of 60 stars, they found that less than 20% of stars host exoplanets with masses greater than 4 $M_{\rm Jup}$ and separations between 20 and 100 au at a 95% confidence level. In other words, wide-orbit giant exoplanets are rare. As expected from the observational results already presented in §1.3.1 and 1.3.2, this general conclusion has not changed from the results of any more recent surveys to date (Bowler, 2016; Nielsen et al., 2019).

Although similar more recent analyses have corroborated the Nielsen et al. (2008) results (e.g., Bowler 2016 find an occurrence rate of $0.8^{+1.2}_{-0.5}$% for 5-13 $M_{\rm Jup}$ exoplanets between 10 and 100 au at a 95% confidence level), an additional caveat remains: all current models have assumed a hot start formation. As already illustrated in Figure 1.2, a factor of ∼5-10 improvement in sensitivity of current high contrast imaging instruments to dimmer $L_{\rm bol}/L_{\odot}$ values will provide more powerful constraints on the giant exoplanet occurrence rate compared to current surveys. Although we know that hot start giant exoplanets are rare, we do yet have the sensitivity to make the same conclusions for exoplanets formed by core accretion. Or, if these cold start exoplanets are instead more abundant than their hot start analogues at a given mass and separation, only one data point (Macintosh et al., 2015) has been collected at the very tip of an "iceberg" that will be explored further by future surveys with more sensitive instruments.

1.3.4 The Debris Disk-Exoplanet Connection

A debris disk is an astrophysical analogue to the Asteroid and Kuiper belts in our Solar System; as we will see below, recent work suggests that exoplanets detected by direct imaging may be more common in systems with a debris disk than in systems without one. As described in Wyatt (2008), star formation over a timescale of a few Myrs will form an accretion disk of gas and dust within which rocky planetesimals will form; regardless of the formation pathway (§1.3.1), on this ∼few Myr time scale gas will be blown out of a newborn stellar system by a combination of radiation pressure and stellar winds, revealing a residual disk of rocky planetesimals. Then on a time a scale of a few Gyrs, a debris disk will form over time from collisions between remaining large rocky planetesimals (with diameters greater than about 1000 km), ultimately creating a grain size distribution ranging from sub-μm- to mm-sized dust particles that typically span out to a few hundred au in separation from the host star.

Debris disks can be detected and spatially resolved in both the optical/NIR , sensitive to $\sim$ μm-sized particles from reflected stellar light (e.g., Draper et al., 2016), and at sub-mm wavelengths, sensitive to ∼mm-sized grains from thermal emission (e.g., Matthews et al., 2014). Due to limited spatial resolution of current facilities, either imaging technique is only capable of resolving the Kuiper belt analogue in other systems; instead, the presence of a second warmer asteroid belt analogue can only be detected through an excess of sub-mm emission in an object's spectral energy distribution (SED). With this more classical technique (e.g., Sadakane & Nishida, 1986), a single Kuiper belt analogue is detected by observing a "bump" on the Rayleigh-Jeans tail of the stellar SED at sub-mm wavelengths, revealing a separate temperature component originating from the debris disk; when two such bumps are detected, the hotter of the two indicates the presence of an asteroid belt analogue whereas the colder component is a Kuiper belt analogue.

Of particular interest to high contrast imaging is the possible exoplanet-debris disk connection; four of the five stars with directly imaged exoplanetary systems within 100 au have debris disks—β Pictoris, HR 8799, HD 95086, and 51 Eridani—with the latter three hosting warm and cold debris disk components interior and exterior to the exoplanets, respectively (Matthews et al., 2014; Su et al., 2015; Patel et al., 2014). Meshkat et al. (2017) have recently presented the first study of this statistical significance, using their own Keck/NIRC2 observations and a compilation of literature results from previous surveys to compare 130 stars with debris disks—

defined by their WISE W1-W4 colours, indicating a mid-infrared excess—to 277 stars with no detected mid infrared excess. The statistical analysis used by Meshkat et al. (2017) to determine the occurrence rates of exoplanets in debris disk systems is similar to the methods used in Nielsen et al. (2008) from §1.3.3. Overall the main results from Meshkat et al. (2017) are generally consistent with Nielsen et al. (2008) and subsequent similar analyses but illustrate an important new distinction regarding the correlation between debris disks and directly imaged exoplanets: at a 68% confidence level, the occurrence rate for the debris disk sample is $6.3^{+3.5}_{-2.6}$ % and for the control sample is $0.73^{+1.07}_{-0.53}$%, suggesting for the first time a tentative distinctness between these two samples. However, although considerable attention was made to match a similar set of ages, spectral type, and a uniform detection threshold across multiple surveys, it is important to note that these results are dominated by the choice of which previously detected exoplanets to include as detections in the disk vs. control sample. As a result, in compiling different detections from different surveys (with different biases) and due to the overall low number of individual exoplanetary targets, many unavoidable subjective decisions are made by Meshkat et al. (2017) on which targets to include/not include in the study, influencing the ultimate occurrence rates between the debris disk and control sample. Thus, although the study by Meshkat et al. (2017) is an important first step towards understanding the debris-disk exoplanet connection, as stated by the authors this distinctness should be viewed as "tentative." Also, as discussed in §1.3.3, this Meshkat et al. (2017) analysis only uses hot start cooling curves; no statistical occurrence rates can can be measured for any cold start giant exoplanets until future instruments/surveys can reach a factor of ∼5-10 better in sensitivity.

1.3.5 Analysis of Individual Targets

The analysis of individual targets has also provided important and exciting results, a few of which are highlighted below:

HR 8799 b,c,d, and e (Marois et al., 2008b, 2010b),

This was the first directly imaged exoplanetary system on Solar System scales, and it is one of only two imaged multi-planetary systems thus far. The system has four giant exoplanets with masses of about 5 to 7 M_{Jup} at separations between about 14 and 68 AU. A number of subsequent analyses have been conducted, a few of which are highlighted below:

- Konopacky et al. (2016) published results from the longest monitoring campaign of a single directly imaged exoplanetary system on a single telescope, using the Keck/NIRC2 high contrast imaging instrument, with observations running between 2009-2014. In this approach all data is acquired and reduced in the same way, removing any differential systematic effects which may be more problematic from compilations of of different literature results, each of which reduce the data differently (e.g., Bowler, 2016). Orbital acceleration is detected at $> 3\sigma$ for exoplanets d and e. This work has also produced perhaps the most popular public movie of an exoplanetary system, showing visually apparent orbital motion of all four exoplanets.
- This detection of four young co-orbiting giant exoplanets has generated debate about the dynamical stability of the system. Fabrycky & Murray-Clay (2010) simulated that the system would have become unstable after only ∼0.1 Myrs without invoking mass limits and resonances. However, Götberg et al. (2016) have recently simulated long term stability of the system without any resonances. Both claims are currently consistent with the data from Konopacky et al. (2016).
- Konopacky et al. (2013) and Barman et al. (2015) have obtained $R \sim 4000$ spectra of exoplanets c and b, respectively, using the Keck/OSIRIS integral field spectrograph (IFS; see §1.4.3.3). At these resolutions, individual absorption lines are detected from carbon monoxide, water, and (in b) methane. Although a C/O ratio is measured and can in principle be used to distinguish a GI vs. CA formation scenario (GI indicated by a similar value to the host star, CA indicated by a different value from the host star), both results are ultimately consistent with the data.

β Pictoris b (Lagrange et al., 2009, 2010),

Imaging of a warped debris disk around the star β Pictoris by Smith & Terrile (1984) suggested that this system may contain an unresolved inner exoplanet (Lecavelier Des Etangs et al., 1995). This young (∼23 Myr) massive (∼13 $M_{\rm Jup}$) exoplanet was first detected with the VLT in 2003 on the northeast side of the star (Lagrange et al., 2009), and then detected again in 2009 on the southwest side of the star (Lagrange et al., 2010), confirming orbital motion. A few highlights of β Pic b analyses are below:

- Morzinski et al. (2015), similar to the initial work from Bonnefoy et al.

(2013), used the observations from the Magellan telescope combined with re-calibrated literature values to obtain the full 0.9 to 4.8 μm SED, similar to the theoretical SEDs from Spiegel & Burrows (2012, as discussed in §1.3.1) but obtained entirely from observations. With observations covering $> 80\%$ of β Pic b's total energy, extending the SED to other wavelengths via the blackbody function yields an empirical bolometric luminosity of $\log(L_{\rm bol}/L_{\odot}) = 3.78 \pm 0.03$, consistent with extrapolations from atmospheric models but inconsistent by $\sim$0.1 dex with typical field BD bolometric corrections (e.g., Filippazzo et al., 2015), similar to the systematic discrepancies between atmospheric properties of field BDs vs. bound companions (see §1.3.2).

- Wang et al. (2016) presented a high-signal-to-noise ratio (S/N) astrometric analysis to determine if the β Pic b exoplanet would transit its host star in the process of orbiting back to the northeast side of the star. They ruled out the possibility of a transit with a 10σ significance but confirmed that the exoplanet's Hill sphere (i.e., the region of influence where objects can be gravitationally bound to the exoplanet) will transit. Accordingly, photometric campaigns, which are sensitive to any rings or moons in the exoplanetary system, have recently finished observations between the ingress and egress of this event, April 2017-January 2018, with no detections found (Lous et al., 2018). The most recent, post-conjunction orbital solutions are presented in Gravity Collaboration et al. (2020).
- Snellen et al. (2014) used high spectral resolution observations to detect CO and measure the effects of rotational line broadening. The promise of this "high dispersion spectroscopy" technique, where high resolution ($R \equiv \lambda/\Delta\lambda \sim 100,000$) raw spectroscopic data is cross-correlated with a template to enable detection of individual molecular lines, has initiated a new perspective in the field of direct imaging, prompting subsequent research and development in combining this approach with high contrast imaging (Snellen et al., 2015; Wang et al., 2017; Mawet et al., 2017). Many current and planned high contrast imaging instruments are now pursuing this high dispersion spectroscopy approach (e.g., Jovanovic et al., 2019).

51 Eridani b (Macintosh et al., 2015).

To date, this detection from GPI has produced either the lowest mass directly

imaged exoplanet within 100 au, according hot start models (2 ± 1 $M_{\rm Jup}$), or the only exoplanet consistent with Marley et al. (2007) cold start models (7 ± 5 $M_{\rm Jup}$). It is also the only T-type directly imaged exoplanet within 100 au. An analysis by Rajan et al. (2017) presents spectro-photometry spanning 1-5 μm from both the GPI and Keck/NIRC2 instruments. A comparison to field BDs shows that 51 Eri b is systematically redder, again suggesting an effect of different atmospheric physics that influence the evolutionary processes in comparison to unbound BDs (as discussed in §1.3.2), although clearly more detections are still needed to confirm these discrepancies. A spectroscopic analysis from the GPI 1.1-2.4 μm data suggests that the exoplanet has a cloudy but patchy atmosphere.

1.3.6 Future Goals: Habitable Exoplanets

The detection and characterization of habitable exoplanets has long been seen as a motivating goal towards understanding our place in the Universe. However, even though many terrestrial-mass habitable zone (HZ)[2] exoplanets have thus far been detected by transit and RV techniques (e.g., Kane et al., 2016), none so far have measured spectra and none have been detected through direct imaging. Over the last few decades, the astrobiology community has demonstrated that spectroscopy is necessary to determine whether or not an exoplanet in the HZ is actually habitable. Spectroscopy enables the identification and characterization of biomarkers, such as oxygen, water, methane, and carbon dioxide; these relative abundances may also evolve on geological timescales (The LUVOIR Team, 2019). Although, exoplanet imaging instrumentation is not yet sensitive enough to detect and characterize such targets (e.g., see Fig. 1.12), direct imaging remains the primary detection method capable of achieving this endeavour in the future; compared to transit spectroscopy (see §1.2), direct imaging is less sensitive to orbital inclination and more efficiently reaches the exoplanet surface (enabling detection of biomarkers from vegetation in addition to the atmosphere), increasing the robustness of habitable exoplanet candidate validation (The LUVOIR Team, 2019).

Many ground- and space-based observatories have identified these advantages enabled by the direct imaging method as priorities for future telescopes and/or instru-

[2]The HZ is defined as the region where liquid water can exist on the exoplanet's surface; stellar mass, exoplanet mass, and distance from the host star all impact the definition of a HZ (Kopparapu et al., 2014).

ments, although in different regimes of detection space, illustrated in Table 1.1. As shown, future space missions will search for "Earth twins" (i.e., around FGK stars) from reflected starlight at visible wavelengths (e.g., The LUVOIR Team, 2019), while future ground-based extremely large telescope (ELT) high contrast imaging instruments will search for habitable exoplanets around both M stars from reflected starlight in the NIR (e.g., Guyon, 2011) and FGK stars from thermally emitted exoplanet light at 10 μm (e.g., Quanz et al., 2015). These regimes are separated due to different requirements in angular resolution and contrast (or instrument sensitivity; see §1.4.2 for a more formal definition), which will be discussed further in §1.4; accessing FGK star HZs in the visible will require a space-based observatory to provide the stability requirements needed to reach deeper contrasts, while FGK stars at 10μm and M stars in the NIR require more modest contrasts achievable from the ground but higher angular resolutions only accessible by ELTs. Research and development on this front is now rapidly expanding, both to show that reaching these contrast goals is feasible once the relevant ground- and space-based observatories come online in the next ∼10-30 years (Guyon, 2018) and to define the relevant spectroscopic biomarkers that these instruments will need to detect (Schwieterman et al., 2018).

Table 1.1: Detection parameter space of habitable exoplanets for different regimes.

observatory type / emission mechanism	**ground-based ($D \gtrsim 25$ m)**	**space-based ($D \lesssim 15$ m)**
reflected light (visible/NIR)	M stars (NIR)	FGK stars (visible)
thermal emission (10 μm)	FGK stars	n/a

1.4 Observational Methods: Direct Imaging of Exoplanets

Given that direct imaging has the potential to revolutionize our understanding of exoplanetary systems, the technical aspects of this method are discussed below.

1.4.1 Astronomical Image Formation

The electromagnetic field of an unresolved astronomical point source as it hits the primary mirror of a telescope can be modelled as:

$$\text{(pupil plane electromagnetic field)}_{(t,x,y,\lambda)} = A_{(t,x,y,\lambda)}\, e^{i\,\phi_{(t,x,y,\lambda)}}. \tag{1.1}$$

Equation 1.1 is a scalar plane wave that consists of a phase, ϕ (in units of radians), and amplitude, A (in units of $\sqrt{\text{energy}}$), of the complex field from electromagnetic radiation, distributed spatially across the telescope pupil in x and y and at a given instance in time, t, and wavelength λ. From now on I will refer to "electromagnetic field" as "electric field" (assuming for the latter the use of detectors that are insensitive to the magnetic field component). The relationship between the electric field in equation 1.1 (i.e., the "entrance pupil") and the primary focal plane of the telescope is a two dimensional Fourier transform, called "the Fraunhofer far field approximation" (Steck, 2015, Chapter 12):

$$\begin{aligned}\text{focal plane electric field}_{(t,\alpha_x,\alpha_y,\lambda)} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(A_{(t,x,y,\lambda)}e^{i\phi_{(t,x,y,\lambda)}}\right)e^{2\pi i(\alpha_x x+\alpha_y y)}dx\,dy \\ \text{focal plane electric field}_{(t,\alpha_x,\alpha_y,\lambda)} &\equiv \text{FT}\left\{A_{(t,x,y,\lambda)}e^{i\phi_{(t,x,y,\lambda)}}\right\}\end{aligned} \tag{1.2}$$

where α_x and α_y are the spatial coordinates in the focal plane, and the FT$\{\}$ symbol represents the Fourier transform operator. For simplicity, in future notation I will drop the use of "(t, x, y, λ)" and "$(t, \alpha_x, \alpha_y, \lambda)$." Lastly, a detector in either the pupil or focal plane measures energy, or the square modulus of the electric field, given by

$$\text{energy} = |\text{electric field}|^2\,. \tag{1.3}$$

Accordingly, the image recorded on a detector in the focal plane is

$$\text{focal plane image} = \left|\text{FT}\left\{Ae^{i\phi}\right\}\right|^2, \tag{1.4}$$

which is also called the point spread function (PSF). For a circular telescope with diameter D and a flat wavefront with no amplitude or phase aberrations (i.e., $A = 1\ \forall\ r \leq D/2,\ A = 0\ \forall\ r > D/2,\ \phi = 0$), Equation 1.4 is an Airy function with an angular full width at half maximum (FWHM) of λ/D in radians, where λ is the wavelength of light (e.g., for a 1 m telescope at $\lambda = 1\mu$m, $\lambda/D = 10^{-6}$ radians $\approx$ 0.2 arcseconds). The separation between the first two minima of the Airy function is approximately 1.22 λ/D. Similarly, the physical size of the FWHM resolution element in the focal plane is $f_{\text{num}} \times \lambda$, where f_{num} is the focal ratio ([focal length]/D) at the focal plane of interest (e.g., for an $f10$ beam at $\lambda = 1\mu$m, $\lambda/D = 10\mu$m). This principle is also known as the Rayleigh criterion: a point source at infinity passing through a circular aperture will be resolved as an Airy function with a finite angular and physical size of λ/D and $f_{\text{num}} \times \lambda$, respectively. An Airy function is shown in the bottom left panel of Fig. 1.4.

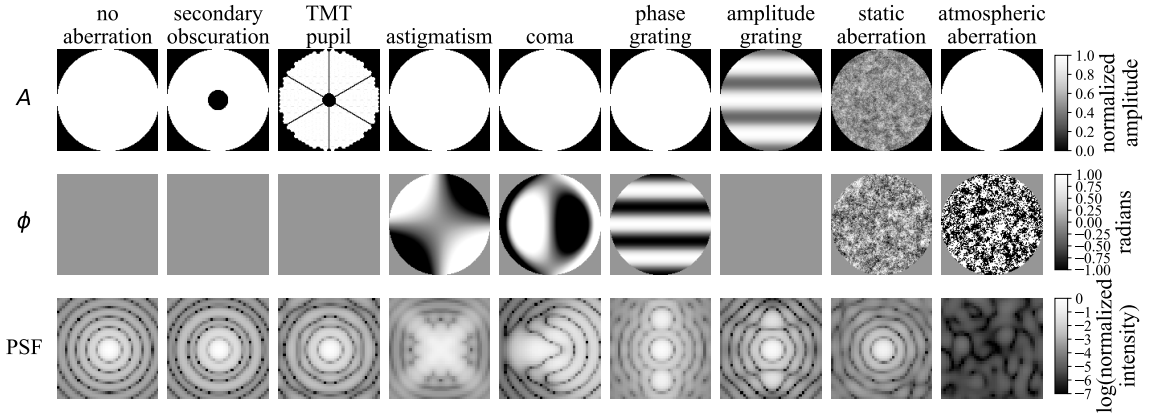


Figure 1.4 An illustration of the effect of amplitude and phase patterns on the PSF. In the relationship PSF= $\left|\text{FT}\left\{Ae^{i\phi}\right\}\right|^2$, A, ϕ, and PSF are shown in the top, middle and bottom rows, respectively, for a variety of cases common in astronomical imaging (the Airy function is in the bottom left). For each row, the scale on the right is the same for all images. The illustrated PSFs all cover a 10×10 λ/D field of view (FOV) around the central optical axis.

When the phase and/or amplitude in the pupil plane deviate from a perfectly flat position, the PSF deviates from a perfect Airy function, as illustrated in Fig. 1.4.

Common forms of amplitude variation include discontinuous pupil illumination, e.g., from secondary obscuration, support structures, and a segmented aperture; a grating pattern; and static aberration. In phase, Zernike polynomials (shown for astigmatism and coma) are commonly used to describe optical alignment and wavelength-dependent deviations from an Airy function for a circular aperture. The phase grating pattern is similar to the amplitude grating; a sine wave, in phase or amplitude, across the pupil of an infinitely large aperture will generate a PSF of two Dirac delta functions, located at separations of $\pm f(\lambda/D)$, where f is the sine wave "spatial frequency" in cycles per pupil along the corresponding axis in the pupil and focal planes ($f = 3$ along the y axis for the amplitude and sine grating cases in Fig. 1.4). For a finite aperture, this delta function is convolved with an Airy function to generate two "copies" of the on-axis PSF, enabling additional constructive and destructive interference of the focal plane electric field between the two "spots," the central core, and each corresponding Airy pattern. Lastly, static optical aberrations and atmospheric turbulence are ultimately described by a full power law, meaning a linear combination of all spatial frequencies, each with relative weights. Static aberration occurs in both phase and amplitude, each typically with a -2 power law (i.e., the intensity of the Fourier transform of the phase or amplitude screen, or spatial power spectral density (PSD), is modelled by a r^{-2} polynomial, where r is a radial grid at the centre of the Fourier plane of the phase/amplitude screen), while atmospheric phase aberration is modelled by a -11/3 power law, called a "Kolmogorov" PSD (Tyson, 2011).

It is important to note here the difference between the PSF for the static vs. atmospheric phase aberration cases in Fig. 1.4. For the former, the electric field, with a standard deviation of 0.2 rad root mean square (RMS) over the pupil,[3] can be approximated by a second order Taylor expansion ($A\, e^{i\,\phi} \approx A\,(1 + i\,\phi - \phi^2/2)$) such that the PSF includes a zeroth order Airy function and three higher order "speckle" terms (Perrin et al., 2003). This Taylor expansion generally holds for $\sigma\{\phi\} < 1$ rad RMS, where here $\sigma\{\}$ is a standard deviation operator over the aperture of the entrance pupil. In contrast, for the latter case of atmospheric aberration (a standard deviation over two orders of magnitude higher) in Fig. 1.4, no Airy function is visible

[3] In future notation, I will also refer to the amplitude of phase aberration in nanometres. Assuming a wavelength, λ, this can be converted to radians via the equation [phase aberration in rad] = [phase aberration in nm]/[λ (in nm)] $\times 2\pi$. Unless otherwise noted, I will hereafter use $\lambda = 1.65$ μm for this conversion, a standard NIR "H-band" observing wavelength.

in the PSF, which is instead dominated by speckles out to angular scales typically at least an order of magnitude larger than λ/D in the NIR for a 10 m-class telescope (i.e., seeing-limited instead of diffraction-limited). In other words a diffraction-limited resolution can only be reached for phase aberration with $\sigma\{\phi\} \lesssim 1$ radian RMS.

1.4.2 Goals of High Contrast Imaging

The main goals of exoplanet imaging are twofold:

1. *Spatially resolve (i.e., separate) an exoplanet from its host star*

   This feat can be achieved by using the largest possible telescopes on the ground using adaptive optics (AO; see §1.4.3.1) or in space, both of which, as discussed above in §1.4.1, are limited to a resolution of λ/D. From this principle, at the same wavelength a larger ground-based telescope using AO achieves a better resolution than a smaller space-based telescope. An exoplanet at a separation of 10 au, a distance of 100 parsecs, and a face-on orbital inclination will lie at a 0.1" separation from its host star. For a 10 metre ground-based telescope at 1.65 μm, this separation is $\sim$2.9 λ/D, but for a 1 metre space-based telescope at 600 nm this same separation is $\sim$0.8 λ/D, illustrating that only the ground-based telescope could resolve such an exoplanet.

2. *Achieve the necessary contrast to perform photometry, spectroscopy, and/or astrometry of the exoplanet.*

   Although definitions of "contrast" vary in the literature, two main principles generally apply:

   (a) An astrophysical flux ratio indicates the ratio of the star-to-exoplanet flux. As illustrated in Figure 1.2, the brightest gas giant exoplanets are $\gtrsim 10^4$ dimmer than their host star (assuming host star luminosities within a factor of a few to 1 $L_\odot$). An Earth-Sun analogue has an astrophysical flux ratio of $\sim 10^7$ at 10 μm and $\sim 10^{10}$ at visible wavelengths.

   (b) Contrast is the necessary sensitivity, or noise floor, of an instrument that is required to detect an exoplanet at a given astrophysical flux ratio; depending on the science case and sources of aber-

> ration that are limiting detection sensitivity, point 2a can also be referred to as "contrast," often in the context of space telescope applications (e.g., Seo et al., 2019). In the aberration-limited regime (e.g., with ground-based telescopes), contrast is often better represented by computing the standard deviation, rather than intensity, at a given separation. In a simplistic example, assuming Gaussian noise,[4] a 5σ detection of a gas giant exoplanet at an astrophysical flux ratio of 10^4 requires a 1σ contrast of 5×10^{-5} at the separation and position angle of the exoplanet. Unless otherwise noted, I will use a standard deviation-based contrast definition in this dissertation.

Since the early stages of high contrast imaging research and development (e.g., Racine et al., 1999), obtaining the necessary contrast for exoplanet detections has been difficult, even with the advent of 10-m class telescopes. This challenge has driven the invention of new AO, coronagraphic, imaging processing, and optical correction methods, all of which will be both introduced below (§1.4.3.1, §1.4.3.2, §1.4.4.1, and §1.4.4.2, respectively) and subsequently presented as the main emphasis of this dissertation.

1.4.3 High Contrast Imaging Subsystems

Figure 1.5 illustrates the main components typically included in a current state-of-the-art ground-based[5] high contrast imaging instrument, which are also outlined below.

1. **an AO system** (§1.4.3.1), which is used to reach the λ/D resolution limit described above. Without AO, optical aberrations from atmospheric turbulence can degrade this limit by more than an order of magnitude. Astronomical AO systems include a wavefront sensor (WFS) to measure atmospheric turbulence, a deformable mirror (DM) to correct for atmospheric turbulence, and a real time controller (RTC) to connect the DM-WFS feedback loop (Fig. 1.5) at ∼kHz frame rates in order to "keep up" with the atmosphere. A dichroic typ-

[4]Note that this assumption is wrong and often the origin of varying contrast definitions (e.g., Soummer & Aime, 2004; Marois et al., 2008a; Mawet et al., 2014).

[5]Although many concepts in this dissertation also apply to space-based high contrast imaging systems, unless otherwise stated, all subsequent figures and text will refer to ground-based systems.

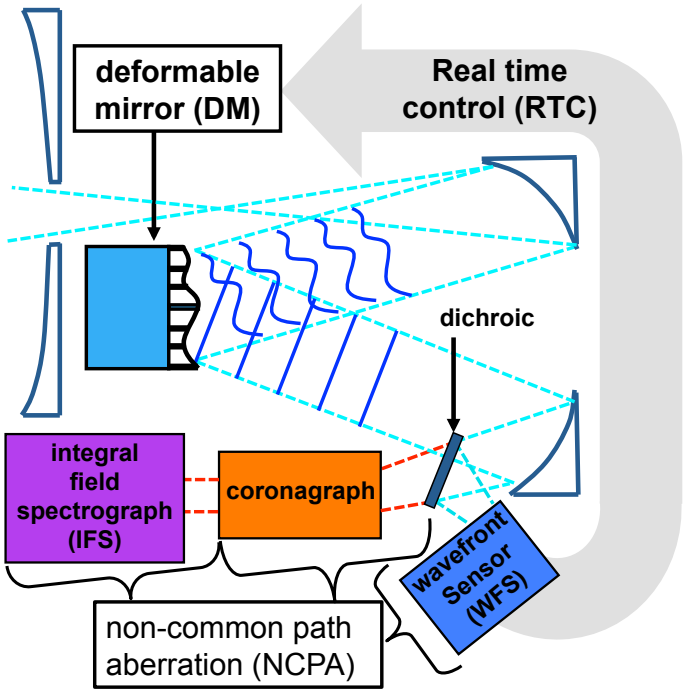


Figure 1.5 Typical subsystems of a high contrast imaging system.

ically splits the WFS and science path between visible and NIR wavelengths, respectively.

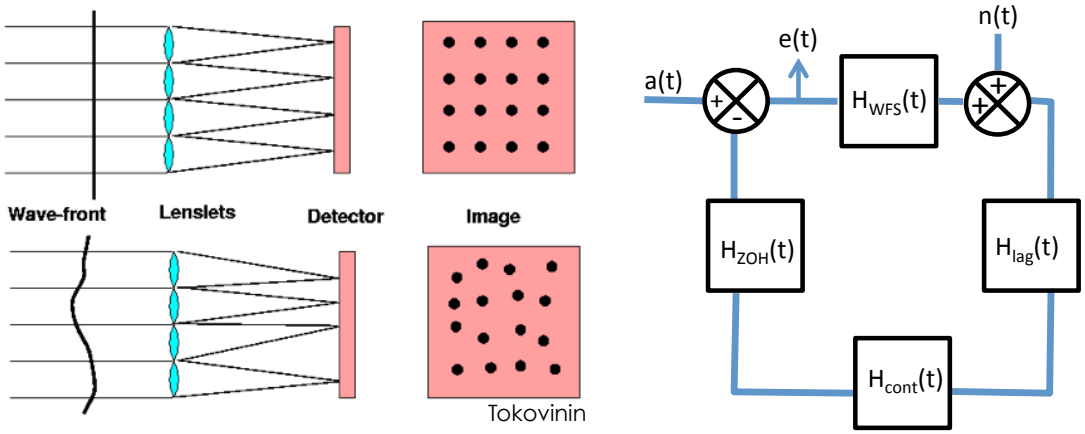
2. **a coronagraph** (§1.4.3.2), which optically attenuates star light without removing exoplanet light and improves the "raw" contrast initially recorded on a detector, compared to a non-coronagraphic image.

3. **an IFS** (§1.4.3.3), which generates a coronagraphic image at each wavelength sampled across the chosen bandpass. This detector plane can also be re-imaged as an input to a high dispersion spectrograph to enable the "high dispersion coronagraphy" method as discussed in §1.3.5.

After recording IFS images, additional image processing and/or wavefront control can be used to improve contrast (these concepts are not illustrated in Fig. 1.5, but are discussed further in §1.4.4.1 and §1.4.4.2, respectively). An additional important component of Figure 1.5 is the presence of non common path aberrations (NCPAs), which will be discussed in §1.4.4. Briefly, NCPAs arise from wavefront errors (WFEs) imparted by any optics downstream from the dichroic; WFEs from optics in the AO WFS path are not common to the coronagraph-IFS path, and WFEs from optics in the coronagraph-IFS path are not measurable by the AO WFS, both of which will

cause (quasi) static aberrations in the IFS image, called "speckles" (§1.4.4), that are not correctable by a conventional AO system, ultimately requiring additional image processing and/or optical correction methods. Additionally, aberrations in the common path of the AO WFS and IFS (i.e., upstream of the dichroic) also generate speckles in the coronagraphic image, both through insufficient sensitivity of the AO WFS to measure nanometre-level wavefront aberrations (separate from the capability of a DM to correct for these aberrations) and wavelength-dependent wavefront evolution, called "chromaticity" (see §1.4.4), between the typically visible AO WFS and NIR IFS.[6]

1.4.3.1 Adaptive Optics



(a) (b)

Figure 1.6 Schematics showing two components of a typical AO system. (a) A Shack Hartmann wavefront sensor (SHWFS), from (Tokovinin, 2001), and (b) an RTC (from Gerard 2017b; see text for description).

In addition to a DM, which is already illustrated in Figure 1.5, Figure 1.6 illustrates additional details about the two other essential components of an AO system (Tyson, 2011), which are outlined below:

- a WFS (Fig. 1.6a)

  Figure 1.6a, from Tokovinin (2001), illustrates the concept of a SHWFS, a common WFS in many AO systems. For a flat wavefront, a two dimensional

[6] In addition to wavelength-dependent quasi-static errors (Marois et al., 2000), these chromaticity effects are also a limitation for residual AO errors (Guyon, 2005).

grid of lenses, called a "lenslet array," focuses an aligned grid of spots, or micro-PSFs, onto the detector, which can be used as a calibrated "zero point" for a corrected wavefront (top row). When phase aberration caused by atmospheric turbulence propagates through the same setup, however, the grid of spots on the detector will no longer be aligned (bottom row). Using the calibrated grid of spots from the top row, a WFS measurement like the one in the bottom row sends correction information to the DM (through the RTC) so that each spot is driven back to the centre of its sub-aperture (i.e., a square grid centred around each spot in the presence of a flat wavefront). The benefit of a closed-loop AO system is therefore that every recorded WFS image will continuously instruct the DM to try to drive the spots back to their calibrated positions, allowing iterative corrections during the time that the wavefront remains "static." This coherence timescale, called τ_0, is typically ∼2-5 ms at astronomical observatories (Schöck et al., 2009), thus motivating the importance of running the AO RTC at kHz speeds. More sensitive AO WFSs are also an active area of research; a few are discussed further below and later in this dissertation (§3.5.3, §3.5.4, §3.5.6.2).

- A RTC (Fig. 1.6b)

  This illustration shows the concept of a controller from Gerard (2017b), which is typical for most AO systems. The block diagram is a standard notation for closed-loop systems modelled in the time or frequency domain (e.g., Ogata, 2009); different components of temporal response are modelled in the complex plane for each box (with each component called an "open loop transfer function" and represented by an "H") , and the circular junctions ($\bigotimes$) indicate where a signal is added/subtracted from the AO feedback loop. The open loop transfer functions are models of temporal effects from Alloin & Mariotti (1994) and Véran & Herriot (2009) and are as follows: H_{WFS} is for the WFS, H_{lag} is for the computation time required to convert a WFS measurement to DM commands (typically of order ∼a few hundred μs for current AO systems), H_{cont} is for the controller used in the system (see below), and H_{ZOH} is for a zero order hold, representing the effect that the DM is not continuously deforming but rather only changing shape in "step function" responses at the frame rate of the AO system. The signals $a(t)$, $e(t)$, and $n(t)$ represent the input, output, and WFS noise signals of the AO loop, respectively. The quantity $e(t)/a(t)$ is called

the rejection transfer function, or H_{rej}, and represents how well the AO system is able to correct for atmospheric turbulence in the temporal domain (i.e., separate from the spatial domain of wavefront correction), whereas $n(t)/e(t)$ is called the noise transfer function, or H_{n}, and represents the temporal effects of how WFS measurement noise is attenuated/amplified through the AO feedback loop. Ideally, we want both $|H_{\mathrm{rej}}|$ and $|H_{\mathrm{n}}|$ to be less than one over all temporal frequencies, ensuring both a good correction and good system stability. If $|H_{\mathrm{n}}| > 1$, noise recorded on the WFS is being amplified through the feedback loop, which will lead to instability. Achieving good system stability will depend entirely on the controller used, H_{cont}. A common controller used in an AO system is an integrator, which determines what commands to send to the DM by adding some percentage of the most recently calculated DM commands to the previous DM commands; the "some percentage" component of the controller is called the "gain," and can be adjusted specifically to minimize both $|H_{\mathrm{rej}}|$ and $|H_{\mathrm{n}}|$ over all temporal frequencies. More sophisticated AO controllers are an active area of research; a few are further discussed below in this section.

In general, extreme adaptive optics (ExAO), which is used for exoplanet imaging, is often distinguished from other forms of AO by the following properties:

- Typical targets are nearby ($\lesssim$200 pc, as motivated at the beginning of §1.4) and relatively bright (I mag $\lesssim 10$). In other applications of AO that observe more distant/dimmer stars, a nearby bright guide star must be used (either natural or artificial by a laser); in ExAO the science target is the guide star.

- The typical scientific FOV for ExAO correction is $\lesssim$ a few arcseconds (also motivated at the beginning of §1.4; as discussed in §1.3.3, observed exoplanets are rare beyond a few hundred au). Other non-ExAO "flavours" aim to correct a FOV $\gtrsim$ a few arcminutes.

- The AO correction is generally "higher order," meaning ExAO uses

    1. At least one $\gtrsim$1000 actuator DM (called a "tweeter") and typically also a second lower actuator count DM with more stroke (i.e., the maximum physical range a DM actuator can "push" or "pull") than the tweeter, called a "woofer" (Lavigne et al., 2007).
    2. A WFS that is more sensitive than a standard SHWFS (Guyon, 2005), meaning the measurement is less corrupted by

(a) aliasing,

(uncorrectable spatial wavefront modes beyond the number of DM actuators that are instead measured by the WFS as correctable modes),

(b) non-linearities, and

(the residual WFE level at which the assumption of a linear relationship between aberration measured by the WFS and commands applied to the DM breaks down)

(c) photon/detector/background noise.

(the level at which uncorrelated noise corrupts the measurement of the desired noiseless wavefront signal).

More sensitive WFSs include

- A Zernike WFS (Zernike, 1934; N'Diaye et al., 2016), which uses an intermediate focal plane optic that provides a piston phase shift of $\phi = \pi/2$ within a $(\lambda/D)/2$ radius from the optical axis, enabling phase aberration in the entrance pupil to be converted to intensity (i.e., pixel values on a detector) in the re-imaged pupil plane downstream of this optic,
- A Pyramid WFS (Ragazzoni, 1996), which uses a four-sided pyramidal optic in an intermediate focal plane, re-imaging four off-axis pupils and also converting phase into intensity aberration in the downstream pupil plane,
- A focal plane WFS (Jovanovic et al., 2018, and references therein, see §1.4.4.2 and §3.5.6.2), which uses the science focal plane image as a WFS (e.g., the IFS in Fig. 1.5).

3. More sophisticated RTCs, including
   - gain optimization (Gendron & Lena, 1994), where an integral controller is still used but the gain is constantly changing with time in order to optimize atmospheric rejection and WFS noise attenuation, and
   - predictive control (Poyneer et al., 2007; Guyon & Males, 2017), where previous WFS measurements are used to look for patterns to predict future WFS measurements, improving the temporal response of the $H_{\rm lag}$ and $H_{\rm cont}$ transfer functions from Fig. 1.6.

1.4.3.2 Coronagraphy

If the AO system were able to provide a perfectly flattened wavefront (in reality there are residual atmospheric and quasi-static aberrations remaining; see §1.4.3.2.2 and §1.4.4), the best achievable contrast for an Airy function is illustrated by the blue curve in Fig 1.7 a. Separate from the λ/D resolution limit, this blue curve shows that the Airy pattern also sets a fundamental limit to achievable contrast as a function of separation, motivating the need for a coronagraph. As Illustrated in Figure 1.7 b, a coronagraph is designed to achieve two goals:

1. Occult the central core of the non-coronagraphic PSF (i.e., described by an Airy disk; equation 1.4) using a focal plane mask (FPM), without blocking light from an off-axis exoplanet.

   If no FPM is used, the central core of a (non-coronagraphic) Airy function is over 1000 times brighter than the off-axis "wings" at just a few λ/D. A FPM enables imaging these wings without, e.g., saturating the central core due to limited detector dynamic range. The inner working angle (IWA) of a coronagraph is the radial separation at which 50% of the exoplanet light is transmitted through to the coronagraphic image detector.[7]

2. Improve the contrast using a Lyot stop and/or additional pupil plane optics.

   The use of Lyot stop is the same coronagraph design invented by Bernard Lyot to observe the Solar corona (Lyot, 1939). As illustrated in Figure 1.7 b (viewed from above/below the optical axis), a Lyot stop is a washer-shaped object placed in the pupil plane down stream from the FPM designed to block the light diffracted by the FPM. The apodizer pupil plane upstream of the FPM, included in some coronagraph designs, is also illustrated in Fig. 1.7 b.

By comparing the "no coronagraph" and "Lyot coronagraph" curves in Figure 1.7 a, the two effects of a FPM and Lyot stop are apparent: compared to a non-coronagraphic PSF, the coronagraphic PSF

- is significantly attenuated within the IWA (enabled by the FPM) and

[7]For example, if an exoplanet is separated by 1 au from its host star that is 20 pc away (i.e., 50 mas) and a FPM blocks all the light within 2 λ/D (86 mas for an 8 m telescope at 1.65 μm) from the on-axis host star, even though the exoplanet is separated beyond the λ/D resolution limit the FPM will still block most of its light; instead the FPM IWA governs the innermost separation at which an exoplanet could be detected.

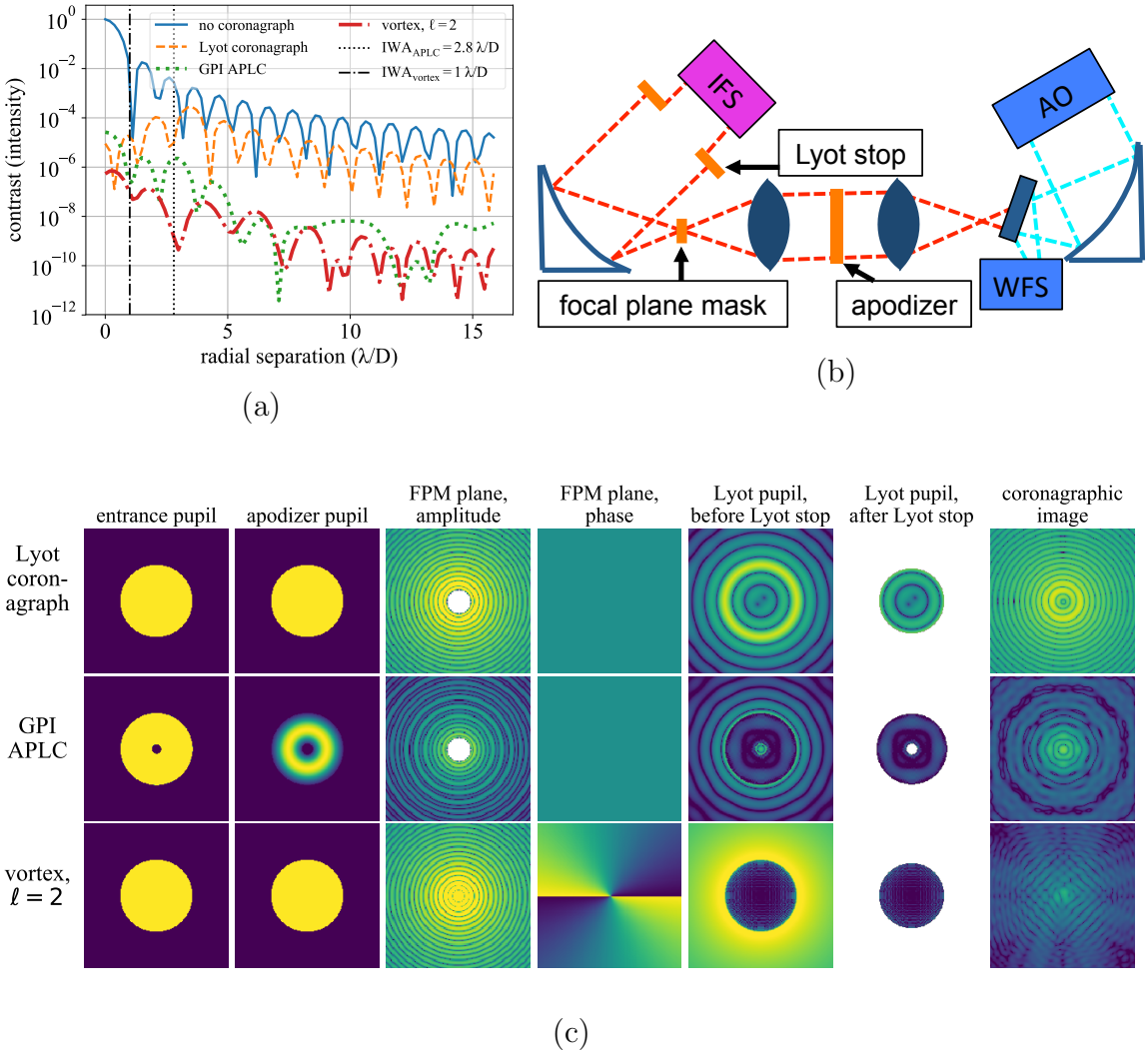


Figure 1.7 (a) Diffraction-limited contrasts (in intensity) from a non-coronagraphic (Airy) PSF (Fig. 1.4), a classical Lyot coronagraph (Lyot, 1939), the GPI apodized Lyot coronagraph (APLC) (Soummer et al., 2006), and a charge 2 vortex coronagraph (Mawet et al., 2005). (b) A geometrical outline of a "three-plane" coronagraph (including an entrance apodizer, focal plane mask, and Lyot stop), similar to the GPI APLC in (a). (c) The electric field intensity (and phase for the FPM plane) in the pupil and focal planes between the telescope entrance pupil and coronagraphic image for the three cononagraphs in (a). Each column is shown on the same relative intensity and spatial scale for comparison between the three coronagraphs. The entrance pupil, apodizer pupil, and FPM phase planes are each shown on a linear scale, while the remaining columns are shown on a log scale.

- reaches a generally lower contrast at separations beyond the IWA (enabled by the Lyot stop).

Many more sophisticated coronagraph designs have been implemented since the original invention by Lyot (1939), some of which do not include any FPM or Lyot stop; however, the unified goal of any coronagraph is illustrated in Figure 1.7 a: attenuate the PSF core within the IWA and improve the contrast beyond the IWA. Two additional optimized coronagraphs—the APLC (Soummer et al., 2006) and the charge 2 vortex coronagraph (Mawet et al., 2005, the "topological charge," ℓ, indicates an azimuthal phase ramp between 0 and $\ell(2\pi)$ radians over 360°)—are shown in Figure 1.7 (a) and (c), illustrating that, with more optimized designs beyond a simple Lyot coronagraph, contrast can be further improved by orders of magnitude. Note that the vortex coronagraph design is particularly relevant to my PhD research and will be discussed later in §3.5.4.2. Many phase mask FPM designs in addition to the vortex have been developed, pioneered by Roddier & Roddier (1997) and Rouan et al. (2000). A detailed review of these coronagraph designs, along with many other "flavours," is presented in Guyon (2007). Modern coronagraph designs are now often developed through numerical optimization procedures (i.e., unlike the earlier proposed analytical functions, e.g., in Mawet et al. 2005), many of which are summarized in Ruane et al. (2018).

1.4.3.2.1 Low Order Wavefront Control As discussed in §1.4.1, at NIR wavelengths λ/D has a physical size of tens to hundreds of microns for a typical optical design of a coronagraphic instrument. The FPMs for most coronagraphs need to be aligned well within this margin, typically $\lesssim 0.1\lambda/D$ (Guyon, 2007). Residual low order WFEs (including tip/tilt) from closing the AO loop are usually insufficient to reach this tolerance requirement for a handful of reasons:

1. Differential chromatic WFEs remain between the visible arm of the AO WFS and the NIR arm of the coronagraph.

2. Differential vibration, drift, and turbulence between the WFS and the coronagraph arm can "trick" the WFS into applying tip/tilt offsets that appear correct with respect to the WFS but are ultimately systematically incorrect with respect to the coronagraph, both on quasi-static (minutes to hours) and dynamic (milliseconds to seconds) timescales.

3. Even if 1 and 2 above are not significant issues, random noise (e.g., detector noise, sky background, photon noise) may limit the AO WFS tip/tilt WFE residuals from achieving sufficient tolerances, meaning a more sensitive WFS is ultimately needed.

For these reasons, most Lyot coronagraph designs typically also require using a low order wavefront sensor (LOWFS) to align the FPM to the tolerances needed to reach sufficient contrast in the coronagraphic image. There are a variety of LOWFS flavours, as summarized in Mawet et al. (2012). A commonality among all designs is that (a) they operate at the same wavelength as the coronagraphic image and (b) are closer to or at the plane of the FPM or coronagraphic image, mitigating points 1 and 2 above, respectively. The issue of LOWFS vs. AO WFS sensitivity to tip/tilt and/or other low order aberrations is more complex, and is still an open area of research; depending on the type of AO WFS (see §1.4.3.1 above) and coronagraph being used, it could be that meeting points 1 and 2 above are sufficient to meet the coronagraph alignment tolerances at the level of achievable WFE residuals. I will explore this topic further in the context of my PhD research in §3.5.5.

1.4.3.2.2 Atmospheric Speckle Lifetime In ExAO Coronagraphic Images

Revisiting the more realistic scenario that AO systems deliver a non-zero closed-loop residual WFE as input into the coronagraph, Fig. 1.8 illustrates the impact of this non-perfect correction on a coronagraphic vs. non-coronagraphic image. For the same input WFE in the entrance pupil, speckles (i.e., uncorrected aberrations in the focal plane image that deviate from the diffraction-limited case; see §1.4.4) are more easily visible in the coronagraphic image. This enhancement is enabled by the lower level of diffraction in the coronagraphic case; if a speckle's intensity in the focal plane, which is set by a corresponding pupil plane amplitude of a given Fourier mode, is at a contrast below that of the diffraction-limited contrast, this speckle will appear "pinned" to the diffraction ring, while for brighter speckles above this level such interference with the Airy intensity pattern will be a second-order effect (Perrin et al., 2003).[8] Because diffraction is attenuated by a coronagraph, un-pinned speckles above the diffraction limit now set the contrast level in a single wavefront realization. However, as the wind moves and evolves phase aberration across the pupil of the telescope, the residual AO wavefront (which is shown for a single instance in time

[8]For example, in Fig. 1.4 this pinned speckle effect is stronger for the amplitude grating and weaker for the phase grating, as the latter generates relatively brighter off-axis spots.

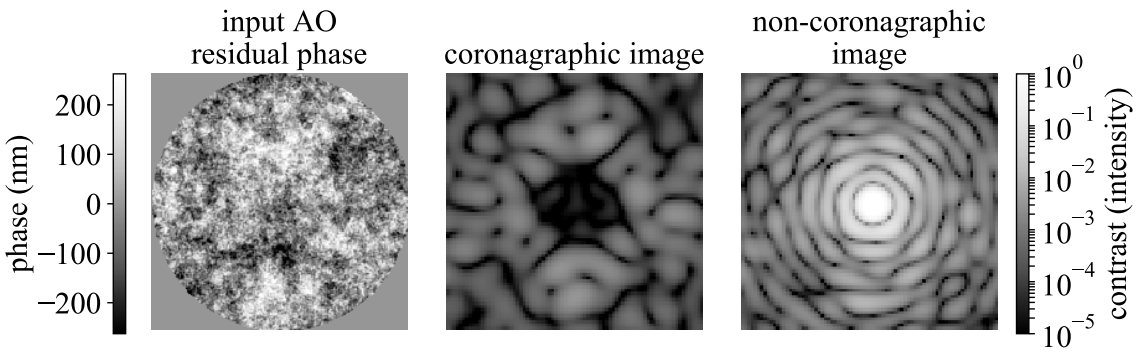


Figure 1.8 An illustration showing the impact of typical ExAO residual WFEs (assuming $\lambda = 1.65$ μm) on a coronagraphic (using the same APLC setup as in Fig. 1.7) vs. non-coronagraphic image. The same input phase realization (left) is used to simulate the images in the middle and right panels, which are both displayed on the same scale. The brighter diffraction rings in an Airy function generate speckles "pinned" to the Airy rings (Perrin et al., 2003); in comparison, these same speckles are un-pinned in the coronagraphic image. Subtracting these unpinned speckles is essential for reaching the diffraction-limited contrast level in a coronagraphic image. Such a subtraction procedure is therefore crucially dependent on the speckle lifetime.

in Fig. 1.8) and the corresponding unpinned speckle pattern in the coronagraphic image is constantly changing. Knowing the coronagraphic speckle "lifetime" (i.e., the length of time before a given speckle in the coronagraphic image evolves into a new uncorrelated realization), is thus of critical importance for designing an efficient speckle subtraction system (through active DM correction and/or post processing).

Unlike the atmospheric coherence time (τ_0; §1.4.3.1), which is typically on the order of a few milliseconds, Macintosh et al. (2005) and subsequent papers have shown that speckle lifetimes are typically on the order of a few hundred milliseconds. This discrepancy is due to the relationship between the electric field intensity of a speckle in the coronagraphic focal plane and its corresponding pupil plane phase aberration (i.e. a single Fourier mode in the pupil). A pupil plane sine and cosine wave (both at the same spatial frequency and amplitude) will generate PSF copies with identical intensities in the coronagraphic image (if the amplitudes are sufficient to generate unpinned speckles, as is the case for the phase grating in Fig. 1.4); the phase difference between the two will be"hidden" in the phase of the focal plane electric field and is not detectable by a standard imaging camera. However, if the sine/cosine amplitude instead generates pinned speckles (from applying a lower amplitude in the entrance pupil and/or switching from a coronagraphic to a non-coronagraphic imager), the

phase of these speckles *will* interfere will the diffracted light, changing intensity accordingly. Based on the typical level of ExAO-corrected phase residuals (∼100 nm RMS; Poyneer et al. 2016), Macintosh et al. (2005) showed that the coronagraphic speckle lifetime, independent of spatial frequency, is set by the telescope pupil atmospheric crossing time: for $D = 8$ m, a windspeed of v=10 m/s, and accounting for wavefront evolution (i.e., the atmospheric wavefront should be completely decorrelated after a full pupil crossing), coronagraphic speckle lifetimes are about 200-600 ms, depending on location in the image and other closed-loop AO parameters. For an AO residual speckle subtraction procedure done only in post-processing, these results illustrate that integrations must be taken on less than or equal to these ∼few hundred millisecond timescales, which I will revisit in application to my research in §3.5.2. However, for active wavefront control, also knowing the phase of these speckles is essential to enable destructive interference using a DM, setting stricter requirements on frame rate for such a system, which I will also revisit in application to my research in §3.5.3.

1.4.3.3 Integral Field Spectroscopy

An IFS for high contrast imaging is used for algorithmic speckle subtraction (§1.4.4.1) and to ultimately (if sufficient contrast and resolution can be achieved) extract an exoplanetary spectrum (§1.3.2). All high contrast imaging instruments currently use a lenslet-based IFS design, originally proposed and implemented by Bacon et al. (1995). Indeed, the same lenslet array technology used for a SHWFS (§1.4.3.1, Figure 1.6 a) is now used here for a spectrograph; lenslet arrays can be manufactured at the typical ∼tens of microns scale to sample the focal plane of a diffraction-limited 10 metre-class telescope, which is not possible for other IFS designs such as fibre bundles or image slicers. A typical optical design for an IFS of an ExAO system from Larkin et al. (2014) is illustrated in Figure 1.9 and described below.

1. Light transmitted through the Lyot stop of the coronagraph is focused onto a lenslet array, which samples the focal plane into sub-apertures (as opposed to a SHWFS, which samples the pupil plane).

2. The relay optics re-collimate the beam downstream from the lenslet array. A prism (i.e., the dispersive element) is then placed in a collimated pupil plane.

3. Optics downstream of the prism then focus the light onto a detector.

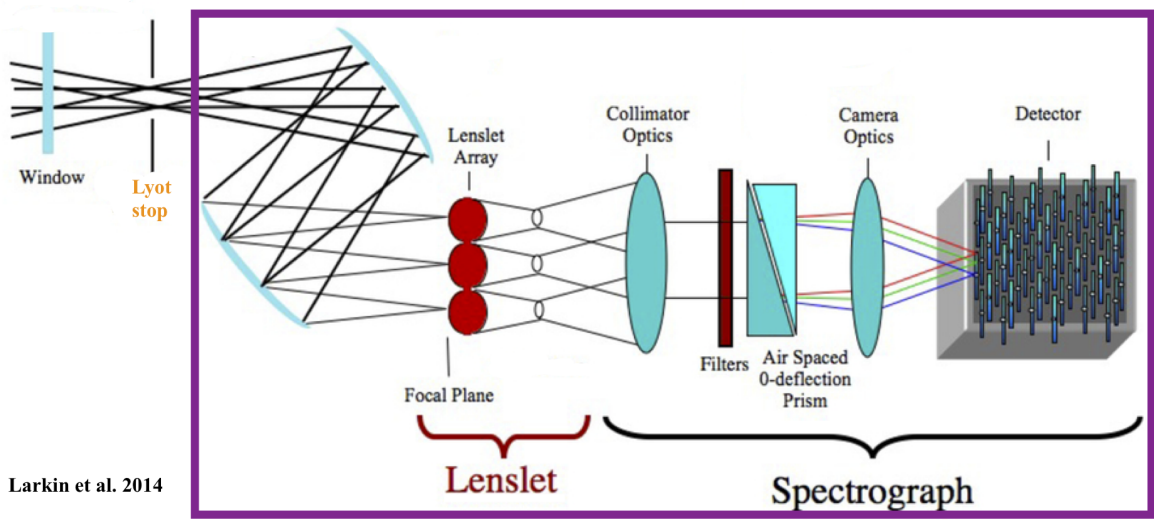
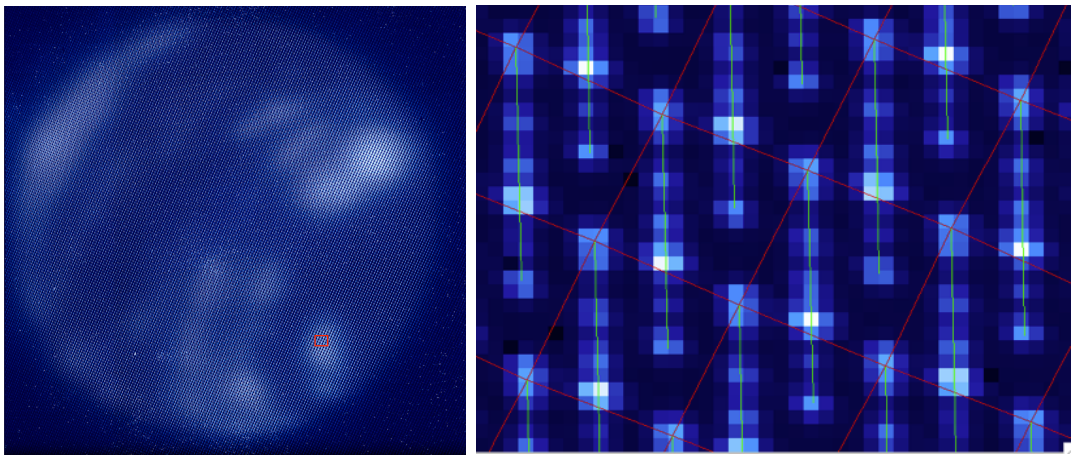


Figure 1.9 An outline of a lenslet array IFS, from Larkin et al. (2014).

A crucial component of the optical design is the position angle of the lenslet array x-y grid relative to the dispersive direction of the prism: the two must be offset from one another by $\sim$5°-20° to prevent spectra from overlapping (e.g., Hinkley et al., 2011; Larkin et al., 2014). Another important design component is the sampling at the lenslet array, which will ultimately set the plate scale in the extracted data cube and should thus be $\geq$2 lenslets/(λ/D). Then, as illustrated in Figure 1.9, a grid of dispersed micro-PSFs is imaged on the detector. Once the image is recorded, a data reduction pipeline (DRP) is required to convert the raw image into a data cube (e.g., Perrin et al., 2014a), providing an image at each wavelength sampled by the IFS in steps typically of one resolution element (see §1.4.3.3.1 below). The spectral resolution (i.e., $R \equiv \lambda/\Delta\lambda$) of a lenslet-based IFS depends on the number of pixels on the detector, the desired wavelength range, and the FOV; it can vary between $R \sim 5000$ (Larkin et al., 2006) and $R \sim 18$ (Groff et al., 2015). Typical ExAO instruments imaging in the NIR with a FOV$\lesssim$ 3 arcseconds require $R \lesssim 100$.

1.4.3.3.1 IFS DRP Extraction Fig. 1.10 illustrates how a datacube is extracted from raw IFS images, showing that spatially-dependent (low resolution) spectral features can be extracted at the diffraction-limited resolution of an 8m-class telescope (in this case, the Gemini S. Observatory). However, the main GPI IFS calibration and data cube extraction procedure (described in Perrin et al. 2014a and papers therein) may cause non-negligible differential chromatic and/or temporal speckle evolution effects. In other words, for a wavefront that is physically achromatic and



(a) (b)

Figure 1.10 (a) A raw GPI IFS H band image of Neptune (a FOV of about 3"×3"), before extraction to a data cube, from (Macintosh et al., 2014). (b) A sub-window (illustrated by the red box) of (a), showing the position (red grid) and wavelength (green lines) solutions, which were obtained from pre-calibrated measurements (Perrin et al., 2014a).

static, the algorithmic DRP calibration and extraction procedure will still decorrelate frames over wavelength and time, respectively. These effects may depend on the relative systematic offsets in calibration of the lenslet PSFs wavelength solutions, overlapping "cross talk" between neighbouring PSFs, robustness of the data cube extraction algorithm, and more. A more detailed analysis of the possible impact of these effects will be presented in application to my PhD research in §2.1 and §2.2.

1.4.4 Speckle Evolution and Subtraction

In addition to diffraction (§1.4.1) and AO residuals (§1.4.3.1, 1.4.3.2.2), additional contrast limitations in coronagraphic images arise from the polishing and reflectivity errors of individual optical elements within a high contrast imaging instrument (as illustrated in the "static aberration" case of Fig. 1.4), which generate (quasi-) static phase and amplitude WFEs, respectively. The "static aberration" case has long been known to be quasi-static in nature (e.g., Marois et al., 2003), evolving on timescales of minutes to hours due to slower thermal and flexure-related changes in the instrument and ultimately requiring daily or often hourly calibrations. All of these effects are illustrated in Fig. 1.11, which shows speckle evolution as a function of time (i.e., stability) and wavelength (i.e., chromaticity).

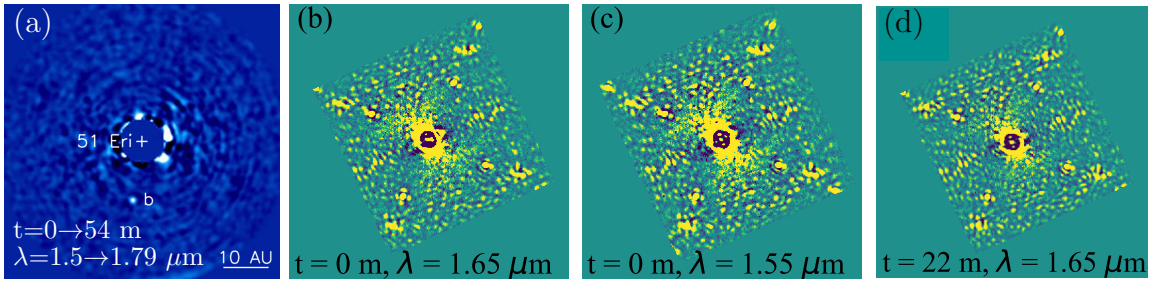


Figure 1.11 (a) The processed image of 51 Eri b (the lowest mass directly imaged exoplanet to date; see §1.3.5), from Macintosh et al. (2015). (b-d) Raw images of the same sequence used to generate (a), illustrating why we cannot yet detect lower mass exoplanets. Images in (b)-(d) are flux-normalized, registered, and magnified to the same spatial scale to align speckles, using the grid of four off-axis PSF copies (which were generated from a pair of amplitude gratings, similar to the illustration in Fig. 1.4). Speckle noise that is both quasi-static (throughout the image) and dynamic (closer to the star and stronger in one direction, related to the wind direction across the entrance pupil) limits the achievable contrast in these raw images. The exoplanet 51 Eri b in panel (a) is not visible in these raw images, which have a contrast at least 100 times worse than the processed image. The evolution of speckles over time and wavelength (stability and chromaticity, respectively)—both of which are viewable as an animations in (b) and (c), respectively, in Adobe Reader version 7 or greater—is what currently limits the achievable final contrast in (a).

This speckle stability and chromaticity from (1) AO residuals and (2) quasi-static aberration are typically mitigated in raw images, such as Fig. 1.11 (b)-(d), as follows:

1. taking long ($\gtrsim$ tens of seconds) exposures with typical $\sim$100 nm RMS ExAO residuals, allowing uncorrelated residual atmospheric speckles to "average out"

over time (§1.4.3.2.2), and

2. using high quality, ∼few nm RMS optics in the common path of the coronagraphic image (e.g., Marois et al., 2008c) and active wavefront control with a LOWFS (§1.4.3.2.1; Mawet et al. 2012).

However, as the above methods are already utilized in the raw images of Figure 1.11 (i.e., where 51 Eri b is not yet detected), additional speckle subtraction methods are necessary to improve contrast beyond the limits set by these approaches. Figure 1.12 further highlights this discrepancy, illustrating the contrast needed to directly

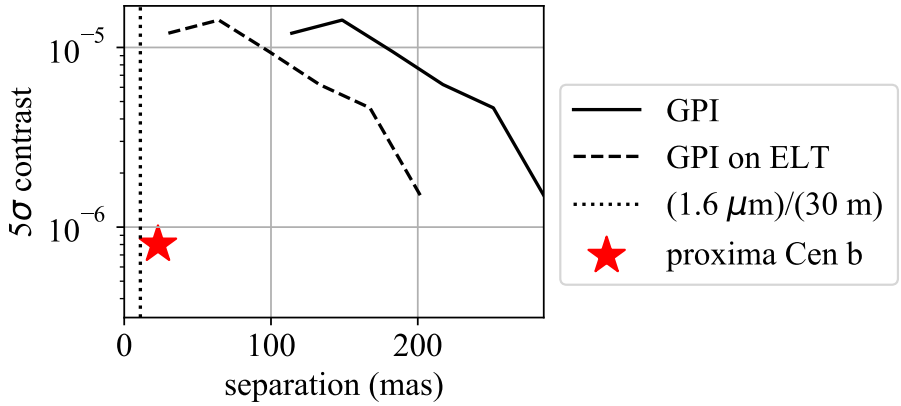


Figure 1.12 A comparison between the deepest contrasts currently reached (after post-processing) at the innermost separations with GPI (Gerard & Marois, 2016b) and the contrasts needed to directly image an Earth-mass HZ exoplanet around the nearest M star, Proxima Centauri, on future ELTs (Guyon, 2011; Anglada-Escudé et al., 2016). Although the current GPI contrast curve would achieve better angular resolution on an ELT, the analagous achievable contrasts would still be about the same as current levels due to the same limitations set by speckle evolution (Fig. 1.11). Thus, although these HZs can be resolved by ELTs, current state-of-the-art technology is still about two orders of magnitude away from reaching the necessary contrasts for these science cases (also see §1.3.6).

image an Earth-mass HZ exoplanet around Proxima Centauri (also see §1.3.6 and §3.5.8) with a future ELT, compared to the final contrasts we would be getting with current state-of-the art instrumentation. In the remainder of this section I will first summarize the standard speckle subtraction methods used in Fig. 1.11 (§1.4.4.1) and then highlight a path forward to enable reaching even deeper contrasts in the future (§1.4.4.2).

1.4.4.1 Classical Differential Imaging

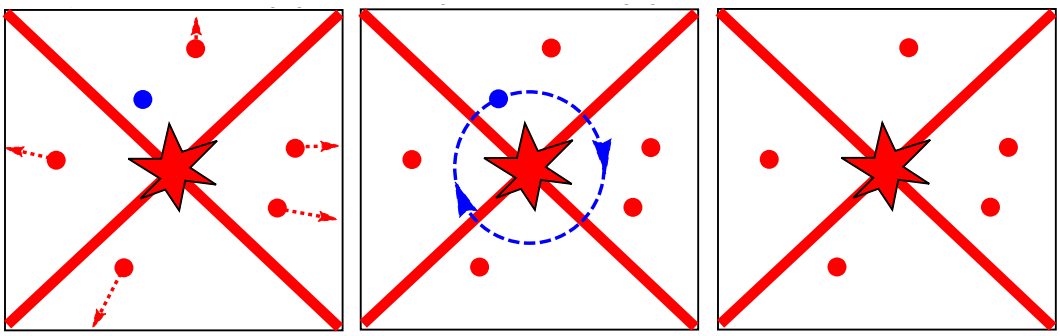
Three main post-processing/observing strategies that are used to further attenuate the speckle pattern in images like Figure 1.11 are described below and shown in Figure 1.13:

1. Spectral Differential Imaging (SDI; Racine et al. 1999; Marois et al. 2000; Sparks & Ford 2002), where coronagraphic images are recorded simultaneously at different wavelengths in either multiple narrow bands or with an IFS. After registering and magnifying the recorded images to a common spatial scale using a reference wavelength, λ_0 (i.e., accounting for radial magnification as a function wavelength), both atmospheric and quasi-static speckles in a target image at wavelength λ_1 can be subtracted by a similar speckle measurement in a reference image at wavelength λ_2. The exoplanet light is not subtracted because radial magnification has sufficiently changed the exoplanet separation at λ_2 from ρ to $\rho|\lambda_1 - \lambda_2|/\lambda_0$ (i.e., to no longer overlap with the exoplanet separation at λ_1) and/or the exoplanet luminosity at λ_2 differs significantly compared to λ_1.

2. Angular Differential Imaging (ADI; Marois 2004; Marois et al. 2006a), where the instrument rotator of an altitude-azimuth telescope is configured to fix the telescope's pupil orientation with respect to the Celestial Sphere while tracking a star over time. This adjustment causes the position angle of an exoplanet to rotate with time, simply due to the rotation of the Earth, while speckles originating from the telescope and instrument remain at a fixed location in the image over the same sequence. For a Cassegrain instrument, this effect is achieved by fixing the instrument rotator position over an observing sequence, which would normally track the sky rotation to cancel out this effect; the same effect is achieved for a Nasmyth instrument by adjusting the pupil tracking speed. Thus, quasi-static (but not atmospheric) speckles in a target image taken at time t_1 can be subtracted from a similar measurement taken at time t_2, while an exoplanet is not subtracted because the varying position angle in between t_1 and t_2, $\Delta\theta$ (in radians), has rotated the exoplanet at t_2 and separation ρ by a distance of $\rho(\Delta\theta)$ from the exoplanet position at t_1. This distance, $\rho(\Delta\theta)$, is sufficient for neither exoplanet PSF to be overlapping in the subtracted image if enough FOV rotation has occurred.

3. Reference Differential Imaging (RDI), where a reference star that generates a

similar speckle pattern but has no exoplanet is observed (or constructed from the target sequence) and then subtracted from the target image(s). This technique, although older in origin (e.g., Smith & Terrile, 1984) has been used for recent high contrast imaging campaigns (e.g., Gerard & Marois, 2016b; Hagan et al., 2018; Currie et al., 2019; Bohn et al., 2019).



(a) SDI (b) ADI (c) RDI

Figure 1.13 An illustration of classical differential imaging observing modes for high contrast imaging. An exoplanet is represented by a blue dot. Speckles from quasi-static aberration and AO residuals, which are "masquerading" as exoplanets, are shown with red dots. Three speckle subtraction techniques are shown: (a) SDI, where speckles are diffracted radially with wavelength while an exoplanet's position remains fixed; (b) ADI, where an exoplanet's position angle changes with time while speckles remain fixed; and (c) RDI, where a separate reference star with no expected exoplanet is either observed or reconstructed to subtract the target star speckle pattern but not the expected exoplanet.

1.4.4.1.1 Least-Squares-Based PSF Subtraction To subtract the speckle pattern in a single region of a single image, the observing strategies in Fig. 1.13 are often combined with more robust subtraction algorithms. These algorithms utilize techniques such as a least-squares (Lafrenière et al., 2007), singular value decomposition (Marois et al., 2010a), and/or principal component analyses (Soummer et al., 2012), as illustrated in Fig. 1.14.

In addition to the process of generating least-squares coefficients for subtracting a region of the target image, Fig. 1.14 also illustrates the principle of a FM, where the same coefficients are also used to model the algorithmic exoplanet throughput during the subtraction. As shown, a selection criterion, or aggressiveness, specifies that reference images be selected beyond a radial and/or azimuthal gap in wavelength and/or time, respectively, from the target image in order to minimize algorithmic

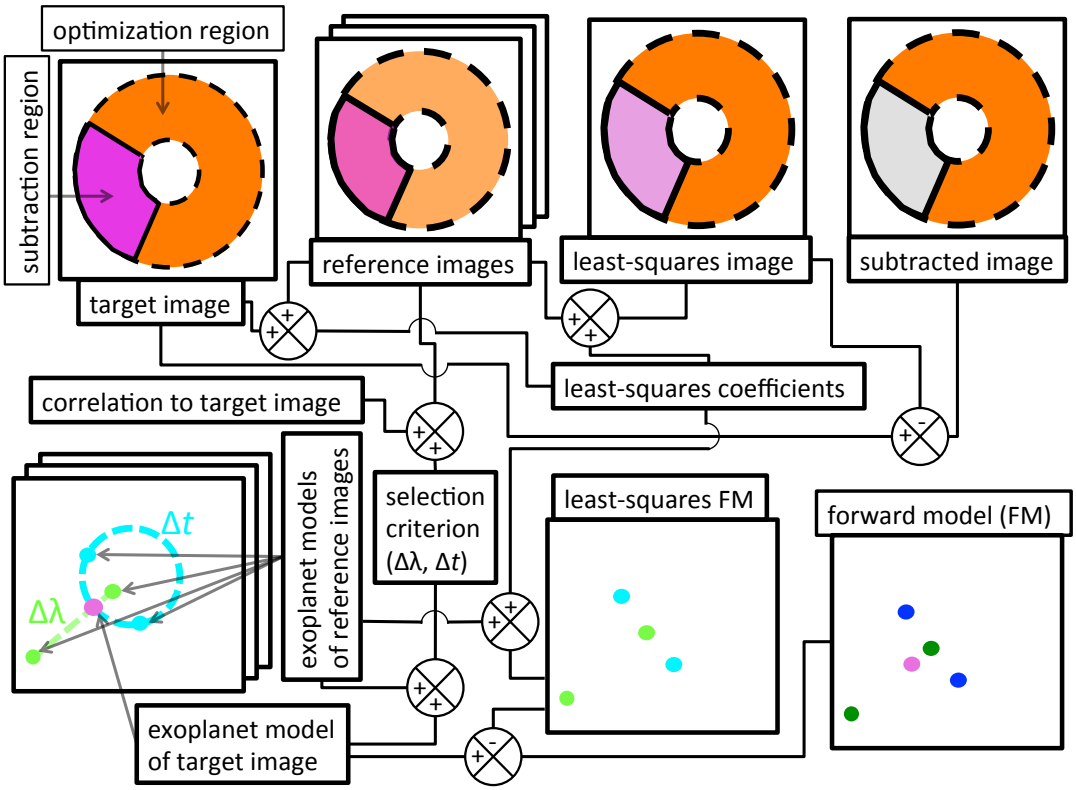


Figure 1.14 A block diagram outlining the steps of least-squares PSF subtraction and forward modelling. The junctions, represented by a ⊗ symbol, combine two boxes (connected to a "+" or "-") and output to a third box. The junctions with a "-" sign indicate that the output box is the difference between the two input images. All other boxes, where the two junctions are "+" signs, represent a more abstract process of combining two input concepts (i.e., an image, a vector of images, a vector of coefficients, or a selection criterion) to produce an output concept. The forward model (FM) inputs, on the lower left of the diagram, include a model of the exoplanet target image and all other reference images in the sequence, both in time and wavelength, including the assumed spectrum of the exoplanet. These models are then used to define a selection criterion (i.e., a chosen minimum value of $\Delta\lambda$ and Δt with which to define usable reference images in the subtraction procedure), which is combined with the correlation of each reference image in the sequence with the target image in order to select the best reference images to use in subtracting the target image. After a few intermediate steps involving the calculation of least squares coefficients, the outputs are illustrated on the right side of the diagram, including a subtracted image along with a noiseless forward model of the exoplanet signal in the subtracted image.

exoplanet throughput losses. For the highest aggressiveness of 1 (using dimensionless units calculated from relative aperture photometry between the target and reference image models), any reference image in the full target sequence, including the target image, could be used. In contrast, the lowest aggressiveness of 0 would not allow any reference images in the target sequence to be selected. Aggressiveness is also illustrated in Fig. 1.15, using high contrast imaging data from this dissertation that will be presented later in §2.2. In this figure, the target image is highlighted in the centre of this region with a correlation of one. Additionally, a range of times and wavelengths near the target image are not selectable because, if used, an exoplanet that is present in the target image would be partially subtracted by the same exoplanet in the reference images. This un-selectable region of Fig. 1.15, which is illustrated in white, would be either larger or smaller for a lower or higher value of user-defined aggressiveness, respectively. The size of this un-selectable region also depends on the separation between the central star and the given subtraction region in the image, the

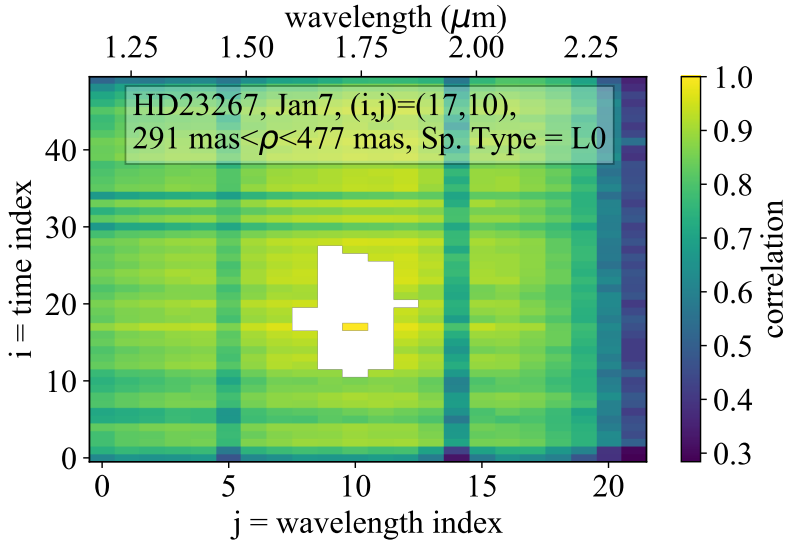


Figure 1.15 The correlation of selectable reference images with a single target image, using broadband (JHK) IFS data from Gerard et al. (2019b, as discussed further in §2.2). The target image is illustrated with a correlation of 1 in the center of the white region. Various correlation trends illustrate the stability and chromaticity of the sequence. Darker vertical lines illustrate wavelength-dependent effects, including the atmospheric water bands at wavelength slices 5 and 14 and a degradation of performance towards the red edge of K band, while darker horizontal lines illustrate time-dependent effects from variable AO and/or quasi-static stability. Each time index has increments of 60s.

assumed exoplanet spectrum, the wavelength, and the parallactic angle relative to the target image. Then, reference images are chosen from the remaining sample based on their correlation, also illustrated in Fig. 1.15. Generally, the most correlated reference images are near the target image in time and wavelength, although this depends on the stability and chromaticity of the sequence, respectively. A more correlated set of reference images (i.e., those with a more similar speckle pattern to the target image) provides a better basis for the least-squares algorithm and will yield a deeper final contrast without changing any other free parameters in the speckle subtraction pipeline (Marois et al., 2014).

1.4.4.1.2 Limitations of Classical Differential Imaging As illustrated by Figures 1.11 and 1.15, SDI and ADI are still fundamentally limited by chromaticity and temporal stability, respectively, and RDI is limited by both, illustrated below in Table 1.2.

Table 1.2: Comparison of limitations from classical differential imaging modes.

method	**limited by chromaticity?**	**limited by stability?**
SDI	✓	✗
ADI	✗	✓
RDI	✓	✓

AO residuals cause an uncorrelated "halo" between long exposure frames that will limit the achievable contrasts with ADI.[9] Fresnel effects from optics in between the pupil and focal plane (i.e., effects that are not described by equation 1.4) cause quasi-static aberrations as a function of wavelength (e.g., Marois et al., 2008c), which limit SDI. Flexure and temperature-related changes as a function of time (e.g., Vigan et al., 2010), also further limit ADI. The only way to overcome these limitations is to (1) subtract speckles on timescales faster than evolving AO residuals and quasi-static speckles (which is not possible with ADI), and (2) subtract monochromatic speckles ($\Delta\lambda/\lambda \lesssim 1\%$) to mitigate chromaticity (which is not possible with SDI). Thus, this fast, achromatic speckle subtraction strategy is not possible with classical differential imaging techniques, as these techniques require using multiple wavelengths (namely,

[9]To first order, this halo should be achromatic and therefore subtractable by SDI. However, to second order, AO residuals contribute additional chromaticity through scintillation and atmospheric dispersion effects (Guyon, 2005)

SDI and RDI) and/or long time periods between usable frames (namely, ADI,[10] and RDI).

1.4.4.2 Coherent Differential Imaging and Focal Plane Wavefront Control

The enormous potential of coherent differential imaging (CDI) and focal plane wavefront control[11] arises from the self-calibrated design of this technique, where by recording one monochromatic coronagraphic image, coherent starlight is used to measure and subtract its own optical aberrations without removing any incoherent exoplanet light. Thus, because there are no requirements on the operational chromatic bandpass and/or exposure time, *CDI is in principle not limited by chromaticity or stability*, opening the door to more efficient speckle subtraction and deeper contrasts in the future. Over the past 15 years, a number of such techniques have been proposed, developed, and tested in the lab (e.g., Guyon, 2004; Bordé & Traub, 2006; Baudoz et al., 2006; Give'On et al., 2007; Serabyn et al., 2011; Sauvage et al., 2012). A few of these pioneering papers that lay the ground work for my research in §3.5 are described below.

1.4.4.2.1 Focal Plane Wavefront Control Guyon (2004) first proposed the use of optical coherence for exoplanet imaging. This main concept involves a Mach-Zehnder interferometer design, where a sub-λ/D spatial filter at the focal plane of the entrance pupil redirects the PSF core into a separate reference channel that subsequently recombines with the main beam, enabling simultaneous interferometric images on two detectors and, through a difference of the two images, a direct measurement of the complex electric field of the star. An important concept of this design is that the electric field of an off-axis exoplanet, which does not pass through the spatially-filtered reference arm, is incoherent with the star, allowing interferometric measurement of the stellar speckle field without any bias from or subtraction of exoplanet light. Such a Mach-Zehnder interferometer design is similar to the current GPI calibration system (Wallace et al., 2010). However, this style of interferometry is now

[10]Note that the time between usable frames with ADI is governed by the rotation rate of the parallactic angle (i.e., the rotation of the Celestial Sphere as observed from a given latitude on Earth) and the separation in the image, but typically requires at minimum a few minutes between frames so that the exoplanet position in a reference frame with respect to quasi-static speckles has rotated sufficiently to not overlap with the exoplanet position in the target frame.

[11]In this dissertation, "CDI" specifically refers to post-processing, while "focal plane wavefront control" refers to active control with a DM. In subsequent sections, for simplicity I will occasionally use "CDI" to refer generically to both techniques unless otherwise stated specifically.

known to be problematic for high contrast imaging due to fast differential vibration ($\gtrsim$ tens of Hz at $\gtrsim$ a few μm amplitudes) between the two interferometer arms that prevents the resolution of any fringes on the detector (B. Macintosh, T. Hayward, private communication), ultimately motivating a common path interferometer design to mitigate these effects (§3.5).

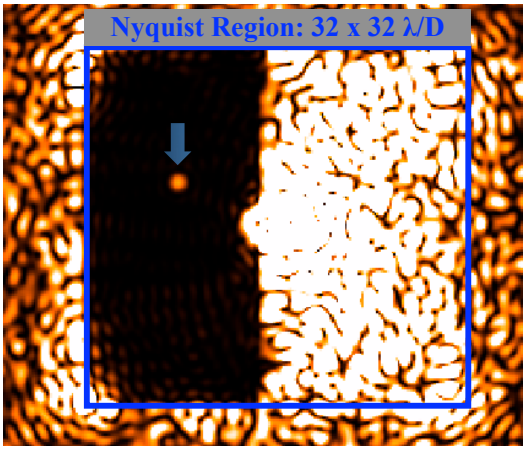


Figure 1.16 An illustration of the fundamental limitations derived by Bordé & Traub (2006): only half of the Nyquist region (here simulating 32 DM actuators across the entrance pupil diameter) is correctable for complex electric field measurement and correction in the low aberration regime with one DM.

A second pioneering paper by Bordé & Traub (2006) has paved the way for all subsequent CDI papers on the fundamental limits of complex wavefront measurement and correction with a DM in the low aberration regime (i.e., the WFE over the entrance pupil of ϕ, from equation 1.4, is less than one radian RMS, enabling a first order Taylor expansion; see §1.4.1). Bordé & Traub (2006) show that a single DM can, at most, correct for half of the nominal focal plane AO control region, also called a dark hole (DH) or Nyquist region,[12] when both phase and amplitude aberrations are present and each are limiting the achievable contrast, as illustrated in Figure 1.16. Although conventional AO (and even current ExAO) systems normally only correct for phase aberrations from atmospheric turbulence, static amplitude aberrations at the $\lesssim 1\%$ RMS reflectivity level (as simulated in Fig. 1.4) reach a quasi-static noise floor at contrasts of $\lesssim 10^{-4}$, comparable to $\lesssim$ tens of nm RMS from quasi-static phase aberration (e.g., Marois, 2004; Gerard & Marois, 2016b). Bordé & Traub (2006) also developed a new ideology towards measurement and DM control of the focal plane electric field known as focal plane wavefront control. Before this seminal paper, the goal of AO systems was to flatten the phase of the wavefront, whereas now the goal of focal plane wavefront sensing and control is to optimize contrast. Bordé & Traub

[12]For a square DM, the size of the area in the focal plane that can be corrected is a square of size $N_{\rm act}(\lambda/D) \times N_{\rm act}(\lambda/D)$ centred around the optical axis, where $N_{\rm act}$ is the number of DM actuators across the entrance pupil diameter (e.g., see Gerard & Marois, 2016a).

(2006) pioneered a new ideology of wavefront sensing and control that drives the DM to minimize the complex focal plane electric field (in half of the Nyquist region, i.e., a half DH), which does not necessarily flatten the wavefront (particularly when correcting for amplitude aberrations and/or diffraction with the DM). This type of "contrast minimization" control algorithm is typically accomplished by equating the derivative of an expression for the complex focal plane electric field (which contains a model of the DM actuator commands, typically called "influence functions") to zero and solving a linear equation for the DM actuator commands (e.g., Bordé & Traub, 2006; Give'On et al., 2007).

Martinache et al. (2014) have also more recently presented the first ground-based on-sky test of the speckle nulling algorithm originally proposed by Bordé & Traub (2006). Due to small discrepancies between modelling and real images in translating coronagraphic images into DM commands, minimizing the focal plane intensity in a region of interest ultimately converged after 20 iterations of five second exposure frames. This iterative process limited contrast improvement to the first bright diffraction ring at 2.3 λ/D, a relatively stable quasi-static feature for these timescales on the order of a minute. Residual atmospheric speckles and quasi-static speckles at higher spatial frequencies were not correctable at these frame rates, ultimately yielding a factor of 3 contrast improvement and motivating the need for shorter exposures and fewer iterations to converge.

1.4.4.2.2 CDI Separate from active DM control, the main goal of CDI is instead to record one or more images while measuring/probing the coherent speckles, subsequently enabling a post-processing subtraction of the speckle field while preserving exoplanet algorithmic throughput. In principle, this approach does not prevent also performing focal plane wavefront control; ideally, these two approaches can be combined to reach deeper limits than either individual approach. However, research and development in CDI has been comparatively unexplored to date. A few developments are summarized below:

- Baudoz et al. (2012a) presented the first CDI processing results, using the Self-Coherent Camera (SCC) technique (see §3.1). With this approach, similar to (Guyon, 2004), a model of a spatially filtered reference beam was used to reconstruct an interference pattern that encodes the coherence information, enabling a contrast gain on laboratory data of up to a factor of 10 at separations less than 15 λ/D (i.e., out to the AO control radius).

- Bottom et al. (2017) demonstrated a CDI approach using phase-shifting interferometry, where one spatially filtered path in a Mach-Zehnder interferometer (again similar to the approach described in Guyon 2004 and Wallace et al. 2010) is advanced through four piston offsets between 0 and $3\pi/2$ radians. This approach was demonstrated on-sky at Palomar leading to the first CDI-based detection of a substellar companion, the brown dwarf HD 49197b. However, due to (1) the requirements of acquiring multiple images to make a single wavefront measurement and (2) hardware that limited this procedure from occurring on millisecond timescales, on-sky wavefront measurements were only acquired on timescales longer than a few seconds. Thus, correction of residual atmospheric and quasi-static speckles that generate temporal instability on shorter timescales was not feasible, ultimately yielding a factor of about 1-4 contrast improvement depending on the post-processing methods that were used.

- Walter et al. (2019) recently proposed a high frame rate statistical sampling algorithm, taking advantage of new NIR photon-counting (i.e., millisecond time resolution) detector technology (Mazin et al., 2012; Goebel et al., 2018); this approach is the temporal analog to the high dispersion spectroscopy + high contrast imaging approach (see §1.3.5), where a static off-axis exoplanet signal should ultimately be distinguishable from time-varying AO residuals and quasi-static aberration. Although it is dependent on a statistical model for these various components, this technique has promising potential, with future planned laboratory and on-sky tests.

- Frazin (2016) and Rodack et al. (2018) have also recently proposed a high frame rate statistical processing approach, where acquiring coronagraphic images at a frame rate synchronized to AO WFS images allows a separate estimate of both the stellar speckles and the exoplanet signal, enabling a subtraction of the former without removing the later. Future tests are also planned for this method.

1.4.4.2.3 Fast Focal Plane Wavefront Control and CDI The current contrast limitations, both in stability and chromaticity, clearly motivate the need to run focal plane wavefront control and CDI on millisecond timescales. With this approach, "freezing" ExAO residuals and quasi-static aberrations to a single wavefront realization will enable speckle subtraction (in post-processing for CDI and/or with a WFS

measurement and DM correction for focal plane wavefront control)[13] on a mainly static, achromatic speckle pattern. However, for focal plane wavefront control, integration of this approach with existing ExAO systems is not a trivial task, as illustrated in Fig. 1.17. Such integration of high order focal plane wavefront control/CDI on millisecond timescales in control of a single and/or multiple DMs has not yet been implemented in an ExAO system, and is still in the research and development phase. In this dissertation I will present my research on this topic (§3.5).

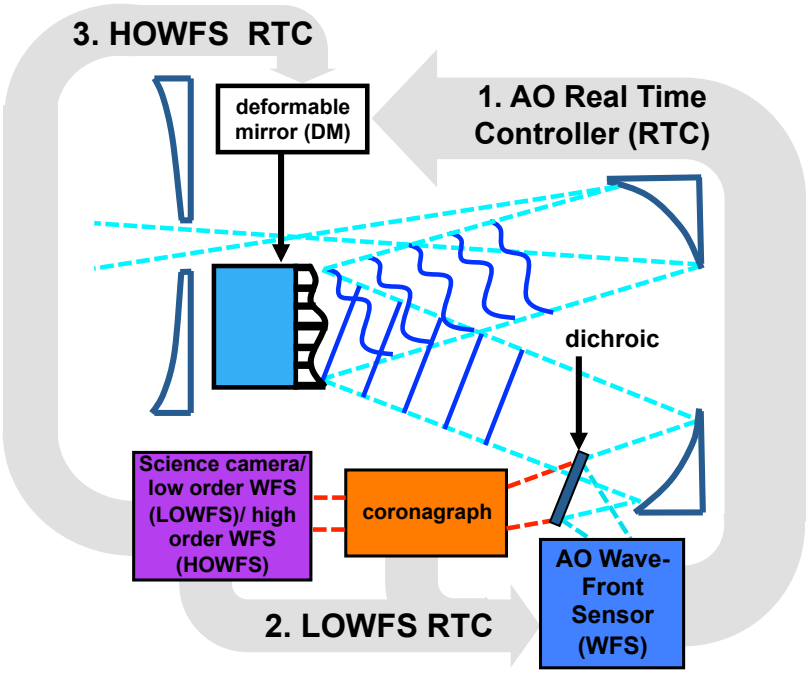


Figure 1.17 A modification of Fig. 1.5, showing the different feedback loops involved in an ExAO + high contrast imaging system that implements fast focal plane wavefront control (i.e., a high order WFS). The LOWFS control is enabled using the coronagraphic image and/or a separate detector (§1.4.3.2.1).

[13]Note that the above-mentioned fast CDI methods proposed by Walter et al. (2019) and Frazin (2016) have been proposed but not yet validated, although neither involve active DM control (i.e., these are post-processing CDI solutions).

1.5 Agenda

In this chapter I have outlined many scientific motivations for ground-based exoplanet imaging, the limitations faced by current instruments in being more sensitive to imaging fainter, closer-in planets, and a path forward for overcoming these limitations. I will continue this theme in presenting my PhD research:

In Chapter 2, I will present my research on speckle chromaticity and stability and the nature of its limitations on current high contrast imaging instruments,

in Chapter 3, I will present my research on speckle subtraction using CDI and focal plane wavefront control, which is designed to overcome the current limitations described in Chapter 2,

in Chapter 4, I discuss future work, and

in Chapter 5, I summarize my work in this dissertation.

In Appendix A, I present my work on modelling of optical communication links, which was completed during an internship as part of my PhD studies.

Chapter 2

Current Limitations: Speckle Evolution

In this chapter I present both unpublished work completed during my PhD studies (§2.1) and work from Gerard et al. 2019b (this section and §2.2).

As discussed in §1.4.4, speckle evolution over time (temporal stability) and wavelength (chromaticity) is currently the main factor limiting current exoplanet imaging instrumentation from reaching deeper contrasts. Overcoming these limitations may ultimately enable future studies of habitable exoplanets (1.3.6). Although significant efforts have been made to understand the impacts of temporal stability and chromaticity from a numerical perspective (e.g., Marois et al. 2008c and Vigan et al. 2010, respectively), a similar comprehensive laboratory and/or on-sky analysis with second generation high contrast imaging instruments that are equipped with an IFS has not yet been completed. Hinkley et al. (2007), and Milli et al. (2016) have carried out temporal analyses on the quasi-static speckle lifetime in coronagraphic images, but with no corresponding chromaticity analysis. Following from the work of Marois et al. (2010a), temporal correlation analyses have informed some new ADI+SDI PSF subtraction pipelines to select reference images ranked by correlation (e.g. Currie et al., 2012; Wang et al., 2015). Marois et al. (2014) presented an initial analyses of time- vs. wavelength-dependent correlation in individual target images. In this chapter I will extend this analysis to multiple target images and configurations; later I will show that such correlation can vary significantly depending on observing conditions and on location in the coronagraphic image.

The main focus of this chapter is thus to present a detailed chromaticity analysis, using laboratory data from GPI in §2.1 and observations from the Subaru adaptive optics system (AO188; Minowa et al. 2010), Subaru Coronagraphic Extreme Adaptive Optics (SCExAO; Jovanovic et al. 2015b), and Coronagraphic High Angular Resolution Imaging Spectrograph (CHARIS; Groff et al. 2016, 2017) in §2.2. In §2.2 I also present new modifications to the ADI+SDI-based template locally optimized combination of images (TLOCI; Marois et al. 2014) PSF subtraction pipeline for my datasets. My SCExAO/CHARIS datasets use a Lyot coronagraph in the broadband JHK mode, covering the full 1.15 - 2.39 μm range at R$\sim$18; this new low resolution broadband mode, currently unique to CHARIS amongst current high contrast imaging instruments, is thought to provide a potential gain in achievable contrast through SDI PSF subtraction for an achromatic system (Marois et al., 2008c, 2014). The GPI will be upgrading its IFS to a similar broadband JHK mode (Chilcote et al., 2018). Thus, my performance analysis and PSF subtraction methods in this chapter are crucial to understanding the impacts of chromaticity in using this broadband mode.

2.1 GPI Chromaticity

Returning to the limitations of wavelength-dependent speckle evolution, chromaticity simulations of atmospheric and quasi-static speckles (Guyon 2005 and Marois et al. 2008c, respectively) have shown that the difference between two wavelength slices within a given bandpass, as in Fig. 1.11c (after registration and magnification to align speckles as a function of wavelength), should be many orders of magnitude below the raw contrast achieved in the individual images. For quasi-static speckles, these simulated limits are illustrated in Fig. 2.1, which shows the results of end-to-end GPI simulations, as described in Marois et al. (2008c). These simulations utilize the Fresnel propagation code PROPER (Krist, 2007), where the chromatic effects of the position and aberration on each individual optical surface (most of which are in between the pupil and focal planes) are taken into account (also see appendix A.3 for a more detailed description of Fresnel propagation, where I use the same PROPER software for a separate application). The "single difference" dashed line in Fig. 2.1 shows that when subtracting two "raw" GPI images at two different wavelengths (after aligning to the same spatial scale), Fresnel instrument chromaticity limits contrast gain to a factor of about 100. In the remainder of this section I will present the data I acquired (§2.1.1) and analysis I performed (§2.1.2) to test these predictions.

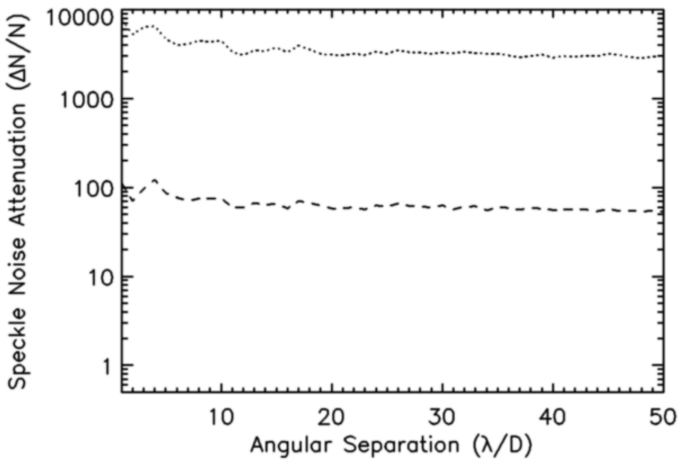


Figure 2.1 Figure from C. Marois (private communication), showing with the dashed line the achievable contrast gain after subtraction of two monochromatic PSFs 1.515 and 1.570 μm. The additional dotted line is not relevant to the analysis in this dissertation. Simulated images to used to produce these results assume a 50 nm RMS entrance pupil phase error, no image magnification with wavelength, and the chromaticity Fresnel propagation setup from Marois et al. (2008c).

2.1.1 Data Acquisition and Processing

I acquired and processed a dataset of off-sky GPI images, obtained using the internal light source on June 13, 2017, in order to compare measurements with the simulations in Fig. 2.1. The light source was adjusted to maximum intensity without saturating the SHWFS (datasets were still acquired in closed loop to stabilize the beam), enabling a high S/N detection of speckles in the coronagraphic image relative to the detector background and photon noise. A sequence of images was then acquired using the H band FPM, Lyot stop, and bandpass filter, and a series of dithering patterns and apodizer masks. Data were acquired using the J, H, and K1 apodizer masks; for each mask, images were acquired in a sequence of four dithers, using pointing and centring mirrors located downstream of the FPM. For a single extracted slice of the IFS data, the apodizer sequence for a single pointing is shown in Figure 2.2, and the dithering sequence for a single apodizer is shown in Figure 2.3. For a given apodizer, the dithering sequence was acquired to attenuate any detector/flat field effects that could potentially modify the measured chromaticity of a given speckle, and Fig. 2.3 clearly shows that such effects are present in the data. Different apodizers were used to measure the impact of this individual optical surface on chromaticity, particularly as the only transmissive optic upstream of the FPM. Note that the apodizer pattern

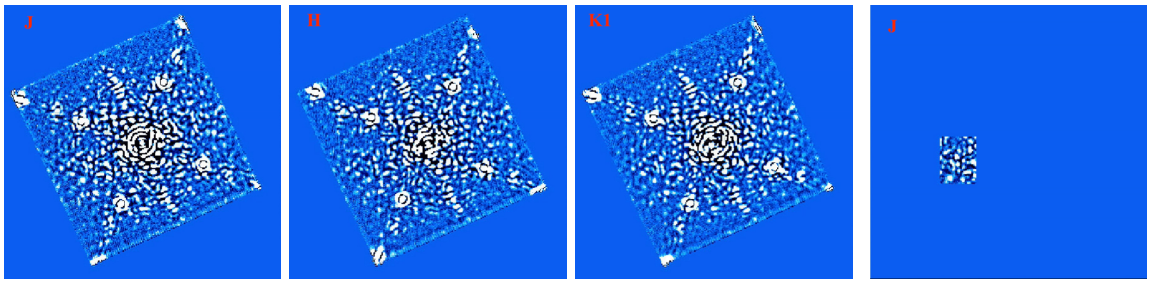


Figure 2.2 Sequence of data acquired using the J, H, and K1 apodizers (labeled accordingly), all using the H band FPM, Lyot stop, and bandpass filter, shown for slice 17 of 37. The right panel shows an isolated region of the images used for future analysis. The left and right panels are viewable as an animation in Adobe Reader version 7 or greater.

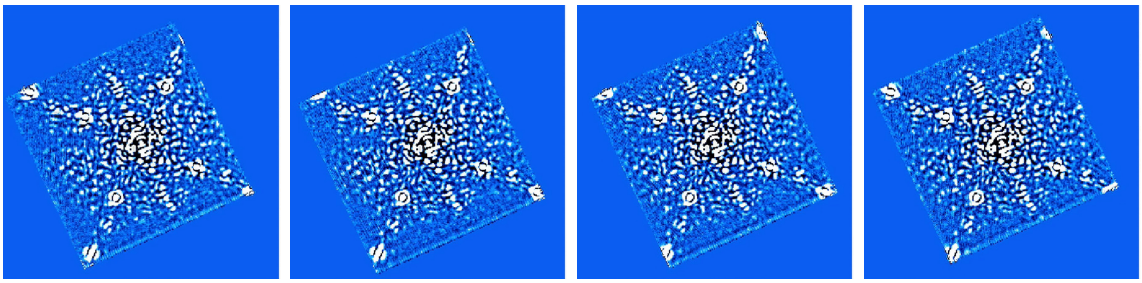


Figure 2.3 Sequence of dithering positions, acquired using the nominal H band coronagraph configuration, shown for slice 17 of 37. The left panel is viewable as an animation in Adobe Reader version 7 or greater.

for the J, H, and K1 apodizers are the same, just each with a different anti-reflective coating (Soummer et al., 2009), and so the diffraction level should in principle be at the same contrast for all three apodizers. Ignoring the first and second order satellite spots (Wang et al., 2014),[1] as illustrated in the right panel of Fig. 2.2, some speckles are clearly changing with the different apodizer masks, while others remain relatively static, warranting further investigation into specific effects from this optic. The region illustrated in the right panel of Fig. 2.2 will be used for future analysis.

After recording the above-described sequence, the raw data was reduced and extracted into data cubes using the GPI DRP (Perrin et al., 2014b). Using the satellite spots, each image was then registered, magnified to align speckles, and flux normalized. Images were then filtered with a 3×3 pixel median boxcar to remove bad pixels. The images displayed in Figures 2.2 and 2.3 are all processed according to

[1]The amplitude grating on the apodizer mask generating these spots was not clocked to the same position angle for the different masks, biasing the eye away from other evolving speckles.

this procedure.

2.1.2 Analysis

First, I calculated a direct comparison with the predictions in Fig. 2.1. Slice 4 (at 1.52 μm) and 10 (at 1.57μm) are the closest to the 1.515 and 1.570 μm simulated in Fig. 2.1, respectively. Using the H apodizer sequence, medianed across the four dithers, I found that

$$\frac{\sigma\{(\text{slice } 4)\}}{\sigma\{(\text{slice } 4) - (\text{slice } 10)\}} = 4, \tag{2.1}$$

where $\sigma\{\}$ is a standard deviation operator over the region illustrated in the right panel of Fig. 2.2. This result is more than an order of magnitude off from the predictions in Fig. 2.1, suggesting that either (1) there are significantly stronger physical chromaticity effects present in the GPI instrument that were not simulated in the models from Marois et al. (2008c), and/or (2) there are additional systematics, other than speckle noise, limiting the achievable contrast gain, warranting future investigation on both of these possible limitations (see below).

Next, I reconfirmed the above results from equation 2.1 in Fig. 2.4. Note that although slightly better results are obtained for a narrow range of slices in calculating the median slice, using an infinitely narrow range (i.e., a single slice) would provide a contrast gain of infinity at that slice and adjacent slices (microspectra are highly correlated across adjacent slices, as there are only $\sim$20 pixels across the full spectrum extracted into 37 slices; Perrin et al. 2014b), thus less physically representative of true chromaticity effects. Clearly, in general, the gains measured in Fig. 2.4 and equation 2.1 are more than an order of magnitude less than the predicted in Fig. 2.1.

At the time of this writing (February 2020), factors causing this apparent chromaticity degradation remain largely untested (but see §4.1.1). In addition to unanticipated chromaticity generated from individual optical surfaces, many possible DRP calibration and extraction errors could also be contributing to the decorrelation. These and other possible effects are further discussed below in §2.2.2.4 in application to my SCExAO chromaticity analysis, and may also be more generally common to all ExAO lenslet-based IFS instruments. Regardless of the source of this observed chromaticity, the effects are likely limiting contrast gains by SDI subtraction (which I will confirm in §2.2.3), warranting further analysis, solutions, and strategies to mitigate these effects.

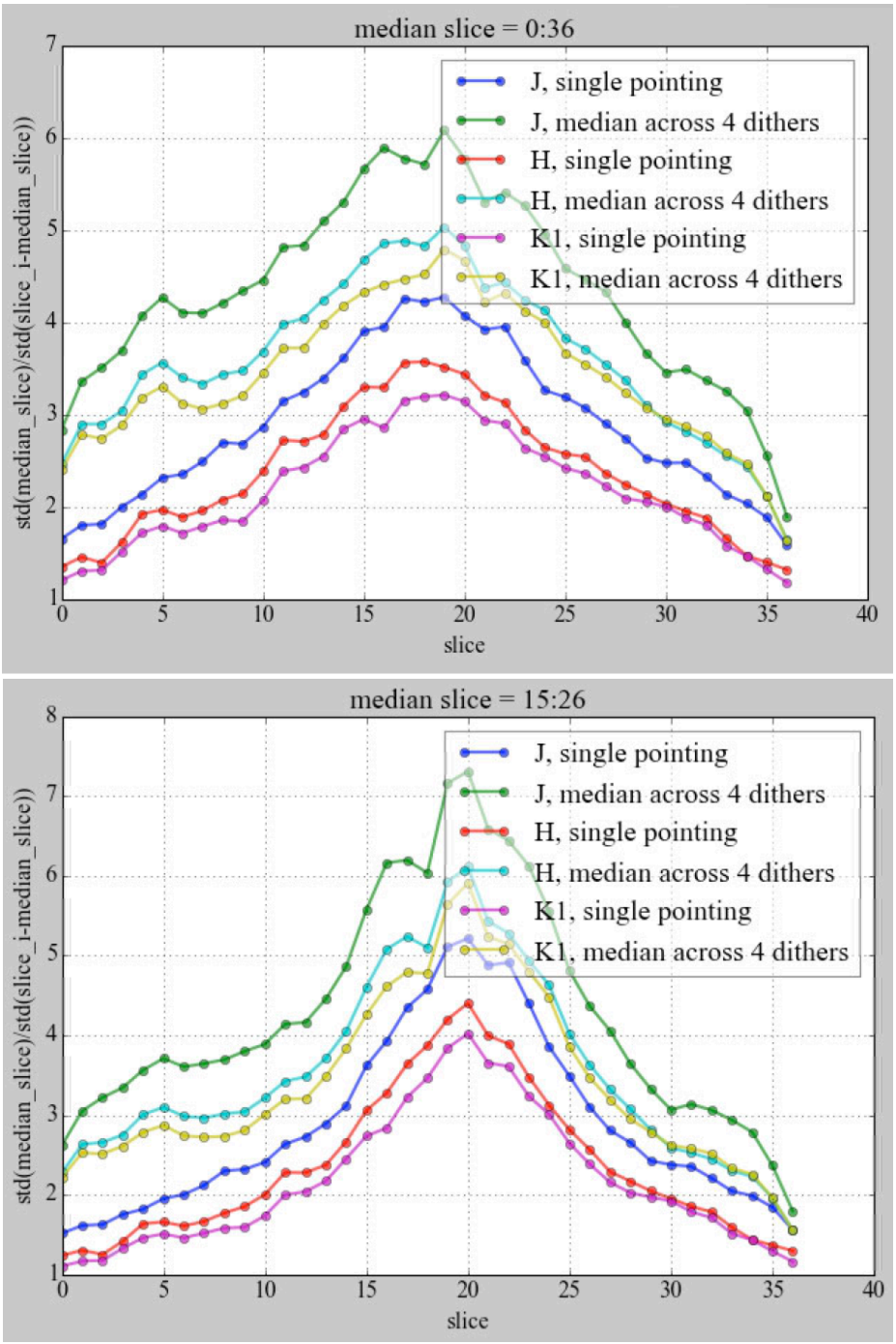


Figure 2.4 Contrast gain measured across the H band for the different data configurations: the J, H, and K1 apodizers, and with or without a median of the four dither positions for a given apodizer. "std" (on the y-axis) is analagous to $\sigma\{\}$ in equation 2.1. The "median slice" is a median stack across a given number of slices in the data cube, defined in the title of each panel.

2.2 SCExAO Chromaticity

The structure of this section is as follows: in §2.2.1 I present details of my target observations, in §2.2.2 I present the architecture of my PSF subtraction pipeline (§2.2.2.2 and §2.2.2.3) and then use the same reference image selection and correlation calculation procedures to carry out a detailed chromaticity analysis (§2.2.2.4), and in §2.2.3 I present final contrast curves and discuss instrument performance implications for the future. Lastly, using these final contrast curves, in §2.2.4 I present a mass upper limit analyses to constrain the parameter space in which possible exoplanetary companions could exist in these systems.

2.2.1 Observations

Name (HD)	Sp. Type	Age (Myr)	Distance (pc)	Warm Disk (arcsec)	Cold Disk (arcsec)
70313	A3V	200	50	0.17	1.48
98673	A 5-7V	737	83	0.06	0.54
109085	F2 V	1380	18	0.15	6.26
125162	A3 V λ Boo	313	30	1.1	7.01
141378	A5IV	478	53	0.25	1.7
23267	A0	60	140	0.030	0.16
38206	A0V	30	77	0.099	1.2

Table 2.1: Target parameters. The ages and temperatures are used to derive a distance and the disk component locations using relations from Wyatt (2008). The results of the distance derivations are taken from Kennedy & Wyatt (2014), who also provide the temperatures to calculate the disk component locations.

Utilizing open use PI time (via Gemini-Subaru exchange time) in 2017A (Gerard, 2017a), I observed a handful of young, nearby two temperature-component debris disks, the extrasolar analogues of the Asteroid and Kuiper belts in our Solar System (Wyatt, 2008, see §1.3.4). The physical properties of these targets are shown in Table 2.1, including derived locations for the warm and cold debris disk components. I identified these targets in Kennedy & Wyatt (2014) as those had not yet been observed by GPI or Spectro-Polarimetric High-contrast Exoplanet REsearch instrument (SPHERE; Beuzit et al. 2019) and/or are inaccessible to either for being too far north. I also observed two additional targets in 2018A through ancillary time (Hodapp, 2018). Note that many of the disk component locations in Table 2.1 are

well matched to the typical FOV for that 8 m-class high contrast imaging instruments ($\lesssim$ a few arcseconds) to be spatially resolved. Although this does not place constraints on the contrast or surface brightness needed for a detection, this more generally motivates need for high contrast imaging observations of these targets.

The observing parameters for these targets are listed in Table 2.2, including dates of observations, exposure times, parallactic angle (i.e., FOV) rotation, and optical seeing. Note that as indicated the seeing levels on UT 9/4/17 varied between 0.6 and 1 arcseconds, with target-specific values not recorded throughout the night. Other telemetry data, such as H band Strehl ratio (SR), wind speed, and/or τ_0, was not acquired for these observations. All observations were made in the standard pupil tracking mode for ADI. The coronagraphic setup in this broadband mode utilized the standard Lyot coronagraph, applying the 217 mas diameter FPM, for which the IWA is 113 mas; at the time, no other coronagraph options offered sufficient performance over the full broadband mode of CHARIS. The DM waffle spots (Marois et al., 2006b; Sivaramakrishnan & Oppenheimer, 2006; Jovanovic et al., 2015a), or "satellite spots," are applied with a non-standard amplitude for the April 2017 data, and so instead an on-sky non-coronagraphic data sequence (using a neutral density filter but with the same Lyot stop) is taken to measure the star-to-spot ratio directly (see §2.2.2.1). The January 2018 data uses the standard 50 nm amplitude spots and corresponding pre-computed star-to-spot ratio as in Currie et al. (2018a).

UT date (d/m/yy)	Name (HD)	m_V	$t_{\rm exp}$ (s)	$n_{\rm frames}$	$\Delta_{\rm PA}$ (deg)	Seeing (arcsec)
9/4/17	70313	5.51	15	107	25	0.6-1
9/4/17	98673	6.43	31	91	22	0.6-1
9/4/17	109085	4.3	13	93	18	0.6-1
9/4/17	125162	4.18	20	93	21	0.6-1
9/4/17	141378	5.53	31	70	21	0.6-1
7/1/18	23267	6.88	60	50	32	0.5
8/1/18	"	"	"	40	24	0.7
8/1/18	38206	5.73	31	118	27	1.4

Table 2.2: Observing parameters, including apparent V band magnitude (m_V), exposure time for individual frames ($t_{\rm exp}$), number of frames taken ($n_{\rm frames}$), the FOV rotation ($\Delta_{\rm PA}$), and the optical seeing values.

2.2.2 Data Analysis

2.2.2.1 Data Cube Extraction and Setup

I used the CHARIS DRP (Brandt et al., 2017) to generate data cubes from recorded IFS detector images, which uses a least-squares method to extract the flux in each slice of the data cube (Draper et al., 2014). A background subtraction, flat field correction, and bad pixel mask are also applied in the extraction process as described in Brandt et al. (2017). All images are then prepared for subsequent PSF subtraction (§2.2.2.2) and correlation analysis (§2.2.2.3 - 2.2.2.4). The satellite spots are used for image registration, magnification with wavelength to align speckles, flux-normalization to remove the wavelength dependence of the broadband transmission filter and stellar spectrum, and contrast calibration. I chose to magnify the images with wavelength centred around slice 16 of 22 ($\lambda_0 = 2.0\mu$m). Slice 16 was chosen instead of slice 11 ($\lambda = 1.7\mu$m; in the middle of the bandpass) because of wavelength-dependent sampling issues. The CHARIS PSF in J and H band is slightly under-sampled, and I found that cubic spline interpolation, used to rotate and magnify the images for my ADI+SDI PSF subtraction algorithm (§2.2.2.2), provided a more accurate centroiding precision from numerical interpolation in the K band. The four satellite spots are all separated by 15.91 λ_0/D from the star, and so measured satellite spot positions in the registered, magnified images are used to calculate a plate scale for contrast curve measurements (§2.2.3). I was not able to use any on-sky astrometric data to validate these derived plate scale values. Images are then high-pass-filtered using a 4.1×4.1 λ_0/D median boxcar filter to attenuate the smooth AO halo while retaining a close-to-unity algorithmic throughput of λ/D-scale static speckles and any possible exoplanet(s). The λ^2 dependence of the satellite spot intensity is also removed during the flux normalization procedure. The satellite spots are omitted from reference image selection in §2.2.2.2 and correlation calculations for analysis in §2.2.2.4. Finally, all registered, magnified, flux-normalized, high-pass-filtered images are converted into "raw" contrast units (i.e., the pixel value-to-star ratio throughout the whole image) using measured and pre-calibrated star-to-satellite spot ratios, whose data acquisition is described at the end of §2.2.1. The same registration, magnification, flux normalization, and high-pass filtering procedure is applied to the custom April 2017 star-to-spot calibration data before computing the ratios.

2.2.2.2 PSF Subtraction

My ADI+SDI PSF subtraction algorithms build on the methods described in §1.4.4.1. These subtraction algorithms originate from the locally optimized combination of images (LOCI; Lafrenière et al., 2007) algorithm, further developed by Marois et al. (2010a) and Marois et al. (2014) as TLOCI. An overview of the procedure used in my TLOCI pipeline to subtract a single region of a single target image, including the use of a FM, was discussed in §1.4.4.1.1 and illustrated in Fig. 1.14. Fig. 1.15 and corresponding text in §1.4.4.1.1 also illustrated the concept of correlation-based reference image selection and aggressiveness as input parameters into my PSF subtraction pipeline. Additional details are described below about the setup and new modifications of my TLOCI pipeline.

To help define, identify, and remove uncorrelated images from each dataset, I will introduce a new free parameter into my PSF subtraction pipeline called a correlation cut. For a single target image, the value of a correlation cut will define the minimum level of correlation that is rendered acceptable in a reference image used to subtract the target image speckle pattern. A given target image/region will be discarded from the subtracted sequence if there are not at least $N_{\rm ref}$ reference images above the correlation cut value, all of which must pass a selection criterion, where $N_{\rm ref}$ and aggressiveness are also free parameters. For example, if correlation cut = 0.5 and $N_{\rm ref}$ = 10 but there are only five reference images that are greater than 50% correlated to the target image using a moderate aggressiveness, the target image/region will be discarded from the sequence. The target image would be discarded in this case even if, e.g., one of those five reference images is 99% correlated to the target image. Thus, it is clearly important to use a physically motivated correlation cut so that a maximal number of target images in an observing sequence can be subtracted while still rejecting a maximal number of uncorrelated reference images. Marois et al. (2010a) and subsequent papers have shown that using correlated reference images is crucial to prevent unnecessary noise propagation in the covariance matrix inversion step of a least-squares/principal component analysis subtraction algorithm. The correlation cut values are dimensionless by definition and have a minimum and maximum of 0 and 1, respectively. My subsequent analysis in §2.2.2.3 informs the decision to use the following correlation cut values for the four different radial sections of an image, progressing from the innermost to outermost separations: 0.8, 0.74, 0.67, and 0.59. Although the specific angular separations of these regions varies for each target based on my method

of measuring the pixel scale using the satellite spots (§2.2.2.1), the average separation and standard deviation across all targets is: 104 ± 2 mas $< \rho < 286 \pm 4$ mas, 286 ± 4 mas $< \rho < 468 \pm 7$ mas, 468 ± 7 mas $< \rho < 649 \pm 10$ mas, and $\rho > 649 \pm 10$ mas. A graphical illustration of these zones is illustrated later in §2.2.2.2. Note that I chose to fix the radial separation of subtraction regions based on pixel separation, which is why the derived angular separation varies slightly from target to target. In pixels, the radial separations for each annulus for all targets are $7 < \rho < 19$, $19 < \rho < 32$, $32 < \rho < 44$, and $\rho > 44$.

Based on the definitions and work in Marois et al. (2010a) and Johnson-Groh (2016), for all subsequent analyses in this section I use an L0 or T6 input spectrum,[2] an aggressiveness of 0.5 and 15 reference images per least-squares subtraction. In Gerard & Marois (2016b), I found that optimizing aggressiveness and number of reference images yielded a negligible gain in contrast, and so here I use the same aggressiveness, number of reference images per subtraction, and input spectral types (to define a selection criterion and for subsequent forward modelling). I use a singular value decomposition (SVD) cut-off value for the covariance matrix inversion of 1×10^{-4}, as in Gerard & Marois (2016b), where a maximum value of 1 would set all the least-squares coefficients to zero. Although this is a low value compared to other algorithms (e.g., Currie et al., 2018b), note that there are many degeneracies between the number of reference images used, reference image correlation, and SVD cut-off (Marois et al., 2010a); my approach here uses a lower SVD cutoff but compensates by selecting more correlated reference images.

In a new addition to my TLOCI pipeline, the central region of each subtraction region is defined by both the parallactic angle and wavelength at any given timestamp, and so the location of a single region will both rotate with time and magnify with wavelength over a single observing sequence, similar to the approach in Wang et al. (2015). These definitions are used so that the subtraction regions line up after de-rotation and de-magnification, which allows the generation of a FM map for each target. The optimization regions are defined for each subtraction region as the other two azimuthal regions at the same radial separation. This subtraction and optimization region geometry ensures that the least-squares coefficients are never generated from the subtraction zone where a exoplanet would be. S/N maps are calculated

[2] J02281101+2537380 and J02281101+2537380 from Burgasser et al. 2008 and Burgasser et al. 2004, respectively, subsequently convolved and binned to the appropriate CHARIS spectral resolution; see §1.3.2 for an overview of giant exoplanet spectral types and their evolutionary processes.

using back-rotated images, as in Gerard & Marois (2016b). A few example data products of a single target sequence from my PSF-subtraction pipeline are shown in Figure 2.5, including a subtracted collapsed data cube, corresponding S/N map, and corresponding FM map.

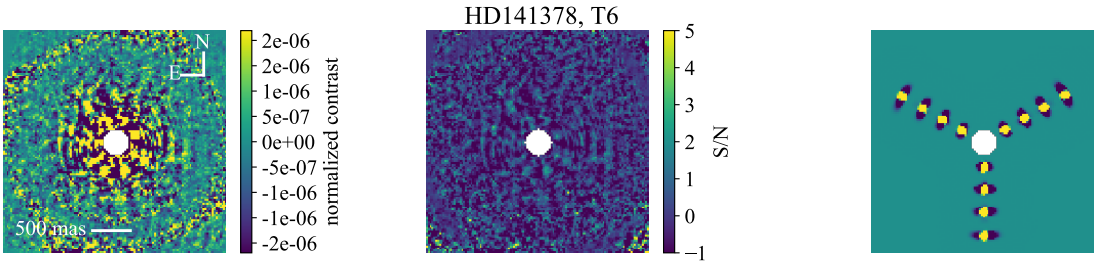


Figure 2.5 Example PSF-subtracted data products for a single target, HD141378, assuming a T6 spectral type. Left: an image of a collapsed PSF-subtracted sequence. Middle: a S/N map of the left image. Right: a collapsed FM map. Note the relative absence of dark shadows in the radial relative to azimuthal direction. This suggests that images at the same wavelength and different time are more correlated than images at the same time and different wavelength, causing the former to be selected as reference images more often than the latter.

For each target I also run a bootstrapping procedure, illustrated in Figures 2.6 and 2.7. As in Marois et al. (2010a) and Marois et al. (2014), I implement a FM algorithmic throughput correction procedure in each subtraction region of the image, including a model for the radial variation of the algorithmic throughput within each region. In addition to the centre, four additional FM throughput corrections using the same least-squares coefficients are computed at smaller and larger separations within each subtraction region. A fifth order polynomial is then fit to the radial variation of throughput correction values to generate a continuous function across the region. As a result of this throughput correction procedure, a bootstrapped planet at any location in the image (i.e., not necessarily at the radial or azimuthal centre of a subtraction region), is extracted at the input value and does not need any additional throughput correction.[3] These results are illustrated in Fig. 2.7. The simulated (i.e., bootstrapped) exoplanets were injected throughout the image, one per subtraction region, such that none lies at the radial or azimuthal centre of the region, as in the

[3]although this correction will break down when the noise distribution deviates significantly from azimuthal symmetry, such as the wind-driven halo/"butterfly" effect (e.g., Madurowicz et al., 2019). However, the apparent lack of these discrepancies in my data, illustrated in Fig. 2.7, suggests that such butterfly effects are negligible here.

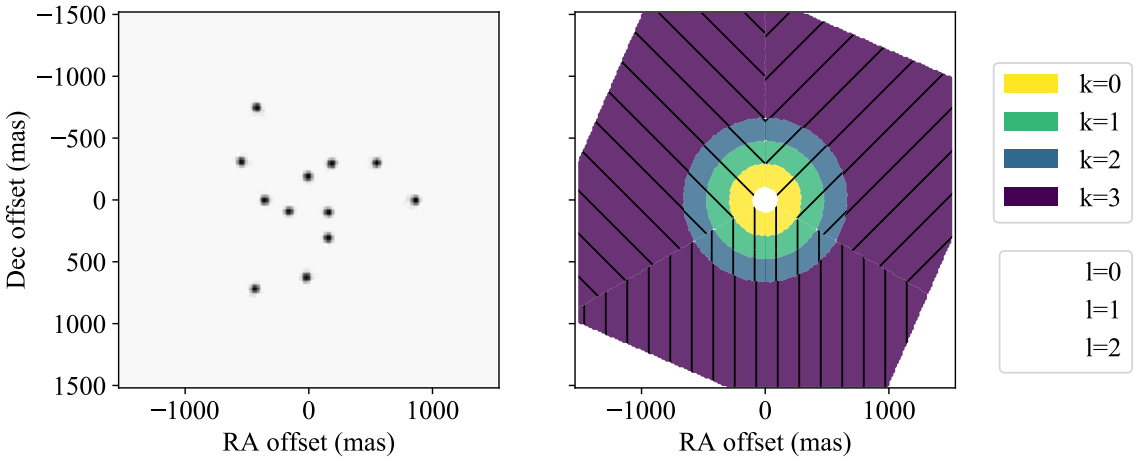


Figure 2.6 Right: definitions of the different subtraction regions defined in the final collapsed image. "k" and "l" represent the radial and azimuthal indices for each subtraction region, respectively. Left: the locations of simulated planets (one per subtraction).

left panel of Fig. 2.6. Almost all recovered fluxes match the input value within their expected scatter, materialized by the 5σ contrast value at the location of the planet.

2.2.2.3 Correlation Cut Analysis

Additional correlation analysis informed my choice of values for a correlation cut, as discussed in §2.2.2.2. Still using the HD 23267 Jan. 7 dataset, frame 17, slice 10 as an example (as in Fig. 1.15), Fig. 2.8 illustrates the trends in correlation of selected reference images at either the same time or same wavelength (i.e., i = 17 or j = 10, respectively). The top two panels show the correlation trends of the 15 most correlated images (as defined in §2.2.2.2, N_{ref}=15) at large separations as a function of time (a) or wavelength (b), using an aggressiveness of 0.5, and using an assumed L0 input spectrum. Uncorrelated frames can be identified as vertical "streaks" in these images (e.g., the water bands at j = 5 and j = 14, as in Fig. 1.15). The chosen correlation cut value at this separation should reject these uncorrelated frames from the sequence in order to prevent unnecessary noise amplification in the least-squares algorithm (Marois et al., 2010a). Because I require that all 15 reference images lie above my defined correlation cut value, the bottom panels of Fig. 2.8 show a one dimensional slice of the top two panels: the correlation of reference image number 14 (i.e., the 15th most correlated selectable reference, using a zero-index counting

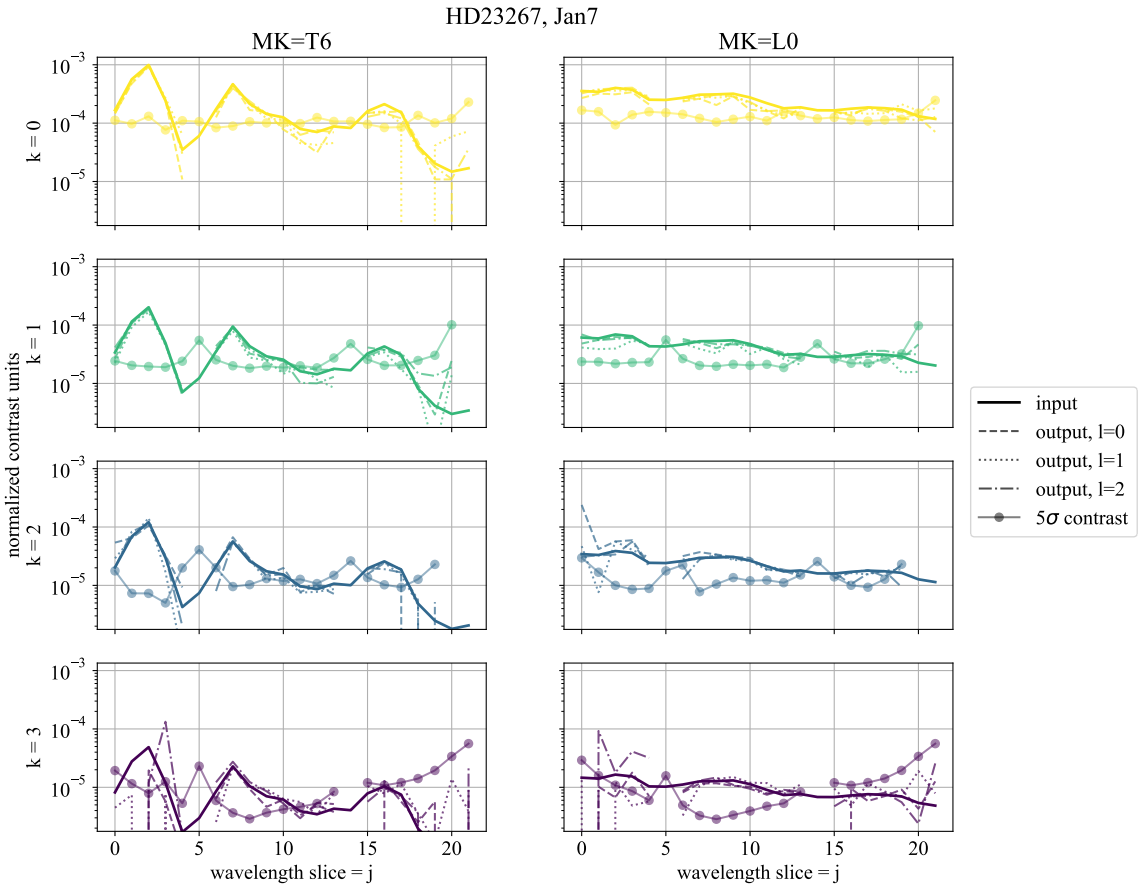


Figure 2.7 The injected and recovered fluxes of the simulated exoplanets illustrated in Fig. 2.6, as well as the 5σ contrasts corresponding to the separation of each simulated exoplanet (calculated from the back-rotated images). The same conventions for "k" and "l" are used here as in Fig. 2.6.

system) as a function of time (c) and wavelength (d) for both spectral types and all separations. Similarly, at all separations and spectral types I would like to discard the frames where there are "dips" in these plots in order to optimally reject uncorrelated images and avoid propagating them through the least-squares algorithm. Fig. 2.8 c and d also suggest that larger separations may be less correlated overall and that a different correlation cut value will be needed for each separation. Note that the low correlation at larger separations may be a result of wavefront decorrelation at those spatial frequencies and/or increased background noise because the speckles at those separations are detected at a lower S/N. At a given separation, using a correlation cut value that is too high will reject too many images at that separation and return only

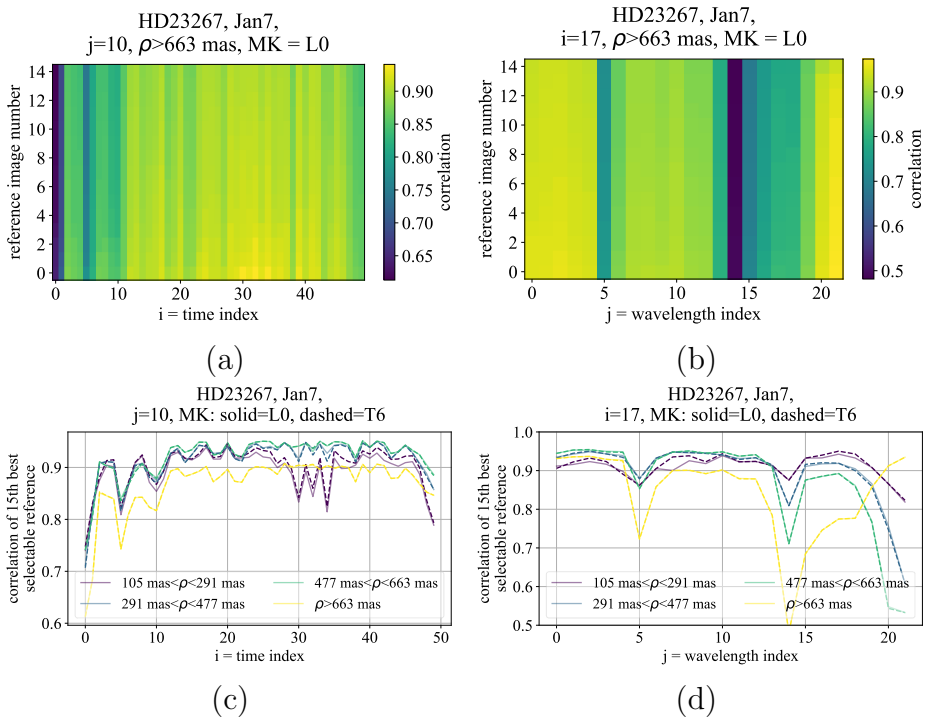


Figure 2.8 (a) and (b): the 15 most correlated selectable reference images (i.e., reference image numbers 0 through 14, using a zero-index counting system) at a single wavelength slice as a function of time and at a single time stamp as a function of wavelength, respectively, both for large separations and assuming a L0 template selection criterion. (c) and (d): The 15th most correlated selectable reference image at all separations and for two input spectral types (c) at a single wavelength as a function of time and (d) at a single time as a function of wavelength (i.e., reference image number 14 as a function of time and wavelength).

a minimal number of subtracted frames to the final collapsed cube; the "optimal" correlation cut value should balance rejecting too many images with propagating too much noise through the least-squares algorithm.

The extension of Fig. 2.8 c and d is to calculate the median behaviour over a full sequence and for all target images. Thus, in Fig 2.9 I compute the median and robust standard deviation at a given separation for the 15th most correlated reference image over all target images (time and wavelength) and input spectral types. Fig. 2.9 illustrates that, for most target images, the 15th most correlated image decreases in correlation with separation, motivating the use of different correlation cut values at each separation. The horizontal dashed lines in Fig. 2.9 illustrate my chosen correlation cut values at each separation. I set these values to generally lie near the bottom of the corresponding 1σ error bars over all targets. Using this

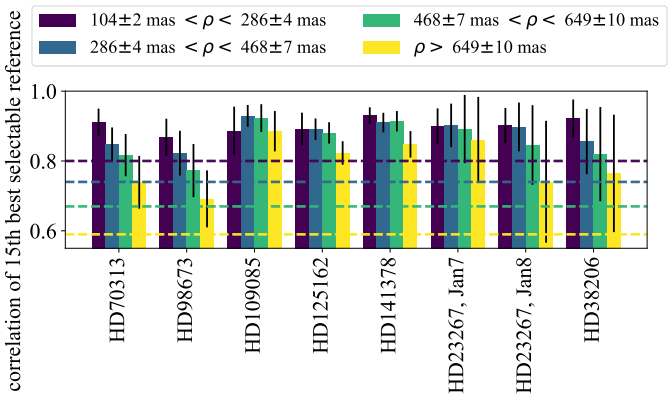


Figure 2.9 The median correlation of the 15th most correlated, selectable reference image and corresponding robust standard deviation over all times, wavelength, and input spectral types, at each separation for each target. My adopted correlation cut values are illustrated by the horizontal dashed lines, colour coded to each separation.

definition to set the correlation cut values will balance the rejection of bad frames relative to the median correlation value without discarding too many frames from the sequence. In general, the optimal correlation cut value will vary for each least-squares subtraction and is degenerate with the SVD cutoff and $N_{\rm ref}$ parameters (Marois et al., 2010a); future work will ultimately require a more optimized approach to balance these degeneracies (e.g. Gerard & Marois, 2016b), but is beyond the scope of this section.

2.2.2.4 Chromaticity Analysis

As discussed in §2, in a perfectly achromatic system, the larger "lever arm" of wavelength coverage over the broadband JHK filter of CHARIS should provide a better algorithmic exoplanet throughput than a narrower bandpass, and so if both cases provide the same level of speckle subtraction, the final throughput-corrected contrast should be better in the former case. However, as discussed in §1.4.4.1.2, in reality broadband systems are not achromatic, and the level of wavefront decorrelation, or chromaticity (i.e., separate from wavelength magnification), will ultimately determine the limits of speckle subtraction via SDI processing (e.g., Marois et al., 2008c). In this section I will measure the chromaticity in my CHARIS datasets, which will then inform the speckle subtraction limits in §2.2.3.

A generalization of Figure 1.15 extends to Figure 2.10, which shows the median correlation of all of my targets as a function of time and wavelength. Fig. 2.10 is

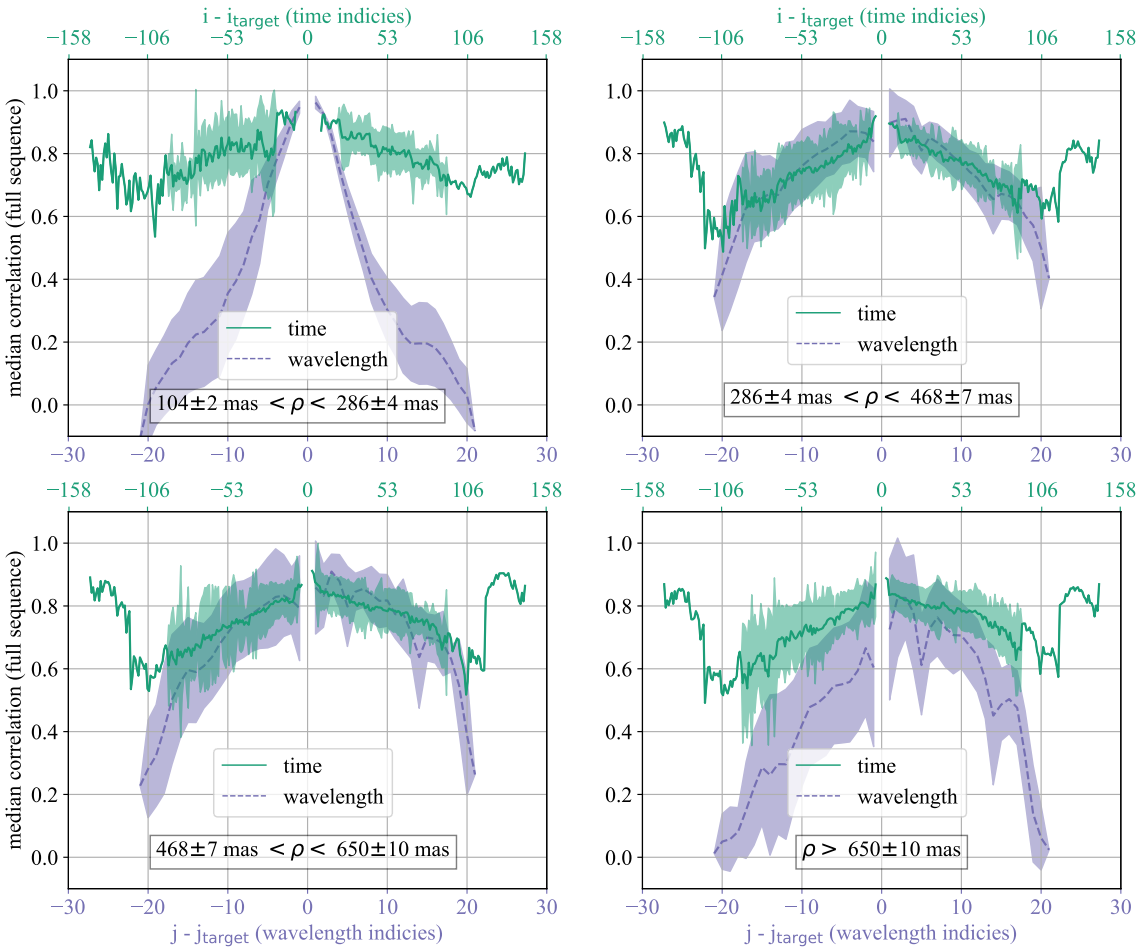


Figure 2.10 The median correlation of selectable reference images over all of my observed targets, separated into time and wavelength. The shaded regions show the robust standard deviation over all targets. Each panel corresponds to a different separation in the image. For every target image, each selectable reference image is identified as either the target wavelength and a different time ($j = j_{\rm target},\ i \neq i_{\rm target}$; top x-axis) or the target time and a different wavelength ($i = i_{\rm target},\ j \neq j_{\rm target}$; bottom x-axis). The main conclusion from this figure is that there is significant chromaticity at small and large separations over the full JHK bandpass; at adjacent wavelength channels SDI may still be advantageous, but this figure suggests that this is not the case for a large wavelength lever arm across the full bandpass (see §2.2.3 for a validation of this hypothesis).

generated as follows:

1. For a single target image at a single time, wavelength, and user-defined separation, spectral type, and aggressiveness, a selection criterion is applied to all of the images in the observing sequence; images that do not meet this criterion are discarded.

2. Correlation of the target image with every remaining image from step 1 is computed, normalized to unity using the standard deviation in both the target image and reference images. Steps 1-2 are illustrated in Fig. 1.15.

3. In order to create a common zero-point for every subsequent target image (i.e., so the central index of Fig. 2.10 is zero for every target image), each correlation value from step 2 is saved in a two dimensional array of size $2i_t \times 2j_t$, where i_t and j_t are the total number of time stamps and wavelength slices, respectively (i.e., so both the negative and positive indices can be shown in Fig. 2.10). The x and y indices that fill a quarter of the total area of the two dimensional array for each correlation value from step 2 are $(2i_t - i_{\text{target}})$ and $(2j_t - j_{\text{target}})$, respectively, where i_{target} and j_{target} are the indices of time and wavelength for the given target image. This index definition ensures that the target image is in the centre, or zero point, of each $2i_t \times 2j_t$ array. Both the region removed by the selection criterion and the other three quarters of the array remain empty and are filled with "not a number" (NaN) placeholders.

4. Steps 1 - 3 are repeated over time and wavelength indices for every target image within a single observing sequence, generating a vector of two dimensional arrays of size $i_t j_t \times 2i_t \times 2j_t$. The autocorrelation of each target image is in the centre of each $2i_t \times 2j_t$ array.

5. A median is taken across the zeroth axis of the vector in step 4 (i.e., the axis of dimension $i_t j_t$), ignoring NaN values, to generate a two dimensional array of size $2i_t \times 2j_t$, representing the median correlation between the target image and every other image in the sequence. A slice along the x direction of this array at $y = j_t$ (i.e., the zero point of the wavelength direction) reveals the median correlation of the target image vs. time, whereas a slice along the y direction of this array at $x = i_t$ reveals the median correlation vs. wavelength.

6. Steps 1-5 are repeated over each user-defined spectral type.

7. Steps 1-6 are repeated over each user-defined separation.

8. Steps 1-7 are repeated for each target, computing the median and standard deviation of correlation over time and wavelength indices for each separation.

The outcome of the above procedure is illustrated in Figure 2.10, whereas steps 1-7 are illustrated in Fig. 2.11, for which I used a separate dataset from my sample, the κ And data from Currie et al. (2018b). Rather than a single target in my sample, the Currie et al. (2018b) dataset was chosen for Fig. 2.11 to understand how the conclusions from Fig. 2.10 change with higher quality AO correction.

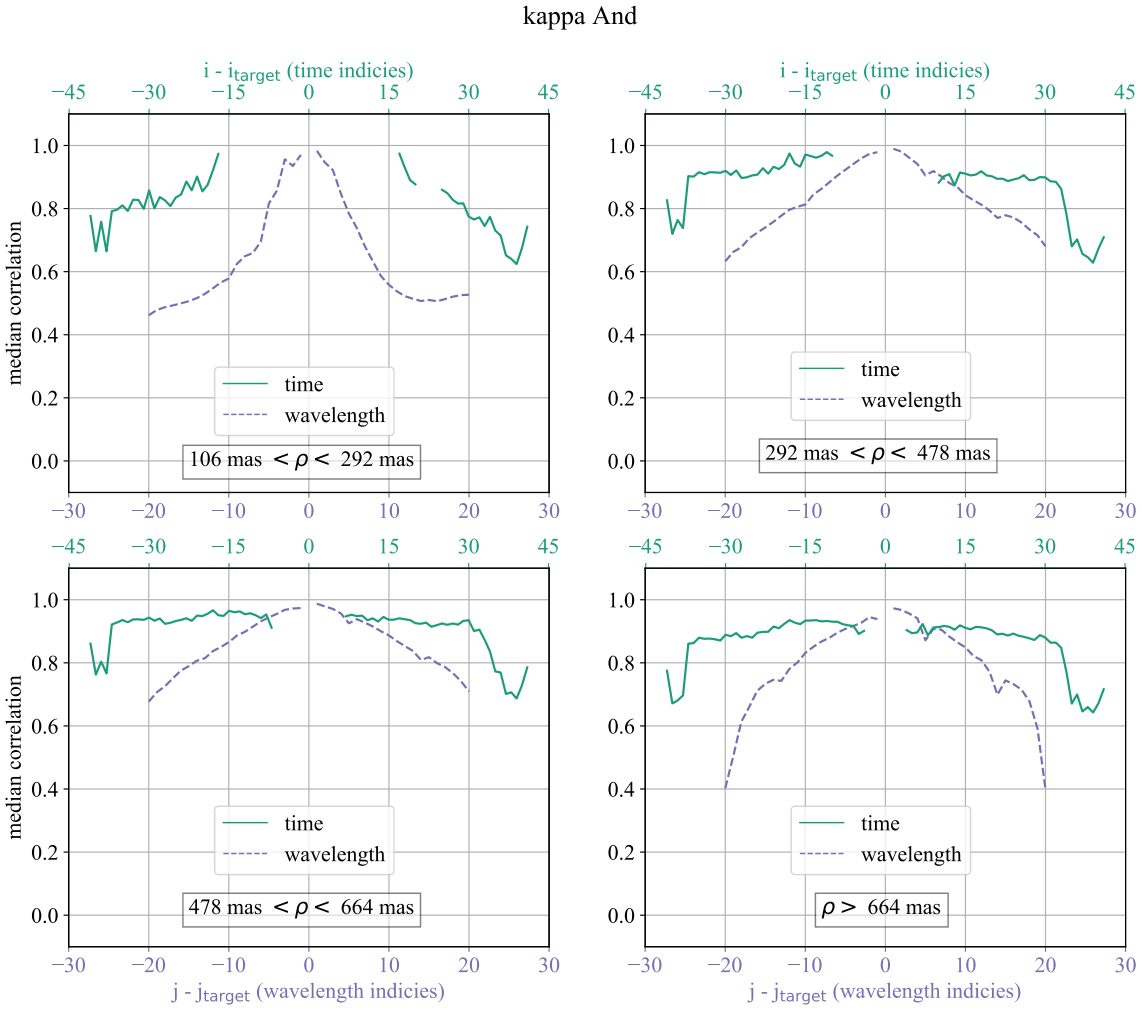


Figure 2.11 Median correlation with time and wavelength, analogous to Fig. 2.10, for the κ And dataset from Currie et al. (2018b).

The main empirical conclusion from Figures 2.10 and 2.11 is that SCExAO/CHARIS datasets are generally more correlated as a function of time than as a function of wavelength, most notably at small and large separations. This is particularly important at small separations, where PSF subtraction algorithms generally underperform compared to the rest of the image (e.g., Gerard & Marois, 2016b). For a given difference in parallactic angle and/or wavelength between a target and reference image, the respective azimuthal and/or radial movement of an exoplanet in between the two images is larger at larger separations. In other words, for a given aggressiveness, more reference images near the target image, both in time and wavelength, have to be discarded at smaller separations compared to larger separations. Although a longer wavelength lever arm would help overcome this problem, Figures 2.10 and 2.11 illustrate that this would generally add unnecessary noise in the covariance matrix inversion. A more correlated image at the same wavelength and different time will be selected instead. Later in §2.2.3 I will test the predictions from Fig. 2.10 on my PSF subtraction pipeline.

With that said, at small separations in Fig. 2.10 and at all separations in Fig. 2.11, reference images are actually more correlated to a given target image at the closest wavelengths than at the closest times. Although these images will be preferentially selected for SDI processing, this approach will not necessarily reach the deepest (throughput-corrected) final contrasts. Selecting a closer wavelength will enable reaching a better non-throughput-corrected contrast but will also cause more self-subtraction, therefore requiring a larger throughput correction. Thus, the optimal reference selection based on final, throughput-corrected contrast will ultimately depend on optimizing the trade off between correlation and self-subtraction, as in Gerard & Marois (2016b). Additionally, although the chromaticity is still worse than the temporal-stability in Fig. 2.11, note that there is a significant improvement in both temporal stability and chromaticity between Fig. 2.10 and 2.11. Although the improvement from temporal stability is most likely from improved AO performance (optical seeing and H band SRs for the Currie et al. 2018b dataset are around 400 mas and 0.92, respectively), the origin of chromaticity improvement is not clear and will require further investigation and analysis, beyond the discussion below.

Despite an apparent bias in the chromaticity measurements towards more stable sequences (i.e., chromaticity values at all separations in Fig. 2.11 are more correlated than in Fig. 2.10, a measurement which should ideally be independent of varying temporal performances), chromaticity remains a stronger effect than temporal sta-

bility in both Figures 2.10 and 2.11, still suggesting that additional factors are at play in causing such additional decorrelation with wavelength. Although the specific origin of the observed chromaticity is beyond the scope of this section, I discuss a few possibilities here that could explain these results. In general, Fresnel propagation generates a chromatic evolution of the wavefront from any out-of-pupil plane optics, particularly for optics that are transmissive and/or near the focal plane (e.g., Marois et al., 2008c), as well as from atmospheric scintillation (Guyon, 2005), although both of these effects are expected to be very weak at the current level of raw contrasts (but see Madurowicz et al. 2019, who show that in some cases scintillation effects are observed in GPI data). Additional sources of the observed behaviour may be algorithmic in nature, arising from either numerical interpolation errors and/or DRP extraction errors. These numerical interpolation errors are wavelength-dependent when aligning data cubes, since, as discussed in §2.2.2.1, the full CHARIS bandpass goes from near the Nyquist sampling limit in J band to more oversampled in K band. Although numerical interpolation errors are more likely to disproportionally effect the J band (from under-sampling effects) while DRP extraction errors are more likely to affect the K band (from increased thermal background effects), further investigation is needed to understand the relative impact of either effect. At small separations there are two additional possible origins of the observed chromaticity: the atmospheric dispersion corrector (ADC) and the FPM IWA. The ADC does not sufficiently offset differential atmospheric tip/tilt to centre the FPM on the star at the red edge of the K band. For the FPM IWA, because my chromaticity analysis uses magnified data cubes to align speckles as a function of wavelength, the IWA on magnified data cubes varies between about 2 and 4 λ/D at the red and blue ends of the bandpass, respectively, which could explain the observed chromaticity in the inner-most radial zone.

2.2.3 Contrast Curves

My 5σ contrast curves are shown in Fig. 2.12. The corresponding exoplanet mass limits for these final contrast curves are presented in §2.2.4. Final contrasts are calculated using the standard deviation at each separation in a 3 pixel-wide annulus on the collapsed, back-rotated images. Raw contrasts are calculated on a median image of the stack of the registered, magnified, flux-normalized, high-pass-filtered images (i.e., the same image preparation used before the correlation analysis in addition to a median stack of the full sequence). Also note that the frames used to calculate raw

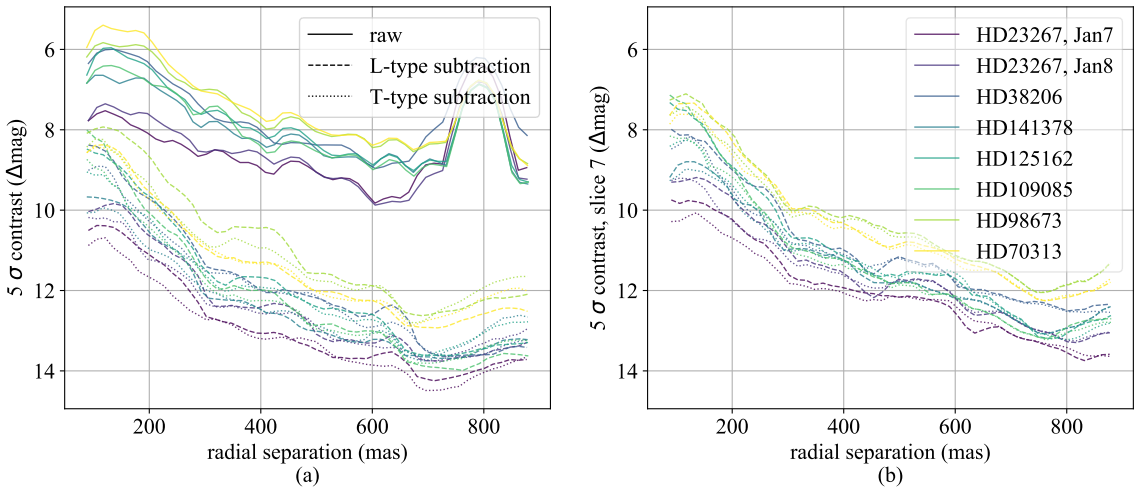


Figure 2.12 (a): contrast curves for all of my targets. (b): Final contrasts at a single wavelength, slice 7 = 1.5 μm. Each target is colour coded (the same in both figures); the solid, dashed, and dotted lines indicate the raw, final L-type, and final T-type contrasts, respectively.

contrast are not de-rotated to align the parallactic angle, thus keeping quasi-static aberration aligned during the median collapse across time and wavelength (the latter only for the panel a of Fig. 2.12). Although deeper contrasts have been reached with CHARIS on more stable datasets (e.g., Currie et al., 2018b; Rich et al., 2018), the main purpose of this section is instead to provide a detailed correlation analysis (§2.2.2.4) that can be used as a reference for future CHARIS observations (e.g., considering the impact of SDI in broadband mode).

With that said, Fig. 2.12 a and b does show that final T-type subtractions generally reach up to ∼50% deeper contrasts than the corresponding L-type subtractions for the same target, particularly at small separations, illustrating that SDI *is* improving contrast.[4] Contrasts in a single slice are shown in addition to the final broadband contrasts to control for the effects of noise averaging in comparing an L- vs. T-type reduction. In applying a weighted average using the assumed planet spectral type to collapse the final subtracted cube, noisier frames that lie further away from the bright part(s) of the spectrum are given less weight to the final collapsed cube (Marois et al., 2014), and so looking only at a single wavelength slice at the peak of the H band T-

[4]Although this gain could be described by the relative lack of chromaticity at small compared to large separations (as shown in Fig. 2.10), stable datasets with a larger FOV rotation than acquired here (Table 2.2) could also enable better contrasts for an ADI-only reduction at these smaller separations, motivating the need for additional high SR "deep dive" datasets in the future.

type spectrum removes this effect. Although other wavelengths will not show as much of a performance gain further away from the T-type spectrum peak, a deeper contrast reached by a T-type vs. L-type, at any wavelength, illustrates that SDI is being used to improve contrast. This improvement occurs because, within the 15 reference images that are chosen to subtract a target image at a given time and wavelength, some are at the same time but different wavelength as the target image. These results are consistent with my analysis in §2.2.2.4 which showed that chromaticity generally prevents using a reference image for SDI with a large wavelength lever arm, but that reference images closer in wavelength were often more correlated than images closer in time.

Along these lines, because the main conclusions of Figures 2.10 and 2.11 indicate that reference images at a large wavelength difference from the target image are unlikely to be selected (i.e., separate from the hypothesis of whether or not SDI is improving contrast), I carried out the exact same PSF subtraction procedure as described in §2.2.2.2 on all seven targets using only the H band wavelengths within the broadband data cubes (i.e., slices 6 - 13). The idea here is to see what level of contrast improvement is lost by removing the broadband feature of CHARIS while still using SDI in a standard ∼20% bandpass. Note that using the same parameters as in §2.2.2.2 (e.g., N_{ref}=15) means that there are less available reference images as a function of wavelength to select for the H band-only sequence. However, because reference images are selected by correlation, my chromaticity analysis in §2.2.2.4 and Figure 2.10 has determined that images at these wider wavelength differences are generally not selected, and so I expect this discrepancy to have a negligible effect on final contrast.

Fig. 2.13 illustrates the contrast difference on a single slice in the centre of H band for all of my targets between the full broadband mode and the H band mode. Note that comparing the collapsed cubes between the broadband and H-only subtractions would introduce a systematic bias; in the final wavelength collapse the broadband dataset would average more slices than the H-only dataset, producing a deeper final contrast for the former even if the individual subtracted cubes are the same for both sequences. In general, all of my April 2017 observations show little to no gain using the broadband dataset, as expected from Fig. 2.10. Interestingly, the January 2018 observations, which reach the deepest raw and final contrasts in Fig. 2.12, instead consistently show a ∼0.5 - 1 magnitude gain from using the broadband mode. Also when using the broadband mode, in a few cases at separations less than about 200

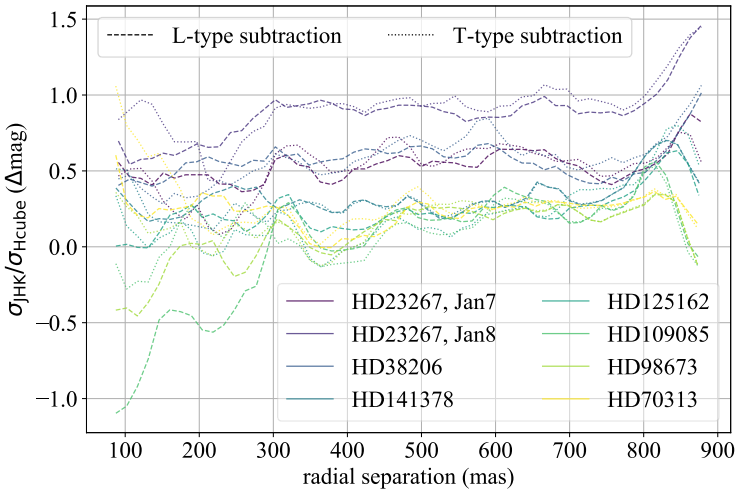


Figure 2.13 The difference in contrast for each target when using the full JHK dataset vs. only the H band to do PSF subtraction, illustrated here for a single wavelength slice at the centre of H band (slice 11 in the full JHK cube). The color and line styles are the same representation as in Fig. 2.12.

mas, the L-type subtractions appear to decrease in contrast by up to a magnitude.

Ultimately, the origin of this observed chromaticity that may be improving and/or degrading contrast will need to be better understood before a detailed analysis leading to mitigation strategies can be done. As discussed in §2.2.2.4, many different possible sources of chromaticity could reproduce the results in Figures 2.10, 2.11, and 2.13. These sources may be algorithmic and/or optical in nature, and could be quasi-static and/or dynamic. However, independent of the origin, in this section I have illustrated that chromaticity in my broadband datasets is limiting the final achievable contrast, thus motivating the need for further work towards more achromatic ExAO systems.

2.2.4 Mass Limits

In this section, the final contrast curves from Fig. 2.12 a are converted into upper limits on exoplanet mass using models from Spiegel & Burrows (2012) and Pickles (1998). It is important to note here that the observations from this section will not be used for a stand-alone occurrence rate analysis as in Meshkat et al. (2017, see §1.3.4 for a summary of this work). I did not observe a control sample and therefore cannot perform a similar analysis. Furthermore, I do not detect any new exoplanets, and so folding my eight targets into the Meshkat et al. (2017) sample of 130 stars would not significantly change their results.

The process to convert contrast curves into mass upper limits, shown in Figure

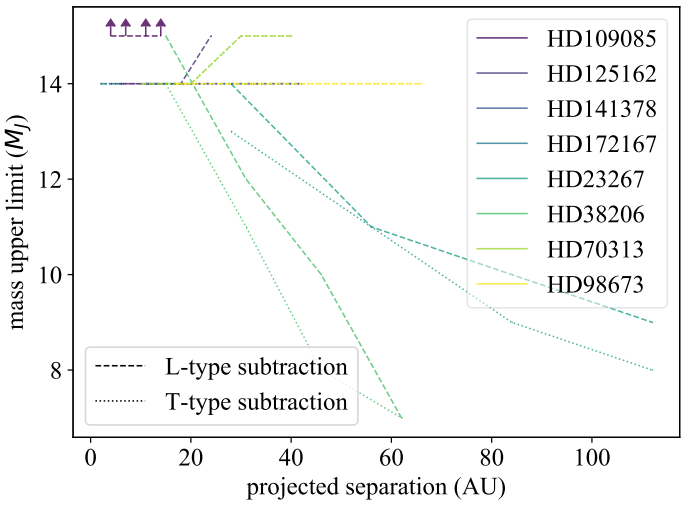


Figure 2.14 The minimum detectable planet mass for each target at a given separation, assumed spectral type, atmosphere, and entropy. Masses are computed using cooling curves from Spiegel & Burrows (2012), synthetic spectra from Pickles (1998), and target ages, distances, V band magnitude, and stellar spectral type from Kennedy & Wyatt (2014).

2.14, is as follows. First, target ages, distances, V band magnitude, and stellar spectral type are obtained from Kennedy & Wyatt (2014). For each target, a hot start spectrum from Spiegel & Burrows (2012) is chosen[5] for a user-defined model atmosphere type–either hybrid clouds for an L-type reduction or cloud-free for a T-type reduction, both at solar abundance–and planet mass based on the target age, flux normalized using the target distance. The spectrum is then integrated over the full 1.154 - 2.387 μm range of the CHARIS broadband bandpass. Separately, a Pickles spectrum (Pickles, 1998) is chosen to represent the host star, closest matching in spectral type. The synthetic stellar spectrum is flux normalized by the target V band magnitude and similarly integrated over the CHARIS broadband bandpass. The ratio of the integrated flux of the exoplanet over the integrated flux of the star represents the astrophysical flux ratio of an exoplanet at the user-defined model atmosphere type and exoplanet mass. At a given atmosphere and target separation, the exoplanet mass is adjusted until the difference between the 5σ contrast at that separation (for a given L- or T-type subtraction) and the astrophysical flux ratio is minimized. If this minimum does not converge on the Spiegel & Burrows (2012) model grid out to 15 Jupiter masses (M_J), the upper limit is shown as"$>$ 15 M_J" in Fig. 2.14. This process is repeated for each T- and L-type subtraction; the full process is then repeated for contrasts at 200, 400, 600, and 800 mas, all of which is repeated for every target (for HD23267, I choose whichever contrast is deeper at any given separation). Note that these limits are not particularly constraining when compared to deeper contrasts obtained by SCExAO/CHARIS and/or other similar instruments. Also see

[5]I.e., initial entropy = 13.0 k_B/baryon; these observations are insensitive to cold start planetary masses.

§4.2 for a discussion of future work from this initial upper limit analysis.

Chapter 3

A Path Forward: Fast Focal Plane Wavefront Sensing and Subtraction

Given the limitations set by speckle evolution, as described in Chapter 2, a new path forward is needed to reach deeper contrasts, particularly for the requirements needed to image habitable rocky exoplanets on future ELTs (§1.3.6). In §1.4.4.2, I illustrated a possible solution: fast focal plane wavefront control and CDI. Despite the relatively unexplored parameter space in application to ground-based high contrast imaging instruments, this approach has enormous potential: in theory, *fast focal plane wavefront control and CDI are not limited by chromaticity or stability*, opening the door to unprecedented contrast gains. Although still labeled as "differential imaging," CDI differs from other classical differential imaging techniques (§1.4.4.1), in principle requiring only one monochromatic image to extract the coherent wavefront information without bias from the incoherent exoplanet signal. This architecture enables millisecond-timescale speckle subtraction (which is not possible with ADI) and, if fast enough, removes the limitation of stability, while the design of CDI as monochromatic (which is not possible with SDI) removes the limitation of chromaticity.

In practice, the required temporal and chromatic bandwidth is completely dependent on the level of stability and chromaticity relative to the photon noise limit, respectively. The former will be explored in detail later in this chapter, while the latter still remains relatively unexplored and planned for future work (see §4.3.1).

In the bulk of this chapter (§3.5) I will highlight my work from Gerard et al. (2018a), Gerard et al. (2018b), Gerard et al. (2019a), and Gerard & Marois (2020), where I develop the framework for a specific CDI method, called the SCC, to be

applied to ground-based telescopes in order to correct both residual atmospheric and quasi-static aberration, hereafter referred to as the Fast Atmospheric SCC Technique (FAST). I will also present additional unpublished work, carried out both before (§3.3, §3.4) and after (§3.5.5, §3.5.6, §3.5.7.4) my published work on FAST. I will first begin with a general introduction to the SCC method in §3.1.

3.1 The Self-Coherent Camera

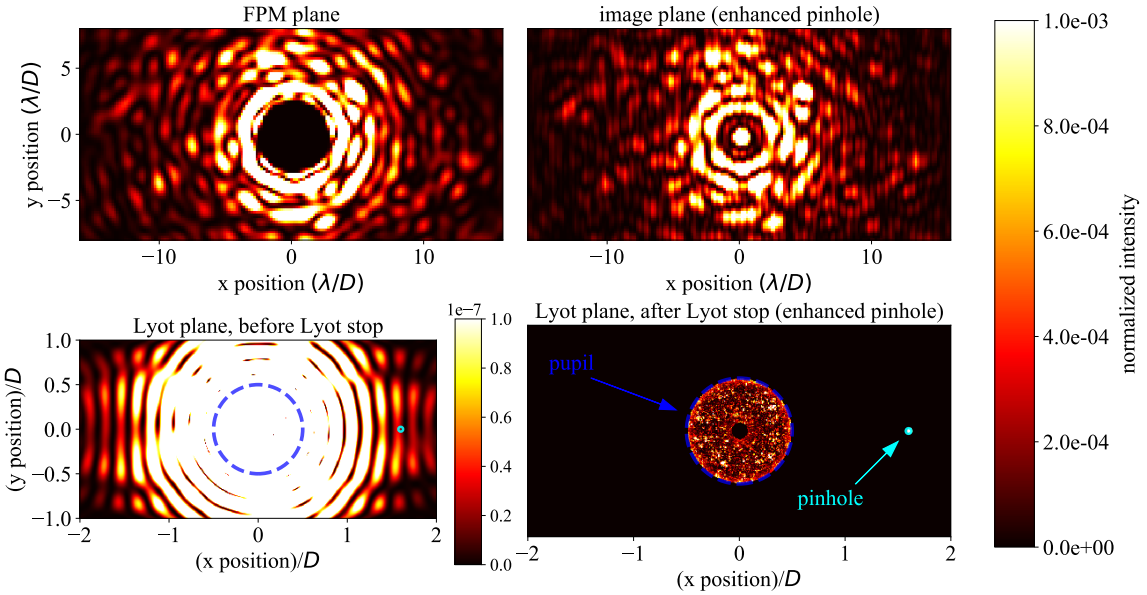


Figure 3.1 Upper left: a FPM is used to occult the PSF core. Lower left: the FPM diffracts light in the downstream pupil plane beyond the original entrance pupil footprint (indicated by a dotted blue circle). This effect occurs from any FPM, either in phase and/or amplitude (e.g., see Fig. 1.7 c). Lower right: a Lyot stop is used to suppress these diffraction effects, but a small off-axis pinhole transmits some of this diffracted light (which is of order 10^6 times dimmer compared to the central pupil for most FPM designs, shown in the lower left). Upper right: intensity in the downstream image plane after this Lyot stop is applied, showing sub-λ/D scale fringes from the interference between the Lyot plane pinhole and central pupil. These fringes provide a proxy measurement of the complex electric field in the image plane, despite only imaging electric field intensity. The relative intensity in the Lyot pinhole is enhanced by a factor of 2.5×10^5 in the upper and lower right images for illustration purposes. All images are normalized to the peak image plane intensity when no FPM is present.

The SCC, which is illustrated in Fig. 3.1, was originally developed by Baudoz et al. (2006) and then modified by Galicher et al. (2010) specifically to be incorpo-

rated into a high contrast imaging instrument, utilizing the principle of coronagraphic diffraction. As presented in §1.4.3.2, by adding a FPM, either in amplitude and/or in phase, to suppress the core of the PSF, light is diffracted outside the footprint of the entrance pupil in the down-stream pupil plane, called the Lyot plane. An aperture mask, called a Lyot stop, is then used in the Lyot plane to attenuate such diffraction effects that would ultimately be seen in the coronagraphic image (i.e., the intensity of the focal plane wavefront that is downstream of the Lyot plane). However, the SCC modifies this classical Lyot coronagraph design by adding an off-axis pinhole, outside of the entrance pupil footprint, in the Lyot stop. The diffracted light transmitted through this pinhole will then interfere in the downstream focal plane with the light that has been transmitted through the central aperture of the Lyot stop, enabling a camera placed in this focal plane—with the necessary sampling—to resolve fringes.

These recorded fringes in the SCC image measure the complex focal plane electric field in a single image; the amplitude is determined from the fringe intensity, while the phase is determined from the relative fringe position, as illustrated in Fig. 3.2. As shown, a normal coronagraphic image is mostly insensitive to the relative phase of a single Fourier mode; small intensity variations at the location of the PSF copy ($\pm 9.5\ \lambda/D$ along the x axis for this 9.5 cycle/pupil sine wave) occur due to interference between the PSF copy Airy function and the background diffracted light (see §1.4.3.2.2), a highly non-linear phenomenon dependent on the relative levels of sinespot intensity and background diffraction (Perrin et al., 2003). However, this same relative phase in the SCC image is linearly mapped to the relative fringe position on either PSF copy; in other words, as the sine wave translates across the pupil, fringes translate across the PSF copy.

Because only diffracted starlight passes through the off-axis Lyot stop pinhole,[1] fringes are generated in the coronagraphic image only on coherent starlight and not incoherent exoplanet light. In other words, a speckle from the on-axis star located at the same position as an off-axis exoplanet will be fringed, while the exoplanet will not; measuring the amplitude and phase of only the fringed speckle (see below) and applying the same amplitude but opposite phase to the DM will attenuate the speckle but *not* the exoplanet.

[1] In reality, some diffracted exoplanet light will be transmitted through the pinhole, although typically at a negligible level unless the exoplanet is close to or smaller than the coronagraph IWA. Subsequent SCC exoplanet algorithmic throughput tests in §3.5.2 will validate this principle.

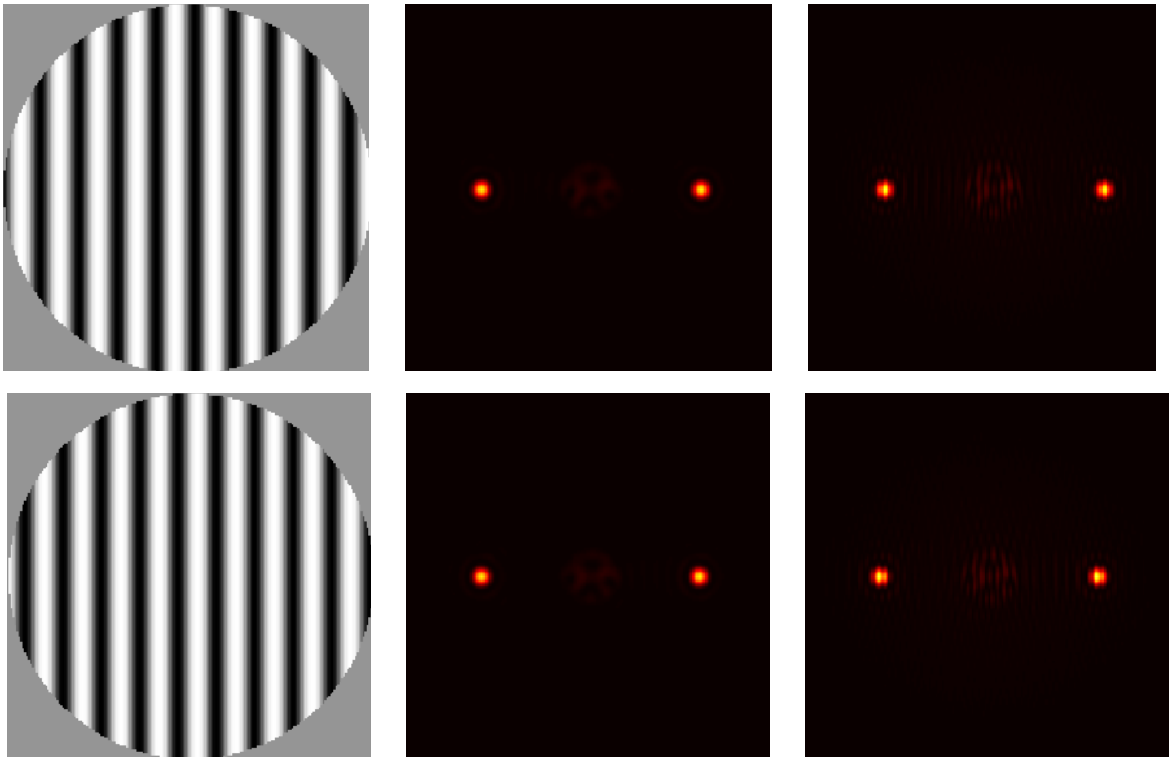


Figure 3.2 Upper left: a Fourier mode applied to the phase of the electric field in the entrance pupil (20 nm amplitude at $\lambda = 1.65$ μm). Upper middle: the corresponding coronagraphic image, using the Tip/tilt+Gaussian+Vortex (TGV) coronagraph (see §3.5.4.2), with no pinhole in the Lyot stop. Upper right: the same coronagraphic image as the middle panel but with a pinhole in the Lyot stop. Lower panel: the same as the upper panel, but with the sine wave phase in the lower left panel translated by π radians relative to the upper left panel (i.e., the sine wave in the upper left panel is converted into a cosine wave in the lower left panel). The animated top panel of this figure, showing a continuous translating sine wave and its impact on the coronagraphic and SCC images, is viewable in Adobe Reader version 7 or greater.

With this setup, a recorded monochromatic SCC image at wavelength λ is given by (Baudoz et al., 2006)

$$I(\vec{\alpha}) = |A_S(\vec{\alpha})|^2 + |A_P(\vec{\alpha})|^2 + |A_R(\vec{\alpha})|^2 + 2\ Re\left\{A_S(\vec{\alpha})A_R(\vec{\alpha})^*\ e^{\left(\frac{2\ i\ \pi\ \vec{\alpha}\ \xi_0}{\lambda}\right)}\right\}, \quad (3.1)$$

where $A_S(\vec{\alpha})$, $A_P(\vec{\alpha})$, and $A_R(\vec{\alpha})$ are the focal plane complex electric field components of the star, planet, and light from the Lyot stop pinhole, respectively, ξ_0 is the separation between the optical axis and the center of the pinhole in the plane of the Lyot stop, and $\vec{\alpha}$ defines the (x, y) position in the image. The fringe term in equation 3.1, $2\ Re\left\{A_S(\vec{\alpha})A_R(\vec{\alpha})^*\ e^{\left(\frac{2\ i\ \pi\ \vec{\alpha}\ \xi_0}{\lambda}\right)}\right\}$, contains a range of sub-λ/D spatial scales, defined by the off-axis Lyot stop pinhole size and separation from the central pupil. I will remove the use of $\vec{\alpha}$ from subsequent notation in this dissertation for simplicity; similarly, I will replace the fringe term by $2\ M(|A_R|\ |A_S|)$, where M is a dimensionless "fringe function," varying between 0 and 1. I will also use subsequent notation of $I_S \equiv |A_S|^2$, $I_R \equiv |A_R|^2$, and $I_P \equiv |A_P|^2$. These "stellar speckle" (I_S), "pinhole PSF" (I_R), and fringe ($2M\sqrt{I_R I_S}$) terms are all illustrated in Fig. 3.3.

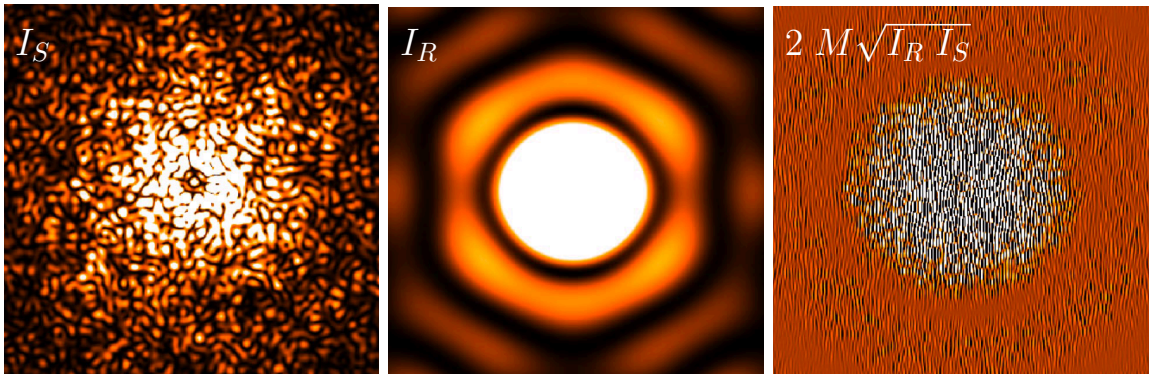


Figure 3.3 SCC image components, including the stellar speckle (left), pinhole PSF (middle), and fringe (right) terms, each shown on the same relative spatial scale.

Because the fringe term, $2\ M(|A_R|\ |A_S|)$, is linearly dependent on both $|A_R|$ and $|A_S|$, in order to prevent $|A_R|$ from reaching a minimum inside the DM Nyquist control region of the SCC image (see §1.4.4.2.1, Fig. 1.16)—which would attenuate most of the fringes at that separation and prevent a measurement/correction at sufficient

S/N—the maximum off-axis Lyot stop pinhole diameter, d, is (Mazoyer et al., 2014)

$$1.22\ \lambda/d \leq \sqrt{2}\ N_{\rm act}(\lambda/D),\ \text{or}$$
$$d \leq 1.22\sqrt{2}\ D/N_{\rm act}, \tag{3.2}$$

where $N_{\rm act}$ is the number of DM actuators across the entrance pupil. In general, equation 3.2 is only a limitation for DM control; a subtraction via CDI image processing may utilize a smaller or larger pinhole, correcting for a larger or smaller area in the focal plane, respectively, than the AO control region (although the subtraction at the separation of the pinhole minima will still be worse due to the lower S/N detection of the pinhole PSF).

3.1.1 Post-Processing Algorithms

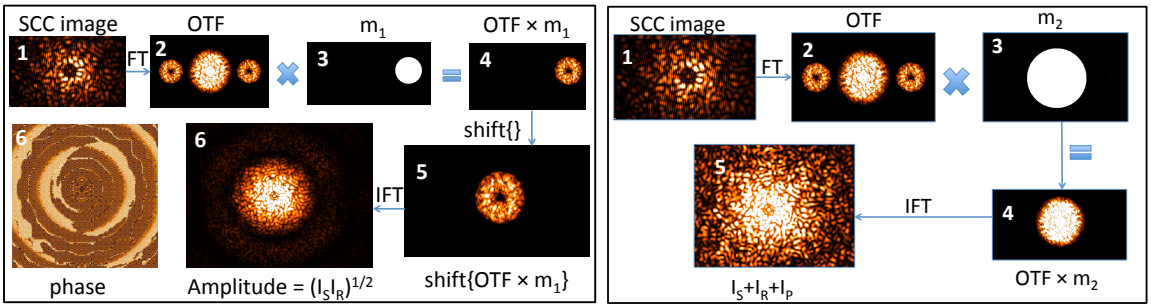
The fringe term from equation 3.1 can then be isolated in spatial frequency to prevent any loss of exoplanet light. The full processing algorithm to obtain this information from the recorded image is described in Galicher et al. (2010, and references therein) and illustrated in Figure 3.4, utilizing the complex-valued Fourier transform of the image—the optical transfer function (OTF)—and showing the OTF amplitude—the modulation transfer function (MTF).
In summary, the fringe term in equation 3.1 can be isolated as follows.

1. Take the Fourier transform of the recorded image to obtain the complex-valued OTF.

2. Multiply the amplitude of OTF by a binary mask, m_1, to isolate a side lobe, which is generated from sub-λ/D spatial scales in the image and which contains only the fringed starlight. In order for this side lobe to not overlap in the OTF plane with the main beam, the center of the pinhole in the Lyot plane must be placed at minimum separation, ξ_0, from the center of the pupil of (Galicher et al., 2010)
$$\xi_0 \geq (D/2)\ (3 + d/D). \tag{3.3}$$

3. Shift the unmasked complex array to the centre of the image.

4. Apply an inverse Fourier Transform back to the image plane.



(a) (b)

Figure 3.4 Two different SCC Fourier filtering algorithms. For both panels, "FT" and "IFT" represent the Fourier transform and inverse Fourier transform, respectively. The binary masks "m_1" and "m_2" are shown in panels a and b, respectively. The "shift{}" operator repositions the isolated OTF sidelobe to the center of the Fourier plane. (a) From Baudoz et al. (2006), the SCC OTF yields a main beam and two higher spatial frequency side-lobes, the latter of which represents the electric field of the SCC image fringe term and can be isolated as illustrated here. (b) A separate but similar filtering algorithm to isolate the lower spatial frequency image components, removing the fringe term.

This wavefront sensing algorithm, or Fourier sidelobe isolation algorithm, is also the standard wavefront sensing algorithm used for wavefront control and is crucial to preserving optical exoplanet throughput. As long as no exoplanet light is sent through the reference pinhole, all of the exoplanet information will reside in the main beam of the OTF, whose amplitude is shown in Figure 3.4; by isolating the OTF sidlobe, we are directly measuring the electric field of the star without confusion/contamination from the exoplanet.

The product produced from step 4 above (and step 6 of Fig. 3.4 a), referred to in the literature as I_-, is an algorithmic solution to isolate the complex amplitude of the modulation term in equation 3.1. By masking only one side lobe in the OTF, this removes both the factor of two in the fringe term and the modulated component of that term, M, since all other baselines are also removed (shifting the OTF side-lobe to the center of the Fourier plane after applying a binary mask will only change the phase of I_- in the focal plane, but not the amplitude). Thus, taking the absolute value of I_- gives

$$|I_-| = |\mathrm{IFT}\{\mathrm{FT}\{I\}\ m_1\}| = \sqrt{I_S\ I_R}. \tag{3.4}$$

Although the phase term from this filtering algorithm is used for wavefront control by Baudoz et al. (2012a), Mazoyer et al. (2014), and subsequent papers, we do not

need to know the speckle phase to do post-processing (see §3.5.2).

Another image term can be separately reconstructed by placing a different binary mask in the OTF plane, as illustrated in Figure 3.4 b: removing the modulation term by isolating the main beam with a binary mask, m_2, then doing an inverse Fourier transform and taking the absolute value provides the un-modulated terms of the image:

$$\mathrm{IFT}\{\mathrm{FT}\{I\}\, m_2\} = I_S + I_R + I_P. \tag{3.5}$$

No absolute value or real component operator is needed on the output of equation 3.5; $\mathrm{IFT}\{\mathrm{FT}\{I\}\, m_2\}$ is entirely a real image, since the Fourier filtering mask m_2 is centro-symmetric. If there was a way to estimate the pinhole PSF (as will be discussed later in §3.5.2.1), this would then allow a reconstruction of I_S using equation 3.4 and thus a full subtraction of all terms in equation 3.5 other than I_P. Assuming we have obtained a noiseless image of the pinhole PSF, from the same wavefront realization as equation 3.5, the subtracted image, $\mathrm{im}_{\mathrm{subt}}$ is then

$$\begin{aligned}
\mathrm{im}_{\mathrm{subt}} &= \mathrm{IFT}\{\mathrm{FT}\{I\}\, m_2\} - \left(\frac{|I_-|^2}{I_R} + I_R\right) \\
&= (I_S + I_R + I_P) - \left(\frac{|I_-|^2}{I_R} + I_R\right) \\
&= I_P + \left(I_S - \frac{|I_-|^2}{I_R}\right).
\end{aligned} \tag{3.6}$$

Using the simulation parameters described below in §3.2, using a single atmospheric and static phase screen realization I obtain a starting 1σ contrast across the full 32×32 λ/D AO control region of 6.3×10^{-4}. Using this noiseless target image to calculate $|I_-|$ and then, additionally using the noiseless pinhole PSF, the resulting subtracted image from equation 3.6 reaches a 1σ contrast of 2.0×10^{-18} across the full AO control region (when no companion is simulated); this limit is from unphysical numerical noise, which illustrates that these effects should be negligible after incorporating photon noise into my simulations, and that instead analyzing noisy images will isolate the physical effects of photon noise propagation through my coronagraphic image

reconstruction algorithm. Thus, with no photon noise,

$$\left(|A_S|^2 - \frac{|I_-|^2}{I_R}\right) = 0, \text{ or}$$
$$\frac{|\mathrm{IFT}\{\mathrm{FT}\{I\}\ m_1\}|^2}{I_R} = |A_S|^2. \tag{3.7}$$

However, in reality photon noise sets a fundamental limit on the achievable contrast in equation 3.6. These effects will be discussed later in in §3.5.1.2.

3.1.2 Agenda

Using the above framework developed and tested by Baudoz et al. (2006) and subsequent papers, the goal of my research in this chapter is to present a new strategy to enable efficient SCC operation on ground-based telescopes. The simulation parameters and numerical setup used throughout this chapter are listed in §3.2. In §3.3 I first present an analysis of tolerances and solutions for vibration and drift effects using the classical SCC design. Then in §3.4 I present an initial analysis of methods to increase the SCC fringe S/N in order to access faster ground-based operational timescales. In §3.5, I present my main solution to boost the fringe S/N, including new FPM designs (§3.5.1.1, §3.5.4), post-processing algorithms (§3.5.2), DM control algorithms (§3.5.3, §3.5.4.3), WFS linearity and sensitivity analyses (§3.5.5, §3.5.6), initial laboratory testing results (§3.5.7), and predicted scientific yield (§3.5.8).

3.2 Simulation Parameters and Assumptions

Throughout Chapter 3, unless explicitly stated, simulations are run at 1.65 μm, using an 8.2 meter telescope, and a 1 % bandpass filter to calculate flux but otherwise using a monochromatic Fraunhofer simulation; Mazoyer et al. (2014) show that chromaticity effects on the SCC performance should be negligible over this narrow bandwidth and have similar performance to the monochromatic case (but see §4.3.1). For a classical Lyot coronagraph design, I use a 5 λ/D diameter amplitude FPM, and a circular Lyot stop, 4% undersized relative to the entrance pupil. The term "algorithmic exoplanet throughput" refers to the ratio of the peak exoplanet flux between the output and input of a PSF subtraction algorithm, with a recorded image as the input and a PSF-subtracted image (i.e., using image processing) as the output; this expression does

not refer optical/instrumental throughput of exoplanet light. I use the Fraunhofer approximation to propagate the electromagnetic field between pupil and focal planes. I assume use of a 32×32 actuator square DM across the entrance pupil to define the AO control region as a 32×32 λ/D box around the optical axis of the image plane. I use a pinhole along the $+x$ axis of the Lyot plane with a relative size (equation 3.2) of $d/D = 1.22\sqrt{2}/32 \approx 5.4\%$ and relative separation (equation 3.3) of $\xi_0 = 1.6\ D + d/2$ (I use a factor of 1.6 instead of 1.5 to provide additional zero padding to between the MTF central lobe and sidelobes, easing the m_1 radius tolerances in the algorithmic reconstruction). I use a beam ratio (image size divided by pupil diameter, or number of pixels per λ/D) of 5.36. When simulating exposures to estimate photon noise, I calculate flux in number of photons per second for the above parameters from Bessell et al. (1998) and then assume a transmission through the atmosphere of 90%, transmission through the telescope and instrument of 20%, and detector quantum efficiency of 80%. Simulated stellar magnitudes and exposure times are noted in each section. The 5σ contrast curves are calculated from an image as follows:

1. Normalize the image: divide the propagated coronagraphic image by the peak pixel value of the non-coronagraphic image, obtained from the propagated focal plane intensity before applying a FPM but with a Lyot stop placed in the entrance pupil.

2. Calculate the standard deviation of the image output from step 1 in a 0.6 λ/D wide annulus centered at radial separation ρ, and then multiply the result by 5 to obtain a "5σ contrast."

3. Repeat step 2 at each radial separation, increasing by 0.2 λ/D increments between 3 and 16 λ/D, and then plot each contrast value vs. each ρ value to obtain a "5σ contrast curve."

To simulate quasi-static aberration, I tune the entrance pupil WFE PSD amplitude and power law of both phase and amplitude aberration to match a raw GPI contrast curve at of single IFS slice ($\lambda = 1.65\mu$m) from a single frame of the 51 Eri b detection sequence (Macintosh et al., 2015); in this sequence, consecutive exposures are about 90% correlated over time, suggesting that we are seeing mostly static wavefront error and a small amount of AO residuals. Additionally, GPI images show no/very little symmetry; as in Marois (2004), pure phase or pure amplitude aberration should be approximately symmetric in a coronagraphic image for a small

amount of aberration (i.e., so that a first order Taylor expansion of the PSF from Perrin et al. 2003 is a valid approximation), and so the absence of symmetry implies an approximately equal contribution of phase and amplitude aberration to generating quasi-static speckles. With this in mind, I use a 25 nm RMS, -1.5 power law phase aberration and a 1% RMS, -2 power law (in intensity) amplitude aberration to generate a static phase and amplitude aberration. For both phase and amplitude components, tip and tilt are subsequently removed by a least-squares subtraction (other than in §3.5.5).

To approximate an ExAO system, I use a 100 nm RMS phase screen conjugated to the entrance pupil, or by the Maréchal approximation a SR of 0.87, with a -2 power law, which is typical for AO-corrected residual turbulence (J.P. Véran, private communication), subsequently removing tip and tilt to simulate the effect of a LOWFS and remove lower order pointing effects (again, other than in §3.5.5). I do not consider the effects of residual uncorrected aberrations (i.e., the phase screen PSD amplitude and shape should be significantly larger, on the order of tens of microns RMS, and closer to -11/3, respectively, beyond the spatial frequencies corresponding to the DM control radius) and therefore will not analyze performance beyond the 16 λ/D control radius. I use the framework from Srinath et al. (2015), where the α parameter, a dimensionless number between 0 and 1, is used to simulate atmospheric "boiling," or deviating random effects from pure frozen flow; e.g., $\alpha = 0.95$ means that 95% of the current phase screen (or $\sqrt{100^2(0.95)}$ ≈97 nm RMS) will be translated to the next phase screen realization via a Fourier shift based on the time interval, telescope diameter, and wind speed and direction, whereas as the other $\sqrt{100^2(1-0.95)} \approx 22$ nm RMS component of the next phase screen realization will be randomly generated, but with the same -2 power law. Unless otherwise noted I use a windspeed of 10 m/s in the +x pupil direction and $\alpha = 0.95$ (i.e., 5% random turbulence is added at each time interval).

I use the Gemini entrance pupil (with secondary obscuration but no spiders) and the GPI APLC apodizer mask (Soummer et al., 2006) for the figures in §3.1 and in Sections 3.3 and 3.4. For the simulations in Sections 3.5.2 and 3.5.3, I use the GPI APLC apodizer mask from Soummer et al. (2006) in addition to the same Tip/tilt+Gaussian (TG) FPM as in §3.5.1.1. I use the TGV coronagraph in Sections 3.5.4, 3.5.5, and 3.5.8.

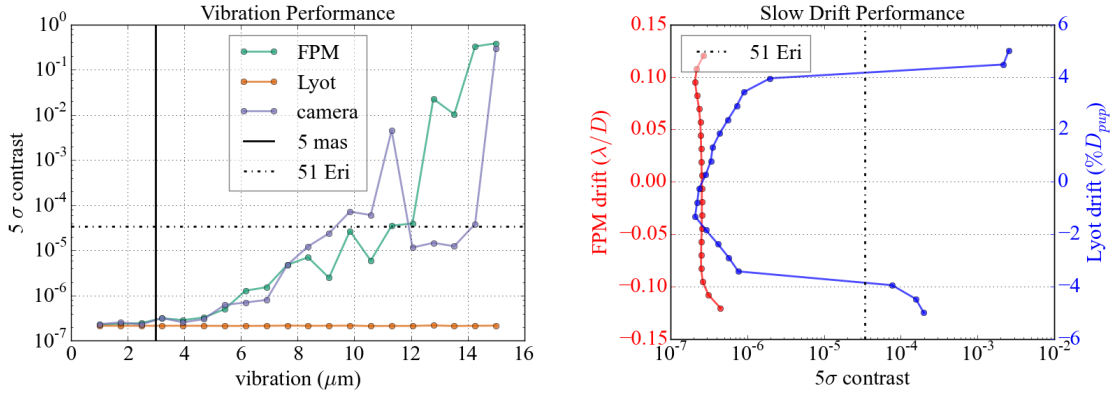
3.3 SCC Vibration and Drift Tolerancing

This section presents unpublished work completed during my PhD studies.

Before my invention of FAST (§3.5), in the first ∼8 months of my PhD I ran detailed numerical simulations using the GPI APLC (Soummer et al., 2006) with the classical SCC design (Baudoz et al., 2006; Galicher et al., 2010) to test the tolerances of vibration and drift. One motivation for using the SCC (often called a "common path interferometer"), is that it is thought to be more robust to vibration than other "non-common-path interferometer" designs, such as a Mach-Zehnder or Michelson interferometer (e.g., Steck, 2015, chapter 5). For the latter designs, a single input beam is split into separate optical arms before recombining; if differential vibration between these two arms is larger than $\sim \lambda$ RMS, fringes will no longer be detected due to fringe smearing. The SCC, however, acts as an interferometer in a single optical path; Lyot stop vibration will not change the separation between the pupil and the pinhole, as these apertures are on the same component, thus providing more robust tolerances to vibration. The GPI calibration system (Wallace et al., 2010, similar to the initial CDI design proposed in Guyon 2004)—a Mach-Zehnder interferometer design—has not been operational since commissioning (B. Macintosh, private communication) due to the above-described fringe smearing effects. Accelerometer measurements have shown that these instrument vibrations are occurring at frequencies greater than about ten Hz and amplitudes greater than a few μm, particularly affected by strong vibrations near the FPM where the beam is initially separated (T. Hayward, private communication). Without additional hardware modifications, future operation of this sub-system would require a faster operational frame rate (i.e., to free the turbulence and vibrations to a single wavefront realization) and/or achieving differential tip/tilt stability between the two interferometer arms (via passive and/or active wavefront control) below the level of λ RMS.

In this section, I ran simulations, using the setup described in §3.2, to understand how similar vibrations would impact the SCC operability, shown in Figure 3.5a. To simulate vibrations in the x, y, and z directions I added tip, tilt, and focus phase shifts in random directions but with varying amplitudes in the upstream pupil/focal plane of each optic/detector in the coronagraph path and measured the ultimate effect on achievable contrast throughout a half DH using an SCC DM correction as described in Mazoyer et al. (2014) and later in this chapter (§3.5.3). The results of these vibration

simulations are illustrated in Fig. 3.5 a, clearly showing that the initial contrast using the SCC can be maintained at vibration amplitudes of less than about 5μm, despite working at λ=1.65μm.



(a) (b)

Figure 3.5 Vibration (a) and drift (b; with methods in Fig. 3.6 applied for the Lyot stop) performance of the SCC. The dashed dotted line labeled "51 Eri," using the Macintosh et al. (2015) dataset, shows the best achievable raw contrasts with GPI.

Additionally, although the the GPI FPM is typically well aligned to within 0.1 λ/D using a specially designed low order LOWFS (an 11 $\times$ 11 sub-aperture NIR SHWFS that measures the light rejected by the FPM in a downstream pupil plane; Wallace et al. 2010), the apodizer and Lyot stop (although at a significantly lower tolerance requirement in the pupil plane compared to the FPM) rely on open loop flexure models to stay aligned on each target (Dunn et al., 2008). The pupil viewing camera used to align the apodizer and Lyot stop (Larkin et al., 2014) at the beginning of each night only has $\gtrsim$1% precision (i.e., relative to the pupil diameter) in positional alignment. Unknown drifts may then occur throughout the night, potentially deviating from the open loop models at greater than the nominal $\sim$1 % precision. Thus, in Figure 3.5 b I show the results from drift simulations for the FPM and Lyot stop. Although Fig. 3.5 b shows that the FPM is relatively insensitive to drifts within the 0.1 λ/D LOWFS precision (D. Savransky, private communivation), I initially found that the Lyot stop, when used in the SCC, was extremely sensitive to $\gtrsim$1% shifts, immediately degrading the achievable contrast by $\gtrsim$100 (left panel of Figure 3.6 c); these effects are not included in Figure 3.5 b, which instead use the results after applying my drift correction method to this problem, shown in Figure 3.6 and summarized below.

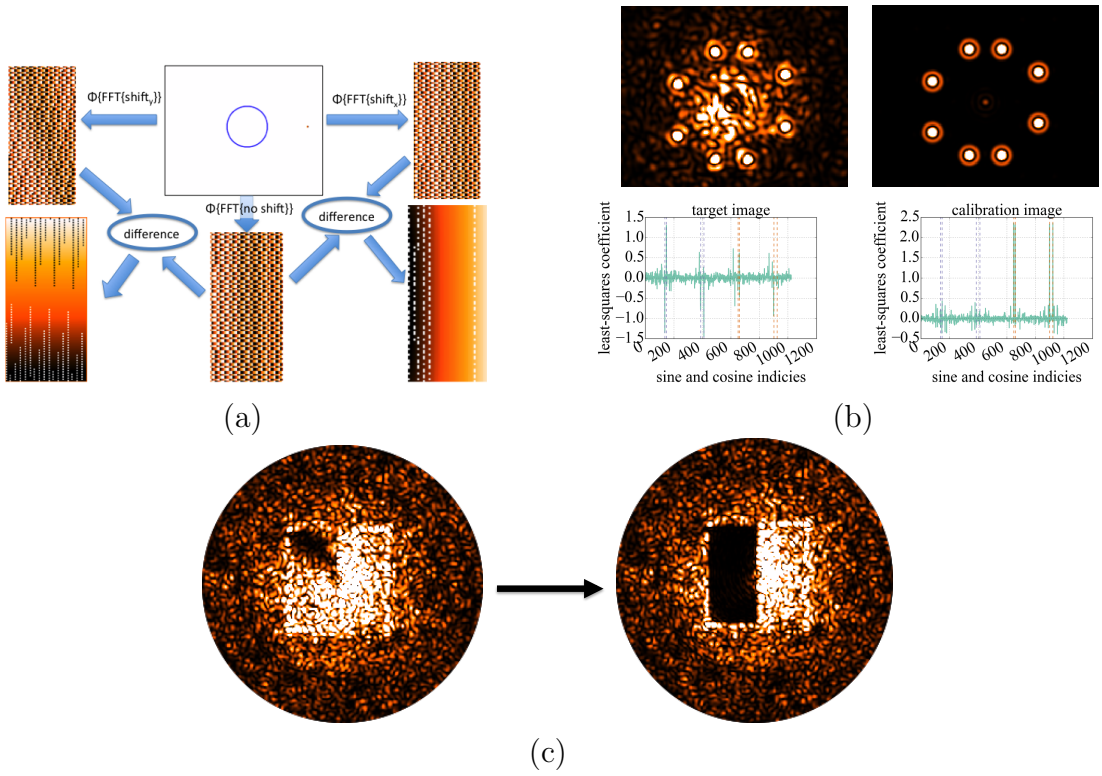


Figure 3.6 My methods to diagnose and correct for a misaligned Lyot stop using the SCC. Panel a examines the phase of the pinhole PSF (the pupil location is illustrated by a blue circle, but is assumed occulted) . "FFT" stands for "fast Fourier transform," which is used in my numerical simulations, "Φ{}" represents an operator used to determine the phase of the focal plane complex wavefront, and the x and y shifts are both 1% of the pupil diameter (i.e., the diameter of the blue circle). In panel b, the DM commands generated from four pure sine waves (right panel) are compared to those for a shifted Lyot stop (left panel). In panel c, a degraded DH due to the same 1 % Lyot drift as panel b is corrected using the information in panels a and b (see text).

I found that the main reason causing the contrast degradation from a Lyot stop shift was the relative electric field phase discrepancy going through the pinhole between the misaligned and aligned images, as illustrated in Figure 3.6a. Looking only at the complex electric field transmitted through the Lyot stop pinhole, the difference of the phase in the downstream focal plane shows a linear gradient following the direction of the Lyot stop shift (ignoring the numerical phase wrapping effects). Thus, placing four pure sine waves at different locations within a half dark hole in Figure 3.6 b shows that the least-squares coefficients (or DM commands, generated from a least-squares fit between the target image and a series of calibrated reference images

that places DM sines and cosines at λ/D separations throughout the half DH, as in Mazoyer et al. 2014 and §3.5.3) in the aberrated target image on the left (for which the Lyot stop has been shifted in a random combination of x and y by 1% of the pupil diameter) indicate a combination of both sines and cosines compared to the pure sine wave coefficients generated from the right panel of Fig. 3.6 b (for which no Lyot stop shift is applied). The ratio of least-squares coefficients between the left and right panels of Fig. 3.6 b enables a direct measurement of the phase discrepancy in Figure 3.6 a, thereby allowing a correction with the same relative phase offsets applied to the entire basis of calibrated sine and cosine reference images, the result of which is illustrated in Figure 3.6 c, clearly demonstrating that after correction is performed on the right panel the factor of $\gtrsim 100$ contrast degradation caused on the left panel from a 1% shifted Lyot stop is then corrected with the DM (i.e., no additional stages/open loop models are required).

3.4 Boosting Fringe S/N: Initial Failures

This section presents both unpublished work completed during my PhD studies and text from Gerard et al. (2018a).

In addition to limitations from vibration and drift, the classical SCC design presented in §3.1 will only enable long exposure measurements of the static wavefront component. For short ∼ms exposures, photon noise can hide any fringes from detection unless there is enough light going through the pinhole. Assuming that we want to detect and subtract speckles as soon as they are visible and before they disappear, when one photon from a speckle is recorded, we also want at least one photon recorded from the pinhole PSF at the same location so that the fringe amplitude is not zero (see §3.5.4.1.2 for a formal definition of "fringe S/N"). Or, when a few photons are recorded for a speckle but photon noise is still a dominant factor, we ideally want the fringe amplitude to be at least as large as the speckle amplitude. If this requirement is not met, fringes on a speckle of interest will be detected at a relatively lower S/N than the speckle (with respect to photon noise). If the fringe S/N is too low, we may not be be able to sufficiently measure/subtract the speckle; photon noise will propagate through the measurement algorithm (and in some cases may even be amplified) and bias the applied correction. Integrating longer will increase the fringe S/N to mitigate this problem, but this integration time must always in principle be

shorter than the speckle lifetime (see §1.4.3.2.2); otherwise a sufficient S/N measurement/correction of a speckle can never be made before it "disappears," or evolves into another pattern. Thus, from this argument it is clear that we want to put as much light as possible through the pinhole; ideally, we want the pinhole power (i.e., [unit energy]/[unit time] integrated over the pinhole area), $P_{\rm pin}$, to be comparable to the power in the main beam of the Lyot stop, $P_{\rm Lyot}$. And, since the pinhole is considerably smaller, this requires that the flux going through the pinhole be considerably greater, $\propto (D/d)^2 \sim 340$ times greater than the flux going through the main Lyot beam. Instead, using the standard Lyot coronagraph and classical SCC described in §3.2 (which will be denoted by a "0" subscript), $(P_{\rm pin})_0/(P_{\rm Lyot})_0 \approx 2.7 \times 10^{-6}$, as illustrated in the lower left panel of Fig. 3.1. In other words, with a standard Lyot coronagraph, SCC fringe detection requires collecting more than 10^5 photons, clearly limiting operation on millisecond timescales. As a result, this discrepancy limits on-sky SCC operation to bright stars and/or long exposures. Although I have un-physically increased $P_{\rm pin}$ in the upper and lower right panels of Fig. 3.1 for visual display, we need a physical (i.e., optical) solution to realistically enable an optimal fringe S/N for millisecond timescale SCC operation. With this goal of increasing the pinhole power by a factor of ∼a million, I first tried a number of ideas, outlined below, but all of which still provided an insufficient "boost."

- **sine chop**: placing bright DM sine spots at 2.5 λ/D, right on the edge of the FPM. The rationale here is that by "chopping" a sine wave in half with the FPM, this diffracts additional light into the Lyot pinhole. However, increasing this "chopped" sine wave amplitude eventually generates nonlinear effects (e.g., second order spots); I found this amplitude limit to be around 70 nm, which set the limit on additional light that can be diffracted into the pinhole, ultimately not being enough light. Additionally, this method generated increased residual aberration around the 2.5 λ/D "chopped" sine spot location in the image, an undesirable feature for small IWA coronagraphy.

- **grating**: placing a sine wave grating (in phase) in the plane of the FPM, but only beyond the AO control region. A sine wave in the focal plane creates two copies of the coronagraphic pupil in the downstream Lyot plane (one cycle per λ/D is the lowest spatial frequency at which the two pupil copies will not be overlapping). By placing this grating beyond the AO control region, no exoplanet light within the AO control region will be sent to the pinhole. This

super-Nyquist grating also acts as a spatial filter, creating pupil copies in the Lyot plane with only super-Nyquist spatial frequencies (i.e., the PSDs of these pupil copies are ∼zero inside the Nyquist region). The amplitude of the sine wave grating was optimized to 0.3 μm; lower amplitudes produced dimmer pupil copies that put less light through the pinhole, while higher amplitudes also produced dimmer first order pupil copies in exchange for brighter second order copies (i.e., higher amplitudes caused nonlinear effects). This method was then combined with each of the two options below.

- **super-Nyquist power law phase screen (SNPS)+grating**: adding a SNPS in the coronagraph entrance pupil, upstream of the FPM (i.e., this phase screen PSD is zero at spatial frequencies inside the DM control radius). This causes an additional, controlled amount of diffracted light to to be redistributed from the PSF core to outside the DM control radius. This diffracted light in the focal plane is then again diffracted by a super-Nyquist grating as described above, creating brighter pupil copies in the Lyot plane, ultimately adding more light to the off-axis pinhole. I found this method to be successful in increasing the amount of pinhole light relative to the main Lyot beam, but still too inefficient; since the pupil copy is not concentrated into the pinhole, a significant amount of this light is still blocked by the Lyot stop. The values shown for this method use a SNPS WFE of 1.6 μm rms (the pupil copies no longer get any brighter with increased aberration, again reaching the edge of the linear regime).
- **Gauss+grating**: adding a single symmetric 2 dimensional Gaussian phase shift (with the σ parameter of this Gaussian function set to $d/2$) in the entrance pupil, upstream of the FPM. This could be enabled by a custom phase screen, and/or a DM actuator poke. The focal plane sine wave will then copy the entrance pupil Gaussian information to the off-axis pupil copy in the Lyot plane. A careful choice of where to place the Gaussian (e.g., which actuator to poke) and the focal plane sine wave frequency causes the initial aberration (which is converted from phase in the enterance pupil to amplitude in the Lyot plane by means of the FPM) in the Lyot plane pupil copy to overlap exactly at the pinhole location, similar to Figure 3.7. However, increasing the Gaussian amplitude beyond 2.6 μm ultimately did not further increase the actuator poke amplitude in the Lyot

plane pupil copy, again reaching the end of the linear regime. In this case this limit is also due to the finite amount of power in a pupil plane Gaussian influence function at super-Nyquist spatial frequencies; a focal plane grating covering the full control radius would provide more pinhole light in the Lyot plane, although this would then fringe the exoplanet light.

- **tilt**: adding a phase shift of tilt in the plane of the FPM, but only applied outside the DM control radius. Like the focal plane grating, this will also make an off-axis pupil copy in the Lyot plane, again acting as a spatial filter so that this copy only contains super-Nyquist spatial frequencies. However, there will now only be one copy and minimal higher order effects. With this approach I considered the same two combinations as with the above-described focal plane grating option.
  - **SNPS+tilt**: I used the same SNPS setup as in the above-described SNPS+grating scenario. In addition to adding tilt, in an attempt to concentrate the light from this off-axis pupil copy through the pinhole, I tried adding annular Zernike polynomials (Mahajan, 1981), similarly applying a phase shift outside the control radius. The rationale here is that classical Zernike polynomials are not optimized for this super-Nyquist geometry and do not have enough power in low order modes at these super-Nyquist spatial frequencies to significantly concentrate the pupil copy in the Lyot plane, whereas this may work better with annular Zernikes. However, my simulations did not show any significant improvement in concentrating the light with annular Zernikes in the focal plane after a grid search using focus or spherical aberration. My general findings here are thus consistent with the hypothesis that there is simply not enough low-order information at these super-Nyquist separations ($> 16\lambda/D$) to concentrate this light in the Lyot plane to sub-pupil scales, although an additional range or basis of low order modes with this geometry could be further explored in a future paper.
  - **Gauss+tilt**: the same setup as the Gauss+grating method; the only difference is that the super-Nyquist focal plane grating is replaced with a super-Nyquist focal plane tilt, as illustrated in Figure 3.7.

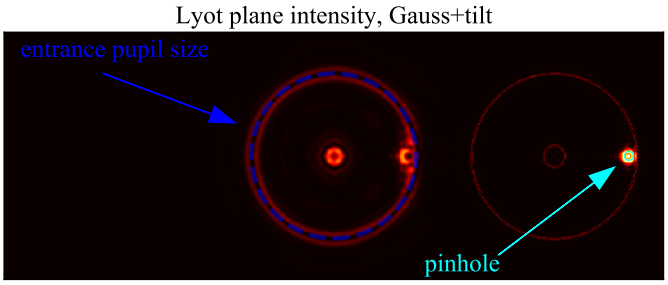


Figure 3.7 The Lyot plane intensity of the Gauss+tilt method. The Gaussian aberration in the entrance pupil is copied into the off-axis location in the Lyot plane by means of a super-Nyquist tilt phase shift applied in the plane of the FPM. The entrance pupil Gaussian aberration location is chosen so that it overlaps with the pinhole position in the Lyot plane. A residual Gaussian aberration is still present in the main beam of the Lyot plane, but this could be blocked by a custom Lyot stop.

Table 3.1: The power transmitted through the SCC pinhole relative to the pupil in the same Lyot plane for a given method (top row) and relative to the pinhole from the simple Lyot coronagraphic setup described in §3.2 (bottom row). The performance metrics are shown from simulations without any additional entrance pupil phase or amplitude aberration. All methods other than the TG FPM use the simple Lyot coronagraph as described in §3.2.

method	**normal**	**sine chop**	**Gauss+ grating**	**Gauss+ tilt**	**SNPS+ grating**	**SNPS+ tilt**	**TG FPM**
$\frac{P_{\text{pin}}}{P_{\text{Lyot}}}$	2.7e-06	2e-05	0.02	0.088	0.0014	0.0048	0.42
$\frac{P_{\text{pin}}}{(P_{\text{pin}})_0}$	1	8.4	1.1e+04	4.4e+04	8.1e+03	2.1e+04	1.4e+05

Table 3.1 summarizes the results from these above ideas in addition to my first optimized solution later presented in §3.5.1.1, called the TG FPM. It shows that a few other contenders are within an order or magnitude of my TG FPM solution: Gauss+grating, SNPS+tilt, and Gauss+tilt. Of particular interest is the Gauss+tilt method, within a factor of ∼3 in $P_{\text{pin}}/(P_{\text{pin}})_0$ from the TG FPM. The $P_{\text{pin}}/P_{\text{Lyot}}$ value for this method is low only because of continued use of the generic Lyot stop in these basic simulations; Figure 3.7 shows the Lyot plane for this method and that a custom Lyot stop could block the Gaussian poke at the edge of the main Lyot beam, likely increasing $P_{\text{pin}}/P_{\text{Lyot}}$ to a value comparable to the TG FPM. Also note that all methods in Table 3.1 (i.e., other than the TG FPM) are in principle compatible with any other focal plane amplitude/phase mask coronagraphs; the close performance of the Gauss+tilt method to the TG FPM has important implications in terms of coronagraphic compatibility with both current and future instruments.

3.5 The Fast Atmospheric SCC Technique

3.5.1 Principle

In this section I will both present and motivate the need for a FAST solution, which includes my newly proposed TG FPM and a speckle subtraction strategy designed for speckle evolution on millisecond timescales. Detailed FAST solutions for post-processing with CDI and active focal plane wavefront control will be presented later in §3.5.2 and §3.5.3, respectively.

3.5.1.1 The TG FPM

As introduced in §3.1, the conventional SCC design will only work for long exposures to measure quasi-static aberration. This restriction arises from the amount of light transmitted through the pinhole in the Lyot plane. Previous SCC papers have focused on correcting quasi-static aberration on the order of a few nanometers rms because of this limitation; here, I now address wavefront control with the SCC more generically, by measuring and correcting for speckles on all time scales, including millisecond-lifetime atmospheric speckles.

Assuming that we want to detect and subtract speckles as soon as they are visible and before they disappear, when only a few photons are recorded for a speckle, we ideally want the fringe amplitude to be at least as large as the amplitude of the speckle that we want to remove, requiring that the envelope of the pinhole PSF match the amplitude and power law of the stellar speckles, as discussed in §3.4. If this requirement is not met, fringes on a speckle of interest will be detected at a relatively lower S/N than the speckle, which will generate a noisier/biased calibration. Integrating longer will increase the fringe S/N to mitigate this problem, but this integration time must always in principle be shorter than the speckle lifetime (§1.4.3.2.2); otherwise a sufficient S/N measurement of a speckle's fringes can never be made before the speckle "disappears."

I ultimately found that the most successful way to reach sufficient fringe S/N was to design a new coronagraph, called the TG FPM, which is outlined below. I replaced the amplitude mask of the FPM in a simple Lyot coronagraph and instead applied a tilt to separations between zero and three λ/D, creating an off-axis pupil copy in the Lyot plane. This FPM acts as a spatial filter for the downstream off-axis pupil copy, filtering only low order modes ($\leq$ 3 cycles/pupil). After applying a grid search of

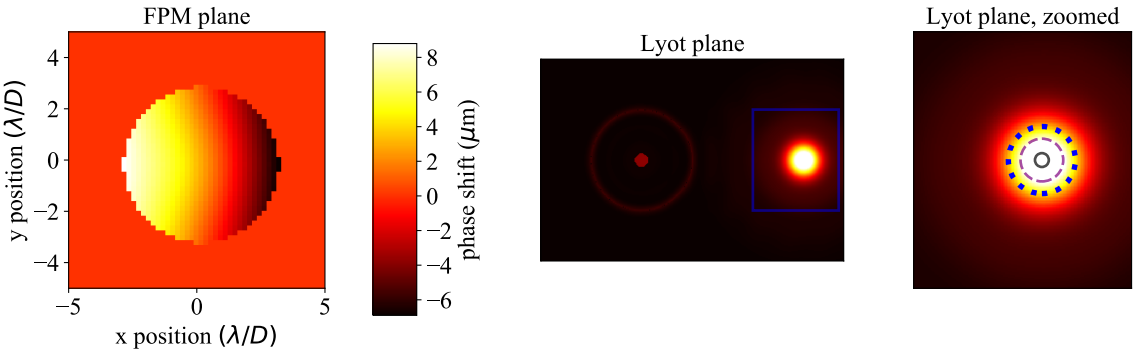
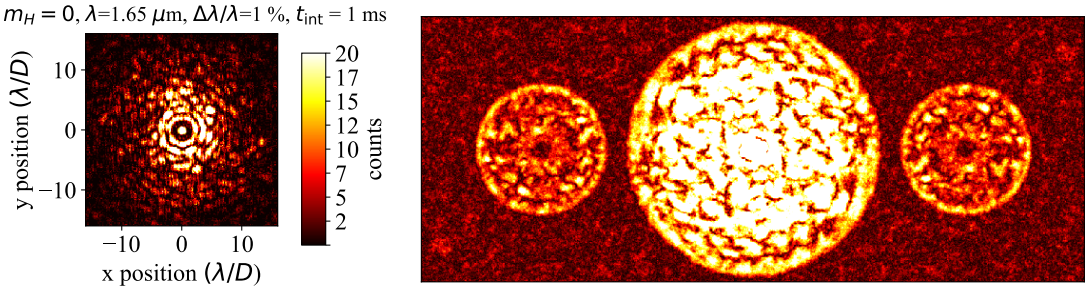


Figure 3.8 Simulations of the TG FPM. Left: instead of using an amplitude FPM, a phase shift is applied in the focal plane using a coronagraphic mask called the TG FPM. Middle: the Lyot plane intensity after using the TG FPM. The main pupil is on the left, and the off-axis pupil created from the TG FPM is on the right. Right: a zoomed region of the blue rectangle in the middle panel. The dashed purple circle shows the theoretical minimum size of this off-axis pupil copy coming from the 6 λ/D diameter TG FPM compared to our measured FWHM (the dotted blue circle), illustrating that we are within a factor of about 2 from this limit. The solid black circle shows the size of the Lyot stop pinhole for comparison, still well below the diffraction limit (the FPM inner working angle for a diffraction limit corresponding to the maximum pinhole size is 9.3 λ/D). The integrated intensity in the Lyot pinhole with this FPM is increased by a factor of about 3.2×10^5 compared to the normal (unapodized) Lyot coronagraph.

amplitude and width parameters for various additional low order aberrations, I found the best fringe amplitude results when using a two-dimensional symmetric Gaussian with a 2.9 μm amplitude and 4.7 λ/D FWHM, shown in combination with the FPM tilt in Figure 3.8 along with the corresponding Lyot plane intensity. The right panel of Figure 3.8 shows the diffraction limit of a 6 λ/D diameter FPM (i.e., for a telescope with diameter $D/6$), and illustrates that my best solution is within a factor of 2 from this limit. Although my TG FPM solution redistributes more light towards the centre of the off-axis pupil, I am not actually "shrinking" this pupil; this would require either a smaller telescope and/or a mask that completely cancels the smallest spatial scales of the complex electric field in the focal plane.

A significant amount of light is still blocked from going through the pinhole. The pinhole cannot be made larger due to the limitations of equation 3.2 (which requires that the 1.22 λ/d minimum of the pinhole PSF lie outside the AO control region, where d is the pinhole diameter remapped to the entrance pupil). This maximum pinhole size is also impossible to reach with my 6 λ/D diameter TG FPM design,

illustrated in the right panel of Figure 3.8; the diffraction limit is $D/6$ whereas the largest pinhole size is (from equation 3.2) $D/18.5$ (i.e., reaching this diffraction limit would require an 18.5 λ/D diameter FPM, or a 9.3 λ/D IWA). However, despite these limitations, utilizing the TG FPM design in Figure 3.8 causes a sufficient amount of light to be transmitted through the pinhole to enable fringe detection for exposures on the order of milliseconds; the image and MTF for a 1 ms exposure of a $m_H = 0$ star using the new TG FPM design are shown in Figure 3.9, illustrating that the higher spatial frequency MTF sidelobes are detected above the photon noise floor.



(a) (b)

Figure 3.9 (a) A simulated SCC image, with photon noise, for a 1 ms exposure of a $m_H = 0$ star using the TG FPM design shown in Figure 3.8 and the setup described in §3.2. (b) The MTF of (a), showing that fringes (i.e., the higher spatial frequency side-lobes), are detected above the background photon noise floor.

3.5.1.2 Photon Noise

To understand how photon noise is propagated through the subtraction algorithm in equation 3.6, independent of the pinhole PSF estimation, I use a noiseless I_R but in addition a noisy target image, im_{noisy} (i.e., photon noise is not simulated for I_R but is simulated for the "im_{noisy}" SCC image), to estimate the fringe amplitude, $|I_-|_{\text{est}}$. By decomposing the noisy image into a noiseless component and a noise-only component

(i.e., $\mathrm{im}_{\mathrm{noisy}} = \mathrm{im}_{\mathrm{noiseless}} + [\mathrm{im}_{\mathrm{noisy}} - \mathrm{im}_{\mathrm{noiseless}}]$) the subtracted noisy image is then

$$\begin{aligned}
\mathrm{im}_{\mathrm{subt,\ noisy}} &= \mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_2\} - \left(\frac{|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_1\}|^2}{I_R} + I_R\right) \\
&= \mathrm{IFT}\{\mathrm{FT}\{(\mathrm{im}_{\mathrm{noiseless}} + [\mathrm{im}_{\mathrm{noisy}} - \mathrm{im}])\}\, m_2\} \\
&\quad - \left(\frac{|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_1\}|^2}{I_R} + I_R\right) \\
&= \mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noiseless}}\}\, m_2\} + \mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}} - \mathrm{im}_{\mathrm{noiseless}}\}\, m_2\} \\
&\quad - \left(\frac{|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_1\}|^2}{I_R} + I_R\right) \\
&= (I_S + I_R + I_P) + \mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}} - \mathrm{im}_{\mathrm{noiseless}}\}\, m_2\} \\
&\quad - \left(\frac{|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_1\}|^2}{I_R} + I_R\right) \\
&= I_P + \mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}} - \mathrm{im}_{\mathrm{noiseless}}\}\, m_2\} \\
&\quad - \left(\frac{|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_1\}|^2}{I_R} - I_S\right) \\
&= I_P + \mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}} - \mathrm{im}_{\mathrm{noiseless}}\}\, m_2\} \\
&\quad - \left(\frac{|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\, m_1\}|^2 - |\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noiseless}}\}\, m_1\}|^2}{I_R}\right) \\
&\equiv I_P + T_1 - T_2, \qquad (3.8)
\end{aligned}$$

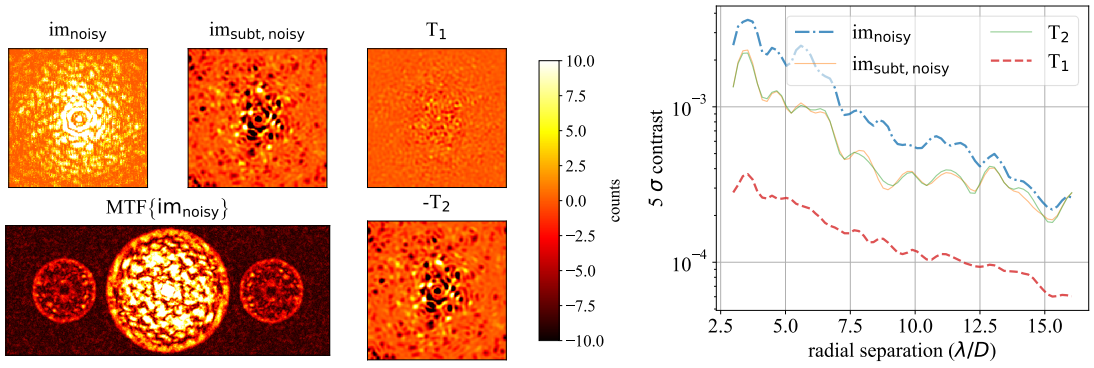
where the second to last line of equation 3.8 uses the numerical result from equation 3.7. Equation 3.8 illustrates that the subtraction algorithm presented above (when using a simultaneous noiseless pinhole PSF and if no exoplanet is simulated) *is* the photon noise limit and that two different terms contribute to this limit:

1. T_1 is the photon noise limit from a noisy SCC image, Fourier filtered to remove fringes according to Figure 3.4 b so that no spatial scales exist in the image that are smaller than λ/D, and

2. T_2 is the photon noise limit of the reconstructed I_S (i.e., starlight) term from a noisy SCC image, Fourier filtered to isolate the fringe amplitude according to Figure 3.4 a.

Images and contrast curves for $\mathrm{im}_{\mathrm{noisy}}$, $\mathrm{im}_{\mathrm{subt,\ noisy}}$, T_1, and T_2 are shown in Figure 3.10 a and b, respectively, for a 1 ms exposure on a $m_H = 0$ star along with the default parameters from §3.2. A traditional amplitude FPM, Lyot stop, and classical

SCC does not diffract enough light into the off-axis Lyot pinhole to see any fringes above the photon noise for a ms exposure, and so, as in Figure 3.1, I unphysically enhance the pinhole intensity in Figure 3.10 by a factor of 2.5×10^5, analogous to the TG FPM solution presented in §3.5.1.1.



(a) (b)

Figure 3.10 (a) Assuming perfect knowledge of the pinhole PSF, various simulated terms from equation 3.8 are shown for a 1 ms exposure, with photon noise, for a $m_H = 0$ star, with the pinhole intensity enhanced by a factor of 2.5×10^5, and with all other parameters as the default values from §3.2. The panel shows the recorded noisy image (upper left), the subtracted image/photon noise limit (upper middle), the photon noise limit from a noisy SCC image, Fourier filtered to remove fringes so that no spatial scales exist that are smaller than λ/D (T_1, upper right), and the photon noise limit of the reconstructed I_S term from a noisy SCC image, Fourier filtered to isolate the modulation amplitude ($-T_2$, lower right). For all these images the FOV is 32×32 λ/D and a colour bar is shown on the right, illustrating the number of photons recorded and remaining in the subtracted images. The MTF of the upper left image is also shown in the lower left, illustrating the relatively higher impact of photon noise on the sidelobes (i.e., the fringe term) compared to the central beam. (b) The contrast curves for the images in panel a.

Figure 3.10 illustrates that photon noise in exposures on the order of milliseconds, even on bright stars, sets a fundamental limit on the achievable contrast with this coronagraphic image reconstruction technique. Comparing the two terms that contribute to this photon noise limit, the noise from T_2 dominates over the noise from T_1, illustrating that higher spatial frequencies of the image at the MTF sidelobe positions are more affected by photon noise (i.e., in terms of contrast) than the central beam of the MTF. This greater impact of photon noise at higher spatial frequencies can be explained by the comparing the S/N of the MTF sidelobe vs. the central beam in Figure 3.10 a: although the MTF sidelobes are clearly detected above the background

photon noise, the central beam is still significantly brighter than the sidelobes and thus detected at a relatively higher S/N. Accordingly, this higher spatial frequency, pixel-to-pixel noise in the image ultimately propagates through the $|I_-|$ reconstruction algorithm to λ/D spatial-scales in $|I_-|_{\text{est}}$, causing λ/D-scale residuals in T_2 that dominate over T_1 instead of just pixel-to-pixel dominated photon noise. Ultimately, the relative impact of T_1 vs. T_2 will depend on the relative amplitude of the fringe term and thus the intensity through the Lyot stop off-axis pinhole (i.e., if the MTF sidelobes in Figure 3.10 a were brighter, or detected at a higher S/N, T_2 would have a relatively lower contribution to the overall $T_1 - T_2$ photon noise limit). Although the pinhole light for the simulations in this section is enhanced by a factor of 2.5×10^5 (in intensity) for illustration, this is set to a similar value as provided by the TG FPM (as shown in Table 3.1). Thus, the illustration here showing that the noise from T_2 dominates over T_1 in millisecond exposures at the 100 nm rms aberration level will also hold throughout this chapter.

Figure 3.10 b and equation 3.8 show that a direct measurement of the pinhole PSF at infinite S/N would allow subtraction to the photon noise limit (by definition), since $\text{im}_{\text{subt, noisy}} = T_1 - T_2$ (if no exoplanet is simulated). With a noiseless pinhole PSF, the photon noise limit is generated entirely from Poisson noise and linear operators, suggesting that the $\text{im}_{\text{subt, noisy}}$ output of each new recorded image should be completely uncorrelated from one another and that a continuous stacking of these subtracted ms images would improve contrast proportional to $t^{-0.5}$. Thus, $T_1 - T_2$ or its filtered version (see Fig. 3.11) is shown in plots as "the photon noise limit" throughout this chapter. Regarding additional noise propagated through my subtraction algorithms (§3.5.2.1) as the result of an imperfect estimate of the pinhole PSF, looking at the second to last line of equation 3.8, we can see that as long as my estimate of the pinhole PSF is

1. symmetric (in pixel value distribution) around the true noiseless I_R value, and

2. uncorrelated from frame to frame each time a new noisy image is generated,

the contrast should still improve proportional to $t^{-0.5}$ simply by stacking subtracted images, although at a slightly worse contrast from the photon noise limit. However, as in §3.5.2.2 and 3.5.2.3, in some cases one or both of these conditions may not be met, ultimately limiting the achievable contrast even after an infinite exposure time.

3.5.1.3 Long Exposure Simulations and Analysis

In this section I argue that even with the advent of my new TG FPM design, the typical CDI and/or focal plane wavefront control strategy of taking long exposures to measure and correct for quasi-static aberration is still limited by the atmosphere. Using the simulation parameters and TG FPM design presented in §3.2 and §3.5.1.1, respectively, Figure 3.11 illustrates the contrast limitations for a 30 second exposure of a $m_H = 0$ star; it shows (a) the recorded image, (b) a perfect subtraction of everything that is fringed in panel a (i.e., using equation 3.8, assuming a noiseless, simultaneous measurement of the pinhole PSF during that 30 second exposure), (c) a high-pass filter of this subtracted image using a 2×2 λ/D median boxcar filter (see Eq. 3.15), (d) a contrast curve of these three images compared to the input and photon noise limit images, and (e) the contrast at 10 λ/D vs. time of subtracted, high-pass-filtered images compared to the photon noise limit for both $m_H = 0$ and $m_H = 5$ stars.

I find that the contrast in Figure 3.11 b is about 50% of the contrast in the input image (a), a non-negligible component of the image that is 'invisible" to the SCC (or any CDI algorithm) because fringed atmospheric speckles are blurred out over the long exposure. The spatial scales in the residual atmospheric halo of Figure 3.11 b, labeled "output," suggests that a high-pass filter of this image will improve the contrast, and is thus shown to the right as "$F_{\mathrm{HP}}\{\mathrm{output}\}$." Indeed, Figure 3.11 d shows that high-pass filtering does improve the contrast by up to a factor of about 5 throughout the AO control region, although this is still a factor of about 3-20 above the photon noise limit. This contrast curve also shows that applying the same high-pass filter to the normal photon noise limit, derived from Eq. 3.8, has a negligible effect.

As discussed in §1.4.3.2.2, Figure 3.11 e illustrates the conclusions from Macintosh et al. (2005). For a 5th magnitude star, $F_{\mathrm{HP}}\{\mathrm{output}\}$ is photon noise-dominated and so contrast continually improves proportional to $t^{-0.5}$. For the 0th magnitude star, $F_{\mathrm{HP}}\{\mathrm{output}\}$ is flat and dominated by atmospheric speckles at $t < \tau_{\mathrm{spec}} \approx 200$ ms, and averaging proportional to $(t/\tau_{\mathrm{spec}})^{-0.5}$ at $t > \tau_{\mathrm{spec}}$. More generally, note that although the level of residual atmospheric correlation/speckle lifetime will change with varying observing conditions (causing a variable gap in contrast), the main point of Figure 3.11 is to demonstrate the following averaging properties for atmospheric speckles during long exposures:

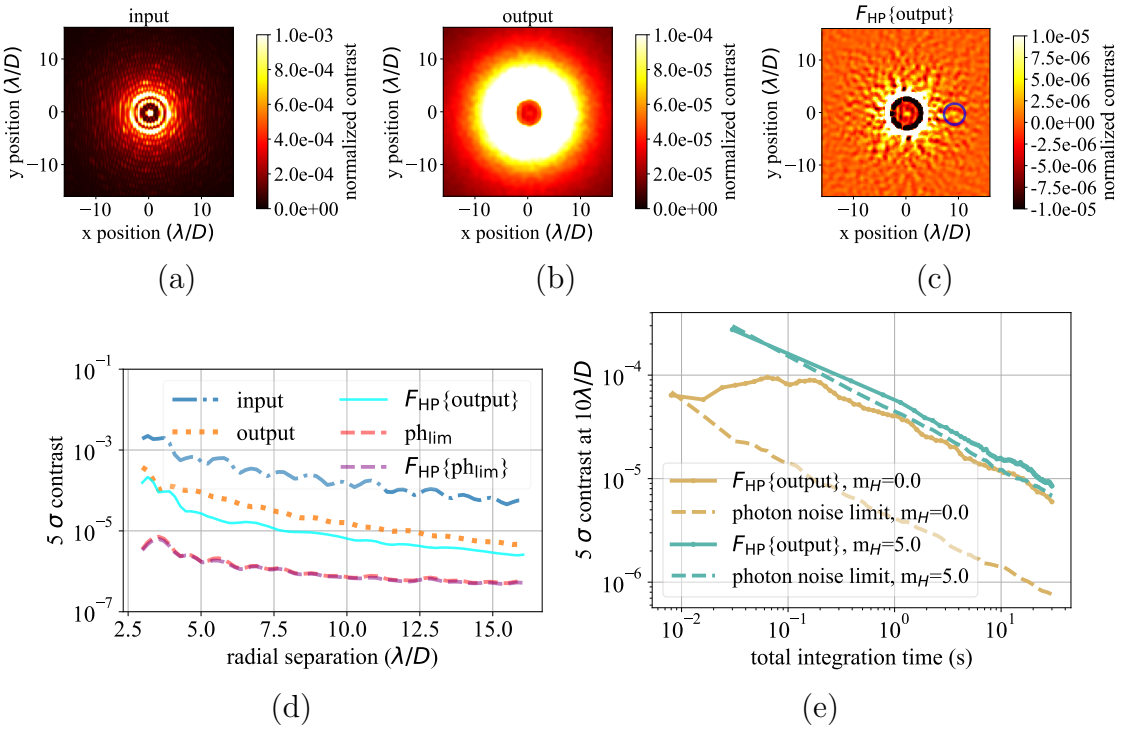


Figure 3.11 (a) A simulated image for a 30 second exposure of a $m_H = 0$ star. (b) A PSF subtraction of the left image using equation 3.6, assuming a perfect simultaneous measurement of the pinhole PSF during that 30 second exposure. (c) a high-pass filter of panel b; a simulated exoplanet at $(x, y) = (10, 0)$ λ/D at a flux-normalized contrast of 10^{-5} is identified by the blue circle. (d) a contrast curve of panels a - c compared to the photon noise limit images (labeled "ph_{lim}"). (e) Contrast at 10 λ/D vs. cumulative integration time after applying the same subtraction algorithm as used on panel c for both 0th and 5th magnitude stars; the photon noise limit is also shown for comparison to each star.

1. Atmospheric speckles leave a residual halo that cannot be measured optically by the SCC or any CDI algorithm.

2. Point 1 limits the achievable contrast, even after post processing, to a non-negligible factor above the photon noise limit.

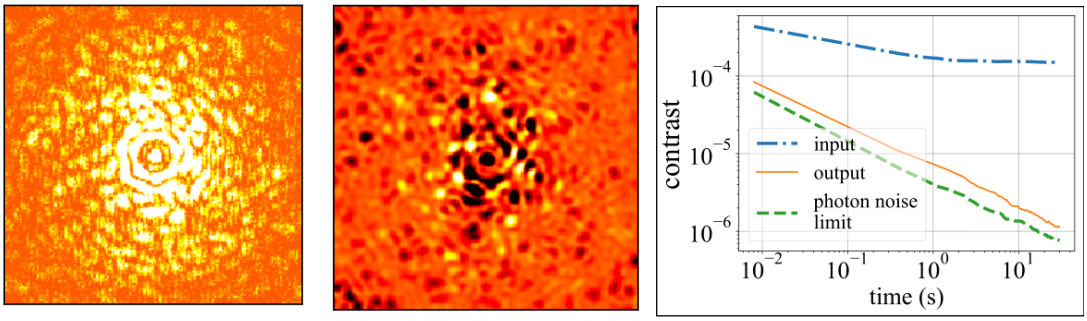
3. On dim stars, a negligible amount of photons are recorded per speckle lifetime, and so 1 is either at or buried within the photon noise limit.

Points 1-2 above illustrate a fundamental limitation on achievable raw contrast with any CDI algorithm during a long exposure (e.g., Sauvage et al., 2012). Thus, my next FAST approach is designed to minimize this effect by running fast exposures

every few milliseconds, subtracting these recorded images, and then stacking the subtracted residuals to obtain contrast improvement proportional to $t^{-0.5}$.

3.5.1.4 FAST Speckle Subtraction Strategy

The FAST solution, in contrast to the limitations set by AO residuals presented in §3.5.1.3, is illustrated in Fig. 3.12 and outlined below:



(1) (2) (3)

Figure 3.12 The FAST solution: (1) detect SCC fringes in a millisecond exposure (as in Fig. 3.9), freezing AO residuals and quasi-static speckles, (2) subtract any stellar speckles (i.e., a combination of AO residuals and quasi-static speckles) detected in step 1 using the coherent fringes also detected in step 1, and (3) stack the subtracted residuals from step 2 over time, continually gaining in contrast near the photon noise limit (in comparison to saturated raw contrasts from un-subtracted quasi-static speckles over long exposures, as illustrated by the dash-dotted blue curve).

1. A new coronagraphic mask (§3.5.1.1, §3.5.4.2) enables SCC fringe detection for millisecond-timescale exposures, freezing AO residuals and quasi-static speckles in a single wavefront realization.

2. The fringes from point 1 enable a subtraction of all coherent stellar speckles, either via CDI (§3.5.2) and/or active wavefront control (§3.5.3).

3. Stacking the residuals from point 2 over time continually improves contrast near the photon noise limit (§3.5.2.3, §3.5.3.2.2), leading to images that are no longer limited by a "halo" of residual AO and/or quasi-static speckles.

3.5.2 CDI

In this section I will consider a few different approaches to FAST CDI post-processing strategies, first presented in Gerard et al. (2018a). Note that an additional CDI post-

processing strategy is presented later in §3.5.5.3.

3.5.2.1 Estimation of the Pinhole PSF

As described in §3.1.1, two different Fourier filtering algorithms on the recorded SCC image, I, provide

$$|I_-| \equiv |\mathrm{IFT}\{\mathrm{FT}\{I\}\, m_1\}| = \sqrt{I_S\, I_R}, \text{ and} \tag{3.9}$$

$$\mathrm{IFT}\{\mathrm{FT}\{I\}\, m_2\} = I_S + I_R + I_P. \tag{3.10}$$

From these two equations with three unknowns, we cannot retrieve the exoplanet image I_P. Baudoz et al. (2012a) approach this problem by recording a calibration image of I_R (i.e., the pinhole PSF), using an internal source and occulting the main beam of the Lyot stop. This is a viable approach to reconstructing the speckle pattern I_S induced by small and/or static aberrations. However, it may not be a viable approach for measuring atmospheric speckles on millisecond timescales unless the electric field transmitted through the pinhole is relatively stable across multiple realizations of a translating atmospheric phase screen. I will discuss this calibrated reference imaging approach later in §3.5.2.1.2. First, in §3.5.2.1.1, I will discuss the approach of direct, on-sky measurement of the pinhole PSF. Finally, in §3.5.2.1.3 I will present a filtering algorithm, using only the recorded on-sky SCC image to reconstruct the pinhole PSF.

3.5.2.1.1 Direct Measurement Obtaining a high, simultaneous S/N image of the pinhole PSF along with a noisy target image should result in a contrast improvement proportional to $t^{-0.5}$ (see §3.1.1 and §3.5.1.2 for a further discussion). However, in reality this measurement of the pinhole PSF will also be affected by photon noise due to the finite amount of light transmitted through the SCC pinhole. Using the TG FPM, I simulate a separate pinhole PSF measurement assuming use of a 50/50 beam splitter placed on the pinhole: 50% of the light going through the pinhole is transmitted through the beam splitter and focused onto the science camera, while the other 50% is redirected by a small angle (i.e., just downstream of the pinhole) and focused onto the same camera. This would also require a field stop in an upstream focal plane (either at or downstream from the FPM plane) to block any stellar background that would otherwise prevent any pinhole PSF light from being recorded with a reasonable S/N. The beam splitter will induce quasi-static and chromatic aberra-

tions on the reference beam, but I do not simulate them here because their effects should be second order as the beam splitter is downstream from the FPM and can be super-polished to extremely high level precision.

With the setup discussed above and in §3.2, the pinhole PSF for a one millisecond exposure of a m_H=0 star is shown in Figure 3.13 a. Although as many as four

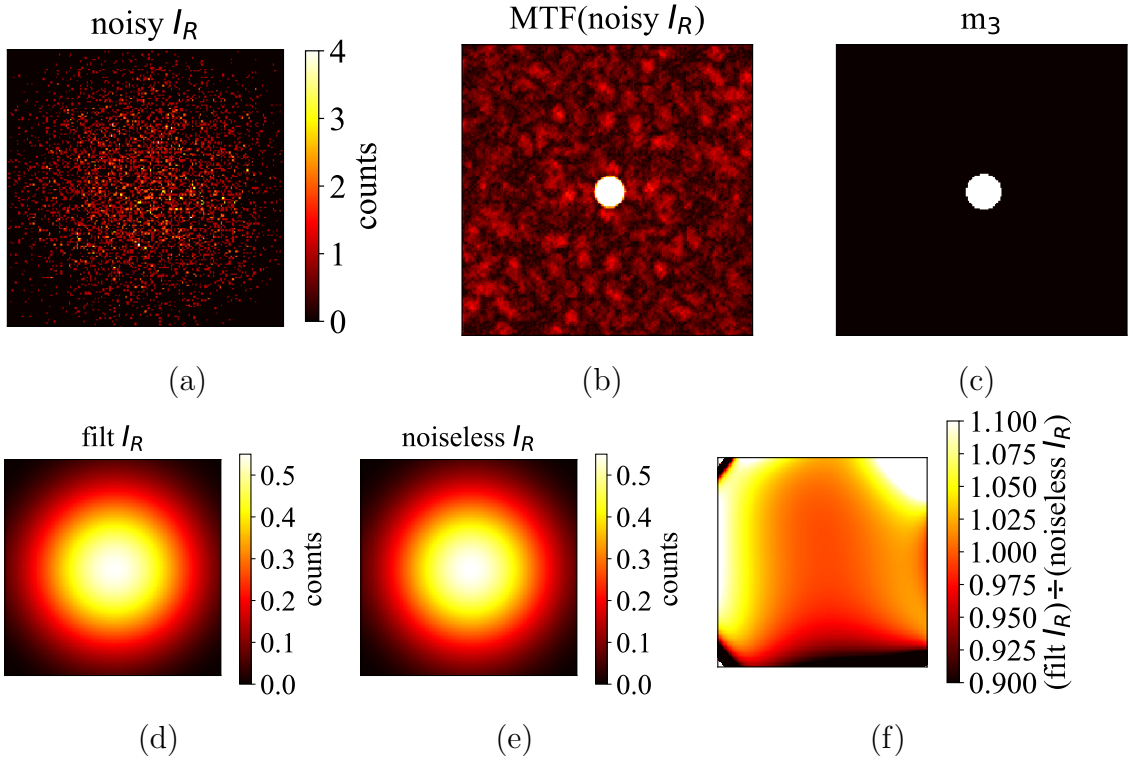


Figure 3.13 (a) a recorded pinhole PSF, with photon noise, for a 1 ms integration for a m_H=0 star using the TG FPM. (b) the MTF of panel a. (c) a binary mask used to Fourier filter panel a so that photon noise is suppressed on spatial-scales smaller than λ/d. (d) the result of Fourer filtering panel a: filt $I_R = \mathrm{IFT}\{\mathrm{FT}\{\mathrm{noisy\ I}_R\}m_3\}$. (e) a recorded pinhole PSF where no photon noise is simulated. (f) the ratio of panels d to e.

photons per pixel are recorded in a one millisecond exposure, photon noise still clearly dominates compared to the image without photon noise in Fig. 3.13 e. However, looking at the MTF of Fig. 3.13 a—in Fig. 3.13 b—shows that only the central region of Fourier space has physical representation of spatial-scales larger than λ/d (i.e., the central region of the MTF), while all other higher spatial frequencies in the MTF are only from recorded photon noise. Thus, using the Fourier filter m_3 in Fig. 3.13 c allows the noisy image in Fig. 3.13 a to be converted to the Fourier-

filtered image in Fig. 3.13 d. The ratio of Fig. 3.13 d to e, which is illustrated in Fig. 3.13 f, shows that this reconstruction, even for an extremely low number of recorded photons is still accurate to about ±10%. As long as this offset is random on a frame-by-frame basis (i.e., generated from photon noise), stacking the output of a subtracted image using equations 3.9 and 3.10 (i.e., where a filtering algorithm as in 3.13 d is used to estimate the pinhole PSF) will continue to average randomly (see §3.5.1.2 for a further discussion). Only a static offset generating speckles that are consistently over- or under-subtracted will create static residuals after stacking. This same principle applies to the other two pinhole PSF reconstruction methods to be presented in §3.5.2.1.2 and §3.5.2.1.3. Intuitively, there is no reason to believe here that any static offset would be present in Figure 3.13 e, since these residuals are generated from photon noise present at the same low spatial frequencies as the pinhole PSF, and that low-order diffraction and/or atmospheric effects from the FPM are unlikely to cause such static offsets (but see §3.5.2.2).

3.5.2.1.2 Calibration Algorithm Similar to the approach in Baudoz et al. (2012a), a daytime pinhole PSF can be recorded with very high S/N using a bright internal source and masking the main beam in the Lyot plane. However, this recorded calibration pinhole PSF will be without

1. an intrinsic flux normalization, unlike the direct measurement in §3.5.2.1.1, and
2. effects from residual AO atmospheric turbulence.

The implications of point 2 above are that additional low order aberrations from the atmospheric phase screen (i.e., at spatial frequencies of less than 3 cycles/pupil for a 6 λ/D diameter TG FPM) that are not present in the calibration pinhole PSF will ultimately limit the achievable contrast with this method. This limitation arises from non-centro-symmetric features of the on-sky pinhole PSF relative to the static pinhole PSF (discussed further in §3.5.2.2). Furthermore, I simulate the static phase screen differences between a daytime calibration and on-sky measurement by using a different specific wavefront map of static phase and amplitude aberration for the calibration pinhole PSF vs. the on-sky static component, where both are still normalized to the same RMS values as described in §3.2 but are fully de-correlated from one another. Then, assuming the pinhole PSF—I_R—is properly calibrated with this approach, equations 3.9 and 3.10 allow a subtraction of I_S and I_R without subtracting I_P.

3.5.2.1.3 Reconstruction Algorithm In this section I will revisit the underdetermined system of equations 3.9 and 3.10 (i.e., two equations and three unknowns). If I_P could be removed from equation 3.10, we would have a determined system. Noting that the spatial scale of I_R is limited to λ/d, running a low-pass filter on equations 3.9 and 3.10 to retain spatial scales larger than or equal to λ/d should have the effect of "filtering the exoplanet light out of the image," thus providing two equations with two unknowns and a solution for I_R. These steps are outlined below, where $F_{\mathrm{LP}}\{\}$ is a linear low-pass filter operator:

$$\begin{aligned} \mathrm{I}_{\mathrm{m,\,lf}} \equiv F_{\mathrm{LP}}\{|I_-|\} &= F_{\mathrm{LP}}\{\sqrt{I_S\, I_R}\} \\ &\approx F_{\mathrm{LP}}\{\sqrt{I_S}\}\sqrt{I_R},\ \text{and} \qquad (3.11) \\ \mathrm{im}_{\mathrm{lf}} \equiv F_{\mathrm{LP}}\{\mathrm{IFT}\{\mathrm{FT}\{I\}m_2\}\} &= F_{\mathrm{LP}}\{I_S + I_R + I_P\} \\ &\approx F_{\mathrm{LP}}\{I_S\} + F_{\mathrm{LP}}\{I_P\} + I_R \\ &\approx F_{\mathrm{LP}}\{I_S\} + I_R. \qquad (3.12) \end{aligned}$$

I found that using a 3.9×3.9 λ/D median boxcar filter for $F_{\mathrm{LP}}\{\}$ provided both a good contrast and algorithmic exoplanet throughput (see §3.2 for my formal definition), although these metrics were not optimized with any tolerance requirements.

The $F_{\mathrm{LP}}\{\}$ is removed from I_R because a 3.9 λ/D median boxcar kernel has a negligible impact on the smallest λ/d scales ($\lambda/d = 37\ \lambda/D$ in my simulations) of I_R. Additionally, in equation 3.12 I assume that $F_{\mathrm{LP}}\{I_P\} \approx 0$ because the central λ/D-sized core of the exoplanet PSF is removed by the low pass filter (only about 2% of the peak flux in an exoplanet core remains after applying the low pass filter), and I ignore the higher order halo terms of the exoplanet PSF in the context of point source detection (Perrin et al., 2003). However, I do not make the same argument for the low-pass filtered starlight term, $F_{\mathrm{LP}}\{I_S\}$, because this term is of similar magnitude to the I_R term; otherwise, ignoring the low spatial frequency starlight information would bias my estimate of I_R. Ultimately, the exoplanet PSF halo term is absorbed into $F_{\mathrm{LP}}\{I_S\}$ and/or I_R, which will eventually cause static biases in the regime of high exoplanet S/N. Thus, combining equations 3.11 and 3.12 (now two equations and two unknowns), I find that

$$I_R = \frac{\left(\mathrm{im}_{\mathrm{lf}} - \sqrt{\mathrm{im}_{\mathrm{lf}}^2 - 4\ \mathrm{I}_{\mathrm{m,\,lf}}^2}\right)}{2}. \qquad (3.13)$$

Finally, combining equations 3.9, 3.10, and 3.13 then allows subtraction of all of the terms in equation 3.10 except for I_P.

3.5.2.2 Flux Normalization and Static Limitations

In this section I present a flux normalization procedure and discuss static limitations of the subtraction algorithms presented in §3.5.2.1. For all algorithms, I found additional static systematic offsets that ultimately prevented contrast improvement below a certain level. For this reason, I implemented two additional steps to flux normalize the reconstructed image without subtracting any additional exoplanet flux:

1. Annular flux normalization by summation:

$$\begin{aligned} \mathrm{im_{subt}}(r) &= \mathrm{IFT}\{\ \mathrm{FT}\{\mathrm{I_{noisy}}\}\mathrm{m_2}\} - c_{\mathrm{norm}}(r)\left[|\mathrm{IFT}\{F\{\mathrm{I_{noisy}}\}\mathrm{m_1}\}|^2/\widehat{I}_R + \widehat{I}_R\right], \text{ where} \\ \mathrm{c_{norm}}(r) &= \sum_r \left[\mathrm{IFT}\{\ \mathrm{FT}\{\mathrm{I_{noisy}}\}\mathrm{m_2}\}\right] / \sum_r \left[|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{I_{noisy}}\}\mathrm{m_1}\}|^2/\widehat{I}_R + \widehat{I}_R\right], \end{aligned} \tag{3.14}$$

   where $\mathrm{I_{noisy}}$ represents the recorded noisy SCC image, $\widehat{I}_R$ represents the estimated/calibrated pinhole PSF via any of the algorithms presented in §3.5.2.1, and $\sum_r$ denotes a sum of all the pixels within 0.3 λ/D of a given separation r. Equation 3.14 is equivalent to forcing the mean of the output subtracted image, $\mathrm{im_{subt}}$, to 0. To note, even though the target image, $\mathrm{IFT}\{\ \mathrm{FT}\{\mathrm{I_{noisy}}\}\mathrm{m_2}\}$, includes exoplanet light and therefore biases the normalization coefficient at the separation of the exoplanet, I found about 0-15% algorithmic exoplanet throughput losses from this effect for my simulated exoplanets at flux-normalized contrasts levels of 10^{-5} and 5×10^{-5} (see §3.5.2.3), and otherwise found significant improvement of normalization by summation vs., e.g., robust standard deviation.

2. High-pass filtering the already-flux-normalized image from equation 3.14, but only after stacking images to the desired integration time. The annular normalization approach in step 1 will not remove symmetric, low-order residual aberrations that can build up over time because of small calibration errors, e.g., as discussed in §3.5.2.1.2. Thus, I use an aggressive 2.1 × 2.1 λ/D median boxcar filter to generate a low-pass-filtered version of the flux-normalized

subtracted image, $F_{\mathrm{LP}}\{\mathrm{im}_{\mathrm{subt}}\}$, and then high-pass filter the image via

$$F_{\mathrm{HP}}\{\mathrm{im}_{\mathrm{subt}}\} = \mathrm{im}_{\mathrm{subt}} - F_{\mathrm{LP}}\{\mathrm{im}_{\mathrm{subt}}\}. \tag{3.15}$$

Equation 3.15 is the final form of stacked images that are displayed and shown in contrast curves of §3.5.2.3 for the calibration and reconstruction algorithms. I found that annular normalization degraded contrast in the case of the direct pinhole PSF measurement algorithm, and so for this algorithm I only use a high-pass filter (i.e., step 2 but not step 1 above). However, interestingly, I did find static limitations for the direct pinhole PSF measurement algorithm when a high-pass filter was was not used, revealing the presence of static low order aberrations as a result of my Fourier filtering algorithm, which is illustrated in Figure 3.13. For an analogous comparison I also subsequently high-pass filter the photon noise limit images (see §3.5.1.2 but also Figure 3.11 which shows this has a negligible effect).

Even in the absence of photon noise, there is still a fundamental limit on the performance of the calibration and reconstruction algorithms, shown in Figure 3.14. The labels "algo," "cal," and "filt" represent the reconstruction, calibration, and direct pinhole PSF algorithms presented in sections 3.5.2.1.3, 3.5.2.1.2, and 3.5.2.1.1, respectively. Without photon noise, the direct pinhole PSF is limited by numerical noise at the 10^{-18} level (see §3.5.1.2), and so this subtraction algorithm does not suffer from the same level of performance degradation as the other two in terms of achievable contrast improvement.

Figure 3.14 a shows the factor by which contrast improves compared to an input noiseless target image (but still with atmospheric and static wavefront components as described in §3.2) and thus illustrates that we cannot improve contrasts in a single subtraction by better than a factor of about 20 and 6 at 10 λ/D with the algo and cal algorithms, respectively. Because no photon noise is simulated here, these limits are related to additional imperfections in the estimate of the pinhole PSF, which ultimately causes an asymmetric speckle pattern in the output image governed by equation 3.14. Specifically, this bias arises from the $\left[|\mathrm{IFT}\{\mathrm{FT}\{\mathrm{im}_{\mathrm{noisy}}\}\mathrm{m}_1\}|^2/\widehat{I}_R + \widehat{I}_R\right]$ term generated for the pinhole PSF calibration/reconstruction, since a combination of phase and amplitude aberration in the pupil plane produces a non-centro-symmetric speckle pattern (see §3.2 for a further discussion). This asymmetry will limit the achievable contrast in equation 3.14 by using a single normalization coefficient at a fixed separation. However, over a long exposure the limits from Figure 3.14 a may

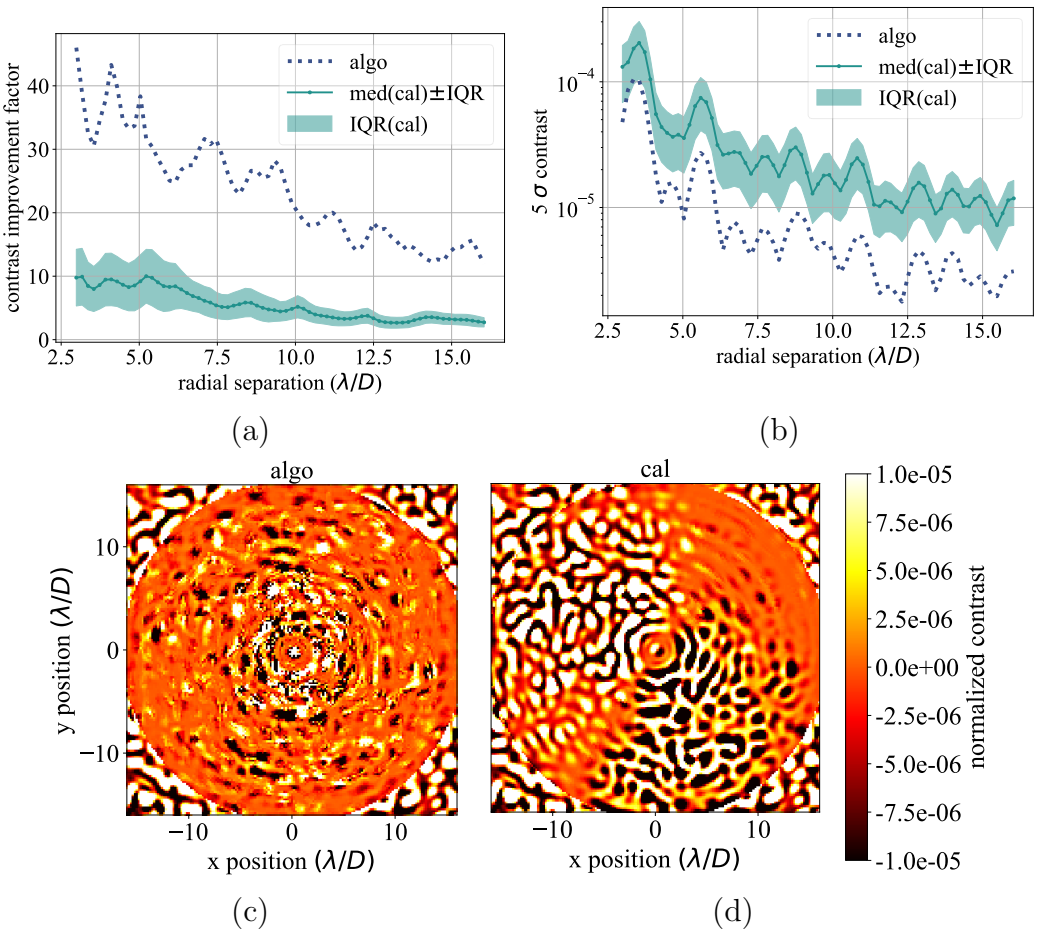


Figure 3.14 Performance limitations of the calibration (shown as "cal") and pinhole PSF reconstruction (shown as "algo") algorithms, in the absence of photon noise. (a) contrast improvement factor vs. radial separation for the algo and cal algorithms (i.e., a ratio of contrast curves between the input target image and output subtracted image). The target image includes both atmospheric and static wavefront components (as described in §3.2). The cal algorithm is shown as the median and interquartile range (shown as "IQR") from 100 different iterations that each use a different static phase screen to generate the calibration pinhole PSF (but still with the same RMS amplitude). (b) contrast curves for the algo and cal algorithms applied to a SCC image with only static aberration, predicting the ultimate limits for either algorithm during a long exposure. (c) and (d): the algo and cal images, respectively, used to generate the contrast improvement curves in panel a; a single random realization is selected from the 100 different cal iterations.

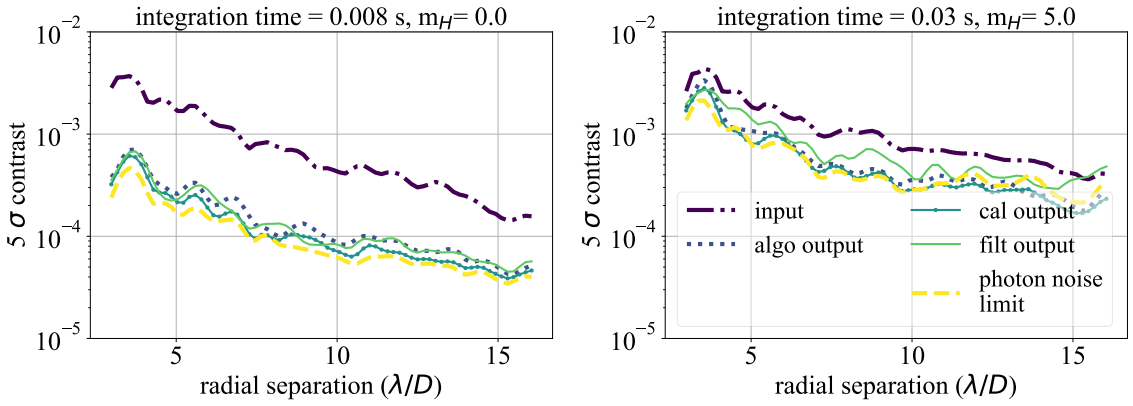
not be the same; a large number of uncorrelated residual atmospheric speckle patterns may average to allow reaching greater improvement than the limits for a single atmospheric phase screen realization.

The predicted contrast limits for a long exposure are illustrated in Figure 3.14 b, showing the subtraction of a target image with only static aberration (i.e., no atmospheric or photon noise component is simulated). The atmospheric averaging effect described above should in principle not reach below the limits in Figure 3.14 b. Thus, once adding photon noise, which should in principle not create additional static effects, the contrast improvement proportional to $t^{-0.5}$ from stacking images is expected to flatten out around these levels.

Figure 3.14 also shows a variation in contrast improvement for the calibration algorithm depending on the specific phase screen realization used to generate the pinhole PSF. This illustrates the sensitivity of this technique to pupil modes of any spatial frequency that vary in each phase screen realization (although, as described in §3.2, tip and tilt are removed by a least-squares subtraction algorithm from each static phase screen) and/or low-order modes diffracted from the TG FPM into the off-axis Lyot stop pinhole; one instance of this effect is clearly seen in Figure 3.14 d, which shows residual low order features that span the full AO control region (i.e., every individual subtracted cal image shows a different residual low order pattern). This sensitivity can either improve or degrading the image quality, varying contrast by factors of ± 10 in different parts of the image, and thus warrants a further study of how to favourably optimize this variation (e.g., using a deconvolution procedure to remove the dominant modes that are causing the contrast degradation).

3.5.2.3 FAST Simulations and Analysis

With the framework developed in §3.5, I consider the FAST approach to measuring and subtracting both quasi-static and atmospheric speckles over short exposures, in contrast to the long exposure strategy in §3.5.1.3. Figure 3.15 shows the results of recording and subtracting a 8 millisecond exposure of a 0th magnitude star and 30 millisecond exposure of a 5th magnitude star according to each CDI method described in §3.5.2.1. Figure 3.15 illustrates that I am able to reach close to the photon noise limit in a relatively short exposure for both a 0th and 5th magnitude star, although this limit is only a factor of about 10 times below the input target image. Assuming the use of a photon counting camera, photon noise will dominate in these short exposures. As discussed in §1.4.3.2.2 and §3.1 and shown in Figure 3.11 e, the speckle lifetime in my simulations is about 200 milliseconds. Thus, the choice of a 30 millisecond integration time for the $m_H = 5$ simulations does not risk the AO halo averaging



(a) (b)

Figure 3.15 Contrast curves showing the recorded input image, output subtracted images, and photon noise limit for (a) an 8 millisecond exposure on a 0th magnitude star, and (b) a 30 millisecond exposure on a 5th magnitude star. The legend in panel b is the same for panel a. The labels "algo," "cal," and "filt" represent the reconstruction, calibration, and direct pinhole PSF algorithms presented in sections 3.5.2.1.3, 3.5.2.1.2, and 3.5.2.1.1, respectively.

over multiple speckle lifetimes. Relatedly, I also choose to use 8 ms exposures on the 0th magnitude simulation. Although a 1 kHz frame rate also reaches the same photon noise limit after 30 seconds, the former 125 Hz frame rate provides a larger contrast improvement in individual subtractions, as illustrated in the left vs. right panels of Figure 3.15.

Using this subtraction procedure on individual exposures, Figures 3.16 and 3.17 show the results of averaging out to a cumulative exposure time of 30 seconds.

As predicted in §3.5.2.2 (Figure 3.14), the contrast curves for the pinhole PSF calibration and reconstruction algorithms flatten out, likely due to static limitations from a combination of how well the pinhole PSF is estimated and how the unrealistic assumptions of azimuthal symmetry in the images impact my flux normalization procedure. The calibration algorithm for the $m_H = 5$ simulation does not yet flatten out because the photon noise limit only improves in contrast by a factor of about 10 over 30 seconds; this photon noise limit for the $m_H = 5$ simulation is still above the static limits that are illustrated in the $m_H = 0$ simulation. The predicted limits from §3.5.2.2, Figure 3.14 b—the best achievable contrasts for algo and cal of 4×10^{-6} and 2×10^{-5}, respectively—are mostly consistent with the results from Figure 3.16: the measured 30 s contrasts for algo and cal are 1×10^{-5} and 6×10^{-6}, respectively. Interestingly, the addition of atmospheric speckles and/or photon noise decreases

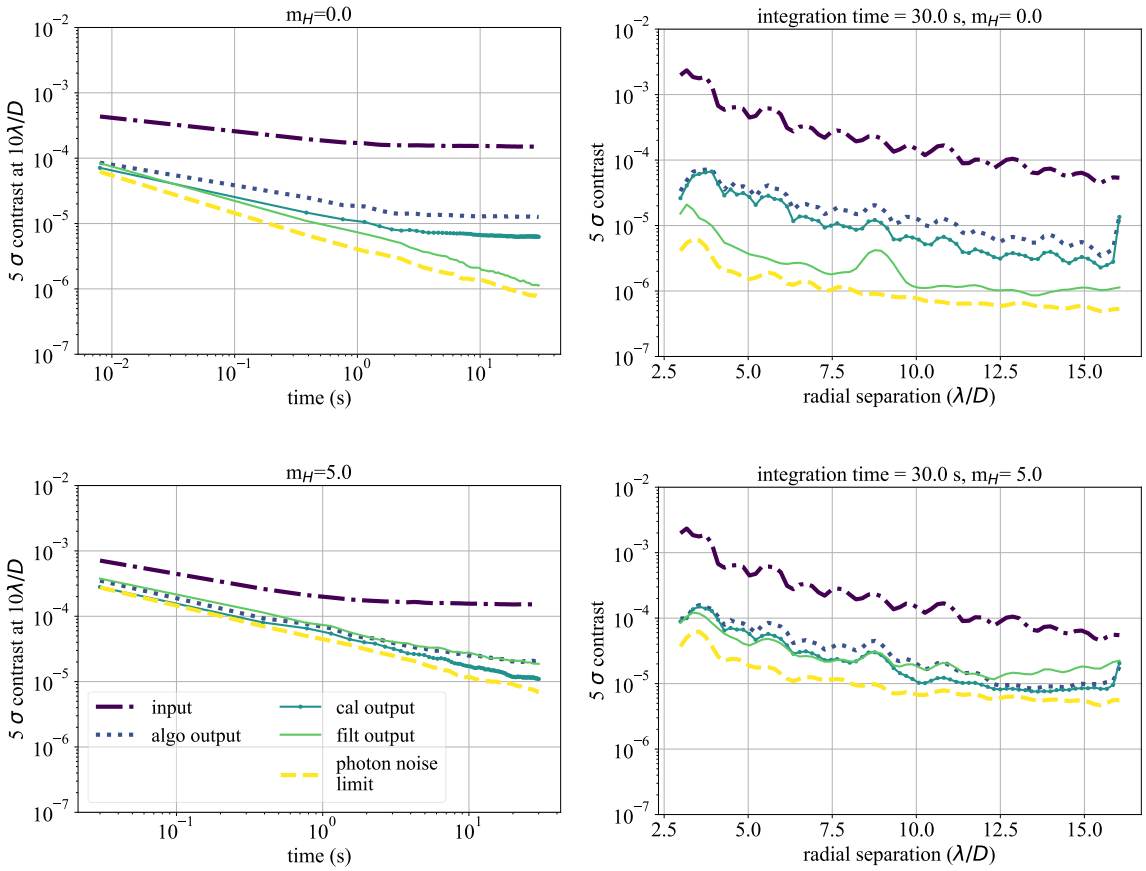


Figure 3.16 Simulation results for a 30 second stacked exposure of 0th and 5th magnitude stars (top and bottom rows, respectively). The left and right columns show the contrast at 10 λ/D vs. time and full contrast curves after 30 seconds, respectively. The legend in the lower left pannel is the same for all panels.

the reconstruction algorithm limit by a factor of 2.5 but increases the calibration algorithm limit by a factor of 3.3.

Figure 3.16 shows that contrast in the direct pinhole PSF measurement algorithm continues to improve proportionally to $t^{-0.5}$ out to 30 seconds in both simulations. This is an important result and shows that, in principle, a direct, simultaneous measurement of the full SCC image and pinhole PSF is not limited by static aberration, even if each exposure can only record a few photons. For example, extrapolating the $m_H = 0$ results predicts that 84 hours of telescope time would reach a 5σ contrast of 10^{-8} at $10\lambda/D$ for a 1% bandpass.

Running simulations longer than 30 s would not provide additional insight into the physical limitations considered in this section, since the limiting and continuous effects

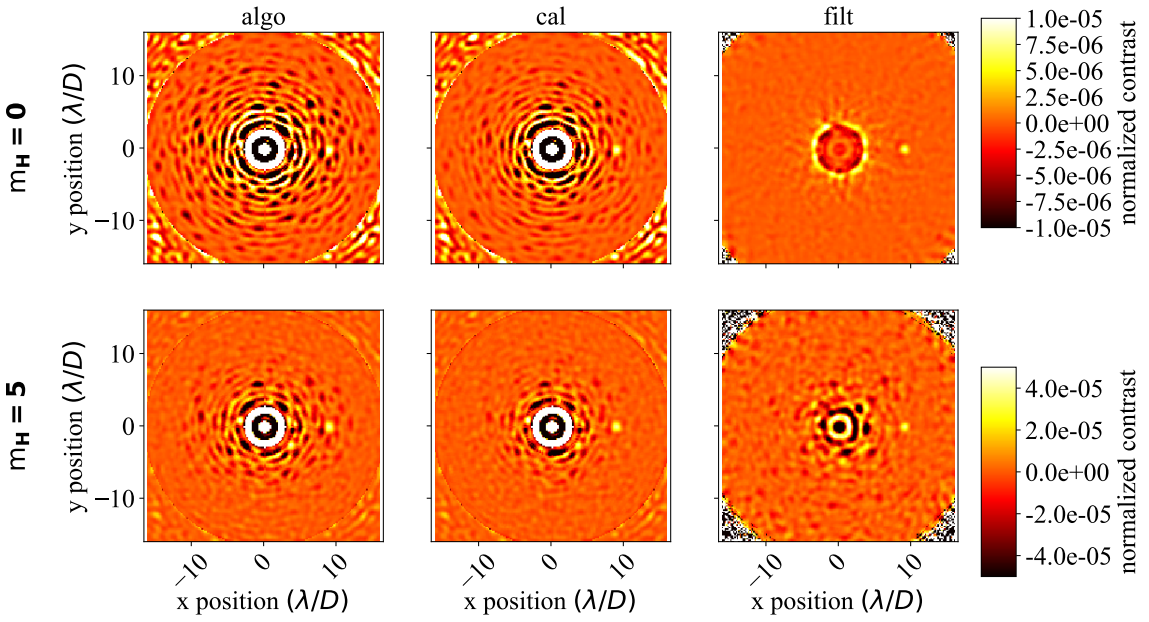


Figure 3.17 The stacked images for 0th and 5th magnitude stars (top and bottom row, repsectively) after 30 seconds from each PSF subtraction method presented in §3.5.2.1: the algo, cal, and filt algorithms (left, middle, and right column, respectively). An exoplanet is simulated in each image at $(x, y) = (+9, 0)$ λ/D with a flux-normalized contrast of 10^{-5} and 5×10^{-5} for the 0th and 5th magnitude stars, respectively. The respective algorithmic exoplanet throughput for the algo, cal, and filt algorithms are 1.18, 1.0, 0.8 for the $m_H = 0$ star and 0.88, 0.84, 0.77 for the $m_H = 5$ star.

observed in the upper left panel of Figure 3.16 provide sufficient information to predict the behaviour out to longer exposure times for either the 0th or 5th magnitude case. These effects are also predicted and discussed in a noiseless analysis of the limitations from static aberration in §3.5.2.2. For the 5th magnitude case, the algo and cal algorithms will flatten out at around the same contrast improvement factors as the 0th magnitude case (approximately 12 and 20, respectively; we can already begin to see this flattening for the algo algorithm after 30 seconds in the lower left panel of Figure 3.16) while the filt algorithm will continuously improve (close to the photon noise limit) all the way out to the end of the simulation. The same behaviour applies to the 0th magnitude case, for which all three subtraction algorithms clearly predict contrasts that are continuously flat (for algo and cal) or improving proportional to $t^{-0.5}$ (for filt) out to any exposure time.

3.5.3 Wavefront Control

This section includes work completed during my PhD studies, which was published in Gerard et al. (2018b).

In addition to CDI post-processing, the second complementary approach to fast coherent speckle subtraction utilizes active control of a DM, as first discussed in §1.4.4.2.3. This approach is considerably more challenging than only applying post-processing, as Fig. 1.17 illustrates that the coronagraphic image is used to control the DM as a "second stage" AO WFS (i.e., meaning the "first stage" AO WFS→DM loop must first be closed before the second downstream fast focal plane WFS→DM loop can be closed). In addition to the limitations that only half of the AO control region is correctable in the presence of amplitude aberrations (§1.4.4.2.1), WFS linearity, sensitivity, and closed-loop stability (all standard parameters used to define tolerance requirements for an AO WFS) must be evaluated to ensure the conditions in which fast focal plane wavefront control can operate effectively (linearity and sensitivity will be evaluated later in §3.5.6). However, there is one main advantage of active wavefront control over to post-processing: the photon noise limit. For the former, speckles can be subtracted *optically*, so that significantly less photons are recorded in the coronagraphic image compared to the latter. As a result, active wavefront control can in principle enable reaching deeper photon noise-limited contrasts than is achievable only with post-processing. In this section I will demonstrate the benefit of this advantage, also showing a feasible architecture for optimized closed-loop DM control with FAST.

3.5.3.1 SCC Calibration Procedure

The standard SCC DM control algorithm (Baudoz et al., 2012a; Mazoyer et al., 2014) involves placing a series of sines and cosines on the DM, which correspond to spots in the detector plane that are separated by λ/D throughout the half DH. For each recorded sine/cosine image, the standard SCC wavefront sensing algorithm is used to isolate the complex fringe term (Baudoz et al., 2006; Galicher et al., 2010; Baudoz et al., 2012a; Mazoyer et al., 2014; Delorme et al., 2016; Gerard et al., 2018a), called I_- (equation 3.9, Fig. 3.4 a), generating a "vectorized" image of the real and imaginary components of I_- inside the DH. Instead of using the recorded sine/cosine image to calculate I_-, I use the difference between a sine/cosine image and an image with a

flat DM, as first proposed in Baudoz et al. (2012a). This differential approach allows significantly better linearity of the least-squares subtraction; without this approach I require using 30 nm amplitude sines/cosines, whereas with the differential approach I was able to use 5 nm amplitudes for the same spot locations. Although quasi-static speckles can be corrected only for a half DH with a single DM because of the presence of both phase and amplitude aberrations (Bordé & Traub, 2006), the same correction will subtract residual atmospheric speckles over a full DH because there is ∼no atmospheric amplitude aberration (i.e., ignoring effects from scintillation and diffraction). Then, for an $N \times N$ actuator DM, each vector is multiplied and summed by every other vector to generate a $N^2 \times N^2$ covariance matrix. Using a pseudo inverse to invert the covariance matrix, I tuned the SVD cutoff to optimize contrast after one iteration, finding an optimal value of 0.15 (where a value of 1 would set the inverse covariance matrix to zero). After the pseudo inverse covariance matrix is calculated in a daytime calibration procedure, during on-sky operation it is multiplied by the target image correlation vector to generate least-squares coefficients (Lafrenière et al., 2007) for every sine/cosine reference position. Each least squares coefficient is subsequently multiplied by the corresponding entrance pupil sine/cosine phase shift and summed in a linear combination to produce the DM phase screen reconstruction, which is then multiplied by negative one (i.e., an optical subtraction), added by the DM to the wavefront phase and propagated through the coronagraph to the detector plane.

I also modified some steps of the above SCC calibration procedure that required additional tuning to optimize the final contrast in a calibrated image:

1. Sine/cosine spots are placed at λ/D intervals but at 0.5 λ/D offset from the centre of the star. Thus, in x and y offsets from the optical axis in the detector plane, the spot positions are between 0.5 and 15.5 λ/D at 1 λ/D increments, still yielding N^2 reference spots and a $N^2 \times N^2$ covariance matrix.

2. I found better linearity and contrast of the DM correction by setting the reference sines/cosines with radial separation $\leq 5\lambda/D$ equal to zero. This is expected for separations $\leq 3\lambda/D$ (the IWA), for which any exoplanet would be blocked by the FPM; however, I also found that subtracting speckles pinned to the bright diffraction rings between 3 and 5 λ/D degraded contrast and linearity of the DM correction. Thus, the effective IWA in my simulations throughout §3.5.3 is 5 λ/D.

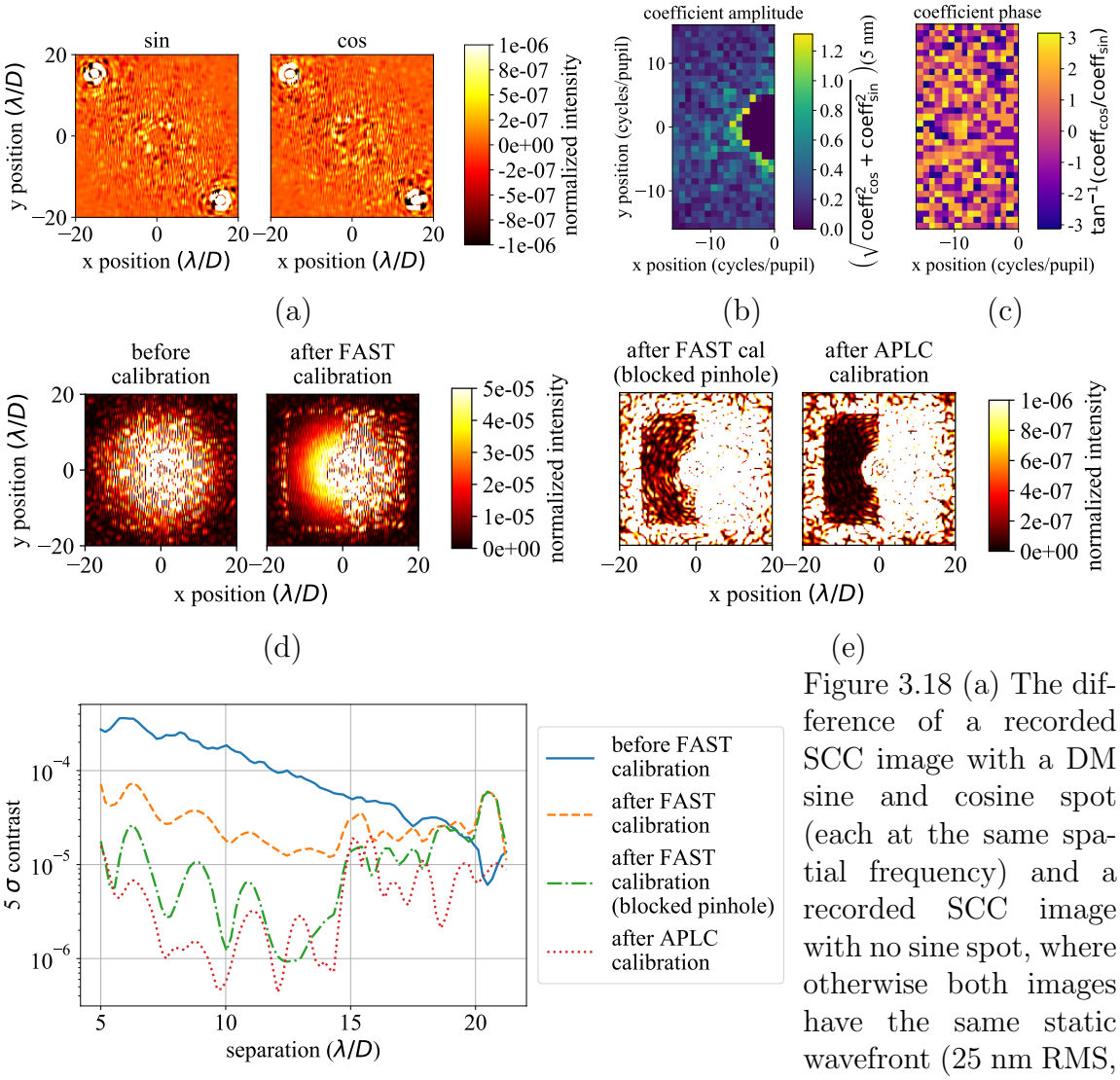


Figure 3.18 (a) The difference of a recorded SCC image with a DM sine and cosine spot (each at the same spatial frequency) and a recorded SCC image with no sine spot, where otherwise both images have the same static wavefront (25 nm RMS, 1 % amplitude RMS). (b) and (c) The least-squares coefficients in phase and amplitude, respectively, generated from an uncalibrated target image to be applied to the 5 nm reference sine/cosine DM spots, shown as an amplitude and phase, respectively. (d) An uncorrected image with the same static aberration used to generate (a), (b), and (c), labeled "before calibration," and a calibrated image after applying a FAST correction with one DM, labeled "after FAST calibration." (e) left: the same DM commands used to generate the image in the right panel of (d) but with the reference pinhole blocked. right: a calibrated correction with one DM using the GPI APLC (Soummer et al., 2006) (an amplitude FPM instead of the TG FPM, but still with the same 3 λ/D IWA); the same calibration procedure used to generate the right panel of (d) is followed. (f) contrast curves for images in (d) and (e).

Figure 3.18 illustrates the process and results of my SCC calibration procedure after including the additional modifications described above, showing that the pinhole PSF of the TG FPM limits achievable contrast. My current design concentrates too much light through the reference pinhole; because there is nine times more flux going through the pinhole than the central pupil, the pinhole PSF is now above the speckle noise floor and limiting the achievable contrast. The "raw" 5σ contrast in Fig. 3.18 d converges to 7×10^{-5} over the DH, a contrast improvement by a factor of 3. When the same DM correction from Fig. 3.18 is applied with the pinhole blocked in 3.18 e, the DH contrast instead improves by a factor of 16. Even if the pinhole PSF was optimally matched to the speckle amplitudes, the same limiting physical principle would ultimately apply: assuming the zeroth order diffraction noise floor is sufficiently below the second order speckle halo term (Perrin et al., 2003, i.e., using a coronagraph optimized to suppress diffraction), a calibrated image with the TG FPM will ultimately be limited by the pinhole PSF, not the speckle halo. At this time I do not address this problem; in this section I will only consider 2 second exposures to demonstrate the basic framework for FAST DM control, and thus do not expect to reach better than a factor of 3 contrast improvement.

Fig. 3.18 e illustrates a second problem: compared to a 3 λ/D IWA amplitude FPM, the 3 λ/D IWA TG FPM cannot reach the same contrast levels. Calibrated speckles in the DH of the right panel of Fig. 3.18 e improve by a factor of 1.8 compared to left panel. Because I do not simulate photon or detector noise here (i.e., assuming a daytime calibration using a bright internal source), the physical origin of this discrepant noise floor must arise from the difference between the wavefronts transmitted through the pinhole when using an amplitude FPM vs. TG FPM. At this time I do not address this discrepancy and simply highlight the potential limitation, although the contrast curve in Fig. 3.18 f shows that this problem is a second order effect compared to the limiting pinhole PSF amplitude effect discussed above.

3.5.3.2 Numerical Simulations

3.5.3.2.1 Framework I ran simulations of a fast DM control loop using the optimized calibration procedure described in §3.5.3.1. Temporal PSD plots are estimated by computing a periodogram of the time-series data (Press et al., 2002). To remove low frequency under-sampled data I first apply a linear de-trending procedure to the time series and then implement a Welch routine (Welch, 1967), using 256 segments

and a Hanning window to suppress noise in the PSD. On-sky images use a $m_H = 0$ star, 250 Hz frame rate ($T_s = 4$ ms), and a 1 millisecond delay ($\tau = 1$ ms). This delay is not unrealistic. For a typical CPU, the speed for a fast Fourier transform (FFT) of a real two dimensional image at single precision is about 15,000 mflops.[2] For a 256×256 image (a beam ratio of 8 if sampling only the DM control region), this would take 175 μs per FFT.[3] Two FFTs to compute I_- would take 350 μs. Combining the 1024×1024 (rows×columns) inverse covariance matrix, 1024×3610 reference image matrix, and 3610×1 target image vector (the latter two are typically dotted to produce the target image correlation vector) into a single matrix operation produces a $(1024 \times 3610) \cdot (3610 \times 1)$ matrix multiplication. One Narrow Field InfraRed Adaptive Optics System server processing one laser guide star WFS (using entirely off-the-shelf CPUs) multiplies a matrix of 7000×5400 by a 5400×1 vector in 500 μs (Smith et al., 2016), and so my FAST approach should be less than this amount. Finally, the FFT cannot start before the last pixel is received, so adding a few hundred microseconds to read out the image yields (350 μs) + (less than 500 μs) + (about 200 μs) = (less than about 1.05 ms).

I tested two different methods of temporal DM control using an integrator controller: a constant gain and a optimized modal gain approach (Gendron & Lena, 1994). I used a standard model for the open loop, closed loop, and noise transfer functions for an integral controller (Alloin & Mariotti, 1994; Véran & Herriot, 2009), hereafter respectively denoted as H_{OL}, H_{rej} and H_n, which are a function of the gain (g), WFS exposure time (T_s), and servo lag (τ). For the constant gain controller I will demonstrate FAST DM control results for both $g = 1.1$ and $g = 0.2$. For $T_s = 4$ ms, $\tau = 1$ ms, and $g = 1.1$, H_{OL} has a 45° phase margin, a common tolerance requirement to provide the best balance of temporal rejection and system stability for an unoptimized controller (Véran & Herriot, 2009). However, the optimal gain is ultimately governed by the balance of both high atmospheric rejection and low WFS noise amplification, or

$$\min\left\{\mathrm{WFE}_i(g = g_{\mathrm{opt}})\right\} = \min\left\{\sqrt{\int_0^{f_n} df\ \mathrm{PSD}_i^{'}\ |H_{\mathrm{rej}}(T_s, \tau, g)|^2 + \int_0^{f_n} df\ \mathrm{PSD}_{(n,\,i)}\ |H_n(T_s, \tau, g)|^2}\right\}, \tag{3.16}$$

where "min{}" is a minimization operator over the integrator gain, i represents a single Fourier mode of the wavefront, $f_n = 1/(2T_s)$ is the Nyquist frequency of the system

[2] `http://www.fftw.org/speed/E31220V3-3.1GHz/`

[3] `http://www.fftw.org/speed/`

frame rate, the open loop PSD of pure atmospheric turbulence is PSD_i', $\mathrm{PSD}_{(n,i)}$ is the open loop PSD of pure WFS noise (i.e., flat at all temporal frequencies), and g_{opt} is the optimal gain governed by the above equation. However, because we do not have direct access to either PSD_i' or $\mathrm{PSD}_{(n,i)}$, it can be shown (Poyneer & Véran, 2005) that instead equation 3.16 is analogous to

$$\min\left\{\mathrm{WFE}_i(g=g_{\mathrm{opt}})\right\} = \min\left\{\sqrt{\int_0^{f_n} df\ \mathrm{PSD}_i\ |H_{\mathrm{rej}}(T_s,\tau,g)|^2}\right\}, \tag{3.17}$$

where PSD_i is now the PSD from noisy time series measurements of open loop coefficients at a single Fourier mode, measuring the temporal statistics of both atmospheric turbulence and WFS noise. With this framework, my implementation of FAST modal gain optimization is outlined below and in Figure 3.19:

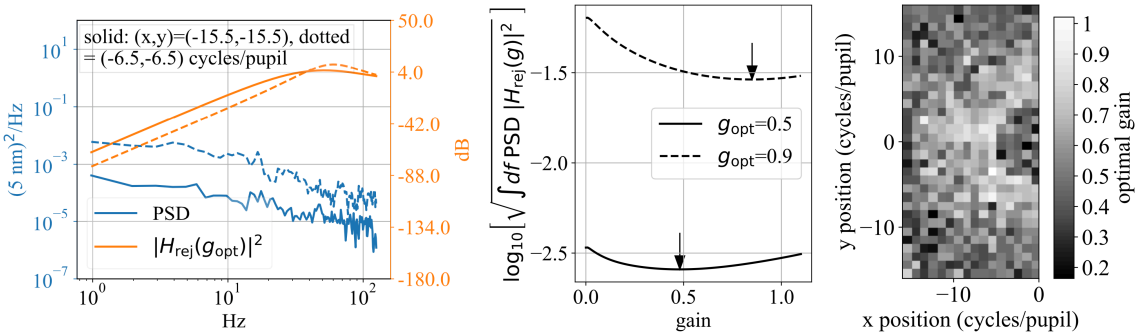


Figure 3.19 An outline of the modal gain optimization procedure used for FAST DM control. (a) For two different Fourier modes, plots of both open loop temporal PSDs and the square modulus of the rejection transfer function for the optimal gain found in panel b. (b) Gain optimization curves for the same two Fourier modes from panel a, governed by equation 3.17; the optimal gain for each mode is shown in the legend and labeled by an arrow. (c) the optimal gain for every Fourier mode in the DH.

1. Measure open loop SCC DM coefficients to sufficiently sample the temporal statistics. I used 2000 frames (8 seconds).
2. For the coefficient amplitude of a single Fourier mode (i.e., $\sqrt{\mathrm{coeff}^2_{\cos,\,i} + \mathrm{coeff}^2_{\sin,\,i}}$, or a single pixel in Figure 3.18 b),
   (a) construct open loop PSDs (Figure 3.19 a) and then
   (b) find the optimal integrator gain using equation 3.17 (Figure 3.19 a and b).

3. Repeat step 2 for every Fourier mode to produce an optimal gain map (Figure 3.19 c).

4. After the open loop measurements from step 1 are finished recording, apply the optimal gain map from step 2b in closed loop.

If the rejection transfer function in step 2b above is not known, it can instead be estimated empirically via (Poyneer et al., 2016)

$$|\tilde{H}_{\mathrm{rej}}(g_{\mathrm{CL}})|^2 = \mathrm{PSD}_{(\mathrm{i,\ CL})}\left(g_{\mathrm{CL}}\right)/\mathrm{PSD}_i, \tag{3.18}$$

where $\mathrm{PSD}_{(\mathrm{i,\ CL})}$ is the closed loop PSD for mode i, g_{CL} is the gain used during those closed loop measurements (not necessarily g_{opt}), and $\tilde{H}_{\mathrm{rej}}$ is the estimated rejection transfer function that can replace H_{rej} in equation 3.17.

Figure 3.19 illustrates the potential advantage of gain optimization. The optimal gain map in Fig. 3.19 c clearly shows a dependence on Fourier mode in the coronagraphic image, illustrating that speckles at the edge of the DH are detected at a relatively lower S/N than speckles closer in, and so the optimal gains are adjusted accordingly. As expected, the PSDs in Fig. 3.19 a show a sloped power law at low temporal frequencies (from atmospheric speckles) and a flat spectrum at high temporal frequencies (from photon noise). The corresponding optimal rejection transfer function balances the maximal rejection of atmospheric speckles at low temporal frequencies and minimal amplification of photon noise at high temporal frequencies. For a Fourier mode closer to the star compared to a mode further away, an atmospheric speckle is brighter (causing a relatively higher PSD amplitude at low temporal frequencies), but as a result the photon noise is also larger (also causing a relatively higher PSD amplitude at high temporal frequencies). However, because the S/N of an atmospheric speckle will increase proportional to $t^{0.5}$ (when t is less than the speckle lifetime), as expected the optimal gain is still higher for the Fourier mode closer to the star.

It is also important to note here that equation 3.17 generalizes to any fast wavefront sensing algorithm and any controller. As the temporal statistics change, H_{rej} can change (ideally matched to the inverse of the open loop PSD) but still remain optimized by equation 3.17. This optimization procedure may be over multiple free parameters for more sophisticated controllers (e.g., using a leak controller; Dessenne et al. 1999; Poyneer et al. 2016). Although slower "drifting" effects from quasi-static

aberration and/or diffraction may not meet Nyquist limit sampling the requirement of equation 3.17, the detrending of open loop temporal PSDs (described at the beginning of this section) should thus not affect the temporal rejection of atmospheric speckles.

3.5.3.2.2 Results Figure 3.20 shows the results after two seconds of integration time in closed loop operation for both a low and high constant gain compared to modal gain optimization. Open loop images (i.e., with no FAST DM control) are denoted as $I_{g=0.0}$. The photon noise limit for each simulation is shown as a dashed line; the open loop photon noise limit (the pink dashed line) is represented by the limit for a perfect algorithmic subtraction from §3.5.1.2. Figure 3.20 illustrates both the general potential advantages of optical subtraction over algorithmic subtraction as well as the advantage of gain optimization over a constant gain.

First, as expected the photon noise limits of all optically subtracted images lie below the algorithmic photon noise limit. The clear advantage of optical subtraction is thus when the contrast can reach below the algorithmic-only subtraction limit. If this can be achieved, deeper contrasts can be reached faster. E.g., if optical subtraction is a factor of two below the open loop photon noise limit, reaching a 5σ contrast of 10^{-8} at 10 λ/D will take 84 hours of telescope time for the algorithmic-only subtraction (§3.5.2.3) but only 21 hours for the aforementioned optical subtraction. Although the actual contrast curves for the optically subtracted images are still far from the algorithmic photon noise limit, additional modifications to FAST DM calibration and control (some of which are presented later in §3.5.6) will bring optical subtraction closer to this limit, including addressing the static pinhole PSF-related problems discussed in §3.5.3.1 and more sophisticated controllers. For the former problem, I showed in §3.5.3.1 that "blocking" the pinhole PSF from the image improves contrast, which could be done algorithmically by filtering the MTF (§3.5.2.1). Additionally, algorithmic subtraction methods from §3.5.2.3 are fully compatible with optically subtracted images, enabling a "two-stage" (optical+algorithmic) subtraction procedure to help surpass the algorithmic-only photon noise limit (e.g., a simultaneous pinhole PSF measurement as in §3.5.2.1.1 would allow continuous algorithmic subtraction down to the corresponding optical photon noise limit).

Second, independent of being able to reach below the algorithmic photon noise limit or not, modal gain optimization clearly outperforms both a low and high constant gain. Compared to the results for modal gain optimization, the upper right

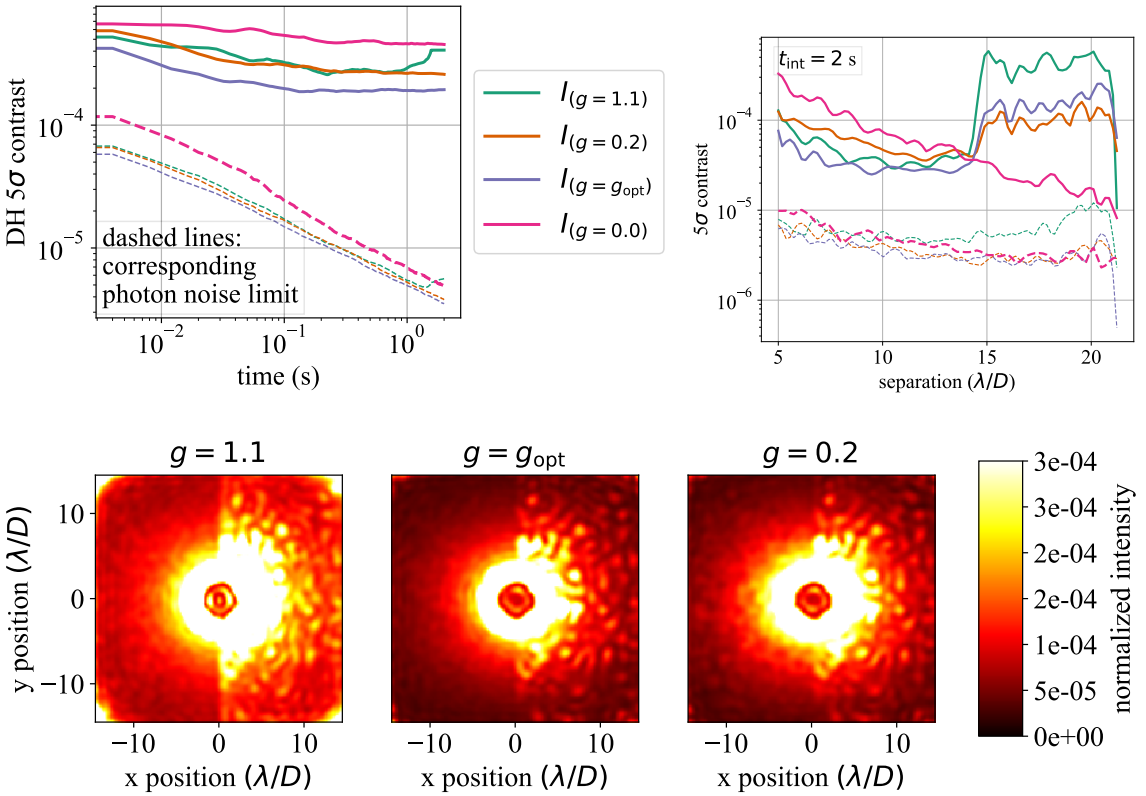


Figure 3.20 Results from my two second FAST DM control simulations. Upper left: 5σ contrast over the half DH vs. time for a low, high, and optimized gain, compared to the respective photon noise limits and the algorithmic photon noise limit from §3.5.1.2. Upper right: contrast curves after two seconds of integration time for the images from the upper left panel. Lower left, middle and right: images corresponding to the upper right contrast curves for the high, optimized, and low gain simulations, respectively.

panel of Fig. 3.20 illustrates that a high gain reaches a similar contrast closer to the star but is worse further from the star, while a low gain reaches a similar contrast further from the star but is worse closer to the star, explaining the results over the half DH in the upper left panel. Along with the images in the lower panel, these results clearly illustrate that a higher gain is required closer to the star (where speckles are detected with a higher S/N) for optimal speckle rejection, while a lower gain is required further away from the star (where speckles are detected with a lower S/N) to prevent noise amplification, consistent with the computed optimal gain map Fig. 3.19 c. With that said, although the gain optimizer performs better overall compared to either constant gain, noise is clearly still amplified compared to open loop images

beyond about $16\lambda/D$. Thus, additional approaches and/or modifications to this setup will ultimately be required to minimize noise amplification, such as a smaller pinhole and/or an anti-aliasing spatial filter (Poyneer & Macintosh, 2004).

3.5.4 The TGV Coronagraph

The work presented in this section is from Gerard & Marois (2020).

The FPM proposed in §3.5.1.1, was not optimized for contrast, following a simple Lyot coronagraph design that is limited by bright diffraction rings in the coronagraphic image. Although speckles "pinned" by diffraction rings (Perrin et al., 2003) can be well-subtracted by post-processing with this design (§3.5.2), it still poses significant difficulties in DM control and adds additional photon noise to the coronagraphic image that cannot be removed by post-processing (§3.5.3). In this section I propose a modification of the original FAST TG FPM that is better optimized for both fringe S/N and attenuation of diffraction: the TGV FPM. First, in addition to the numerical setup for my simulations as described in §3.2, further setup and parameter definitions are provided §3.5.4.1.1 - 3.5.4.1.2. In §3.5.4.2 I then introduce the design of the TGV mask and present a new framework behind FAST coronagraph design, now treating the FPM as both a diffraction attenuator *and* a WFS. Then in §3.5.4.3 I show the relative advantages of dark hole generation using the TGV mask compared to previous work, and finally in §3.5.4.4 I discuss the future outlook from this section.

3.5.4.1 Setup

3.5.4.1.1 Contrast and Throughput Definitions In §3.5.4.3, a contrast curve is produced by computing five times the standard deviation in an azimuthal annulus (of width 3 pixels) at a given separation in the coronagraphic image (i.e., I_S), normalized by the peak value of the same wavefront(s) if the FPM is removed. In §3.5.4.2, normalized intensity curves are computed with the median intensity instead of standard deviation, otherwise using the same normalization and azimuthal bins. I chose to show intensity in §3.5.4.2 instead of standard deviation to allow comparison with the diffraction-limited case in Fig. 3.23; in this case, because a diffraction-limited PSF for a flat entrance pupil wavefront phase is azimuthally symmetric, computing azimuthally averaged standard deviation is less physically representative of contrast in this regime. Also note that I compute contrast and intensity curves on only I_S, as

opposed to the full SCC PSF, to isolate the effects of diffraction and speckle suppression vs. fringe S/N; although I_R will ultimately play a role in contrast, this effect can in principle be fully attenuated in post-processing by measuring the "live" pinhole PSF (§3.5.2.1.1), and so I do not discuss this impact here.

In §3.5.4.2 and §3.5.4.1.2 below I calculate the number of photons collected at the telescope entrance pupil by simulating a $m_H = 0$, 10 millisecond exposure time, with all other simulation parameters as in §3.2.

3.5.4.1.2 Fringe S/N and Photon Noise Throughout §3.5.4 I will represent the Fourier filtering algorithms shown in Figures 3.4 a and b, by the operator $F_{I_-}\{\}$ (i.e., used to generate both I_- and the "un-fringed" SCC image, respectively). With these SCC Fourier filtering algorithms in mind, "fringe S/N" is the ratio between the signal and noise components of the SCC fringes (i.e., considering only the spatial frequencies isolated by m_1 in Fig. 3.4 a). Accordingly, I will define the y-axis of Figure 3.23 c and d as

$$\text{fringe S/N} \equiv \frac{|F_{I_-}\{I_{\text{noiseless}}\}|}{\sigma\{|F_{I_-}\{I_{\text{noisy}}\}| - |F_{I_-}\{I_{\text{noiseless}}\}|\}}, \tag{3.19}$$

where I_{noisy} and $I_{\text{noiseless}}$ are SCC images from equation 3.1 simulated with and without photon noise, respectively, and $\sigma\{\}$ is a numerical standard deviation operator. The Fourier plane equivalent of Equation 3.19 is illustrated in Fig. 3.21.

Note that because of the absolute value signs,

$$\sigma\{|F_{I_-}\{I_{\text{noisy}}\}| - |F_{I_-}\{I_{\text{noiseless}}\}|\} \neq \sigma\{|F_{I_-}\{I_{\text{noisy}} - I_{\text{noiseless}}\}|\}.$$

Relatedly, computing $I_{\text{noisy}} - I_{\text{noiseless}}$ is less physically meaningful in this context; many pixels can detect zero photons in individual 1 millisecond frames, thus producing $-I_{\text{noiseless}}$ as the "noise" component in these cases, which is less physically representative of photon noise in this "quantum regime" than the expression in the denominator of equation 3.19 (thus motivating my choice for a 10 ms instead of 1 ms exposure in §3.5.4.1.1). Similar to the intensity curves in Fig. 3.23 a and subsequent contrast curves, the numerator and denominator of equation 3.19 are calculated as the median and standard deviation, respectively, in three pixel wide annuli as a function of image plane separation. Also note that equation 3.19 only considers the S/N of the fringe amplitude but not the fringe phase (i.e., for a single fringe on a single speckle, the detectability of its intensity, relative to photon noise, but not its relative

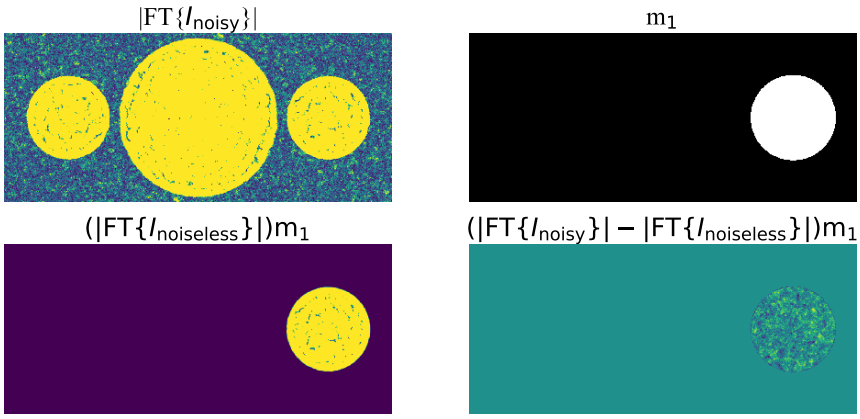


Figure 3.21 An illustration of the OTF plane components of Equation 3.19, using the TG FPM for an input static phase screen with 25 nm RMS in phase and 1% RMS in intensity. Upper left: the MTF, from an SCC image with simulated photon noise. Upper right: an algorithmic binary mask to isolate the spatial frequencies of the fringes (as in Fig. 3.4 a). Lower left: the noiseless (i.e., signal) component—at the isolated spatial frequencies of the fringes—of the upper left image, representing the numerator in equation 3.19. Lower right: the noise component—again at the isolated spatial frequencies of the fringes—of the upper left image, representing the denominator in equation 3.19.

position on the speckle, respectively); as illustrated in the lower left panel of Fig. 3.4 a, numerical phase wrapping prevents an analogous S/N analysis for fringe phase. However, there is no reason to believe that this "fringe phase S/N" would draw any conclusions that deviate from the results in Fig. 3.23 c and d, as the fringes still need to be detected above the photon noise in order to measure their relative position.

Next, as in derived in §3.5.1.2 (Eq. 3.8), the "photon noise limit," or photon noise limited-contrast, is determined by the combined effects of photon noise propagation through the two Fourier filtering algorithms in Fig. 3.4. With this context in mind and utilizing the definitions in Equations 3.1 and 3.19 and in Fig. 3.4, the y-axis of Fig. 3.23 e is given by

$$\text{photon noise limit} = 5\,\sigma\left\{\left(\text{IFT}\left\{\text{FT}\left\{I_{\text{noisy}} - I_{\text{noiseless}}\right\} m_2\right\}\right) + \left(\frac{|F_{I_-}\{I_{\text{noisy}}\}|^2 - |F_{I_-}\{I_{\text{noiseless}}\}|^2}{I_R}\right)\right\}. \tag{3.20}$$

Relatedly, the images labeled as "output photon noise limit" in Fig. 3.24 are given

by $(\mathrm{IFT}\{\mathrm{FT}\{I_{\mathrm{noisy}} - I_{\mathrm{noiseless}}\} m_2\}) + \left(\frac{|F_{I\_}\{I_{\mathrm{noisy}}\}|^2 - |F_{I\_}\{I_{\mathrm{noiseless}}\}|^2}{I_R}\right)$. Note that I_R in the above equation is an assumed simultaneous noiseless measurement of the pinhole PSF. As described in §3.5.1.2, although in reality the simultaneous measurement of an on-sky pinhole PSF will be noisy, this will only increase the noise contribution from the second term in equation 3.20, therefore rendering equation 3.20 as a lower limit.

Lastly, related to the photon noise calculations in Fig. 3.23, panels d and e and supporting text utilize the term "fringe ratio" and "modified fringe ratio" which I respectively define as

$$\begin{aligned} \text{fringe ratio} &\equiv I_R/I_S, \text{ and} \\ \text{modified fringe ratio} &\equiv \tilde{I_R}/I_S, \end{aligned} \tag{3.21}$$

where $\tilde{I_R}$ in equation 3.21 is produced by numerically adjusting the intensity in the off-axis pinhole of the SCC Lyot stop plane by a piston "fudge factor" (without changing the wavefront in the central Lyot stop pupil) between values both smaller and larger than the natural, unadjusted fringe ratio value.

3.5.4.2 The TGV FPM

Building on the design proposed in Gerard et al. (2018a), the main goal of my proposed FPM modification in this section is improve the diffraction-limited contrast in the coronagraphic image while still maintaining a sufficient fringe S/N. The latter fringe S/N requirement remains essential in order to operate the FAST technique on millisecond timescales for bright stars. Although the raw contrast, dominated by unpinned quasi-static and/or atmospheric speckles, may not show an improvement over the previous design, we ultimately want to reach a deeper contrast in the subtracted image by post-processing and/or DM control. With this in mind, the TGV FPM is a focal plane phase mask with three components, each of which has the goals described below:

Tip/tilt: generate a spatially filtered, off-axis pupil in the Lyot plane (the off-axis pupil is limited to Fourier modes less than $2e$ cycles/pupil, where e represents the radius of the central Tip/tilt+Gaussian region in λ/D),

Gaussian: concentrate, in the Lyot plane, the intensity of the off-axis pupil generated by the Tip/tilt component, and

Vortex: redistribute diffracted star light from inside to outside the central Lyot pupil using a vortex phase ramp (Mawet et al., 2005).

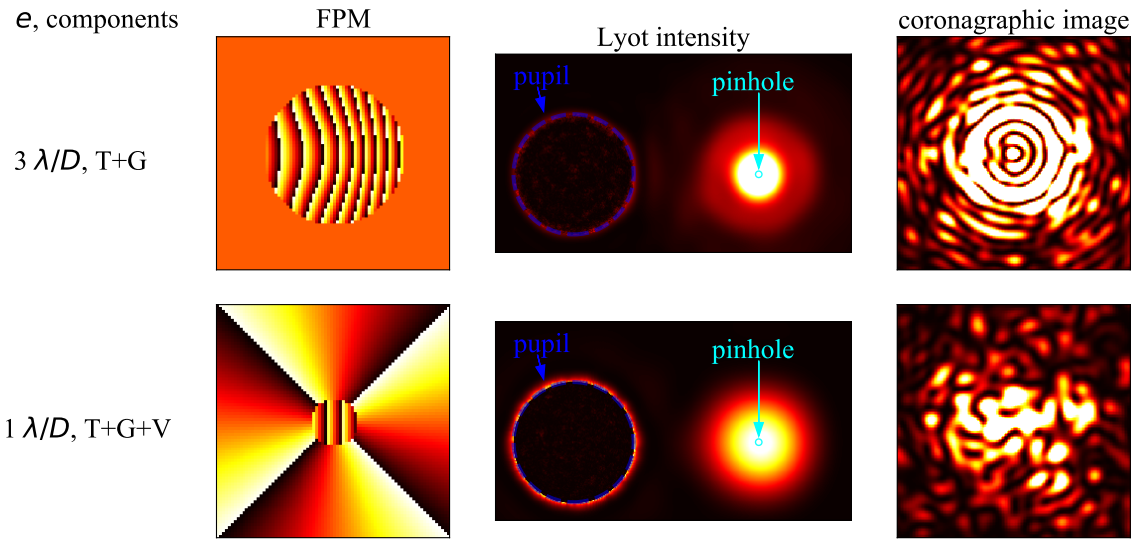


Figure 3.22 A comparison of the old (TG, top row) and new (TGV, bottom row) coronagraphic phase masks, showing the phase-wrapped optical path difference (left "FPM" column), corresponding Lyot plane intensity before a Lyot stop is applied (middle "Lyot intensity" column), and corresponding coronagraphic image after a Lyot stop is applied with no SCC pinhole (right column). The pinhole and entrance pupil apertures in the Lyot plane are illustrated in cyan and blue, respectively. Each column is on the same contrast scale, and the same single aberrated wavefront realization, applied in the entrance pupil, is used to generate both Lyot intensity and coronagraphic images above.

Figure 3.22 conceptually illustrates the differences between my old TG mask and my new TGV mask. First, considering only the intensity in the off-axis Lyot pupil, a larger e value concentrates more light into the SCC pinhole. This effect occurs from the optical relationship between the FPM and pupil plane diameters; for $e = 1\ \lambda/D$, light concentration of the off-axis Lyot pupil, enabled by the G term, can only decrease the full width at half maximum (FWHM) by a factor of 2 (i.e., the diffraction limit), whereas for $e = 3\ \lambda/D$, the downstream pupil FWHM can instead shrink by a factor of 6. In addition to this "compresibility" diffraction limit, a larger e value collects more light from the on-axis star that is relayed into the off-axis pupil. Second, considering only the intensity distribution around the central Lyot pupil, more diffracted light is sent outside of the pupil by the TGV mask than the TG mask, illustrated by the bright ring around the edge of the pupil generated by the TGV mask. This effect is then seen

in the coronagraphic image generated using the TGV mask, producing speckles that are no longer pinned to bright diffraction rings. The attenuation of pinned speckles by the TGV mask occurs because more diffracted light is redistributed outside of the central pupil and blocked by the Lyot stop, while dynamic and quasi-static aberrations from the telescope and instrument, respectively, are still transmitted through the central pupil.

The mathematical prescription for the TGV mask, ϕ_{TGV}, is defined as follows:

$$\begin{aligned} T &\equiv 3.16(\xi_0)(x\,\cos\theta_0 + y\,\sin\theta_0),\ \forall\ r < e \\ G &\equiv g\ e^{-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2},\ \forall\ r < e \\ V &\equiv l_p\ \theta,\ \forall\ r > e, \\ \phi_{\mathrm{TGV}} &= T + G + V\ [\mathrm{rad}], \end{aligned} \tag{3.22}$$

where x and y are a linear ramp in units of λ/D along each respective axis with the zero point corresponding to the optical axis, $r \equiv \sqrt{x^2+y^2},\ \theta \equiv \tan^{-1}\left(\frac{y}{x}\right)$, ξ_0 is the distance between the center of the Lyot pupil and the center of the pinhole in units of pupil radii, and θ_0 is the position angle of the tip/tilt direction applied to the T component (counter-clockwise from the $+x$ direction) and optically matched to the corresponding position angle between the centers of the pupil and SCC pinhole in the Lyot plane. Conceptually, the effect of the G term is to emulate a speckle or PSF core whose resolution limit is larger than λ/D, corresponding to a spatially filtered pupil in the Lyot plane whose FWHM is smaller than the re-maped entrance pupil. I found that a symmetric 2D Gaussian function is effective at concentrating light in the Lyot plane, although others similar functions could be examined in future global optimizations (see §3.5.4.4). Values for g, σ, and l_p represent the Gaussian amplitude (in radians), Gaussian width (in λ/D), and the integer-valued topological charge of V (dimensionless; Mawet et al. 2005), respectively, and are all free parameters to be optimized. For a chosen value of e I performed a grid search optimization of g and σ using the Lyot plane intensity distribution; integrated intensity over the pinhole divided by integrated intensity over the pupil was computed for every grid value, and the chosen optimal values were set to optimize this integrated Lyot fringe ratio metric. I did not run an additional grid search for e and l_p. Using this procedure I found an optimal TGV prescription of $e = 1\ \lambda/D$, $g = 6\ rad$, $\sigma = 5\ \lambda/D$, and $l_p = 4$. In Gerard et al. (2018a) and Gerard et al. (2018b), the TG mask was defined with

$e = 3\ \lambda/D$, $g = 11.5\ rad$, $\sigma = 2.0\ \lambda/D$, and $l_p = 0$.

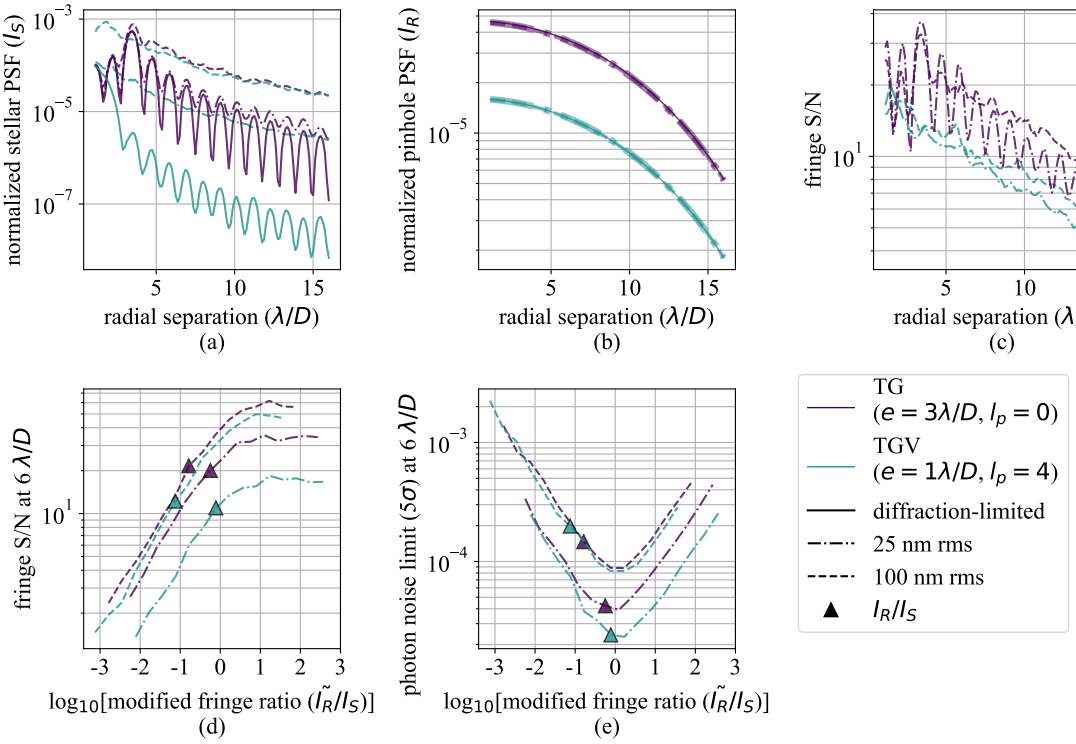


Figure 3.23 An illustration of the tradeoffs between coronagraphic image intensity (a), pinhole PSF intensity (b), fringe S/N (c and d), and photon noise-limited contrast (e), utilizing the definitions in §3.5.4.1.1 - 3.5.4.1.2. The purple and teal curves represent the old TG and new TGV masks, respectively. The solid, dotted, and dashed-dotted lines represent performance with no wavefront error (i.e., diffraction-limited in both phase and amplitude), on-sky conditions for a 10 millisecond exposure with both AO residuals and quasi-static wavefront errors (i.e., 100 nm RMS phase aberration and 1 % RMS intensity aberration), and on-sky conditions for a 10 millisecond exposure with a perfect AO correction but remaining uncorrected non-common path errors (i.e., 25 nm RMS phase aberration, 1 % RMS intensity aberration), respectively. For the curves simulating on-sky conditions, each line shown is the median contrast or fringe S/N curve, each first individually determined from 10 different uncorrelated entrance pupil wavefront realizations. The fringe ratio values in panels d and e are indicated for each curve by a triangle symbol of the corresponding colour.

Utilizing the setup and definitions in §3.5.4.1.1 - §3.5.4.1.2, Fig. 3.23 provides a more robust illustration of the tradeoffs between contrast and fringe S/N, and illustrates a new approach to coronagraph design, simultaneously considering diffraction attenuation, WFS sensitivity, and photon noise-limited contrast, shown in Fig. 3.23 a, c/d, and e, respectively. The main conclusion from Fig. 3.23, shown in panel e, is

that photon noise-limited contrast is optimized at a fringe ratio of $I_R/I_S \approx 1$. This concept is illustrated further in Fig. 3.24.

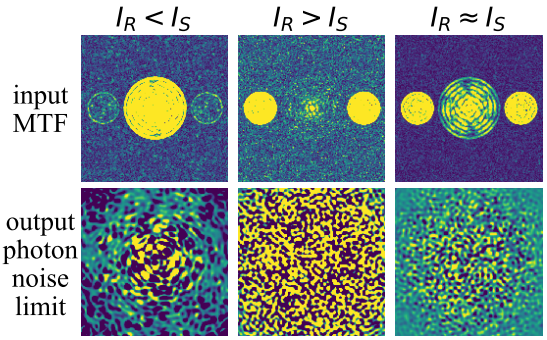


Figure 3.24 An illustration of three different modified fringe ratio cases, building off of Fig. 3.23 d and e, showing for each case the input image MTF and the output photon noise limit (images in this bottom row are all shown on the same linear contrast scale), both as defined in §3.5.4.1.2. SCC images are generated by simulating a 10 ms on-sky exposure (i.e., with photon noise and the setup described in §3.5.4.1.1), using the TGV FPM and adjusting the Lyot plane pinhole intensity as described in §3.5.4.1.2.

Summarizing the results from Fig. 3.23 d and e and Fig. 3.24, the two less optimal regimes of photon noise-limited contrast are:

$I_S > I_R$: An insufficient amount of light is sent through the Lyot stop pinhole, such that excess photon noise from I_S both decreases the fringe S/N and increases the photon noise-limited contrast (left column of Fig. 3.24). This scenario is the most common for typical coronagraph designs, where the fringe ratio is not optimized in the design procedure.

$I_S < I_R$: Too much light is being sent through the Lyot stop pinhole, such that excess photon noise from I_R degrades the photon noise-limited contrast (middle column of Fig. 3.24; any exoplanet which could be detected in the central lobe of the MTF in the upper left or right panels is now buried in the photon noise generated from I_R), although Fig. 3.23 d clearly shows that this effect does not decrease fringe S/N, which instead asymptotes as I_R continually increases in this regime. However, if new coronagraph designs can enable this regime, note that this case can be mitigated by modifying the size (Mazoyer et al., 2014), transmission, and/or complex electric field imparted through the Lyot stop pinhole, adjusting I_R to best match the amplitude of I_S.

Thus, incorporating the tradeoffs with contrast (see below, which will determine the WFS sensitivity to non-linearities; Guyon 2007), $I_R/I_S = 1$ should be adopted as a

coronagraph design parameter to optimize WFS sensitivity to photon noise (Guyon, 2005) without degrading the achievable photon noise-limited contrast.

Additional conclusions from Fig. 3.23 are similar to those from Fig. 3.22, showing that

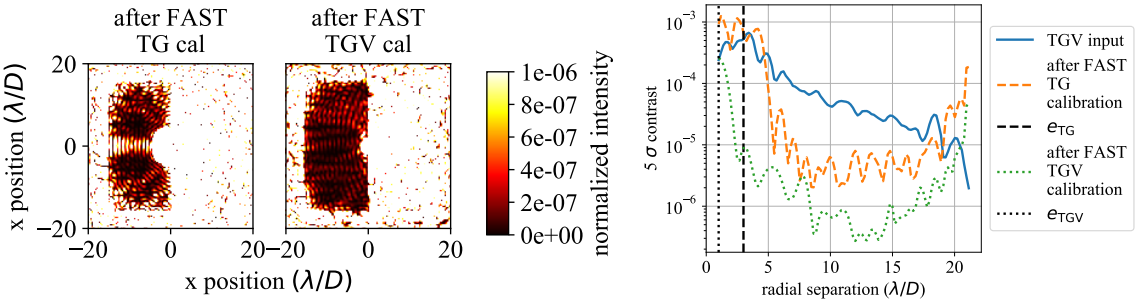
1. for open-loop (i.e., FAST loop open, AO loop closed) FAST on-sky exposures that "freeze" both the atmospheric residuals and quasi-static aberration (i.e., curves labeled "100 nm RMS"), the fringe S/N is generally higher for the old TG design than for the new TGV design due to the higher level of I_R in panel b while contrast is the same for both (panel a), but

2. the diffraction-limited contrast is orders of magnitude better for the new TGV vs. old TG mask design.

As a result, raw contrasts for the "25 nm RMS" case are lower for the TGV than the TG design, and accordingly the TGV mask reaches a deeper photon noise-limited contrast for this case. For the 100 nm RMS case, photon noise-limited contrast is instead lower for the TG mask, due to the aforementioned higher fringe S/N but equal contrast levels compared to the TGV mask. This suggests that, for on-sky DM control of un-pinned speckles (i.e., minimizing entrance pupil WFE), there is a crossing point once the FAST loop is closed where the achievable photon noise-limited contrast of the TGV mask surpasses the values of the TG mask. Thus, although fringes for the TGV design would be detected at a relatively lower S/N in open loop millisecond frames, deeper contrasts are expected if the FAST loop can close (see §3.5.4.4 for further discussion).

3.5.4.3 Quasi-Static DM Control

In this section I examine the performance of the TGV mask in generating a half DH via DM control of quasi-static speckles. As a reminder, the main error terms that FAST addresses in enabling deeper detections are quasi-static and residual AO speckles. Note that the SCC command matrix relies on a linear assumption to transform SCC images into DM commands in a single least-squares-based matrix multiplication (Baudoz et al., 2012a). Although here I am only considering correction of quasi-static aberration, I have not considered other iterative non-linear DM control algorithms that are more optimized for diffraction attenuation (e.g., Bordé & Traub, 2006; Give'On et al., 2007; Pueyo et al., 2009) as the ultimate goal of this approach



(a) (b)

Figure 3.25 (a) Calibrated half DH I_S images (i.e., Lyot stop pinhole closed) using the old TG and new TGV mask design (left and right panels, respectively). (b) Contrast curves for the two images in (a), calculated on pixels only within the half dark hole, as well as for the TG input image before calibration. The same static wavefront realization (i.e., 25 nm RMS phase and 1 % RMS intensity aberration) is input into both the TG and TGV calibration procedures.

will to be to run the same correction on-sky on noisy millisecond exposures. A detailed description of the SCC DM calibration procedure can be found in Gerard et al. (2018b), §3.5.3, and references therein. In Fig. 3.25 I compare the results of this procedure for the TG and TGV masks using a single input static wavefront realization with 25 nm RMS of phase aberration and 1% intensity RMS of amplitude aberration. The calibration results are shown after three iterations using an integrator controller and a unity gain. The control algorithm linking the Fourier modes recorded in the SCC image and the DM commands relies on a Taylor expansion of the wavefront assuming small phase and amplitude defects (Baudoz et al. 2012a; i.e., linearizing $a\ e^{i\,\phi}$ to $a(1 + i\ \phi)$, where a and ϕ represent the spatial distributions of amplitude and phase, respectively, in the complex electric field of the entrance pupil plane). Multiple iterations are therefore still required, even with a unity gain, to address non-linearities between the two planes.

A comparison of the calibrated DH generated from these two coronagraphs yields a few important results:

1. the DH contrast with the TGV mask is about 200 times deeper than the TG mask at 2 - 5 λ/D, enabled by the smaller value of e and deeper diffraction-limited contrast at these separations, and

2. up to 10 times deeper at 5 - 20 λ/D.

Note that contrast curves in Fig. 3.25 b are shown within the value of e; although

planet throughput at these separations is ∼0, the curves are still shown to illustrate the independent concept of diffraction suppression between the two coronagraphs. As discussed in Gerard et al. (2018b), even though $e = 3\ \lambda/D$ for the TG mask, I found that I had to use an algorithmic mask to block the central 5 λ/D in radius because of bright diffraction in the coronagraphic image that otherwise biased the least-squares algorithm. If I instead used an algorithmic mask down to, e.g., 3 λ/D, I could not reach the same contrasts from 5 - 20 λ/D as in Fig. 3.25 b. I also found the same effect for the TGV mask, requiring an algorithmic DH mask with an innermost radial separation IWA of 2 λ/D instead of the TGV e value of 1. Regardless of this limitation, the TGV mask clearly provides a gain in achievable DH contrast over the TG mask.

3.5.4.4 Discussion

Although I did not yet specifically address the achievable contrasts for closed-loop FAST operation using millisecond-timescale on-sky images, the framework already presented in this section provides a promising outlook for expected on-sky performance. Even though in Fig. 3.23 I showed that TGV fringe S/N is above 10 at separations less than about 6 λ/D for a 10 ms exposure with 100 nm RMS WFE, as soon as the FAST DM control loop is closed the WFE should decrease to a much lower value, thereby boosting the fringe ratio to a more optimal value and improving the achievable photon noise-limited contrast. If the fringe ratio is boosted to greater than one (thereby degrading the photon noise limit), the pinhole size and/or throughput can be adjusted to set $I_R = I_S$. Thus, the main potential limitation will be whether or not the FAST DM control loop can close at the lower frame rate needed to detect fringes in the raw images; this will be investigated in detail in a forthcoming paper (also see §4.3.4).

I have also illustrated that coronagraph mask design optimization is clearly a crucial step in optimizing the achievable contrast using FAST post-processing and/or DM control, which I will explore further in §4.3.2. Many factors need to be considered in the design process, such as optimizing the tradeoffs between diffraction-limited contrast, on-sky millisecond-timescale fringe S/N, and sensitivity to low-order aberrations. The initial study in this section is meant to provide the conceptual framework for optimization of a more instrument-ready FAST coronagraph design. Future work on this topic will consider additional factors that would influence realistic coronagraph

design, such as sensitivity to secondary obscuration and supports, chromaticity, and a full Monte Carlo analysis of TGV free parameters. Additionally, such optimizations will need to consider AO performance, a new approach to coronagraph design; Figure 3.23 has illustrated that the requirements for fringe S/N (which will trace the WFS sensitivity photon noise propagation; Guyon 2005) and contrast (which will trace the WFS non-linearities; Guyon 2007) are inherently tied to FAST coronagraph design. Such future FAST optimizations will therefore also need to consider the relative tradeoffs of these factors in a focal plane wavefront control AO error budget analysis, which I will explore in detail in §3.5.6.2.

3.5.5 Low Order Wavefront Sensing

This section presents unpublished work, completed mostly during my Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Training Experience (CREATE) Technologies for Exo-Planetary Science (TEPS) international internship at Paris Observatory in Meudon, France.

In this dissertation thus far, for all simulated phase screens I have removed tip/tilt in the entrance pupil by a least-squares subtraction, before propagating through to the coronagraphic image. In reality, uncorrected tip/tilt, particularly in a long exposure, will blur SCC fringes, lowering the fringe S/N; active tip/tilt control will thus be essential to successful FAST operation. In this section I will present simulations of a FAST wavefront control architecture to correct for tip/tilt and other low order modes using a DM and/or CDI strategy.

3.5.5.1 Lyot-based Low Order Wavefront Sensing

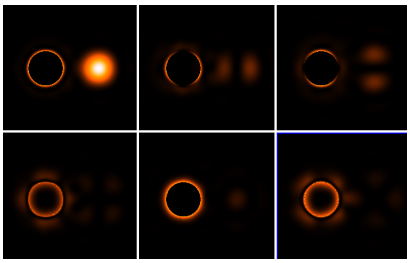


Figure 3.26 Lyot plane intensity images using the TGV FPM for a flat input wavefront (upper left) and of the differential electric field (i.e., the intensity of the difference between: (1) a simulated complex-valued Lyot plane electric field for a given input phase screen and (2) the electric field corresponding to upper left image) for input low order Zernike modes with peak-to-valley amplitudes of 1 nm, including tip (upper middle), tilt (upper right), left astigmatism (lower left), focus (lower middle), and right astigmatism (lower right).

Figure 3.26 shows the TGV Lyot plane intensity distribution in response to low order Zernike modes, revealing that most of the signal from these modes will come from light outside the central Lyot pupil, although some of this information will still

be transmitted to the SCC through the off-axis pinhole, particularly (as illustrated) for tip/tilt and focus. For this reason I will consider two modes of low order wavefront sensing in this section:

1. A Lyot-based low order wavefront sensor (LLOWFS) (Singh et al., 2014, illustrated in Fig. 3.27).

   This approach is commonly used for phase mask coronagraphs, where it is not possible to utilize other custom focal plane masks that are designed for low order wavefront sensing, such as the coronagraphic LOWFS (Guyon et al., 2009) or a SHWFS that is spatially filtered by an amplitude FPM (Wallace et al., 2010).

2. The SCC.

   Although the classical SCC design is known to be insensitive to low order aberrations (Galicher et al., 2010, instead utilizing a LLOWFS design, Baudoz et al. 2018), as discussed above, Fig. 3.26 clearly shows that this is not true for FAST, where spatially-filtered low-order modes from the "TG" component of any FAST FPM are sent through the Lyot stop pinhole and transmitted to the coronagraphic image, thus warranting further analysis for this application.

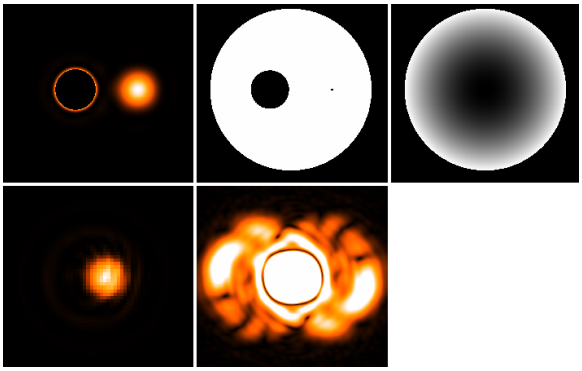
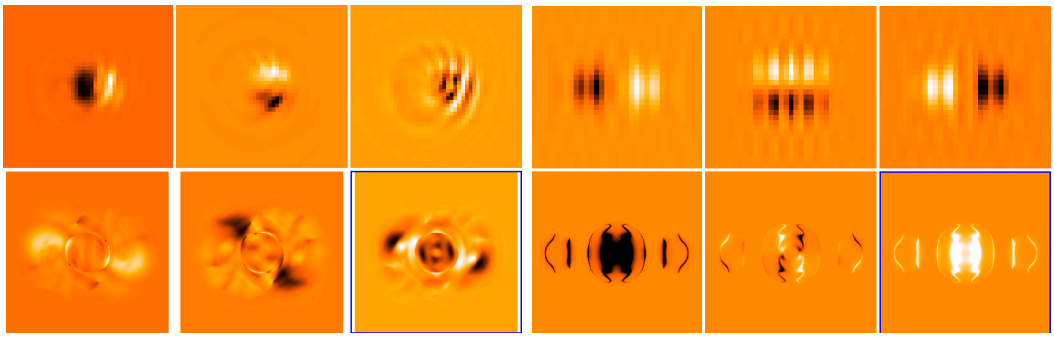


Figure 3.27 An illustration of the LLOWFS mask and image formation from the TGV FPM. Upper left: Lyot plane intensity. Upper middle: the amplitude profile of a FAST LLOWFS mask, transmitting (illustrated as black) the central pupil and off-axis pinhole to the SCC detector while reflecting (illustrated as white) the light nominally blocked by the Lyot stop to a separate detector. Upper right: the LLOWFS detector is defocused in the downstream focal plane to remove the wavefront sensing sign ambiguity of symmetric Zernike modes (see Guyon et al. 2009); this defocus is simulated by adding a focus term (in phase) in the Lyot plane, which is defined by the aperture diameter of the LLOWFS mask. Lower left: the resulting defocused LLOWFS image. Lower middle: the MTF of the LLOWFS image, illustrating that "fringes" are formed in the LLOWFS image as a result of interference between the off-axis Lyot pupil "aperture" and the "aperture" from the bright ring around the central Lyot pupil.



(a) (b)

Figure 3.28 Differential images (top row; between input and image with a flat entrance pupil wavefront) and corresponding differential MTFs (bottom row) for the LLOWFS (a) and SCC (b) in response to a 1 nm peak-to-valley (PV) amplitude of entrance pupil tip, tilt, and focus (left, middle, and right colums, respectively, for each panel).

An illustration of the response to tip, tilt, and focus for each of these LOWFS approaches is shown in Fig. 3.28. Although tip and tilt can clearly be identified by eye for both sensors, note that one potential advantage of the LLOWFS is that both the fringes and central lobe of the MTF can be used for wavefront sensing (as the light redirected by the LLOWFS mask should be only starlight, with negligible contamination from exoplanet light beyond the FPM IWA), whereas the SCC can only use the MTF sidelobe to prevent exoplanet contamination. Also note that unlike the LLOWFS detector the SCC cannot be defocused, as this is also the science camera; however, the focus sign ambiguity is resolved by the fringes of the SCC, potentially removing the need for this defocus (but see below). These differential images in Fig. 3.28 (and for Zernike modes up to three radial orders) form the bases for an interaction/covariance matrix, which are then inverted to compute a command matrix and least-squares DM coefficients for each mode.

Next, Fig. 3.29 shows the LLOWFS linearity of focus as a function of different simulated defocus values, clearly illustrating that larger defocus values make the sensor more linear. For this reason I will adopt a 2 μm defocus for subsequent LLOWFS simulations in this section.

Fig. 3.30 compares the linearity of tip, tilt, focus, and astigmatism between the LLOWFS (a) and SCC (b). The tip/tilt linear range is indistinguishable between sensors, both showing about $\pm$500 nm PV, which should be well within the range of AO residuals for a 100 nm RMS input (Noll, 1976). However, the focus linearity clearly deviates between the two sensors, with a significantly greater linear range for

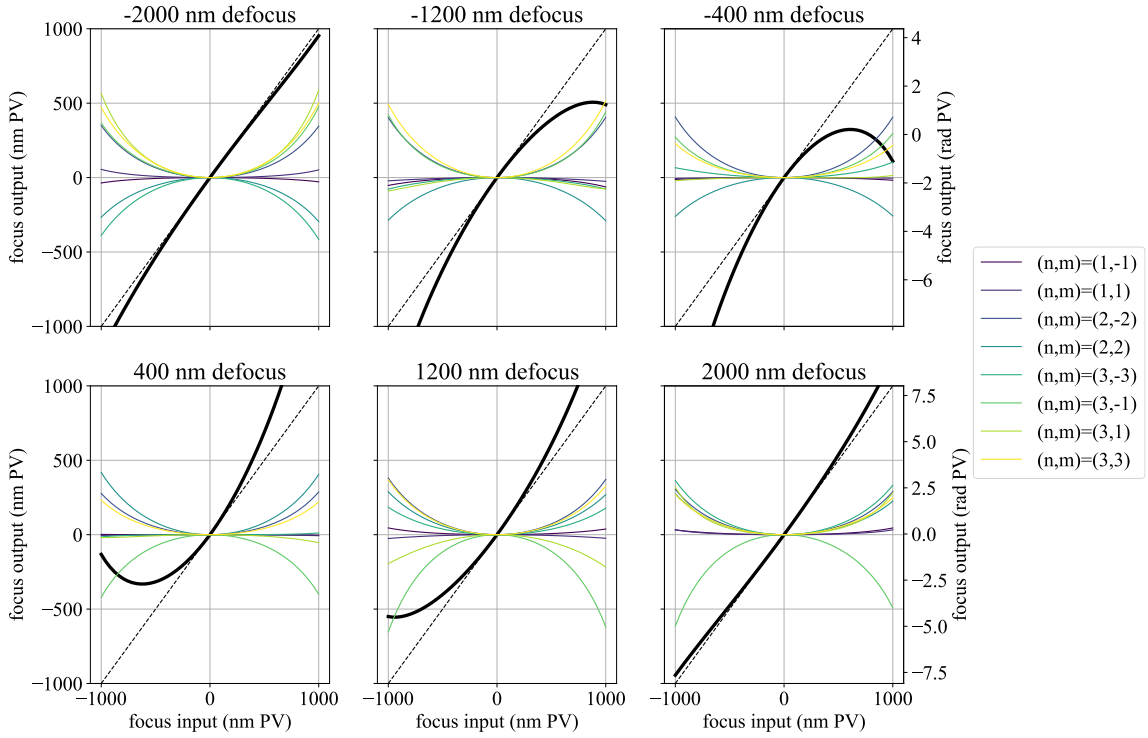


Figure 3.29 TGV LLOWFS linearity (solid black line) as a function of defocus amplitude in each panel. The dashed black line shows the curve for a perfectly linear sensor. Other colours show the cross terms from other low order modes.

the LLOWFS, highlighting the benefit of defocusing the LLOWFS detector. Although the SCC focus linear range is similarly still about ±500 nm PV, AO residuals should still not disqualify this LOWFS from operation in a second stage AO loop. However, if static offsets (e.g., during regular off-sky alignment procedures) are expected to reach these levels, Figure 3.30 clearly shows that the LLOWFS is the better sensor to use.

Lastly, I compare the WFS sensitivity to photon noise between the LLOWFS and SCC in Fig. 3.31. These results clearly illustrate that the LLOWFS can reach a deeper level of residual WFE for the same total number of photons as input to the entrance pupil over a range of possible input amplitudes. This lower residual from the LLOWFS is consistent with Fig. 3.26, showing that for a given entrance pupil input low order phase aberration, most of the signal will be directed outside the Lyot pupil and pinhole, thus with much more signal reaching the LLOWFS detector relative to the SCC. Note that the absolute values of the level of residuals in Fig. 3.31 should not be considered a realistic estimate of achievable closed-loop tip/tilt performance, as

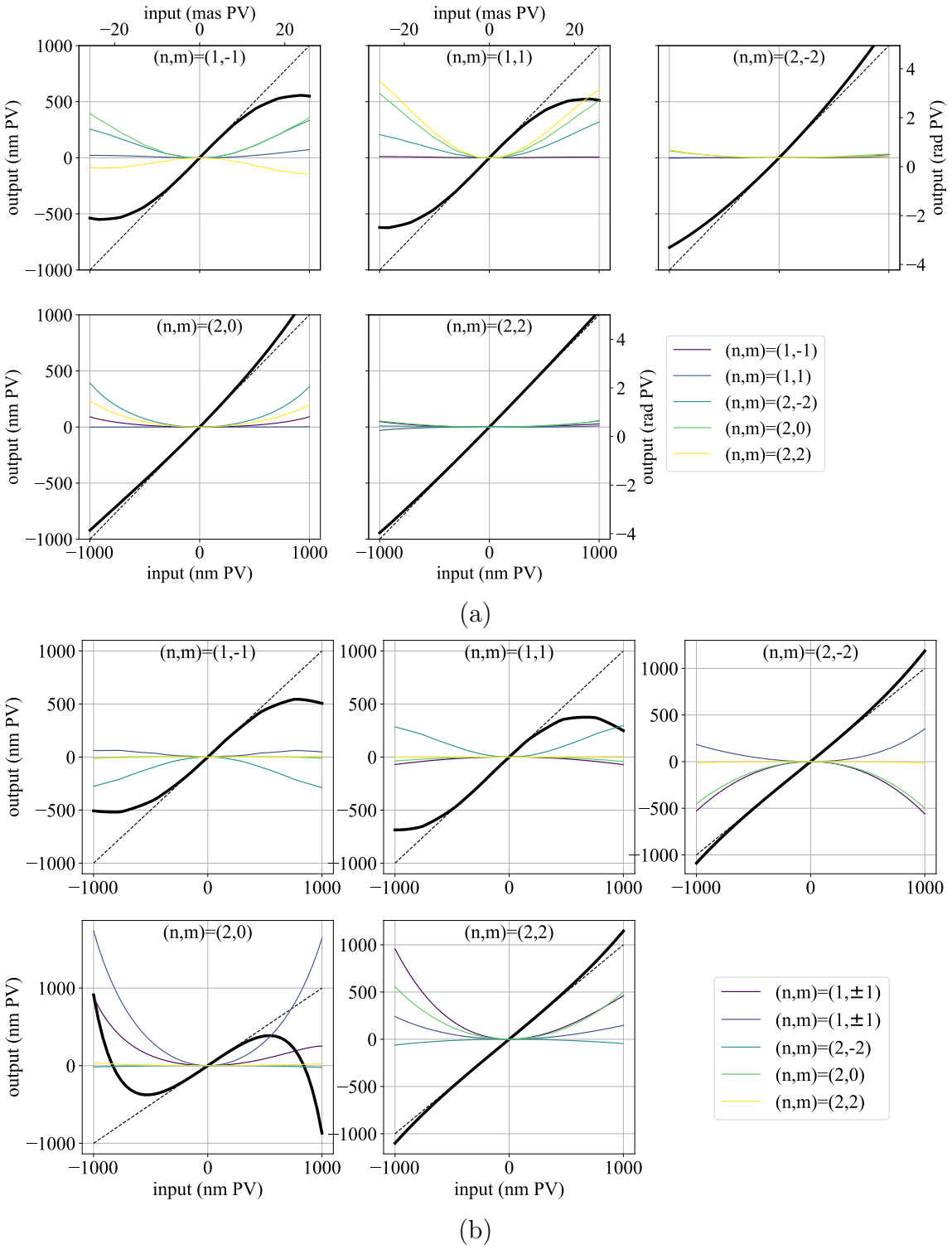


Figure 3.30 The linearity and cross terms (as in Fig. 3.29) of the first five Zernike modes for the LLOWFS (a) and SCC (b). Units in "mas" assume a 10m telescope.

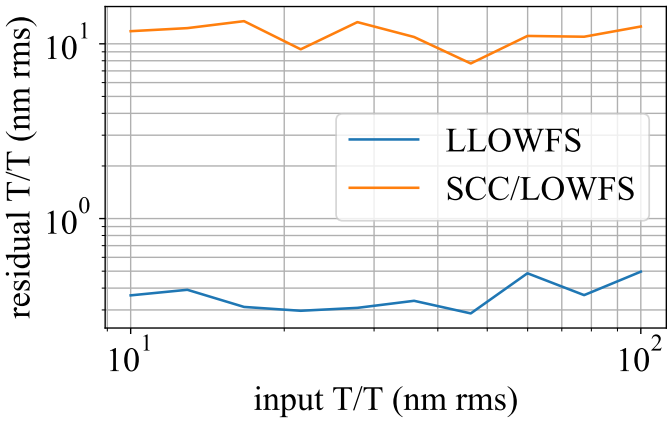


Figure 3.31 Sensitivity to photon noise for the LLOWFS vs. SCC, showing input vs. output tip/tilt ("T/T") WFE for a simulated (with photon noise included) 1 ms exposure, m_H=0 star, with all other parameters as in §3.2. Random directions of tip and tilt are combined (normalized to the given RMS input value) to produce an output residual WFE after DM correction; this process is repeated 10 times for each input RMS value, with the median output values then plotted here. Output residuals for each realization are computed from a single least-squares matrix multiplication (i.e., an open loop correction with a gain of 1) using the differential image as an input, as for Figures 3.29 and 3.30.

these simulations do not include any dynamic inputs and/or closed-loop controllers. Regardless, Fig. 3.31 can be interpreted as showing that the LLOWFS measurement of tip/tilt is less impacted by photon noise than the TGV+SCC LOWFS setup, with all other factors being equal, and so future detailed closed-loop simulations should clearly benefit from utilizing a LLOWFS.

3.5.5.2 Lyot-based High Order Wavefront Sensing

Extending the analysis from Fig. 3.26, Fig. 3.32 shows both (1; top row) the intensity of the differential Lyot plane electric field and (2; bottom row) the differential Lyot plane intensity images using the TGV FPM for three different 1 nm PV entrance pupil Fourier modes. In contrast to low order modes, the majority of the electric field signal for high order modes remains within the central Lyot pupil. This is to be expected, as these high order modes produce a signal similar to an off-axis exoplanet in the focal plane electric field, for which coronagraphs are designed to largely transmit such signals through the Lyot pupil to retain a high exoplanet throughput beyond the IWA (but see §4.3.2 for a further discussion). However, the bottom row of Fig. 3.32

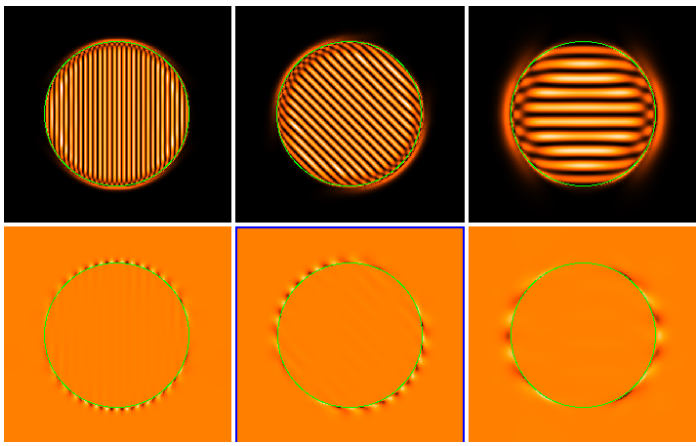


Figure 3.32 Top row: the intensity of the differential Lyot plane electric field for three different Fourier modes at high (left), mid (middle), and low (right) spatial frequencies. Bottom row: the differential Lyot plane intensity of the same three modes as the top row. The entrance pupil footprint is illustrated by a green circle in each panel.

shows that the small amount of electric field signal that is transmitted just beyond the central pupil generates a strong signal in differential Lyot plane intensity, suggesting that the LLOWFS detector could also serve as a high order WFS. This strategy is a particularly promising approach, since it has the potential, in principle, to optically measure and subtract the light from a given speckle before ever being recorded in the SCC image. Even though the SCC also has the ability to measure and subtract a given speckle, because this is also the science image, recording photons to measure a speckle with the SCC inherently increases the photon noise limit in contrast, thereby requiring more integration time to reach a given contrast needed to detect an exoplanet (again, see §4.3.2 for a further discussion on this principle).

In this section and dissertation I only consider high order wavefront sensing and control with pupil plane LLOWFS images. The challenge of the nominal focal plane LLOWFS configuration for high order wavefront sensing is more difficult, additionally imaging both the phase component of diffracted light in the Lyot plane and the diffraction effects of the complex apertures imparted on the wavefront (i.e., apertures defined by both the existing Lyot intensity distribution and and LLOWFS mask). I found that such effects, without additional coronagraph design optimization, generate strong non-linearities between the LLOWFS focal plane and DM plane, and thus I save this approach for the future work discussion in §4.3.2.

Thus, in Fig. 3.33 I consider an approach using LLOWFS pupil images to measure and subtract high order Fourier modes with the DM. In short, Fig. 3.33 shows that this LLOWFS pupil image cannot be used to sense the aberrations that form speckles in the SCC image. The corrected LLOWFS image in panel 4 does not subtract speckles in panel 6. This is also consistent with the differential Lyot intensity

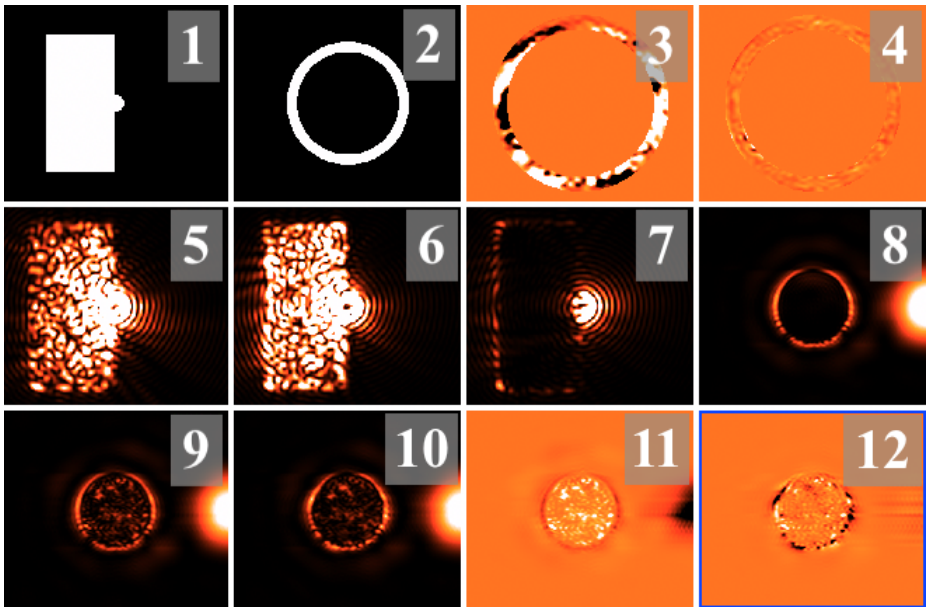
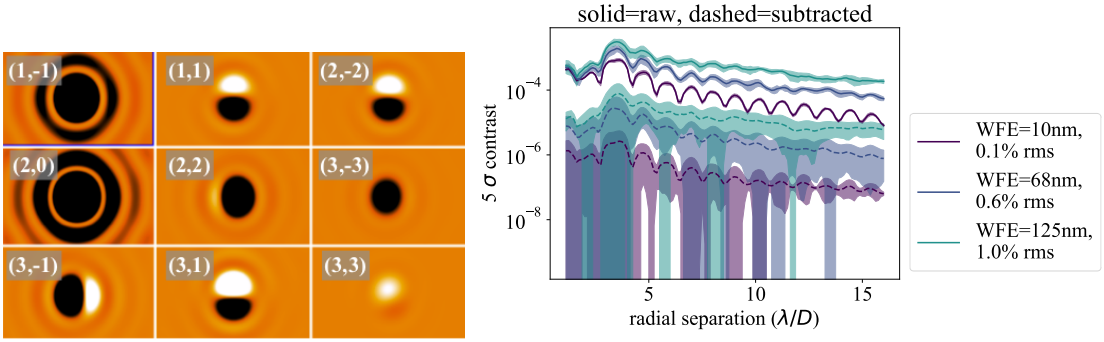


Figure 3.33 (1) A field stop simulated in the FPM plane, along with the TGV phase mask (not pictured), to isolate spatial frequencies only within a half DH, (2) a LLOWFS mask applied in the Lyot plane, reimaging the light hitting the white annulus (just outside the central pupil) to a separate LLOWFS pupil detector, (3) a differential LLOWFS pupil image, with the input as a 100 nm RMS phase screen and the reference as the DM commands used to generate the DH in panel 7, (4) a differential LLOWFS pupil image after measurement of panel 3 and DM correction (shown on the same scale as panel 3), (5) a coronagraphic image (with the SCC pinhole blocked) of the same input phase screen as in panel 3, (6) a coronagraphic image (on the same scale and setup as panel 5) with the same DM correction as used in panel 4, (7) a reference coronagraphic image with a DH generated by SCC DM correction of panel 5, (8) the Lyot pupil intensity with the DM correction from panel 7 applied, (9) the Lyot pupil intensity with the DM correction from panel 4 applied, (10) the Lyot pupil intensity with the atmospheric phase screen from panels 3 & 5 applied, (11) the difference between panels 9 & 8, and (12) the difference between panels 9 & 10.

images in panels 11 and 12, which confirm that the LLOWFS measurement and DM correction are indeed subtracting speckles "pinned" to the bright pupil "ring" (comparing the pupil edges between the two panels), but that this correction does not actually subtract speckles transmitted through to the SCC image (comparing the pupil centres between the two panels). With these results in context, future coronagraph design optimizations should consider approaches to improve the linearity of high order modes between the LLOWFS and SCC planes (§4.3.2).

3.5.5.3 CDI Strategy

Given that the most promising CDI strategy presented in §3.5.2, the direct pinhole PSF measurement, requires some form of additional hardware/setup to make feasible, here I consider another alternative CDI approach that requires only utilizing the on-sky SCC image. The principle is illustrated in Fig. 3.34 and described below:



(a) (b)

Figure 3.34 Illustration of a FAST CDI subtraction strategy, recording an off-sky interaction matrix of the pinhole PSF for low order Zernike modes (a; shown here for $n = 1 \rightarrow 3$ in the form "(n,m)") to enable on-sky pinhole PSF reconstruction and stellar speckle subtraction. (b) Simulation results using the TG mask and the on-sky SCC telemetry (as opposed to the synchronized on-sky LLOWFS telemetry). The plotted contrast curves show the median and standard deviation values for 10 different random input wavefront realizations.

1. In an off-sky calibration procedure, apply a series of low order Zernike modes to the DM, recording in the coronagraphic image plane for each mode:
   (a) the pinhole PSF and
   (b) the full SCC image and/or LLOWFS image.
2. Build an interaction matrix of differential images for both step 1a (illustrated in Fig. 3.34 a) and 1b.
3. Once on-sky, record SCC and/or synchronized LLOWFS images at a high frame rate.
4. (This and subsequent steps can be done offline.) Use each recorded SCC/LLOWFS image from step 3 to generate least-squares coefficients for low order Zernike modes from the interaction matrix taken in step 1b.

5. Apply the coefficients from step 4 to the offline pinhole PSF interaction matrix from step 1a to reconstruct the on-sky pinhole PSF.

6. Use the reconstructed pinhole PSF in equation 3.6 to subtract the coherent speckles in the recorded on-sky image (Fig. 3.34 b).

Fig. 3.34 b shows a series of simulated inputs with different WFE levels and the resulting output contrast levels using the procedure described above, demonstrating that the contrast gains appear to be a function of input WFE, with higher gains possible from lower input WFE. The more impactful gains, however, should come from averaging residuals, as in §3.5.2 for the on-sky pinhole PSF measurement approach. Figure 3.35 illustrates this principle. First, the input contrasts level out with time due to static features from diffraction; even with uncorrelated entrance pupil wavefront realizations, these un-subtracted diffraction rings will remain at the same position in the coronagraphic image for each new wavefront. Thus, it is to be expected that the input curve levels out with time. In principle, the two output subtraction curves should continue to gain in contrast with time if there are no static residuals. However, Fig. 3.34 clearly shows that there are static residuals. These limits were investigated for the cal algorithm in §3.5.2.2 and in Fig. 3.14, arising due to static

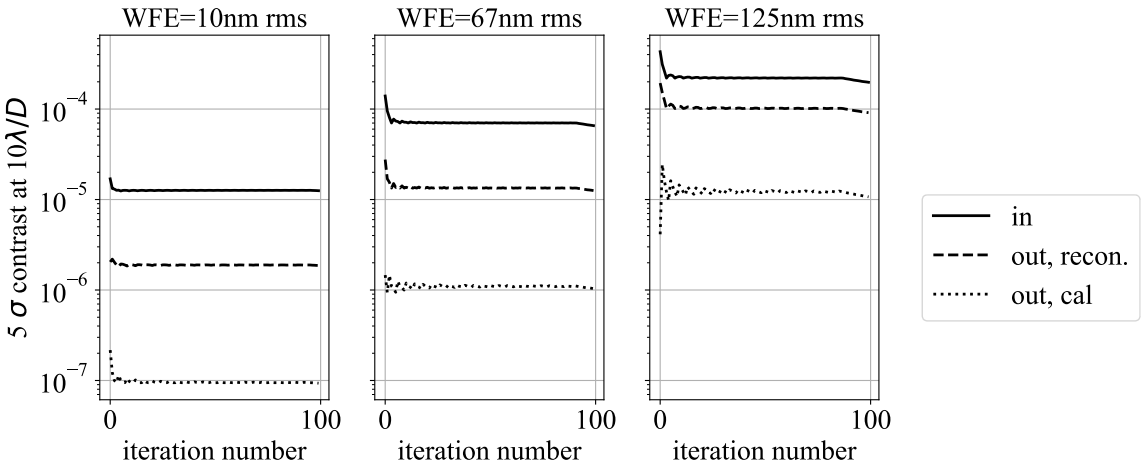


Figure 3.35 The impact of averaging residuals after CDI subtraction. Each panel shows a different input WFE level (and corresponding solid curve). The x-axis shows a proxy for time, where each "iteration number" represents a new uncorrelated wavefront realization from the given input WFE level, which is equivalent to a full pupil crossing assuming single layer frozen flow. The y-axis shows cumulative contrast at 10 λ/D as a function of iteration number. The "out, cal" curve is the same CDI method as used in §3.5.2.

offsets between low order aberrations in the calibrated pinhole PSF and on-sky pinhole PSF. However, Fig. 3.35 now shows these same limits for the proposed pinhole PSF reconstruction algorithm in this section, suggesting that similar static limits are still present. Furthermore, these limits are clearly stronger than the cal algorithm. Many factors could be causing this, such as the choice of basis for the pinhole PSF reconstruction,[4] the linear range where the assumption in step 5 above is valid (and its modal dependence for different low order Zernike polynomials), dependence on different FAST coronagraph architectures (e.g., TGV vs. TG FPM), and more. Future papers will consider these factors in investigating the limits of this approach, as this subtraction algorithm remains a promising CDI solution for on-sky operation without additional hardware modification to the nominal SCC operation.

3.5.6 Linearity and Sensitivity

This section presents unpublished work completed during my PhD studies.

WFS linearity and sensitivity are the two pillars of active wavefront sensing and control (Guyon, 2005). As the connection between the WFS and DM planes becomes more linear, the wavefront control loop can (1) increase operational dynamic range, being able to close the loop on higher RMS input WFE and (2) reach deeper levels of diffraction-limited residual WFE. As a WFS becomes more sensitive, a wavefront control loop can reach deeper levels of photon noise-limited residual WFE. In Guyon (2005), the Zernike wavefront sensor (ZWFS) (Zernike, 1934; N'Diaye et al., 2016) is characterized as the most sensitive WFS, albeit with limited linear range. The ZWFS is constructed by applying a $\pi/2$ phase shift in the focal plane, within only a λ/D diameter circular region centred around the optical axis, and re-imaging the down stream pupil plane, which because of this mask converts entrance pupil phase aberration into an intensity pattern on the WFS detector. In this section I will explore the principles of WFS linearity and sensitivity in application to the FAST TGV and ZWFS.

[4]For example, Fig. 3.34 appears to show that differential pinhole PSF images are mostly realized as local tip/tilt offsets in the Lyot stop pinhole, whereas higher order aberrations are still not seen, suggesting that higher order Zernike modes could instead be degrading achievable contrast.

3.5.6.1 Linearity

Fig. 3.36 shows the linearity of each WFS at low, mid, and high spatial frequencies. As in Fig. 3.30, the output value for a given mode comes from the value of the least-squares coefficient, in this case produced from a covariance matrix using the calibration procedure described in §3.5.3.1. Fig. 3.36 clearly shows that the TGV WFS is a more linear sensor than the ZWFS, particularly considering the cross terms. With that said, Table 3.2 catalogs the minimum and maximum for the linear range of each mode, showing that it is not that different between the two WFSs. In principle, this gain for the TGV could be advantageous over the ZWFS in regimes where the input turbulence is strong enough to be inside the linear range of the former but outside the latter, although Table 3.2 shows that this is a narrow window. Also considering the shape of the TGV linearity curve in Fig. 3.36, the turnover into the non-linear regime is more immediate than for the ZWFS. This non-linear TGV inflection point is clearly defined and identical for all three Fourier modes, occurring at the amplitudes at which second order sine spots appear above the on-axis TGV diffraction-limited contrast levels, thereby removing the linear relationship between first order sine spot intensity and DM sine wave amplitude (as in this non-linear case, increasing the DM amplitude increases the intensity of the second order sine spot rather than the first order spot).

Table 3.2: Linear range for the Fourier modes in Fig. 3.36. Fourier mode labels in the top row of this table are analogous to the "(x position, y position)" labels in Fig. 3.36 a for the sine spot on the left side of the image plane control region. Linear range maximum and minimum values, shown as "$^{+\text{max}}_{-\text{min}}$ nm", are measured from the respective maximum and minimum inflection points of each curve.

WFS	(x,y)=(-9.5,+7.5) cycles/pupil	(x,y)=(-6.5,+3.5) cycles/pupil	(x,y)=(-3.5,-0.5) cycles/pupil
TGV	$^{+388}_{-388}$ nm	$^{+388}_{-388}$ nm	$^{+388}_{-388}$ nm
ZWFS	$^{+385}_{-60}$ nm	$^{+304}_{-265}$ nm	$^{+265}_{-305}$ nm

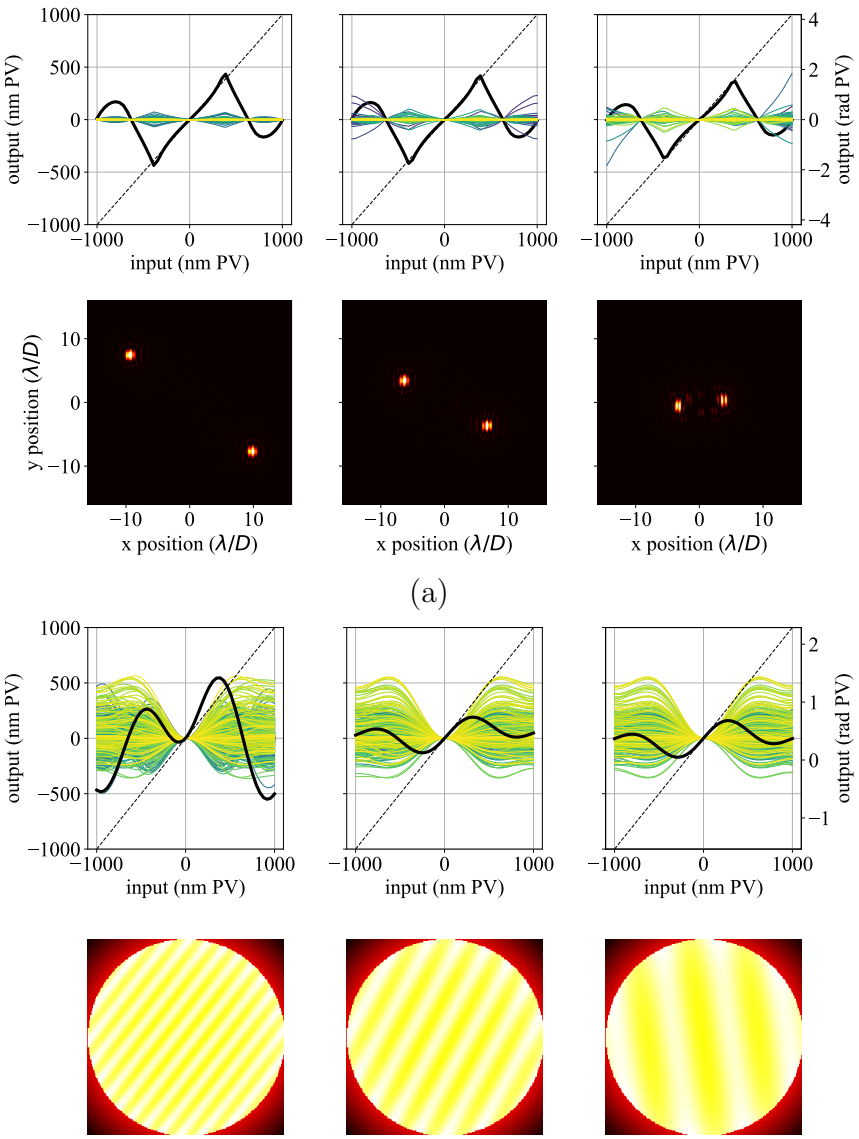
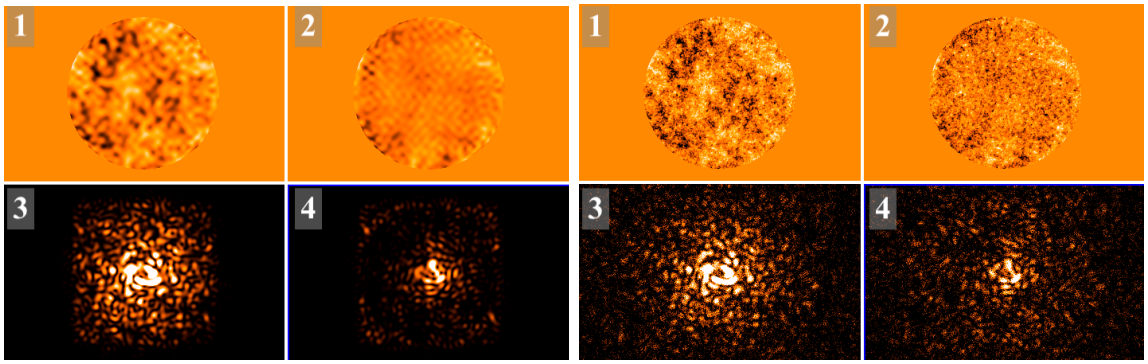


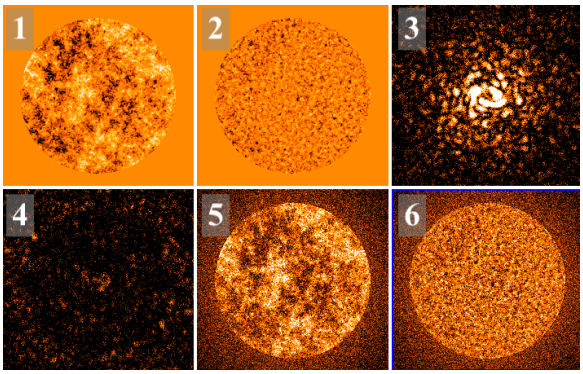
Figure 3.36 Linearity, as in Fig. 3.30, of three different Fourier modes (as illustrated in the bottom row for each panel) of high, mid, and low spatial frequencies (left, middle, and right column, respectively) for the TGV WFS (a) and ZWFS (b).

3.5.6.2 Sensitivity

Fig. 3.37 illustrates the achievable residual WFE and contrast for a single 100 nm RMS input phase screen, and Figure 3.38 corroborates these results through error budget simulations over multiple realizations. The three terms used in these error budget calculations are non-linearities, aliasing, and photon noise (typically the latter is referred to as "sensitivity"). The process used to generate these terms is described below, still using the same interaction matrix/calibration procedure as described in



(a) (b)



(c)

Figure 3.37 An illustration of WFS sensitivity for a single wavefront realization, before and after WFS measurement and DM correction with the TGV (a and b) and ZWFS (c). When photon noise is simulated, observations assume a $m_H = 0$ star and 1 ms exposure, with all other parameters as in §3.2. For each panel (a - c), each number represents the input entrance pupil wavefront (1), the output (DM-corrected) entrance pupil wavefront (2), the input coronagraphic image (i.e., with the Lyot stop pinhole blocked) generated from propagating the input wavefront in frame 1 (3), and the output coronagraphic image generated from propagating the output wavefront in frame 2 (4). In panel a, the input wavefront has been projected onto the DM Fourier modes (i.e., anti-aliased), and the output shows the residual only for TGV non-linearities (i.e., no photon noise is simulated). In panel b, the input includes a full power law (without any filtering) and photon noise in the coronagraphic images, and so the output includes the TGV non-linearities, aliasing, and photon noise components. In panel c, the inputs and outputs include the same setup as in panel b, except that correction is instead made by a ZWFS. Panels c.5 and c.6 show the corresponding input and output (i.e., with non-linearities, aliasing, and photon noise simulated) ZWFS images.

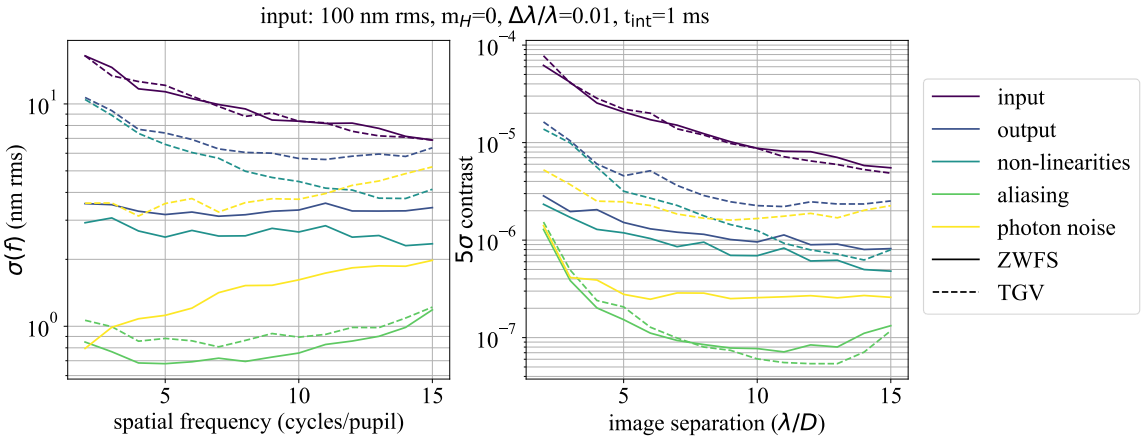


Figure 3.38 A comparison of WFS sensitivity between the ZWFS and TGV, showing the error budget components of each output. The y-axis of the left panel is defined by equation 1 from (Vigan et al., 2019). Each curve is generated from an average over 10 different uncorrelated wavefront realizations. The ZWFS output and photon noise curves assume a 30/70 beam splitter between the ZWFS and coronagraphic image path. The coronagraphic image path for the ZWFS uses a TGV coronagraph.

3.5.3.1:

1. For an input wavefront with a continuous power law, project this phase screen onto the DM Fourier modal basis using a least-squares (as illustrated in between Fig. 3.37 b.1 and a.1), functioning as an anti-aliasing filter.

2. propagate the wavefront from step 1 to the WFS plane, not simulating photon noise on the detector.

3. generate DM commands (assuming an open loop, unity gain correction) from the image in step 2. This subtracted entrance pupil plane wavefront represents the "non-linearities" component, as illustrated in Fig. 3.37 a.2.

4. propagate the residual pupil phase from step 3 to generate the coronagraphic image non-linearities component, as in Fig. 3.37 a.4.

5. repeat steps 1 - 4 but without the anti-aliasing filter, generating a corresponding residual phase screen for "non-linearities + aliasing."

6. compute the difference between the phase screens in step 5 and 3, generating the "aliasing" phase screen component.

7. propagate step 6 to generate the coronagraphic image aliasing component.

8. repeat steps 5 - 7, but instead of adding aliasing keep the anti-aliasing filter and add photon noise, generating the "photon noise" phase screen and coronagraphic image components.

9. the "total" output wavefront and coronagraphic image components are generated from a simulated image without the anti-aliasing filter and with photon noise, which is also analogous in WFE to the sum in quadrature of the three non-linearities, aliasing, and photon noise components.

Clearly, on almost all fronts, the ZWFS remains a more sensitive WFS than the FAST TGV WFS. First, the non-linearities component is lower for the ZWFS,[5] as this pupil plane WFS is not affected by diffraction nearly as much as the focal plane TGV WFS. Although the linearity analysis in §3.5.6.1 suggests that the TGV WFS should reach a lower level of non-linearities for a large enough input WFE, the 100 nm RMS input used in this section (reaching between ∼tens and ∼ones of nm RMS for different spatial frequencies as illustrated in Fig. 3.38) is still within the ZWFS linear range outlined in Table 3.2.

Next, the photon noise component is also clearly lower for the ZWFS. This discrepancy can be explained by the difference in absolute throughput to the detector plane between the two WFSs, as illustrated in Fig. 3.37 (b.3 vs. c.3) and previously in Fig. 3.32. Fig. 3.32 showed that, in both absolute and differential intensities, high order modes are mostly blocked by the Lyot stop and not transmitted to the SCC image, while for the ZWFS most of the light from high order modes is instead transmitted through to the detector plane. With this context in mind, Fig. 3.37 illustrates that although the TGV can effectively subtract speckles down to the photon noise limit in panel b.4, the total throughput in panel b.3 is significantly lower than in panel c.5 (i.e., even with the 30/70 beam splitter simulated in panel c but not panel b), thus enabling a much higher S/N wavefront measurement and correspondingly lower RMS WFE DM correction for the ZWFS. Because of this effect, the ZWFS is ultimately able to subtract the speckles in panel c.3 to a level in c.4 that is significantly below the achievable photon noise limit in panel b.4.

Lastly, aliasing appears similar between the two sensors, both with negligible impact on the total WFE and contrast. With that said, if future ZWFS and/or TGV

[5] although note that the gap for non-linearities between the two WFSs is relatively larger in WFE space than in contrast space, as contrast incorporates the TGV measurement and subtraction of diffraction (as illustrated in Fig. 3.37 a.4) whereas pupil plane WFE does not.

mask designs can bring both the non-linearities and photon noise components down to this sub-nm RMS level, further developments should then also consider optical anti-aliasing methods, such as a spatial filter (Poyneer & Macintosh, 2004, and see §4.3.2).

3.5.7 Laboratory Simulations and Results

This section utilizes text from Gerard et al. (2018b) (§3.5.7.1) and Gerard et al. (2019a) (§3.5.7.2, 3.5.7.3), as well as unpublished work completed during my PhD (§3.5.7.4).

Laboratory testing of FAST is the next step towards on-sky deployment, complementary to the simulations discussed so far. In this section I will present published results on this front, including fabrication tolerance simulations for my TG FPM (§3.5.7.1) and the first laboratory tests of FAST (§3.5.7.2, §3.5.7.3, and §3.5.7.4).

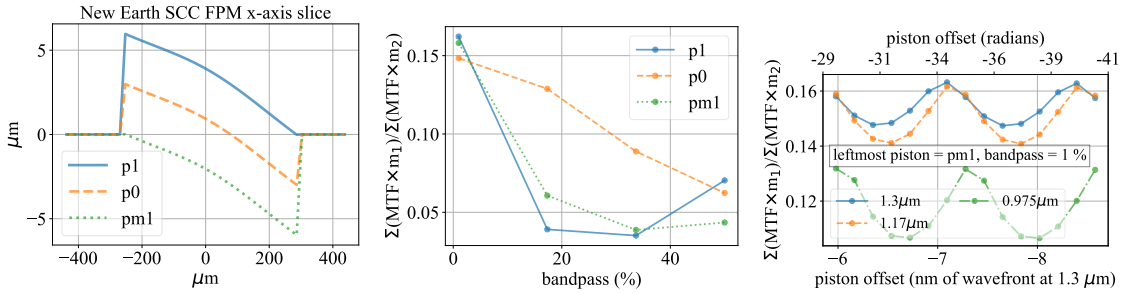
3.5.7.1 Manufacturing Simulations

3.5.7.1.1 Numerical Simulation Parameters For my simulations in this section, I will use $\lambda_0 = 1.3$ μm (where λ_0 is the central wavelength optimized for the TG FPM), and an accordingly optimized TG FPM Gaussian FWHM of $4.7\lambda_0/D$ and amplitude of 2.22 μm. I sample and integrate over 5 different wavelengths across each bandpass to simulate broadband performance. The simulations in this section do not include photon noise.

Chromatic magnification effects are simulated in the FPM plane by spatially rescaling the phase shift applied by the TG FPM (i.e., the IWA, and Gaussian FWHM are scaled by λ/λ_0, where λ is a simulated wavelength within the desired bandpass). This effect causes a relatively less-optimal throughput of intensity through the Lyot stop pinhole at wavelengths other than λ_0. In sections 3.5.7.1.2 and 3.5.7.1.3 I will want to understand how chromatic differences of differential piston between the Lyot stop pupil and pinhole affect integrated fringe ratio (defined in §3.5.7.1.2). Because this degradation is separate from how magnification with wavelength affects integrated fringe ratio, I do not simulate magnification with wavelength in the detector plane. Therefore, the integrated fringe ratio values will in reality be worse than the results presented in this section at wavelengths further away from λ_0.

3.5.7.1.2 Integrated Fringe Ratio Although my initial goal to do FAST atmospheric wavefront sensing optimized the integrated intensity going through the reference pinhole using the TG FPM (Gerard et al., 2018a), I instead consider here the metric of integrated fringe ratio. I define integrated fringe ratio as the cumulative flux in the MTF side lobe of an SCC image divided by the cumulative flux in the central beam of the same MTF, which is denoted as $\Sigma\left(\mathrm{MTF} \times \mathrm{m}_1\right) / \Sigma\left(\mathrm{MTF} \times \mathrm{m}_2\right)$, where m_1 and m_2 are binary masks which filter the MTF side lobe and central beam, respectively, as in Fig. 3.4. This value is similar to my previous metric of power through the pinhole relative to the central pupil (Table 3.1); in both cases higher values correspond to higher fringe amplitudes at a single wavelength, where in an optimal regime these numbers are close to unity so a speckle can be measured as soon as it is detected. However, once we start to consider fringe detection over a larger bandpass, these two metrics will differ. Because the phase shift applied on the complex wavefront by the TG FPM is chromatic, the downstream complex wavefront transmitted through the pinhole will vary as a function of wavelength. Accordingly, the integrated fringe ratio will be lower over a broadband. The intensity pattern in the Lyot plane will only be optimized (i.e., the most concentrated around the pinhole) for a single wavelength, and differential phase offsets between the pupil and the pinhole could cause nulling effects that vary as a function of wavelength. Rather than measuring flux in the Lyot plane, cumulative flux in the MTF is therefore a better metric to measure this type of integrated fringe ratio degradation.

With these integrated fringe ratio effects and metric in mind, I considered three different possible TG FPM designs to be manufactured for laboratory testing at NRC's Extreme Wavefront lab for Exoplanet Advanced Research Topics at Herzberg (NEW EARTH) laboratory at the National Research Council of Canada in Victoria (see §3.5.7.4), illustrated in Figure 3.39 a. I have actively been working with collaborators and vendors to determine options for manufacturing the TG FPM from Gerard et al. (2018a). A four step grey scale lithography process—involving (1) fabrication of a greyscale mask, (2) an ultraviolet lithographic process to transfer the grey scale pattern into photoresist, and (3) etching to transfer the photoresist pattern into fused silica, and (4) coating—would require using the pm1 design (S. Thibault, private communication). However, another liquid crystal-based fabrication process could allow using the p0 design (F. Snik, private communication). The p1 and pm1 designs create a chromatic piston phase discrepancy between the pinhole and central pupil in the Lyot plane. Figure 3.39 b illustrates that this effect significantly



(a) (b) (c)

Figure 3.39 (a) Three different possible reflective TG FPM designs for the NEW EARTH testbed FPM plane at f/67 and λ_0=1.3 μm, showing a slice along the x-axis zoomed in around the central TG region. (b) Integrated fringe ratio vs. bandpass ($\Delta\lambda/\lambda_0 \times 100$) for the three different designs from panel a. (c) Integrated fringe ratio vs. additional negative piston offset for the pm1 design from panel a, cycling through 4π radians below the pm1 piston value.

decreases the integrated fringe ratio over a large bandpass. Thus, the pm1 design will not be a viable option for fabrication of a broadband TG FPM. However, for a narrow bandpass, which I will consider in this section, all three designs provide a similar integrated fringe ratio.

Figure 3.39 c shows the simulation results of how additional piston for a 1% bandpass affects integrated fringe ratio. Varying the pm1 design by an additional 4π radians, constructive and destructive interference cycle the integrated fringe ratio through 2 full periods of oscillation. This curve is shown for three different wavelengths, but for all cases I use the same TG FPM, optimized for λ_0 =1.3 μm. Thus, integrated fringe ratio is highest across all piston phase offsets for $\lambda = \lambda_0$, lower for $\lambda = 1.17$ μm, and lowest for $\lambda = 0.975$ μm (i.e., more light goes through the pinhole closer at wavelengths closer to λ_0). Additionally, the amplitude of integrated fringe ratio oscillation as a function of continuous piston offset is higher for wavelengths further away from λ_0, again illustrating that the TG FPM is a inherently chromatic design.

3.5.7.1.3 Surface Figure Error Now informed by the results of §3.5.7.1.2, Figure 3.40 shows integrated fringe ratio vs. phase WFE on the surface of the pm1 TG FPM design using a 1 % bandpass centred around λ_0. I also assume a 1% RMS, -2 power law amplitude error on this surface. Imposing a integrated fringe ratio tolerance requirement of better than 0.075 (i.e., better than a factor of two degradation),

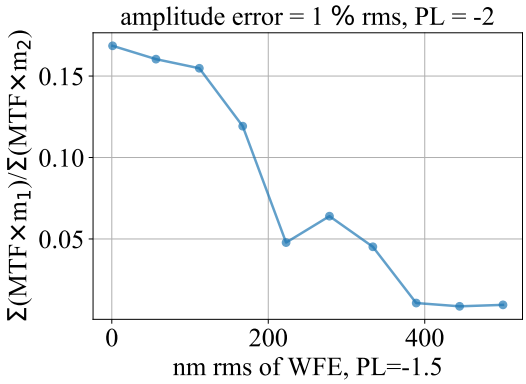


Figure 3.40 Integrated fringe ratio vs. phase WFE on the surface of the pm1 TG FPM design, assuming an additional static level of amplitude aberration. Assuming a integrated fringe ratio tolerance requirement of less than about a factor of two from the optimal value, a reflective pm1 TG FPM design would require less than about 140 nm rms of surface figure error.

Figure 3.40 illustrates that this would require $\lesssim$200 nm rms of WFE, or $\lesssim$100 nm rms of surface figure error for a reflective TG FPM design, motivating the rough precision needed for fabrication.

3.5.7.2 First laboratory Tests at LESIA

The work described in this section was completed during my NSERC CREATE TEPS internship at Paris Observatory in Meudon, France in February - March, 2019.

A transmissive coronagraphic mask was designed for operation at 700 nm on the Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique—Laboratory for Space Studies and Instrumentation in Astrophysics (LESIA) high contrast imaging testbed (THD2; Baudoz et al. 2018) and fabricated by Zeiss Optics. Fabrication was done by ion etching into a SiO_2 wafer, followed by anti-reflective coating for 700 nm. Eleven different samples were all fabricated from the same wafer; I used microscope images to identify the best mask by inspection, as shown in Figure 3.41 a. Due to constraints set by other mask designs also being fabricated on the same wafer, the peak-to-valley optical path difference was limited to 2π radians; using the TG FPM design from §3.5.1.1 and the position of the off-axis Lyot pinhole on the THD2 bench, the IWA of the mask was limited to 0.7 λ/D. With no apodizer and/or phase mask

design outside this central 1.4 λ/D diameter TG region, my goals here, rather than to demonstrate a deeply corrected contrast, were to provide an initial measurement and characterization of fringe ratio in the pupil and focal planes downstream of the aligned FPM. With this in mind, Fig. 3.41 b - e shows a compilation of FAST results from the LESIA high contrast testbeds, obtained in March 2019.

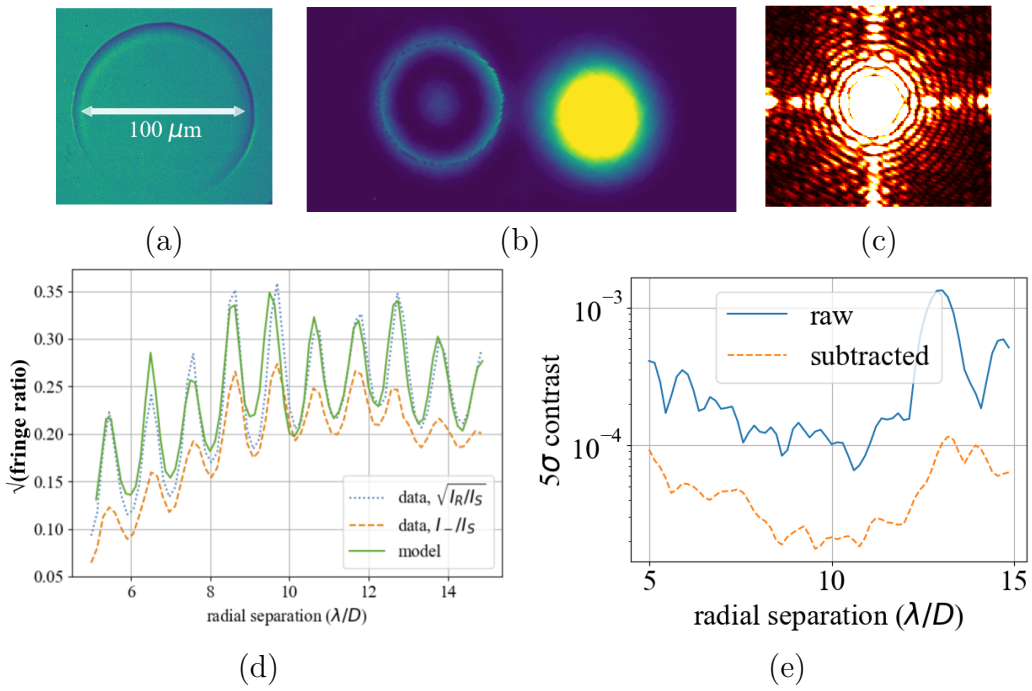


Figure 3.41 Laboratory tests at LESIA. (a) Microscope image of the fabricated TG FPM, (b) Lyot pupil image, (c) SCC image, (d) fringe ratio curves, and (e) contrast curves showing the image from panel c (i.e., raw) and my best subtraction.

Initial testing at LESIA was carried out in the "Corono Lab" (Baudoz et al., 2012b), formerly the original THD lab and now used for initial inspection, alignment, and testing of coronagraphic masks before transferring them to the main THD2 lab. After microscope inspection and mask alignment on the bench, the Lyot plane pupil image using this mask is shown in Fig. 3.41 b; the THD2 bench cannot be configured to take analogous coronagraphic pupil images, with the FPM aligned and Lyot stop removed (i.e., taken on the same camera as the focal plane images using a deployable lens to re-image the pupil), although coronagraphic pupil images can still be acquired with the Lyot stop in (Baudoz et al., 2018). The Corono lab uses a standard Lyot stop coronagraph design with no off-axis pinhole, and so coronagraphic images with SCC fringes can only be obtained on the THD2 bench. Note that the f number in the

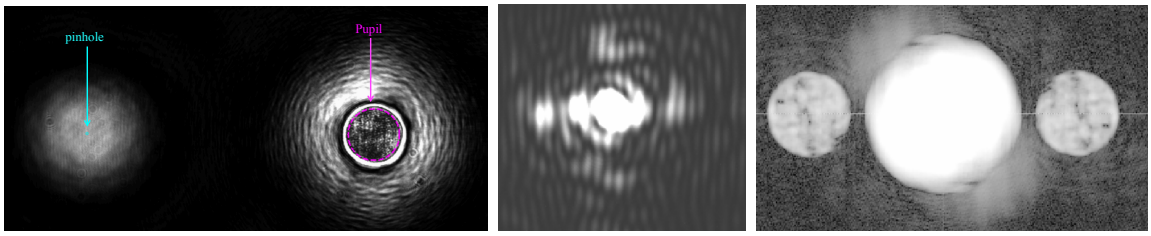
FPM plane is 30% smaller in the Corono lab than in the THD2 lab (i.e., providing a TG FPM tilt angle that is 30% shallower and an IWA than is 30% larger). Assuming an off-axis Lyot stop pinhole lies at the peak intensity of the off-axis pupil in Fig. 3.41 b, I measure a rough integrated Lyot plane fringe ratio (i.e., cumulative of the intensity in the pinhole divided by cumulative intensity in the central pupil) of 4%, consistent with simulations that account for the mask design and bench setup.

Next, bringing the same mask to the THD2 lab produced Figures 3.41 c - e. In Fig. 3.41 d I calculate fringe ratio as a function of separation by recording the fringed SCC image (Fig. 3.41 c), the coronagraphic image with the off-axis Lyot pinhole closed (I_S), and the pinhole PSF image with the central Lyot pupil blocked (I_R). With the these three images, fringe ratio can be calculated in two different ways, both of which are shown in Fig. 3.41 c. One way uses the two unfringed images (I_S and I_R), while the other uses the fringed image and the unfringed coronagraphic image (I_S); the latter isolates the fringe amplitude in the Fourier plane of the SCC image (OTF) to generate I_- (equation 3.9). Mathematically and numerically, these two methods should yield exactly the same curve; observationally they do not. I will return to a discussion of this discrepancy in the following paragraph. Finally, 5σ contrast curves are shown in Fig. 3.41 e for the raw input image (Fig. 3.41 c) and my best subtraction (using a direct measurement of the pinhole PSF on the same wavefront realization as in §3.5.2.1), illustrating that contrast gain is limited to a factor of about 5 at all separations.

The main factor causing the discrepancies in Fig. 3.41 d and the subtracted contrast limitations in Fig. 3.41 e arise from my coronagraph design and limited dynamic range in the coronagraphic image. As described above, because my IWA was limited to 0.7 λ/D, the center of the image in Fig. 3.41 c is about 300 times brighter than speckles at mid to high separations from the central star. As a result, to prevent saturation in the PSF core (filtering saturated SCC images in the Fourier domain can cause significant systematic effects, as discussed below), speckles from 5 to 15 λ/D are only detected at about a hundred counts above the read noise of the Andor s-CMOS detector (Baudoz et al., 2018). Thus, the marginal gain in Fig. 3.41 e is actually reaching the background noise limit; in order to go deeper, I would need a better coronagraph design and/or a higher dynamic range detector. In future laboratory demonstrations and papers I will use the TGV design (§3.5.4.2, 3.5.7.4), which rejects much more light outside of the central pupil in the Lyot plane and should therefore significantly improve over these current limitations. Similarly,

in Fig. 3.41 d, the discrepancy between the two measured fringe ratio curves are attributed to the same "low S/N effect." Photon noise from the bright central core of the coronagraphic image is about 10 times higher than noise at mid to high spatial frequencies; as a result, the side lobes in the image MTF are detected at a S/N about 10 times lower than it could be with a better coronagraph. Instead, a better algorithmic approach is to use Fourier filtering in the image plane to attenuate this effect (typically a Butterworth filter), masking the central star and its associated photon noise. However, systematic effects, trading off between decreasing photon noise from the central star and decreasing signal from the fringes, still limit this approach. The orange dashed curve in Fig 3.41 d uses a particular set of Butterworth parameters to highlight this systematic discrepancy.

3.5.7.3 ETH Zürich Tests



(a) (b) (c)

Figure 3.42 Laboratory images from the Swiss Federal Institute of Technology Zurich (ETH Züruch) High Contrast Imaging Lab (Kühn et al., 2018). (a) Lyot plane intensity, before a Lyot stop is applied. As labeled, the pinhole is located on the left, while the coronagraphic pupil is located on the right. (b) SCC coronagraphic Image. (c) MTF of panel b.

Additional tests were performed using my TGV design in the ETH Züruch high contrast imaging lab (Kühn et al., 2018), which includes an adaptable coronagraphic setup using a liquid crystal on-silicon spatial light modulator (SLM) in the focal plane. The TGV design was applied to this SLM to produce the images in Fig. 3.42. Using the illustrated "pupil" and "pinhole" apertures in Fig. 3.42 a, I measure an integrated Lyot plane fringe ratio of 7.6%, again consistent with simulations.

Similar to §3.5.7.2, I was not able to subtract the recorded SCC image due to limited dynamic range in the coronagraphic image. In this case, a bright non-coronagraphic PSF is present in the coronagraphic image due to back-reflections inside the SLM structure, notably at the glass/liquid crystal interface which cannot

be anti-reflection coated (J. Kühn, private communication). To prevent saturation from this "leakage term," the MTF sidelobes in Fig. 3.42 c are again detected close to the background noise limit. Interestingly, this leakage term is not coherent with the wavefront for which the TGV phase shift is applied, suggesting that the optical path difference from the aforementioned back-reflections is not negligible vs. the coherence length. This is illustrated by (1) the absence of fringes on the leakage term in the coronagraphic image and (2) the absence of the low order features in the MTF sidelobe compared to the central lobe. However, saturating the leakage term in the coronagraphic image would create systematic ringing and aliasing effects in the MTF at the higher spatial frequencies of the fringes (some of which are already seen in Fig. 3.42 c), and so algorithmic filtering of this incoherent light is still not an optimal solution to reaching deeper contrasts. Future demonstrations of FAST using liquid crystal technology will implement methods to increase the attenuation of this leakage term, such as solutions proposed by Doelman et al. (2017) or Janin-Potiron et al. (2019).

3.5.7.4 NEW EARTH Laboratory

The NEW EARTH laboratory is Canada's first dedicated testbed to develop new high contrast imaging technologies, a strategic investment of over $500k CAD by the National Research Council of Canada, Herzberg Astronomy and Astrophysics (NRC-HAA) Astronomy Technology Program. The NEW EARTH lab was introduced earlier in this dissertation (§3.5.7.1.2) and is briefly summarized in §6.1 of the unpublished Canadian Long Range Plan 2020 White Paper number W059. The NRC-HAA exoplanet imaging and AO research groups have been working on developing this laboratory facility since October 2016, retrofitting a section of the NRC-HAA library into a lab facility and designing, developing and installing hardware and software to test FAST, the first main goal of the NEW EARTH lab. In this section I will present some of our initial preliminary results (as of this writing in February 2020) to which I have contributed. Note that as of this writing the NEW EARTH lab is still being assembled, tested, and optimized, and as a result I was not able to include more laboratory data in this dissertation.

Figure 3.43 illustrates some of the first NEW EARTH lab results, related to testing and validation of a fabricated TGV FPM. By depositing four metal layers (binned from the continuous TGV design in §3.5.4.2), this mask fabrication method—

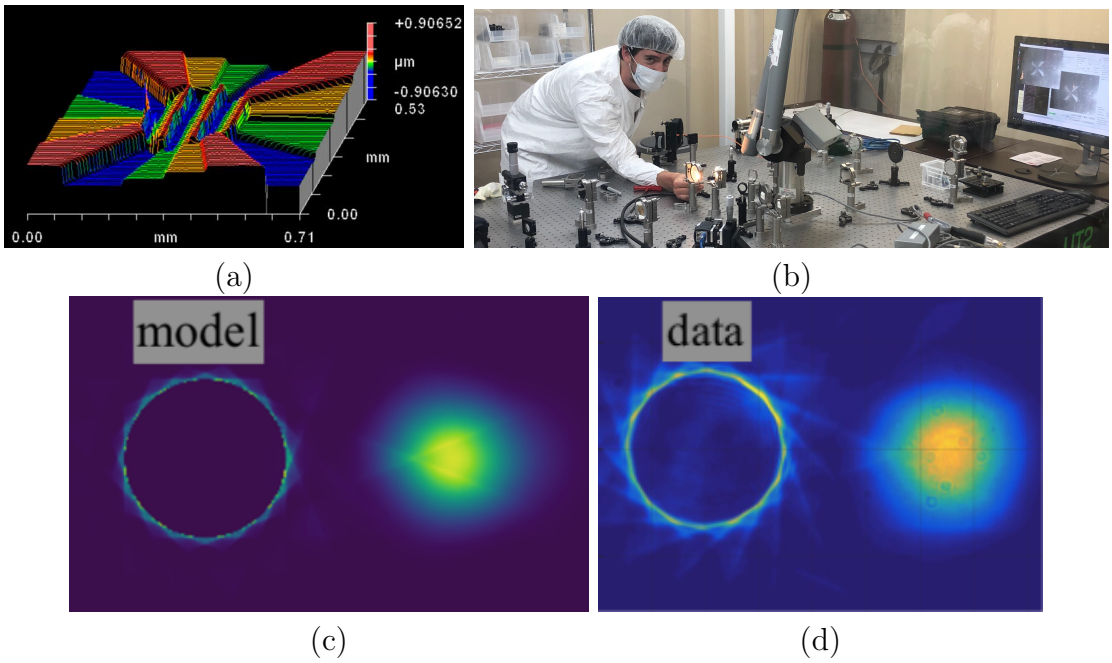
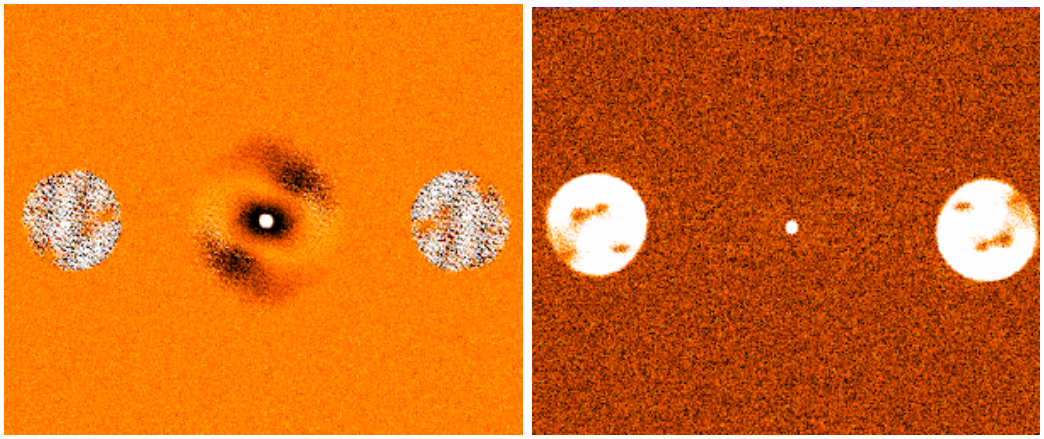


Figure 3.43 (a) Interferometer measurement of surface profile for the TGV FPM (3.5.4.2), fabricated by 4-layer aluminum deposition at King Abdullah University of Science and Technology's Computational Imaging Group (PI: W. Heidrich) (b) Laboratory image of me during the initial optical alignment of this coronagraphic mask. (c) and (d) simulated and measured images, respectively, of Lyot plane intensity distribution with the TGV mask aligned.

developed and made by King Abdullah University of Science and Technology's Computational Imaging Group (PI: W. Heidrich)—yields an inherently reflective surface, consistent with the GPI FPM design (Soummer et al., 2009; Macintosh et al., 2014) and enabling an achromatic tilt phase shift to support future broadband endeavours as discussed in §4.3.1 (a transmissive FPM would yield a chromatic tilt phase shift, un-optimally shifting the off-axis pupil position in the Lyot plane as a function of wavelength). Note the phase wrapping of the tilt angle in the central TG region; this four layer phase-wrapped reflective design and fabrication method were chosen to meet monochromatic tolerance requirements (§3.5.7.1) after first unsuccessfully trying a higher layer, higher dynamic range design via lithographic etching. More details about the fabrication procedure will be presented in a future paper. An image of me aligning the mask on the NEW EARTH bench using a flat field source at the FPM plane is shown in Fig. 3.43 b. In panels c - d, the simulated vs. measured Lyot plane images (i.e., after alignment of the mask on the PSF) are compared, showing good agreement of both diffraction structure and concentrated intensity of the off-

axis pupil. Note that currently due to an insufficiently bright source and absence of a DM in the beam, a detailed contrast curve analysis in comparing simulated vs. measured coronagraphic images as in §3.5.7.2 has not yet been completed. With the current setup, acquired coronagraphic images only detect the diffraction level around the first ∼few λ/D before hitting the SCC detector read noise, preventing a detailed characterization out to further separations without unrealistic exposure times; installation of the DM (to generate brighter speckles) and a brighter light source are both planned for the near future.



(a) (b)

Figure 3.44 A comparison of the differential MTF between a fringed and unfringed image, showing results from initial data (left), where the two images are acquired 30 s apart from one another, compared to the stack of consecutive differential images acquired using an optical chopper on the Lyot stop pinhole running at 60 Hz (right).

Next, Fig. 3.44 shows preliminary results suggesting that use of an optical chopper to quickly "blink" the Lyot stop pinhole on and off enables "freezing" the local turbulence on timescales longer than the difference between two images (i.e., 17 ms for the right panel), whereas differential turbulence on the timescale of 30 s for the left panel clearly shows a residual in the central MTF lobe above the photon noise. This pinhole "chopping" approach is not a new idea, first proposed by Give'on et al. (2012) and then later by Martinez (2019), although this is the first laboratory demonstration run at high frame rate to demonstrate the benefit of freezing the turbulence. For the same reasons discussed above, I was not yet able to measure the contrast gains in the coronagraphic image by CDI and/or DM control with this chopping mode operational, although such tests are planned in the near future. Regardless, these initial results in Fig. 3.44 clearly illustrate encouraging potential for use of both the TGV

FPM and optical chopper in reaching deeper contrast limits with a FAST speckle subtraction scheme.

3.5.8 Scientific Potential

This section utilizes text from Gerard et al. (2019a).

In §1.3.6 I discussed a driving future goal of the direct imaging field: habitable exoplanets. In this section I will show how my FAST technology could enable this on future telescopes and also illustrate additional new science that FAST could enable with current facilities.

Fig. 3.45 illustrates a few possible new science cases that FAST could enable for both current 10 m-class telescopes and future ELTs. The following paragraph describes the setup and for the simulations used to generate the science cases presented in this paragraph. While probably optimistic, the simulations show that fast focal plane wavefront sensing could enable the first direct images of indirectly detected exoplanets. Although host star radial velocities have revealed a plethora of such giant exoplanets and measured their masses, directly imaging their reflected starlight would provide an atmospheric characterization and crucial dynamical mass

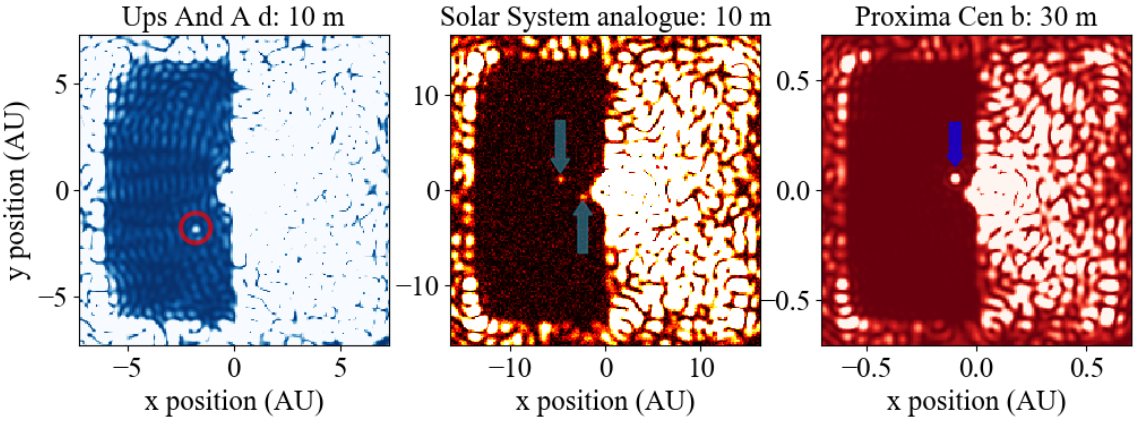


Figure 3.45 Simulated optimistic FAST science cases for current 10 m- (left and middle) and future 30 m- (right) class telescopes. Left: A 0.85 μm, 5 hour exposure of the radial velocity exoplanet Ups And A d (10 $M_{\rm Jup}$, 2.5 AU at 13.5pc; Butler et al. 1997). Middle: A 1.65 μm, 1 hour exposure of two 1 $M_{\rm Jup}$ planets at 5 and 10 AU orbiting a 30 Myr star at 30 pc. Right: A 1.22 μm, 30 minute exposure of the radial velocity exoplanet Proxima Cen b (1-3 Earth masses in the habitable zone at 1.3pc; Anglada-Escudé et al. 2016).

calibration (since mass can only be inferred through imaging). This synergy and improved calibration will also further increase with new astrometric detections from Gaia (Sozzetti, 2017) that can also similarly be directly imaged (both for young stars in thermal emission and old stars in reflected light); e.g., such calibrations would provide a direct window into measuring the mass-luminosity relation for young gas giants to resolve existing discrepancies between cold and hot start cooling curves. Such detections would also serve as a technology demonstration of the tolerances and requirements for reflected light imaging of habitable exoplanets on ELTs. Fast near-infrared focal plane wavefront control with current facilities could also detect thermal emission from young ($\lesssim$ 100 Myr), self-luminous exoplanets down to 1 $M_{\rm Jup}$ at 5 au around 40 targets, as illustrated in Fig. 3.46. Since current direct imaging surveys are not yet sensitive to these Solar System scales (Nielsen et al., 2019), fast focal plane wavefront sensing methods, such as FAST, could provide the first esti-

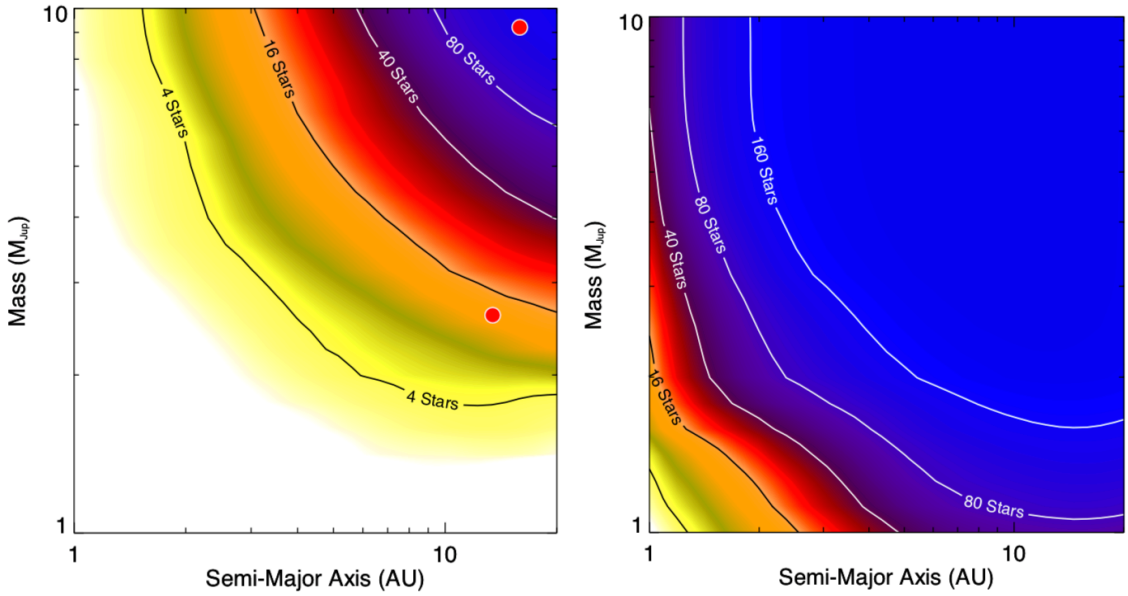


Figure 3.46 Occurrence rate simulations (as described in §1.3.3) without (left, from Nielsen et al. 2019; red points represent detected exoplanets) and with FAST (right; E. Neilsen, private communication). In the right panel, 1 $M_{\rm Jup}$ at 5 au lies along the 40 star contour (i.e., contours represent the number of stars around which an instrument would be sensitive to detecting an exoplanet at a given mass and semi-major axis), illustrating that an instrument with FAST would be sensitive to Jupiter analogs around at least 40 stars, whereas current sensitivity (left panel) is insensitive to this regime of mass-separation parameter space. Note that these simulations use the hot start BT-Settl model atmospheres (Allard, 2014), for which models are not available below 1 $M_{\rm Jup}$, explaining the abrupt cut off in the right panel.

mates of the commonality of Jupiter "twins" beyond our Solar System. Atmospheric characterizations for these low-mass/cold exoplanet science cases would open a new window into understanding the formation and evolution mechanisms of the coldest "Y-type" gas giants exoplanets, currently seen only beyond the Solar System in analogue field brown dwarfs (Dupuy & Kraus, 2013). As a result of these observations and calibrations with giant exoplanets, once the ELTs are available in the late 2020s methods such as FAST will enable the direct detection of exoplanets in the habitable zone around the nearest stars.

The simulations used to generate Fig. 3.45 are run as follows. All simulations use my TGV coronagraph design (Gerard & Marois, 2020, §3.5.4). As described in §3.2, speckles are generated from an assumed 100 nm rms residual AO error, 25 nm rms static error, and 1% rms static intensity error, all normalized in the entrance pupil. The bandpass central wavelength for each panel is assumed to scale the aforementioned wavefront error, and I assume use of the full bandpass to count the number of photons reaching the coronagraphic image (from left to right in Fig. 3.45: I band, H band, and J band). However, a monochromatic Fraunhofer propagation at the center of each bandpass is otherwise used to simulate an image. Although with this approach I do not include effects of wavefront chromaticity, focal plane mask chromaticity, and/or fringe smearing with wavelength, the point of these simulations is not to demonstrate end-to-end broadband performance. Instead, I use these simulations to illustrate the potential science cases that fast focal plane wavefront sensing methods, such as fast FAST, could optimistically enable. With this in mind, I assume that the monochromatic performance gain simulated in §3.5.2 (i.e., continually reaching the photon noise limit with increasing exposure time) will be the same in my broadband setup here; a separate forthcoming analysis will address these assumptions with detailed end-to-end broadband simulations (see §4.3.4). Transmission values are the same as §3.2. A 10 or 30 m diameter unobscured circular telescope pupil is assumed, titled in each panel of Fig. 3.45. Sky background levels for each filter are obtained from Keck/NIRC2 measurements.[6] The middle panel in Fig. 3.45 assumes a $m_H = 5$ host star; the cooling curves from Spiegel & Burrows (2012) are used to calculate the exoplanet H band magnitude, assuming a hot start (highest entropy) formation scenario and a 30 Myr age. Ups And A b has an assumed planet-to-host star astrophysical flux ratio of 10^{-8}, assuming a cloudy atmosphere without water, phase function of 0.5 at 60 degrees, and geometric albedo of 0.7 (M. Marley, private

[6] `https://www2.keck.hawaii.edu/inst/nirc2/filters.html`

communication). Proxima Cen b has an assumed planet-to-host star astrophysical flux ratio of 8×10^{-7} (Guyon, 2011). As in §3.5.2 I assume that the photon noise limit can be reached by post-processing, and as in §3.5.3 I assume that this limit can be decreased further by DM control. Thus, the assumed exposure times in Fig. 3.45 are used to determine the photon-noise limit image, calculated from the difference between a propagation with and without photon noise. The propagation only uses a single complex wavefront realization and does not include a full translating residual atmospheric phase screen as in §3.5.2 or §3.5.3; again, here I am simply illustrating the potential advantage of FAST in a simple model. The photon noise limit image is then divided by $\sqrt{50}$, assuming that DM control improves contrast by a factor of 50 at all image separations within the DH. The coronagraphic image for the on-axis propagated wavefront (with photon noise) is then normalized to the contrast of the aforementioned photon noise limit image. Contrast is computed as the standard deviation over the half DH of the DM control region. The propagated off-axis wavefront (also with photon noise), scaled to the aforementioned planet-to-star ratio, is then added to the scaled on-axis star in the coronagraphic image plane and displayed in Fig. 3.45.

Chapter 4

Future Work

In this section I will present a perspective on future work needed to better understand speckle evolution (§4.1), analysis of my debris disk targets (§4.2), and the next steps for FAST (§4.3).

4.1 Speckle Analyses

Beyond the work presented in Chapter 2, as long as ExAO systems remain limited by temporal stability (i.e., preventing subtraction to the photon noise limit using ADI and/or RDI post-processing, thus motivating the additional need for SDI-processing), chromaticity will also limit the deepest contrasts that can be reached. Although fast CDI/focal plane wavefront sensing is a promising solution to overcome both problems, further development to push the limits of classical differential imaging is also warranted. On the temporal side, it is not yet well understood how advancements in active AO control and/or passive open-loop control (e.g., steering the beam with pointing and centring mirrors to compensate for flexure errors over an observing sequence; Dunn et al. 2008) can lead to improvements in final post-processed contrasts (although some initial analyses do show correlations; Bailey et al. 2016), which I will discuss further in §4.1.1. On the chromatic side, detailed characterizations of colder/lower mass exoplanet atmospheres will continually push the ExAO community towards reaching deeper broadband contrasts, in which case chromaticity will ultimately become a fundamental limitation to this goal; thus, the future push towards reaching deeper contrasts will therefore also require a push towards understanding the limitations currently set by chromaticity and implementing more achromatic

broadband systems. In my view, both problems are ultimately characterized by the following question:

What level of WFE stability and chromaticity is needed to reach the photon noise limit after post-processing?

Although this photon noise limit will be dependent on the host star magnitude, λ, observing conditions, throughput, coronagraph design, etc., numerical simulations will be able to answer this question for typical targets/observing configurations and, as a result, will set tolerance requirements on future goals for instrument performance. These tolerance requirements are defined as the level of wavefront decorrelation over time and/or wavelength at which a corresponding residual halo is *below* the photon noise limit (i.e., for a single speckle, less than one photon at a given wavelength is recorded until the wavefront evolves into a new realization over time and/or wavelength, similar to the 5th magnitude star case illustrated in Fig. 3.11). Such simulations can then translate into tolerance requirements for, e.g., SR and/or chromaticity from instrument-related Fresnel propagation effects. Thus, for GPI, SCExAO, and other future ExAO instruments, an optimal approach moving forward is to run numerical simulations, including ADI+SDI PSF subtraction, where wavefront evolution over time and wavelength are tuneable parameters (e.g., similar to the α parameter from Srinath et al. 2015 introduced in §3.2) to be optimized/characterized according to the simulated final contrast results.

4.1.1 GPI

Beyond the general ideology discussed above, at the time of this writing (spring 2020) specific off-sky tests with the GPI instrument are also now planned in the near future:

chromaticity: (PI: C. Marois) A double notch bandpass filter is planned for installation in front of the IFS during the GPI 2.0 integration and testing at Notre Dame. This filter will be designed such that the two monochromatic PSFs can be extracted by the DRP with minimal crosstalk between slices, enabling both a more detailed characterization how the algorithmic extraction process limits contrast gains by SDI and a more robust comparison of how intrinsic wavefront chromaticity compares to Fresnel simulations, building on the work in §2.1.

stability: (I am the PI) Before GPI leaves the Gemini South telescope at the end of the 2020B semester, WFS telemetry and high frame rate IFS data will be

recorded over a simulated observing sequence (i.e., with the telescope tracking the sky rotation) using the internal light source. Acknowledging that many speckles in the raw GPI images may originate from *common*-path aberration (in contrast to the standard paradigm that most speckles are from NCPA) due to limited sub-nm level WFS sensitivity to aliasing and non-linearities, the AO RTC may not yet be optimized to stabilize common-path quasi-static aberration. Using numerical simulations in comparison with the acquired WFS telemetry and IFS images from this dataset, future RTC optimizations may enable reaching deeper final contrasts (e.g., via ADI processing) by improving the stabilization of common-path quasi-static speckles.

4.2 Debris Disk Analysis

The contrast and mass limits from my SCExAO observations (§2.2.3 and 2.2.4, respectively) could be used in future analyses of two-component debris disk targets. For example, the Meshkat et al. (2017) occurrence rate analysis of dusty systems (see §1.3.4 for a summary of this work) could be extended to my non-detections (some of which are not included in the Meshkat et al. 2017 sample), although this addition is likely to yield a negligible difference in the Meshkat et al. (2017) results, as my sample is an order of magnitude smaller than theirs. Additionally, an analysis similar to Matthews et al. (2018) could be performed, combining my mass lower limits with dynamical simulations in two-component debris disk systems, producing mass upper limits, to place constraints on any possible companions in these systems.

4.3 FAST

Although in this dissertation I have demonstrated the innovative potential of FAST to overcome the issues of wavefront stability and chromaticity, more work is still needed to deploy a FAST-enabled instrument on-sky.

4.3.1 Chromaticity

Although the FAST simulations in this dissertation have only used a monochromatic wavefront to simulate performance, broadband performance must be addressed before

a realistic instrument design and operation can be proposed. New problems to be addressed in broadband operation can be separated into two categories:

1. **wavefront chromaticity**, referring to how the wavefront, as projected in the entrance pupil, evolves with wavelength, and

2. **PSF magnification with wavelength**, referring to how the plate scale in the focal plane changes with wavelength, purely due to wavelength-dependent diffraction (i.e., independent of wavefront chromaticity).

Without addressing each of the above points, broadband SCC images will be "smeared," resulting from fringes at a given physical separation in the image differing between wavelengths across the bandpass.

Several solutions for SCC operation have been proposed to mitigate point 2 above, which could be modified/adopted to work for FAST, including:

- A Wynne corrector (Wynne, 1979), which optically magnifies wavelengths just before the detector to effectively cancel the effect of PSF magnification entirely.[1]

- The Multi-Reference SCC (MRSCC; Delorme et al. 2016), which uses multiple off-axis pinholes in the Lyot stop, forming fringes along multiple directions in the coronagraphic image to mitigate fringe smearing effects from magnification. The MRCC concept is illustrated in Fig. 4.1 along with a possible modification of the TG FPM that would be compatible with this setup to enable a broadband version of FAST.

However, even if point 2 above can be mitigated, point 1 will remain a fundamental limitation, as I have shown in Chapter 2 of this dissertation. In other words, *there is no active solution to mitigate speckle evolution with wavelength*, also illustrated in Fig. 4.2. With that said, passive solutions have been shown in simulation to set these limits many orders of magnitude below the final contrasts we are reaching now, both for AO residuals (Guyon, 2005) and quasi-static aberration (Marois et al., 2008c, §2.1). However, as we continue to improve current technology to reach closer to these limits, new solutions will eventually be necessary.

With this future context in mind, a unique solution to these problems that is worth pursuing is a FAST IFS to enable narrowband CDI post-processing of individual

[1] but note that this magnification effect will still occur at the FPM plane, still generating chromatic effects on achievable broadband contrast in the detector plane.

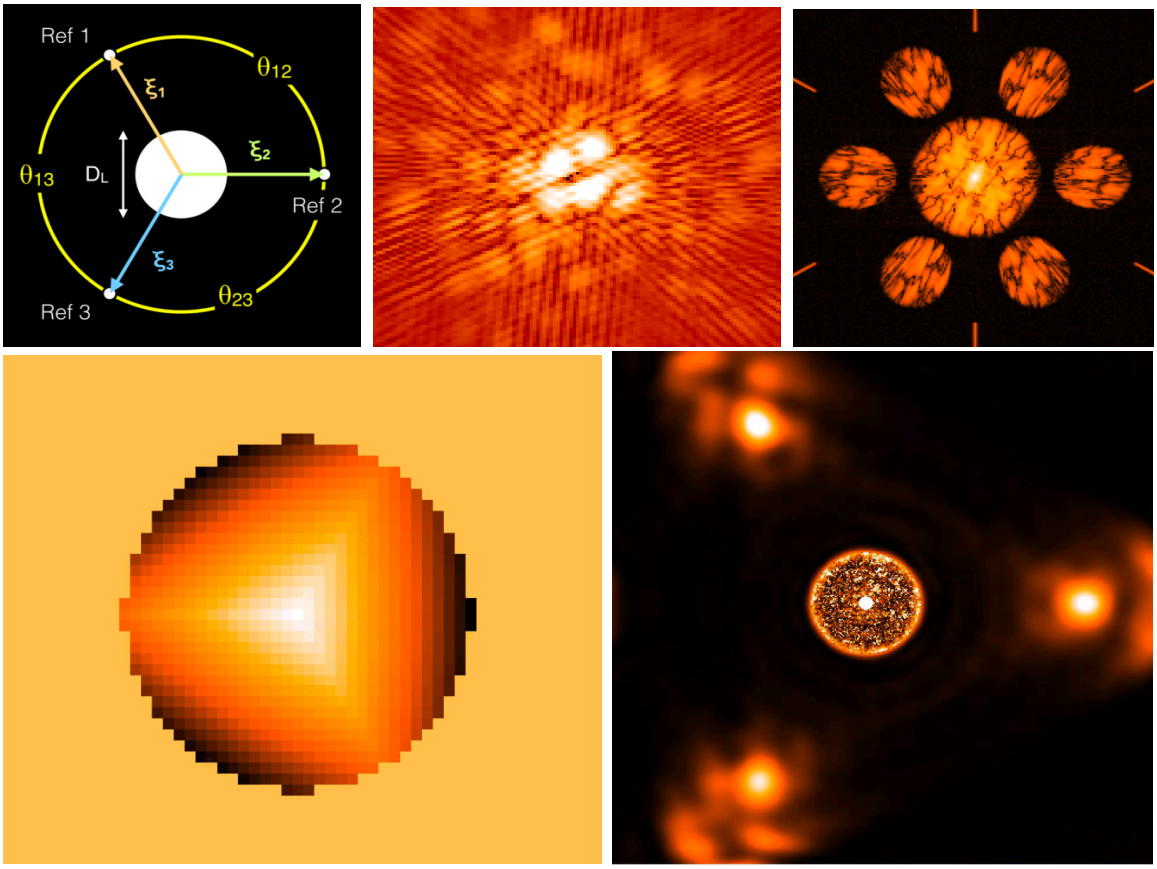


Figure 4.1 An illustration of the multi-reference SCC (MRSCC) and possible FAST compatibility. Upper panels (left to right, from Delorme et al. 2016): the Lyot stop, coronagraphic image, and MTF of a MRSCC setup. As in Fig. 3.1, the pinhole intensity is unphysically enhanced for illustration to generate the image and MTF above. Lower panel: a possible modification of my TG FPM design (left; using the same design as §3.5.1.1 other than splitting the tilt angles equally into three different directions) , to physically enhance the intensity and fringe S/N in all three off-axis pinholes off the Lyot stop (right).

wavelength slices taken during millisecond-timescale exposures. If this capability could be enabled with new technology (see below), CDI processing of each slice could be used to directly attenuate (but not completely remove)[2] the impacts of wavefront chromaticity, where the contrast gain relative to the limits from a single broadband image—shown in Fig. 4.2—can be estimated by the number of wavelength resolution elements across a given bandpass (e.g., if there are 20 wavelength slices across the

[2]Note that wavefront chromaticity would still be present in individual wavelength slices, but that stacking the CDI-subtracted slices is just a way of algorithmically attenuating this impact on the final contrast, rather than optically attenuating/cancelling this effect.

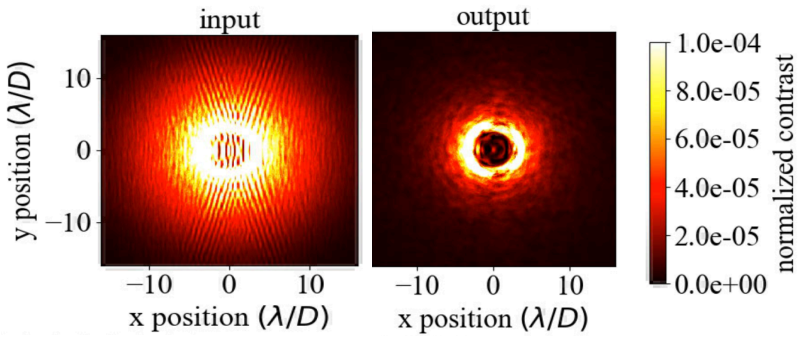


Figure 4.2 Left: a simulated broadband SCC image using the TG FPM, with $\lambda_0 = 1.65\mu$m and $\Delta\lambda/\lambda_0$=20%. Although PSF magnification with wavelength is not simulated, 100 different uncorrelated wavefront realizations have been used across the bandpass to generate the illustrated broadband image. As a result, using an assumed noiseless, simultaneous measurement of the pinhole PSF along with the isolated fringes via filtering of the MTF sidelobe (i.e., using the broadband version of equation 3.6 for this setup), the subtraction limit of the "input" image in the left panel is shown as the "output" image in the right panel. Separated from the effects of PSF magnification with wavelength, this residual "halo" in the right panel sets a fundamental limit to achievable broadband contrast with any CDI and/or active wavefront control speckle subtraction method due to wavefront chromaticity. Although realistic atmospheric and quasi-static wavefront chromaticity are much weaker than what is simulated here (Guyon, 2005; Marois et al., 2008c), the level of this residual "halo" due to wavefront chromaticity ultimately sets a fundamental limit on achievable broadband contrasts.

bandpass, compared to the limits set by wavefront chromaticity in a single broadband image, FAST CDI can in principle reach final contrasts that are 20 times deeper, assuming each slice is decorrelated from one another and individual subtractions are still limited by wavefront chromaticity).[3] This approach can be thought of as the chromatic analogy to the temporal FAST solution for subtracting AO residuals (Fig. 3.11e): just as you need to run on fast millisecond-timescales to subtract AO residuals, you need to run over many narrow resolution elements to subtract "chromaticity residuals" that will similarly average into a halo if collapsed over a full spectral bandpass.[4] A few different approaches to how this strategy could be

[3]Note that this millisecond "CDI+IFS" approach is impossible with active wavefront control using a single DM; multiple wavefront realizations from wavefront chromaticity would require multiple DMs to correct for each realization. Ignoring the limits from photon noise, ∼20 DMs would be required to reach the same gains achieved by 20 wavelength slices with an IFS+CDI approach, clearly an impractical solution.

[4]although note that for either case (temporal or chromatic), this strategy is only necessary if the residual "halo" (from either AO residuals or wavefront chromaticity, respectively) is above the

implemented are discussed below:

1. A lenslet array and/or image slicer-based IFS

   Although this is currently the most common form of IFS in high contrast imaging (see §1.4.3.3 for an overview), a millisecond-timescale FAST IFS application with this approach would be challenging. Current 4k×4k detectors, which are necessary with this approach to sample micro spectra at R∼50 over a 2×2 arcsecond FOV (e.g., Larkin et al., 2014), take at least a few seconds to read out a full frame. Accounting for the additional super-Nyquist sampling requirements of the SCC ($\gtrsim$5 lenslets/(λ/D)), the analogous setup for FAST would require a photon-counting $\gtrsim$10k×10k detector to readout full frames on millisecond-timescales, far beyond the near-term capabilities of detector technology. With that said, sacrifices could be made in FOV and/or spectral resolution to make this feasible with current detector technology, and warrants further study on feasibility.

2. A Microwave Kinetic Inductance Detector (Mazin et al., 2012)

   This detector is, in theory, able to measure both the arrival time and wavelength of a photon for individual pixels, functioning inherently as both an IFS and a photon-counting detector (i.e., zero read noise/dark current). The potential of this technology is enormous, enabling the same temporal resolution at every IFS wavelength slice as point 1 but without additional optics beyond the detector. However, in practice, real prototypes of these detectors deployed in laboratory and on-sky settings have been limited to very low spectral resolution and operation at less than about few hundred Hz, not yet reaching photon-counting capabilities. Further development of this capability (Mazin et al., 2019) will provide an encouraging option for the FAST IFS CDI strategy.

3. A tuneable narrowband filter(s)

   If FAST CDI remains near photon noise-limited with time, this approach could be enabled by (a) recording a sequence of images with a single narrowband filter where the total integration time is set to match the desired contrast, and then (b) repeating this process over N other narrowband filters, ultimately producing $N+1$ slices across the bandpass. If FAST CDI instead reaches static

photon noise limit, as illustrated in Fig. 3.11e.

limits that still require combined use with SDI, an adjustable filter can be used so that consecutive images on millisecond-timescales are recorded at different wavelengths. For example, running at 1 kHz, 20 spectral resolution elements across a given bandpass would take 20 ms to acquire (but with each frame still only able to reach photon noise-limited contrasts for a 1 ms exposure). Individual exposures would still freeze AO residuals and quasi-static errors, but with each frame at a different wavelength (i.e., enabling subsequent SDI processing). As with the former approach, a higher spectral resolution and/or a better contrast would require longer integration times. Either approach *is* feasible with current manufacturing technology, with the main disadvantage being a decreased duty cycle/observing efficiency relative to points 1 or 2.

4.3.2 Coronagraph Design

Building off of the Lyot-based high order WFS concept presented in §3.5.5.2, consider the following thought experiment:

- A single speckle from a single entrance pupil Fourier mode is detected at the same raw contrast in the SCC image between coronagraph A and B.
- However, due to differences in throughput between coronagraph A and B, for a millisecond exposure on a given star (with all other throughput parameters being equal) this speckle receives 10 photons for coronagraph A but 100 photons for coronagraph B.
- As a result, the SCC can subtract (via DM control and/or CDI) this speckle with coronagraph A down to $\sqrt{10}$ ~3 photons, while with coronagraph B the same speckle is subtracted down to $\sqrt{100} \sim 10$ photons, meaning that ultimately a deeper contrast is reached with coronagraph A even though less photons are recorded initially compared to coronagraph B.
- However, for the same comparison, the LLOWFS detector for the same Fourier mode receives 10 photons for coronagraph A but 10^4 photons for coronagraph B (i.e., the different LLOWFS and/or FPM masks between the two coronagraphs change the throughput to the LLOWFS detector). As a result, using the LLOWFS detector as a high order WFS, coronagraph A can subtract the speckle as seen on the SCC image down to $10/\sqrt{10} \sim 3$ photons, whereas coronagraph B reaches $100/\sqrt{10^4} = 1$ photon.

Thus, an ideal coronagraph design should enable a linear relationship between a given Fourier mode as seen by the LLOWFS detector and the SCC detector (building on the work from §3.5.5.2), where a maximal amount of a given mode is sent to the LLOWFS detector and *not* the SCC detector. This setup will enable both better sensing and correction for the LLOWFS and better photon noise-limited contrast for the SCC. In §3.5.6.2, the ZWFS (using a 30/70 beam splitter upstream of the FPM) is instead a temporary solution to this same optimization problem; although further work could be done on optimizing the ZWFS (and/or similar Fourier-based pupil plane WFSs; Fauvarque et al. 2017) sensitivity to photon noise, this approach will inherently decrease exoplanet throughput to the SCC detector (i.e., for the 30/70 beam splitter, exoplanet throughput is also decreased by 30%), whereas the Lyot-based high order WFS approach still has the potential for 100% exoplanet throughput. Thus, even though it is tempting to optimize a coronagraph design to increase the throughput to the SCC for both high- and low-order modes, thereby enabling a better WFS sensitivity to photon noise, this will only inherently decrease the achievable photon-noise limited contrast and should therefore be avoided.[5] Accordingly, along with Lyot-based high order WFS linearity (which can be independently optimized via custom LLOWFS mask designs), future FAST coronagraph designs (i.e., custom entrance apodizer and/or FPM masks) should prioritize maximizing exoplanet relative to stellar throughput in the coronagraphic image.

Next, as discussed in §3.5.4.2 and §3.5.6, future FAST coronagraph designs should still be optimized to improve WFS sensitivity, non-linearities, and linear range, albeit with the primary constraints discussed above. Reaching fringe ratios closer to $I_R/I_S = 1$ (without increasing coronagraph throughput) will improve WFS sensitivity to photon noise and thus enable reaching both diffraction-limited contrasts via DM control in a fewer number of iterations a deeper photon noise-limited contrast via CDI. Better diffraction attenuation will improve WFS non-linearities beyond the limits in §3.5.6.2 and enable deeper achievable diffraction-limited contrasts via DM control (§3.5.4.3). It may also be possible to increase the FAST linear range (§3.5.6.1) by coronagraph design optimization, governed by the aberration level at which second (and higher) order sine spots impact the linear assumption between the SCC detec-

[5]But with that said, it is perhaps also worth considering wavefront sensing architectures where the SCC is used purely for a second stage correction in the WFS path (i.e., a separate coronagraph and detector is used for the science path). E.g., if coronagraph design can instead improve throughput, and accordingly improve photon noise sensitivity, relative to the ZWFS, this may warrant using FAST in future systems purely as a WFS.

tor and DM planes. Lastly, anti-aliasing methods, such as a spatial filter (Poyneer & Macintosh, 2004) and/or design methods inherent to the coronagraph optimization procedure, could be considered in the future if the former photon noise and non-linearity components are optimized to reach levels comparable to the alaising limits in Fig. 3.38.

4.3.3 NEW EARTH

In the near future, I will focus on end-to-end laboratory demonstrations of a narrow-band FAST module to be deployed on-sky, including possible applications at Gemini/GPI, Keck, and/or Subaru/SCExAO. Although a more detailed broadband setup may be implemented in the more distant future, a possible compatible solution for broadband operation in this "narrow band mode" was discussed in §4.3.1, utilizing a tuneable narrowband filter (which is also still compatible with the optical chopper strategy presented in §3.5.7.4). This simplified operational approach of performing narrowband speckle subtraction will enable a low-risk option to deploy the first FAST-enabled instrument on-sky, with further modifications to enable broadband operation still possible subsequently.

4.3.4 End-to-End Simulations and Science Cases

Complementary to laboratory demonstrations, performance predictions and motivating science cases of a FAST-enabled instrument will also require detailed, high-fidelity simulations. By predicting achievable contrast in different realistic observing conditions and as a function of stellar magnitude, many science cases can be updated, including characterizations of individual targets (§1.3.5) and survey design for occurrence rate analyses (§1.3.3), both of which will help inform our understanding of giant exoplanet formation (§1.3.1), atmospheric evolution (§1.3.2), and the connection between exoplanets and debris disks (§1.3.4).[6] Relatedly, although the occurrence rate analysis results presented in §3.5.8 (Fig. 3.46) suggest that a FAST instrument upgrade using current 10 m-class facilities could enable a vast range of new parameter space to search for lower-mass, closer-in, young exoplanets around nearby stars, it remains crucial to update these instrument performance simulations (which are an

[6]Note that FAST CDI processing is in principle insensitive to disk inclination, whereas current ADI+SDI approaches are significantly less sensitive to face-on disks (Esposito et al., submitted, AJ), potentially enabling new unbiased disk surveys and science cases to be developed.

input into the Monte Carlo simulations used to produce conclusions as in Fig. 3.46). More detailed end-to-end simulations (e.g., including realistic atmospheric residuals output from AO simulator codes, simulating realistic chromaticity, vibration, flexure, and other quasi-static effects, understanding how these effects vary in different observing conditions, etc.) will update these results in order to help optimize the instrument's potential scientific impact before going on-sky, informing, e.g., future survey design and preparation of which individual stars to target. Alongside the FAST laboratory testing and instrument development efforts discussed in §4.3.3, it will remain essential to continue further efforts in developing these simulations and corresponding preparations of optimal science cases.

Chapter 5

Conclusions

In this dissertation I have provided a detailed view into both the current limitations of exoplanet imaging instruments in reaching deeper contrasts (Chapter 2) and a path forward to remove these limitations (Chapter 3). In this chapter I will summarize these results.

5.1 Current Limitations: Speckle Evolution

In Chapter 2 I illustrated the impacts of two main limitations currently faced by state-of-the-art exoplanet imaging instruments: speckle evolution over time (stability) and wavelength (chromaticity).

5.1.1 GPI

In §2.1 I presented an initial chromaticity analysis using off-sky GPI data acquired with the internal light source. My main findings from this analysis are that measured contrast gains via a simple single difference between two wavelengths are more than an order of magnitude worse than expected from end-to-end Fresnel simulations. These results were validated across multiple apodizer and dithering configurations, reinforcing the robustness of my measurements. However, additional tests and analyses are needed before mitigation strategies can be implemented (see §4.1.1). Note that these additional tests require custom optics installed in GPI that cannot be implemented easily at the Gemini South telescope, which is why further work on this topic was not completed during my PhD studies.

5.1.2 SCExAO

In §2.2, I presented a detailed correlation analysis, data processing architecture, final contrast curves, and corresponding exoplanet mass limits using the SCExAO/CHARIS ExAO instrument. My main findings are as follows:

1. A correlation analysis of the SCExAO/CHARIS system across all seven of my targets showed that for small ($\lesssim$300 mas) and large ($\gtrsim$650 mas) separations from the central star, chromaticity across the full JHK bandpass is worse than temporal stability across the full observing sequence (Fig. 2.10). This result suggests that at these separations, the large wavelength "lever arm" of the CHARIS broadband datasets could not be used advantageously to improve SDI post-processing (e.g., slices in J band can generally not be used to subtract slices in K band, and vice versa).

2. With that said, I found that for the same target my T-type PSF subtraction generally reaches up to $\sim$50% deeper final contrasts than my L-type subtraction (Fig. 2.12), illustrating that more correlated wavelength slices close to the target image wavelength are being used to improve contrast. This is consistent with the findings of point 1 above (i.e., Fig. 2.10); together, these results illustrate that improved contrast from SDI processing of my targets comes from selecting reference slices closer in wavelength to the target image where chromaticity is less significant.

3. I presented a modified framework for exoplanet algorithmic throughput correction using forward modelling that is designed for the CHARIS broadband mode (§2.2.2.2). As a result, simulated planets are recovered in good agreement with their input values even when close to the 5σ noise floor (Fig. 2.7).

4. PSF-subtracted sequences using only the H-band slices within the broadband dataset generally reach final contrasts that are similar to the final contrasts of the same slices in the broadband sequences (Fig. 2.13), consistent with the hypothesis made in points 1 and 2 above.

5. Using my contrast curves and assuming a hot start formation scenario, I presented lower limits on exoplanet mass for of all my observations (§2.2.4), reaching down to 7 $M_{\rm Jup}$ at 62 au.

5.2 A Path Forward: Fast CDI/Focal Plane Wavefront Sensing

In Chapter 3, I presented a solution to the speckle stability and chromaticity limitations analyzed in Chapter 2, called FAST.

5.2.1 Initial Work

First, in §3.3, I presented an analysis of SCC tolerances to drift and vibration. I showed that:

1. Working at 1.6 μm, the SCC is insensitive to vibration amplitudes as large as about 5 μm (Fig. 3.5 a).

   This robustness to vibrations at amplitudes much larger than the operational wavelength (which currently renders the GPI Mach-Zehnder interferometer/high order WFS in-operational) can be attributed to the common-path interferometer design of the SCC, fixing the Lyot stop pinhole-to-pupil separation independent of vibration amplitude, and therefore presenting a promising solution for future instruments operating in a similar vibration environment.

2. Drifts of the FPM and Lyot stop, using the GPI APLC design, are insensitive to performance degradation up to about 0.1 λ/D and 2% of the pupil diameter, respectively (Fig. 3.5 b).

   While the former is within the achievable tolerances for existing LOWFS technologies, the latter required new developments. I outlined my new active SCC correction method in Fig. 3.6, enabling an on-sky measurement and DM correction procedure to compensate for Lyot stop drift.

Then, in §3.4, I investigated initial methods to increase the SCC fringe S/N, prior to my adopted FAST solution below. I showed that many different approaches improved over the classical SCC design, all of which were possible to combine with different existing coronagraph designs. Although these approaches did not ultimately provide a sufficient fringe ratio boost in comparison to my custom FAST solution (Tab. 3.1), testing these ideas was essential in influencing and developing the subsequent framework for my TG FPM design.

5.2.2 TG FPM and CDI

In sections 3.5.1 and 3.5.2, I outlined my initial FAST solution and a post-processing CDI speckle subtraction strategy, summarized below:

1. I designed a new FPM, called the TG FPM, specialized for my FAST approach that allows fringes to be detected in exposures that are of order a few milliseconds and boosts fringe ratio by a factor of $\sim 10^6$ relative to a standard SCC design (§3.5.1.1)

2. I illustrated that the standard approach to ground-based CDI/focal plane wavefront control (i.e., taking a long exposure to average out atmospheric speckles and measure only the remaining static speckles) is limited in achievable contrast by residual atmospheric speckles that are above the photon noise limit but not measureable/correctable during long exposures (§3.5.1.3).

3. Developing on the framework of Baudoz et al. (2012a), I showed that a post-processing algorithm can be used to reconstruct and subtract both atmospheric and static components of an on-sky SCC image without subtracting an incoherent, off-axis exoplanet, but that this process benefits from additional knowledge of the SCC pinhole PSF (§3.5.2.1)

4. I tested three different algorithms to estimate the missing pinhole PSF information:

   (a) a direct, "live" simultaneous measurement of the pinhole PSF along with the SCC image (§3.5.2.1.1),

   (b) use of a daytime pinhole PSF calibration image (§3.5.2.1.2), and

   (c) a pinhole PSF reconstruction algorithm that uses only the recorded SCC image and a series of filtering algorithms (§3.5.2.1.3).

5. In §3.5.2.3, with the new TG FPM design from §3.5.1.1 and post-processing algorithms from §3.5.2.1, I illustrated that obtaining fast exposures on the order of a few milliseconds to freeze the atmosphere is a solution to the long exposure limitations from §3.5.1.3. With this approach, I found that the direct pinhole PSF measurement algorithm enabled reaching near-photon noise-limited contrasts all the way out to the end of my 30 second simulations for both 0th and 5th magnitude stars. This result suggests high contrast imaging instruments

using this technique would no longer be limited by static or atmospheric aberration, and that "raw" contrast would improve simply by integrating longer. If implemented, this would be a radical improvement compared to current instrument capabilities.

5.2.3 Wavefront Control

In §3.5.3 I presented a FAST speckle subtraction strategy using active wavefront control with a DM, complementary to the CDI post-processing approach in §3.5.2. My findings were:

1. In §3.5.3.1 I presented several new modifications to the classical SCC DM calibration procedure from Baudoz et al. (2012a) and subsequent papers, illustrating that the much higher pinhole PSF intensity enabled by the TG FPM is limiting achievable contrasts after generating a DH, and that the classical GPI APLC coronagraph still reaches a deeper unfringed DH contrast than with the same IWA TG mask. The latter can be attributed to use of an apodizer mask that is not optimized for the TG FPM, motivating future coronagraph design optimization for this case (§4.3.2).

2. In §3.5.3.2.1 I provided a framework to optimize the performance of FAST closed-loop wavefront control, using the modal gain optimization procedure developed by Gendron & Lena (1994).

3. In §3.5.3.2.2 I showed that this modal gain optimization framework improves speckle subtraction relative to a constant gain integral controller and that FAST DM control generally lowers the photon noise limit relative to the CDI-only limits in §3.5.2.

5.2.4 TGV Coronagraph

In §3.5.4 I introduced the TGV FPM. In doing so, I also introduced a new methodology towards coronagraph design, where contrast and fringe S/N are considered simultaneously. I showed that the TGV mask has a number of advantages over the TG mask previously proposed in §3.5.1.1, including

1. a better balance of diffraction-limited contrast, fringe S/N, and photon noise-limited contrast (§3.5.4.2), and

2. almost 6 magnitudes deeper contrasts obtained at 2 - 5 λ/D by DM control of quasi-static aberration (§3.5.4.3).

Lastly, I discussed how future coronagraph designs, influenced by my initial approach in §3.5.4, can be optimized as both a WFS and high contrast imager (§3.5.4.4, §4.3.2).

5.2.5 Low Order Wavefront Control

In §3.5.5 I presented the concept of a LLOWFS in application to the FAST TGV phase mask. In comparing the response to low order aberrations between the LLOWFS and the SCC, I found that:

1. Using the advantage of defocus optimization (Fig. 3.29), the LLOWFS linearity of focus is increased (about ± 1 μm PV; Fig. 3.30 a) relative to the SCC (about ± 500 nm PV; Fig. 3.30 b), but otherwise with negligible differences in tip/tilt linearity (both about about ± 500 nm PV).

2. The LLOWFS sensitivity to photon noise is better than the SCC (Fig. 3.31), consistent with the majority of light from an entrance pupil low order aberration being sent to the LLOWFS detector rather than the SCC (Fig. 3.26).

In §3.5.5.2 I found that high order wavefront sensing with the LLOWFS, using an annular mask around the edge of the TGV pupil, was effective in the common path of the LLOWFS detector, but was non-linearly related to the SCC detector plane; therefore the LLOWFS was not effective at subtracting speckles that impact the SCC. Lastly, in §3.5.5.3 I presented a new CDI strategy, utilizing the on-sky LOWFS telemetry to reconstruct the on-sky pinhole PSF, although preliminary results suggest that static limitations still prevent this reconstruction approach from continually reaching deeper contrasts.

5.2.6 High Order WFS Linearity and Sensitivity

In §3.5.6 I presented a comparison of WFS linearity and sensitivity between the TGV WFS and the ZWFS, showing that

1. the TGV WFS has a slightly larger linear range than the ZWFS (§3.5.6.1), and

2. the ZWFS remains significantly more sensitive than the TGV WFS (§3.5.6.2).

With that said, future FAST coronagraph design optimization (§4.3.2) may be able to improve both linearity and sensitivity beyond the limits presented in §3.5.6.

5.2.7 Laboratory Tests

In §3.5.7 I presented the first laboratory tests of FAST. In the LESIA laboratories I tested the first FAST focal plane mask designed to boost the SCC fringe visibility to be used in millisecond exposures (§3.5.7.2). Along with additional tests at the ETH Züruch laboratories (§3.5.7.3), the measured fringe visibility boost for my new FAST FPM designs are consistent with simulations. I also presented some initial results from the NEW EARTH laboratory, showing that my reflective TGV FPM appears to have been fabricated at sufficient tolerances and that a fast optical chopper shows promising potential to enable FAST CDI for individual wavefront realizations. Now that I have validated the basic concept of the FAST coronagraphic mask, subsequent NEW EARTH laboratory demonstrations will focus on generating deeper contrast gains and end-to-end instrument demonstrations (§4.3.3).

5.2.8 Science Cases

In §3.5.8 I presented preliminary optimistic simulations illustrating the potential science cases that FAST could enable on both current and future telescopes, including:

1. visible reflected light imaging of known RV giant exoplanets on current 10 m-class telescopes,

2. NIR imaging of thermal emission from Jupiter analogues on current 10 m-class telescopes, and

3. NIR reflected light imaging of habitable exoplanets around M dwarf stars on future 30 m-class ELTs.

Many additional cases have yet to be developed, including applications to direct imaging of circumstellar disks (both for total intensity and polarized emission), spectroscopic (and/or polarimetric) characterization of existing and new substellar objects, occurrence rate analyses, and more. In the future, these science cases will also be integrally connected with high fidelity end-to-end simulations, as discussed in §4.3.4, providing realistic performance estimates to support survey design and preparation for individual target analyses.

5.3 Future Work

In Chapter 4, I presented a detailed perspective on the future outlook of work presented in this dissertation, including the next steps for speckle stability and chromaticity analyses (§4.1), related future laboratory tests with GPI (§4.1.1), and further occurrence rate and/or dynamical analyses my SCExAO debris disk observations (§4.2). I outlined a comprehensive strategy for future developments of FAST (§4.3), including addressing limitations set by chromaticity (§4.3.1), motivating and outlining the goals for future coronagraph design optimization procedures (§4.3.2), laboratory testing to develop the first on-sky FAST-enabled instrument (§4.3.3), and using high fidelity end-to-end simulations to inform and update FAST science cases (§4.3.4). Despite the plethora of developments in speckle evolution and subtraction to pursue in the future, these endeavours remain promising; with no major physical limitations to FAST at this time, the future appears bright for exoplanet imaging.

Bibliography

Allard, F. 2014, in IAU Symposium, Vol. 299, Exploring the Formation and Evolution of Planetary Systems, ed. M. Booth, B. C. Matthews, & J. R. Graham, 271–272

Alloin, D. M., & Mariotti, J.-M. 1994, C: Mathematical and Physical Sciences, Vol. 423, Adaptive Optics for Astronomy (Kluwer Academic Publishers)

ALMA Partnership, Brogan, C. L., Pérez, L. M., et al. 2015, The Astrophysical Journal Letters, 808, L3

Anglada-Escudé, G., Amado, P. J., Barnes, J., et al. 2016, Nature, 536, 437

Apai, D., Karalidi, T., Marley, M. S., et al. 2017, Science, 357, 683

Bacon, R., Adam, G., Baranne, A., et al. 1995, A&AS, 113, 347

Bailey, V. P., Poyneer, L. A., Macintosh, B. A., et al. 2016, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9909, Status and performance of the Gemini Planet Imager adaptive optics system, 99090V

Barman, T. S., Konopacky, Q. M., Macintosh, B., & Marois, C. 2015, ApJ, 804, 61

Basri, G., & Brown, M. E. 2006, Annual Review of Earth and Planetary Sciences, 34, 193

Baudoz, P., Boccaletti, A., Baudrand, J., & Rouan, D. 2006, in IAU Colloq. 200: Direct Imaging of Exoplanets: Science & Techniques, ed. C. Aime & F. Vakili, 553–558

Baudoz, P., Galicher, R., Patru, F., Dupuis, O., & Thijs, S. 2018, arXiv e-prints, arXiv:1801.06600

Baudoz, P., Mazoyer, J., Mas, M., Galicher, R., & Rousset, G. 2012a, in Proc. SPIE, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV, 84468C

Baudoz, P., Mazoyer, J., Mas, M., Galicher, R., & Rousset, G. 2012b, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8446, Dark hole and planet detection: laboratory results using the self-coherent camera, 84468C

Bessell, M. S., Castelli, F., & Plez, B. 1998, A&A, 333, 231

Beuzit, J. L., Vigan, A., Mouillet, D., et al. 2019, A&A, 631, A155

Bohn, A. J., Kenworthy, M. A., Ginski, C., et al. 2019, arXiv e-prints, arXiv:1912.04284

Bonnefoy, M., Boccaletti, A., Lagrange, A. M., et al. 2013, A&A, 555, A107

Bordé, P. J., & Traub, W. A. 2006, ApJ, 638, 488

Boss, A. P. 2000, The Astrophysical Journal Letters, 536, L101

Bottom, M., Wallace, J. K., Bartos, R. D., Shelton, J. C., & Serabyn, E. 2017, MNRAS, 464, 2937

Bowler, B. P. 2016, ArXiv e-prints, arXiv:1605.02731

Bozza, V., Mancini, L., & Sozzetti, A., eds. 2016, Astrophysics and Space Science Library, Vol. 428, Methods of Detecting Exoplanets

Brandt, T. D., Rizzo, M., Groff, T., et al. 2017, Journal of Astronomical Telescopes, Instruments, and Systems, 3, 048002

Burgasser, A. J., Liu, M. C., Ireland, M. J., Cruz, K. L., & Dupuy, T. J. 2008, ApJ, 681, 579

Burgasser, A. J., McElwain, M. W., Kirkpatrick, J. D., et al. 2004, AJ, 127, 2856

Burrows, A., Hubbard, W. B., Lunine, J. I., & Liebert, J. 2001, Reviews of Modern Physics, 73, 719

Butler, R. P., Marcy, G. W., Williams, E., Hauser, H., & Shirts, P. 1997, ApJ, 474, L115

Campbell, B., Walker, G. A. H., & Amor, J. 1984, JRASC, 78, 206

Castro-Almazán, J., Lorenzo, B., & Muñoz-Tuñon, C. 2017

Chauvin, G., Desidera, S., Lagrange, A.-M., et al. 2017, A&A, 605, L9

Chilcote, J. K., Bailey, V. P., De Rosa, R., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Proc. SPIE, 1070244

Currie, T., Debes, J., Rodigas, T. J., et al. 2012, ApJ, 760, L32

Currie, T., Kasdin, N. J., Groff, T. D., et al. 2018a, Publications of the Astronomical Society of the Pacific, 130, 044505

Currie, T., Brandt, T. D., Uyama, T., et al. 2018b, AJ, 156, 291

Currie, T., Marois, C., Cieza, L., et al. 2019, ApJ, 877, L3

Dali Ali, W., Ziad, A., Berdja, A., et al. 2010, A&A, 524, A73

Delorme, J. R., Galicher, R., Baudoz, P., et al. 2016, A&A, 588, A136

Dessenne, C., Madec, P. Y., & Rousset, G. 1999, Optics Letters, 24, 339

Doelman, D. S., Snik, F., Warriner, N. Z., & Escuti, M. J. 2017, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10400, Proc. SPIE, 104000U

Draper, Z. H., Marois, C., Wolff, S., et al. 2014, in Proc. SPIE, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, 91474Z

Draper, Z. H., Duchêne, G., Millar-Blanchaer, M. A., et al. 2016, ArXiv e-prints, arXiv:1605.02771

Dunn, J., Wooff, R., Smith, M., et al. 2008, in Proc. SPIE, Vol. 7019, Advanced Software and Control for Astronomy II, 701910

Dupuy, T. J., & Kraus, A. L. 2013, Science, 341, 1492

Fabrycky, D. C., & Murray-Clay, R. A. 2010, ApJ, 710, 1408

Fauvarque, O., Neichel, B., Fusco, T., Sauvage, J.-F., & Girault, O. 2017, Journal of Astronomical Telescopes, Instruments, and Systems, 3, 019001

Filippazzo, J. C., Rice, E. L., Faherty, J., et al. 2015, ApJ, 810, 158

Frazin, R. A. 2016, Journal of the Optical Society of America A, 33, 712

Galicher, R., Baudoz, P., Rousset, G., Totems, J., & Mas, M. 2010, A&A, 509, A31

García-Lorenzo, B., & Fuensalida, J. J. 2011, Monthly Notices of the Royal Astronomical Society, 410, 934

Gendron, E., & Lena, P. 1994, A&A, 291, 337

Gerard, B. 2017a, S17A-TE261: Search for exoplanets sculpting debris disks with SCExAO+CHARIS, Gemini Proposal: Subaru Exchange Time, ,

—. 2017b, Astronomy Directed Studies in Adaptive Optics Control, Spring 2017, University of Victoria/National Research Council of Canada, Astronomy and Astrophysics

Gerard, B. L., & Marois, C. 2016a, in Proc. SPIE, Vol. 9909, Adaptive Optics Systems V, 990950

Gerard, B. L., & Marois, C. 2016b, in Proc. SPIE, Vol. 9909, Adaptive Optics Systems V, 990958

Gerard, B. L., & Marois, C. 2020, PASP, 132, 064401

Gerard, B. L., Marois, C., & Galicher, R. 2018a, AJ, 156, 106

Gerard, B. L., Marois, C., Galicher, R., et al. 2019a, in Proc. AO4ELT6, arXiv:1910.04554

Gerard, B. L., Marois, C., Galicher, R., & Véran, J.-P. 2018b, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10703, Proc. SPIE, 1070351

Gerard, B. L., Marois, C., Currie, T., et al. 2019b, AJ, 158, 36

Give'On, A., Belikov, R., Shaklan, S., & Kasdin, J. 2007, Optics Express, 15, 12338

Give'on, A., Shaklan, S., Kern, B., et al. 2012, in Proc. SPIE, Vol. 8442, Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave, 84420B

Goebel, S. B., Hall, D. N. B., Guyon, O., Warmbier, E., & Jacobson, S. M. 2018, Journal of Astronomical Telescopes, Instruments, and Systems, 4, 026001

Götberg, Y., Davies, M. B., Mustill, A. J., Johansen, A., & Church, R. P. 2016, A&A, 592, A147

Gravity Collaboration, Lacour, S., Nowak, M., et al. 2019, A&A, 623, L11

Gravity Collaboration, Nowak, M., Lacour, S., et al. 2020, A&A, 633, A110

Groff, T., Chilcote, J., Brandt, T., et al. 2017, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10400, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1040016

Groff, T. D., Kasdin, N. J., Limbach, M. A., et al. 2015, in Proc. SPIE, Vol. 9605, Techniques and Instrumentation for Detection of Exoplanets VII, 96051C

Groff, T. D., Chilcote, J., Kasdin, N. J., et al. 2016, in Proc. SPIE, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 99080O

Gruneisen, M. T., Sickmiller, B. A., Flanagan, M. B., et al. 2016, Optical Engineering, 55, 026104

Guesalaga, A., Neichel, B., Correia, C., et al. 2016, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9909, Online estimation of atmospheric turbulence parameters and outer-scale profiling, 99093C

Guyon, O. 2004, ApJ, 615, 562

—. 2005, ApJ, 629, 592

—. 2007, Comptes Rendus Physique, 8, 323

Guyon, O. 2011, in Second International Conference on Adaptive Optics for Extremely Large Telescopes. Online at `http://ao4elt2.lesia.obspm.fr/spip.php%3Farticle511.html`, 11

—. 2018, ARA&A, 56, 315

Guyon, O., & Males, J. 2017, ArXiv e-prints, arXiv:1707.00570

Guyon, O., Matsuo, T., & Angel, R. 2009, ApJ, 693, 75

Haffert, S. Y., Bohn, A. J., de Boer, J., et al. 2019, Nature Astronomy, 3, 749

Hagan, J. B., Choquet, É., Soummer, R., & Vigan, A. 2018, AJ, 155, 179

Hemani Kaushal, V. K. Jain, S. K. 2017, in Free Space Optical Communication (Springer India)

Hinkley, S., Oppenheimer, B. R., Soummer, R., et al. 2007, ApJ, 654, 633

Hinkley, S., Oppenheimer, B. R., Zimmerman, N., et al. 2011, PASP, 123, 74

Hodapp, K. 2018, UH-07B, UH Proposal: Subaru Time, ,

IAU. 2006, https://www.iau.org/static/resolutions/Resolution_GA26-5-6.pdf

Ida, S., & Lin, D. N. C. 2004, ApJ, 604, 388

Janin-Potiron, P., Chambouleyron, V., Schatz, L., et al. 2019, Journal of Astronomical Telescopes, Instruments, and Systems, 5, 039001

Jensen-Clem, R., Millar-Blanchaer, M., Mawet, D., et al. 2018, in American Astronomical Society Meeting Abstracts, Vol. 231, American Astronomical Society Meeting Abstracts #231, # 211.05

Johansen, A., & Lambrechts, M. 2017, Annual Review of Earth and Planetary Sciences, 45, 359

Johnson-Groh, M. 2016, Master's thesis, 111 pages; Canada: University of Victoria (Canada)

Jovanovic, N., Guyon, O., Martinache, F., et al. 2015a, ApJ, 813, L24

Jovanovic, N., Martinache, F., Guyon, O., et al. 2015b, PASP, 127, 890

Jovanovic, N., Absil, O., Baudoz, P., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10703, Adaptive Optics Systems VI, 107031U

Jovanovic, N., Delorme, J. R., Bond, C. Z., et al. 2019, arXiv e-prints, arXiv:1909.04541

Kane, S. R., Hill, M. L., Kasting, J. F., et al. 2016, ApJ, 830, 1

Kaushal, H., Jain, V., & Kar, S. 2017, Free Space Optical Communication, 1st edn. (Springer Publishing Company, Incorporated)

Kennedy, G. M., & Wyatt, M. C. 2014, Monthly Notices of the Royal Astronomical Society, 444, 3164

Keppler, M., Benisty, M., Müller, A., et al. 2018, A&A, 617, A44

Kim, I. I., Hakakha, H., Adhikari, P., Korevaar, E. J., & Majumdar, A. K. 1997, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 2990, Scintillation reduction using multiple transmitters, ed. G. S. Mecherle, 102–113

Kirkpatrick, J. D. 2005, ARA&A, 43, 195

Konopacky, Q. M., Barman, T. S., Macintosh, B. A., & Marois, C. 2013, Science, 339, 1398

Konopacky, Q. M., Marois, C., Macintosh, B. A., et al. 2016, AJ, 152, 28

Kopparapu, R. K., Ramirez, R. M., SchottelKotte, J., et al. 2014, ApJ, 787, L29

Krist, J. E. 2007, PROPER: an optical propagation library for IDL, , , doi:10.1117/12.731179

Kühn, J., Patapis, P., Lu, X., & Arikan, M. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10706, Proc. SPIE, 107062N

L Edwards, B., Israel, D., Wilson, K., Moores, J., & Fletcher, A. 2012, SpaceOps 2012 Conference, doi:10.2514/6.2012-1261897

Lacour, S. 2019, in The Very Large Telescope in 2030, 31

Lafrenière, D., Marois, C., Doyon, R., Nadeau, D., & Artigau, É. 2007, The Astrophysical Journal, 660, 770

Lagrange, A.-M., Gratadour, D., Chauvin, G., et al. 2009, Astronomy and Astrophysics, 493, L21

Lagrange, A.-M., Bonnefoy, M., Chauvin, G., et al. 2010, Science, 329, 57

Larkin, J., Barczys, M., Krabbe, A., et al. 2006, in Proc. SPIE, Vol. 6269, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 62691A

Larkin, J. E., Chilcote, J. K., Aliado, T., et al. 2014, in Proc. SPIE, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, 91471K

Lavigne, J.-F., Véran, J.-P., & Poyneer, L. A. 2007, in Adaptive Optics: Analysis and Methods/Computational Optical Sensing and Imaging/Information Photonics/Signal Recovery and Synthesis Topical Meetings on CD-ROM (Optical Society of America), AWB5

Le Bouquin, J.-B., Berger, J.-P., Beuzit, J.-L., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10703, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1070371

Lecavelier Des Etangs, A., Deleuil, M., Vidal-Madjar, A., et al. 1995, Astronomy and Astrophysics, 299, 557

Lorenzo, B., & J. Fuensalida, J. 2011, Monthly Notices of the Royal Astronomical Society, 416, doi:10.1111/j.1365-2966.2011.19186.x

Lous, M. M., Weenk, E., Kenworthy, M. A., Zwintz, K., & Kuschnig, R. 2018, A&A, 615, A145

Lyot, B. 1939, MNRAS, 99, 580

Macintosh, B., Poyneer, L., Sivaramakrishnan, A., & Marois, C. 2005, in Proc. SPIE, Vol. 5903, Astronomical Adaptive Optics Systems and Applications II, ed. R. K. Tyson & M. Lloyd-Hart, 170–177

Macintosh, B., Graham, J. R., Ingraham, P., et al. 2014, Proceedings of the National Academy of Science, 111, 12661

Macintosh, B., Graham, J. R., Barman, T., et al. 2015, Science, 350, 64

Madurowicz, A., Macintosh, B., Bailey, V. P., et al. 2019, arXiv e-prints, arXiv:1909.12981

Mahajan, V. N. 1981, Journal of the Optical Society of America (1917-1983), 71, 75

Mahajan, V. N. 1983, J. Opt. Soc. Am., 73, 860

Maire, J., Ziad, A., Borgnino, J., & Martin, F. 2007, Monthly Notices of the Royal Astronomical Society, 377, 1236

Marley, M. S., Fortney, J. J., Hubickyj, O., Bodenheimer, P., & Lissauer, J. J. 2007, The Astrophysical Journal, 655, 541

Marley, M. S., & Leggett, S. K. 2009a, Astrophysics and Space Science Proceedings, 10, 101

—. 2009b, Astrophysics and Space Science Proceedings, 10, 101

Marley, M. S., & Robinson, T. D. 2015, Anual Review of Astronomy And Astrophysics, 53, 279

Marois, C. 2004, PhD thesis, Ph.D dissertation, 2004. 275 pages; Canada: Universite de Montreal (Canada); 2004. Publication Number: AAT NR00082. DAI-B 66/02, p. 945, Aug 2005

Marois, C., Correia, C., Galicher, R., et al. 2014, in Proceedings of the SPIE, Vol. 9148, Adaptive Optics Systems IV, 91480U

Marois, C., Doyon, R., Nadeau, D., Racine, R., & Walker, G. A. H. 2003, in EAS Publications Series, Vol. 8, EAS Publications Series, ed. C. Aime & R. Soummer, 233–243

Marois, C., Doyon, R., Racine, R., & Nadeau, D. 2000, PASP, 112, 91

Marois, C., Lafrenière, D., Doyon, R., Macintosh, B., & Nadeau, D. 2006a, The Astrophysical Journal, 641, 556

Marois, C., Lafrenière, D., Macintosh, B., & Doyon, R. 2006b, The Astrophysical Journal, 647, 612

—. 2008a, The Astrophysical Journal, 673, 647

Marois, C., Macintosh, B., Barman, T., et al. 2008b, Science, 322, 1348

Marois, C., Macintosh, B., Soummer, R., Poyneer, L., & Bauman, B. 2008c, in Proc. SPIE, Vol. 7015, Adaptive Optics Systems, 70151T

Marois, C., Macintosh, B., & Véran, J.-P. 2010a, in Proceedings of the SPIE, Vol. 7736, Adaptive Optics Systems II, 77361J

Marois, C., Zuckerman, B., Konopacky, Q. M., Macintosh, B., & Barman, T. 2010b, Nature, 468, 1080

Martinache, F., Guyon, O., Jovanovic, N., et al. 2014, PASP, 126, 565

Martinez, P. 2019, A&A, 629, L10

Matthews, B., Kennedy, G., Sibthorpe, B., et al. 2014, The Astrophysical Journal, 780, 97

Matthews, E., Hinkley, S., Vigan, A., et al. 2018, MNRAS, 480, 2757

Mawet, D., Riaud, P., Absil, O., & Surdej, J. 2005, ApJ, 633, 1191

Mawet, D., Pueyo, L., Lawson, P., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8442, Proc. SPIE, 844204

Mawet, D., Milli, J., Wahhaj, Z., et al. 2014, ApJ, 792, 97

Mawet, D., Ruane, G., Xuan, W., et al. 2017, ApJ, 838, 92

Mayor, M., & Queloz, D. 1995, Nature, 378, 355

Mazin, B., Bailey, J., Bartlett, J., et al. 2019, in BAAS, Vol. 51, 17

Mazin, B. A., Bumble, B., Meeker, S. R., et al. 2012, Optics Express, 20, 1503

Mazoyer, J., Baudoz, P., Galicher, R., & Rousset, G. 2014, Astronomy and Astrophysics, 564

Mazoyer, J., Baudoz, P., Galicher, R., & Rousset, G. 2014, Astronomy and Astrophysics, 564, L1

Meshkat, T., Mawet, D., Bryan, M. L., et al. 2017, AJ, 154, 245

Milli, J., Banas, T., Mouillet, D., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9909, Adaptive Optics Systems V, 99094Z

Minowa, Y., Hayano, Y., Oya, S., et al. 2010, in Proc. SPIE, Vol. 7736, Adaptive Optics Systems II, 77363N

Mordasini, C. 2013, A&A, 558, A113

Morzinski, K. M., Males, J. R., Skemer, A. J., et al. 2015, ApJ, 815, 108

Muterspaugh, M. W., Lane, B. F., Kulkarni, S. R., et al. 2010, AJ, 140, 1657

N'Diaye, M., Vigan, A., Dohlen, K., et al. 2016, A&A, 592, A79

Nielsen, E. L., Close, L. M., Biller, B. A., Masciadri, E., & Lenzen, R. 2008, The Astrophysical Journal, 674, 466

Nielsen, E. L., De Rosa, R. J., Macintosh, B., et al. 2019, AJ, 158, 13

Noll, R. J. 1976, J. Opt. Soc. Am., 66, 207

Ogata, K. 2009, Modern Control Engineering, 5th edn. (Pearson)

Patel, R. I., Metchev, S. A., & Heinze, A. 2014, ApJS, 212, 10

Perrin, M. D., Sivaramakrishnan, A., Makidon, R. B., Oppenheimer, B. R., & Graham, J. R. 2003, The Astrophysical Journal, 596, 702

Perrin, M. D., Maire, J., Ingraham, P., et al. 2014a, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9147, Gemini Planet Imager observational calibrations I: Overview of the GPI data reduction pipeline, 91473J

Perrin, M. D., Maire, J., Ingraham, P., et al. 2014b, in Proceedings of the SPIE, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, 91473J

Perryman, M. 2011, The Exoplanet Handbook (Cambridge University Press)

Pickles, A. J. 1998, Publications of the Astronomical Society of the Pacific, 110, 863

Poyneer, L. A., & Macintosh, B. 2004, Journal of the Optical Society of America A, 21, 810

Poyneer, L. A., Macintosh, B. A., & Véran, J.-P. 2007, J. Opt. Soc. Am. A, 24, 2645

Poyneer, L. A., & Véran, J.-P. 2005, Journal of the Optical Society of America A, 22, 1515

Poyneer, L. A., Palmer, D. W., Macintosh, B., et al. 2016, Appl. Opt., 55, 323

Poyneer, L. A., Palmer, D. W., Macintosh, B., et al. 2016, Appl. Opt., 55, 323

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2002, Numerical recipes in C++ : the art of scientific computing

Pueyo, L., Kay, J., Kasdin, N. J., et al. 2009, Appl. Opt., 48, 6296

Quanz, S. P., Crossfield, I., Meyer, M. R., Schmalzl, E., & Held, J. 2015, International Journal of Astrobiology, 14, 279

Racine, R., Walker, G. A. H., Nadeau, D., Doyon, R., & Marois, C. 1999, PASP, 111, 587

Ragazzoni, R. 1996, Journal of Modern Optics, 43, 289

Rajan, A., Rameau, J., De Rosa, R. J., et al. 2017, AJ, 154, 10

Rameau, J., Chauvin, G., Lagrange, A.-M., et al. 2013, The Astrophysical Journal Letters, 772, L15

Raymond, S. N. 2006, ApJ, 643, L131

Rich, E. A., Wisniewski, J. P., Currie, T., et al. 2018, ArXiv e-prints, arXiv:1811.07785

Rodack, A. T., Males, J. R., Guyon, O., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10703, Proc. SPIE, 107032N

Roddier, F., & Roddier, C. 1997, PASP, 109, 815

Rouan, D., Riaud, P., Boccaletti, A., Clénet, Y., & Labeyrie, A. 2000, Publications of the Astronomical Society of the Pacific, 112, 1479

Rowlands, N., Vila, M. B., Evans, C., et al. 2008, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7010, JWST fine guidance sensor: guiding performance analysis, 701036

Ruane, G., Riggs, A., Mazoyer, J., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10698, Proc. SPIE, 106982S

Sadakane, K., & Nishida, M. 1986, Publications of the Astronomical Society of the Pacific, 98, 685

Saumon, D., & Marley, M. S. 2008, The Astrophysical Journal, 689, 1327

Sauvage, J.-F., Mugnier, L., Paul, B., & Villecroze, R. 2012, Optics Letters, 37, 4808

Schneider, J., Dedieu, C., Le Sidaner, P., Savalle, R., & Zolotukhin, I. 2011, A&A, 532, A79

Schöck, M., Els, S., Riddle, R., et al. 2009, PASP, 121, 384

Schwieterman, E. W., Kiang, N. Y., Parenteau, M. N., et al. 2018, Astrobiology, 18, 663

Seo, B.-J., Patterson, K., Balasubramanian, K., et al. 2019, in Techniques and Instrumentation for Detection of Exoplanets IX, ed. S. B. Shaklan, Vol. 11117, International Society for Optics and Photonics (SPIE), 599 – 609

Serabyn, E., Wallace, J. K., & Mawet, D. 2011, Appl. Opt., 50, 5453

Siemion, A., Benford, J., Cheng-Jin, J., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 116

Singh, G., Martinache, F., Baudoz, P., et al. 2014, PASP, 126, 586

Sivaramakrishnan, A., & Oppenheimer, B. R. 2006, The Astrophysical Journal, 647, 620

Smith, B. A., & Terrile, R. J. 1984, Science, 226, 1421

Smith, M., Kerley, D., Chapin, E. L., et al. 2016, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9909, Thirty Meter Telescope narrow-field infrared adaptive optics system real-time controller prototyping results, 99094N

Snellen, I., de Kok, R., Birkby, J. L., et al. 2015, A&A, 576, A59

Snellen, I. A. G., Brandl, B. R., de Kok, R. J., et al. 2014, Nature, 509, 63

Snellen, I. A. G., & Brown, A. G. A. 2018, Nature Astronomy, 2, 883

Soummer, R., & Aime, C. 2004, in Proceedings of the SPIE, Vol. 5490, Advancements in Adaptive Optics, ed. D. Bonaccini Calia, B. L. Ellerbroek, & R. Ragazzoni, 495–503

Soummer, R., Aime, C., Ferrari, A., et al. 2006, in IAU Colloq. 200: Direct Imaging of Exoplanets: Science and Techniques, ed. C. Aime & F. Vakili, 367–372

Soummer, R., Pueyo, L., & Larkin, J. 2012, The Astrophysical Journal Letters, 755, L28

Soummer, R., Sivaramakrishnan, A., Oppenheimer, B. R., et al. 2009, in Proceedings of the SPIE, Vol. 7440, Techniques and Instrumentation for Detection of Exoplanets IV, 74400R

Sozzetti, A. 2017, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10400, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 104001E

Sparks, W. B., & Ford, H. C. 2002, ApJ, 578, 543

Spiegel, D. S., & Burrows, A. 2012, The Astrophysical Journal, 745, 174

Srinath, S., Poyneer, L. A., Rudy, A. R., & Ammons, S. M. 2015, Optics Express, 23, 33335

Stahler, S. W., Shu, F. H., & Taam, R. E. 1980, The Astrophysical Journal, 241, 637

Steck, D. A. 2015, Classical and Modern Optics, available online at `http://steck.us/teaching` (revision 1.5.2, 28 June 2015)

Su, K. Y. L., Morrison, S., Malhotra, R., et al. 2015, The Astrophysical Journal, 799, 146

The LUVOIR Team. 2019, arXiv e-prints, arXiv:1912.06219

Tokovinin, A. 2001, http://www.ctio.noao.edu/~atokovin/tutorial/part3/wfs.html

Tyson, R. K. 2011, Principles of adaptive optics

Unwin, S. C., Shao, M., Tanner, A. M., et al. 2008, PASP, 120, 38

Véran, J.-P., & Herriot, G. 2009, in Frontiers in Optics 2009/Laser Science XXV/Fall 2009 OSA Optics & Photonics Technical Digest (Optical Society of America), JTuC2

Vigan, A., Moutou, C., Langlois, M., et al. 2010, MNRAS, 407, 71

Vigan, A., N'Diaye, M., Dohlen, K., et al. 2019, A&A, 629, A11

Wallace, J. K., Burruss, R. S., Bartos, R. D., et al. 2010, in Proc. SPIE, Vol. 7736, Adaptive Optics Systems II, 77365D

Walter, A. B., Bockstiegel, C., Brandt, T. D., & Mazin, B. A. 2019, PASP, 131, 114506

Wang, J., Mawet, D., Ruane, G., Hu, R., & Benneke, B. 2017, AJ, 153, 183

Wang, J. J., Ruffio, J.-B., De Rosa, R. J., et al. 2015, pyKLIP: PSF Subtraction for Exoplanets and Disks, Astrophysics Source Code Library, , , ascl:1506.001

Wang, J. J., Rajan, A., Graham, J. R., et al. 2014, in Proceedings of the SPIE, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, 914755

Wang, J. J., Graham, J. R., Pueyo, L., et al. 2016, AJ, 152, 97

Weeks, A. R., Xu, J., Phillips, R. R., et al. 1998, Appl. Opt., 37, 4782

Welch, P. D. 1967, IEEE Trans. Audio & Electroacoust, 15, 70

Winker, D. M. 1991, J. Opt. Soc. Am. A, 8, 1568

Wyatt, M. C. 2008, ARA&A, 46, 339

Wynne, C. G. 1979, Optics Communications, 28, 21

Zarka, P., Lazio, J., & Hallinan, G. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 120

Zernike, F. 1934, MNRAS, 94, 377

Ziad, A. 2016, Review of the outer scale of the atmospheric turbulence, , , doi:10.1117/12.2231375

Ziad, A., Blary, F., Borgnino, J., et al. 2013, A&A, 559, L6

Appendix A

Adaptive Optics and Atmospheric Modelling for Optical Communications

In this appendix I will summarize my work completed on applications of AO and atmospheric modelling to optical communications. This work was completed working as an intern at Com Dev/Honeywell Aerospace, in Ottawa, ON, from Aug. - Dec., 2018, as part of an industrial internship program enabled by the NSERC CREATE TEPS program. I will show that transmission improvement by nominal AO correction is limited to only geocentric orbit (GEO) satellite links. I will also present an overview of the simulation framework I have developed for link budget modelling of atmospheric losses due to beam wander, scintillation, and diffraction. With this framework I will show that in some cases a multi-aperture array of smaller telescopes may have better link stability than a single monolithic aperture. Overlapping topics and ideologies connecting the work presented in this appendix to the other exoplanet imaging research presented in this dissertation are also further discussed.

A.1 Introductory Material

Variables used in this appendix are shown in Table A.1. A conceptual illustration of a ground-to-satellite optical communication link is illustrated in Fig. A.1. The discrepant beam footprint between the uplink and downlink will cause different levels of end-to-end throughput losses and variation over time, directly influencing the

Table A.1: Variables used in this appendix.

variable	**description**
λ	electromagnetic wavelength
θ_0	isoplanatic angle
C_n^2 profile	atmospheric refractive index structure profile
r_0	atmospheric coherence length, or Fried parameter
τ_0	atmospheric coherence time
L_0	von Kármán outer scale
L_T	Talbot Length

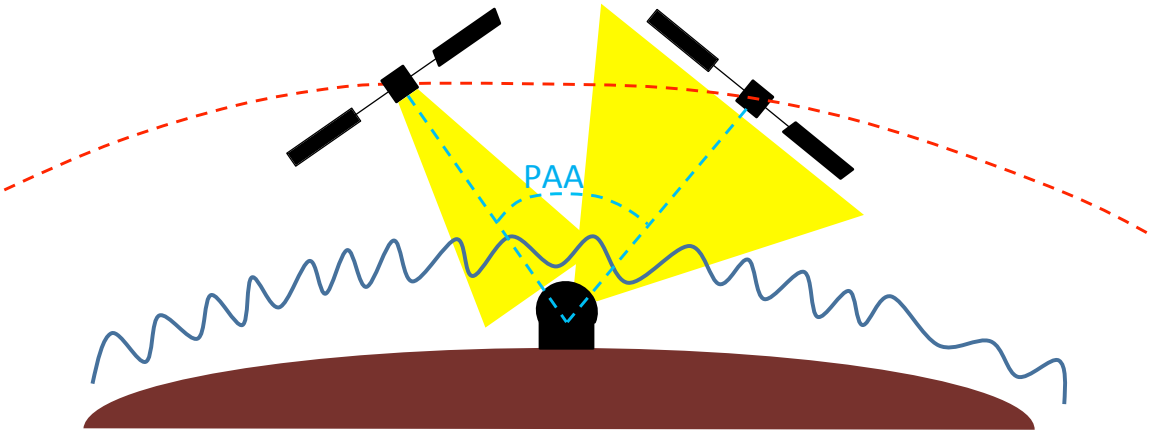


Figure A.1 Illustration of an optical communication link between a satellite and optical ground receiver (OGR), shown not to scale. A satellite orbits along the path of the dotted red line, at altitudes on the order of ∼hundreds of km, while atmospheric turbulence is located much closer to the OGR, at altitudes on the order of ∼tens of km. As a result, this near field vs. far field effect, which will be explored further in this appendix, enables a smaller downlink footprint at the OGR compared to the uplink footprint at the satellite. Orbital motion of the satellite also causes the downlink and uplink to propagate through different atmospheric phase offsets, seen by the OGR as an angular offset called the point ahead angle (PAA).

achievable transmission bandwidth and link stability. The PAA effect will also set limitations for an AO system on the OGR to pre-compensate for the uplink using information measured by the downlink, depending on θ_0, illustrated in Fig. A.2. The isoplanatic angle is given by $\theta_0 \equiv r_0/h$, and can be thought of as the angle at which the projected telescope pupil no longer goes through the same patches of turbulence, and is thus directly dependent on r_0–the typical size of a single turbulent atmospheric "cell"– and h, the characteristic height above the ground of atmospheric turbulence. The atmospheric coherence time, first introduced in §1.4.3.2.2, is given by $\tau_0 \equiv r_0/v$,

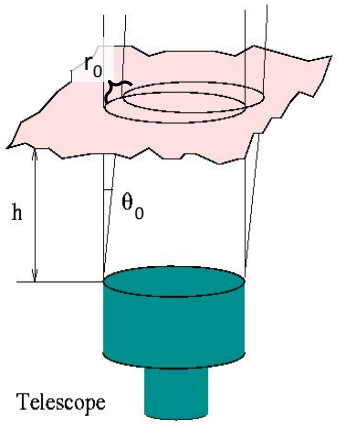


Figure A.2 An illustration of the concept of the isoplanatic angle (θ_0), adapted from Tokovinin (2001), defined by the angle, at the characteristic height (h) in the atmosphere, at which two phase screens become decorrelated, no longer seeing the same turbulence.

where v is the characteristic atmospheric windspeed. In reality, the strength of the atmosphere is not concentrated at a single height of h, but instead distributed over a range of altitudes, defined by the C_n^2 parameter and illustrated in Fig. A.3 Lastly,

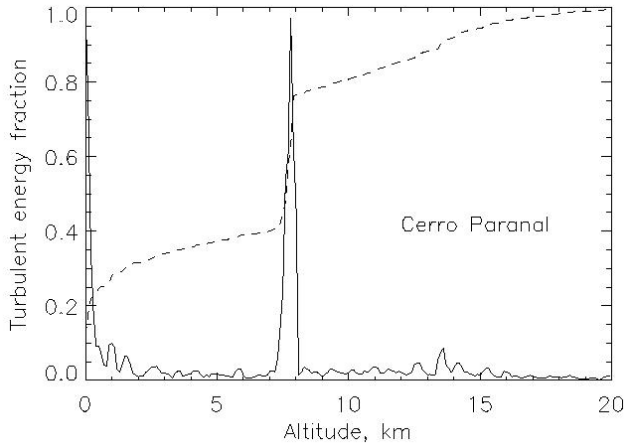


Figure A.3 An illustration of the concept of the atmospheric refractive index structure profile, from Tokovinin (2001), showing an example of how atmospheric turbulence is distributed over a range of altitudes.

the spatial distribution of WFE over the telescope pupil (in contrast in the altitude distribution in Fig. A.3), is illustrated in Fig. A.4, showing the difference between Kolmogorov- and von Kármán-scaled turbulence. In addition to §A.6.5, further details and equations on atmospehric trubulence and AO are taken from Tokovinin (2001) and Tyson (2011), respectively.

A.2 Atmospheric Models for LEO Satellites

In this section I use parametric models of atmospheric turbulence to motivate the required performance of an AO system for optical communications, for the downlink

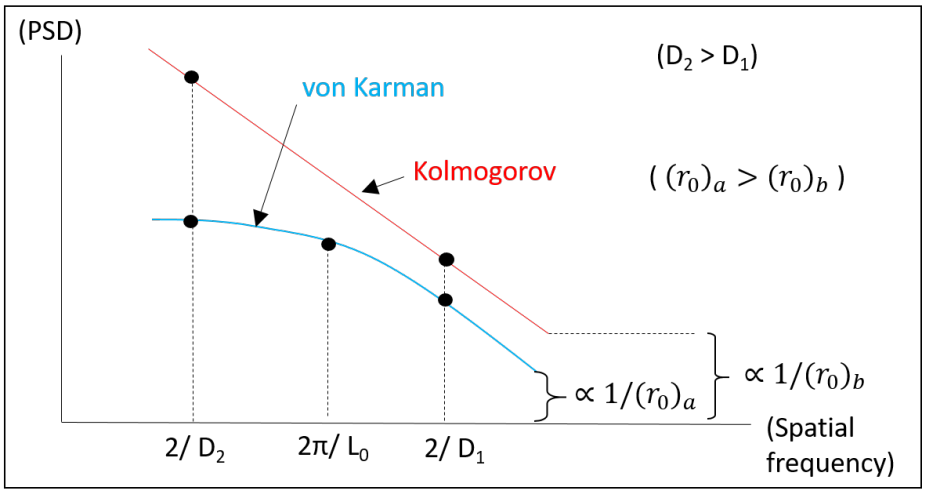


Figure A.4 An illustration of the difference between a Kolmogorov and von Kármán PSD. The PSD normalization is scaled by $1/r_0$; an additional "knee" is defined in the von Kármán PSD by the outer scale (L_0), at a spatial frequency of $2\pi/L_0$. Atmospheric WFE is determined from the atmospheric PSD via: WFE= $\sqrt{\int_{2/D}^{\infty} df\ \mathrm{PSD}(f)}$, where f is spatial frequency and D is a telescope diameter "collecting" atmospheric turbulence over the full atmosphere. This equation assumes the light source is at infinity (i.e., a star), so that the wavefront is diffraction-limited and perfectly flat as it enters the atmosphere.

and/or uplink configurations.

A Hufnagel-Valley boundary C_n^2 profile (Tyson, 2011, eq. 2.18) is used to generate r_0 and θ_0. A Bufton wind profile (Tyson, 2011, , eq. 2.17) is then used to calculate τ_0. For scintillation, the Roytov variance is calculated from (Kaushal et al., 2017, eq. 2.91). Open loop WFE can be computed from r_0 using a corresponding spatial PSD and assumed telescope size (Tyson, 2011, eq. 2.28-2.29), as in Fig. A.4. Using a von Kármán PSD, a L_0 value must be assumed (Tyson, 2011, eq. 2.11). With these models the following can be estimated:

1. λ/r_0, providing an estimate of the seeing and Roytov variance,

2. θ_0 which can be compared to the expected PAA to understand the feasibility of pre-compensated OGR AO correction for the uplink,

3. τ_0 which provides an estimate of the required AO system frame rate needed for optimal performance, and

4. open loop WFE, which directly translates to DM stroke requirements and can

also set dynamic range requirements for the AO WFS.

These parameters are plotted in Fig. A.5 for a wavelength of 1560 nm.

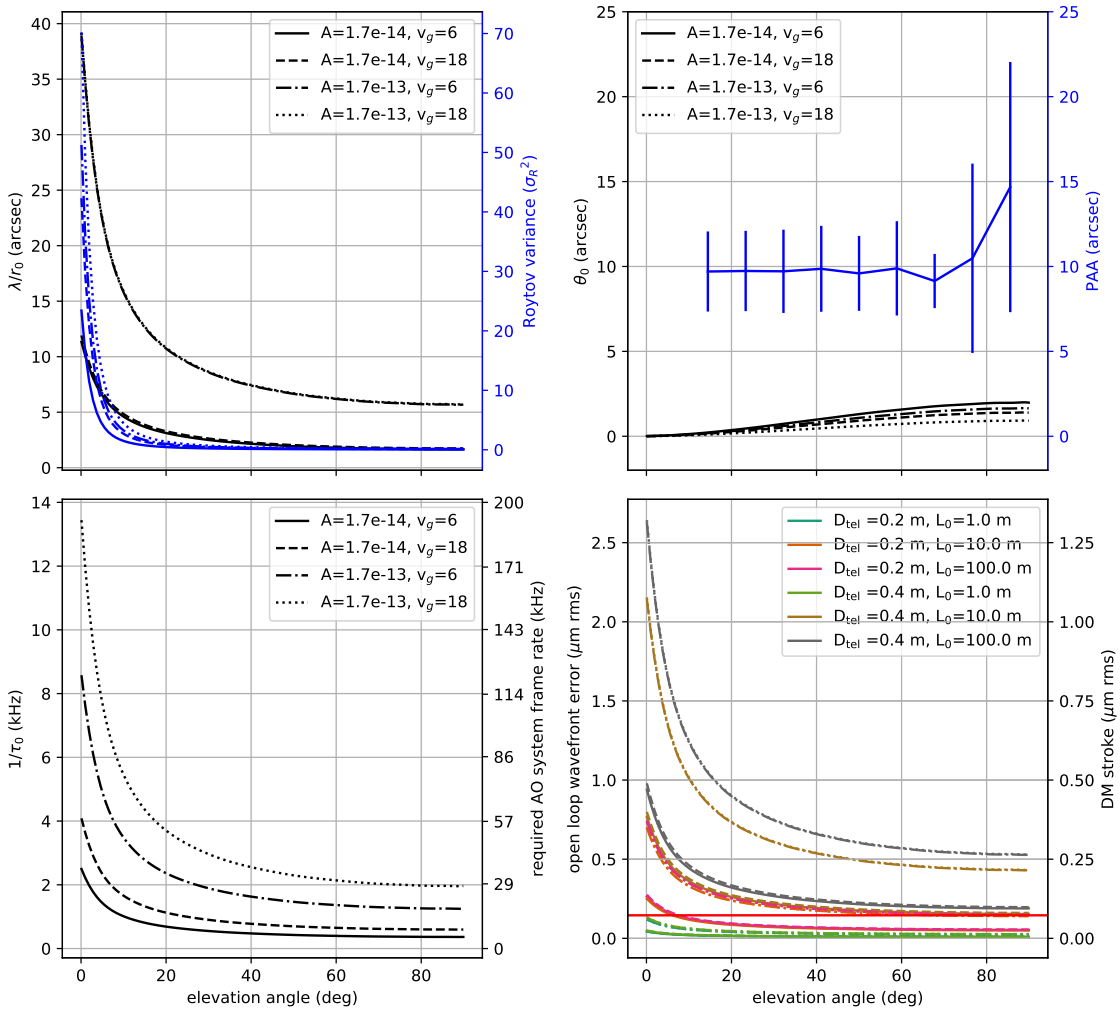


Figure A.5 upper left: seeing (black) and scintillation (blue). Upper right: θ_0 (black) and PAA (blue). Lower left: inverse of τ_0 (left y-axis) and corresponding AO system frame rate (right y-axis, assuming an integral controller with a gain of 0.55). lower right: open loop WFE (left y-axis), assuming a von Kármán power law, and corresponding DM stroke (right y-axis; assuming a reflective surface, this is just the left y-axis values divided by 2). The horizontal red line shows the WFE for a diffraction limited telescope (i.e., a SR of 0.7) via the Maréchal approximation (Mahajan, 1983). For all panels parameters are shown for a typical range of A (free parameter in the Hufnagel Valley C_n^2 profile), and v_g (free parameter in the Bufton wind model), and are plotted vs. elevation angle.

For all models, Fig. A.5 shows the given value vs. elevation angle. An elevation

angle "fudge factor" function computes the path length through the atmosphere as a function of elevation angle relative to zenith. Most AO textbooks typically list this fudge factor function as the secant of the zenith angle, but such an expression deviates from the correct answer at low elevation (i.e., the path length is not infinity at an elevation angle of zero). This "slant range" equation is instead implemented here; this is a transcendental equation, requiring a numerical solver and polynomial fit to produce an analytical form.

The lower left and right panels of Fig. A.5 set typical requirements for an AO system in a range of conditions: required frame rate can vary anywhere between about 1 and 200 kHz, and required DM stroke can vary between about 0 and 1.2 microns RMS. The former requirement could potentially limit accessible elevation angles (i.e., assuming a 200 kHz frame rate is not achievable with current technology), while the latter requirement is not particularly constraining, since current commercials DMs can be made with tens of microns of stroke (e.g., Le Bouquin et al., 2018). This AO frame rate requirement is much higher than typical systems for astronomical application because of the setup for a low earth orbit (LEO) satellite: the angular velocity observed at the OGR, which is highest at zenith and lowest at the horizon, is much higher than the angular velocity of the celestial sphere due to rotation of the Earth. As a result, the LEO angular velocity must be added to the Bufton wind velocity profile as a second term (e.g., as in Gruneisen et al., 2016, §3.3), constantly increasing the wind velocity with increasing elevation to values much larger than normal atmospheric conditions as seen from a stationary observer at the OGR.

Figure A.6 illustrates an extension from the conclusions of Fig. A.5. The main conclusion from Fig A.6 is that in some cases AO is not necessary because the open loop WFE is already diffraction limited. For example, at 1.7 acrsecond seeing (at 1560 nm), telescopes with a diameter smaller than about 25 centimetres are generally always diffraction-limited, although larger telescopes meet this criteria for smaller outer scales. In reality the seeing and outer scale will constantly vary at different site locations and from night to night (see §A.6.2.2). The main point of Figure A.6 is simply to illustrate that at some of these locations for some of those nights, AO may or may not be necessary depending on the telescope diameter.

The upper right panel of Figure A.5 clearly shows that the typical PAA for LEO satellites[1] is too large for pre-compensated AO correction of the uplink to be feasible,

[1]The PAA calculation illustrated in Fig. A.5 is from the output of LEO orbit simulations provided by Honeywell.

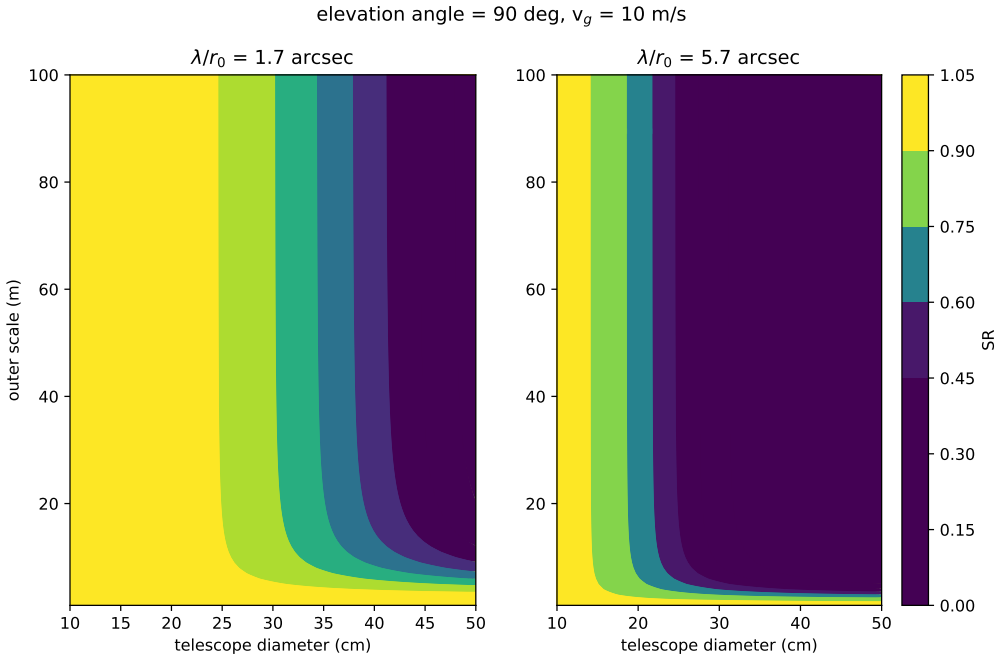


Figure A.6 SR, calculated via the Maréchal approximation, at λ= 1560 nm as a function of L_0 and telescope diameter for two different seeing values (corresponding to different values of A in the Hufnagel Valley boundary model (left and right panels).

illustrating why AO correction is typically limited to GEO satellites. For the downlink, turbulence is in the far field and will generally have a weaker effect on signal attenuation compared to the same turbulence realization for the uplink (which is why surveillance satellites generally do not benefit from AO). This issue could be resolved by using a point-ahead laser guide star and/or satellite constellation, but for now, this plot motivates my main work in §A.3, which will instead focus on open loop atmospheric models without AO correction.

A.3 Fresnel Propagation Simulations

In this section I will discuss the various setup, implementation, and results of my MATLAB simulations using PROPER (Krist, 2007), an open source Fresnel propagation code. Although PROPER is commonly used for detailed end-to-end polychromatic simulations of exoplanet imaging instruments (e.g., Marois et al., 2008c), here I will apply the same software to simulate scintillation effects from atmospheric turbulence in a simulated optical communication link setup between an OGR and

LEO satellite. Fresnel wavefront propagation is based on the Talbot effect (Krist, 2007, and references therein), illustrated in Fig. A.7 and described below.

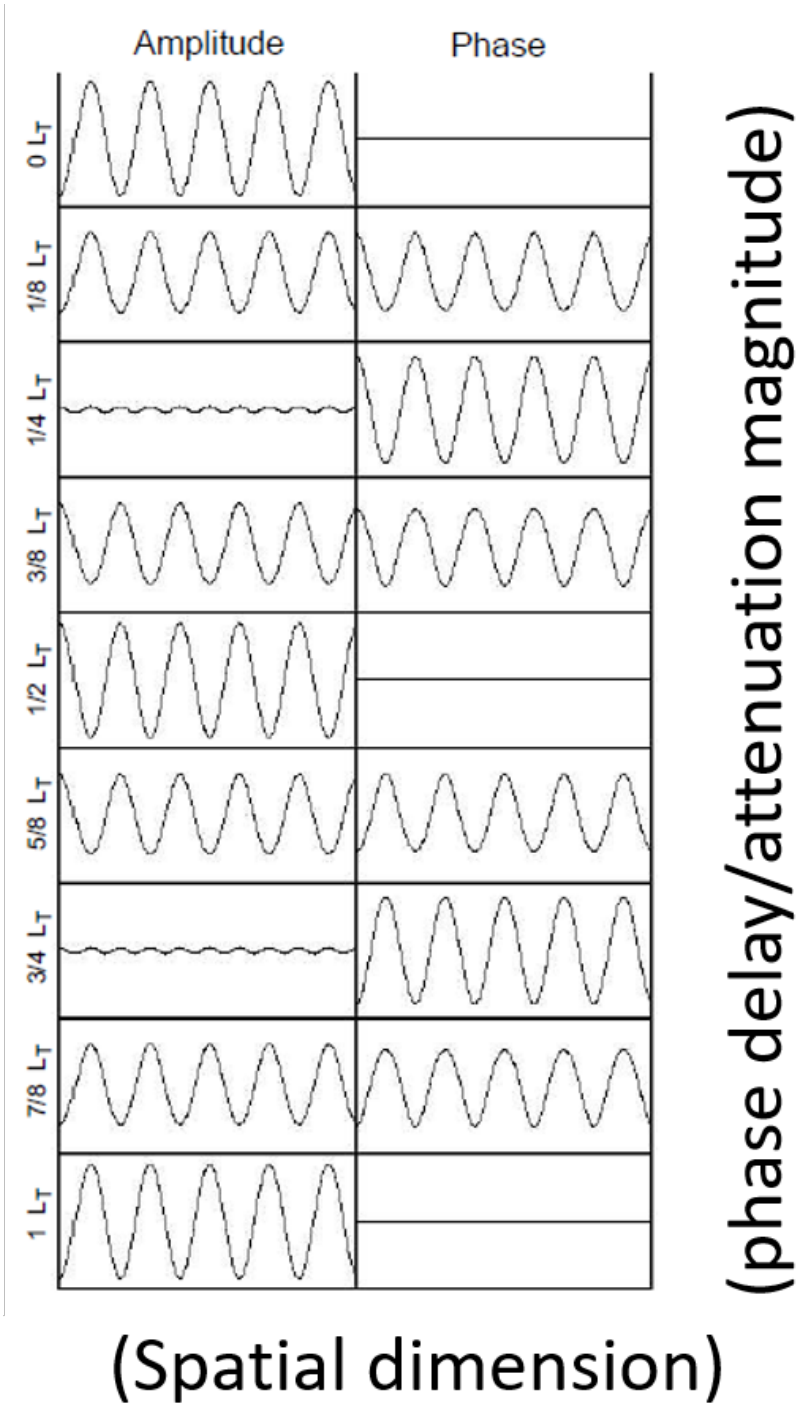


Figure A.7 An illustration of the Talbot effect, adapted from (Krist, 2007, Fig. 21). Wavefront propagation is in the downwards direction, beginning from the top, and illustrated as a function of L_T. The x-axis is a slice of the spatial direction, and the y-axis represents the wavefront amplitude (left) or phase (right) along that slice. Over a period of L_T the wavefront evolves from a pure amplitude sine wave to a mixture of phase and amplitude (including a pure amplitude cosine wave) and ultimately back to a pure amplitude sine wave.

The main principle of the Talbot effect is that propagation of a pure sine wave

(in either phase or amplitude) will convert between phase and amplitude, ultimately returning to the original state periodically. The periodic length over which this evolution happens is called L_T and is given by

$$L_T = 2P^2/\lambda, \tag{A.1}$$

where P is the period of the sine wave. One main principle of Fresnel propagation is that the Talbot effect prevents a wavefront from ever returning to the same state. E.g., for an amplitude sine wave with $P = 10$ cm period and λ= 735 nm (i.e., the Nyquist limit for a 10 cm telescope a operating at 735 nm) $L_T \approx 27$ km, whereas for $P = 5$ cm period and $\lambda = 735$ nm $L_T \approx 7$ km; because these two Talbot lengths are not integer multiples of one another, at any given multiple of the Talbot length for one period, the other period will be some mixture of amplitude and phase, never returning to the original state of two pure amplitude sine waves. Atmospheric scintillation is a result of this effect; because light propagating through phase aberration in the atmosphere (i.e., a continuous power spectrum of spatial frequencies) is converted between phase and amplitude aberration according to equation A.1, time-variable atmospheric turbulence of purely phase aberration converts into a time-variable power spectrum of amplitude aberration.

Note that here I am not considering atmospheric losses due to particle scattering and/or absorption. These simulations are meant to understand and isolate the effects of wavefront propagation through turbulent atmospheric phase screens of pure phase aberration (i.e., pockets of air with variable indicies of refraction).

A.3.1 Setup

The same atmospheric and wind models from §A.2 are used here to simulate a dynamic model of atmospheric turbulence. Unless otherwise specifically mentioned, I use a ground wind speed of $v_g = 10$ m/s, $A = 1.3 \times 10^{-13}$ (which is constrained from measurements provided by the Ottawa division of the Royal Astronomical Society of Canada to the typical 1.7 arcsecond seeing in a narrow band Hα filter) and a seven layer atmospheric phase screen model is used, duplicating the L_0 and wind direction values from Guyon & Males (2017).

Fig. A.8 shows two different time stamps of the above described seven layer atmospheric model at zenith. I chose arbitrarily to "end" the atmosphere at 32 km, at which point the WFE (labeled σ in Fig. A.8) has significantly decreased below the

ground layer values.

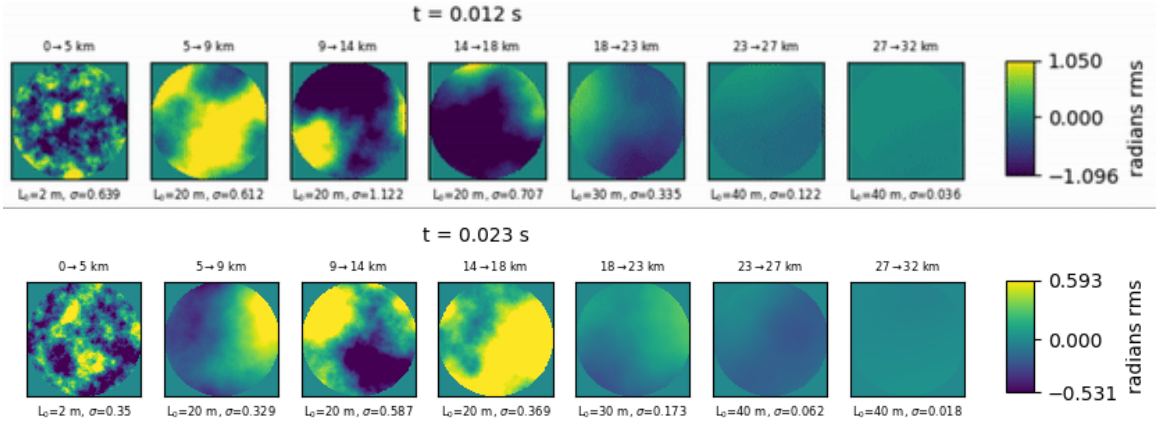


Figure A.8 Two timestamps of a model for a translating atmospheric phase screen at zenith, assuming a 40 cm aperture. Ranges of elevations, von Kármán outer scale, and corresponding WFE are shown for each layer. Panel a is viewable as an animation in Adobe Reader version 7 or greater.

Next, going to a full scale wavefront Fresnel propagation model, I model the wavefront phase and amplitude distribution starting from the exit pupil of the OGR/satellite (for the uplink/downlink, respectively) and ending at the entrance pupil on the other end. Illustrated simulations of this end-to-end simulation procedure for the downlink and uplink are in Figures A.9 and A.10, respectively, and outlined below:

1. Starting at the exit pupil, a symmetric 2D Gaussian intensity illumination profile is added, with the exit pupil 4.5 times larger than the Gaussian standard deviation. Then, static phase and amplitude aberration are added to the same exit pupil (i.e., on top of the existing Gaussian amplitude profile), where the WFE and power law are free parameters. In my simulations I use 56 nm RMS phase aberration, 0.17 % RMS intensity aberration, and a -2 power law for both amplitude and phase.

2. The aberrated wavefront is then propagated from the exit pupil to the center of the first atmospheric layer (top of the atmosphere for downlink; bottom of the atmosphere for uplink). This propagation causes the Gaussian beam to expand due to diffraction; this expansion will increase both with distance and with larger static WFE.

3. At the first layer, an atmospheric phase screen is generated and then added to the phase of the wavefront over the full image. The sampling at that layer

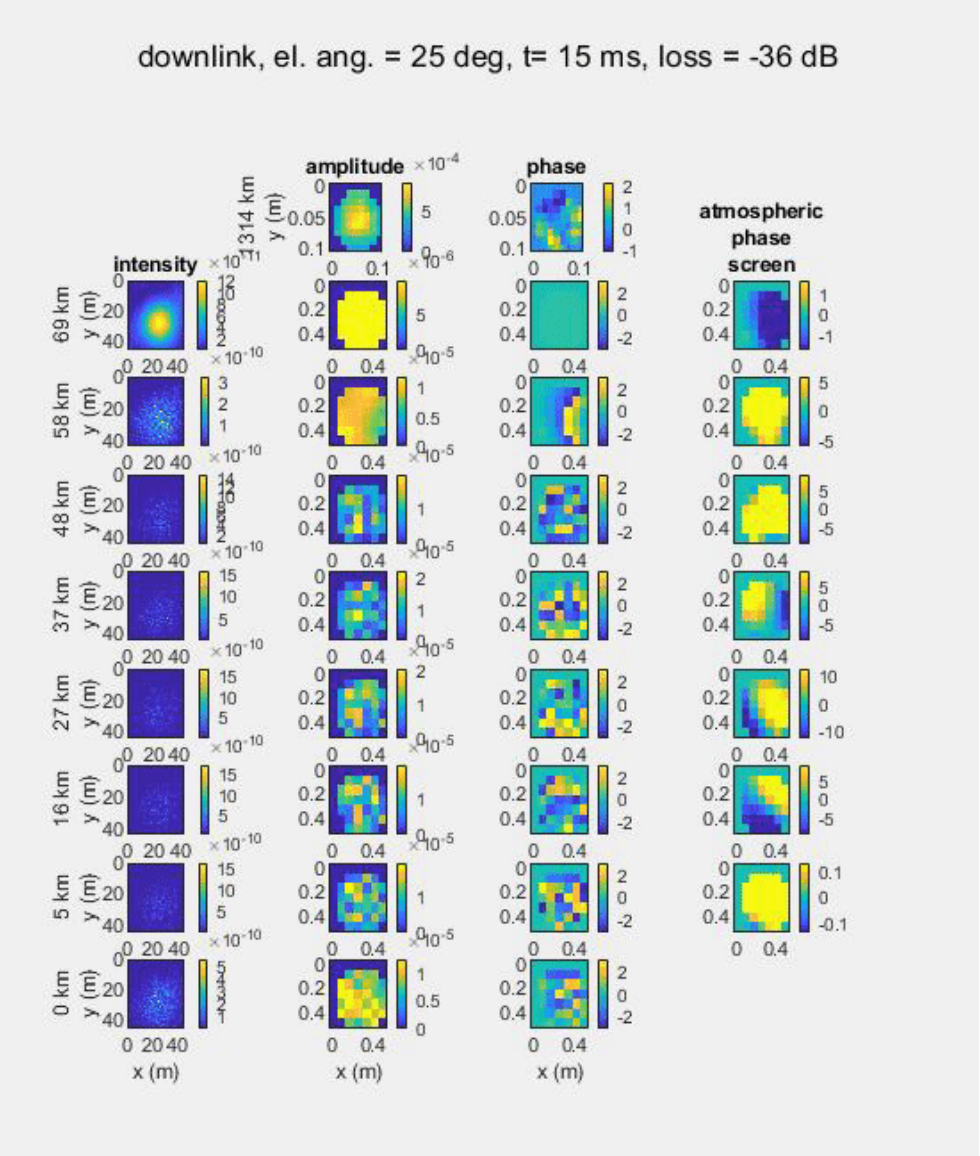


Figure A.9 A downlink illustration at a single time stamp and elevation angle. The rows represent different altitudes; the satellite is on the top row, and OGR is on the bottom row, and the centre of each atmospheric layer is in between. In columns, the left is intensity of the wavefront, the left middle is amplitude of the wavefront over the projected entrance pupil (OGR for the downlink), the right middle is phase of the wavefront over the projected entrance pupil, and the right is the atmospheric phase screen over the projected entrance pupil. An atmospheric phase screen is added to the phase of the wavefront at each atmospheric layer before propagating to the next layer and ultimately arriving at the entrance pupil. The figure is viewable as an animation in Adobe Reader version 7 or greater.

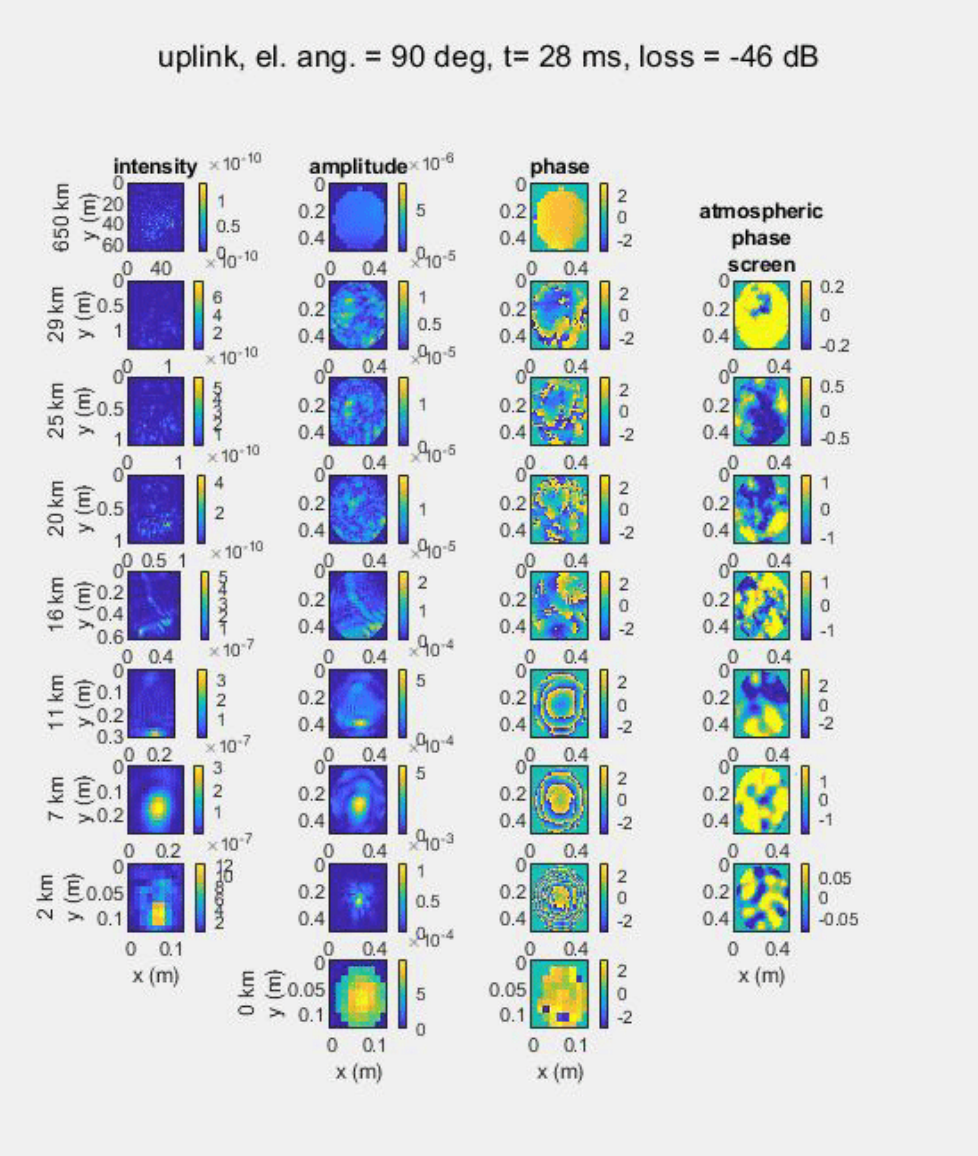


Figure A.10 The same illustration as Fig A.9 for the uplink. The exit and entrance pupils are now the OGR and satellite, respectively. The figure is viewable as an animation in Adobe Reader version 7 or greater.

defines the spatial scales present in the phase screen, while the upper and lower height bounds of the atmospheric layer define the strength of the turbulence.

4. The wavefront is again propagated to the centre of the next layer, and a phase screen is again generated and added to the wavefront at this layer given the sampling. This process is repeated over each atmospheric layer until reaching the top or bottom layer (for the uplink or downlink, respectively).

5. The wavefront is propagated through free space from the last atmospheric layer to either the satellite (uplink) or OGR (downlink). A pupil mask for the entrance pupil is defined, given the sampling of the wavefront and the entrance pupil?s physical diameter, and then atmospheric loss is computed as the ratio of intensities between the exit and entrance pupil.

6. Steps 2 - 5 are then repeated after translating the atmospheric phase screen in each layer given a user-defined time difference between frames and a ground wind velocity (which is used by the aforementioned velocity profile (§A.2) to compute the wind velocity at each atmospheric layer).

7. The average and standard deviation of the loss values after a user-defined number of iterations (i.e., integration time) are saved, and all above steps are repeated over any additional free-parameters of interest.

A.3.2 Convergence tests

Many additional modifications/tests were made to the multi-step procedure described above in §A.3.1. In this section I consider and analyze possible numerical limitations for the following parameters and/or concepts in my simulations: number of atmospheric layers (§A.3.2.1), the distribution of spatial scales in my numerically generated atmospheric phase screens and their effect on transmission values (§A.3.2.2), and the numerical implementation of tip and tilt Zernike modes (§A.3.2.3).

A.3.2.1 Number of Atmospheric Layers

Investigating how many atmospheric layers are needed can be done by increasing the number of layers until both the transmission and standard deviation converge. The minimum number of layers at which convergence is reached should then in principle be sufficient to use going forward. By deviating from the seven layer setup presented in Guyon & Males (2017), I can no longer use the same prescribed L_0 values, and so for this setup I use the worst case scenario $L_0 = 100$ m at every layer (but see §A.6.1). Additionally, in these simulations the atmospheric phase screen at each new time step is randomly generated according to the prescribed power law and wavefront error (i.e., it is completely de-correlated to the atmospheric phase screen of the previous time step); although randomizing the atmospheric phase screens is not physical on small timescales, this effect should allow convergence to a realistic long

exposure transmission value faster than translating the phase screens and therefore require less iterations. The results of these convergence simulations are shown in Fig. A.11.

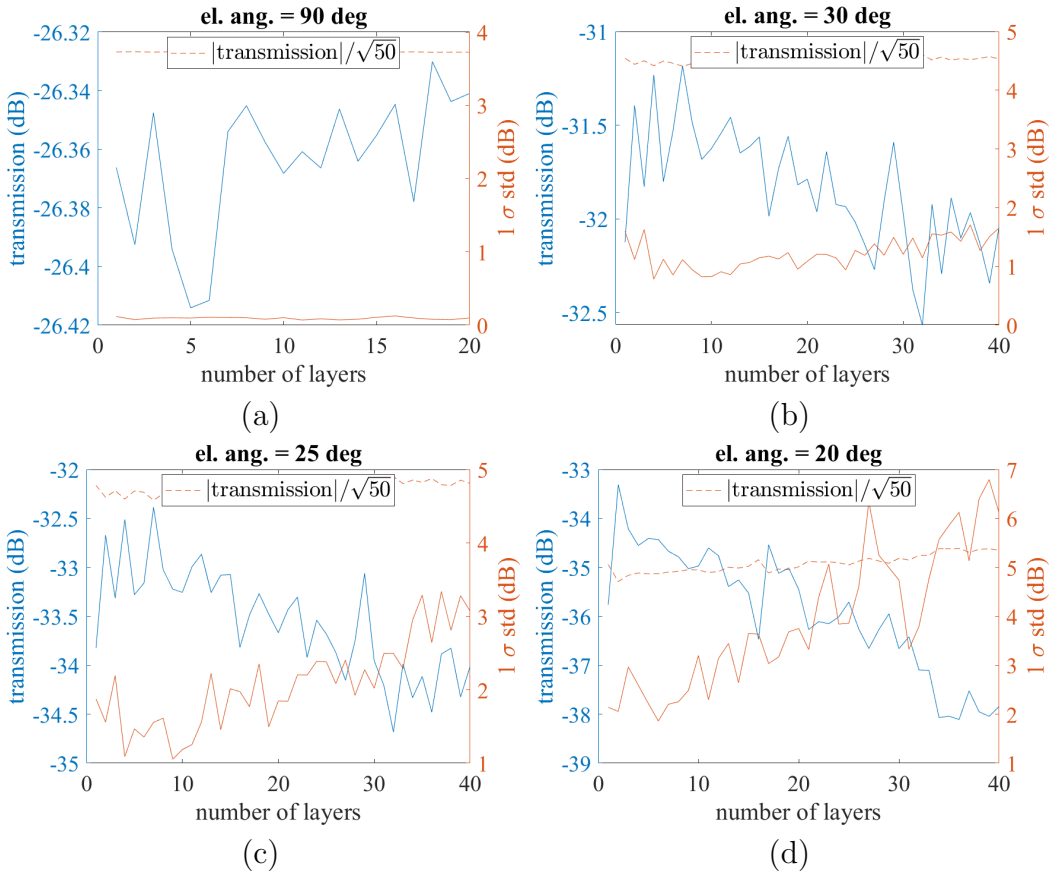


Figure A.11 Convergence simulations over number of atmospheric layers for four different elevation angles (a-d). Transmission and standard deviation of transmission as a function of number of atmospheric layers is shown on the left and right y-axis, respectively. The standard deviation expected from a Poisson noise distribution (using 50 iterations per simulation) for each simulation is also shown.

Fig A.11 illustrates that both standard deviation and transmission converge immediately, even for one layer, for elevation angles greater than about 20 degrees. At 20 degrees (Fig. A.11 d), both the transmission and standard deviation do not converge all the way out to 40 layers. This apparent divergence may be an un-physical result of how the Talbot effect is applied in the strong turbulence regime (J.-P. Véran, private communication). The Talbot effect occurs at every elevation in these simulations, but at lower elevations the layer spacing definitions appear to impact the simulation results. For a single layer, you would propagate to the middle of the atmosphere,

whereas for two layers you would propagate to the first third and second third of the atmosphere; at lower elevations, where these distances are larger and the aberration is stronger, the amount of phase converted into amplitude is different enough between the one and two layer case so that the amount of scintillation is different, and the same again is true going from two to three layers, from three to four, etc. Regardless of the underlying cause, because of the results from these simulations, all subsequent analysis will only consider elevation angles above 25 degrees.

A.3.2.2 Atmospheric Wavefront Error

As described in §A.2, a telescope diameter must be assumed to compute the integrated WFE at a given atmospheric layer. Although initially I assumed that this diameter was the projected entrance pupil onto the given atmospheric layer, this assumption yielded un-physical WFE on larger spatial scales, illustrated in Fig. A.12 a.

Fig. A.12 shows that, depending on the user-defined choice of image and pupil size, telescope diameter, and outer scale, the measured WFE on different spatial scales may not match the expected value from integrating theoretical von Kármán turbulence (Tyson, 2011, eq. 2.11). Particularly, Fig. A.12 a shows that by using the above-mentioned "projected entrance pupil" for WFE normalization (i.e., assuming a satellite diameter on the scale of a metre despite the full beam radius being on the scale of tens of metres), the resultant atmospheric phase screen WFE is orders of magnitude larger than it should be at spatial scales larger than the projected entrance pupil. These effects are a result of

1. spatial sampling limited to scales (relative to the user-defined pupil diameter and outer scale) larger than the image and smaller than two pixels, and

2. un-physical numerical artifacts that deviate the generated phase screen from the true PSD shape (see §A.6.2.1).

Point 1 is an unavoidable limitation that can only be addressed by using larger images. This is illustrated by the improvement between Fig. A.12 b and c, which shows that going from a 2048 to 4096 image size can more accurately reach the expected WFE on smaller spatial scales. However, even with the 4096 pixel image size, smaller spatial scales are still off from theoretical values by at least an order of magnitude. With limited computing resources, images larger than this scale are not practical to simulate due to limited memory, and so this sets limitations on the realistic outer

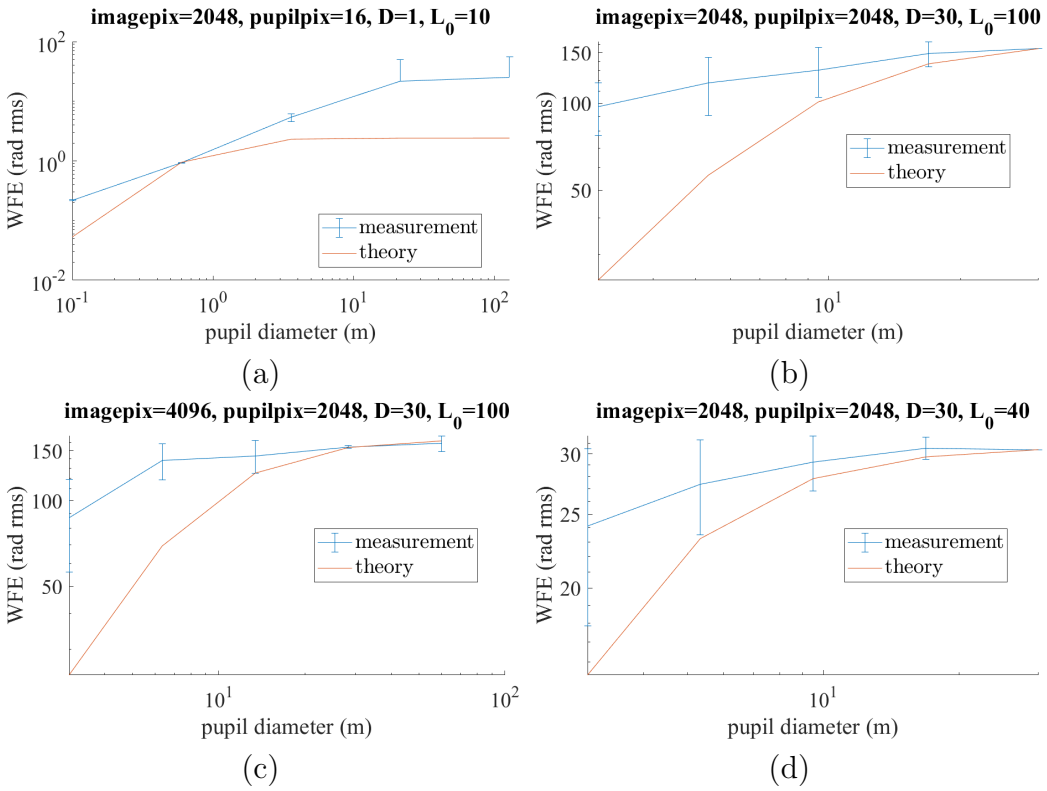


Figure A.12 Measured WFE on a von Kármán atmospheric phase screen as a function of user-defined "pupil diameter," or aperture, over the phase screen, compared to expected theoretical values. "imagepix" and "pupilpix" are the image and pupil size in units of pixels, respectively, while "D" and "L_0" are the pupil diameter and outer scale in units of meters. Ten different random, uncorrelated phase screens are used to plot the mean and standard deviation for the blue "measurement" line. The "theory" line is generated from integrating the theoretical von Kármán PSD given the value of L_0, as illustrated in Fig. A.4. Integrated WFE is computed in a single layer over the full atmospheric C_n^2 profile.

scales that can be used. Fig. A.12 d, however, shows that for a realistic beam size and outer scale (L_0=40 m is the largest value used in the stratification defined by Guyon & Males 2017), measured WFE is only off by a factor of two, at most, from the theoretical values on the smallest spatial scales; this smaller discrepancy on the smaller spatial scales is much better than the orders of magnitude discrepancy in Fig. A.12 a and should have a negligible impact on transmission losses.

Fig. A.12 also implies that phase screens need to be normalized to the largest spatial scales to realistically match theoretical values. To test this claim, I ran simulations to measure the effects of atmospheric loss on the above-mentioned user-defined

"pupil diameter" on a given atmospheric phase screen. These simulations used a single atmospheric layer, informed by the results from Fig. A.11, and, as in §A.3.2.1, generated completely de-correlated phase screens at each new time step to allow more efficient numerical convergence. A 100 metre outer scale is used; although Fig. A.12 showed that WFE is overestimated on smaller spatial scales in this case, the larger spatial scales to which the WFE is normalized should still dominate the transmission effects (but see §A.6.1). Note that as a result of using this larger outer scale, the total WFE is larger, and so atmospheric transmission losses will be larger than if using a smaller outer scale; the point of these simulations is just to show convergence of the atmospheric "pupil diameter" size, not to realistically simulate atmospheric losses.

The convergence simulation results in Fig. A.13 illustrate that the entire image should be used to normalize phase screens at each atmospheric layer. It is not surprising that the atmospheric losses increase when increasing the user-defined aperture of the phase screen; this effect increases the total WFE at a given layer and therefore increases losses from scintillation. Although the magnitude of scintillation losses may be unrealistically large here, as discussed above, this conclusion is still particularly important in understanding the relative impact of different effects contributing to atmospheric loss. With no atmospheric turbulence, loss would be governed entirely by the expansion of a Gaussian beam in free-space propagation. With strong enough atmospheric turbulence, this expansion due to free-space diffraction may be negligible compared to additional "expansion" effects from scintillation (see §A.4 and Fig. A.17).

A.3.2.3 Tip/Tilt implementation

Atmospheric tip/tilt (T/T), or "beam wander," will be considered separately from beam expansion due to scintillation. PROPER, which is not a ray trace program, may not be correctly simulating T/T when integrating over the full propagation distance; further investigation showed that although PROPER does simulate T/T, the sampling is not re-calculated when the T/T angle would otherwise steer the beam out of the image, as which point T/T is incorrectly simulated. This concept is illustrated in Fig. A.14.

As a result of this limitation, the following procedure was implemented to properly account for T/T amplitudes:

1. At each atmospheric layer, T/T is removed from each phase screen by a least-

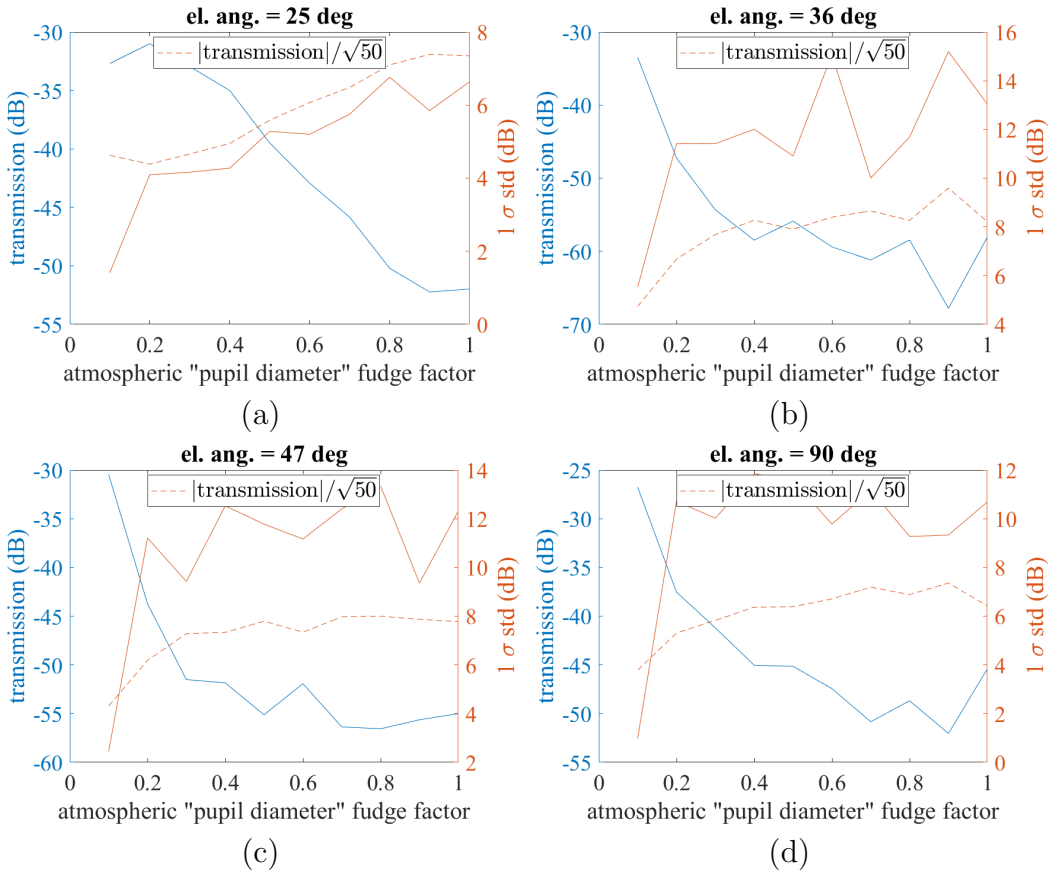


Figure A.13 Uplink convergence simulations to motivate the spatial scale to which WFE should be normalized. The left and right y-axes are the mean and standard deviation of transmission over 50 uncorrelated atmospheric realizations. The standard deviation for a Poisson noise distribution is also shown. The x-axis is a unit-less "fudge factor" representing the size of the pupil used to normalize the phase screen relative to the size of the full image: 1 and 0 represent an assumed diameter of the full image and smaller than 1 pixel, respectively. Results for four different elevation angles (a-d) are shown.

squares subtraction algorithm (Lafrenière et al., 2007) before adding the phase screen to the wavefront and propagating the next layer.

2. The T/T least-squares coefficients are saved at each layer.

3. After propagating the full distance between the exit and entrance pupil, the T/T-removed wavefront is then manually translated in the plane of the entrance pupil (i.e., before applying any aperture mask) using the sign of the coefficients from step 2, the magnitude from theoretical open loop T/T coefficients predicted by Noll (1976) (but see §A.6.2.1), and the distances between each layer and the

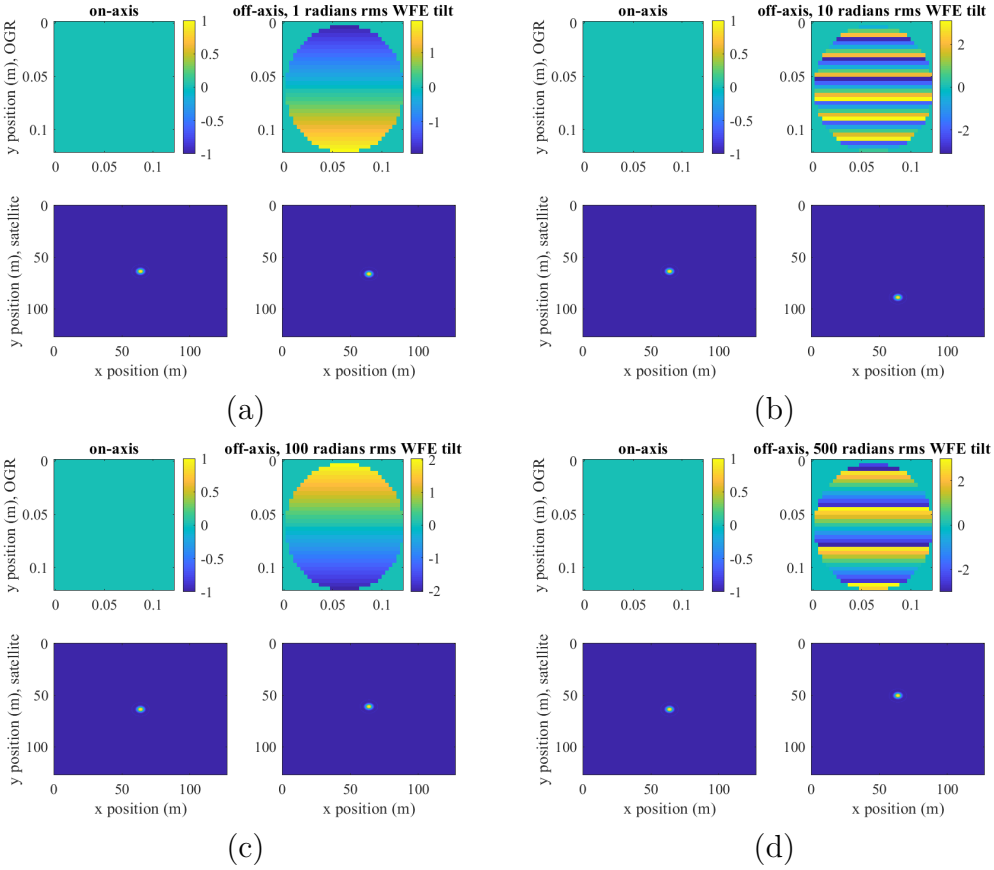


Figure A.14 A comparison between a flat wavefront (in both amplitude and phase; only the phase is shown) exiting the pupil of a 12 cm unobscured OGR and a tilted wavefront (in phase, still flat in amplitude), both propagated 650 km to the satellite. Four different tilt amplitudes are simulated (a - d). As soon as the peak-to-valley tilt amplitude increases above 2 π radians (b-d) there is phase wrapping, but in c and d there is also aliasing (i.e., 2 π radians is smaller than the distance between two pixels). The result in c and d are that tilt is aliased back into lower angles in the opposite direction.

entrance pupil plane. Translations are typically on the order of tens of metres for the uplink and tens of centimetres for the downlink.

4. The shifted wavefront from step 3 is multiplied by a binary mask, representing the entrance pupil, in the centre of the image. Atmospheric transmission is then computed (as before) as the ratio of integrated of intensities between the entrance and exit pupils.

A.3.3 Multi-aperture setup

Many previous papers have proposed the use of optical communication transmitters and receivers using multiple apertures to improve transmission limitations from scintillation effects (e.g., Kim et al., 1997; Weeks et al., 1998). L Edwards et al. (2012) describe a more recently implemented system that includes a dedicated setup of multi-aperture transmitters and receivers. The potential gain from a multi-aperture array over a monolithic telescope is to mitigate scintillation effects that can cause "brown outs."

First, considering a single OGR, if the diameter is smaller than r_0, it should only see one "cell," or spatial scale, of atmosphere, and can thus be considered diffraction-limited. Accordingly, scintillation will generally only exist as "piston modes" over the OGR pupil, and so the standard deviation should be considerably lower than for a larger telescope with higher order modes. For the same reason, an array of telescopes with each diameter smaller than r_0 and each separtion larger than r_0 should see less impact from scintillation than from a monolithic aperture whose total collecting/receiving area is the same as the sum of the smaller telescope array. As a receiver, this array is not co-phased to act as a single telescope and would thus require the light from all the separate apertures to be coupled into/onto a single fibre/detector to reach the same mean transmission values as the monolithic case with the same total receiving area. For a multi-aperture transmitter, the light from each small telescope must be incoherent with one another in order to prevent additional "amplitude aberration" from amplifying scintillation effects instead of attenuating them; e.g., L Edwards et al. (2012) design the four-aperture uplink array to transmit at four separate wavelengths, with each aperture at a slightly different wavelength to maintain incoherence.

With these requirements in mind, an illustration of my multi-aperture transmitter setup is in Fig. A.15. After about 20 km, the separate beam footprint from each of the transmitters is no longer apparent; at this point, scintillation effects have redistributed the five Gaussian intensity profiles over a larger spatial scale, "erasing" their initial footprints.

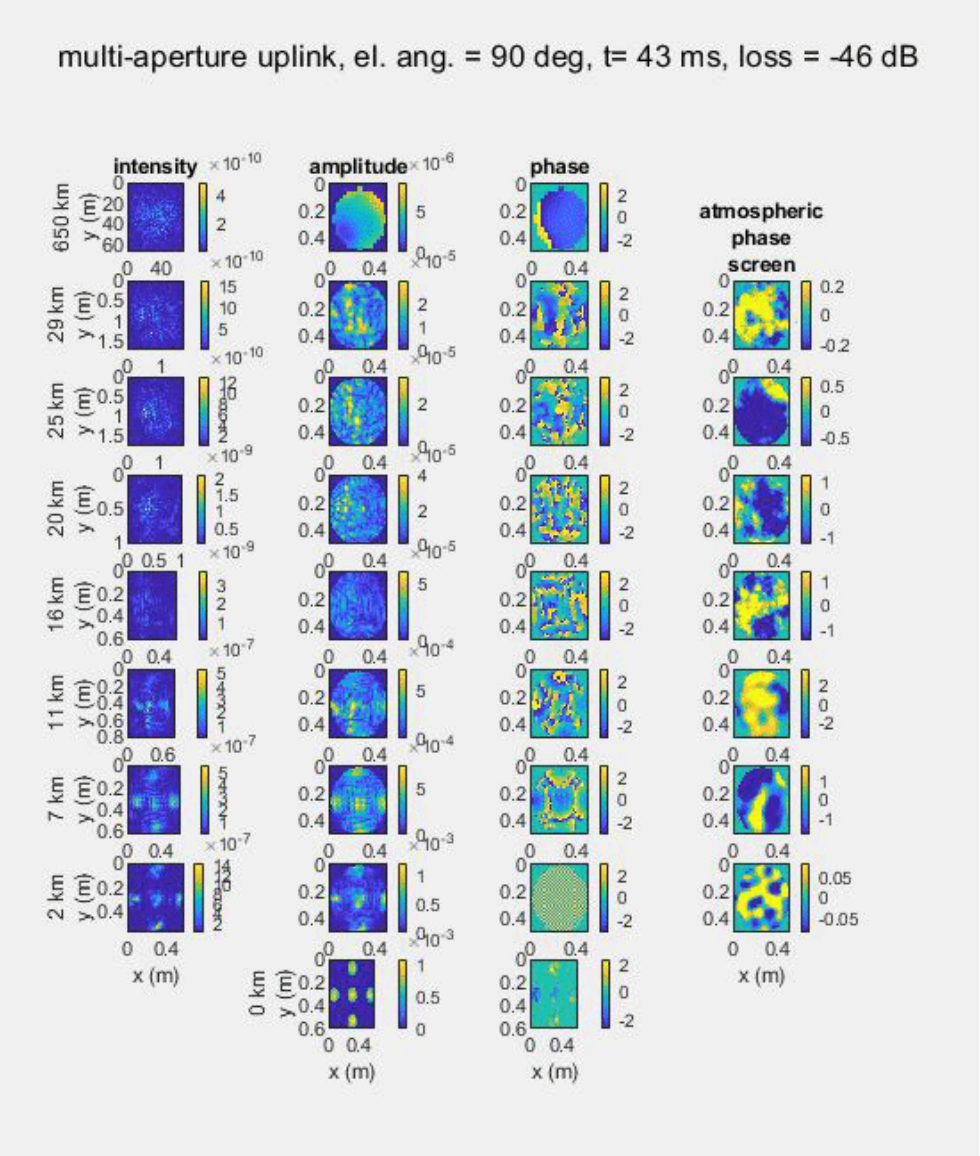


Figure A.15 A snapshot from the uplink multi-aperture movie, using five 12 cm OGR transmitters and a 50 cm monolithic receiver. To simuate incoherence, the beam from each OGR is separately propagated through the same atmospheric phase screen realization (with four of the five beams offset from the optical axis as shown) and summed at each atmospheric layer for illustration. The exit pupil intensity used to compute total loss is calculated using the sum of intensities from each beam propagation after accounting for T/T (§A.3.2.3) and multiplying by a binary mask. The figure is viewable as an animation in Adobe Reader version 7 or greater.

A.4 Results

Uplink and downlink simulation results are shown in Fig. A.16 for three different OGR diameters and secondary obscurations (all monolithic). Fig. A.17 further illustrates the relative contribution of diffraction, scintillation, and beam wander.

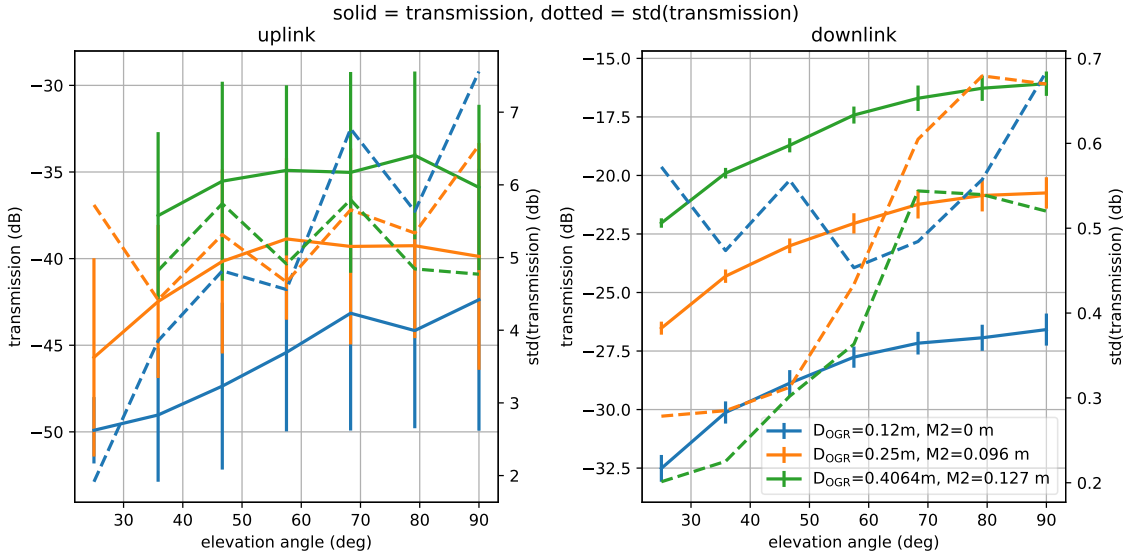


Figure A.16 Uplink (left) and downlink (right) mean atmospheric transmission and standard deviation for three different OGR diameters.

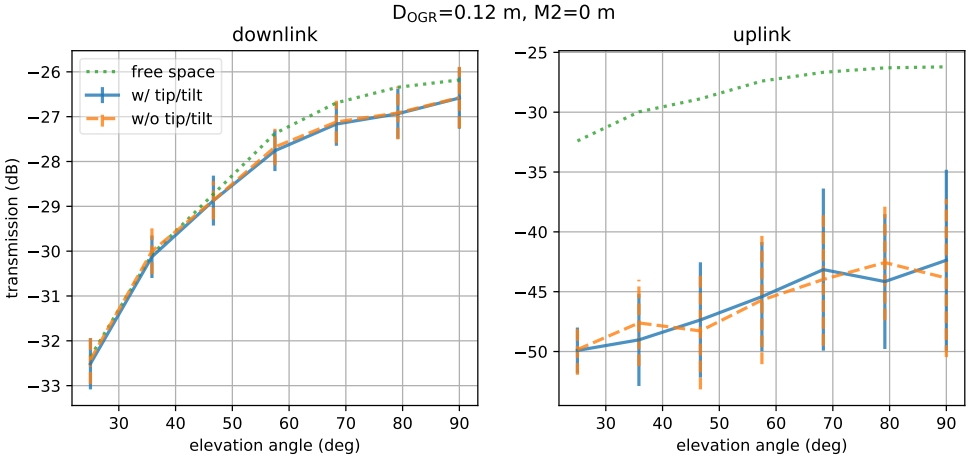


Figure A.17 Downlink (left) and uplink (right) mean atmospheric transmission and standard deviation, comparing results from propagation through free space, without T/T, and with T/T.

The uplink results in Fig. A.16 are not immediately intuitive why a larger diameter OGR produces less atmospheric loss than a smaller diameter OGR. Although the exit pupil of a larger OGR starts out as a larger beam, in the diffraction-limited case the Gaussian envelope expands at a shallower angle for the larger OGR. At around 100 km the beam from the larger OGR becomes smaller than the beam from the smaller OGR; by 650 km at the satellite, the Gaussian beam footprint from the smaller OGR should be larger than the beam from the larger OGR. If global tip and tilt were a dominant contribution to the losses in atmospheric transmission (still ignoring atmospheric turbulence from any other modes), the result from Fig. A.16 would be concerning: for the same effective atmospheric T/T angle, the wider "wings" of the Gaussian intensity distribution at the satellite from the smaller OGR should mean that more light is transmitted relative to the narrower "wings" from the larger OGR. However, Figure A.17 illustrates that the contribution of global atmospheric tip/tilt is negligible relative to the total transmission losses. Instead, this illustrates that most of the losses are from scintillation in higher order modes and that the intensity distribution remains relatively on-axis (i.e., relative to the beam size). Considering an on-axis footprint of a free-space Gaussian beam after propagating from the ground to the satellite, the footprint from the larger OGR should be smaller, and therefore the transmission should be better. The uplink results of Fig. A.16 show that this effect does not change with scintillation; in other words, atmospheric turbulence only amplifies the transmission loss behaviour expected from free-space propagation (but does not reverse it).

The downlink results from Fig. A.16 are more intuitively expected, and illustrate that a larger diameter OGR enables a better atmospheric throughput. As Fig. A.17 shows that the downlink losses are mostly from diffraction, this result is not surprising; the beam footprint at the ground is independent of OGR diameter, and so the larger OGR collects more light from the same beam.

Fig. A.18 shows the possible gains in standard deviation of transmission for a multi-aperture array OGR with five small telescopes (as described in §A.3.3) compared to a monolithic telescope with the same total collecting/transmitting area (i.e., $\sqrt{5}$ times larger in diameter than any one of the individual telescopes in the multi-aperture array). The effective area of an atmospheric cell is shown at the top of Fig. A.18 (i.e., corresponding to r_0= 16 cm). Although the middle and right panel show that a multi-aperture array has lower transmission than from a monolithic telescope with the same total collecting/transmitting area for the downlink/uplink,

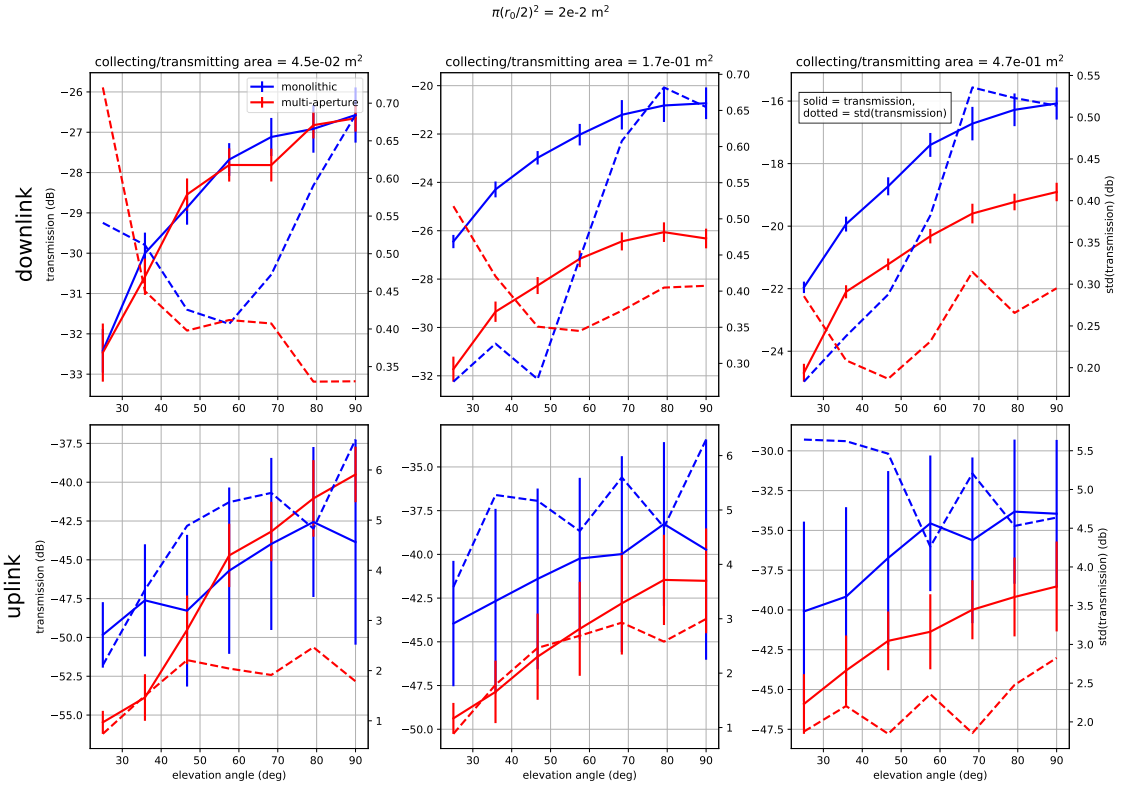


Figure A.18 Uplink and downlink transmission (solid lines) and standard deviation (dashed lines) for three different total collecting areas, comparing a monolithic OGR (blue) with a multi-aperture array OGR (red).

respectively, this is to be expected; the collecting area of each individual telescope in the array must be smaller than $\pi(r_0/2)^2$; this criteria is met for the left panel (i.e., $(4.5 \times 10^{-2})/5 = 0.9 \times 10^{-2} < 2 \times 10^{-2}$) but not the middle and right panels. Accordingly, the individual telescopes within the multi-aperture arrays in the middle and right panels are not "diffraction-limited," and therefore do not see the same gain as in the left panel. The monolithic telescopes for the middle and right panels typically collect more light than the multi-aperture arrays because of scintillation averaging over a larger aperture (i.e., for the smaller multi-aperture telescopes whose diameters are larger than r_0, scintillation diffracts more light outside of the aperture than for the corresponding monolithic case).

With that said, the left panel of Fig A.18 shows that if the multi-aperture array is designed to match the atmospheric conditions, standard deviation can improve by $\sim$0.5 dB for the downlink and $\sim$4 dB for the uplink. These gains could potentially

increase even further by going to larger total collecting areas (i.e., increasing the number of small telescopes in the array; see §A.6.3.2)).

A.5 Conclusions

My main findings from this work are as follows:

- Pre-compensated AO correction for the uplink of LEO satellites, using the downlink as a beacon for the WFS measurement, is not feasible to implement in a standard orbital configuration due to limitations from
  - a high angular velocity with respect to the OGR, creating effective windspeeds that would require unrealistic frame rates for an AO control system to keep up with, and
  - the PAA being significantly larger than the isoplanatic angle, preventing the downlink from enabling a WFS measurement for the same cone of turbulence that would need correcting for the uplink.
- I developed a MATLAB Fresnel simulation software to explore how atmospheric transmission for an uplink and/or downlink (without AO) depend on a wide variety of parameters, including basic design choices for the OGR, satellite, and atmospheric conditions. With these simulations I showed that
  - uplink losses from scintillation are significant, of order 20 dB beyond expected losses from free space propagation (with negligible losses from beam wander), while downlink losses remain mostly diffraction-limited (i.e., there is little to gain from an AO-corrected downlink),
  - a multi-aperture array that is designed to match atmospheric conditions can improve the standard deviation of link transmission by a few tenths of a dB for the downlink and a few dB for the uplink (but see §A.6.3.2) relative to a monolithic aperture with the same total collecting area.

A.6 Future Work

During the internship in which I completed this work, most of my time was spent developing and testing my atmospheric propagation code. Many additional higher level questions have yet to be answered, and are therefore summarized in this section.

A.6.1 Convergence

The convergence simulations in Figures A.11 and A.13 were done before the spatial scale analysis in Fig. A.12. The results from Fig. A.12 should have realistically informed the use of an outer scale less than ∼40 metres in the convergence simulations for number of layers and atmospheric "pupil diameter." These simulations should be repeated using an outer scale value less than 40 metres to confirm that the convergence behaviour is still the same. Intuitively, there is no reason to believe that a a phase screen with a smaller WFE and more realistic power law should change the conclusions that one layer is sufficient down to an elevation angle of 25 degrees and that the full image should be used to normalize WFE, but simulations should be redone to confirm this.

A.6.2 Explore Parameter Space

The dependence on how uplink and downlink atmospheric transmission are affected by additional parameters—including wavelength, satellite diameter, satellite and OGR static wavefront error, outer scale values (§A.6.2.1), and C_n^2 profile models (§A.6.2.2)—should be further explored.

A.6.2.1 Tip/Tilt and Higher Order Modes

In §A.3.2.3 I described the procedure for proper implementation of T/T in my simulations, using theoretical open loop coefficients from Noll (1976). These open loop coefficients are derived assuming a Kolmogorov power spectrum of turbulence (i.e., an infinite outer scale). Realistically, these coefficients should include the effects from a finite outer scale, which will lower the overall effect of beam wander. Open loop atmospheric coefficients of Zernike modes that include the impact from a finite outer scale are computed in Winker (1991).

Beyond just T/T, there are additional methods to generate higher fidelity phase screens with less numerical artifacts but that are more computationally expensive. One such method involves a linear combination of Zernike polynomials. Normalizing each Zernike mode to the coefficients expected from Noll (1976) or Winker (1991) before summation will then generate a Kolmogorov or von Kármán PSD, respectively, with the correct normalization of all Zernike modes by definition. Realistically, a truncation must be made at a certain point between higher order modes and infinity;

also at a certain point the covariance matrix between Zernike modes is no longer diagonal in Noll (1976) and Winker (1991), but both of these effects are small and probably negligible. Despite the results from my convergence tests in §A.3.2, this more robust method of phase screen generation should be investigated to see if it changes convergence/transmission results.

Additionally, further investigation/validation of the uplink results from Fig. A.16 is warranted. Considering my interpretation of these results presented in §A.4, if the beam radius of the larger OGR remains larger than the smaller OGR until crossing at 100 km, this means that the beam radius at each atmospheric layer (each of which lie below the 100 km crossing point) should be larger for the larger OGR, and that at the satellite the beam radius of the larger OGR should be smaller, *with* scintillation added. This validation would need to be performed on the long exposure intensity profiles at each layer and the satellite plane, since instantaneous scintillation would make it difficult to accurately measure a beam radius for a single time-step.

A.6.2.2 Site-dependent performance

Moving away from the Hufnagel-Valley model, it is important to understand the possible impacts of realistic weather conditions on the results of my simulations. A number of relevant papers presenting differential image motion monitor (DIMM), multi aperture scintillation sensor (MASS), and scintillation detection and ranging (SCIDAR) measurements describe more realistic estimates of C_n^2 profiles and L_0 values, including (M. Shoeck, private communication) Castro-Almazán et al. (2017); Lorenzo & J. Fuensalida (2011); García-Lorenzo & Fuensalida (2011); Guesalaga et al. (2016); Dali Ali et al. (2010); Ziad, A. et al. (2013); Ziad (2016); Maire et al. (2007). Although no DIMM, MASS, or SCIDAR measurements exist for any location in Canada, a number additional tests remain important, even without real profile measurements, outlined below:

- At astronomical observatory sites, what is the order of magnitude level of time variability of the C_n^2 profile (both in amplitude and in shape) and outer scales? Some rough, order of magnitude scaling to the Hufnagel-Valley C_n^2 profile that matches this level of temporal variability should be applied in my simulations (e.g., changing the A and or v_g parameters) to see what effect this has on transmission. In other words, even if measured C_n^2 profiles do not exist at a typical Canadian site to be used for optical communications, the relative extent

to which profile variability could change atmospheric transmission can still be understood.

- Use another C_n^2 profile other than Hufnagel-Valley to see how much transmission results can change from systematic assumptions about the atmospheric profile. Many other models exist; e.g., see (Hemani Kaushal, 2017, section 2.2, Figure 2.19, and Table 2.5)

A.6.3 Additional Multi-Aperture Analysis

A.6.3.1 Fibre Coupling

My analysis on atmospheric transmission for multi-aperture transmitters and receivers was only for propagation between the exit and entrance pupils. Ultimately the wavefront in the entrance pupil must be coupled into either a single or multi-mode fibre in the focal plane. This requires using the full entrance pupil wavefront to propagate to the focal plane. My computation of transmission only used intensity in the entrance pupil and did not consider the effects phase aberration on transmission. For the multi-mode case, transmission should be an aperture sum of the point spread function (including phase in the entrance pupil) over the size of the fiber core. For the single-mode case, fibre coupling will require use of the phase of the focal plane wavefront. In either case, gains in standard deviation by using a multi-aperture array should only increase, since scintillation effects in phase on spatial scales larger than r_0 are analogous to the same effects in amplitude.

A.6.3.2 Larger Arrays

Fig. A.18 showed that an array of five telescopes, each smaller than r_0, produced a similar mean atmospheric transmission but lower standard deviation, by about 0.5 and 4 dB for the downlink and uplink, respectively. Additional apertures should be added to increase the total collecting/transmitting area of the array while still keeping each aperture smaller than r_0 and being separated from one another by larger than r_0 in order to reach the total collecting area of the middle and right panels of Fig. A.18. Because standard deviation will continually increase with larger apertures due to increased scintillation (separate from the effect of mitigating scintillation losses from aperture averaging, which affects the mean), the discrepancy in standard deviation between the monolithic and multi-aperture OGR should increase with total

collecting/transmitting area (again, assuming that every telescope remains smaller than r_0 for the multi-aperture array, which is not the case in the middle and right panels of Fig. A.18). If so, a rough scaling relation could then be made between the improvement in standard deviation by going to a multi-aperture array as a function of total collecting/transmitting area, which may be a useful parameterization to use in the future if there are specific requirements set on standard deviation of atmospheric transmission.

A.6.4 Acquisition Tracking Sensor Simulations/Integration

In section §A.3.2.3 I described the implementation of open loop tip/tilt coefficients predicted by Noll (1976), and in §A.6.2.1 I proposed that this should be upgraded to the coefficients from Winker (1991), incorporating a finite outer scale. However, realistically, closed loop performance of the acquisition and tracking sensor (ATS) will govern/define the relative effects of beam wander for the uplink. The ATS is an AO system for T/T, and so simulating closed loop performance would require the following setup

1. Starting with an open loop downlink (i.e., before the ATS loop on the satellite is closed) received at the OGR, the OGR ATS will act as a WFS measurement for the required T/T correction of atmospheric turbulence for closed-loop downlink correction and pre-compensated open-loop uplink correction (the uplink is open loop for LEO satellites because of the large PAA as discussed in §A.2; see below).

2. In closed-loop, at minimum both the current and previous measurement must be used to compute the next T/T correction (i.e., the wavefront from multiple time steps must be saved). A T/T correction is computed using the measurements from step 1 over multiple time steps.

3. This correction is then applied both to the uplink (at the required PAA) and the downlink.

4. The same closed- and open-loop point-ahead procedure is also applied on the satellite ATS to the uplink and downlink, respectively.

Unfortunately, the same PAA problem discussed in §A.2 and illustrated in the upper right panel of Fig. A.5 applies to T/T correction: the OGR ATS cannot measure or correct atmospheric T/T for the uplink using the downlink as a WFS because

the PAA is larger than θ_0. With this in mind, closed-loop performance may actually *amplify* the effects of beam wander for the uplink and/or expected residual closed-loop WFE for the downlink (although this PAA problem is not an issue for the downlink T/T correction). Two amplitude effects may increase ATS-corrected WFE and/or decrease stability:

- For a diffraction-limited Gaussian beam from the downlink with a small amount of atmospheric T/T from phase aberration, the Gaussian beam footprint will have a non-zero T/T component in amplitude (i.e., just by sampling an off-axis diffraction-limited Gaussian beam).

- Scintillation (i.e., from phase aberration converting to amplitude aberration) may also have a non-zero T/T component in amplitude. Although this component traces the T/T originally from atmospheric phase aberration, the difference in response of WFS to amplitude vs. phase aberration will depend on the design/response of the sensor, in this case the ATS.

Using the above-described procedure to simulate these effects will help define requirements for ATS performance and closed loop stability due to scintillation effects from atmospheric turbulence.

A.6.5 AO Simulations/Integration

Apart from going to a GEO satellite setup, the first steps to continued investivation of AO performance can be evaluated with my existing PROPER code as follows:

1. Using the phase and amplitude in the entrance pupil from a standard open-loop downlink propagation, apply the negative of that phase along with a standard Gaussian beam in amplitude in the exit pupil for the uplink.

2. Propagate the uplink through the same atmospheric realizations used to simulated the downlink, translating each layer for the uplink accounting for only light travel time (i.e., assume an instantaneous measurement and open loop, unity gain correction is made at the OGR exit pupil). The light travel time should be of order τ_0.

3. See what happens during the uplink. Does this correction actually "cancel" atmospheric beam spreading effects from scintillation and/or T/T? If not, an AO

correction can only be made for the downlink (i.e., unless you have a separate AO system on the satellite for the uplink).

4. Repeat steps 1 - 3, but instead of applying the opposite phase from the downlink as in step 1 apply the opposite amplitude in addition to the Gaussian amplitude (i.e., assume use of a WFS/AO system that can measure/correct for amplitude aberration, e.g., Gerard et al. 2018b, §3.5.3).

Note that if this procedure does improve the uplink atmospheric transmission, further investivation/development is warrented; applying the negative of the WFS measurement is not the same as flattening the wavefront, and in classical AO terms would require a non-standard operation of re-defining the WFS offset slopes to new values at each iteration.

A.7 Overlap with Exoplanet Imaging Research

Although the work presented in this appendix was completed for a non-astronomical application, potential developments in related areas of optical communications that could benefit from and/or support advancements in exoplanet imaging include:

- **WFSs designed to measure amplitude aberration.** Even with a phase-only DM correction, amplitude aberration can be measured and corrected for in the focal plane. Fast focal plane wavefront sensing and control, e.g., as being developed by me, is thus a potential solution to enable scintillation correction in existing AO systems for optical communications, where these effects are much stronger than they are for astronomical observations.

- **RTC technology.** Even without AO, ATS RTCs are being developed for optical communications that could be advantageous to astronomical ExAO systems. Com Dev/Honeywell has been developing ATS systems for high temporal bandwidth fine pointing control for both optical communications and space-based astronomy (e.g. Rowlands et al., 2008); further applications to ground-based AO should be explored.

- **detector technology.** NIR optical communication receivers could potentially benefit from the high speed photon-counting detectors currently being developed for astronomical applications (e.g. Mazin et al., 2012; Goebel et al., 2018).