

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Finding exo-Earths and potentially life elsewhere: the need for extreme adaptive optics systems

Miguel Filipe Miranda dos Santos



Mestrado em Inteligência Artificial

Supervisor: Prof. Dr. Carlos M. Correia

Second Supervisor: Dra. Ana Filipa Sequeira

July 30, 2026

Finding exo-Earths and potentially life elsewhere: the need for extreme adaptive optics systems

Miguel Filipe Miranda dos Santos

Mestrado em Inteligência Artificial

Approved in oral examination by the committee:

President: Prof. Miriam Santos

Examiner: Prof. Jalo Nousiainen

Member: Prof. Carlos Correia

July 30, 2026

Resumo

O *Extremely Large Telescope* (ELT) de 40m, atualmente em construção, contará com o *Planetary Camera and Spectrograph* (PCS) para detetar e estudar exoplanetas, que precisará de um sistema de ótica adaptativa extrema (XAO) excepcionalmente rápido e robusto, com um número sem precedentes de graus de liberdade e uma correção por múltiplos espelhos ainda não demonstrada.

Este trabalho tem como objetivo melhorar o desempenho e a robustez da correção, ao mesmo tempo que relaxa os requisitos de controlo em tempo real do sistema XAO do PCS. Para tal, aplicamos o *Policy Optimization for Adaptive Optics* (PO4AO), que é um algoritmo de aprendizagem por reforço (RL), com adaptações específicas para melhorar a robustez e ter em conta o atraso inerente de dois *frames* do sistema. O algoritmo foi testado em duas arquiteturas candidatas para o controlo do PCS: o *woofer-tweeter*, que contém um sensor de frente de onda (WFS) e dois espelhos deformáveis (DMs), e o *cascade*, com dois ciclos em sequência, cada um com o seu próprio WFS e DM. O desempenho de cada esquema é avaliado para diversos níveis de complexidade do sistema, desde uma configuração reduzida num telescópio de 10m até uma simulação à escala do ELT, de forma a avaliar a escalabilidade da abordagem de RL.

As simulações demonstram que, na maior parte dos testes, a abordagem RL obtém um contraste igual ou superior ao de um integrador tradicional a operar ao dobro da frequência, com um fator de 2 a 6 em pequenas separações angulares. O melhor contraste é obtido com o RL na arquitetura *woofer-tweeter*, com um desempenho superior ao RL na arquitetura *cascade* por um fator de 1.5 a 3 em pequenas separações angulares.

Palavras-chave: Aprendizagem por Reforço • Óptica Adaptativa Extrema • *Extremely Large Telescope*

Abstract

The 40m Extremely Large Telescope (ELT), now under construction, will rely on the Planetary Camera and Spectrograph (PCS) to detect and study exoplanets, requiring an exceptionally fast and robust eXtreme Adaptive Optics (XAO) system featuring an unprecedented number of degrees of freedom with as-yet-unproven multi-mirror wavefront correction.

This work aims to improve correction performance and robustness while relaxing the real-time control requirements of the PCS's XAO system. To do this, we apply the Reinforcement Learning (RL) algorithm Policy Optimization for Adaptive Optics (PO4AO) with targeted adaptations to improve robustness and account for the system's inherent two-frame delay. We tested it with two candidate PCS control architectures: woofer–tweeter, featuring one Wavefront Sensor (WFS) and two Deformable Mirrors (DMs), and cascade, with two sequential loops, each with its own WFS and DM. The performance of each scheme is evaluated across a range of system complexities, from a scaled-down configuration on a 10 m-class telescope to a full 40 m ELT-sized simulation, in order to assess the scalability of the RL approach.

The simulations show that, in most tested cases, the RL controller achieves contrast equal to or better than a traditional integrator operating at twice the loop frequency, by a factor of 2 to 6 at small angular separations. The best overall contrast is obtained with RL in the woofer–tweeter architecture, exceeding RL in the cascade architecture by a factor of 1.5 to 3 at small angular separations.

Keywords: Reinforcement Learning • Extreme Adaptive Optics • Extremely Large Telescope

Acknowledgements

Gostaria de agradecer, em primeiro lugar, aos meus orientadores, o Doutor Carlos Correia e a Doutora Ana Sequeira, pela paciência com que tiveram e pelos ensinamentos que me deram para a realização de esta tese.

Agradeço aos meus irmãos e irmãs pelo apoio e pela ajuda que me prestaram. Aos meus amigos, que me acompanharam e proporcionaram memórias que tornaram esta etapa mais do que a conclusão de um mestrado.

E, acima de tudo, quero agradecer aos meus pais, que, apesar da distância, estiveram sempre presentes para me ajudar, ouvir e guiar, tornando este trabalho possível.

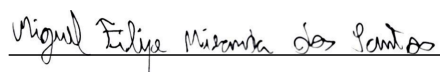
Miguel dos Santos

Declaração de Honra

Declaro que a presente dissertação é de minha autoria e não foi previamente apresentada e avaliada nesta ou em outra instituição. As referências a outros autores (afirmações, ideias, pensamentos), assim como a utilização de trabalhos publicados respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referência. Tenho consciência de que a prática de plágio ou de autoplágio constituem ilícitos acadêmicos, conforme estabelecido no Código de Ética da U.Porto.

Declaro, ainda, que utilizei ferramentas de Inteligência Artificial para a realização de parte(s) da presente dissertação, e essa utilização e os seus resultados estão devidamente assinalados e foram objeto de cuidada verificação para deteção de erros, imprecisões, atribuições infundadas ou outras formas de enviesamento de conteúdos e argumentos, bem como de determinação de autoria.

Miguel Filipe Miranda dos Santos



Contents

1	Introduction	1
2	Background and State of the art	3
2.1	Imaging Through Atmospheric Turbulence	3
2.2	Adaptive Optics	4
2.2.1	Wavefront Representation	5
2.2.2	Wavefront Sensors	6
2.2.3	Deformable Mirror	10
2.2.4	Real-Time Controller	11
2.2.5	Performance Metrics	14
2.2.6	Extreme Adaptive Optics	16
2.3	Brief Introduction to Artificial Intelligence	18
2.4	AI in AO	23
2.4.1	Wavefront Reconstruction	23
2.4.2	Wavefront Prediction	23
2.4.3	Control	24
2.4.4	Double Stage Control	26
2.5	Research Gap	26
3	Methodology	27
3.1	Algorithm Setup	29
3.1.1	Architecture Overview	29
3.1.2	Dynamics Model	29
3.1.3	Policy Model	30
3.1.4	Training Procedure	30
3.1.5	Algorithm Adaptations	31
3.1.6	Parameters	32
3.2	AO Control	33
3.3	Performance Metrics	35
4	Results	36
4.1	Single-Stage Results	36
4.2	Double-Stage Results	38
4.2.1	Scaled-down system	39
4.2.2	Intermediate-scale system	42
4.2.3	ELT-sized system	46
4.2.4	Supplementary Results	49

5	Discussion	50
5.1	Results Analysis	50
5.2	Real-World Application and Limitations	51
5.3	Challenges	51
6	Future work	53
	References	54
A	SPIE Poster	59
B	SPIE Proceedings	61

List of Figures

2.1	Typical adaptive optics system. Reproduced from [55].	5
2.2	Examples of phase-distortion representations. Images reproduced from [21] and [8], respectively.	6
2.3	Examples of wavefront sensors. Images reproduced from [1] and [49], respectively.	6
2.4	PyWFS detector intensity patterns from which the measurements in Eq. (2.4) are derived.	7
2.5	PyWFS and SH sensitivity. Extracted from [46].	8
2.6	PyWFS response curve, extracted from [56].	9
2.7	PyWFS with modulation from 0 to 5 and SH sensitivity curves, extracted from [12].	10
2.8	Deformable mirror. Image reproduced from [48].	11
2.9	Simplified temporal diagram of operations in Adaptive Optics (AO), extracted from [10].	13
2.10	Comparison of corrected (Strehl = 0.9) and uncorrected (Strehl < 0.1) PSF examples.	14
2.11	2D PSD example.	15
2.12	Comparison of PSF and PSD radial averages after scale alignment.	16
2.13	XAO system in action. Image reproduced from [23].	17
2.14	Woofer–tweeter control concept diagram for PCS.	17
2.15	Cascade control concept diagram for PCS.	18
2.16	Types of machine learning. Reproduced from [36].	19
2.17	Model-free versus model-based reinforcement learning. Reproduced from [26].	20
2.18	CNN. Reproduced from [41].	20
2.19	RNN. Reproduced from [2].	21
2.20	Autoencoder. Reproduced from [6].	21
2.21	U-Net. Reproduced from [53].	21
2.22	Transformer. Reproduced from [60].	22
2.23	Hybrid U-Net and reinforcement learning architecture [50].	25
3.1	Policy Optimization for Adaptive Optics (PO4AO) pseudocode, taken from [45].	31
3.2	Reconstruction quality of second stage modes by the first stage in the scaled down system.	34
4.1	Strehl ratio in the I band for the reinforcement learning (RL) and linear reconstructor (RC) approaches at a loop frequency of 500 Hz.	36
4.2	WFE for the RL and RC approaches at a loop frequency of 500 Hz.	37
4.3	Strehl ratio in the I band for the Reinforcement Learning (RL) and Reconstructor (RC) approaches at a loop frequency of 1 kHz.	37
4.4	WFE for the RL and RC approaches at a loop frequency of 1 kHz.	38

4.5	Contrast for the RL and RC approaches in the CAO configuration at the scaled-down system.	39
4.6	Contrast for the RL and RC approaches in the WT configuration at the scaled-down system.	40
4.7	Contrast comparison between the RC_WT and RC_CAO configurations at 2 kHz and 4 kHz for the scaled-down system.	40
4.8	Contrast of the RL approach for the WT vs. CAO configurations at the scaled-down system.	41
4.9	Contrast for all experimental configurations at the scaled-down system.	41
4.10	Ideal post-coronagraphic image for all experimental configurations at the scaled-down system.	42
4.11	Contrast for the RL and RC approaches in the WT configuration.	43
4.12	Contrast for the RL and RC approaches in the CAO configuration.	43
4.13	Contrast of the RC approach for the WT vs. CAO configurations.	44
4.14	Contrast of the RL approach for the WT vs. CAO configurations.	44
4.15	Contrast for all experimental configurations.	45
4.16	Ideal post-coronagraphic image for all experiments at the intermediate-scale system.	45
4.17	Contrast for the RL and RC approaches in the WT configuration at the ELT -sized system.	46
4.18	Contrast for the RL and RC approaches in the CAO configuration at the ELT -sized system.	46
4.19	Contrast of the RC approach for the WT vs. CAO configurations at the ELT -sized system.	47
4.20	Contrast of the RL approach for the WT vs. CAO configurations at ELT -sized system.	47
4.21	Contrast for all experimental configurations at the ELT -sized system.	48
4.22	Ideal post-coronagraphic image for all experiments at the ELT -sized system.	48
4.23	New contrast comparison for the CAO vs WT configurations at the ELT -sized system.	49

List of Tables

3.1	Setup Parameters	28
3.2	Inference time comparison	29
3.3	RL parameters	33
4.1	Strehl ratio results for RL and RC controllers.	38

List of Acronyms

AI	Artificial Intelligence
AO	Adaptive Optics
ANN	Artificial Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
CAO	Cascade Adaptive Optics
CNN	Convolutional Neural Network
DDPG	Deep Deterministic Policy Gradient
DM	Deformable Mirror
ELT	Extremely Large Telescope
FWHM	Full Width at Half-Maximum
GPI	Gemini Planet Imager
GHOST	GPU-based High-order adaptive OpticS Testbench
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
KL	Karhunen–Loève
LSTM	Long Short-Term Memory
ML	Machine Learning
NCPA	Non-Common Path Aberration
NIR	Near Infrared
OOPAO	Object-Oriented Python Adaptive Optics
OPD	Optical Path Difference
PCA	Principal Component Analysis
PO4AO	Policy Optimization for Adaptive Optics
PCS	Planetary Camera and Spectrograph
PSD	Power Spectral Density
PSF	Point Spread Function
PyWFS	Pyramid Wavefront Sensor
RC	Reconstructor
RL	Reinforcement Learning
RMS	Root Mean Square

RNN	Recurrent Neural Network
RTC	Real-Time Computer
SAC	Soft Actor–Critic
SCAO	Single-Conjugate Adaptive Optics
SH	Shack–Hartmann
SPHERE	Spectro-Polarimetric High-contrast Exoplanet Research instrument
SVD	Singular Value Decomposition
VLT	Very Large Telescope
WFE	Wavefront Error
WFS	Wavefront Sensor
WT	Woofer–Tweeter
XAO	eXtreme Adaptive Optics

Chapter 1

Introduction

Direct imaging of exoplanets remains one of the most demanding challenges in astronomy. This difficulty comes from the extreme brightness contrast between a planet and its host star, which can exceed 10^9 at visible wavelengths. In addition, the small angular separation between the two sources causes the starlight to dominate the image, obscuring the faint planetary signal.

For ground-based telescopes, this challenge is amplified by atmospheric turbulence, which distorts incoming wavefronts and spreads the starlight across the focal plane. As a result, the already weak planetary signal becomes further buried within the residual stellar glare. High-contrast imaging systems therefore rely on Adaptive Optics (AO) to correct these distortions in real time and restore diffraction-limited performance. However, conventional AO architectures are insufficient to meet the required levels of correction accuracy and temporal responsiveness expected for direct planet imaging.

To address this problem, the Extremely Large Telescope (ELT), which is currently under construction, will include a yet-to-be-developed Planetary Camera and Spectrograph (PCS) dedicated to the detection and characterization of exoplanets. To achieve this, it will employ a two-stage eXtreme Adaptive Optics (XAO) system with a very high-density actuator count at frame rates four times higher than the highest available today (i.e., up to 4 kHz). These specifications significantly exceed those of existing systems and introduce substantial technological challenges, as the required hardware does not yet exist. To address these limitations, recent research has explored the use of Artificial Intelligence (AI) for AO control. In particular, Reinforcement Learning (RL) has shown great promise, offering the ability to learn control policies directly from interaction with the system and to adapt continuously to changing observing conditions. However, most existing approaches apply to single-stage AO systems and do not account for the coupled dynamics inherent to multi-stage XAO architectures such as that of the PCS.

The objective of this thesis is to investigate the application of reinforcement learning to two-stage XAO systems, with the aim of improving both peak and average correction performance, in terms of Strehl ratio and contrast, while relaxing hardware and real-time control requirements. By enabling more efficient and adaptive AO control, this approach seeks to bridge the gap between current technological limitations and the performance demands of future high-contrast imaging

instruments such as the **PCS**.

The remainder of the work is structured as follows. Chapter 2 provides background on adaptive optics and artificial intelligence and reviews the state of the art in **AI**-based **AO** control. Chapter 3 presents the methodology developed in this work. Chapter 4 reports the results obtained, Chapter 5 discusses these results, and Chapter 6 outlines directions for future work.

Chapter 2

Background and State of the art

2.1 Imaging Through Atmospheric Turbulence

A question that has been asked for centuries is whether life exists on other planets. One of the most widely used methods to answer this question is to search for biosignatures through direct imaging of exoplanets and analysis of their spectral data, which may contain elements that support life. However, direct imaging is extremely challenging, as there are several factors that limit its capabilities. The first is that a planet's host star is much brighter than the planet itself, meaning that we need a telescope with extreme angular resolution, which represents the smallest angular separation at which two objects can still be distinguished.

Telescopes have a maximum angular resolution inherent to their design: for a telescope of diameter D , the diffraction-limited angular resolution is:

$$\theta_{\text{diff}} \approx \frac{\lambda}{D} \quad (2.1)$$

where λ is the observing wavelength. This happens because, as light is emitted from a star, we can trace its optical path using an imaginary surface on which every point has the same phase, called the wavefront. This wavefront is spherical, but given the long distances involved, only a small fraction of it is collected, so it can be considered locally planar.

When this wavefront encounters an obstacle or aperture, it undergoes an effect called diffraction, whereby it spreads outward from that point as if it were a spherical wave. To illustrate the diffraction effect produced by a circular aperture, we can consider the intensity pattern resulting from a point-like source, known as the Point Spread Function (**PSF**). Diffraction produces a pattern of rings at different radial distances from the center. The width of the central peak is characterized by the Full Width at Half-Maximum (**FWHM**), given by $\sim \lambda/D$, while the radius of the first dark ring is given by $1.22 \lambda/D$. It is this diffraction pattern that imposes a fundamental limit on the angular resolution, since the PSFs of two distinct points can only be resolved if the angular distance between them is at least λ/D .

This is one of the main motivations behind the construction of large telescopes, since a larger telescope can achieve a finer diffraction-limited angular resolution. Modern facilities range from

6 to 10 m in diameter, with the **ELT** planned to reach 39 m.

It is here, however, that the second major challenge for direct imaging arises: even with a large aperture, a telescope cannot operate at its diffraction limit in practice, due to atmospheric turbulence. As a wavefront propagates through the atmosphere, it passes through regions of varying refractive index, which locally bend and delay different parts of the wavefront; this effect is known as refraction. The atmospheric turbulence, often modelled using Kolmogorov theory [25], is composed of multiple layers, each with its own refractive index, mainly due to differences in temperature. Its principal characterizing parameter is the Fried parameter r_0 [15], which describes the spatial scale over which the wavefront phase remains coherent. At visible wavelengths, r_0 at good astronomical sites typically ranges from 10–20 cm, with:

$$r_0 \propto \lambda^{6/5} \quad (2.2)$$

meaning that turbulence effects are less severe at near-infrared wavelengths, which is a key reason most **AO** systems operate in the Near Infrared (**NIR**).

Over apertures smaller than r_0 , the wavefront phase varies by less than 1 rad Root Mean Square (**RMS**); over larger apertures, the phase variance grows according to Kolmogorov statistics. For $D \gg r_0$ (the regime of all modern 8–40 m telescopes), resolution is limited by turbulence to approximately λ/r_0 .

2.2 Adaptive Optics

Adaptive optics (**AO**) is a technique that aims to restore diffraction-limited performance by sensing and correcting wavefront distortions in real time.

In its simplest form, Single-Conjugate Adaptive Optics (**SCAO**), consists of [52]:

- **Wavefront Sensor (WFS)** that measures the distorted wavefront.
- **Deformable Mirror (DM)** that correct the wavefront by taking its inverse shape.
- **Real-Time Computer (RTC)** that computes **DM** commands from **WFS** measurements.

The **AO** system can operate in two different modes. The first is open-loop, where the **WFS** senses the full effect of the atmospheric turbulence and the **DM** only compensates for it. The second is closed-loop, where the **WFS** measures the residual after correction by the **DM**, and the resulting commands are fed back into the loop. The mode chosen for this thesis is closed-loop, as it generally provides better performance, since it can also sense errors introduced by the **DM** itself. Fig. 2.1 shows a typical closed-loop operation.

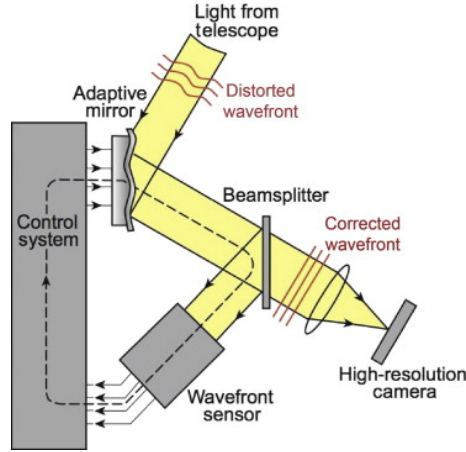


Figure 2.1: Typical adaptive optics system. Reproduced from [55].

2.2.1 Wavefront Representation

Since the wavefront phase cannot be measured directly, it can instead be represented in a finite modal basis, as in Eq. (2.3), with the most common being Zernike polynomials [42] (Fig. 2.2a) and Karhunen–Loève (KL) modes [11] (Fig. 2.2b).

$$W(r) = \sum_{i=1}^N c_i M_i(r) \quad (2.3)$$

Zernike polynomials form an orthogonal basis on a circular pupil and map to classical optical aberrations (tip/tilt, defocus, astigmatism, coma, etc.). They are intuitive and widely used for low-order calibration but are not statistically matched to atmospheric turbulence; high-order Zernike modes are strongly correlated under Kolmogorov statistics, making them a suboptimal basis for independent modal control.

Karhunen–Loève modes are the eigenvectors of the turbulence phase covariance matrix. They are uncorrelated and ordered by the amount of variance each mode captures.

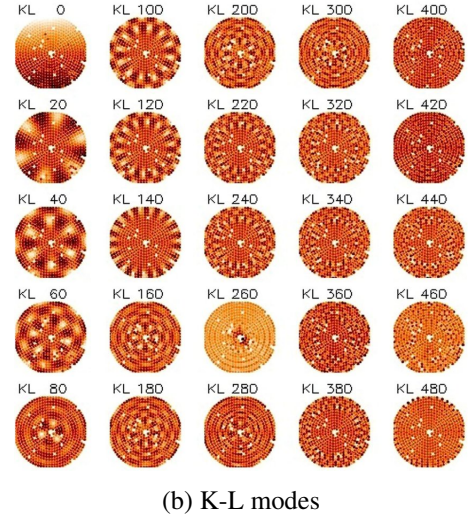
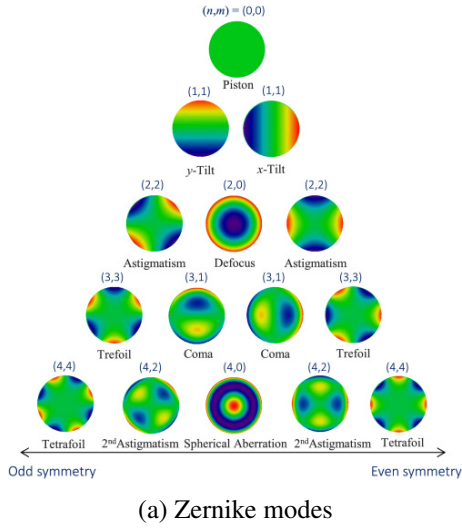


Figure 2.2: Examples of phase-distortion representations. Images reproduced from [21] and [8], respectively.

2.2.2 Wavefront Sensors

With the distortion defined, the next step is to measure it. As mentioned earlier, the wavefront phase cannot be directly observed; instead, sensors are used to extract information about the distortion, which is then used to reconstruct the wavefront.

The design of a **WFS** involves a trade-off between sensitivity (the responsiveness to changes in the wavefront), dynamic range (the span of aberration amplitudes the sensor can measure before saturating), and linearity of the response.

The two most widely used wavefront sensors are the Shack–Hartmann (**SH**) wavefront sensor and the Pyramid Wavefront Sensor (**PyWFS**) [51], depicted in Fig. 2.3.

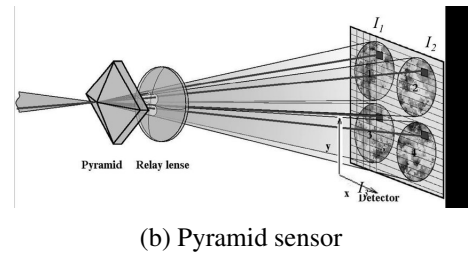
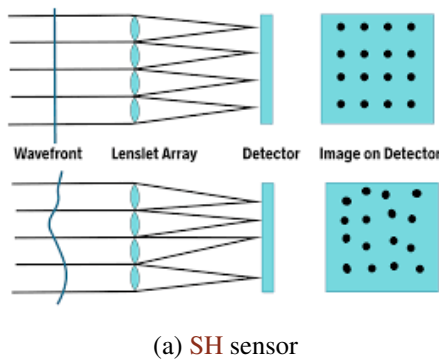


Figure 2.3: Examples of wavefront sensors. Images reproduced from [1] and [49], respectively.

The **SH** wavefront sensor consists of an array of small lenslets placed in front of a detector. Each lenslet focuses incoming light onto a spot on the detector. Deviations of these spots from their reference positions are used to estimate local wavefront slopes. By combining these slope

measurements, the complete wavefront can be reconstructed. **SH** sensors are simple, robust, and widely used; however, their sensitivity is limited, particularly when observing faint sources.

In its original design, the **PyWFS** employs a four-sided pyramidal prism placed at the focal plane of the telescope. The tip of the pyramid acts as a beam splitter: the incoming stellar light is divided into four separate beams, each directed toward a different face of the prism, producing four pupil images on the detector, as shown in Fig. 2.4. Wavefront aberrations are reconstructed by analyzing the intensity differences among these four pupil images [61], as defined in Eq. (2.4):

$$S_x = \frac{(I_1 + I_2) - (I_3 + I_4)}{I_1 + I_2 + I_3 + I_4}, \quad S_y = \frac{(I_1 + I_4) - (I_2 + I_3)}{I_1 + I_2 + I_3 + I_4} \quad (2.4)$$

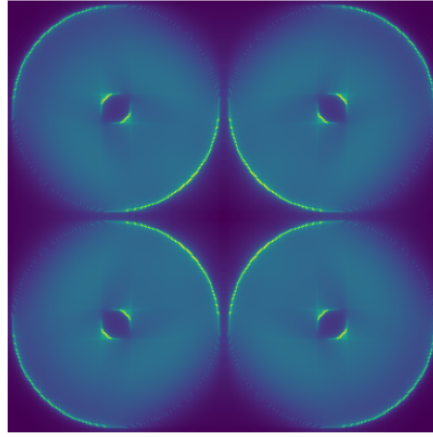


Figure 2.4: **PyWFS** detector intensity patterns from which the measurements in Eq. (2.4) are derived.

Compared to the classical **SH** wavefront sensor, the **PyWFS** offers significantly higher sensitivity, as shown in Fig. 2.5.

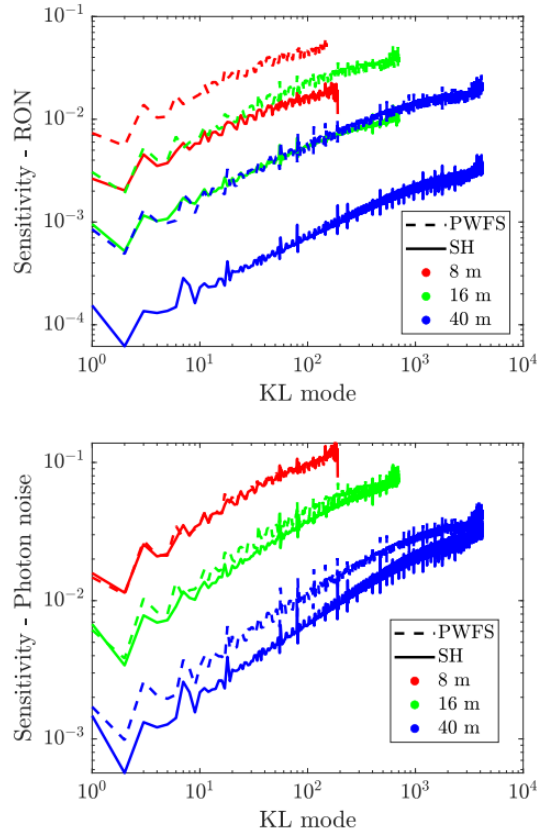


Figure 2.5: **PyWFS** and **SH** sensitivity. Extracted from [46].

In the small-aberration limit, the normalized intensity differences of Eq. (2.4) are proportional to the local wavefront slope, and the sensor response is approximately linear. However, this linear regime is only valid as long as the focal-plane, the image of the star, remains centered near the pyramid apex and illuminates all four faces simultaneously.

When wavefront aberrations are large, the focal-plane image shifts away from the apex and may fall predominantly on one or two faces of the pyramid rather than all four, leaving some pupils either over-illuminated or entirely deprived of light. In this regime the relationship between the measured intensity differences and the true wavefront slopes becomes strongly non-linear, and the reconstruction can fail or yield biased estimates.

As shown in Fig. 2.6, this non-linearity manifests as a deviation between the true wavefront amplitude and the sensor response: equal increments in input phase do not produce equal increments in the reconstruction.

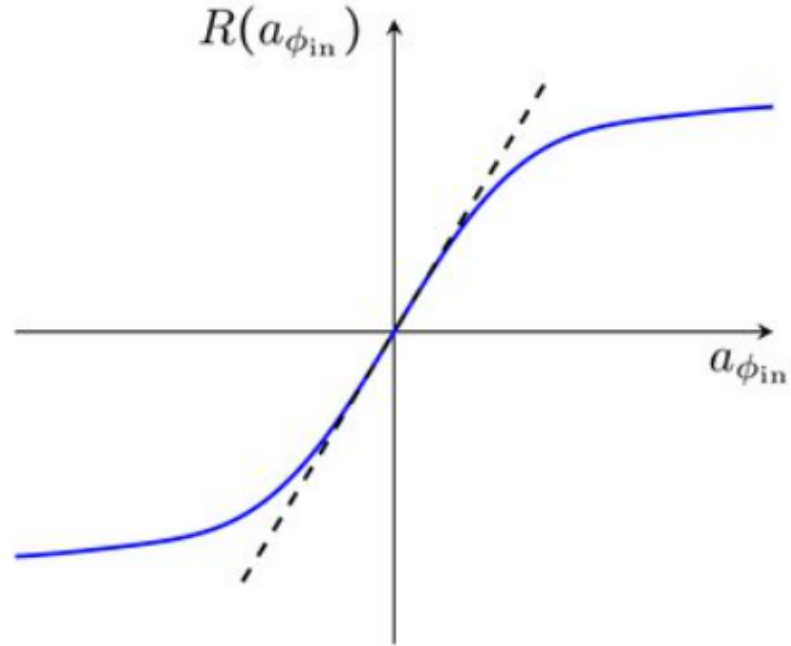


Figure 2.6: **PyWFS** response curve, extracted from [56].

A standard approach to mitigating this nonlinearity is tip-tilt modulation[3]: a fast steering mirror placed upstream of the pyramid introduces a circular motion around the pyramid apex, with a modulation radius r_{mod} typically expressed in units of λ/D . Over the course of each modulation cycle, the spot sweeps across all four faces, ensuring that every pupil receives a time-averaged flux proportional to the incoming wavefront. The sensor signal is then integrated over the full modulation period. The disadvantage of this approach is that it reduces the sensor sensitivity, as shown in Fig. 2.7.

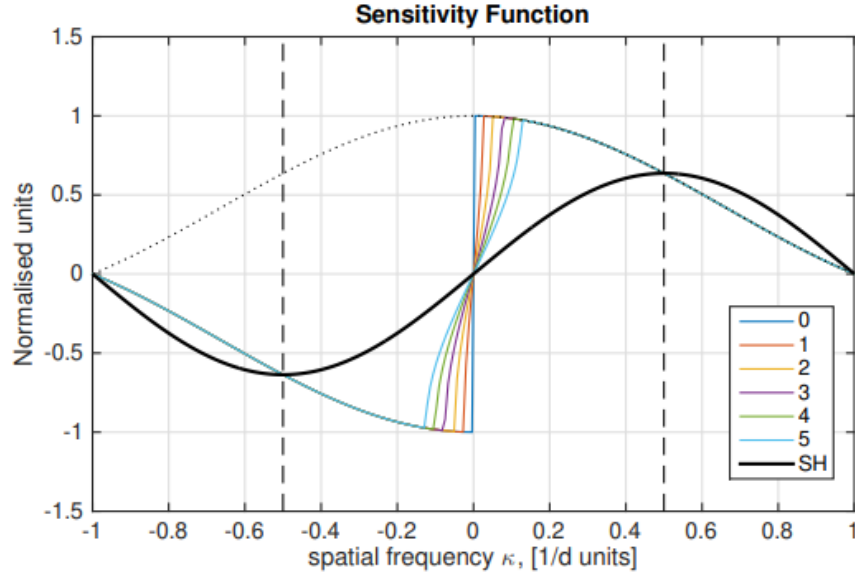


Figure 2.7: **PyWFS** with modulation from 0 to 5 and **SH** sensitivity curves, extracted from [12].

2.2.3 Deformable Mirror

Having described how to measure the distortion, the next step is to correct it. This can be done using a deformable mirror (shown in Fig. 2.8) that changes its shape to apply the opposite of the incoming distortion. Deformable mirrors are characterized by the following parameters [47] [35]:

- **Number of actuators:** The number of the degrees of freedom with which the mirror surface can be deformed.
- **Actuator pitch:** The spacing between actuators, which sets the spatial resolution of the correction. Pitch depends on where the **DM** sits in the optical path and the size of the pupil image.
- **Actuator stroke:** The maximum range of motion an actuator can provide, which sets the largest aberration amplitude the **DM** can correct. Stroke requirements depend on turbulence strength, plus extra margin for flattening the mirror, correcting telescope aberrations, and compensating vibrations.
- **Coupling:** The fraction of one actuator's stroke that affects its nearest neighbour. This is typically around 20–40%.
- **Response time:** How fast the **DM** can react to control commands, which sets the achievable **AO** loop frequency.

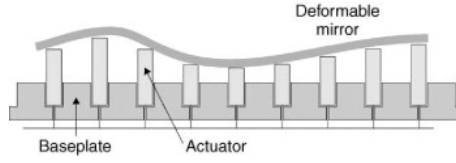


Figure 2.8: Deformable mirror. Image reproduced from [48].

2.2.4 Real-Time Controller

The **RTC** is the computer system responsible for driving the **AO** system. Its main functions consist of processing the measurement taken from the **WFS**, reconstruct the wavefront and compute the commands to send to the **DM**.

2.2.4.1 Reconstruction

Prior to observations, the **AO** system is illuminated by a calibration source (typically a laser), and the **DM** actuators are displaced individually to probe the system's response. By pushing and pulling the actuators with amplitude $\mathbf{A}$, the corresponding **WFS** slope measurements $\mathbf{s}_i$ (for actuator index i) are obtained, with push and pull values s^+ and s^- , respectively, as shown in Eq. (2.5).

$$\mathbf{s}_i = \frac{s^+ - s^-}{2\mathbf{A}} \quad (2.5)$$

Repeating this process for each actuator, one can gather a dataset of slopes, $\mathbf{s}$, and corresponding actuator values, $\mathbf{a}$. This establishes a linear relationship using the so-called interaction matrix, $\mathbf{D}$, as in Eq. (2.6).

$$\mathbf{s} = \mathbf{D}\mathbf{a} \quad (2.6)$$

If $\mathbf{D}$ were square and invertible, the wavefront representation could be obtained directly as $\mathbf{a} = \mathbf{D}^{-1}\mathbf{s}$. However, $\mathbf{D}$ is typically rectangular, since the number of wavefront sensor measurements usually differs from the number of **DM** actuators.

Furthermore, $\mathbf{D}$ is often ill-conditioned, containing modes that are poorly sensed or entirely unobservable (e.g., the “waffle” mode). We therefore compute the Moore–Penrose pseudoinverse $\mathbf{D}^+$ using Singular Value Decomposition (**SVD**), $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. This enables for a robust reconstruction by filtering out the smallest singular values, which corresponds to noise-dominated or unobservable modes, thereby preventing instability and ensuring that the control effort is focused only on reconstructible wavefront aberrations. The command matrix above is computed using a dataset obtained from zonal influence functions (IFs), i.e., poking one actuator at a time. Another possibility is to use a modal representation, in which the system response is calibrated and reconstructed directly in a basis of modes (e.g., Zernike or Karhunen–Loève modes) rather than individual actuator pokes. Modal reconstruction has the practical advantage of letting different modes be reconstructed, gained, and even excluded independently.

2.2.4.2 Control Law

Based on the reconstruction ε , the control law converts it into the commands applied to the **AO** system. For a closed-loop **AO** system, a common control law is the integrator. At timestep k , given a reconstruction ε_k , the command u_k is given by Eq. (2.7):

$$u_k = u_{k-1} - g \varepsilon_k \quad (2.7)$$

where $0 \leq g \leq g_{\max}$ is the loop gain. The maximum gain $g_{\max}$ is constrained by closed-loop stability. The gain controls the rate at which the correction is applied, and its optimal value depends on factors such as noise, delay, and the type of wavefront sensor. A lower gain, for instance, helps average out measurement noise at the cost of a slower response to genuine atmospheric changes. The basic integrator can be further improved by using modal control [16], assigning a different gain to each mode so that lower-order modes can be controlled more aggressively than noisier high-order ones.

The method above corresponds to an idealized scenario in which the reconstruction, command computation, and **DM** correction all occur instantaneously and at the same timestep. In practice, however, each of these operations takes a finite amount of time to complete, as illustrated by the chronogram in Fig. 2.9. A typical **AO** loop has an overall two-frame delay between the moment a wavefront distortion is sensed and the moment it is corrected, arising from two separate one-frame contributions.

The first frame of delay accounts for **WFS** integration and readout: the reconstruction at timestep k is computed not from measurements collected during frame k , but from the slope measurements obtained during the previous frame, s_{k-1} :

$$\varepsilon_k = R s_{k-1} \quad (2.8)$$

where R is the reconstruction matrix. The integrator in Eq. (2.7) then produces the command u_k from this reconstruction during frame k .

The second frame of delay accounts for **RTC** computation latency and the finite time required to apply the command to the **DM**: although u_k is computed during frame k , it is only physically applied to the **DM** during the following frame. Overall, this means that a wavefront distortion sensed during frame $k - 1$ is only corrected on the **DM** during frame $k + 1$, corresponding to a two-frame delay.

This pipelined structure means that a command is applied to the **DM** at every single frame, not only every other frame: while the command for frame $k + 1$ is being computed from ε_k , the previously computed command u_{k-1} is simultaneously being applied to the **DM**. This distinction is important, since if it were not the case, the loop would effectively operate at half its nominal frequency.

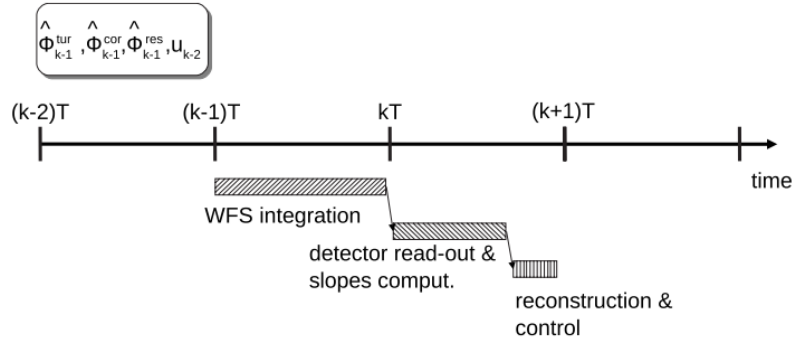


Figure 2.9: Simplified temporal diagram of operations in AO, extracted from [10].

All of these steps must be performed at high frequency, because another challenge posed by atmospheric turbulence is that it changes over time. The temporal evolution of turbulence is characterized by the Greenwood frequency f_G [17], which represents the control bandwidth required for effective AO correction and depends on wind velocity and the turbulence strength profile in the atmosphere. The corresponding Greenwood time constant, $\tau_G = 1/(2\pi f_G)$, defines the timescale over which atmospheric conditions remain correlated.

When the control loop operates much faster than f_G , atmospheric distortions change only slightly between measurements, and increasing the loop speed further yields diminishing returns. On the other hand, system performance degrades significantly when the response speed falls below the Greenwood frequency, since the distortions evolve faster than the system can adapt.

2.2.4.3 Error Budget

When designing an AO system, one usually builds an error budget, which encompasses the possible sources of error and is expressed as a sum of approximately independent residual phase variance terms. This provides a framework for estimating which errors can be addressed by the various components of the AO system, and for predicting overall performance before the system is built.

Total error budget. The total residual wavefront variance is commonly decomposed into approximately independent contributions:

$$\sigma_{\text{total}}^2 = \sigma_{\text{fitting}}^2 + \sigma_{\text{lag}}^2 + \sigma_{\text{noise}}^2 + \sigma_{\text{aliasing}}^2 + \sigma_{\text{cal}}^2 + \sigma_{\text{non-linear}}^2 + \sigma_{\text{other}}^2 \quad (2.9)$$

- $\sigma_{\text{fitting}}^2$: Error due to the finite number of actuators of the DM which makes it impossible to perfectly reconstruct the continuous wavefront shape.
- σ_{lag}^2 : The temporal error, related to the timing constraints of the RTC, appears due to the loop delay. The delay is the combination of several effects, such as the time taken to compute the commands or move the DM.
- σ_{noise}^2 : WFS measurement noise (photon and detector/read noise) propagated through the reconstructor.

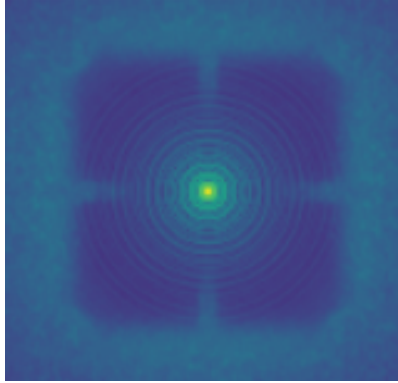
- $\sigma_{\text{aliasing}}^2$: Spatial frequencies above the **WFS** sampling Nyquist limit folding back into the measured low-order modes.
- σ_{cal}^2 : Static or slowly varying calibration errors, such as interaction-matrix mismatch.
- $\sigma_{\text{non-linear}}^2$: Error arising from non-linear effects in the sensing or actuation chain, such as centroid gain compression for large spot displacements or partially illuminated subapertures in a Shack–Hartmann **WFS**, or modulation-dependent optical gain in a pyramid **WFS**.
- σ_{other}^2 : Other errors, such as Non-Common Path Aberration (**NCPA**) or anisoplanatism, that are not taken into account in this thesis.

2.2.5 Performance Metrics

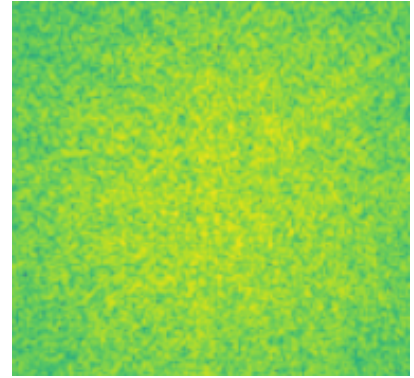
The Strehl ratio S is the primary metric for evaluating adaptive optics performance. It is defined as the ratio of the actual peak intensity of a point source image to the theoretical maximum intensity achievable with a perfect, diffraction-limited optical system.

In simpler terms, the Strehl ratio quantifies how close the corrected image is to the ideal diffraction-limited case. A Strehl ratio of 1 (or 100%) means perfect correction, while lower values indicate residual aberrations. For example, a Strehl ratio of 0.8 indicates that the image achieves 80% of the ideal peak intensity.

$$S_r = \frac{I_{\text{corrected}}(0,0)}{I_{\text{diffraction-limited}}(0,0)} \quad (2.10)$$



(a) Corrected PSF



(b) Uncorrected PSF

Figure 2.10: Comparison of corrected (Strehl = 0.9) and uncorrected (Strehl < 0.1) **PSF** examples.

For phase errors with $\sigma < 1$ radian rms (corresponding to a Strehl ratio of approximately > 0.3), the Strehl ratio can be approximated by the Maréchal approximation [52]:

$$S_r \approx e^{-\sigma^2} \quad (2.11)$$

where σ is the **RMS** of the phase error across the aperture.

Another important metric is contrast. Formally, contrast is defined as the ratio of the intensity at a given point in the focal-plane image to the peak intensity at the center of the host star:

$$C(\mathbf{r}) = \frac{I_{\text{corrected}}(\mathbf{r})}{I_{\text{corrected}}(0,0)} \quad (2.12)$$

where $I_{\text{corrected}}(\mathbf{r})$ is the intensity at position $\mathbf{r}$ in the focal plane. A contrast of 10^{-6} , for example, means that the target location is one million times fainter than the star, the typical regime in which giant exoplanets become detectable through direct imaging.

In an ideal scenario, a coronagraph is used to suppress the on-axis stellar contribution, so that the residual intensity in the image is dominated by astrophysical sources or instrument noise rather than the overwhelming glare of the star. In practice, however, no coronagraph achieves perfect suppression. Wavefront errors, introduced by optical surface imperfections, vibrations, or atmospheric turbulence, scatter light from the stellar **PSF** into the surrounding focal plane, creating speckles. These quasi-static speckle patterns mimic the spatial and photometric signature of a planetary companion, fundamentally limiting the achievable contrast. The suppression of speckle noise is therefore one of the central challenges in high-contrast imaging.

Power Spectral Density (**PSD**), as shown in Fig. 2.11, is a complementary metric: rather than quantifying the residual intensity at a single point or radial separation, the **PSD** decomposes the wavefront error map into its constituent spatial frequencies, revealing how much wavefront error power is associated with each characteristic spatial scale across the pupil. It is computed by taking the Fourier transform of the Optical Path Difference (**OPD**) map, which describes the deviation of the wavefront from a perfect reference at every point (x,y) across the pupil, and then taking the squared modulus of the result:

$$\text{PSD} = |\mathcal{F}\{\text{OPD}(x,y)\}|^2 \quad (2.13)$$

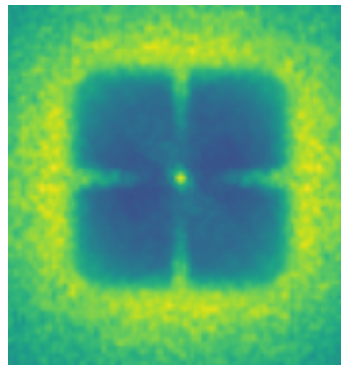


Figure 2.11: 2D **PSD** example.

The **PSD** provides a diagnostic of the errors affecting the **AO** reconstruction. By examining Fig. 2.11, several features can be identified: the butterfly-like pattern along the diagonal is a signature of servo-lag error. The cross-shaped feature is associated with the use of a modulated **PyWFS**

sensor. The radius of the corrected region is determined by the size of the **DM**, which sets the spatial limit of the **AO** control.

Physically, low spatial frequencies correspond to broad, slowly varying wavefront errors, such as residual tip-tilt or other low-order aberrations. High spatial frequencies correspond to fine-scale errors, such as actuator print-through on a deformable mirror.

Since the simulation used in this thesis does not include an actual coronagraph, we instead compute an approximation of the post-coronagraph contrast from the **PSD**:

$$C(\mathbf{r}) = \frac{PSD(\mathbf{r})}{I_{\text{corrected}}(0,0)} \quad (2.14)$$

The main reason behind this approximation is that, for a perfect coronagraph, which removes all coherent light, the main contribution to contrast degradation would be the residual wavefront errors [19]. As noted above, the **PSD** maps these residual errors, so in this idealized scenario contrast can be approximated from the **PSD**, although at a different scale. As shown in Figure 2.12, the residual **PSD** and the corrected **PSF** exhibit similar curves beyond the **DM** cutoff frequency.

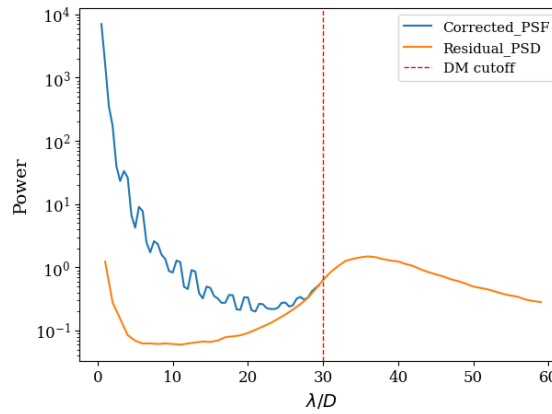


Figure 2.12: Comparison of **PSF** and **PSD** radial averages after scale alignment.

2.2.6 Extreme Adaptive Optics

Direct imaging of exoplanets requires extremely high contrast to detect faint companions in the close vicinity of a bright host star. In practice, this is extremely challenging because the stellar signal dominates the image, and even small residual wavefront errors can significantly reduce detectability.

For this reason, an **AO** system intended for exoplanet imaging must do more than maximize the Strehl ratio, it must also improve contrast at the angular separations where planets are expected to be found.

Compared to conventional **AO** systems, **XAO** features significantly faster control loops (around 1 to 4 kHz)[23], achieves much higher correction quality (Strehl ratios of approximately 90% versus 60% at visible wavelengths [37]), and employs more complex deformable mirrors, typically with more than 1,000 actuators.

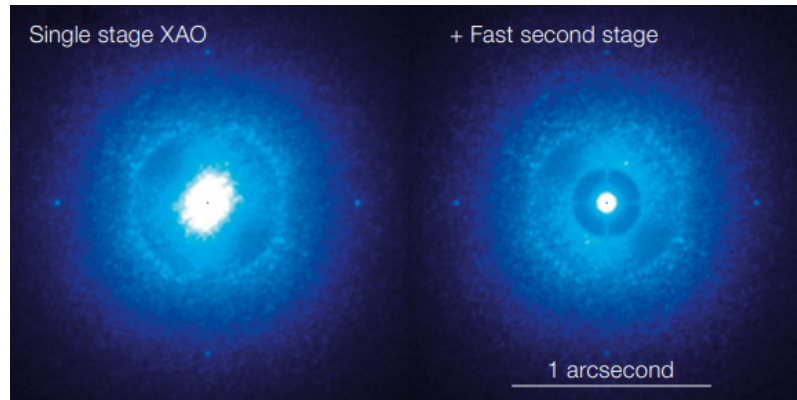


Figure 2.13: **XAO** system in action. Image reproduced from [23].

The first generation of instruments built around this concept are already operational. Spectro-Polarimetric High-contrast Exoplanet Research instrument (**SPHERE**)[5], installed at the Very Large Telescope (**VLT**), and Gemini Planet Imager (**GPI**)[34], both came online in the early 2010s and remain the two most productive facilities for direct exoplanet imaging to date. The first relies on a single high-order deformable mirror, while the second uses a woofer–tweeter configuration, driven at kHz-class rates, to deliver the wavefront quality a coronagraph needs. Building on this success, both instruments are now being upgraded to a second generation, **SPHERE+**[7] and **GPI 2.0**[9] respectively, with the goal of pushing contrast performance further and reaching closer to the host star.

The next step beyond these upgrades is a new instrument altogether: the Planetary Camera and Spectrograph (**PCS**), planned for the Extremely Large Telescope. **PCS** is designed to reach contrasts and inner working angles well beyond what **SPHERE+** or **GPI 2.0** can offer. For this reason, **PCS** adopts a two deformable mirror architecture.

Two ways of combining the two mirrors have been proposed for **PCS**: Woofer–Tweeter (**WT**) and Cascade Adaptive Optics (**CAO**).

In the **WT** configuration, represented in Fig. 2.14, the two mirrors work in parallel on the same measured wavefront, but split the correction by amplitude and spatial frequency. The woofer is a mirror with fewer actuators but a much larger stroke, and it absorbs the large, slowly varying distortions. The tweeter then only has to correct the smaller, high-order residual left behind.

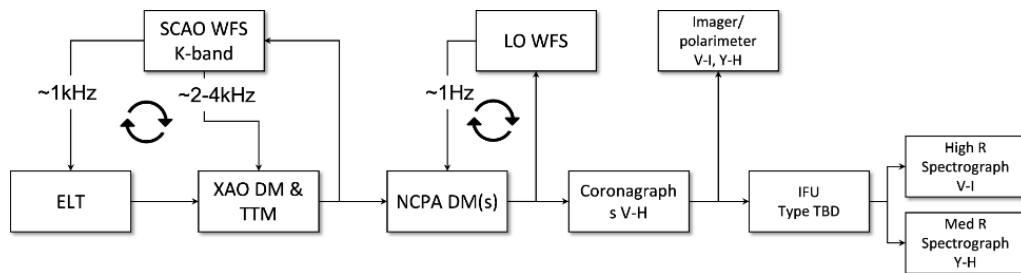


Figure 2.14: Woofer–tweeter control concept diagram for **PCS**.

In the **CAO** configuration, shown in Fig. 2.15, the two stages operate in series. The first stage performs a coarse correction using its own wavefront sensor, and a second wavefront sensor then measures only what remains afterward. The second mirror corrects this residual rather than the full original aberration. Because the second stage only has to handle what the previous one left behind, smaller aberrations that would otherwise stay buried under larger ones become detectable and correctable, enabling a more precise overall correction.

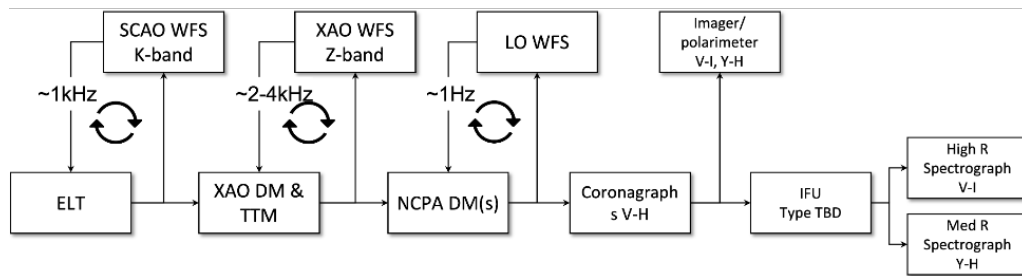


Figure 2.15: Cascade control concept diagram for **PCS**.

Both approaches achieve the same end goal, sharing the correction burden across two mirrors, but at the cost of a more complex system: more hardware to control, and more intricate control strategies to coordinate the stages. It is precisely this added complexity that makes **AI**-based control approaches promising for **XAO** systems such as **PCS**.

2.3 Brief Introduction to Artificial Intelligence

Artificial intelligence (**AI**) is a broad field within computer science focused on developing systems capable of performing tasks that typically require human intelligence. These tasks include perception, learning, reasoning, problem-solving, and decision-making. Unlike traditional computational approaches that rely on explicitly defined rules or deterministic algorithms, **AI** systems are designed to learn and adapt their behaviour based on data and experience.

This adaptive capability allows **AI** to recognize patterns, make predictions, and optimize decisions without requiring complete prior knowledge of the physical processes or environments in which they operate. Today, **AI** technologies have already achieved remarkable success across diverse domains, including image recognition, speech recognition, machine translation, autonomous driving, natural language processing, and strategic gameplay in complex board games.

At the core of modern **AI** lies Machine Learning (**ML**), a subfield that focuses on algorithms that can learn from and make predictions based on data. **ML** can be broadly divided into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning relies on datasets containing input-output pairs, where the correct output is known for each input. The goal is to learn a mapping that can generalize to unseen data, accurately predicting outputs for new inputs. Applications include regression, classification, and time series forecasting.

Unsupervised learning deals with unlabeled data. Its goal is to uncover hidden structures or relationships within the dataset, such as clusters, latent variables, or anomalies. Common techniques include clustering algorithms, Principal Component Analysis (PCA), and autoencoders.

Reinforcement learning (RL) involves an autonomous agent interacting with an environment. The agent learns by trial and error, receiving feedback through rewards or penalties, with the objective of maximizing long-term accumulated reward. RL has been instrumental in developing systems capable of mastering complex tasks such as robotic control and game-playing strategies.

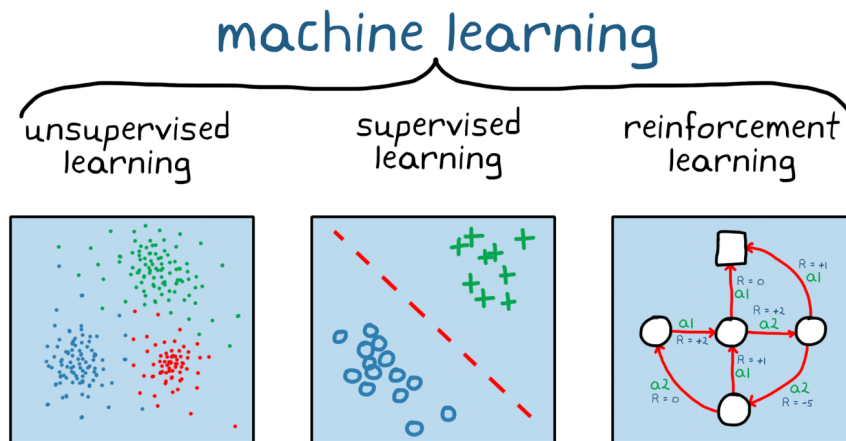


Figure 2.16: Types of machine learning. Reproduced from [36].

Reinforcement learning approaches can be categorized into two main paradigms: model-based and model-free methods. Model-based RL algorithms explicitly learn a state-space or dynamics model of the environment. Using this model and the currently measured state, they aim to find the control command that maximizes the optimization objective at each iteration. In contrast, model-free RL algorithms do not learn an explicit dynamics model. Instead, they directly learn control policies through trial and error. Actor-critic methods like Soft Actor-Critic (SAC) and Deep Deterministic Policy Gradient (DDPG) combine a policy network (actor) that selects actions with a value network (critic) that estimates expected returns, enabling gradient-based policy optimization.

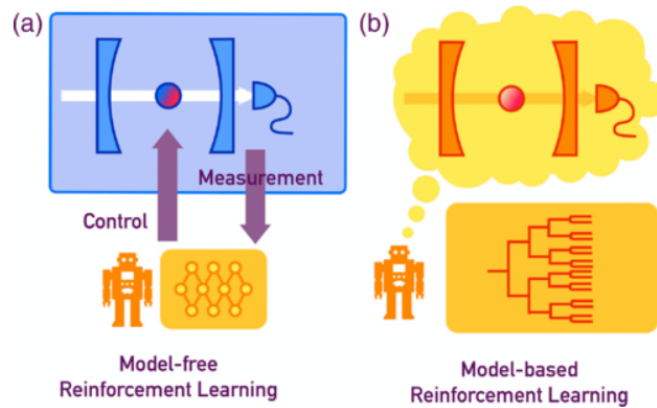


Figure 2.17: Model-free versus model-based reinforcement learning. Reproduced from [26].

A major breakthrough in AI has come from the development of Artificial Neural Networks (ANNs), computational models inspired by biological neural systems. ANNs consist of interconnected nodes (neurons) arranged in layers: an input layer, one or more hidden layers, and an output layer. Each neuron computes a weighted sum of its inputs, applies a nonlinear activation function, and passes the result to the next layer. During training, the network adjusts its internal parameters, weights and biases, to minimize prediction errors, typically using algorithms such as gradient descent and backpropagation.

Deep learning, a subset of machine learning, refers to neural networks with many hidden layers. These deep architectures can automatically extract hierarchical features from raw data, capturing complex and non-linear relationships. Depending on the task, specialized architectures have been developed:

Convolutional Neural Networks (CNNs) are particularly effective for spatial data such as images and videos, as they exploit local connectivity and shared weights to detect hierarchical visual patterns.

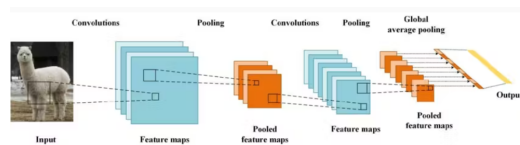


Figure 2.18: CNN. Reproduced from [41].

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are designed for sequential data such as speech or time series, capturing temporal dependencies.

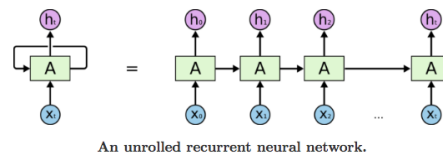


Figure 2.19: RNN. Reproduced from [2].

Autoencoders are unsupervised neural networks that learn efficient, compact representations of data by encoding inputs into a latent space and reconstructing them with the objective of minimizing reconstruction error. Variants like Variational Autoencoders (VAEs) are used in generative modeling and data synthesis.

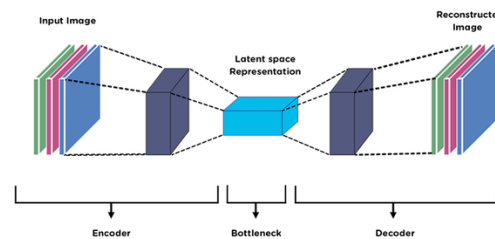


Figure 2.20: Autoencoder. Reproduced from [6].

U-Net, originally developed for biomedical image segmentation, employs an encoder–decoder structure with skip connections. The encoder compresses spatial information, while the decoder reconstructs detailed spatial maps, making it highly effective for pixel-level prediction tasks.

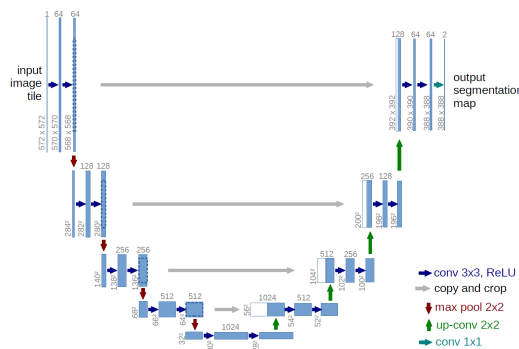


Figure 2.21: U-Net. Reproduced from [53].

Transformers, based on self-attention mechanisms, have revolutionized natural language processing (e.g., Generative Pre-trained Transformer (GPT), Bidirectional Encoder Representations from Transformers (BERT)) and are increasingly applied to computer vision and multimodal tasks due to their scalability and capacity to model long-range dependencies. Unlike RNNs, transformers process sequences in parallel and can directly attend to any position, making them highly efficient and effective at capturing long-range dependencies.

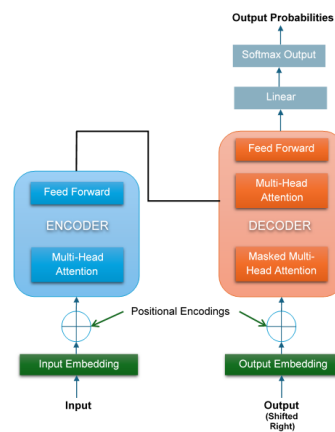


Figure 2.22: Transformer. Reproduced from [60].

Together, these developments have transformed **AI** into a rapidly evolving discipline capable of solving previously intractable problems across science, industry, and society. As computational power and data availability continue to expand, so too will the potential of **AI** systems to learn, reason, and augment human capabilities in increasingly sophisticated ways.

2.4 AI in AO

In this section, we review prior work on applying **AI** to adaptive optics, starting with wavefront reconstruction, followed by wavefront prediction. We then move to the main themes of this thesis: single and double-staged control.

2.4.1 Wavefront Reconstruction

The main motivation for applying **AI** to wavefront reconstruction is to improve the accuracy of the wavefront estimate, thereby providing better commands to the control loop. These approaches focus on reducing errors introduced by sensor noise, non-linearity, and aliasing.

Early attempts to exploit **ANNs** for adaptive optics (**AO**) date back to the early 1990s. Initial studies used in-focus and out-of-focus images to estimate tip-tilt aberrations [4, 32] or Zernike coefficients [54], using data from the Multiple Mirror Telescope [63]. These works established that a purely data-driven mapping from sensor data to wavefront coefficients is feasible, but they were restricted to low-order aberrations and small, hand-curated datasets.

Subsequent work explored classical machine learning techniques for wavefront sensor (**WFS**) processing. Montera et al. [38] and Lloyd-Hart et al. [31] reported promising results compared with traditional reconstruction methods, although their approaches were not deployed in on-sky systems. The common weakness across these methods is structural rather than a matter of insufficient training: fully connected architectures treat each subaperture measurement independently and therefore cannot exploit the strong spatial correlations between neighbouring subapertures that characterize atmospheric turbulence [57].

CNNs directly address this spatial-correlation limitation by preserving the 2D structure of the sensor data. DuBose et al. [13], building on Swanson et al. [58], introduced the Intensity/Slopes Network, which fuses **SH** slopes with raw subaperture intensities rather than discarding the latter. They showed that this fusion improves reconstruction accuracy specifically in the low-signal-to-noise regime where slope estimation is least reliable [18]. Weinberger et al. [62] took a complementary approach for unmodulated **PyWFS** data, where the central difficulty is not noise but the sensor's inherently non-linear, range-limited response. Their Visual Transformer architecture extends the usable dynamic range without sacrificing the high sensitivity that makes the unmodulated **PyWFS** attractive, trading the locality bias of **CNN** for the global attention mechanism better suited to a non-linear, non-local sensor response.

What these works do not yet address is the temporal dimension: even an accurate reconstructor estimates the wavefront at the time the **WFS** frame was captured, not at the time the **DM** command is applied. This delay is the subject of the next section.

2.4.2 Wavefront Prediction

This section focuses on studies that aim to reduce servo-lag error, mainly by predicting the future state of the wavefront. Wong et al. [63] employed feed-forward neural networks for this purpose,

demonstrating improvements over classical integrator controllers. To better capture the temporal dynamics of atmospheric turbulence, Hafeez et al. [20] and Liu et al. [30] introduced **LSTM** networks, which are better suited to modeling sequential dependencies.

Swanson et al. [57] extended neural-network-based prediction to closed-loop **AO** systems and identified an important challenge: a distribution mismatch can occur between the training data and the real-time observations encountered during closed-loop operation. To mitigate this effect, they introduced an adversarial prior during training [14].

The main limitation of these works is the gap between training on simulated data and deployment in real-world control. Reinforcement learning (**RL**) emerges as a promising direction to address this problem, since it can learn directly from online data and thereby mitigate this issue.

2.4.3 Control

This section covers works that learn directly from online data, which is the main feature that distinguishes them from the methods discussed previously. These approaches are particularly well suited to correcting servo-lag and calibration errors, owing to their ability to learn from and adapt to real-time data; they can also help reduce non-linearity, noise, and aliasing errors. Reinforcement learning (**RL**) has emerged as a promising framework for closed-loop **AO** control. Unlike classical control techniques, which typically rely on quadratic cost functions such as mean squared error, **RL** enables direct optimization of performance metrics [14].

RL approaches can be broadly categorized into model-based and model-free methods.

Model-free **RL** methods learn control policies directly from interaction data without an explicit system model. For example, Landman et al. [28] applied Deep Deterministic Policy Gradient (**DDPG**) to optimize an **RNN** architecture, enabling adaptation to changing conditions without online retraining and reducing errors caused by vibrations. This work was initially limited to tip-tilt control, but subsequent work [29] extended it to a high-order **DM**. The drawback of this approach is the complexity of implementing the **RNN** architecture on an actual **RTC**.

Pou et al. [39] took a hybrid approach combining control and reconstruction, as shown in Fig. 2.23. Their method uses a U-Net model, trained on previously collected data, that takes the raw intensity fields from the pyramid sensor and combines this reconstruction with that of a linear reconstructor. The combined reconstruction is then passed to a model-free reinforcement learning agent using Soft Actor-Critic (**SAC**) in a multi-agent framework that is trained online.

This approach demonstrates robustness across different seeing conditions and data mismatch, as well as the ability to control the unmodulated pyramid sensor. The main drawback is the need for previously collected data to train the U-Net model, along with the fact that performing reconstruction with the U-Net can be too time-consuming for closed-loop operation at higher frequencies.

Overall model-free methods generally offer faster inference but require a large number of interactions with the environment to converge, which can limit their applicability to real-world systems.

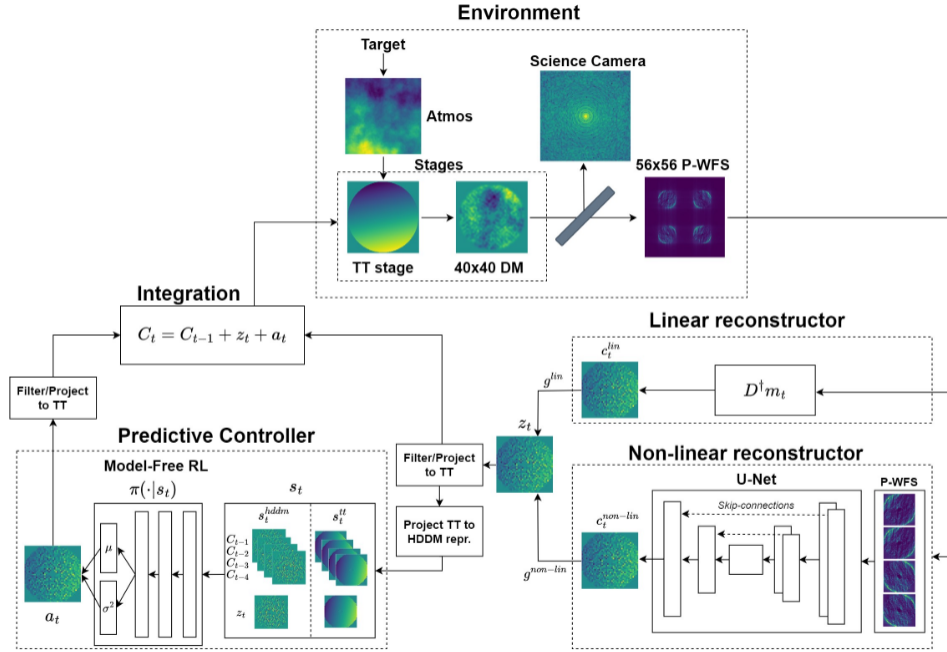


Figure 2.23: Hybrid U-Net and reinforcement learning architecture [50].

As for model-based methods, model-based **RL** learns an explicit representation of the system dynamics and uses it for planning and control. Nousiainen et al. [43] applied model-based **RL** to predict the temporal evolution of turbulence and to compensate for misregistration between the **DM** and the sensor. Although model-based methods typically require fewer interactions with the environment compared to model-free methods, they are computationally demanding and sensitive to modelling errors.

A follow-up proposal combined these two methods into a hybrid approach, as in the work of Nousiainen et al. [45], which combines a model-free (policy) component with a model-based (dynamics) component. The main idea is to use the dynamics model to learn a representation of the environment, and to use this representation to train the policy model. This reduces the number of environment interactions required to train the model-free component, while still allowing fast inference for use in the control loop. The main drawback is that, although this approach is shown to be robust to distribution mismatch and can control the unmodulated **PyWFS**, robustness across different seeing conditions is not demonstrated.

These works show promising control for single-stage or simple multi-mirror setups (using an extra tip-tilt mirror), but they do not demonstrate applicability to future instruments such as the **PCS**, which will use more than one complex deformable mirror. This aspect is addressed in the next section.

2.4.4 Double Stage Control

At the time of writing, only a few works address this aspect; we therefore divide them according to architecture concept, namely Woofer–Tweeter (**WT**) and Cascade Adaptive Optics (**CAO**).

2.4.4.1 Cascade Control

Nousiainen et al. [44] apply their method to control the second stage of an **XAO** system on the GPU-based High-order adaptive OpticS Testbench (**GHOST**) testbench. The first stage uses a 41×41 **DM** running at a loop frequency of 1000 Hz, while the faster second stage (2000 Hz) is simulated with a 24×24 **DM** on an 8 m telescope. During operation, the integrator controls the first stage while the **RL** agent controls only the second. The approach showed improvement over the integrator in terms of correcting noise errors.

2.4.4.2 Woofer–Tweeter Control

Krokberg et al. [27] apply the method of Nousiainen et al. [45] to control a woofer–tweeter configuration on a 6.5 m telescope, with a 12×12 **DM** for the first stage and a 48×48 **DM** for the second stage (these values are estimated from the number of active actuators, since the number across the diameter is not stated). The first stage runs at 1000 Hz and is controlled by an integrator, while the second runs at 2000 Hz and is controlled by the **RL** agent. Its performance is compared against an integrator controlling only the tweeter stage, and it shows improvement over this baseline.

2.5 Research Gap

Although these works begin to demonstrate the applicability of reinforcement learning to multi-mirror control, their scope remains limited to relatively simple implementations and does not account for the scalability required by the **PCS** instrument.

This thesis therefore applies reinforcement learning to control two double-staged architectures, namely Cascade Adaptive Optics and Woofer–Tweeter, described in Fig. 2.15 and Fig. 2.14, respectively, across three experimental scales ranging from a scaled-down version to an **ELT**-sized architecture, with the aim of evaluating control performance and scalability.

Chapter 3

Methodology

To perform the simulations required for this thesis, we used the Object-Oriented Python Adaptive Optics (OOPAO) framework [59]. OOPAO is an open-source, Python-based, end-to-end adaptive optics (AO) simulation tool currently under active development. It models the full optical path, from atmospheric turbulence through wavefront sensing and deformable mirror (DM) correction to scientific image formation.

With this setup, two types of experiments were conducted. The first consisted of a single-stage AO loop using a 20-meter telescope and a PyWFS with a modulation of $2\lambda/D$. Initially, an unmodulated PyWFS was considered; however, the algorithm proved to be too unstable under those conditions. This setup was used to evaluate the robustness of the model under varying observing conditions. Specifically, simulations were performed for different guide star magnitudes (0 and 9), Fried parameters r_0 (0.08 m and 0.15 m), and loop frequencies (500 Hz and 1000 Hz).

The second experiment used two-stage AO architectures, where in all cases, a guide star with a magnitude of 5 was used. The first stage consisted of a lower-density deformable mirror, while the second stage used a higher-density mirror. The complete set of parameters is provided in Table 3.1. For this part, we tested the algorithm at three different scales of the ELT: a scaled-down version, an intermediate-scaled version, and a ELT-sized version.

The scaled-down version was used to validate the initial approach and establish baseline performance comparisons, benefiting from its lower computational cost and faster simulation times. The intermediate-scale version served to test initial scalability and identify any issues arising from the increase in system size. The full ELT-scale version was used to assess the algorithm's performance under the full complexity of a PCS implementation.

One caveat about the ELT-scale configuration. Unlike the smaller-scale versions, where the number of wavefront sensor subapertures was matched to the DM actuator grid (e.g. 40×40 subapertures for a 41×41 actuator DM), the first-stage sensor at ELT scale used 120×120 subapertures, the same as the second stage, rather than the 80×80 that would correspond to the 81×81 DM. This was a consequence of a simulation constraint: the telescope pupil resolution must be at least twice the number of subapertures of each sensor and must also be evenly divisible by that number. An configuration with 81×81 and 121×121 DMs would have required a pupil

resolution of 480 pixels, which proved too computationally expensive. The resolution was therefore reduced to 240 pixels, which in turn required increasing the first-stage subaperture count to 120×120 to satisfy the divisibility constraint.

Table 3.1: Setup Parameters

Telescope		
Parameter	Value	Units
Telescope diameter	10/20/40	m
Obstruction ratio	14/14/30	percent
Sampling frequency, first stage	1000	Hz
Active actuators, first stage	356/1332/4788	actuators
Active actuators, second stage	768/2920/10612	actuators
PWFS subapertures, first stage	$20 \times 20/40 \times 40/120 \times 120$	apertures
PWFS subapertures, second stage	$30 \times 30/60 \times 60/120 \times 120$	apertures
PWFS modulation	2/2/4	λ/D
DM coupling	0.3	percent
DM influence functions	Gaussian	–
WFS wavelength	1.7	μm
Science camera wavelength	0.79	μm

Atmosphere parameters		
Parameter	Value	Units
Fried parameter	15	cm at 500 nm
Number of layers	3	–
Layer altitudes	0/4/10	km
C_n^2	50/35/15	percent (%)
Wind speeds	10/26/35	m/s
Wind directions	0/45/180	degrees
L_0	25	m

The overall architecture of the two-stage system consists of a first-stage **AO** loop and a second-stage **XAO** loop correcting at a higher temporal frequency ($2\times$ to $4\times$ faster than the first stage).

In this case, two control architectures were tested:

- **Cascade control:** each stage has its own wavefront sensor and independent integrator.
- **Woofers–tweeters control:** a single wavefront sensor is used; low-frequency components are offloaded from the second stage to the first stage via an interaction-matrix-based mapping.

In the woofers–tweeters approach, actuator commands from the second-stage **DM** are projected into a modal basis and mapped onto equivalent modes of the first-stage **DM** using an interaction

matrix. These modal coefficients are converted into voltages and applied to the first-stage **DM**. Since the second stage operates at a higher frequency, offloading is performed only when the first stage updates.

3.1 Algorithm Setup

The choice of algorithm was guided by the main objectives of this work: controlling very dense deformable mirrors, meeting the fast inference requirements imposed by the loop frequency, and reducing the mismatch between training data and the application environment. **RL**-based control is the class of methods that best satisfies all three requirements. Within this category, we considered U-Net+RL from Pou et al. [39] and Policy Optimization for Adaptive Optics (**PO4AO**) from Nousiainen et al. [45].

Both U-Net+RL and **PO4AO** showed comparable Strehl performance and demonstrated potential for multi-mirror control: **PO4AO** has already been applied to two different two-stage configurations, while U-Net+RL splits control between two mirrors through its multi-agent architecture. Both methods were also robust to training-application data mismatch and capable of operating with a non-modulated **PyWFS**. The deciding factor between the two was therefore the scalability of inference time with telescope size.

Table 3.2: Inference time comparison

Method	VLT	ELT
U-Net+RL	0.414 ms	1.21 ms
PO4AO	0.30 ms	0.37 ms

As shown in Table 3.2, scaling from an 8-meter telescope **VLT** to a 40-meter telescope (**ELT**) increases the inference time of U-Net+RL by almost a factor of three, while **PO4AO**’s inference time remains nearly constant. Based on this scalability, the reinforcement learning algorithm used to control the loop was adapted from **PO4AO**, with the original code publicly available at [22].

3.1.1 Architecture Overview

PO4AO combines model-based and model-free reinforcement learning in a two-component architecture consisting of a dynamics model and a policy model. This hybrid approach exploits the complementary strengths of both paradigms.

3.1.2 Dynamics Model

The dynamics model is a model-based **RL** component implemented as an ensemble of **CNN**, each with leaky ReLU activation functions [33].

Given the current state s_t and action a_t , the dynamics model $\hat{p}_\omega(s_t, a_t)$ predicts the next state s_{t+1} .

The state is represented in actuator space as a 3D tensor of shape $(N \times N \times (k + m))$, where:

- $N \times N$ is the **DM** actuator grid,
- k is the number of previous states,
- m is the number of previous actions.

The model parameters ω , namely the neural network weights and biases, are optimised using the Adam optimiser [24].

To prevent the well-known problem of model overfitting in the early stages of training [40], an ensemble of dynamics models is employed. Each model in the ensemble is trained on a different bootstrap subset of the collected observations, leading to diverse approximations of the environment dynamics. During policy optimisation, predictions are averaged across all ensemble members, providing a more robust estimate of the next state[45].

3.1.3 Policy Model

The policy is a model-free component, also implemented as a fully convolutional neural network with the same architectural design as the dynamics model. It takes the current state as input, represented by the 3D tensor described above, and outputs a 2D tensor representing the **DM** actuator voltages to be applied.

Since the **WFS** measurement is insensitive to certain **DM** shapes, such as the well-known waffle mode and boundary actuator modes, a projection step is applied to ensure that these uncontrolled modes are excluded from the control commands. Specifically, the 2D output of the policy network is reshaped into a vector, multiplied by a modal filter matrix that projects it onto the controllable subspace, and then reshaped back into a 2D actuator map before being sent to the **DM**. [45]

3.1.4 Training Procedure

The policy is optimised by sampling initial states from previously collected observations, computing actions using the current policy, and rolling out the dynamics model to simulate the resulting state trajectory up to a planning horizon h . The cumulative predicted loss over this horizon is then backpropagated to update the policy parameters, without requiring any additional real environment interactions during this phase.

Algorithm 1 Policy Optimization for Adaptive Optics (PO4AO)

```

1: Initialize policy and dynamics model parameters  $\theta$  and  $\omega$ 
   randomly
2: Initialize gradient iteration length  $K$ , batch size  $B < |\mathcal{D}|$  and
   planning horizon  $H$ 
3: while not converged do
4:   Generate samples  $\{s_{t+1}, s_t, a_t\}$  by running policy  $\pi_\theta(a_t|s_t)$ 
   for  $T$  timesteps (an episode) and append to  $\mathcal{D}$ 
5:   Fit dynamics by minimizing Eq. (9) w.r.t  $\omega$  using Adam
6:   for iteration  $k = 1$  to  $K$  do
7:     Sample a mini batch of  $B < |\mathcal{D}|$  states  $\{s_\tau\}$  from  $\mathcal{D}$ 
8:     for each  $s_\tau$  in the mini batch do
9:       Set  $\tilde{s}_1^\tau = s_\tau$ 
10:      for  $t = 1$  to  $H$  do
11:        Predict  $a_t = \pi_\theta(s_t)$ 
12:        Predict  $s_{t+1} = \hat{p}_\omega(s_t, a_t)$ 
13:        Calculate  $R_t = \hat{r}_\omega(s_t, a_t)$ 
14:      end for
15:    end for
16:    Update  $\theta$  by taking a gradient step according to
     $\nabla_\theta \sum_{t=\tau}^{\tau+H} R_t$  with Adam.
17:  end for
18: end while

```

Figure 3.1: PO4AO pseudocode, taken from [45].

A full experiment proceeds as:

1. Run a noisy integrator to collect initial data. Noise is deliberately injected to increase sample diversity.
2. Train the dynamics model on the collected data.
3. Use the dynamics model to optimize the policy model.
4. During an episode, collect additional data so that both models can be trained in parallel at the episode end.
5. Repeat steps 2 – 4 until the experiment ends.

This architecture combines the strengths of both reinforcement learning approaches: the policy, being model-free, is lightweight and responds at loop frequency without the computational overhead of model inference at each step. The dynamics model, being model-based, can accurately represent the environment from relatively few real interactions, allowing the policy to train against a large number of simulated rollouts rather than requiring equivalent real AO loop interactions.

3.1.5 Algorithm Adaptations

One thing we noticed when applying the algorithm to the simulation is that, for different r_0 values and loop frequencies, the method was not robust and, in most cases, ended up diverging the entire control loop.

So to correct this several modifications were introduced. First, the loss function was adapted to account for the system's two-frame delay, specifically the fact that state s_{t+1} results from applying a_{t-1} to s_t , rather than a_t . Accordingly, instead of computing the reward based on action a_t and the next state produced by applying a_{t+1} to the dynamics, s_{t+1} is now obtained by applying a_{t-1} to the dynamics model, and the reward is computed from that state. This allows the model to better account for the delay inherent to the system.

However, this modification alone led to excessive instability, requiring two additional loss terms. The first, present in the original code implementation, penalises large actuator outputs, preventing the model from producing physically unrealisable actions. The second corrects the tendency for actions to overcompensate previous commands: a smoothing term penalises large differences between consecutive actions, stabilising the control loop. The resulting loss function is given in Eq. (3.1):

$$\hat{r}_\omega(s_t, a_t) = -\|\tilde{o}_{t+1}\|^2 - \lambda_a \|a_t\|^2 - \lambda_s \|a_t - a_{t-1}\|^2 \quad (3.1)$$

where:

- $\tilde{o}_{t+1}$ is the predicted next state, obtained from $\hat{p}_\omega(s_t, a_{t-1})$;
- λ_a penalises large actuator voltages;
- λ_s penalises rapid changes between consecutive actions, stabilising the control loop against overcompensation.

Second, the initial training data collection was changed so that rather than injecting random noise directly in actuator space, perturbations are applied in a modal basis, enabling the model to learn from more physically meaningful distortions.

Finally, the dynamics model was reformulated to predict the change in state rather than the next absolute state. This increases training stability because state transitions are typically much smaller in magnitude than the full state, allowing the model to focus on the effects of a given action.

3.1.6 Parameters

After the adaptations above, the algorithm's performance was found to be relatively insensitive to fine-tuning of hyperparameters. We therefore adopted the parameter values reported in the original work. The complete list is given in Table 3.3.

Table 3.3: RL parameters

PO4AO parameters		
Parameter	Value	Units
Planning horizon (H)	4	steps
Past DM commands (m)	15	commands
Past WFS measurements (k)	15	frames
CNN ensemble size	5	–
Dynamics iterations/episode	30	steps
Policy iterations/episode	15	steps
Training mini-batch size	32	–

Fixed CNN parameters		
Parameter	Value	Units
Number of conv. layers	3	layers
Filter size	3×3	pixels
Padding	1	pixels
Activation functions	Leaky ReLU	–

3.2 AO Control

For both experiments, data were collected using a noisy integrator over 10 episodes of 200 interactions each. After training, the model kept running for an additional 50 episodes.

This number of episodes was chosen because it was the minimum number of frames needed for the **PSD** computation to converged to the same values outside the **DM** zone of influence for experiments with the same loop frequency.

The integrator gain was selected as the highest value delivering good performance without destabilizing the system, resulting in a gain of 0.5. This gain was used for all integrator actions in both single-stage and double-stage configurations. The only exception was during data collection for cascade control, where a gain of 0.2 was used for the first stage instead. A higher gain caused the model to chase residuals at low frequencies and conflict with the first-stage correction; however, a gain of zero was also not viable, as in low- r_0 environments the model could not learn the contribution of the first stage and would similarly enter a fighting regime.

For all tests, we used the maximum number of modes after truncation to remove modes that cannot be well sensed by the **WFS**. This approach was preferred over fine-tuning or the common 85% rule of thumb because a higher mode count theoretically allows for better correction quality and one of the primary motivations for applying reinforcement learning is to reduce the need for manual tuning by allowing the model to learn how best to adapt to its environment.

The woofer–tweeter configuration requires a different approach. In this architecture, offloading is based on the interaction matrices IM_1 and IM_2 , that are computed as defined in 2.2.4.1, they describe the relationship between **WFS** slopes and the modal coefficients of the first and second

stages, respectively. The goal is to identify second-stage **DM** modes that can be reproduced by the first-stage **DM** while generating an equivalent **WFS** response.

The mapping between the two stages is obtained via a least-squares fit between the corresponding interaction matrices. Specifically, the projection matrix C_r is computed as:

$$C_r = \arg \min_x |IM_1 x - IM_2|^2 \quad (3.2)$$

The resulting matrix C_r provides the least-squares approximation of these second stage modes in the modal basis of first stage, thereby enabling the transfer of low-order corrections between the two stages.

Mode selection for this mapping is based on reconstruction quality. Since the integrator directly controls only the second stage, the offloading process must introduce minimal reconstruction error. The reconstruction quality of mode i is defined as:

$$Q_i = 1 - \frac{\|IM_{2,i} - IM_1 C_{r,i}\|^2}{\|IM_{2,i}\|^2} \quad (3.3)$$

where $IM_{2,i}$ denotes the i -th mode of the second-stage interaction matrix and C_i is the corresponding column of the projection matrix. Only modes satisfying $Q_i \geq 0.99$ are retained for offloading, as exemplified in Fig. 3.2.

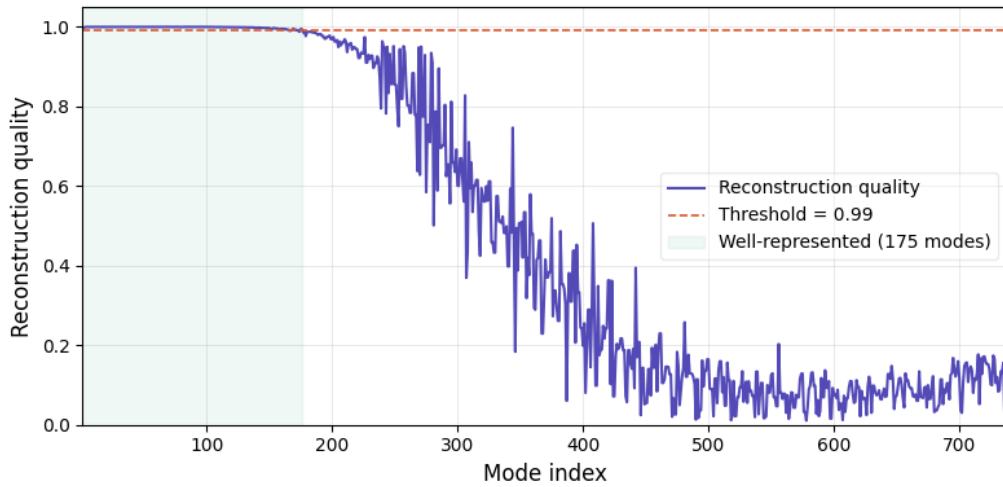


Figure 3.2: Reconstruction quality of second stage modes by the first stage in the scaled down system.

As mentioned before, the two stages operate at different loop frequencies; therefore, the offloading procedure is only performed when the first stage updates. The offload is determined by the current state of the second stage **DM**: the actuator values are converted into second stage modes, and only those modes that satisfy the quality criterion are transferred. An offload factor of 0.5 was used in all experiments. We did not offload all mode values, because even with high reconstruction quality there is still some noise that could be propagated. The first stage control

loop operates with a one-frame delay, since it only needs time to compute the offload and apply it to the **DM**, as illustrated in Fig. 2.9.

3.3 Performance Metrics

For the single-stage experiments, performance is evaluated using the Strehl ratio and the wavefront error, since the primary objective is to assess robustness under varying conditions.

For the two-stage systems, the main metric is contrast, which is the most important performance metric for high-contrast instruments such as **PCS** and is approximated using Eq. (2.14).

Chapter 4

Results

This chapter presents the main results obtained from the simulations. The results are organized into two sections, corresponding to the two experimental configurations: the single-stage and the two-stage architectures.

4.1 Single-Stage Results

Figures 4.1 and 4.2 show the Strehl ratio and the Wavefront Error (WFE) for the reinforcement learning (RL) and linear Reconstructor (RC) approaches across the different parameter combinations, at a loop frequency of 500 Hz. In all cases, RL outperforms RC.

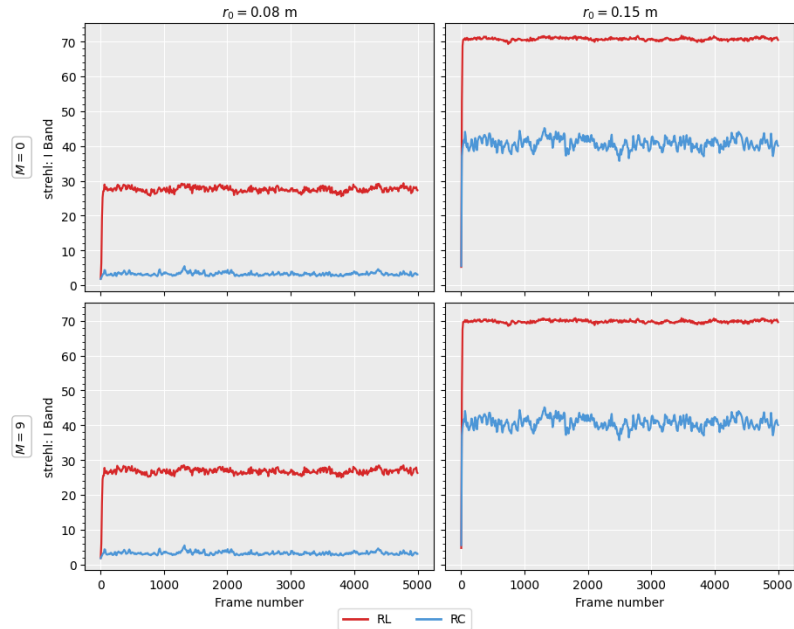


Figure 4.1: Strehl ratio in the I band for the reinforcement learning (RL) and linear reconstructor (RC) approaches at a loop frequency of 500 Hz.

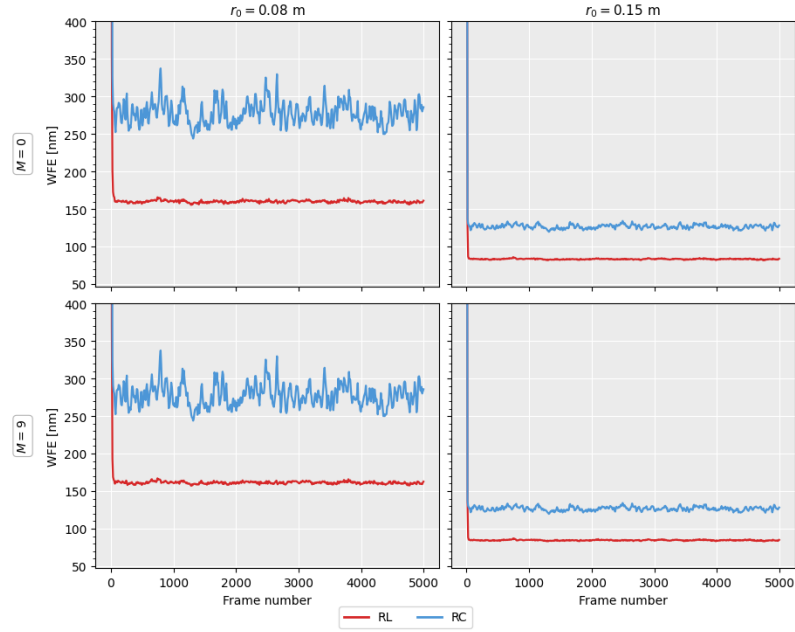


Figure 4.2: WFE for the **RL** and **RC** approaches at a loop frequency of 500 Hz.

Figures 4.3 and 4.4 show the same results for a loop frequency of 1 kHz. The **RL** approach again achieves higher performance than **RC** across all parameter combinations. The trend observed at 500 Hz is preserved: performance improves with larger r_0 , and the 1 kHz loop consistently outperforms the 500 Hz loop.

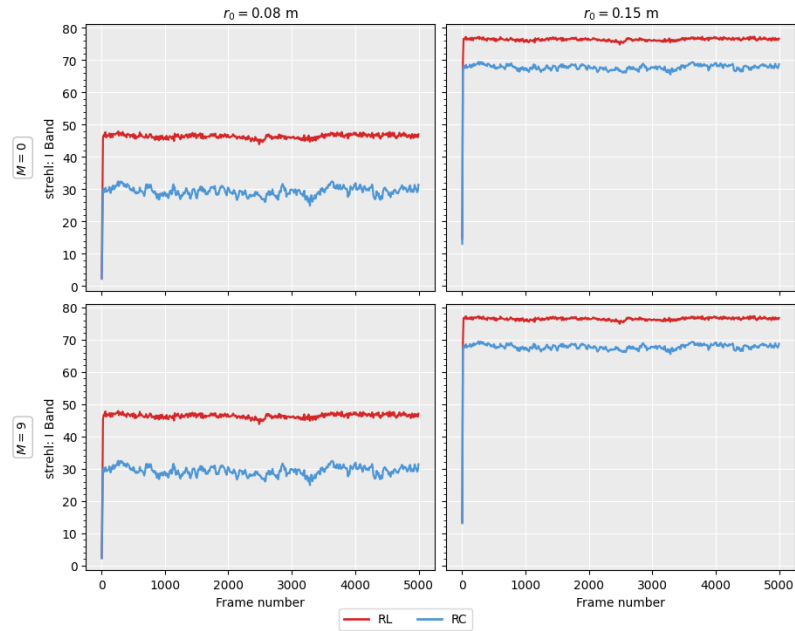


Figure 4.3: Strehl ratio in the I band for the **RL** and **RC** approaches at a loop frequency of 1 kHz.

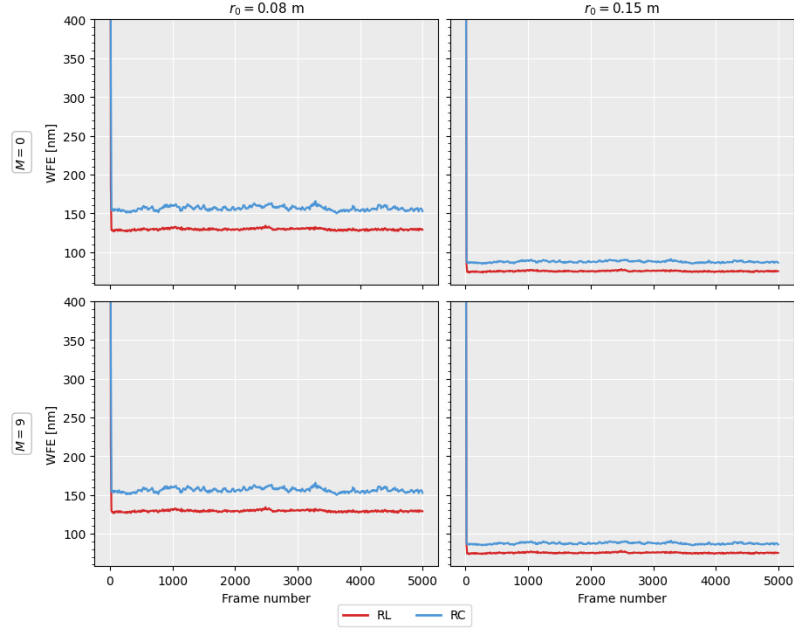


Figure 4.4: **WFE** for the **RL** and **RC** approaches at a loop frequency of 1 kHz.

Table 4.1: Strehl ratio results for **RL** and **RC** controllers.

Method	M	r_0 (m)	500 Hz	1000 Hz
RL	0	0.08	27.5456 ± 0.9331	46.3592 ± 0.8269
		0.15	70.7775 ± 0.5221	76.3987 ± 0.4774
	9	0.08	26.8028 ± 0.9270	46.3697 ± 0.8135
		0.15	69.8582 ± 0.5290	76.4329 ± 0.4757
RC	0	0.08	3.2504 ± 0.4795	29.2078 ± 1.3569
		0.15	40.7749 ± 1.6819	67.7064 ± 0.7672
	9	0.08	3.2504 ± 0.4795	29.2078 ± 1.3569
		0.15	40.7749 ± 1.6819	67.7064 ± 0.7672

From these results, we can conclude that the adaptations made to the algorithm make it robust to diverse seeing conditions (r_0), with **RL** maintaining a clear advantage over the linear reconstructor across all tested cases. The size of this advantage is not uniform, however: it is largest under the most challenging conditions (small r_0 , low loop frequency) and narrows as conditions become more favourable.

4.2 Double-Stage Results

This section presents the results for the two-stage architecture. Instead of using the Strehl ratio to compare performance, this section uses contrast. The results are organized according to the scale of the experiment.

4.2.1 Scaled-down system ($D = 10\text{ m}$, $20 \times 20 / 30 \times 30$)

This part presents the results for the small-scale experiments. Fig. 4.5 shows that, in the **CAO** configuration, **RL** achieves performance comparable to **RC** at 4 kHz, with both outperforming **RC** at 2 kHz.

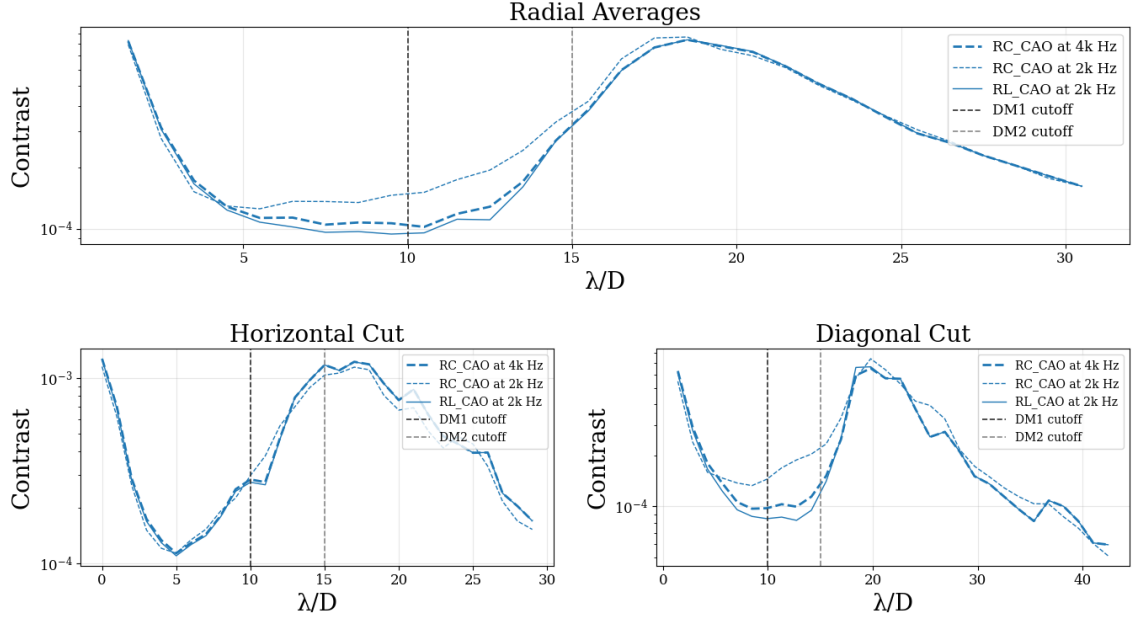


Figure 4.5: Contrast for the **RL** and **RC** approaches in the **CAO** configuration at the scaled-down system.

Fig. 4.6 shows that **RL** achieves better performance than both **RC** configurations, with **RC** at 4 kHz outperforming **RC** at 2 kHz.

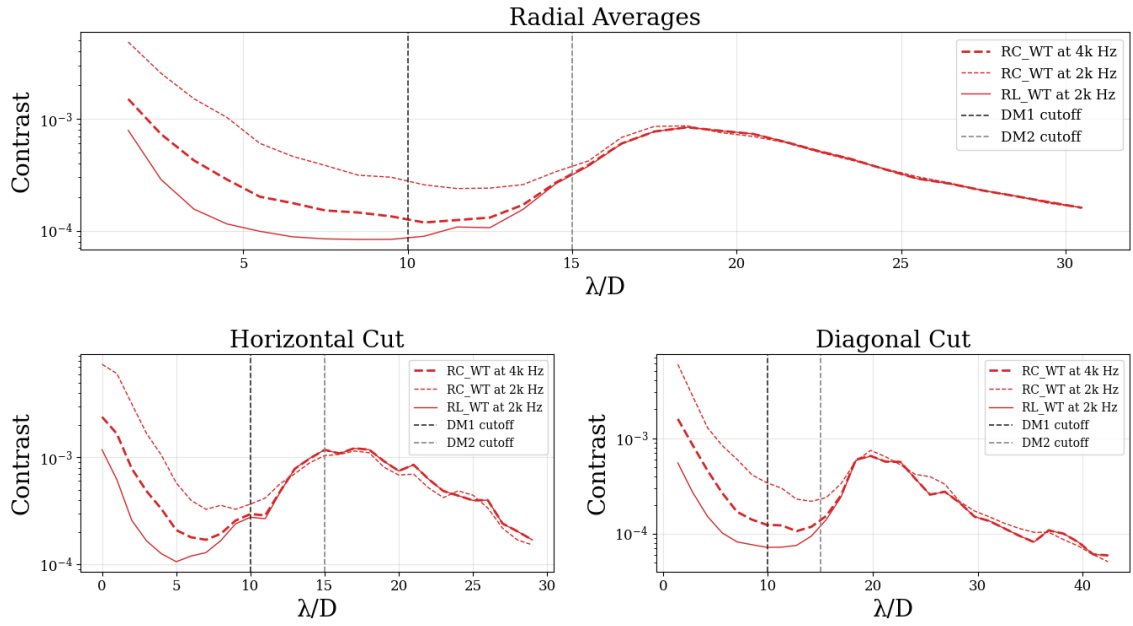


Figure 4.6: Contrast for the **RL** and **RC** approaches in the **WT** configuration at the scaled-down system.

Fig. 4.7 shows that, overall, **RC_WT** at 2 kHz has the lowest performance among all cases, while **RC_CAO** at 4 kHz achieves the best performance overall. It can also be seen that **RC_CAO** at 2 kHz outperforms **RC_WT** at 2 kHz.

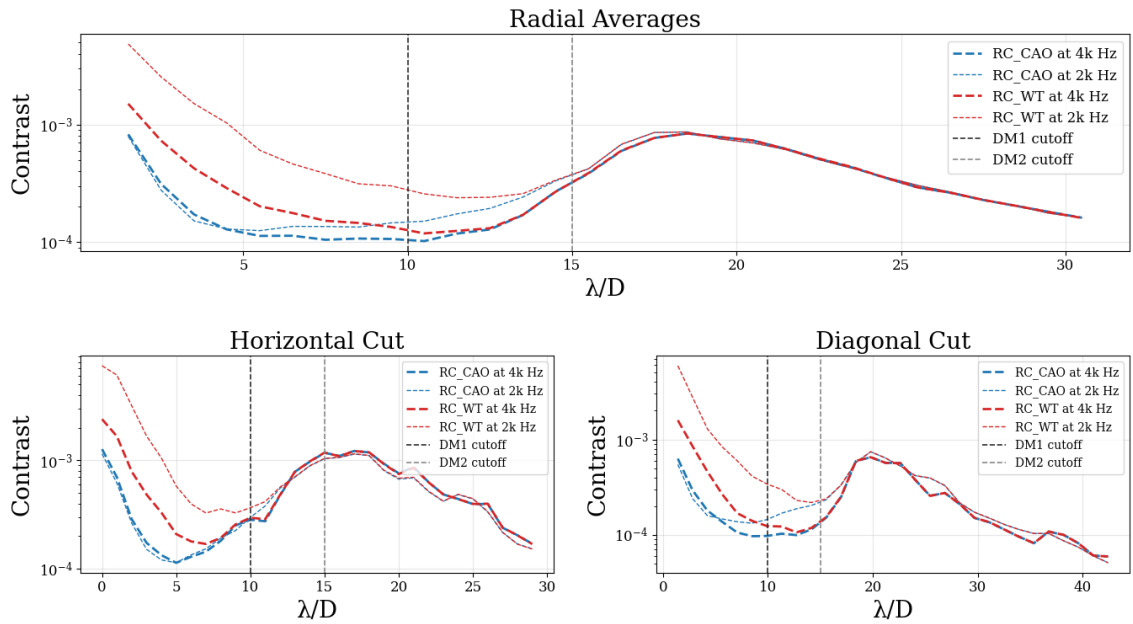


Figure 4.7: Contrast comparison between the **RC_WT** and **RC_CAO** configurations at 2 kHz and 4 kHz for the scaled-down system.

Fig. 4.8 shows that both **RL_WT** and **RL_CAO** achieve similar performance for this scaled-down experiment.

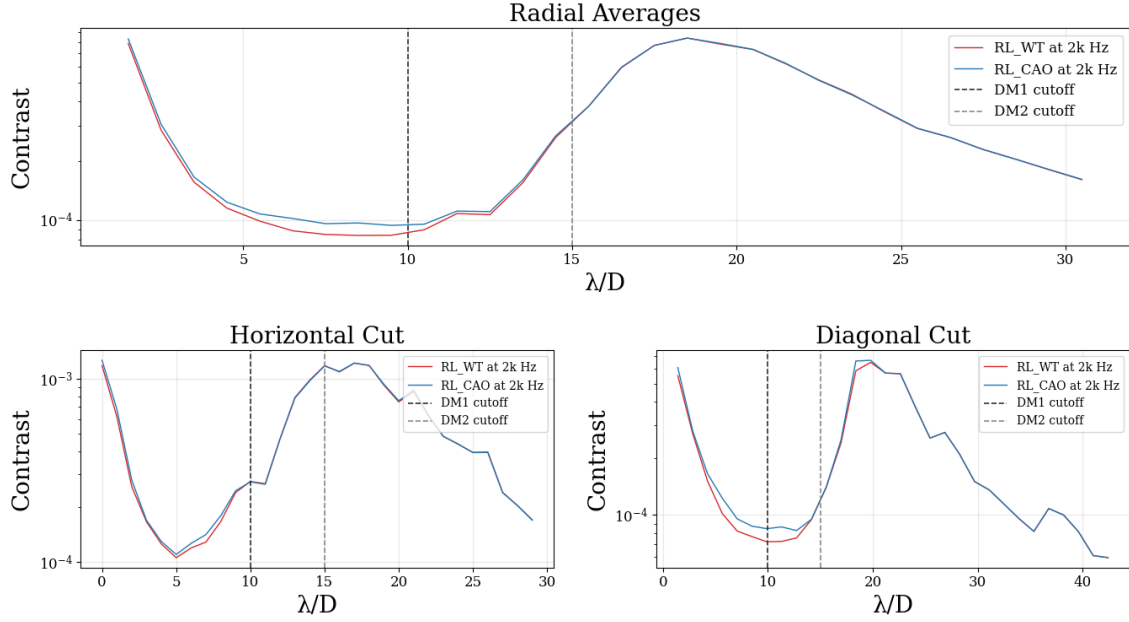


Figure 4.8: Contrast of the **RL** approach for the **WT** vs. **CAO** configurations at the scaled-down system.

Fig. 4.9 summarizes the results discussed above, showing that both **RL** and **RC_CAO** achieve the best overall performance.

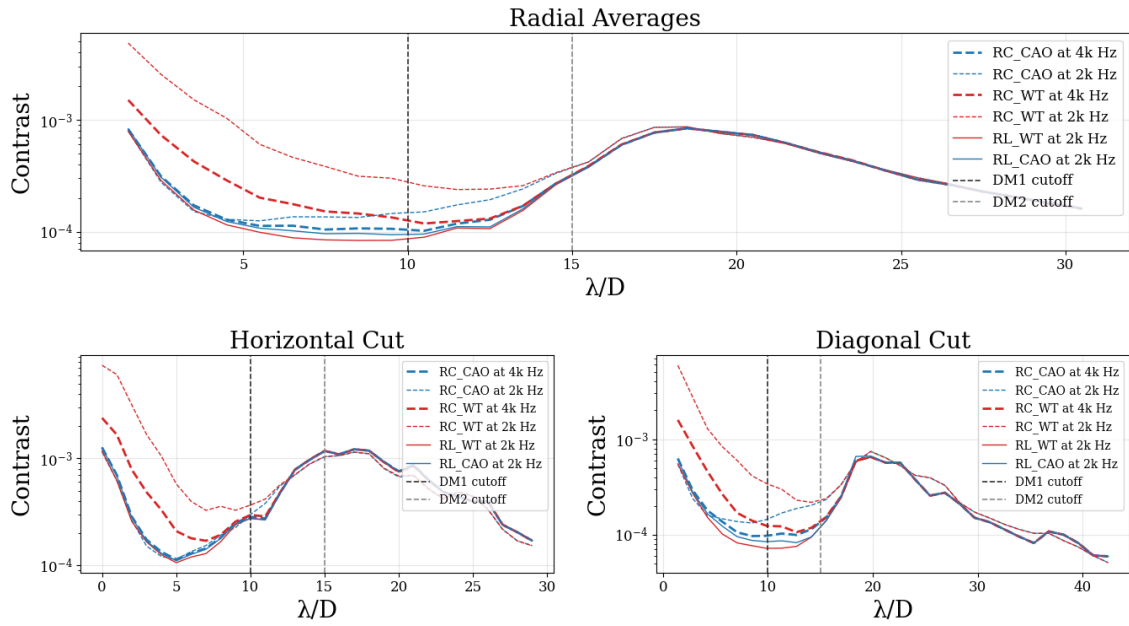


Figure 4.9: Contrast for all experimental configurations at the scaled-down system.

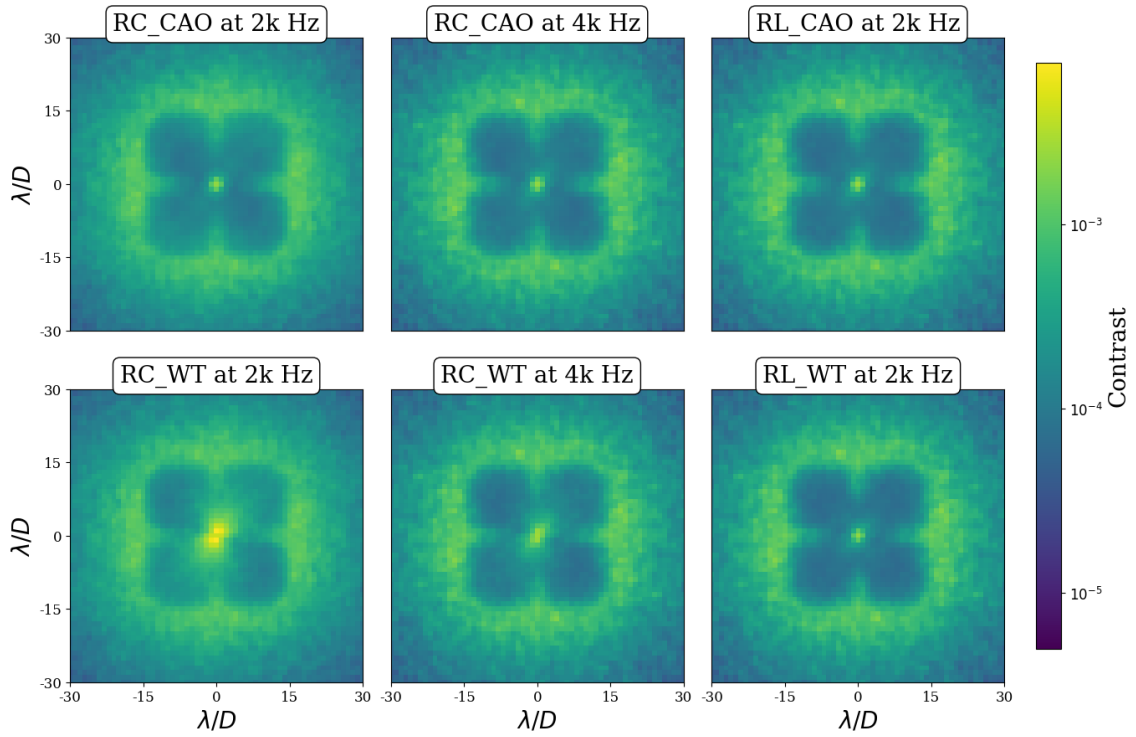


Figure 4.10: Ideal post-coronagraphic image for all experimental configurations at the scaled-down system.

For this scaled-down system, the CAO architecture achieved the best performance, and this best case can be closely approximated by the RL approaches.

4.2.2 Intermediate-scale system ($D = 20$ m, $40 \times 40 / 60 \times 60$)

Fig. 4.11 shows that RL at 2 kHz achieves better contrast than RC at all tested frequencies on the WT configuration.

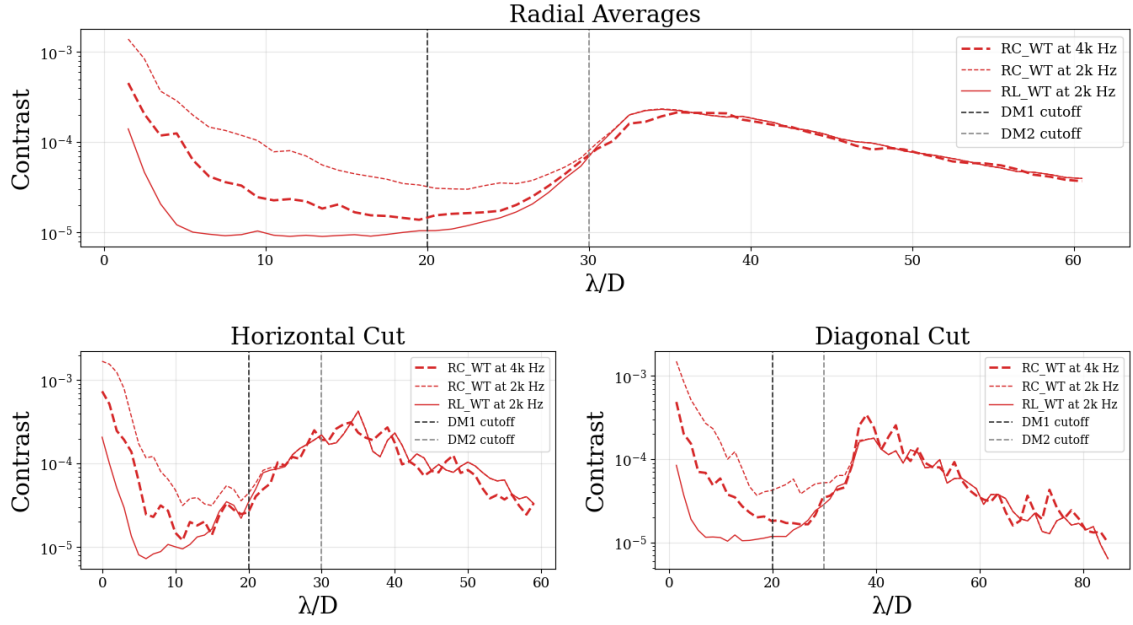


Figure 4.11: Contrast for the **RL** and **RC** approaches in the **WT** configuration.

Fig. 4.12 shows the results for the **CAO** configuration. In this case, **RL** and **RC** at 4 kHz achieve nearly identical performance, with both methods outperforming **RC** at 2 kHz.

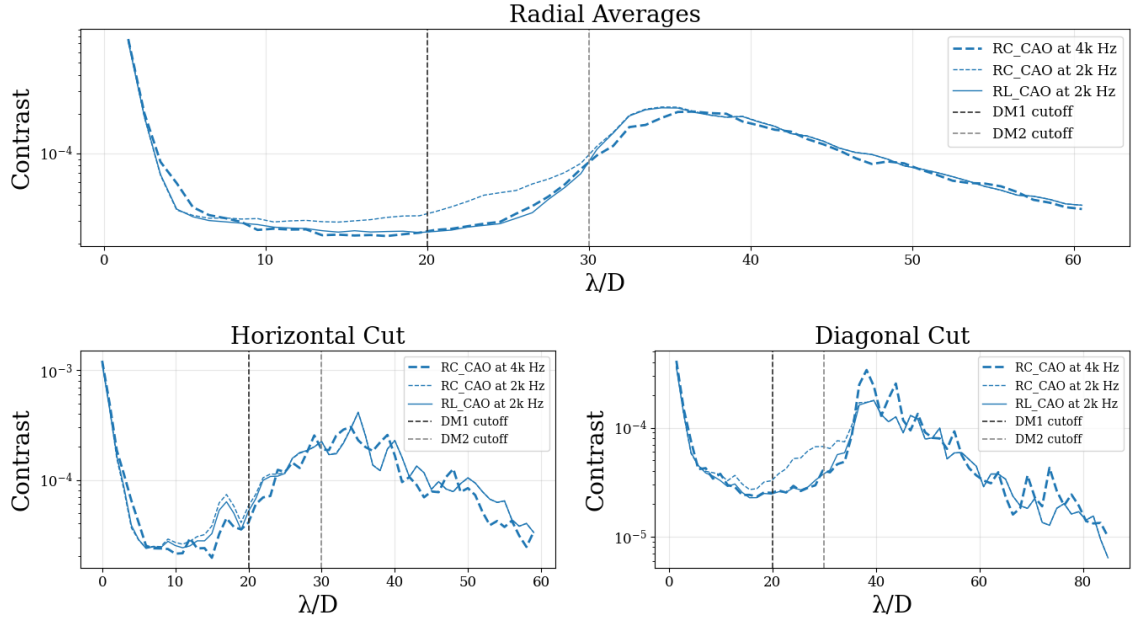


Figure 4.12: Contrast for the **RL** and **RC** approaches in the **CAO** configuration.

Fig. 4.13 compares the contrast of **RC** for the **WT** and **CAO** configurations. **CAO** achieves higher contrast at smaller angular separations than **WT**, with this difference decreasing as the loop frequency increases. At larger separations, **WT** performs better than **CAO**, and this disparity increases as the loop frequency increases.

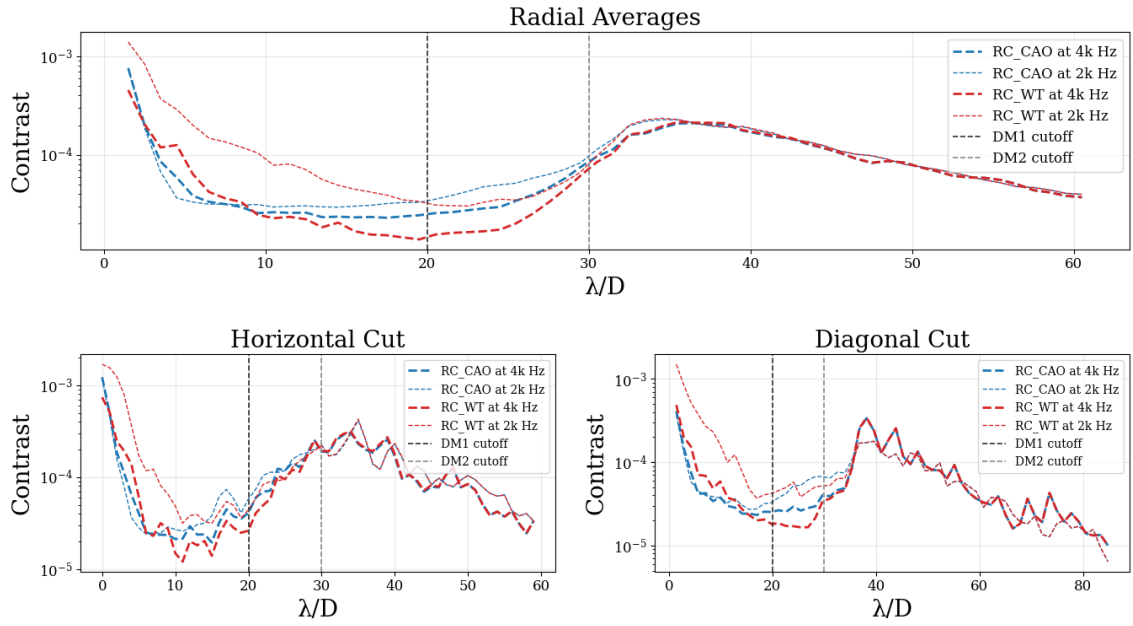


Figure 4.13: Contrast of the **RC** approach for the **WT** vs. **CAO** configurations.

Fig. 4.14 shows that the **RL**-based **WT** configuration achieves better contrast than **RL**-based **CAO** across all correction frequencies.

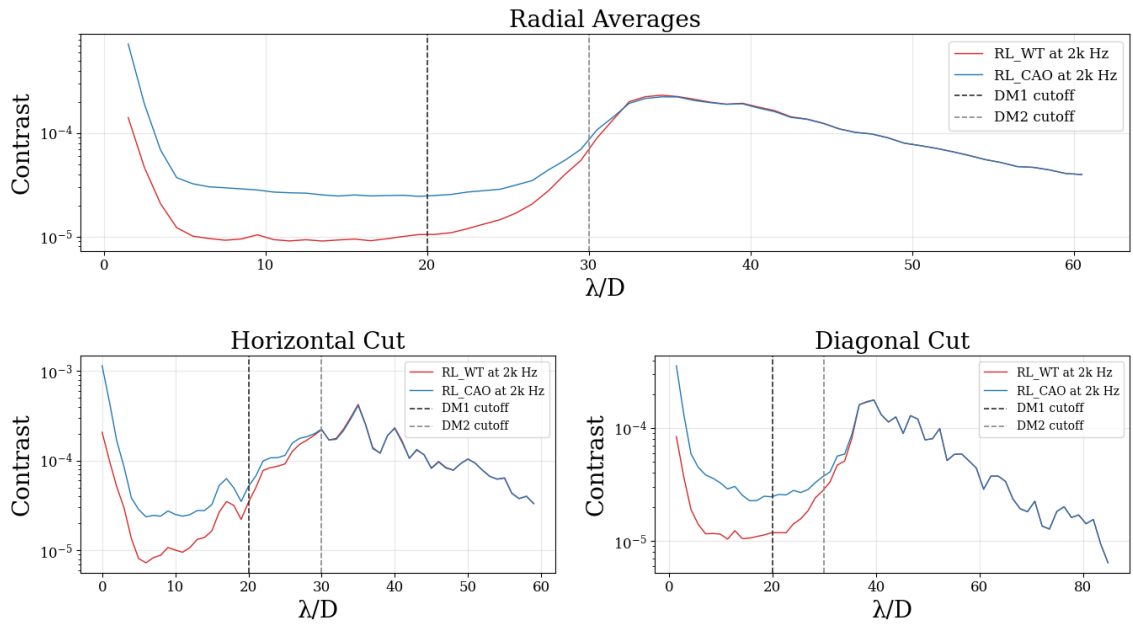


Figure 4.14: Contrast of the **RL** approach for the **WT** vs. **CAO** configurations.

Fig. 4.15 shows the contrast achieved by all experimental configurations. The **RL**-based **WT** configuration achieves the highest contrast among all tested cases.

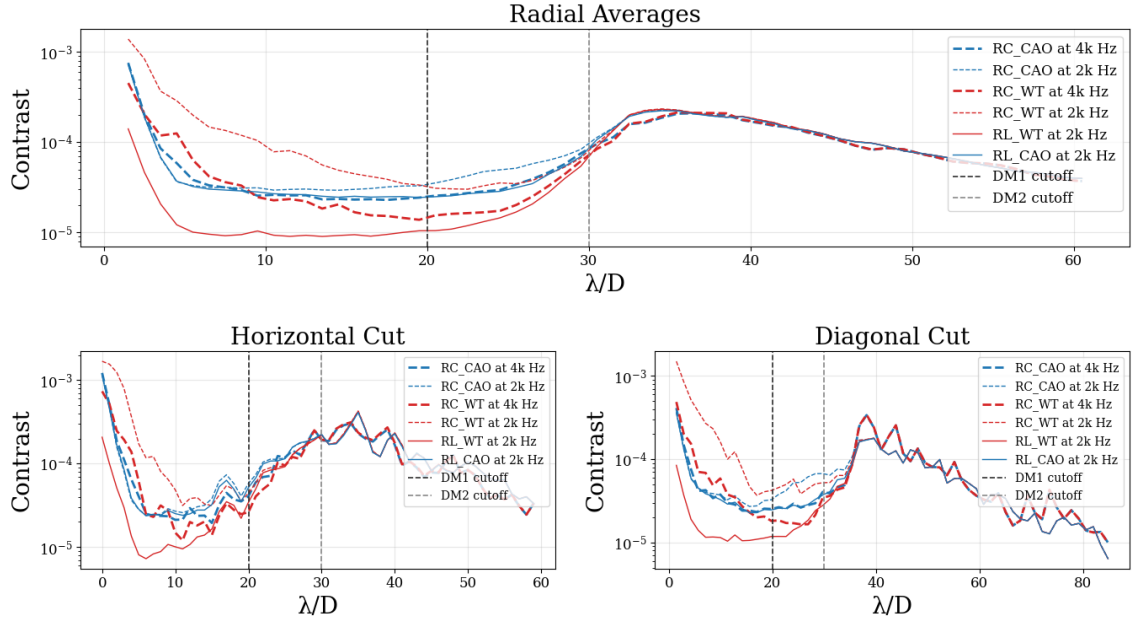


Figure 4.15: Contrast for all experimental configurations.

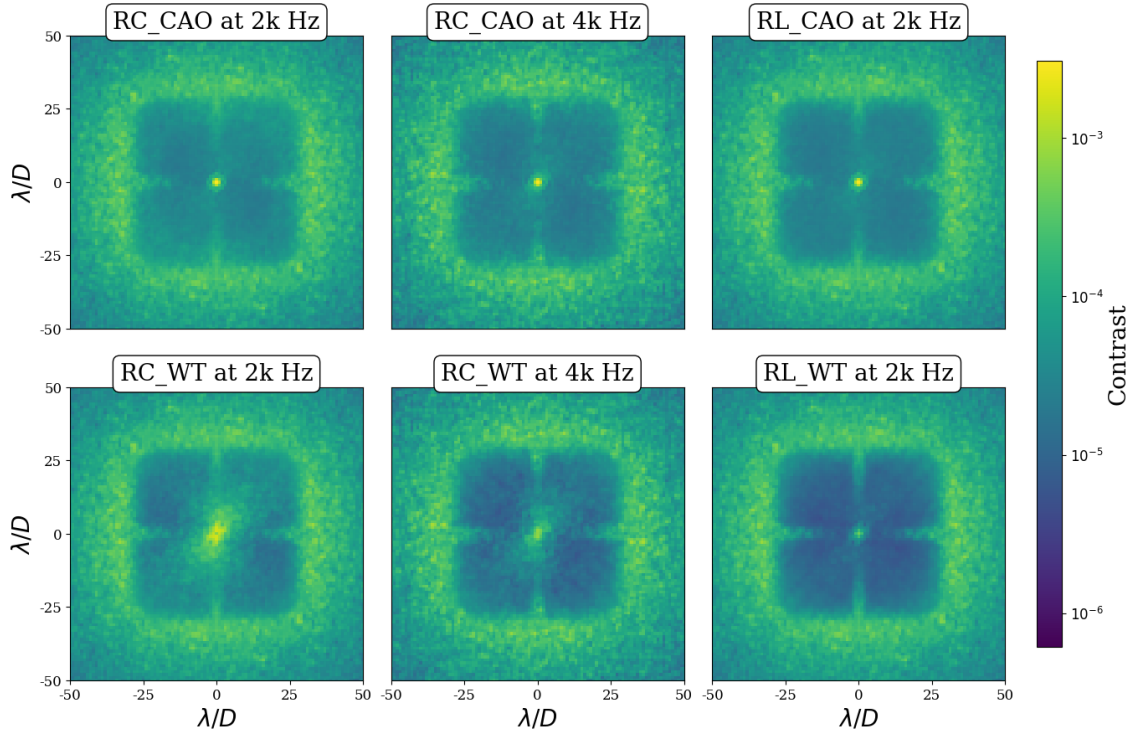


Figure 4.16: Ideal post-coronagraphic image for all experiments at the intermediate-scale system.

In this experiment, we begin to observe that the **WT** configuration achieves the best performance when combined with **RL**. Compared to the scaled-down system, **RC_WT** appears to outperform **RC_CAO** at higher frequencies, when compared at the same loop frequency.

4.2.3 ELT-sized system

Here, we scale the approach to an **ELT**-sized system. Fig. 4.17 shows a comparison of results similar to that of the previous sections. Fig. 4.18 shows a different result: this time, **RL** performance is no longer comparable to **RC** at 4 kHz, and instead more closely matches **RC** at 2 kHz.

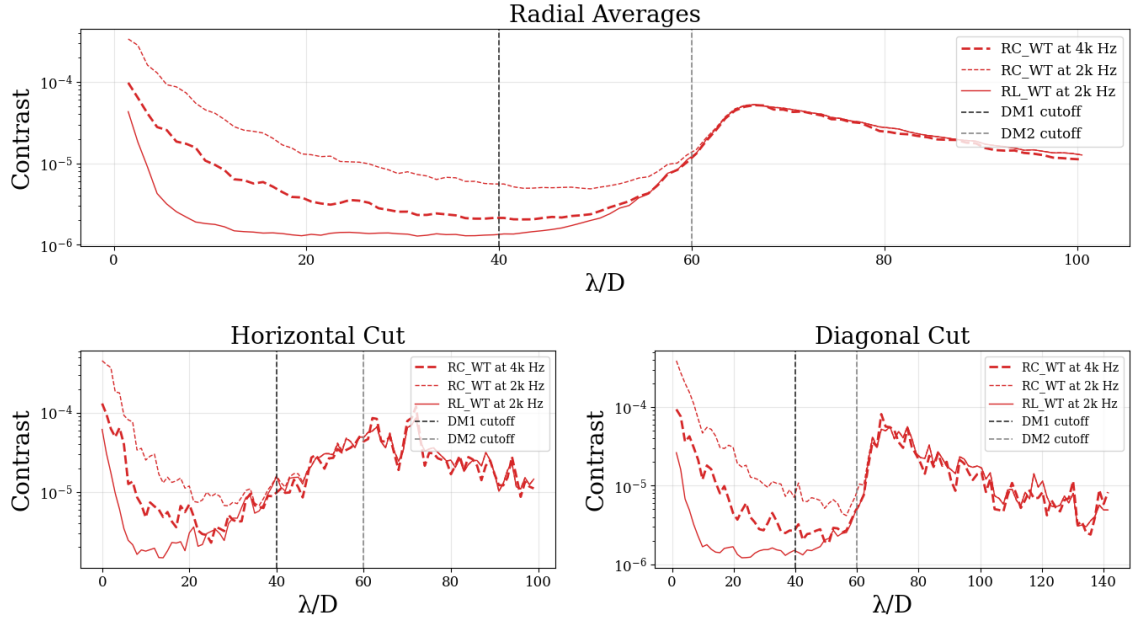


Figure 4.17: Contrast for the **RL** and **RC** approaches in the **WT** configuration at the **ELT**-sized system.

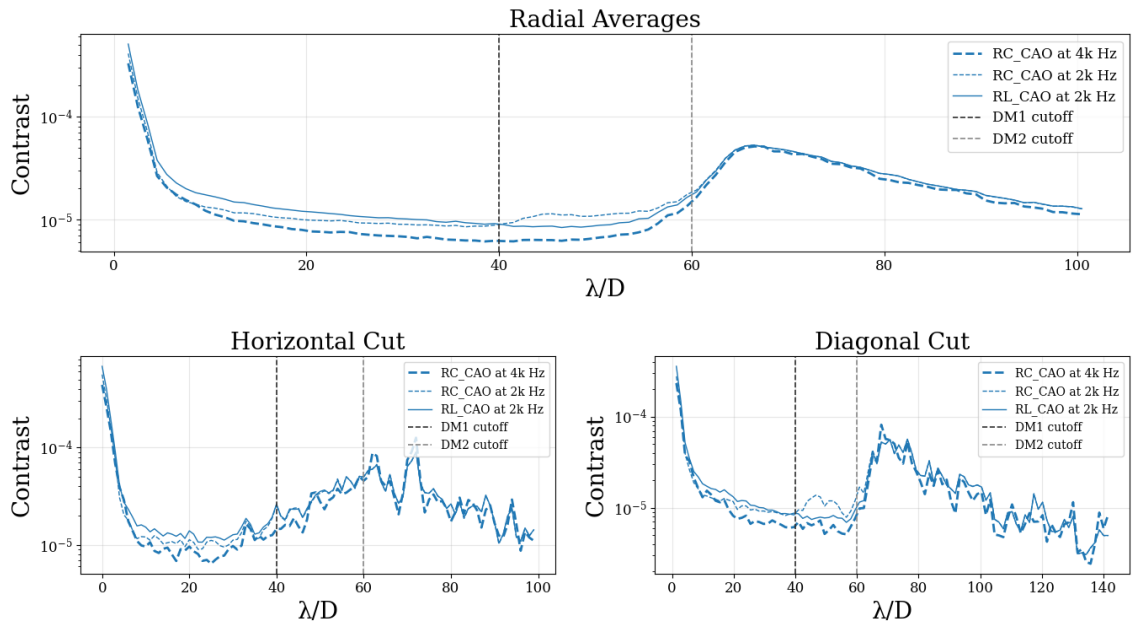


Figure 4.18: Contrast for the **RL** and **RC** approaches in the **CAO** configuration at the **ELT**-sized system.

Fig. 4.19 shows a pattern similar to that of the previous sections: the **WT** configuration performs even better at larger separations than the **CAO** configuration, with **WT** at 4 kHz outperforming **CAO** at 4 kHz across almost all separations.

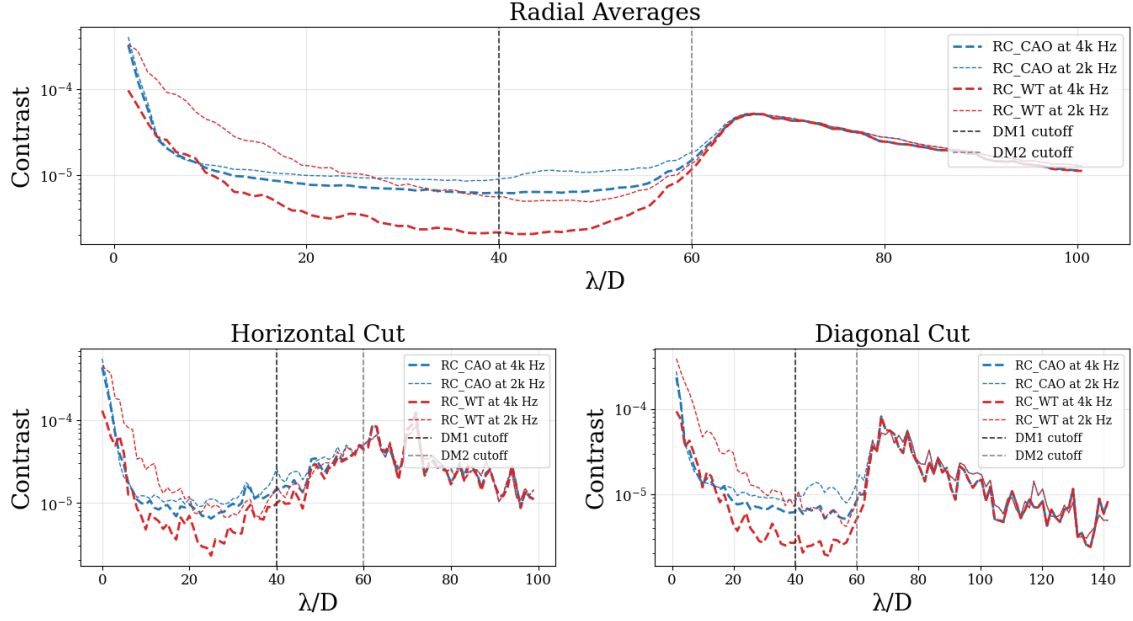


Figure 4.19: Contrast of the **RC** approach for the **WT** vs. **CAO** configurations at the **ELT**-sized system.

Fig. 4.20 shows results consistent with those of the previous sections.

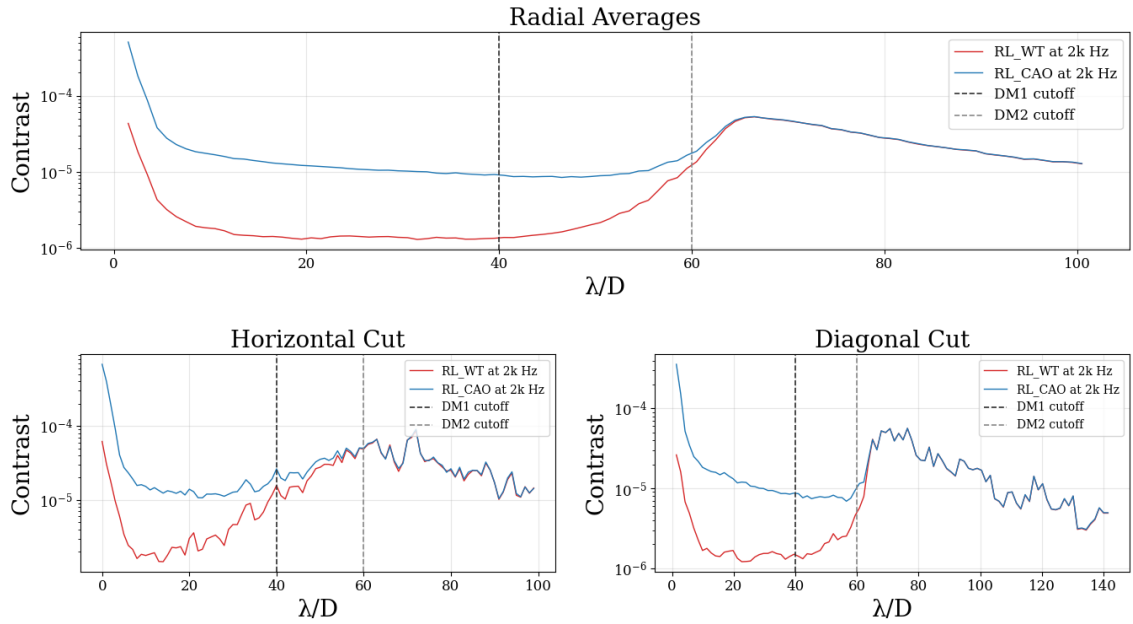


Figure 4.20: Contrast of the **RL** approach for the **WT** vs. **CAO** configurations at **ELT**-sized system.

Fig. 4.21 shows that **RL_WT** still achieves the best performance among all experiments.

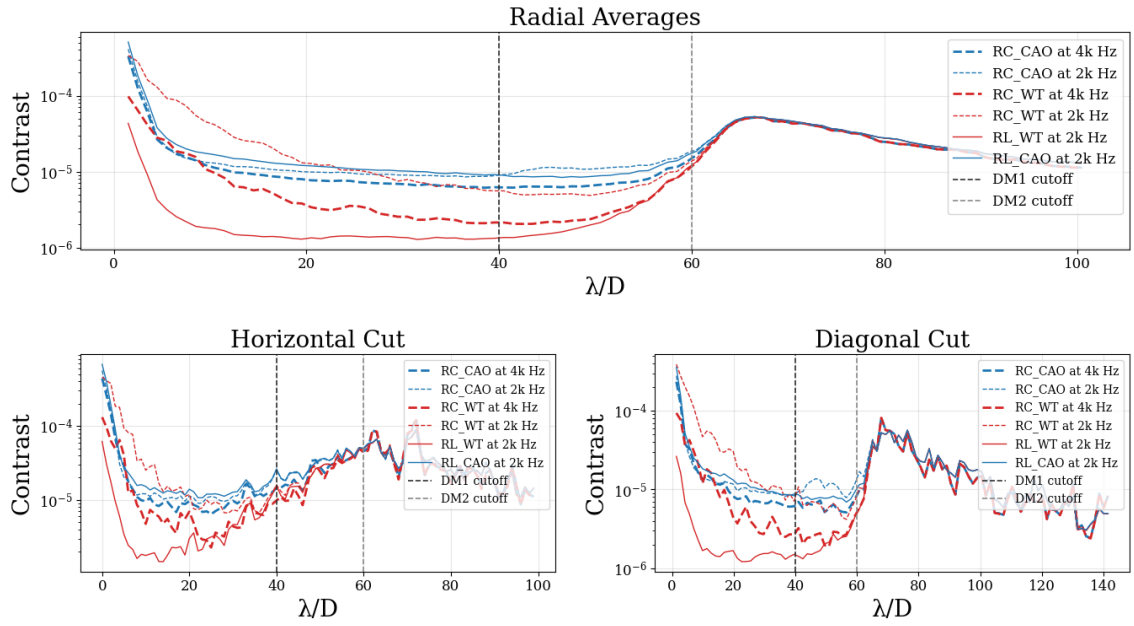


Figure 4.21: Contrast for all experimental configurations at the **ELT**-sized system.

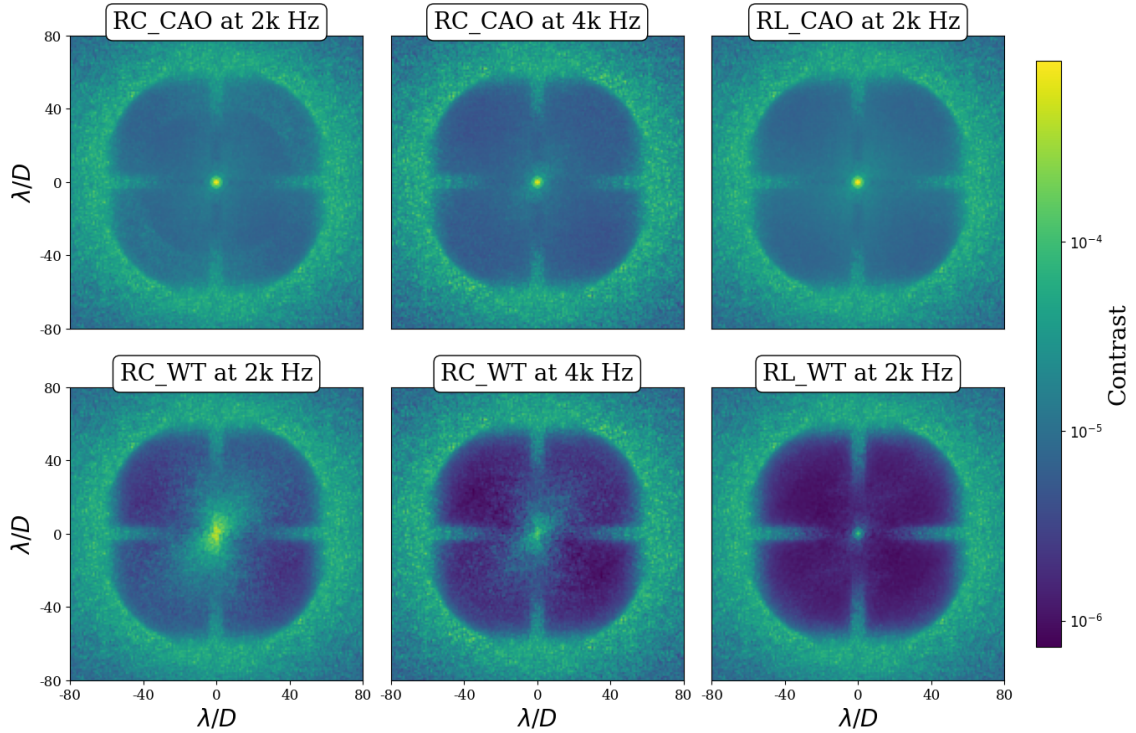


Figure 4.22: Ideal post-coronagraphic image for all experiments at the **ELT**-sized system.

Unlike at the smaller scales, the **CAO** configuration does not scale as favourably to the **ELT** case: **RL** performance, which matched or exceeded **RC_CAO** at 4 kHz for the smaller systems, here only matches **RC_CAO** at 2 kHz. The **WT** configuration, in contrast, continues to perform well at this scale and remains the best-performing configuration overall when combined with **RL**.

4.2.4 Supplementary Results

To better understand the disparity in results observed in the previous sections, particularly the difference in contrast between the **WT** and **CAO** configurations, we test changing the band of the second **WFS** at **CAO** configuration. Previously set to I band, it is now tested at H band, matching the band used by the **WFS** in the **WT** configuration.

As shown in Fig. 4.23, both **RC_CAO** and **RL_CAO** at H band achieve higher contrast than both **RC_WT** and the **CAO** controllers at I band, although their performance remains lower than that of **RL_WT**. A more detailed discussion of these results is provided in the paper reproduced in Annex B.

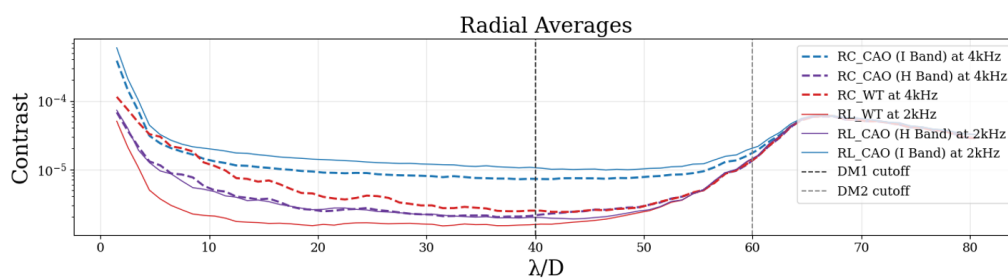


Figure 4.23: New contrast comparison for the **CAO** vs **WT** configurations at the **ELT**-sized system.

Chapter 5

Discussion

5.1 Results Analysis

This work aimed to apply reinforcement learning to improve the performance of two-stage control while relaxing hardware requirements. Overall, the results show that this objective was achieved: in most tested configurations, **RL** matched or outperformed a linear reconstructor operating at twice the loop frequency, suggesting that hardware requirements could be relaxed. In the **WT** configuration, the improvement ranged from a factor of 2 to 6 at smaller angular separations and reached up to a factor of 2 at larger separations. In the **CAO** configuration, **RL** performance was similar to that of **RC** at twice the loop frequency. Across all configurations, the **RL**-based **WT** approach delivered the best overall performance, improving by a factor of 4 to 6 at smaller separations compared with the best-performing **CAO** approach.

An interesting pattern emerged: the **CAO** configuration with a linear reconstructor typically yields better performance at smaller angular separations than the **WT** configuration. This is expected by design: the second stage operates on the residuals of the first stage, providing stronger correction at those spatial frequencies. However, this advantage did not carry over to the **RL** controllers, which instead performed better in the **WT** configuration.

Initially, we believed that this divergence stemmed from a need for further tuning between the two stages of the **CAO** controller. Further testing showed, however, that the disparity between the **RC** controllers did not originate from this, but rather from the operating band of the second **WFS** in the **CAO** configuration. When this band was changed to match that of the **WT** configuration, the **CAO** configuration with **RC** outperformed the **WT** configuration.

This band effect, however, does not fully explain the difference in **RL** performance, as **RL_WT** still showed the best performance even after accounting for it; we believe the main factor behind this remaining gap is a difference in causality between the two architectures. In the **WT** configuration, although the **RL** controller only acts on the second **DM**, it has full causal knowledge of both **DM** states, so the algorithm only needs to adapt to the evolution of the environment. In the **CAO** configuration, by contrast, the **RL** controller must adapt both to the evolution of the environment and to the effect of the independent, first-stage controller. It should also be noted that

the **CAO** tests did not employ a fully tuned first-stage controller; further tuning could narrow the performance gap with the **WT** configuration.

5.2 Real-World Application and Limitations

While these results demonstrate the capability of **RL** to control complex, multi-mirror systems, translating this approach to real-world deployment introduces additional challenges beyond scalability alone.

One aspect that was not measured in this work is inference time. In a real deployment, the **RL** controller must execute within the available loop period, and it remains to be verified whether the current implementation can meet these loop-frequency requirements.

Another limitation is the duration of the initial training phase, which is performed on data collected during the warm-up period. A mismatch between the conditions present during data collection and those at deployment time is possible; however, one of the key advantages of the algorithm is its robustness to distributional shift. Even if the initial model is deployed under slightly different conditions, it can still provide a comparable level of loop control. Subsequent retraining cycles do not share this limitation: they are considerably faster and could, in principle, be performed in parallel with the active controller, with updated weights applied seamlessly.

Several aspects of real-world deployment could not be fully accounted for in the simulations presented here. Although the action delay was included in the loop control, each command was applied instantaneously. In practice, since the second stage operates at a much higher frequency than the first, commands issued mid-frame by the first stage may be applied during an ongoing correction cycle, introducing sub-frame timing effects that were not modeled. Other factors, such as readout and photon noise in real sensor responses or errors associated with the physical components of the telescope, were not simulated in this work. Moreover, even a high-fidelity simulation cannot capture all real-world effects, and only on-sky testing would reveal the true performance of the algorithm.

Despite these limitations, the results are encouraging. **RL**-based control offers capabilities that are particularly well matched to the demands of two-stage **AO**: the ability to anticipate future wavefront evolution across coupled stages, the potential to eliminate periodic interaction-matrix recalibrations, and the capacity to handle the nonlinear response of pyramid wavefront sensors in ways that linear reconstructors cannot.

5.3 Challenges

The most significant challenge was the computational cost of the simulations, particularly at **ELT** scale. The most demanding operations, namely the computation of interaction matrices and **KL** modes, restricted both the scale and the number of experiments that could be carried out.

Another challenge was that the original objective was to control the unmodulated pyramid wavefront sensor. However, this configuration proved too unstable to operate reliably across diverse seeing conditions.

Chapter 6

Future work

For future work, there are diverse avenues that can be pursued. The first would be to extend **RL** control to both stages of the cascade architecture. Although this would increase the overall complexity of the algorithm, it could help mitigate the causality issues present in the **CAO** configuration and improve its performance at small angular separations, as initially suggested by the linear reconstructor results.

A second direction would be to optimise the code and benchmark inference time, to ensure that the algorithm can operate within the loop frequencies required for real-world deployment.

Finally, the algorithm should be tested on a representative **PCS** architecture that accounts not only for scale, as in this work, but also for the **ELT**'s actual pupil geometry and other relevant system characteristics.

References

- [1] 4D Technology. *Shack-Hartmann Sensor vs an Interferometer*. 2025. URL: <https://4dtechnology.com/products/shack-hartmann-vs-interferometer/> (visited on 02/03/2026).
- [2] Dan Abolafia. *A Recurrent Neural Network Music Generation Tutorial*. 2016. URL: <https://magenta.withgoogle.com/2016/06/10/recurrent-neural-network-generation-tutorial> (visited on 02/03/2026).
- [3] Guido Agapito et al. “Non-modulated pyramid wavefront sensor-use in sensing and correcting atmospheric turbulence”. In: *Astronomy & Astrophysics* 677 (2023), A168.
- [4] J. R. P. Angel et al. “Adaptive optics for array telescopes using neural-network techniques”. In: *Nature* 348 (1990), pp. 221–224. DOI: [10.1038/348221a0](https://doi.org/10.1038/348221a0). URL: <https://doi.org/10.1038/348221a0>.
- [5] J.-L. Beuzit et al. “SPHERE: the exoplanet imager for the Very Large Telescope”. In: *aap* 631, A155 (Nov. 2019), A155. DOI: [10.1051/0004-6361/201935251](https://doi.org/10.1051/0004-6361/201935251). arXiv: 1902.04080 [astro-ph.IM].
- [6] Deepak Birla. *Autoencoders*. 2019. URL: <https://medium.com/@birla.deepak26/autoencoders-76bb49ae6a8f> (visited on 02/03/2026).
- [7] A. Boccaletti et al. *Upgrading the high contrast imaging facility SPHERE: science drivers and instrument choices*. 2022. arXiv: 2209.02092 [astro-ph.IM]. URL: <https://arxiv.org/abs/2209.02092>.
- [8] Runa Briguglio et al. “Optical calibration and performance of the adaptive secondary mirror at the Magellan telescope”. In: *Scientific Reports* 8 (July 2018). DOI: [10.1038/s41598-018-29171-6](https://doi.org/10.1038/s41598-018-29171-6).
- [9] Jeffrey Chilcote et al. “GPI 2.0: upgrading the Gemini Planet Imager”. In: *Ground-based and Airborne Instrumentation for Astronomy VIII*. Ed. by Christopher J. Evans, Julia J. Bryant, and Kentaro Motohara. Vol. 11447. International Society for Optics and Photonics. SPIE, 2020, 114471S. DOI: [10.1117/12.2562578](https://doi.org/10.1117/12.2562578). URL: <https://doi.org/10.1117/12.2562578>.
- [10] Carlos Correia et al. “On the optimal reconstruction and control of adaptive optical systems with mirror dynamics”. In: *Journal of the Optical Society of America A* 27.2 (2010), pp. 333–349.
- [11] Carlos Correia et al. “Optimal modal basis for ELT M4 mirror Force/Position control”. In: URL: <https://api.semanticscholar.org/CorpusID:272831561>.
- [12] Carlos M Correia et al. “Performance limits of adaptive-optics/high-contrast imagers with pyramid wavefront sensors”. In: *Monthly Notices of the Royal Astronomical Society* 495.4 (June 2020), pp. 4380–4391. ISSN: 1365-2966. DOI: [10.1093/mnras/staa843](https://doi.org/10.1093/mnras/staa843). URL: <http://dx.doi.org/10.1093/mnras/staa843>.

- [13] Theodore B. DuBose, Dennis F. Gardner, and Abbie T. Watnik. “Intensity-enhanced deep network wavefront reconstruction in Shack–Hartmann sensors”. In: *Optics Letters* 45.7 (2020), pp. 1699–1702. DOI: [10.1364/OL.45.001699](https://doi.org/10.1364/OL.45.001699). URL: <https://doi.org/10.1364/OL.45.001699>.
- [14] J. Fowler and Rico Landman. “Tempestas ex machina: A review of machine learning methods for wavefront control”. In: *Techniques and Instrumentation for Detection of Exoplanets XI*. Vol. 12680. Proceedings of SPIE. 2023, 126800E. DOI: [10.1117/12.2676536](https://doi.org/10.1117/12.2676536). URL: <https://doi.org/10.1117/12.2676536>.
- [15] David L Fried. “Statistics of a geometric representation of wavefront distortion”. In: *Journal of the optical society of America* 55.11 (1965), pp. 1427–1435.
- [16] E. Gendron and P. Léna. “Astronomical adaptive optics. 1: Modal control optimization”. In: *Astronomy and Astrophysics* 291 (Oct. 1994), pp. 337–347.
- [17] Darryl P Greenwood. “Bandwidth specification for adaptive optics systems”. In: *Journal of the Optical Society of America* 67.3 (1977), pp. 390–393.
- [18] Y. M. Guo et al. “Adaptive optics based on machine learning: a review”. In: *Opto-Electronic Advances* 5 (2022), p. 200082. DOI: [10.29026/oea.2022.200082](https://doi.org/10.29026/oea.2022.200082). URL: <https://doi.org/10.29026/oea.2022.200082>.
- [19] Olivier Guyon. “Limits of Adaptive Optics for High-Contrast Imaging”. In: *The Astrophysical Journal* 629 (Aug. 2005), pp. 592–614. DOI: [10.1086/431209](https://doi.org/10.1086/431209).
- [20] Rehan Hafeez et al. “Forecasting wavefront corrections in an adaptive optics system”. In: *Journal of Astronomical Telescopes, Instruments, and Systems* 8, 029003 (Apr. 2022), p. 029003. DOI: [10.1117/1.JATIS.8.2.029003](https://doi.org/10.1117/1.JATIS.8.2.029003). arXiv: 2112.01437 [astro-ph.IM].
- [21] YH Hsieh et al. “Integral-based parallel algorithm for the fast generation of the Zernike polynomials”. In: *Optics express* 28.2 (2020), pp. 936–947.
- [22] jnousi. *PO4AO*. <https://github.com/jnousi/PO4AO>. Accessed: 2026-06-07. 2023.
- [23] M. Kasper et al. “PCS — A Roadmap for Exoearth Imaging with the ELT”. In: *The Messenger* 182 (Mar. 2021), pp. 38–43. DOI: [10.18727/0722-6691/5221](https://doi.org/10.18727/0722-6691/5221). arXiv: 2103.11196 [astro-ph.IM].
- [24] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2014). URL: <https://api.semanticscholar.org/CorpusID:6628106>.
- [25] Andrey Nikolaevich Kolmogorov. “The local structure of turbulence in incompressible viscous fluid for very large Reynolds number”. In: *Cr Acad. Sci. USSR* 30 (1941), p. 301.
- [26] Mario Krenn et al. “Artificial intelligence and machine learning for quantum technologies”. In: *Phys. Rev. A* 107 (1 Jan. 2023), p. 010101. DOI: [10.1103/PhysRevA.107.010101](https://doi.org/10.1103/PhysRevA.107.010101). URL: <https://link.aps.org/doi/10.1103/PhysRevA.107.010101>.
- [27] Tomi Krokberg. “Reinforcement learning in multi-mirror adaptive optics”. In: (2022).
- [28] R. Landman et al. “Self-optimizing adaptive optics control with reinforcement learning”. In: *Adaptive Optics Systems VII*. Ed. by Laura Schreiber, Dirk Schmidt, and Elise Vernet. Vol. 11448. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Dec. 2020, 1144849, p. 1144849. DOI: [10.1117/12.2560053](https://doi.org/10.1117/12.2560053). arXiv: 2012.01997 [astro-ph.IM].

- [29] Rico Landman et al. *Self-optimizing adaptive optics control with Reinforcement Learning for high-contrast imaging*. 2021. arXiv: 2108.11332 [astro-ph.IM]. URL: <https://arxiv.org/abs/2108.11332>.
- [30] Xuewen Liu et al. “Wavefront prediction using artificial neural networks for open-loop adaptive optics”. In: *Monthly Notices of the Royal Astronomical Society* 496.1 (2020), pp. 456–464. DOI: 10.1093/mnras/staa1558. URL: <https://doi.org/10.1093/mnras/staa1558>.
- [31] M. Lloyd-Hart and P. McGuire. “Spatio-temporal prediction for adaptive optics wavefront reconstructors”. In: *European Southern Observatory Conference and Workshop Proceedings*. Ed. by M. Cullum. Vol. 54. European Southern Observatory Conference and Workshop Proceedings. Jan. 1996, p. 95.
- [32] M. Lloyd-Hart et al. “First Results of an On-Line Adaptive Optics System with Atmospheric Wavefront Sensing by an Artificial Neural Network”. In: *apjl* 390 (May 1992), p. L41. DOI: 10.1086/186367.
- [33] Andrew L. Maas. “Rectifier Nonlinearities Improve Neural Network Acoustic Models”. In: 2013. URL: <https://api.semanticscholar.org/CorpusID:16489696>.
- [34] Bruce Macintosh et al. “Adaptive optics for direct detection of extrasolar planets: the Gemini Planet Imager”. In: *Comptes Rendus Physique* 8.3 (2007). Optical techniques for direct imaging of exoplanets, pp. 365–373. ISSN: 1631-0705. DOI: <https://doi.org/10.1016/j.crhy.2007.04.007>. URL: <https://www.sciencedirect.com/science/article/pii/S1631070507000709>.
- [35] P.-Y. Madec. “Overview of deformable mirror technologies for adaptive optics and astronomy”. In: *Adaptive Optics Systems III*. Ed. by Brent L. Ellerbroek, Enrico Marchetti, and Jean-Pierre Véran. Vol. 8447. International Society for Optics and Photonics. SPIE, 2012, p. 844705. DOI: 10.1117/12.924892. URL: <https://doi.org/10.1117/12.924892>.
- [36] MathWorks. *What Is Reinforcement Learning?* 2026. URL: <https://ww2.mathworks.cn/discovery/reinforcement-learning.html> (visited on 02/03/2026).
- [37] Julien Milli et al. “Performance of the extreme-AO instrument VLT/SPHERE and dependence on the atmospheric conditions”. In: *Proceedings of the Adaptive Optics for Extremely Large Telescopes 5*. AO4ELT5. Instituto de Astrofísica de Canarias (IAC), 2017. DOI: 10.26698/ao4elt5.0034. URL: <http://dx.doi.org/10.26698/AO4ELT5.0034>.
- [38] Dennis A. Montera et al. “Processing wave-front-sensor slope measurements using artificial neural networks”. In: *Applied Optics* 35.22 (1996), pp. 4238–4251. DOI: 10.1364/AO.35.004238.
- [39] Bartomeu Pou Mulet. “Supervised and reinforcement learning for extreme adaptive optics with application to exoplanet imaging”. PhD thesis. Université Paris sciences et lettres; Universitat politècnica de Catalunya ..., 2025.
- [40] Anusha Nagabandi et al. “Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning”. In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 7559–7566.
- [41] Nikolaj Buhl. *Convolutional Neural Networks (CNN) Explained*. 2023. URL: <https://encord.com/blog/convolutional-neural-networks-explained/> (visited on 02/03/2026).

- [42] Robert J Noll. “Zernike polynomials and atmospheric turbulence”. In: *Journal of the Optical Society of America* 66.3 (1976), pp. 207–211.
- [43] Jalo Nousiainen et al. “Adaptive optics control using model-based reinforcement learning”. In: *Optics Express* 29 (2021), pp. 15327–15344. DOI: [10.1364/OE.426470](https://doi.org/10.1364/OE.426470). URL: <https://doi.org/10.1364/OE.426470>.
- [44] Jalo Nousiainen et al. “Laboratory experiments of model-based reinforcement learning for adaptive optics control”. In: *Journal of Astronomical Telescopes, Instruments, and Systems* 10.1 (2024), p. 019001. DOI: [10.1117/1.JATIS.10.1.019001](https://doi.org/10.1117/1.JATIS.10.1.019001). URL: <https://doi.org/10.1117/1.JATIS.10.1.019001>.
- [45] Jalo Nousiainen et al. “Toward on-sky adaptive optics control using reinforcement learning: Model-based policy optimization for adaptive optics”. In: *Astronomy & Astrophysics* 664 (2022), A71. DOI: [10.1051/0004-6361/202243311](https://doi.org/10.1051/0004-6361/202243311). URL: <https://doi.org/10.1051/0004-6361/202243311>.
- [46] Francisco Oyarzun et al. “Performance comparison of the Shack-Hartmann and pyramid wavefront sensors with a laser guide star for 40 m telescopes”. In: *Astronomy & Astrophysics* 691 (2024), A205.
- [47] R. Paschotta. *Deformable Mirrors*. RP Photonics Encyclopedia. Available online at https://www.rp-photonics.com/deformable_mirrors.html. 2019. DOI: [10.61835/43m](https://doi.org/10.61835/43m). URL: https://www.rp-photonics.com/deformable_mirrors.html (visited on 06/15/2026).
- [48] C. Pernechele. “IMAGING | Adaptive Optics”. In: *Encyclopedia of Modern Optics*. Ed. by Robert D. Guenther. Elsevier, 2005, pp. 127–134. ISBN: 9780123693952. DOI: [10.1016/B0-12-369395-0/00702-8](https://doi.org/10.1016/B0-12-369395-0/00702-8). URL: <https://www.sciencedirect.com/science/article/pii/B0123693950007028>.
- [49] D. Peter et al. “PYRAMIR: Exploring the On-Sky Performance of the World’s First Near-Infrared Pyramid Wavefront Sensor”. In: *pasp* 122.887 (Jan. 2010), p. 63. DOI: [10.1086/649647](https://doi.org/10.1086/649647).
- [50] B. Pou et al. “Adaptive optics control with multi-agent model-free reinforcement learning”. In: *Optics Express* 30 (2022), pp. 2991–3015. DOI: [10.1364/OE.451270](https://doi.org/10.1364/OE.451270). URL: <https://doi.org/10.1364/OE.451270>.
- [51] Roberto Ragazzoni. “Pupil plane wavefront sensing with an oscillating prism”. In: *Journal of Modern Optics* 43.2 (Feb. 1996), pp. 289–293. DOI: [10.1080/09500349608232742](https://doi.org/10.1080/09500349608232742).
- [52] François Roddier. “Adaptive optics in astronomy”. In: (1999).
- [53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: vol. 9351. Oct. 2015, pp. 234–241. ISBN: 978-3-319-24573-7. DOI: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [54] D. G. Sandler et al. “Use of a neural network to control an adaptive optics system for an astronomical telescope”. In: *Nature* 351 (1991), pp. 300–302. DOI: [10.1038/351300a0](https://doi.org/10.1038/351300a0). URL: <https://doi.org/10.1038/351300a0>.
- [55] Tilman Spohn, Doris Breuer, and Torrence Johnson. *Encyclopedia of the Solar System*. 3rd. Amsterdam: Elsevier, 2014. ISBN: 978-0124158450.
- [56] Arnaud Striffling et al. “Using a pyramid wavefront sensor off-zero: from theory to practice”. In: Aug. 2024, p. 58. DOI: [10.1117/12.3019000](https://doi.org/10.1117/12.3019000).

- [57] Robin Swanson et al. “Closed loop predictive control of adaptive optics systems with convolutional neural networks”. In: *Monthly Notices of the Royal Astronomical Society* 503.2 (2021), pp. 2944–2954.
- [58] Robin Swanson et al. “Wavefront reconstruction and prediction with convolutional neural networks (Conference Presentation)”. In: July 2018, p. 52. DOI: [10.1117/12.2312590](https://doi.org/10.1117/12.2312590).
- [59] Cedric Taissir Heritier, Christophe Verinaud, and Carlos M. Correia. “OOPAO: Object Oriented Python Adaptive Optics”. In: *Adaptive Optics for Extremely Large Telescopes (AO4ELT7)*. June 2023, 50, p. 50. DOI: [10.13009/AO4ELT7-2023-052](https://doi.org/10.13009/AO4ELT7-2023-052).
- [60] Tutorialspoint. *Architecture of Transformers in Generative AI*. 2025. URL: <https://www.tutorialspoint.com/gen-ai/architecture-of-transformers-in-generative-ai.htm> (visited on 02/03/2026).
- [61] Christophe Vérinaud. “On the nature of the measurements provided by a pyramid wavefront sensor”. In: *Optics Communications* 233.1 (2004), pp. 27–38. ISSN: 0030-4018. DOI: <https://doi.org/10.1016/j.optcom.2004.01.038>. URL: <https://www.sciencedirect.com/science/article/pii/S0030401804000628>.
- [62] Camilo Weinberger et al. “Transformer neural networks for closed-loop adaptive optics using nonmodulated pyramid wavefront sensors”. In: *Astronomy & Astrophysics* 687 (2024), A202. DOI: [10.1051/0004-6361/202349118](https://doi.org/10.1051/0004-6361/202349118). URL: <https://doi.org/10.1051/0004-6361/202349118>.
- [63] Alison P. Wong et al. “Predictive control for adaptive optics using neural networks”. In: *Journal of Astronomical Telescopes, Instruments, and Systems* 7, 019001 (Jan. 2021), p. 019001. DOI: [10.1117/1.JATIS.7.1.019001](https://doi.org/10.1117/1.JATIS.7.1.019001).

Appendix A

SPIE Poster



INTRODUCTION

The European Extremely Large Telescope (ELT) will contain the yet-to-be developed Planetary Camera and Spectrograph (PCS), an instrument that has as its main mission the detection and study of exoplanets. This requires a high-performance extreme adaptive optics (XAO) system with fast loop rates, pyramid wavefront sensors, and the control of multiple deformable mirrors (DMs). These requirements introduce several challenges, including nonlinear sensor response, higher system complexity and demanding hardware that not-yet exists.

In this work, we investigate reinforcement learning (RL) as a possible solution to some of these challenges, testing it in two candidate PCS control architectures: Cascade and Woofer-Tweeter.

METHODS

We apply the RL algorithm PO4AO [2] to two dual-mirror control architectures for the PCS [3]: Cascade (CAO) and Woofer-Tweeter (WT).

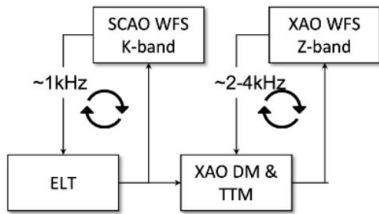


Fig. 1: Concept diagram of PCS Cascade control.

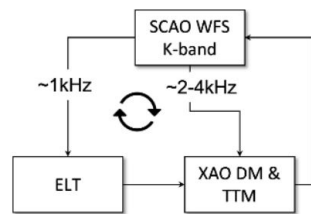


Fig. 2: Concept diagram of PCS Woofer-Tweeter control.

PO4AO consists of two main blocks:

- Dynamics model: learns the environment and predicts its evolution from previous states and actions
- Policy model: selects the best actions based on the current state, its trained on the dynamic model.

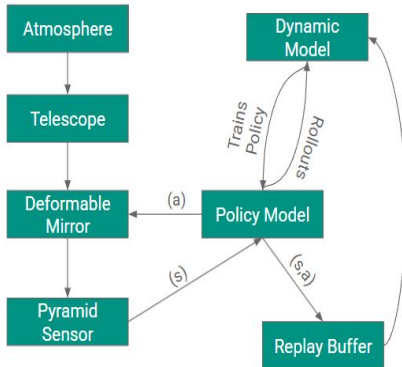


Fig. 3: PO4AO simplified diagram.

ADVANTAGES

- **Reduces recalibration needs:** as the models learn through interaction with the environment, they adapt continuously with the system.
- **Predicts system evolution:** by modelling the system dynamics, RL can anticipate future distortions, reducing servo-lag errors.
- **Faster convergence:** using the dynamics model to train the policy enables faster convergence with fewer environment interactions, essential in environments in constant motion.

SIMULATION SETUP

Simulations were carried out using OOPAO [1] on an ELT-sized dual-mirror system:

- First DM: 80×80 actuators at 1000 Hz loop rate.
- Second DM: 120×120 actuators at 2000 Hz and 4000 Hz loop rates.
- Control: RL in both configurations only directly controls the second DM.

RL contrast performance is compared against a classical linear reconstructor(RC).

RESULTS

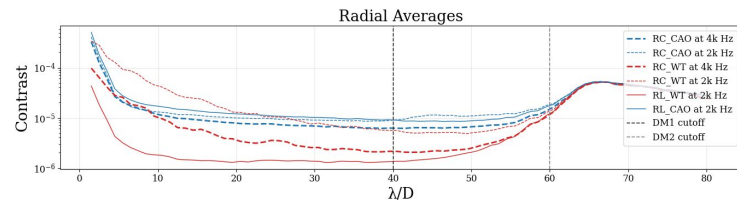


Fig. 4: Contrast for all experimental configurations.

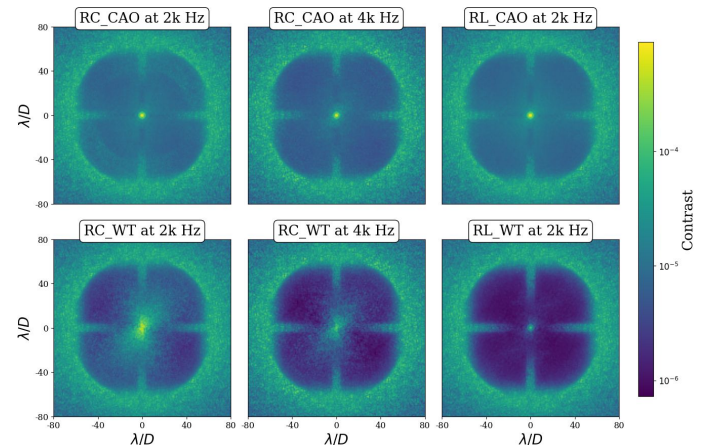


Fig. 5: Ideal post-coronagraphic image for all experimental configurations.

In Woofer-Tweeter architecture:

- The RL controller outperforms the classical integrator operating at twice the loop frequency by a factor of 2–6 at small angular separations.
- This improvement appears to come from correcting servo-lag errors, visible in Fig. 5 by the butterfly effect.

In Cascade architecture:

- Performance may be degraded by overcorrection, because the first-stage action may already be accounted for by the faster second stage, leading to overcorrection when new commands are applied.
- Also may come from the RL not having full causality over both DMs states, needing to adapt to both the environment's evolution and the contribution of the first-stage integrator.

OUTLOOK

- Cross-check and benchmark Woofer-Tweeter and Cascade performance against other simulations.
- Resolve inter-stage interaction issues in the Cascade architecture.

REFERENCES

1. Heritier et al, OOPAO: Object Oriented Python Adaptive Optics, 2023
2. Nousiainen et al, Toward on-sky adaptive optics control using reinforcement learning, 2022
3. Kasper et al, PCS — A Roadmap for Exoearth Imaging with the ELT, 2021

Appendix B

SPIE Proceedings

Assessing Reinforcement Learning in XAO Control for PCS

Miguel Santos^{a,c}, Ana F. Sequeira^{a,b}, and Carlos M. Correia^{a,b}

^aFaculty of Engineering, University of Porto, Porto, Portugal

^bINESCTEC, Porto, Portugal

^cCENTRA, Center for Astrophysics and Gravitation, IST, University of Lisbon, Lisbon, Portugal

ABSTRACT

The 40m Extremely Large Telescope (ELT), now under construction, will rely on the Planetary Camera and Spectrograph (PCS) to detect and study exoplanets, requiring an exceptionally fast and robust eXtreme Adaptive Optics (XAO) system featuring an unprecedented number of degrees of freedom with as-yet-unproven multi-mirror wavefront correction.

This work aims to investigate the application of Reinforcement Learning (RL) to improve performance and robustness while relaxing the real-time control requirements of the PCS's XAO system.

To do this, we apply the RL algorithm PO4AO (Nousiainen et al. 2022) adapted to work on multi-mirror control systems at the ELT's dimensionality, with further modifications to improve robustness and account for the system's inherent two-frame delay. We tested it with two candidate PCS control architectures: woofer-tweeter (WT), featuring one WFS and two DMs, and cascade(CAO), with two sequential control loops, each with its own WFS and DM. The performance of each scheme is evaluated across a range of system complexities, from a scaled-down configuration on a 10 m-class telescope to a full 40 m ELT-sized simulation, in order to assess the scalability of the RL approach.

The simulations show that, in most tested cases, the RL controller achieves contrast equal to or better than least-squares reconstruction + integrator control operating at twice the loop frequency. Our preliminary results show that the best overall contrast is obtained with RL in the woofer-tweeter architecture, exceeding RL in the cascade architecture by a factor of 1.5 to 3 at small angular separations, which we attribute to the RL controller's ability to operate in a fully causal mode in the WT configuration but not in CAO.

Keywords: Extremely Large Telescope, Extreme Adaptive Optics, Reinforcement Learning, Planetary Camera and Spectrograph

1. INTRODUCTION

Direct imaging of exoplanets remains one of the most demanding challenges in astronomy.¹ This challenge comes from the extreme brightness contrast between a planet and its host star, which can exceed 10^9 at visible wavelengths for Earth-like rocky planets.¹ In addition, the small angular separation between the two sources causes the starlight to dominate the image, obscuring the faint planetary signal.

For ground-based telescopes, this challenge is amplified by atmospheric turbulence, which distorts incoming wavefronts and spreads the starlight across the focal plane. As a result, the already weak planetary signal becomes further buried within the residual stellar glare. High-contrast imaging systems therefore rely on Adaptive Optics (AO) to correct these distortions in real time and restore diffraction-limited performance. However, conventional AO architectures are insufficient to meet the correction accuracy and temporal responsiveness required for direct planet imaging. Servo-lag is one of the dominant error terms in this context, and achieving extreme AO correction at the visible wavelengths targeted by the PCS requires a very high density DM to limit fitting error, improve the Strehl ratio, and ensure the coronagraph can operate effectively.

Further author information:

M.S.: E-mail: mfmnds@outlook.com

To address this problem, the Extremely Large Telescope (ELT), currently under construction, will include a yet-to-be-developed Planetary Camera and Spectrograph (PCS)¹ dedicated to the detection and characterisation of exoplanets. To achieve this, it will employ a multi-mirror eXtreme Adaptive Optic (XAO) system with a very high actuator density, operating at frame rates up to four times higher than the highest currently available (i.e., up to 4 kHz), with two of the proposed schemes being woofer-tweeter (WT) and cascade (CAO) configurations. These specifications significantly exceed those of existing systems and introduce substantial technological challenges, as the required hardware does not yet exist.

Recent research has explored the use of Artificial Intelligence (AI) for AO control, with various works focusing on improving wavefront reconstruction to enhance correction, as well as wavefront prediction to compensate for the system’s inherent delay. RL, in particular, has shown great promise for this application, offering the ability to learn control policies directly from interaction with the system and to continuously adapt to changing observing conditions. These methods can be divided into two main categories: model-free methods,^{2–4} which learn purely through trial and error while interacting with the environment, and typically offer faster inference times but lower sample efficiency, as the algorithms used tend to be less complex; and model-based methods,⁵ which learn an explicit representation of the system dynamics and use it for planning and control, generally achieving higher sample efficiency but at the cost of slower inference, as they tend to be more complex in order to learn the system representation. These two approaches can be combined into a hybrid strategy⁶ that achieves both fast inference and high sample efficiency, as the model-based algorithm is used to train, or optimise, a model-free policy that then performs the real-time control.

Although these approaches have shown promising results, only a few studies^{7,8} have begun testing them in multi-mirror control XAO systems like the PCS, and even then only for comparatively simpler configurations. The aim of this work was to investigate the application of RL to two proposed schemes for the PCS’s XAO system and test their scalability to ELT-sized configurations, with the objective of relaxing hardware and real-time control requirements while maintaining or improving contrast performance.

2. METHODOLOGY

To test the scalability of the method, we conduct simulations at three scales: a scaled-down system, an intermediate-scale system, and an ELT-sized system. Each of these systems consists of two deformable mirrors (DMs), the second of which has a higher actuator density than the first, and operates at a loop frequency of either 2 kHz or 4 kHz while the first operates at 1 kHz.

As stated previously, we evaluate the method on two control architectures proposed for the PCS: cascade control, in which there are two independent correction loops, each with its own wavefront sensor (WFS) and deformable mirror; and woofer-tweeter control, in which a single wavefront sensor drives both mirrors, with low-frequency components offloaded from the second mirror to the first at a rate of 1 kHz.

We performed all simulations using the OOPAO framework.⁹ The full simulation parameters can be found in Table 1.

2.1 Algorithm Setup

We use the PO4AO algorithm,⁶ which combines a dynamics model and a policy model. The dynamics model learns the evolution of the environment from past observations and actions; the policy model, also conditioned on past observations and actions, is optimized on the dynamics model to output the action that maximizes correction. A full run of the algorithm proceeds as:

1. Run a noisy integrator to collect initial data. Noise is deliberately injected to increase sample diversity.
2. Train the dynamics model on the collected data.
3. Use the dynamics model to optimize the policy model.
4. During an episode, collect additional data so that both models can be trained in parallel at the episode end.

5. Repeat steps 2 – 4 until the experiment ends.

To adapt the original code¹⁰ to work with our simulations, we needed to change the way it interacted with the environment. This was achieved by modifying the `step` function so that it sent the action, updated the simulation state, and returned the new measurements.

Table 1. Setup Parameters.

Telescope		
Parameter	Value	Units
Telescope diameter	10/20/40	m
Obstruction ratio	14/14/30	percent
Active actuators, 1 ^o DM	356/1332/4788	actuators
Active actuators, 2 ^o DM	768/2920/10612	actuators
1 ^o PWFS subapertures (CAO)	$20 \times 20/40 \times 40/120 \times 120$	apertures
2 ^o PWFS subapertures (CAO,WT)	$30 \times 30/60 \times 60/120 \times 120$	apertures
PWFS modulation	2/2/4	λ/D
DM coupling	0.3	percent
DM influence functions	Gaussian	–
1 ^o WFS wavelength (CAO)	1.7	μm
2 ^o WFS wavelength (CAO WT)	1.7/0.79 1.7	μm
Science camera wavelength	0.79	μm

Atmosphere		
Parameter	Value	Units
Fried parameter	15	cm at 500 nm
Number of layers	3	–
Layer altitudes	0/4/10	km
C_n^2	50/35/15	percent (%)
Wind speeds	10/26/35	m/s
Wind directions	0/45/180	degrees
L_0	25	m

2.2 Algorithm Modifications

Like most reinforcement learning algorithms, PO4AO can be highly sensitive to hyperparameter values and typically requires time-consuming tuning to remain robust across diverse seeing conditions. To reduce this tuning factor while preserving performance, we introduce the following modifications:

1. The dynamics model target was changed from predicting the next observation to predicting the change in observation.
2. The policy model loss function was adapted to account for the system’s two-frame delay, such that the observation used to compute the loss at time step t is obtained by sending the action a_{t-1} to the dynamics model instead of the action a_t , better matching what happens in the real system.
3. A smoothing term defined as $\lambda \|a_t - a_{t-1}\|^2$, was added to the loss function to counteract the instability introduced by the previous modification, ensuring that two consecutive actions cannot differ too greatly from one another.

4. Random actuator-space noise applied during the initial data collection was replaced with mode-based noise for more physically meaningful exploration.

The effect of these adaptations can be seen in Fig. 1; the increase in performance comes mainly from the ability to control more modes without diverging, going from 400 Karhunen–Loève (KL) modes in the original implementation to 750 with all modifications applied.

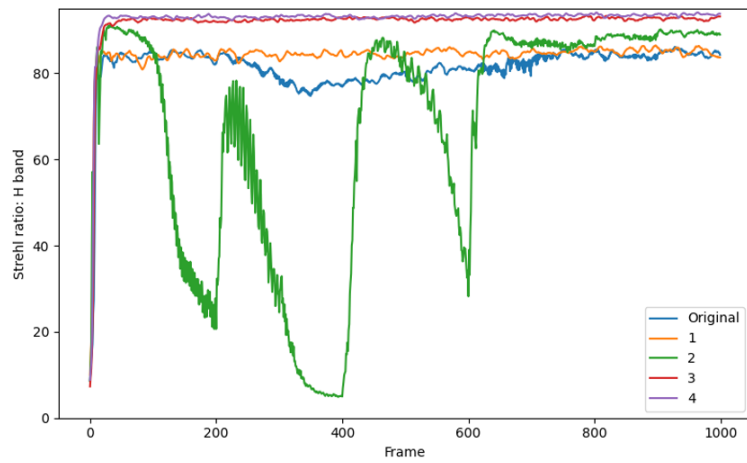


Figure 1. Effect of the algorithm adaptations on performance.

2.3 AO Control Architectures

For the cascade architecture, each run first operates the first stage alone until convergence, after which the second stage is engaged. In the woofer-tweeter architecture, the controller acts directly only on the second DM, and part of its correction is offloaded to the first DM at a rate of 1 kHz, as described above.

For all experiments, the noisy integrator used to collect the initial data is run for 10 episodes of 200 steps each, followed by 50 evaluation episodes, each also of 200 steps.

We evaluate the performance based on contrast. Given that, at the time of this work, the simulation does not include a coronagraph, we approximate the post-coronagraph contrast using the Power Spectral Density (PSD), as per Eq. (1):

$$C(\mathbf{r}) = \frac{\text{PSD}(\mathbf{r})}{I_{\text{corrected}}(0, 0)}, \quad (1)$$

where $I_{\text{corrected}}(0, 0)$ is the peak intensity of the point spread function after AO correction.

The main reason behind this approximation is that, for a perfect coronagraph, which removes all coherent light, the main contribution to contrast degradation would be the residual wavefront errors.¹¹

3. RESULTS

In this section, we test the algorithm across three different scales, as mentioned in Section 2, in order to demonstrate the scalability of the approach.

In all cases, the performance of the algorithm is compared against a least-squares minimization reconstruction (RC) algorithm. The results shown here correspond only to the RC running at 4 kHz. To simplify the nomenclature, results are abbreviated in the following format: first, the type of controller (RC or RL); then, the architecture used (CAO or WT); and, in the case of CAO, followed by the band used on the second WFS (H or I).

3.1 Scaled-down

In Figure 2, we observe that RL_WT, RL_CAO_I, and RC_CAO_I all achieve similar performance, with an improvement factor of 2 to 4 at smaller angular separations compared to RC_WT. Both controllers on CAO_H outperform the previous configurations by a factor of 1.5.

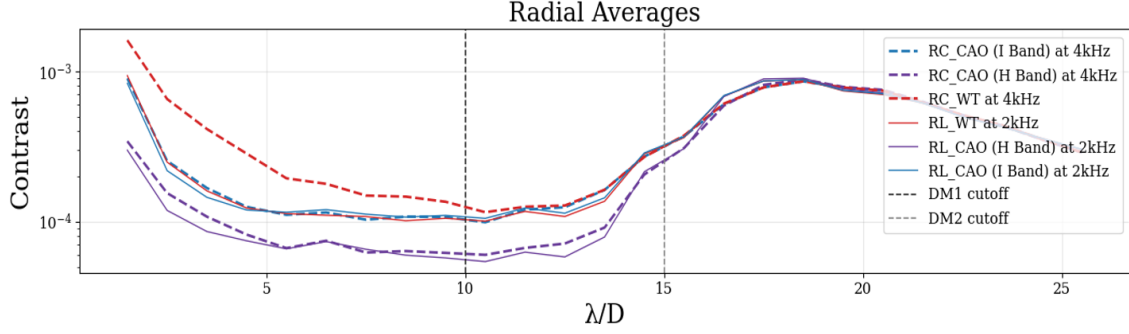


Figure 2. Contrast for the Scaled-down configuration.

3.2 Intermediate

As observed in Figure 3, two patterns begin to emerge. First, although RC_WT performs worse than RC_CAO_I and RL_CAO_I at smaller separations, it performs better at larger separations, by up to a factor of 2, becoming comparable to RC_CAO_H. Second, both RL_CAO configurations perform similarly to their respective RC_CAO counterparts, while RL_WT shows an improvement factor of 3 to 9 at smaller separations compared to RC_WT, and a factor of 1.5 to 2 compared to RL_CAO_H, making it the best performer overall.

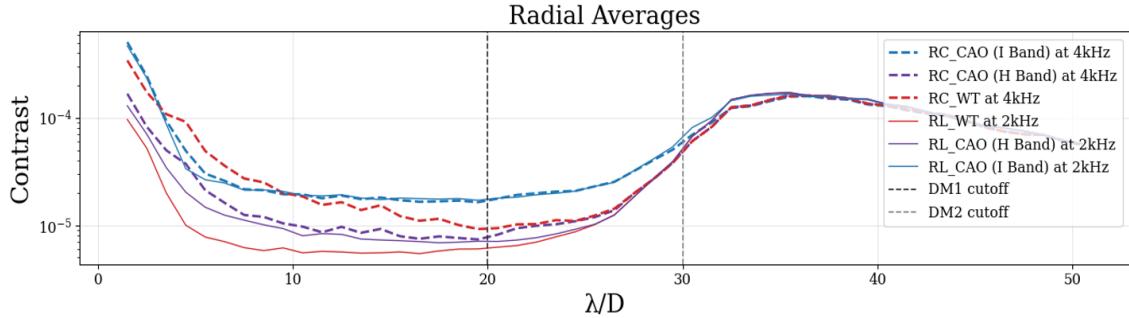


Figure 3. Contrast for Intermediate configuration.

3.3 ELT-sized

It should be noted that, although this experiment is ELT-sized, it is not yet a true representation of the final ELT configuration. Several system parameters, including the DM geometry, the number of WFS subapertures, and other relevant design choices, have not yet been optimized to match the final telescope architecture.

Figure 4 shows that, for the ELT-sized simulation, the previously mentioned patterns are maintained: RC_WT achieves even higher performance than RC_CAO_I, approaching that of RC_CAO_H, while RL_WT still maintains the best overall performance, with a factor of 1.5 to 3 at smaller separations compared to the second-best performers (both controllers on CAO_H). The new divergence in these results is that, this time, RL_CAO_I did not achieve performance similar to or better than RC_CAO_I.

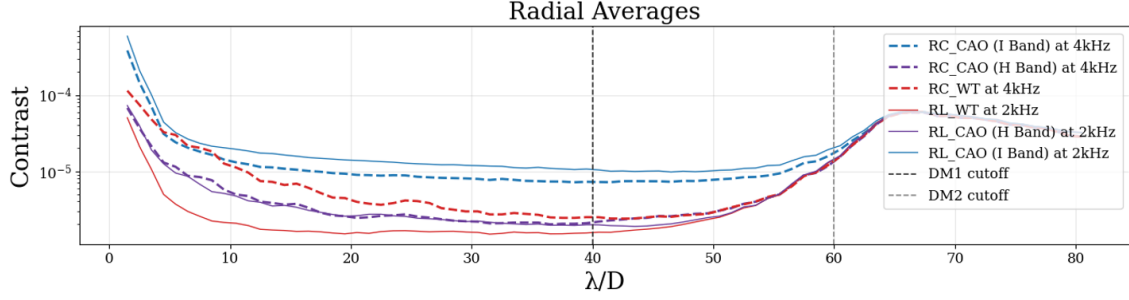


Figure 4. Contrast for the ELT-sized configuration.

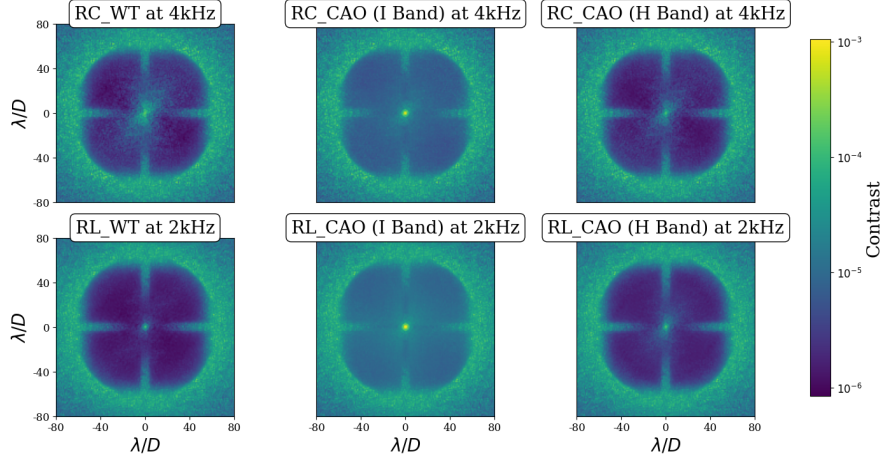


Figure 5. Ideal post-coronagraphic images for the ELT-sized configuration.

4. DISCUSSION AND OUTLOOK

This work aimed to apply reinforcement learning to improve the performance of multi-mirror XAO

control while relaxing hardware requirements. Overall, the results show that this objective was achieved: in most tested configurations, RL matched or outperformed a classical approach operating at twice the loop frequency, suggesting that hardware requirements can indeed be relaxed.

An interesting pattern emerged from the comparison between architectures. With a classical approach, the CAO configuration typically yields better performance than the WT configuration at smaller angular separations. This can be attributed to the fact that CAO, consisting of two independent single-integrator loops in series, effectively behaves as a double integrator, providing stronger rejection of residual errors at those spatial frequencies than a single integrator alone. Our WT implementation, by contrast, only a single integrator is used, which inherently limits its rejection performance relative to CAO. This is not a fundamental limitation of the WT architecture itself: comparable rejection can in principle be achieved with a more sophisticated classical WT control law.¹²

However, this advantage did not carry over to the RL controllers, which instead performed better in the WT configuration. This behaviour can be explained by a difference in causality between the two architectures. In the WT configuration, although the RL controller only acts on the second DM, it has full causality of both DM's states, so the algorithm only needs to adapt to the evolution of the environment. In the CAO configuration, by contrast, the RL controller must adapt both to the evolution of the environment and to the effect of the independent, first-stage controller. It should also be noted that the CAO tests did not employ a fully tuned first-stage controller; further tuning could narrow the performance gap with the WT configuration.

Regarding the classical controllers, the operating band of the second wavefront sensor had a strong effect on the performance of the CAO architecture, contrary to expectations: the CAO configuration performed better in

H band than in I band, despite I band having a shorter wavelength. This can be explained by the fact that, although diffraction-limited performance should improve at shorter wavelengths, the effect of turbulence on the final correction also increases as wavelength decreases, and does so at a higher rate than the corresponding gain from operating at shorter wavelengths. This is why the performance of the WT configuration converges toward that of the CAO configuration in H band as the scale increases, since its sensor also operates in H band. This same effect can be seen in other works,¹³ where, despite different system characteristics and optimizations, the classical approach still achieves better performance in H band compared to I band.

Future work could proceed along three main avenues. The first is to adapt the RL framework to control both DMs simultaneously in the CAO architecture, in order to address the causality issue described above. The second is to adapt the simulations to more closely match the full ELT configuration, not only in terms of scale, but also in terms of pupil geometry and other relevant system characteristics. Finally, we could apply RL to other promising multi-mirror control schemes, such as integrated AO,¹⁴ which, like CAO, has two DMs and two wavefront sensors, but where both wavefront sensors see the full correction of both DMs, rather than one sensor seeing only the residual of the other stage's correction.

ACKNOWLEDGMENTS

This work acknowledges the use of the Claude Sonnet 5 model for language and grammar clean-up.

REFERENCES

- [1] Kasper, M., Cerpa Urra, N., Pathak, P., Bonse, M., Nousiainen, J., Engler, B., Heritier, C. T., Kammerer, J., Leveratto, S., Rajani, C., Bristow, P., Le Louarn, M., Madec, P.-Y., Ströbele, S., Verinaud, C., Glauser, A., Quanz, S. P., Helin, T., Keller, C., Snik, F., Boccaletti, A., Chauvin, G., Mouillet, D., Kulcsár, C., and Raynaud, H.-F., “PCS — A Roadmap for Exoearth Imaging with the ELT,” *The Messenger* **182**, 38–43 (Mar. 2021).
- [2] Landman, R., Haffert, S. Y., Radhakrishnan, V. M., and Keller, C. U., “Self-optimizing adaptive optics control with reinforcement learning for high-contrast imaging,” (2021).
- [3] Mulet, B. P., *Supervised and reinforcement learning for extreme adaptive optics with application to exoplanet imaging*, PhD thesis, Université Paris sciences et lettres; Universitat politècnica de Catalunya ... (2025).
- [4] Landman, R., Haffert, S. Y., Radhakrishnan, V. M., and Keller, C. U., “Self-optimizing adaptive optics control with reinforcement learning,” in [Adaptive Optics Systems VII], Schreiber, L., Schmidt, D., and Vernet, E., eds., *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **11448**, 1144849 (Dec. 2020).
- [5] Nousiainen, J., Rajani, C., Kasper, M., and Helin, T., “Adaptive optics control using model-based reinforcement learning,” *Optics Express* **29**, 15327–15344 (2021).
- [6] Nousiainen, J., Rajani, C., Kasper, M., Helin, T., Haffert, S. Y., Verinaud, C., Males, J. R., Van Gorkom, K., Close, L. M., Long, J. D., Hedglen, A. D., Guyon, O., Schatz, L., Kautz, M., Lumbres, J., Rodack, A., Knight, J. M., and Miller, K., “Toward on-sky adaptive optics control using reinforcement learning: Model-based policy optimization for adaptive optics,” *Astronomy & Astrophysics* **664**, A71 (2022).
- [7] Nousiainen, J., Engler, B., Kasper, M., Rajani, C., Helin, T., Heritier, C. T., Quanz, S. P., and Glauser, A. M., “Laboratory experiments of model-based reinforcement learning for adaptive optics control,” *Journal of Astronomical Telescopes, Instruments, and Systems* **10**(1), 019001 (2024).
- [8] Krokberg, T., “Reinforcement learning in multi-mirror adaptive optics,” (2022).
- [9] Taissir Heritier, C., Verinaud, C., and Correia, C. M., “OOPAO: Object Oriented Python Adaptive Optics,” in [Adaptive Optics for Extremely Large Telescopes (AO4ELT7)], 50 (June 2023).
- [10] jnousi, “PO4AO.” <https://github.com/jnousi/PO4AO> (2023). Accessed: 2026-06-07.
- [11] Guyon, O., “Limits of adaptive optics for high-contrast imaging,” *The Astrophysical Journal* **629**, 592–614 (08 2005).
- [12] Dinis, I., Galland, N., Raynaud, H.-F., Béchet, C., Tallon, M., Tallon-Bosc, I., Langlois, M., Ferreira, F., Boccaletti, A., Wildi, F., Ségransan, D., Kasper, M., and Kulcsár, C., “SAXO+, the second-stage adaptive optics for SPHERE: single vs cascade control loop performance assessment,” in [Adaptive Optics Systems X], (14150-272). (July 7, 2026).

- [13] Correia, C. M., Langlois, M., Kasper, M., Taras, A., Rahim, A., Boccaletti, A., Sevin, A., Magniez, A., Graf, C., Kulcsár, C., Bond, C. Z., Gratadour, D., Diolaiti, E., Cortecchia, F., Ferreira, F., Di Rico, G., Orban de Xivry, G., Raynaud, H.-F., Nousiainen, J., Schreiber, L., Desdoigts, L., N'Diaye, M., Chebbo, M., Feldt, M., Lombini, M., Tallon, M., Le Louarn, M., Galland, N., Schwartz, N., Vermot, P., Landman, R., Briguglio, R., Haffert, S. Y., Bertram, T., Morris, T. J., Fonteneau, V., and Chambouleyron, V., “Closing the technology gap for the Extremely Large Telescope’s Planetary Camera and Spectrograph (ELT-PCS): its extreme adaptive optics system conceptual design,” in [*Adaptive Optics Systems X*], (July 9, 2026).
- [14] Menessini, M., Busoni, L., Esposito, S., and Agapito, G., “Performance comparison of multistage ExAO systems for high contrast imaging,” in [*Adaptive Optics Systems X*], (14150-311). (July 8, 2026).