JATIS Journal of Astronomical Telescopes, Instruments, and Systems

RESEARCH PAPER

Machine-learning approach for optimal self-calibration and fringe tracking in photonic nulling interferometry

Barnaby R. M. Norris[a,b,c,*] **Marc-Antoine Martinod**[d] **Peter Tuthill**[a,b,c] **Simon Gross**[e] **Nick Cvetojevic**[f] **Nemanja Jovanovic**[g] **Tiphaine Lagadec**[a] **Teresa Klinner-Teo**[a] **Olivier Guyon**[h] **Julien Lozi**[h] **Vincent Deo**[h] **Sebastien Vievard**[h] **Alex Arriola**[e] **Thomas Gretzinger**[e] **Jon S. Lawrence**[c] **and Michael J. Withford**[e]

[a]University of Sydney, Sydney Institute for Astronomy, School of Physics, Sydney, New South Wales, Australia
[b]University of Sydney, School of Physics, Sydney Astrophotonic Instrumentation Laboratories, Sydney, New South Wales, Australia
[c]Astralis, Astronomical Instrumentation Consortium, Sydney, New South Wales, Australia
[d]KU Leuven, Institute of Astronomy, Leuven, Belgium
[e]Macquarie University, MQ Photonics Research Centre, Department of Physics and Astronomy, Sydney, New South Wales, Australia
[f]Université Côte d'Azur, Laboratoire Lagrange, Observatoire de la Côte d'Azur, Nice, France
[g]California Institute of Technology, Pasadena, California, United States
[h]National Institutes of Natural Sciences, Subaru Telescope, National Astronomical Observatory of Japan, Hilo, Hawaii, United States

ABSTRACT. Photonic technologies have enabled a generation of nulling interferometers, such as the guided light interferometric nulling technology instrument, potentially capable of imaging exoplanets and circumstellar structure at extreme contrast ratios by suppressing contaminating starlight, and paving the way to the characterization of habitable planet atmospheres. But even with cutting-edge photonic nulling instruments, the achievable starlight suppression (null-depth) is only as good as the instrument's wavefront control and its accuracy is only as good as the instrument's calibration. Here, we present an approach wherein outputs from non-science channels of a photonic nulling chip are used as a precise null-depth calibration method and can also be used in real time for fringe tracking. This is achieved using a deep neural network to learn the true *in-situ* complex transfer function of the instrument and then predict the instrumental leakage contribution (at millisecond timescales) for the science (nulled) outputs, enabling accurate calibration. In this method, this pseudo-real-time approach is used instead of the statistical methods used in other techniques (such as null self calibration, or NSC) and also resolves the severe effect of read-noise seen when NSC is used with some detector types.

Keywords: nulling; photonics; machine learning; calibration; null self-calibration; fringe tracking

Paper 22118G received Dec. 4, 2022; revised Oct. 27, 2023; accepted Oct. 30, 2023; published Nov. 30, 2023.

1 Introduction

Nulling interferometry is a key technology in the quest to directly image high-contrast objects at angular resolutions at and higher than the telescope diffraction limit, such as the case of directly

*Address all correspondence to Barnaby R. M. Norris, barnaby.norris@sydney.edu.au

imaging exoplanets in the habitable zone. As with conventional interferometry, light from separate telescopes or sub-apertures is coherently combined, and the visibility and phase of the resulting fringes is used to determine the source intensity map of the target. But in nulling interferometry, differential phase delays are carefully tuned such that the central star is subject to maximal destructive interference, removing its otherwise overwhelming photon noise and allowing the faint, off-axis science object's light to be detected.

Since the concept was originally suggested,[1] a wide variety of implementations have been proposed and realized, including multiple re-combinations of baselines[2] and multi-element space-based instruments,[3] while applications have been extended to the detection of exo-zodiacal disks.[4] A standard mathematical formalism has also been developed.[5] Notable instruments that employ nulling interferometry, such as the Keck Interferometer Nuller[6] and the Large Binocular Telescope Interferometer,[7] perform the necessary manipulation and interference of light using conventional bulk optics. However, spatial structure in the wavefront, induced by seeing, limits the null depths achievable with these methods, and restricts the types of output signals accessible. Subsequently, photonic technologies, using either single-mode fibers (such as with the Palomar Fibre Nuller[8,9]) or a more complex set of waveguides inscribed within a photonic chip [such as the guided light interferometric nulling technology (GLINT) nuller[10–13]], were used to create nulling instruments. Here, the single-mode nature of the waveguides removes all higher-order spatial structure, with the null created and controlled entirely by a single phase and amplitude value for each input (for a given wavelength and polarization). Photonic chips enable sophisticated architectures with multiple beam combiners and splitters, and multiple simultaneous outputs encoding photometry, bright (constructive interference) channels, and so on.

There are two central challenges to be met in nulling interferometry: (1) creating and maintaining a deep null and (2) calibrating the null depth. The former is critical to achieve maximum suppression of stellar photon noise and is dependent both on instrument and photonic chip design and dynamically on fringe tracking and wavefront correction. The latter of these challenges—null-depth calibration, which is essential for science measurements to be made—will be the main focus of this paper.

2 Null-Depth Calibration Challenge

2.1 Contributions to Null Depth

For an ideal nuller, the light from an unresolved source would be perfectly nulled. It would be entirely coherent so its fringe visibility would be unity and with the appropriate phase delay applied to a baseline destructive interference would be complete and no light would emerge from the "null" output of the instrument.

Any spatial extension of the source, however, would reduce the degree to which the light could be destructively interfered (since the source is now partly spatially incoherent), and some light would emerge through the instrument's "null" output no matter the phase offset applied. This null-depth N is the key science observable and is defined as

$$N = \frac{I_-}{I_+}, \tag{1}$$

where I_- and I_+ are the intensity of the destructive and constructive fringes, respectively. This is fundamentally the same property as the visibility V familiar to interferometrists,[14] and the two are related[8] as per

$$N = \frac{1 - |V|}{1 + |V|}. \tag{2}$$

In this ideal model, the phase-delay across a baseline would be adjusted until the starlight is maximally reduced, and the residual null-depth then measured to provide an interferometric measurement of the source intensity distribution. As with conventional interferometry, null-depths from multiple baseline lengths and angles could be used to construct a more detailed image of the source—all free from the host star's polluting photon noise.

However, in real life things are not so straightforward. Much of the light that emerges from the "nulled" output is not due to spatial incoherence (the science signal) but due to instrumental

leakage—that is, starlight that has not been fully nulled due to wavefront and instrumental effects. This instrumental leakage term arises from constant sources (non-ideal beam combiner design/fabrication, chromatic dependencies, asymmetric throughputs, etc) and rapidly varying sources arising from seeing. These variable components are particularly problematic as they cannot simply be calibrated out using a laboratory characterization of the optical/photonic system. The instantaneous null depth is a function of the differential phase and differential amplitude across each baseline, both of which are being rapidly modulated by uncorrected seeing (note that for a single-mode photonic device, injection efficiency is a strong function of wavefront error (WFE), and so rapidly varying baseline amplitude is a significant component). In some cases, differential polarization can also be a source of leakage, though in the GLINT instrument light is passed through a common linear polarizer prior to injection to avoid this.

The instrumental leakage term can easily be of the same magnitude as the science signal, so obtaining a useful science measurement is contingent on accurately knowing the leakage term. If you know the leakage then you know the true astrophysical null, and for small astrophysical nulls, it can be shown[15] that the observed null depth N_{obs} is given as

$$N_{\text{obs}} = N_{\text{astro}} + N_{\text{inst}}, \tag{3}$$

where N_{astro} and N_{inst} are the true astrophysical null and instrumental leakage terms, respectively.

The classical method to calibrate the null depth was to observe a separate point spread function (PSF) reference star, as is common in interferometry, measure its average observed null depth, and subtract this from the average observed null depth of the science target. But this assumes that all properties of the seeing, the telescope, adaptive optics (AOs) system, etc., remain identical between these observations and it has been shown[15] that this is not an accurate method. Instead, recent nulling interferometry has made use of a different technique, null self-calibration (NSC).

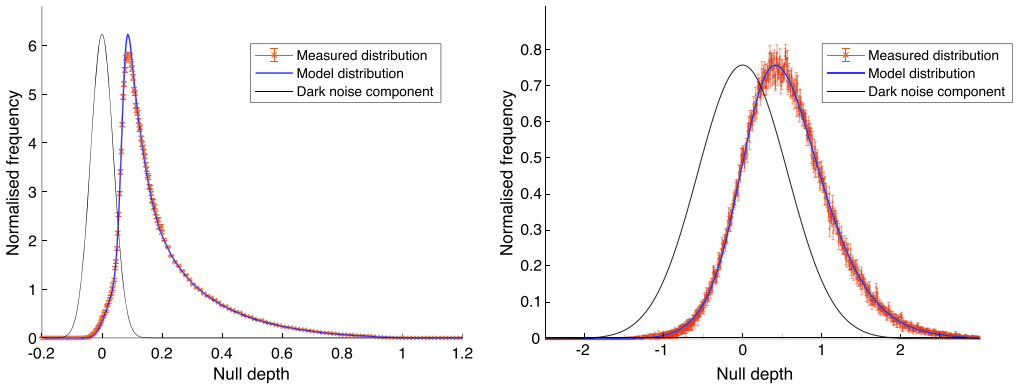
2.2 Null Self-Calibration

The NSC method[7,10,15] is a statistical method, relying on the fact that no matter the applied wavefront and amplitude errors the observed null depth cannot go deeper than the fundamental limit imposed by the sources spatial incoherence. A histogram of the null-depths for an entire observation is calculated and this is compared with a probability distribution function (PDF) created by a forward model. This model must include a priori knowledge of the chip's various chromatic coupling coefficients, etc., and can also draw on simultaneous or quasi-simultaneous on-sky measurements of injection efficiency (from the chip's photometric outputs) and detector noise/background (via chopping).

The model's predicted PDF of the observation is then fitted to the observed histogram by fitting the model's remaining free parameters. This includes the differential phase error across the baseline—a dominant source of time-varying instrumental leakage. This is assumed to be normally distributed over the observation and so is simply fitted with a mean and a standard deviation parameter. The quantity of interest—the astrophysical null-depth—is also fitted. Figure 1(a) shows an example of the observed histogram and fitted NSC PDF for GLINT observations of α Tau.[10]

However, NSC has some important limitations. First, it assumes that the differential phase errors are normally distributed and so can be parameterized by a single mean and standard deviation, an assumption that does not match the reality of the residual WFE from a complex AO and/or fringe-tracking system. Moreover, it has been found that fitting these two parameters to the observed PDF is somewhat degenerate with other noise processes,[10] especially in low signal-to-noise ratio (SNR) regimes.

This assumption of normally distributed phase errors is especially problematic in the presence of low-wind-effect/island modes/petaling modes[16–19] (hereafter referred to as LWE). These severe aberrations are caused by phase discontinuities across the spiders (secondary-mirror supports) in the telescope pupil, exacerbated by thermal effects that these structures create when the wind is low. This phase shear causes a severely broken PSF and is a major issue in current high-contrast imaging. However, pupil-plane wavefront sensors, such as a pyramid wavefront sensor (PWFS) used with a conventional wavefront reconstruction algorithm, have poor sensitivity to these modes. Worse still, if the phase offset across a spider is greater than $\lambda/2$ the wavefront sensor may jump to a semi-stable correction but with a 1λ offset between pupil segments,



(a) Bright target - asymmetric distribution seen

(b) Faint target - read noise dominates

Fig. 1 Example histograms of null-depths and NSC fits for GLINT observations. In both cases, the center of the distribution (measured null depth) is >0, but to distinguish between instrumental leakage and astrophysical contributions the detailed, asymmetric shape of the distribution must be fitted. In panel (a), the camera read noise distribution (black) is small compared to the overall null distribution, so this information can be recovered. But in panel (b), the star is fainter and read noise dominates, washing out this required detail. Note that no matter how much data are acquired, the width of the read noise contribution does not decrease. From GLINT data published in Refs. 10 and 12.

causing the broken PSF to persist for some time. The effect of this "phase lockup" is very pronounced in GLINT—since it operates at twice the wavelength of SCExAO's WFS, phase lockup results in a π phase offset being applied across a baseline that spans the spider, effectively swapping the null and antinull outputs.

Instead, it would be advantageous if the actual baseline phase was known at each instant in time, and this could then be fed directly into the model rather than fitted. Even then, this method assumes an accurate forward-model of the interferometric chip (or system) exists, which is difficult to produce for a real-life non-ideal photonic component over a wide range of wavelengths.

Another problem faced by NSC is that it works very poorly when there is a large amount of camera detector noise (i.e., read noise or dark noise) or IR background. The reason for this can be seen by reference to Fig. 1. In Fig. 1(a), the mean of the distribution is clearly offset from zero (i.e., the measured null-depth is >0), but this does not distinguish between the instrumental and astrophysical nulls. Instead, the unique, asymmetric shape of the PDF helps distinguish between these sources. Note the black distribution showing the contribution from detector noise—the measured histogram will be a convolution of this with the wavefront- and amplitude-induced null distributions.

However, this is a very high SNR example. Often, the histogram appears more like that shown in Fig. 1(b).[12] Here, the width of the detector noise distribution dominates and it is very difficult to distinguish between null-depth contributions. Note that no matter how long the target is observed, the width of the noise component never becomes narrower and the shape of the histogram will not change (although it will be more precisely defined).

Another limitation of the NSC method is that it has no cognizance of correlations between these error terms, which may occur due to optical factors (e.g., a moment of poor wavefront correction would likely affect baseline phase and injection efficiency) and instrumental factors (such as cross-coupling, either intentional or unintentional, between baselines in a photonic chip).

Here, an alternative method is proposed, which avoids using a statistical analysis and instead directly determines the actual instrumental leakage for each baseline for each instant in time.

3 Direct Determination of Instantaneous Null Leakage

Instead of analyzing the overall statistics of the observation, here, a model of the chip and optical system is created, which predicts the instantaneous instrumental leakage $N_{\text{inst}}(t)$ for each null output for each moment in time as a function of wavelength. The model uses as its input the

various other, bright outputs of the chip. These include the photometric channels, the bright (anti-null) outputs of all baselines, and if applicable, the "null" outputs for baselines, which are not currently in a null configuration (as is usually the case with GLINT). Crucially, since these outputs are bright they all have a high SNR compared to the null outputs when detector noise is a concern and so addresses the fitting problem encountered when the null channel is noisy as described in Sec. 2.2.

3.1 Model Description

To create this model, a data-driven approach is used, where the model is entirely constrained by actual data acquired from the chip, rather than an analytical or forward model which requires prior knowledge of all aspects of the chip's complex optical properties. This training data should be obtained from observations of as diverse as possible wavefront conditions, so as to maximally probe all regions of the chip's transfer function.

If the model's output is to be used to calibrate the null outputs, then this data should be from an unresolved source, such as in the lab using the instrument's inbuilt light source and a range of turbulence applied to its deformable mirror (DM) or on-sky by observing an unresolved star.

It should be emphasized that even if this training data are acquired on-sky from an unresolved source, this is fundamentally different to the classical method of calibration with a PSF reference star. In that case, one is assuming that the seeing statistics, AO properties, etc., remain the same between calibrator and science target. But here, there is no assumption that any of these things remain consistent. This is simply a means to obtain a diversity of data to probe the chip's transfer function. The only assumption is that the physical transfer function of the chip itself does not change, which is true (up to the limits of photonic stability when encountering temperature or strain changes). It should also be noted that in the case of large WFE or shallow astrophysical null depths, this data-driven model can be used to calibrate the bright output (as demonstrated in an NSC context in Ref. 11) to avoid the use of small-value approximations (e.g., having I_+ approximated by the total measured flux).

The model that is learned from the data is implemented using a neural network (NN).[20] NNs and their application to AOs is explained in detail in Wong et al.,[21] but essentially an NN is a method that learns and reproduces any non-linear function[22] based only on a set of examples. These examples (training data) must include the inputs (independent variables) and outputs of the function and should span as wide a region as possible of the parameter space in which the function will be applied. The fidelity with which the NN reproduces this function depends on the hyperparameters of the network (its architecture, complexity, training methods, etc.) and the quantity and quality of the training data provided.

An NN is closely analogous to a matrix and its training process analogous to the standard method of finding a matrix's pseudo-inverse using a singular value decomposition, with the key difference being that the NN is non-linear. This property is required in the present application, since the observed quantities are intensities, which are a non-linear function of the (un-observed) complex electric fields and complex coupling functions that describe the chip's (or optical system's) transfer function.

A crucial aspect of the deployment of NNs is to avoid over-fitting, and a large amount of machine learning research and methodology has been developed to avoid this problem. In the case of overfitting, the model learns to describe only the training data (essentially "remembering" this data) and does not generalize to new data. Before any training begins, standard practice is to split the data (usually randomly shuffled) into training data, used to train the network, and validation data, which is never seen by the training process and used as an independent test of the success of the NNs performance. The performance metric of the NN (the loss function, often the mean-squared error between true and predicted values) for both training data and previously unseen validation data is closely monitored as training occurs, and if signs of overfitting are observed, then network hyperparameters (such as regularization) are adjusted to prevent this.

To provide a straight-forward demonstration, the NN used here is a simple architecture—a fully connected (a.k.a. dense) feed-forward NN. Here, inputs and outputs are a vector of numbers (waveguide fluxes), and each unit in the hidden layers has a connection to every unit in the subsequent layer. In this study, we slowly increased the network complexity until the point of diminishing returns (for our relatively small data set) was reached, which led to a model

consisting of three layers of 1000 units each using an rectified linear unit (ReLU) activation function, plus the output layer. It was found that strong regularization was important to avoid overfitting, with both dropout and L2 regularization being used. A slow learning rate of 10^{-5} was found to provide the best convergence, likely because of the noisiness of the training data, and 500 epochs (with batch size 128) was used.

In the GLINT instrument, all chip output fibers are spectrally dispersed via a prism and imaged onto the detector, producing null, antinull, and photometric measurements as a function of wavelength (see Ref. 11 for further technical details). Due to the dependence of baseline phase as a function of wavelength [for a given optical path difference (OPD)], there is important wavefront information contained in the spectral domain. For example, ambiguity arising from phase wrapping is resolved. To leverage this, for each chip output, the NN model is given wavelength-dependent values (as vector of fluxes for each wavelength channel and it predicts the null-channel leakage as a function of wavelength. In the current fully connected model, the vectors are simply concatenated at the NN's input layer, and wavelength interdependence is learned empirically, but in a future refinement, the spectral correlation could be enforced by, for example, a spectral-domain convolutional kernel in a convolutional neural network (CNN).

Figure 2 shows a diagram of this method. The inputs to the NN are the ensemble of bright outputs of the nulling chip described above, for each measured wavelength channel. The outputs of the network are the null channels for which the leakage term is desired. Note that the actual inputs to the nulling chip (i.e., the light from telescopes or sub-apertures) are not a measurable value for our model.

To train the model [Fig. 2(a)], training data produced by applying some varied set of WFEs to the instrument (as described above) is used. The set of bright chip outputs are taken as the NN's inputs, and the model's outputs (null depths) are compared to the true chip null outputs for each data point. A loss function is defined, here just the mean-squared-error between the predicted and true values, and the model is trained to minimize this loss value. Once trained, the model is used in inference mode [Fig. 2(b)] where the bright channels of new science data are fed into the network, and the predicted null outputs used to calibrate the data.

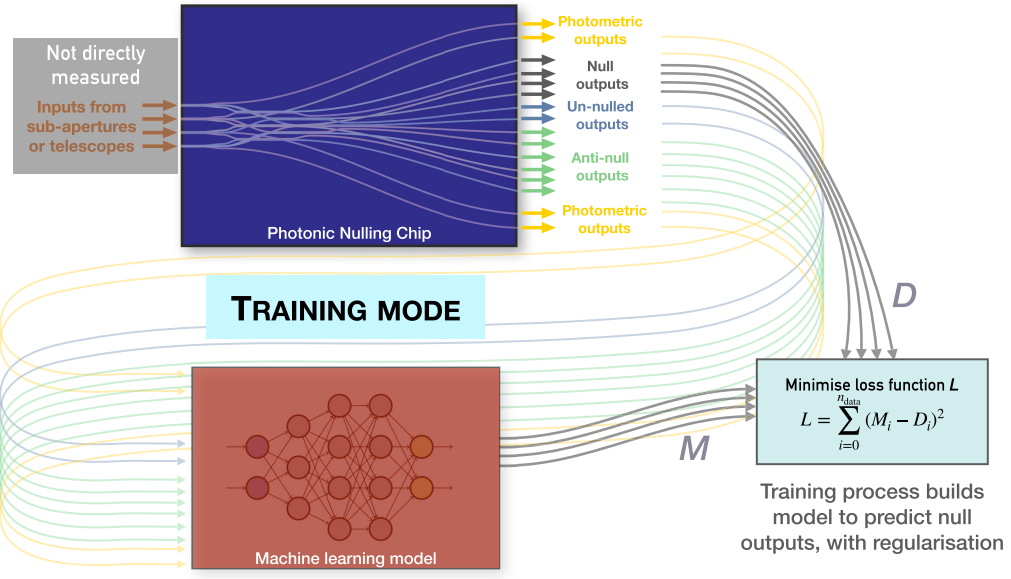
However a diverse choice of data sources (or better still, a combination thereof) could also be used to predict the null leakage, as long as there is some mapping between that measurement space and the null outputs.[23] In Sec. 6, the prediction of null leakage from the system's wavefront sensor telemetry and observed PSF is explored.

3.2 Fringe Tracking and Other Real-time Uses

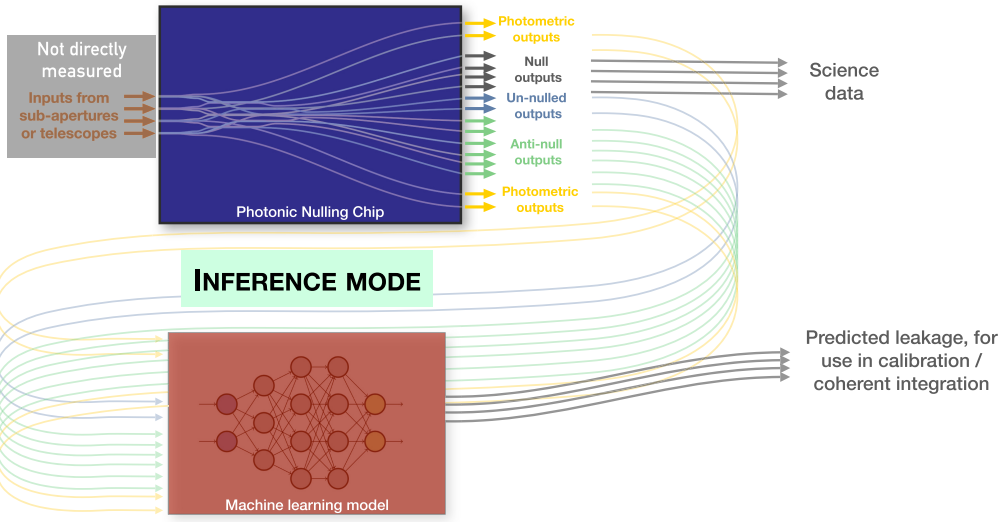
Another application is to use this model to produce real-time baseline OPD measurements to use for fringe-tracking. Driving the fringe-tracker directly from the nuller chip itself, rather than from a separate fringe-tracker instrument, is ideal as it removes the effects of non-common path error. Moreover, it means fringe-tracking and other AO measurements are performed at the same wavelength as the science measurements, mitigating the effect of atmospheric angular dispersion. It has been observed that these types of WFEs (especially those due to vibration and temperature drifts) are considerable with GLINT. This concept could also be applied to real-time measurement (and correction via the AO loop) of higher order terms, such as low-wind-effect/petaling, global tip/tilt, and others. Even with just two-channel beam combiners for each baseline, as long as multi-wavelength data is used, then the information required for fringe-tracking is present.

Since these quantities are functions of the chip inputs, some labeled data must be introduced into these inputs to obtain measurements in the desired space (coefficients for OPD, tip/tilt, etc.). In other words, even though the OPD information is indeed contained within the leakage predictions discussed thus far, for fringe-tacking use, we need to obtain a representation of these predictions projected onto the OPD space. This process is essentially equivalent to measuring a low-order response-matrix as in usual AOs. But in this case, the chip output is a non-linear function of these applied modes and the current incident wavefront (since these wavefronts are coherently combining, and the chip output intensities are the square of this complex sum).

Figure 3 shows a proposed method. During training [Fig. 3(a)], the chosen aberration space (e.g., differential OPD, tip/tilt, etc.) is modulated by some randomly chosen coefficients. This can



(a) Training mode



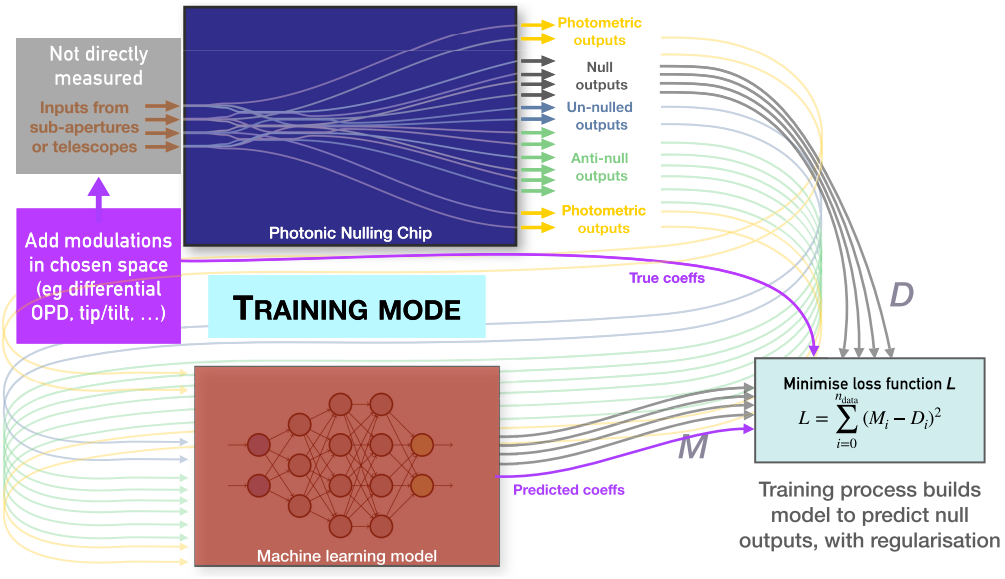
(b) Inference mode

Fig. 2 (a) and (b) A diagram of the proposed method, wherein an NN model is trained to predict the null-depth (instrumental leakage) using as inputs the remaining high SNR "bright" outputs of the photonic nulling chip. See text for details.

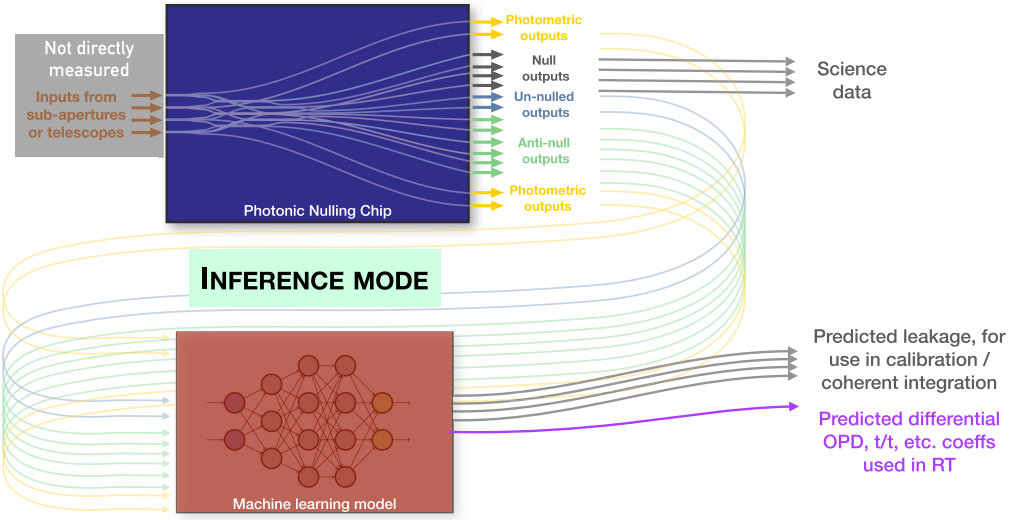
be achieved by adding these modes to the AO system's DM or using separate micro-electro-mechanical systems piston mirrors in the case of differential OPD. These quantities are then included in the model's output space, and the difference between the predicted and actually applied coefficients is included in the loss function. The model is trained to predict both the null-depths and the coefficients of interest.

During observations, the model is run in inference mode [Fig. 3(b)] in realtime. The predicted null-depths can be saved for later calibration, but the predicted baseline OPD mismatch (or other coefficients) is used in closed loop by the AO system to keep the fringes steady and injection high. If used to sense and correct low-wind effect, this has the bonus effect of benefiting all other imaging instruments that are operating at the same time.

If desired, the differential OPD data could be generated off-line instead and used to conduct an NSC-like analysis of the data, but with the phase errors no longer being a fitted parameter.



(a) Training mode



(b) Inference mode

Fig. 3 (a) and (b) A diagram of a modified method, where additional coefficients describing baseline OPD mismatch (or other aberrations) are directly predicted to be used in real time for fringe-tracking or AO. See text for details.

4 Demonstration of Instrumental Leakage Prediction

4.1 Method

To evaluate this process, a number of experiments were performed using the GLINT photonic nulling interferometer,[10–12] deployed on the SCExAO AOs system[24,25] at the Subaru telescope. This instrument, built around an integrated-optics nulling chip, has successfully demonstrated on-sky measurements of objects well beyond the telescope diffraction limit,[10,11] but its sensitivity is largely limited by camera detector noise and its null calibration precision by the performance of NSC (especially under noisy conditions).

For each experiment, 100 s of data was used, split into 80% training, 19% validation data and a separate 1% of contiguous holdout data. Since the instrument samples at speeds comparable to the atmospheric coherence time, it is expected there will be some correlation between consecutive frames. Since the data are randomly shuffled, it is conceivable that the model could still

slightly "overfit" even if validation data loss is low, since strongly correlated frames may occur in both training and test sets. The purpose of this additional holdout-data is to check that this is not occurring. It is contiguous data taken from the end of the data set, and so should not have correlated frames present in the main data set, and thus its loss function will reveal overfitting even if not apparent in the validation data loss. Moreover, since this data are contiguous its predicted wavefront can be viewed in a time-domain diagram (or as a movie) alongside the corresponding true values, enabling a human "sense-check" that the wavefront prediction is working as expected (e.g., to detect if an unsuitable loss function was used). This methodology was used for all experiments in this paper, and the holdout data are used to create all figures and movies presented. Due to non-optimal path-length matching in the current prototype chip, all four nullable baselines cannot be simultaneously nulled, so for each experiment two sets of data were taken, each with the phase offset for two baselines set to achieve a null.

Network hyperparameters were manually optimized to prevent overfitting and achieve pleasing prediction accuracy, though for real-world deployment a rigorous automated hyperparameter optimization should be performed. Hyperparameters were tuned simultaneously on all datasets. This resulted in a single set of hyperparameters that were used for all experiments, to check that a common architecture should work independent of source brightness. As described in Sec. 3.1, a three layer (plus output layer) fully connected network was used, using an ReLU activation function, and trained with the Adam optimizer with a batch size of 128 and learning rate of 10^{-5}.

Of central importance in building such a model is regularization—that is, preventing the model from overfitting (e.g., fitting to noise or the stochastic composition of the training set) and not generalizing. It was found that, especially when training on noisy (on-sky) data, strong regularization was required to avoid overfitting while still maintaining a network complexity large enough to provide accurate predictions over diverse WFEs. Here, dropout[26] was found to be most successful, which was used between each hidden layer, with a dropout rate of 50%. It was also found that L2 kernel regularization was helpful and applied with a regularization factor of 0.01.

4.2 Results

In the first experiment, high SNR data were obtained off-sky using SCExAO's internal broadband light source and a sliding Kolmogorov phase screen applied to the system's 2K actuator DM, simulating an on-sky observation. The data were acquired at 1400 frames/s, with the applied turbulence having an amplitude of 1000 nm root mean square (RMS) and wind-speed of 5 m/s. This large amplitude was used to maximally probe the transfer function of the chip over a large WFE domain and be well outside the linear-approximation regime.

Figure 4 shows the actual and predicted measurements for a null output for this laboratory phase-screen test (using holdout data), for a single wavelength. Data is shown at two zoom levels and it can be seen that the predicted null depth (leakage) is highly consistent with the true, measured values.

In the lower plot, vertically zoomed by a factor of ~40 to show the null, the predicted values do not perfectly lie within the 1σ detector noise band, illustrating the precision-level of the prediction for this test. This region is at the "turning point" of the null, where the relationship between delta-phase and output intensity is maximally non-linear. Performance may be increased by rigorous hyperparameter tuning (including network dimensions) to maximize non-linearity handling. This region also consists of the lowest SNR training examples (i.e., since the null output is ~zero), making it the slowest for the NN to learn. But due to the strong regularization used in training, the values predicted here are in the middle of the true range, rather than blowing up from noise. The prime requirement is that these errors do not introduce a systematic bias (i.e., they are noise with zero-mean). The impact of this is quantified by the experiments in Sec. 5.

In Fig. 5, results are shown for laboratory turbulence where a large wavelength range is considered, for four "nulled" baselines. In the time-windows shown, periodic vibration-induced leakage can be clearly seen in null baseline 1. In all cases, the residual is consistent with noise, as would be expected from an ideal prediction.

A subsequent test was performed using on-sky data, obtained from an observation of the star α Bootis with GLINT in June 2020, as shown in Fig. 6. Even though the delays on baselines 1 and 4 were correctly set to produce nulls, the observation suffered from severe LWE/petaling and were high enough that phase-lockup (where the PyWFS intermittently locks with a 1λ phase

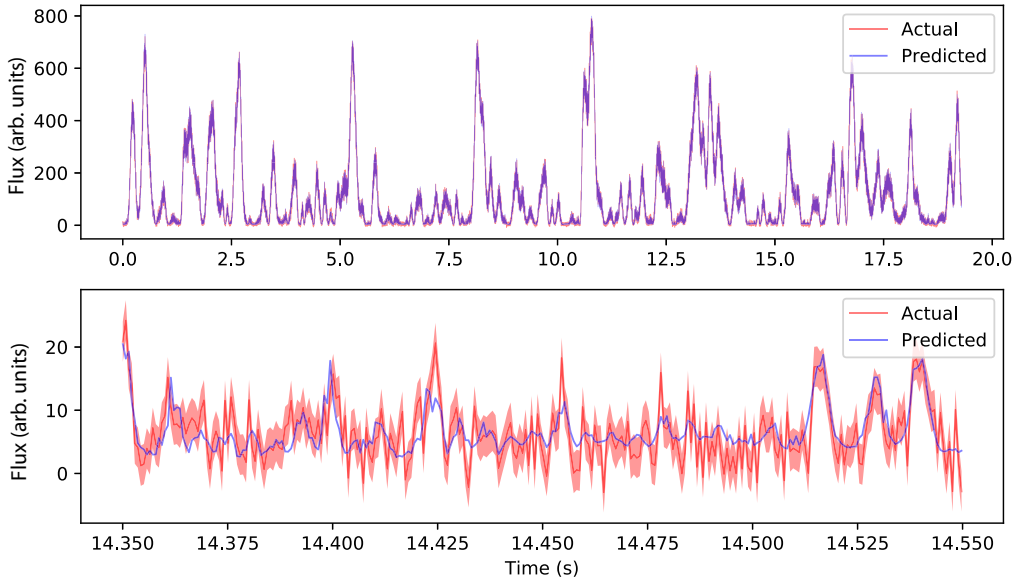
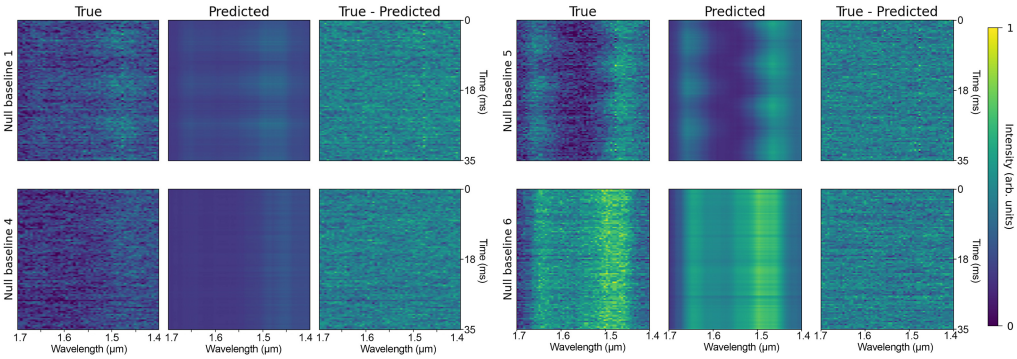


Fig. 4 Predicted-by-model (blue) and measured (red) null leakage for null output 1 of the GLINT chip, for a Kolmogorov phase screen applied in the laboratory. The intrinsic detector noise (relatively small for this bright laboratory data) is shown by the red shaded region. (b) Zoomed-in region of (a), at a time when the instantaneous WFE allowed a good null. The generated WFE (1 μm RMS) produces large variation in null depth, which is well predicted by the model.

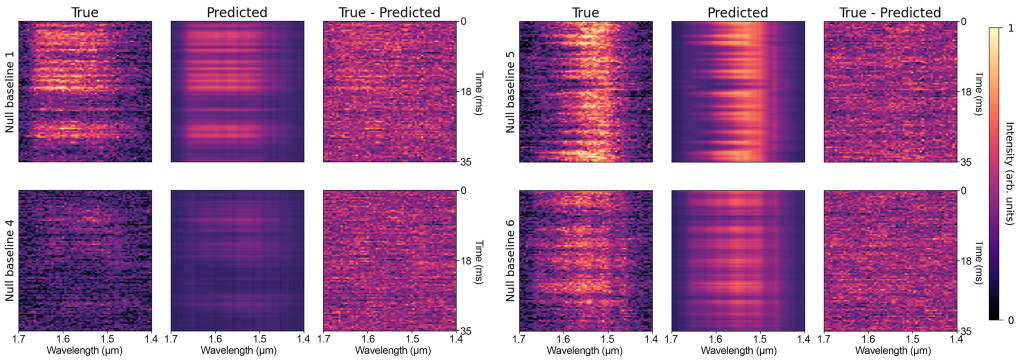


(a) Nulled baselines 1 and 4 (b) Nulled baselines 5 and 6

Fig. 5 (a) and (b) True, predicted, and residual leakage for four nulled baselines of the GLINT chip, for a Kolmogorov phase screen applied in the laboratory, shown as a function of wavelength over a 35 ms time period. In baselines 1 and 4, the null is relatively deep, but intermittent leakage (arising largely from vibration-induced WFE) is visible, which is well-predicted by the model. Baselines 5 and 6 are not at the true white-light null, so leakage is high and strongly chromatic, and this is still well predicted. Note that while the true data is noisy, the predicted data is not, thanks to the high SNR of the bright outputs used for prediction. Color stretch is the same across all panels. Video 1 is an animated version of this figure. (Video 1, 45.2 MB, MP4 [URL: https://doi.org/10.1117/1.JATIS.9.4.048005.s1]).

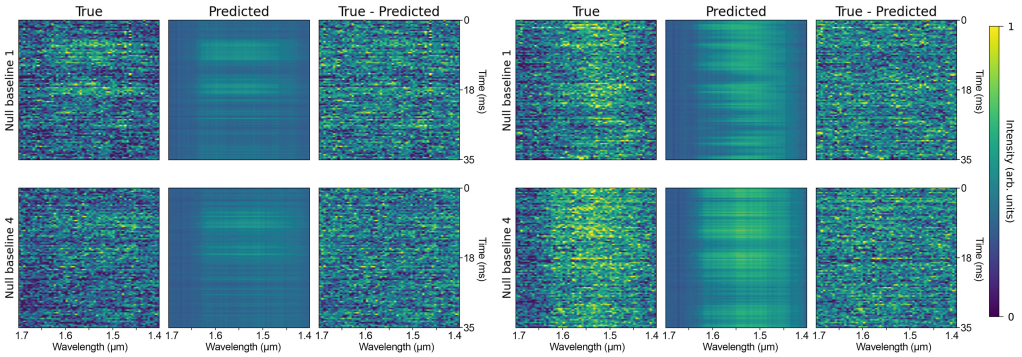
offset between segments, as described in Sec. 2.2) often occurred. This is clearly seen in the measured null-depths, especially in the stripe-like patterns appearing in null baseline 1. The model successfully predicts these, demonstrating it can access sufficient information to sense these modes and predict them as a function of wavelength. It also demonstrates that the model's effectiveness is not limited to the linear regime.

The star δ Virginis was also observed and a model used to predict its leakage, as shown in Fig. 7. As with α Bootis, LWE is present and successfully sensed and its corresponding effect on the leakage predicted. It is clearly seen here that the predictions have much higher SNR than the measured null outputs, since the predictions are based on the bright (high SNR) chip outputs.



(a) Nulled baselines 1 and 4 (b) Nulled baselines 5 and 6

Fig. 6 (a) and (b) True, predicted, and residual leakage for four nulled baselines of the GLINT chip for an on-sky observation of α Bootis. The observations suffered from severe low-wind-effect, leading to PWFS lock-up, which is especially well seen in the "striping" in baselines 1 and 4. Baselines 5 and 6 are not at the true white-light null and so show strongly chromatic leakage. In all cases, the model provides a good high-SNR prediction of the leakage. Color stretch is the same across all panels. Video 2 is an animated version of this figure. (Video 2, 38.0 MB, MP4 [URL: https://doi.org/10.1117/1.JATIS.9.4.048005.s2]).



(a) Nulled baselines 1 and 4 (b) Nulled baselines 5 and 6

Fig. 7 (a) and (b) True, predicted, and residual leakage for four nulled baselines of the GLINT chip for an on-sky observation of δ Virginis. As with α Bootis, the observations encountered strong LWE, and PWFS phase lockup occurred. As before, this was well predicted by the model and subtracted cleanly. It should be noted that the predictions are far higher SNR than the null-output measurements (since they are built from the bright, high SNR outputs). Color stretch is the same across all panels. Video 3 is an animated version of this figure. (Video 3, 45.1 MB, MP4 [URL: https://doi.org/10.1117/1.JATIS.9.4.048005.s3]).

Hence this method of calibrating using the predicted instantaneous null leakage does not suffer from the same detector noise limitation as NSC.

5 Experimental Comparison to NSC

5.1 Method

To quantify the performance advantage of the NN method compared to the traditional NSC method, the same dataset was analyzed with both methods and their ability to accurately predict the null depth—and the associated uncertainties—was examined. It was not possible to perform this test using on-sky data, since there was no on-sky data available for an unresolved star (to use for training data), which meant the absolute value of the mean null depth in their resulting analyses has an unknown offset. Instead, laboratory data (that shown in Fig. 5) where Kolmogorov turbulence has been applied via the SCExAO DM is used. This dataset used an attenuated light

Table 1 Overview of the noise properties (combined read-noise and dark-noise) of the test data used to compare NSC and NN calibration. Both datasets contain data for the N1 and N4 null channels. The "low noise" data just contain the actual camera read-noise and dark-noise, while the "high noise" data have had additional Gaussian noise injected into the raw signal. Noise and SNR given here is per wavelength-channel per frame. All values are expressed in flux units (derived from camera analog-digital units).

Dataset	RMS noise	Mean I_-	Mean I_+	I_- SNR	I_+ SNR
Low noise - N1	5.25	4.85	48.11	0.92	9.17
Low noise - N4	5.25	6.97	46.74	1.33	8.91
High noise - N1	11.29	4.84	48.12	0.43	4.26
High noise - N4	11.29	6.99	46.73	0.62	4.14

source to approximately match the on-sky fluxes observed in the on-sky tests. To test performance on even fainter targets, a second dataset was created wherein high read noise and dark current contribution was simulated by adding Gaussian noise to the raw lab data—see Table 1 for details.

In these tests, the light source is a broadband supercontinuum light source injected via a single mode fibre and thus the true null depth is known to be zero. But the raw measured null depths are high (of order 10^{-2}) due to instrumental leakage. Calibration performance is judged by the precision by which the true zero null depth is recovered, and the uncertainties placed on it.

First, the NSC method (using the Barnacle package[27]) was used to measure the calibrated null depth. This implementation handles multi-wavelength data and does not assume small WFE approximations. In this method, a PDF of the chip's output signals is produced via a histogram of all data. Then a simulated PDF is fitted to it as a function of parameters such as the average and variance of phase error and amplitude error, as well as the parameter of interest (the astrophysical null). The gradient descent algorithm was given initial parameter guesses close to the expected null depth value and then the fit was re-run ~250 times with randomized starting positions each time (a.k.a. basin-hopping) to avoid the problem of starting within a local, not global, minimum.

Then, analysis was conducted using the NN method presented here. To correctly calculate the null depth, predicted leakage values for both I_- and I_+ are needed. In low WFE cases when I_- is extremely small, the true I_+ can simply be approximated as the total measured flux,[5,15] but in the present WFE regime we cannot make that approximation and thus the I_+ value is calibrated in the same way as described previously. In both cases, the data used to train the model is separate to the data used to perform these tests to avoid potential overfitting leading to an overestimate of performance.

For this dataset, the instrumental null depth N_{inst} was defined as the ratio of the predicted I_- to I_+ outputs. Similarly, the observed null N_{obs} is the ratio of raw measurements of the I_- and I_+ outputs. Then, as per Eq. (3), the real "astrophysical null" $N_{\text{astro}} = N_{\text{obs}} - N_{\text{inst}}$. Since in this data, the source is unresolved (a single-mode fiber), for both analysis methods, we would hope to see $N_{\text{astro}} = 0$.

5.2 Results

The results of this analysis are shown in Fig. 8. As shown in the left-hand panel, the NN method produces a calibrated null depth for each wavelength channel, and the mean of these over wavelength is taken to be the final null depth estimate. The uncertainties for each data point are the standard error in the mean of the null depths predicted for each time step. The null-depths produced are very small, of order 10^{-4}, and in most cases, their uncertainties are consistent with the true null depth of zero. The exception is the 1.477 μm measurement for null channel 1, which was affected by a slowly varying bad-pixel on the detector [also visible in Fig. 5(a)], leading to the statistical errors (shown here) underestimating the total error by a factor of ~2 for this measurement.

Notably, the accuracy of this prediction is not obviously affected by the degree of noise present. For the lower-noise data, the measured null-depths for the two channels were

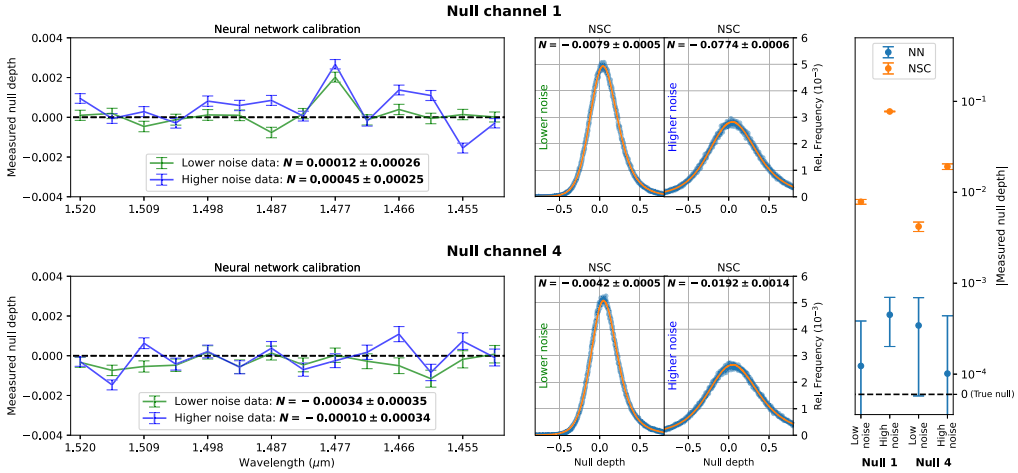


Fig. 8 Results of the comparison of NSC and NN calibration methods, for laboratory data with moderate WFE and an unresolved source (so true null-depth of 0), for datasets with different noise levels (see Table 1) and for two baselines. Left: the resulting calibrated null depths using the NN method, plotted as a function of wavelength. The null depth is measured to be of order 10^{-4} and in most cases with estimated uncertainties consistent with null depth of zero. Center: the measured histograms and resulting PDF fit using the NSC method, along with the resulting null-depths and uncertainty estimation. The histogram is hard to distinguish from a Gaussian distribution (especially for higher-noise data), resulting in poor estimation of null depth and uncertainties. Right: summary of resulting null depths from the two methods, with the absolute difference between true and measured nulls plotted on a hybrid-log scale (vertical axis $<10^{-4}$ is linear). The NN method outperforms that NSC method by ~2 orders of magnitude in accuracy and has far more realistic uncertainty estimations.

$1.2 \pm 2.6 \times 10^{-4}$ and $-3.4 \pm 3.5 \times 10^{-4}$, respectively, while for the higher-noise data, they were $4.5 \pm 2.5 \times 10^{-4}$ and $-1.0 \pm 3.4 \times 10^{-4}$.

On the other hand, the NSC method performed poorly in the presence of noise, with null-depths of $\sim 10^{-3}$ and $\sim 10^{-2}$ for lower-noise and higher-noise data, respectively. Moreover, the estimated uncertainties on these fitted parameters—derived from the diagonals of the covariance matrix returned by the gradient descent algorithm—are underestimated by 1 to 2 orders of magnitude. In the case of NSC, the accuracy of the predicted null is seen to be strongly influenced by the noise level. For the lower noise data, the calibrated null depths were measured to be $-7.9 \pm 0.5 \times 10^{-3}$ and $-4.2 \pm 0.5 \times 10^{-3}$, and for the higher noise data, they were $-7.7 \pm 0.06 \times 10^{-2}$ and $-1.9 \pm 0.1 \times 10^{-2}$. The underlying problem encountered by NSC can be seen in the histograms and fitted model PDF in the centre panel of Fig. 8. Despite the fact that a very good fit to the data has been found, as described in Sec. 2.2, higher dark/read noise broadens the PDF and washes out the tell-tale asymmetries, which allows static and WFE-induced leakage to be disambiguated from true null depth. Note that the NSC method still fits to the data at multiple wavelengths, but only a single wavelength's histogram is plotted for clarity.

6 Prediction of Leakage from Diverse Data Sources

In addition to the bright outputs of the nuller chip, there are various other sources of real-time data available in the SCExAO system, which may contain useful information determining the null leakage. One such data stream is the PWFS. An experiment was performed where the PWFS telemetry was used as the sole input to a model to predict the null leakage, rather than the chip's bright outputs. Tests using the raw PWFS image [flattened to a one-dimensional (1D) vector] and also using SCExAO's reconstructed wavefront were performed, with no clear difference seen in the quality of prediction between these two methods.

Figure 9 shows the results of this experiment (in this case using SCExAO modes), from May 2021 on-sky observations of α Bootis. At first glance, it appears the PWFS-based prediction does not perform as well as the previous examples. However, it is informative to note that

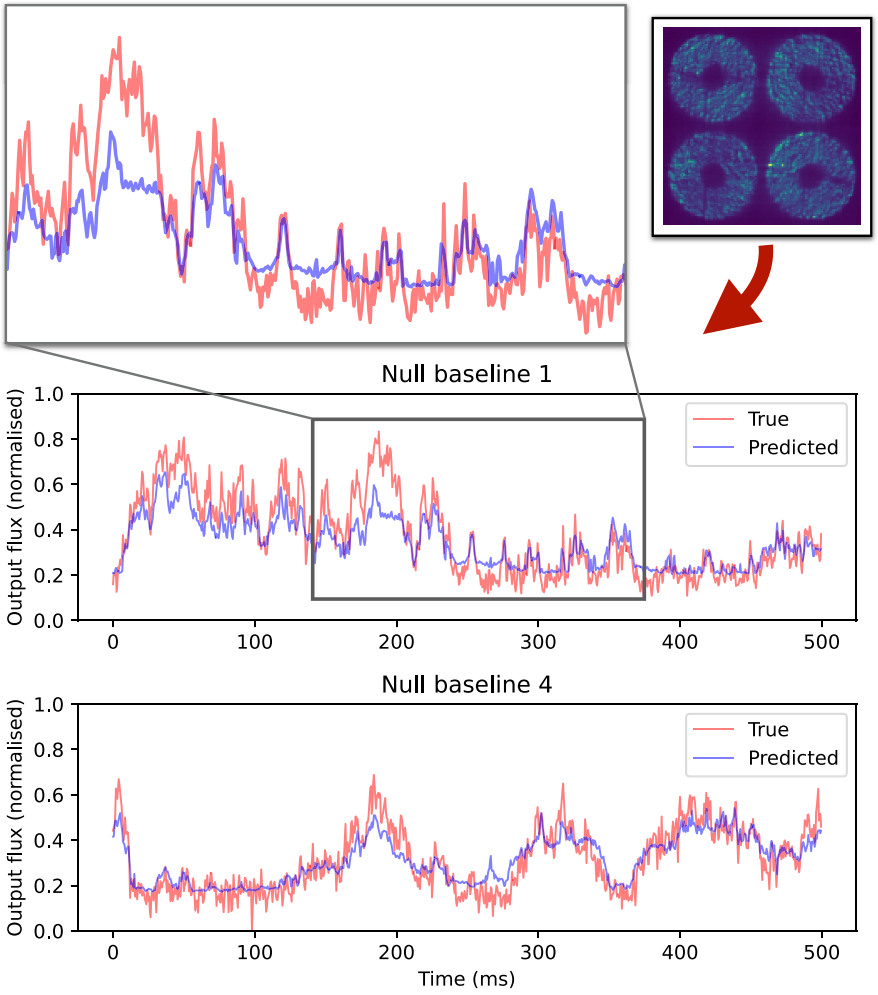


Fig. 9 Predicted (blue) and measured (red) null leakage for null outputs 1 and 4 of the GLINT chip using only the PWFS (inset) telemetry as input to the model. Data are on-sky observations of α Bootis in May 2021. Prediction using only PWFS data does not work quite as well. However, as most clearly seen in the zoomed portion, often the prediction includes correct features but is missing larger offsets. This is consistent with the large offsets being due to inter-segment phase offset, such as from LWE, to which the PWFS is insensitive.

the prediction appears to work relatively well for small-amplitude, short period WFS, but has large systematic offsets from the true values. This is consistent with the fact that the PWFS is insensitive to LWE or other inter-segment phase-shear modes. In this case, the small successfully predicted perturbations correspond to "normal" WFE, but the large offsets arise from LWE modes.

The use of the simultaneously recorded PSF as a data source was also investigated. Here, the image from SCExAO's infrared high-speed camera, flattened into a 1D vector, was used as the sole input to the model. As seen in the results in Fig. 10, this data-source enabled a much better prediction of null depth than PWFS data. The PSF clearly shows LWE aberrations (with a splitting PSF), and the large offsets in the null leakage are correctly predicted. However, it is not perfect, and one interesting issue can be seen in the zoomed region in the figure. In some places, such as here, the variation in null leakage is successfully predicted but the sign is reversed. This is consistent with the fact that a focal-plane image has sign degeneracies for even modes (for example, a PSF cannot show a difference between a positive and negative defocus aberration of the same amplitude). It is therefore not unexpected that the PSF alone cannot unambiguously determine the null leakage.

Introducing phase diversity to the PSF, such as including a defocused image or using multiple wavelengths may break this degeneracy. While using the PWFS or PSF alone to predict the

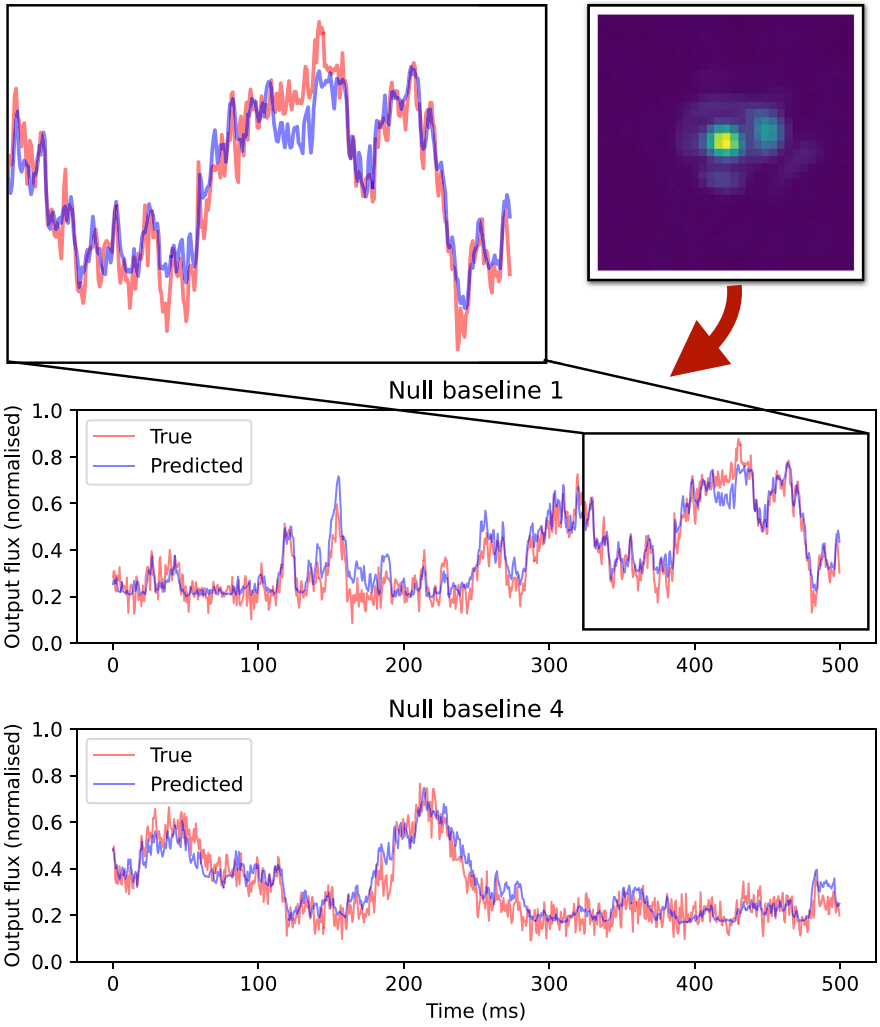


Fig. 10 Predicted (blue) and measured (red) null leakage for null outputs 1 and 4 of the GLINT chip using only the infrared PSF (inset) as input to the model. Data are on-sky observations of α Bootis in May 2021. While the prediction is more successful than using only PWFS data, one interesting problem should be noted. As emphasized in the zoomed version, at some times, the predicted variation in null leakage is correct but of inverted sign. This is consistent with the sign degeneracy present in any focal-plane image of the PSF.

null is informative, a key goal is to maximize the input space and SNR by simultaneously utilizing all data streams (nulling chip bright outputs, PWFS telemetry, PSFs at multiple wavelengths, and other sensors) to perform the optimal null leakage prediction. This will require careful weighting and regularization of the model, to ensure degeneracies in one source to not bias the model.

7 Conclusion and Next Steps

Along with maintaining a deep null, calibrating the null depth to extract accurate science observables is a key challenge in nulling interferometry. The measured output of a nulled baseline is a combination of the astrophysical null (the science quantity of interest) and instrumental leakage, which is rapidly varying in time as a function of seeing. The instrumental leakage must be precisely known to perform science measurements. Simply subtracting the time-averaged null depth of an unresolved target from the science data works poorly, since it requires seeing and AO parameters to remain very consistent. Often a statistical approach—NSC—is used, but this assumes normally distributed phase errors (fitted by a single mean and standard deviation) and does not work well when detector noise or background contributions are high.

Here, an approach using an NN model to predict the instrumental leakage for every instance in time is proposed. The model is built entirely using empirical data taken by the instrument (either on-sky or in the laboratory by applying turbulence to the DM). Using the bright outputs of the chip as input to the model, the instrumental null depth can be predicted with a high SNR. An extended version of the model could also produce differential OPD (or other aberration) measurements in real time, for use in closed-loop fringe-tracking or AO. Diverse data sources (such as the system's WFS or camera) could also be used.

A model was trained and tested using several datasets, including laboratory data and two on-sky targets, representing brighter and fainter cases. In all cases, the model successfully predicted the null leakage as a function of wavelength and with high SNR.

To deploy this in a science context, several aspects require further investigation. First, the robustness of a single model to different observing conditions or epochs must be evaluated. Ideally, a single model would be trained, using multiple sets of on-sky and laboratory data. Whether a single model will give accurate predictions in all cases, or whether a model needs to be additionally fine-tuned or trained for each observation, remains to be seen. The actual accuracy of the calibration using this method must be investigated and improved if necessary. While a noisy prediction is acceptable, a bias in the prediction of instrumental null directly translates to miscalibration. Evaluating the hardware and model in the laboratory using incoherent sources of precisely known sizes should be performed.

Beyond the basic model demonstrated here, additions such as real-time prediction of differential OPD for fringe tracking should be implemented and tested. It may also be advantageous to combine data from multiple sources (WFS, PSF, etc.) but this must be done in a way to avoid degeneracies in one sensor space affecting the overall inference.

The model architecture here was very simple (a fully connected NN). Gains may be found in using other architectures, for example, a CNN where 1D convolutional kernels in the wavelength domain are used. Furthermore, taking into account, the time domain may be highly advantageous. Consecutive measurements are highly correlated in time (due to temporal sampling at rates comparable to the atmospheric coherence time) but this is currently ignored. A time domain model, such as a recurrent NN or time-domain CNN, would enable this correlation to be exploited to potentially improve calibration accuracy and SNR. A transformer type network could also prove useful thanks to its positional encoding, and more complex architectures can take into account the interconnected spectral/spatial/temporal relationships. Finally, it is hoped that the general concept presented here will find utility in the calibration of other types of measurements, such as long-baseline interferometry, speckle nulling, and adaptive coronagraphy.

Code and Data Availability

The data and code utilized in this study are available from the authors upon request.

Acknowledgments

Barnaby R. M. Norris is the recipient of an Australian Research Council Discovery Early Career 230 Award (Grant No. DE210100953) funded by the Australian Government. The development of SCExAO was supported by the Japan Society for the Promotion of Science (Grant-in-Aid for Research Nos. 23340051, 26220704, 23103002, 19H00703 and 19H00695); the Astrobiology Center of the National Institutes of Natural Sciences, Japan; the Mt. Cuba Foundation; and the director's contingency fund at Subaru Telescope. The authors would like to thank Dr. Eckhart Spalding for his work on the GLINT instrument upgrade. The authors wish to recognize and acknowledge the very significant cultural role and reverence that the summit of Mauna Kea has always had within indigenous Hawaiian communities and are most fortunate to have the opportunity to conduct observations from this mountain.

References

1. R. N. Bracewell, "Detecting nonsolar planets by spinning infrared interferometer," *Nature* **274**, 780–781 (1978).
2. J. R. P. Angel and N. J. Woolf, "An imaging nulling interferometer to study extrasolar planets," *Astrophys. J.* **475**, 373–379 (1997).

3. A. Léger et al., "Could we search for primitive life on extrasolar planets in the near future?" *Icarus* **123**, 249–255 (1996).
4. O. Absil et al., "Performance study of ground-based infrared Bracewell interferometers. Application to the detection of exozodiacal dust disks with GENIE," *Astron. Astrophys.* **448**, 787–800 (2006).
5. E. Serabyn, "Nulling interferometry: symmetry requirements and experimental results," *Proc. SPIE* **4006**, 328–339 (2000).
6. M. M. Colavita et al., "Keck interferometer nuller data reduction and on-sky performance," *Publ. Astron. Soc. Pac.* **121**, 1120 (2009).
7. D. Defrère et al., "Nulling data reduction and on-sky performance of the large binocular telescope interferometer," *Astrophys. J.* **824**, 66 (2016).
8. B. Mennesson et al., "High-contrast stellar observations within the diffraction limit at the Palomar Hale Telescope," *Astrophys. J.* **743**, 178 (2011).
9. J. Kühn et al., "Exploring intermediate (5-40 AU) scales around AB Aurigae with the Palomar Fiber Nuller," *Astrophys. J.* **800**, 55 (2015).
10. B. R. M. Norris et al., "First on-sky demonstration of an integrated-photonic nulling interferometer: the GLINT instrument," *Mon. Not. R. Astron. Soc.* **491**, 4180–4193 (2020).
11. M.-A. Martinod et al., "Scalable photonic-based nulling interferometry with the dispersed multi-baseline GLINT instrument," *Nat. Commun.* **12**, 2465 (2021).
12. T. Lagadec et al., "The GLINT South testbed for nulling interferometry with photonics: design and on-sky results at the Anglo-Australian Telescope," *Publ. Astron. Soc. Aust.* **38**, e036 (2021).
13. B. R. M. Norris et al., "Optimal self-calibration and fringe tracking in photonic nulling interferometers using machine learning," *Proc. SPIE* **12183**, 121831J (2022).
14. A. Labeyrie, S. G. Lipson, and P. Nisenson, *An introduction to Optical Stellar Interferometry*, Cambridge University Press (2006).
15. C. Hanot et al., "Improving interferometric null depth measurements using statistical distributions: theory and first results with the Palomar Fiber Nuller," *Astrophys. J.* **729**, 110 (2011).
16. J.-F. Sauvage et al., "Tackling down the low wind effect on SPHERE instrument," *Proc. SPIE* **9909**, 990916 (2016).
17. J. Milli et al., "Low wind effect on VLT/SPHERE: impact, mitigation strategy, and results," *Proc. SPIE* **10703**, 107032A (2018).
18. M. N'Diaye et al., "Calibration of the island effect: experimental validation of closed-loop focal plane wavefront control on Subaru/SCExAO," *Astron. Astrophys.* **610**, A18 (2018).
19. S. Vievard et al., "Overview of focal plane wavefront sensors to correct for the low wind effect on SUBARU/SCExAO," https://doi.org/10.48550/arXiv.1912.10179 (2019).
20. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**, 436–444 (2015).
21. A. P. Wong et al., "Machine learning for wavefront sensing," *Proc. SPIE* **12185**, 121852I (2022).
22. K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks* **2**(5), 359–366 (1989).
23. O. Guyon et al., "High contrast imaging at the photon noise limit with self-calibrating WFS/C systems," *Proc. SPIE* **11823**, 1182318 (2021).
24. N. Jovanovic et al., "The Subaru coronagraphic extreme adaptive optics system: enabling high-contrast imaging on solar-system scales," *Publ. Astron. Soc. Pac.* **127**, 890 (2015).
25. J. Lozi et al., "SCExAO, an instrument with a dual purpose: perform cutting-edge science and develop new technologies," *Proc. SPIE* **10703**, 1070359 (2018).
26. N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).
27. M.-A. Martinod and B. R. M. Norris, "SydneyAstrophotonicInstrumentationLab/GLINTPipeline: Barnacle code first realease," https://doi.org/10.5281/zenodo.4563370 (2021).

Barnaby R. M. Norris is a DECRA Fellow at the University of Sydney. His work focuses on instrumentation design, implementation, and observations. He specializes in astrophotonics in high-angular resolution imaging, including direct imaging, adaptive optics (especially machine learning) and interferometry. He led the design and construction of the VAMPIRES instrument and GLINT photonic nulling interferometer, deployed the Subaru Telescope. New research focuses on the use of photonic lanterns as wavefront sensors and imaging devices, and planet formation and stellar mass-loss from evolved stars.

Biographies of the other authors are not available.