PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Data-driven subspace predictive control: lab demonstration and future outlook

Haffert, Sebastiaan, Males, Jared, Close, Laird, Long, Joseph, Schatz, Lauren, et al.

Sebastiaan Y. Haffert, Jared R. Males, Laird Close, Joseph Long, Lauren Schatz, Kyle van Gorkom, Alexander Hedglen, Jennifer Lumbres, Alexander Rodack, Olivier Guyon, Justin Knight, Maggie Kautz, Logan Pearce, "Data-driven subspace predictive control: lab demonstration and future outlook," Proc. SPIE 11823, Techniques and Instrumentation for Detection of Exoplanets X, 1182306 (1 September 2021); doi: 10.1117/12.2594875

SPIE.

Event: SPIE Optical Engineering + Applications, 2021, San Diego, California, United States

Data-driven subspace predictive control: lab demonstration and future outlook.

Sebastiaan Y. Haffert[a], Jared R. Males[a], Laird M. Close[a], Kyle Van Gorkom[a,b,c], Joseph D. Long[a], Alexander D. Hedglen[a,b], Olivier Guyon[a,b,d,e], Lauren Schatz[a,b], Maggie Kautz[a,b], Jennifer Lumbres[a,b], Alexander Rodack[a,b], Justin M. Knight[a,b], He Sun[f], and Kevin Fogarty[g,h]

[a]Steward Observatory, University of Arizona, Tucson, Arizona, United States
[b]Wyant College of Optical Science, University of Arizona, 1630 E University Blvd, Tucson, AZ 85719, USA
[c]NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[d]Astrobiology Center, National Institutes of Natural Sciences, 2-21-1 Osawa, Mitaka, Tokyo, JAPAN
[e]National Astronomical Observatory of Japan, Subaru Telescope, National Institutes of Natural Sciences, Hilo, HI 96720, USA
[f]Department of Computing and Mathematical Science, California Institute of Technology, Pasadena, CA 91125, USA
[g]The Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA
[h]NASA Ames Research Center, Moffett Field, California 94035 USA

ABSTRACT

The search for exoplanets is pushing adaptive optics systems on ground-based telescopes to their limits. A major limitation is the temporal error of the adaptive optics systems. The temporal error can be reduced with predictive control. We use a linear data-driven integral predictive controller that learns while running in closed-loop. This is a new algorithm that has recently been developed. The controller is tested in the lab with MagAO-X under various conditions, where we gain several orders of magnitude in contrast compared to a classic integrator. We will present the lab results, and we will show how this controller can be implemented with current hardware for future extremely large telescopes.

Keywords: high-contrast imaging, high-resolution spectroscopy, exoplanets, adaptive optics

1. INTRODUCTION

Atmospheric turbulence severely degrades the image quality of telescopes. Adaptive optics systems are required to restore the spatial resolution of ground-based telescopes. The next generation of Giant Segmented Mirror Telescopes (GSMT) will have the largest telescope apertures to date and will need adaptive optics to operate. One of the primary science goals of these telescopes will be the imaging and characterization of Earth-like planets around M-type stars. However, these planets are several orders of magnitude fainter than their host star at separations of a couple times the diffraction-limit. This makes the planets difficult to distinguish from their host star.[1] High-contrast imaging instruments are used to remove the influence of the star by employing extreme adaptive optics and coronagraphs.[2]

Current direct imaging instruments routinely reach post-processed contrast levels of 10^{-4} to 10^{-6} at angular separations between 0.1 arcsec and 1.0 arcsec.[3] This sensitivity is enough to image and characterize massive self-luminous planets[4] that emit the majority of their emission in the near-infrared part of the spectrum. And even

Further author information: (Send correspondence to S.Y.H.)
S.Y.H.: E-mail: shaffert@arizona.edu
S.Y.H. is a NASA Hubble fellow

Techniques and Instrumentation for Detection of Exoplanets X, edited by Stuart B. Shaklan, Garreth J. Ruane,
Proc. of SPIE Vol. 11823, 1182306 · © 2021 SPIE · CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2594875

though these instruments are sensitive enough to detect Jupiter-like planet, very few are discovered. Analysis of direct imaging surveys and radial velocity surveys hint that there is a turn over where the occurrence rate of exoplanets starts to drop.[5–8] This turnover happens between 1 to 10 AU, which is the expected position of the snow line.The sensitivity closer in to the star has to be improved if we want to observe more planets.

A major limitation at small angular separations is caused by time lag in the adaptive optics (AO) system.[9–11] The correction of the atmosphere is always trying to catch up because the wavefront that has been measured has also already passed through the system. This causes a delay that can not be corrected anymore. There are multiple approaches to solve this issue. The first is to run the AO system at a high enough speeds that the atmospheric turbulence is frozen. This requires measuring the wavefront at speeds of several kHz. This approach are being tried at several instruments.[12] The other approach is to predict how the atmosphere is going to evolve and correct the wavefront errors before they are measured. Predictive control can lead to significant gains in post-processed contrast for high-contrast imaging. The post-processed contrast could be improved by a factor 100 to 1000, if predictive control is used and the temporal evolution of the atmosphere is predictable.[13–15]

We have recently developed the data-driven subspace predictive control (DDSPC) algorithm. This algorithm only uses the wavefront measurements and the past DM commands to determine the new optimal command. The DDSPC algorithm directly uses the closed-loop residuals, without reconstructing the full turbulence. The advantage of this approach is that we side step any reconstruction error due to model errors. However, that algorithm had only been implemented in Python which prohibited its use on-sky. In this proceeding we describe an implementation for the GPU with the Compute Unified Device Architecture (CUDA)[16] and show the speed gain. In Section 2, we discuss the CUDA implementation of the closed-loop data-driven subspace identification algorithm and its controller. In Section 3 we show the verification of the CUDA implementation in a lab setting and the speed improvement. And Section 4 concludes the manuscript and gives an outlook for on-sky demonstrations.

2. HIGH-SPEED IMPLEMENTATION WITH CUDA

2.1 Summary of the algorithm

The DDSPC algorithm is called data-driven and model free. This is a slight misnomer, because there is no controller that is truly model free. If models are called model free then what is meant is that there is no underlying parameteric model of which the parameters are optimized. The DDSPC algorithm uses an auto-regressive structure as its backbone. This results in the following model structure for a single DM mode or actuator,[17]

$$y_f^i = Ay_p^i + Bu_p^i + Cu_f^i. \tag{1}$$

Here y_f^i and u_f^i are the future vectors that contains the future N wavefront sensor measurements and DM commands at time step i. And y_p^i and u_p^i are the vectors that contain the M past wavefront sensor measurements and DM commands at time step i. This is a model that is completely linear in A, B and C. Therefore, these matrices can be estimated with a Linear-Least Squares (LLS) approach. The LLS problem is then given by,

$$y_f^i = \begin{bmatrix} A^i & B^i & C^i \end{bmatrix} \begin{bmatrix} y_p^i \\ u_p^i \\ u_f^i \end{bmatrix} = \Theta^i \phi^i \tag{2}$$

Here Θ^i is the concatenation of all model matrices, and ϕ^i is the concatenation of y_p^i, u_f^i and u_p^i. We have chosen to use recursive LLS (RLS) to learn the model because we want to learn the system dynamics online and have the ability to track changes. The mathematical details of the RLS implementation can be found in.[17] With the models in hand, the controller has to be found. The cost function for the controller is the quadratic sum of all future measurements and control commands,

$$J_i = y_f^{iT} y_f^i + \lambda u_f^{iT} u_f^i = \begin{bmatrix} y_p^{iT} & u_p^{iT} & u_f^{iT} \end{bmatrix} \begin{bmatrix} A^TA & A^TB & A^TC \\ B^TA & B^TB & B^TC \\ C^TA & C^TB & C^TC \end{bmatrix} \begin{bmatrix} y_p^i \\ u_p^i \\ u_f^i \end{bmatrix} + \lambda u_f^{iT} u_f^i. \tag{3}$$

Here J_i is the cost at iteration i and λ is a regularization parameter that determines how much the DM commands have to be dampened. After some algebra we find the optimal control signal as,

$$u_f = -\left(C^{iT}C^i + \lambda I\right)^{-1}\begin{bmatrix}A^{iT}C & B^{iT}C^i\end{bmatrix}\begin{bmatrix}y_p^i \\ u_p^i\end{bmatrix} = -K^i\begin{bmatrix}y_p^i \\ u_p^i\end{bmatrix}. \tag{4}$$

Here K_i is the controller at time step i. The derivation that was presented here assumes that all measurements and commands are for a single actuator or mode. This assumes that there is no cross-coupling between the modes. This allows us to create a distributed controller. First the wavefront sensor reconstructs the modal coefficients, generally with an interaction matrix, and the DDSPC filters are applied to each mode independently. This decoupling makes this algorithm naturally parallel. However, there is no reason why a model with all modal coupling added back would not work. The same equations will govern this model. The only downside is the increased computational complexity.

The system is first trained before the algorithm is deployed on realistic disturbance. This is necessary to make sure the algorithm will not get stuck in Null spaces of the model. A System identification (SI) approach is used to let the algorithm get familiar with the system it is controlling. Here we excite the system with a known disturbance by adding noise to the controller commands. For high-order systems, it is important to construct information-rich signals that are able to persistently excite all relevant frequencies. A popular choice is a random binary signal (RBS). During the learning sequence, a randomly generated binary signal will be added to the final computed commands.

2.2 Implementation in CUDA

The implementation as demonstrated in[17] was written in the Python programming language. This was the major limiting factor in speed to move from demonstration speeds ($\sim$ 100 Hz) to real production speeds ($\sim$ 1 kHz). Additionally, the python implementation could only run at 100 Hz with at most several hundred of controlled modes. Therefore, the most important step to move to an on-sky demonstration was the implementation in a lower-level programming language with less overhead. The main bread and butter of the algorithm consists of matrix-vector multiplications and linear algebra operations. Using GPUs for the calculations is therefore a natural choice. Additionally, the distributed nature of the algorithm makes it easier for GPUs because the problem itself is already completely parallel.

To make efficient use of the GPUs the data has to be formatted in a specific way in the GPU's memory. The CUDA library (and others) have a special implementation for distributed problems of the most common linear algebra operations. These are the xxxStridedBatched operations. An overview of usage and perfomance of the StridedBatched operations show that this almost reaches maximum performance compared to operations on a single large matrix.[18] The first speed and timing tests with a GPU implementation of the DDSPC have already shown that it is possible to control 1600 modes faster than 1.5 kHz in double precision and faster than 3 kHz in single precision. We ran several tests to investigate the actual gain in computational performance on the Real Time Computer (RTC) of MagAO-X. The RTC is equipped with several NVIDIA RTC 2080Ti, for the presented measurements we only used a single GPU. For the computational scaling we varied the number of controlled modes. The algorithm is expected to scale linearly with the number of modes because it is distributed and each mode will have equal computational burden. The prediction horizon is chosen to be $N = 3$ and we take $M = 4$ number of step into account from the past. The results are shown in Figure 1. The single precision CUDA implementation is the fastest as is expected. The python implementation is 3 to 4 orders of magnitude slower than the CUDA implementation. The double precision is a factor two slower compared to the single precision implementation if many modes (> 1000) are controlled, which is the expected degradation based on the number of bits that are used for the computation. The 1 kHz limit is reached for the double precision implementation at $\sim$ 4000 modes. At single precision 4000 modes can be controlled with less than 0.4 ms per iteration, which is a limiting loop speed of 2.5 kHz. All presented computations include online training of the model and the matrix-vector product that is required to reconstruct the modal coefficients. The number of measurement pixels is chosen to be 4400, which correspond to the number of active pixels in the reconstructed slope-maps of the MagAO-X pyramid wavefront sensor. These results are encouraging and show that this algorithm is expected to run above 1 kHz on-sky.

Because of the distributed nature of the algorithm, we expect that it is trivial to switch from a single GPU to multiple GPUs, which would allow us to run DDSPC for the ExAO systems of the future GSMTs on current hardware. For example, GMagAO-X (the visible direct imaging instrument proposed for the GMT) is expected to have 21000 thousand actuators with 3000 actuators per segment. The DDSPC algorithm would be able to run this xAO system by using 1 GPU per mirror segment.

Algorithm 1: Overview of the DDSPC algorithm

initialization;
while *closed-loop* **do**
  `// Reconstruct the wavefront modal coefficients with an interaction matrix.`
  $\vec{I}_i = \mathrm{WFS}(t_i)$;
  $\Delta\vec{y}_i = R\vec{I}_i$;
  add new measurement to history buffer($\Delta\vec{y}_i$);
  update predictor: $\Theta_{i-1} \rightarrow \Theta_i$;
  update controller: $K_{i-1} \rightarrow K_i$;
  if *learning* **then**
    $\vec{n}_i$ = random number $\in (-1, 0, 1)$;
  else
    $\vec{n}_i = 0$
  end
  $\Delta\vec{u}_i$ = calculate new command + $\tilde{\mathrm{n}}_\mathrm{i}$;
  add new command to history buffer($\Delta\vec{u}_i$);
end

3. OUTLOOK AND CONCLUSION

Here we discussed and presented the CUDA implementation of the data-driven subspace predictive control algorithm. Implementing the algorithm in CUDA decreased the computational time with more than 3 orders. This makes it possible to run the algorithm under realistic conditions on the telescope. The current implementation can make use of the distributed nature of the algorithm to run it at actuator numbers for future GSMTs. The DDSPC has been able to close the loop on 2000 modes with MagAO-X. We are currently fine tuning the hyper parameters of the controller to increase its performance. An on-sky demonstration of the new predictive controller is scheduled for December 2021.

ACKNOWLEDGMENTS

Support for this work was provided by NASA through the NASA Hubble Fellowship grant #HST-HF2-51436.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. This research made use of HCIPy, an open-source object-oriented framework written in Python for performing end-to-end simulations of high-contrast imaging instruments.[19]

REFERENCES

[1] Traub, W. A. and Oppenheimer, B. R., [*Direct Imaging of Exoplanets*], 111–156, University of Arizona Press (2010).
[2] Guyon, O., "Extreme adaptive optics," *Annual Review of Astronomy and Astrophysics* **56**, 315–355 (2018).
[3] Beuzit, J. L., Vigan, A., Mouillet, D., Dohlen, K., Gratton, R., Boccaletti, A., Sauvage, J. F., Schmid, H. M., Langlois, M., Petit, C., Baruffolo, A., Feldt, M., Milli, J., Wahhaj, Z., Abe, L., Anselmi, U., Antichi, J., Barette, R., Baudrand, J., Baudoz, P., Bazzon, A., Bernardi, P., Blanchard, P., Brast, R., Bruno, P., Buey, T., Carbillet, M., Carle, M., Cascone, E., Chapron, F., Charton, J., Chauvin, G., Claudi, R., Costille, A., De Caprio, V., de Boer, J., Delboulbé, A., Desidera, S., Dominik, C., Downing, M., Dupuis, O., Fabron, C., Fantinel, D., Farisato, G., Feautrier, P., Fedrigo, E., Fusco, T., Gigan, P., Ginski, C., Girard, J., Giro,

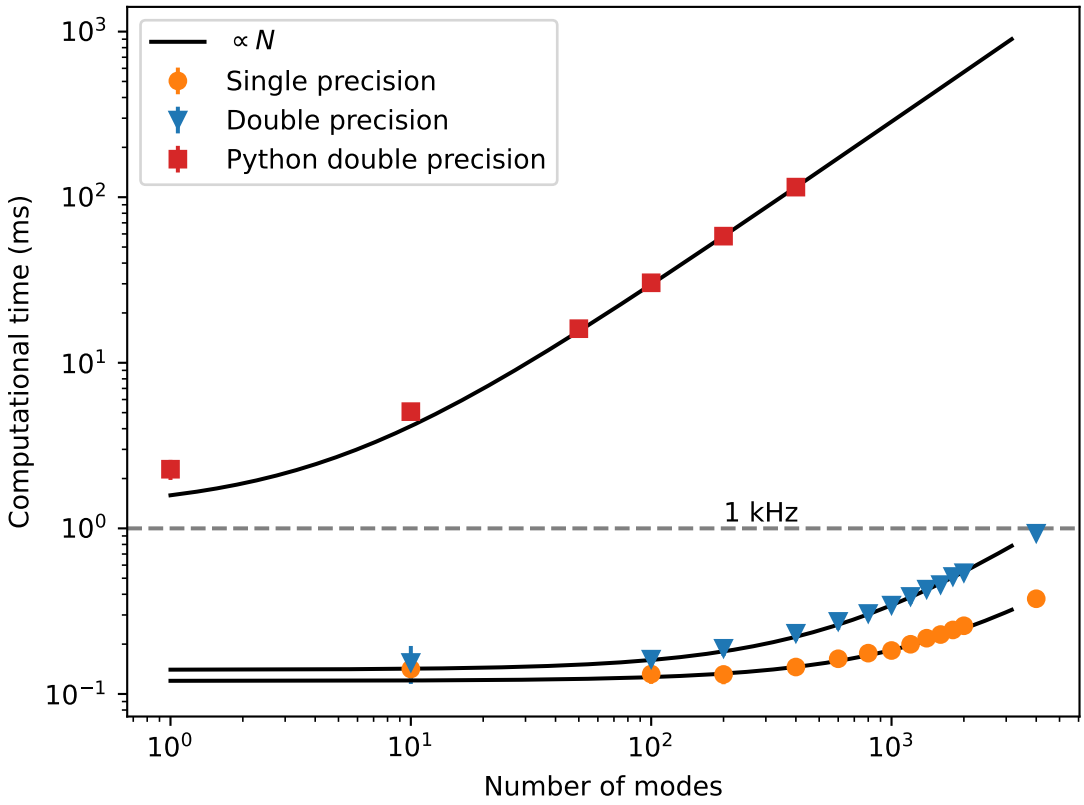

Figure 1. The computational time for a single iteration against the number of modes. The red squares show the timing measurements for the double precision Python implementation. The blue triangles and orange dots show the double and single precision implementation in CUDA, respectively. Each point has associated errorbars which are smaller than the marker sizes, which means that there is a small spread in computational performance between different iterations. The single precision CUDA implementation is the fastest with the double precision being a factor two slower if many modes (> 1000) are controlled. The python implementation is 3 to 4 orders of magnitude slower than the CUDA implementation.

E., Gisler, D., Gluck, L., Gry, C., Henning, T., Hubin, N., Hugot, E., Incorvaia, S., Jaquet, M., Kasper, M., Lagadec, E., Lagrange, A. M., Le Coroller, H., Le Mignant, D., Le Ruyet, B., Lessio, G., Lizon, J. L., Llored, M., Lundin, L., Madec, F., Magnard, Y., Marteaud, M., Martinez, P., Maurel, D., Ménard, F., Mesa, D., Möller-Nilsson, O., Moulin, T., Moutou, C., Origné, A., Parisot, J., Pavlov, A., Perret, D., Pragt, J., Puget, P., Rabou, P., Ramos, J., Reess, J. M., Rigal, F., Rochat, S., Roelfsema, R., Rousset, G., Roux, A., Saisse, M., Salasnich, B., Santambrogio, E., Scuderi, S., Segransan, D., Sevin, A., Siebenmorgen, R., Soenke, C., Stadler, E., Suarez, M., Tiphène, D., Turatto, M., Udry, S., Vakili, F., Waters, L. B. F. M., Weber, L., Wildi, F., Zins, G., and Zurlo, A., "SPHERE: the exoplanet imager for the Very Large Telescope," *A&A* **631**, A155 (Nov 2019).

[4] Marley, M. S., Fortney, J. J., Hubickyj, O., Bodenheimer, P., and Lissauer, J. J., "On the Luminosity of Young Jupiters," *ApJ* **655**, 541–549 (Jan. 2007).

[5] Bowler, B. P., Liu, M. C., Shkolnik, E. L., and Tamura, M., "Planets around Low-mass Stars (PALMS). IV. The Outer Architecture of M Dwarf Planetary Systems," *ApJS* **216**, 7 (Jan 2015).

[6] Nielsen, E. L., De Rosa, R. J., Macintosh, B., Wang, J. J., Ruffio, J.-B., Chiang, E., Marley, M. S., Saumon, D., Savransky, D., Ammons, S. M., Bailey, V. P., Barman, T., Blain, C., Bulger, J., Burrows, A., Chilcote, J., Cotten, T., Czekala, I., Doyon, R., Duchêne, G., Esposito, T. M., Fabrycky, D., Fitzgerald, M. P., Follette, K. B., Fortney, J. J., Gerard, B. L., Goodsell, S. J., Graham, J. R., Greenbaum, A. Z., Hibon, P., Hinkley, S., Hirsch, L. A., Hom, J., Hung, L.-W., Dawson, R. I., Ingraham, P., Kalas, P., Konopacky, Q., Larkin, J. E., Lee, E. J., Lin, J. W., Maire, J., Marchis, F., Marois, C., Metchev, S., Millar-Blanchaer, M. A., Morzinski, K. M., Oppenheimer, R., Palmer, D., Patience, J., Perrin, M., Poyneer, L., Pueyo, L., Rafikov, R. R., Rajan, A., Rameau, J., Rantakyrö, F. T., Ren, B., Schneider, A. C., Sivaramakrishnan, A., Song, I., Soummer, R., Tallis, M., Thomas, S., Ward-Duong, K., and Wolff, S., "The Gemini Planet Imager Exoplanet Survey: Giant Planet and Brown Dwarf Demographics from 10 to 100 au," *AJ* **158**, 13 (July 2019).

[7] Fernandes, R. B., Mulders, G. D., Pascucci, I., Mordasini, C., and Emsenhuber, A., "Hints for a Turnover at the Snow Line in the Giant Planet Occurrence Rate," *ApJ* **874**, 81 (Mar. 2019).

[8] Wagner, K., Apai, D., and Kratter, K. M., "On the Mass Function, Multiplicity, and Origins of Wide-orbit Giant Planets," *ApJ* **877**, 46 (May 2019).

[9] Kasper, M., "Adaptive optics for high contrast imaging," in [*Proc. SPIE*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **8447**, 84470B (2012).

[10] Milli, J., Mouillet, D., Fusco, T., Girard, J. H., Masciadri, E., Pena, E., Sauvage, J. F., Reyes, C., Dohlen, K., Beuzit, J. L., Kasper, M., Sarazin, M., and Cantalloube, F., "Performance of the extreme-AO instrument VLT/SPHERE and dependence on the atmospheric conditions," *arXiv e-prints* , arXiv:1710.05417 (Oct. 2017).

[11] Cantalloube, F., Absil, O., Bertram, T., Brandner, W., Delacroix, C., Feldt, M., Kenworthy, M., Kulas, M., Milli, J., Neureuther, P., Orban de Xivry, G., Pathak, P., Por, E., Scheithauer, S., Steuer, H., and van Boekel, R., "High contrast imaging with ELT/METIS: The wind driven halo, from SPHERE to METIS," *arXiv e-prints* , arXiv:1911.11241 (Nov. 2019).

[12] Males, J. R., Close, L. M., Guyon, O., Hedglen, A. D., Van Gorkom, K., Long, J. D., Kautz, M., Lumbres, J., Schatz, L., Rodack, A., et al., "Magao-x first light," in [*Adaptive Optics Systems VII*], **11448**, 114484L, International Society for Optics and Photonics (2020).

[13] Guyon, O. and Males, J., "Adaptive Optics Predictive Control with Empirical Orthogonal Functions (EOFs)," *arXiv e-prints* , arXiv:1707.00570 (July 2017).

[14] Males, J. R. and Guyon, O., "Ground-based adaptive optics coronagraphic performance under closed-loop predictive control," *Journal of Astronomical Telescopes, Instruments, and Systems* **4**, 019001 (Jan. 2018).

[15] Correia, C. M., Fauvarque, O., Bond, C. Z., Chambouleyron, V., Sauvage, J.-F., and Fusco, T., "Performance limits of adaptive-optics/high-contrast imagers with pyramid wavefront sensors," *MNRAS* **495**, 4380–4391 (June 2020).

[16] nVIDIA, "CUDA Toolkit." https://developer.nvidia.com/cuda-toolkit (2021). [Online; accessed 9-August-2021].

[17] Haffert, S. Y., Males, J. R., Close, L. M., Van Gorkom, K., Long, J. D., Hedglen, A. D., Guyon, O., Schatz, L., Kautz, M. Y., Lumbres, J., et al., "Data-driven subspace predictive control of adaptive optics for high-contrast imaging," *Journal of Astronomical Telescopes, Instruments, and Systems* **7**(2), 029001 (2021).

[18] Cecka, C., "Pro Tip: cuBLAS Strided Batched Matrix Multiply." https://developer.nvidia.com/blog/cublas-strided-batched-matrix-multiply/ (2017). [Online; accessed 27-July-2021].

[19] Por, E. H., Haffert, S. Y., Radhakrishnan, V. M., Doelman, D. S., van Kooten, M., and Bos, S. P., "High contrast imaging for python (hcipy): an open-source adaptive optics and coronagraph simulator," in [*Adaptive Optics Systems VI*], **10703**, 1070342, International Society for Optics and Photonics (2018).