

The EXPRES Stellar-Signals Project II. State of the Field in Disentangling Photospheric Velocities

LILY L. ZHAO,^{1,2} DEBRA A. FISCHER,¹ ERIC B. FORD,^{3,4,5,6,7} ALEX WISE,^{3,7} MICHAEL CRETIGNIER,^{8,9} SUZANNE AIGRAIN,^{10,11} OSCAR BARRAGAN,^{10,11} MEGAN BEDELL,^{2,12} LARS A. BUCHHAVE,^{13,11} JOÃO D. CAMACHO,^{14,15,16} HEATHER CEGLA,^{17,18} JESSI CISEWSKI-KEHE,¹⁹ ANDREW COLLIER CAMERON,^{20,21} ZOE L. DE BEURS,^{22,23,24} SALLY DODSON-ROBINSON,^{25,26,27} XAVIER DUMUSQUE,^{8,9} JOÃO P. FARIA,^{14,15,16} CHRISTIAN GILBERTSON,^{3,4,5,7} CHARLOTTE HALEY,^{28,27} JUSTIN HARRELL,^{25,27} DAVID W. HOGG,^{29,2,30,31,12} PARKER HOLZER,³² ANCY ANNA JOHN,^{20,21} BAPTISTE KLEIN,^{10,11} MARINA LAFARGA,^{17,18} CATHERINE LEMBO,^{25,27} FLORIAN LIENHARD,³³ VINESH MAGUIRE-RAJPAUL,^{33,11} ANNELIES MORTIER,^{33,34} BELINDA NICHOLSON,^{10,11} MICHAEL L. PALUMBO III,^{3,7} VICTOR RAMIREZ DELGADO,^{25,27} CHRISTOPHER J. SHALLUE,^{35,24} ANDREW VANDERBURG,^{36,23,24} PEDRO T. P. VIANA,^{14,15,16} JINGLIN ZHAO,^{3,7} NORBERT ZICHER,^{10,11} SAMUEL H. C. CABOT,¹ GREGORY W. HENRY,³⁷ RACHAEL M. ROETTENBACHER,^{38,1} JOHN M. BREWER,³⁹ JOE LLAMA,⁴⁰ RYAN R. PETERSBURG,⁴¹ AND ANDREW E. SZYMKOWIAK¹

¹Department of Astronomy, Yale University, 52 Hillhouse Ave., New Haven, CT 06511, USA

²Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York, NY 10010, USA

³Department of Astronomy & Astrophysics, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA, 16802, USA

⁴Center for Exoplanets and Habitable Worlds, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA, 16802, USA

⁵Institute for Computational & Data Sciences, The Pennsylvania State University, University Park, PA, 16802, USA

⁶Institute for Advanced Sciences

⁷PennState Team

⁸Astronomy Department of the University of Geneva, 51 Chemin de Pegasi 51, 1290 Versoix, Switzerland

⁹Geneva Team

¹⁰Sub-department of Astrophysics, Department of Physics, University of Oxford, Oxford OX1 3RH, UK

¹¹OxBridGen Team

¹²CCA Team

¹³DTU Space, National Space Institute, Technical University of Denmark, Elektrovej 328, DK-2800 Kgs. Lyngby, Denmark

¹⁴Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762, Porto, Portugal

¹⁵Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

¹⁶Porto Team

¹⁷Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

¹⁸Warwick Team

¹⁹Department of Statistics, University of Wisconsin-Madison, 1300 University Ave., Madison, WI 53706, USA

²⁰University of St Andrews, Centre for Exoplanet Science, SUPA, School of Physics & Astronomy, North Haugh, St Andrews KY16 9SS, UK

²¹Andrews Team

²²Department of Astronomy, University of Texas at Austin, 2515 Speedway, Austin, Texas 78712, USA

²³Department of Astronomy, University of Wisconsin-Madison, Madison, WI, 53706, USA

²⁴ML_EPRVs Team

²⁵Department of Physics and Astronomy, University of Delaware, 217 Sharp Lab, Newark, DE 19716, USA

²⁶Bartol Research Institute, Sharp Lab, 104 The Green, Newark, DE, 19716, USA

²⁷Sidera Team

²⁸Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL

²⁹Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA

³⁰Center for Data Science, New York University, 60 Fifth Avenue, New York, NY 10011, USA

³¹Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

³²Department of Statistics and Data Science, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA

³³Astrophysics Group, Cavendish Laboratory, University of Cambridge, J.J. Thomson Avenue, Cambridge CB3 0HE, UK

³⁴Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

³⁵Center for Astrophysics—Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA

³⁶Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

³⁷Center of Excellence in Information Systems, Tennessee State University, Nashville, TN 37209, USA

³⁸Yale Center for Astronomy and Astrophysics, Yale University, 46 Hillhouse Avenue, New Haven, CT 06511, USA

³⁹San Francisco State University University, 1600 Holloway Ave., San Francisco, CA 94132, USA

⁴⁰Lowell Observatory, 1400 W. Mars Hill Rd., Flagstaff, AZ 86001, USA

⁴¹Department of Physics, Yale University, 217 Prospect St, New Haven, CT 06511, USA

ABSTRACT

Measured spectral shifts due to intrinsic stellar variability (e.g., pulsations, granulation) and activity (e.g., spots, plages) are the largest source of error for extreme precision radial velocity (EPRV) exoplanet detection. Several methods have been developed to disentangle stellar signals from true center-of-mass shifts due to planets. The EXPRES Stellar Signals Project (*ESSP*) presents a self-consistent comparison of 21 different methods tested on the same extreme-precision spectroscopic data from EXPRES. Methods either derived new activity indicators, constructed new models for mapping an indicator to the needed RV correction, or separated out shape- and shift-driven RV components. Method results were compared based on the total and nightly scatter of returned RVs, agreement with the results of other methods, and correlation with activity indicators. Though nearly all submitted methods do better than the classic linear decorrelation method for mitigating stellar signals, no method is yet consistently reducing the RV RMS to the necessary sub-meter-per-second levels. There exists a concerning lack of agreement between the RVs returned by different methods, even for those returning similar final RV RMS values. This highlights the incompleteness of using the RMS alone to assess method performance, which is a common practice that should be used with more caution. Our results suggest that continued progress in this field necessitates increased interpretability of methods, high-cadence data to capture stellar signals at all time scales, and continued tests like the *ESSP* using consistent data sets with more advanced metrics for method performance. Future comparisons should make use of well-characterized data sets, such as solar data or data with known injected planetary and/or stellar signals, to better understand method performance and whether planetary signals are preserved.

Keywords: Exoplanet detection methods (489), Radial velocity (1332), Planet hosting stars (1242), Stellar activity (1580), Spectrometers (1554)

1. INTRODUCTION

With the new generation of extreme-precision spectrographs, sub-meter-per-second radial velocity (RV) measurement precision has become achievable (Pepe et al. 2013; Schwab et al. 2016; Jurgenson et al. 2016; Blackman et al. 2020; Petersburg et al. 2020; Suárez Mascaño et al. 2020; Brewer et al. 2020; Pepe et al. 2021). Photospheric velocities from intrinsic stellar variability and activity features are now the dominant source of RV scatter.

A star’s radial velocity is measured by modeling Doppler shifts in absorption lines of stellar spectra. Different forms of stellar variability will change spectra such that lines will appear shifted, deeper/shallower, or asymmetric. These line shape changes can be mistaken for true center-of-mass shifts in the RV analysis. In this way, stellar signals will add errors to the resultant

RV measurements and can even masquerade as periodic, false planet signals.

With RV precision better than one meter per second, we must contend with obscuring photospheric velocities that arise from stellar oscillations (Mayor et al. 2003; Bouchy et al. 2005; Kjeldsen et al. 2005; Arentoft et al. 2008; Chaplin et al. 2019), granulation (Dravins 1982; Kjeldsen & Bedding 1995; Lindegren & Dravins 2003; Dumusque et al. 2011b; Meunier et al. 2015; Cegla et al. 2018; Lanza et al. 2019), super granulation (Rieutord & Rincon 2010; Rincon & Rieutord 2018; Meunier & Lagrange 2019), and activity features such as spots, faculae, or plages (Saar & Donahue 1997; Hatzes 2002; Saar 2003; Desort et al. 2007; Huélamo et al. 2008; Boisse et al. 2011; Dumusque et al. 2011a; Lewis et al. 2011; Jeffers et al. 2013; Cabot et al. 2021; ?). The various types of photospheric velocities imprint on a star’s spectrum in different, potentially quasi-periodic ways and evolve on a range of timescales.

Pressure gradients in the convective zones of stars result in p-mode oscillations with time scale of a few minutes, where the frequency and amplitude of these oscillations increases slightly with T_{eff} (Mayor et al. 2003; Bouchy et al. 2005; Kjeldsen et al. 2005; Arentoft et al. 2008). This movement can cause RV variations from 10 cm s $^{-1}$ up to 1 m s $^{-1}$ (Dumusque et al. 2011c; Chaplin et al. 2019).

Solar-type stars will also exhibit granulation patterns, which arise from convection in the outer layers of the star (Dravins 1982; Lanza et al. 2019; Kjeldsen & Bedding 1995; Lindegren & Dravins 2003; Dumusque et al. 2011b; Cegla et al. 2018). Upflow in the middle of granulation cells appear blueshifted while the downflow in the narrow, dimmer edge regions appear redshifted. This creates a net RV blueshift, known as convective blueshift, which can lead to asymmetries in spectral lines.

Granulation changes on timescales of a few minutes to hours, which integrates to different net RV shifts. These changes result in varying magnitudes of the convective blueshift and its resultant spectral line shape changes. This effect can introduce random RV variations of 0.4 to 0.8 m s $^{-1}$ and increases with the T_{eff} of the star (Meunier et al. 2015). Supergranulation describes large cells outlined by the magnetic network, which can persist for hours to up to two days (Rieutord & Rincon 2010; Rincon & Rieutord 2018; Meunier & Lagrange 2019). They give rise to similar issues as granulation and can introduce RV variations of 0.3 to 0.7 m s $^{-1}$ (Meunier et al. 2015).

Strong magnetic activity can also generate activity features in the photosphere of a star, i.e., darker starspots or brighter faculae and plages (Saar & Donahue 1997; Hatzes 2002; Saar 2003; Desort et al. 2007; Huélamo et al. 2008; Boisse et al. 2011; Dumusque et al. 2011a; Lovis et al. 2011; Jeffers et al. 2013). This magnetic activity will suppress convection in a star and change the magnitude of convective blueshift relative to a quiet photosphere, resulting in an integrated RV change of 0.4 to 1.4 m s $^{-1}$ (Meunier et al. 2010).

Activity features rotate in and out of view as the star rotates. Spots, with a lower temperature, suppress flux while faculae and plages, with a higher temperature, increase flux. The presence of activity features therefore changes the integrated flux distribution of the star. As a star rotates, the side of the star rotating towards the observatory appears blueshifted while the side rotating away appears redshifted. If the same amount of flux is coming from both sides (i.e., the star is featureless), these effects cancel each other out. Changes in flux due to activity features can break that balance and introduce up to 0.4 m s $^{-1}$ variations on the Sun (Saar & Donahue

1997; Meunier et al. 2010). The different temperatures of activity features locally modify absorption and emission processes and produce asymmetry in the integrated spectral line profiles that vary with stellar rotation.

Traditionally, stellar signals have been decorrelated from radial-velocity measurements with the use of activity indicators. These indicators aim to gauge magnetic activity on the target star and/or presence of activity features for each exposure (e.g., Boisse et al. 2009; Dumusque et al. 2011c; Figueira 2013; Holzer et al. 2021). Popular indicators include properties of the cross-correlation function (CCF) used to derive RVs, such as various CCF bisector asymmetry measurements (e.g., Queloz et al. 2001; Povich et al. 2001) or the full-width half max of the CCF (e.g., Queloz et al. 2009). The CCF can be thought of as an average of all line shapes in the spectrum, and is therefore only sensitive to line shape changes that appear in most lines. Magnetic activity on the star has also been shown to correlate with localized spectral features including emission in the core of Ca II H&K lines (396.96 nm and 393.47 nm respectively; Saar et al. 1998; Meunier & Lagrange 2013), the Ca infrared triplet (849.8, 854.2, and 866.2 nm; Saar & Fischer 2000), and the H- α line (656.28 nm; Skelly et al. 2008; Robertson et al. 2014; Giguere et al. 2016).

Linearly decorrelating RVs against activity indicators has not been successful at disentangling stellar signals to sub-meter-per-second precision (Fischer et al. 2016). Recently, more advanced methods have been proposed to disentangle stellar signals from true center-of-mass RV shifts. Gaussian process (GP) models have been used to more flexibly model stellar signals (Haywood et al. 2014; Rajpaul et al. 2015; Faria et al. 2016; Rajpaul et al. 2017; Angus et al. 2018; Jones et al. 2021; Gilbertson et al. 2020a). Methods using different activity indicators and a Bayesian framework were found to deliver more robust results (Dumusque et al. 2017).

There has also been a move towards capturing the effects of stellar activity at the level of the 1D spectrum, i.e., before calculating the CCF and extracting RVs (e.g., Davis et al. 2017; Dumusque 2018; Meunier et al. 2017; Jones et al. 2021). The use of pixel-level statistical techniques has revealed that different lines show different behaviors and levels of sensitivity to stellar activity.

With many promising methods being developed to address the issue of stellar signals, we present here a head-to-head comparison of many of these methods on real data. For four stars—HD 101501, HD 26865, HD 10700, and HD 34411—the EXPRES Stellar-Signals Project released high-fidelity Extreme Precision Spectrograph (EXPRES) data that are representative of next-

Table 1. Stellar Parameters

	HD 101501	HD 26965	HD 10700	HD 34411
Spectral Type	G8V B	K1V	G8V	G0V
V	5.34 (e)	4.43 (d)	3.49 (d)	4.71 (e)
$B-V$	0.74 (e)	0.82 (g)	0.72 (e)	0.62 (e)
$\log R'_{\text{HK}}$	-4.483 (g)	-4.928 (g)	-4.976 (g)	-5.085 (g)
Dist. [pc]	9.541 (f)	4.98 (f)	3.65 (k)	12.484 (f)
RV [km s $^{-1}$]	-5.6 (f)	-42.269 (f)	-16.597 (f)	66.57 (i)
$L_{\text{star}} [L_{\odot}]$	0.609 ± 0.009 (b)	0.457 ± 0.002 (f)	0.52 ± 0.03 (j)	1.732 (b)
$R_{\text{star}} [R_{\odot}]$	0.86 ± 0.02 (c)	0.83 ± 0.02 (c)	0.82 ± 0.02 (c)	1.28 ± 0.04 (c)
$M_{\text{star}} [M_{\odot}]$	0.9 ± 0.12 (c)	0.8 ± 0.11 (c)	0.99 ± 0.13 (c)	1.08 ± 0.14 (c)
$T_{\text{eff}} [\text{K}]$	5502 (c)	5092 (c)	5333 (c)	5873 (c)
$\log g$	4.52 (c)	4.51 (c)	4.6 (c)	4.26 (c)
[Fe/H]	-0.04 (c)	-0.3 (c)	-0.53 (c)	0.1 (c)
Age [Gyr]	3.5 (c)	12.8 (c)	12.4 (c)	4.8 (c)
$v \sin i$ [km s $^{-1}$]	2.2 (c)	0.5 (c)	1.6 (c)	0.1 (c)
P_{rot} [days]	17.1 (h)	40 (g)	34 (a)	

NOTE—(a) Baliunas et al. (1996); (b) Boyajian et al. (2012); (c) Brewer et al. (2016); (d) Brewer & Fischer (2016); (e) Ducati (2002); (f) Gaia Collaboration (2018); (g) Isaacson & Fischer (2010); (h) Maldonado et al. (2010); (i) Nidever et al. (2002); (j) Pijpers (2003); (k) van Leeuwen (2007)

generation spectrographs as well as differential photometry from the Fairborn Automatic Photoelectric Telescopes (APTs) (Zhao et al. 2020). Eleven teams tested 15 different methods (with different variations for a total of 21 different methods) on the data provided. All methods used the provided data products (i.e., spectra, CCFs, RVs, and/or derived activity indicators), which allows us to compare the performance of methods on exactly the same data. No methods made use of the provided photometry.

The data and targets are described more in Section 2. Section 3 gives an overview of all methods tested and highlights commonalities between methods (with longer descriptions included in the Appendix). The resulting RVs produced by different methods are compared in Section 4. Section 5 gives a summary of all methods and the pertinent results. Section 6 discusses the different assumptions made by methods that define the current state of the field. From there, we make suggestions for future method development and data challenges. We conclude in Section 7.

2. DATA

The data sets for the ESSP include spectroscopic data from EXPRES and ground-based photometric measurements from the APTs for four targets—HD 101501 (61 UMa), HD 26965 (omi² Eri), HD 10700 (τ Ceti), and HD 34411 (λ Aur). Here, we describe the EXPRES and APTs instruments, as well as the four targets. We provide benchmarks for the amount of RV scatter that is expected for the EXPRES instrument and pipeline with minimal contributions from stellar signals. Stellar parameters for each target are tabulated in Table 1.

2.1. Spectroscopic Data From EXPRES

EXPRES is an optical (390 – 780 nm), fiber-fed spectrograph with a median resolution of $R \sim 137,000$ (Jurgenson et al. 2016). The instrument was fully commissioned at the 4.3-m Lowell Discovery Telescope (LDT) (Levine et al. 2012) near Flagstaff, AZ in January 2019 and is being used for a Doppler planet survey on about 125 (partial) nights per year. The spectrograph is housed in a vacuum enclosure to achieve temperature and pressure stabilization. A Menlo Systems laser frequency comb (LFC; Wilken et al. 2012; Molaro et al. 2013; Probst et al. 2014, 2020; Milaković et al. 2020) ranging from \sim 490–730 nm is used for precise wavelength calibration.

The instrument calibration stability for EXPRES ranges between 3–7 cm s $^{-1}$ using consecutive LFC spectra taken for over thirty minutes to an hour (Blackman et al. 2020). Figure 1 shows the RV scatter of over an hour of consecutive LFC exposures. The RMS is 3.21 cm s $^{-1}$ after a linear trend is removed. The linear trend is thought to be due to changing instrument temperature and is accounted for in precision RV work via the wavelength calibration. The instrument calibration stability can be thought of as the minimum RMS achievable by the EXPRES hardware as it measures the degree of scatter that cannot be calibrated out.

An exposure meter picks off 2% of the light from behind the fiber entrance to the spectrograph to monitor the photon flux for chromatic barycentric corrections. This exposure meter system also terminates exposures when the target signal-to-noise ratio (SNR) of 250 per pixel at a wavelength of about 550 nm is reached.

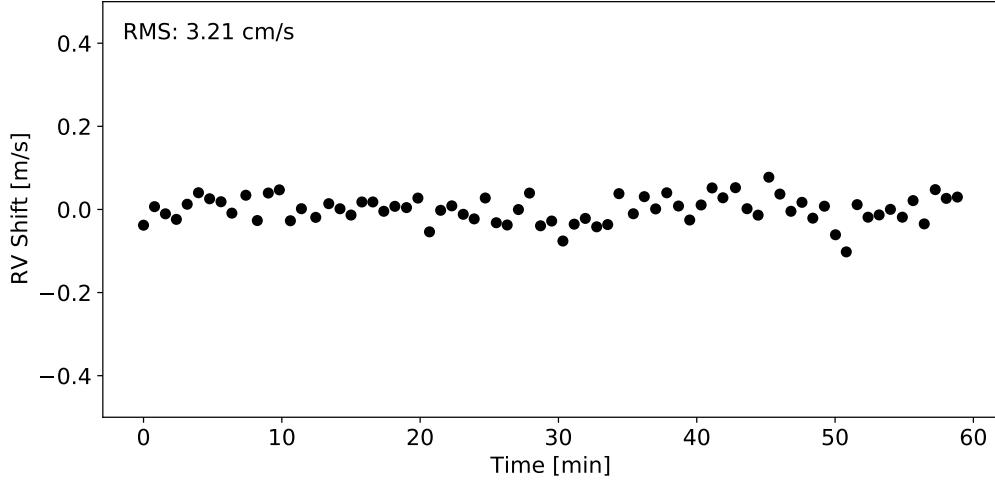


Figure 1. Perceived shift in LFC spectra in units of m s^{-1} across an hour of consecutive LFC exposures. These perceived shifts are attributed to variations in the instrument and therefore give a measure of the instrument stability. The RMS of shifts across this hour is given in the top-left corner.

Two or three consecutive exposures separated only by read-out time are obtained for each target star per night to improve the nightly-binned precision (Brewer et al. 2020). The on-sky, analytical single-measurement precision for exposures with a SNR of 250 (per pixel at $\lambda=550$ nm) is about 0.3 m s^{-1} (Petersburg et al. 2020). This matches the typical intranight RMS scatter for consecutive observations.

One-dimensional spectra are extracted using a flat-relative, optimal extraction pipeline (Zechmeister et al. 2014; Petersburg et al. 2020). Extracted spectra were made available to the ESSP participants along with chromatic barycentric-corrected wavelengths (Blackman et al. 2017). Two sets of wavelengths are provided, one set with a classic polynomial wavelength solution, and one set generated using *excalibur*, a hierarchical, non-parametric wavelength solution (Zhao et al. 2021). The provided spectra also included a model of telluric lines generated using *SELENITE* (Leet et al. 2019), a continuum model, and the associated blaze function to recover photon counts.

In addition to extracted spectra, cross-correlation functions (CCF), forward-modeled RVs, and classic activity indicators were provided for each observation. Un-

less otherwise stated, teams used the provided spectra, CCFs, RVs, and activity indicators as inputs to their methods, thereby ensuring a consistent comparison between the different method results. Table 2 gives the number of RV measurements, the number of nights where spectra were acquired, and the time baseline for each data set.

2.1.1. Default RVs

The standard EXPRES pipeline derives RV measurements using a forward-modeling technique (Brewer et al. 2020). A template spectrum is constructed using three consecutive observations taken on one night of a given target star. Each observed spectra is then broken into two-angstrom chunks that are shifted and scaled to match this template spectrum. Chunk-by-chunk (CBC) RVs are derived for each exposure by finding the weighted average of all chunks in a spectra. The weights for this average are empirically generated based on the stability of each chunk across all observations. Chunks that are more stable over time are weighted higher while chunks that return higher RV scatter are down weighted. CBC RVs for all four targets are given in Table 3.

CBC RVs derived from on-sky EXPRES data regularly return sub-meter-per-second RMS and intra-night scatter (INS) that matches the analytical 30 cm s^{-1} errors. Figure 2 depicts RVs from six photospherically quiet stars observed with EXPRES (which are not a part of this study). The nightly-binned RMS of these pipeline RV measurements range from 0.5 to 0.8 m s^{-1} . The average INS over nights (using only nights with more than one observation) ranges from 0.1 to 0.4 m s^{-1} . These stars demonstrate the RV precision achievable by EXPRES data in the absence of strong photospheric veloc-

Table 2. Spectroscopic Observations

Target	No. Obs.	Nights	Date Range
HD 101501	45	22	Feb. 10, 2019 - Nov. 26, 2020
HD 26965	114	37	Aug. 20, 2019 - Nov. 27, 2020
HD 10700	174	34	Aug. 15, 2019 - Nov. 27, 2020
HD 34411	188	58	Oct. 8, 2019 - Nov. 27, 2020

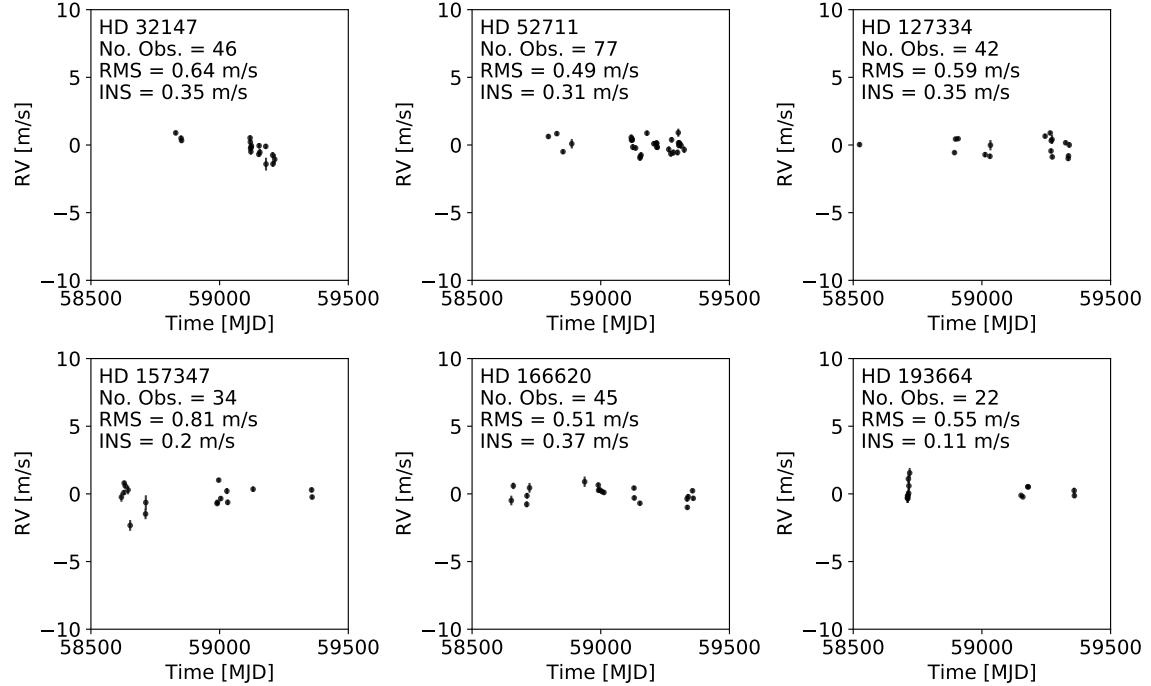


Figure 2. EXPRES RVs for six quiet stars. Shown RVs are derived using a chunk-by-chunk (CBC) forward modeling scheme and binned by night. The RMS of these nightly-binned RVs are given in the top-right corner along with the average intra-night scatter (INS).

ties adding scatter. Complete mitigation of RV contribution from stellar signals should result in a similar final RMS value.

Table 3. CBC RVs

Target	Time [MJD]	RV [m s^{-1}]	Err. [m s^{-1}]
HD 101501	58524.466	-0.338	0.322
HD 101501	58524.491	2.38	0.325
HD 101501	58524.497	2.66	0.308
:	:	:	:
HD 26965	58715.487	-0.101	0.435
HD 26965	58719.469	-1.85	0.368
HD 26965	58719.472	-1.44	0.408
:	:	:	:
HD 10700	58710.457	0.075	0.388
HD 10700	58710.458	-2.25	0.377
HD 10700	58710.46	-3.03	0.387
:	:	:	:
HD 34411	58764.475	3.47	0.324
HD 34411	58764.477	1.98	0.34
HD 34411	58764.479	4.8	0.314
:	:	:	:

NOTE—A stub of this table is provided here for reference; the full RV data sets are published online.

2.1.2. Default CCFs

We provided CCFs as well as the resultant CCF RVs for each spectra. These CCFs were generated using the code described in Ford et al. (2021), CCF masks based on the publicly available ESPRESSO masks of the closest matching spectral type, and a rectangular window function. We found the CBC RVs to have consistently lower RV scatter than the CCF RVs, and so the default RVs for methods were the CBC RVs. Due to differences in the weighting schemes between the CBC Pipeline and construction of a CCF, it is expected that the two methods carry different sensitivities to changes in the spectra.

Since the EXPRES pipeline returns flat-relative extractions, it was important to account for the varying SNR of each line. Lines for the CCF were weighted using the product of the ESPRESSO-mask-provided weight and a constructed weight factor based on the median signal-to-noise (SNR) ratio (assuming only photon-noise). For lines that show up in multiple orders, the SNR weight factor was computed separately for the line in each order.

Lines which overlapped with a telluric feature (as identified by SELENITE) during any observation were rejected for all observations. Lines that were shifted beyond the edge of a given order during any observation were excluded from use within that order for all observations. Only pixels with a wavelength calibration

from the LFC (\sim 490-730 nm) were used to construct the CCFs.

2.1.3. Default Activity Indicators

Each observation was also accompanied by several common activity indicators and their empirically determined errors. Spectroscopic activity indicators provided include the S -value, a measure of the emission in the core of the Ca II H&K lines (Meunier & Lagrange 2013; Saar et al. 1998), and measures of changes in the $H\alpha$ line core emission (Skelly et al. 2008; Robertson et al. 2014; Giguere et al. 2016). Both the $H\alpha$ emission (i.e., a measure of the depth of the normalized $H\alpha$ line), and the equivalent width of the $H\alpha$ line were provided.

We also provided a number of activity indicators derived from the CCF as described here. The difference in the center of the CCF at the top of the CCF as compared to the bottom of the CCF is a measure of the skew in the CCF bisector (CCF BIS) (Queloz et al. 2001) and the velocity span indicator (V_{span}) (Boisse et al. 2011). Varying spectral line profiles will result in changes to the CCF full width at half maximum (FWHM), which is often used as an activity indicator (e.g., Queloz et al. 2009). We also provide the results of fitting the CCF to various, asymmetric profiles, such as a bi-Gaussian (Figueira 2013) or a skew normal probability density function (Simola et al. 2019), where the asymmetry parameter of these profiles is reported as an activity indicator.

Analytical errors are provided where possible; otherwise, empirical errors were determined by finding the spread in calculated indicators for nine chromospherically quiet stars¹. Using a total of approximately 400 observations of these seven stars, a histogram of indicator values was plotted to reveal a Gaussian shape. The standard deviation of a Gaussian fit to the histogram is taken to be the empirical error for the given activity indicator.

2.2. Photometry from APT

Ground-based photometry for all four *ESSP* target stars was obtained with either the T4 0.75 m or T12 0.8 m Automatic Photoelectric Telescopes (APTs) at Fairborn Observatory in southern Arizona. T4 observed HD 101501, HD 26965, and HD 10700, while T12 observed HD 34411. Table 4 describes the number of photometric observations for each target.

The T4 APT is equipped with a single channel photometer that uses an EMI 9124QB bi-alkali photomul-

tiplier tube to measure the difference in brightness between the program star and three nearby comparison stars in the Strömgren b and y passbands. The T12 APT has a two-channel photometer that uses a dichroic filter to separate the Strömgren b and y passbands so that separate EMI 9124QB photomultiplier tubes can measure the two colors simultaneously. To improve the photometric precision, we combine the differential b and y magnitudes into a single $(b+y)/2$ “passband”. The right hand column of Figure 3 shows the light curves of each star, covering between 13 and 28 observing seasons.

The precision of a single observation with these APTs, as measured from pairs of constant comparison stars, is around 0.0010 to 0.0015 mag on good nights. The T4 and T8 (a twin of T12) APTs are described in Henry (1999), where further details of the telescope, precision photometers, and observing and data reduction procedures can be found.

In each photometric data set, we identify a long-term trend that is modeled by applying Gaussian smoothing to the light curve with a 100-day window. A window of 100 days was chosen to find trends on the order of observing seasons while preserving the signal on the timescale of individual stellar rotations. These trends can be subtracted off to remove large-scale variation in the photometry from stellar signals (e.g., variations across activity cycles). Photometry and this smooth trend is given for all four targets in Table 5.

The photometric data were interpolated to the time stamps of the given spectroscopic data and associated RVs using a Gaussian process (GP) model with a quasi-periodic kernel (Rasmussen & Williams 2006) and implemented with the george package (Ambikasaran et al. 2015). This kernel depends on four hyperparameters $\phi = \{a^2, \lambda_e, \lambda_p, P_{\text{GP}}\}$, which correspond to the covariance amplitude, decay timescale (which is related to the typical spot evolution timescale), a smoothing parameter for the periodic term, and periodic timescale (which is related to the stellar rotation period), respectively. This kernel is used frequently for photometric modeling and stellar activity characterization in the literature (e.g. Haywood et al. 2014; Angus et al. 2018).

A GP was trained on the most recent six years of APT data for each star after first determining the best-fit hyperparameters via nested sampling. While the GP regression returned reasonable interpolated median values and 1σ uncertainties, it failed to estimate well-principled extrapolated photometric values for RV timestamps taken after the last photometric measurement. This behavior is expected past a few times the typical spot lifetimes on stars, which is typically of

¹ HD 32923, HD 34411, HD 84737, HD 86728, HD 158633, HD 166620, HD 182488, HD 186427, HD 217014

Table 4. Photometric Observations

Target	No. Obs.	Nights	Date Range
HD 101501	3290	2113	Apr. 18, 1993 - Jun. 22, 2020
HD 26965	1631	1500	Sep. 9, 1993 - Feb. 20, 2020
HD 10700	1369	1007	Nov. 5, 1996 - Jan. 24, 2020
HD 34411	1214	816	Nov. 25, 2005 - Apr. 3, 2018

Table 5. Photometry and Long-Term Trend

Target	Time [MJD]	$(b + y)/2$ [mag]	Trend [mag]
HD 101501	49095.696	-0.00145	-0.653
HD 101501	49095.782	-0.0023	-0.653
HD 101501	49096.783	0.00425	-0.653
	:	:	:
HD 26965	49239.941	-0.00231	-2.29
HD 26965	49245.933	0.00084	-2.29
HD 26965	49246.93	0.00139	-2.29
	:	:	:
HD 10700	50392.762	-0.00435	-2.63
HD 10700	50396.743	0.00115	-2.63
HD 10700	50397.735	0.00325	-2.63
	:	:	:
HD 34411	53699.829	0.00075	-1.11
HD 34411	53700.842	0.00265	-1.11
HD 34411	53702.821	-0.00085	-1.11
	:	:	:

NOTE—A stub of this table is provided here for reference; the full photometric data sets are published online.

der tens of days, although may be longer for young stars (Bradshaw & Hartigan 2014).

2.3. Targets

The four *ESSP* stars, as described in Table 1, are being observed as part of the EXPRES 100 Earths survey (Brewer et al. 2020). The targets range in activity level as predicted by $\log R'_{\text{HK}}$ values. Figure 3 shows the measured RVs and photometry.

HD 101501 is the most chromospherically active (Isaacson & Fischer 2010, $\log R'_{\text{HK}} = -4.483$) of the four *ESSP* targets. The EXPRES RVs exhibit an RMS scatter of 4.89 m s^{-1} . A GP model of this data preconditioned on photometry found a statistical preference for an activity-only model (Cabot et al. 2021). The provided HD 101501 data set has the fewest number of observations of the four though a longer time baseline.

HD 26965 is a K1V star with $\log R'_{\text{HK}} = -4.928$ (Isaacson & Fischer 2010) that exhibits an RV RMS scatter of 3.19 m s^{-1} . Previous RV analysis of HD 26965 using

HIRES (Vogt et al. 1994), PFS (Crane et al. 2006), CHIRON (Tokovinin et al. 2013), and HARPS (Mayor et al. 2003) RV data from 2001 to 2016 revealed a periodic signal of about 42.364 days (Díaz et al. 2018). Additional data from the Dharma Planet Survey, which added RVs collected from 2014 to 2015, concluded that there exists a 42.38 day periodic signal from a $K = 1.8 \text{ m s}^{-1}$ planet, and that the stellar rotation rate of the star measured from stellar activity indicators is between 39-44.5 days (Ma et al. 2018). Analysis using the complete set of RV data from the California Legacy Survey (CLS), taken from 1987 through 2020, attributes the periodicity to stellar signals (Rosenthal et al. 2021).

HD 10700, i.e., τ Ceti, is an older (Brewer et al. 2016, 12.4 Gyr) and chromospherically quiet (Isaacson & Fischer 2010, $\log R'_{\text{HK}} = -4.976$) star. The EXPRES RVs exhibit an RMS scatter of 1.8 m s^{-1} dominated by a five-minute periodic variation that matches what we would expect from p-modes. Seven planet candidates have been published, though three of these signals (planet candidates b, e, and d) were later retracted (Tuomi et al. 2013; Feng et al. 2017). Typically, three to five consecutive EXPRES observations are taken of τ Ceti. On August 25, 2019 and October 8, 2019, more than twenty back-to-back observations were taken within the night (covering a span of approximately 40 minutes) to resolve for p-mode oscillations.

HD 34411 is most similar to the Sun of all four targets; it is a 4.8 Gyr, G0V star (Brewer et al. 2016; Gaia Collaboration 2018). The star has low chromospheric activity, with $\log R'_{\text{HK}} = -5.085$ (Isaacson & Fischer 2010). The EXPRES RVs show an RMS scatter of 1.78 m s^{-1} .

3. METHODS

The *ESSP* included fifteen different methods (some with variations) all with the goal of isolating true center-of-mass shifts. Table 6 lists all methods along with variations on each method. The “Input” column specifies the primary type of provided data that was input into the method (i.e., the extracted spectra, the CCF, or the RVs along with activity indicators). The “Run Time” column gives an estimate of the computational expense of the method by specifying what the method was run on and the order of magnitude of time it took to run. Related publications to each method are given where available; otherwise, the name of the most pertinent author to contact for each method is listed.

A short description of each method is given below including any specifics to the implementation represented here and data requirements. Similar methods are compared and contrasted. Longer descriptions of each indi-

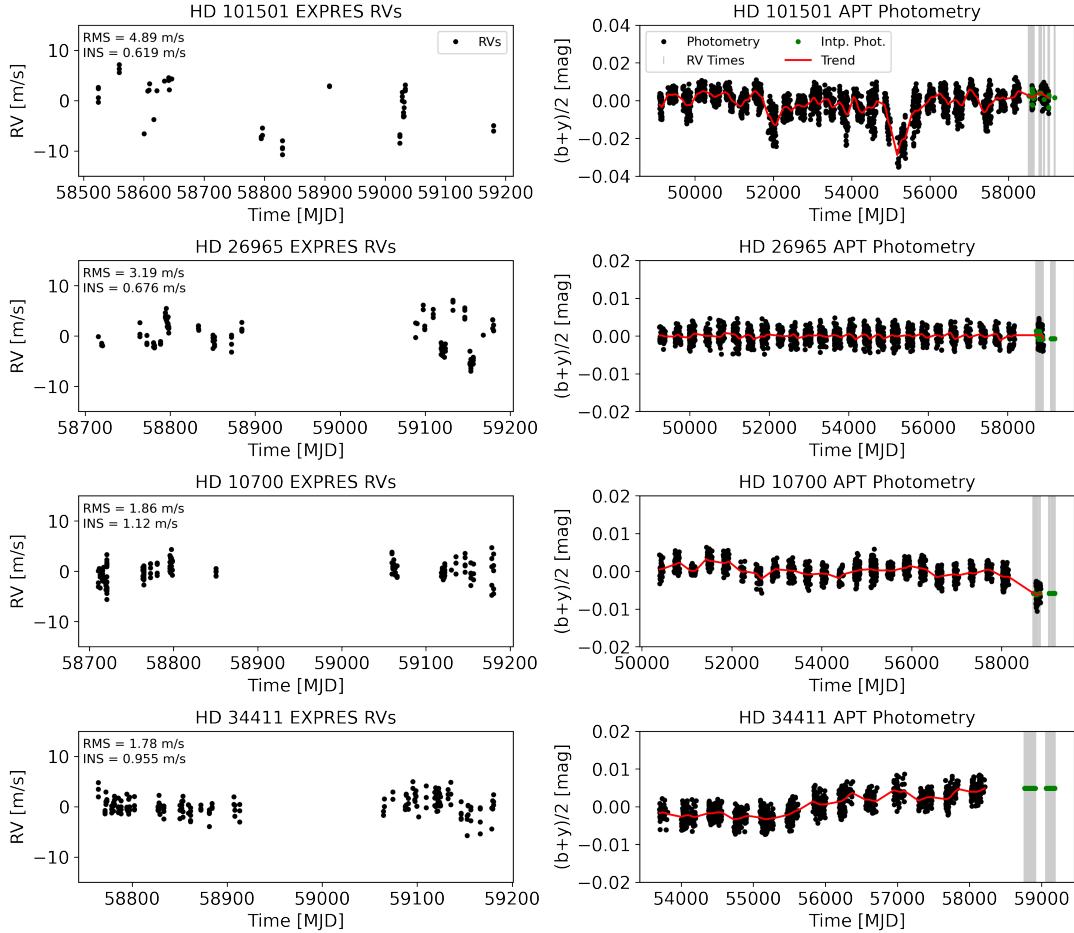


Figure 3. (Left column) EXPRES radial velocities obtained in 2019 and 2020 and (right column) several years of APT ground-based differential photometry obtained for the four ESSP stars: HD 101501, HD 26965, HD 10700, and HD 34411. Both RVs and photometry time series are relative to the median value subtracted. The photometric observations were smoothed in 100-day windows to capture long-term trends; the red line shows the smoothing model. The time intervals of the EXPRES RVs are marked in the right column of photometry points by gray, vertical lines. The GP interpolation/extrapolation of the photometric data to the RV timestamps are over plotted as green points.

vidual method can be found in the Appendix and the provided appropriate references.

Methods are grouped into subsections according to the type of input data used: the RVs with global indicators (§3.1), the CCFs (§3.2), or the extracted (pixel-level) spectra (§3.3 and 3.4).

3.1. Methods That Use RVs and Classic Activity Indicators as Input

Activity indicators aim to gauge the magnetic activity present on the target star, presence of activity features, or the expected amplitude of stellar signals. These indicators are global parameters; one value is determined for each spectrum. One can fit a model relating the activity indicators and apparent RVs in an attempt to remove or mitigate the effect of stellar signals on measured RVs. Classically, this was done using a simple linear fit.

We present the results of a classic linear decorrelation with the provided activity indicators to serve as a baseline result. RVs are plotted against the different indicators independently and fit to a line as a function of indicator value. The fitted line evaluated at the value of the different indicators is then subtracted from the RV measurements. There is evidence of a phase shift between some indicator variation and corresponding RV variation (e.g., Santos et al. 2014; Collier Cameron et al. 2019; ?), adding scatter to a direct comparison between indicator and RV and limiting the efficacy of this linear decorrelation method.

More recent work has developed novel ways of linking indicators to RV measurements and modeling out the stellar component of the RV measurements (Rajpaul et al. 2015; Barragán et al. 2019; Gilbertson et al. 2020b; ?; ?; ?; Barragán et al. 2021). Indicator-dependent

Table 6. Teams and Methods

Team	Method	Input	Run Time	Reference/Contact
PennState	<i>GLOM</i>	RVs/Indicators	laptop, minutes	Gilbertson et al. (2020b)
Sidera	<i>FDPCA</i>	RVs/Indicators	laptop, seconds	Ramirez Delgado et al., in prep Dodson-Robinson et al., in prep
Porto	<i>GPRN</i>	RVs/Indicators	cluster, hour	Camacho et al., in prep.
Andrews/PennState	<i>SCALPELS</i>	CCFs	laptop, seconds	Collier Cameron et al. (2020)
PennState	<i>SCALPELS+GLOM</i>	CCFs	laptop, minutes	Gilbertson et al. (2020b)
OxBridGen	<i>CCF Prime</i>	CCF	desktop, minutes	Baptiste Klein
PennState	<i>FIESTA+GLOM</i>	CCFs	laptop, minutes	Jinglin Zhao
ML_EPRVs	CCF Linear Regression (LR)	CCFs	laptop, seconds	de Beurs et al. (2020)
ML_EPRVs	CCF LR + H α	CCFs/Indicators	laptop, seconds	de Beurs et al. (2020)
ML_EPRVs	CCF LR + Keplerian	CCFs	laptop, seconds	de Beurs et al. (2020)
PennState	CCF Mask-VALD	Spectra	laptop, minutes	Alex Wise
Warwick	CCF Mask-BIS	Spectra	laptop, minutes	Lafarga et al. (2020)
Warwick	CCF Mask-RV	Spectra	laptop, minutes	Lafarga et al. (2020)
Geneva	<i>LBL+PCA_{Spec.}</i>	Spectra	laptop, hours	Dumusque (2018)
Geneva	<i>LBL+PCA_{RV}</i>	Spectra	laptop, hours	Cretignier et al., submitted Cretignier et al. (2021)
Geneva	<i>LBL+PCA_{Spec./RV}</i>	Spectra	laptop, hours	Cretignier et al., submitted Cretignier et al. (2021)
OxBridGen	<i>PWGP</i>	Spectra	desktop, day	Rajpaul et al. (2020)
PennState	<i>DCPCA</i>	Spectra	laptop, seconds	Jones et al. (2017)
PennState	<i>DCPCA+GLOM</i>	Spectra	laptop, minutes	Gilbertson et al. (2020b) Holzer et al. (2021)
CCA	ResRegGen Self	Spectra	laptop, hour	Lily Zhao
CCA	ResRegGen	Spectra	laptop, hour	Lily Zhao
CCA	ResRegDis	Spectra	laptop, minutes	Lily Zhao

methods will only be sensitive to signals that are reflected in the provided indicators; for example, if the used indicators do not track the effects of oscillation or granulation, then these methods will not return models sensitive to these effects. Teams who used RVs and indicators as input were asked to use the provided forward-modeled CBC RVs and the given indicators.

A Gaussian Process Linear Ordinary Differential Equation (ODE) Maker (*GLOM*) developed by the PennState team, is a Julia package that uses a shared, latent Gaussian Process (GP) to model both RV and indicator time series concurrently. This makes use of the flexibility of a GP model while also constraining the model with indicator time series to capture only stellar signal related variations. *GLOM* can be thought of as a generalization of the multi-dimensional GP method implemented in *pyaneti* (Rajpaul et al. 2015; Barragán et al. 2019, 2021). This method requires dense sampling throughout the characteristic time scale of the signal being modeled (e.g., the stellar rotation period). It is utilized for many of the submitted methods here that generate an indicator for the presence of stellar signals. More information can be found in Gilbertson et al. (2020b) or Appendix A.1.

Fourier Domain Principal Component Analysis (*FDPCA*), developed by the Sidera team, Fourier transforms RV and indicator time series using nonuniform methods to identify coherent oscillations between multiple series regardless of their relative phases. The Fourier transformed series are decomposed using principal component analysis (PCA) to derive orthogonal axes of variation in the activity indicators and their associated weights. The results presented here were trained on the RV, H α emission, and CCF FWHM time series. The model incorporated increasing numbers of principal components until 95% of the total variation was captured. This method requires observations to cover the entire phase range of the signal being modeled. Observations should be dense in phase space, not just time. A more in-depth description can be found in Appendix A.2.

The Gaussian Process Regression Network (*GPRN*) method, developed by the Porto team, models RVs and indicators through a neural net framework where each node is an independent GP model and the weights of each node are also determined by a GP model. While each node and weight can be represented by an independent GP, hyper-parameters and priors may be shared between models to reduce the number of free parameters. For the results presented here, one node was defined

by a GP with a quasi-periodic kernel while GPs with squared-exponential kernels were used for the weights with no shared hyper-parameters. The models were trained on the RV and CCF FWHM time series. The *GPRN* method is still being developed; preliminary results are included here. A more in-depth description can be found in Appendix A.3.

3.2. Methods That Use the Cross Correlation Function (CCF) as Input

The CCF has long been used in endeavors to mitigate the effects of stellar signals. CCFs are computed by cross correlating a given spectra with a mask tuned to where spectral line centers are expected to appear. The mask can either be binary (i.e., 1 where there is a line, 0 where not) or incorporate different line widths and window functions.

As this mask is shifted relative to a stellar spectrum, the convolution of the two will give larger or smaller values depending on how well the mask lines up with the spectral absorption lines. A perfect alignment of the mask with the bottom of every spectral line will result in the lowest cross-correlation value. The shift that results in the lowest CCF point is taken to be the RV shift of the spectrum.

In shifting, the CCF will sample the shape of all the spectral lines included in the mask, including the wings of these lines. Lines can be weighted differently according to their depth or their SNR. The CCF therefore provides a sort of weighted average of all the line shapes in the spectrum. This makes the CCF a powerful tool for capturing line shape distortions. On the one hand, averaging over all lines in a mask strengthens the signal of any line shape changes that are common to many lines; on the other, this averaging may blur out the different changes seen in individual lines.

Four methods used the CCF as input. They differ in their approach to modeling shape deformations within the CCF and how to separate these from translational shifts that can be attributed to true center-of-mass motion of the target star from orbiting planets.

The Self-Correlation Analysis of Line Profiles for Extraction of Low-Amplitude Shifts (*SCALPELS*) method, submitted by the Andrews and PennState teams, uses PCA to model the variations in a CCF's auto-correlation function. Because the auto-correlation function is intrinsically insensitive to translational differences, *SCALPELS* is not sensitive to true shifts in the CCFs. The measured RV time series can then be projected onto the identified principal components to derive and subtract out the shape-driven component of the measured RV while preserving the shift-driven com-

ponent. The results presented here use only the first two principal components to guard against incorporating noise into the model. *SCALPELS* operates in the velocity-domain and as such does not require any additional information about the star or dense time sampling. Using PCA means the model benefits from wider ranges of stellar activity states producing a large range of variation within the CCFs.

SCALPELS uses PCA in a similar way to *FDPCA*, where the principal components are used as a new basis with which to construct denoised measured RV shifts due to stellar signals. *SCALPELS* uses PCA on the auto-correlation function of the CCFs while the *FDPCA* method implements PCA on the Fourier series of the RV and indicator time series. Note that while there is a description of a leave-one-out-framework with *SCALPELS* in Collier Cameron et al. (2020), no cross-validation framework is implemented for the results submitted here.

The *SCALPELS+GLOM* method is another use of PCA. The amplitudes of the first two principal components for each observation, which describe the magnitude of the two largest axes of variation, are treated as activity indicators and input into *GLOM* to be co-modeled with the RV measurements. For the results presented here, the latent GP model used the sum of two Matérn $\frac{5}{2}$ kernels. This introduction of a GP model introduces relevant data requirements, such as dense-sampling in time, to the method. More information about both implementations of *SCALPELS* can be found in Collier Cameron et al. (2020) as well as Appendix B.1.

The *CCF Prime* method, submitted by the OxBridGen team, uses higher order derivatives of a GP modeled reference CCF (here a mean combined CCF) to fit shape changes. While the first GP derivative is sensitive to translational differences, the second derivative and above instead model shape changes. These higher-derivatives are used to recreate the shape-driven component of the measured RVs, which can then be subtracted out. The *CCF Prime* method is still being developed; preliminary results are included here. A more in-depth description of the *CCF Prime* method can be found in Appendix B.2.

The FourIEr Phase SpecTrum Analysis (*FIESTA*) method, submitted by the PennState team, isolates line shape changes using a Fourier basis with respect to velocity. Horizontal, translational differences will manifest as a constant shift at all frequencies in this basis. Shape-driven shifts can therefore be isolated as frequency-dependent shifts. The results presented here run a PCA on the derived shifts for each frequency and uses the amplitudes from this PCA as input into *GLOM*. This is sim-

ilar to how PCA is used within the *SCALPELS+GLOM* framework, which is distinct from the use of PCA in the *SCALPELS* or *FDPCA* methods. *FIESTA* requires careful normalization of the CCFs for each observation, as vertical translational differences could be mistaken for a shape change. More information can be found in Zhao & Tinney (2020) and Appendix B.3.

The *SCALPELS*, *CCF Prime*, and *FIESTA* methods all implement a change of basis to separate out the shape- and shift-driven components of the measured RVs. While these methods are conceptually similar, they make different assumptions on the appropriate basis and dimensionality of the variations being modeled. High SNR for observations are more necessary with *CCF Prime* (for more accurate GP derivatives) and *FIESTA* (to allow for incorporating higher frequencies) than *SCALPELS*, but the use of PCA by *SCALPELS* is more dependent upon the assumption that the shape-driven changes are the dominant source of variation.

The CCF Linear Regression method, submitted by the ML_EPRVs team, uses machine learning to model variations in the residuals of each CCF as compared to a reference CCF, constructed by median-combining all CCFs. Differential CCFs are normalized (in terms of amplitude and overall variance) and then sampled at a small-number of locations across the velocity range of the CCFs. ~~More observations allow for more sampling locations. For the results presented here, the CCFs were sampled at four to six locations. For each observation, an associated weight was fit for these different CCF locations.~~ This method does not use timing information, and so does not care about the sampling of observations, but does benefit from more observations.

For all four targets, a slightly more complicated CCF Linear Regression model was also implemented, wherein the H α emission value for each observation was included in the model with its own fitted weight parameter to help predict variations due to stellar signals. For HD 26965, which hosts a proposed planet, a third model that incorporates a weighted Keplerian was also implemented. ~~More information can be found in de Beurs et al. (2020) and Appendix B.4.~~

All methods attempting to model line shape changes, such as the four described here, will be helped by data with high resolution. Higher resolution spectra contain more information about the shape of each spectral line and will therefore more faithfully capture the shape deformations being modeled. This is true whether the shape changes are being modeled as averaged in the CCF or in the spectra itself.

3.3. Line-by-Line Methods

The remaining methods take the full, extracted spectra as input. Several methods, described in this section, use the spectra to determine preferred lines or regions of spectra to use in deriving RV measurements. Methods that model variation throughout the complete spectra are described in the following section (§3.4).

Three methods focused on carefully picking which lines to include when constructing a CCF. It has recently become clear that spectral lines will respond in different ways to stellar variation both in terms of behavior and magnitude of response (Davis et al. 2017; Thompson et al. 2017; Wise et al. 2018; Dumusque 2018; Cretignier et al. 2020a). Isolating lines that are less swayed by stellar signals or other occluding effects will help in calculating CCFs and RVs resilient to other variations and ultimately more representative of true, center-of-mass shifts in the spectra.

The CCF Mask-*VALD* method, submitted by the PennState team, used line center information from the Vienna Atomic Line Database (*VALD*) to vet line blends and tested a range of CCF mask window widths, where a truncated Gaussian window function was used for all lines. The optimal cutoffs for distance between line centers and mask window width were found empirically by minimizing the RMS of the resultant CCF RVs. More details can be found in Appendix C.1.

The CCF Mask-BIS and CCF Mask-RV methods, both submitted by the Warwick team, use correlations with the BIS activity indicator or the provided CCF RVs to vet lines. RVs for individual lines were found by fitting each line to a Gaussian and using the mean of this fit as the line center. Lines were excluded if their RVs were found to scatter greater than 10–15 m s⁻¹ or their RVs were found to be strongly correlated with the BIS or CCF RV (i.e., the Pearson correlation coefficient is greater than some cutoff, where the cutoff depends on the specific indicator used and target). The RVs of the remaining lines are averaged to compute a combined RV for each observation. More details can be found in Lafarga et al. (2020) and Appendix C.2.

Note that CCF Mask-RV is not the only method to use the RV as an activity indicator (see, for example, the discussion of the *ResRegGen* and *ResRegDis* methods below). This use case assumes that all variation in the measured RVs is dominated by stellar signals. We know that instrument systematics are not the dominant source of error in these data sets, as seen from EXPRES data of quiet stars (see Figure 2). While there are no obvious planetary signals, this does not preclude planet signals on the order of or smaller than the stellar signal amplitude adding variation to the RVs.

All three of the above methods fit lines to a Gaussian profile to determine line parameters—such as line center, width, etc.—or change in line parameters. The provided *SELENITE* telluric model was also used in all three cases to remove lines within $\sim 30 \text{ km s}^{-1}$ of a telluric feature.

The Geneva team also implemented a line-by-line (*LBL*) RV analysis. The *LBL* RVs were derived relative to a master spectrum using post-processed spectra (Dumusque et al. 2011b). The provided EXPRES spectra were (1) merged (i.e., all echelle orders were combined), (2) continuum normalized using *RASSINE* (Cretignier et al. 2020b), and then (3) further cleaned of tellurics and first-order morphological variations using *YARARA* (Cretignier et al. 2021). Lines returning a poor fit to the master spectrum or exhibiting larger scatter than expected from the median RV error were not included in the final combined RV calculation.

PCA was used to de-noise the results at either the spectral level, denoted here as *LBL+PCA_{Spec.}*, or produce a metric of variation at the line-by-line RV level, *LBL+PCA_{RV}*. At the spectral level, the first three components of a weighted PCA are used to recreate a denoised representation of the spectra. These de-noised spectra are then used to construct a master spectrum and derive *LBL* RVs.

PCA was also run on the *LBL* RVs themselves to identify variations across all lines and across all observations. Rather than denoising, here PCA is instead used to determine the magnitude of variation that is then treated like an activity indicator against which the combined *LBL* RVs are decorrelated with a multi-linear regression. *LBL* RVs derived and decorrelated using RV-level PCA are described as the *LBL+PCA_{RV}* method.

Both methods can also be combined, which are here represented by the *LBL+PCA_{Spec./RV}* method. Though both *LBL+PCA_{Spec.}* and *LBL+PCA_{RV}* use PCA, it is important to note that PCA is used on different data products for the two methods and to different ends. The difference is similar to the difference between how PCA is utilized in the *SCALPELS* method versus the *SCALPELS+GLOM* method. More details about deriving *LBL* RVs and the *RASSINE* and *YARARA* methods can be found in Dumusque (2018); Cretignier et al. (2020b, 2021). More information about the *LBL+PCA_{Spec.}*, *LBL+PCA_{RV}*, and *LBL+PCA_{Spec./RV}* implementations represented in this report can be found in Appendix C.3.

The Pairwise Gaussian Process RV Extraction (*PWGP*) method, submitted by the OxBridGen team, breaks the spectrum into chunks and uses GPs to model and align pairs of chunks. The behavior of each chunk across all observations is used to determine which ar-

eas of the spectrum are more or less sensitive to variation from telluric contamination or stellar signals. In the limit where each chunk contains one line, which the implementation presented here approaches, the *PWGP* method can be thought of as an approximate line-by-line approach. A Matérn $\frac{5}{2}$ kernel is used to model and align each chunk. Chunks exhibiting unusually high scatter or strong correlation with activity indicators are excluded. The RV for each observation is then calculated as a weighted average of the remaining chunks, where the RV error for each chunk is determined via a MCMC analysis. More information can be found in Rajpaul et al. (2020) and Appendix C.4.

For all these methods, there exists a trade off. Increasing the selectivity of lines or chunks to include will better mitigate the effects of stellar signals and other possible causes of line shape variation. Using less data, however, will increase the photon noise. These methods would all benefit from high SNR observations, which decreases the photon noise that must be contended with. This allows for confident RV estimates from relatively few, very stable lines.

3.4. Full-Spectrum Methods

While the methods described in the previous section treated each line/chunk as independent, the below methods model the entire spectra as a whole. Of course, in some ways the methods of the previous section do take into account information across the whole spectra, for example when setting cutoffs using all lines or running PCA on all lines. Unlike previously presented methods, though, these “full-spectrum” methods generally operate on all spectral pixels rather than isolating spectral lines of interest.

The Doppler-Constrained Principal Component Analysis (*DCPCA*) method, submitted by the PennState team, runs PCA on spectra shifted by the maximum-likelihood RV and uses the resultant PCA amplitudes, a measure of the amount of primary variation present in each exposure, as activity indicators. Though the PCA is run on the spectra, this use case of PCA is more similar to the *LBL+PCA_{RV}* method (or *SCALPELS+GLOM*), where the amplitude of the variation, not the axes of variation (i.e., the principal components), are the result of interest. To cut down on the noise that gets input into the PCA, only the spectral regions surrounding lines included in the default ESPRESSO masks used are fed into the PCA. These indicators are then either linearly decorrelated against RVs (*DCPCA*) or co-modeled with RVs using *GLOM* (*DCPCA+GLOM*). More information can be found in Jones et al. (2021) or Appendix D.1.

The *ResRegGen* and *ResRegDis* methods, both submitted by the CCA team, use the residuals of observed spectra from a template spectrum to regress against different housekeeping data—such as activity indicators—and derive the stellar photosphere contribution to the measured RV shifts. *ResRegGen* uses a generative framework for which H α equivalent width and CBC RVs are used as labels to derive the activity-component of the measured RVs. *ResRegDis*, on the other hand, uses a discriminative framework where the full residuals of each observation are used to inform the appropriate correction to the measured RVs. The discriminative framework is slightly more agnostic to the labels used, meaning *ResRegDis* is less tied to the information available in the activity indicators included in the model as *ResRegGen*. Both methods use a linear, first-order model and residuals to a reference template constructed using *wobble* (Bedell et al. 2019).

Both *ResRegGen* and *ResRegDis* implement a “cross-validation” (CV) framework. This guards against overfitting as the model is constructed without information from the subset of data that the model is then evaluated at. For *ResRegDis*, each observation is left out one at a time to construct an independent model. For *ResRegGen*, one eighth of the data is left out at a time to speed up the computation time. For reference, the “self” test variant for *ResRegGen* (*ResRegGen Self*) is included, wherein all data are used to construct the model. The only difference, then, between *ResRegGen* and *ResRegGen Self* is removing the cross-validation framework that is used to prevent over-fitting. Because seven eights of the data are still used to construct the model for the cross-validation version of *ResRegGen* and the validation data is chosen at random across the time baseline, we do not expect the *ResRegGen* method to be less informed than the *ResRegGen Self* method that uses all data points; the cross-validation step only ensures the resultant model is general. The *ResRegGen Self* method is presented as a more direct comparison to the RMS metrics reported by other methods, none of which employed cross-validation in the reported results. The *ResRegGen* and *ResRegDis* methods are still being developed; preliminary results are included here. More information about *ResRegGen* and *ResRegDis* can be found in Appendix D.2 and D.3 respectively.

4. RESULTS

For each method, teams submitted “clean RVs” representing the measured RV shift of the star cleaned of stellar signals and other modeled variations determined to be separate from true center-of-mass shifts. Where directly modeled, the RV signal due to these variations,

which we will refer to as “activity RVs,” was also submitted along with any indicators that the method derived.

We acknowledge that this chosen name of “activity RVs” is imperfect. The variations being traced by different methods may source from stellar activity features, such as spots, faculae, etc., but could also be due to inherent stellar variation, such as pulsations or granulation, or trace other sources of variation in the spectra from, for example, uncorrected tellurics, instrumental changes, etc.

For some methods, the clean and activity RVs represent different components of the model and so do not sum to the original RVs provided. The clean RVs and provided activity RVs from all methods are plotted in Appendix E along with their Lomb-Scargle periodograms (Lomb 1976; Scargle 1982).

4.1. *RV RMS of Method Results*

Table 7 gives the change in overall and nightly RMS for each method as compared to the RMS of the provided, uncorrected CBC RVs. The nightly RMS, or intra-night scatter (INS), represents the average scatter over all nights with more than one observation. Positive Δ RMS values indicate that the method returned a lower RMS than the original. Negative Δ RMS values mean there was more spread in the returned RVs than in the original provided RVs. Methods are ordered in the same order as described in the Methods section above (§3).

The final RMSs of the clean RVs for all methods are plotted in Figure 4. The height of each bar as well as its position along the x-axis scales with the overall RMS of the returned clean RVs. Each bar is mapped to its corresponding method across the x-axis, along which the methods are ordered by decreasing RMS from left to right.

Each bar is colored by the type of data the method takes in as input, corresponding to the break down of methods in Section 3. Note that here, all methods that use a sort of activity indicator, classic or newly derived (e.g., amplitudes from PCA, etc.), are grouped together regardless of the input data needed to derive the indicator used.

The baseline method of decorrelating RVs against classic activity indicators, shown in Figure 4 as brown bars, does not produce a significant decrease in RMS. The decrease is modestly significant for HD 101501, the most active of the targets given.

The relative returned RMS of each method differs across the four stars. Methods that return low RMS for one or some of the targets do not necessarily return low RMS for all of the targets. The relative RMS of different methods is most different for HD 26965, for which a few

methods (*GPRN*, *ResRegDis*, and *LBL+PCA_{Spec./RV}*) return much lower relative RMS values than for the other three targets. For HD 101501, *LBL+PCA_{Spec./RV}* and *GPRN* also return among the lowest RMS, as with HD 26965, but here CCF LR and CCF LR+H α return much lower relative RMS than with any other target. The relative orders of the methods are fairly consistent between HD 10700 and HD 34411. Recall that the four targets differ in expected activity level, total number of observations, sampling of observations, and number of proposed planet candidates.

For each of the four targets, there are one or more clusters of methods returning a similar RV RMS, which can be seen as overlapping bars in Figure 4. For HD 101501, there is a cluster of methods returning a final RMS of approximately 2.5 m s^{-1} , i.e., a 48% decrease in RMS. The HD 26965 results exhibits a cluster at 2.7 m s^{-1} (16% decrease) and 2.3 m s^{-1} (28% decrease). Note these are slightly greater than the 1.8 m s^{-1} semi-amplitude of the proposed planet (Ma et al. 2018). The HD 10700 (τ Ceti) results cluster around 1.6 m s^{-1} (13% decrease). The HD 34411 results cluster at 1.5 m s^{-1} (15% decrease) and 1.4 m s^{-1} (22% decrease). The methods that are returning similar RMS values and forming these clusters differ in their approach to disentangling stellar signals, and in fact the methods that are clustered together even differ from target to target.

Table 7. RMS and INS of Cleaned RVs from each Method in Units of m s^{-1}

Method	HD 101501			HD 26965			HD 10700			HD 34411		
	INS	RMS_{all}	INS	RMS_{all}	INS	RMS_{all}	INS	RMS_{all}	INS	RMS_{all}	INS	RMS_{all}
Original EXPRES CBC RVs	0.62	4.887	0.65	3.195	1.071	1.864	0.944	1.78				
S-Value	0.02	0.26	0.021	0.526	0.195	0.186	-0.005	0.072				
H α Emission	-0.317	0.564	-0.051	0.209	0.005	0.003	0.0	0.0				
H α Equiv. Wid	0.027	0.031	-0.001	0.001	-0.001	0.072	-0.005	0.049				
CCF BIS	0.09	1.118	-0.011	0.048	-0.006	0.005	-0.024	0.058				
CCF FWHM	-0.005	0.534	0.001	0.022	0.002	0.001	-0.008	0.118				
V _{span}	-0.076	0.567	-0.007	0.009	0.016	0.018	-0.007	0.006				
Bi-Gaussian Fit	0.082	1.498	-0.005	0.015	-0.006	0.038	-0.008	0.154				
Skew Normal Fit	-0.483	0.206	-0.025	0.014	0.001	0.006	0.0	0.0				
FDPCA	-0.001	2.418	0.0	0.775	0.0	-0.017	0.0	0.255				
GPRN					0.0	0.054	0.0	0.362				
SCALPELS	-0.42	2.079	-0.473	0.859	0.122	0.458	0.245	0.547				
SCALPELS+GLOM	-0.206	2.31	-0.202	1.217	0.143	0.496	0.271	0.571				
CCF Prime	0.182	1.76	0.0	0.222	-0.01	0.032	0.124	0.18				
FIESTA+GLOM	0.234	2.355	0.1	0.713	0.13	0.24	0.129	0.088				
CCF Linear Regression	0.001	2.607	-0.176	0.56	0.196	0.297	0.065	0.196				
CCF LR + H α	-0.13	2.863	-0.193	0.679	0.168	0.352	0.065	0.196				
CCF LR + Keplerian			-0.015	0.484								
CCF Mask-VALID	0.202	1.336	-0.01	-0.001	-0.142	0.021	-0.002	-0.029				
CCF Mask-BIS	-0.231	1.421	-0.289	0.505	-0.292	-0.134						
CCF Mask-RV	-0.232	2.292	-0.605	0.905	-0.285	-0.136						
LBL+PCA _{Spec.}	-0.182	2.36	-0.226	0.85	-0.027	-0.098	0.088	0.261				
LBL+PCA _{RV}	-0.421	2.46	-0.354	0.51	0.122	0.286	0.101	0.33				
LBL+PCA _{Spec./RV}	-0.238	3.159	-0.032	1.549	0.066	0.205	0.115	0.394				
PWGP	-0.022	2.132	0.179	0.857	0.232	0.374	0.251	0.376				
DCPCA	0.012	1.942	0.109	0.998	0.146	0.235	0.144	0.145				
DCPCA+GLOM	0.027	2.368	0.108	1.125	0.147	0.241	0.144	0.111				
ResRegGen Self	0.123	2.86	0.117	2.01	0.629	1.041	0.496	0.644				
ResRegGen	-0.165	0.374	-0.172	0.251	-0.1	0.076	-0.041	0.053				
ResRegDis	-0.104	1.957	-0.263	1.785	-0.21	0.123	-0.243	0.271				

Note—All RMS values given in units of m s^{-1} .

The self test version of *ResRegGen*, *ResRegGen Self*, always returns a lower RMS than the cross-validation implementation of *ResRegGen*. Furthermore, *ResRegGen Self* is often among the methods returning the lowest RMS value. *ResRegGen* and *ResRegGen Self* only differ in whether there is a safe guard built into the method against over-fitting the model. The difference in their resultant RMS therefore highlights the difference between an appropriately general model with *ResRegGen* and a likely over-fitted model with *ResRegGen Self*. We note that while the *ResRegGen* class of methods has many free parameters and is therefore particularly vulnerable to over-fitting, some degree of over-fitting may be in play for other methods presented in this paper. Many methods trained model components on the same data that they reported results for with no data held out, as was done in *ResRegGen Self*. For future challenges, we would prefer to hold out a validation set of data for the final comparison of methods to ensure that any over-fitting is diagnosed uniformly across methods.

Similarly, the use of *GLOM* to co-model RVs and indicators almost always results in a lower RV RMS than the alternative (i.e., *SCALPELS+GLOM* as compared to *SCALPELS* results or *DCPCA+GLOM* as compared to *DCPCA* results). In some cases, the use of *GLOM* across methods returns RVs with a similar RMS (see *SCALPELS+GLOM*, *FIESTA+GLOM*, and *DCPCA+GLOM* for HD 101501 and likewise *FIESTA+GLOM* and *DCPCA+GLOM* for HD 10700). This suggests that *GLOM* is a powerful tool for modeling out variation in a time series regardless of the indicator(s) it is given.

Methods that operate along very similar principles often return very different RV RMS results. For instance, the three methods that model the CCF with a change of basis (i.e., *SCALPELS*, *CCF Prime*, and *FIESTA*) give resultant RVs RMS values with a spread of 0.15 to 0.45 m s⁻¹ depending on the target. The same is true of the different line-by-line methods (shown as light blue bars in Figure 4), whose results have a spread of 0.15 to 0.60 m s⁻¹ in resultant RV. In both cases, the HD 34411 results have the lowest spread while the HD 101501 results have the highest. On the other hand, all methods that use *GLOM* (i.e., *SCALPELS+GLOM*, *FIESTA+GLOM*, and *DCPCA+GLOM*) return similar resultant RMS values that have a spread of only 0.02 to 0.22 m s⁻¹ across the four targets.

We see here that the different methods do have a notable impact on the resultant RMS of the clean RVs, yet it is impossible to say from this one-dimensional metric what exactly is being modeled out by each method. Just

because a method is returning a lower RMS does not necessarily mean it is doing better at mitigating stellar signals specifically; this cannot be inferred from the RMS alone.

4.2. Comparing Methods

Through the *ESSP*, all teams were given the same set of EXPRES data to use with their respective methods and model out stellar signals. When working with real data, we do not know what stellar signals are present for each target. Because the data are consistent among all methods, we would expect the stellar signal being removed to be consistent between methods successfully modeling photospheric velocities. Hence, the activity RVs for each method should be correlated with one another.

We perform a pairwise comparison of the activity RVs returned by each method. For methods that do not naturally derive the RVs due to stellar signals, we approximate these activity RVs as the RV shift removed—i.e., we take the difference between the provided, EXPRES CBC RVs and the submitted clean RVs to be the activity RVs. For each pair of activity RV time series, which we expect to have a one-to-one correspondence, we use the Pearson correlation coefficient (PCC) to gauge the strength of correlation between the activity RVs derived by two different methods.

Figure 5 shows markers for each pair of methods colored by the PCC between the activity RVs each pair of methods returned. The first column of each plot shows the PCC of each set of activity RVs against the provided EXPRES CBC RVs. A PCC of > 0.4 with an associated p-value of < 0.05 (square markers in Figure 5) is considered statistically significant, as established by the spread of PCCs returned from comparing series of randomly generated numbers of the same length as the RV data sets.

Comparisons between two methods that both submitted activity RVs are shown as filled in markers. Comparisons involving activity RVs that were recreated as the difference between the provided RVs and submitted clean RVs are not filled in. Only submitted methods are included; the results from classic linear decorrelation with standard activity indicators are not shown.

The top of each plot recreates a scaled bar graph of the final RMS of the clean RVs for each method. These insets are meant to help associate each method with their final returned clean RV RMS. Methods that returned similar final RMS values, as well as the relevant correlation markers, are highlighted in shades of green, mirroring the shading in Figure 4.

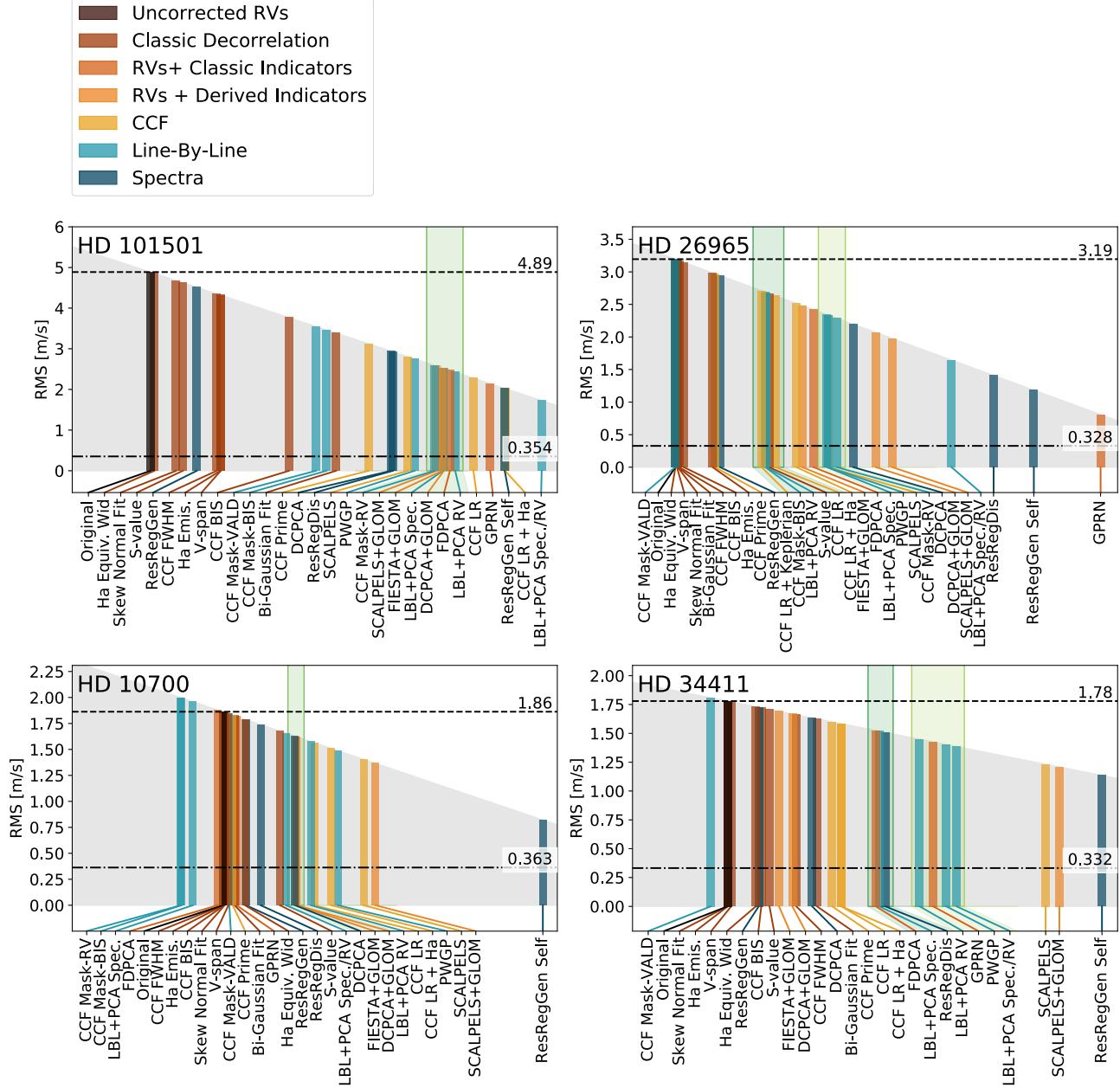


Figure 4. Final overall RMS of the clean RVs submitted for each of the four targets. The height and x-position of each bar scales with the final RMS. Bars are colored by the type of input data used. For each target, the RMS of the original, uncorrected EXPRES CBC RVs is shown as a black bar with its height emphasized by a horizontal dashed line across the full plot. The average intra-night scatter of the EXPRES CBC RVs is also marked with a horizontal dash-dotted line. Methods returning similar RMS values to each other that cluster together are emphasized via green shading.

As expected, most PCCs are positive, but there is limited strong (> 0.4) correlation. Even methods returning similar RMS values to one another (i.e., markers close to the diagonal) are often not deriving activity RVs that are significantly correlated with one another. The methods returning the most similar RVs (as highlighted via green shading) are correlated for HD 10700 and HD 34411, but

not for HD 26965. HD 101501, the most chromospherically active of the four stars, has the most correlation amongst the activity RVs returned.

Methods returning lower clean RV RMS (i.e., methods closer to the bottom or further to the right of each subplot) are more likely to have activity RVs significantly correlated with other methods'. These methods are even

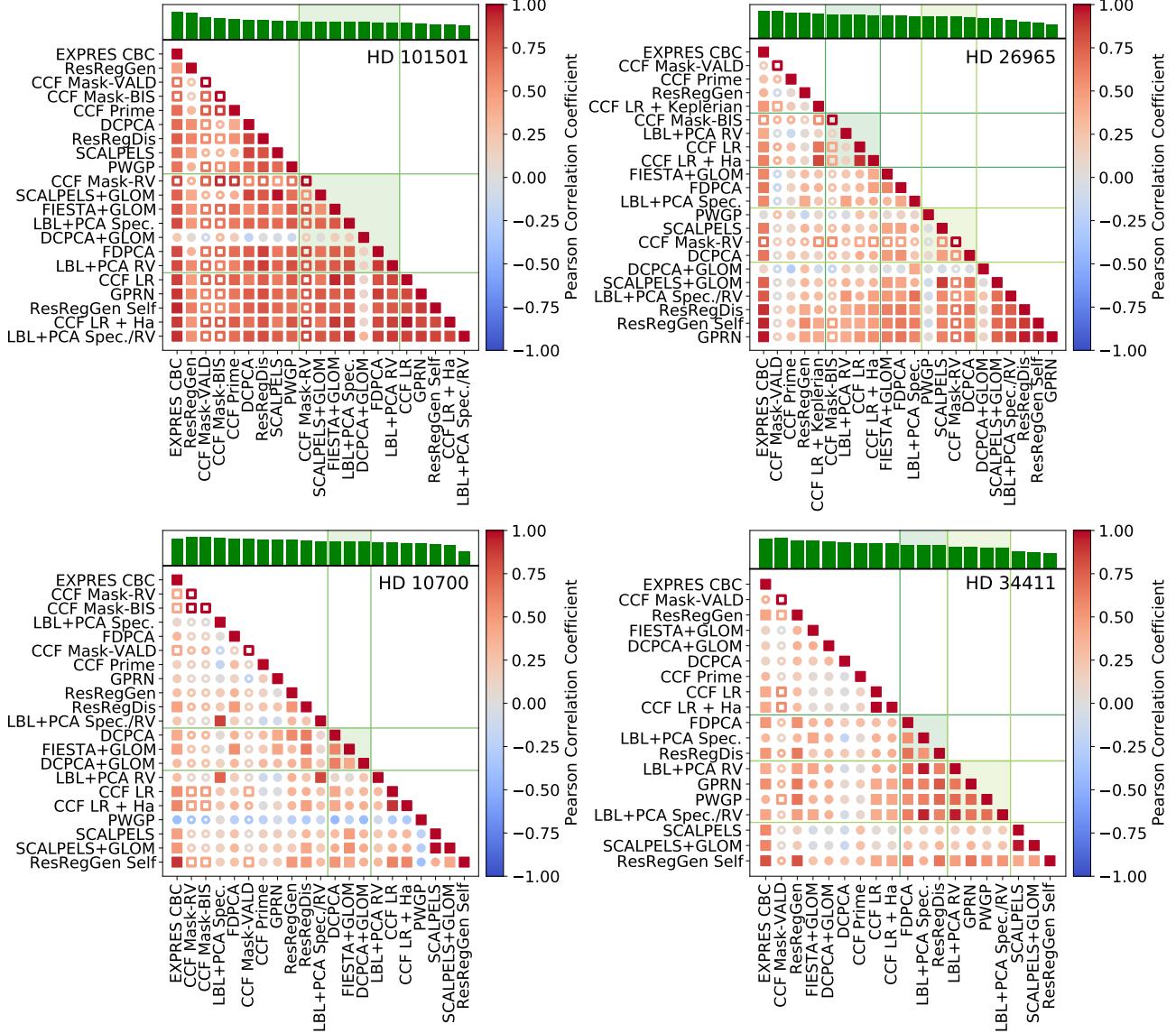


Figure 5. Pairwise comparisons between the activity RVs of all submitted methods. Each marker represents a pairing; the color of the marker gives the PCC between the activity RVs of two methods. Methods that did not submit activity RVs (for which the difference between the original and clean RVs was used instead) are shown as marker outlines. Significant PCC values (i.e., $PCC > 0.4$) are shown as squares while a $PCC < 0.4$ is depicted with a circle. The first column of each plot gives the PCCs with the provided CBC RVs. The following rows/columns have methods ordered from top to bottom and left to right by decreasing total RMS, the same as how methods are ordered in Figure 4. At the top of each subplot in green is a scaled bar-graph of the resultant RMS for each method. As mirrored from Figure 4, methods that returned similar final RMS values are shaded in green along with their associated correlation coefficients.

more likely to be significantly correlated with the provided EXPRES CBC RVs (see the first column of each plot). If the derived activity RVs of these lower-RMS methods are subsuming much of the signal in the provided EXPRES RVs, then we would expect to see them show greater correlation with the provided RVs and all other methods that use the provided RVs as a starting point.

Variations on a method are nearly always significantly correlated with one another. Namely, the activity RVs returned by SCALPELS and SCALPELS+GLOM are significantly correlated for all four targets. The same is true for the three CCF Linear Regression variations (i.e., CCF LR, CCF LR + $\text{H}\alpha$, and CCF LR + Keplerian) and the three residual-regression based methods (i.e., ResRegGen, ResRegGen Self, and ResRegDis).

Variations on line-by-line methods also agree with each other. The results of CCF Mask-BIS and CCF Mask-RV are always correlated as are the results of $LBL+PCA_{Spec.}$, $LBL+PCA_{RV}$, and $LBL+PCA_{Spec./RV}$. However, the activity RVs returned by these similar methods do not correlate strongly with each other. Correlation with the *PWGP* activity RVs is particularly troublesome in the case of HD 26965 and HD 10700, where they are not correlated with the activity RVs returned from any other method.

The results of the *DCPCA+GLOM* method are only correlated with the results of the *DCPCA* method for HD 10700 despite both methods being informed by the same indicator. The *DCPCA+GLOM* activity RVs are not correlated with the activity RVs of any other method for HD 101501 and HD 34411. They are at most correlated with three other methods for the other two targets.

4.3. Correlation with Indicators

While we do not know exactly what the stellar signal is in these real data sets, we do have access to activity indicators, both classic and those derived from submitted methods, that aim to capture the level of magnetic activity, or amplitude of stellar signal, for each exposure. Of course, we have no assurances that any single indicator is a perfect tracer of the presence or magnitude of stellar signals. All the same, it is instructive to investigate whether the returned activity RVs of the various methods are correlated with any indicator. While not necessary, a correlation may be sufficient to lend interpretability to method results.

It has been established that activity indicators should not be expected to share a strict linear relation with the activity RV as phase differences are liable to blur out any linear relation (Santos et al. 2014; Collier Cameron et al. 2019). In an attempt to allay this mismatch, we use the Spearman’s rank correlation coefficient (SCC) to determine the correlation between activity RVs and activity indicators. The SCC gives a measure of the commensurable monotonicity of two data sets and so does not require linearity (as the PCC does). However, both the PCC and SCC are only sensitive to mappings between indicators and RVs that are one-to-one (for instance, if the response of an indicator is out-of-phase with the RV response, the resultant PCC and SCC will likely both show low correlation strength).

Since activity indicators themselves are imperfect, as is our ability to define the exact relation between indicators and stellar signals, a low SCC between an indicator time series and activity RV time series is not unexpected. At the same time, the existence of a strong correlation between indicator and activity RVs adds interpretabil-

ity as it provides a link between the model results and stellar signals.

Figure 6 shows markers similar to those in the correlation matrices of Figure 5, but here denoting SCC values for indicator/method pairs. The activity RVs from each method are compared to classic activity indicators *S*-value, $H\alpha$ equivalent width, CCF BIS, and CCF FWHM. Method results are also compared with indicators generated as the result of one of the submitted methods. For *SCALPELS+GLOM*, *FIESTA*, and *DCPCA*, the first two PCA amplitudes (denoted PCA1 and PCA2) are included as indicators. For *CCF Prime*, the coefficient corresponding to the second derivative term of the linear model serves as an indicator (see Equation B5).

We want to establish whether the activity RVs from a method are significantly correlated with any of the activity indicators more so than the clean RVs returned by the method. To place the SCCs in context, the SCC for both the clean and activity RVs as compared with the indicator time series (SCC_C and SCC_A respectively) are shown connected by a line for each indicator/method pair. The marker associated with the SCC_C is on the left while the SCC_A marker is on the right of each pair. Each marker is vertically offset by their SCC value, meaning the slope of the connecting line scales with the change in correlation strength from clean to activity RVs.

Some of the submitted methods are specifically guided by a given indicator. This relationship is highlighted by purple boxes in Figure 6. For example, the *ResRegGen* results shown here use the $H\alpha$ equivalent width as a label, and so the $H\alpha$ equivalent width and *ResRegGen* pair of markers has a purple box around it. As the highlighted methods make use of the indicator in question, there is more reason to expect a strong correlation between the two, though it is still possible that the exact relation is not captured by an SCC value.

Methods returning activity RVs that exhibit significant correlation with indicator time series by and large do not exhibit a similar significant correlation between the method’s returned clean RVs and indicators. This provides context and lends confidence to the correlation between activity RVs and indicators in these cases. Clean RVs are rarely correlated with indicators, and only ever when the provided EXPRES RVs are themselves correlated with the indicator in question. The amplitude of the first principal component from *SCALPELS* (*SCALPELS* PCA1) is correlated with the provided EXPRES RVs for all four targets.

As before, the activity RVs returned for HD 101501 exhibit the most consistent correlation with nearly all indicators. The provided, EXPRES CBC RVs for HD 101501

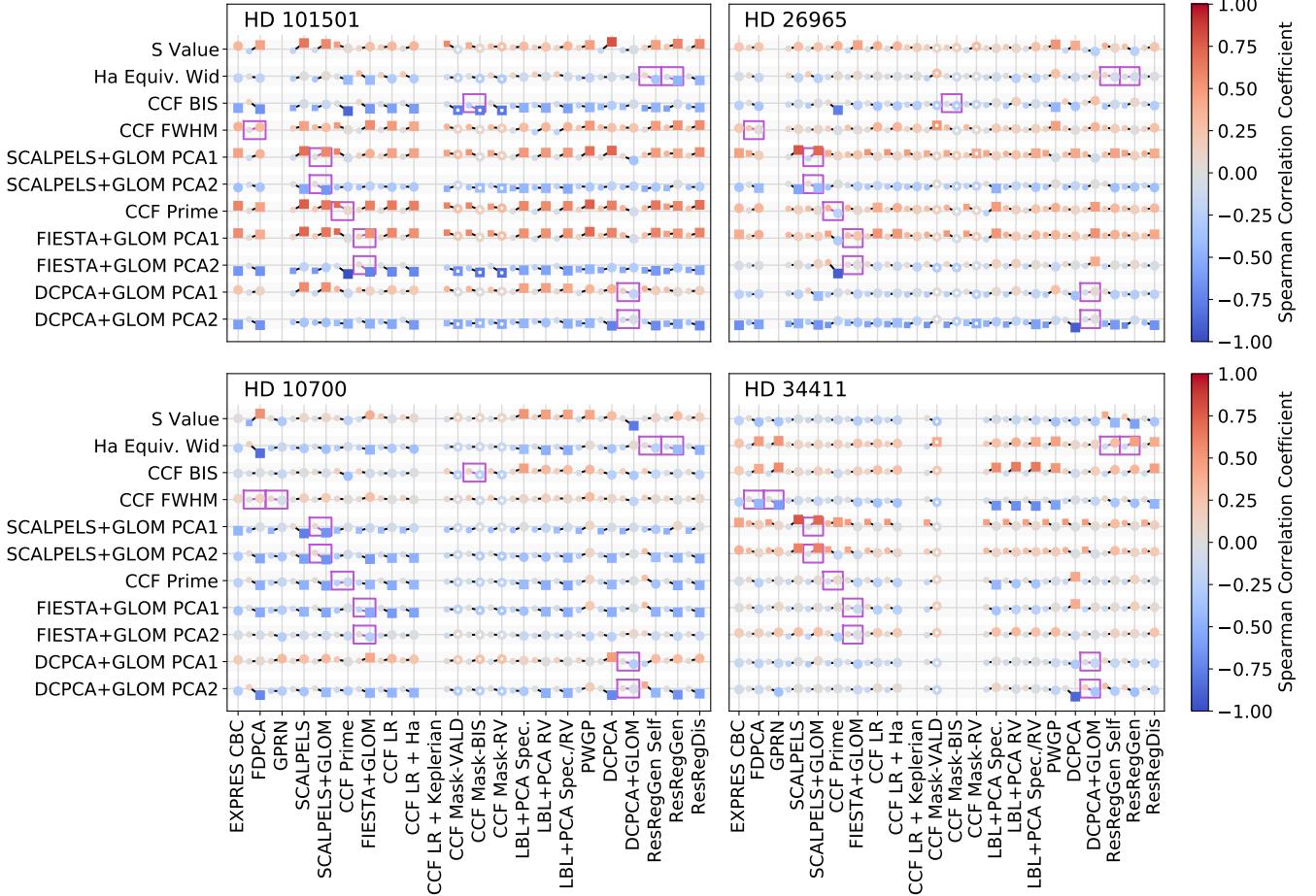


Figure 6. Correlation strength between activity RVs of submitted methods and activity indicators both classic (top four rows) and derived (bottom seven rows). Similar to Figure 5, here the color of each marker denotes the Spearman’s rank correlation coefficient (SCC) with strong values (i.e., $|SCC| > 0.4$) shown as square markers and constructed activity RVs (i.e., provided RVs - clean RVs) shown as marker outlines. The first column in each plot shows the SCC for each indicator time series with the provided EXPRES CBC RVs. Markers are also shifted vertically by their SCC. For each indicator/method pair, two markers are given connected by a line; the smaller, left marker in each pair gives the SCC with the clean RVs (SCC_C) while the larger, right marker gives the SCC with the activity RVs (SCC_A). The vertical offset and difference in color between the two markers alludes to the change in correlation strength. Purple boxes are drawn around indicator/method pairs where the method in question is directly informed by the given indicator.

are themselves correlated with most indicators. Many of the clean RVs for HD 101501 remain correlated with indicators. In fact, for certain methods, the clean RVs exhibit significant SCC values with indicators while the activity RVs do not.

The activity RVs returned by the *SCALPELS+GLOM* method are significantly correlated with the PCA amplitudes being modeled by *GLOM* for all four targets. Recall that *SCALPELS+GLOM* was also strongly correlated with *SCALPELS* results from Figure 5. Neither is true for the *DCPCA+GLOM* results. The results of *ResRegGen* are correlated with H α equivalent width for all targets except HD 26965 while the results of *ResRegGen Self*, which does not include a safe-guard

against overfitting, are not correlated with H α equivalent width for any target.

The amount of correlation with classic vs. derived indicators changes from target to target. For HD 34411, some methods return activity RVs correlated with classic indicators while the derived indicators result in very few significant correlation. For HD 10700, on the other hand, the derived indicators tend to show more significant correlations than do classic indicators. The *S*-Value shows the most correlation with activity RVs for HD 10700. CCF BIS and CCF FWHM exhibit more significant correlations for HD 101501 and HD 34411 than for the other two targets.

Some classic indicators are inversely correlated with activity RVs. For the CCF BIS, all significant SCC val-

ues for HD 101501 and HD 26965 are negative while all significant SCC values for HD 10700 and HD 34411 are positive. On the flip side, all significant SCC values between CCF FWHM and activity RVs are positive for HD 101501 and HD 26965 but all negative for HD 34411². There are some indicators, both classic and submitted, that change from positively to negatively correlated depending on that activity RVs, though no indicators have both significant positive SCCs and significant negative SCCs.

4.4. HD 26965 Results

One of the hopes of including the HD 26965 data as part of the *ESSP* was to gain a deeper understanding of the ~ 40 day, periodic signal. This period had previously been associated with both the stellar rotation rate of the star and with a potential orbiting planet with a RV semi-amplitude of 1.8 m s^{-1} (Díaz et al. 2018; Ma et al. 2018; Rosenthal et al. 2021).

For each of the submitted methods, we compare the periodogram of the clean RVs and the activity RVs, as shown in Figure 7 in blue and orange respectively. We also include periodograms of the provided EXPRES RVs and all RVs from the California Legacy Survey (CLS) (Rosenthal et al. 2021) in the top row for reference. We focused on the power for periodicities between 39 and 44.5 days, the proposed stellar rotation rates for HD 26965, which bookend the proposed 42.38 day planet period (Ma et al. 2018). The maximum power in this period range along with the corresponding p-value is given in the top-left corner of each subplot. Note that the EXPRES data peaks at 42.67 days, close to the proposed planet period. The CLS data, on the other hand, peaks at 41.52 days, slightly lower than the proposed planet period, and features a much higher peak at 52.13 days.

Methods with more power (within the highlighted period range) in the clean RV periodogram are shown in blue subplots while methods with more power in the activity RV periodogram is shown in orange. Four methods either have no significant peaks for those periods or return similar power in both the clean and activity periodograms.

Six out of twenty methods attribute the ~ 40 day period to stellar signals while eleven of the methods produced clean RVs that still contain the ~ 40 day period. Five of the six methods that believe the signal is due to stellar variations returned the five lowest RMS values for their clean RVs. As we saw with the lack of correlation between activity RVs from different methods in Figure

5, here we see again that the different methods do not agree on what the stellar signal is.

5. SUMMARY

Using EXPRES data as a test bed for several different methods, the *ESSP* is able to make a direct comparison between the results of twenty-one methods (including method variants) for disentangling stellar signals from true center-of-mass shifts. Methods returned clean RVs, with stellar signals removed, and activity RVs, which capture the variation that was removed.

The different methods varied in the type of data read in, metric for the presence of photospheric velocities, and mitigation of these signals once detected. We compared method results based on the total and nightly RMS of the returned clean RVs, agreement between activity RVs, and correlation between activity RVs and activity indicators.

5.1. Categories of Methods

Methods for disentangling stellar signals operate along three broad lines. Some methods innovate on the idea of activity indicators and use different models to derive a metric for the amplitude of the stellar signal present in an observation. Other methods instead use such indicators and construct models for mapping this amplitude measurement to the appropriate RV correction.

The last category of method separates the data into components that inform the true bulk shift of the star and components that add variability. For instance, line-by-line methods separate variable lines from more stable lines that are assumed to be a better tracer of the true bulk shift of a star. Many of the methods that model the CCF determine the shape-driven component of the measured RVs as opposed to the shift-driven component.

Table 8 summarizes all submitted methods along these three lines. Variations on the same method idea are not included. Some methods naturally produce a metric as well and so operate along more than one of the three lines.

5.2. Method Results

The historical standard of linearly decorrelating RVs against activity indicators rarely changes the resultant RV RMS significantly. This method of mitigating stellar signals is not sufficient in an EPRV context.

Most of the submitted methods reduce the RV RMS for all targets. However, no method is able to completely model out the contribution from stellar signals. EXPRES data of quiet stars exhibit an RMS of 0.5–0.8

² There are no significant SCC values for HD 10700

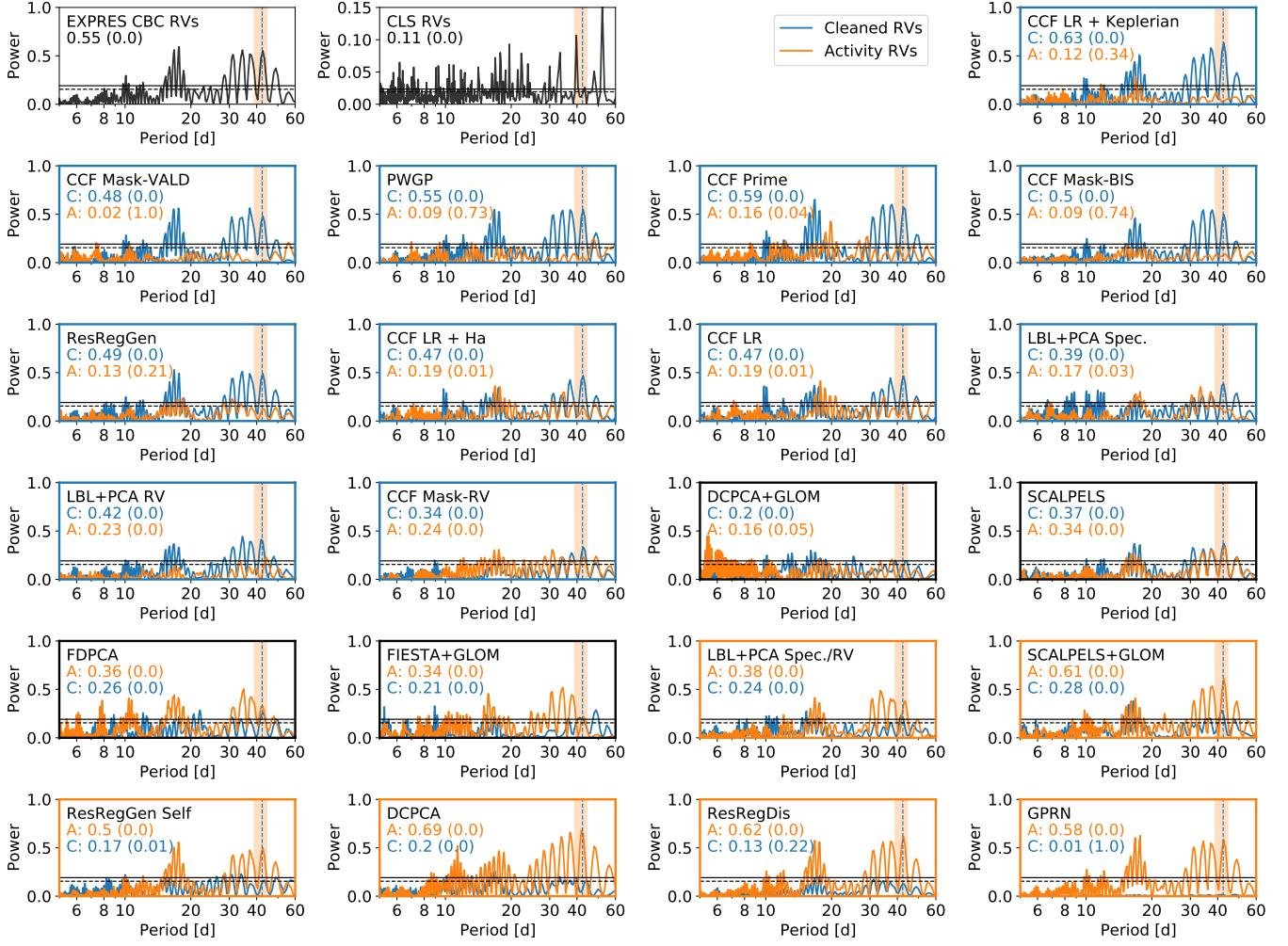


Figure 7. Periodograms of RVs for HD 26965. The top-left two plots show the periodogram for the provided EXPRES CBC RVs (left) and over 30 years of RVs from the California Legacy Survey (CLS) on the right (Rosenthal et al. 2021). Periodograms of the clean (blue) and activity (orange) RVs are given for all twenty-one of the methods that submitted ESSP results for HD 26965. Horizontal dashed and solid black lines denote p-values of 0.1 and 0.01 respectively. The proposed period for HD 26965 b, 42.38 days, is marked by a vertical, dashed blue line while the range given for the stellar rotation rate, 39-44.5 days, is shaded orange (Ma et al. 2018). The maximum power in this shaded region for both the clean (C) and activity (A) RV periodogram is given in the top-left corner of each subplot along with the corresponding p-value of the peak. Methods are ordered left-to-right and top-to-bottom by the difference between the clean and activity periodogram peaks. Subplots for methods with a stronger peak with the clean RVs have blue axes, Orange axes indicate a stronger periodicity in the activity RVs.

m s^{-1} ; no method³ was able to reduce the RV RMS to less than 1.2 m s^{-1} except for the GPRN method for HD 26965 only.

The relative RMS of a method changed from target to target. HD 101501 and HD 26965 saw the most variation in relative method performance as a few methods returned much lower relative RMS values than they did

for other targets. HD 101501 is the most chromospherically active of the four targets. HD 26965 was complicated by a proposed planet signal very close to the measured rotation rate of the star. Relative method RMS was much more consistent between HD 10700 and HD 34411, which had comparable original RV RMS values despite the dense time sampling on two nights for HD 10700. This change in behavior between the different stars may suggest that methods may perform differently depending on the amplitude or dominant type of variation exhibited by different stars. This may also con-

³ We do not include the results of the ResRegGen Self method as it is not considered to be statistically rigorous. This result was mainly included as a test of the importance of incorporating cross-validation into models.

Table 8. Method Philosophies

Method	Metric	Mitigation	Separation
<i>GLOM</i>		Multi-Dimensional GP Modeling	
<i>FDPCA</i>		Commonalities in Fourier Space	
<i>GPRN</i>		GP Neural Net Modeling	
<i>SCALPELS</i>	PCA Amplitudes (CCF)		Shape/Shift-Driven RVs
<i>CCF Prime</i>	GP Model Coefficients		Shape/Shift-Driven RVs
<i>FIESTA+GLOM</i>	Fourier Model Coefficients		Shape/Shift-Driven RVs
CCF Linear Regression			
CCF Masks			Variable/Stable Lines
<i>LBL+PCA_{Spec}</i>			Variable/Stable Lines
<i>LBL+PCA_{RV}</i>	PCA Amplitudes (<i>LBL</i> RVs)		Variable/Stable Lines
<i>PWGP</i>			
<i>DCPCA</i>	PCA Amplitudes (Spectra)		
<i>ResRegGen</i>		Regression w/ Spectral Residuals	
<i>ResRegDis</i>		Regression w/ Spectral Residuals	

tribute to the lack of correlation seen between method results for the same star.

The average intra-night scatter changes very little, but does increase for some methods. Whether the INS increases or decreases can also change for different targets with the same method. We do not expect the magnetic field of a star to change on the timescale of a single night and even less so for consecutive observations taken on the same night. This means that any signal from magnetically-driven stellar variability should be nearly the same for all observations taken within a night. Methods that increase the INS may benefit from taking this information into account.

The activity RVs returned by the different methods often do not agree with one another. All methods were used on the same data set and so should be capturing the same stellar signal. Of course, different methods may have varying levels of success in modeling the observed stellar signal or be more/less sensitive to different types of photospheric velocities. For methods that did not provide activity RVs, we instead compared constructed activity RVs from differencing the originally provided CBC RVs and the submitted cleaned RVs. These constructed activity RVs may be less correlated with methods that directly generated activity RVs as there is no guarantee such a construction will contain only stellar signals. If the method modeled out more variation than just due to stellar signals, those variations will persist in the constructed activity RVs.

Some of the methods, most notably *DCPCA+GLOM* for all targets and *PWGP* for HD 26965 and HD 10700, are not correlated with the activity RVs of any of the other methods. In the case of *DCPCA+GLOM*, it is interesting to note that the *DCPCA* method results do not have the same issue, suggesting the *GLOM* implementa-

tion for this method was maybe poorly constrained. It would be interesting to investigate why the *PWGP* results exhibit such lack of correlation only for two of the four stars.

The lack of agreement between methods makes it difficult to confidently state what signal is being modeled and removed by each method to result in the observed RMS reduction. It is also a demonstration of why RV RMS alone is an incomplete metric for method performance, as it fails to establish the nature of what a method is capturing. Many methods invoke tunable parameters to control how much variability is fitted out as due to stellar signals. In optimizing these parameters, the resultant RV RMS should be used as a goodness metric with caution.

This disagreement is further highlighted with the HD 26965 results. Whether or not HD 26965 hosts a planet changes depending on the method used. Some methods model out the \sim 40 day period as due to photospheric velocities while others attribute that periodicity to true center-of-mass shifts. Many methods found it difficult to account for the planet signal as well as the close-by stellar rotation rate. A test with an injected Keplerian signal of known period would give clearer results.

For some methods, the derived activity RVs do exhibit a significant correlation with activity indicators. This suggests that those methods are in fact modeling out photospheric velocities. However, it would be irresponsible to expect all activity RVs to correlate with imperfect indicators using an imperfect correlation metric that is only sensitive to monotonically related series. Such a correlation is only one small step towards understanding exactly what signal a method is modeling and

whether that signal is truly due to photospheric velocities, an orbiting planet, or other.

6. DISCUSSION

An increasing number of EPRV instruments are already online returning sub-meter-per-second single-measurement precision (e.g., Pepe et al. 2013; Jurgensohn et al. 2016; Schwab et al. 2016; Carmona et al. 2018; Seifahrt et al. 2018; Gilbert et al. 2018; Blackman et al. 2020; Petersburg et al. 2020; Suárez Mascañó et al. 2020; Pepe et al. 2021) with many more optical and infrared spectrographs being commissioned, built, or planned (e.g., Szentgyorgyi et al. 2014; Thompson et al. 2016; Bouchy et al. 2017; Gibson et al. 2018). The impressive engineering feat of these different instruments is opening up a new regime of extremely stable and precise spectroscopic data. However, each of these instruments and the data they take will have to contend with added RV scatter due to stellar noise unless we can mitigate these effects to sub-meter-per-second levels. None of the methods presented in this paper were able to consistently achieve that across the data sets provided.

Though there is no one single method clearly performing the best, this collection of methods and results brings clarity to the approaches and assumptions that define current progress. Here, we will highlight some of the commonalities between methods to frame the current state-of-the-field. From this, we derive suggested future directions both for method development and continued data challenges like this one.

6.1. Common Approaches and Assumptions Between Methods

The choice of input data changes the information made available to each method. For instance, indicator-driven methods will only be able to pick up on stellar signals that are tracked by the indicators used. Similarly, CCF-based methods will only be able to account for variations that are present in the CCF. None of the methods made use of the provided photometry. Some methods were not able to use the photometry because it was not simultaneous with the RVs.

Both CCF-based methods and methods that use the full spectra can only account for line shape variations at the level of the resolution of the spectrograph. Higher resolution data will contain more information about line shape changes. Methods modeling shape changes may benefit from implementing low-pass filtering tuned to the resolution of the spectrograph. This would prevent methods from being swayed by higher-frequency variations than is allowable by the spectrograph resolution and therefore must be due to noise.

Currently, methods are tracing stellar signals within spectral data as either tied to a classic activity indicator, giving rise to a spectral line shape variation, or a cause of increased scatter. It is worth considering if stellar signals may manifest in a way that is not captured by these metrics. We know that indicators are imperfect. Stellar variation that introduces a shift rather than an asymmetry would currently be completely missed. Taking increased scatter to be synonymous with stellar variation/activity may prove a dangerous parallel as we have seen that a reduction in RMS does not necessarily equate with mitigating a stellar RV component.

Many methods assume a large range of activity states or, more specifically, that these activity states manifest in the data in a large range of ways. Methods that model the effects of stellar signals using PCA assume that stellar signals are the primary source of variation and are therefore traced by the first few/several principal components. Using correlations with indicators or increased scatter to determine the presence of stellar signals is also helped by having a large range of activity states sampled.

Template CCFs and spectra are needed for many methods. Methods varied in whether they used the mean, median, or optimization (e.g., wobble) to construct this template. These templates are most often used to highlight variations away from the template, which are then attributed to the presence of stellar signals. The ideal template will not carry any significant variations so that it can be used as a reference to isolate the effects of stellar signals in each individual observation. Constructing a mean or median template CCF/spectrum and using it to highlight changes due to chromospheric velocities therefore assumes an even sampling of activity states that will average out. It would be worthwhile to investigate how dependent method results are on the template used.

On a similar note, methods that ascribe deviations from a Gaussian fit as an indication of stellar signals inherently assume that the line shape and CCF shape is well-described by a Gaussian fit. We see no evidence otherwise with the EXPRES data, for which great pains were taken to stabilize the instrument LSF across the detector. Were this not the case, however, any instrumental deviations from a Gaussian profile could be mistaken for shape changes due to stellar signals.

Many methods make the assumption that Gaussian processes and principal component analysis are good models for stellar signals. Different methods, however, implement GPs and/or PCA in distinct ways. For instance, *GLOM* uses a GP to model a time series while the *GPRN* model uses GPs to define a neural net frame-

work. *CCF Prime* forms a basis out of the derivatives of a GP model. In each case, a GP is implemented towards different ends and requires different assumptions of the appropriate kernel, hyper parameter priors, etc.

PCA can be used to construct a variation specific basis or as a measure of the amplitude of variation. Roughly speaking, the distinction can be made based on what aspect of the PCA is used. Some methods (e.g., *FIESTA*, *DCPCA*, *LBL+PCA_{RV}*, *SCALPELS+GLOM*) use just the amplitudes for each component derived from PCA as indicators of variation and therefore photospheric velocities. Other methods (e.g., *FDPCA*, *SCALPELS*, *LBL+PCA_{Spec.}*) make use of the principal components themselves to denoise spectra or model the RV shift tied to the variation being modeled by the PCA. As PCA is agnostic to the source of the variation, caring only about the amplitude, implementations of PCA may also be picking up on variations from the instrument, the extraction, tellurics, etc. which are not stellar in nature but important to correct for nonetheless.

Derived RVs are often used to align CCFs/spectra (for example for template construction), thereby implicitly assuming that true center-of-mass shifts from orbiting planets have been or can be removed leaving only stellar signals. We know, however, that these measured RVs are swayed by stellar signals. Methods should consider iterating with clean RVs produced by methods given different results, provided we are confident the corrections are truly removing only stellar signals (e.g., Cretignier et al. 2021).

Methods also all operate under the self test framework (i.e., no built-in cross-validation), meaning all data is used to construct the model, unless otherwise stated. From comparing the results between *ResRegGen* and *ResRegGen Self*, we saw that the *ResRegGen Self* method always returned a lower RMS but returned activity RVs that were not even correlated with the activity indicator used to guide the model. This suggests that the *ResRegGen Self* model is over-fitting as it contains signals that are not informed by the indicator, something the cross-validation aspect of *ResRegGen* guards against. Implementing leave-one-out, such as is done here by *LBL+PCA_{Spec.}*, *ResRegDis*, and described for *SCALPELS* in Collier Cameron et al. (2019), or other cross-validation tests, such as the framework for *ResRegGen*, should be a default of methods disentangling stellar signals when applicable in order to ensure the stability of the model being used.

6.2. Future Directions for Methods

While the reduction in RMS for the cleaned RVs of the different methods is encouraging, as a one-dimensional

metric of method performance, it is not clear what exactly is resulting in this reduced scatter. This is especially worrisome given the lack of agreement between method results. To progress, methods should be held to a higher level of interpretability. Understanding what exactly methods are tracing will be helpful in developing them further and build confidence that potential planetary signals are preserved.

Interpretability is easier to establish when there is a known ground truth—i.e., what the stellar signal is expected to be, and what is a true center-of-mass shift. One such test would be to inject simulated, center-of-mass shifts into real data. Methods that are truly only picking up on line shape changes and asymmetries will preserve these injected shifts. The most informative simulations will be shifts of the magnitude similar to the RMS of the data and at periods near the stellar rotation rate or its harmonics, as these signals will be the hardest to disentangle.

A kind of ground truth is also known for well-characterized systems, the prime example of which is our Sun. The Sun remains one of the few stars for which we can definitively remove all planet shifts⁴. Any remaining variation in the solar spectra will be from stellar signals or instrumental variation. We are also able to trivially image the surface of the Sun. With several solar telescopes expected to accompany next-generations instruments coming on line, simultaneous observations using different instruments along with photometry and surface maps will help isolate stellar signals from unique instrumental variation. Dense sampling and high cadence will additionally be immensely more achievable for the Sun than with other stars.

At the same time, the field should be careful not to lean on solar data, or simulations constructed with only solar data, exclusively. Stellar signals and their spectral manifestations differ for different types of stars. It is necessary to build up the ability to convincingly simulate or thoroughly characterize stellar signals that arise from a range of spectral types to ensure that method performance is universal.

All the different metrics being used to trace the presence of stellar signals should be tried with all the different methods being employed to mitigate the effects of stellar signals (i.e., as outlined in columns one and two of Table 8). Rather than trying to find one, “best”

⁴ where we are assuming the RV signal from the proposed Planet 9 would be below the white noise level; most constraints on Planet 9’s orbit correspond to an RV semi-amplitude of $\sim 4 \text{ cm s}^{-1}$ Batygin & Brown (2016); Batygin et al. (2019); Millholland & Laughlin (2017); Brown & Batygin (2021)

method as they are currently named, we should instead be testing all combinations of metrics and mitigation strategies to fully explore the parameter space.

Results of the different separation methods (i.e., methods outlined in column three of Table 8) should be compared with one another to see if any ground truth can be established. For instance, all the line-by-line methods work to identify lines that are more or less variable. It would be informative to understand which lines the methods agree on and for which lines they differ. Using physical information about the different lines, e.g., the line's element, transition specifics, formation level in the stellar photosphere, etc., can lend interpretability to these line-by-line methods and other methods that identify variation in the spectra.

Line-by-line methods have thus far primarily used scatter in returned RV, correlation with different activity indicators (classic or otherwise), and error of resultant RV to vet lines or chunks. More advanced methods for vetting may be interesting to explore. For instance, a periodogram of the RVs returned could be used to vet for time series that show power at troubling periods, e.g., the stellar rotation rate. Clustering analysis may also be useful in identifying lines or chunks with similar properties and help link problematic chunks with one another.

The axes of variation revealed by the different PCA methods could be picking up on the same variations. Commonalities between methods lends significance to the variations captured, which could be traced back to effects we would expect from an understanding of stellar physics. Different methods decomposing the CCF should have some commonalities even if the basis used varies greatly.

None of the methods analyzed here made use of photometry, though such efforts exist and have shown success (e.g., Aigrain et al. 2012; Cabot et al. 2021; ?). As an independent probe of activity on the stellar surface, photometry has proved useful for linking the signal being modeled by a method with changes on a star's surface (Kosiarek & Crossfield 2020). In this way, incorporating photometric information into more methods would help with method interpretability by tying the modeled RV signals to a separate measure of activity. Simultaneous photometry is most immediately useful for most current methods.

Currently, we do not have a good understanding of the precision or cadence of photometry needed to inform EPRV work. Future research should work to understand the type of photometry needed to guide methods for disentangling stellar signals. Simultaneous photometry should be prioritized.

6.3. Future Directions for Data Challenges

Comparing methods with consistent data sets will grow increasingly important as EXPRES and other next-generation spectrographs continue collecting high-fidelity data. For this report, we carried out only a few fairly simplistic test using the RMS of submitted RVs and correlation coefficients. We have seen that RMS is not sufficient to capture exactly what a model is doing, and we know that activity indicators are neither perfect nor expected to be linearly correlated with stellar signals.

The field would greatly benefit from the development of more representative comparison metrics that do a more complete job of diagnosing the extent to which methods are capturing the effects of stellar signals. Invoking a periodicity dependence or expectation for the effects of stellar signals beyond increasing scatter would be a good start. Establishing a standard suite of assessments for all methods will help place old and new methods in context.

Future data can serve as the truest validation set for methods trained on the already provided data. More data will also likely sample a greater range of activity states, resulting in additional variation in the observed spectra that will help method performance.

The existing data along with any future data can also be used to empirically determine data requirement limits for the different methods. We can synthetically degrade the data sets provided to methods to establish how method performance depends on different aspects of the data sets. For example, in addition to total number of data points, the cadence of the data (e.g., n observations in a month vs. n observations over a year) or nightly sampling (e.g., three observations per night or only one) can be adjusted. The SNR of the observations can also easily be degraded or the resolution.

There are currently several data pipelines and methods for extracted spectra and removing instrumental signals (Petersburg et al. 2020; Zhao et al. 2021; Cretignier et al. 2021). It is worth considering the effect different extraction pipelines may have on the ability to model out stellar signals. Method performance could change with the degree to which instrument variations are addressed, wavelength calibration, whether the echelle orders are merged, continuum normalization, etc.

Similarly, adjusting CCF masks and construction methods is ongoing research, as we saw with the various CCF Mask methods. The best CCF line list, mask window, and pipeline may differ for different stars but also for different use cases. For instance, the method results given here chose quiet lines to return quiet RVs, but there may be a use case for choosing the identified

variable lines instead to construct a CCF mask meant to highlight the signatures of stellar variability. Though we requested that all CCF methods use the provided CCFs for this report, exploration is warranted as to how different CCFs may change the results of these methods.

Currently, the focus of many methods and indicators lie in tracing activity features or magnetic field strength; less emphasis is placed on inherent stellar variability, such as p-mode oscillations or (super-)granulation. Pulsations and changes in granulation pattern persist on the time scale of minutes while super-granulation has a time-scale of hours to days. Pulsations may cause lines to shift rather than change in shape. These types of variation will have a different diagnostic than activity features.

Before we can disentangle the effect from granulation, we must understand it. This will require very densely sampled observations at high-enough resolution. “High-enough” resolution scales with the size of activity feature we are interested in. Given the time-scale of (super-)granulation, the ideal data set will have very dense sampling over the course of a night for four to five consecutive nights in order to capture both short-term granulation features and potentially day-long super-granulation effects.

7. CONCLUSIONS

Twenty-one different methods (including variations) were tried on EXPRES data to produce a consistent comparison of method results on data that are representative of extreme-precision instruments. The methods tested return lower RMS values than the classic linear decorrelation methods in nearly all cases. Though EXPRES data of quiet stars regularly return RMS values of $0.5\text{--}0.8 \text{ m s}^{-1}$, no method is yet consistently reducing the RMS of more chromospherically active stars to sub-meter-per-second levels across all four stars. Lack of agreement between the signals being modeled out by different methods makes it difficult to determine what exactly is being modeled by each method and whether the signals being picked out are truly stellar in origin.

Current and future methods should consider:

- increasing method interpretability in order to establish the source of the signals being picked out by the method,
- ensuring models are appropriately general by implementing cross-validation tests,
- iterating when aligning CCF/spectra with derived RVs, and

- making methods robust to the assumption that a large range or equal distribution of activity states is covered within the data set.

Methods currently work at identifying the presence of stellar signals by using either a derived activity indicator, changes in line shape, or increased scatter. Future work should investigate whether those diagnostics are comprehensive.

Next steps for establishing method performance include:

- developing more holistic metrics for how well a method disentangles stellar signals,
- cross-pollinating methods that generate activity indicators with methods that are informed by indicators,
- comparing and contrasting results of similar methods, e.g., *LBL* methods, derived PCA components, GP hyperparameters, etc.,
- testing methods on well-characterized systems, e.g., solar data, dynamically packed planetary systems, data with injected Keplerian or stellar signals, etc., and
- testing methods on EXPRES data sets degraded in terms of SNR, resolution, observing cadence, etc.

Note that care must be taken when injecting a Keplerian signal to ensure telluric lines are not also shifted. Injecting stellar signals will require developing simulations capable of faithfully reproducing all flavors of stellar variability and activity.

The design of RV surveys should consider whether to prioritize phase coverage of potential planets or fully characterizing the effects of stellar signals. An EPRV data set that fully resolves all time scales of stellar signals may be needed to completely understand the effects of chromospheric velocities on spectra. Such a data set for a Sun-like star would likely need to span 4–5 consecutive nights with at least 2–3 hours of continuous, densely sampled observations per night.

While progress is being made in mitigating stellar signals, more work remains to be done. We will not be able to successfully detect Earth-like planets until photospheric velocities from inherent stellar variability and activity features can be disentangled to below the 50 cm s^{-1} level.

Facilities: LDT, TSU:AST

ACKNOWLEDGMENTS

These results made use of the Lowell Discovery Telescope at Lowell Observatory. Lowell is a private, non-profit institution dedicated to astrophysical research and public appreciation of astronomy and operates the LDT in partnership with Boston University, the University of Maryland, the University of Toledo, Northern Arizona University and Yale University. LLZ gratefully acknowledges support from the NSF GRFP under Grant No. DGE1122492 and the Green family. DAF acknowledges support for the design and construction of EXPRES from NSF MRI-1429365, NSF ATI-1509436 and Yale University. DAF gratefully acknowledges support to carry out this research from NSF 2009528, NSF 1616086, NASA 80NSSC18K0443, NSF AST-2009528, the Heising-Simons Foundation, and an anonymous donor in the Yale alumni community. This work was partially supported by NASA Exoplanet Research Program Grant #80NSSC18K0443 (DAF, EBF, AW, JZ). This work was supported by a grant from the Simons Foundation/SFARI (675601, E.B.F.). This research was partially supported by Heising-Simons Foundation Grant #2019-1177 (E.B.F.). The Center for Exoplanets and Habitable Worlds is supported by the Pennsylvania State University and the Eberly College of Science.

G. W. H. acknowledges long-term support from NASA, NSF, Tennessee State University, and the State of Tennessee through its Centers of Excellence program. RMR acknowledges support from the Yale Center for Astronomy & Astrophysics (YCAA) Prize Postdoctoral

Fellowship and the Heising-Simons 51 Pegasi b Postdoctoral Fellowship. X.D. and M.C. acknowledge that this project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement SCORE No 851555). X.D. and M.C. also recognise that this work has been carried out within the framework of the NCCR PlanetS supported by the Swiss National Science Foundation. AM acknowledges support from the Cambridge Kavli Institute Fellowships. Work by SDR, JH, and VRD was supported by Bartol Research Institute. VRD received additional support from the University of Delaware Summer Scholars Program. JDC, JPJ and PTPV were supported by the following grants, awarded by FCT - Fundação para a Ciência e Tecnologia and FEDER through COMPETE2020: UIDB/04434/2020; UIDP/04434/2020; PTDC/FIS-AST/32113/2017 and POCI-01-0145-FEDER-032113; PTDC/FIS-AST/28953/2017 and POCI-01-0145-FEDER-028953. JPJ is further supported in the form of a work contract funded by national funds through FCT with reference DL57/2016/CP1364/CT0005. SA, BK and OB acknowledge support from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 865624). NZ is supported by studentship no. 1947725 under Grant Code ST/N504233/1 from the UK Science and Technology Facilities Council (STFC). H.M.C. and M.L. acknowledge support from the UKRI FLF grant MR/S035214/1.



REFERENCES

- Aigrain, S., Pont, F., & Zucker, S. 2012, MNRAS, 419, 3147, doi: [10.1111/j.1365-2966.2011.19960.x](https://doi.org/10.1111/j.1365-2966.2011.19960.x)
- Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O’Neil, M. 2015, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 252, doi: [10.1109/TPAMI.2015.2448083](https://doi.org/10.1109/TPAMI.2015.2448083)
- Angus, R., Morton, T., Aigrain, S., Foreman-Mackey, D., & Rajpaul, V. 2018, MNRAS, 474, 2094, doi: [10.1093/mnras/stx2109](https://doi.org/10.1093/mnras/stx2109)
- Arentoft, T., Kjeldsen, H., Bedding, T. R., et al. 2008, ApJ, 687, 1180, doi: [10.1086/592040](https://doi.org/10.1086/592040)
- Baliunas, S., Sokoloff, D., & Soon, W. 1996, ApJL, 457, L99, doi: [10.1086/309891](https://doi.org/10.1086/309891)
- Barragán, O., Aigrain, S., Rajpaul, V. M., & Zicher, N. 2021, arXiv e-prints, arXiv:2109.14086. <https://arxiv.org/abs/2109.14086>
- Barragán, O., Gandolfi, D., & Antoniciello, G. 2019, MNRAS, 482, 1017, doi: [10.1093/mnras/sty2472](https://doi.org/10.1093/mnras/sty2472)
- Batygin, K., Adams, F. C., Brown, M. E., & Becker, J. C. 2019, PhR, 805, 1, doi: [10.1016/j.physrep.2019.01.009](https://doi.org/10.1016/j.physrep.2019.01.009)
- Batygin, K., & Brown, M. E. 2016, AJ, 151, 22, doi: [10.3847/0004-6256/151/2/22](https://doi.org/10.3847/0004-6256/151/2/22)
- Bedell, M., Hogg, D. W., Foreman-Mackey, D., Montet, B. T., & Luger, R. 2019, AJ, 158, 164, doi: [10.3847/1538-3881/ab40a7](https://doi.org/10.3847/1538-3881/ab40a7)
- Blackman, R. T., Szymkowiak, A. E., Fischer, D. A., & Jurgenson, C. A. 2017, The Astrophysical Journal, 837, 18, doi: [10.3847/1538-4357/aa5ead](https://doi.org/10.3847/1538-4357/aa5ead)
- Blackman, R. T., Fischer, D. A., Jurgenson, C. A., et al. 2020, AJ, 159, 238, doi: [10.3847/1538-3881/ab811d](https://doi.org/10.3847/1538-3881/ab811d)
- Boisse, I., Bouchy, F., Hébrard, G., et al. 2011, A&A, 528, A4, doi: [10.1051/0004-6361/201014354](https://doi.org/10.1051/0004-6361/201014354)

- Boisse, I., Moutou, C., Vidal-Madjar, A., et al. 2009, *A&A*, 495, 959, doi: [10.1051/0004-6361:200810648](https://doi.org/10.1051/0004-6361:200810648)
- Bouchy, F., Bazot, M., Santos, N. C., Vauclair, S., & Sosnowska, D. 2005, *A&A*, 440, 609, doi: [10.1051/0004-6361:20052697](https://doi.org/10.1051/0004-6361:20052697)
- Bouchy, F., Doyon, R., Artigau, É., et al. 2017, *The Messenger*, 169, 21, doi: [10.18727/0722-6691/5034](https://doi.org/10.18727/0722-6691/5034)
- Boyajian, T. S., McAlister, H. A., van Belle, G., et al. 2012, *ApJ*, 746, 101, doi: [10.1088/0004-637X/746/1/101](https://doi.org/10.1088/0004-637X/746/1/101)
- Bradshaw, S. J., & Hartigan, P. 2014, *ApJ*, 795, 79, doi: [10.1088/0004-637X/795/1/79](https://doi.org/10.1088/0004-637X/795/1/79)
- Brewer, J. M., & Fischer, D. A. 2016, *ApJ*, 831, 20, doi: [10.3847/0004-637X/831/1/20](https://doi.org/10.3847/0004-637X/831/1/20)
- Brewer, J. M., Fischer, D. A., Valenti, J. A., & Piskunov, N. 2016, *ApJS*, 225, 32, doi: [10.3847/0067-0049/225/2/32](https://doi.org/10.3847/0067-0049/225/2/32)
- Brewer, J. M., Fischer, D. A., Blackman, R. T., et al. 2020, *AJ*, 160, 67, doi: [10.3847/1538-3881/ab99c9](https://doi.org/10.3847/1538-3881/ab99c9)
- Brown, M. E., & Batygin, K. 2021, arXiv e-prints, arXiv:2108.09868. <https://arxiv.org/abs/2108.09868>
- Cabot, S. H. C., Roettenbacher, R. M., Henry, G. W., et al. 2021, *AJ*, 161, 26, doi: [10.3847/1538-3881/abc41e](https://doi.org/10.3847/1538-3881/abc41e)
- Carmona, A., Donati, J. F., Moutou, C., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII, ed. C. J. Evans, L. Simard, & H. Takami, 1070215, doi: [10.1117/12.2313777](https://doi.org/10.1117/12.2313777)
- Cegla, H. M., Watson, C. A., Shelyag, S., et al. 2018, *ApJ*, 866, 55, doi: [10.3847/1538-4357/aaddfc](https://doi.org/10.3847/1538-4357/aaddfc)
- Chaplin, W. J., Cegla, H. M., Watson, C. A., Davies, G. R., & Ball, W. H. 2019, *AJ*, 157, 163, doi: [10.3847/1538-3881/ab0c01](https://doi.org/10.3847/1538-3881/ab0c01)
- Collier Cameron, A., Mortier, A., Phillips, D., et al. 2019, *MNRAS*, 487, 1082, doi: [10.1093/mnras/stz1215](https://doi.org/10.1093/mnras/stz1215)
- Collier Cameron, A., Ford, E. B., Shahaf, S., et al. 2020, arXiv e-prints, arXiv:2011.00018. <https://arxiv.org/abs/2011.00018>
- Crane, J. D., Shectman, S. A., & Butler, R. P. 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 6269, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. I. S. McLean & M. Iye, 626931, doi: [10.1117/12.672339](https://doi.org/10.1117/12.672339)
- Cretignier, M., Dumusque, X., Allart, R., Pepe, F., & Lovis, C. 2020a, *A&A*, 633, A76, doi: [10.1051/0004-6361/201936548](https://doi.org/10.1051/0004-6361/201936548)
- Cretignier, M., Dumusque, X., Hara, N. C., & Pepe, F. 2021, arXiv e-prints, arXiv:2106.07301. <https://arxiv.org/abs/2106.07301>
- Cretignier, M., Francfort, J., Dumusque, X., Allart, R., & Pepe, F. 2020b, *A&A*, 640, A42, doi: [10.1051/0004-6361/202037722](https://doi.org/10.1051/0004-6361/202037722)
- Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. A., & Ford, E. B. 2017, *ApJ*, 846, 59, doi: [10.3847/1538-4357/aa8303](https://doi.org/10.3847/1538-4357/aa8303)
- de Beurs, Z. L., Vanderburg, A., Shallue, C. J., et al. 2020, arXiv e-prints, arXiv:2011.00003. <https://arxiv.org/abs/2011.00003>
- Desort, M., Lagrange, A. M., Galland, F., Udry, S., & Mayor, M. 2007, *A&A*, 473, 983, doi: [10.1051/0004-6361:20078144](https://doi.org/10.1051/0004-6361:20078144)
- Díaz, M. R., Jenkins, J. S., Tuomi, M., et al. 2018, *AJ*, 155, 126, doi: [10.3847/1538-3881/aaa896](https://doi.org/10.3847/1538-3881/aaa896)
- Dravins, D. 1982, *ARA&A*, 20, 61, doi: [10.1146/annurev.aa.20.090182.000425](https://doi.org/10.1146/annurev.aa.20.090182.000425)
- Ducati, J. R. 2002, *VizieR Online Data Catalog*
- Dumusque, X. 2018, *A&A*, 620, A47, doi: [10.1051/0004-6361/201833795](https://doi.org/10.1051/0004-6361/201833795)
- Dumusque, X., Santos, N. C., Udry, S., Lovis, C., & Bonfils, X. 2011a, *A&A*, 527, A82, doi: [10.1051/0004-6361/201015877](https://doi.org/10.1051/0004-6361/201015877)
- Dumusque, X., Udry, S., Lovis, C., Santos, N. C., & Monteiro, M. J. P. F. G. 2011b, *A&A*, 525, A140, doi: [10.1051/0004-6361/201014097](https://doi.org/10.1051/0004-6361/201014097)
- Dumusque, X., Lovis, C., Ségransan, D., et al. 2011c, *A&A*, 535, A55, doi: [10.1051/0004-6361/201117148](https://doi.org/10.1051/0004-6361/201117148)
- Dumusque, X., Borsa, F., Damasso, M., et al. 2017, *A&A*, 598, A133, doi: [10.1051/0004-6361/201628671](https://doi.org/10.1051/0004-6361/201628671)
- Faria, J. P., Haywood, R. D., Brewer, B. J., et al. 2016, *A&A*, 588, A31, doi: [10.1051/0004-6361/201527899](https://doi.org/10.1051/0004-6361/201527899)
- Feng, F., Tuomi, M., Jones, H. R. A., et al. 2017, *AJ*, 154, 135, doi: [10.3847/1538-3881/aa83b4](https://doi.org/10.3847/1538-3881/aa83b4)
- Figueira, P. 2013, Astronomical Society of the Pacific Conference Series, Vol. 472, Stellar Noise in Exoplanet Searches, ed. M. Chavez, E. Bertone, O. Vega, & V. De la Luz, 137
- Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, *PASP*, 128, 066001, doi: [10.1088/1538-3873/128/964/066001](https://doi.org/10.1088/1538-3873/128/964/066001)
- Ford, E., Wise, A., & Palumbo, M. 2021, *RvSpectML/EchelleCCFs.jl*: v0.1.11, v0.1.11, Zenodo, doi: [10.5281/zenodo.4593963](https://doi.org/10.5281/zenodo.4593963)
- Gaia Collaboration. 2018, *VizieR Online Data Catalog*, I/345
- Gibson, S. R., Howard, A. W., Roy, A., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII, 107025X, doi: [10.1117/12.2311565](https://doi.org/10.1117/12.2311565)

- Giguere, M. J., Fischer, D. A., Zhang, C. X. Y., et al. 2016, ApJ, 824, 150, doi: [10.3847/0004-637X/824/2/150](https://doi.org/10.3847/0004-637X/824/2/150)
- Gilbert, J., Bergmann, C., Bloxham, G., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII, ed. C. J. Evans, L. Simard, & H. Takami, 107020Y, doi: [10.1117/12.2312399](https://doi.org/10.1117/12.2312399)
- Gilbertson, C., Ford, E. B., Jones, D. E., & Stenning, D. C. 2020a, ApJ, 905, 155, doi: [10.3847/1538-4357/abc627](https://doi.org/10.3847/1538-4357/abc627)
- . 2020b, ApJ, 905, 155, doi: [10.3847/1538-4357/abc627](https://doi.org/10.3847/1538-4357/abc627)
- Hatzes, A. P. 2002, Astronomische Nachrichten, 323, 392, doi: [10.1002/1521-3994\(200208\)323:3/4<392::AID-ASNA392>3.0.CO;2-M](https://doi.org/10.1002/1521-3994(200208)323:3/4<392::AID-ASNA392>3.0.CO;2-M)
- Haywood, R. D., Collier Cameron, A., Queloz, D., et al. 2014, International Journal of Astrobiology, 13, 155, doi: [10.1017/S147355041300044X](https://doi.org/10.1017/S147355041300044X)
- Henry, G. W. 1999, PASP, 111, 845, doi: [10.1086/316388](https://doi.org/10.1086/316388)
- Holzer, P. H., Cisewski-Kehe, J., Zhao, L., et al. 2021, AJ, 161, 272, doi: [10.3847/1538-3881/abf5e0](https://doi.org/10.3847/1538-3881/abf5e0)
- Huélamo, N., Figueira, P., Bonfils, X., et al. 2008, A&A, 489, L9, doi: [10.1051/0004-6361:200810596](https://doi.org/10.1051/0004-6361:200810596)
- Isaacson, H., & Fischer, D. 2010, ApJ, 725, 875, doi: [10.1088/0004-637X/725/1/875](https://doi.org/10.1088/0004-637X/725/1/875)
- Jeffers, S. V., Barnes, J. R., Jones, H., & Pinfield, D. 2013, in European Physical Journal Web of Conferences, Vol. 47, European Physical Journal Web of Conferences, 09002, doi: [10.1051/epjconf/20134709002](https://doi.org/10.1051/epjconf/20134709002)
- Jones, D. E., Stenning, D. C., Ford, E. B., et al. 2017, arXiv e-prints, arXiv:1711.01318.
<https://arxiv.org/abs/1711.01318>
- Jones, D. E., Stenning, D. C., Ford, E. B., et al. 2021, Improving Exoplanet Detection Power: Multivariate Gaussian Process Models for Stellar Activity.
<https://arxiv.org/abs/1711.01318>
- Jurgenson, C., Fischer, D., McCracken, T., et al. 2016, in Proc. SPIE, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 99086T, doi: [10.1117/12.2233002](https://doi.org/10.1117/12.2233002)
- Kjeldsen, H., & Bedding, T. R. 1995, A&A, 293, 87.
<https://arxiv.org/abs/astro-ph/9403015>
- Kjeldsen, H., Bedding, T. R., Butler, R. P., et al. 2005, ApJ, 635, 1281, doi: [10.1086/497530](https://doi.org/10.1086/497530)
- Kosiarek, M. R., & Crossfield, I. J. M. 2020, AJ, 159, 271, doi: [10.3847/1538-3881/ab8d3a](https://doi.org/10.3847/1538-3881/ab8d3a)
- Lafarga, M., Ribas, I., Lovis, C., et al. 2020, A&A, 636, A36, doi: [10.1051/0004-6361/201937222](https://doi.org/10.1051/0004-6361/201937222)
- Lanza, A. F., Gizon, L., Zaqarashvili, T. V., Liang, Z. C., & Rodenbeck, K. 2019, A&A, 623, A50, doi: [10.1051/0004-6361/201834712](https://doi.org/10.1051/0004-6361/201834712)
- Leet, C., Fischer, D. A., & Valenti, J. A. 2019, AJ, 157, 187, doi: [10.3847/1538-3881/ab0d86](https://doi.org/10.3847/1538-3881/ab0d86)
- Levine, S. E., Bida, T. A., Chylek, T., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8444, Ground-based and Airborne Telescopes IV, 844419, doi: [10.1117/12.926415](https://doi.org/10.1117/12.926415)
- Lindgren, L., & Dravins, D. 2003, A&A, 401, 1185, doi: [10.1051/0004-6361:20030181](https://doi.org/10.1051/0004-6361:20030181)
- Lomb, N. R. 1976, Ap&SS, 39, 447, doi: [10.1007/BF00648343](https://doi.org/10.1007/BF00648343)
- Lovis, C., Dumusque, X., Santos, N. C., et al. 2011, arXiv e-prints, arXiv:1107.5325.
<https://arxiv.org/abs/1107.5325>
- Ma, B., Ge, J., Mutterspaugh, M., et al. 2018, MNRAS, 480, 2411, doi: [10.1093/mnras/sty1933](https://doi.org/10.1093/mnras/sty1933)
- Maldonado, J., Martínez-Arnáiz, R. M., Eiroa, C., & Montes, D. 2010, Astrophysics and Space Science Proceedings, 14, 419, doi: [10.1007/978-3-642-11250-8_108](https://doi.org/10.1007/978-3-642-11250-8_108)
- Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20
- Meunier, N., Desort, M., & Lagrange, A. M. 2010, A&A, 512, A39, doi: [10.1051/0004-6361/200913551](https://doi.org/10.1051/0004-6361/200913551)
- Meunier, N., & Lagrange, A. M. 2013, A&A, 551, A101, doi: [10.1051/0004-6361/201219917](https://doi.org/10.1051/0004-6361/201219917)
- . 2019, A&A, 625, L6, doi: [10.1051/0004-6361/201935099](https://doi.org/10.1051/0004-6361/201935099)
- Meunier, N., Lagrange, A. M., & Borgniet, S. 2017, A&A, 607, A6, doi: [10.1051/0004-6361/201630328](https://doi.org/10.1051/0004-6361/201630328)
- Meunier, N., Lagrange, A. M., Borgniet, S., & Rieutord, M. 2015, A&A, 583, A118, doi: [10.1051/0004-6361/201525721](https://doi.org/10.1051/0004-6361/201525721)
- Milaković, D., Pasquini, L., Webb, J. K., & Lo Curto, G. 2020, MNRAS, 493, 3997, doi: [10.1093/mnras/staa356](https://doi.org/10.1093/mnras/staa356)
- Millholland, S., & Laughlin, G. 2017, AJ, 153, 91, doi: [10.3847/1538-3881/153/3/91](https://doi.org/10.3847/1538-3881/153/3/91)
- Molaro, P., Esposito, M., Monai, S., et al. 2013, A&A, 560, A61, doi: [10.1051/0004-6361/201322324](https://doi.org/10.1051/0004-6361/201322324)
- Nidever, D. L., Marcy, G. W., Butler, R. P., Fischer, D. A., & Vogt, S. S. 2002, ApJS, 141, 503, doi: [10.1086/340570](https://doi.org/10.1086/340570)
- Pepe, F., Cristiani, S., Rebolo, R., et al. 2013, The Messenger, 153, 6
- . 2021, A&A, 645, A96, doi: [10.1051/0004-6361/202038306](https://doi.org/10.1051/0004-6361/202038306)
- Petersburg, R. R., Ong, J. M. J., Zhao, L. L., et al. 2020, AJ, 159, 187, doi: [10.3847/1538-3881/ab7e31](https://doi.org/10.3847/1538-3881/ab7e31)
- Pijpers, F. P. 2003, A&A, 400, 241, doi: [10.1051/0004-6361:20021839](https://doi.org/10.1051/0004-6361:20021839)
- Povich, M. S., Giampapa, M. S., Valenti, J. A., et al. 2001, AJ, 121, 1136, doi: [10.1086/318745](https://doi.org/10.1086/318745)

- Probst, R. A., Lo Curto, G., Avila, G., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, ed. S. K. Ramsay, I. S. McLean, & H. Takami, 91471C, doi: [10.1117/12.2055784](https://doi.org/10.1117/12.2055784)
- Probst, R. A., Milaković, D., Toledo-Padrón, B., et al. 2020, *Nature Astronomy*, 4, 603, doi: [10.1038/s41550-020-1010-x](https://doi.org/10.1038/s41550-020-1010-x)
- Queloz, D., Henry, G. W., Sivan, J. P., et al. 2001, *A&A*, 379, 279, doi: [10.1051/0004-6361:20011308](https://doi.org/10.1051/0004-6361:20011308)
- Queloz, D., Bouchy, F., Moutou, C., et al. 2009, *A&A*, 506, 303, doi: [10.1051/0004-6361/200913096](https://doi.org/10.1051/0004-6361/200913096)
- Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, *MNRAS*, 452, 2269, doi: [10.1093/mnras/stv1428](https://doi.org/10.1093/mnras/stv1428)
- Rajpaul, V., Buchhave, L. A., & Aigrain, S. 2017, *MNRAS*, 471, L125, doi: [10.1093/mnrasl/slx116](https://doi.org/10.1093/mnrasl/slx116)
- Rajpaul, V. M., Aigrain, S., & Buchhave, L. A. 2020, *MNRAS*, 492, 3960, doi: [10.1093/mnras/stz3599](https://doi.org/10.1093/mnras/stz3599)
- Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian Processes for Machine Learning (MIT Press)
- Rieutord, M., & Rincon, F. 2010, *Living Reviews in Solar Physics*, 7, 2, doi: [10.12942/lrsp-2010-2](https://doi.org/10.12942/lrsp-2010-2)
- Rincon, F., & Rieutord, M. 2018, *Living Reviews in Solar Physics*, 15, 6, doi: [10.1007/s41116-018-0013-5](https://doi.org/10.1007/s41116-018-0013-5)
- Robertson, P., Mahadevan, S., Endl, M., & Roy, A. 2014, *Science*, 345, 440, doi: [10.1126/science.1253253](https://doi.org/10.1126/science.1253253)
- Rosenthal, L. J., Fulton, B. J., Hirsch, L. A., et al. 2021, *ApJS*, 255, 8, doi: [10.3847/1538-4365/abe23c](https://doi.org/10.3847/1538-4365/abe23c)
- Saar, S. H. 2003, *Astronomical Society of the Pacific Conference Series*, Vol. 294, The Effects of Plage on Precision Radial Velocities, ed. D. Deming & S. Seager, 65–70
- Saar, S. H., Butler, R. P., & Marcy, G. W. 1998, *ApJL*, 498, L153, doi: [10.1086/311325](https://doi.org/10.1086/311325)
- Saar, S. H., & Donahue, R. A. 1997, *ApJ*, 485, 319, doi: [10.1086/304392](https://doi.org/10.1086/304392)
- Saar, S. H., & Fischer, D. 2000, *ApJL*, 534, L105, doi: [10.1086/312648](https://doi.org/10.1086/312648)
- Santos, N. C., Mortier, A., Faria, J. P., et al. 2014, *A&A*, 566, A35, doi: [10.1051/0004-6361/201423808](https://doi.org/10.1051/0004-6361/201423808)
- Scargle, J. D. 1982, *ApJ*, 263, 835, doi: [10.1086/160554](https://doi.org/10.1086/160554)
- Schwab, C., Rakich, A., Gong, Q., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 99087H, doi: [10.1117/12.2234411](https://doi.org/10.1117/12.2234411)
- Seifahrt, A., Stürmer, J., Bean, J. L., & Schwab, C. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII, ed. C. J. Evans, L. Simard, & H. Takami, 107026D, doi: [10.1117/12.2312936](https://doi.org/10.1117/12.2312936)
- Simola, U., Dumusque, X., & Cisewski-Kehe, J. 2019, *A&A*, 622, A131, doi: [10.1051/0004-6361/201833895](https://doi.org/10.1051/0004-6361/201833895)
- Skelly, M. B., Unruh, Y. C., Collier Cameron, A., et al. 2008, *MNRAS*, 385, 708, doi: [10.1111/j.1365-2966.2008.12917.x](https://doi.org/10.1111/j.1365-2966.2008.12917.x)
- Suárez Mascareño, A., Faria, J. P., Figueira, P., et al. 2020, arXiv e-prints, arXiv:2005.12114, <https://arxiv.org/abs/2005.12114>
- Szentgyorgyi, A., Barnes, S., Bean, J., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, ed. S. K. Ramsay, I. S. McLean, & H. Takami, 914726, doi: [10.1117/12.2056741](https://doi.org/10.1117/12.2056741)
- Thompson, A. P. G., Watson, C. A., de Mooij, E. J. W., & Jess, D. B. 2017, *MNRAS*, 468, L16, doi: [10.1093/mnrasl/slx018](https://doi.org/10.1093/mnrasl/slx018)
- Thompson, S. J., Queloz, D., Baraffe, I., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, ed. C. J. Evans, L. Simard, & H. Takami, 99086F, doi: [10.1117/12.2232111](https://doi.org/10.1117/12.2232111)
- Tokovinin, A., Fischer, D. A., Bonati, M., et al. 2013, *PASP*, 125, 1336, doi: [10.1086/674012](https://doi.org/10.1086/674012)
- Tuomi, M., Jones, H. R. A., Jenkins, J. S., et al. 2013, *A&A*, 551, A79, doi: [10.1051/0004-6361/201220509](https://doi.org/10.1051/0004-6361/201220509)
- van Leeuwen, F. 2007, Hipparcos, the New Reduction of the Raw Data, Vol. 350, doi: [10.1007/978-1-4020-6342-8](https://doi.org/10.1007/978-1-4020-6342-8)
- Vogt, S. S., Allen, S. L., Bigelow, B. C., et al. 1994, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 2198, Instrumentation in Astronomy VIII, ed. D. L. Crawford & E. R. Craine, 362, doi: [10.1117/12.176725](https://doi.org/10.1117/12.176725)
- Wilken, T., Curto, G. L., Probst, R. A., et al. 2012, *Nature*, 485, 611, doi: [10.1038/nature11092](https://doi.org/10.1038/nature11092)
- Wilson, A. G., Knowles, D. A., & Ghahramani, Z. 2012, in Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML'12 (USA: Omnipress), 1139–1146, <http://dl.acm.org/citation.cfm?id=3042573.3042720>
- Wise, A. W., Dodson-Robinson, S. E., Bevenour, K., & Provini, A. 2018, *AJ*, 156, 180, doi: [10.3847/1538-3881/aadd94](https://doi.org/10.3847/1538-3881/aadd94)

- Zechmeister, M., Anglada-Escudé, G., & Reiners, A. 2014, *A&A*, 561, A59, doi: [10.1051/0004-6361/201322746](https://doi.org/10.1051/0004-6361/201322746)
- Zhao, J., & Tinney, C. G. 2020, *MNRAS*, 491, 4131, doi: [10.1093/mnras/stz3254](https://doi.org/10.1093/mnras/stz3254)
- Zhao, L., Fischer, D. A., Ford, E. B., et al. 2020, *Research Notes of the American Astronomical Society*, 4, 156, doi: [10.3847/2515-5172/abb8d0](https://doi.org/10.3847/2515-5172/abb8d0)
- Zhao, L. L., Hogg, D. W., Bedell, M., & Fischer, D. A. 2021, *AJ*, 161, 80, doi: [10.3847/1538-3881/abd105](https://doi.org/10.3847/1538-3881/abd105)

APPENDIX

A. IN-DEPTH DESCRIPTIONS OF METHODS THAT USE RVs AND CLASSIC ACTIVITY INDICATORS AS INPUT

A.1. *GLOM*

GLOM, developed by members of the PennState Team, is a software package for joint GP modeling of several parameters, such as Doppler shifts along with one or more activity indicator time series (Gilbertson et al. 2020b). The model is based on the assumption that all time series can be modelled using a latent variable $G(t)$, which is described by a Gaussian process and a covariance function γ . The *GLOM* implementation can also incorporate a non-zero mean function, $m_n(t)$ for each set of variables being modeled.

RVs and activity indicators are then modeled together using the latent GP $G(t)$, its derivatives, and this mean function. For N total number of parameter time series, the framework is as follows:

$$\begin{aligned} q_0(t) &= m_0(t) + a_{0,0}G(t) + a_{0,1}\dot{G}(t) + a_{0,2}\ddot{G}(t) + \epsilon_0(t) \\ q_1(t) &= m_1(t) + a_{1,0}G(t) + a_{1,1}\dot{G}(t) + a_{1,2}\ddot{G}(t) + \epsilon_1(t) \\ &\vdots \\ q_N(t) &= m_N(t) + a_{N,0}G(t) + a_{N,1}\dot{G}(t) + a_{N,2}\ddot{G}(t) + \epsilon_N(t) \end{aligned} \quad (\text{A1})$$

Each $q_n(t)$ is the time series of the variables being modeled. The latent GP is given by $G(t)$, while the variables $a_{n,0}$, and $a_{n,1}$ where $n = 1, \dots, N$ are free parameters. The mean function is included as $m_n(t)$ and $\epsilon_n(t)$ represents measurement uncertainties.

GP models are a powerful tool for modeling stochastic behavior and therefore very apt for modeling photospheric velocities. However, they are liable to vacuum up all signals in a data set, including, for instance, planet signals. By modeling several time series simultaneous, this method places constraints on the GP model by incorporating the information from activity indicators into the GP modeling. This guides the model to only pick up on signals that can be tied to the provided indicators. Introducing indicators into the modeling does increases the size of the correlation matrix, making the method more computationally expensive.

The method requires RVs and corresponding indicator time series for each observation. Photometry can be used to establish a constraint on the stellar rotation period of the target. *GLOM* can be run for different combinations of RV and indicator time series, with different results depending on the indicator used. *GLOM* is incorporated as a part of many later methods that establish different indicators of activity.

The success of the method is dependent on the sampling of the data, which should be relatively close in time and appropriateness of the chosen GP kernel. It would be better to have less observations but a denser sampling throughout the characteristic time scale of the signal being modeled (i.e., the stellar rotation rate). The GP model adopts a quasi-periodic kernel along with constant offset and jitter terms for each time-series. Some care must be taken in choosing the priors for the GP hyper-parameters, which will change for different data sets.

A.2. *FDPCA*

Fourier Domain Principal Component Analysis, submitted by the Sidera team, detects common patterns in the Fourier coefficients of RV and activity-indicator time series and uses this to predict the stellar signal component of the RV. Moving to the Fourier domain allows the method to identify and remove correlated signals even if they are out of phase.

First, the non-uniform Fourier transforms of all activity-indicator time series and RVs are computed. Next, the activity-indicator Fourier series are scaled so that they have unit variance in the time domain. The Fourier series for each activity indicator are then stacked into a matrix to form a set of explanatory variables for the RV Fourier series:

$$\left[\Re(\mathcal{F}\{\text{H}\alpha\text{EW}\}) \ \Im(\mathcal{F}\{\text{H}\alpha\text{EW}\}) \ \Re(\mathcal{F}\{\text{CCF FWHM}\}) \ \Im(\mathcal{F}\{\text{CCF FWHM}\}) \ \dots \right] \quad (\text{A2})$$

where $\Re(\mathcal{F})$ and $\Im(\mathcal{F})$ are the real and imaginary parts of the Fourier transform, respectively. The matrix is then run through PCA.⁵

With activity principal components in hand, the real and imaginary parts of the RV Fourier series can be regressed onto these principal components. The regression coefficients are used to determine the contribution of the RV Fourier series that is related to the activity indicators. This measurement of the chromospheric contribution to the RV Fourier series can then be inverse transformed back into the time domain to find the stellar signal correction needed for each RV. Parseval's theorem is used to recover the correct variance of the RV activity contribution.

Implementing this method requires RVs and indicators taken at the same time stamps. The power of this method comes from identifying coherences between the provided indicators and the RV measurements.

In order to use this method to measure a signal, the observations must completely cover the phase of the signal. For example, to capture the effects of a rotating activity feature, the observations must completely sample the star's rotation. It is not just a question of dense sampling of observations, the observations must cover the entire phase range.

As with all methods that invoke PCA, there is always the question of how many principal components to incorporate. For the results presented here, principal components were included until 95% of the total variance was captured.

A.3. GPRN

The Gaussian Process Regression Network method, submitted by the Porto team, adaptively combines GP models to jointly describe variations in the RVs and activity indicators. The structure of a *GPRN* share some similarities to an artificial neural network, with independent GPs acting as both nodes and weights. Following the work of Wilson et al. (2012), a GPRN can model a function $\mathbf{y}(\mathbf{x})$ as

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\mathbf{f}(\mathbf{x}) + \sigma_y \mathbf{z}(x). \quad (\text{A3})$$

On this network $\mathbf{f}(\mathbf{x})$ and $\mathbf{W}(\mathbf{x})$ are independent GPs,

$$\begin{aligned} f_j(x) &\sim \mathcal{GP}(0, k_f) \text{ for } j = 1, \dots, q, \\ W_{ij}(x) &\sim \mathcal{GP}(0, k_w) \text{ for } i = 1, \dots, p \text{ and } j = 1, \dots, q. \end{aligned} \quad (\text{A4})$$

This framework is capable of accommodating input dependent signal and noise correlations between multiple output variables, input dependent length-scales and amplitudes, and leads to heavy tailed predictive distributions.

The method requires RVs and activity indicators as inputs, where each RV measurement must have a corresponding activity indicator taken at the same time stamp. For instance, non-simultaneous photometry could not be used as an indicator. The number of nodes and weights, as well as the associated co-variance functions, can be decided a priori or a posteriori based on marginal likelihood comparison.

In principle, each one of the GPs that form a node or weight of the regression network has its own set of associated hyper-parameters and respective priors. However, it is possible to share hyper-parameters to reduce the number of free parameters, for example between the GPs acting as weights. For the results presented here, only one node was assumed defined by a GP with a quasi-periodic co-variance function. GPs with squared-exponential kernels were used for the weights with no shared hyper-parameters.

B. IN-DEPTH DESCRIPTIONS OF METHODS THAT USE THE CCF AS INPUT

B.1. SCALPELS

SCALPELS, submitted by the Andrews and PennState teams, makes use of autocorrelation functions to separate out Doppler shifts from shape changes that arise due to stellar signals (Collier Cameron et al. 2020). The autocorrelation function of either the spectra itself or its CCF can be used. In the velocity domain, the autocorrelation function is invariant to translation. Projecting the measured velocity time series onto the principal components of the autocorrelation function isolates shape changes. Because it is translationally invariant, these projected perturbations can be subtracted from the original velocities with the dynamical shifts preserved.

⁵ *FDPCA* was implemented with the following python packages: Flatiron Institute's `finufft` for non-uniform FFTs, `sklearn.preprocessing.StandardScaler` for scaling Fourier series to have unit variance, `sklearn.decomposition.PCA` for the PCA, and `sklearn.linear_model.LinearRegression` for the linear regression.

Applying the method requires either the spectra or the CCF to derive the autocorrelation function as well as the barycentric corrected time stamps, RVs, and RV errors for each observation. From this, *SCALPELS* will output velocity variations that are driven by shape changes. Subtracting out these shape-driven velocities leaves the true dynamical shifts preserved.

Since *SCALPELS* operates in the wavelength-domain, it does not require any information about the star’s behavior (i.e., rotation rate, pulsation timescale, etc.) nor does it need very dense sampling of the stellar rotation cycle. Ideally, there should be at least 40 observations of a target over a full range of stellar activity states. Observations taken at different activity states help the PCA of the autocorrelation function identify variations due to shape changes.

All *SCALPELS* results presented here use the autocorrelation function of the provided EXPRES CCFs. Results can vary with number of principal components incorporated. The submissions given here used two principal components to minimize the risk of over-fitting.

The PCA results from *SCALPELS* were also input into *GLOM*, where the amplitudes of the principal components, i.e., the magnitude of the shape variation modeled in the CCF, were treated as activity indicators and modeled along with the RV shifts. This process was run using the sum of two Matérn $\frac{5}{2}$ kernels for the latent GP model.

B.2. CCF Prime

The *CCF Prime* method, submitted by the OxBridGen team, is an exploratory approach to decomposing the CCF by linearly modeling variations in each spectra’s CCF. A reference CCF is first constructed by modeling the mean CCF of all observations using a GP with a square-exponential kernel. Let this reference CCF be denoted by $C(v)$ where v are the velocities at which the CCF is sampled. The quotient of each CCF against this reference CCF is then linearly modeled.

Let $c_i(v)$ denote the quotients of each CCF against the reference CCF, i.e., $c_i(v) = \frac{C_i(v)}{C(v)}$, where i indexes over all exposures and $C_i(v)$ is the CCF for exposure i . The linear model is then defined by the following equation

$$c_i(v) = a_i + \sum_{k=0}^3 b_{ik} C^{(k)}(v) \quad (\text{B5})$$

where k corresponds to the different derivatives of $C(v)$ with respect to velocity. In this case, $C^{(0)}(v) = C(v)$. The parameters a_i and b_{ik} are the linear parameters of the model.

The first derivative term in equation B5 is sensitive to shift-induced variations on the CCF. The second derivative and higher picks up on shape distortions instead. In this way, decomposing the CCF variations into different terms separates out changes due to dynamic shifts versus changes due to differences in shape. Using only derivatives of two or higher to recreate the time series will result in CCFs with only the variations due to shape changes. The effects of these shape changes can then be removed from the time series. The coefficients of the derivative terms (i.e., $k \geq 2$) can also be used as activity indicators, as they reflect the magnitude of CCF variations due to changes in shape.

This method is conceptually similar to the *SCALPELS* method described in Section B.1. In this framework, the quotients ($c_i(v)$) of each observation’s CCF over a reference CCF is modeled whereas in *SCALPELS* the autocorrelation function of the CCF or spectrum is used. For *SCALPELS*, the autocorrelation function is intrinsically insensitive to transitional shifts. For *CCF Prime*, the higher-order (> 2) derivatives are insensitive to transitional shifts. These higher-order derivatives and their coefficients in the linear model capture the variation in the CCF and the magnitude of the variation, much as PCA does for *SCALPELS*. The coefficients of the linear model can also act as an activity indicator (much as the amplitudes from the PCA are used for *SCALPELS+GLOM*). As the *CCF Prime* method remains exploratory, more work needs to be done to establish whether the different derivatives create an orthonormal basis as PCA does.

The *CCF Prime* method requires only normalized CCFs and is straightforward to implement. Higher resolution data will contain more information on the line profile distortions being modeled. Higher SNR observations will give more accurate derivatives. The observations should sample a broad range of activity states. This ensures that changes in the CCF due to stellar signals are not reflected in the combined, reference CCF. With many different manifestations of stellar signals in the range of CCFs, the specific features of any given activity state will be blurred out.

B.3. FIESTA

The *FIESTA* method, submitted by the PennState team, decomposes the CCF of a spectrum into Fourier basis functions (Zhao & Tinney 2020, Zhao et al., in prep.). The shifts of each of these basis functions are then calculated for a range of Fourier frequencies. A pure CCF shift will manifest as a constant shift in all Fourier frequencies and can

easily be subtracted out. Shape deformations, however, will be frequency dependent. This decomposition therefore parameterizes the effects of stellar signals as a series of shifts at each frequency for each CCF. These frequency-dependent shifts can be used together as a multi-dimensional activity indicator.

The *FIESTA* method reads in CCFs for each observation. These CCFs must be properly normalized as a vertical offset could also produce a frequency-dependent shift that would be mistaken for a shape deformation. Observations with greater SNR allow for more frequencies to be incorporated.

The activity indicators produced by *FIESTA* were post-processed using principal component analysis (Zhao et al. 2021, in prep.) and modeled jointly with dynamical RV shifts using GLOM (as described in Section A.1).

B.4. CCF Linear Regression

The CCF Linear Regression method, submitted by the ML_EPRVs team, makes use of machine learning to model variations in the CCF that are expected to be due to stellar signals (de Beurs et al. 2020). Specifically, the machine learning model predicts the difference between a Gaussian fit to the CCF and the true velocity shift. This prediction can then be subtracted from the input RVs to give corrected RVs.

This method requires CCFs for each exposure and best fit RVs. The CCFs are first shifted by the best-fit RVs so there are no translational difference between the different CCFs. This allows the model to instead focus on shape variations. The model is fed differential CCFs, i.e., the residuals from subtracting a reference CCF (made by taking the median of all CCFs) from each CCF. These differential CCFs are normalized by the median and standard deviation of each point in the CCF across all observations. This normalization serves to scale the variations so they are roughly equal 

In order to reduce the complexity of the model, only about four to six locations across the residual CCFs are modeled using a linear regression model. The more observations there are, the more locations can be used without the risk of over fitting. The base model for a single CCF and associated RV is given by:

$$RV = w_1 \cdot CCF_1 + w_2 \cdot CCF_2 + \cdots + w_v \cdot CCF_v \quad (B6)$$

where CCF_v is the value of the differential CCF at velocity v and w_v is the associated weight parameter that is fit for.

Two slightly more complicated models were also tested. For all targets, H α information was added to the model to give:

$$RV = w_1 \cdot CCF_1 + w_2 \cdot CCF_2 + \cdots + w_v \cdot CCF_v + b \cdot H\alpha \quad (B7)$$

where H α is the derived H α emission for the given exposure and b is the associated weight that is fit for like the w_v weights are. For HD 26965, a fitted Keplerian was also added with a fitted weight parameter d as follows:

$$RV = w_1 \cdot CCF_1 + w_2 \cdot CCF_2 + \cdots + w_v \cdot CCF_v + b \cdot H\alpha + d \cdot Keplerian \quad (B8)$$

 This CCF Linear Regression method does not use timing information. Though it benefits from more observations, the cadence of these observations does not matter. More observations allow for more locations in the differential CCFs to be included in the model, allowing it to potentially pick up on more shape variations. The method can be sensitive to choice of location across the differential CCFs, which require some fine-tuning.

C. IN-DEPTH DESCRIPTIONS OF LINE-BY-LINE METHODS

C.1. Generating CCF Masks with VALD Input (CCF Mask-VALD)

The PennState team generated and tested a series of different CCF masks. These masks aim to generate cleaner CCFs by mitigating the effects of variable lines, blended lines, telluric contamination, and lines strongly affected by stellar variability and activity.

In generating these CCF masks, first an automatic line-fitting code finds all spectral lines and fits them to a Gaussian with a linear offset. Fitted line depths are used as mask weights for each line. Any spectral line with a line center falling within 30 km s $^{-1}$ of features in the provided SELENITE telluric model were removed.

A line list from the Vienna Atomic Line Database (VALD) is used to vet lines too near each other in order to avoid line blends. For each target, an optimal definition of “too near” was empirically determined, where any lines with centers closer than a given line blend cutoff were removed. Cutoffs ranging between 0 to 27 km s $^{-1}$ in intervals of 3 km s $^{-1}$ were tested.

Masks used a Gaussian window function. Different mask widths were tried where the sigma of the Gaussian window function ranged from one to eight pixels. The optimal mask window width and line blend cutoff was decided by the combination that gave the lowest resultant RV RMS.

Generating these masks requires the spectra along with a telluric model. The approximate RV shift of each spectra as well as the expected line velocity width makes line-fitting easier. The target star's stellar temperature and $\log g$ are needed for the VALD line list.

C.2. CCF Masks with Less Activity Sensitive Lines(CBC Mask-BIS and CCF Mask-RV)

The Warwick team constructed weighted, binary masks to remove the contributions from blended lines or lines particularly sensitive to stellar signals (Lafarga et al. 2020). Spectral lines are found by identifying relative minima in a high SNR stellar template built by coadding the EXPRES observations and parametrized by fitting a Gaussian function. This gives an initial line list with rest wavelengths for all lines. Only lines with widths, depths, and asymmetry that fall between a specified range (as specified in Lafarga et al. (2020)) are kept. This ensures that the included lines are clear, sharp lines with no obvious blends. The provided SELENITE telluric model is used to vet for any lines too near a telluric feature.

RVs are then computed for each individual line in each of the observations. Each line is fit to a Gaussian. The mean of this Gaussian is taken to be the line center, which is then compared to the initial line list. Lines are determined to be either sensitive or insensitive to photospheric velocities based on how correlated they are with a given activity indicator. The Pearson correlation coefficient is used to gauge the degree of correlation. Lines were established as inactive if they had a coefficient less than 0.2-0.4 and spread in RVs less than $10\text{-}15 \text{ m s}^{-1}$ (with the specific cutoff depending on the target). Active lines had correlation coefficients greater than 0.3-0.5 with RVs or a correlation coefficient less than or equal to -0.3 in the case of the BIS-guided mask.

Very correlated lines are likely to be strongly affected by stellar signals. If a line's RVs exhibit a lot of scatter, it becomes difficult to tell whether a line is truly uncorrelated with an activity indicator, or if the correlation is merely lost among the scatter. Therefore, lines that exhibit a large RV scatter are also discarded. The remaining lines that exhibit little to no correlation with activity indicators are averaged to compute a final RV for each exposure.

The results presented in this report used either the CCF BIS (CCF Mask-BIS) or the CCF RV (CCF Mask-RV) as an indicator to establish what lines are strongly correlated with stellar signals. Note, the CCF RV and individual line RV are not fully independent, which could bias the correlations measured. Other than choice of indicator, there is no specific tuning required for this method.

For this method, the data must be high enough resolution to resolve line blends. The data should also be stable enough that the dominate variations in lines are due to stellar signals and not instrumental or other non-astrophysical effects. More observations, especially over a greater range of activity states, will result in a better measure of correlation.

C.3. Spectral Cleaning, Line-by-Line RVs, and PCA (LBL RVs+)

The Geneva team used a combination of spectral cleaning techniques and line-by-line RVs. The provided spectra were first continuum normalized using *RASSINE*, an open source python package that makes use of convex hulls to determine continuum points (Cretignier et al. 2020b). YARARA was then used to clean the spectra of tellurics and first-order morphological variations away from a median spectra (Cretignier et al. 2021). Using this post-processed spectra, a master spectrum and tailored stellar mask (to avoid line blends) was developed for each star.

Line-by-line RVs were extracted, where RVs for each spectral line are derived relative to the star-specific master spectrum (Dumusque 2018). The degree to which lines are affected by stellar signals or observational systematics varies from line to line, as reflected in the spread of each line-specific RV across all observations.

PCA was then used to identify variations across all lines in all observations, where each observation has been corrected by its average RV. The first three principal components are used to decorrelate the average RV signal for each observation using a multi-linear regression.

This method is run using merged spectra, where all echelle orders of a spectrum have been merged to form one, long spectrum. The basic method described here requires little tweaking to run, but implementing YARARA can get increasingly more complex if it is used to do a more tailored job of removing instrumental systematics. Because each line now stands alone, this analysis does require higher SNR spectra in comparison with a classic CCF. In order to use YARARA to disentangle telluric features, the input set of observations must have a good coverage of different barycentric shifts in order to separate the stellar lines from the telluric lines. For best performance from the PCA, it is ideal if the observations also cover a wide range of stellar activity states.

This method outputs RVs for every line as well as the principal, non-Dopplerian variations in these RVs from the PCA. The PCA here is run directly on the line RVs or the spectra itself rather than chromospheric proxies, such as more classic activity indicators. The PCA step might be swayed by outliers or the presence of large variation, e.g., hardware changes, abnormal observing conditions, etc. By using the whole spectrum and treating each line independently, LBL RVs reveal how individual lines are affected by variations from either stellar signals or instrument systematics. This gives a better picture of how these affects are manifesting in the spectra.

There are three flavors of LBL results presented here. The *SHELL* LBL RV results uses PCA at the spectral level while the *YARARA* LBL results implement PCA on the recovered RVs for each line. Both methods can also be combined by first applying the PCA decomposition to the spectra, extracting LBL RVs from that spectra, and then decomposing the resultant LBL RVs with another PCA. These results are included as *SHELL+YARARA* LBL RVs.

C.4. PWGP

The Pairwise Gaussian Process RV Extraction method, submitted by the OxBridGen team, uses GPs to model and then align all pairs of spectra with each other (Rajpaul et al. 2020). These pairwise RVs can then be combined to establish differential stellar RVs without having to construct a master template. The pairwise matching is done on a highly localized basis—i.e., each spectra is broken up into many different “chunks” with each chunk containing one to a few spectral features.

These smaller chunks can then be treated as independent measures of the spectral shift, where some chunks will be more affected by stellar variability than others. More sophisticated implementations are possible, for example modifying the GP modeling of spectral chunks to model stellar variability in addition to Doppler shifts. For the results presented in this report, spectral chunks that appeared “contaminated” by stellar variability were simply not used when computing final RVs.

The PWGP method reads in spectra. A Matérn $\frac{5}{2}$ kernel is used to model and align each chunk, with different hyper-parameters returned for each chunk. This can get quite computationally expensive, but is helped by the pairwise framework. Though the method requires little tuning to run, some thought must go into deciding which chunks are considered “contaminated” and what to do with them.

There are many possible metrics to use in determining which chunks appear to be contaminated. The chunk itself may exhibit unusually large variation from one exposure to another, suggesting there are stellar signals or tellurics present in the chunk that is causing it to return such a large range of RV measurements. Similarly, the RV error of a chunk may be higher than typical. The RVs of a chunk may also show statistically significant correlation with an activity indicator, suggesting the RV from that chunk is mostly due to stellar signals rather than true dynamical shifts.

Tuning the cut offs for which chunks to include requires balancing between the RMS of the final RVs and the error bars on these measurements. Removing too many chunks will exclude too much data from the process, thereby increasing the error bars for each RV measurement. Not removing enough chunks means noise will continue to be incorporated into the final RV measurements, thereby resulting in greater RV scatter.

After cutting contaminated chunks, the RV measurements of the remaining chunks are combined to recover final RVs. The RV from each chunk is inversely weighted by the scatter in returned RVs for that chunk as determined via a Markov chain Monte Carlo (MCMC) analysis. By using a MCMC, the resultant weight incorporates both the photon noise and uncertainty from the GP fit.

Using GP modeling to align spectra should perform better (as compared to non-GP models) with lower-resolution and lower-SNR spectra. However, having higher SNR/resolution spectra is needed in the following step of identifying contamination.

This method benefits from using a principled, GP modeling framework for spectral interpolation and alignment. This precludes the need to generate a master template and indeed does not require any information about where lines are, what they may look like (i.e., depth, width, etc.), or how they might change with stellar signals. On the other hand, the model also can not incorporate any prior knowledge of stellar or telluric contamination and does not distinguish between different forms of contamination whether stellar, terrestrial, or instrumental.

D. IN-DEPTH DESCRIPTIONS OF METHODS THAT MODEL THE SPECTRA

D.1. DCPCA

The Doppler-Constrained Principal Components Analysis method, submitted by the PennState team, identifies the largest variations in RV shifted spectral data using PCA (Jones et al. 2017). The resultant principal components

highlight where the spectra is changing the most while the corresponding amplitudes of each principal component captures the magnitude of this change for each observation. By feeding the PCA the full spectral format, the PCA is able to pick up on changes at the pixel level. The principal component amplitudes can be used as an activity indicator.

The *DCPCA* method requires spectra and initial guess RVs for each observation. Since the observations were provided with wavelengths in a barycentric frame and there were no hot-Jupiters present, a constant was used for the initial guess. The spectra were first shifted by the best-fit RV for each observation and then interpolated onto a common wavelength grid using a GP with a Matérn $\frac{5}{2}$ kernel. Some tuning of what parts of the spectra to include in the PCA will help ensure the PCA is not picking up on variations from the instrument or tellurics. While the method can be run on the full spectrum, the results reported here used the areas of spectra near lines specified in different CCF masks (see Section C.1 for a discussion of the diffrent masks). This helps to avoid telluric contamination and blended lines.

The number of principal components to incorporate into the analysis can be chosen in a number of ways. As always, only principal components with significant features (i.e., are not only noise) should be used. With enough exposures, a classic cross-validation test can be used to gauge the performance of incorporating different numbers of components. More observations will likely result in more significant components. A component can also be tied to photospheric velocities if the amplitudes of the component are correlated with activity indicators.

Data with a high SNR and high resolution makes variations in the spectra clearer. A broad wavelength coverage would also help, as it would encompass more changes.

For the results presented in this report, the amplitudes of the first two principal components were used as indicators. The publically available ESPRESSO masks (also used to generate the provided CCFS) were used to determine which segments of the spectra were fed into the PCA. The RVs were decorrelated against the resultant principal component amplitudes via both a simple linear regression and using the *GLOM* framework with the sum of two Matérn $\frac{5}{2}$ kernels.

D.2. ResRegGen

ResRegGen, submitted by the CCA team, takes the residuals of each observed spectrum against a Doppler-shifted template spectrum and regresses these residuals against housekeeping data, such as the derived RVs, activity indicators, or instrumental measurements. *ResRegGen* operates under a generative framework, as in it constructs a model using a finite number of housekeeping data sets to predict what the residuals will look like. In doing so, *ResRegGen* establishes what properties of the residuals can be tied to the different effects being traced by the housekeeping data, be it stellar signals, instrument systematics, or whatever else it is given. The effects not due to an orbiting planet can then be removed.

For N observations, let F represent all residuals for each pixel of a spectra to a model while Q represents all housekeeping data being used including the RV. Then we use Δf_n to denote the residuals of a given observation n and \hat{q}_n to represent the predicted RV correction for that observation. For a realistic test, for each observation n or validation set, the Δf_n residuals should be left out of F . RV corrections can then be calculated as follows:

$$\hat{q}_n = \Delta f_n \frac{dF}{dQ} \cdot \left[\frac{dF}{dQ} \cdot \frac{dF}{dQ} \right]^{-1} \quad (\text{D9})$$

where $\frac{dF}{dQ}$ represents the spectral residuals being regressed against the housekeeping data. This is a first-order regression model. The housekeeping data can vary depending on what is needed to give a complete, orthogonal representation of the variations being modeled.

Implementing this method requires spectra of each observation and housekeeping data associated with each spectra. The template spectrum can be generated in any number of ways. Higher resolution spectra will preserve more evidence of stellar variability in the residuals. The regression itself is trivial to implement.

By incorporating all the spectral residuals, *ResRegGen* is able to incorporate information from every pixel of the spectral data. The housekeeping data is then used to try and predict the behavior of different pixels. Incorporating more data that traces different effects makes the method more sensitive to different causes of spectral variations. On the flip side, the method is also incapable of tracing any variation not associated with the provided housekeeping data. The regression will be poorly constrained if the data used are not all independent and affect a real change on the residuals being modeled.

The method outputs the magnitude of change to the RVs expected from these variations, though it is not straight-forward to interpret exactly what variations the regression is picking up on.

For the results presented in this report, a model spectrum was generated using *wobble*, a data-driven method for extracting RVs by inferring the underlying spectral components (Bedell et al. 2019). The CBC RVs and H α equivalent width are used as housekeeping data. Expected RV offsets are calculated using a cross-validation framework where an eighth of the data at a time, where that eighth is left out of the model construction. For reference, the results where all data is used is given as *ResRegGen* self results. For both the self and cross-validation frameworks, all observations are used to construct the model spectrum with *wobble*.

D.3. *ResRegDis*

ResRegDis, submitted by the CCA team, is similar to *ResRegGen* and also regresses spectral residuals to a shifted template against housekeeping data; *ResRegDis*, however, operates under a discriminative framework as opposed to the generative framework with *ResRegGen*. With a discriminative framework, *ResRegDis* uses the residuals to predict the housekeeping data. The result is a prediction of the magnitude of RV shift due to observation-specific spectral variations (as captured in the residuals to a spectral model).

As with *ResRegGen*, let F represent the array of all spectral residuals, Δf_n the residuals for a given observation n , and Q be the array of RVs acting as labels. The predicted RV correction for each observation, \hat{q}_n , can then be calculated

$$\hat{q}_n = \Delta f_n \cdot (F^T F + \alpha I)^{-1} F^T Q \quad (\text{D10})$$

where α represents an opportunity to introduce expected amount of information, for example uncertainties on the spectral residuals or spectral resolution.

The inputs, implementation, and output for the *ResRegDis* method is the same as for the *ResRegGen* method described above. After acquiring the residuals to a template spectra and associated RVs for each spectra, the method takes seconds to run. The only housekeeping data used for *ResRegDis* are the CBC RVs for each exposure.

The discriminative framework is more agnostic about precisely what housekeeping data is included. The regression itself works to construct an orthogonal transformation that can be mapped onto the derived RVs. This framework is more appropriate in the regime where the spectra is varying in more ways than can be captured by the provided housekeeping data. Since it is not clear whether known activity indicators trace all possible spectral variations due to stellar signals, the discriminative framework may be more appropriate for disentangling photospheric velocities from true center-of-mass shifts.

In truth, there is a latent model that produces both the housekeeping data and the spectral variations, namely the activity and intrinsic variability of the target stars. Both the generative and discriminative frameworks move between the products of this latent model, just in different directions. Both the *ResRegGen* and *ResRegDis* methods are ongoing work; the results presented here are an initial implementation of the two methods.

E. SUBMITTED RVs OF ALL METHODS

The following section show the submitted RVs, both clean and activity RVs where available, as well as their periodograms. Given the large nature of the figures, their content is described here in the text.

The top-left plot shows the originally provided EXPRES RVs (first column) along with the periodogram (second column) in black and the periodogram of the time sampling in green. The rest of the rows show the submitted clean RVs in blue. Each figure is labeled by the team and method name.

The periodogram subplots for each method shows a periodogram of the clean RV in blue. If provided, the periodogram of submitted activity RVs are also shown in orange. A significance level of p-value = 0.01 is shown as a horizontal, black line across the periodograms. A p-value of 0.1 is shown as a dashed black line. Axes with the words “No Submission” are shown for methods still awaiting submissions.

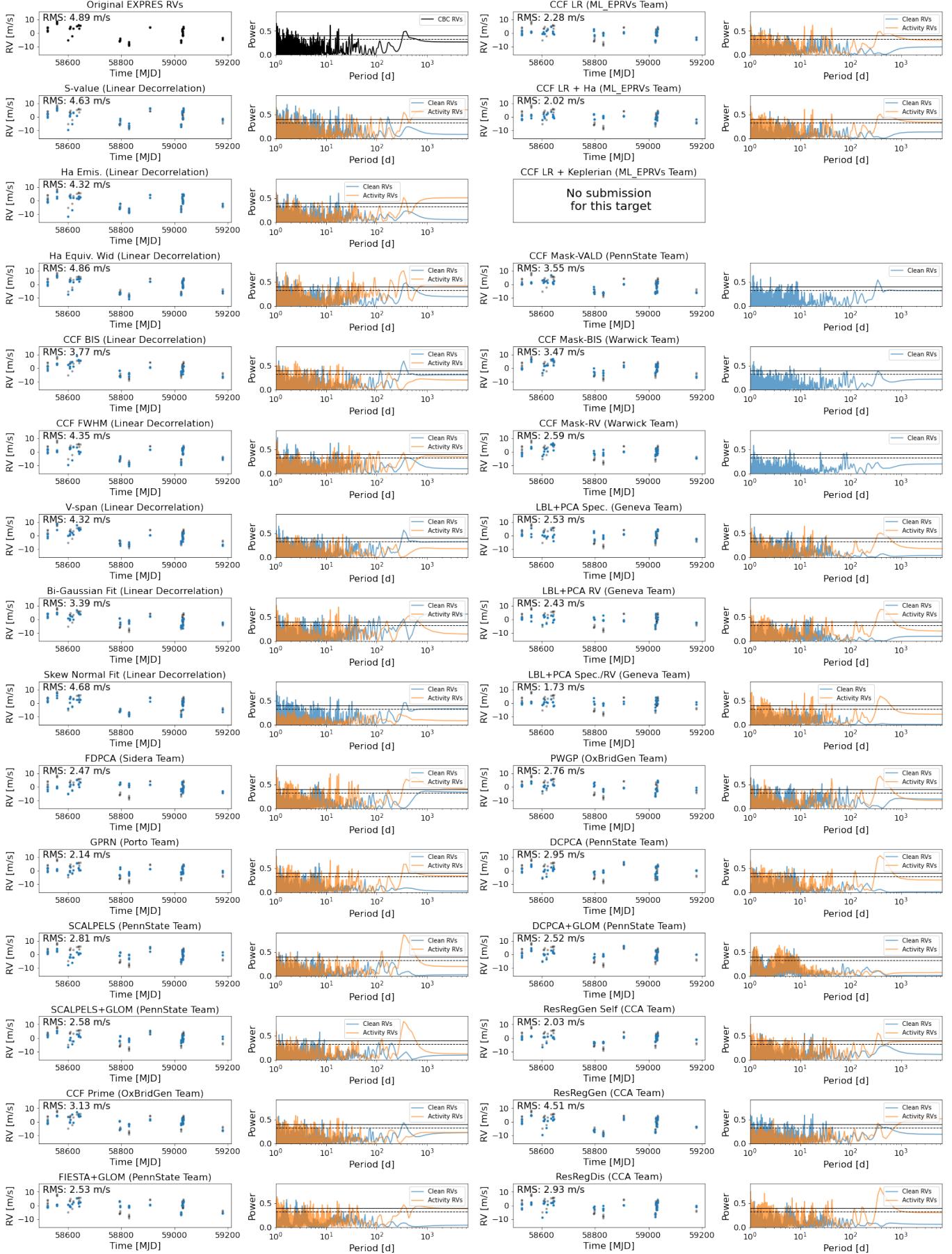


Figure 8. Submitted results for HD 101501. For each periodogram, p-values of 0.01 and 0.001 are shown as horizontal solid and dashed black lines respectively.

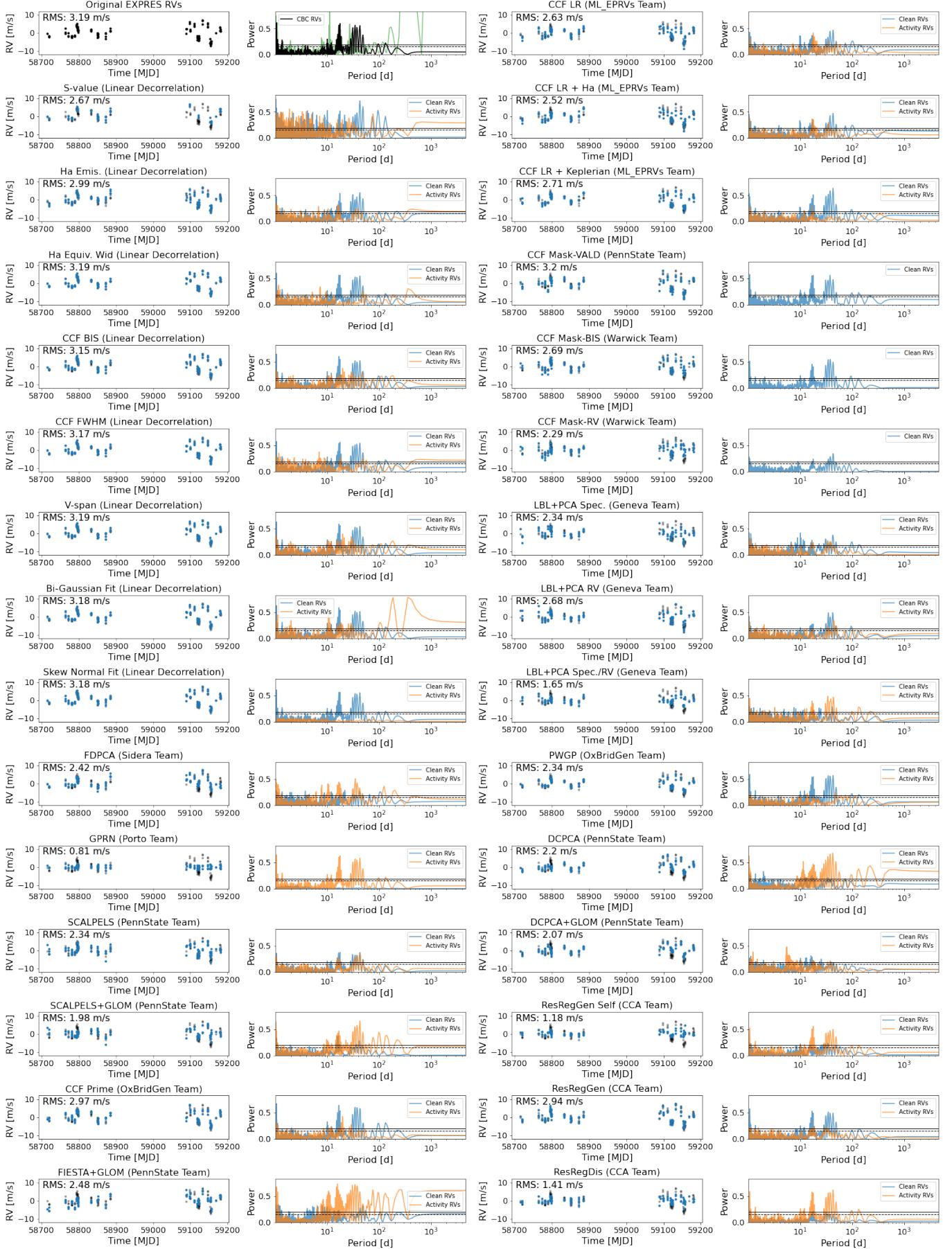


Figure 9. Submitted results for HD 26965.

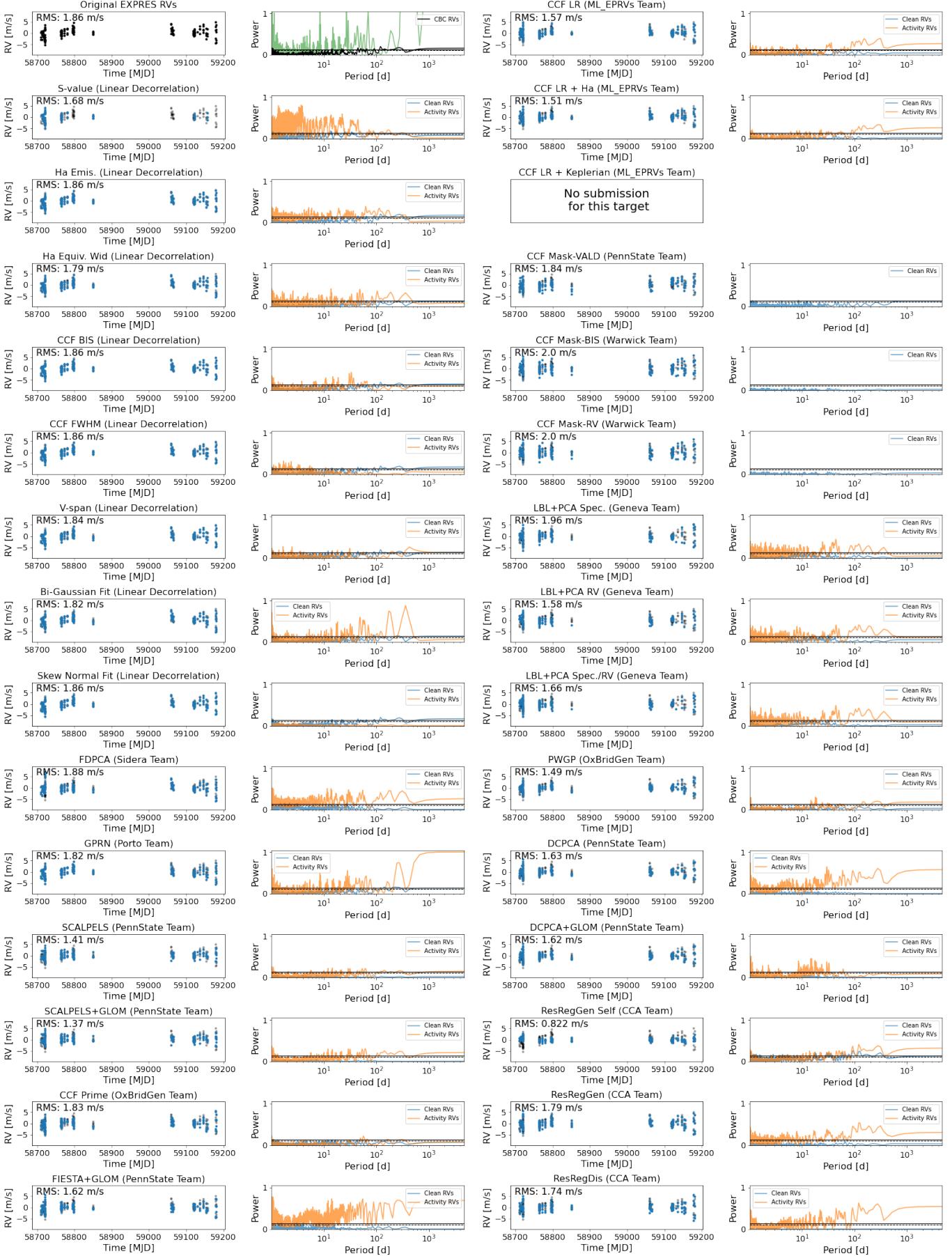


Figure 10. Submitted results for HD 10700.

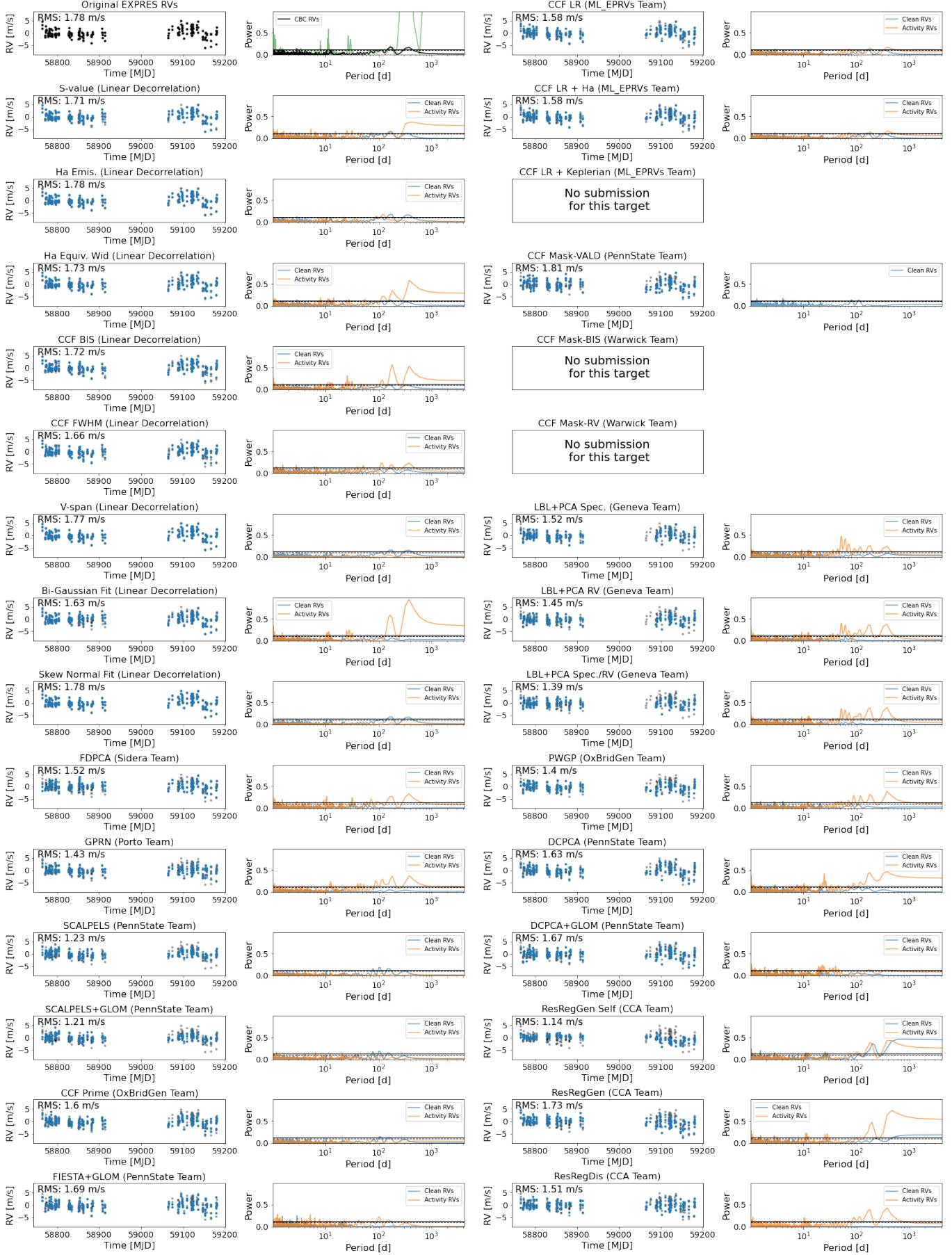


Figure 11. Submitted results for HD 34411.