Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Yale Graduate School of Arts and Sciences Dissertations

Fall 10-1-2021

The Path to Extreme Precision Radial Velocity With EXPRES

Lily Ling Zhao
Yale University Graduate School of Arts and Sciences, lily27zhao@gmail.com

Follow this and additional works at: https://elischolar.library.yale.edu/gsas_dissertations

Recommended Citation

Zhao, Lily Ling, "The Path to Extreme Precision Radial Velocity With EXPRES" (2021). *Yale Graduate School of Arts and Sciences Dissertations*. 446.
https://elischolar.library.yale.edu/gsas_dissertations/446

This Dissertation is brought to you for free and open access by EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Yale Graduate School of Arts and Sciences Dissertations by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

Abstract

The Path to Extreme Precision Radial Velocity With EXPRES

Lily Ling Zhao

2021

The field of exoplanets is currently poised to benefit hugely from improved radial velocity (RV) precision. Extreme precision radial-velocity (EPRV) measurements, capable of detecting planetary signals on the order of 10-30 cm s^{-1}, will deliver integral planetary parameters, be sensitive to a missing category of lower-mass planets, grant a deeper understanding of multi-planet architectures, and support both current and future space missions such as *TESS* and *JWST*. The ability of EPRV to deliver mass estimates is essential for comprehensively characterizing planets, understanding formation histories, and interpreting atmospheric spectra.

Until recently, RV precision had stalled at around 1 m s^{-1}, i.e. signals with a semi-amplitude of less than 1 m s^{-1}could not be faithfully detected. We demonstrate with *HARPS*, *UVES*, and *CHIRON* observations of α Cen the need for better data, not just more data. Even with over a decade of observations at around 1 m s^{-1} precision, large areas of mass/period parameter space remained unprobed. Higher-fidelity data is needed to significantly push down detection limits.

EXPRES, the EXtreme PREcision Spectrograph, was one of the first next-generation spectrographs to go on sky. Installed at the 4.3-m Lowell Discovery Telescope in 2017 and commissioned through 2019, *EXPRES* is a fiber-fed, ultra-stabilized, echelle spectrograph with a high median resolving power of $R \sim 137,000$ and an instrument calibration stability of 4-7 cm s^{-1}, a factor of 10 better than previous instruments.

The stringent requirements of EPRV measurements along with the stability of *EXPRES* and similar instruments changes how we must extract, calibrate, and model the resultant spectral data. This dissertation discusses the work that must be done in this new

regime in terms of data pipelines and modeling stellar signals and showcases some initial progress. We present *EXPRES*' data pipeline, a new data-driven method for wavelength calibration, and the current state of the field for disentangling stellar signals.

The *EXPRES* extraction pipeline implements a flat-relative, optimal extraction model and *excalibur* for wavelength calibration. *Excalibur* is a hierarchical, non-parametric method for wavelength calibration developed as part of this thesis work. Calibration line-positions are de-noised by using all calibration images to construct a model of the accessible calibration space of the instrument. This denoising returns wavelengths a factor of five more precise than previous polynomial-based methods. With *EXPRES* data, *excalibur* reduced the overall RMS of RV data sets for all targets tested by 0.2-0.5 m s^{-1}. This consistent reduction in overall RMS implies that *excalibur* is addressing an instrumental, red-noise component that would otherwise permeate all exposures.

With instrumental noise lowered and extraction error reduced, intrinsic stellar variability and the resulting apparent RVs now dominate the error budget for EPRV measurements. The *EXPRES* Stellar Signals Project (*ESSP*) released high-fidelity, spectroscopic data from *EXPRES* and photometric observations from the automatic photoelectric telescopes (*APTs*) for four different stars. This allowed for a self-consistent comparison of the 19 different methods submitted, which represent the current state of the field in disentangling stellar signals. The analysis of results is ongoing work.

Currently, the best performing method give a final RV RMS of 1.2 m s^{-1}. Submitted methods nearly always do better than classic methods of decorrelating RVs from stellar signals. We found that methods returning the lowest RV RMS often used the full spectra and/or flexible statistical models such as Gaussian processes or principal component analysis. However, there was a concerning lack of agreement between methods. If we hope to improve on current advancements and develop methods achieving sub-meter-per-second RMS, we must introduce more interpretability to methods to understand what is and is not working. A densely sampled, high-resolution data sensitive to all categories of stellar variation is needed to understand all types of stellar signals.

This dissertation work centers on the question of achieving EPRV capabilities for detecting planets incurring reflex velocity signals on the order of 10-30 cm s^{-1}. We consider what needs to be done, describe current development towards this goal, and discuss the future work that remains before sub-meter-per-second precision can become a regular reality. We emphasize the power of data-driven pipelines to account for variations in data for EPRV applications and beyond. Empirically backed conclusions for mitigating photospheric velocities are summarized from the results of the *ESSP* along with next steps and additional data requirements. Progress is being made, but there remains much work to be done.

The Path to Extreme Precision Radial Velocity With EXPRES

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Lily Ling Zhao

Dissertation Director: Debra Ann Fischer

December 2021

Copyright © 2021 by Lily Ling Zhao

All rights reserved.

Acknowledgments

Thank you to DAF for being a tremendous advisor. If DAF were merely a good or even a great advisor, I simply would not have made it to the end. Their intelligence and ability to always ask the right question kept me searching for answers. Their high standard of rigor kept me confident and proud of the answers we found together. Their compassion and humanity helped me hope I could one day be as cool as them. I will keep trying.

This work is primarily the result of the heroic efforts of the *EXPRES* team. Thank goodness for DAF, JMB, AES, JL, CAJ, TMM, DS, RTB, ABD, RMR, and SHCC.

Thank you to my thesis committee for lending their time and expertise towards improving this work. Thank you to MG for the sincerity and the confidence. Thank you to GPL for the lunches and the hustling. Thank you to DWH for the whacky ideas and the smooth jazz. Thank you to MEB for the chai and the wags.

Thank you to the National Science Foundation Graduate Research Fellowship for the funding, the opportunities, and the conceit.

Thank you to SCM for being the best person I know. It is completely inappropriate how often I brag to people about having been her office mate, but it is only because I loved being her office mate so very much. Thank you to NS for the bowls. Thank you to DZS for the wings. Thank you to RMR for the frozen treats. Thank you to TBM for excelling at a task that only he can fully understand how hard it is. I am not sure it will get easier, but I will say that if it makes you feel better.

Thank you to SM for your in-depth knowledge of my work and powers of predictive modeling.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

On February 17, 1600, Giordano Bruno was burned at the stake for (among many other things) suggesting that the universe is filled with stars hosting planets much like our solar system (Aquilecchia, 2021). In 2019, Michel Mayor and Didier Queloz were awarded the Nobel Prize in Physics for their discovery of the first exoplanet orbiting a solar-type star.

This first discovery, 51 Peg b, was found using the radial velocity (RV) method (Mayor & Queloz, 1995), alternatively referred to as Doppler spectroscopy or the wobble method. RVs are a measure of how Doppler shifted a star's spectrum is. A star hosting planet(s) will orbit the system's common center of mass; this movement gives rise to characteristic, periodic Doppler shifts in the star's spectrum.

The resultant RVs are a function of stellar, planetary, and orbital parameters as given by Equation 1.1 below.

$$RV = K \cdot cos(v + \omega) + e \cdot cos(\omega) \tag{1.1}$$

Here, v is the true anomaly, ω is the argument of periastron, e is the eccentricity of the orbit, and K is the semi-amplitude. K is defined in Equation 1.2 and will be described in more detail below. The true anomaly gives the angle between the direction of periapsis and the position of the orbiting body at the focus of the ellipse being orbited (see Figure 1.1). It can easily be calculated using the guiding center approximation. The argument of

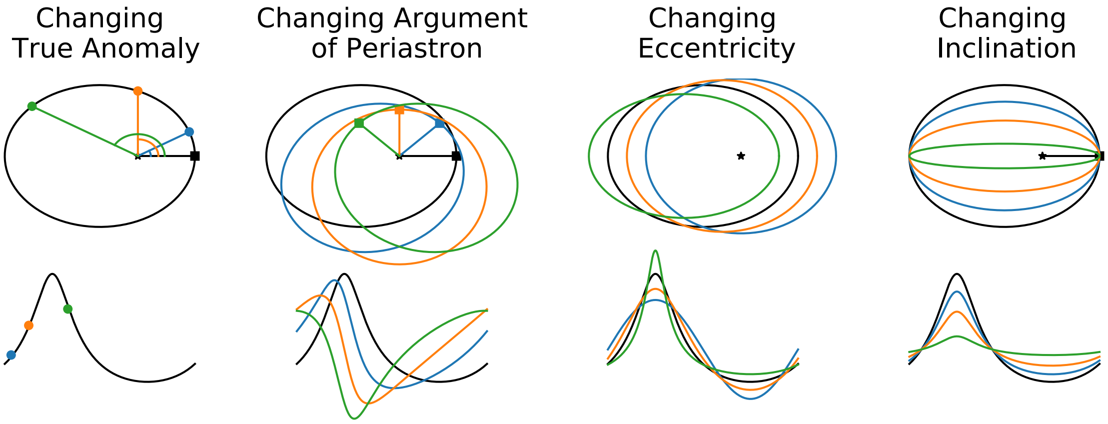

Figure 1.1: Change in orbital parameters and accompanying RV curve. Each column shows three different values for either (left to right) true anomaly v, argument of periastron ω, eccentricity e, or inclination i. The associated RV value for the different changes, if viewing the orbit from the bottom of the page, is given below each orbit in the corresponding color. Different true anomaly values trace a planet's position along its orbit. The true anomaly along with eccentricity, which describes the oblateness of the orbit, defines the shape of the orbit. Changing arguments of periastron or inclination change the orientation of the orbit but not the shape.

periastron, ω, gives the angle of the periastron point from the plane of reference.

The true anomaly describes where a planet is along its orbit and, along with the eccentricity, gives the shape of the orbit. The argument of periastron describes the orientation of this orbit. These variables therefore uniquely describe an orbiting body's position relative to the point it is orbiting. This allows us to derive a planet's orbital parameters from fitting its RV measurements.

Figure 1.1 shows an example of an orbit with different true anomaly (column 1), argument of periastron (column 2), and eccentricity (column 3) values. The accompanied RV curve for the different values, when viewing the orbit from the bottom of the page, is given below each orbit. RV measurements allow us to characterize planet orbits by deriving these three parameters from the shape of the RV time series.

The magnitude of the signal is given by K, also known as the RV semi-amplitude. It represents the largest movement the planet will invoke in its host star if on a circular orbit.

The value of K in units of m s^{-1} is given by:

$$K = 28.4329 m\, s^{-1} \frac{1}{\sqrt{1-e^2}} \cdot \frac{M_{pl}\sin(i)}{M_{jup}} \cdot \left(\frac{M_\star + M_{pl}}{M_\odot}\right)^{-2/3} \cdot \left(\frac{P}{1yr}\right)^{-1/3} \tag{1.2}$$

where M_{pl} is the mass of the planet in Jupiter masses (M_{Jup}), i is the inclination of the orbit, $M_\star$ is the mass of the host star in solar masses ($M_\odot$), and P is the period of the planet in years. The inclination of the orbit with respect to the observer is essentially degenerate with the mass of the orbiting planet. RV measurements can only pick up on movement towards and away from the observer.

The right-most column of Figure 1.1 shows the 2D project of an orbit with different inclinations. The plane of reference for the inclination angles given is perpendicular to the face of the page. An inclination of 90°, i.e. an edge-on orbit, would lie flush with the page while an inclination of 0°, i.e. a face-on orbit, would be perpendicular to the page and produce no measurable RV signal. The sine of the inclination scales the magnitude of RV signal.

Close-in, large planets around lower-mass stars have larger K values, meaning they produce a larger RV signal that will be easier to detect. Less massive, far-away planets are harder to detect with RVs. For example, 51 Peg b is a short-period planet almost half the size of Jupiter and invokes a RV signal with a semi-amplitude of 56.24 m s^{-1} (Petersburg et al., 2020). It was discovered with *ELODIE*, a spectrograph with a measurement precision of 7 m s^{-1} (Mayor & Queloz, 1995).

Earth, on the other hand, at less than one 300th the mass of Jupiter and at a distance of 1 AU, invokes a reflex velocity on the Sun with a semi-amplitude of only 0.09 m s^{-1}. Depending on the mass of the host star, Earth-like planets in the habitable zone give K values of approximately 0.1-0.3 m s^{-1}.

1.1 Building RV Precision

The RV method is still one of the two most prolific ways for discovering exoplanets. As the field has progressed, more and more planets with smaller and smaller semi-amplitudes have been discovered. Figure 1.2 plots exoplanets discovered with either transits or RVs as a function of their mass and period. Points are colored by their discovery method. Diagonal lines mark constant semi-amplitude values assuming a host-star mass of 1 $M_{\odot}$.

Until recently, RV precision had stalled at around 1 m s^{-1}. The measurements from available instruments and data analysis methods at the time were unable to faithfully measure the movement of stars due to orbiting planets with K less than 1 m s^{-1}. This limitation prevented RV surveys from detecting lower-mass or further-out planets.

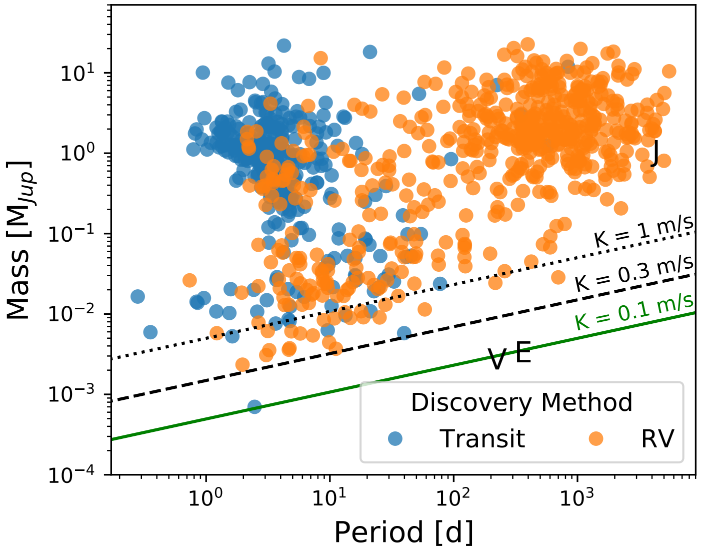

Figure 1.2: Planets discovered by transit (blue) or RVs (orange) plotted by mass and period. Planet properties, including discovery method, are pulled from `exoplanets.org` with over 98.8% of all discovered planets shown. Lines of constant semi-amplitude, K, assuming a solar-mass host star are shown for K = 1 m s^{-1} (black, dotted), K = 0.3 m s^{-1} (black, dashed), and K = 0.1 m s^{-1} (green, solid). Note, RV points < K = 1 m s^{-1} line orbit stars less than one solar mass. Venus, Earth, and Jupiter are also plotted as V, E, and J respectively.

This is further demonstrated in Chapter 2 using the α Centauri star system as an example. Even decades of observations at the 1 m s^{-1} level leave large areas of mass/period parameter space un-probed (Zhao et al., 2018). Additional observations of the same quality will add little. It will take higher-fidelity data to significantly push down detection limits.

We know from *Kepler* and other survey missions that nearly all stars have one or more exoplanets, with Earth-size planets among the most common (Borucki et al., 2011; Batalha, 2014). Of Sun-like stars, 10-25% could play host to Earth-sized worlds orbiting in the habitable zone (Petigura et al., 2013; Burke et al., 2015; Christiansen et al., 2016; Fulton et al., 2017). Lower-mass planets are abundant in the galaxy, but they are missing from RV surveys because the measurements were not sensitive enough to measure these smaller-amplitude signals.

The radial velocity method remains the primary way to measure mass estimates for exoplanets, which is needed in characterizing planets. The transit method, which measures a dip in light as a planet orbits between its host star and the telescope, imposes strict limitations on the orbital alignment of the planet with respect to the observer. Transit timing variations from planet-planet interactions may provide rough estimates of a planet's mass, but will only work for transiting planets in multi-planet systems. Astrometric measurements could deliver mass estimates similar to RVs, but must be done with space-based instruments and has yet to yield any confirmed planets. Without more precise RV measurements, we are incapable of deriving mass estimates and gaining a deeper understanding of the most common and easily most popular population of planets.

The push for extreme-precision radial-velocity (EPRV) capabilities is being driven by this need for better RV measurements to characterize planets. EPRV work centers on the goal of breaking the 1 m s^{-1} precision boundary, striving instead to achieve the 10-30 cm s^{-1} precision needed to detect Earth-like planets.

1.2 The Power of EPRV

Furthering EPRV capabilities will impact all areas of exoplanetary science. EPRVs will not just discover more planets, but the mass estimates and orbital characterization from this work will help characterize these planets, their dynamics, and possible formation histories. Alongside planets, EPRV work will also further solar physics and understanding of stellar variability and activity. EPRV characterization of planets is additionally essential for the success and efficiency of many current and future space telescopes.

Enabling the RV detection of lower-mass planets will build the number of available mass estimates and orbital parameters for this under-probed population. Coupled with radius-measurements for transiting planets, we can begin to fill in the mass-radius diagram for lower-mass planets. Mass and radius measurements together allow us to derive planet density, construct models of planet composition, and interpret atmosphere spectroscopy.

EPRV measurements will establish a more complete dynamical picture of planetary systems. In addition to low-mass planets, the increased sensitivity of EPRV measurements will be able to find previously missed planets in multi-planet systems. More precise measurements will give better constraints on measured orbital parameters.

In the case of transiting planets, more sensitive RV measurements will be able to pick up on smaller Rossiter-McLaughlin signals, which inform the alignment of a planet's orbit with the stellar spin axis. Orbital misalignment has been shown to be increasingly common for a range of planet types (Schlaufman, 2010; Winn et al., 2010; Louden et al., 2021). Measuring this misalignment is therefore critical for understanding star and planet formation. An understanding of the complete dynamics of multi-planet systems along with orbital alignments will help inform migration models and planet formation simulations.

As with all RV measurements, any observation is of course to first-order an observation of a star. EPRV work will produce a catalog of extremely stable, high resolution stellar spectra. In addition to deriving stellar properties, the success of EPRV work will depend on understanding how changes from intrinsic stellar variability and activity features imprint on stellar spectra. This related work will help build our understanding of stellar physics and surface features.

Ongoing and future missions will benefit from EPRV characterization of planets. Transit surveys, like the ongoing *TESS* and *CHEOPS* missions and the planned *PLATO* mission, make use of RVs to help confirm planet candidates, return orbital parameters, and calculate planet bulk densities. Precursor observations and RV characterization will allow for better informed target selection for future missions, such as *JWST*, the Nancy Grace Roman Space Telescope, *LUVOIR*, or *HabEx*. This preliminary work is expected to dramatically raise the efficiency of these missions.

1.3 The Requirements of EPRV

Though young, the need for better RV characterization of planets and their orbits has made EPRV a rapidly growing field. Several “next-generation" instruments are already online returning sub-meter-per-second single-measurement precision (e.g. Pepe et al., 2013; Jurgenson et al., 2016; Petersburg et al., 2020; Suárez Mascareño et al., 2020; Carmona et al., 2018; Seifahrt et al., 2018; Gilbert et al., 2018; Schwab et al., 2016a) with many more optical and infrared spectrographs being commissioned, built, or planned (e.g. Thompson et al., 2016; Bouchy et al., 2017; Gibson et al., 2018; Szentgyorgyi et al., 2014).

Spectroscopic RV measurements are in essence measuring the change in spectral line position as a function of wavelength. A Doppler shift of 10 cm s^{-1} corresponds to a line shift of approximately two-ten-thousandths (0.0002) of a pixel. Given $9\,\mu m$ pixels made of Si atoms that are approximately $0.2\,nm$ in diameter, the 10 cm s^{-1} Doppler shift from an Earth-like planet is roughly equivalent to a change in line position over the width of 9 Si atoms. Measuring such a signal requires many spectral lines and a through understanding of every other source of potential changes to these lines.

Achieving this *extreme* precision enters into a new regime of measurement requirements. The hardware and engineering specifics of instruments must be held to a new standard of stability. The data from these next-generation spectrographs will have to be carefully extracted to produce high-fidelity spectra. The resultant spectra will need to be modeled in innovative ways to produce RV measurements that truly capture center-of-mass Doppler shifts invoked by orbiting planets rather than other changes present in the spectra.

1.3.1 Hardware

Echelle spectrographs are the chosen optical framework for EPRV work due to their ability to generate higher resolution spectra with greater separation of spectral features. An echelle grating is used to split light into many different wavelengths that are then aligned via collimating lenses. The collimated light then passes through cross-dispersing prisms that split the light into a series of echelle "orders." Each order represents a subset of the total wavelength range, with higher echelle orders corresponding to shorter, bluer wavelengths. The higher the echelle order, the more wavelength overlap there will be between the end of one order and the beginning of the next.

Controlling the stability of data for EPRV work begins not at the instrument, but at the telescope. The front-end module must ensure observations are trained on the same part of a star throughout an integration and that the light from this star is entering into the instrument optics in the same way at the same angle. Variation in the illumination of the optical elements can introduce instabilities to the line positions. In very much the opposite way, light must not be allowed to travel the same way through multi-mode fibers, which can generate coherent wave fronts. This gives rise to speckle patterns that can introduce variations to line shapes that could be mistaken for line shifts.

Instruments must be stabilized in terms of temperature, pressure, and vibrations in order to produce consistent, precise data. Movement due to vibrations or expansions from temperature and/or pressure changes will change the nature of the optical path and ultimately the appearance of the light that gets projected onto the detector. For EPRV work, the temperature should be stable to roughly 1 mK and the pressure stable to 10^{-7} Torr (Fischer et al., 2016).

The detector recording light information must be well characterized. The process of fabricating CCDs, the typical detector used for optical spectrographs, introduces small variations in pixel position and size that can introduce errors in an EPRV context. Spectral lines could appear to shift not because they are actually shifting in wavelength, but because the detector's pixels are moving under them (Dumusque et al., 2015). Pixel position can vary up to $90\,nm$ and pixel size can vary up to $150\,nm$ depending on the pixel stitching stepper used (Fischer et al., 2016). The charge transfer efficiency (CTE) of each pixel will also vary, requiring care when interpreting the raw data into photon counts.

To determine wavelength-dependent RV shifts, we must carefully characterize the wavelengths corresponding to each detector pixel. This is typically done using a calibration source that gives rise to a set of lines with known wavelengths. Classically, gas cells containing a purified gas, such as ThAr or I_2, were used. When excited, these gas cells produce a set of emission lines whose wavelengths are known from their molecular structure and atomic physics. By observing where these lines fell on the detector, a model can be constructed relating detector position to wavelength.

More precise and stable wavelength calibration is now possible with the recent development of laser frequency combs (LFCs) and etalons (Wilken et al., 2012; Molaro et al., 2013; Probst et al., 2014). These devices use femtosecond mode-locked lasers to generate a dense set of uniformly spaced, narrow emission lines with wavelengths known to 20-digit precision (Murphy et al., 2007; Li et al., 2008; Steinmetz et al., 2008). These LFCs and etalons make it possible to cover large wavelength ranges with lines of known wavelength, producing a well characterized, stable data set to use when constructing a model for wavelength as a function of detector pixel.

EXPRES

The EXtreme PREcision Spectrograph, *EXPRES*, was one of the first EPRV spectrographs to be commissioned and start collecting on-sky data. Built and supported by a Yale-based team, *EXPRES* was installed at the 4.3-m Lowell Discovery Telescope (*LDT*) (Levine et al., 2012) near Flagstaff, AZ in December 2017. As an optical spectrograph, *EXPRES* has a wavelength range of $390-780\,nm$. It is on-sky about 125 (partial) nights per year conducting a RV survey of the brightest, nearby stars (Brewer et al., 2020).

A custom front-end module stabilizes the position of the star relative to the input fiber, corrects for atmospheric dispersion of the image up to an airmass of about 1.2, and controls how the observed light is injected into the instrument. The spectrograph itself is housed in a vacuum enclosure that regulates the temperature and pressure of the optical elements.

The optical bench is made of invar, a particularly thermally inert material. Vibration dampeners isolate the optical bench from the enclosure and the enclosure from the floor, which itself is an isolated slab to guard against vibrations from telescope movement. An additional thermal enclosure made of polyurethane foam panels surrounds the vacuum chamber.

Light from the spectrograph is projected onto a Semiconductor Technology Associates (STA) detector, which suffers less from pixel position errors and has a measured quantum efficiency variation of 3.3%. Science light travels down a rectangular fiber ($33 \times 33\,\mu m$) to preserve resolution. This fiber is aligned so the long, $33\,\mu m$ edge lines up with detector columns, thereby minimizing the number of columns light of a given wavelength falls on.

A custom designed camera barrel minimizes the effect of variations in light injection and reduces changes in the line spread function. The line spread function (LSF) describes the shape of the light being projected by the optical elements onto the detector. The *EXPRES* LSF is well described by a Gaussian. The lenses in *EXPRES*'s camera barrel are specifically designed to stabilize the LSF so it changes very little across the detector. The width of the PSF does change some across the detector, but the change is slow. This insures that the LSF change for lines as they shift between neighboring pixels is minimal.

Figure 1.3 plots the full-width half max (FWHM) of ThAr and LFC lines across the detector. Each marker is colored by the measured FWHM of an emission line fit to a Gaussian. The resolution of *EXPRES* ranges from 120,000 to 150,000 with a median resolution of $R \sim 137,000$ (Blackman et al., 2020). Higher resolution means *EXPRES* can separate out more spectral features and provide fine sampling of spectral lines. There are approximately four pixels per resolution element in *EXPRES* data.

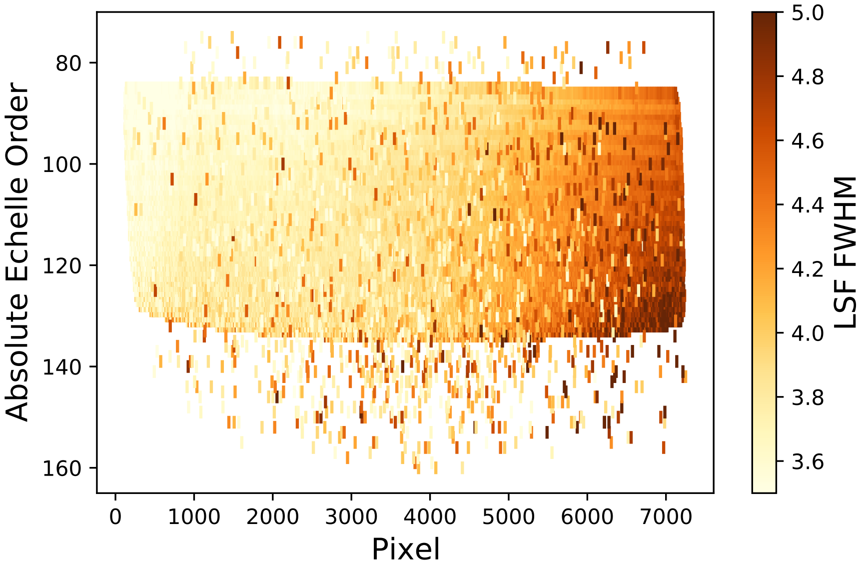

Figure 1.3: Resolution of *EXPRES* across the detector. ThAr and LFC lines are plotted by absolute echelle order (y-axis) and detector x pixel (x-axis). Each line, which represents a delta function convolved with the instrument's line spread function, is colored by the measured FWHM from fitting the calibration line to a Gaussian. The FWHM of lines do change across the detector, but the change is small and occurs slowly across the detector.

EXPRES uses a Menlo Systems laser frequency comb for precise wavelength calibration (Steinmetz et al., 2008; Probst et al., 2014, 2020). The light from this calibration source is injected after the front-end module directly into the spectrograph. It follows the same path as the science light and is projected onto the same pixels.

Measuring the shift between several consecutive LFC exposures provides a measurement of the instrument calibration stability, i.e. how much a spectra appears to be shifting due to instrument variations alone. *EXPRES* exhibits an instrument calibration stability of 4-7 cm s^{-1}. Figure 1.4 shows the shifts measured over approximately 30 minutes of consecutive LFC exposures.

1.3.2 Data Analysis

EPRV measurements necessitate high-fidelity pipelines that return clean spectra. Pipelines take the raw data from the instrument, i.e. an array of counts per detector pixel, and return

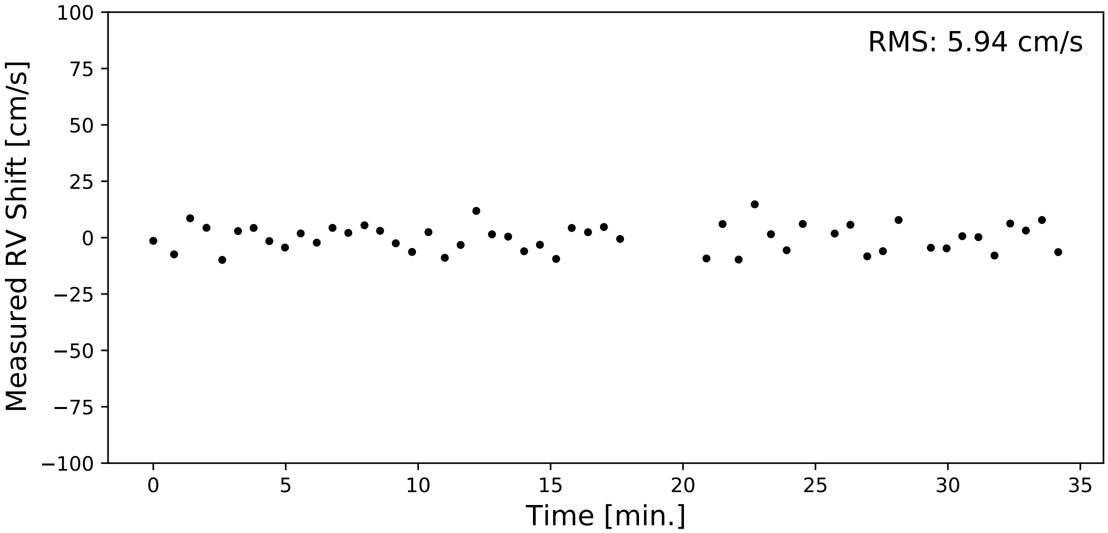

Figure 1.4: The instrument calibration stability of *EXPRES* as measured by over 30 minutes of consecutive LFC exposures. Because the calibration light only travels through the instrument optics, any apparent shifts in consecutive LFC exposures represent shifts due to instrument variation. The LFC exposures shown here exhibit a final RMS of approximately 6 cm s^{-1} after a linear trend, which is easy to calibrate out, has been removed.

spectra, a measure of the emission from the observed source as a function of wavelength. Stable hardware helps produce stable raw data. An understanding of the hardware and raw data can inform pipeline construction to accommodate any defects or variations in the hardware.

Differences in CTE, for example, are commonly corrected using a flat-field image. A flat-field image is an exposure of white light, provided by a tunable LED in the case of *EXPRES*, that illuminates all detector pixels that light from the fiber falls on. Because there are no absorption or emission features in the smooth, white-light source being injected, any variations in this flat-field exposure can be attributed to pixel-level CTE changes. Measuring these changes allows us to learn about the behavior of detector pixels and enables us to correct for them.

Flat-fielding is an excellent example of using a measurement of known variation in raw data that is then corrected for through the data pipeline. *EXPRES*'s data pipeline extracts spectra by scaling a composite, nightly flat-field to recreate the observed data from other sources (Petersburg et al., 2020). This flat-relative optimal extraction inherently folds in CTE corrections, removes the blaze function introduced by the echelle optics, and mitigates the effects of localized cosmic-ray hits. More specifics of this pipeline are given in Chapter 3.

Though flat-fielding is hardly new, recent work has shown that a similar philosophy can be used to address changes in the LSF or the calibration state of the spectrograph. Spectro-perfectionism is an algorithm for extracting data that incorporates a thorough understanding of the LSF across the detector in order to forward model this effect in the raw data (Bolton & Schlegel, 2010; Cornachione et al., 2019). By collecting data that characterize the LSF behavior across the data, spectro-perfectionism can use this information to correct for LSF changes.

Wavelength calibration, the process of matching each pixel to the wavelength of life being projected onto it, is also improved by using the calibration data itself to learn more about the instrument. RV measurements are inherently a measure of line shifts with respect to wavelength. The dispersion, or difference in wavelength between one pixel and a neighboring pixel, changes with wavelength. This makes wavelength calibration an essential yet nontrivial step of the pipeline.

Chapter 4 describes *excalibur*, a hierarchical, non-parametric method for wavelength calibration. *Excalibur* uses calibration images to construct the complete, accessible calibration space of an instrument. It then uses local calibration images or time to pinpoint an instrument's location in this space for each exposure. The calibration images inform a model of how the instrument can vary. Using this model, we can re-create more faithful calibration line positions that suffer less from measurement error. This gives a better sense of the calibration state and corresponding wavelength calibration for the instrument at each observation.

The cm s^{-1} stability of EPRV instruments has changed how we must extract, model, and calibrate this new regime of data. The future of extraction pipelines lies in methods that take advantage of the information such stable raw data provides of the instrument itself, and turn this knowledge into an opportunity to correct for any variations in the instrument's optical elements or detector via the data pipeline. Only with precise, high-fidelity spectra will the subtler effects of stellar signals become clear.

1.3.3 Stellar Signals

Discovering and characterizing planets with the RV method depends on being able to measure true Doppler shifts in observed spectra. Independent of the instrument and data pipeline, stellar photospheres themselves are evolving and changing, sometimes in periodic ways. These photospheric velocities manifest as spectral line shape changes and make it significantly trickier to measure small shifts in these changing lines.

Photospheric velocities arise from both intrinsic stellar variations, such as stellar pulsations or granulation, as well as activity features, such as darker star spots or brighter faculae and plages. The effect of these variations change in both behavior and magnitude for different spectral lines and evolve on a range of timescales (Davis et al., 2017; Dumusque, 2018; Meunier et al., 2017). The resultant line shape changes skew measurements of line shift and add potentially periodic noise to derived RV measurements. In order to increase RV precision, we must be able to disentangle these stellar signals from true center-of-mass shifts due to orbiting planets.

The different types of stellar variability are summarized in Table 1.1 along with their timescale and impact on RVs. We briefly describe each category below.

P-mode oscillations arise when pressure waves propagate to the surface of stars, causing the envelope of the star to expand and contract. This occurs on the timescale of minutes for main-sequence stars, with hotter stars experience oscillations of higher frequency and amplitude (Arentoft et al., 2008; Kjeldsen et al., 2005; Bouchy et al., 2005; Mayor et al., 2003). These pulsations can introduce RV variations of 0.1 to 1 m s^{-1} for main-sequence stars (Dumusque et al., 2011b; Chaplin et al., 2019).

In solar-type stars, convection gives rise to a granulation pattern that induces a "convective blueshift" (Lanza et al., 2019; Cegla et al., 2018; Dumusque et al., 2011b; Lindegren & Dravins, 2003; Kjeldsen & Bedding, 1995; Dravins, 1982). The center of granulation cells are formed from convective upflow and appear blueshifted. The darker edges

Table 1.1: Types and Effects of Stellar Variability and Activity

Variation Type	Timescale	RV Variation [m s^{-1}]
Oscillations	minutes	0.1-1
Granulation	minutes to hours	0.4-0.8
Super-Granulation	hours to two days	0.3-0.7
Activity Features:		
Suppressing Convective Blueshift	days to month	0.4-1.4
Rotation	days to month	1-10

of the granules represent the convective downflow, which would appear redshifted as it is moving away from the surface of the star. The unequal area between the middle upflow and edge downflow regions creates a net RV blueshift, i.e. the convective blueshift, and changes spectral lines to appear asymmetric.

The surface of a star will feature different realizations of granulation that change on the order of a few minutes to hours. Different granulation patterns will integrate to different net RV shifts, meaning the magnitude of the convective blueshift in stars and its effect on stellar lines is also changing. These differences can result in random RV variations on the order of 0.4 to 0.8 m s^{-1}. The amplitude tends to increase with the T_{eff} of the star (Meunier et al., 2015).

Supergranulation is a similar effect caused by large granulation cells outlined by the magnetic network of a star. These cells can persist for hours or up to two days and manifest as RV variations between 0.3 and 0.7 m s^{-1}(Meunier et al., 2015; Meunier & Lagrange, 2019; Rincon & Rieutord, 2018; Rieutord & Rincon, 2010)

Activity features, such as darker star spots and brighter faculae and plage regions, can arise in the presence of strong magnetic activity on a star (Jeffers et al., 2013; Lovis et al., 2011; Dumusque et al., 2011a; Boisse et al., 2011; Huélamo et al., 2008; Desort et al., 2007; Saar, 2003; Hatzes, 2002; Saar & Donahue, 1997). Magnetic activity can act to suppress convection in a star, which will change the convective blueshift signal. This change can result in an RV variation of 0.4 to 1.4 m s^{-1} (Meunier et al., 2010).

Activity features are localized regions on the surface of a star. The features, as well as their effects, will rotate along with the star and come in and out of view. Darker, lower temperature spots will give off less flux than brighter, hotter faculae and plages. The presence of activity features can therefore change the flux distribution of a star. As a star rotates, half of the star will appear to be blueshifted as it rotates towards us, while the other half will appear redshifted as it rotates away from us. Typically, this effect will even out, but the different flux from activity features could break this balance, introducing potentially 1 to 10 m s^{-1} variations depending on the size of the feature and stellar rotation rate (Meunier et al., 2010; Aigrain et al., 2012; Haywood et al., 2014).

Traditionally, stellar signals have been decorrelated from radial-velocity measurements using global activity indicators. These indicators aim to gauge magnetic activity on the target star, presence of activity features, or magnitude of line shape changes at the time of each exposure (Figueira, 2013; Dumusque et al., 2011c; Boisse et al., 2009, e.g.). For example, magnetic activity on the star has been shown to correlate with emission in the core of Ca II H&K lines (Meunier & Lagrange, 2013; Saar et al., 1998), the CA infrared triplet (Saar & Fischer, 2000), and the H-α line (Giguere et al., 2016; Robertson et al., 2014; Skelly et al., 2008).

Other favorites include properties of the cross-correlation function (CCF) often used to derive RVs. The CCF can be thought of as an average of all line shapes in the spectrum. It is therefore sensitive to line shape changes that appear in most lines, but will blur out lines behaving in different ways. The CCF bisector provides a measure of asymmetry in the CCF (e.g. Queloz et al., 2001; Povich et al., 2001) while the FWHM of the CCF can be used as a measure of the amount of line changes that result in a blurred, wider CCF (e.g. Queloz et al., 2009).

Decorrelating RV measurements against these global activity indicators has not shown success in disentangling stellar signals at the sub-meter-per-second level (Fischer et al., 2016). Recent years have seen an explosion in different methods attempting to address the issue of stellar signals.

Methods vary in the type of data used. Some methods present more advanced ways of modeling RVs and activity indicators together (e.g. Rajpaul et al., 2015; Barragán et al., 2019; Gilbertson et al., 2020). Other methods try new decompositions of the CCF to isolate shape changes (e.g. Collier Cameron et al., 2020; Zhao & Tinney, 2020; de Beurs et al., 2020). A growing class of methods are using the full spectra in their analysis to identify variations on a line-by-line basis or model expected changes from stellar variability (e.g.

Jones et al., 2017; Dumusque, 2018; Lafarga et al., 2020; Holzer et al., 2020; Rajpaul et al., 2020; Holzer et al., 2021). These new methods make use of a range of methodology, such as Gaussian process (GP) modeling, principal component analysis (PCA), fourier transforms, etc.

Many of the most recent methods being employed and developed are discussed in more detail in Chapter 5 and compared based on results using *EXPRES* data as part of the *EXPRES* Stellar Signals Project (*ESSP*) (Zhao et al., 2020). As part of the *ESSP*, 19 different methods were tested on a consistent set of *EXPRES* data, allowing for a direct comparison of the performance of these different methods. Though ongoing work, the *ESSP* has helped establish the current state of the field for disentangling stellar signals and has clarified what is needed to push progress forward.

RV measurements are essential for discovering and characterizing planets, deriving their composition, and informing formation simulations. We need EPRV measurements to move the field forward (Chapter 2). As more and more EPRV spectrographs come on line striving for sub-meter-per-second precision, they will all face the same hardware, pipeline, and stellar issues.

These instruments will have to make use of data-driven pipelines that use the data to characterize and correct for variations in the raw data, such as flat-relative extraction (Chapter 3) or *excalibur* for wavelength calibration (Chapter 4). Interpreting the resultant spectra requires diagnosing how stellar variability and activity features can change the observed spectra (Chapter 5). We can only confidently detect planets with $K < 1$ m s^{-1} if true center-of-mass shifts from orbiting planets are faithfully disentangled from stellar signals.

To work towards EPRV measurements is to work in a new regime. Errors that were negligible before become dominant. Correcting for such sources reveals new errors that previous data was not sensitive enough to perceive. This thesis encompasses a variety of work that addresses some error sources, reveals others, and takes stock of the current state of the field. Chapter 6 will give a summary of the work and discuss future directions to continue down this extreme precision road.

Chapter 2

Planet Detectability in the Alpha Centauri System

Lily L. Zhao[1], *Debra A. Fischer, John Brewer, Matt Giguere, Bárbara Rojas-Ayala*

We use more than a decade of radial velocity measurements for α Cen A, B, and Proxima Centauri from *HARPS*, *CHIRON*, and *UVES* to identify the $M \sin(i)$ and orbital periods of planets that could have been detected if they existed. At each point in a mass-period grid, we sample a simulated, Keplerian signal with the precision and cadence of existing data and assess the probability that the signal could have been produced by noise alone. Existing data places detection thresholds in the classically defined habitable zones at about $M \sin(i)$ of 53 $M_{\oplus}$ for α Cen A, 8.4 $M_{\oplus}$ for α Cen B, and 0.47 $M_{\oplus}$ for Proxima Centauri. Additionally, we examine the impact of systematic errors, or “red noise" in the data. A comparison of white- and red-noise simulations highlights quasi-periodic variability in the radial velocities that may be caused by systematic errors, photospheric velocity signals, or planetary signals. For example, the red-noise simulations show a peak above white-noise simulations at the period of Proxima Centauri b. We also carry out a spectroscopic analysis of the chemical composition of the α Centauri stars. The stars have super-solar metallicity

---

[1] Originally published as: Zhao, L., Fischer, D. A., Brewer, J., Giguere, M., & Rojas-Ayala, B. 2018, The Astronomical Journal, 155, 24. I processed the different data sets and ran simulations of planet signals to determine what types of planets were detectable with the data available at the time.

with ratios of C/O and Mg/Si that are similar to the Sun, suggesting that any small planets in the α Cen system may be compositionally similar to our terrestrial planets. Although the small projected separation of α Cen A and B currently hampers extreme-precision radial velocity measurements, the angular separation is now increasing. By 2019, α Cen A and B will be ideal targets for renewed Doppler planet surveys.

2.1 Introduction

Over the past two decades, hundreds of exoplanets have been detected with the radial velocity technique, opening a new subfield of astronomy. In 2009, the NASA *Kepler* mission (Borucki et al., 2011; Batalha et al., 2013) used the transit technique to dramatically advance our understanding of exoplanet architectures, especially for low-mass planets. Burke et al. (2015) used the Q1-Q16 *Kepler* catalog (Mullally et al., 2015) with the Christiansen et al. (2015) pipeline completeness parameterization to assess planet occurrence rates for *Kepler* G and K dwarfs. For exoplanets with radii $0.75 \leq R_{\rm planet} \leq 2.5\ R_\oplus$ and orbital periods, $50 \leq P_{\rm orb} \leq 300$ days, they find an occurrence rate, $F_0 = 0.77$ planets per star, with an allowed range of $0.3 \leq F_0 \leq 1.9$. The Burke et al. (2015) *Kepler* data analysis suggests that most GK stars have rocky exoplanets and portends a bright future for the discovery of low-mass planets orbiting nearby GK stars with the radial velocity technique, once precision is improved.

At a distance of 1.3 parsecs, the three stars in the α Centauri system are our closest neighbors. The stars of the central, α Cen AB binary system orbit each other with a semi-major axis of 24 AU and an eccentricity of 0.524 (Pourbaix & Boffin, 2016). Though planets are now known to be common, there has been theoretical speculation about whether planets would form in such a close binary system (Thébault et al., 2006, 2008, 2009). Simulations have shown that if planets do form in this system (Quintana & Lissauer, 2006; Quintana et al., 2007; Guedes et al., 2008), there are regions where they can reside in dynamically stable orbits (Wiegert & Holman, 1997; Quarles & Lissauer, 2016) around either α Cen A and α Cen B. Furthermore, approximately 20% of known planets orbit stars that are components of binary star systems. Particularly interesting is the case of HD 196885 AB, a stellar binary system with a semi-major axis of 24 AU and an eccentricity of 0.409, similar to the orbit of α Cen AB, with a known planet orbiting the primary star (Correia et al., 2008; Fischer et al., 2009; Chauvin et al., 2007). The case of HD 196885 Ab provides empirical evidence that the formation of planets is not precluded around α Cen A or B.

The third star, Proxima Centauri, is a smaller M dwarf and orbits this pair with a semi-major axis between 8,700 and greater than 10,000 AU (Wertheimer & Laughlin, 2006; Kervella et al., 2017b). The α Centauri system has long been a key target for Doppler planet searches from southern hemisphere observatories(Murdoch et al., 1993; Endl et al., 2001; Dumusque et al., 2012; Endl et al., 2015). While no planets have yet been discovered around α Cen A or B (c.f. Dumusque et al., 2012; Hatzes, 2013; Rajpaul et al., 2016), an Earth-mmass planet has been detected orbiting Proxima Centauri using the Doppler technique (Anglada-Escudé et al., 2016). This recent discovery has increased interest in the system and the proximity of these stars is an enormous advantage for missions that aim to obtain images of any exoplanets. As human exploration ventures beyond our solar system, these closest stars will surely be our first destination.

In this work, we publish radial-velocity observations of α Cen A and B, obtained at the Cerro Tololo Interamerican Observatory (*CTIO*) with the Echelle Spectrograph (es) from 2008 - 2010 and the *CTIO* High Resolution (*CHIRON*) spectrograph. These data, together with archived data from the High Accuracy Radial Velocity Planet Searcher (*HARPS*) and the Ultraviolet and Visual Echelle Spectrograph (*UVES*) of α Cen B and Proxima Centauri are used to test planet detectability and place constraints on the mass and orbital periods of putative planets that may remain undetected around these three stars.

2.2 The alpha Centauri System

Alpha Centauri is a hierarchical triple-star system. The primary and secondary components, α Cen A and B, are main-sequence stars with spectral types G2V and K1V, respectively, that are gravitationally bound in an eccentric orbit with a semi-major axis of about 24 AU. The two stars currently have an angular separation of about 5 arcseconds, which is not resolvable with the naked eye. Their combined brightness of -0.27 magnitudes makes α Cen AB one of the brightest objects in the southern hemisphere. The third star in this system, α Cen C or Proxima Centauri, was discovered in 1915 (Innes, 1915) and is a relatively faint $V = 11.1$ magnitude M6V dwarf at a projected angular separation of 2.2 deg from α Cen AB.

The recent astrometric analyses of α Cen A (van Leeuwen, 2007; Pourbaix & Boffin, 2016; Kervella et al., 2017a) yield an orbital parallax between 743 and 754 mas, corresponding to a distance of 1.33 to 1.35 pc away. The three stars in the α Centauri system are our closest stellar neighbors.

2.2.1 Doppler Analysis

Observations of α Centauri A and B were obtained with the 1.5-m telescope at the Cerro Tololo Interamerican Observatory (*CTIO*) in Chile. From 2008 - 2010, the refurbished Echelle Spectrometer (*ES*) was used to collect data. The *ES* was located in the Coudé room; however no other attempt was made to stabilize the thermal environment of the spectrograph and both diurnal and seasonal variations resulted in temperature changes of several degrees in the Coudé room. Light from the telescope was coupled to this instrument with an optical fiber and a slit was positioned at the focus to set the resolution to $\sim 48,000$. However, the slit width was manually set with a micrometer and was not very precise, therefore, we expect that slight variations in the resolution occurred over time.

In 2011, we replaced the *ES* with the *CHIRON* spectrograph (Tokovinin et al., 2013). This instrument was also placed in the Coudé room and the optical fiber was changed to an octagonal fiber to reduce modal noise in our spectra. *CHIRON* was not in a vacuum enclosure, however the combination of thermal insulation and a thermally controlled space inside the Coudé room stabilized the temperature drifts to +/- 2 K. There are four observing modes with *CHIRON*; for our observations of α Centauri A and B we adopted a fixed-width slit at the focus of the optical fiber that provided an instrumental resolution of R $\sim$ 90,000 at the expense of a $\sim$ 30% light loss. A small fraction of light was picked off from the light path inside the spectrograph and sent to a photomultiplier tube to determine the photon-weighted midpoint and correct for the barycentric velocity during our observations.

The *ES* and *CHIRON* both use an iodine cell to provide the wavelength solution and to model Doppler shifts (Butler et al., 1996). The iodine cell is inserted into the light path for all of the program observations where radial velocities will be measured. The forward modeling process that we use also requires high-SNR, high-resolution template observations and a very high resolution Fourier transform spectrum (FTS) of the iodine cell, obtained at the Pacific Northwest National Labs (PNNL) Environmental Molecular Sciences Laboratory (EMSL). Template observations are made without the iodine cell and are bracketed by several observations of bright, rapidly rotating B-stars through the iodine cell. The B-star observations are used to model the wavelength solution and the spectral line spread function (SLSF) of the instrument. The template observation is deconvolved with the SLSF, providing a higher resolution, iodine-free spectrum for modeling Doppler shifts. With the template observation, T_s and the FTS iodine spectrum, I_2, the model is constructed as:

$$(T_s \times I_2) * SLSF \tag{2.1}$$

and a Levenberg-Marquardt least squares fitting is used to model the program observations. The error budget for the *CHIRON* radial velocity (RV) measurements accounts for instrumental errors (including variations in temperature, pressure, and vibrations), modal noise in the octagonal fiber, algorithm errors in the analysis, and the inclusion of velocity effects (granulation, spots, faculae) from the surface of the stars. For the α Centauri AB stars, flux contamination from the companion star turned out to be the most significant error source.

2.2.2 Spectroscopic Analysis

The stellar properties and chemical abundances of α Cen A and B were determined by using the spectral synthesis modeling code, Spectroscopy Made Easy (SME), described in Brewer et al. (2016), to analyze several iodine-free spectra obtained with the *CHIRON* spectrograph in 2012. The stellar parameters that we derive, as well as some comparison data that represent the range of values from the published literature with available uncertainties, are listed in Table 2.1.

Because we have analyzed 28 α Cen A and B spectra, the rms of those spectroscopic parameters is one way to assess uncertainties. However, for all spectroscopic parameters, we find that the rms is too small to provide a plausible estimate of uncertainties. Instead, we adopt the more conservative model parameter uncertainties that were established using the same SME modeling technique for more than 1600 stars observed with the Keck *HIRES* spectrograph (Brewer et al., 2016). Following Brewer et al. (2016), small empirical corrections were applied to the elemental abundances of α Cen AB to account for slight systematic trends that occur as a function of temperature with our analysis method.

Our spectroscopic analysis yields an effective temperature of 5766 ± 25 K for α Cen A, and 5218 ± 25 K for α Cen B. The effective temperature for α Cen A is consistent with the effective temperature measurement derived from angular-diameter measurements by Boyajian et al. (2013) and consistent with the G2V spectral classification (Perryman et al., 1997; van Leeuwen, 2007). The calculated effective temperature for α Cen B is similarly consistent with the results of Boyajian et al. (2013) and the K1V spectral classification.

Both stars have a super-solar metallicity, [Fe/H]= 0.22 ± 0.03 and 0.24 ± 0.03 for α Cen A and B, respectively, consistent with other published values, e.g., Anderson & Francis (2011). We measure a C/O ratio of 0.47 ± 0.05 and Mg/Si of 1.05 ± 0.03 for α Cen A, similar to the solar value. The results for α Cen B are nearly identical with a C/O ratio of 0.49 ± 0.05 and a Mg/Si ratio of 1.05 ± 0.03, the same as α Cen A. Because the ratios of abundances in stellar photospheres evolve slowly over main-sequence lifetimes (Pinsonneault et al., 2001; Turcotte & Wimmer-Schweingruber, 2002), we can use the C/O and Mg/Si ratios as a proxy for disk compositions. Brewer & Fischer (2016) showed that most stars have low C/O ratios, leaving the Mg/Si ratio important for regulating the geology of planetesimals. The implication is that any rocky planets forming around α Cen A or B could have a composition and internal structure that may be similar to the solar system terrestrial planets.

The temperature and [Fe/H] for Proxima Centauri were derived from infrared K-band features in XSHOOTER spectra available from the ESO Public Archive. The observations were carried out in Period 92 using a slit width of 0.4" (R ~9100) and were reduced following the standard recipe described in the XSHOOTER pipeline manual[2] (Vernet et al., 2011). The wavelength calibration for the spectra was based on telluric lines, using a modified version for XSHOOTER data of the IDL-based code *xtellcor_general* by Vacca et al. (2003). The Proxima Cen spectra were convolved with a Gaussian kernel to degrade the resolution to R ~2700, in order to use the Na I, Ca I, and H_2O-K2 indices calibrated to provide metallicity estimates for M dwarf stars by Rojas-Ayala et al. (2012). With this technique, we derive T_{eff} = 2879 ± 50 K and [Fe/H]= 0.08 ± 0.12 and a spectral type of M5.5V. The metallicity for Proxima Cen is slightly lower than that of α Cen A and B; however, the uncertainty in the Proxima Cen measurement is four times the uncertainty for α Cen A or B. Proxima Centauri should share the same chemical composition as α Cen A

---

[2] http://www.eso.org/sci/software/pipelines/

and B unless those stars had a significantly different accretion history than Proxima.

2.2.3 Isochrone Analysis

Using spectroscopic parameters (T_{eff}, [Fe/H], [Si/H]) and distance, we derive the best fit models to the Yale-Yonsei (Y^2) isochrones (Demarque et al., 2004) to estimate the stellar mass, radius, and age for α Cen A and B. Our stellar masses (listed in Table 2.1) agree well with other published values (Lundkvist et al., 2014; Pourbaix & Boffin, 2016) and the radius is consistent with the angular-diameter measurement by Kervella et al. (2017a). The isochrone-derived age for α Cen A is $5.17^{+1.03}_{-0.97}$ Gyr, slightly older than the Sun and consistent with previous age estimates. Our isochrone model for α Cen B gives a younger age with large uncertainties, $2.53^{+3.12}_{-1.89}$ Gyr. The posterior in the isochrone fit shows a peak at younger ages for α Cen B that is ill-constrained by $\log g$ and distance. However, the ages for the two stars do agree within their uncertainties.

The log g, stellar mass, and radius of Proxima Cen were determined by adopting the age of ~5 Gyr that we estimate for α Cen AB, and interpolating the temperature onto a solar-metallicity isochrone for main-sequence, low-mass stars from Baraffe et al. (2015). Because M dwarfs evolve very slowly after the pre-main-sequence phase, any errors in the adopted age of the star will not significantly affect the derived stellar model. The Baraffe et al. (2015) isochrones were only calculated for solar metallicity; therefore, the isochrone model parameters will not account for the slightly super-solar metallicity of Proxima Centauri. The isochrone model parameters for Proxima Cen are also compiled in Table 2.1.

2.2.4 Chromospheric Activity

The chromospheric activity of α Cen A and B was monitored Henry et al. (1996) by measuring emission in the cores of the Ca II H & K lines relative to continuum bandpasses (i.e., the $S_{\rm HK}$ values), scaled to the long-term Mount Wilson H & K study (Wilson, 1978; Vaughan et al., 1978; Duncan et al., 1991; Gray & Baliunas, 1995). The $S_{\rm HK}$ values together with the $B - V$ of the star can then be transformed to $\log R'_{\rm HK}$, which is the fraction of bolometric luminosity from the lower chromosphere after subtracting off photospheric contributions (Noyes et al., 1984). Using $\log R'_{\rm HK}$ instead of $S_{\rm HK}$ allows for a straightforward comparison of chromospheric activity that is independent of spectral type.

Chromospheric activity provides a good way to estimate rotation periods and ages (Noyes et al., 1984), even for older and more slowly rotating stars. Both α Cen A and B are chromospherically quiet stars with estimated rotation periods of about 22 and 41 days, respectively (Morel et al., 2000). This is typical of stars that are about the age of the Sun. Coronal cycles have been measured at X-ray and UV wavelengths with periods of 19 and 8 years for α Cen A and B, respectively (Ayres, 2014, 2015).

A recalibration of chromospheric activity-age relation and a calibration of stellar rotation to stellar age (gyrochronology) was carried out by Mamajek & Hillenbrand (2008). Their revised calibration returns activity and gyrochronology ages of 6.6 and 4.4 Gyr for α Cen A and 5.2 and 6.5 Gyr for α Cen B. Mamajek & Hillenbrand (2008) estimate uncertainty in these ages of about 1.5 Gyr for the activity calibration and 1.3 Gyr for their gyrochronology technique.

2.2.5 Stellar Ages

Stellar ages can be estimated in several ways: from isochrone fitting, stellar activity, stellar rotation speed (gyrochronology), dynamical measurements of the visual binary orbit, or galactic kinematics. In the case of a binary star system, we expect that the stars are co-eval; both stars should yield an independent estimate for the age of the binary system. Taking the average of the stellar ages for both α Cen A and B estimated from the ages tabulated in Table 2.1, we calculate a weighted mean age for the α Centauri system of 5.03 ± 0.34 Gyr.

Table 2.1: Stellar Parameters for the α Cen Stars

Parameter	α Cen A	α Cen B	α Cen C	Source
ID	HD128620 HIP71683 GJ559A	HD128621 HIP71681 GJ559B	HIP70890 GJ551	
Spectral Type	G2V	K0V	M5Ve	Perryman et al. (1997)
V_{mag}	−0.01	1.13	11.1	Perryman et al. (1997)
Parallax [mas]	754.81 ± 4.11	796.92 ± 25.9	771.64 ± 2.6	van Leeuwen (2007)
Parallax [mas]	743 ± 1.3	743 ± 1.3	–	Pourbaix & Boffin (2016)
Parallax [mas]	747.17 ± 0.61	747 ± 0.61	–	Kervella et al. (2017a)
μ RA [mas/yr]	−3679.25	−3614.39	−3775.75	Perryman et al. (1997)
μ RA [mas/yr]	–	–	−3773.8 ± 0.4	Kervella et al. (2016)
μ Dec [mas/yr]	473.67	802.98	765.65	Perryman et al. (1997)
μ Dec [mas/yr]	–	–	770.5 ± 2.0	Kervella et al. (2016)
M_V	4.36	5.5	15.5	using Kervella et al. (2017a) parallax
$M_\star / M_\odot$	1.10 ± 0.03	0.97 ± 0.04	–	Lundkvist et al. (2014)
$M_\star / M_\odot$	1.13 ± 0.007	0.97 ± 0.04	–	Pourbaix & Boffin (2016)
$M_\star / M_\odot$	1.1055 ± 0.0039	0.9373 ± 0.0033	–	Kervella et al. (2016)
$M_\star / M_\odot$	–	–	0.1221 ± 0.0022	Mann et al. (2015)
$M_\star / M_\odot$	–	–	0.120 ± 0.015	Anglada-Escudé et al. (2016)
$M_\star / M_\odot$	1.1 ± 0.02	0.91 ± 0.02	0.106 ± 0.005	(this work)
θ_{LD} [mas]	8.511 ± 0.020	6.001 ± 0.0034	–	Kervella et al. (2003)
θ_{LD} [mas]	–	–	1.044 ± 0.08	Ségransan et al. (2003)

Continued on next page...

Table 2.1 – continued from previous page

Parameter	α Cen A	α Cen B	α Cen C	Source
$R_\star/R_\odot$	1.22 ± 0.01	0.88 ± 0.01	–	Lundkvist et al. (2014)
$R_\star/R_\odot$	1.231 ± 0.0036	0.868 ± 0.0052	–	Pourbaix & Boffin (2016)
$R_\star/R_\odot$	1.2234 ± 0.0053	0.8632 ± 0.00537	–	Kervella et al. (2017a)
$R_\star/R_\odot$	–	–	0.1542 ± 0.0045	Mann et al. (2015)
$R_\star/R_\odot$	–	–	0.141 ± 0.02	Anglada-Escudé et al. (2016)
$R_\star/R_\odot$	1.23 ± 0.04	0.84 ± 0.02	0.131 ± 0.005	(this work)
$L_\star/L_\odot$	1.5159 ± 0.0051	0.5014 ± 0.0017		Boyajian et al. (2013)
$L_\star/L_\odot$	–	–	0.00155 ± 0.00005	Anglada-Escudé et al. (2016)
$L_\star/L_\odot$	–	–	0.0011 ± 0.0002	(this work)
Age [Gyr]	$5.17^{+1.03}_{-0.97}$	$2.53^{+3.12}_{-1.89}$	–	(this work)
Age [Gyr]	5.2 ± 1	4.5 ± 1.2	–	(isochrone) Boyajian et al. (2013)
Age [Gyr]	4.85 ± 0.5	–	–	Thévenin et al. (2002)
Age [Gyr]	6.6 ± 1.6	5.2 ± 1.6	–	(activity) Mamajek & Hillenbrand (2008)
Age [Gyr]	4.4 ± 1.3	6.5 ± 1.3	–	(gyro) Mamajek & Hillenbrand (2008)
$\log R'_{\mathrm{HK}}$	–	−4.9 ± 0.08	–	Dumusque et al. (2012)
$\log R'_{\mathrm{HK}}$	−5.002	−4.923	–	Henry et al. (1996)
P_{rot} [days]	22	41	–	Morel et al. (2000)
T_{eff} [K]	5793 ± 7	5232 ± 9	–	Boyajian et al. (2013)
T_{eff} [K]	5766 ± 25	5218 ± 25	2879 ± 50	(this work)
$v\sin(i)$[km s^{-1}]	2.51 ± 0.5	1.9 ± 0.5		(this work)
$\log g$ [cm s^{-1}]	4.31 ± 0.05	–	–	Lundkvist et al. (2014)
$\log g$ [cm s^{-1}]	4.3117 ± 0.0015	–	–	Kervella et al. (2017a)
$\log g$ [cm s^{-1}]	4.27 ± 0.05	4.44 ± 0.05	5.23 ± 0.01	(this work)

Continued on next page...

Table 2.1 – continued from previous page

Parameter	α Cen A	α Cen B	α Cen C	Source
[Fe/H]	0.20	0.21	–	Anderson & Francis (2011)
[Fe/H]	0.22 ± 0.03	0.24 ± 0.03	0.10 ± 0.12	(this work)
[C/H]	0.19 ± 0.03	0.19 ± 0.03	–	(this work)
[O/H]	0.25 ± 0.04	0.23 ± 0.04	–	(this work)
[Si/H]	0.21 ± 0.03	0.20 ± 0.03	–	(this work)
[Mg/H]	0.19 ± 0.03	0.19 ± 0.03	–	(this work)
[C/O]	0.47 ± 0.05	0.49 ± 0.05	–	(this work)
[Mg/Si]	1.05 ± 0.03	1.05 ± 0.03	–	(this work)
[Si/Fe]	1.11 ± 0.04	1.04 ± 0.03	–	(this work)

2.2.6 Stellar Orbits

Pourbaix et al. (1999) used published astrometry and radial velocity (RV) data from the European Southern Observatory Coudé Echelle Spectrograph to derive the orbital parameters for the α Cen A and B stellar binary system. Pourbaix & Boffin (2016) refined this binary star orbit by supplementing their previous analysis with 11 years of high-precision RV measurements from the *HARPS* spectrograph and some additional astrometric data from the Washington Double Star Catalog (Hartkopf et al., 2001). They derive an orbital period of 79.91 ± 0.013 years and eccentricity of 0.524 ± 0.0011, and masses $M_{\mathrm{A}} = 1.133 \pm 0.005$ $M_{\odot}$ and $M_{\mathrm{B}} = 0.972 \pm 0.0045$ $M_{\odot}$. We use their published orbital parameters to plot the projected orbit of α Cen B around α Cen A in Figure 2.1. The projected separation between the two stars reaches a local minimum in 2016, but will increase to observable levels by 2019.

Proxima Centauri has a projected separation of $15,000 \pm 700$ AU from α Cen AB and a relative velocity with respect to α Cen AB of 0.53 ± 0.14 km s^{-1} (Wertheimer & Laughlin, 2006). Wertheimer & Laughlin (2006) used Hipparcos kinematic information and carried out Monte Carlo simulations to determine the binding energy of Proxima Cen relative to α Cen AB. They found a high probability that Proxima Cen is gravitationally bound and near apastron in a highly eccentric orbit. More recently, Kervella et al. (2017b) added published *HARPS* RV measurements and likewise concluded that Proxima Centauri is gravitationally bound to the α Cen AB stars, traveling in an orbit with eccentricity of $0.50^{+0.08}_{-0.09}$ with an orbital period of $\sim 550,000$ years.

2.3 Exoplanet Searches

The three stars in the α Centauri system have been targets of different precision, radial-velocity surveys to search for exoplanets from southern hemisphere observatories (Endl

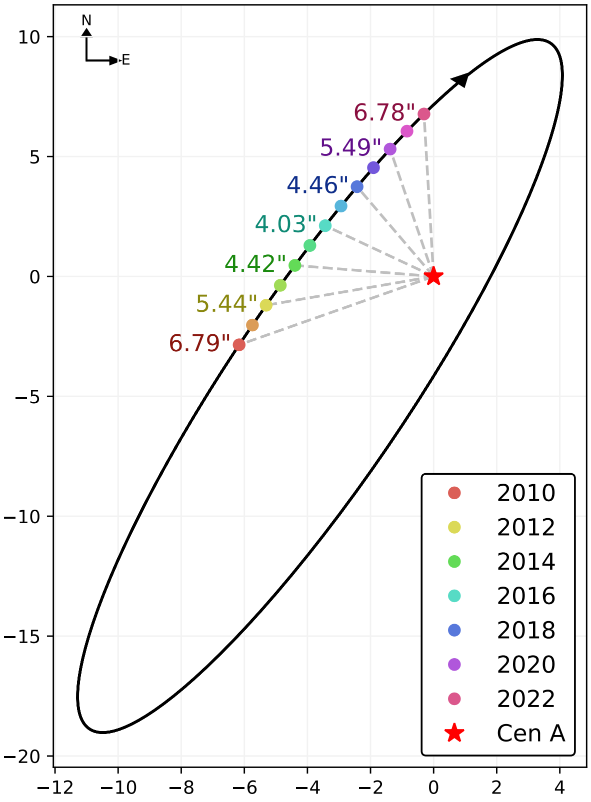

Figure 2.1: Projected orbital plane of α Cen A and B. The angular separation reaches a temporary minimum just under 4” in 2017 and the angular separation begins to increase in 2018. By 2020, the separation exceeds 5."5 and ground-based, radial-velocity searches can resume without suffering significant contamination from the companion star.

et al., 2001; Dumusque et al., 2012; Endl et al., 2015). In 2012, a planet was announced orbiting α Cen B (Dumusque et al., 2012) using data from the *HARPS* spectrograph. While that putative signal was later shown to be a sampling alias in the time series data (Rajpaul et al., 2016), Anglada-Escudé et al. (2016) subsequently discovered a low-mass planet orbiting Proxima Centauri, a M5.5V star. The orbital period of Proxima Cen b is 11 days, which places this planet at the appropriate distance from its host star to fall within the habitable zone. This detection was a record-breaking discovery because of the low mass of the planet, although the habitability of this world is now being debated. Airapetian et al. (2017) find that the planet orbiting Proxima Centauri will incur a significant atmospheric loss of oxygen and nitrogen in addition to a massive loss of hydrogen because of the high-energy flux from this relatively active M dwarf (Airapetian et al., 2017).

There are several reasons why Doppler planet searches around the binary stars α Cen A and B are well-motivated. The stars are bright, allowing for high cadence and signal-to-noise spectra. The declination of the stars is -60 degrees, close to a southern polar orbit, so that the observing season stretches between nine months and a year depending on the position of the observatory. Dynamical simulations (Wiegert & Holman, 1997) show that any planets in the system are likely to be nearly aligned with the binary-star orbit; this implies that any RV amplitude would not be strongly attenuated by orbital inclination.

However, there are some challenges for planet detection, formation, and long-term stability around α Cen A or B. One key concern is that the semi-major axis of the binary star orbit is only about 24 AU (Pourbaix & Boffin, 2016) and the orbital eccentricity of 0.524 means that the separation of the stars is only 16.3 AU at periastron passage. While Wiegert & Holman (1997) demonstrate that any existing planets would be dynamically stable if they orbit within a few AU of either star, the close proximity of the stars has led to theoretical speculation about whether planets could have formed in the first place around α Cen A or B (Barbieri et al., 2002; Quintana et al., 2002; Quintana & Lissauer, 2006; Quintana et al., 2007; Thébault et al., 2008, 2009). Encouragingly, 20% of detected exoplanets have been found in binary star systems orbiting one or the other star. An especially interesting case is the binary star HD 196885 AB. With a semi-major axis of 24 AU and an eccentricity of 0.409, this is a close analog of the α Cen AB binary pair. HD 196885 A is known to host a gas-giant planet with $M\sin(i)$ of $\sim 3\ M_{Jup}$ and an orbital period of 3.69 years (Correia et al., 2008; Fischer et al., 2009; Chauvin et al., 2007).

Doppler surveys generally avoid binary stars with separations less than ~5 arcseconds because additional RV errors can be incurred by flux contamination from the companion star. At the next periastron passage of α Cen AB (May 2035) the projected separation of the two stars will be less than 2 arcseconds. However, with an orbital plane that is only 11 degrees from an edge-on configuration, the projected separation of α Cen AB reached a secondary minimum of $\sim$ 4 arcseconds in 2017. Figure 2.1 shows the relative orbit of α Cen B orbiting α Cen A, projected onto the plane of the sky. Beginning in 2012, the angular separation between the two stars decreased to 5.44" and flux contamination from the binary-star companion was observed in the radial-velocity measurements and was exacerbated on nights of poor seeing conditions.

For the *CHIRON* data, while there was code developed to scale the flux taking into account contamination, the improvement was insufficient for precision radial velocity measurements. The RVs listed in the Dumusque et al. (2012) paper were restricted to observations obtained through 2011 that had better than one arcsecond seeing. No *HARPS* radial velocities were published for 2012 because the seeing conditions were not adequate to avoid flux contamination during that year. Wittenmyer et al. (2014) and Bergmann et al. (2015) have presented methods for modeling flux-contaminated spectra to reach an rms of a few meters per second. However, this more complex modeling does not reach sub-meter-per-second precision, the precision needed to contribute to the detection of planets with velocity semi-amplitudes less than one or two meters per second. The current small projected angular separation of α Cen AB may force a hiatus in ground-based Doppler programs for this system until 2019 or 2020.

2.3.1 Constraints from Existing Data

The existing Doppler planet searches allow us to place constraints on the mass-period parameter space where planets would have been detected if they existed. Conversely, we can see what type of planets would have escaped detection.

Data from the Echelle Spectrograph (ES), *CHIRON*, *HARPS*, and *UVES* were compiled to constrain exoplanet detections for α Cen A, B, and C (Proxima). Radial velocities of both α Cen A and B were obtained by our team using the ES between 2008 - 2011 and the *CHIRON* spectrograph between 2011 - 2013 at the 1.5-m Cerro Tololo Interamerican Observatory (*CTIO*) in Chile. The *HARPS* spectrograph is located at the 3.6-m ESO La Silla telescope. *HARPS* radial velocities of α Cen B were obtained between February 2008 and July 2011 and published by Dumusque et al. (2012). We also use published RVs of Proxima Centauri from *HARPS* that span 2005 - 2016, and published RVs from 2010 - 2016 at the Ultraviolet and Visual Echelle Spectrograph (*UVES*) on the Very Large Telescope at Cerro Paranal in Chile (Anglada-Escudé et al., 2016). Both *CHIRON* and *UVES* are calibrated using the iodine cell technique while *HARPS* is calibrated using the simultaneous Thorium-Argon reference method Tokovinin et al. (2013); Anglada-Escudé et al. (2016).

The 1.5-m *CTIO* telescope is part of the Small to Moderate Aperture Research Telescopes (SMARTS) consortium. The ES was a recommissioned, fiber-fed spectrograph located at the 1.5-m *CTIO* telescope. The typical, single-shot precision of the ES was about 7 m s^{-1}. This spectrograph was replaced in 2011 with the *CHIRON* spectrograph, which was immediately upgraded and recommissioned in 2012 with new optical coatings, a new CCD, better temperature control, and octagonal fibers (Tokovinin et al., 2013). While the short term velocity rms reached 0.5 m s^{-1} for bright, single stars observed with *CHIRON* (Tokovinin et al., 2013), a serious short-coming for RV measurements of α Cen A and B RV measurements is that the front end fiber feed was designed with a 2."7 field of view to maximize the number of collected photons during poor seeing conditions on the 1.5-m telescope. When *CHIRON* was re-commissioned in 2012, the angular separation of α Cen A and B was only 5."5 and there was significant flux contamination from the companion star on nights when the seeing was worse than one arcsecond. By 2013, the angular separation of AB had decreased so the flux contamination increased and there were few nights when the rms of the RV measurements was less than three times the average scatter per night. We tested a new Doppler code that included a scaled flux from the companion star, as described by Bergmann et al. (2015); however, we were only able to reach a single-shot precision of $\sim$ 15 m s^{-1} from the flux contaminated spectra for α Cen A and B. We prefer to retain the original velocities, rather than velocities from our scaled flux analysis, because they more clearly identify nights with spectral contamination that should be rejected.

Figure 2.2 shows all of the binned RV measurements collected by the ES (left of the vertical dashed line) and *CHIRON* (right of the vertical dashed line) for α Cen A (top panel) and α Cen B (bottom panel). Flux contamination from the companion stars causes the velocities for α Cen A to decrease (shifting toward the velocity of α Cen B) and velocities for α Cen B to increase. The effect of flux contamination is apparent in Figure 2.2. The nights with poor seeing conditions that resulted in flux contamination were excluded from the published *HARPS* data (Dumusque et al., 2012). To eliminate nights at the *CTIO* with significant flux contamination, we determined an acceptable threshold for the measured contamination. After subtracting out the binary trend from both data sets, the resulting RV measurements should be Gaussian distributed about zero. However, contaminated data will lie far away from the mean. We fit a Gaussian curve to the distribution of each data set and consider any data point more than 3σ away from the distribution's mean as suffering from considerable contamination. Figure 2.3 shows the distribution (black), fitted Gaussian curve (blue), and subsequent cuts (orange). The 3σ cuts frame the bulk of the observations, thereby excluding only nights that deviate significantly from the mean.

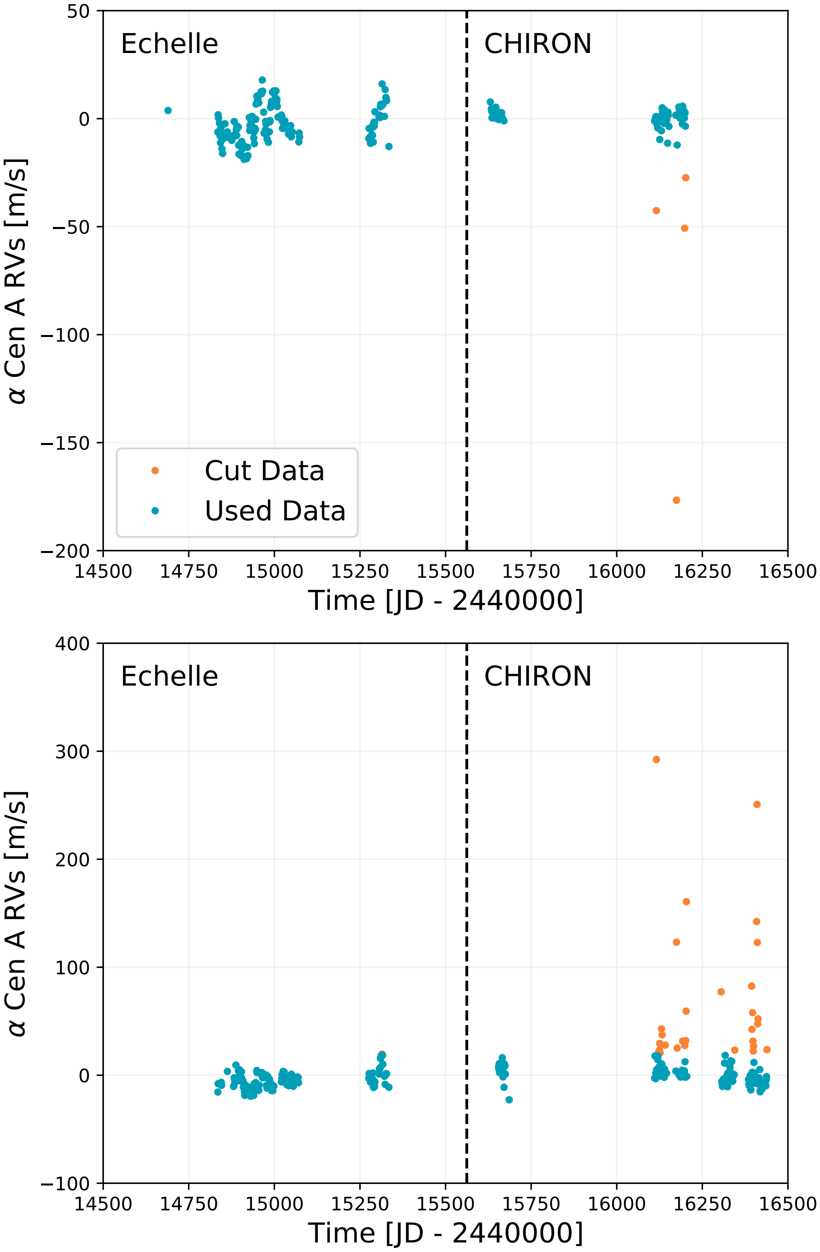

Figure 2.2: Data of α Centauri A (top) and α Centauri B (bottom) taken at the CTIO Telescope from 2009 to 2012. The Echelle Spectrograph was switched to CHIRON in 2011, as shown by a dashed line on each graph. Observations are binned by night. Blue points represent the data points used in the simulations. Observations falling more than 3σ away from the average due to contamination were cut, shown here in orange.

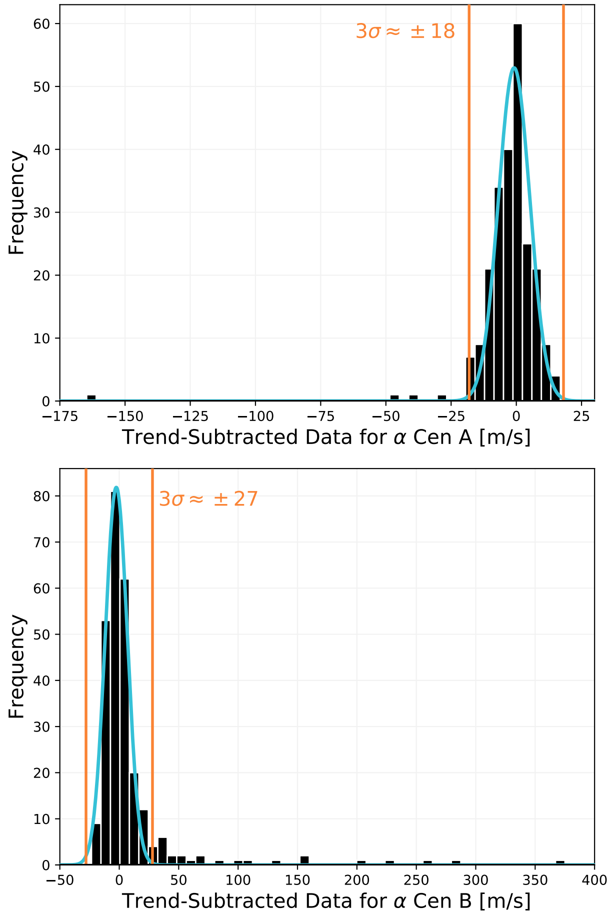

Figure 2.3: A histogram of the de-trended radial velocity measurements for each night at the *CTIO* Telescope of α Cen A (top) and α Cen B (bottom). Each histogram is fit to a Gaussian, shown in blue. Nights where data fall more than 3σ away (shown by red, vertical lines) most likely suffer contamination from the other star and are cut. Retained nights are shown in blue in Figure 2.2.

Table 2.2: Relative, Binned RV Data from the Telescope. A stub of this table is provided in the printed version of this paper and the complete table is available in the online version of this paper

Star	JD-2440000	RV m s^{-1}	Err m s^{-1}	Source
A	14689.5270	-239.82	7.20	Echelle
A	14834.8477	-187.13	4.64	Echelle
B	14834.8350	119.94	4.05	Echelle
B	14835.8154	126.48	4.74	Echelle

Data that is retained and used in the simulations are shown in Figure 2.2 in blue while cut data is plotted in orange. This contamination is visually obvious and increases with time following 2011 as the stars orbit closer and closer together (see Figure 2.1). All velocities removed are skewed in the direction to be expected from contamination (e.g. down for α Cen A and up for α Cen B). Additionally, this choice of cut vets more data points from the set of α Cen B observations, which further suggests contamination since α Cen A is the brighter star and would therefore cause more significant contamination in α Cen B observations than the other way around.

In Table 2.2, we list the nightly-binned, radial-velocity measurements on nights where there was not significant flux contamination from the companion star for α Centauri A and B, resulting in 228 data points for α Cen A and 241 points for α Cen B. The uncertainty on our single measurements is of the order 5 m s^{-1} for ES data and 1.1 m s^{-1} for *CHIRON* data with regards to α Cen A. The uncertainty for α Cen B observations is approximately 4.4 m s^{-1} for ES data and 1.2 m s^{-1} for *CHIRON* data. Because the errors are not pure white noise, the error for each binned observations is taken to be the average of the formal errors of every point that night. The rms of the final, nightly-binned radial velocities is 7.2 m s^{-1} for α Cen A over 4.13 years and 8.9 m s^{-1} for α Cen B over 4.38 years.

2.4 Simulations

Using the cleaned and nightly-binned velocities for α Cen A and B from the ES and *CHIRON* spectrographs at *CTIO*, the published *HARPS* velocities for α Cen B, and the *HARPS* and *UVES* velocities for Proxima Cen, we carried out Monte Carlo simulations to assess whether planets of a given mass with orbital periods between 2 and 1000 days would have been detectable. The maximum orbital period of 1000 days was chosen because we expect that dynamical influences from the binary orbit of the α Cen AB stars would destabilize orbits of putative planets beyond about 2 AU (Wiegert & Holman, 1997). We restricted the detectability simulations for Proxima Centauri (α Cen C) to the same time baseline, searching for significant signals well beyond the habitable zone of the low-mass star. The minimum orbital period of 2 days is arbitrary, but avoids spurious 1-day sampling aliases in the *CHIRON* and *HARPS* data sets.

For the detectability simulations, we established a grid in planet mass and orbital period parameter space for each of the stars (α Cen A, B, and Proxima Centauri) and injected a simulated Keplerian signal at each grid point, adopting stellar mass values from Pourbaix & Boffin (2016). For simplicity, our simulations assume circular orbits and single-planet architectures. Grid points are spaced on a hybrid log-linear scale to adequately sample the parameter space.

The simulations reveal the detectable $M \sin(i)$ of the planet. While our simulations test $M \sin(i)$ rather than planet mass, planets are expected to inherit the 79° orbital inclination of the binary star system (Wiegert & Holman, 1997). Therefore, we expect that the $M \sin(i)$ value is close to the true mass of any planets around either α Cen A or B.

The statistical significance of the injected signal was determined by assuming the null hypothesis. In other words, we assess the probability that a signal of similar strength to our injected Keplerian signal would be produced by random errors in our data. Planets that are more massive or in closer orbits will produce stronger reflex velocities in the host stars that give rise to stronger, coherent signals. These planets are more easily detected as their signals are harder to reproduce by noise alone. Our simulations test what strength of signal is necessary to overcome the inherent noise in the data and produce a coherent, detectable signal from a planet. We tested planet detectability in the presence of both white noise and the red noise present in the reported RVs.

2.4.1 White-Noise Simulations

We simulated Keplerian RVs with identical temporal sampling and error bars as the observed data sets, preserving any window functions in the observations. Our white-noise simulations assume that the radial velocity scatter is completely captured by white noise that is scaled to the quoted error bars. The simulated radial velocities for the white-noise simulations were created with a random draw from a Gaussian distribution that was scaled to the formal error at the time of each observation. The mean of the formal errors for the binned ES and *CHIRON* data is 0.48 m s^{-1} for α Centauri A and 0.51 m s^{-1} for α Centauri B (but, as we show later, the systematic errors in the ES and *CHIRON* data are significantly larger). The binned *HARPS* data of α Centauri B have a mean error of 1.0 m s^{-1}. The standard error for the combined, binned *HARPS* and *UVES* data of Proxima Centauri are on average 0.94 m s^{-1}. We generated 1500 sets of time-series, white-noise RV data. The simulated radial velocities were created by adding realizations of white noise to theoretical Keplerian models at each mass-period grid point.

Using a Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982), the periodogram power for the simulated Keplerian velocities was then compared to the periodogram power of the 1500 white noise data sets at each grid point. The data sets that are dominated by white noise produce a power spectrum with multiple low peaks at many periods, while a detectable Keplerian signal will produce a tall peak at the correct orbital period. Examples

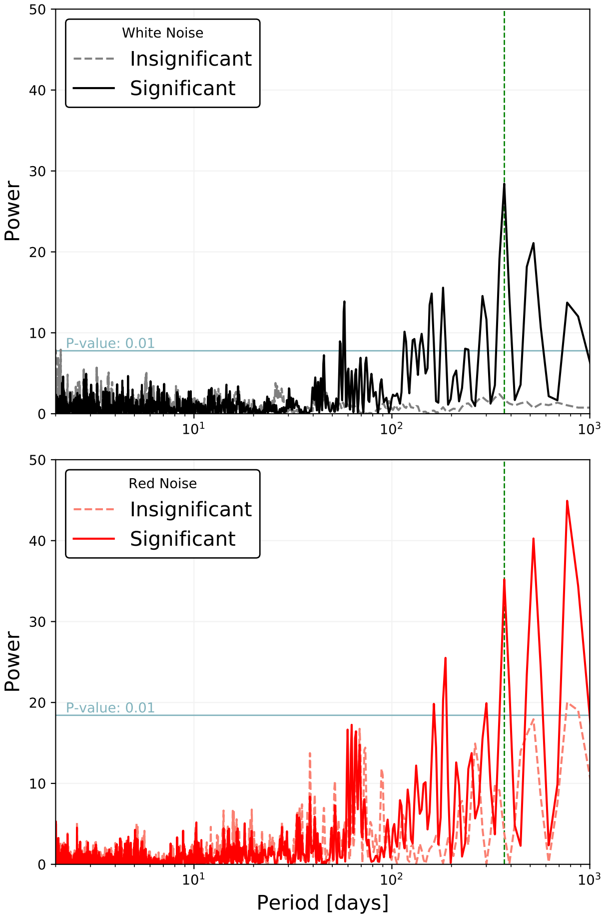

Figure 2.4: Comparing generated Keplerian signals to noise. An example periodogram of a significant detection (solid line) and an insignificant planetary signal (dashed-line) are given for white noise (black, top) and red noise (red, bottom). On both graphs, a blue, horizontal line marks the peak height that is greater than the maximum peak height in 99% of 1500 instances of pure noise. This corresponds to a p-value of less than 0.01 for the signal. Therefore, periodograms with peaks higher than this line are considered significant. The vertical, green, dashed line on both graphs marks the period of the generated signal.

of significant vs. insignificant periodograms are given in Figure 2.4 for both white noise (top) and red noise (bottom).

To decide whether the RVs with an injected Keplerian signal would be detectable with current observations, we calculate a p-value at each grid point. The p-value gives the probability that the injected radial-velocity data produces the same signal as only random noise. The p-value is defined as the fraction of comparisons where the white-noise simulations yield a greater maximum peak height than the simulated Keplerian signal. Planets producing significant signals will more consistently give stronger periodogram peaks, resulting in lower p-values. Larger p-values indicate that the signal produced by the planet has no more significance than white noise alone. We adopt an arbitrary but often used threshold p-value of 0.01, meaning that fewer than 1 of 100 white noise simulations produced a periodic signal that was stronger than a simulated Keplerian signal. [3]

Figure 2.5 shows the white noise detectability simulations for α Cen A, α Cen B, and Proxima Cen. The p-values and color gradients are scaled so that a boundary appears where Keplerian signals yield a p-value of 0.01. Signals with lower p-values (above this boundary) would have likely been detected if they existed, while Keplerian signals with p-values greater than 0.01 would be buried in the white noise given the stated errors of the *CHIRON* and *HARPS* programs. The *CHIRON* and *HARPS* data sets for α Cen B are kept separate as they are unique in their sampling and would lead to different aliasing as well as exhibit different instrumental errors. Analyzing the two data sets separately allowed our results to capture these differences. Additionally, while combining the two data sets helps to push white-noise detection limits lower, the red-noise simulations suffer instead. The *CHIRON* data, with more systematic errors, serve to reduce the sensitivity of the *HARPS* data rather than give better results. The upper right panel of Fig 2.5 has a dot indicating

---

[3] The purpose of this analysis is to assess the strength an injected Keplerian signal requires to produce a signal significantly distinct from what would be produced by pure noise in the current observations. We consider each scenario individually and therefore do not make any adjustments to account for the multiple comparisons problem.

the mass and period of Proxima Cen b (Anglada-Escudé et al., 2016).

2.4.2 Red-Noise Simulations

Our analysis using the white-noise simulations described above will not account for any systematic or quasi-periodic instrumental errors, analysis errors, photospheric jitter, or even actual planets. To investigate the impact of systematic errors or red noise sources, we treat the reported residual velocities from subtracting out the binary orbit from the observations as coherent noise. This is a worst-case scenario and we note that it is possible to improve detectability by de-correlating some of these noise sources using techniques like line bisector variations or FWHM variations to estimate photon noise (e.g. Dumusque et al., 2012; Rajpaul et al., 2016; Anglada-Escudé et al., 2016).

These residual velocities are assumed to capture uncorrected observational errors, including instrumental errors and stellar jitter. The residual velocities would also contain any potential planetary signals. For this red noise simulation, we simply interpreted the residual velocities as pure red noise, and continued the same simulation described for white noise, adding Keplerian signals parameterized by each point of a mass-period grid. A comparison of the red and white noise analysis can be useful for highlighting possible planetary signals as well as quasi-periodic errors in our radial velocity data.

For the Monte Carlo, red noise simulations, the radial-velocity residuals were added to 1500 freshly generated white noise realizations and to the theoretical Keplerian signal at each point of a mass-period grid. The periodograms of the red-noise simulations now contain stronger power than the white-noise simulations, meaning the simulated Keplerian signal must generally have a larger amplitude to reach a p-value of 0.01 (see Figure 2.4). Figure 2.6 shows the red noise simulations for α Cen A from *CHIRON* (upper left), Proxima Centauri (upper right), α Cen B from *CHIRON* (lower left) and α Cen B from *HARPS* (lower right). Solid, black lines on each plot show a power law that was fit to the

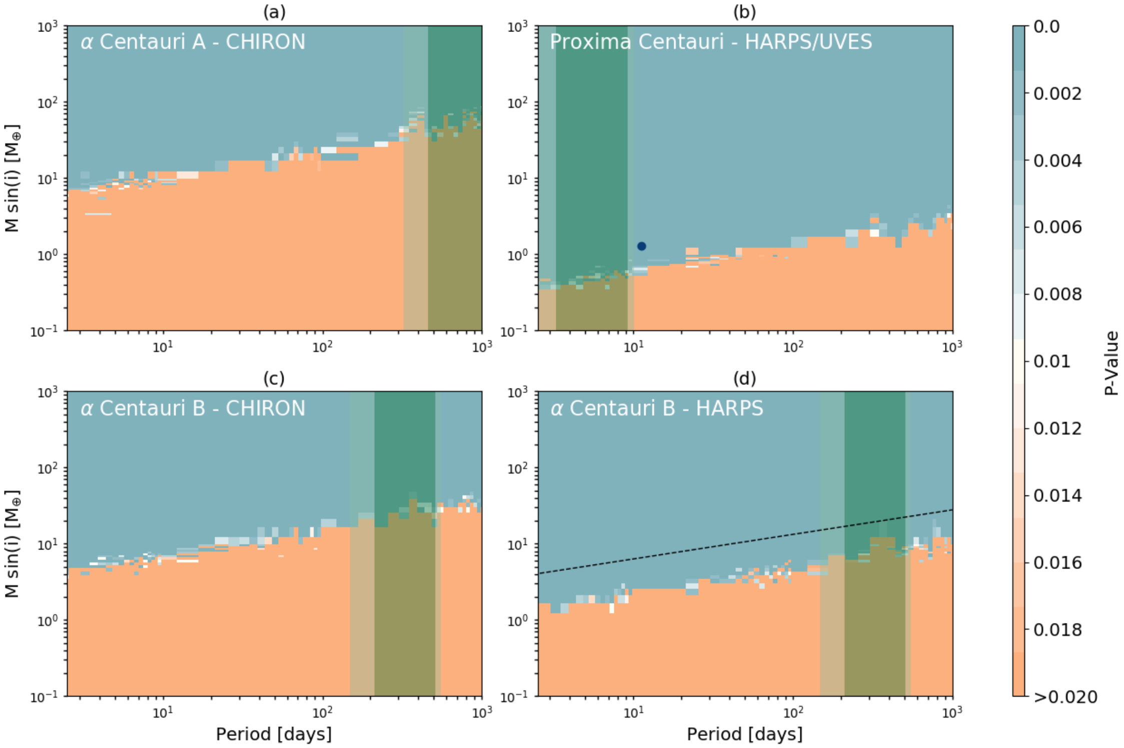

Figure 2.5: White-noise simulations. Mass vs. period grids showing the significance at which a planet of such a mass and period would have been detected assuming only the reported errors for observations of (a) α Centauri A from ES and *CHIRON*, (b) Proxima Centauri from *HARPS* and *UVES*, (c) α Centauri B from ES and *CHIRON*, and (d) α Centauri B from *HARPS*. A p-value of less than 0.01 (indicated by shades of blue) is considered significant. Green vertical bands mark the conservative habitable zone where liquid water could persist for most of the stellar lifetime and the lighter green covers the optimistic habitable zone (as defined by Kopparapu et al. (2013)). A power law was fit to the detectability border of the α Cen B ES and *CHIRON* data and is plotted on the α Cen B *HARPS* grid as a dashed line. The location of Proxima Cen b is indicated with a dot.

detectability border from the white-noise simulations for each data set.

2.5 Results

Our white-noise simulations are summarized in Figure 2.5. Because both the original error bars and times of observation are retained, the white-noise simulations will preserve our ability to identify window functions in the sampling of our data. These simulations exclude planets in the conservative habitable zone of each star with a $M \sin(i)$ of greater than 53±13 $M_\oplus$ for α Cen A, 8.4±1.5 $M_\oplus$ for α Cen B, and 0.47±0.08 $M_\oplus$ for Proxima Centauri. However, this is an overly optimistic scenario. Doppler measurements are known to have contributions to the derived radial velocities that arise from instability of the instrument, errors in the analysis, and velocities in the stellar photosphere from spots, faculae, granulation, p-mode oscillations, or meridional flows (e.g., Santos et al., 2000; Saar & Fischer, 2000; Queloz et al., 2001; Wright, 2005; Lagrange et al., 2010; Meunier et al., 2010; Borgniet et al., 2015). These velocities can obscure the Doppler signals that arise from orbiting exoplanets. The white noise simulations will not capture these noise sources because of an implicit assumption that the measurement errors are captured by the formal RV uncertainties.

In contrast, our red-noise simulations, shown in Figure 2.6, represent a worst-case scenario. For these simulations, we assume that the time-series radial velocities contain only coherent noise. This noise is added directly to random white noise and the generated Keplerian signals, effectively preserving any temporal coherence in the noise.

In practice, radial velocities can be treated with Gaussian Process Regression (Rajpaul et al., 2016) or decorrelated using line bisectors or the FWHM of the cross correlation function (Dumusque et al., 2012) to mitigate the impact of non-Keplerian radial velocities on exoplanet detectability. Red noise has also been empirically modelled (Tuomi et al., 2013) Therefore, our red-noise simulations slightly underestimate planet detectability. The

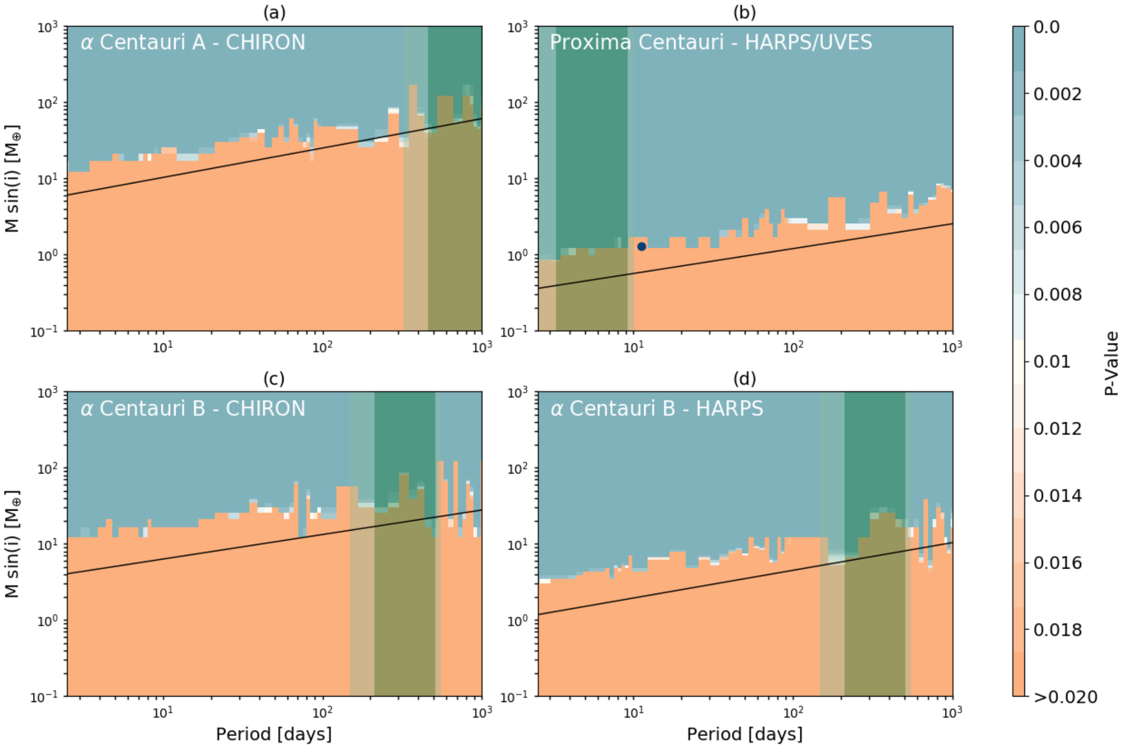

Figure 2.6: Red-noise simulations. Mass vs. period grids showing the significance at which a planet of such a mass and period would have been detected assuming that the current data of (a) α Centauri A from ES and *CHIRON*, (b) Proxima Centauri from *HARPS* and *UVES*, (c) α Centauri B from ES and *CHIRON*, and (d) α Centauri B from *HARPS* is simply red noise. The color scale to p-value is the same as for Figure 2.5. A power law was fit to the detectability border given by the white-noise simulations and is plotted here as a black line. The orange parameter space indicates areas where planets could still remain undetected. The conservative and optimistic habitable zones are the same as Fig 2.5. Proxima Centauri b is indicated on subfigure (b) by a dot.

white-noise and red-noise simulations together frame the mass-period boundary where existing Doppler surveys constrain the existence of planets orbiting the α Centauri stars.

The boundary between parameter space where planets would be detected or missed in the presence of white noise (Fig. 2.5) or red noise (Fig. 2.6) is approximately defined by the lowest mass at each period for which the p-value of the generated Keplerian RV exceeds our threshold of 0.01. This border for the white-noise simulations is both lower and smoother compared to the red-noise simulations. To more closely investigate these differences, we subtract the detectability border of the white-noise simulations, M_{WN}, from the detectability border of the red-noise simulations, M_{RN}. This difference is plotted as a function of period in Figure 2.7 for α Cen A (top), α Cen B (middle), and for Proxima Cen (bottom). Peaks in this difference plot will occur due to quasi-periodic noise sources or planetary signals. Interesting to note is the peak at around 675-720 days that can be seen in both the *CHIRON* and *HARPS* observations of α Cen B. Additionally, similar peaks appear in both the α Cen A and α Cen B *CHIRON* data near 65, 150, and 575 days. The bottom plot in Figure 2.7 includes a vertical, dashed line at the period of Proxima Centauri b, around which a clear peak can be seen.

2.6 Discussion

2.6.1 Detectability

We have carried out simulations to show how past Doppler surveys of the α Centauri stars constrain the probability of exoplanets over the mass-period parameter space shown in Figures 2.5 and 2.6. While the Doppler technique can only derive $M \sin i$, rather than the true planet mass, the dynamical influences of the binary star system mean that any stable planets are more stable and therefore more likely to be nearly co-planar with the 79° inclination of the stellar binary system (Wiegert & Holman, 1997; Quarles & Lissauer, 2016). This

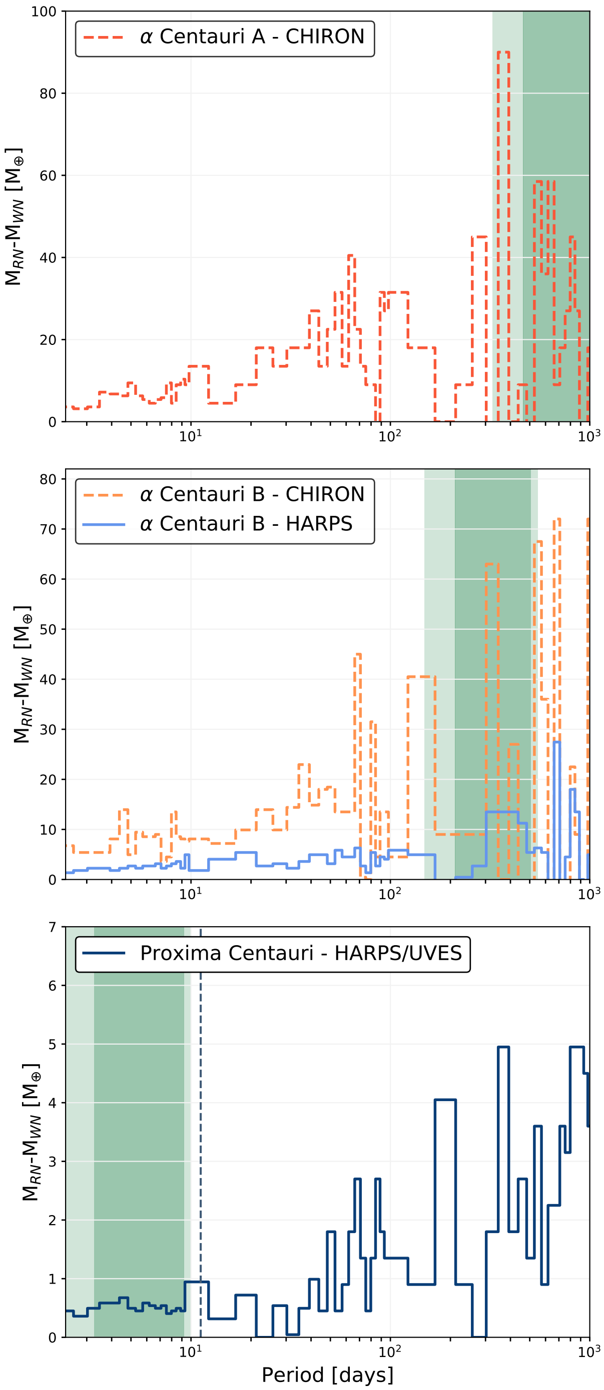

Figure 2.7: Difference in detectability between the red-noise simulations and the white-noise simulations from subtracting the two detectabliity borders for α Centauri A (top), α Centauri B (middle) and α Centauri C (bottom). Peaks indicate periodicities in the residual radial velocities that could correspond to stellar noise, systematic errors, or even planetary signals. Negative differences are not shown. We assume negative values to arise from over-fitting the data or white noise alone and so hold no physical meaning. A blue, dashed line in the bottom, Proxima Centauri plot indicates the orbital period of Proxima Cen b.

suggests that the $M\sin(i)$ is approximately the actual planet mass for prospective planets around α Cen A or B. We show that Earth analogs could still exist around either α Cen A or B and would not have been detected by the past decade of precision radial velocity searches. Continued, high-cadence, high-precision radial velocity observations could still reveal Earth-sized planets within this star system, even within the habitable zones of each of the three stars.

At each point in the parameter space of $M\sin(i)$ and orbital period, we sample a Keplerian signal at the actual time of the observations with added white noise scaled to the errors to provide a baseline of planet detection space. These simulations exclude planets within the conservative habitable zone of each planet with a $M\sin(i)$ of greater than 53 $M_\oplus$ for α Cen A, 8.4 $M_\oplus$ for α Cen B, and 0.47 $M_\oplus$ for Proxima Centauri on average. This result for α Cen B comes from the *HARPS* data set; the *CHIRON* data set excludes planets in the habitable zone of α Cen B to greater than 23.5 $M_\oplus$. We then repeat our analysis using the actual velocity scatter after subtracting the binary star orbit as "red" noise in addition to the white noise. We assess the probability that this signal could have been produced by noise alone by calculating a p-value, the fraction of comparisons where noise-only simulations yield greater periodogram power than the Keplerian signal at any grid point. The color scales of Figures 2.5 and 2.6 pivot around a p-value of 0.01, which would be marginally detectable.

Both the white-noise and the red-noise simulations preserve the cadence of observations. Because observations are a discrete sampling of a continuous signal, aliases appear in periodograms that can be mistaken for true, astrophysical signals. These commonly correspond to periodicities of the sidereal year, sidereal day, solar day, and synodic month (Dawson & Fabrycky, 2010). For example, reduced sensitivity can be seen in all four data sets presented in Figure 2.5 around 300-400 days, which likely corresponds to the annual constraints the Earth's orbit around the Sun places on observations.

We subtract the white-noise mass-period boundary that occurs at a p-value of 0.01

from that same boundary in the red-noise simulations to highlight periodicities present in the residual radial velocities. Even after implementing the cuts described in section 3.1, we acknowledge that some of the remaining RV measurements may still be affected by small amounts of contamination, which effectively contributes to the red noise. The peaks apparent in Figure 2.7 could correspond to quasi-periodic systematic errors, stellar jitter, or even planetary signals. For example, peaks that appear consistently in the *CHIRON* data for both α Cen A and α Cen B (e.g. at 65, 150, and 575 days), are indicative of instrumental or systematic errors since it is improbable that both stars will exhibit the same astrophysical velocity signals. Also potentially interesting are the periods where peaks in the *CHIRON* and *HARPS* data of α Cen B align (e.g. at around 700 days). Because these peaks appear in observations from two different instruments with different data reduction pipelines, it seems unlikely that the same peaks would arise in both data sets from instrumental or systematic error; however, these peaks could still be the result of astrophysical velocity signals. In the case of Proxima Centauri, it is illustrative to note a distinct signal at the period of the recently discovered Proxima Cen b. A Keplerian signal would produce a red noise source in the velocities of that star; this peak is likely due to the signal produced by Proxima Cen b that is retained in the residual radial velocities.

Radial velocity precision approaching 10 centimeters per second will ultimately be needed to detect exoplanets with smaller masses and longer orbital periods in the yet to be probed parameter space around α Cen A and B. There are several challenges for reaching such high RV precision. Some of the issues should be relatively straightforward to address. For example, the *p*-mode oscillations of α Cen A have a radial velocity amplitude of $1 - 3$ m s^{-1} (Butler et al., 2004),which adds random scatter to radial velocities. The *p*-mode amplitudes in α Cen B are much weaker with a semi-amplitude of only 0.08 m s^{-1} (Kjeldsen et al., 2005); however, changing granulation patterns also introduce radial velocity scatter at the level of 0.6 m s^{-1} for both stars on timescales ranging from 15 minutes to several hours (Dumusque et al., 2012; Del Moro, 2004). Both *p*-mode oscillations and granulation change on relatively short periods, allowing the observing strategy to be tailored to dramatically reduce these noise sources. For example, a series of exposures over ten minutes is sufficient to average over the high frequency *p*-mode signals.

Additional challenges to higher RV precision include requirements of higher stability for next generation spectrographs (temperature and pressure stability), improved wavelength calibration, calibration of both CCD stitching and random pixel position errors, and mitigation of modal noise for multi-mode fibers (Fischer et al., 2016). It seems likely that ongoing efforts to address these engineering challenges will be successful. Techniques for modeling or decorrelating Doppler velocities that arise from stellar photospheres are less mature. Significant progress on disentangling stellar noise sources is required so that clean orbital velocities can be obtained. Currently, the two stars are separated by less than 5”, giving rise to cross contamination between the two stars and preventing high-precision, radial-velocity measurements. As the separation between α Cen A and B begins to increase in 2019, radial velocity measurements will help to push constraints even lower and could ultimately lead to the discovery of Earth-like planets.

2.7 Acknowledgements

We gratefully acknowledge the anonymous referee whose insightful and specific comments served to significantly improve the paper. DAF gratefully acknowledges support from NASA NNH11ZDA001 and MJG thanks the NASA Earth and Space Fellowship 13-ASTRO13F-0011. LLZ gratefully acknowledges support from the NSF GRFP. B.R-A acknowledges the support from CONICYT PAI/CONCURSO NACIONAL INSERCIÓN EN LA ACADEMIA, CONVOCATORIA 2015 79150050. Based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme 092.D-0300(A). Based on data obtained from the ESO Science Archive Facility under request number 275002. We thank Tom Blake and the

PNNL EMSL for obtaining the FTS scan of our iodine cell. We also thank contributors to Matplotlib, the Python Programming Language, and the free and open-source community.

Facilities

Pacific Northwest National Labs (PNNL) Environmental Molecular Sciences Laboratory (EMSL), CTIO:1.5m

Chapter 3

An Extreme Precision Radial Velocity Pipeline: First Radial Velocities from EXPRES

Ryan R. Petersburg, J. M. Joel Ong, Lily L. Zhao[1]*, Ryan T. Blackman, John M. Brewer, Lars A. Buchhave, Samuel H. C. Cabot, Allen B. Davis, Colby A. Jurgenson, Christopher Leet, Tyler M. McCracken, David Sawyer, Mikhail Sharov, René Tronsgaard, Andrew E. Szymkowiak, Debra A. Fischer*

The EXtreme PREcision Spectrograph (*EXPRES*) is an environmentally stabilized, fiber-fed, $R = 137,500$, optical spectrograph. It was recently commissioned at the 4.3-m Lowell Discovery Telescope (LDT) near Flagstaff, Arizona. The spectrograph was designed with a target radial-velocity (RV) precision of 30cm s^{-1}. In addition to instrumental innovations, the *EXPRES* pipeline, presented here, is the first for an on-sky, optical, fiber-fed spectrograph to employ many novel techniques—including an "extended flat" fiber used for wavelength-dependent quantum efficiency characterization of the CCD, a flat-relative optimal extraction algorithm, chromatic barycentric corrections, chromatic calibration offsets, and an ultra-precise laser frequency comb for wavelength calibration. We describe the reduction, calibration, and radial-velocity analysis pipeline used for *EXPRES*

---

[1] Originally published as: Petersburg, R. R., Ong, J. M. J., Zhao, L. L., et al. 2020, The Astronomical Journal, 159, 187. I co-wrote and developed the *EXPRES* pipeline along with the other two co-first authors and continue to help with upkeep and running of the pipeline.

and present an example of our current sub-meter-per-second RV measurement precision, which reaches a formal, single-measurement error of 0.3 m s^{-1} for an observation with a per-pixel signal-to-noise ratio of 250. These velocities yield an orbital solution on the known exoplanet host 51 Peg that matches literature values with a residual RMS of 0.895 m s^{-1}.

3.1 Introduction

Results from the NASA *Kepler* mission show that small planets with radii between 1–4 $R_\oplus$ are found orbiting 20–50% of main-sequence stars (Winn & Fabrycky, 2015). While such transit surveys, such as *Kepler*, *K2*, and *TESS*, have revealed a wealth of planets, few of these planets have had their masses measured; mass estimates which do exist are typically radial-velocity (RV), dynamical masses. More precise RV measurements are required to determine mass estimates for these planets, particularly small rocky ones, than are possible with pre-existing RV spectrographs. Were they available, these mass measurements would shed light on planetary structure, bulk density, and the mass-radius relation for sub-Neptune-mass planets.

To meet these needs, a new generation of RV spectrographs is now emerging: the Echelle SPectrograph for Rocky Exoplanets Search and Stable Spectroscopic Observations (*ESPRESSO*: Pepe et al., 2013) and The EXtreme PREcision Spectrograph (*EXPRES*: Jurgenson et al., 2016) are now on sky taking data, while the NN-explore Exoplanet Investigations with Doppler spectroscopy spectrograph (*NEID*: Schwab et al., 2016b) is in the commissioning phase. These new extreme-precision radial velocity (EPRV) spectrographs are driving towards the sub-10-centimeter-per-second radial-velocity precision needed to detect a true Earth twin around a Sun-like star.

Extremely precise spectrographs in turn demand extremely precise data reduction pipelines. The fidelity of the data and error estimates returned by these data reduction pipelines is paramount if groundbreaking discoveries returned from such instruments are to be credible. To this end, a complete science reduction, extraction, and analysis pipeline was newly developed and tailored for *EXPRES* data. In what follows, we begin with a brief description of the instrument and the calibration strategy. We then describe the analysis performed by our pipeline on *EXPRES* data, and present the first radial velocity measurements from the instrument.

3.2 Instrument Description

The EXtreme PREcision Spectrograph (*EXPRES*) is a fiber-fed, white-pupil EPRV spectrograph with a design resolution of $R = 150,000$ and a wavelength range of 3800 – 7800 Å. In practice, Blackman et al. (2020) show that the median resolution is better characterized as $R \sim 137,500$ with a maximum resolution reaching $R \sim 150,000$ in some regions of the detector. *EXPRES* is environmentally stabilized in a vacuum enclosure and is situated at Lowell Observatory's 4.3-m Lowell Discovery Telescope (*LDT*) near Flagstaff, Arizona. The multi-instrument port configuration of the *LDT* allows for high-cadence, flexible scheduling of stars (up to 280 partial nights per year).

Owing to various changes to the hardware configuration during commissioning, we have divided radial velocities from *EXPRES* into different calibration epochs, with an independent RV offset fitted for each epoch. Each epoch demarcates changes introduced to either the configuration of the echellogram on the CCD, the shape of the PSF, or the stability of our calibration sources. A summary of the changes that delineate the beginning of each epoch can be found in Table 3.1.

The inputs to *EXPRES* are a $33\times132\mu$m science fiber as well as a $60\times180\mu$m extended fiber. The extended fiber is wider than the science fiber in the cross-dispersion direction, allowing for higher SNR flat-fielding for all pixels illuminated by light from the science fiber, particularly at the cross-dispersion edges. Both fibers are also supplemented by their

Table 3.1: Instrumental Epochs

Epoch	Start	End	Changes Before Epoch Start
0	-	2018-04-15	Commissioning
1	2018-04-15	2018-06-15	Increased CCD pre-settle time
2	2018-06-15	2018-11-08	Fiber change; CCD rotation
3	2018-11-08	2019-02-07	Original science fiber replaced; Calibration unit rebuilt
4	2019-02-07	2019-08-04	LFC beat frequency mitigated; fiber agitator repaired
5	2019-08-04	-	realignment of FEM; replacement of LFC PCF

own square simultaneous fibers, though these are not used in normal operation.

The flat field light source is provided by a custom solution: a collection of 25 LEDs integrated on a single compact chip. The wavelength range of the LEDs cover the entire bandwidth of *EXPRES*, and the relative power of each LED is tuned to approximately match the inverse response of the spectrograph. Light from these LEDs, averaging 12.5 W, is coupled into an integrating sphere and subsequently injected into both the extended and science fibers.

Wavelength calibration is carried out with a Menlo Systems laser frequency comb (LFC; similar to those in Steinmetz et al., 2008; Probst et al., 2016). Three LFC exposures are taken through the science fiber every 15-30 minutes throughout a night. Two ThAr exposures are taken each night, once at the beginning and once at the end. The LFC is set to standby between each set of calibrations to reduce wear on nonlinear optical elements, and is only turned on approximately one minute before the next exposure is needed, to suppress turn-on transients.

Barycentric corrections are derived from the *EXPRES* exposure meter, a $R \approx 100$ spectrograph with an EMCCD detector and a bandpass that covers the spectral range of the LFC. During each science observation, the exposure meter takes a continuous series of 1 s exposures. Further technical details can be found in Blackman et al. (2019).

3.3 Analysis of *EXPRES* Data

The *EXPRES* pipeline is written in Python and makes heavy use of the SciPy stack (Virtanen et al., 2020). We show a schematic representation of it in Figure 3.1. In brief, the following steps are taken:

- At the start of each calibration epoch, several hundred extended flat images are taken. These are used to construct a master extended flat-field image, which is divided out from all exposures in the corresponding calibration epoch.
- Each night, 30 dark and 30 science flat images are taken. They are used to reduce and extract the science frames taken the same night.
- Echellogram orders are traced using the reduced science flats, a scattered light model is removed, and a flat-relative optimal extraction is performed.
- Wavelength solutions are interpolated for all science frames using bracketed LFC exposures, seeded by a nightly Thorium Argon source, as a calibration reference.
- Telluric lines are identified empirically with SELENITE (Leet et al., 2019) and labelled for later analysis.
- For exposures marked for RV analysis, we obtain radial velocities with two methods: cross-correlation against a line mask and a forward model.

We describe each of these steps in some detail below.

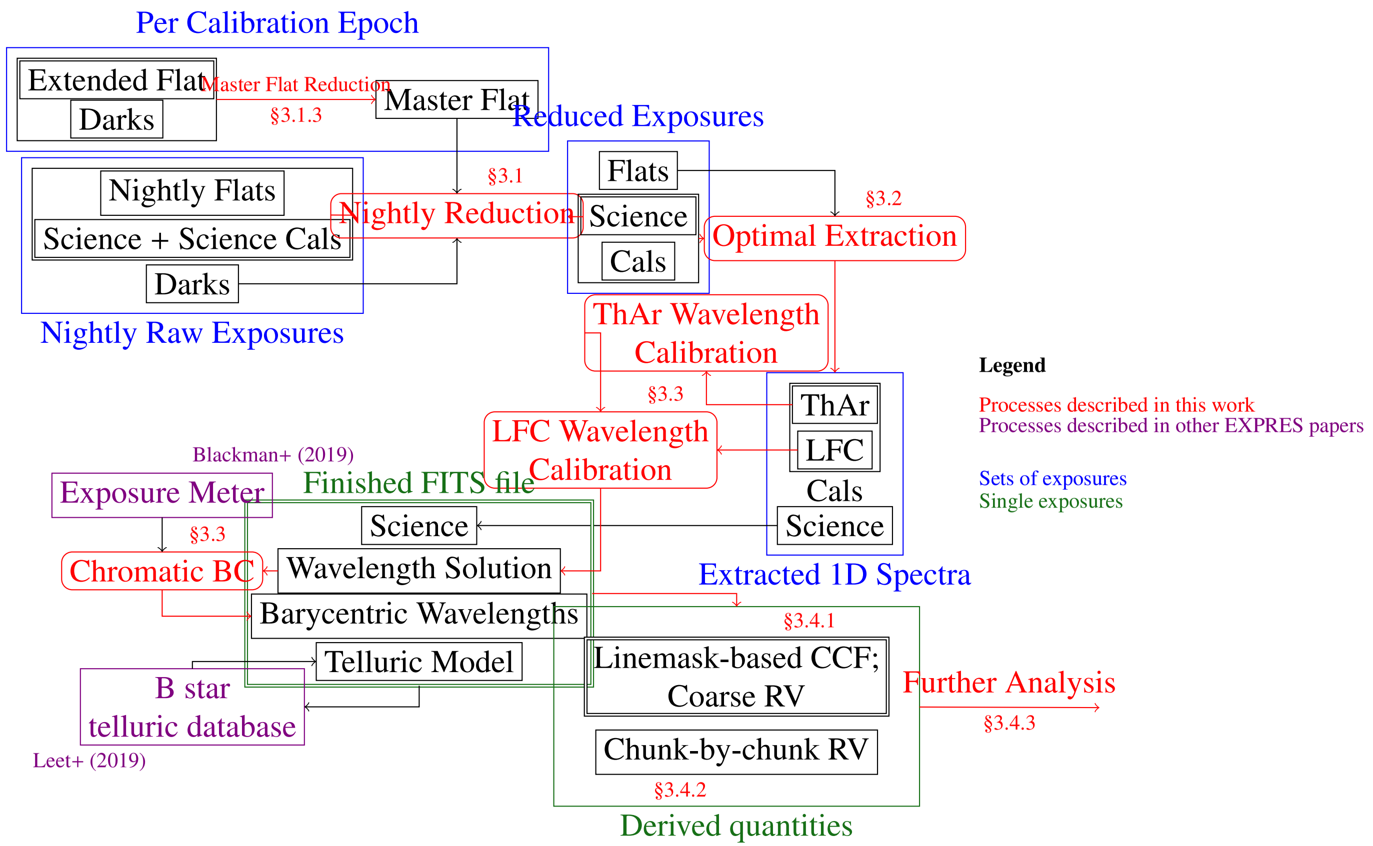

Figure 3.1: Data flow for the *EXPRES* pipeline. Small black boxes represent different kinds of exposures (if in blue boxes) or different data associated with an exposure (if in green boxes). Double-boxes indicate finished data products, ready for use in subsequent science analysis.

3.3.1 Reduction

The *EXPRES* detector is an STA1600LN CCD backside-illuminated image sensor with a $10,560 \times 10,560$ array containing $9\mu\text{m} \times 9\mu\text{m}$ pixels. The CCD is divided into 16 equal 5280×1320 pixel sections (two rows of eight) each with their own independent output amplifier and, thus, corresponding gain (see Table 3.2). Further information about the *EXPRES* CCD can be found in Blackman et al. (2020).

In the context of the *EXPRES* pipeline, “reduction” refers to the conversion of these 16 independent regions read in analog-to-digital units (ADU) to a full two-dimensional frame in units of photoelectrons. The reduction steps are as follows, dependent on the type of image being reduced (science, dark, science flat, or extended flat):

1. Subtract a bias frame constructed from the overscan regions (all).
2. Multiply each amplifier region by the corresponding gain coefficient (all).
3. Median combine calibration frames (dark, science flat, extended flat).
4. Subtract the reduced dark image (science, science flat, extended flat).
5. Divide by the reduced master extended flat (science, science flat).
6. Approximate the noise model using photon (Poisson) and read noise (science, science flat).
7. Trace the echelle orders (science flat).
8. Approximate the scattered light using a two-dimensional b-spline model (science, science flat).

In the following subsections, we go into detail about each of the above steps. Information about the reduction of exposure meter data can be found in Blackman et al. (2019).

3.3.1.1 Overscan

Each of the *EXPRES* CCD amplifier regions have overscans along both the serial (horizontal) and parallel (vertical) registers. The serial overscan is 5300 × 180 pixels along the right side of each amplifier region and the parallel overscan is 20 × 1320 pixels along the center line of the CCD (at the bottom of the upper regions and the top of the lower regions). Because these overscans contain virtual pixels read out by the same electronics as the real amplifier region, the *EXPRES* reduction pipeline uses them to approximate the bias of the CCD. The process is as follows:

1. Calculate the mean of the serial overscan region along its horizontal axis.
2. Smooth this mean using a cubic b-spline with knots every $\sim$ 100 pixels.
3. Correct the rows in the parallel overscan region by subtracting the overlapping smoothed serial overscan region.
4. Calculate the mean of the parallel overscan region along its vertical axis.
5. Smooth this mean using a cubic b-spline with knots every $\sim$ 100 pixels.
6. Construct a bias for each pixel in the amplifier region by summing the corresponding row from the serial overscan and column from the parallel overscan.

The mean and subsequent spline fit for the overscan regions of a single amplifier region are shown in Figure 3.2. This process is executed for all exposures, including darks.

3.3.1.2 Gain

The independent amplifier gains for the *EXPRES* CCD (Table 3.2) were determined empirically by matching the edges of neighboring amplifier regions based on stacked bias-subtracted extended flat images. We median combine 20 columns along each edge of two

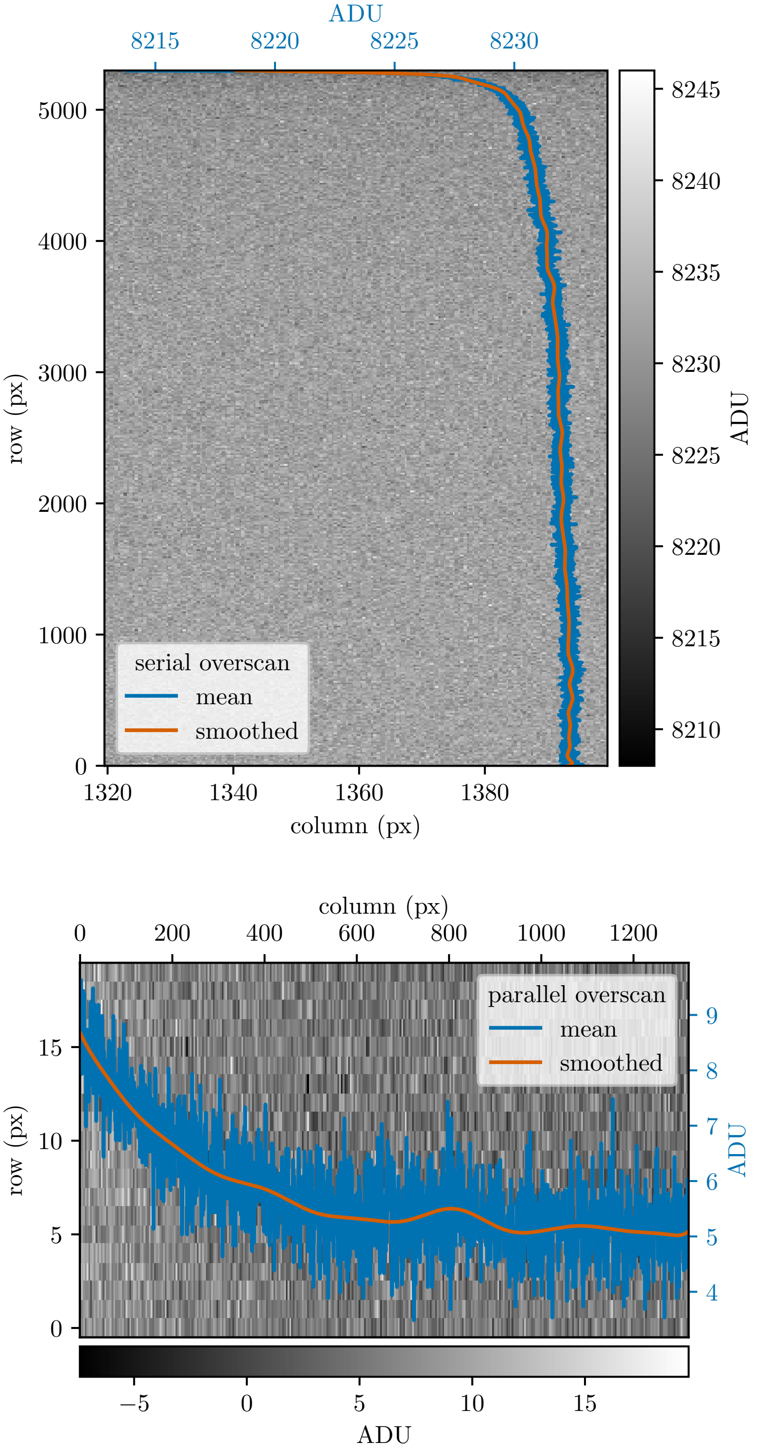

Figure 3.2: Bias approximation using the serial and parallel overscans. The mean and subsequent smoothed mean for each overscan are shown in blue and orange, respectively. Note that the counts shown in the parallel overscan were first subtracted by the first 20 rows of the serial overscan smoothed mean, generating values close to zero.

adjacent amplifier regions and determine the factor that minimizes the difference between them. We repeat this process along each row of amplifier regions, yielding a set of relative corrections for each row. Then, using the 20 median-combined rows along the center line of the detector, we determine the single factor that relates the top row of corrected amplifier regions to the bottom row.

Since these corrections are merely relative, we assume that the mean gain of the 16 amplifier regions given by the manufacturer is approximately correct. Thus, we match the mean of the empirically determined gain corrections to that given by the manufacturer. We note that the gains given by the manufacturer were not sufficient at matching the boundary conditions of the *EXPRES* CCD: some gain corrections were tweaked by more than 3%.

Table 3.2: *EXPRES* CCD Gain

Amplifier	Empirical	STA	Amplifier	Empirical	STA
0	2.57980	2.6645	8	2.69787	2.6945
1	2.55171	2.5352	9	2.65100	2.6422
2	2.53844	2.5218	10	2.64354	2.6367
3	2.52444	2.4065	11	2.60344	2.5502
4	2.51480	2.6024	12	2.60497	2.5571
5	2.49382	2.5686	13	2.59183	2.5630
6	2.52745	2.4816	14	2.62881	2.5691
7	2.53478	2.4960	15	2.72938	2.6111

We were not able to calculate the gains in the typical manner—relating the variance of each pixel to its mean—because our flat-fielding LED source has an intrinsic variability that cannot be modeled by photon noise alone. See Blackman et al. (2020) for further details.

3.3.1.3 Master Extended Flat

In order to measure the quantum efficiency (QE) variations with high signal across the relevant areas of the CCD, *EXPRES* employs an "extended flat" fiber. This rectangular fiber has slightly larger dimensions of $60 \times 180 \mu$m—as compared to the "science" rectangular fiber with dimensions 33×132μm—and is aligned with the center of the science fiber using a motor-driven mirror (see Figure 3.3. Periodically throughout the operation of *EXPRES*, and at least once per epoch (Table 3.1), the flat-fielding LED—typically used for order tracing and optimal extraction slit modeling—is injected into the extended flat fiber and a series of more than 100 images are taken of the resultant spectrum.

Using a median combination of these extended flat images, we construct a master extended flat-field image by dividing out a smooth fit to its echellogram. For each column of each order, we fit a parametric slit function: the squared convolution of a rectangle function with a Gaussian function, which has the analytic form

$$P_{x,y}(A, d, \sigma) = A\left[\Phi\left(\frac{y + d/2}{\sqrt{2}\sigma}\right) - \Phi\left(\frac{y - d/2}{\sqrt{2}\sigma}\right)\right]^2 \tag{3.1}$$

where Φ is the error function, d is the width of the rectangle, σ is the standard deviation of the Gaussian, and A is the amplitude.

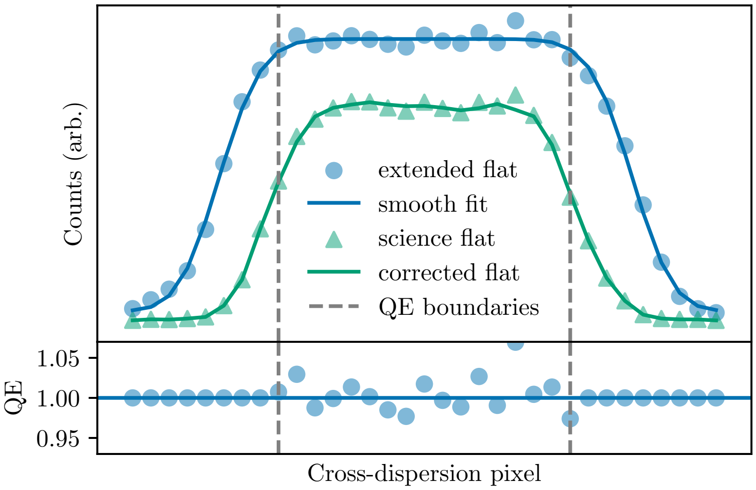

Figure 3.3: Example cross-dispersion, cross-section of an extended flat as compared to a science flat. The best fit of Equation (3.1) is given for the extended flat. The resultant quantum efficiency (QE) corrections are shown in the lower plot and the bounds of the corrected region are demarcated. Note that these are corrections *relative* to the mean QE. The QE corrected science flat is also shown in the upper plot.

This choice of functional PSF form was motivated by the physical nature of *EXPRES* and Fourier optics approximations. The input near field of the spectrograph is a rectangle function generated by the rectangular fiber. Taking the Fourier transform of the near field yields an approximation for the far field, which is subsequently morphed as it travels along the optical path of the instrument. We approximate this morph as a Gaussian function. Therefore, the resultant near field that is captured by the detector is the inverse Fourier transform of the modified far field. Invoking the convolution theorem, this detected near field is simply the convolution of the input rectangular function with a Gaussian function, which we subsequently square to approximate the total energy of the light as it hits the detector, yielding Equation 3.1.

Finally, we smooth the best-fit values for A, d, and σ along each order using a cubic spline with ten equally spaced knots. The parameters from this smooth fit yield the profile with which we can divide the original extended flat image to generate the master extended flat-field template.

We only use pixels along the approximately flat portion of the extended slit in this constructed smooth profile. This decision was made for two reasons:

- there is less signal along the top and bottom edges of the slit, thus there is inherently more scatter in the flat corrections for these pixels, and
- even with the analytic function that we use, the steep edges of the slit function are difficult to fit, which leads to systematic problems in the master extended flat.

Since the slit function of the extended flat was designed to be only 50% larger than the science fiber slit function, the range of pixels covered by the flat portion of the extended slit does not correct the entire cross-dispersion profile of the science order. Thus, we also change the motor position of the extended flat injection mirror, which moves the extended flat slit function along the cross-dispersion direction of the echellogram. We thereby expand the number of pixels included in the master extended flat by generating a master extended flat image for each mirror position and then mean-combining these images. This process is completed for each epoch.

3.3.1.4 Noise Model

The two largest contributions to the noise model for any given pixel on the *EXPRES* CCD are photon noise and read noise, where these two quantities are measured and summed in quadrature for each pixel. Photon noise is assumed to be Poisson, such that the standard deviation is equal to the square root of the photoelectron counts. Read noise, on the other hand, is calculated empirically for each amplifier. First, the standard deviation of the nightly stack of dark frames is determined for every bias-subtracted gain-corrected pixel. Then, the median of these standard deviations is assigned as the read noise for each amplifier region. We assume that the read noise is consistent throughout each night of observation.

For the median-combined science flat frame, there is an additional noise term: intrinsic variability in the flat-fielding LED. As described in Blackman et al. (2020), the LED brightness varies by about 0.5% over time. The uncertainty from this variability is proportional to the counts (as opposed to photon noise, which is proportional to the square root of the counts). Therefore, it too is summed in quadrature to produce the total noise for the science flats. As with all median- or mean-combined frames, the noise model for the science flat is divided by the square root of the number of frames.

3.3.2 Spectral Extraction

"Extraction," in the context of the *EXPRES* pipeline, refers to the process of converting reduced two-dimensional CCD data into a series of one-dimensional, normalized spectra, one for each order of the echelleogram. This process involves tracing the echelle orders, removing scattered light, executing the optimal extraction, and continuum normalizing the resultant spectra. Details for these steps are found in the following sections.

3.3.2.1 Order Tracing

The orders of the echellogram are traced using the reduced science flat frame. First, the orders are detected using a peak-finding algorithm along the mean-combined center three columns. Then, for each order, moving from this center-line outward one column at a time, triplets of neighboring columns are mean-combined and the centroid of the resultant array is calculated. The right and left ends of each order are determined by setting a SNR $>$ 30 threshold and stopping the trace once 50 subsequent columns do not reach this threshold. Finally, these centroids are smoothed along each echelle order using a 6th degree polynomial. A single set of traces is calculated for each night of observations.

3.3.2.2 Scattered Light

Due to imperfections in the *EXPRES* optics, some scattered light from the instrument hits the CCD. This diffuse, scattered light is assumed to be smoothly varying across the detector and is estimated from the counts in the regions between the orders of the echellogram. First, a variance-weighted mean and associated uncertainty is calculated for each column of each inter-order region (including those immediately above and below the traced echellogram) using the seven pixels set halfway between adjacent traced orders. These inter-order background approximations are then smoothed using a cubic b-spline with knots set every $\sim$100 columns. Cosmic rays that could potentially skew this smoothed fit are iteratively rejected using a 5σ outlier cut.

The 2D scattered light image is generated through a quadratic interpolation along each column of the smoothed inter-order backgrounds. The calculated scattered light is subtracted from the reduced image before any further extraction. This process is completed for the science flats, stellar frames, and all wavelength calibration frames. An example of a subregion of the resultant scattered light approximation is shown in Figure 3.4.

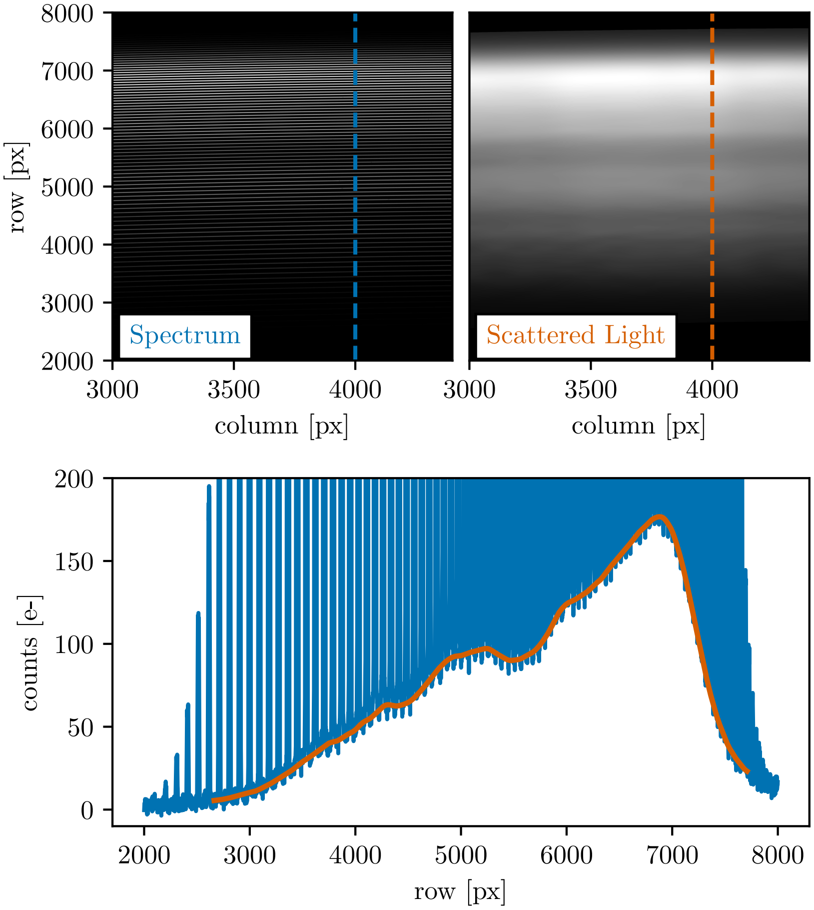

Figure 3.4: In the upper plot, two-dimensional scattered light approximation for reduced and median-combined science flat frames taken on October 24, 2019 are shown. The scales for the two images are not matched in order to empahsize the scattered light. The cross section of a single column (4000) for both the spectrum and calculated scattered light is shown in the lower plot. Note that the range of traced orders sets the upper and lower limits of the calculated scattered light.

3.3.2.3 Optimal Extraction

The optimal extraction algorithm that we implement in the *EXPRES* pipeline is an updated version of the algorithms developed by Horne (1986); Piskunov & Valenti (2002); Zechmeister et al. (2014). Following along the traces (from Section 3.3.2.1) of each echelle order, we construct a least-squares estimator for each 33-pixel tall column x:

$$\chi_x^2 = \sum_y (D_{x,y} - P_{x,y} s_x)^2 w_{x,y} \tag{3.2}$$

where $D_{x,y}$ are the photoelectron counts for each pixel (x, y) in the reduced data, $P_{x,y}$ is the model of the slit function corresponding to those same pixels, s_x is the extracted spectral intensity, and $w_{x,y}$ are the weights for each pixel (see Figure 3.5. These weights are inversely proportional to the variance of $D_{x,y}$ ($\sigma^2_{x,y}$, as further discussed in Section 3.3.1.4) and include a binary cosmic ray mask, $M_{x,y}$, i.e. $w_{x,y} = M_{x,y}/\sigma^2_{x,y}$ (see Zechmeister et al. 2014). The minimization of Equation (3.2) has an analytic solution:

$$s_x = \frac{\sum_y w_{x,y} D_{x,y} P_{x,y}}{\sum_y w_{x,y} P_{x,y} P_{x,y}}. \tag{3.3}$$

The propagated uncertainty of the extraction

$$\sigma_{s_x} = \sqrt{\frac{1}{\sum_y w_{x,y} P_{x,y} P_{x,y}}} \tag{3.4}$$

is rescaled by $\chi_{\mathrm{red},x}$ as in Zechmeister et al. (2014). However, we smooth $\chi_{\mathrm{red},x}$ across each order using a 3rd-order polynomial before applying it to σ_{s_x} to avoid low number statistic variance.

Cosmic rays are rejected by iteratively adding to the cosmic ray mask, $M_{x,y}$, based on a tiered outlier rejection algorithm. After calculating s_x for an entire order, the residual $(D_{x,y} - P_{x,y} s_x) w_{x,y}$ for each pixel is calculated. For each column in the order with a pixel that

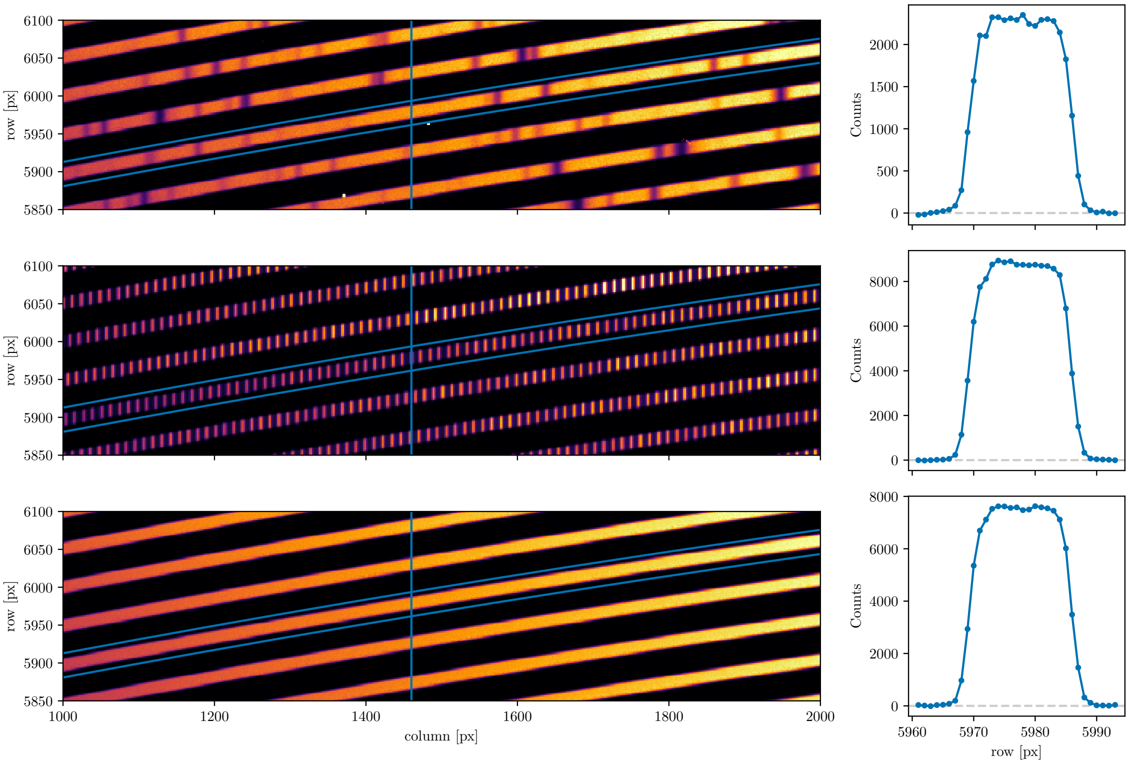

Figure 3.5: Section of raw images for a science exposure (above; HD 217014), laser frequency comb (middle), and calibration flat (below) taken with *EXPRES* on 24 October 2019. The extraction aperture for our optimal extraction of echelle order 100 is shown on the flat images as the intersection of the traced order (with vertical extent of 33 pixels) and the single-column slit at $x = 1460$, both shown in blue. The reduced counts in the extraction aperture for each image (i.e. removing hot pixels and QE variations) are shown in the right panels, as a function of pixel row position y.

exceeds 8σ, the pixel with the largest residual is rejected and s_x is re-calculated for that column. This process repeats until all 8σ outliers are rejected. Next, this same rejection process repeats for 2σ outliers that also neighbor previously rejected pixels, thus rejecting the dim tails of otherwise bright cosmic rays. Importantly, this second step is executed using two-dimensional pixel information, meaning that a bright cosmic ray on a pixel in a given column can have its tail rejected in an adjacent column.

The *EXPRES* pipeline has two distinct methods of approximating the model $P_{x,y}$ in Equation (3.3):

1. fit each column of the science flat with a parametric slit function (Equation (3.1)), smooth the parameters along each order using a b-spline, and then normalize the function; or
2. use the science flat, without normalization, as in Zechmeister et al. (2014).

This yields two modes of operation when extracting data, wherein (1) keeps the echelle blaze function of each order intact while (2) intrinsically removes the blaze.

We note that using method (2) requires inclusion of the variance prior in the flat $\sigma^2_{P_{x,y}}$, as well as that for the data $\sigma^2_{D_{x,y}}$, when determining the weights $w_{x,y}$ of the least-squares estimator. Therefore, the variance prior for each pixel should instead be constructed as

$$\sigma^2_{x,y} = \sigma^2_{D_{x,y}} + s_x^2 \sigma^2_{P_{x,y}}, \tag{3.5}$$

and an additional nonlinear cost term must be added to Equation (3.2) to prevent $s_x \to \infty$. In our case we choose $\sum_y \ln \sigma^2_{x,y}$.

Thus, the minimization of Equation (3.2) must be solved numerically:

$$s_x^{(n+1)} = \frac{\sum_y w_{x,y}^{(n)} D_{x,y} P_{x,y}}{\sum_y w_{x,y}^{(n)} P_{x,y} P_{x,y} + R_x^{(n)}} \tag{3.6}$$

where the weights $w_{x,y}^{(n)}$ for each iteration are calculated using the previous iteration's solution $s_x^{(n)}$ and $s_x^{(0)}$ is calculated assuming all $\sigma^2_{P_{x,y}} = 0$. The relaxation factor $R_x^{(n)}$ is also chosen to be

$$R_x^{(n)} = \sum_y \sigma^2_{P_{x,y}} w_{x,y}^{(n)} - \sum_y \sigma^2_{P_{x,y}} \left(w_{x,y}^{(n)}\right)^2 \left(D_{x,y} - s_x^{(n)} P_{x,y}\right)^2 \tag{3.7}$$

from the minimization of the modified Equation (3.2). This numerical process is repeated until a relative tolerance of 10^{-10} is met. Thus, method (2) is operationally slower than method (1), but intrinsically accounting for un-corrected QE variations and automatically removing the blaze without relying on a blaze model outweigh this minor increase in computational cost.

Removal of the blaze through method (2) can be naively understood by recognizing Equation (3.2) is approximately solved by $s_x \approx \frac{D_x}{P_x}$, where P_x is the spectral intensity of the LED source times the blaze and D_x is the spectral intensity of the star times the blaze. Both P_x and D_x describe their respective intensity at the given column x of the order. Therefore, the resultant continuum of a stellar spectrum extracted using method (2) is simply the continuum of the star divided by the spectrum of the LED source. Since both of these are slowly varying functions, we can model the extracted stellar continuum with a simple linear model for each order, calculated iteratively with 2σ outlier rejection for those pixels contained in absorption lines. See Figure 3.6 for examples of extracted spectra and the approximated continua.

In the *EXPRES* pipeline, we strictly use (2) to extract all data and only use (1) as a secondary check for our extracted RVs. However, we note that maximizing the SNR of the cross-correlation function described in Section 3.3.4 requires weights, which increase with signal-to-noise of the stellar continuum, to be assigned to different portions of the spectrum. Unfortunately, the posterior uncertainties of the extraction have too much scatter to provide these weights. We find that using the blaze function is a good proxy for a smooth weighting function; therefore, we approximate a blaze function model for each

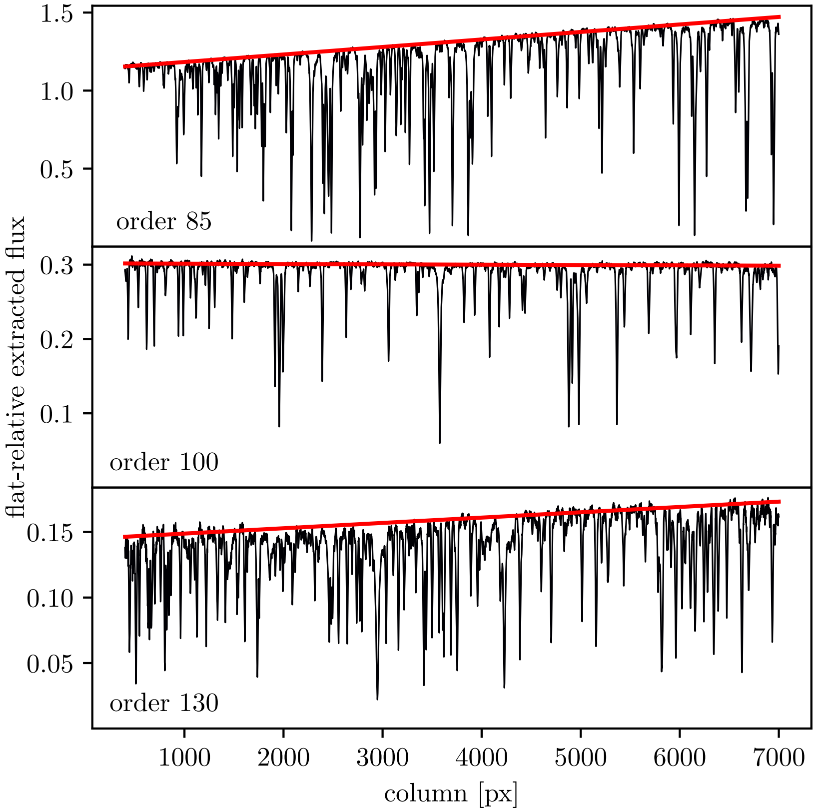

Figure 3.6: Flat-relative optimally extracted flux (black) and associated calculated continua (red) for three orders from HD217014 observed on 24 October 2019. All continua were calculated using a linear model. Notice that the spectral flux relative to the flat-fielding LED source flux is significantly different between orders, but the resultant continuum across each order varies slowly enough to remain essentially linear.

epoch using the α-hull method (Xu et al., 2019) applied to a science flat extracted with (1). We find that RV results obtained with (2) including the separately defined blaze function are superior to those results from spectra extracted with (1).

Extraction method (2) is also the most appropriate choice for wavelength calibration spectra, such as those from the ThAr lamp and the LFC. The calibration methods used in Section 3.3.3 are linear and can make use of the scattered posterior uncertainties as weights when fitting each emission line without systematically shifting results as in cross-correlation. Additionally, light from the LED, ThAr lamp, and LFC all travel through approximately the same lengths of fiber, meaning removal of an instrumental response function (i.e. a calibration "continuum") should not be necessary when using method (2).

The signal-to-noise ratio (SNR) of a given observation is reported here as the maximum s_x/σ_{s_x} in echelle order 111. This is effectively the per-pixel SNR at 550 nm, conforming to the metric used by Fischer et al. (2016) to compare many contemporaneous spectrographs. The resolution element of *EXPRES* contains approximately 4 pixels (Jurgenson et al., 2016), thus the per-resolution-element SNR is simply twice the per-pixel SNR.

3.3.3 Wavelength Calibration

EXPRES uses a laser frequency comb (LFC) as its primary calibration source, which generates a series of spectral lines evenly spaced in frequency, whose nominal frequencies ν_n satisfy the relation

$$\nu_n = \nu_{\text{rep}} \times n + \nu_{\text{offset}}, \tag{3.8}$$

for integers n. The repetition rate ν_{rep} and offset frequency ν_{offset} are referenced against a GPS-disciplined quartz oscillator, providing calibration stability corresponding to a fractional uncertainty of less than 8×10^{-12} for integration times greater than 1 s.

While it is possible to obtain an absolute calibration from the LFC once the free spectral range of the echellogram has been adequately characterized, the LFC suffers from poor throughput in very blue and very red orders. In particular, although our instrumental throughput is sufficient to permit order tracing and extraction from echelle orders 75 through 160 (for 3800 Å$\leq \lambda \leq$ 8220 Å), the LFC only sufficiently illuminates echelle orders 82 through 135 (for 4500 Å$\leq \lambda \leq$ 7500 Å). The photonic crystal fiber (PCF) of the LFC was then replaced in July 2019 due to the decreasing stability of the LFC. With the replacement, the polarization of the LFC was switched, making the LFC redder and thus changing the orders illuminated by the LFC to echelle orders 82 through 130 (with the blue edge at (5300 Å). The polarization switch should significantly increase the lifetime and stability of the LFC. Consequently, we use ThAr lamp exposures taken at the beginning and end of each night as a secondary calibration source to provide well-constrained wavelength solutions for orders outside the range of the LFC.

Calibration triplets (3 LFC's) are taken through the science fiber at roughly 15-30 minute intervals throughout observing, interwoven with science exposures. The exposure times of these calibration frames are chosen to match the target SNR of the science exposures. While *EXPRES* is equipped with a secondary square fiber to permit simultaneous wavelength calibrations, we choose to take calibrations through the science fiber so that our calibration data sample the same pixels and optical elements as the science exposures. This strategy aims to homogenize our exposures to pixel-level, uncalibratable systematic errors. Also, as shown by Blackman et al. (2020), the instrumental stability of *EXPRES* is such that sampling the LFC every 15-30 minutes provides enough information to correct for any instrumental changes throughout the night, as simultaneous calibration would. Calibration images are taken while the telescope is slewing and so typically cost little additional time (less than 2 minutes an hour).

A ThAr wavelength solution is generated from each ThAr exposure using the IDL code `thid.pro`, developed by Jeff Valenti, which identifies ThAr lines by matching lines in an exposure against a line atlas. A 6th-order, 2D polynomial is then fitted over pixel location x and the absolute echelle order m against the scaled wavelength $m\lambda$. Matching lines against an atlas is performed manually once at the beginning of each calibration epoch; otherwise, the wavelength solution from the immediately preceding ThAr exposure is used as an initial guess for the locations of atlas lines in a given ThAr exposure, allowing this process to be automated. Since the LFC lines are sparse relative to the precision of the ThAr calibration (1 LFC line every 10 pixels on average with the ThAr calibration accurate to the nearest pixel), this is sufficient to permit unambiguous mode identification for the LFC lines.

For any given LFC exposure, the locations of modes are identified by fitting Gaussians to each peak after a smooth background has been subtracted. An initial, trial wavelength solution is generated by linearly interpolating the ThAr solutions from the beginning and end of the night. These are used to determine the mode number n corresponding to the frequency of each mode. Once again, a 2D polynomial is fitted for $m\lambda$ as a function of m and x. Since the LFC produces a far denser set of lines—typically about 20,000 lines across 50 orders are identified in an LFC exposure, compared to about 4,000 lines across 82 orders in a ThAr exposure—we use a 2D polynomial described by a 10x10 matrix of coefficients (9th-order in each dimension) for the fit. The locations of the ThAr lines are also included in the fit in order to constrain the behaviour of this polynomial in echelle orders that are otherwise inaccessible to the LFC.

For each of the polynomial coefficients describing the wavelength solution, we fit a smooth function (a cubic polynomial) in time. We have found that interpolating the polynomial coefficients, rather than directly interpolating the pixel-wise wavelength solutions, is more robust to imperfections in individual calibration frames. This set of 100 functions is evaluated at the photon-weighted midpoint time of each science exposure to generate a wavelength solution. This differs from the standard practice at other spectrographs (e.g. *HARPS* and ESPRESSO), where a single velocity offset, rather than a time-dependent wavelength solution, are assigned to each science exposure. We choose to do this in order to accommodate time-dependent variations in the characteristics of the instrument, which may lead to calibration shifts that cannot be adequately described by a single average velocity offset.

We illustrate this in Figure 3.7, where we show the calibration drift over the course of a single night in the LFC-constrained region of the echellogram. Our radial velocity solutions (see Section 3.3.4) sample spectral lines that are non-uniformly distributed across the detector; therefore, it is advantageous to characterize local variations to prevent overall systematic offsets. Of course, the precise differential velocity imparted by this calibration drift ultimately depends on the spectral type as well as both barycentric and systemic velocity of the star under consideration.

As the final step in wavelength calibration, the *EXPRES* pipeline applies a barycentric correction to the wavelength solution of each stellar observation using the method described by Blackman et al. (2017). For each 1 s exposure of the exposure meter, a barycentric correction is calculated using BARYCORR (Wright & Eastman, 2014). The photon-weighted average barycentric correction is calculated for each of 8 wavelength bins of the spectra. A third-degree polynomial is then fit to these averages, yielding a smoothly-varying wavelength-dependent barycentric correction $z_{\mathrm{B}}(\lambda)$ for the observation. As shown by Tronsgaard et al. (2019), it is important to distinguish this from a photon-weighted midpoint time used to calculate an overall chromatically-dependent barycentric correction (e.g., Landoni et al., 2014) as this can impart a $\sim$10 cm s^{-1} systematic error to the radial velocity especially in cases of longer exposure times, high airmass, or poor seeing.

Finally, we apply z_{B} directly to the wavelength solution:

$$\lambda_{\mathrm{bary}} = \lambda_{\mathrm{lab}}^{(\mathrm{vac})}\left(1 + z_{\mathrm{B}}\left(\lambda_{\mathrm{lab}}^{(\mathrm{air})}\right)\right) \tag{3.9}$$

where λ_{lab} is the LFC-generated lab-frame wavelength solution and λ_{bary} is the wavelength

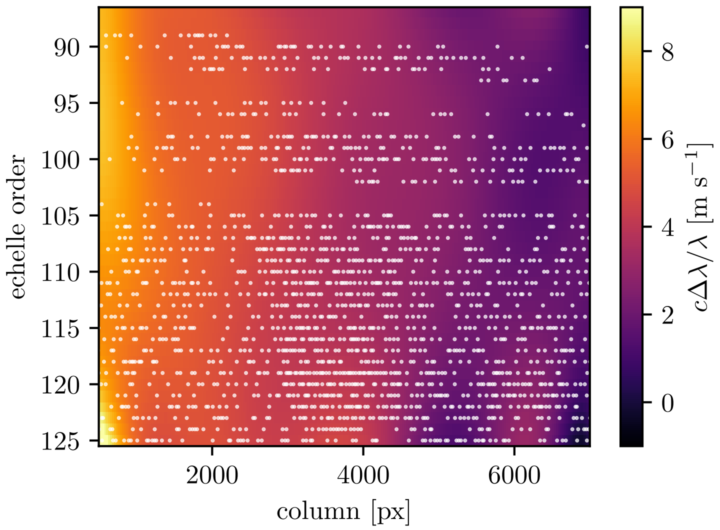

Figure 3.7: Calibration drift over the course of a single night (24 October 2019) for the LFC-constrained region of the echellogram, plotted in terms of the absolute echelle order number and the pixel column of the CCD. The average calibration drift for the whole night (~4.0 m s^{-1}) is of similar magnitude to local variations in the drift, therefore using a single average velocity offset would necessarily incur significant additional calibration error. In other words, the wavelength solution along the left and right edges of the shown spectral format would be offset by -4.0 and +4.0 m s^{-1} respectively if a single velocity offset were to be used. Spectral lines used for radial velocity solutions (shown as white dots drawn from the *ESPRESSO* G2 linelist) sample the detector in a nonuniform fashion, and result in different overall velocity offsets depending on spectral type.

solution in the frame of the solar system barycenter. Because the *EXPRES* exposure meter is not in vacuum (as opposed to *EXPRES* itself), $z_{\mathrm{B}}(\lambda)$ is measured using air-wavelengths. Therefore, λ_{lab} is converted from vacuum to air using the algorithm and parameters derived by Ciddor (1996) before applying the barycentric correction in Equation (3.9).

3.3.4 Radial Velocity Solutions

The data analysis pipeline of *EXPRES* employs two distinct computational techniques to independently extract radial velocities from stellar spectra:

1. a "cross-correlation function" method (CCF; see Baranne et al., 1979) is used to determine a rough estimate of the absolute radial velocity for each observation (Section 3.3.4.1), and
2. a forward model based on a morphed NSO solar spectrum is used to derive a more precise relative radial velocity curve (Section 3.3.4.2).

Both of these methods are currently implemented in the *EXPRES* pipeline for self-validation. Simultaneous results from both methods are presented in Section 3.4. The methods as implemented in the *EXPRES* pipeline are described as follows.

3.3.4.1 Cross-Correlation

As the first step in our analysis, a CCF method estimates the absolute RV of *EXPRES* science targets precise to several tens of cm s^{-1} (depending on photon noise). We also use the CCF method to diagnose drifts and instabilities in our calibration sources, using line lists given by the comb parameters following Equation (3.8) for the LFC or a ThAr line atlas for the ThAr lamp.

The CCF is constructed from the input spectrum $f(\lambda)$ as well as a spectral-type linelist—a set of spectral lines at rest vacuum wavelengths $\{\lambda_i(0)\}$ associated with contrast weights $\{c_i\}$ and widths $\{h_i\}$. For a given trial radial velocity v, the wavelength of each line in the linelist is redshifted appropriately to

$$\lambda_i(v) = \lambda_i(0)\sqrt{\frac{c+v}{c-v}}. \tag{3.10}$$

The CCF is then computed as a numerical approximation to the integral

$$\mathrm{CCF}(v) = \int \mathrm{d}\lambda f(\lambda) \sum_i c_i w\left(\frac{\lambda - \lambda_i(v)}{h_i}\right) \tag{3.11}$$

where w is an arbitrary window function approximating a Dirac δ function and λ is with respect to the barycentric-corrected wavelength solution from Equation (3.9).

The CCF in Equation (3.11) is computed independently for each echelle order with a variety of trial velocities, and the CCFs for all relevant orders are co-added before a velocity model is fitted. This is a similar practice to other CCF-based RV pipelines (e.g. Brahm et al., 2017). When deriving extreme precision radial velocities, we only include orders falling within the spectral range of the LFC, since in principle it affords considerably better sampling density and calibration stability than those regions covered by the ThAr lamp alone.

An appropriate functional model is then fitted against the co-added values of the CCF. The position parameter and posterior uncertainties of the fitted model are returned as the reported velocity and formal errors. Other quantities of astrophysical interest (e.g., rotational broadening width, bisector inverse slope) are also computed from the co-added CCF.

Our construction of the CCF incorporates the ability to use an arbitrary window function w. In the current iteration of the *EXPRES* pipeline, we use a cosine function, matching other contemporary CCF implementations (e.g. Freudling et al., 2013; Modigliani et al., 2019; Brahm et al., 2017). We also use a Gaussian functional model to fit the CCF for our reported radial velocities. Other possible combinations of window functions and CCF models are discussed in Section 3.3.4.3.

3.3.4.2 Forward Modeling

In addition to our CCF RV solution, we have developed a new forward-modeling technique by adapting algorithms developed for the iodine RV technique (Butler et al., 1996) as well as ideas from the "line-by-line" method developed by Dumusque (2018). Forward-modelling from empirical stellar spectral templates is known to produce velocities with less statistical scatter than the CCF method, and typically measures relative rather than absolute radial velocities (Anglada-Escudé & Butler, 2012). Our modeling process is simplified relative to the iodine method because the optical design of *EXPRES* was optimized to provide stability in the line spread function (LSF) of the spectrograph (Jurgenson et al., 2016; Blackman et al., 2020), eliminating the need to model the instrumental LSF with several free parameters. In addition, free parameters for wavelength solution and dispersion are eliminated since the barycentric wavelength solution (Equation (3.9)) is provided as part of the nightly optimal extraction.

First, we construct a spectral template for each stellar target. An ideal template will have very high SNR and will be a good spectral match to the program stars. Our starting point is to obtain a set of four consecutive spectra—each with SNR of about 250—providing an effective SNR of 500 per pixel or SNR of 1000 per resolution element. As described by Dumusque (2018), we prefer to use spectra with low barycentric velocities so that the program spectra shift around the approximate zero point wavelengths. The telluric contamination is then modeled in each spectrum using SELENITE (Leet et al., 2019) and divided out. Finally, the set of spectra are co-added.

However, even the co-added spectrum will not provide a high enough SNR for a robust template. Therefore, we take the additional step of morphing the NSO solar spectrum (see Figure 3.8) with a native SNR ~10,000 and resolution ~500,000 to match the co-added, telluric-cleaned spectra for each of our program stars using the following pro-

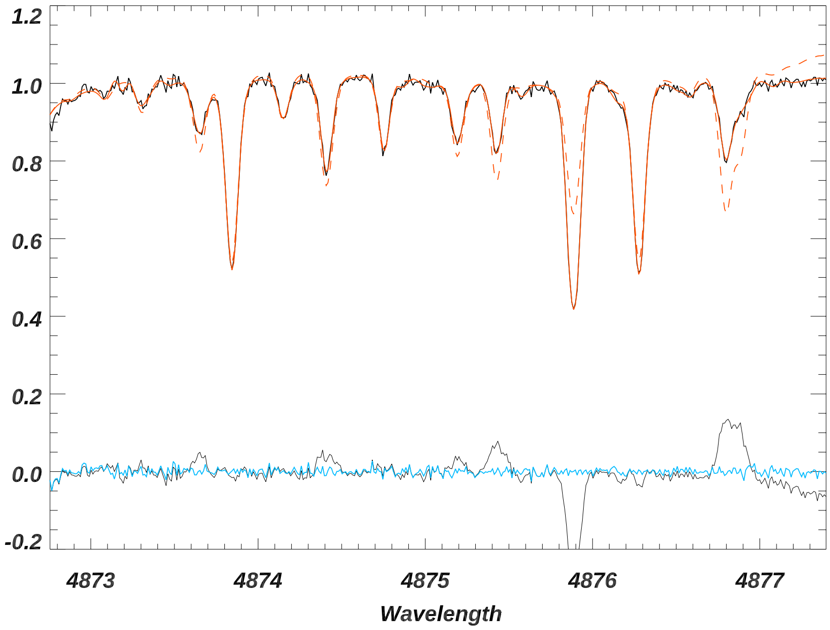

Figure 3.8: The high-resolution, high-SNR NSO spectrum (red dashed line) is shifted and cubic-spline interpolated to the wavelength scale of a program observation (upper black line). The difference spectrum (bottom black line) is used to identify discrepancies between the spectra above the photon noise threshold (bottom blue line). A Levenburg-Marquardt algorithm drives the growth of pseudo-lines until the NSO spectrum has morphed to the match the spectrum of the program star (solid red line). This morphed spectrum is then used as a template for forward modeling.

cedure:

1. The co-added program star spectrum is divided into ~2000 individual chunks that are 140 pixels (~2Å) wide within the orders of the spectrum covered by the LFC.

2. The barycentric wavelengths of the program star are used to extract a segment of the NSO spectrum with generous padding of 200 pixels for modeling shifts. This segment of the NSO spectrum is shifted to the barycentric frame of the co-added program star spectrum.

3. A Levenburg-Marquardt (L-M) least squares algorithm is used to (i) determine the best-fit width for a Gaussian convolution kernel to rotationally broaden the NSO spectrum, (ii) refine the Doppler shift of the NSO spectrum, and (iii) apply a vertical shift to align the continuum of the NSO and the co-added spectrum.

4. The rotationally broadened and shifted NSO spectrum is cubic-spline interpolated onto the wavelength scale of the co-added spectrum.

5. A difference spectrum is calculated. Nodes are dropped down consecutively at points where the absolute value of the difference spectrum exceeds a threshold, characterized by the photon noise of the co-added spectrum. Pixels with the largest residuals in the difference spectrum are modeled first. The maximum number of nodes is 60, but depending on the chunk there are typically about a dozen nodes required to model the NSO spectrum for each 130-pixel chunk.

6. At each node a positive or negative Gaussian feature with a width characterized by the line spread function of *EXPRES* is used to perturb the NSO spectrum; the depth of the morphing feature is determined by L-M fitting of the residuals.

7. Iterative growth of the morphing lines stops when the residuals of the difference spectrum are consistent with photon noise.

8. Each chunk of the template is weighted according to the amount of spectral information, using the SNR and the derivative of intensity I with pixel:

$$\sum \frac{\delta I}{c}\frac{\lambda}{\delta\lambda}\frac{1}{\mathrm{SNR}} \tag{3.12}$$

Once the template for each star has been generated, a L-M fit is executed for every 140-pixel chunk of the program star spectra. Telluric-affected pixels in each observation are assigned zero weight in the fit. There are only two free parameters for each chunk: a Doppler shift and continuum normalization scale factor. Thus, each chunk—45 LFC-calibrated orders each with about 47 chunks yielding ~2,000 total chunks—provides an independent measurement of the relative RV for the star. The RV for each chunk is subsequently subtracted by the mean of that chunk over all observations, thus removing any offsets that might occur because of geometric anomalies in the detector while preserving the spread in RV variations.

Weights for each chunk are determined using empirical arguments, the χ^2 of the L-M fit, and a chunk-specific modifier based on its relative temporal scatter. Chunks that do not contain any absorption lines in the stellar template as well as chunks that yield relative velocities greater than ± 1000 m s^{-1} are assigned zero weight. Moreover, chunks that have $\chi^2 > 5.0$ (typically occurring if an incorrect stellar template was used or a telluric line was missed, for example) and remaining chunks that are among those with the largest 3% of reduced χ^2 are all assigned zero weight.

Because some chunks have less spectral information, there will be more scatter in the RVs derived from these chunks. For example, chunks in the blue part of the spectrum typically have several absorption lines, but chunks in the red part of the spectrum may have only one spectral line meaning the L-M fitting will not be well-constrained. Likewise, telluric contamination within a given chunk can manifest as large scatter in the RV over time. Therefore, the non-zero weight for given chunk i within an observation j is

$$w_{i,j}^{-1} = \chi_{i,j}^2 \frac{\sum_j^n (v_{i,j} - \bar{v}_j)^2}{n-1} \tag{3.13}$$

where $\bar{v}_j$ is the median velocity for all chunks of observation j and n is the total number of observations for a given stellar target. The reported RV measurement for each observation is thus a weighted mean of the individual chunk velocities and the formal error is the corresponding standard error of the weighted mean.

3.3.4.3 Further Analysis

Once RVs have been derived, the extracted spectra and CCFs are passed down the pipeline for more sophisticated analysis. The spectral range of *EXPRES* is intended to permit characterisation of stellar activity and planetary atmospheric absorption lines. For chromospheric activity in particular, we extract the Ca II line core emission ratio index S_{HK} (using the parametric model of Isaacson & Fischer, 2010), calibrated to yield results consistent with the Mt. Wilson Observatory catalogue (Duncan et al., 1991).

We also aim to incorporate spectroscopic activity indicators directly into the RV solution methodology. For example, in addition to using a rectangular “box” function and truncated cosine in the CCF, which are implemented in other similar velocity analysis codes (e.g. Freudling et al., 2013; Modigliani et al., 2019; Brahm et al., 2017), the *EXPRES* analysis code implements CCF computation using Gauss-Hermite window functions of the form

$$w(x) = \frac{1}{\sqrt{2^n n! \sqrt{\pi}}} H_n(x) \exp\left[-\frac{x^2}{2}\right], \tag{3.14}$$

where H_n is the n^{th} (physicists’) Hermite polynomial. Computing higher-order CCFs as coefficients in a Hermite-functional decomposition, and more generally with respect to different orthogonal basis functions, will permit more sophisticated analysis of stellar activity (via a sparse description of variations in the CCF line profile), as an alternative parameterization to current derived observables (such as the CCF bisector inverse slope/FWHM). Alternatively, data-driven decorrelation of stellar activity from bulk radial velocities, or alternative template-based RV solution methodologies (Holzer et al., in prep), may be possible once we have built up an archive of stellar spectra.

3.4 Initial Results - HD 217014

We now present velocities derived with this pipeline based on 47 observations of HD 217014 (51 Peg) over the ten month span between the beginning of Epoch 4 and December 1, 2019. We do so to examine various sources of uncertainty and error in the velocimetric pipeline, while avoiding the known instrumental instabilities inherent in Epochs 1-3 (see Blackman et al., 2020; Szymkowiak et al., in preparation, for details), and to compare the two radial velocity methods outlined in Section 3.3.4. Observations with an SNR less than 160 are not included in this analysis. We construct our CCF using the G2 linemask from the *ESPRESSO* pipeline (Freudling et al., 2013; Modigliani et al., 2019) and a cosine window function. We also fit all RVs with a single planet Keplerian model, constrained by the literature value of the orbital period (4.2308 days, Wang & Ford, 2011). This Keplerian model is parameterized by the velocity semi-amplitude (K), the eccentricity (e), the argument of periastron, and a phase of periastron.

Table 3.3: *EXPRES* Commissioning RVs of 51 Peg (full data set available online)

BMJD	V_{ccf} cm s^{-1}	V_{fm} cm s^{-1}	RV Epoch	SNR
58639.45844	−3320697 ± 20	5739 ± 32	4	385
58641.45174	−3331424 ± 44	−5035 ± 42	4	179
58643.46218	−3321644 ± 34	4854 ± 34	4	225
58644.46095	−3322776 ± 35	3527 ± 34	4	233
58646.45596	−3330577 ± 39	−4045 ± 38	4	203
	⋮			

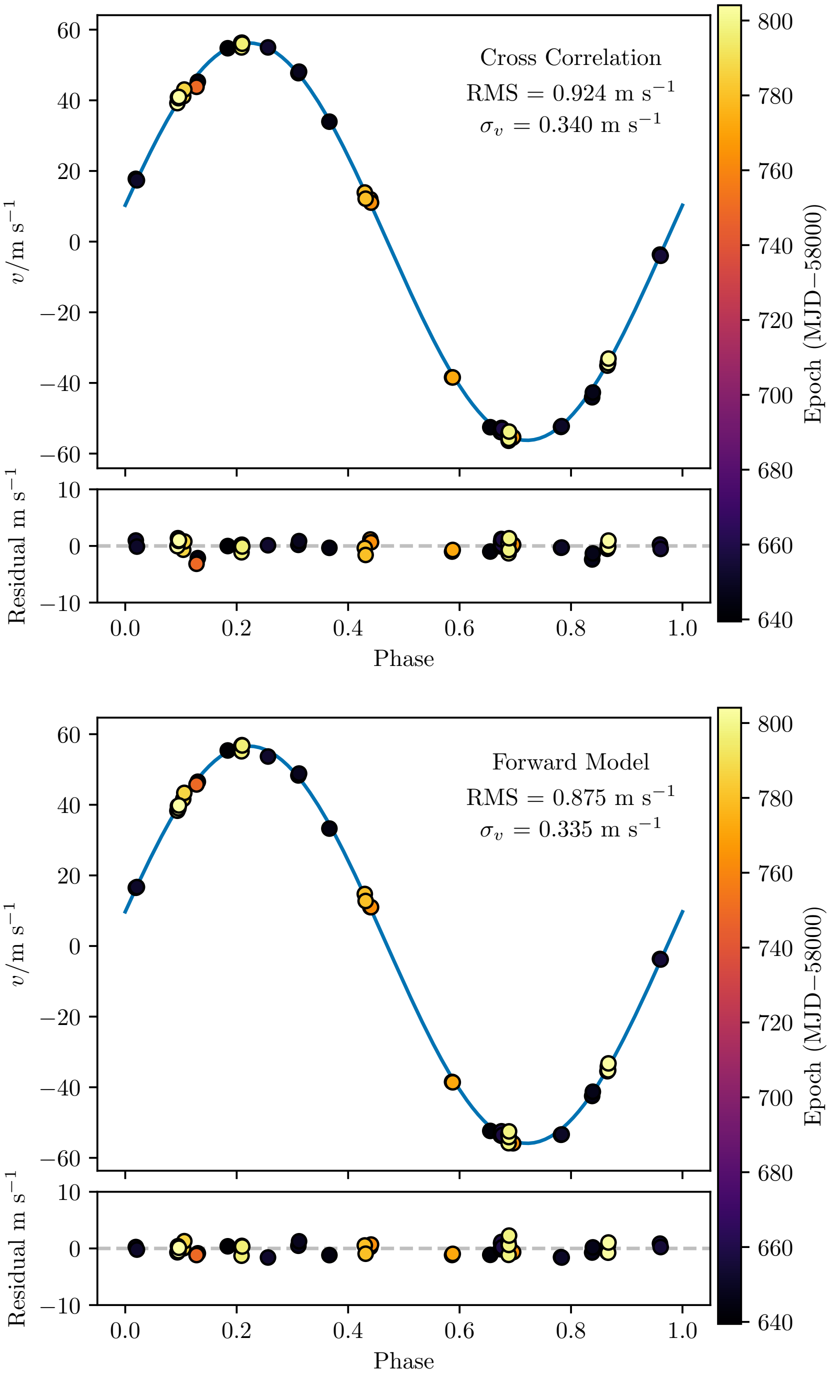

Figure 3.9: Phased radial velocities, Keplerian orbital fits, and residuals for *EXPRES* observations of 51 Peg b. The figure is labelled with the RMS residual to the fitted orbital solution, as well as the median formal error σ_v of all data points shown.

Figure 3.9 shows the resulting radial velocities from both the CCF and Forward Model methods along with their respective orbital fits. Because the CCF uses a linelist with absolute wavelengths, the fit systemic velocity of $-33.2603(5)$ km s^{-1} (in excellent agreement with Gaia Collaboration, 2018) has been removed. For each of the RV epochs, we also fit an independent velocity offset relative to the overall offset. For epochs 4 and 5, these are $-1.5(4)$ m s^{-1} and 1.2(5) m s^{-1} when using the CCF method, and $-1.2(8)$m s^{-1} and 0.8(7) m s^{-1} when using the forward model. These offsets account for modifications to the instrumental systematics owing to the various fiber changes and realignments. The offsets differ slightly in magnitude between the two RV analysis methods because they intrinsically weigh regions of the detector differently, accentuating or mitigating certain instrumental systematics.

We show the values of the Keplerian fit parameters in Table 3.4 along with the RMS residual and the median formal error (σ_v). The parameter uncertainties were derived by taking the square root of the product of the posterior variances and reduced χ^2 of the least squares fit. By way of comparison, we also perform the same procedure with eight years of archival velocities from the *HIRES* instrument on the Keck I telescope (corrected for instrumental systematics per Tal-Or et al., 2019) and four months of data from the *HARPS* DRS (Trifonov et al., 2020). The *EXPRES* orbital solution parameters for 51 Peg are consistent with those returned from these previous studies, but with higher precision due to the improved formal errors. The two *EXPRES* RV methods are also internally consistent, with the forward model producing a slightly more favorable RMS. Finally, we note that

Table 3.4: Fit parameters for 51 Peg b

Instr.	K m s^{-1}	e	RMS m s^{-1}	σ_v m s^{-1}
EXPRES CCF	56.24 ± 0.14	0.000 ± 0.002	0.924	0.340
EXPRES FM	56.26 ± 0.13	0.007 ± 0.003	0.875	0.335
HARPS DRS	53.4 ± 1.6	0.062 ± 0.010	0.941	1.023
HIRES	56.7 ± 0.4	0.020 ± 0.007	2.74	1.169

the RMS of the *EXPRES* fit residuals approximately matches that of the *HARPS* DRS, with the *EXPRES* fit returning parameters more comparable with literature values (Bedell et al., 2019; Wilson et al., 2019; Wang & Ford, 2011).

3.5 Discussion

Each step of the pipeline contains several free parameters — for instance, the degree of the fitting polynomials to use for the spatial wavelength solution fits and temporal smoothing, as well as various SNR threshholds for calibration line identification and order inclusion in the CCF. To assist community users of the instrument, we have opted, as far as possible, to preselect reasonable default values for most of these parameters, which may be overridden at runtime. In what follows, we document some nonobvious but critical aspects of these systematics, and we illustrate some aspects of the decision-making process for choosing our default values for some of these parameters.

3.5.1 Formal vs. True Velocity Errors

Since the formal velocity errors returned from the CCF fitting procedure are constructed only from the co-added CCFs and their propagated errors, they do not account for effects like wavelength calibration error (inducing spurious velocity shifts) or time estimation error (via erroneous barycentric corrections). Instead, they mostly reflect velocity estimation error due to photon noise being propagated to the CCF.

On the other hand, the formal velocity errors returned from the forward model fitting procedure does include some information about relative uncertainty in certain regions of the *EXPRES* detector. For instance, a chunk that tends to have a telluric line will naturally incur more spread in the measured RV for that chunk. Therefore, even though this chunk is down-weighted by our analysis, its spurious effect still propagates to the RV error.

These assumptions are bourne out in Figure 3.10, showing these formal errors as a function of the observation SNRs. The CCF points depend essentially only on photon noise and potentially CCD readout and optimal extraction systematics—which we are confident of having adequately accounted for—up to some constant that may depend on e.g. the choice of CCF linemask or window function, or intrinsic astrophysical properties of the target. Conversely, the Forward Model formal errors contain much more scatter that we believe folds in some uncertainty from telluric contamination and, potentially, stellar noise. Thus, our estimation of the true photon noise limit of *EXPRES* is better described by the propagated errors of the CCF analysis, though the two analyses yield quite similar results. As shown in Figure 3.10, we define this limit to be 30cm s^{-1} for a single observation at the *EXPRES* target SNR of 250.

Following our diagnosis and repair of the LFC beat frequency noted in Table 3.1, we measured the remaining uncalibratable velocity errors arising from wavelength calibration in particular to be relatively small—between 4 and 6cm s^{-1} RMS (Szymkowiak et al., in

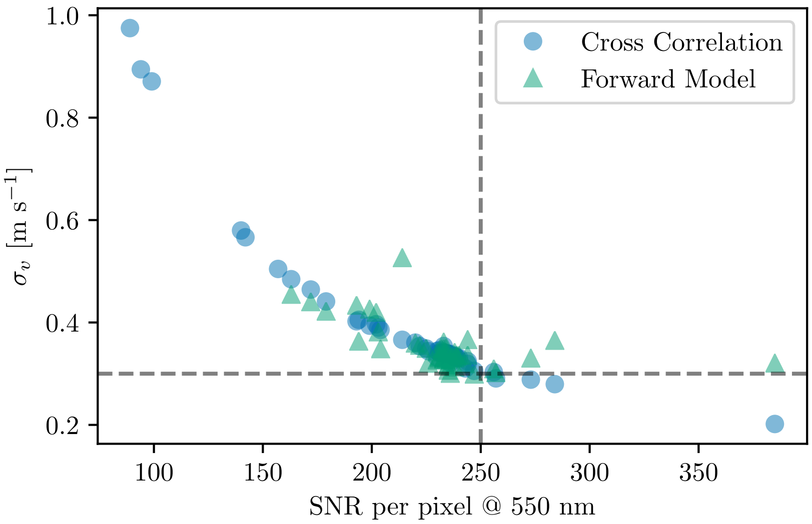

Figure 3.10: Formal RV errors returned from both RV fitting procedures, plotted against the per-pixel SNR (as a characterization of photon noise) for velocities from Section 3.4. *EXPRES*'s target SNR of 250 is shown with a vertical line and the approximate associated formal error of 30 cm s^{-1} is shown with a horizontal line. Note that data here with SNR less than 160 are not included in Figure 3.9.

prep). There also exist several other sources of error (e.g. from uncalibratable instrumental systematics and guiding errors), that constitute additional contributions to the RV error budget. To correctly estimate the velocimetric error, one needs to appropriately account for and then combine these error terms (e.g. by adding them in quadrature) with the formal value reported from the RV analysis. A detailed inventory of these error sources (Blackman et al., 2020) estimates the combined instrumental and guiding errors of *EXPRES* at ~10cm s^{-1}. Thus the single observation error of *EXPRES* is clearly dominated by the apparent photon noise.

3.5.2 Chromatic Dependences

Many of the novel techniques that we have adopted in the *EXPRES* pipeline involve the introduction of chromatic dependences into quantities that have previously been considered to be uniform with wavelength, such as calibration offsets and the barycentric offset velocity. It therefore behooves us to investigate possible chromatic effects that emerge at the end of the CCF velocity-solving and orbit-fitting procedure.

The CCF analysis in Section 3.4 was performed by co-adding CCFs derived from echelle orders 126 through 86 (4850 Å$\leq \lambda \leq$ 7150 Å) before fitting an absorption-line model to derive a velocity. These orders are those for which at least $N_{min} = 19$ LFC lines are detected that pass both the SNR threshhold and all quality checks imposed by our peakfitter (N_{min} depends on the degree of the polynomial fitted to the wavelength solution, which is a free parameter in our code, as are the threshhold values for these quality checks).

In Figure 3.11, we show the RMS residuals from the orbital solutions that arise when we repeat these analyses while varying the range of echelle orders used when co-adding CCFs; our default parameter selections are indicated with the red dotted lines. In particular, this means extrapolating the wavelength solution and chromatic barycentric correction beyond the spectral range of the LFC, which covers echelle orders 130 through 82, and of the chromatic exposure meter, which covers orders 135 through 86 (~ 4650 Å$\leq \lambda <$ 7150 Å).

For this data set, we see that there is a sharp dependence on the bluest order co-added into the CCF. Moreover, we see that the introduction of orders redder than the LFC cutoff also slightly increases the RMS error to the fit. Recalling that the wavelength solution outside of the LFC region is largely constrained by the ThAr lamp, this potentially implies a calibration offset between the LFC and the ThAr lamp, despite both sources illuminating the instrument through the same fiber.

On the other hand, we do not see any similarly sharp cutoff when extrapolating the chromatic barycentric correction to outside of the wavelength range covered by the exposure meter. This suggests that the wavelength dependences of our barycentric corrections (detailed more fully in Blackman et al. 2019) are generally smooth enough for robust extrapolation.

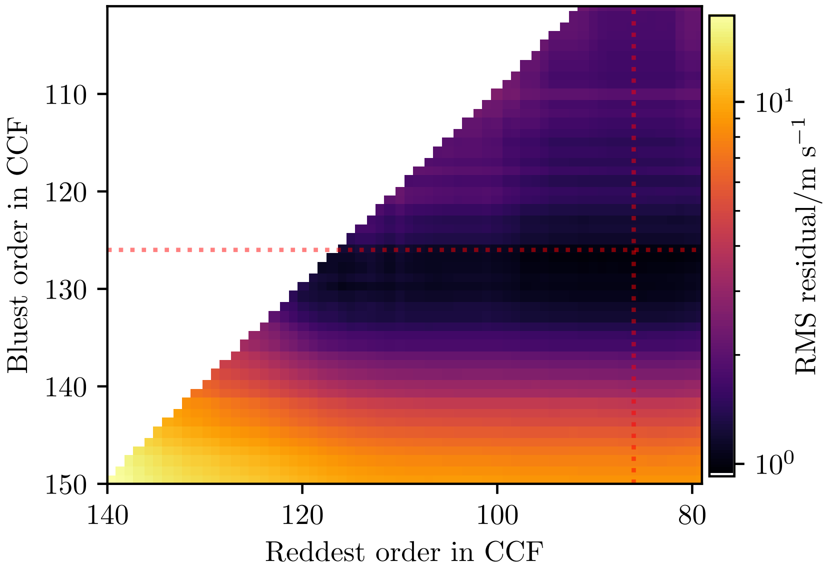

Figure 3.11: RMS scatter from orbital solutions fitted to HD 217014 as a function of bluest and reddest echelle order included in the CCF computation. Points along each diagonal indicate sets of velocities computed with the same number of echelle orders. At least 10 echelle orders have been included in all CCF computations. Dotted red lines show the default parameters used in the preceding sections.

Finally, it is possible to choose narrower ranges of echelle orders that yield smaller RMS errors than our default parameter selection. However, we note that this is potentially dependent on the specifics of the CCF linelists used in the computation and also possibly on underlying astrophysical properties of the science targets and other fortuitous factors. We have opted to include as many orders as can be accurate, so as to minimize photon noise, while still leaving the option to include fewer orders available to end-users.

3.6 Conclusion

The commissioning process on the *EXPRES* instrument is essentially complete, along with the development of an optimal extraction pipeline that we have been using for preliminary RV analysis through both CCF and forward modelling techniques. Within the instrumental back-end (i.e. limiting ourselves to the calibration unit and the spectrograph proper), we have determined our photon-noise-limited RV errors to be approximately 0.3 m s^{-1} for a single observation with SNR of 250. With *EXPRES*'s current observing strategy of four observations per night per target, this result implies a nightly measurement error of only 0.15 m s^{-1}. While our on-star measurement error appears to be ~0.9 m s^{-1}—based on residual RMS to an orbital fit of 51 Peg b—we must also note that our RV analysis pipeline does not fully address photospheric velocity sources, telluric contamination, or longer-term instrumental errors. These other sources of RV scatter are beyond the scope of this paper although we are actively investigating them. The RV precision presented in this paper, therefore, represents our worst-case scenario in the absence of further improvements. We anticipate upcoming hardware improvements and more sophisticated RV solution methodologies to only enhance our measurement precision and long-term instrumental stability.

Moreover, there are multiple parts of the pipeline that remain under active development. We are investigating the use of spectro-perfectionism (Bolton & Schlegel, 2010;

Cornachione et al., 2019) as an alternative to optimal extraction. We are also exploring a hierarchical, non-parametric wavelength solution that takes advantage of the low degrees of freedom allowed in a stabilized instrument and the density of lines offered by new wavelength calibrators (Zhao et al., 2021). There are also plans to implement a more data-driven approach (as in Bedell et al. 2019, with modifications to permit chromatic barycentric corrections) as yet a third RV analysis technique.

These caveats notwithstanding, we have demonstrated that the technical innovations that have been invested into the development of novel instrumentation and software analysis techniques for *EXPRES* have largely paid off — they have permitted us to unambiguously attain sub-meter-per-second on-sky radial velocity precision. Presently, this makes *EXPRES* the most precise EPRV spectrograph in the northern hemisphere.

Finally, we hope the lessons we have learned in the process of commissioning this instrument, and the techniques we have developed, to be of some value to the community of EPRV instrument builders moving forward.

3.7 Acknowledgements

We thank the anonymous referee for comments that significantly improved this paper. We gratefully acknowledge Yale University, the Heising-Simons Foundation, and an anonymous Yale alumni donor for providing telescope time for *EXPRES* used to obtain the data in this paper. We also acknowledge the extensive work that went into the RePack spectral extraction code, written by Lars Buchhave; this code helped to benchmark the *EXPRES* optimal extraction described in this paper. JO thanks X. Dumusque for illuminating and productive discussions. We especially thank the NSF for funding that allowed for precise wavelength calibration and software pipelines through NSF ATI-1509436 and NSF AST-1616086 and for the construction of *EXPRES* through MRI-1429365. This material is based upon work supported by the National Science Foundation Graduate Research

Fellowship under Grant No. DGE1122492 (RRP, LLZ, ABD).

Software

SciPy library (Virtanen et al., 2020), NumPy (Oliphant, 2006–; van der Walt et al., 2011), Astropy (Astropy Collaboration et al., 2013; Price-Whelan et al., 2018).

Facilities

LDT

Chapter 4

Excalibur: A Non-Parametric, Hierarchical Wavelength-Calibration Method for a Precision Spectrograph

Lily L. Zhao[1], *David W. Hogg, Megan Bedell, Debra A. Fischer*

Excalibur is a non-parametric, hierarchical framework for precision wavelength-calibration of spectrographs. It is designed with the needs of extreme-precision radial velocity (EPRV) in mind, which require that instruments be calibrated or stabilized to better than 10^{-4} pixels. Instruments vary along only a few dominant degrees of freedom, especially EPRV instruments that feature highly stabilized optical systems and detectors. *Excalibur* takes advantage of this property by using all calibration data to construct a low-dimensional representation of all accessible calibration states for an instrument. *Excalibur* also takes advantage of laser frequency combs or etalons, which generate a dense set of stable calibration points. This density permits the use of a non-parametric wavelength solution that can adapt to any instrument or detector oddities better than parametric models, such as a polynomial. We demonstrate the success of this method with data from the *EXtreme PREcision Spectrograph* (*EXPRES*), which uses a laser frequency comb. When wave-

---

[1] Originally published as: Zhao, L. L., Hogg, D. W., Bedell, M., & Fischer, D. A. 2021, The Astronomical Journal, 161, 80. I developed an implementation of *excalibur* to run on *EXPRES* data and ran all tests to assess *excalibur* performance.

lengths are assigned to laser comb lines using *excalibur*, the RMS of the residuals is about five times lower than wavelengths assigned using polynomial fits to individual exposures. Radial-velocity measurements of HD 34411 showed a reduction in RMS scatter over a 10-month time baseline from 1.17 to 1.05 m s^{-1}.

4.1 Introduction

Precise, radial-velocity programs have been fruitful in finding and characterizing extra-solar planets (e.g. Mayor et al., 2011; Bonfils et al., 2013; Plavchan et al., 2015; Butler et al., 2017). These programs typically make use of spectrographs with resolutions on the order of $50,000 - 100,000$, which correspond to line widths on the order of 3000 m s^{-1}. The state of the art RV precision reached 1 m s^{-1} by 2016 (Fischer et al., 2016). The newest generation of instruments aim to reach 0.1 m s^{-1} precision, the required precision to detect terrestrial worlds. This requires new spectrographs to be calibrated or stabilized to better than 10^{-4} of a pixel (assuming that the spectrographs are well sampled). Two next-generation spectrographs, *EXPRES* and *ESPRESSO*, have been commissioned for more than a year and are demonstrating < 0.1 m s^{-1} instrumental errors and ~ 0.2 m s^{-1} errors on stars (Pepe et al., 2013; Jurgenson et al., 2016; Blackman et al., 2020; Petersburg et al., 2020; Brewer et al., 2020; Suárez Mascareño et al., 2020).

Traditionally, wavelength solutions are constructed by fitting a polynomial to lines from a calibration source in order to describe the relationship between wavelength and pixel for each echelle order (Butler et al., 1996; Lovis & Pepe, 2007; Cersullo et al., 2019). In this framework, each calibration image is treated independently. The returned wavelength solutions worked well at the level of 1 m s^{-1} precision.

The move towards 0.1 m s^{-1} RV precision, necessitates higher-fidelity calibration data and wavelength models. These models need to account for high-order spatial variations that can arise from small imperfections in the optics of an instrument and non-uniformity in detector pixel sizes/spacing. There has been significant effort in using an entire set of calibration images to identify incongruous ThAr lines (Coffinet et al., 2019) or obtain high-resolution Fourier transform spectra of reference cells (Wang et al., 2020). It has also been found that using multiple polynomials in the dispersion direction, tuned to capture detector defects, better describes the wavelength solution than a single, continuous polynomial (Milaković et al., 2020).

Here, we propose to simplify and improve calibration programs for EPRV hardware systems with two practical yet innovative ideas. The first flows from the fact that calibration sources—which include arc lamps (in some wavelength ranges), etalons, and laser-frequency combs (LFCs)—illuminate the spectrograph with very stable, very dense sets of lines–almost every location in the spectrograph image plane is surrounded by nearby, useful calibration lines. This enables use of a calibration methodology that is *non-parametric*, or not defined by a prescribed, analytic function described by a finite number of parameters: If every point in the spectrograph detector is sufficiently surrounded by nearby calibration lines, the wavelength solution can, for example, be made simply as an interpolation of the calibration data. The density of lines removes the need to enforce any functional form for the wavelength solution (such as a continuous ninth-order polynomial). In some ways, this is a generalization of recent work that has demonstrated the efficacy of constructing a wavelength solution as multiple, segmented polynomials (Milaković et al., 2020). A non-parametric approach will improve calibration accuracy by not forcing the choice of a parametric form that may bias the calibration, especially when the chosen function is inappropriate (as, for example, polynomials are at detector edges).

The second simple idea follows from the observation that most physical systems have only a few dominant degrees of freedom, meaning most spectrographs vary along only a small number of axes in "calibration space", or the (very high-dimensional) space of all possible wavelength solutions. This is particularly true of EPRV instruments, which are equipped with stringent environmental stabilizing. The thermomechanical stability of these instruments reduces the variations they experience to something that can be represented by a low-dimensional framework. That is, spectrographs, especially stabilized ones, should have few environmentally accessible degrees of freedom. This renders it inadvisable to fit each calibration exposure or calibrate each science exposure independently. Instead, all the calibration data (or all the data) should be used to determine the calibration space in which the instrument can and does vary. Subsequent calibration work then need only determine where in the small, accessible part of calibration space the spectrograph was located for each exposure.

In the context of probabilistic models, this structure is *hierarchical*: The calibration data are used not just to determine the wavelength solution at one moment, but also to determine the possible *calibration space* of wavelength solutions at all moments. In statistics, this concept is often described as *de-noising*: we can improve calibration by recognizing that every calibration exposure contains information about every other calibration exposure. Thus, every exposure can be improved (i.e., de-noised) with information from every other exposure.

The method we propose here—*excalibur*—embodies these ideas. It is a non-parametric, hierarchical, data-driven method to generate a wavelength model. By being non-parametric, it delivers enormous freedom to the wavelength solution to match or adapt to any instrument or detector oddities. By being hierarchical, it restricts that freedom tremendously, but it does so appropriately for the empirically determined variations in a spectrograph.

The method *excalibur* is designed for temperature-controlled, fiber-fed spectrographs with good calibration sources, such as laser-frequency combs, or etalons. We have in mind EPRV instruments and EPRV science cases, primarily because the need for good wavelength calibration is so great in this field. Irregardless, we expect *excalibur* to have applications for other kinds of spectrographs in other contexts. *Excalibur* should be applicable to other spectrograph with low-dimensional variability, though the precision of the returned wavelengths will depend on the available calibration sources (more discussion in

Section 4.6 below) .

4.2 Method

The *excalibur* method is designed to take many calibration images, each containing a series of calibration lines with known wavelengths and well-fit detector positions, and de-noise and interpolate this information into a full wavelength model applicable to all exposures taken‘ with the instrument. It operates on two core ideas; The wavelength solution should be allowed flexibility, but it lives in a very low-dimensional calibration space where the degrees of freedom are set by the limited kinematics of the spectrograph hardware. *Excalibur* therefore assumes that the space of possible calibration states for an instrument is low-dimensional but assumes very little about the forms of those states.

Excalibur also assumes dense-enough calibration line coverage with well-fit line centers to provide sufficient constraints on an interpolated wavelength solution across an echelle order. Upstream errors in line center positions may propagate through *excalibur* wavelength models. The required line density is dependent on the required precision of the returned wavelength model; larger spacing between lines offer less constraint and are likely to return worse wavelengths. We revisit and quantify these conditions in Section 4.6.

Wavelength calibration is usually posed in the following way: Given an exposure n, and echelle order m, there is a relationship between the two-dimensional (x, y)-position on the detector and the wavelength λ

$$\lambda(x, y, m, n) = f(x, y, m; \theta_n) \quad , \tag{4.1}$$

where θ_n represents the parameters describing the wavelength solution for a given exposure.

Classically, pipelines employ polynomials to construct smooth wavelength solutions for each exposure. For example, the *EXPRES* pipeline sets the function $f(x, y, m; \theta_n)$ from Equation 4.1 to a 2D, 9^{th}-order polynomial, where θ_n represents the polynomial coefficients, c_{nij}, unique to each exposure n (Petersburg et al., 2020).

$$\lambda(x, m, n) = \sum_{i=0}^{9} \sum_{j=0}^{9} c_{nij}\, x^i\, m^j + \text{noise} \quad , \tag{4.2}$$

Here, the y-dependence is dropped as, in our framing, this dependence is carried by spectral order m. The line position on the detector is therefore uniquely identified by echelle order m and pixel in the dispersion direction, x. The coefficients c_{nij} are interpolated from the time of calibration exposures to time t_n of a science exposure n by a third-order polynomial with respect to time. This third-order polynomial is evaluated at the time of non-calibration, science exposures to re-construct the coefficients for a 2D, 9th-order polynomial wavelength solution for that exposure. Each calibration image obtains its c_{nij} coefficients independently.

Given a stabilized instrument with low degrees of freedom, however, the calibration of any image can be reliably informed by the calibration of every other image. The calibration data themselves can be used to develop a low-dimensional basis for expressing the space of all possible calibrations for a spectrograph with few degrees of freedom.

If the space of all calibration possibilities is in fact K-dimensional (where K is a small integer, i.e 2 or 8 or thereabouts), and if the calibration variations are so small that they can be linearized, then the function $f(x, m; \theta_n)$ from Equation 4.1 should be low-dimensional. In *excalibur*, we transpose the calibration model—making the position x a function of λ—into the following form

$$x(\lambda, m, n) = g_0(\lambda, m) + \sum_{k=1}^{K} a_{nk}\, g_k(\lambda, m) \quad , \tag{4.3}$$

where $g_0(\lambda, m)$ is the fiducial or mean or standard calibration of the spectrograph, the a_{nk} are K scalar amplitudes for each exposure n, and the $g_k(\lambda, m)$ are basis functions expressing the "directions" in calibration space that the spectrograph can depart from the fiducial calibration. The resultant $x(\lambda, m, n)$, a list of calibration line positions for a given exposure, can be regarded as the calibration state of the spectrograph for that exposure. When this calibration structure is used to deliver a wavelength solution, $x(\lambda, m, n)$ can be inverted back into $\lambda(x, m, n)$ to recover wavelengths for each detector position x and echelle order m (Bauer et al., 2015).

The challenge is to learn these basis functions, $g_k(\lambda, m)$, from the data and get the K amplitudes, a_{nk}, for every exposure n. There are many ways to discern the basis functions. In this paper, we present one implementation of *excalibur* using principal component analysis (PCA) (Pearson, 1901; Jolliffe & Cadima, 2016). A PCA is justifiable in the limit where exposures have very high signal-to-nose ratio, as is usually the case with typical calibration images. There are many alternatives to PCA for this dimensionality reduction; we return to this point in Section 4.5 below.

4.2.1 Dimensionality Reduction: De-Noising of Calibration Frames

Excalibur will use calibration images to determine 1) the space in which an instrument varies and 2) where in the accessible calibration space the spectrograph existed for each exposure. For each calibration exposure, n, *excalibur* requires a full list of lines, (λ, m) that are expected to appear in each calibration exposure. Each line is uniquely defined by a combination of echelle order, m, and "true" or theoretical wavelength, λ. There are many strategies for identifying calibration line positions and matching them to their assigned wavelengths; this problem is left out of scope for this work.

Excalibur assumes that line positions have been identified "correctly," as in the position of a calibration line is determined the same way as the position of a stellar line when

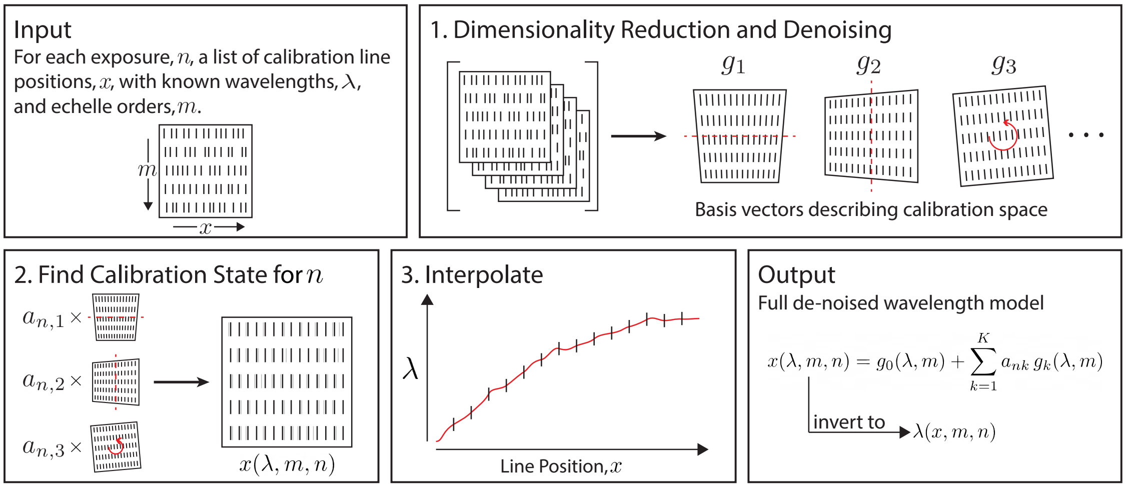

Figure 4.1: A cartoon representation of the *excalibur* method, as described in Section 4.2. We exaggerate variations in measured line position, changes in calibration space, and interpolation deviations for clarity. In step one, dimensionality reduction and denoising (§4.2.1), the complete set of line positions for all exposures is analyzed to return a set of K basis vectors, g_n, which represent different ways the spectrograph calibration changes. These basis vectors span the K-dimensional calibration space of the spectrograph, which includes all possible wavelength solutions. In step two (§4.2.2), the amplitude of each basis vector, $a_{n,k}$, is interpolated to return the calibration state for a specific science exposure, returned as a set of de-noised calibration lines. The assigned wavelengths of these de-noised line positions are then interpolated onto other pixels in step three (§4.2.3) to construct a full wavelength model that returns wavelength as a function of detector position x and echelle order m.

extracting RVs. This also inherently assumes that the calibration lines are not subject to an effect that the science exposures are not, for example differences in charge transfer inefficiency, non-linearities in the PSF, etc. We caution that systemic errors or large uncertainties in fitting line positions easily propagate through to biases in the wavelength models returned by *excalibur*. For more discussion, see Section 4.7.

For each exposure, n, every line, (λ, m), has an associated fitted (measured) detector position, $x(\lambda, m, n)$, for example x-pixel in an 2D extracted echelle order. Fitted line centers that are missing from an exposure (e.g. because the fit failed due to noise, the line is not in its usual echelle order, etc.) can be assigned a `NaN` (hardware not-a-number) for that exposure instead. Let there be P lines per exposure. *Excalibur* reads in a $N \times P$ matrix of line positions for each of P lines for each of N exposures.

The mean of measured line position over the set of calibration exposures represents the fiducial, or standard calibration of the spectrograph, $g_0(\lambda, m)$. In this implementation of *excalibur*, principal component analysis is performed on the difference between this fiducial calibration and each individual line position. The returned principal components serve as basis functions, $g_k(\lambda, m)$, expressing the possible deviations of the spectrograph from this fiducial calibration. The magnitude of each principal component for each exposure, a_{nk}, represents the scalar amplitude of these deviations for each exposure. *Excalibur* then uses a small number, K, of principal components to reconstruct a de-noised version of the line positions as formulated in Equation 4.3.

Missing line center measurements, which were previously marked by `NaN`s, are replaced with de-noised estimates. This is done iteratively until the estimates of missing line centers change by less than 0.01%. This process can be repeated on line centers deemed as outliers by some metric, to account for lines that may have been mis-identified or misfit. The principal components from the final iteration are used to define the spectrograph's calibration space, while the associated amplitudes for each component pinpoint where in that calibration space the spectrograph is located for each calibration exposure.

Algorithm 1: Dimensionality Reduction and De-Noising

```
Data: line positions $x(\lambda, m, n)$ for each exposure $n$, with wavelengths $\lambda$ and
      echelle orders $m$
Result: Basis vectors of the low-dimensional calibration space $g_k(\lambda, m)$ and
        location of exposures in calibration space expressed by amplitudes $a_{n,k}$
while change in missing or outlier line centers > 0.01% do
    $g_0(\lambda, m) = \overline{x(\lambda, m, n)}$;
    using Singular-Value Decomposition, find $U, \Sigma, V$ s.t.
      $U\Sigma V^* = (x(\lambda, m, n) - g_0(\lambda, m))$;
    let $a_{n,k} = U \cdot \Sigma$ and $g_k(\lambda, m) = V$;
    $x(\lambda, m, n) = g_0(\lambda, m) + \sum_{k=1}^{K} a_{nk}\, g_k(\lambda, m)$ for $x(\lambda, m, n) =$ NaN where $K$ is a a
      small integer
end
```

4.2.2 Interpolating Calibration Position

In *excalibur*, the amplitude, $a_{n,k}$, of each principal component is interpolated to determine the calibration state of the spectrograph. For example, the amplitude can be interpolated with respect to time to recreate the calibration state of the spectrograph at different times. The choice of what to interpolate against depends on the dominant contribution to variation in the calibration of the instrument.

In the implementation of *excalibur* presented here, the amplitudes of the principal components are interpolated linearly with respect to time. This is discussed more in Section 4.5.2. Let a prime denote values related to a science exposure n' for which we want wavelengths. We use linearly interpolated magnitudes, $a_{n',k}$ at time $t_{n'}$ to construct the calibration state of the spectrograph for that point in time. Using interpolated amplitudes, $a_{n',k}$, and the basis vectors, $g_k(\lambda, m)$, returned by the de-noising process, a new set of calibration lines, $x'(\lambda, m, n')$ can be constructed for any exposure as formulated in Equation 4.3.

4.2.3 Interpolating a Wavelength Solution

From the de-noising step, *excalibur* can now construct a set of calibration lines, $x'(\lambda, m, n')$ for any exposure n'. To construct a wavelength solution, we invert $x'(\lambda, m, n')$ to $\lambda(x', m, n')$ by interpolating known wavelengths of the calibration lines over detector position. For instance, interpolating the known wavelengths vs. line centers onto every integer x will generate wavelengths for each pixel in an echelle order.

After experiments, we found that a cubic-spline interpolation that enforces monotonicity, such as a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolator, works well for interpolating wavelengths onto pixels. A cubic spline allows for more flexibility than a parameterized function, while the enforced monotonicity allows the wavelength solution, $\lambda(x', m, n')$, to be invertible and prevents spurious deviations that may befall a cubic spline. Choices in interpolation scheme, K, and other tests are further discussed in Section 4.5.3.

The implementation of *excalibur* described here is hosted on GitHub[2].

Algorithm 2: Generating Wavelength Solution

```
Data: the fiducial calibration of the spectrograph $g_0(\lambda, m)$; magnitudes of the
      principal components for each exposure $a_{n,k}$; basis vectors spanning the
      calibration space of the spectrograph $g_k(\lambda, m)$;
Result: Wavelengths for detector positions $x'(m, n')$ of exposure $n'$ with time $t_{n'}$,
        where the prime is used to denote values relevant to this new exposure
Find $a_{n',k}$ by interpolating $a_{n,k}$ with respect to $t_{n'}$;
$x'(\lambda, m, n') = g_0(\lambda, m) + \sum_{k=1}^{K} a_{n'k}\, g_k(\lambda, m)$ where $K = 6$;
for each unique $m$ do
    interpolate $\lambda$ with respect to $x'(\lambda, m, n')$ onto pixels $x'(m, n')$;
end
```

---

[2] `https://www.github.com/lilyling27/excalibur`

4.3 Data

We tested *excalibur* using data from *EXPRES*, the EXtreme PRecison Spectrograph. *EXPRES* is an environmentally-stabilized, fiber-fed optical spectrograph with a median resolving power $R = \lambda/\delta_\lambda =\sim 137,000$, over a wavelength range of $390 - 780\,nm$ (Jurgenson et al., 2016; Blackman et al., 2020). *EXPRES* has two different wavelength calibration sources, a thorium argon (ThAr) lamp and a Menlo Systems laser frequency comb (LFC). LFCs are unique in that the wavelengths of their emission lines are stable and exactly known at pico-meter accuracy (Wilken et al., 2012; Molaro et al., 2013; Probst et al., 2014).

Rather than using a simultaneous calibration fiber, two to three LFC exposures are obtained roughly every 30 minutes while the telescope is slewing to new targets. ThAr exposures are taken at the beginning and end of each night. All calibration data are taken through the science fiber, so that calibration light travels down the same optical pathway and is projected onto the same pixels as the science observations. Light passes through a pupil slicer and double scrambler before being injected into a rectangular fiber, which is fed through a mechanical agitator to ensure modal mixing(Petersburg et al., 2018).

LFC lines cover echelle orders 84-124, which contain approximately 19200 calibration lines. Though our results are primarily based on work with LFC data, there will be some discussion of applications to arc lamps below. ThAr lines cover all 86 extracted orders of *EXPRES* (echelle orders 75-160), which include approximately 5300 lines. For both the LFC and ThAr data, lines that appeared in less than 60% of exposures were not included in the analysis. Similarly, exposures with more than 60% of expected lines missing were cut from the analysis. A list of echelle orders m, line wavelengths λ, and pixel positions x were calculated by the EXPRES pipeline (Petersburg et al., 2020) for every line of every exposure and read into *excalibur*.

Line positions from the *EXPRES* pipeline are generated as follows. A ThAr wavelength solution is generated from each ThAr exposure using the IDL code `thid.pro`, developed by Jeff Valenti. This code identifies ThAr lines by matching lines in an exposure against a line atlas. Line matching is carried out in an automated and unsupervised way with a Levenburg-Marquardt minimization routine. Once each line's position is identified and matched to a wavelength from the line atlas, a sixth-order, 2D polynomial is fit over pixel location x, echelle order m, and scaled wavelength $m\lambda$ (wavelengths are scaled in order to distinguish lines that may appear in more than one order).

Flat-relative, optimally extracted LFC data is background-corrected using a univariate spline. Each peak in an echelle order is then fit with a Gaussian. The mean of this fitted Gaussian to a single peak is taken to be the center of the line. For each line, the ThAr wavelength solution is used to estimate the mode number of a line. The precise wavelength is then calculated using

$$f_n = n \times f_r + f_0 \tag{4.4}$$

where the repetition rate, f_r, is known from the design of the LFC, and the offset frequency, f_0, has been determined by Menlo Systems, the manufacturer of the LFC.

In order to comfortably satisfy the assumption that there exists only low-order variation, which is needed for *excalibur*, we used exposures from after the LFC stabilized following servicing in summer 2019, where the photonic crystal fiber was replaced and the polarization was changed to shift the wavelength range of the LFC redwards. In the results presented here, we use 1227 LFC exposures and 78 ThAr exposures taken between October 14 and December 18, 2019 on 29 unique nights.

4.4 Tests

We perform a series of tests to validate the performance of *excalibur* and benchmark *excalibur*-generated wavelengths against wavelengths generated by a classic, non-hierarchical, parametric method. To implement training-validation tests, we leave out a subset of calibration lines with known wavelengths as a "validation" sample, generate wavelengths for these lines using the remaining data, and compare the predicted wavelength to the assigned wavelength of each line. This inherently folds in errors in the measured line center of each calibration line, but this contribution to the residuals will be the same across all tests.

To assess the classic, polynomial-driven method of wavelength calibration, we take each LFC exposure and separate the lines into even- and odd-indexed lines. We then construct a wavelength solution using only the odd-indexed lines and use that wavelength solution to predict wavelengths for the even-indexed lines; i.e. a polynomial is fit to just the odd-indexed lines and then evaluated at the detector positions of the even-indexed lines (see Equation 4.2). We then generate a wavelength solution using only the even-indexed lines and use it to predict wavelengths for the odd-indexed lines.

To test the interpolation step of *excalibur* (§4.2.3), we employed *excalibur* on all LFC exposures with odd-indexed lines removed. The resultant basis vectors, $g_k(x, y, m)$, and amplitudes, a_{nk}, are therefore only informed by the even-indexed lines of each LFC exposure. We then predict wavelengths for the odd-indexed lines that had been excluded and compare these predictions to their assigned wavelengths. This allows us to test how accurately an interpolated wavelength solution can predict wavelengths.

To test the denoising step of *excalibur* (§4.2.1, §4.2.2), we employed *excalibur* on a randomly selected 90% of all LFC exposures. This means the basis vectors, $g_k(x, y, m)$, and weights, a_{nk}, were constructed using only information from 90% of all exposures. We used the results to predict wavelengths for all the lines in the remaining 10% of calibration exposures. This allows us to test how well we can pinpoint the calibration state of the spectrograph using *excalibur*.

The polynomial and interpolation tests remove the same 50% of lines from each exposure while the denoising test completely removes a randomly selected 10% of calibration exposures and their associated line position measurements. Errors from interpolation will be localized, extending only to neighboring lines. We therefore aggressively remove every other line to ensure we are capturing these local effects. The PCA denoising, on the other hand, folds in information of all lines from all exposures. Here, it is sufficient to completely remove 10% of exposures, a traditional training/validation fraction. Since the information being removed varies between each test depending on its focus, we present the results per line, treating each line as an independent test.

The residuals of a wavelength solution represent the difference between the wavelength solution evaluated at the line position of a calibration line, and the assigned theoretical wavelength (i.e. from Equation 4.4 for LFC lines) on a line-by-line basis in every exposure. The reported RMS of a wavelength solution is therefore the per-line RMS, i.e.

$$RMS/line\,[m\,s^{-1}] = \sqrt{\sum_{n=1}^{N}\sum_{p=1}^{P}\frac{[\frac{(\lambda_{n,p,pred.}-\lambda_{p,theory})}{\lambda_{p,theory}}\times c]^2}{N\times P}} \tag{4.5}$$

where $\lambda_{p,theory}$ is the theoretical wavelength for line p, $\lambda_{n,p,pred.}$ is the wavelength predicted by the constructed wavelength solution for line p in exposure n, and residuals from all P lines from all N exposures are used, for a total of $N\times P$ lines. The difference in wavelength is converted to units of m s^{-1}, a more intuitive metric for EPRV work.

4.4.1 Results

Histograms of the per-line residuals for each of the above described polynomial, interpolation, and denoising tests (respectively) are shown in Figure 4.2. Note that the spread

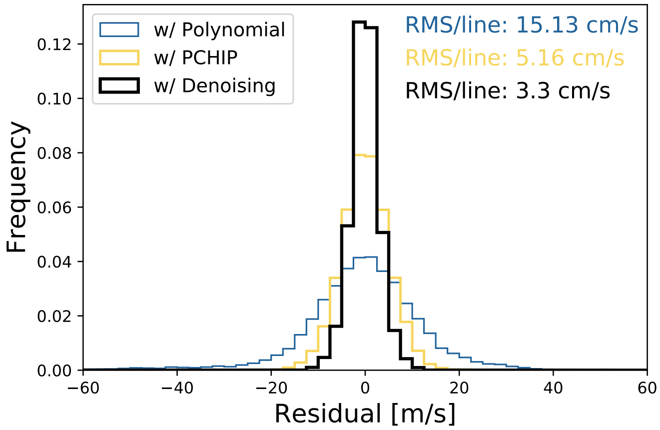

Figure 4.2: Difference in predicted and theoretical wavelengths for the wavelength calibration tests described in Section 4.4. The per line RMS as defined in Equation 4.5 is given in the top-right corner in each method's corresponding color. Incorporating denoising returns the smallest spread in residuals.

in residuals is much smaller for both the denoising and interpolation tests relative to the results of the polynomial wavelength solution.

The per-line residuals from the denoising test also exhibit smaller spread than interpolation alone. This suggests that the spectrograph truly is accurately represented by a low-dimensional model. Recreating line positions using this model gives better line position estimates than treating each exposure independently. The low-dimensional model does not incorporate noise from individual line measurements. Returning more precise, denoised line positions results in smaller per-line residuals.

Excalibur-generated wavelengths also exhibit less structure in the returned residuals. For a randomly selected example LFC exposure, Figure 4.3 plots each line with respect to its echelle order (y-axis) and x-pixel on the detector (x-axis) colored by the difference between the predicted and theoretical wavelength for that line in units of m s^{-1}.

The residuals of the classic, polynomial wavelength solution is shown in the top plot of Figure 4.3. There is a lot of vertical structure and some hints of a periodic, diagonal structure as well. The residuals of the interpolation test for the same exposure is shown in the bottom plot of Figure 4.3. There is no coherent structure here and smaller residuals.

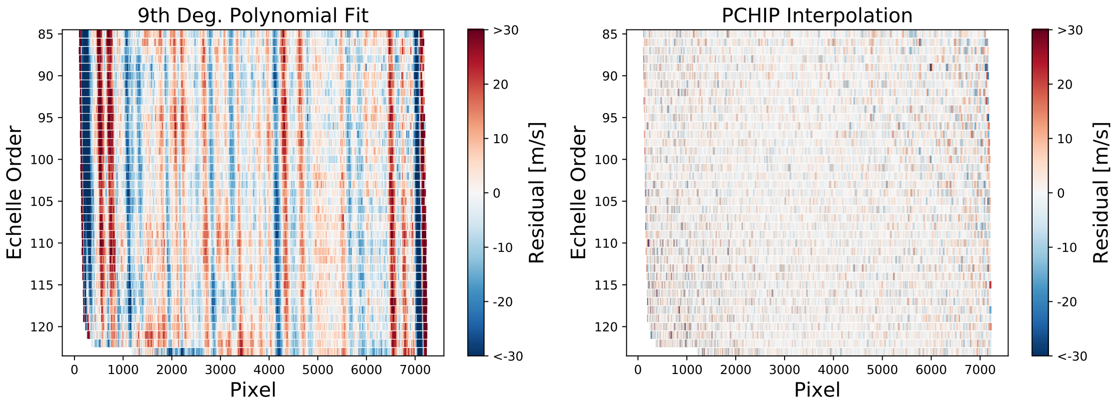

Figure 4.3: Residuals of a single LFC exposure plotted with respect to detector position (as defined by echelle order and x-pixel) for parametric (top) and non-parametric (bottom) wavelength calibration methods. Each line is colored by the difference between the predicted wavelength and the theoretical wavelength for each line, given in units of m s^{-1}. High-order structure, i.e. vertical stripes and patchiness, is apparent in the residuals to a polynomial wavelength solution, which assumes smoothness.

This shows how the flexibility of an interpolated model can account for high-order instrument or detector defects, which emerged as structure in the residuals of the classic, smooth, polynomial-driven wavelength solution. This same flexibility may similarly allow interpolated wavelength solutions to account for position errors in pixel image blocks for different detectors depending on how the interpolation is framed (Fischer et al., 2016; Milaković et al., 2020).

Though the interpolated wavelength solution returns lower, less-structured residuals than the polynomial wavelength solution when guided by LFC lines, the flexibility of an interpolated wavelength solution can result in much worse residuals when not properly constrained, for example in regions between widely separated calibration lines. The left plot of Figure 4.4 shows the residuals when ThAr calibration lines, which are much fewer and less regularly-spaced than LFC lines, are run through *excalibur* and used to predict wavelengths for the (completely independent) LFC exposures taken during the same range of time. Over-plotted in yellow are the positions of the ThAr lines.

Note that running *excalibur* informed by only ThAr lines cannot be regarded as a direct comparison to the LFC runs, as the increased uncertainty and variability in ThAr line positions alone makes the resultant wavelength predictions an order of magnitude worse, hence the different scale of the colorbar in the left plot of Figure 4.4 as compared to Figure 4.3. All the same, the residuals are in general worse where lines are further apart (for example, in redder echelle orders) than where lines are denser.

Figure 4.4 (right) plots the residuals for a subset of order 94 for both a polynomial-based method and a PCHIP-based method guided by either ThAr lines or LFC lines. The PCHIP model with ThAr lines (orange, dashed curve) returns huge residuals between two widely-separated ThAr lines that extends out of frame. The classic, polynomial fit exhibits similar residuals in both amplitude and shape regardless of whether the set of ThAr lines or LFC lines are used. An interpolated wavelength solution using LFC lines (black, solid curve) exhibits the lowest residuals.

The move to an interpolated wavelength solution is driven by the assumption that

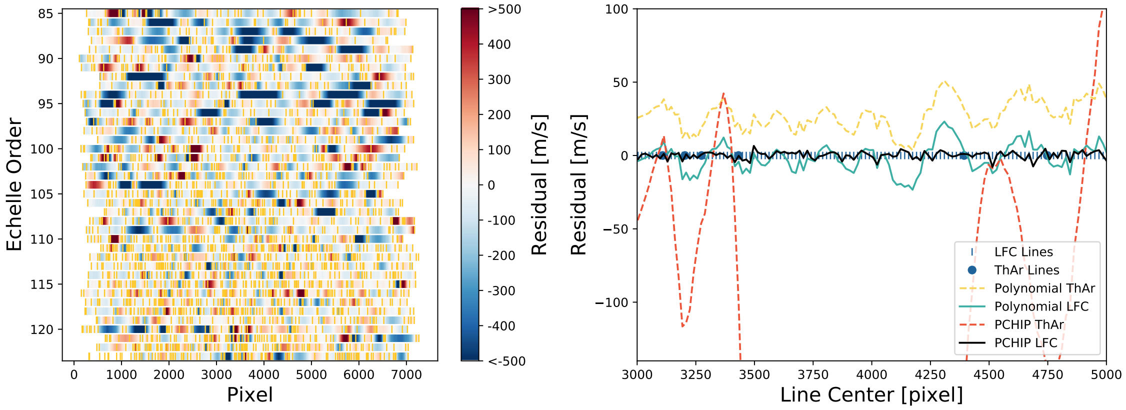

Figure 4.4: Residuals when using ThAr lines to predict wavelengths for LFC lines. **Left**: residuals for a single LFC exposure plotted with respect to detector position and colored by residual, as in Figure 4.3. The positions of ThAr lines are over-plotted in yellow. In general, residuals are greater between ThAr lines with greater separation. **Right**: Comparison of polynomial and interpolated wavelength solutions using either just ThAr lines or LFC lines for a subset of echelle order 94. The shape of the residuals from a polynomial fit are similar whether using ThAr lines or LFC lines. A PCHIP interpolated wavelength model guided by LFC lines returns the smallest residuals.

a high density of calibration lines allows for more freedom in the resultant wavelength solution. This freedom allows the wavelength solution to more accurately ascribe wavelengths. This flexibility, however, is no longer justified in the regime where there are large separations between calibration lines, as this no longer provide sufficient constraint on the interpolated wavelength solution, as is the case in some regions of a classic ThAr lamp.

4.4.2 Impact on Radial Velocity Measurements

We tested *excalibur*-generated wavelengths on RV measurements using *EXPRES* observations of HD 34411, which are presented in Table 4.1. HD 34411 is a G0V star. It is $4.8\,Gyr$ old and relatively quiet ($logR'_{HK} = -5.09$) (Brewer et al., 2016). Because HD 34411 has no known planets and should have a smaller contribution from stellar signals, it is a good star to test the effects of the wavelength calibration on the RMS of the returned RVs. We use 114 observations of HD 34411 taken between October 8, 2019 and March 5, 2020 with SNR 250. Radial-velocity measurements were derived using a chunk-by-chunk, forward-modeling algorithm ran by the *EXPRES* team (Petersburg et al., 2020).

Figure 4.5 compares the resultant RVs when using a classic, ninth-degree polynomial wavelength solution and an *excalibur*-generated wavelength model. Using *excalibur*-generated wavelengths reduces the RMS of the entire data set from 1.17 m s^{-1} with the classic wavelength solution to 1.05 m s^{-1}. This is equivalent to removing an independent,

Table 4.1: *EXPRES* RVs using *Excalibur* Wavelengths. The full data set is available online

JD-2440000	RV m s^{-1}	Error m s^{-1}
18764.4771	3.139	0.335
18764.4791	1.035	0.332
18764.4810	3.074	0.324
18771.4179	-1.927	0.342
18771.4196	2.688	0.357
	⋮	

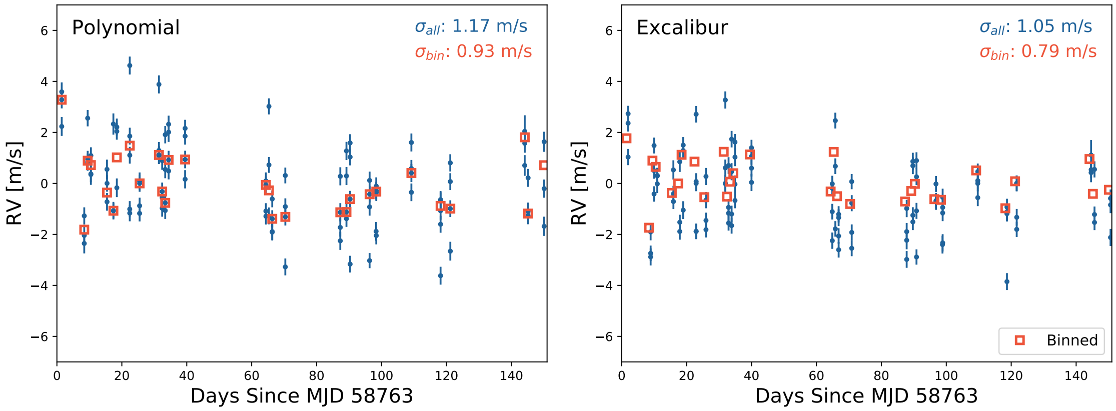

Figure 4.5: HD 34411 RVs measured with *EXPRES*. We zoom in on all but 8 of the exposures, which were taken over 150 days later. All RVs are shown in blue; nightly binned RVs are over-plotted as orange squares. The RMS of the data set and the RMS of the binned values are given in the top-left corner. **Left**: RVs derived using a polynomial-based wavelength solution. **Right**: RVs derived using wavelengths from the implementation of *excalibur* presented in this paper.

additive noise component of $0.52 m\,s^{-1}$ ($= \sqrt{1.17^2 - 1.05^2}$).

We conducted a direct test of a classically-generated wavelength solution with *excalibur*-wavelengths on four other data sets. All targets showed a reduction in or comparable RV RMS. The results from these data sets can not be interpreted as directly as with HD 34411, though, due to larger contributions from stellar variability, known planets, etc. As completely mitigating these different effects is out of scope for this paper, we focus here on the results with HD 34411.

4.5 Choose Your Own Implementation

We have described and tested only one implementation of *excalibur*. Using PCA and an interpolated wavelength solution is a statistically straight-forward step towards a complete hierarchical and non-parametric wavelength model. It is possible to upgrade both the denoising and wavelength solution steps to true models. It is also possible, of course, to implement either step individually. A hierarchical framework can be used to simply denoise the lines before they are fit to a parametric model, and a non-parametric model can be used on lines that have not been denoised.

For the dimensionality reduction and denoising, the PCA could be replaced by a probabilistic PCA model or other probabilistic linear reduction methods, such as heteroscedastic matrix factorization (HMF)(Tsalmantza & Hogg, 2012). It is also possible to move to non-linear reduction methods, like a Guassian process latent-variable model, an auto-encoder, a normalizing flow (e.g. Kramer, 1991; Woodbridge et al., 2020). Using a non-linear denoising model could enable *excalibur* to capture large-scale changes as well as small variations in calibration state.

The wavelength solution could also move past interpolation. For example, a Gaussian process could be used that is constrained to ensure monotonicity. Replacing each step with a model will allow for full, hierarchical Bayesian inference. This means the uncertainty from wavelength calibration could be completely marginalized out. Doing so will have the largest impact if the wavelength calibration is a significant fraction of the error budget.

The implementation of *excalibur* presented here, using PCA for denoising and interpolating a wavelength solution, uses various global variables and methods that we believe are or are close to optimal for constructing a high-fidelity wavelength solution. The following subsections will describe each choice and the associated decision-making process.

4.5.1 Dimensionality of the Calibration Space, K

The dimensionality of the calibration space within which the spectrograph lives is represented by K. In practice, it is the number of principal components, or basis vectors, used to reconstruct the de-noised line centers. K needs to be large enough so that all variability in the spectrograph is captured. Too large, however, and the reconstruction will begin to incorporate noise, thereby defeating the purpose.

Figure 4.6 shows the fiducial calibration state and the first 9 principal components

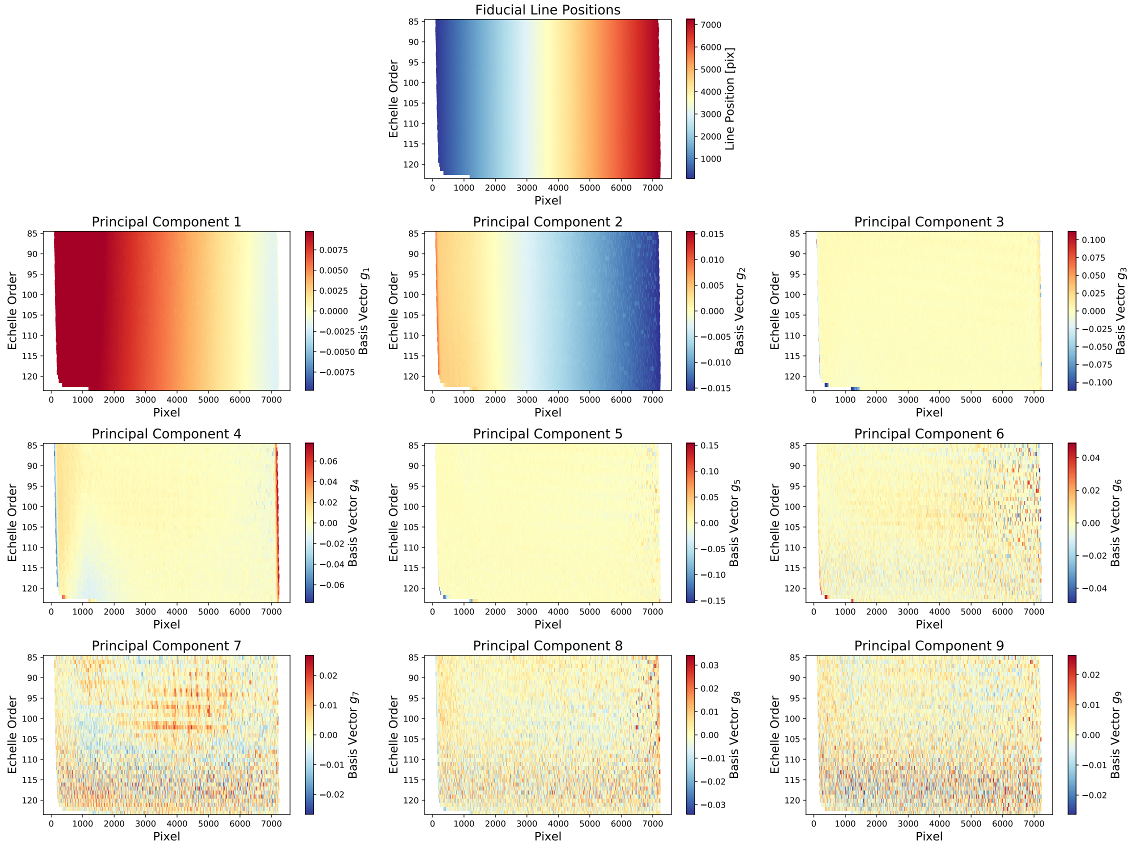

Figure 4.6: **Top**: The fiducial calibration of the spectrograph, i.e. the mean line positions for each line throughout the epoch of stability. **The following** 3×3 **grid of plots** show the first nine principal components constructed using LFC lines. These principal components represent the basis vectors along which the calibration of the spectrograph can deviate from the fiducial calibration. For each principal component, or basis vector, each calibration line is plotted according to its echelle order and x-pixel and colored by the value of the basis vector for that line. Principal components beyond the sixth one become steadily more dominated by noise.

constructed using LFC lines, which represent deviations from the fiducial calibration state. There is clear structure in the first and second principal components. Components three through six show smaller or more localized structure. Components three and four have aberrant behavior on the edges of the two bluest echelle orders, where lower signal results in more variation in the line fits. Later principal components become dominated by noise and show less coherent structure.

In deciding a K value, we ran denoising tests, as described in Section 4.4, for K values spanning from 2 to 512. The resultant per-line RMS for each test is plotted in Figure 4.7. One would expect the returned wavelengths to get increasingly more accurate with larger K until components that represent only noise are incorporated. Residuals might then get worse before ultimately starting to get better again with large K, which marks when the model starts over-fitting the data. Though the returned RMS never gets worse, we find that the improvement plateaus between $K = 6$ and $K = 128$. Comparisons of wavelengths returned by a $K = 6$ model vs. a $K = 32$ model show significant differences in less than 10 bluer lines, which are known to have greater error and variance in their measured line

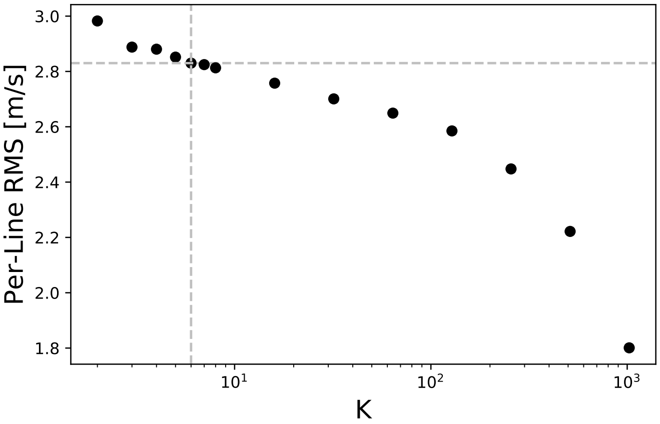

Figure 4.7: Per-line RMS of returned wavelength models for different values of K. Per-line RMS, as defined in Equation 4.5, provides a measure of the accuracy of a wavelength model. There is a dotted, vertical line at $K = 6$, and a dotted, horizontal line is the RMS for $K = 6$. The improvement at around $K = 6$ plateaus.

positions. We therefore settled on a K value of six.

4.5.2 Interpolation of Calibration State to Science Exposures

Figure 4.8 shows the amplitude of the first and second principal component with respect to time on the left. Though there exists a complex overall shape to the amplitudes with respect to time, a clear linear trend exists within each night. This is shown by the right plots in Figure 4.8. As the beginning-of-night and end-of-night calibration sets always include LFC exposures, we use a simple linear interpolation to interpolate principal component amplitudes with respect to time.

The choice in interpolation method can help guide how many wavelength calibration images are truly needed. It is unnecessary to take calibration images at times where the same information can be reconstructed at the desired precision by a well-chosen interpolation scheme. For example, with the *EXPRES* data shown here, it is clear that nightly calibration images are needed, but for a linear trend, only two calibration images throughout the night are strictly required.

We also tested an implementation of *excalibur* where the K principal components within a night were fit to a cubic with respect to time rather than linearly interpolated. This emulates the current, polynomial-based wavelength solution implemented in the *EXPRES* pipeline, where polynomial fits to calibration files are interpolated to science exposure by fitting polynomial coefficients with respect to time to a cubic. We found that using a cubic in place of linear interpolation returns comparable RV RMS for most targets, though appears to do better when a night has sparse calibration data. This suggests that the nightly behavior of *EXPRES* with respect to time is well described by a cubic function, but LFC exposures are typically taken with enough frequency that a linear interpolation provides a good approximation (see Figure 4.8)

The amplitudes $a_{n,k}$ can also be interpolated with respect to any good housekeeping

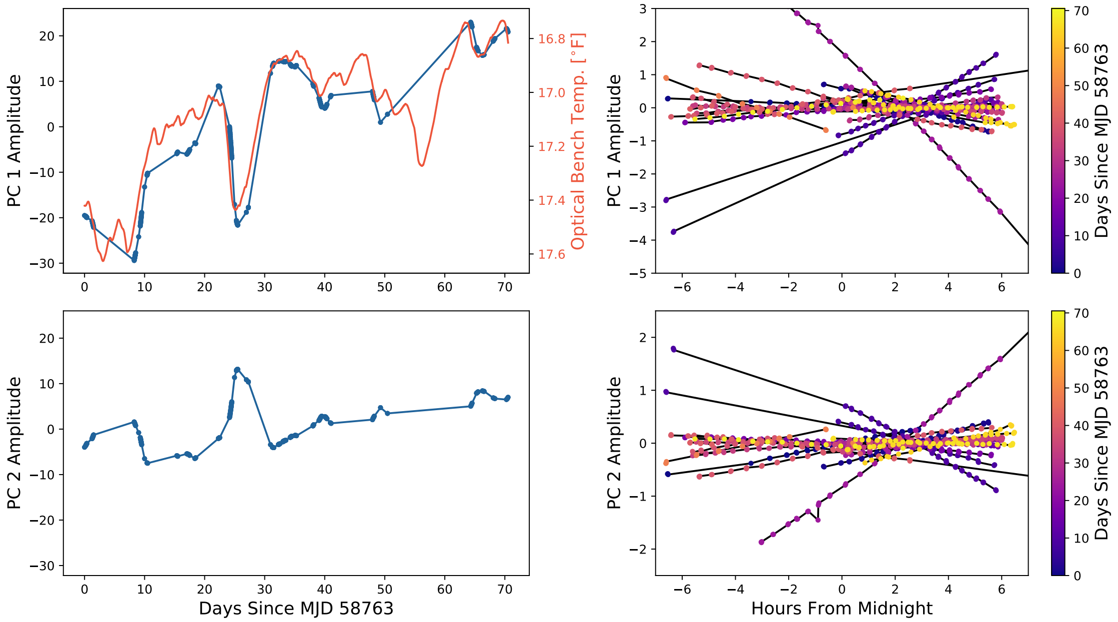

Figure 4.8: Amplitude of the first two principal components shown as a function of time (left) or hour from midnight (right). The top row of plots shows the amplitudes for the first principal component while the bottom row shows the amplitudes for the second principal component. Lines show the result of a linear interpolation. In the top, right plot, the temperature of the optical bench is also plotted in orange. In the right plots, the principal component amplitudes for each night have been artificially offset by the median amplitude per night. All days are therefore roughly on the same scale, but the y-axis is different from the left plots. In the right column, points are colored by the MJD of each exposure.

data, not just time. Best results will come from interpolating with respect to whatever is most strongly correlated with the calibration state of the spectrograph. For example, with *EXPRES*, which is passively temperature controlled, the returned amplitudes $a_{n,k}$ were extremely correlated with the optical bench temperature, as shown in the top-left plot of Figure 4.8, suggesting it would also be possible to interpolate the amplitudes with respect to temperature.

Another pertinent example would be a spectrograph that is mounted on the telescope, and therefore moves with the telescope. In this case, it may be important to interpolate at least in part with respect to the position of the telescope, which enables the resultant calibration to incorporate the gravitational loading experienced by the spectrograph.

4.5.3 Interpolation of Wavelengths with Respect to Pixel

In the implementation described and tested by this paper, interpolation of wavelengths over pixel is done order-by-order using a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolator. This interpolation incorporates the flexibility needed to model the changing dispersion of the spectrograph across an echelle order along with any detector defects while also enforcing monotonicity, which we know must be true across any one echelle order.

A simple linear interpolation would give erroneously low values everywhere. Due to the dispersion intrinsic to echelle spectrographs, the wavelength change between pixels grows greater with greater wavelengths, even within an order. This means that the function of wavelength vs. pixel across an echelle order will always be monotonically increasing and concave down everywhere.

Though less of an issue with LFC lines, a more classic cubic spline interpolation can run into issues with arc lines, which are irregularly spaced or even blended. Close lines appearing at virtually the same pixel location but with different wavelengths could coerce

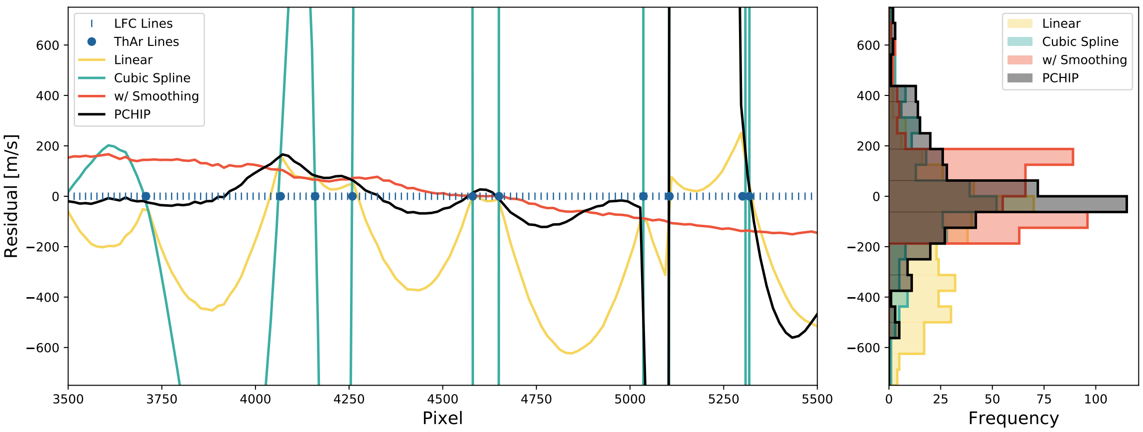

Figure 4.9: Residuals from different interpolation schemes over pixels in echelle order 100. ThAr lines, shown as blue circles, are used to construct a wavelength solution that is then evaluated at each LFC line, shown as blue vertical lines. The residuals of each wavelength solution for a subset of the order is shown on the left. Histograms of the residuals for each method for the complete order is shown on the right. Note: there is a blended ThAr line at approximately pixel 5300, the right most ThAr lines plotted.

a cubic spline into a very high slope. This is demonstrated by the green line in Figure 4.9, which shows the results of interpolating between ThAr lines rather than LFC lines. A line blend appears at approximately pixel 5300, causing the spline to twist to nearly vertical to account for both points. This leads to huge deviations from the correct wavelengths around this line blend as the extraneously high slope of the spline is accounted for.

These huge digressions can be avoided by allowing for some smoothing in the interpolation. In Figure 4.9, we show an example in orange using SciPy’s implementation of a univarate spline. While the result appears to follow the calibration lines much better, the smoothing ultimately causes larger residuals that are spatially correlated (Fig. 4.9, right). In all echelle orders, the edges are overestimated while the middle will be underestimated, shown by the flattened shape of the histogram of residuals. The resultant wavelength solution is underestimating the curvature of the pixel-wavelength relation, giving rise to similar issues as with an inflexible, parametric wavelength solution. Introducing this smoothing parameter is enforcing a smoothness we have no reason to believe is true of the data, thereby re-introducing one of the problems with parametric models.

We instead turn to the PCHIP algorithm, which damps down huge deviations in the traditional cubic spline by requiring the resulting interpolated function be monotonic. Monotonicity is a constraint we know must be true for any one echelle order. Though the PCHIP interpolator shows a similar issue as a classic cubic spline around the ThAr line blend at pixel 5300, the effect is much smaller and affects fewer pixels. Figure 4.9, right, shows that using the PCHIP interpolator returns the lowest spread residuals.

There likely exists an even more fitting model between an overly-constrained polynomial fit or a completely free spline interpolation. For example, there has been success interpolating wavelengths with respect to pixel using a segmented polynomial in the dispersion direction, especially when tuned to known detector defects (Milaković et al., 2020). Stiffer, more constrained splines or carefully chosen knot position may afford the perfect marriage of freedom and constraint that will better describe wavelength changes with pixel.

4.6 Application to Other Spectrographs

We focus on an EPRV use-case here because there is a strong need for wavelength calibration in the EPRV community. *EXPRES* is representative of the newest generation of EPRV spectrographs, and an LFC provides stable, dense calibration lines with known wavelengths, ideal for *excalibur*. The applicability of *excalibur* to any one instrument is a detailed question of the kind of variance experienced by the spectrograph and calibration sources available, but we hope to provide some approximate benchmarks and diagnostics here.

Implementing *excalibur* will require an existing library of calibration exposures that span the range of calibration space accessible by the spectrograph, or at least the space accessed by the science exposures being calibrated. A new set of calibration exposures would be needed for any changes to the instrument that could significantly change the accessible calibration space. While there is no exact cutoff for how many exposures are needed, the number is certainly much larger than K, the dimensionality of the calibration space. More calibration exposures will help with denoising . The number of calibration exposures required throughout a night will depend on how much the instrument varies throughout the night as well as the chosen interpolation scheme, as mentioned in Section 4.5.2.

As an empirical assessment of the needed line density, we removed LFC lines from the *EXPRES* data and calculated the per-line RMS of the returned wavelengths for the removed lines. These tests are similar to the interpolation test explained in Section 4.4, in that a fraction of lines are systematically removed from the analysis. An increasing fraction of lines are removed to simulate different line densities. For these line density tests, though, we are also implementing denoising unlike the pure interpolation test of Section 4.4.

Figure 4.10 plots the scatter of residuals of the wavelengths returned by *excalibur* as a function of separation between lines in units of resolution element. It gives an approximate estimation of the expected error between lines of different separation. Note, however,

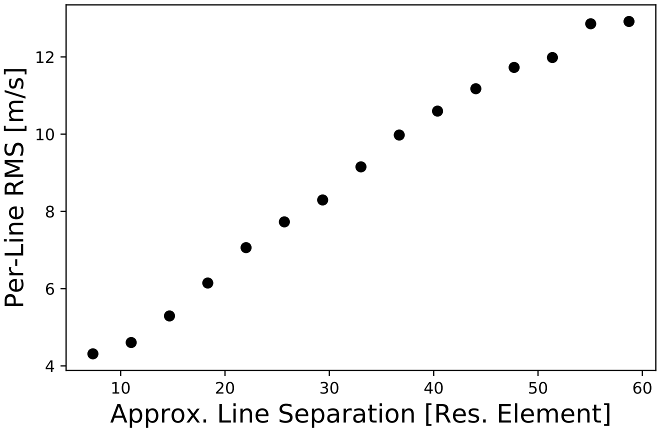

Figure 4.10: Per-line RMS as a function of spacing between lines in units of resolution element. The average line spacing is calculated using the average distance between LFC lines across its wavelength range.

that the stability of the lines and their measured line centers quickly becomes the dominant error source in calibration lamps over the separation between lines. Additionally, in this assessment, the lines remain uniformly spaced. The needed line density depends on the resolution of a given spectrograph, the needed precision of the returned wavelength solution, and the chosen interpolation scheme.

This implementation of *excalibur* to *EXPRES* data interpolates calibration information to the science data using surrounding calibration exposures. Simultaneous calibration data can be used to reinforce the determined calibration state of an instrument at the time of science exposures. This simultaneous calibration data can come from calibration light shined through a reference fiber or any meta-data (e.g. temperature, pressure, time, etc.) that correlates with the calibration. For example, we have seen that the calibration state of *EXPRES* is correlated with the temperature of the optical bench, even with the optical bench temperature varying by less than one degree. This correlation is not seen in the RVs, suggesting the changes with temperature are calibrated out at the wavelength stage.

With a simultaneous reference fiber, the position of calibration lines taken through the reference fiber can simply be concatenated to the array of line positions taken through the science fiber. Both sets of lines will then be used when the low-dimensional basis is constructed. This allows the simultaneous calibration information to contribute to constructing the complete calibration space of the spectrograph and pinpoint where the spectrograph is in that calibration space for any exposure.

The calibration for any science exposure with a simultaneous reference can be determined by finding the amplitude of each basis vector that most closely recreates the calibration line positions through the reference fiber. These amplitudes can then be used to recreate the calibration line positions through the science fiber as well. This replaces the need to interpolate the basis vector amplitudes from calibration exposures to science exposures, something that is done with respect to time in the example implementation described in this paper. The result is likely to be even more precise, as this method incorporates more data. This method, as with all analysis involving a simultaneous reference fiber, will work only as well as the reference fiber's ability to trace changes in the main science fiber.

It is also possible to apply *excalibur* to etalon data with some modifications. The simplest implementation is if the free spectral range (FSR) and therefore wavelength of each line of the etalon is well known. The etalon data can be interpolated onto a set of fiducial lines with set wavelengths. These fiducial lines would therefore be identifiable by echelle order and wavelength alone, with only their line positions varying with different exposures. This returns us to the same framework as developed for the case of an LFC. This marks the simplest implementation of *excalibur* on etalon data, as the uncertainty of a line's wavelength is upstream of the model rather than built in.

Incorporating the FSR as part of the *excalibur* model will require introducing a free parameter to capture changes in the FSR independent of variation in an instrument's calibration state. The calibration state can then be described with respect to mode number, which will be used to uniquely identify a calibration line across exposures rather than wavelength. The FSR is then used to determine how the mode number of each line maps to wavelength for a given exposure. The FSR must not vary so much that the change in this mode-number-to-wavelength mapping becomes non-linear. This model would require a simultaneous reference or other housekeeping data that can be used to determine the FSR for every exposure.

In terms of dimensionality reduction, most physical systems should have only a few dominant axis along which they vary, meaning *excalibur* should be adaptable to a wide range of instrument designs. With PCA, this can be tested by plotting the amplitude of the returned principal components, which should fall quickly after a few components. It should be noted that this only provides a measure of the variance in the PCA space, and is not an explicit test of the end-to-end variation in the resulting model. This condition is therefore necessary but not sufficient if implementing *excalibur* with PCA.

It could still be possible to run *excalibur* on a spectrograph that has a high-dimensional calibration space, meaning a large number of basis vectors are required to capture all the variance in the spectrograph. In this regime, there is always the risk of over-fitting. Regularizing the principal component amplitudes, for example insisting the amplitudes for higher principal components be smaller, can help to return reasonable models (Foreman-Mackey et al., 2015) . Within such a framework, *excalibur* may still deliver good results.

For the results presented here, the data was broken up into different eras of stability based on where the principal component amplitudes showed huge deviations. This was done visually, though there exists many change-point detection algorithms that could be used (Aminikhanghahi & Cook, 2017). There is a trade-off in including more exposures between introducing greater variation, but also more data to offer constraints that may be optimized. Here, an era of stability was chosen manually in order to focus on separating out time-domain intervals in which the data is relatively homogeneous, e.g. most exposures show the same calibration lines. Homogeneity is, of course, implicitly required when implementing PCA. Different denoising models will be able to account for different amounts of stability or lack thereof.

Lastly, we caution that *excalibur* is extremely sensitive to upstream changes that may effect the line centers. For example, PCA is good for detecting variability, but is agnostic to the source of the variability. This is why the principal components shown in Figure 4.6 exhibit errant values for bluer LFC lines, which are lower signal and therefore exhibit more variation in their fitted line centers. It is essential that the line positions being fed to *excalibur* capture only the changes in the spectrograph's calibration state, not potential errors in the fitted line centers.

4.7 Discussion

We show that *excalibur* returns a lower per-line RMS than classic, parametric methods by a factor of 5 (Section 4.4). The residuals were also smoother, exhibiting less spatial correlation (Figure 4.3). Using *excalibur* wavelengths reduced the RMS in RVs of HD 34411 from 1.17 m s^{-1} to 1.05 m s^{-1}(Section 4.4.2).

In implementing *excalibur* on *EXPRES* data, we have successfully constructed a model of *EXPRES*'s accessible calibration space, confirming that *EXPRES* truly is an instrument with low-degrees of freedom. *Excalibur* does not make any claims about what variability each basis vector represents. Those interested in interpreting the variability are encouraged to investigate how the amplitude of the different vectors varies with different housekeeping data to find their source.

Starting with a list of calibration lines with known wavelengths and well-fit line centers for each calibration exposure, *excalibur* will de-noise and interpolate the given lines into a full wavelength solution. *Excalibur* leverages the stability of contemporary EPRV instruments and high density of lines made available by new calibration sources, such as LFCs and etalons, to achieve more accurate wavelengths. *Excalibur* therefore assumes dense enough calibration lines to properly constrain a non-parametric wavelength model, and that the instrument has low degrees of freedom.

Denser calibration lines allow us to move to more flexible wavelength models, which can then account for non-smooth features in the wavelength solution. Stabilized spectrograph hardware makes it more likely that the calibration space of the instrument is low-dimensional. All calibration images in a given generation of stability can therefore be used to constrain the accessible calibration space of the spectrograph as well as where in that calibration space the spectrograph lies. We have described only one, fairly simplistic implementation of *excalibur* here. There are many options for both the de-noising and interpolation steps, as mentioned in Section 4.5.

An advantage of this implementation of *excalibur*, where PCA is applied to line positions from all LFC exposures, is the ability to isolate exposures that exhibit errant variation, which is typically associated with flawed exposures. This allowed us to quickly vet for problematic LFC exposures, which otherwise would have required visual inspection of all 1200+ LFC exposures. In a classic framework where each calibration exposure is treated independently, these aberrant exposures would likely have persisted undetected and are liable to sway the resultant wavelength solutions for at least an entire night.

On the other hand, PCA captures all variance, regardless of source. Though *excalibur* endeavors to capture only variation in the instrument, the PCA is also sensitive to uncertainties and failures in the upstream line-position fitting. For example, we have seen that lower-signal lines that are harder to fit faithfully will have greater variety in returned line positions, which is in turn captured by the PCA. In this sense, *excalibur* is actually a model of not just the instrument, but all the upstream choices used to drive live positions. High-fidelity line positions are essential to ensure the PCA is capturing variations in just the spectrograph's calibration rather than changes in how well a line can be fit or other effects.

Along those lines, we caution that with any wavelength solution, there is a perfect degeneracy between what is defined as the "position of the line" and the resultant wavelength solution. If, for example, a cross correlation method is used to extract RVs from the data, a systematic difference may be introduced depending on what exactly is defined to be the line position, whether it be the mode of a Gaussian fit, the first moment of a complicated PSF, or the output of some particular peak-finding algorithm, etc. In principle, the best way to mitigate this uncertainty would be to use a calibration source that looks like a star.

Compared to traditional methods, which involve fitting high-order polynomials, *excalibur* has several useful statistical properties. *Excalibur* is technically robust to localized issues that arise from either the calibration source or the pipeline used to return line positions. With an interpolated wavelength model, one errant line position will only effect the resultant wavelength model out to the close, neighboring lines. The effect of an outlier is diminished and kept localized. In contrast, an entire parametric fit will be affected by a single errant line, changing the wavelength solution for the entire exposure for a 2D fit. Through de-noising and outlier rejection, *excalibur* adds additional robustness against erroneous line positions.

The locality of the interpolation caries other benefits as well. Manufacturing artifacts in the detector or other optical elements can lead to non-smooth structure in the wavelength solution that can not be captured by polynomials or other smooth functions (see Figure 4.3). An interpolated model introduces greater flexibility, enabling the model to account for such high-order effects. As discussed in Section 4.5.3, there are better and worse interpolators for the task, which may differ for different instruments and different calibration sources. Instead of using an interpolator at all, there might be better results from implementing something more sophisticated, such as a kernel method or a Gaussian process with a kernel adapted for the specifics of an instrument. There is in principle an enormous number of non-parametric methods to explore, which we leave outside the scope of this paper.

Similarly, PCA is just one of many possible dimensionality-reduction methods. We chose to implement *excalibur* using PCA here for simplicity and computational tractability. PCA is a good option because the instrument changes here are small enough that a linear model is an approrpiate representation of the changes. If *excalibur* was updated to a full probabilistic model, the PCA along with the interpolation model would have to be upgraded to something with better probabilistic properties. Other, nonlinear denoising methods may be more robust to large changes, allowing all calibration images ever taken with an instrument to be used to construct the accessible calibration state regardless of hardware adjustments. Further discussion of other implementations of *excalibur* can be found in Section 4.5.

Excalibur can be applied to any data that contains information about the calibration state of the spectrograph (see Section 4.6). For example, though LFC and ThAr exposures are used as an example in this paper, *excalibur* would work similarly for an etalon or any other arc lamp with a list of lines and assigned wavelengths. Simultaneous calibration information can easily be accounted for by simply including the line position information from the simultaneous reference when constructing a low-dimensional basis of the instrument's calibration space.

Once we have defined a calibration space that captures all possible degrees of freedom for a stabilized spectrograph, there are many options for pinpointing where the spectrograph is located within that calibration space. Good housekeeping data, such as temperature or pressure could be used in addition to or instead of time (as mentioned in Section 4.5.2). Telemetry that is seen to be correlated with the calibration state of the spectrograph can even be added to the data used to construct the low-dimensional basis. Furthermore, all exposures taken with the spectrograph in principle contains information about the calibration state of the spectrograph. Theoretically, tellurics, lines in science exposures, or just the trace positions themselves could also be used to determine an instrument's calibration state, thereby providing free simultaneous calibration information.

Excalibur is designed and optimized for extreme-precision, radial-velocity spectrographs. In the battle to construct a high-fidelity data pipeline for extreme-precision radial-velocity measurements, we have shown that *excalibur* represents a step towards mitigating the error from wavelength calibration, as demonstrated by tests using *EXPRES* data (Section 4.4). Though the focus was on EPRV instruments here, *excalibur* should be largely applicable to nearly any astronomical spectrograph.

4.8 Acknowledgements

We gratefully acknowledge the anonymous referee whose insightful and well thought out comments significantly improved the paper. We thank the EXPRES team for their work on the instrument, software, and pipeline. We thank Dan Foreman-Mackey for illuminating discussion. The data presented here 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. We gratefully acknowledge ongoing support for telescope time from Yale University, the Heising-Simons Foundation, and an anonymous donor in the Yale community. We especially thank the NSF for funding that allowed for precise wavelength calibration and software pipelines through NSF ATI-1509436 and NSF AST-1616086 and for the construction of EXPRES through MRI-1429365. LLZ gratefully acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant No. DGE1122492.

Software

SciPy library (Virtanen et al., 2020), NumPy (Oliphant, 2006–; van der Walt et al., 2011), Astropy (Astropy Collaboration et al., 2013; Price-Whelan et al., 2018).

Facilities

LDT

Chapter 5

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

Lily L. Zhao[1]*, Debra A. Fischer, Gregory W. Henry, Eric B. Ford, Alex Wise, Michael Cretignier, Rachael M. Roettenbacher, Samuel H.C. Cabot, Suzanne Aigrain, Oscar Barragan, Megan Bedell, Lars Buchhave, Joãoo Camacho, Heather Cegla, Jessi Cisewski-Kehe, Andrew Collier Cameron, Zoe L. de Beurs, Sally Dodson-Robinson, Xavier Dumusque, João Faria, Christian Gilbertson, Charlotte Haley, Justin Harrell, David W. Hogg, Parker Holzer, Ancy Anna John, Baptiste Klein, Marina Lafarga, Catherine Lembo, Florian Lienhard, Annelies Mortier, Belinda Nicholson, Michael Palumbo, Vinesh Rajpaul, Victor Ramirez Delgado, Christopher J. Shallue, Andrew Vanderburg, Pedro Viana, Jinglin Zhao, Norbert Zicher, John M. Brewer, Andrew E. Symkowiak, Ryan R. Petersburg, Joe Llama*

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

---

[1] The project is originally described in: Zhao, L., Fischer, D. A., Ford, E. B., et al. 2020, Research Notes of the American Astronomical Society, 4 156. I processed all the provided data including spectra, CCFs, and activity indicators. As intellectual lead for the ESSP, I analyzed the results from each method and wrote the summary report. I also developed and contributed results for the *ResRegGen* and *ResRegDis* methods.

radial velocity (EPRV). 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 23 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 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 reducing the RV RMS to the necessary sub-meter-per-second levels. There exists a concerning lack of agreement between methods, even for those returning similar final RV RMS values. This highlights the danger of using the RMS alone to assess method performance, a common practice that must be used with caution. 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.

This work is ongoing.

5.1 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., 2016a; Jurgenson et al., 2016; Blackman et al., 2020; Petersburg et al., 2020; Suárez Mascareño et al., 2020; Brewer et al., 2020). 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), 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; Lovis 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, and causes spectral lines to appear asymmetric.

Different realizations of granulation, which change on the timescale of a few minutes to hours, will integrate 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 change 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). 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 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 shown success in 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; Faria et al., 2016; Rajpaul et al., 2017; Angus et al., 2018). 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). 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 *EXPRES* data that are representative of next-generation spectrographs as well as differential photometry from the *APTs* (Zhao et al., 2020). Eleven teams tested 16 different methods (with different variations for a total of 23 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 5.2. Section 5.3 gives an overview of all methods tested and highlights commonalities between methods. The resultant RVs produced by different methods are compared in Section 5.4. Section 5.5 gives a summary of all methods and the pertinent results. Section 5.6 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 5.7.

5.2 Data

The data sets for the *ESSP* include spectroscopic data from the Extreme Precision Spectrograph (*EXPRES*) and ground-based photometric measurements from the Fairborn Automatic Photoelectric Telescopes (*APTs*) for four targets—HD 101501 (61 UMa), HD 26965 (omi^2 Eri), HD 10700 (τ Ceti), and HD 34411 (lam Aur). Here, we describe the *EXPRES* and *APTs* instruments, as well as the four targets. We provide benchmarks for the amount of scatter that is expected for the *EXPRES* instrument and pipeline independent of contributions from stellar signals. Stellar parameters for each target are tabulated in Table 5.1.

Table 5.1: Stellar Parameters of *ESSP* Targets.
Citations are as follows: (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)

	HD 101501		HD 26965		HD 10700		HD 34411	
Spectral Type	G8V B		K1V		G8V		G0V	
M_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)
$\log R'_{HK}$	-4.483	(g)	-4.928	(g)	-4.976	(g)	-5.085	(g)
π [mas]	104.39	(f)	200.62	(f)	273.96	(k)	80.105	(f)
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_{star} [$L_\odot$]	0.609 ± 0.009	(b)	0.457 ± 0.002	(f)	0.52 ± 0.03	(j)	1.732	(b)
R_{star} [$R_\odot$]	0.86 ± 0.02	(c)	0.83 ± 0.02	(c)	0.82 ± 0.02	(c)	1.28 ± 0.04	(c)
M_{star} [$M_\odot$]	0.9 ± 0.12	(c)	0.8 ± 0.11	(c)	0.99 ± 0.13	(c)	1.08 ± 0.14	(c)
T_{eff} [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)		

5.2.1 Spectroscopic Data From *EXPRES*

EXPRES 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 ~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 5.1 shows the RV scatter of over an hour of consecutive LFC exposures. The resultant RMS is 3.21 cm s^{-1} after a linear trend is removed. The instrument calibration

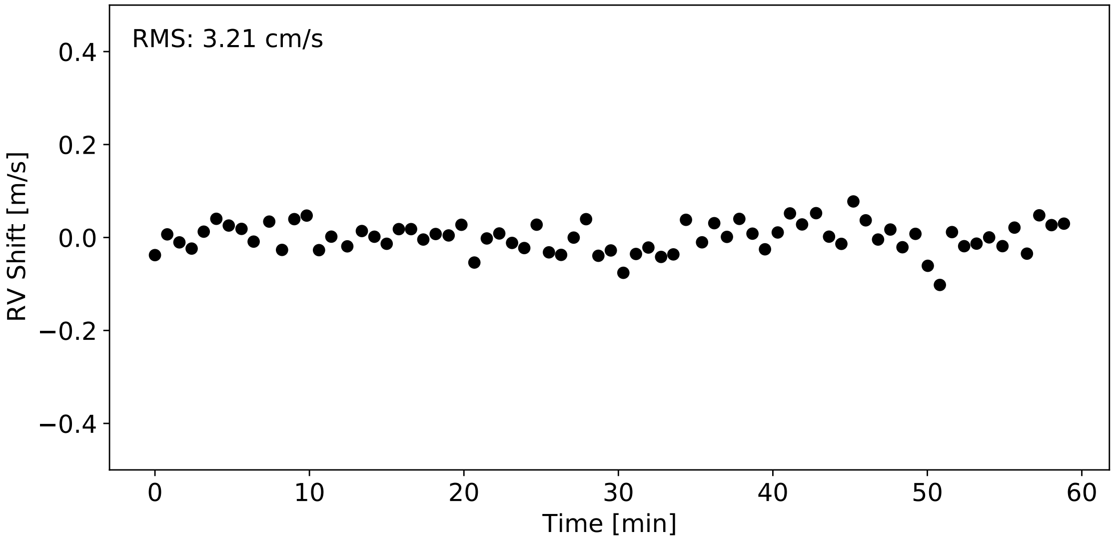

Figure 5.1: Perceived shift in LFC spectra in units of m s^{-1} across an hour of consecutive LFC exposures. These perceived shifts are attributed to variations in the instrument and therefore give a measure of the instrument stability. The RMS of shifts across this hour is given in the top-left corner.

stability can be thought of as the minimum RMS achievable by the *EXPRES* hardware.

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.

Two or three consecutive exposures 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 λ=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. Unless 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.

5.2.1.1 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. Chunks that behave more stably over time are weighted higher while chunks that return a high RV scatter are down weighted. CBC RVs for all four targets are given in Table 5.2.

Table 5.2: CBC RVs. A stub of this table is provided here for reference; the full RV data sets are published online

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
	⋮	⋮	⋮

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 5.2 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

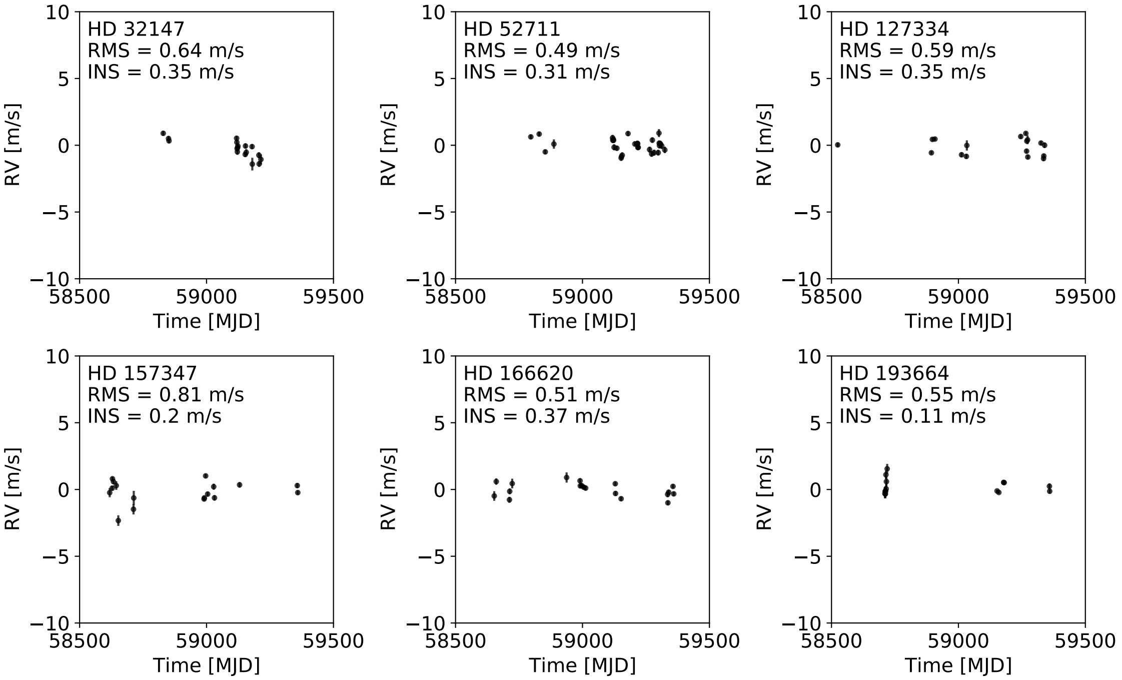

Figure 5.2: *EXPRES* RVs for six quiet stars. Shown RVs are derived using a chunk-by-chunk (CBC) forward modeling scheme and binned by night. The RMS of these nightly-binned RVs are given in the top-right corner along with the average intra-night scatter (INS).

achievable by *EXPRES* data in the absence of strong photospheric velocities adding scatter. Complete mitigation of RV contribution from stellar signals should result in a similar final RMS values.

5.2.1.2 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. Note that 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.

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 (~490-730 nm) were used to construct the CCFs.

5.2.1.3 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 Hα line core emission (Skelly et al., 2008; Robertson et al., 2014; Giguere et al., 2016). Both the Hα emission (i.e. a measure of the depth of the normalized Hα line), and the equivalent width of the Hα line were provided.

We also provided a number of activity indicators derived from the CCF. 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-Gausisan (Figueira, 2013) or a skew normal probability distribution 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 seven chromospherically quiet stars. Using a total of approximately 400 observations of these seven stars, a histogram of indicator values was plotted and fit to a Gaussian. The width of this Gaussian is taken to be the empirical error for the given activity indicator.

5.2.2 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. The data presented here were largely taken with T4, with some of the photometry for HD 34411 taken with T12. The T12 APT went out of operation after the 2017-18 observing season, but operated in the same way as the T4 APT.

The T4 APT is equipped with a single channel photometer that uses an EMI 9124QB bi-alkali photomultiplier tube to measure the difference in brightness between the program star and three nearby comparison stars in the Strömgren *b* and *y* passbands. To improve the photometric precision, we combine the differential *b* and *y* magnitudes into a single $(b + y)/2$ “passband". The precision of a single observation with T4 or T12, as measured from pairs of constant comparison stars, is around ~0.0015 mag on good nights. Both the T4 and T12 APT are described in Henry (1999), where further details of the telescope, precision photometer, and observing and data reduction procedures can be found.

Several years of photometry from the *APTs* were available for each *ESSP* target. 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 5.3.

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 from the *george* package (Ambikasaran et al., 2015). The GP was trained on the most recent six years of APT data. While the GP regression returned reasonable interpolated median values and 1σ uncertainties, we were unable to estimate well-principled extrapolated photometric values for RV timestamps taken after the last photometric measurement.

Table 5.3: Photometry and Long-Term Trend. A stub of this table is provided here for reference; the full photometric data sets are published online.

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
	⋮	⋮	⋮

5.2.3 Targets

The four *ESSP* stars, as described in Table 5.1, are being observed as part of the *EXPRES* 100 Earths survey (Brewer et al., 2020). The targets range in activity level as predicted by $\log R'_{\rm HK}$ values. For each target, Table 5.4 gives the number of RV measurements, the number of nights where spectra were acquired, and the time baseline for each data set. Figure 5.3 shows the measured RVs and photometry.

Table 5.4: Amount of Spectroscopic and Photometric Data for *ESSP* Targets

	RVs		
HD Num.	No. Spec.	Nights	Start/End Date
101501	45	22	Feb. 10, 2019 - Nov. 26, 2020
26965	114	37	Aug. 20, 2019 - Nov. 27, 2020
10700	174	34	Aug. 15, 2019 - Nov. 27, 2020
34411	188	58	Oct. 8, 2019 - Nov. 27, 2020
	Photometry		
HD Num.	No. Phot.	Nights	Start/End Date
101501	3290	2113	Apr. 18, 1993 - Jun. 22, 2020
26965	1631	1500	Sep. 9, 1993 - Feb. 20, 2020
10700	1369	1007	Nov. 5, 1996 - Jan. 24, 2020
34411	1214	816	Nov. 25, 2005 - Apr. 3, 2018

HD 101501 is the most chromospherically active (Isaacson & Fischer, 2010, $\log R'_{\rm HK}= -4.483$) 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).

HD 26965 is a K1V star with $\log R'_{\rm HK}= -4.928$ (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 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).

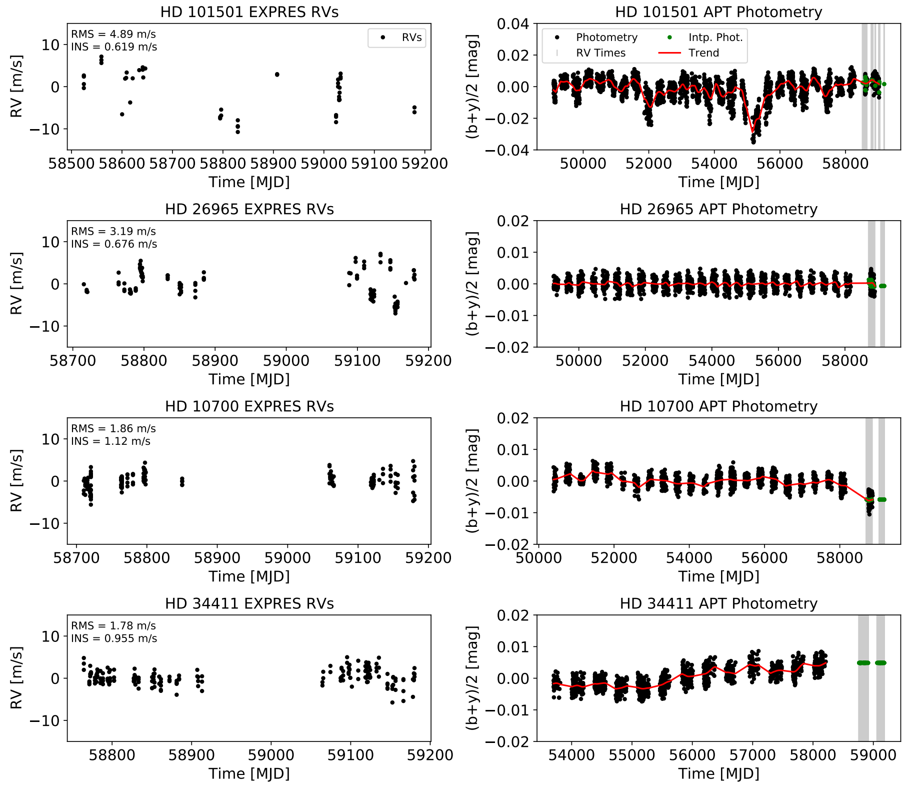
HD 101501 EXPRES RVs
HD 101501 APT Photometry
HD 26965 EXPRES RVs
HD 26965 APT Photometry
HD 10700 EXPRES RVs
HD 10700 APT Photometry
HD 34411 EXPRES RVs
HD 34411 APT Photometry

Figure 5.3: (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 of the photometric data to the RV timestamps are over plotted as green points.

HD 10700, i.e. τ Ceti, is an older (Brewer et al., 2016, 12.4 Gyr) and chromospherically quiet (Isaacson & Fischer, 2010, $\log R'_{\mathrm{HK}} = -4.976$) 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 consecutive observations were taken, covering a span of approximately 40 minutes, to search 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 $\log R'_{\mathrm{HK}} = -5.085$ (Isaacson & Fischer, 2010). The *EXPRES* RVs show an RMS scatter of 1.78 m s^{-1}.

5.3 Methods

Eleven teams tested sixteen different methods with the goal of isolating out true center-of-mass shifts. Table 5.5 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, whether 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 individual method can be found in latter sections and the provided appropriate references.

Methods are grouped into subsections according to the type of input data used: the RVs with global indicators (§5.3.1), the CCFs (§5.3.2), or the extracted (pixel-level) spectra (§5.3.3 and 5.3.4).

Table 5.5: Teams and Methods

Team	Method	Input	Run Time	Reference/Contact
PennState	*GLOM*	RVs/Indicators	?	Gilbertson et al. (2020)
Sidera	*FDPCA*	RVs/Indicators	laptop, seconds	Sally Dodson-Robinson
Porto	*GPRN*	RVs/Indicators	cluster, hour	Camacho et al., in prep.
Andrews/PennState	*SCALPELS*	CCFs	laptop, seconds	Collier Cameron et al. (2020)
PennState	*SCALPELS*+*GLOM*	CCFs	laptop, minutes	Gilbertson et al. (2020)
OxBridGen	*CCF Prime*	CCF	desktop, minutes	Baptiste Klein
PennState	*FIESTA*+*GLOM*	CCFs	laptop, minutes	Zhao & Tinney (2020)
ML_EPRVs	CCF Linear Regression (LR)	CCFs	laptop, seconds	de Beurs et al. (2020)
ML_EPRVs	CCF LR + Hα	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	*LBL*+$\text{PCA}_{\text{Spec.}}$	Spectra	laptop, hours	Dumusque (2018) Cretignier et al., in prep.
Geneva	*LBL*+PCA_{RV}	Spectra	laptop, hours	Cretignier et al., submitted
Geneva	*LBL*+$\text{PCA}_{\text{Spec./RV}}$	Spectra	laptop, hours	Cretignier et al., submitted
OxBridGen	*PWGP*	Spectra	desktop, day	Rajpaul et al. (2020)
PennState	*DCPCA*	Spectra	laptop, seconds	Jones et al. (2017)
PennState	*DCPCA*+*GLOM*	Spectra	laptop, minutes	Gilbertson et al. (2020)
YaleWI	*HGRV*+*SAFE*	Spectra	cluster, minutes	Holzer et al. (2020) Holzer et al. (2021)
LienhardMortier	*ZLSD*	Spectra	laptop, hour	Lienhard et al., in prep.
CCA	*ResRegGen* Self	Spectra	laptop, hour	Lily Zhao
CCA	*ResRegGen*	Spectra	laptop, hour	Lily Zhao
CCA	*ResRegDis*	Spectra	laptop, minutes	Lily Zhao

5.3.1 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.

More recent work has developed novel ways of linking indicators to RV measurements and modeling out the stellar component of the RV measurements. Indicator-dependent 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 *MDGP* method described in previous work and implemented in *pyaneti* (Rajpaul et al., 2015; Barragán et al., 2019). 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. (2020) or Section 5.8.1.

Fourier Domain Principal Component Analysis (*FDPCA*), developed by the Sidera team, translates RV and indicator time series into the Fourier domain to capture coherences between the different, likely periodic time series regardless of whether they are in phase. The Fourier series for the RV and indicator time series are run though a principal component analysis (PCA) to derive activity-indicator-related axes of variation and the magnitude of this variation. The results presented here were trained on the RV, Hα 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 Section 5.8.2.

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 Section 5.8.3.

5.3.2 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 resultant 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 present; 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 component. The results presented here use only the first two principal components to guard against incorporating noise into the model. *SCALPELS* operates in the wavelength-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.

The *SCALPELS+GLOM* method is another use of PCA. The amplitudes of the principal components for each observation, which describe the magnitude 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 Section 5.9.1.

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 Section 5.9.2.

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 similar 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 Section 5.9.3.

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 Hα 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 Section 5.9.4.

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.

5.3.3 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 (§5.3.4).

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 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 Section 5.10.1

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 Section 5.10.2.

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 5.2). 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 ~30 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$_{\text{Spec.}}$, or produce a metric of variation at the line-by-line RV level, *LBL*+PCA$_{\text{RV}}$. 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$_{\text{RV}}$ method.

Both methods can also be combined, which are here represented by the *LBL*+PCA$_{\text{Spec./RV}}$ method. Though both *LBL*+PCA$_{\text{Spec.}}$ and *LBL*+PCA$_{\text{RV}}$ 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$_{\text{Spec.}}$, *LBL*+PCA$_{\text{RV}}$, and *LBL*+PCA$_{\text{Spec./RV}}$ implementations represented in this report can be found in Section 5.10.3.

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 areas 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 Section 5.10.4.

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.

5.3.4 Methods That Model the Spectra

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.

The Doppler-Constrained Principal Component Analysis (*DCPCA*) method, submitted by the PennState team, runs PCA on spectra shifted by the best-guess 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$_{\mathrm{RV}}$ 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. (2017) or Section 5.11.1.

The Hermite-Gaussian Based Radial Velocity (*HGRV*) method and Stellar Activity F-statistic for Exoplanet Surveys (*SAFE*) statistic, submitted by the YaleWI team, similarly uses information from the entire spectrum to derive an activity indicator. The *HGRV* method derives RVs by fitting the difference between observed lines and a template spectrum to a first-order Hermite-Gaussian function, the scaling of which is directly proportional to the magnitude of shift between the observed and template lines.

The *SAFE* statistic uses higher order Hermite-Gaussian functions to try and fit for any remaining variation beyond the bulk shift captured by the first-order Hermite-Gaussian. From this fit, *SAFE* infers the possibility that variation from chromospheric variation is present in the given spectrum. This method assumes spectral lines are approximately Gaussian in shape and all spectra are appropriately normalized, but carries no requirements in terms of cadence of observations or range of activity states sampled. The *SAFE* statistic for each observation can be used as an activity indicator. The *HGRV+SAFE* results presented here come from co-modeling the provided CBC RVs and derived *SAFE* statistics with *GLOM*. More information can be found about the *HGRV* and *SAFE* methods in Holzer et al. (2020) and Holzer et al. (2021) respectively as well as in Section 5.11.2. Results for this method are forthcoming.

The Zeeman Least-Squares Deconvolution (*ZLSD*) method, submitted by the LienhardMortier team, incorporates the effects of magnetic line broadening when modeling each spectrum. Accounting for this effect when deriving RVs mitigates the chances that it will lead to a spurious RV measurement. This method also produces a measure of the magnetic activity on the star at the time of observation, which has been shown to be a powerful activity indicator (Haywood et al., 2020). A stable LSF and high resolution allows more subtle line broadening effects to become detectable. The *ZLSD* method is still being developed; preliminary results are included here. More information can be found in Section 5.11.3. Results for this method are forthcoming.

The *ResRegGen* and *ResRegDis* methods, both submitted by the CCA team, uses the residuals of observed spectra to 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 Hα 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 not as 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 over-fitting 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*, eighths of the data are 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 *ResRegGen* and *ResRegDis* methods are still being developed; preliminary results are included here. More information about *ResRegGen* and *ResRegDis* can be found in Sections 5.11.4 and 5.11.5 respectively.

5.4 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. 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 Section 5.12 along with their periodograms.

5.4.1 RV RMS of Method Results

Table 5.6 gives the change in overall and nightly RMS for each method as compared to the RMS of the provided, uncorrected CBC RVs provided. The nightly RMS, or intra-night scatter (INS), represents the average scatter over all nights with more than one observation. Positive ΔRMS values indicate that the method returned a lower RMS than the original. Negative ΔRMS values means 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 (§5.3).

Table 5.6: Δ RMS of Cleaned RVs for each Method in units of m s^{-1}

	HD 101501		HD 26965		HD 10700		HD 34411	
	RMS_N	RMS_{all}	RMS_N	RMS_{all}	RMS_N	RMS_{all}	RMS_N	RMS_{all}
Original *EXPRES* Data	0.62	4.887	0.65	3.195	1.071	1.864	0.944	1.78
Method	ΔRMS_N	ΔRMS_{all}	ΔRMS_N	ΔRMS_{all}	ΔRMS_N	ΔRMS_{all}	ΔRMS_N	ΔRMS_{all}
S-Value	0.02	0.26	0.021	0.526	0.195	0.186	-0.005	0.072
Hα Emission	-0.317	0.564	-0.051	0.209	0.005	0.003	0.0	0.0
Hα 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	-0.034	0.058
CCF FWHM	-0.005	0.534	0.001	0.022	0.002	0.001	-0.008	0.118
V$_{\text{span}}$	-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 + Hα	-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-*VALD*	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$_{\text{Spec.}}$	-0.182	2.36	-0.226	0.85	-0.027	-0.098	0.088	0.261

Continued on next page...

Table 5.6 – continued from previous page

Method	ΔRMS_N	ΔRMS_{all}	ΔRMS_N	ΔRMS_{all}	ΔRMS_N	ΔRMS_{all}	ΔRMS_N	ΔRMS_{all}
LBL+PCA$_{\mathrm{RV}}$	-0.421	2.46	-0.354	0.51	0.122	0.286	0.101	0.33
LBL+PCA$_{\mathrm{Spec./RV}}$	-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

The final RMSs of the clean RVs for all methods are plotted in Figure 5.4. 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

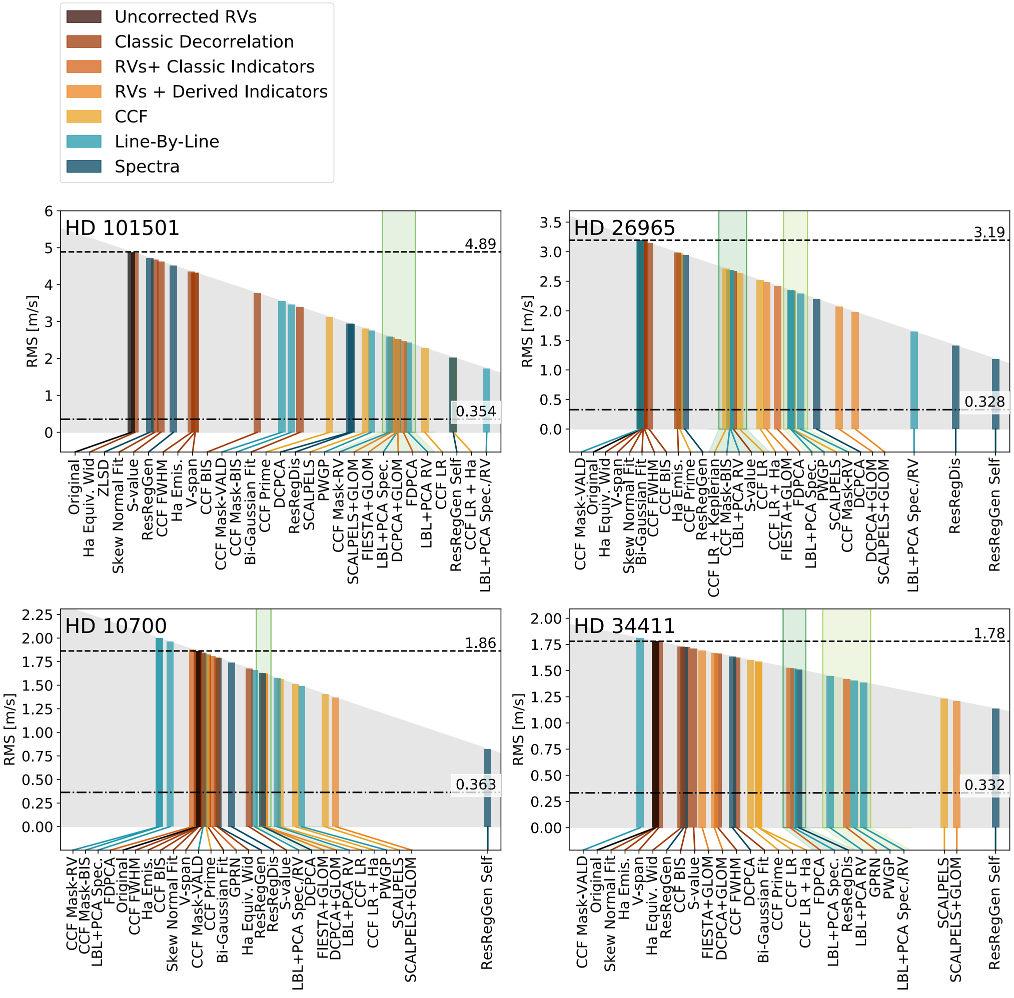

Figure 5.4: Final overall RMS of the clean RVs submitted for each of the four targets. The height and x-position of each bar scales with the final RMS. Bars are colored by the type of input data used. For each target, the RMS of the original, uncorrected *EXPRES* CBC RVs is shown as a black bar with 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.

to the break down of methods in Section 5.3 above. 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 historical method of decorrelating RVs against classic activity indicators, shown in Figure 5.4 as dark-orange bars, does not produce a significant decrease in RMS. The decrease is modestly significant for HD 101501, the most active of the targets given.

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 5.4. 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). 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.

The self test version of *ResRegGen*, *ResRegGen* Self, always returns a lower RMS than the cross-validation implememtation of *ResRegGen*. Furthermore, *ResRegGen* self is often among the methods returning the lowest RMS value. Given the difference between the resultant RMS for the self and cross-validation versions of *ResRegGen*, this is more a statement of the difference between these two frameworks rather than the *ResRegGen* method itself.

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, despite there being three methods that model the CCF with a change of basis (i.e. *SCALPELS*, *CCF Prime*, and *FIESTA*) the resultant RVs from these methods return very different RMS values. The same is true of the different line-by-line methods (shown as light blue bars in Figure 5.4). While some of the LBL methods do return similar RV RMS values, there is a large spread in the values returned.

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. Such a statement cannot be inferred from the RMS alone.

5.4.2 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. This being real data, we do not exactly know what the expected stellar signals should be for each target. As the data are consistent among all methods, though, we would expect the stellar signal being removed to be consistent between methods successfully modeling photospheric velocities. The activity RVs for each method should therefore 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, we use the Pearson correlation coefficient (PCC) to gauge the strength of correlation between the activity RVs derived by two different methods.

Figure 5.5 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 5.5 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 5.4.

As expected, most PCCs are positive, but there is little significant 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 chormospherically 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 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 pro-

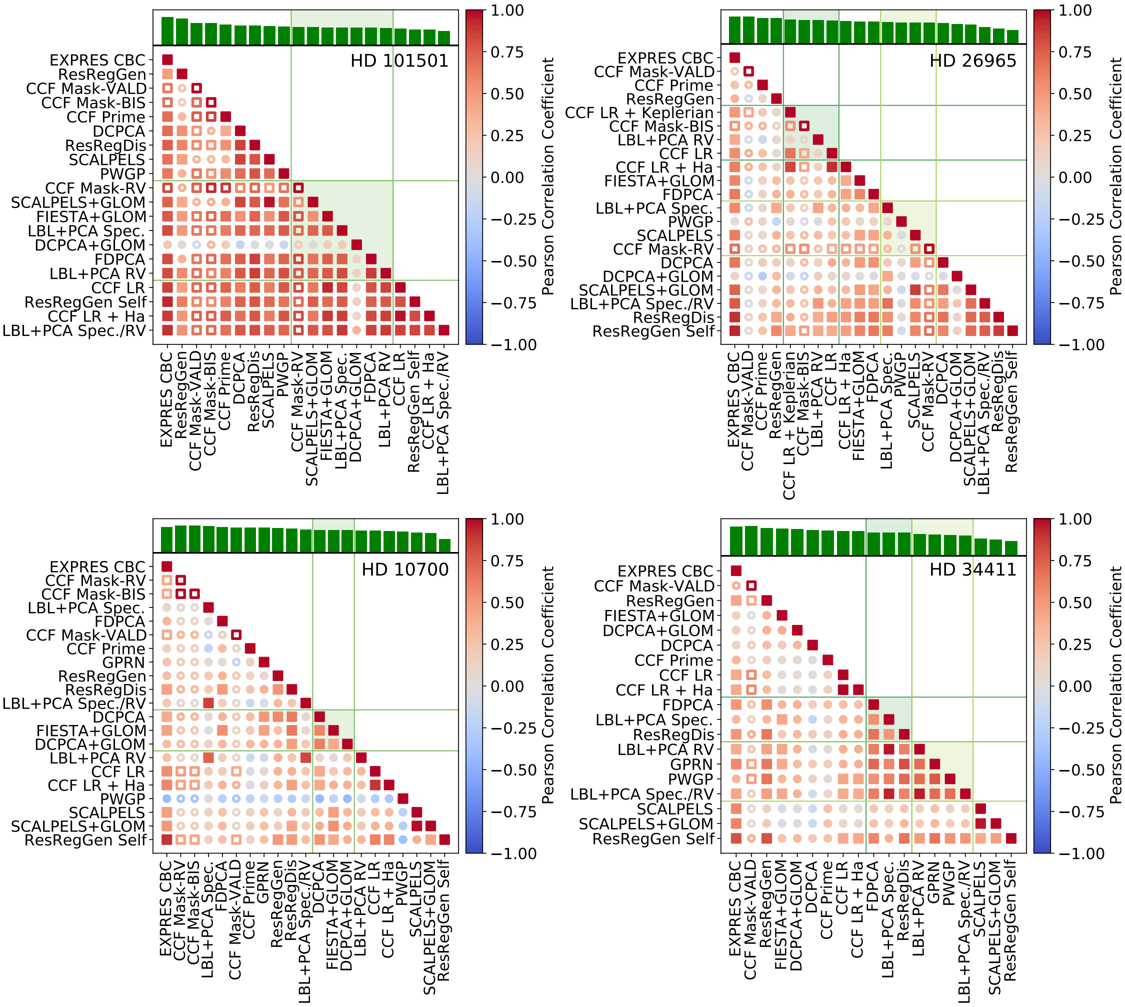

Figure 5.5: Pairwise comparisons between the activity RVs of all submitted methods. Each marker represents a pairing; the color of the marker gives the PCC between the activity RVs of two methods. Methods that did not submit activity RVs (for which the difference between the original and clean RVs was 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 5.4. At the top of each subplot in green is a scaled bar-graph of the resultant RMS for each method. As mirrored from Figure 5.4, methods that returned similar final RMS values are shaded in green along with their associated correlation coefficients.

vided *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 + Hα, 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$_{\mathrm{Spec.}}$, *LBL*+PCA$_{\mathrm{RV}}$, and *LBL*+PCA$_{\mathrm{Spec./RV}}$. 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 often not even correlated with the results of the *DCPCA* method despite both being informed by the same indicator. The *DCPCA+GLOM* activity RVs in fact rarely exhibit significant correlation with the results of any other method across all four targets.

5.4.3 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 monotoncity of two data sets rather their linearity.

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 is encouraging.

Figure 5.6 shows markers similar to those in the correlation matrices of Figure 5.5, but here denoting SCC values for indicator/method pairs. The activity RVs from each method are compared to classic activity indicators S-value, Hα 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 5.5).

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 ($\mathrm{SCC_C}$ and $\mathrm{SCC_A}$ respectively) are shown connected by a line for each indicator/method pair. The marker associated with the $\mathrm{SCC_C}$ is on the left while the $\mathrm{SCC_A}$ 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 5.6. For example, the *ResRegGen*

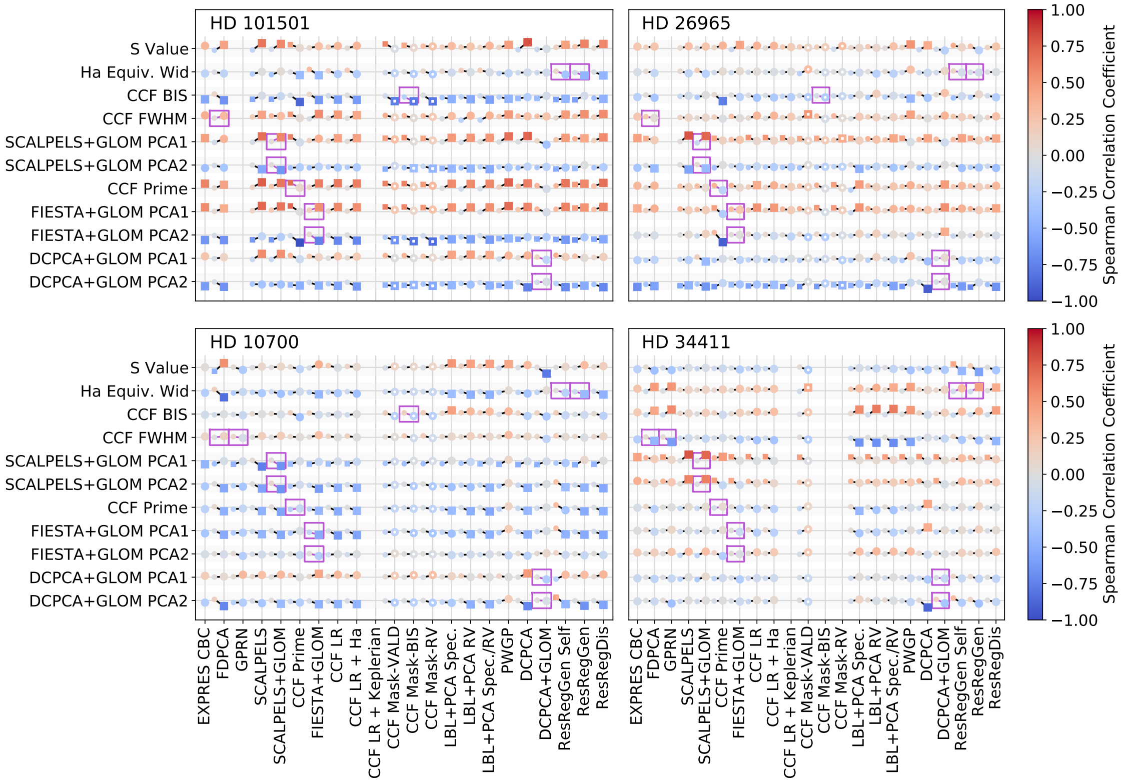

Figure 5.6: Correlation strength between activity RVs of submitted methods and activity indicators both classic (top four rows) and derived (bottom seven rows). Similar to Figure 5.5, here the color of each marker denotes the Spearman's rank correlation coefficient (SCC) with significant values (i.e. $|SCC| > 0.4$) 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 ($\mathrm{SCC_C}$) while the larger, right marker gives the SCC with the activity RVs ($\mathrm{SCC_A}$). 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.

results shown here use the Hα equivalent width as a label, and so the Hα 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 do exhibit significant correlation with indicator time series by and large do not also return significantly correlated clean RVs, lending confidence to the measured coorelation with the activity RVs. 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 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 5.5. Neither is true for the *DCPCA+GLOM* results. The results of *ResRegGen* are correlated with Hα equivalent width for all targets except HD 26965 while the results of *ResRegGen* self are not correlated with Hα 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.

5.4.4 HD 26965 Results

One of the benefits 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 (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 5.7 in blue and orange respectively. We also include periodograms of the provided *EXPRES* RVs and all RVs from the California Legacy Survey (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 (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.

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 five out of the six lowest RMS values for their clean RVs. As we saw with the lack of correlation between activity RVs from different methods in Figure 5.5, here we see again that the different methods do not agree on what the stellar signal is.

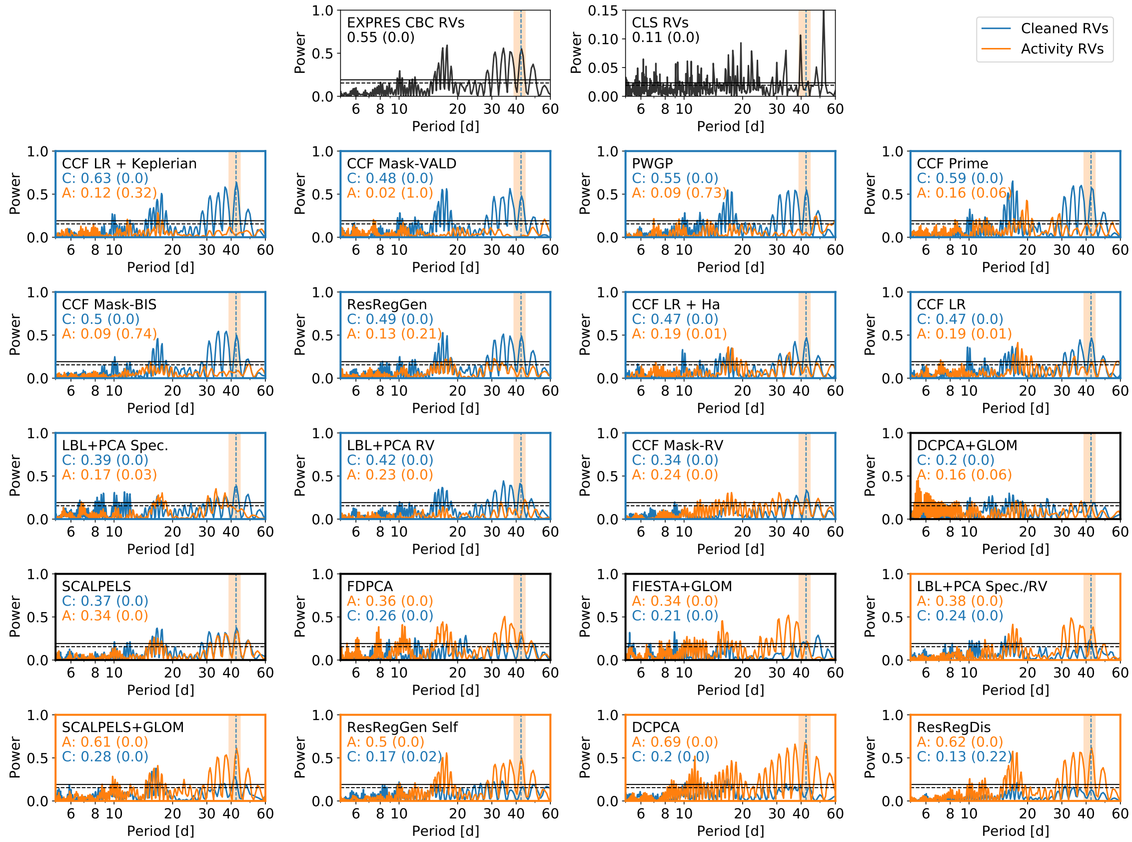

Figure 5.7: Periodograms of clean and activity RVs for HD 26965. The top row shows the periodogram for the provided *EXPRES* CBC RVs (left) and over 30 years of RVs from the California Legacy Survey (CLS) on the right (Rosenthal et al., 2021). Periodograms of the clean (blue) and activity (orange) RVs are given for all twenty of the methods that submitted *ESSP* results for HD 26965. Horizontal dashed and solid blank 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 signal with the clean RVs have blue axes, and orange axes for a stronger periodicity in the activity RVs.

5.5 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-three 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.

5.5.1 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 5.7 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.

Table 5.7: Method Philosophies

Method	Metric	Mitigation	Separation
GLOM		Multi-Dimensional GP Modeling	
FDPCA		Commonalities in Fourier Space	
GPRN		GP Neural Net Modeling	
SCALPELS	PCA Amplitudes (CCF)		Shape/Shift-Driven RVs
CCF Prime	GP Model Coefficients		Shape/Shift-Driven RVs
FIESTA+GLOM	Fourier Model Coefficients		
CCF Linear Regression			Shape/Shift-Driven RVs
CCF Masks			Variable/Stable Lines
LBL+$\text{PCA}_{\text{Spec.}}$			Variable/Stable Lines
LBL+PCA_{RV}	PCA Amplitudes (*LBL* RVs)		
PWGP			Variable/Stable Lines
DCPCA	PCA Amplitudes (Spectra)		
HGRV+SAFE	Likelihood of Shape Changes		
ZLSD	Magnetic Activity Strength		
ResRegGen		Regression w/ Spectral Residuals	
ResRegDis		Regression w/ Spectral Residuals	

5.5.2 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 m s^{-1}; no method[2] was able to reduce the RV RMS to less than 1.2 m s^{-1}.

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 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 exhibit a concerning lack of agreement 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. We found that even methods resulting in similar RV RMS values do not return correlated activity RVs.

This lack of agreement 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,

---

[2] 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.

as is fails to establish the nature of what a method is capturing.

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 ~40 day period as due to photospheric velocities while others attribute that periodicity to true center-of-mass shifts.

For some methods, the derived activity RVs do exhibit a significant correlation with activity indicators. While this is an encouraging result that suggests those methods are in fact disentangling photospheric velocities, it is irresponsible to expect all activity RVs to correlate with imperfect indicators. 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 or an orbiting planet.

5.6 Discussion

An increasing number of EPRV instruments are already online returning sub-meter-per-second single-measurement precision (e.g. Pepe et al., 2013; Jurgenson et al., 2016; Petersburg et al., 2020; Suárez Mascareño et al., 2020; Carmona et al., 2018; Seifahrt et al., 2018; Gilbert et al., 2018; Schwab et al., 2016a) with many more optical and infrared spectrographs being commissioned, built, or planned (e.g. Thompson et al., 2016; Bouchy et al., 2017; Gibson et al., 2018; Szentgyorgyi et al., 2014). 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 achieve that with the data sets provided.

Though there is no one single method clearly performing the best or any sub-meter-per-second RMS result, 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.

5.6.1 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. It is interesting to note that none of the submitted methods made use of the provided photometry.

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.

The effect of stellar signals is, in all cases, tied to either a classic activity indicator, shape variations, or 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 use PCA assume that stellar signals are the primary source of variation. 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 more advanced modeling (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 template itself should therefore not carry any significant variations and would ideally be representative of a quiet CCF/spectrum. Only then can the template be used as a reference to isolate the effects of stellar signals in each individual CCF. Constructing a mean or median template CCF/spectrum and using it in this way 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, many methods assume the line shape and CCF shape is well-described by a Gaussian. 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 variations from a Gaussian fit 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 GP to model a time series while the *GPRN* model uses GPs to define a neural net framework. *CCF Prime* forms a basis out of the derivatives of a GP model. In each case, 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*+$\text{PCA}_{\text{Spec.}}$) make use of the principal components themselves to denoise spectra or model the RV shift tied to the variation being modeled by the PCA.

Unless otherwise specified, methods assume true center-of-mass shifts have been removed or that they can be removed. Most methods make use of the derived RVs to align CCFs/spectra, even while we know these measured RVs are swayed by stellar signals. It may be worth looking into whether iterating with clean RVs produced by methods given different results, provided we are confident the corrections are truly removing only stellar signals.

Methods also 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.

5.6.2 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. It is necessary to understand what exactly methods are tracing in order to develop them further and be confident 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 know we can definitively remove all planet shifts. 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, too exclusively. Stellar signals and their spectral manifestations may 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 5.7). Rather than trying to find one, “best" 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 5.7 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 photopshere, 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 (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 insuring the signal being modeled with a method arises from the star. Incorporating photometric information into more methods will help with method interpretability by tying the modeled RV signals to a separate measure of activity. High-precision, simultaneous photometry should be collected for more stars to this end.

5.6.3 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; Cretignier et al., 2020b; Zhao et al., 2021; Cretignier et al., 2020b). It is worth considering the affect 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, no methods or indicators are directly sensitive to granulation or super-granulation. Granulation persists on the time scale of minutes while super-granulation has a time-scale of hours to days. We know from MHD simulations that the signatures from granulation are present in the spectra, but we remain insensitive to this effect with our current indicators and data.

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.

5.7 Conclusions

Twenty-three 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 decorrelation methods in nearly all cases. Though *EXPRES* data of quiet stars regularly return RMS values of 0.5-0.8 m s^{-1}, no method is yet reducing the RMS of more chromospherically active stars to sub-meter-per-second levels. Lack of agreement between the signal being modeled out by different methods makes it difficult to determine whether models are truly picking up on stellar signals and stellar signals alone.

In order to better mitigate the effects of stellar signals, we must demand greater interpretability of methods. Progress can only be made if we are confident about what signals a method is modeling out. This will also ensure that other signals, such as those from planets, will be preserved.

Method results should be compared for well-characterized data sets with known planetary signals, for example solar data or data with injected Keplerian shifts. Continuing to use consistent data sets will be paramount for understanding the differences between methods, though there is much room for improving the tools used to compare results.

To address all types of stellar signals, we must collect data that are sensitive to all the different sources of stellar variation and activity. In particular, oscillations and (super-)granulation remain poorly characterized; a high-resolution data set of densely sampled observations over four to five nights is needed to understand this signal and make it possible to correct for it. Armed with a high quality data set, the *ESSP* can also be repeated for degraded cadence, resolution, SNR, etc. as an empirical test of the data quality needed to achieve EPRV measurements that are clean of stellar signals.

Currently, there is progress being made in mitigating stellar signals but much work remains to be done. Stable, high-resolution, high-cadence data must be a priority in order to probe all types of stellar signals. 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.

5.8 In-Depth Descriptions of Methods That Use RVs and Classic Activity Indicators as Input

5.8.1 *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., 2020). 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:

$$\begin{aligned} q_0(t) &= m_0(t) + a_{0,0}G(t) + a_{0,1}\dot{G}(t) + a_{0,2}\ddot{G}(t) + \epsilon_0(t) \\ q_1(t) &= m_1(t) + a_{1,0}G(t) + a_{1,1}\dot{G}(t) + a_{1,2}\ddot{G}(t) + \epsilon_1(t) \\ &\vdots \\ q_N(t) &= m_N(t) + a_{N,0}G(t) + a_{N,1}\dot{G}(t) + a_{N,2}\ddot{G}(t) + \epsilon_N(t) \end{aligned} \tag{5.1}$$

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 $n = 1, ..., N$ 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.

5.8.2 *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:

$$\begin{bmatrix} \Re(\mathcal{F}\{\mathrm{H}\alpha\mathrm{EW}\}) & \Im(\mathcal{F}\{\mathrm{H}\alpha\mathrm{EW}\}) & \Re(\mathcal{F}\{\mathrm{CCF\ FWHM}\}) & \Im(\mathcal{F}\{\mathrm{CCF\ FWHM}\}) & \cdots \end{bmatrix} \tag{5.2}$$

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. [3]

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.

---

[3] *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.

5.8.3 *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* is similar to that of 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 $\mathbf{y}$(x) as

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\mathbf{f}(\mathbf{x}) + \sigma_y \mathbf{z}(x). \tag{5.3}$$

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

$$\begin{aligned} &f_j(x) \sim \mathcal{GP}(0, k_f) \text{ for } j = 1, ..., q, \\ &W_{ij}(x) \sim \mathcal{GP}(0, k_w) \text{ for } i = 1, ..., p \text{ and } j = 1, ..., q, \end{aligned} \tag{5.4}$$

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 neural 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.

5.9 In-Depth Descriptions of Methods That Use the CCF as Input

5.9.1 *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.

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.

5.9.2 *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

$$c_i(v) = a_i + \Sigma_{k=0}^{3} b_{ik} C^{(k)}(v) \tag{5.5}$$

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 5.5 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 5.9.1. 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.

5.9.3 *FIESTA*

The *FIESTA* method, submitted by the PennState team, decomposes the CCF of a spectrum into Fourier basis functions (Zhao & Tinney, 2020). 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. in prep) and modeled jointly with dynamical RV shifts using GLOM (as described in Section 5.8.1).

5.9.4 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 assocaited RV is given by:

$$RV = w_1 \cdot CCF_1 + w_2 \cdot CCF_2 + \cdots + w_v \cdot CCF_v \tag{5.6}$$

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, Hα information was added to the model to give:

$$RV = w_1 \cdot CCF_1 + w_2 \cdot CCF_2 + \cdots + w_v \cdot CCF_v + b \cdot H\alpha \tag{5.7}$$

where Hα is the derived Hα 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:

$$RV = w_1 \cdot CCF_1 + w_2 \cdot CCF_2 + \cdots + w_v \cdot CCF_v + b \cdot H\alpha + d \cdot Keplerian \tag{5.8}$$

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.

5.10 In-Depth Descriptions of Line-by-Line Methods

5.10.1 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 withing 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.

5.10.2 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 in active if they had a coefficient less than 0.2-0.4 and spread in RVs less than 10-15 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.

5.10.3 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 the decomposing the resultant LBL RVs with another PCA. These results are included as *SHELL+YARARA* LBL RVs.

5.10.4 *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 line (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.

5.11 In-Depth Descriptions of Methods That Model the Spectra

5.11.1 *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 5.10.1 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.

5.11.2 *HGRV+ SAFE*

The Hermite-Gaussian based Radial Velocity method, submitted by the YaleWI team, estimates the RV shift of a stellar spectrum using Hermite-Gaussian functions. For each line, the difference between a template spectrum and a shifted spectrum can be modeled by scaling a first-degree Hermite-Gaussian function under the assumption that all spectral lines have shapes that are approximately Gaussian density-shaped. The scaling coefficient directly maps to the magnitude of the shift.

The Stellar Activity F-Statistic for Exoplanet surveys statistic uses several, higher order Hermite-Gaussian functions (excluding the first-degree Hermite-Gaussian function) to test for variability in the line shapes that may be attributable to stellar signals. *SAFE* gives an F-statistic that can be used in a hypothesis-testing framework with the null hypothesis that all the coefficients of the Hermite-Gaussian terms are zero, which would suggest that there is not significant line shape variations from photospheric velocities present. The higher the *SAFE* statistic value, the less likely that the null hypothesis—that there are no changes due to stellar signals present—is correct, meaning it is therefore more likely that the effects of stellar signals are present.

Implementing the *HGRV* method and deriving the *SAFE* only requires a set of continuum normalized spectra. *SAFE* is sensitive to poor normalization, so it is important that the normalization is carried out well. Since the method does not incorporate time information, it is agnostic to the cadence and time coverage of observations. More observations with higher SNR help to produce a better estimate of a template spectrum, which is used in the method.

The results presented here represent RVs derived using the *HGRV* method and cleaned of contamination from stellar signals using *GLOM* on the principal component amplitudes from applying PCA to the Hermite-Gaussian coefficients. Note, that in the use of *GLOM*, the method does become sensitive to the cadence of observations. The resultant Hermite-

Gaussian coefficients from deriving *SAFE* are used as an activity indicator to constrain a GP model of the chromospheric contribution present in each RV measurement.

5.11.3 *ZLSD*

The Zeeman Least-Squares Deconvolution method, submitted by the LienhardMortier team, uses a mask that takes into account the magnetic sensitivity of each line as established by the line's Landé factor. By taking this Zeeman broadening into account, *ZLSD* provides an RV measurement that is less sensitive to magnetic variability and also produces an estimate of the magnetic field and filling factor of the star at the time of the observation.

The method takes spectra as input. In order to prepare the spectra for the *ZLSD* algorithm, the blaze, continuum, telluric, and read-noise information is also needed. Though the total number of observations does not really matter, the observations should be of high resolution, high precision, and preferably redder. Higher resolution spectra will allow more subtle line broadening changes to become detectable. Stabilised data will prevent variations in the instrumental profile that could be mistaken for Zeeman broadening. As Zeeman broadening is dependent on wavelength, the effect becomes more pronounced and therefore more measurable towards the red/infrared.

The resultant RV measurements from this modeling will be less sensitive to the magnetic variability that accompanies stellar signals. The estimate of the magnetic field/filling factor of the star can also be used as an activity indicator. Hemispherically-averaged line-of-sight magnetic field has been shown with the Sun to be the best activity indicator known to date (Haywood et al., 2020).

The *ZLSD* method is still under development and improvements are continuously being made. The results cited here represent the performance of an early experimental version which did not capitalize on the full flexibility of the approach. The non-magnetic basis of the *ZLSD* pipeline is presented in Lienhard et al. (2021, in prep.).

5.11.4 *ResRegGen*

ResRegGen, submitted by the CCA team, takes the residuals of each observed spectrum against a 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:

$$\hat{q}_n = \Delta f_n \frac{dF}{dQ} \cdot \left[\frac{dF}{dQ} \cdot \frac{dF}{dQ} \right]^{-1} \tag{5.9}$$

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 straightforward 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 Hα 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*.

5.11.5 *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 a labels. The predicted RV correction for each observation, $\hat{q}_n$, can then be calculated

$$\hat{q}_n = \Delta f_n \cdot \left(F^T F + \alpha I\right)^{-1} F^T Q \tag{5.10}$$

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.

5.12 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-right 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 subsequent column 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.

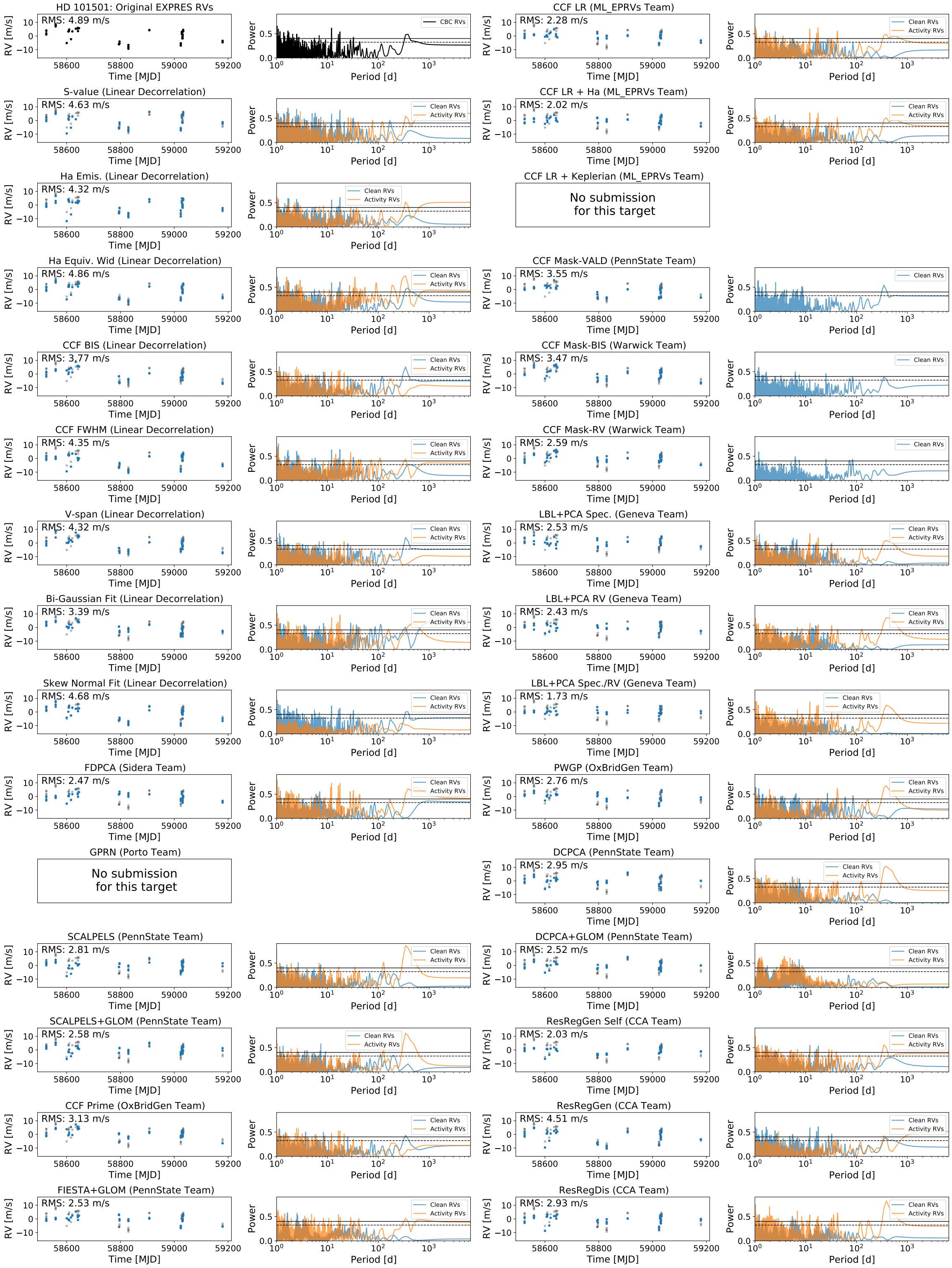

Figure 5.8: Submitted results for HD 101501. For each periodogram, p-values of 0.01 and 0.01 are shown as horizontal solid and dashed black lines respectively.

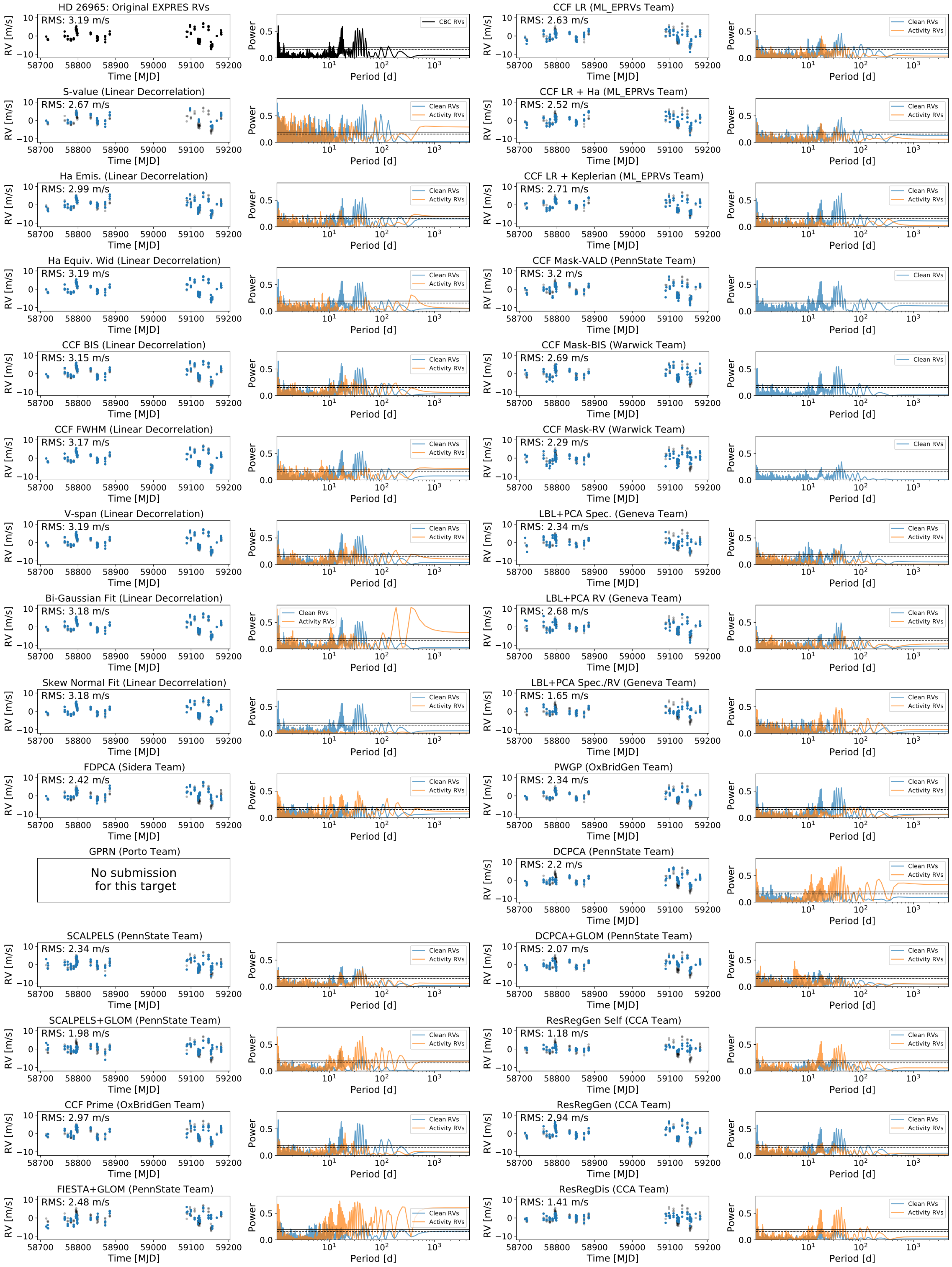

Figure 5.9: Submitted results for HD 26965.

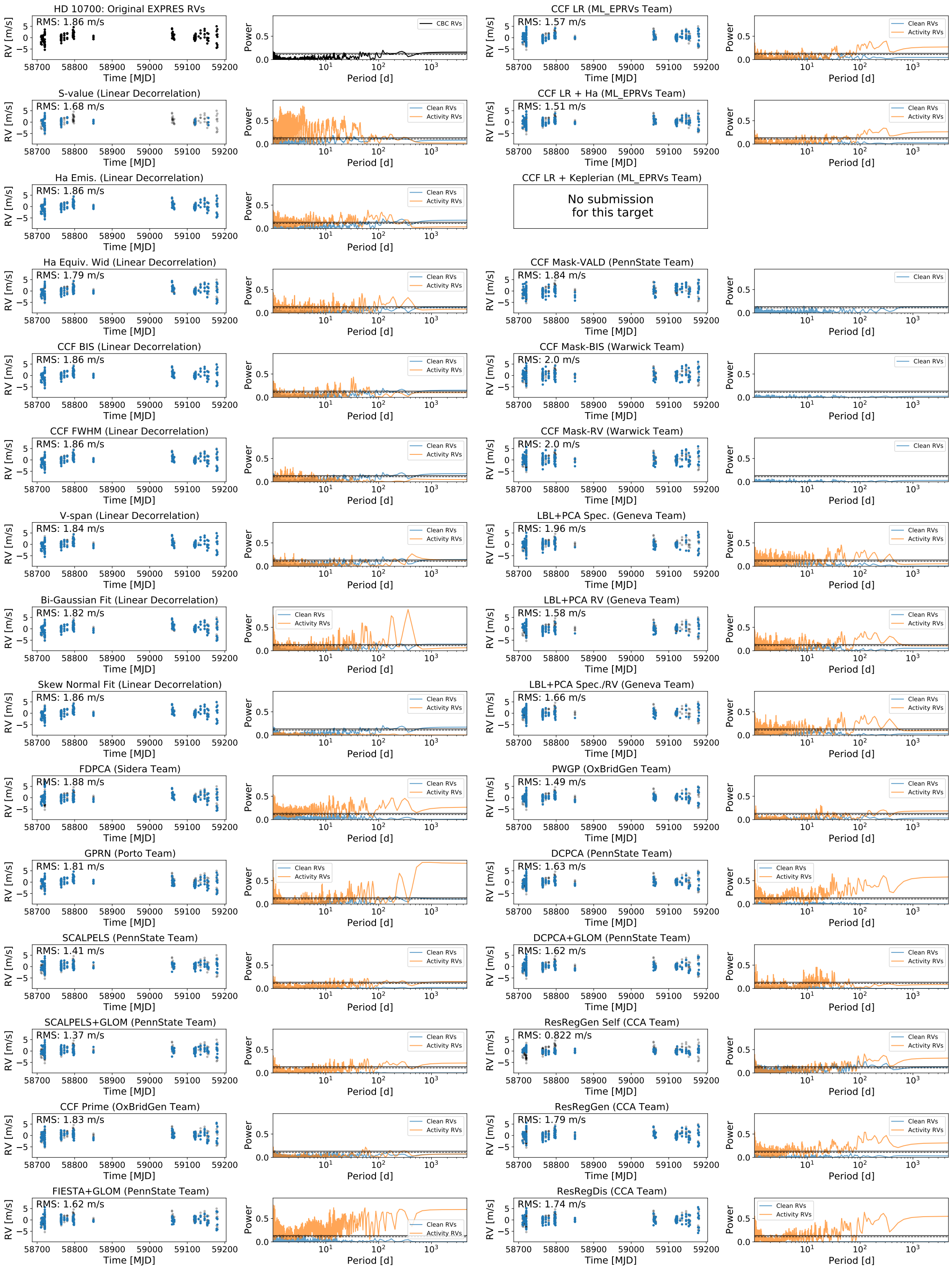

Figure 5.10: Submitted results for HD 10700.

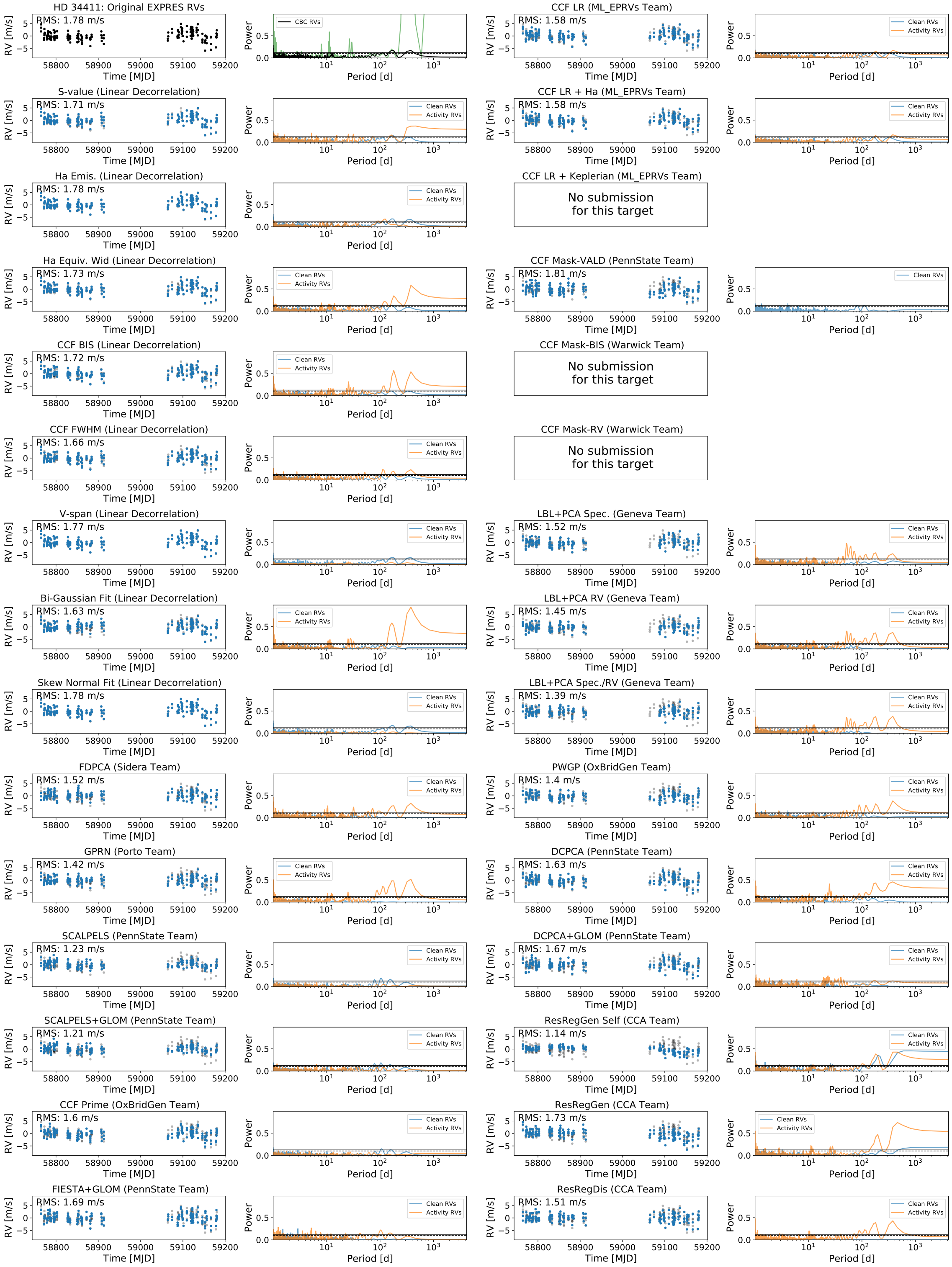

Figure 5.11: Submitted results for HD 34411.

5.13 Acknowledgements

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. 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, the Heising-Simons Foundation, and an anonymous donor in the Yale alumni community. This work was supported by a grant from the Simons Foundation (675601 and EBF). G. W. H. acknowledges long-term support from NASA, NSF, Tennessee State University, and the State of Tennessee through its Centers of Excellence program. LLZ gratefully acknowledges support from the NSF GRFP under Grant No. DGE1122492. RMR acknowledges support from the Yale Center for Astronomy & Astrophysics (YCAA) Prize Postdoctoral Fellowship.

5.13.1 Facilities

LDT, TSU:AST

Chapter 6

Conclusion

This dissertation has demonstrated the need for higher-precision radial-velocity measurements and explores both the data analysis techniques and spectral modeling needed to achieve this standard. Extremely precise RVs will detect a population of lower-mass planets that we know exist from population statistics and begin to characterize these planets by providing mass estimates. EPRV measurements will also inform the dynamics of multi-planet systems and star-planet alignment, which will help guide planet formation and migration theories.

Sub-meter-per-second RV precision will enable a deeper understanding of exoplanet composition and histories, but will require carefully stabilized instrumentation, high-fidelity data pipelines, and innovative methods for disentangling photospheric velocities from the true signals of planets.

6.1 Summary

Previous RV standards with measurement precision on the order of 0.8-1 m s^{-1} are not sufficient for finding low-mass, Earth-like planets (Zhao et al., 2018). Chapter 2 analyzed over a decade of RV measurements from *HARPS*, *UVES*, and *CHIRON* taken of α CenA, B, and Proxima Centauri, the three closest stars to our Sun. Even with years of data of these bright, well characterized stars, we found that there could persist undiscovered planets in the habitable zone with $M\sin(i)$ up to 53 $M_{\oplus}$for α CenA, 8.4 $M_{\oplus}$for α CenB, and 0.47 $M_{\oplus}$for Proxima Centauri. Simply adding more data of the same quality will not significantly push down detection limits; we need more stable, better understood data.

Next-generation spectrographs, like *EXPRES*, have opened up a new regime of hardware stability and resolution. To remain sensitive to the small signals targeted by EPRV work, the raw data from these highly-stabilized instruments demand high-fidelity extraction pipelines. These pipelines must not add any extraneous variations themselves, and additionally should be used as an opportunity to account for any variations in the raw data (Petersburg et al., 2020).

Excalibur's hierarchical framework gives an example of using the data itself to learn about variations in the instrument. *Excalibur* utilizes all calibration images to construct a model of the accessible calibration space of an instrument and de-noise line positions (Zhao et al., 2021). With *EXPRES* data, *excalibur* reduced the overall RMS of RV data sets for all targets tested by 0.2-0.5 m s^{-1}. Implementing data-driven methods for extracting data produces more stable spectra and ultimately more precise RVs.

The resultant spectra must be carefully modeled in order to return true, center-of-mass velocities due to orbiting planets that are not swayed by the effects of photospheric velocities from stellar variability and activity. Many methods have been developed throughout the years to disentangle true shifts from these stellar signals. The *EXPRES* Stellar Signals Project (*ESSP*) provides the first direct comparison of these different methods on a consistent set of real data representative of next-generation instrumentation. This analysis is ongoing work.

Nineteen different methods were submitted that implemented a variety of techniques. The presence of stellar signals or variation in the spectra was gauged using classic indicators, spread in returned RVs or fit parameters, principal component analysis, etc. Different methods also had their own ways of removing the effects of these stellar signals, such as removing identified sources of noise at the cause of greater bias, co-modeling activity metrics and RVs using Gaussian processes, down weighting areas giving more scatter, etc.

Method performance was evaluated based on final and nightly RMS as well as agreement between methods. There was no one clear method or approach that obviously and consistently performed the best. Methods most consistently returning the lowest RV RMS values tended to use the full spectra as input and/or flexible statistical models such as Gaussian processes or principal component analysis to pick up on variations. All tested methods generally performed better than the classic linear decorrelation method. The lowest final RV RMS returned by any method was 1.2 m s^{-1} on the quietest target. Progress is being made in the field, but more work, greater interpretability of methods, and high cadence data is needed to reduce RV RMS down-to sub-meter-per-second levels.

6.2 Discussion

The need for and promise of EPRV measurements has spurred the genesis of many highly stabilized instruments with the goal of 30-50 cm s^{-1} RV precision. This new generation of instruments will be able to produce a growing set of high-fidelity data as they begin collecting data over the time baselines needed to detect planets with RV and probe a range of activity states of target stars.

Many of these new instruments are accompanied by solar telescopes that will collect high SNR, densely sampled, daily data of the Sun. This solar data will allow for more direct comparisons between instruments and provide a good test set for disentangling solar photospheric velocities. The observations from every one of these different instruments will all have to be held up to a high standard of careful data analysis and modeling of stellar signals.

In the regime of EPRV measurements, data extraction must move towards a framework where the pipeline is utilized as an opportunity to account for even small instabilities in the raw data. All exposures from an instrument contain information about the systematic instrumental effects that change how light is projected onto the detector and read out into raw counts. *Excalibur* makes use of this information via a hierarchical framework. It uses the set of all calibration images to construct the space of achievable calibration states of an instrument, i.e. the complete space in which the instrument can and does vary. This analysis uses the data itself to learn about how changes to the hardware is affecting the resultant data. Implementing it on other, similarly calibrated EPRV instruments would be trivial and help reduce a source of red-noise that would otherwise permeate all exposures.

The changes in calibration state identified by *excalibur* can be traced back to hardware variations. For example, when implementing *excalibur* on *EXPRES* data, we used PCA to reduce the dimensionality of the calibration space. The amplitude of the first principal component closely traced the optical bench temperature. The shape of the components themselves also contain some vertical structure across the detector, which may suggest defects related to optical elements prior to the cross-dispersion prisms.

The hierarchical framework of *excalibur* was able to pick up on these changes, which were then corrected for through the flexibility of implementing a non-parametric pixel-to-wavelength model. Future work fully understanding these correlations with hardware properties will make the results of *excalibur* more interpretable. This could in turn help diagnose sources of variability in the instrument hardware.

Excalibur can be expanded to work on different instruments. Many next-generation spectrographs will make use of a simultaneous reference fiber (sim. fiber), which projects a set of calibration lines onto a different set of pixels during science exposures. Though not currently handled by *excalibur*, the idea can easily be adapted to incorporate simultaneous calibration information. The information from this sim. fiber can similarly be used to characterize the calibration state of the instrument and, in constructing the wavelength solution, used to pinpoint the spectrograph's current calibration state through forward modeling.

Preliminary results show that 2D information in the raw data, for example the vertical trace position of the separate echelle orders on the detector, similarly contains information about the calibration state of the spectrograph. Hardware variations affect the projection of light onto the detector in all dimensions, not just the dispersion direction that is typically assessed as a part of wavelength calibration. By characterizing the variations at the 2D level, we introduce a new source of calibration information. If this information can be derived simply from the raw data itself, this will effectively provide every spectrograph with a source of “simultaneous reference" for free.

Using the calibration data itself to characterize instrument variations could potentially carry use cases for any instrument, not just EPRV spectrographs. The hierarchical framework of *excalibur* is applicable to any instrument exhibiting only low-dimensional variation regardless of the amplitude of this variation. In fact, most instruments are expected to vary in a low-dimensional space given the limited degrees of freedom available to physical systems. By introducing statistically principled modeling techniques to data pipelines, extracting spectra from raw data becomes an opportunity to account for variations in the data and recover more confident measurements.

Well characterized, precise data is required if we hope to model the subtle line-shape variations that arise from photospheric velocities. The effects of these stellar variations are known to change in both character and amplitude for different spectral lines. They will be difficult to tease out if there are also instrumental or pipeline variations changing the spectra. Stable data is needed to model stellar signals. High-resolution data helps to better characterize line shape changes.

Apparent RV shifts from photospheric velocities are now the largest source of error in EPRV work. We have seen that the classic method of linearly decorrelating against a standard set of activity indicators is not effective in the regime of sub-meter-per-second percision. More advanced methods that have recently been developed are performing better, but still not reaching sub-meter-per-second precision. The *EXPRES* Stellar Signals Project (*ESSP*) made it possible to benchmark the current state of the field in disentangling stellar signals and make recommendations for next steps.

It is important to start rendering methods for disentangling stellar signals more interpretable. We found that principal component analysis and Gaussian process models are helpful for identifying the presence of activity and returning lower RMS, but these analysis methods are famously un-interpretable and prone to over fitting. Though the methods using these techniques performed well, we could only speculate as to why. This speculation was made all the more difficult by the lack of agreement between methods modeling the same data. In order to truly understand what is working and what is not, we must understand what exactly the different methods are modeling and how this is being corrected for.

One way to build interpretability is to use data sets where we believe we know the proverbial “answer at the back of the book," i.e. what stellar signals are expected and what magnitude of true, center-of-mass shifts. Next steps include simulating data with injected RV shifts, so we know what center-of-mass shifts we expect to find in the resultant cleaned RVs. This would provide a test of whether methods are over fitting and correcting for true shifts as well as stellar signals. True shifts with amplitudes near the RMS of the data set and periods near the stellar rotation rate or its harmonics will be the hardest to disentangle. The true shifts must be carefully simulated so as to not introduce any spurious changes or artifacts, for example shifting telluric lines along with stellar lines.

A more realistic test could use data of a well-characterized star system where we believe we know what stellar signals and true shifts to expect. Of course, this test will only be as useful as the star is well-characterized. The best characterized star will always remain our Sun. Most next-generation spectrographs will be accompanied by a solar telescope; the *EXPRES* solar telescope has already collected almost a year of solar data. The Sun is only just now coming out of a quiet phase. In the next few years, the community will be able to build up a repertoire of high-precision, solar spectra exhibiting stellar signals that will provide a good test bed for different methods.

We must be careful, however, to not become overly reliant on modeling Solar signals. We saw with the *ESSP* results that the greatest agreement between methods returning similar RV RMS values occurred in the case of HD 34411, a very quiet star, but the most Sun-like of the four targets. Solar data from *HARPS* or *SOAP* simulations, constructed of resolved *NSO* observations of the quiet Sun and a Sun spot, have long been the standard for testing methods that aim to disentangle stellar signals. It is necessary to build the ability to convincingly simulate or thoroughly characterize the stellar signals that arise from a range of stellar types to ensure that method performance is universal.

A better understanding of how to disentangle stellar signals will come from forming collaborations between different methods. Different methods used different metrics to gauge the presence of activity. A variety of techniques were then used to mitigate the perceived signal from stellar variability. These two steps were often independent of one another. Different metrics should be tried with different techniques to fully explore the parameter space of current methods correcting stellar signals.

We will learn more about how stellar variability is affecting spectra by comparing line-by-line method results with one another. Three separate teams explored different methods for identifying quiet vs. active lines with *EXPRES* data as part of the *ESSP*. It would be worthwhile to see for which lines these methods agreed and where they differed. Were there lines that all methods agreed are quiet or active? Is there are reason for why methods disagree where they do? The categorization of lines as quiet or active could potentially be linked to what we understand of the atomic transitions that produce each line. Lines may also behave differently according to their depth, which inversely correlates to the depth at which the line forms in the star's photosphere (Cretignier et al., 2020a).

Several methods used PCA to identify variations in the data, both in the CCF (*SCALPELS*) and in the spectra (*DCPCA*, *SHELL* LBL). It would be interesting to compare the resultant PCA decompositions from the different methods. Similar PCA components or amplitudes may suggest that the different methods are picking up on a similar source of variation. The

PCA decomposition could then be rendered interpretable if this variation can be traced back to a signal we would expect to see from stellar variation.

Currently, no method or indicator is directly sensitive to the variation contributed by stellar pulsations or granulation. The effects of these variations occur on minute to day-long time scales. Though MHD simulations have shown that the signatures of these processes should exist in spectra, the currently available observations are likely not sensitive to these changes. The short time scale and small spatial size of granulations requires very dense sampling over the course of several nights and high resolution to measure. The community should work towards collecting an ideal data set for detecting the effects of these oft-ignored contributions to stellar signals. Once we are able to characterize the effects of these variations, it will be easier to account for them.

We need to better understand the data requirements for disentangling stellar signals. Removing the contribution for photospheric velocities is essential to being able to measure true Doppler shifts from low-mass planets. It is only worthwhile to collect data of high enough quality for which it is possible to model out stellar signals. An initial test would be to take the *EXPRES* data provided for the *ESSP* and reduce the cadence/total number of observations, SNR, and resolution, etc. of the data set one at a time. Methods can be tested on the reduced data sets to gauge method performance on data of different quality. This will allow us to empirically figure out when methods break and give estimates of the type of data that must be collected if we hope to use it to model out stellar signals and open up the possibility of detecting planets.

6.3 Conclusion

Enabling EPRV measurements will shed light on low-mass planet compositions, inform planet formation, and constrain migration theories. Working towards this new regime of measurement precision has emphasized error sources that we did not have to care about before and revealed new sources of error that we did not know about. To measure a signal at the sub-meter-per-second level changes the acceptable standards for everything, from the stability and optical design of the instrument to how the resultant data is extracted and calibrated. Disentagling stellar signals at the level of tens of cm s^{-1} has become paramount.

Since the installation of *EXPRES* and *ESPRESSO* in 2017, more precision spectrographs have gone online and more are planned through at least 2022 that will aim to collect high-precision spectra in the optical/infrared at night and stare at the Sun during the day. *TESS* and *CHEOPS* continue to collect light curves revealing planet candidates that will need confirmation and mass measurements. Upcoming missions, like *JWST* and the Roman Space Telescope, as well as mission concepts, such as *LUVOIR* and *HabEx*, are building science campaigns that will necessitate RV characterization to determine good planet candidates for atmospheric characterization.

EPRV is still a relatively new field, but it is growing and it is needed. As with any new field, here at the beginning we are first learning of more problems, and the work lies in finding solutions within this completely new regime of measurement. The work presented in this thesis showcases only the beginning of the push towards EPRV measurements. Progress has been made via more advance data pipelines and an assessment of the state of the field on addressing stellar signals. Progress will continue to be made, ever pushing detection limits lower and revealing more about how planets exist.

Bibliography

Aigrain, S., Pont, F., & Zucker, S. 2012, MNRAS, 419, 3147, doi: `10.1111/j.1365-2966.2011.19960.x` 17, 185

Airapetian, V. S., Glocer, A., Khazanov, G. V., et al. 2017, ApJL, 836, L3, doi: `10.3847/2041-8213/836/1/L3` 34

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` 147

Aminikhanghahi, S., & Cook, D. J. 2017, Knowledge and Information Systems, 51, 339, doi: `10.1007/s10115-016-0987-z` 128

Anderson, E., & Francis, C. 2011, VizieR Online Data Catalog, 5137 26, 31

Anglada-Escudé, G., & Butler, R. P. 2012, ApJS, 200, 15, doi: `10.1088/0067-0049/200/2/15` 82

Anglada-Escudé, G., Amado, P. J., Barnes, J., et al. 2016, Nature, 536, 437, doi: `10.1038/nature19106` 22, 29, 30, 34, 36, 45

Angus, R., Morton, T., Aigrain, S., Foreman-Mackey, D., & Rajpaul, V. 2018, MNRAS, 474, 2094, doi: `10.1093/mnras/stx2109` 138

Aquilecchia, G. 2021, Giordano Bruno. `https://www.britannica.com/biography/Giordano-Bruno` 1

Arentoft, T., Kjeldsen, H., Bedding, T. R., et al. 2008, ApJ, 687, 1180, doi: `10.1086/592040` 15, 136

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, doi: `10.1051/0004-6361/201322068` 96, 133

Ayres, T. R. 2014, AJ, 147, 59, doi: `10.1088/0004-6256/147/3/59` 28

—. 2015, AJ, 149, 58, doi: `10.1088/0004-6256/149/2/58` 28

Baliunas, S., Sokoloff, D., & Soon, W. 1996, ApJL, 457, L99, doi: `10.1086/309891` 140

Baraffe, I., Homeier, D., Allard, F., & Chabrier, G. 2015, A&A, 577, A42, doi: `10.1051/0004-6361/201425481` 27

Baranne, A., Mayor, M., & Poncet, J. L. 1979, Vistas in Astronomy, 23, 279, doi: `10.1016/0083-6656(79)90016-3` 80

Barbieri, M., Marzari, F., & Scholl, H. 2002, A&A, 396, 219, doi: `10.1051/0004-6361:20021357` 34

Barragán, O., Gandolfi, D., & Antoniciello, G. 2019, MNRAS, 482, 1017, doi: `10.1093/mnras/sty2472` 17, 152

Batalha, N. M. 2014, Proceedings of the National Academy of Science, 111, 12647, doi: `10.1073/pnas.1304196111` 5

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, ApJS, 204, 24, doi: `10.1088/0067-0049/204/2/24` 21

Bauer, F. F., Zechmeister, M., & Reiners, A. 2015, A&A, 581, A117, doi: `10.1051/0004-6361/201526462` 103

Bedell, M., Hogg, D. W., Foreman-Mackey, D., Montet, B. T., & Luger, R. 2019, AJ, 158, 164, doi: `10.3847/1538-3881/ab40a7` 90, 95, 162, 208

Bergmann, C., Endl, M., Hearnshaw, J. B., Wittenmyer, R. A., & Wright, D. J. 2015, International Journal of Astrobiology, 14, 173, doi: `10.1017/S1473550414000317` 35, 37

Blackman, R. T., Ong, J. M. J., & Fischer, D. A. 2019, AJ, 158, 40, doi: `10.3847/1538-3881/ab24c3` 58, 61, 93

Blackman, R. T., Szymkowiak, A. E., Fischer, D. A., & Jurgenson, C. A. 2017, The Astrophysical Journal, 837, 18, doi: `10.3847/1538-4357/aa5ead` 78, 142

Blackman, R. T., Fischer, D. A., Jurgenson, C. A., et al. 2020, arXiv e-prints, arXiv:2003.08852. `https://arxiv.org/abs/2003.08852` 11, 57, 61, 64, 67, 76, 82, 87, 92, 98, 108, 135, 141

Boisse, I., Bouchy, F., Hébrard, G., et al. 2011, A&A, 528, A4, doi: `10.1051/0004-6361/201014354` 16, 136, 137, 145

Boisse, I., Moutou, C., Vidal-Madjar, A., et al. 2009, A&A, 495, 959, doi: `10.1051/0004-6361:200810648` 17, 137

Bolton, A. S., & Schlegel, D. J. 2010, PASP, 122, 248, doi: `10.1086/651008` 13, 94

Bonfils, X., Delfosse, X., Udry, S., et al. 2013, A&A, 549, A109, doi: `10.1051/0004-6361/201014704` 98

Borgniet, S., Meunier, N., & Lagrange, A.-M. 2015, A&A, 581, A133, doi: `10.1051/0004-6361/201425007` 47

Borucki, W. J., Koch, D. G., Basri, G., et al. 2011, ApJ, 736, 19, doi: `10.1088/0004-637X/736/1/19` 5, 21

Bouchy, F., Bazot, M., Santos, N. C., Vauclair, S., & Sosnowska, D. 2005, A&A, 440, 609, doi: `10.1051/0004-6361:20052697` 15, 136

Bouchy, F., Doyon, R., Artigau, É., et al. 2017, The Messenger, 169, 21, doi: `10.18727/0722-6691/5034` 7, 180

Boyajian, T. S., McAlister, H. A., van Belle, G., et al. 2012, ApJ, 746, 101, doi: `10.1088/0004-637X/746/1/101` 140

Boyajian, T. S., von Braun, K., van Belle, G., et al. 2013, ApJ, 771, 40, doi: `10.1088/0004-637X/771/1/40` 25, 26, 30

Brahm, R., Jordán, A., & Espinoza, N. 2017, PASP, 129, 034002, doi: `10.1088/1538-3873/aa5455` 81, 86

Brewer, J. M., & Fischer, D. A. 2016, ApJ, 831, 20, doi: `10.3847/0004-637X/831/1/20` 26, 140

Brewer, J. M., Fischer, D. A., Valenti, J. A., & Piskunov, N. 2016, ApJS, 225, 32, doi: `10.3847/0067-0049/225/2/32` 25, 115, 140, 149, 150

Brewer, J. M., Fischer, D. A., Blackman, R. T., et al. 2020, arXiv e-prints, arXiv:2006.02303. `https://arxiv.org/abs/2006.02303` 10, 98, 135, 142, 147

Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8, doi: `10.1088/0004-637X/809/1/8` 5, 21

Butler, R. P., Bedding, T. R., Kjeldsen, H., et al. 2004, ApJL, 600, L75, doi: `10.1086/381434` 52

Butler, R. P., Marcy, G. W., Williams, E., et al. 1996, PASP, 108, 500, doi: `10.1086/133755` 24, 82, 98

Butler, R. P., Vogt, S. S., Laughlin, G., et al. 2017, AJ, 153, 208, doi: `10.3847/1538-3881/aa66ca` 98

Cabot, S. H. C., Roettenbacher, R. M., Henry, G. W., et al. 2021, AJ, 161, 26, doi: `10.3847/1538-3881/abc41e` 136, 148, 185

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` 7, 180

Cegla, H. M., Watson, C. A., Shelyag, S., et al. 2018, ApJ, 866, 55, doi: `10.3847/1538-4357/aaddfc` 15, 136

Cersullo, F., Coffinet, A., Chazelas, B., Lovis, C., & Pepe, F. 2019, A&A, 624, A122, doi: `10.1051/0004-6361/201833852` 98

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` 15, 136

Chauvin, G., Lagrange, A.-M., Udry, S., & Mayor, M. 2007, A&A, 475, 723, doi: `10.1051/0004-6361:20067046` 22, 35

Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2015, ApJ, 810, 95, doi: `10.1088/0004-637X/810/2/95` 21

—. 2016, ApJ, 828, 99, doi: `10.3847/0004-637X/828/2/99` 5

Ciddor, P. E. 1996, ApOpt, 35, 1566, doi: `10.1364/AO.35.001566` 80

Coffinet, A., Lovis, C., Dumusque, X., & Pepe, F. 2019, A&A, 629, A27, doi: `10.1051/0004-6361/201833272` 99

Collier Cameron, A., Mortier, A., Phillips, D., et al. 2019, MNRAS, 487, 1082, doi: `10.1093/mnras/stz1215` 172

Collier Cameron, A., Ford, E. B., Shahaf, S., et al. 2020, arXiv e-prints, arXiv:2011.00018. `https://arxiv.org/abs/2011.00018` 17, 151, 155, 193

Cornachione, M. A., Bolton, A. S., Eastman, J. D., et al. 2019, PASP, 131, 124503, doi: `10.1088/1538-3873/ab4103` 13, 95

Correia, A. C. M., Udry, S., Mayor, M., et al. 2008, A&A, 479, 271, doi: `10.1051/0004-6361:20078908` 22, 35

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` 148

Cretignier, M., Dumusque, X., Allart, R., Pepe, F., & Lovis, C. 2020a, A&A, 633, A76, doi: `10.1051/0004-6361/201936548` 157, 222

Cretignier, M., Dumusque, X., Hara, N. C., & Pepe, F. 2021, arXiv e-prints, arXiv:2106.07301. `https://arxiv.org/abs/2106.07301` 159, 200

Cretignier, M., Francfort, J., Dumusque, X., Allart, R., & Pepe, F. 2020b, A&A, 640, A42, doi: `10.1051/0004-6361/202037722` 158, 159, 186, 200

Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. A., & Ford, E. B. 2017, ApJ, 846, 59, doi: `10.3847/1538-4357/aa8303` 15, 138, 157

Dawson, R. I., & Fabrycky, D. C. 2010, ApJ, 722, 937, doi: `10.1088/0004-637X/722/1/937` 51 de Beurs, Z. L., Vanderburg, A., Shallue, C. J., et al. 2020, arXiv e-prints, arXiv:2011.00003. `https://arxiv.org/abs/2011.00003` 17, 151, 157, 196

Del Moro, D. 2004, A&A, 428, 1007, doi: `10.1051/0004-6361:20040466` 52

Demarque, P., Woo, J.-H., Kim, Y.-C., & Yi, S. K. 2004, ApJS, 155, 667, doi: `10.1086/424966` 27

Desort, M., Lagrange, A. M., Galland, F., Udry, S., & Mayor, M. 2007, A&A, 473, 983, doi: `10.1051/0004-6361:20078144` 16, 136, 137

Díaz, M. R., Jenkins, J. S., Tuomi, M., et al. 2018, AJ, 155, 126, doi: `10.3847/1538-3881/aaa896` 148, 175

Dravins, D. 1982, ARA&A, 20, 61, doi: `10.1146/annurev.aa.20.090182.000425` 15, 136

Ducati, J. R. 2002, VizieR Online Data Catalog 140

Dumusque, X. 2018, A&A, 620, A47, doi: `10.1051/0004-6361/201833795` 15, 18, 82, 138, 151, 157, 159, 200

Dumusque, X., Pepe, F., Lovis, C., & Latham, D. W. 2015, ApJ, 808, 171, doi: `10.1088/0004-637X/808/2/171` 9

Dumusque, X., Santos, N. C., Udry, S., Lovis, C., & Bonfils, X. 2011a, A&A, 527, A82, doi: `10.1051/0004-6361/201015877` 16, 136, 137

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` 15, 136, 158

Dumusque, X., Lovis, C., Ségransan, D., et al. 2011c, A&A, 535, A55, doi: `10.1051/0004-6361/201117148` 17, 136, 138

Dumusque, X., Pepe, F., Lovis, C., et al. 2012, Nature, 491, 207, doi: `10.1038/nature11572` 22, 30, 34, 35, 36, 37, 45, 47, 52

Dumusque, X., Borsa, F., Damasso, M., et al. 2017, A&A, 598, A133, doi: `10.1051/0004-6361/201628671` 138

Duncan, D. K., Vaughan, A. H., Wilson, O. C., et al. 1991, ApJS, 76, 383, doi: `10.1086/191572` 28, 86

Endl, M., Kürster, M., Els, S., Hatzes, A. P., & Cochran, W. D. 2001, A&A, 374, 675, doi: `10.1051/0004-6361:20010723` 22, 32

Endl, M., Bergmann, C., Hearnshaw, J., et al. 2015, International Journal of Astrobiology, 14, 305, doi: `10.1017/S1473550414000081` 22, 34

Faria, J. P., Haywood, R. D., Brewer, B. J., et al. 2016, A&A, 588, A31, doi: `10.1051/0004-6361/201527899` 138

Feng, F., Tuomi, M., Jones, H. R. A., et al. 2017, AJ, 154, 135, doi: `10.3847/1538-3881/aa83b4` 149

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 17, 138, 145

Fischer, D., Driscoll, P., Isaacson, H., et al. 2009, ApJ, 703, 1545, doi: `10.1088/0004-637X/703/2/1545` 22, 35

Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, PASP, 128, 066001, doi: `10.1088/1538-3873/128/964/066001` 9, 17, 53, 75, 98, 113, 138

Ford, E., Wise, A., & Palumbo, M. 2021, RvSpectML/EchelleCCFs.jl: v0.1.11, v0.1.11, Zenodo, doi: `10.5281/zenodo.4593963` 144

Foreman-Mackey, D., Montet, B. T., Hogg, D. W., et al. 2015, ApJ, 806, 215, doi: `10.1088/0004-637X/806/2/215` 128

Freudling, W., Romaniello, M., Bramich, D. M., et al. 2013, A&A, 559, A96, doi: `10.1051/0004-6361/201322494` 81, 86, 87

Fulton, B. J., Petigura, E. A., Howard, A. W., et al. 2017, AJ, 154, 109, doi: `10.3847/1538-3881/aa80eb` 5

Gaia Collaboration. 2018, VizieR Online Data Catalog, I/345 89, 140, 150

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` 7, 180

Giguere, M. J., Fischer, D. A., Zhang, C. X. Y., et al. 2016, ApJ, 824, 150, doi: `10.3847/0004-637X/824/2/150` 17, 138, 145

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` 7, 180

Gilbertson, C., Ford, E. B., Jones, D. E., & Stenning, D. C. 2020, ApJ, 905, 155, doi: `10.3847/1538-4357/abc627` 17, 151, 153, 189

Gray, D. F., & Baliunas, S. L. 1995, ApJ, 441, 436, doi: `10.1086/175368` 28

Guedes, J. M., Rivera, E. J., Davis, E., et al. 2008, ApJ, 679, 1582, doi: `10.1086/587799` 21

Hartkopf, W. I., McAlister, H. A., & Mason, B. D. 2001, AJ, 122, 3480, doi: `10.1086/323923` 32

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` 16, 136, 137

—. 2013, ApJ, 770, 133, doi: `10.1088/0004-637X/770/2/133` 22

Haywood, R. D., Collier Cameron, A., Queloz, D., et al. 2014, International Journal of Astrobiology, 13, 155, doi: `10.1017/S147355041300044X` 17, 138

Haywood, R. D., Milbourne, T. W., Saar, S. H., et al. 2020, arXiv e-prints, arXiv:2005.13386. `https://arxiv.org/abs/2005.13386` 161, 206

Henry, G. W. 1999, PASP, 111, 845, doi: `10.1086/316388` 146

Henry, T. J., Soderblom, D. R., Donahue, R. A., & Baliunas, S. L. 1996, AJ, 111, 439, doi: `10.1086/117796` 27, 30

Holzer, P., Cisewski-Kehe, J., Fischer, D., & Zhao, L. 2020, arXiv e-prints, arXiv:2005.14083. `https://arxiv.org/abs/2005.14083` 18, 151, 161

Holzer, P. H., Cisewski-Kehe, J., Zhao, L., et al. 2021, arXiv e-prints, arXiv:2104.04887. `https://arxiv.org/abs/2104.04887` 18, 151, 161

Horne, K. 1986, PASP, 98, 609, doi: `10.1086/131801` 70

Huélamo, N., Figueira, P., Bonfils, X., et al. 2008, A&A, 489, L9, doi: `10.1051/0004-6361:200810596` 16, 136, 137

Innes, R. T. A. 1915, Circular of the Union Observatory Johannesburg, 30, 235 23

Isaacson, H., & Fischer, D. 2010, ApJ, 725, 875, doi: `10.1088/0004-637X/725/1/875` 86, 140, 148, 149, 150

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` 16, 136, 137

Jolliffe, I., & Cadima, J. 2016, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374, 20150202, doi: `10.1098/rsta.2015.0202` 103

Jones, D. E., Stenning, D. C., Ford, E. B., et al. 2017, arXiv e-prints, arXiv:1711.01318. `https://arxiv.org/abs/1711.01318` 18, 151, 161, 203

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` 7, 56, 75, 82, 98, 108, 135, 141, 180

Kervella, P., Bigot, L., Gallenne, A., & Thévenin, F. 2017a, A&A, 597, A137, doi: `10.1051/0004-6361/201629505` 23, 27, 29, 30

Kervella, P., Mignard, F., Mérand, A., & Thévenin, F. 2016, A&A, 594, A107, doi: `10.1051/0004-6361/201629201` 29

Kervella, P., Thévenin, F., & Lovis, C. 2017b, A&A, 598, L7, doi: `10.1051/0004-6361/201629930` 22, 32

Kervella, P., Thévenin, F., Ségransan, D., et al. 2003, A&A, 404, 1087, doi: `10.1051/0004-6361:20030570` 29

Kjeldsen, H., & Bedding, T. R. 1995, A&A, 293, 87. `https://arxiv.org/abs/astro-ph/9403015` 15, 136

Kjeldsen, H., Bedding, T. R., Butler, R. P., et al. 2005, ApJ, 635, 1281, doi: `10.1086/497530` 15, 52, 136

Kopparapu, R. K., Ramirez, R., Kasting, J. F., et al. 2013, ApJ, 765, 131, doi: `10.1088/0004-637X/765/2/131` 46

Kramer, M. A. 1991, AIChE Journal, 37, 233, doi: `10.1002/aic.690370209` 117

Lafarga, M., Ribas, I., Lovis, C., et al. 2020, A&A, 636, A36, doi: `10.1051/0004-6361/201937222` 18, 151, 158, 199

Lagrange, A.-M., Desort, M., & Meunier, N. 2010, A&A, 512, A38, doi: `10.1051/0004-6361/200913071` 47

Landoni, M., Riva, M., Pepe, F., et al. 2014, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9147, ESPRESSO front end exposure meter: a chromatic approach to radial velocity correction, 91478K, doi: `10.1117/12.2056406` 78

Lanza, A. F., Gizon, L., Zaqarashvili, T. V., Liang, Z. C., & Rodenbeck, K. 2019, A&A, 623, A50, doi: `10.1051/0004-6361/201834712` 15, 136

Leet, C., Fischer, D. A., & Valenti, J. A. 2019, AJ, 157, 187, doi: `10.3847/1538-3881/ab0d86` 59, 82, 142

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` 10, 141

Li, C.-H., Benedick, A. J., Fendel, P., et al. 2008, Nature, 452, 610, doi: `10.1038/nature06854` 9

Lindegren, L., & Dravins, D. 2003, A&A, 401, 1185, doi: `10.1051/0004-6361:20030181` 15, 136

Lomb, N. R. 1976, Ap&SS, 39, 447, doi: `10.1007/BF00648343` 42

Louden, E. M., Winn, J. N., Petigura, E. A., et al. 2021, AJ, 161, 68, doi: `10.3847/1538-3881/abcebd` 6

Lovis, C., & Pepe, F. 2007, A&A, 468, 1115, doi: `10.1051/0004-6361:20077249` 98

Lovis, C., Dumusque, X., Santos, N. C., et al. 2011, arXiv e-prints, arXiv:1107.5325. `https://arxiv.org/abs/1107.5325` 16, 136, 137

Lundkvist, M., Kjeldsen, H., & Silva Aguirre, V. 2014, A&A, 566, A82, doi: `10.1051/0004-6361/201423408` 27, 29, 30

Ma, B., Ge, J., Muterspaugh, M., et al. 2018, MNRAS, 480, 2411, doi: `10.1093/mnras/sty1933` 148, 175, 176

Maldonado, J., Martinez-Arnáiz, R. M., Eiroa, C., & Montes, D. 2010, Astrophysics and Space Science Proceedings, 14, 419, doi: `10.1007/978-3-642-11250-8_108` 140

Mamajek, E. E., & Hillenbrand, L. A. 2008, ApJ, 687, 1264, doi: `10.1086/591785` 28, 30

Mann, A. W., Feiden, G. A., Gaidos, E., Boyajian, T., & von Braun, K. 2015, ApJ, 804, 64, doi: `10.1088/0004-637X/804/1/64` 29, 30

Mayor, M., & Queloz, D. 1995, Nature, 378, 355, doi: `10.1038/378355a0` 1, 3

Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20 15, 136, 148

Mayor, M., Marmier, M., Lovis, C., et al. 2011, arXiv e-prints, arXiv:1109.2497. `https://arxiv.org/abs/1109.2497` 98

Meunier, N., Desort, M., & Lagrange, A. M. 2010, A&A, 512, A39, doi: `10.1051/0004-6361/200913551` 16, 17, 47, 137

Meunier, N., & Lagrange, A. M. 2013, A&A, 551, A101, doi: `10.1051/0004-6361/201219917` 17, 138, 145

—. 2019, A&A, 625, L6, doi: `10.1051/0004-6361/201935099` 16, 136, 137

Meunier, N., Lagrange, A. M., & Borgniet, S. 2017, A&A, 607, A6, doi: `10.1051/0004-6361/201630328` 15, 138

Meunier, N., Lagrange, A. M., Borgniet, S., & Rieutord, M. 2015, A&A, 583, A118, doi: `10.1051/0004-6361/201525721` 16, 136, 137

Milaković, D., Pasquini, L., Webb, J. K., & Lo Curto, G. 2020, MNRAS, 493, 3997, doi: `10.1093/mnras/staa356` 99, 113, 124, 141

Modigliani, A., Sownsowska, D., & Lovis, C. 2019, ESPRESSO Pipeline User Manual, ESO 81, 86, 87

Molaro, P., Esposito, M., Monai, S., et al. 2013, A&A, 560, A61, doi: `10.1051/0004-6361/201322324` 9, 108, 141

Morel, P., Provost, J., Lebreton, Y., Thévenin, F., & Berthomieu, G. 2000, A&A, 363, 675 28, 30

Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2015, ApJS, 217, 31, doi: `10.1088/0067-0049/217/2/31` 21

Murdoch, K. A., Hearnshaw, J. B., & Clark, M. 1993, ApJ, 413, 349, doi: `10.1086/173003` 22

Murphy, M. T., Udem, T., Holzwarth, R., et al. 2007, MNRAS, 380, 839, doi: `10.1111/j.1365-2966.2007.12147.x` 9

Nidever, D. L., Marcy, G. W., Butler, R. P., Fischer, D. A., & Vogt, S. S. 2002, ApJS, 141, 503, doi: `10.1086/340570` 140

Noyes, R. W., Hartmann, L. W., Baliunas, S. L., Duncan, D. K., & Vaughan, A. H. 1984, ApJ, 279, 763, doi: `10.1086/161945` 28

Oliphant, T. 2006–, NumPy: A guide to NumPy, USA: Trelgol Publishing. `http://www.numpy.org/` 96, 133

Pearson, K. 1901, LIII. On lines and planes of closest fit to systems of points in space, Zenodo, doi: `10.1080/14786440109462720` 103

Pepe, F., Cristiani, S., Rebolo, R., et al. 2013, The Messenger, 153, 6 7, 56, 98, 135, 180

Perryman, M. A. C., Lindegren, L., Kovalevsky, J., et al. 1997, A&A, 323, L49 25, 29

Petersburg, R. R., McCracken, T. M., Eggerman, D., et al. 2018, ApJ, 853, 181, doi: `10.3847/1538-4357/aaa487` 108

Petersburg, R. R., Ong, J. M. J., Zhao, L. L., et al. 2020, AJ, 159, 187, doi: `10.3847/1538-3881/ab7e31` 3, 7, 13, 98, 102, 108, 115, 135, 142, 180, 186, 217

Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, Proceedings of the National Academy of Science, 110, 19273, doi: `10.1073/pnas.1319909110` 5

Pijpers, F. P. 2003, A&A, 400, 241, doi: `10.1051/0004-6361:20021839` 140

Pinsonneault, M. H., DePoy, D. L., & Coffee, M. 2001, ApJ, 556, L59 26

Piskunov, N. E., & Valenti, J. A. 2002, A&A, 385, 1095, doi: `10.1051/0004-6361:20020175` 70

Plavchan, P., Latham, D., Gaudi, S., et al. 2015 98

Pourbaix, D., & Boffin, H. M. J. 2016, A&A, 586, A90, doi: `10.1051/0004-6361/201527859` 21, 23, 27, 29, 30, 32, 34, 41

Pourbaix, D., Neuforge-Verheecke, C., & Noels, A. 1999, A&A, 344, 172 32

Povich, M. S., Giampapa, M. S., Valenti, J. A., et al. 2001, AJ, 121, 1136, doi: `10.1086/318745` 17, 138

Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, AJ, 156, 123, doi: `10.3847/1538-3881/aabc4f` 96, 133

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` 9, 12, 108, 141

Probst, R. A., Lo Curto, G., Ávila, G., et al. 2016, in Proc. SPIE, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 990864, doi: `10.1117/12.2231434` 58

Probst, R. A., Milaković, D., Toledo-Padrón, B., et al. 2020, Nature Astronomy, 4, 603, doi: `10.1038/s41550-020-1010-x` 12, 141

Quarles, B., & Lissauer, J. J. 2016, AJ, 151, 111, doi: `10.3847/0004-6256/151/5/111` 22, 49

Queloz, D., Henry, G. W., Sivan, J. P., et al. 2001, A&A, 379, 279, doi: `10.1051/0004-6361:20011308` 17, 47, 138, 145

Queloz, D., Bouchy, F., Moutou, C., et al. 2009, A&A, 506, 303, doi: `10.1051/0004-6361/200913096` 17, 138, 145

Quintana, E. V., Adams, F. C., Lissauer, J. J., & Chambers, J. E. 2007, ApJ, 660, 807, doi: `10.1086/512542` 21, 34

Quintana, E. V., & Lissauer, J. J. 2006, Icarus, 185, 1, doi: `10.1016/j.icarus.2006.06.016` 21, 34

Quintana, E. V., Lissauer, J. J., Chambers, J. E., & Duncan, M. J. 2002, ApJ, 576, 982, doi: `10.1086/341808` 34

Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, MNRAS, 452, 2269, doi: `10.1093/mnras/stv1428` 17, 152

Rajpaul, V., Aigrain, S., & Roberts, S. 2016, MNRAS, 456, L6, doi: `10.1093/mnrasl/slv164` 22, 34, 45, 47

Rajpaul, V., Buchhave, L. A., & Aigrain, S. 2017, MNRAS, 471, L125, doi: `10.1093/mnrasl/slx116` 138

Rajpaul, V. M., Aigrain, S., & Buchhave, L. A. 2020, MNRAS, 492, 3960, doi: `10.1093/mnras/stz3599` 18, 151, 160, 201

Rieutord, M., & Rincon, F. 2010, Living Reviews in Solar Physics, 7, 2, doi: `10.12942/lrsp-2010-2` 16, 136, 137

Rincon, F., & Rieutord, M. 2018, Living Reviews in Solar Physics, 15, 6, doi: `10.1007/s41116-018-0013-5` 16, 136, 137

Robertson, P., Mahadevan, S., Endl, M., & Roy, A. 2014, Science, 345, 440, doi: `10.1126/science.1253253` 17, 138, 145

Rojas-Ayala, B., Covey, K. R., Muirhead, P. S., & Lloyd, J. P. 2012, ApJ, 748, 93, doi: `10.1088/0004-637X/748/2/93` 26

Rosenthal, L. J., Fulton, B. J., Hirsch, L. A., et al. 2021, ApJS, 255, 8, doi: `10.3847/1538-4365/abe23c` 148, 175, 176

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 16, 136, 137

Saar, S. H., Butler, R. P., & Marcy, G. W. 1998, ApJL, 498, L153, doi: `10.1086/311325` 17, 138, 145

Saar, S. H., & Donahue, R. A. 1997, ApJ, 485, 319, doi: `10.1086/304392` 16, 136, 137

Saar, S. H., & Fischer, D. 2000, ApJL, 534, L105, doi: `10.1086/312648` 17, 47, 138

Santos, N. C., Mayor, M., Naef, D., et al. 2000, A&A, 361, 265 47

Santos, N. C., Mortier, A., Faria, J. P., et al. 2014, A&A, 566, A35, doi: `10.1051/0004-6361/201423808` 172

Scargle, J. D. 1982, ApJ, 263, 835, doi: `10.1086/160554` 42

Schlaufman, K. C. 2010, ApJ, 719, 602, doi: `10.1088/0004-637X/719/1/602` 6

Schwab, C., Rakich, A., Gong, Q., et al. 2016a, 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` 7, 135, 180

Schwab, C., Rakich, A., Gong, Q., et al. 2016b, in Proc. SPIE, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 99087H, doi: `10.1117/12.2234411` 56

Ségransan, D., Kervella, P., Forveille, T., & Queloz, D. 2003, A&A, 397, L5, doi: `10.1051/0004-6361:20021714` 29

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` 7, 180

Simola, U., Dumusque, X., & Cisewski-Kehe, J. 2019, A&A, 622, A131, doi: `10.1051/0004-6361/201833895` 145

Skelly, M. B., Unruh, Y. C., Collier Cameron, A., et al. 2008, MNRAS, 385, 708, doi: `10.1111/j.1365-2966.2008.12917.x` 17, 138, 145

Steinmetz, T., Wilken, T., Araujo-Hauck, C., et al. 2008, Science, 321, 1335, doi: `10.1126/science.1161030` 9, 12, 58

Suárez Mascareño, A., Faria, J. P., Figueira, P., et al. 2020, arXiv e-prints, arXiv:2005.12114. `https://arxiv.org/abs/2005.12114` 7, 98, 135, 180

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` 7, 180

Szymkowiak, A., Fischer, D. A., et al. in preparation 87

Tal-Or, L., Trifonov, T., Zucker, S., Mazeh, T., & Zechmeister, M. 2019, MNRAS, 484, L8, doi: `10.1093/mnrasl/sly227` 89

Thébault, P., Marzari, F., & Scholl, H. 2006, Icarus, 183, 193, doi: `10.1016/j.icarus.2006.01.022` 21

—. 2008, MNRAS, 388, 1528, doi: `10.1111/j.1365-2966.2008.13536.x` 21, 34

—. 2009, MNRAS, 393, L21, doi: `10.1111/j.1745-3933.2008.00590.x` 21, 34

Thévenin, F., Provost, J., Morel, P., et al. 2002, A&A, 392, L9, doi: `10.1051/0004-6361:20021074` 30

Thompson, A. P. G., Watson, C. A., de Mooij, E. J. W., & Jess, D. B. 2017, MNRAS, 468, L16, doi: `10.1093/mnrasl/slx018` 157

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` 7, 180

Tokovinin, A., Fischer, D. A., Bonati, M., et al. 2013, PASP, 125, 1336, doi: `10.1086/674012` 24, 36, 148

Trifonov, T., Tal-Or, L., Zechmeister, M., et al. 2020, arXiv e-prints, arXiv:2001.05942. `https://arxiv.org/abs/2001.05942` 89

Tronsgaard, R., Buchhave, L. A., Wright, J. T., Eastman, J. D., & Blackman, R. T. 2019, MNRAS, 489, 2395, doi: `10.1093/mnras/stz2181` 78

Tsalmantza, P., & Hogg, D. W. 2012, ApJ, 753, 122, doi: `10.1088/0004-637X/753/2/122` 117

Tuomi, M., Jones, H. R. A., Jenkins, J. S., et al. 2013, A&A, 551, A79, doi: `10.1051/0004-6361/201220509` 47, 149

Turcotte, S., & Wimmer-Schweingruber, R. F. 2002, J Geophys Res-Space, 107, 1442 26

Vacca, W. D., Cushing, M. C., & Rayner, J. T. 2003, PASP, 115, 389, doi: `10.1086/346193` 26

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science Engineering, 13, 22, doi: `10.1109/MCSE.2011.37` 96, 133

van Leeuwen, F. 2007, Hipparcos, the New Reduction of the Raw Data, Vol. 350, doi: `10.1007/978-1-4020-6342-8` 23, 25, 29, 140

Vaughan, A. H., Preston, G. W., & Wilson, O. C. 1978, PASP, 90, 267, doi: `10.1086/130324` 28

Vernet, J., Dekker, H., D'Odorico, S., et al. 2011, A&A, 536, A105, doi: `10.1051/0004-6361/201117752` 26

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: `https://doi.org/10.1038/s41592-019-0686-2` 59, 96, 133

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` 148

Wang, J., & Ford, E. B. 2011, MNRAS, 418, 1822, doi: `10.1111/j.1365-2966.2011.19600.x` 87, 90

Wang, S. X., Wright, J. T., MacQueen, P., et al. 2020, PASP, 132, 014503, doi: `10.1088/1538-3873/ab5021` 99

Wertheimer, J. G., & Laughlin, G. 2006, AJ, 132, 1995, doi: `10.1086/507771` 22, 32

Wiegert, P. A., & Holman, M. J. 1997, AJ, 113, 1445, doi: `10.1086/118360` 22, 34, 41, 49

Wilken, T., Curto, G. L., Probst, R. A., et al. 2012, Nature, 485, 611, doi: `10.1038/nature11092` 9, 108, 141

Wilson, A. G., Knowles, D. A., & Ghahramani, Z. 2012, in Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML'12 (USA: Omnipress), 1139–1146. `http://dl.acm.org/citation.cfm?id=3042573.3042720` 192

Wilson, M. L., Eastman, J. D., Cornachione, M. A., et al. 2019, PASP, 131, 115001, doi: `10.1088/1538-3873/ab33c5` 90

Wilson, O. C. 1978, ApJ, 226, 379, doi: `10.1086/156618` 28

Winn, J. N., Fabrycky, D., Albrecht, S., & Johnson, J. A. 2010, ApJL, 718, L145, doi: `10.1088/2041-8205/718/2/L145` 6

Winn, J. N., & Fabrycky, D. C. 2015, ARA&A, 53, 409, doi: `10.1146/annurev-astro-082214-122246` 56

Wise, A. W., Dodson-Robinson, S. E., Bevenour, K., & Provini, A. 2018, AJ, 156, 180, doi: `10.3847/1538-3881/aadd94` 157

Wittenmyer, R. A., Endl, M., Bergmann, C., et al. 2014, in IAU Symposium, Vol. 293, Formation, Detection, and Characterization of Extrasolar Habitable Planets, ed. N. Haghighipour, IAU Symposium, 58–64, doi: `10.1017/S1743921313012532` 35

Woodbridge, Y., Elidan, G., & Wiesel, A. 2020, arXiv e-prints, arXiv:2004.10255. `https://arxiv.org/abs/2004.10255` 117

Wright, J. T. 2005, PASP, 117, 657, doi: `10.1086/430369` 47

Wright, J. T., & Eastman, J. D. 2014, Publications of the Astronomical Society of the Pacific, 126, 838–852, doi: `10.1086/678541` 78

Xu, X., Cisewski-Kehe, J., Davis, A. B., Fischer, D. A., & Brewer, J. M. 2019, AJ, 157, 243, doi: `10.3847/1538-3881/ab1b47` 75

Zechmeister, M., Anglada-Escudé, G., & Reiners, A. 2014, A&A, 561, A59, doi: `10.1051/0004-6361/201322746` 70, 72, 142

Zhao, J., & Tinney, C. G. 2020, MNRAS, 491, 4131, doi: `10.1093/mnras/stz3254` 17, 151, 156, 195

Zhao, L., Fischer, D. A., Brewer, J., Giguere, M., & Rojas-Ayala, B. 2018, AJ, 155, 24, doi: `10.3847/1538-3881/aa9bea` 5, 216

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` 18, 138

Zhao, L. L., Hogg, D. W., Bedell, M., & Fischer, D. A. 2021, AJ, 161, 80, doi: `10.3847/1538-3881/abd105` 95, 142, 186, 217