

Modeling Doppler shifts in radial-velocity data with deep learning toward Earth-mass exoplanet detection

Isidro Gómez-Vargas^{1,2,3,*}, Xavier Dumusque¹, Yinan Zhao⁴, Khaled Al Moulla⁵, and Michael Cretignier⁶

¹ Department of Astronomy, University of Geneva 51 chemin de Pegasi, 1290 Versoix, Switzerland

² Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía s/n, 18008 Granada, Spain

³ Institute of Space Sciences (CSIC), Carrer de Can Magrans s/n, 08193 Barcelona, Spain

⁴ Department of Astronomy, University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA

⁵ Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal

⁶ Department of Physics, University of Oxford, OX13RH Oxford, UK

Received 8 February 2026 / Accepted 16 June 2026

ABSTRACT

Context. Detecting the tiny Doppler shifts induced by Earth-mass planets in stellar radial-velocity measurements remains extremely challenging due to stellar activity. Despite substantial progress in statistical and machine-learning techniques, many deep-learning methods performing well on simulated data remain difficult to apply reliably on real stellar spectra.

Aims. The aim of this work is to develop a deep-learning framework that generalizes to real, unseen spectra and improves the detectability of Earth-mass planets in radial-velocity data.

Methods. We train artificial neural networks on HARPS-N solar spectra with injected planetary signals, using physics-motivated spectral representations based on flux and line-formation temperature, together with their velocity gradients. Two training strategies are explored: hold-out testing, which provides a direct assessment of generalization to unseen spectra, and cross-validation, which evaluates performance across multiple folds of the dataset. The robustness of the model is enhanced by optimizing the hyperparameters based on genetic-algorithms, and the predictive uncertainty is quantified using the Monte Carlo dropout.

Results. Our most precise neural network model reliably retrieves, under the cross-validation strategy, the amplitudes, phases, and orbital periods of planetary signals with amplitudes greater than or equal to 25 cm/s and periods between 10 and 550 days. In addition, in all cases tested here, the successfully recovered signals correspond to the most significant peaks in the periodograms of the Doppler-shift predictions. Temperature-based spectral-shell representations consistently outperform flux-based shells, particularly in terms of predictive uncertainty and generalization to unseen data. As a byproduct, we release *doppleriann*, a Python package that implements the proposed framework.

Conclusions. Our results demonstrate that combining physically motivated spectral representations with deep learning provides a promising pathway toward the detection of Earth-mass planets in radial-velocity data from real observations, supported by a modeling framework that is both physically grounded and statistically rigorous, incorporating uncertainty quantification and optimized training strategies.

Key words. methods: data analysis – methods: statistical – techniques: radial velocities – techniques: spectroscopic – planets and satellites: detection

1. Introduction

The Doppler spectroscopy technique, also known as the radial velocity (RV) method, is one of the most relevant ways to detect exoplanets. It was crucial in the discovery of the first exoplanet orbiting a solar-type star in 1995, the 51 Pegasi b hot-Jupiter (Mayor & Queloz 1995), and to date, it has identified over 1000 exoplanets¹. The RV signal of a planet is directly proportional to the mass of the planet and inversely proportional to the cube root of its period. Therefore, detecting Earth-mass planets orbiting solar-type stars in their habitable zone (HZ), at orbital periods of a few hundred days, is extremely challenging, as it requires RV measurements with a precision of ~ 0.1 m/s.

Besides the technological challenge to reach a 0.1 m/s RV precision on timescales ranging from a few minutes to several years, the RV method is sensitive to stellar variability

(e.g., Saar & Donahue 1997; Lindegren & Dravins 2003; Fischer et al. 2016; Dumusque et al. 2017; Meunier et al. 2019; Luhn et al. 2023; Burt et al. 2025), which further complicates the detection of Earth-mass planets. Stellar variability arises from different physical processes: the propagation of pressure waves within the stellar photosphere, manifested as stellar oscillations (e.g., Dumusque et al. 2011; Luhn et al. 2025); convection-driven variations in surface brightness associated with granulation (e.g., Dumusque et al. 2011; Al Moulla et al. 2023; Lakeland et al. 2024; O’Sullivan et al. 2025); and the emergence and evolution of surface magnetic fields responsible for magnetic activity (e.g., Saar & Donahue 1997; Meunier et al. 2010; Dumusque et al. 2014; Korhonen et al. 2015; Zhao et al. 2025; Zhao & Dumusque 2023).

Even with the precision achieved by state-of-the-art high-resolution spectrographs such as ESPRESSO (Pepe et al. 2021), detecting Earth-mass planets in the HZ of solar-type stars remains beyond reach (Figueira et al. 2025). In theory, if measurement noise were independent, a few hundred

* Corresponding author: igomez@iaa.es

¹ See <https://exoplanet.eu/catalog/> for verification.

observations would suffice to detect an Earth-mass planet (Hara & Ford 2023). In practice, however, stellar variability dominates the signal. For instance, solar RV measurements obtained with HARPS-N exhibit a standard deviation of 2.95 m/s over 10 years (Dumusque et al. 2026). Consequently, data analysis, rather than instrumental sensitivity, is likely the primary limiting factor.

Stellar variability mitigation is essential for improving exoplanet detectability when using the RV method. Several studies have addressed this issue using techniques such as principal component analysis (e.g., Davis et al. 2017; Collier Cameron et al. 2021; Cretignier et al. 2022), Gaussian processes (e.g., Rajpaul et al. 2015; Barragán et al. 2022; Delisle et al. 2022; Hara & Delisle 2025), or line-by-line RV analysis (e.g., Dumusque 2018; Artigau et al. 2022; Cretignier et al. 2023; Salzer et al. 2025), all aimed at disentangling planetary signals from the effects of stellar variability.

In recent years, interest in machine-learning techniques for RV exoplanet detection has grown substantially. In particular, artificial neural networks (ANNs), including deep learning (DL) methods, have demonstrated high precision when applied to both simulated and stellar spectral data (de Beurs et al. 2022; Perger et al. 2023; Nieto & Díaz 2023; Liang et al. 2023; Kjærsgaard et al. 2023; Zhao et al. 2024; Gavankar et al. 2026; McWilliam et al. 2026), owing to their ability to model complex nonlinear relationships (Hornik et al. 1990).

Convolutional neural networks (CNNs) have been shown to be effective in several contexts. For instance, de Beurs et al. (2022) applied CNNs to solar data in combination with the cross-correlation function (CCF) technique, while Perger et al. (2023) reports promising results using CNNs on simulated data for two different stars and multiple physical observables. Binary classification of simulated RV time series to determine the presence of planetary signals was explored by Nieto & Díaz (2023).

Other neural-network architectures have also been investigated. The autoencoder-based approach of Liang et al. (2023) demonstrates the recovery of low-amplitude signals in simulated data, and Kjærsgaard et al. (2023) employed autoencoders to improve telluric correction and enhance planet detectability.

Beyond simulated datasets, recent studies have demonstrated promising performance on real stellar data. In particular, the CNN models developed by Zhao et al. (2024) achieve high precision, detecting signals at the 20-centimeter-per-second level in solar data and at 50 cm/s in observations from Alpha Centauri B and Tau Ceti. Likewise, Colwell et al. (2024) reports detections of signals as low as 20 cm/s using an ensemble of CNNs trained on real solar spectra, although the analysis is limited to periods of 50 days and requires computationally intensive models with significant prediction offsets.

Although these approaches show strong performance within their respective settings, most rely on neural network models that must be retrained for each star or simulation, leaving their behavior largely untested across different data distributions, such as new stellar targets, varying observing conditions, or unseen spectral regions. As a result, while these methods are, in principle, applicable to multiple stars, their trained models have not yet demonstrated consistent performance when applied to data with different statistical properties. This limitation is particularly relevant for RV analyses, where real spectra vary in wavelength coverage, noise properties, and levels of stellar activity. These considerations motivate the development of more transferable and interpretable machine-learning frameworks that operate directly on real stellar spectra and maintain robust performance on genuinely unseen data.

Building on these examples, our work presents an alternative approach based on real solar data combined with physically motivated data compression to train neural-network models conceptually similar to those of Zhao et al. (2024). As in Colwell et al. (2024), our method directly predicts the Doppler shift (DS) induced by planetary signals, thereby extracting information that is not directly accessible from classical CCF-based activity indicators. Compared to Colwell et al. (2024), which focuses on short-period signals (up to 50 days) using computationally intensive ensemble models, our framework extends the analysis to a broader range of orbital periods (up to 550 days), while employing physically motivated spectral representations that enable lightweight neural-network architectures. We incorporated advanced hyperparameter optimization and uncertainty quantification to obtain more robust models and prevent overfitting. The trained networks were evaluated on spectral-shell samples that were completely held out from training and validation, providing a rigorous test of their ability to generalize to genuinely unseen temporal and spectral data.

We developed our models using solar data from the HARPS-N telescope (Dumusque et al. 2015; Phillips et al. 2016; Dumusque et al. 2026), based on both stellar flux and line-formation temperature, demonstrating improved performance with the latter. This behavior is consistent with previous findings (Al Moulla et al. 2022, 2024), which show that stellar activity signals depend on the formation temperature of spectral lines, mainly due to the suppression of convective motions in active regions, thereby facilitating the disentanglement of stellar activity and planetary signals.

This paper is structured as follows. In Sect. 2, we describe the spectral-shell representation used as input to our deep-learning framework, considering flux and temperature information. Section 3 details the spectral data processing. In Sect. 4, we present our deep-learning framework, outlining the core methodology as well as the training and prediction strategies, including the hold-out (HO) and cross-validation (CV) schemes. Section 5 reports the results of the planetary recovery tests for both our HO and CV approaches, using flux- and temperature-based representations, together with a discussion of their implications. Finally, Sect. 6 summarizes the main conclusions of this work. Appendix A presents additional validation tests of the methodology, while Appendix B provides an analysis of the predictive uncertainties of the neural-network models to assess their robustness.

2. Spectral-shell representations

In this work, we used spectral shells as input to our deep-learning framework, rather than stellar spectra or CCF-based representations. Spectral shells, initially introduced in Cretignier et al. (2022), are a reduced-dimensionality projection of high-resolution spectra designed to preserve sensitivity to DSs and stellar activity effects. They have a lower dimensionality than stellar spectra, significantly improving neural-network training. Unlike CCFs, which are also low-dimensionality, they have been shown to better probe stellar variability signals (Cretignier et al. 2022). Moreover, this representation compresses the spectral data into a two-dimensional grid defined by normalized flux as a function of the flux gradient. The flux gradient is the same for all spectra and is derived from a high signal-to-noise master spectrum constructed as the average of all available stellar spectra.

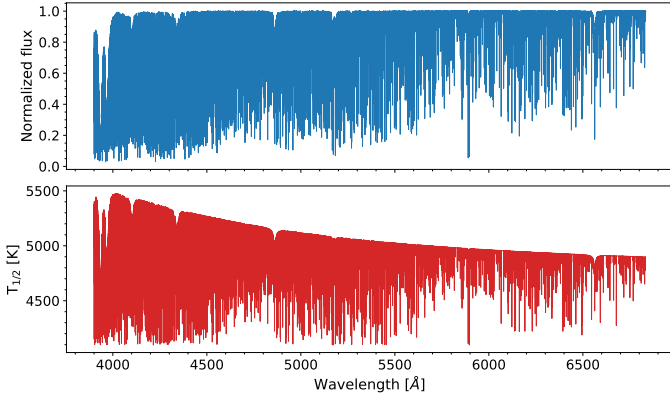


Fig. 1. Example of a single HARPS-N solar spectrum. Top: flux, normalized to its continuum level, as a function of wavelength. Bottom: corresponding average formation temperature, $T_{1/2}$, corresponding to the same wavelengths.

A shell grid is then formed by binning the space defined by the flux gradient and normalized flux. Within each bin, the mean flux variation relative to the master spectrum is recorded. This transformation retains localized spectral variations with respect to the master spectrum that are sensitive to line shape variation, while mitigating high-frequency noise.

In Zhao et al. (2024), rather than using the originally defined gradient of flux with respect to wavelength, the authors show that adopting the gradient of flux with respect to velocity is more appropriate and enables improved modeling of stellar variability. This is because spectral lines from different wavelengths overlap in the spectral-shell representation when expressed in velocity space, thereby correcting for dispersion.

Our methodology builds upon the use of spectral shells with a gradient defined with respect to velocity and extends this approach by replacing normalized flux with formation temperature. This modification was motivated by the work of Al Moulla et al. (2022) (see also Siegel et al. 2022, 2024; Al Moulla et al. 2024), which demonstrates that stellar variability depends on formation height in the stellar photosphere (or, equivalently, on formation temperature). Each sampled spectral point is associated with its average formation temperature, $T_{1/2}$, determined through radiative-transfer modeling and defined as the photospheric temperature at which the cumulative flux contribution function reaches 50% of its maximum. This enables a mapping of the spectral information with respect to photospheric depth. Figure 1 shows a comparison of the same solar spectrum using either normalized flux or formation temperature. This temperature-dependent refinement enhances the capability of the shell representation to probe activity-induced line profile variations. In the following subsections, we describe the formalism for constructing shell representations based on both flux and temperature.

2.1. Flux-based shell representation

Let $F(\lambda)$ and $F_0(\lambda)$ be two spectra defined on the same wavelength grid, with $F_0(\lambda)$ denoting a master or reference spectrum. Under a linear approximation, when $F(\lambda)$ is Doppler-shifted by a small velocity ΔV , the observed flux variation at pixel i is given by (see Bouchy et al. 2001)

$$F(\lambda_i) - F_0(\lambda_i) = \Delta F_i = \frac{\partial F_0(\lambda_i)}{\partial \lambda_i} \cdot \Delta \lambda_i = \frac{\partial F_0(\lambda_i)}{\partial \lambda_i} \cdot \frac{\Delta V_i}{c} \lambda_i, \quad (1)$$

which leads to

$$\Delta V_i = \frac{c}{\lambda_i} \cdot \Delta F_i \cdot \left(\frac{\partial F_0(\lambda_i)}{\partial \lambda_i} \right)^{-1}. \quad (2)$$

To eliminate chromatic effects from dispersion, the shell representation is constructed using the flux derivative with respect to velocity rather than wavelength:

$$\Delta V_i = \Delta F_i \cdot \left(\frac{\partial F_0(v_i)}{\partial v_i} \right)^{-1}. \quad (3)$$

In this formulation, each shell is defined as the set of tuples $(F_0, dF_0/dv)$, and bins are formed in this space. The weight associated with each velocity bin is defined as

$$W_{v,i} = \left[\frac{\partial F_0(v_i)}{\partial v_i} \cdot \frac{1}{\sigma_{F_i}} \right]^2, \quad (4)$$

where σ_{F_i} denotes the noise associated with the observed flux $F(v_i)$. This quantity, equal to $\sqrt{F(v_i) + \sigma_{det,i}^2}$, where $\sigma_{det,i}$ is detector noise at the pixel i , which is generally given by the data extraction pipeline. The total weight in each cell of the shell grid is then calculated as

$$W_{v,cell} = \sum_{i \in cell} W_{v,i}. \quad (5)$$

These weights $W_{v,cell}$ are used to construct the masked shell representation (illustrated in Fig. 2), obtained by the element-wise product between the original spectral shell and the associated weight map. In our experiments, this weighting yields more stable representations and neural-network training by reducing the contribution of low-information or noisy regions.

2.2. Temperature-based shell representation

To account for the fact that stellar activity has a photospheric depth dependence (Al Moulla et al. 2022), we extended the shell framework using a formation temperature-binned approach. Each spectral pixel is associated with its average formation temperature $T_{1/2}$. In analogy with the flux-based formalism, we define an equivalent velocity change in temperature space as

$$\Delta V_T = \frac{\Delta T}{g_T}, \quad \text{with } g_T = \frac{\partial T}{\partial v}, \quad (6)$$

where g_T denotes the local numerical slope, $\frac{\partial T}{\partial v}$, of the temperature spectrum computed along the Doppler velocity grid. This term describes how the temperature values derived from the flux-temperature mapping (described in the following section) vary with velocity and should not be interpreted as a physical temperature gradient.

The corresponding uncertainty in velocity space becomes

$$\sigma_v = \frac{\sigma_T}{g_T}. \quad (7)$$

Here, σ_T is not a directly observed quantity; it is derived by propagating the photon noise in flux, σ_{F_i} , through the flux-to-temperature mapping:

$$\sigma_T = f_{F \rightarrow T}(\sigma_{F_i}), \quad (8)$$

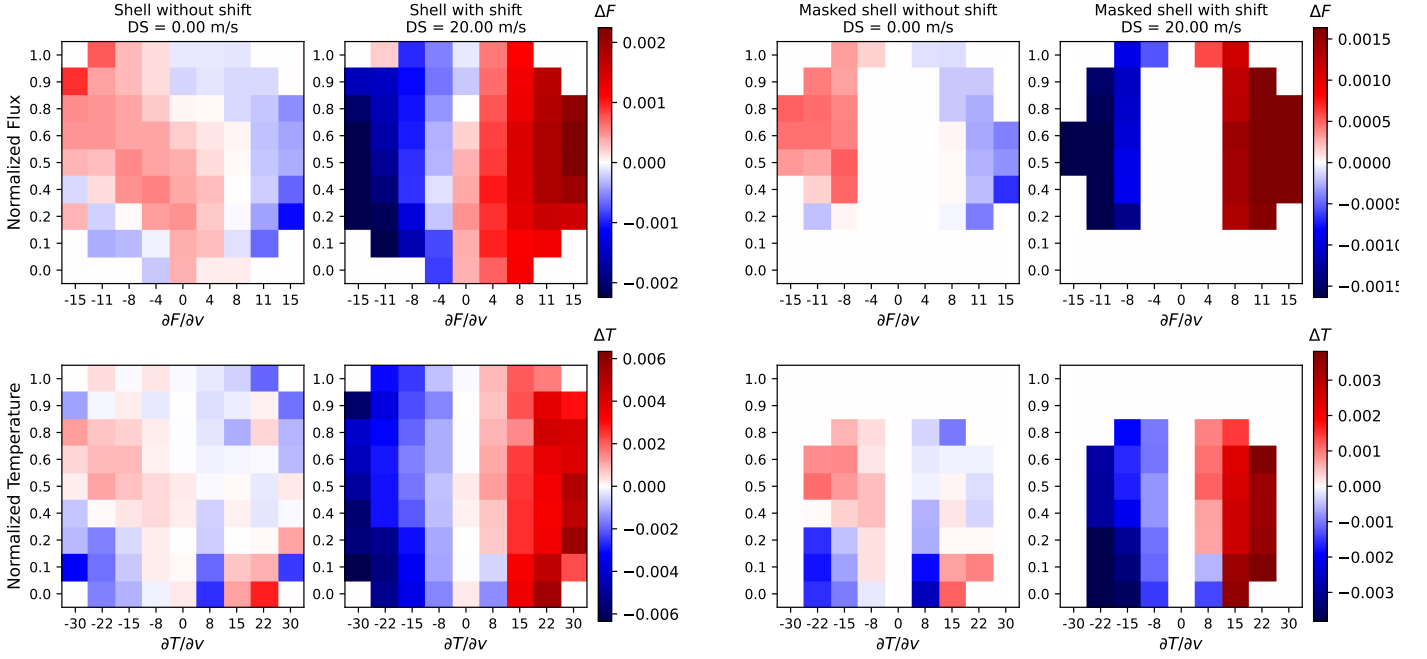


Fig. 2. Shell representations for flux and temperature at the time of a maximum DS of 20 m/s. The masked shell is defined as the element-wise product between the spectral shell and the associated density-weight map. In the shell construction, a cell is assigned zero only when no spectral pixels populate the corresponding bin. Although the same selected spectral pixels are used for both the flux- and temperature-based shells, they are projected onto different shell spaces, $(F, \partial F/\partial v)$ and $(T, \partial T/\partial v)$, so the occupancy of the shell cells is not identical between the two representations. Top row: Flux representations, with the first two panels unmasked and the last two masked. Bottom row: Same configuration for the temperature-based representation.

where $f_{F \rightarrow T}$ denotes the calibration function (either empirical or model-based) relating flux uncertainties to effective temperature uncertainties at each wavelength.

The weight associated with each velocity bin in the temperature shell is thus

$$W_{v,i}^{(T)} = \left[\frac{g_T}{f_{F \rightarrow T}(\sigma_{F_i})} \right]^2, \quad (9)$$

and the total weight in a shell cell becomes

$$W_{v,\text{cell}}^{(T)} = \sum_{i \in \text{cell}} W_{v,i}^{(T)}. \quad (10)$$

This formulation enables the shell representation to retain sensitivity to shifts in stellar activity at different photospheric depths, as traced by spectral line formation temperatures.

3. Data processing

We analyzed a 10-year dataset of HARPS-N solar observations acquired between 2015 and 2024, comprising 2036 daily binned high-resolution spectra (Dumusque et al. 2021; Dumusque et al. 2026). Each spectrum consists of 293 401 flux measurements spanning the wavelength range from 3900 to 6900 Å. The raw spectra were preprocessed using the YARARA pipeline (Cretignier et al. 2021), which corrects for instrumental systematics, telluric absorption, and stellar activity. These corrections are important for mitigating systematic features unrelated to planetary DSs. We note that the YARARA activity correction was reprojected into the spectra we analyzed, forming the so-called “YVA” dataset (e.g., Dalal et al. 2024; O’Sullivan et al. 2025), as our goal is to develop a machine learning framework

that can detect small-mass planetary signals in the presence of stellar signals.

Keeping the same timestamps of the original spectra, we injected synthetic Keplerian planetary signals into them to augment the training dataset and introduce controlled DS. We used the formalism described in Bouchy et al. (2001) to inject a planetary signal into the spectra. For two spectra Doppler-shifted by a certain amount and sampled on the same wavelength grid, the wavelength difference at each point is, to first order, equal to the flux difference divided by the local gradient. This wavelength difference can then be transformed into velocity by dividing by the wavelength and multiplying by the speed of light. Given the spectral gradient derived from the master spectrum, we modified the flux of each spectrum at each wavelength bin according to the velocity of the injected planetary signal. The simulated signals have DS ranging from 0.1 to 5 m/s, uniformly distributed orbital periods between 10 and 100 days, and random phases. For each modified spectrum, we computed the CCF using the ESPRESSO G2 weighted line mask (Pepe et al. 2002, 2021) to derive the total observed RV, which includes both the stellar and the injected planetary components, in addition to residual instrumental and telluric systematics not fully corrected for by YARARA.

Subsequently, we transformed the normalized flux of the spectra into formation temperature values. Following Al Moulla et al. (2022), we used the precomputed solar formation temperatures presented in that work, which can be extracted from the ARVE code (Al Moulla 2025). These were first associated with the master spectrum and subsequently propagated to each individual observation by applying the same wavelength grid, ensuring that each spectral pixel in the time series was assigned a consistent temperature value (see Eq. (6)). This approach avoids full spectral synthesis, remaining computationally efficient while preserving consistency with the solar photospheric structure. The resulting temperatures were then used to build

the temperature-based shell representations (see Sect. 2.2). For stellar sources other than the Sun, the ARVE code can be used to extract precomputed information for a limited set of spectral types; otherwise, equivalent formation temperatures would need to be derived from dedicated spectral synthesis (as done in Al Moulla et al. 2024).

Following the procedure described in Sect. 2, we computed the shell representation and associated weights for each spectrum, by only using spectral lines that can be properly model by spectral synthesis and that are therefore void of systematics induced by tellurics or detector systematics (see Sect. 4.1). For both flux- and temperature-gradient representations, we discretized the shell space at a resolution of ten points per axis, yielding a 9×9 grid of bins. As a result, we generated two different datasets: one comprising flux-based shells and another built from temperature-based shells, each paired with its corresponding CCF-derived RV and injected DS for supervised learning.

It is worth mentioning that the injected DSs were generated using the gradient of the master spectrum and not an analytical model. The latter would imply modifying the wavelengths associated with each spectral bin, differently for each spectrum, which would then significantly increase the complexity in deriving temperature-based spectra and in building a spectral shell. A simple solution could be to resample all the spectra to a common wavelength grid, but this would add some unwanted systematics for each spectrum due to interpolation. We therefore preferred to use the linear approximation described in Bouchy et al. (2001) as in this case, only the systematics involved in the computation of the master spectrum gradient contaminated the data. As this master spectrum has a much higher S/N than individual spectra, the induced systematics are smaller than those obtained by performing a resampling of each spectrum, providing a stable reference for constructing the shell representation. The use of a fixed master spectrum remains an approximation, since stellar activity induces time-dependent spectral variations that are not explicitly incorporated into the shell construction. We note that all precise RV extraction methods are inherently based on relative measurements. Methods such as CCF, template matching, and line-by-line analyses rely on a synthetic or observed reference template, which in our case corresponds to the master spectrum.

4. Deep-learning framework

We trained supervised CNNs that receive as input a physics-informed shell representation, based on flux or temperature, with 9×9 dimensions. Our neural network was designed to predict two outputs: the total RV estimated from the CCF and the DS manually injected into the spectrum. This dual-output configuration was empirically adopted to include the RV as a quantity that can be independently computed from the spectra using the CCF, enabling a direct comparison between traditional and neural-network predictions, while the DS output provides access to the injected planetary signal, which is not directly available from the CCF. A schematic overview of the model architecture is presented in Fig. 3; details are provided below; and Table 1 summarizes the notation used hereafter.

4.1. Training data

Our analysis was based on 2036 HARPS-N solar spectra with corresponding observation times. We restricted the analysis to

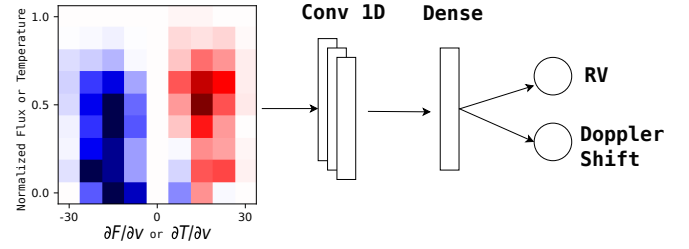


Fig. 3. Schematic of the 1D CNN architecture used to predict RV and DS from shell inputs.

Table 1. Notation used for training, validation, and testing the datasets.

Symbol	Description
$\mathcal{S}$	Full set of HARPS-N solar spectral shells (2036)
$\mathcal{V}$	HO test set (subset of $\mathcal{S}$)
$\mathcal{D}$	Development set, $\mathcal{D} = \mathcal{S} \setminus \mathcal{V}$
$\mathcal{T}$	Training set (subset of $\mathcal{D}$, $\approx 90\%$)
$\mathcal{V}_{\text{int}}$	Internal validation set (subset of $\mathcal{D}$, $\approx 10\%$)
$\mathcal{J}_{\text{train}}$	Injection grid used for training: RV semi-amplitude $\in \{0.1, 0.2, 0.3, 0.5, 1, 2, 5\}$ m/s, $P \in \{20, 40, 60, 80, 100\}$ days, with random phases
$\mathcal{J}_{\text{all}}$	Full injection set used for evaluation: RV semi-amplitude $\in [0.1, 0.45]$ m/s in steps of 0.05, $P \in [10, 100]$ days in steps of 10, plus $P \in [150, 550]$ days in steps of 50, with random phases
M_k	Neural network trained in the k -th CV fold
M	Neural network trained in HO setting
DS	Doppler shift (in m/s)
P	Orbital period (in days)

the wavelength regions selected by the KITCAT line mask from Al Moulla et al. (2022), yielding a total of 31 066 wavelength samples. Those regions were selected because the observed solar spectra agree well with spectral modeling (using pySME Wehrhahn et al. 2023), which is required to derive precise temperature information. In addition, this ensures that the selected spectral lines are not significantly contaminated by tellurics and other instrumental systematics. In these masked spectra, we injected synthetic Keplerian planetary signals with random phases, varying RV semi-amplitudes and orbital periods (P) into the data to augment the training set. We considered two training strategies, whose pseudo-code is detailed in Algorithms 1 and 2.

- Algorithm 1: HO testing. We randomly selected 400 spectra as a fixed test set $\mathcal{V}$, ensuring that these samples are completely unseen during training. The remaining 1636 spectra form the training set $\mathcal{T}$, which is augmented using the injection grid $\mathcal{J}_{\text{train}}$. A single neural network M was then trained on $\mathcal{T}$. For evaluation, we preserved the strict train–test separation by keeping the 400 real test spectra unchanged during training. For each planetary configuration $j \in \mathcal{J}_{\text{all}}$, we created an injected version of the test set $\mathcal{V}_j$ by adding the corresponding synthetic planetary signal. The trained model M was applied to each injected test set $\mathcal{V}_j$ (without retraining), yielding a predicted Doppler-shift time series for that specific configuration. This strategy ensures strict separation between training and evaluation data while enabling an extensive exploration of the injection parameter space.

Algorithm 1 HO testing method.

-
- 1: **Input:** Time series $\mathcal{S}$; training injections $\mathcal{J}_{\text{train}}$; all injections $\mathcal{J}_{\text{all}}$
 - 2: Randomly select 400 spectra as test set $\mathcal{V}$; let $\mathcal{D} = \mathcal{S} \setminus \mathcal{V}$
 - 3: Split $\mathcal{D}$ into training set $\mathcal{T}$ and internal validation set $\mathcal{V}_{\text{int}}$
 - 4: Augment $\mathcal{T}$ with injections from $\mathcal{J}_{\text{train}}$
 - 5: Train a single CNN M on $\mathcal{T}$ (monitoring $\mathcal{V}_{\text{int}}$)
 - 6: **for** each injection $j \in \mathcal{J}_{\text{all}}$ **do**
 - 7: Predict on $\mathcal{V}$ with M (without retraining)
 - 8: **end for**
 - 9: Evaluate detection probability, amplitude, phase, and period errors
-

Algorithm 2 CV method

-
- 1: **Input:** Time series $\mathcal{S}$; injection set $\mathcal{J}_{\text{all}}$; number of folds N_{folds}
 - 2: **for** $k = 1$ to N_{folds} **do**
 - 3: Define evaluation fold $\mathcal{V}_k \subset \mathcal{S}$ with $|\mathcal{V}_k| \approx |\mathcal{S}|/N_{\text{folds}}$
 - 4: Define development set $\mathcal{D}_k = \mathcal{S} \setminus \mathcal{V}_k$
 - 5: Split $\mathcal{D}_k$ into training set $\mathcal{T}_k$ and internal validation set $\mathcal{V}_{k,\text{int}}$
 - 6: Inject signals from $\mathcal{J}_{\text{all}}$ into $\mathcal{T}_k$
 - 7: Train CNN M_k on $\mathcal{T}_k$ (monitoring $\mathcal{V}_{k,\text{int}}$)
 - 8: Predict on $\mathcal{V}_k$ with M_k (without retraining) and save outputs
 - 9: **end for**
 - 10: Concatenate predictions from all folds to recover the full time series
 - 11: Evaluate detection probability, amplitude, phase, and period errors
-

- Algorithm 2: CV training. We implemented an N_{folds} -fold CV strategy, in which the dataset $\mathcal{S}$ is partitioned into N_{folds} complementary subsets. For each fold k , one subset $\mathcal{V}_k$ was used as an unseen evaluation set, while the remaining data defined a development set $\mathcal{D}_k$, further split into training and internal validation subsets. Synthetic planetary signals drawn from the full injection grid $\mathcal{J}_{\text{all}}$ were injected into the training data only. A separate neural-network model M_k was trained for each fold and applied to the corresponding unseen subset $\mathcal{V}_k$, without retraining. This per-fold procedure mirrors the HO strategy in terms of data splitting and synthetic signal injection, with the difference that the role of the evaluation set is rotated across folds.

Detectability in the HO setting was assessed via periodogram analysis of the predicted Doppler-shift time series obtained from the fixed test set of 400 spectra. In the CV setting, after all folds were completed, predictions from the disjoint evaluation subsets were concatenated to reconstruct a time series spanning the full dataset. This ensures that each spectrum was treated as unseen data exactly once, while allowing the final periodogram analysis to exploit all 2036 samples, in contrast to the HO strategy, which relies on a single fixed test set.

In the main analysis, we adopted $N_{\text{folds}} = 5$, yielding evaluation subsets of approximately 400 spectra per fold, comparable in size to the HO test set. For completeness, results obtained with $N_{\text{folds}} = 10$ are presented in Appendix A.2. In both the HO and CV approaches, a wide range of orbital periods, semi-amplitudes, and phases can be explored thanks to the data-augmentation strategy based on synthetic planetary injections. This constitutes a subtle but important difference from Zhao et al. (2024), as our CV approach requires only one

trained model per fold to evaluate multiple injected planetary configurations.

Under both training strategies, for each spectrum, we generated both flux-based and temperature-based shell representations as described in Sect. 2. Radial velocities were computed from the CCF for each spectrum. The CNN input consists of the shell (flux or temperature) multiplied element-wise by its corresponding weight matrix (see Fig. 2), forming a 9×9 feature map. This was done for all augmented spectra, for example, $\approx 57\,200$ elements in a HO test scenario or a single fold of a five-fold CV approach, using ≈ 400 spectral samples for testing. The two target outputs are the CCF-derived RV and the true DS induced by the injected planetary signal, which varies in absolute magnitude between zero and the injected RV semi-amplitude, depending on the observed time of each spectrum.

Although time was used to inject Keplerian signals into the spectra across the original HARPS-N dataset for data augmentation, temporal information was not preserved during training. The spectral shells were randomly selected, and the network was therefore trained on samples corresponding to different injected DSs without explicit temporal ordering. Time information was only reintroduced during testing (see Sect. 4.3), when the predicted DS were associated with the corresponding observation times of the spectral-shell time series and analyzed via periodograms to assess whether the injected planetary signals can be recovered.

The values for the injected RV semi-amplitude and orbital periods were chosen by trial and error, based on how well the neural network performed during training. We find that using a mix of small (e.g., 0.1 m/s) and moderate (1–5 m/s) DS values helped the model learn to recognize both subtle and stronger patterns in the data. Adding very large DS values (e.g., 20 m/s or more) caused the model to focus too much on those signals and made it harder for it to learn the smaller ones. On the other hand, using only very small DS values (0.1 and 0.2 m/s) does not give the model enough variety to learn general patterns. The same idea applies to the choice of orbital periods: we selected them such that several phases of the injected signal are sampled within the full time span of our dataset, ensuring the presence of at least one clear peak in the RV signal. This choice ensures that each injected signal exhibits sufficient temporal variability to provide informative training samples, given the limited temporal sampling of the dataset.

4.2. Neural-network design

We designed two separate neural-network models: one for flux-based shells and another for temperature-based shells. Both models share the same general architecture, and their hyperparameters were optimized independently of each other. All models are implemented using the TensorFlow library (Abadi et al. 2016).

To estimate predictive uncertainty and improve model robustness, we applied the Monte Carlo dropout (MC-DO) technique (Gal & Ghahramani 2016). This method introduces stochasticity during inference by performing multiple forward passes with dropout enabled and then taking the mean prediction. This approach has been shown to provide a Bayesian-like estimate of uncertainty and has been successfully used in other small astrophysical datasets for improving regression performance (Gómez-Vargas et al. 2023b; Mitra et al. 2024).

Each neural network begins with two 1D convolutional layers. Although the shell representation has a 2D layout, its entries are not pixel-based visual features as in classical image

processing. Instead, they contain continuous numerical values with physical meaning, normalized flux, or line-formation temperature gradients, showing correlations primarily along the velocity–gradient axis. Given their small size (9×9) and the structured nature of the data, 1D convolutions offer a more stable and efficient representation than 2D convolutions. Applying 1D convolutions along the rows allows the network to capture local variations across gradient bins while preserving the interpretation of each row as a slice of constant normalized flux or temperature. This configuration also reduces the number of trainable parameters and mitigates overfitting without loss of relevant information, since neighboring gradient values remain correlated due to the smooth variation of normalized flux or temperature with respect to velocity.

To optimize the model architecture and training configuration, we used the Optuna library (Akiba et al. 2019) to employ the Nondominated Sorting Genetic Algorithm II (NSGA-II) (Deb et al. 2000) for exploring the hyperparameter space (Table 2) and minimizing the training loss. This algorithm is well suited for exploring multi-objective optimization problems. It has been mathematically demonstrated that genetic algorithms with elitism, such as NSGA-II, are guaranteed to converge toward the best solution Rudolph (1994). Moreover, they have been successfully applied to astrophysical regression tasks using neural networks (Hetem Jr & Gregorio-Hetem 2007; Gómez-Vargas et al. 2023a; Mitra et al. 2024; Gómez-Vargas & Vázquez 2024).

The hyperparameter search targeted key elements of the architecture: batch size (BS), learning rate (LR), convolutional layer configuration (CLC), and dense layer configuration (DLC). We fixed the number of training epochs at 1000 and employed early stopping with a patience of 40 epochs based on validation loss to prevent overfitting. Learning rates were sampled logarithmically in the 10^{-4} to 10^{-2} range; convolutional layers were explored through combinations of kernel sizes and filter depths; and dense layers followed several compression structures. Dropout was fixed at 0.2, and the SELU (scaled exponential linear unit) activation function (Eq. (11)) was used throughout for its stable behavior during training:

$$\text{SELU}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0. \end{cases} \quad (11)$$

All models were trained using the Adam optimizer and the mean squared error (MSE) loss function, defined in Eq. (12):

$$\text{MSE} = \frac{1}{N} \sum_i \left[(RV_{\text{CCF},i} - RV_{\text{pred},i})^2 + (DS_{\text{inj},i} - DS_{\text{pred},i})^2 \right], \quad (12)$$

where RV_{CCF} denotes the RV measured from the CCF, RV_{pred} is the neural-network prediction, DS_{inj} is the injected DS associated with the synthetic planetary signal, and DS_{pred} is the predicted DS.

Overall, the hyperparameter-optimization procedure involved evaluating a large number of candidate models of order tens to hundreds of configurations when accounting for the combinations in Table 2 and the learning-rate sampling, providing a broad exploration of the hyperparameter space. The best configurations were selected based on validation performance using this loss function. The robustness of the selected models is further supported by the uncertainty and residual analyses presented in Appendix B, which show no evidence of overfitting or unstable behavior.

Table 2. Hyperparameter combinations and best values found for BS, LR, CLC, and DLC.

	Explored values	Flux	Temperature
BS	{128, 256}	256	128
LR	$\log \mathcal{U}(10^{-4}, 10^{-2})$	0.002	0.0002
CLC	{[(128, 3), (256, 3)], [(256, 3), (512, 3)], [(128, 5), (256, 5)], [(256, 5), (512, 5)]}	[(128, 3), (256, 3)]	[(256, 5), (512, 5)]
DLC	{[512], [512, 256], [512, 256, 128], [256], [256, 128], [256, 128, 64]}	[512]	[512]

For both the flux and temperature-based models, the best-performing configurations shared a single dense layer with 512 units. The convolutional layers found for the flux shells were [(128, 3), (256, 3)], and [(256, 5), (512, 5)] for the temperature-based. The optimal batch sizes determined were 256 for the flux model and 128 for the temperature model, with corresponding learning rates of 0.002 and 0.0002, respectively. These configurations consistently yielded the best validation performance and stable convergence across multiple runs. As discussed in Sect. 4.1, although the dataset comprises 2036 independent solar spectra, the training procedure uses 35 synthetic planetary injections, yielding approximately 57 000 augmented training shell samples in each training run both the HO and CV approaches. Each training sample is represented by a 9×9 shell representation matrix. While these realizations are not independent observations, they provide a broad range of planetary amplitudes and phases, substantially increasing the diversity of the training examples. The final architectures selected through the hyperparameter optimization consist of two convolutional layers and one dense layer and contain 759 042 trainable parameters for the flux representation and 931 330 trainable parameters for the temperature representation. Importantly, in neural networks, the raw number of trainable parameters is not, by itself, a sufficient diagnostic of overfitting; generalization may depend on effective capacity measures, regularization, and the training procedure rather than solely on the nominal parameter count (e.g., Bartlett 1996; Ingrassia & Morlini 2005; Belkin et al. 2019; Nakkiran et al. 2021). In our case, dropout and early stopping constrain the effective complexity of the trained models, while the additional tests presented in Appendix A and Appendix B assess their robustness through zero-injection, CV, dominant-peak, chronological HO, injection-recovery, and uncertainty-quantification analyses.

4.3. Model predictions and time-series analysis

Our neural network was trained to detect subtle variations in the shell representation – whether based on flux or temperature – that correspond to a DS induced by planetary signals. This was done independently of time, as no temporal information was encoded in the spectral shells and no time was given during training. Time was only needed to evaluate the performance of the neural network. Indeed, when applied to a time series of spectral shells, our neural network predicts an RV (planets plus stellar activity plus instrumental residuals) and a DS (only planets) time series. Applying a standard periodogram analysis to the predicted RV and DS time series, we can evaluate the performance of the model and its ability to detect planetary signals by performing injection-recovery tests. The predicted RV

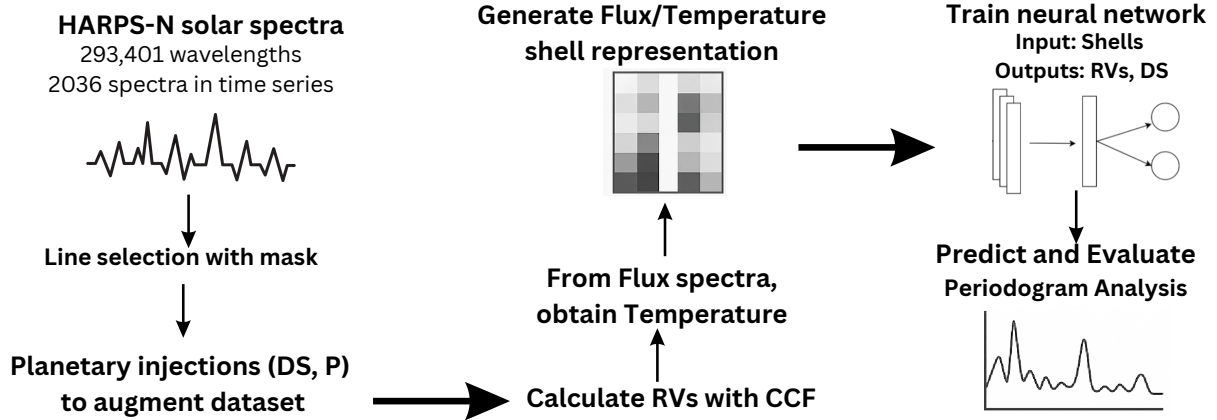


Fig. 4. Schematic workflow of the methodology. HARPS-N solar spectra (293 401 wavelengths, 2036 spectra in a time series) were processed through a line-selection mask to select spectral regions of interest (strong lines with minimum blending). Planetary signals with specified DS and periods (P) are injected to augment the dataset, and their RVs were calculated using the CCF method. Both flux and temperature information were used to generate shell representations, which serve as inputs to a CNN. The trained network predicts RVs and DS, and the predictions were evaluated using periodogram analysis.

time series were used to verify consistency with CCF-derived RVs and, therefore, assess the regression quality of our model. By injecting planetary signals at the level of spectral-shell time series and analyzing the predicted DS time series using a standard periodogram approach, we can assess the ability to recover planetary signals that the traditional CCF-based RVs may not capture due to stellar and instrumental signals.

The shell representations were constructed using a fixed master spectrum. As a result, time-dependent variations in the stellar spectrum induced by activity were not explicitly incorporated into the shell-construction stage. Instead, temporal information entered the framework only through the subsequent periodogram analysis of the predicted DS and RV time series. In particular, we employed the Lomb-Scargle periodogram implementation from the PyAstronomy library (Czesla et al. 2019) to search for planetary signals in the DS time series, and require a detected signal with a false alarm probability (FAP) smaller than 0.1%, as in Zhao et al. (2024). To evaluate the performance of our model, we adopted two strategies: HO testing and CV. In the HO approach (Algorithm 1), a fixed random set of 400 spectra is reserved as an unseen test set, while a single model is trained on the remaining data and evaluated on this held-out subset. In the CV approach (Algorithm 2), the model is trained on the full set of 2036 solar spectra using an N_{folds} -fold scheme, where in each fold a subset of the data is treated as unseen evaluation data and the remaining spectra are used for training with synthetic planetary injections. Considering Algorithms 1 and 2, note that when $N_{\text{folds}} = 5$, the size of the evaluation subset in each fold is approximately 400 spectra; therefore, a single fold of the CV procedure is similar to the HO training.

This framework extends the traditional planetary detection pipeline by incorporating the DS neural network predictions, offering an additional channel of information that may uncover signals the CCF alone cannot resolve. A general summary of the workflow of our analysis is shown in Fig. 4.

4.4. Code availability: *doppleriann*

As part of this work, we developed *doppleriann*², a Python package that implements the spectral-shell representations, computes CCF and extracts RV and CCF-based activity indicators,

² <https://github.com/igomezv/doppleriann>

performs planetary signal injections, applies mask filtering, carries out detectability tests, and integrates the deep-learning framework introduced in this study using the TensorFlow library. The package provides tools for constructing flux-based and temperature-based shell representations, generating synthetic planetary injections, and training CNNs to predict radial velocities and DS. In addition, the package includes the scripts used for the hyperparameter optimization, training, and evaluation of the neural-network models presented in this work.

doppleriann is designed to be modular and extensible, enabling users to adapt the framework to other spectrographs and stellar targets. It includes preprocessing routines for HARPS-N solar spectra, utilities for line selection and temperature assignment, and ready-to-use CNN models consistent with the architectures described in this paper.

The library is publicly available under an open-source license. Documentation and examples are provided to ensure reproducibility and to encourage the application of this methodology to other observational datasets.

5. Results

Using the hyperparameters summarized in Table 2, we defined neural-network architectures for both the flux-based and temperature-based shell representations. Each model was trained and tested under two strategies: HO testing and CV. This results in four distinct model configurations, which we refer to as Flux+HO, Flux+CV, Temp+HO, and Temp+CV, respectively.

For both approaches, each configuration was evaluated over ten independent realizations on $\mathcal{J}_{\text{all}}$. A detection is considered positive if, in at least 70% of the trials (out of ten realizations), the Lomb-Scargle power within $\pm 5\%$ of the injected period exceeds the 0.1% FAP threshold. The top row of Fig. 5 shows the corresponding detection probability. For the remaining metrics, we report the mean values over the successful realizations only, including the mean absolute fractional amplitude difference (in percent), the mean circular phase difference (in cycles), and the mean absolute period difference. Figure 5 summarizes these quantities across the explored parameter space, and each metric is discussed in detail below.

As supplementary material, Appendix A presents additional validation tests of the methodology, including a sanity

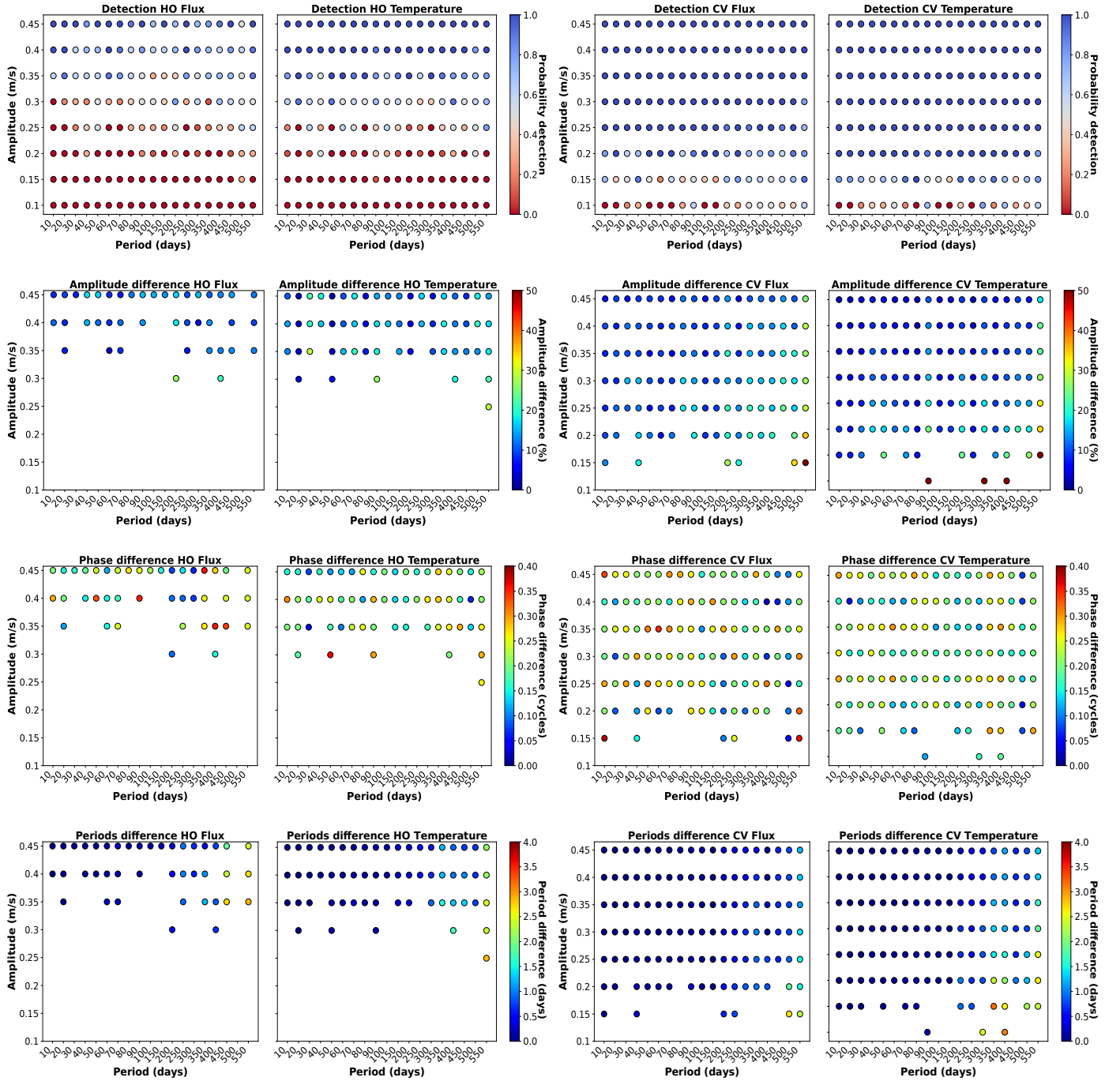


Fig. 5. Detection maps grouped by metric (rows) and model configuration (columns). From top to bottom: detection probability, relative amplitude difference, phase difference, and absolute period difference. From left to right: Flux+HO, Temp+HO, Flux+CV, and Temp+CV. All quantities, except detection probability, correspond to mean values over the successful realizations.

check based on a zero semi-amplitude injection (Appendix A.1), a comparison between five-fold and ten-fold CV strategies (Appendix A.2), an analysis of whether the injected signal is recovered as the dominant peak in the periodogram (Appendix A.3), and a test of the impact of temporal correlations using a chronological HO split (Appendix A.4). Moreover, Appendix A.5 describes specific examples of planetary recovery with small DSs injected at different periods. Appendix B provides additional analysis of the predictive uncertainties of the neural-network models, together with their root mean squared errors (RMSEs) and loss function, offering further support for the robustness of our results.

5.1. Detection probability

The detection probability panels (top row of Fig. 5) show that both HO and CV models successfully recover planetary signals with semi-amplitudes above 40 cm/s across the full range of orbital periods, from 10 to 550 days. At smaller amplitudes, however, differences between the configurations become more evident.

- Hold-out models. The Flux+HO configuration begins to detect injected signals at RV semi-amplitudes of 40 cm/s, whereas this limit is about 5 cm/s better for the Temp+HO

configuration. Overall, the temperature-based approach exhibits consistently higher detection rates.

- Cross-validation models. Overall, CV configurations allows us to detect much smaller planetary signals, with amplitudes as low as 20 cm/s confidently recovered. As in the HO case, using temperature shells as input allows us to detect signals with amplitude about 5 cm/s smaller.

Cross-validation models exhibit higher raw detection sensitivity than their HO counterparts. This is partly a consequence of the evaluation protocol: in the CV approach, planet recovery was assessed over the full time series of 2036 predicted DSs, whereas in the HO case, the evaluation was performed on a fixed set of 400 unseen spectra. As a result, the CV models are able to detect weaker planetary signals, extending sensitivity to lower RV semi-amplitudes. Both evaluation strategies consistently show that temperature-based spectral-shell representations outperform flux-based models in terms of detection sensitivity and recovery rates.

5.2. Amplitude recovery

Amplitude recovery probes the ability of the models to reconstruct the strength of the injected planetary signal and to disentangle it from stellar-induced variability. The amplitude difference is defined as the absolute fractional difference between the recovered and injected Doppler-shift semi-amplitudes, $100|DS_{\text{rec}} - DS_{\text{inj}}|/DS_{\text{inj}}$, averaged over the successful realizations for each injected configuration. The amplitude recovery maps (second row of Fig. 5) highlight systematic biases between flux- and temperature-based models.

- Hold-out models. Both flux- and temperature-based models achieve satisfactory relative amplitude differences, remaining below 30% over most of the explored parameter space. For larger injected RV semi-amplitudes, the amplitude differences decrease further in both cases.
- Cross-validation models. Using the flux-based representation, relative amplitude differences remain below 30% for most injections, except for three cases at RV semi-amplitudes of 15 cm/s and 20 cm/s occurring at long orbital periods (500 and 550 days). In contrast, the temperature-based representation achieves more accurate signal recovery, with relative amplitude differences also remaining below 30% across most of the parameter space. The main exceptions are the 10 cm/s injections and the two cases at 20 cm/s and 25 cm/s at an orbital period of 550 days. These results further confirm the improved robustness of temperature-informed spectral shells, which more effectively capture depth-dependent stellar activity signatures.

In general, the CV models achieve more reliable amplitude recovery than their HO counterparts. Temperature-based representations provide the most stable and accurate reconstruction of the injected Doppler shifts.

5.3. Phase recovery

Phase errors are quantified as the minimum circular distance between the recovered and injected phase offsets, expressed in units of orbital cycles (i.e., fractions of a full period). The phase difference maps (third row of Fig. 5) indicate that all models reproduce orbital phases with generally good fidelity. The values shown correspond to the mean phase difference, averaged over the successful realizations for each configuration.

- Hold-out models. Flux- and temperature-based representations achieve comparable performance, with typical phase

differences remaining below ~ 0.35 cycles, indicating that the recovered phase is usually accurate to better than one third of an orbital period.

- Cross-validation models. Phase differences remain small across most of the explored parameter space, with values typically on the order of ~ 0.3 cycles and systematically lower errors for temperature-based representations, reflecting improved phase stability under CV.

Overall, these results demonstrate that the proposed models preserve the phase information of injected planetary signals with high robustness, even under more conservative CV schemes and that temperature-based inputs provide a modest but consistent advantage in phase recovery accuracy.

5.4. Period recovery

Period errors are expressed as absolute differences between the recovered and injected orbital periods, reported in days. The period recovery maps (bottom row of Fig. 5) show consistently high accuracy across all models.

- Hold-out models. Both flux- and temperature-based representations achieve relative period differences below $\sim 3\%$ over the explored parameter space. The largest discrepancies occur at long orbital periods, particularly at 500 and 550 days, where period recovery becomes more challenging.
- Cross-validation models. Similarly, flux- and temperature-based representations maintain relative period differences below $\sim 3\%$ in most cases. The largest deviations are observed at the longest orbital period (550 days) and for the weakest injected signals, with RV semi-amplitudes of 10–15 cm/s, reflecting the increased difficulty of recovering long-period, low-amplitude signals.

Period recovery remains strong for both training strategies, with comparable levels of accuracy for flux- and temperature-based representations. The performance is slightly worse at longer periods and low signal amplitudes.

5.5. Discussion

Our best neural-network model, based on CV and temperature-based shell (CV-Temp), is capable of reliably recover amplitudes, phases, and orbital periods for planetary signals with amplitude as small as 0.20 m/s and periods ranging from 10 to 550 days. Those result are based on the analysis of 2036 HARPS-N solar spectra and controlled planetary injections. Those detection limits are obtained by analyzing the periodogram of the DS output of our neural network and detecting signals at $\pm 5\%$ of the injected periods with a FAP smaller than 0.1%. As this criterion does not ensure that the injected planetary signal corresponds to the dominant peak in the DS periodogram, we tested this stricter condition (see Appendix A.3). In such a case, the detection limit for the CV and temperature-based shell model only rises by 0.05 m/s to reach 0.25 m/s.

The HO strategy evaluates performance on a fixed set of 400 unseen spectra and therefore provides a conservative assessment of generalization. Despite the reduced evaluation sample, HO models robustly recover signals above ~ 0.35 m/s with accurate amplitude, phase, and period estimates. In contrast, CV models exhibit higher raw detection sensitivity, primarily due to the evaluation protocol, which leverages the full time series of 2036 predicted DSs. More measurements, coupled with an extended temporal baseline, enable the detection of weaker planetary signals at lower RV semi-amplitudes. At the same time, CV models achieve improved recovery of amplitude, phase,

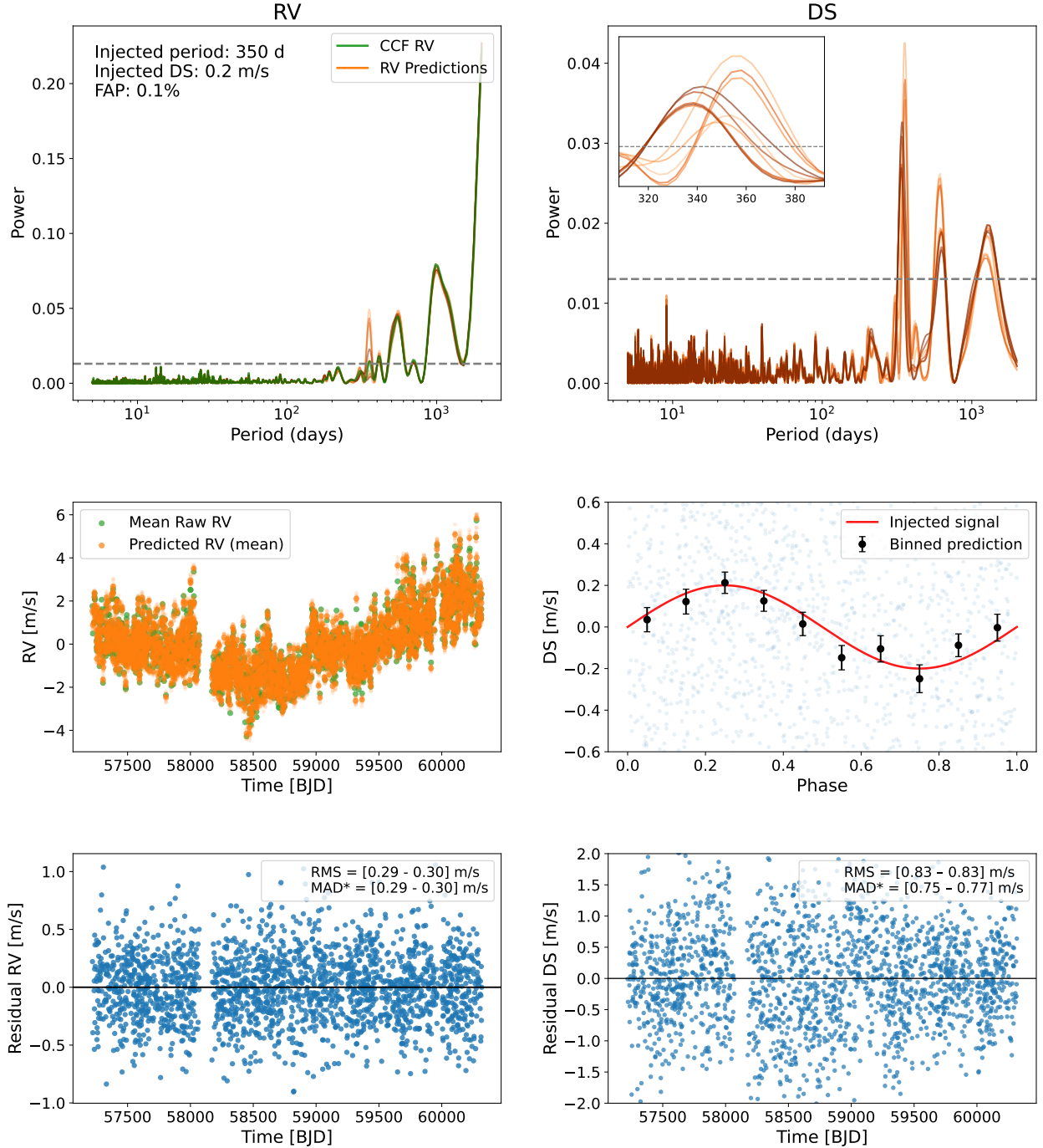


Fig. 6. Example of a 20 cm/s planetary signal recovery using the CV-Temp based model. Top: periodograms of the mean raw CCF RVs (left) and the predicted DSs (right), with the dashed line indicating the 0.1% FAP threshold. The injected period at 350 days is clearly detected in the DS periodogram, while it remains obscured in the RVs. Middle: time series of the mean raw RVs (left) and the phase-folded DS predictions (right), where the black points represent the mean value within the phase bins and the error bars correspond to the standard error of the mean within each bin. For clarity, the y -axis range is restricted in the phase-folded panel; the full dispersion of the predictions is shown in the residual panel. Bottom: corresponding residuals, illustrating that, although individual DS predictions exhibit substantial scatter, the residuals remain centered around zero without a clear systematic structure.

and period across most of the explored parameter space, particularly when using temperature-based representations. The remaining failures are largely confined to long-period and low-amplitude injections, where signal recovery is inherently more challenging.

The superior performance of the neural network model trained with temperature-based shells is consistent with their sensitivity to depth-dependent stellar activity effects. Stellar

line-profile variations are known to depend on formation depth, and encoding formation temperature therefore provides information that is not captured by flux alone. In our experiments, this results in systematically higher detection rates and more stable recovery of amplitudes, phases, and periods when using temperature-based representations. The benefits of this approach are illustrated in Fig. 6, which shows a representative recovery of a low-amplitude planetary signal using the CV-Temp

model. Despite an injected Doppler-shift semi-amplitude of only 20 cm/s and a long orbital period of 350 days, the DS periodogram exhibits its strongest peak at the injected period. This peak is extremely significant with a FAP well below the 0.1% threshold. In contrast, the corresponding RV periodogram is dominated by long-period power and stellar variability, with no evident detection at the injected period. This highlights the advantage of temperature-informed representation in suppressing stellar-induced signals and enhancing planetary detectability. On the other hand, the phase-folded DS time series shows a coherent modulation, with the binned predictions following the injected sinusoidal trend despite significant scatter. Although the residual DS variability remains large, the planetary signal becomes apparent when the data are phase-folded.

Our results complement recent deep-learning approaches applied to solar spectra. Relative to Zhao et al. (2024), we extend the physically motivated spectral representation to include line-formation temperature in addition to flux, increasing the effective training dataset while predicting DSs directly. In contrast, Colwell et al. (2024) focuses on short-period signals, whereas our framework explores a broader range of orbital periods and emphasizes generalization to unseen data using lightweight models, achieving comparable precision over a wider parameter space. These results highlight the potential of combining physically informed representations with DL to improve the detection of low-amplitude planetary signals in high-precision RV surveys.

6. Conclusions

In this work, we presented a deep-learning framework for modeling DS in stellar spectra using physics-motivated spectral-shell representations. By combining data-driven neural networks with advanced hyperparameter optimization, uncertainty quantification, and regularization techniques, and by exploiting both flux and line-formation temperature information, we explored a promising approach for separating planetary Doppler signals from stellar variability in real solar observations.

Using HARPS-N solar spectra with injected planetary signals, we showed that CNNs trained on shell representations recover injected Ds, amplitudes, phases, and orbital periods under controlled injection-recovery experiments. Temperature-based shell representations consistently outperform flux-based shells across all evaluation metrics, highlighting the benefit of encoding depth-dependent stellar information to mitigate activity-induced variability. Under the HO strategy and temperature-based shell, planetary signals with Doppler-shift semi-amplitudes above 0.35 m/s are consistently recovered in our experiments, while CV models demonstrate the confident detection of signals with semi-amplitudes as low as 0.20 m/s. This detection limit is obtained by checking that in the DS output of the neural network, a peak at $\pm 5\%$ of the injected planetary period with a FAP smaller than 0.1% exists in the corresponding periodogram. Using a stricter criterion, in which the planetary signal corresponds to the highest periodogram peak, the detection limit for all periods ranging from 10 to 550 days rises to 0.25 m/s. Such a detection limit thus ensures that the strongest peaks in the DS periodogram correspond to a planetary signal rather than to unrelated systematics.

We note that the lowest detection limits are obtained with the CV strategy and may still be influenced by residual temporal correlations between training and evaluation samples. A more conservative estimate is provided by the chronological HO experiment (Appendix A.4), which preserves temporal

separation between training and test data. The training procedure uses 35 synthetic planetary injections, yielding approximately 57 000 training realizations, all derived from the same underlying solar observations. Therefore, any remaining risk of overfitting is more likely associated with stellar or instrumental patterns than with the injected planetary signals. However, the injection-recovery experiments are limited to circular single-planet systems, and future work should investigate more complex planetary configurations. Nevertheless, a key advantage of the proposed framework is that it operates directly on real stellar data without requiring explicit stellar-activity modeling. The improved performance achieved by temperature-informed representations indicates that physically motivated data compression can substantially enhance the sensitivity of neural-network approaches to low-amplitude planetary signals.

Future work will extend this framework to other stellar types and instruments and explore transfer-learning strategies to enable efficient adaptation across heterogeneous stellar targets and observational conditions. We note that for other stars than the Sun, telluric contamination will be more important as the excursion in barycentric correction for the Sun is minimal, thereby implying a smaller than one pixel shift on the detector between the stellar and telluric features. Although YARARA allows us to strongly mitigate telluric systematics (Cretignier et al. 2021), it is also possible to build spectral shells from selected spectral regions and thus reject regions strongly contaminated by tellurics. Overall, these results highlight the potential of combining DL with physically informed spectral representations to improve the sensitivity of RV analyses, an important step toward detecting Earth-mass planets.

Acknowledgements. IGV and X.D. acknowledge the support from the the Swiss National Science Foundation under the grant SPECTRE (No 200021_215200). K.A. acknowledges support from the Swiss National Science Foundation (SNSF) under the Postdoc Mobility grant P500PT_230225. This work has been carried out within the framework of the NCCR PlanetS supported by the Swiss National Science Foundation under grants 51NF40_182901 and 51NF40_205606. IGV acknowledges current support from the Severo Ochoa Centre of Excellence grant CEX2021-001131-S (MCIN/AEI/10.13039/501100011033), project PID2023-149439NB-C42 from the ‘Proyectos de Generación de Conocimiento’, and the MSCA-COFUND ALLIES-COFUND programme under Horizon Europe (Grant Agreement No. 101126626). This work made significant use of the High-Performance Computing Center of the University of Geneva.

References

- Abadi, M., Barham, P., Chen, J., et al. 2016, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623
- Al Moulla, K. 2025, *A&A*, **701**, A266
- Al Moulla, K., Dumusque, X., Cretignier, M., Zhao, Y., & Valenti, J. 2022, *A&A*, **664**, A34
- Al Moulla, K., Dumusque, X., Figueira, P., et al. 2023, *A&A*, **669**, A39
- Al Moulla, K., Dumusque, X., & Cretignier, M. 2024, *A&A*, **683**, A106
- Artigau, É., Cadieux, C., Cook, N. J., et al. 2022, *AJ*, **164**, 84
- Barragán, O., Aigrain, S., Rajpaul, V. M., & Zicher, N. 2022, *MNRAS*, **509**, 866
- Bartlett, P. 1996, *Adv. Neural Inform. Process. Syst.*, **9**
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. 2019, *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 15849
- Bouchy, F., Pepe, F., & Queloz, D. 2001, *A&A*, **374**, 733
- Burt, J. A., Dumusque, X., & Halverson, S. 2025, arXiv e-prints [arXiv:2511.01954]
- Collier Cameron, A., Ford, E. B., Shahaf, S., et al. 2021, *MNRAS*, **505**, 1699
- Colwell, I., Timmaraju, V., Venkataram, H. S., & Wise, A. 2024, *AJ*, **169**, 24
- Cretignier, M., Dumusque, X., Hara, N., & Pepe, F. 2021, *A&A*, **653**, A43
- Cretignier, M., Dumusque, X., & Pepe, F. 2022, *A&A*, **659**, A68
- Cretignier, M., Dumusque, X., Aigrain, S., & Pepe, F. 2023, *A&A*, **678**, A2

- Czesla, S., Schröter, S., Schneider, C. P., et al. 2019, *Astrophysics Source Code Library* [ascl]
- Dalal, S., Rescigno, F., Cretignier, M., et al. 2024, *MNRAS*, **531**, 4464
- Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. A., & Ford, E. B. 2017, *ApJ*, **846**, 59
- de Beurs, Z. L., Vanderburg, A., Shallue, C. J., et al. 2022, *AJ*, **164**, 49
- Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T. 2000, in *International Conference on Parallel Problem Solving from Nature* (Springer), 849
- Delisle, J. B., Unger, N., Hara, N. C., & Ségransan, D. 2022, *A&A*, **659**, A182
- Dumusque, X. 2018, *A&A*, **620**, A47
- Dumusque, X., Udry, S., Lovis, C., Santos, N. C., & Monteiro, M. J. P. F. G. 2011, *A&A*, **525**, A140
- Dumusque, X., Boisse, I., & Santos, N. C. 2014, *ApJ*, **796**, 132
- Dumusque, X., Glenday, A., Phillips, D. F., et al. 2015, *ApJ*, **814**, L21
- Dumusque, X., Borsa, F., Damasso, M., et al. 2017, *A&A*, **598**, A133
- Dumusque, X., Cretignier, M., Sosnowska, D., et al. 2021, *A&A*, **648**, A103
- Dumusque, X., Al Moulla, K., Cretignier, M., et al. 2026, *A&A*, **706**, A231
- Figueira, P., Faria, J. P., Silva, A. M., et al. 2025, *A&A*, **700**, A174
- Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, *PASP*, **128**, 066001
- Gal, Y., & Ghahramani, Z. 2016, in *International Conference on Machine Learning*, PMLR, 1050
- Gavankar, A., Mittal, T., Ninan, J. P., & Hanasoge, S. 2026, *AJ*, **171**, 33
- Gómez-Vargas, I., & Vázquez, J. A. 2024, *Phys. Rev. D*, **110**, 083518
- Gómez-Vargas, I., Briones Andrade, J., & Vázquez, J. A. 2023a, *Phys. Rev. D*, **107**, 043509
- Gómez-Vargas, I., Medel-Esquivel, R., García-Salcedo, R., & Vázquez, J. A. 2023b, *Eur. Phys. J. C*, **83**, 304
- Hara, N. C., & Ford, E. B. 2023, *Annu. Rev. Stat. Appl.*, **10**, 623
- Hara, N. C., & Delisle, J.-B. 2025, *A&A*, **696**, A141
- Hetem Jr, A., & Gregorio-Hetem, J. 2007, *MNRAS*, **382**, 1707
- Hornik, K., Stinchcombe, M., & White, H. 1990, *Neural Netw.*, **3**, 551
- Ingrassia, S., & Morlini, I. 2005, *Technometrics*, **47**, 297
- Kjærsgaard, R. D., Bello-Arufe, A., Rathcke, A. D., Buchhave, L. A., & Clemmensen, L. K. 2023, *A&A*, **677**, A120
- Korhonen, H., Andersen, J., Piskunov, N., et al. 2015, *MNRAS*, **448**, 3038
- Lakeland, B. S., Naylor, T., Haywood, R. D., et al. 2024, *MNRAS*, **527**, 7681
- Liang, Y., Winn, J. N., & Melchior, P. 2023, *AJ*, **167**, 23
- Lindgren, L., & Dravins, D. 2003, *A&A*, **401**, 1185
- Luhn, J. K., Ford, E. B., Guo, Z., et al. 2023, *AJ*, **165**, 98
- Luhn, J. K., Robertson, P., Halverson, S., et al. 2025, *ApJ*, **987**, 168
- Mayor, M., & Queloz, D. 1995, *Nature*, **378**, 355
- McWilliam, N., De Beurs, Z. L., Vanderburg, A., et al. 2026, *AJ*, **171**, 233
- Meunier, N., Desort, M., & Lagrange, A.-M. 2010, *A&A*, **512**, A39
- Meunier, N., Lagrange, A.-M., Borgniet, S., & Rubiño-Martín, J. A. 2019, *A&A*, **627**, A56
- Mitra, A., Gómez-Vargas, I., & Zarikas, V. 2024, *Phys. Dark Universe*, **46**, 101706
- Nakkiran, P., Kaplun, G., Bansal, Y., et al. 2021, *J. Stat. Mech. Theory Exp.*, **2021**, 124003
- Nieto, L., & Díaz, R. 2023, *A&A*, **677**, A48
- O'Sullivan, N. K., Aigrain, S., Cretignier, M., et al. 2025, *MNRAS*, **541**, 3942
- Pepe, F., Cristiani, S., Rebolo, R., et al. 2021, *A&A*, **645**, A96
- Pepe, F., Mayor, M., Galland, F., et al. 2002, *A&A*, **388**, 632
- Perger, M., Anglada-Escudé, G., Baroch, D., et al. 2023, *A&A*, **672**, A118
- Phillips, D. F., Glenday, A. G., Dumusque, X., et al. 2016, *SPIE Conf. Ser.*, **9912**, 99126Z
- Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, *MNRAS*, **452**, 2269
- Rudolph, G. 1994, *IEEE Trans. Neural Netw.*, **5**, 96
- Saar, S. H., & Donahue, R. A. 1997, *ApJ*, **485**, 319
- Salzer, J., Cisewski-Kehe, J., Ford, E. B., & Zhao, L. L. 2025, *AJ*, **170**, 179
- Siegel, J. C., Rubenzahl, R. A., Halverson, S., & Howard, A. W. 2022, *AJ*, **163**, 260
- Siegel, J. C., Halverson, S., Luhn, J. K., et al. 2024, *AJ*, **168**, 158
- Wehrhahn, A., Piskunov, N., & Ryabchikova, T. 2023, *A&A*, **671**, A171
- Zhao, Y., & Dumusque, X. 2023, *A&A*, **671**, A11
- Zhao, Y., Dumusque, X., Cretignier, M., et al. 2024, *A&A*, **687**, A281
- Zhao, Y., Dumusque, X., Cretignier, M., et al. 2025, *A&A*, **693**, A262

Appendix A: Methodological considerations

This appendix addresses three methodological aspects that deserve further discussion. First, we assess the possibility that the neural-network models could memorize or generate spurious signals in the absence of an injected Doppler-shift signal by means of a zero-signal injection test. Second, we examine whether increasing the number of folds in the CV procedure from five to ten leads to a significant improvement in model performance. In addition, we analyze whether the injected planetary signal is recovered as the dominant peak in the periodogram for our best-performing model. We also investigate the potential impact of temporal correlations in the data by performing a chronological HO test, in which the model is trained on the first 1636 spectra and evaluated on the most recent 400 samples. Finally, we discuss the behavior of the predictions for an injected signal of 20 cm/s in different orbital periods. These three additional analyses provide additional validation of the robustness and reliability of the proposed methodology.

Appendix A.1. The zero signal injection

To verify that the pipeline does not produce spurious detections in the absence of a planetary signal, we perform a zero-injection test on the CV-Temp configuration, in which the injected Doppler semi-amplitude is set to $DS = 0.0$ m/s. This test provides a baseline validation of the model behavior and ensures that detected signals arise from genuine injected variability rather than from artifacts of the neural network or the analysis procedure.

Figure A.1 shows the corresponding periodograms for this case. The RV periodogram computed from the CCF and the neural-network RV prediction both exhibit only long-term variability, with no statistically significant periodic power. Similarly, the DS periodogram displays no peaks exceeding the 0.1% false-alarm probability threshold. The absence of significant peaks across all representations confirms that the pipeline does not introduce false positives and behaves consistently when no planetary signal is present. We emphasize that this experiment should be interpreted as a consistency check of the pipeline rather than as a statistical estimate of the false-alarm rate, since the zero-injection case does not generate multiple independent realizations of the underlying dataset.

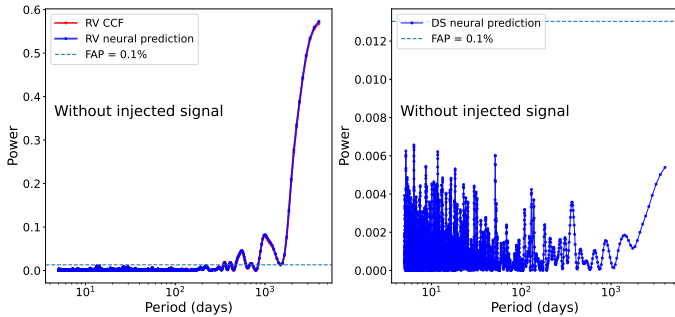


Fig. A.1. Case without injected planetary signal ($DS = 0.0$ m/s). Left: RV periodogram from the CCF (red) and neural prediction (blue), both showing only long-term variability without significant periodic power. Right: Periodogram of the DS prediction, confirming the absence of false planetary detections.

Appendix A.2. Effect of the number of folds in the cross-validation approach

Given that the HARPS-N solar dataset used in this work contains 2036 spectra, we consider a test sample of 400 spectra, corresponding to approximately 25% of the dataset, to be sufficiently large for a statistically meaningful evaluation. Accordingly, in the HO approach, we select this number of spectra for testing, while the remaining data are used for model development. To ensure consistency between the HO and CV strategies, we adopt $N_{\text{folds}} = 5$ in the CV procedure, such that each evaluation fold contains approximately $2036/5 \approx 400$ spectra, making a single CV fold directly comparable to the HO test set.

In principle, increasing the number of folds in the CV procedure results in smaller evaluation subsets and larger training sets, which could intuitively lead to improved predictive accuracy. However, this comes at the expense of increased computational requirements, since the number of independently trained neural-network models scales linearly with N_{folds} . For example, adopting $N_{\text{folds}} = 10$ requires training twice as many models as in the five-fold configuration. Such an increase would be justified only if it led to a substantial improvement in performance. As shown in Fig. A.2, however, the differences between the five-fold and ten-fold configurations are minimal. For injected amplitudes above approximately 15 cm s^{-1} , the ten-fold configuration yields slightly improved amplitude recovery in a limited fraction of cases (of order 5%); this improvement is neither systematic nor sufficient to justify the increase in computational cost.

Appendix A.3. Dominant peak analysis

Planetary recovery tests are typically performed by searching for signals around the injected period. In this work, we consider a signal to be recovered if it is detected within $\pm 5\%$ of the injected period and with a FAP smaller than 0.1%. However, in real observational scenarios, the true period is unknown, and detections must rely on identifying significant peaks in the full periodogram.

To better reflect this situation, we analyze the cases in which our neural network models trained within the CV strategy identify the injected signal as the dominant peak in the periodogram across ten independent realizations, following the same injection setup used in Fig. 5. As shown in Fig. A.3, using this stricter criterion only slightly worsen the detection limits. For the best model which is based on CV and temperature-based shells, the detection limits slightly goes up from 0.20 to 0.25 cm/s. For the CV and flux shell model, the detection limit rises slightly more, going from 0.25 to 0.35 cm/s. This test demonstrate, that when using the CV and temperature-based shell model, planetary signals with amplitude as small as 25 cm/s will correspond to the dominant peak in the analysis of the DS periodogram. We note that even when the injected signal does not correspond to the dominant peak, detections above the false-alarm probability threshold remain informative, as they can be further investigated as candidate signals using complementary analysis methods.

Appendix A.4. Chronological hold-out test

The training and validation sets adopted throughout this work are constructed using random splits, following standard machine-learning practices. However, because solar spectra samples obtained on nearby dates can exhibit temporal correlations induced by stellar activity, a random split may contain observations in the training set that are temporally close to the test

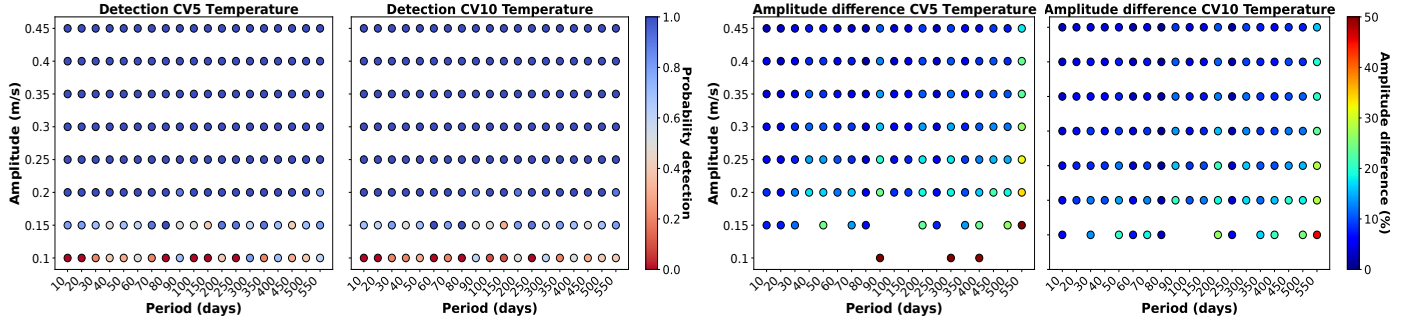


Fig. A.2. Comparison between five-fold and ten-fold CV results for the temperature-shell CNN model. Left: Detection probability as a function of injected orbital period and RV semi-amplitude for $N_{\text{folds}} = 5$ (CV) and $N_{\text{folds}} = 10$ (CV10), respectively. Right: Corresponding relative amplitude differences between injected and recovered signals. Overall, increasing the number of folds from five to ten does not lead to a substantial improvement in detection performance or amplitude recovery, particularly above $\sim 15 \text{ cm s}^{-1}$, while incurring a higher computational cost.

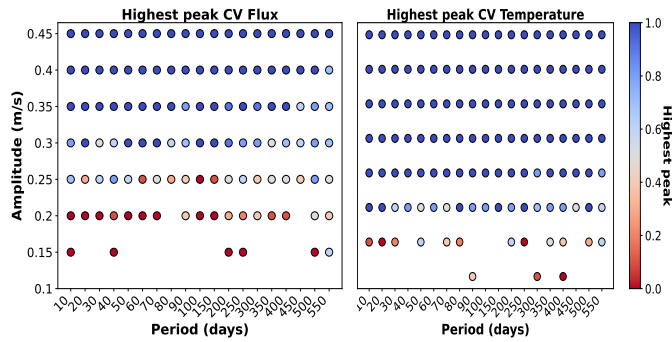


Fig. A.3. Neural-network models using the CV strategy. Here, we assess whether the detected planetary signal corresponds to the dominant peak in the periodogram across ten independent realizations with different random phase injections, for both flux- and temperature-based representations. The color bar indicates the fraction of realizations in which the injected signal is recovered as the dominant peak: a value of 1 (dark blue) corresponds to ten total successful recoveries, while zero (dark red) indicates that the injected signal is not recovered as the dominant peak in any realization.

samples. To evaluate whether the neural network framework discussed in this manuscript could be affected by such correlations, we performed additional chronological held-out stress tests inspired by the temporally grouped validation strategy presented in [Salzer et al. \(2025\)](#).

In this chronological tests, the spectra samples are split chronologically rather than randomly, allowing that training and test sets are fully separated in time. Considering that the neural network generates predictions independently for each spectrum and does not explicitly use temporal information as an input feature, this test evaluate the ability of the model to generalize across different observing periods. They also reduce the possibility that temporally nearby spectra in the training set artificially improve the evaluation performance. To enable a direct comparison with the standard HO strategy presented in Sect. 5, we preserve the same number of training and evaluation samples, using 1636 spectra for training and 400 for testing.

We train our neural network framework on the most recent 1636 spectra and evaluate it on the oldest 400 observations. We selected the oldest 400 spectra for testing as they correspond to a level of activity that exists in the training set; however, they are separated by half the solar cycle periodicity (~ 5 – 6 years), as the first 400 measurements correspond to the middle of the

decreasing phase of solar cycle 24, while the training set includes the middle of the rising phase of solar cycle 25. Figures A.4 and A.5 show the corresponding periodograms and detection maps when testing our trained model on the oldest 400 spectra. Those results are very similar to the ones obtained when the samples are randomly shuffled, with a robust detection limit in the HO experiments based on the temperature shell at a level of 0.35 m/s . However, a small difference is observed for periods equal to or greater than 400 days, where we see that for planetary signals of semi-amplitude 0.35 m/s , the detection rate between the ten different phase evaluations drops, and the relative amplitude difference between the injected and recovered signal increases. This comes from the fact that the testing set is now chronological in time, and thus the 400 spectra selected correspond to a time span of only 600 days. The signal of planets with long periods is therefore not well sampled in phase, inducing the drop in detection sensitivity.

This test demonstrates that our neural network framework retains the ability to recover injected planetary signals even under a strict temporal split between the training and test sets, and thus temporal correlation between nearby spectra when selecting training and testing sets randomly is not inducing overfitting. The prediction of our neural network framework seems therefore, to generalize well across different observing periods.

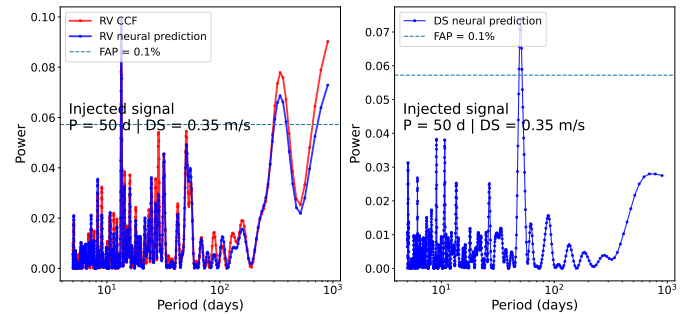


Fig. A.4. RV (left) and DS (right) periodograms generated using a neural network optimized under the chronological HO strategy, trained on the most recent 1636 spectra and tested on the oldest 400. The 0.35 m/s injected planet at 50 days is highly significant in the DS periodogram.

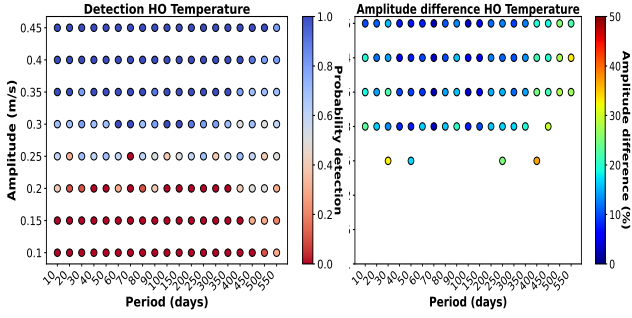


Fig. A.5. Detection probability for the temperature-based model under the chronological HO strategy, trained on the most recent 1636 spectra and tested on the oldest 400. Left: Fraction of successful recoveries over ten realizations. Right: Relative amplitude differences for the detected signals.

Appendix A.5. Planetary signal recovery at different orbital periods for a 20 cm/s signal

To illustrate the behavior of our best-performing neural-network model (CV-Temp), we analyze the recovery of low-amplitude planetary signals at different orbital periods. In analogy with the case shown in Fig. 6, which corresponds to an injected semi-amplitude of 0.2 m/s at 350 days, we perform periodogram analyses of both RV and DS predictions for injections at 10, 50, and 100 days. For each period, the analysis is repeated over ten independent evaluation sets, in accordance with the analysis discussed in Section 5 and with the Figure 5.

Figure A.6 compares the resulting RV and DS periodograms. The second column, corresponding to the DS predictions, shows that for shorter orbital periods, the recovered signals appear as narrower and more sharply defined peaks in the periodogram. This behavior is particularly evident for the 10-day injection. As the orbital period increases, the detected power becomes more dispersed, as seen in the 100-day case, although the spread remains smaller than for the 350-day injection shown in Fig. 6.

In all three cases, the injected signal is successfully detected in all ten total trials using the DS predictions. In contrast, the RV periodograms shown in the first column do not exhibit significant power at the injected periods, neither for the traditional CCF-derived RVs nor for the RVs predicted by the neural network. Even when the FAP threshold is formally satisfied, the peaks associated with the injected signal are overshadowed by stronger power at longer periods, likely driven by magnetic activity.

Overall, in these examples, we can highlight the improved sensitivity of the DS-based analysis for recovering low-amplitude planetary signals across a range of orbital periods.

Appendix B: Model robustness and uncertainty quantification

Figure B.1 shows representative training and validation loss curves for a temperature-based HO model corresponding to the best-performing configuration identified in Section 4.2. The loss decreases smoothly from values of order 10^{-1} at early epochs to a few $\times 10^{-3}$ at convergence, with both curves closely tracking each other throughout training. This behavior indicates stable convergence without signs of overfitting. Similar trends are observed for flux-based models and across the CV folds, providing additional confidence in the robustness of the training procedure.

In addition, we estimate predictive uncertainties using MC-DO, which enables stochastic inference and provides an estimate

of the epistemic uncertainty of the neural-network models. These uncertainties should be interpreted as a measure of model dispersion rather than as calibrated predictive intervals, as would be obtained with methods such as conformal prediction, which we leave for future work. All results reported in this manuscript correspond to the mean prediction over $N = 100$ stochastic forward passes, with uncertainties defined as the standard deviation across these realizations.

For each configuration, we compute the RMSE between predicted and true values for both RV and DS, together with the mean predictive uncertainty across unseen samples. Figure B.2 summarizes these quantities for both the HO and CV approaches. The RMSE values and predictive uncertainties are similar between the HO and CV strategies (Table B.1), indicating consistent predictive behavior across the two validation approaches. The predictive uncertainties, estimated through MC dropout, quantify the variability of the predictions across stochastic forward passes and therefore should not be interpreted as calibrated estimates of the total prediction error.

In all configurations, DS predictions exhibit larger errors and predictive uncertainties than RVs. We emphasize that the neural-network architecture is explicitly designed to model the DS signal rather than the CCF-derived RV. The RV predictions are included primarily as a reference quantity, directly comparable to standard spectroscopic measurements. This distinction reflects the different nature of the two targets: while the RV is directly derived from the observed spectra via the CCF, the DS corresponds to a synthetic signal injected at the spectral level, which may introduce additional sources of noise and mismatch. In this context, the improved performance of temperature-based representations in recovering the injected DS signal indicates that these representations are more sensitive to planetary Ds encoded at the spectral level, whereas flux-based representations are more closely tied to the information content of the CCF and thus to the measured RVs.

Temperature-based shell representations consistently yield lower predictive uncertainties than flux-based shells for both HO and CV strategies, while exhibiting larger residuals in RV and lower residuals in DS (Table B.1). This behavior can be understood in terms of a bias-variance trade-off, given that flux-based models more directly capture the information used to compute the CCF radial velocities, leading to lower RV residuals, whereas temperature-based representations involve an additional transformation to derive temperature that can introduce interpolation-related errors, increasing the bias in RV predictions. At the same time, temperature-based models exhibit significantly lower predictive dispersion in the MC dropout estimates, indicating reduced model variance.

The apparent discrepancy between residuals and predictive uncertainties should therefore not be interpreted as an inconsistency or error in the uncertainty estimation. The RMSE quantifies the total prediction error with respect to the target, while MC dropout provides an estimate of epistemic uncertainty, i.e., the variability of the model predictions under stochastic forward passes. As a result, these two quantities capture different aspects of model performance and are not expected to coincide quantitatively.

We note that the CV results correspond to an average over multiple models, which leads to slightly higher RMSE values compared to the HO case due to the aggregation of independent predictions. For this reason, direct quantitative comparison between HO and CV is not straightforward, although the relative trends between flux and temperature representations remain consistent across both strategies.

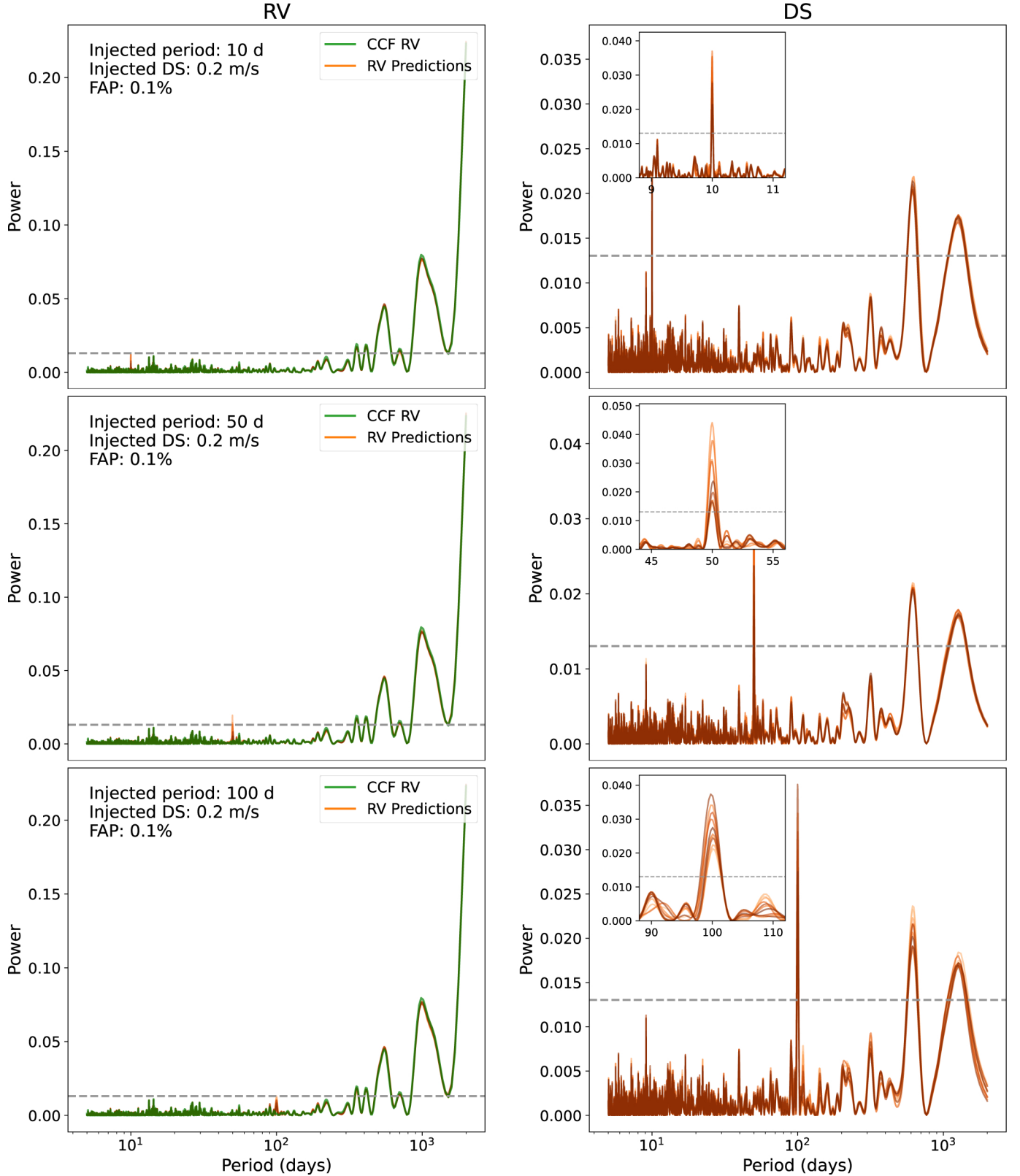


Fig. A.6. Recovery of a 0.2 m/s planetary signal at different orbital periods using the CV-Temp model. Each row corresponds to an injected period of 10, 50, and 100 days, respectively. Left: Periodograms of the traditional CCF RVs (green) and the neural-network-predicted RVs (orange). Right: Periodograms of the predicted DS. The dashed line indicates the 0.1% FAP threshold, and the insets highlight the region around the injected period for the ten independent evaluation sets. The planetary signal is consistently recovered in the DS periodograms for all tested periods, while remaining undetected in the RV-based analyses.

Despite the scatter in individual DS predictions, the periodic structure of the injected signal is preserved across the time series, enabling successful recovery through periodogram-based analyses, as discussed in Section 5.

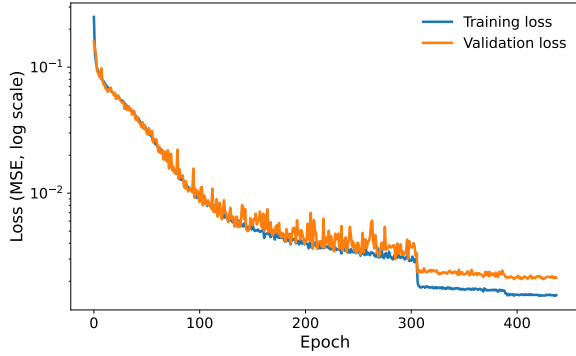


Fig. B.1. Training and validation loss function for an HO CNN model using temperature shells. Both curves decrease steadily and remain close throughout training, with no divergence at later epochs, indicating stable training and no sign of overfitting.

Table B.1. Ranges of RMSE (residuals) and MC dropout predictive uncertainties for RV and DS across HO and CV strategies.

	RV		DS	
	Flux	Temp	Flux	Temp
RMSE (HO)	0.204–0.215	0.294–0.302	0.815–0.824	0.7808–0.7844
Unc. (HO)	0.0926–0.0936	0.042–0.0432	0.1006–0.1023	0.0344–0.0352
RMSE (CV)	0.192–0.1952	0.2886–0.2904	0.894–0.898	0.8244–0.8258
Unc. (CV)	0.0876–0.0888	0.045–0.046	0.093–0.0937	0.036–0.0369

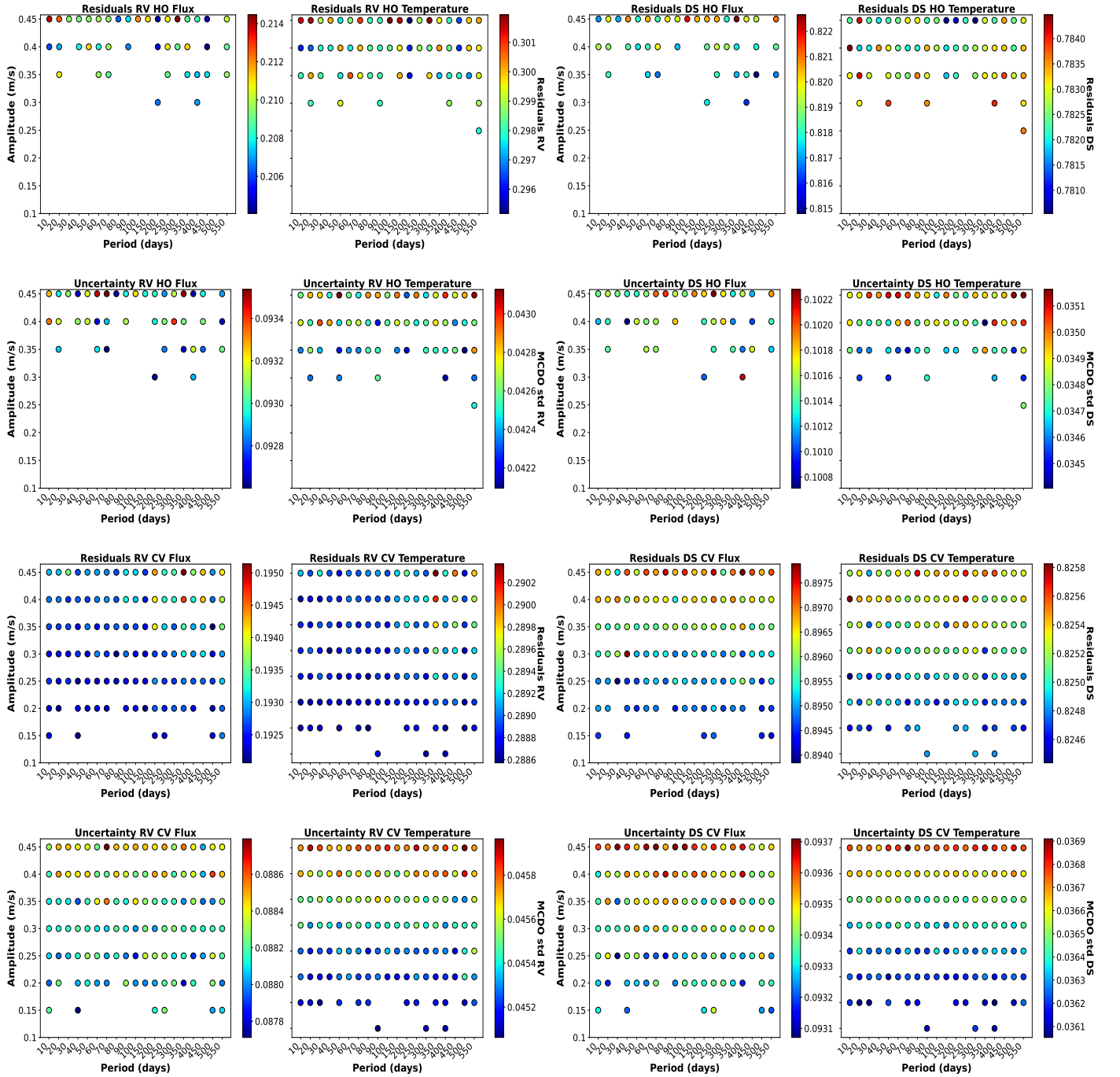


Fig. B.2. RMSEs (residuals) and predictive uncertainties for RV and DS obtained using the HO and CV approaches across the ten different test sets. First row: HO model showing the RMSE between the predicted and true values. Second row: HO model showing the mean predictive uncertainties, defined as the standard deviation over 100 MC-DO. Third row: CV model showing the RMSE obtained with the five-fold CV approach. Fourth row: CV model showing the corresponding mean predictive uncertainties.