Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **529,** 1688–1714 (2024) https://doi.org/10.1093/mnras/stae421
Advance Access publication 2024 February 9

`cecilia`: a machine learning-based pipeline for measuring metal abundances of helium-rich polluted white dwarfs

Mariona Badenas-Agusti [1,2★†] Javier Viaña [2] Andrew Vanderburg [2] Simon Blouin [3] Patrick Dufour [4] Siyi Xu (许偲艺) [5] and Lizhou Sha [6]

[1]*Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2]*Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[3]*Department of Physics and Astronomy, University of Victoria, Victoria, BC V8W 2Y2, Canada*
[4]*Département de Physique, Université de Montréal, Montréal, Québec H3C 3J7, Canada*
[5]*Gemini Observatory/NSF's NOIRLab, 670 N. A'ohoku Place, Hilo, HI 96720, USA*
[6]*Department of Astronomy, University of Wisconsin-Madison, 475 N Charter St, Madison, WI 53706, USA*

Accepted 2024 February 7. Received 2024 February 6; in original form 2023 December 19

ABSTRACT

Over the past several decades, conventional spectral analysis techniques of polluted white dwarfs have become powerful tools to learn about the geology and chemistry of extrasolar bodies. Despite their proven capabilities and extensive legacy of scientific discoveries, these techniques are, however, still limited by their manual, time-intensive, and iterative nature. As a result, they are susceptible to human errors and are difficult to scale up to population-wide studies of metal pollution. This paper seeks to address this problem by presenting `cecilia`, the first machine learning (ML)-powered spectral modelling code designed to measure the metal abundances of intermediate-temperature ($10\,000 \leq T_{\rm eff} \leq 20\,000$ K), Helium-rich polluted white dwarfs. Trained with more than 22 000 randomly drawn atmosphere models and stellar parameters, our pipeline aims to overcome the limitations of classical methods by replacing the generation of synthetic spectra from computationally expensive codes and uniformly spaced model grids, with a fast, automated, and efficient neural-network-based interpolator. More specifically, `cecilia` combines state-of-the-art atmosphere models, powerful artificial intelligence tools, and robust statistical techniques to rapidly generate synthetic spectra of polluted white dwarfs in high-dimensional space, and enable accurate ($\lesssim$0.1 dex) and simultaneous measurements of 14 stellar parameters – including 11 elemental abundances – from real spectroscopic observations. As massively multiplexed astronomical surveys begin scientific operations, `cecilia`'s performance has the potential to unlock large-scale studies of extrasolar geochemistry and propel the field of white dwarf science into the era of Big Data. In doing so, we aspire to uncover new statistical insights that were previously impractical with traditional white dwarf characterization techniques.

Key words: methods: data analysis – methods: observational – techniques: spectroscopic – stars: abundances – stars: atmospheres – stars: white dwarfs.

1 INTRODUCTION

In recent years, the field of exoplanets has undergone remarkable growth, revolutionizing our understanding of the Universe with the discovery of more than 5500 distant worlds.[1] Despite this notable progress, the detection of a potentially habitable Earth-sized planet with an Earth-like interior remains elusive.

In our quest to find exoplanets similar to our own, knowledge of a planet's bulk composition will be crucial to understand its formation and evolution history, geology, climate, and habitability conditions. Traditionally, transit photometry (Charbonneau et al. 2000) and Doppler spectroscopy (Campbell, Walker & Yang 1988) have been used to measure the radii and masses of exoplanets, respectively. These two parameters, when combined, provide the planetary bulk density, which can then be compared to interior models to infer the planet's most likely composition (Zeng, Sasselov & Jacobsen 2016). Nevertheless, this approach is incomplete because it does not account for the specific elemental abundances of the planet. In addition, for planets smaller than $R_p \sim 4\,R_\oplus$, large uncertainties in radius and mass measurements (>10 per cent) pose challenges in distinguishing between different interior compositions (Zeng & Seager 2008). These problems are further complicated by degeneracies in theoretical mass-radius diagrams, which can yield various combinations of metal, silicate, water/ice, and gas for the same density measurement (Seager et al. 2007; Rogers & Seager 2010; Dorn et al. 2015).

★ E-mail: mbadenas@mit.edu
† MIT William Asbjornsen Albert Memorial Fellow.
[1]NASA Exoplanet Data base (https://exoplanets.nasa.gov/). Last Accessed: January 2024.

© 2024 The Author(s).
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

To date, the most direct and only way of overcoming these ambiguities and probing the composition of small, terrestrial exoplanets is through spectroscopy of externally polluted white dwarfs (WDs). These objects are the degenerate stellar cores of low- and intermediate-mass ($\leq$10 $M_\odot$) main-sequence stars. With a typical radius similar to the size of the Earth ($R_{WD} \approx 1\ R_\oplus$), and a mass half of that of the Sun ($M_{WD} \approx 0.6\ M_\odot$), they are extremely dense and compact, which causes their atmospheres to have a stratified chemical structure. In other words, light elements (i.e. H and He) are expected to float in the thin upper layers of their atmospheres, while elements heavier than helium (or 'metals', e.g. Ca, Mg, Si, Fe) should sink rapidly towards the core with diffusion time-scales much shorter than the cooling age of the WD (Paquette et al. 1986; Koester 2009). In reality, however, observations suggest that between 25 per cent and 50 per cent of single WDs are contaminated with heavy elements in their atmospheres (Zuckerman et al. 2003, 2010; Koester, Gänsicke & Farihi 2014), likely from the recent or ongoing accretion of disrupted exoplanetary debris from their surviving outer planetary systems (Jura 2003; Zuckerman et al. 2003, 2010; Jura & Young 2014; Koester, Gänsicke & Farihi 2014).

During the past two decades, high-resolution spectroscopy of these 'polluted' WDs has offered an excellent opportunity to determine the bulk composition of their accreted material. This technique has enabled the detection of more than 20 heavy elements (e.g. see references in table 1 of Klein et al. 2021), revealing that most exoplanetary bodies are dry and rocky (Veras 2021), with some water-rich exceptions (e.g. Farihi et al. 2011; Farihi, Gaensicke & Koester 2013; Raddi et al. 2015; Xu et al. 2017; Hoskin et al. 2020). Despite the significance of these discoveries, conventional analysis techniques of polluted WDs entail time-consuming, manual, and iterative work, which makes them prone to human errors. As a result, these techniques may lead to wrong or sub-optimal results, potentially biasing our understanding of the geology and chemistry of WD pollutants. With these challenges in mind, it has been difficult to transition from individual to population-wide studies of metal pollution. As we move into an era of massive multi-object spectroscopic surveys, when tens of thousands of WDs will be observed and many of them are expected to show multiple heavy elements, conventional analysis techniques will thus struggle to keep up.

In this paper, we address the limitations of 'classical' WD characterization methods by leveraging the power of machine learning (ML). Recently, multiple ML-based tools have been developed to interpolate and efficiently model spectra with the goal of estimating stellar abundances. Notable examples include `The Cannon` (aimed at deriving stellar properties based on a large training set of reference stars with well-known abundances; Ness et al. 2015), `The Payne` (designed to estimate 25 elemental abundances of red giants from low-resolution APOGEE spectra; Ting et al. 2019), `wdtools` (dedicated to the analysis of H-rich WD spectra with no metal pollution; Chandra et al. 2020), and `TheHotPayne`, which is an adaptation of `The Payne` to OBA stars (Xiang et al. 2022). These ML pipelines have undoubtedly shown promising results, but they have not been designed to fit the complex and diverse spectra of polluted WDs. This paper seeks to bridge this gap by presenting the first ML-based code capable of achieving this goal.

Inspired by the groundbreaking contributions of astrophysicist Cecilia Payne and building upon the ML framework of the `The Payne`, we have chosen to name our code '`cecilia`' – or **C***lues of* **E***xoplanet* **C***ompositions* **I***nferred from po***L***luted Wh***I***t***E*** *dw***A***rfs*. With `cecilia`, we combine state-of-the-art atmosphere models with innovative ML methods to improve the performance of conventional WD analysis techniques and ultimately enable large-scale studies of metal pollution. From a design standpoint, `cecilia`'s ML architecture contributes to this goal by adhering to three fundamental principles: automation, efficiency, and flexibility. As an automated pipeline, `cecilia` solves the dependence of current techniques on human-based, time-intensive, and iterative procedures. In terms of efficiency, our code harnesses the speed of ML tools to rapidly estimate elemental abundances of polluted WDs with a retrieval accuracy similar to that of traditional methods. Lastly, `cecilia` has a flexible ML architecture that can be re-trained at any time with new, better, and/or larger models. This also makes our pipeline a particularly useful feature extraction tool beyond the fields of WD and exoplanetary science.

In this work, we describe the building blocks of `cecilia`, its current performance and limitations, and potential opportunities for improvement. The organization of the paper is as follows. In Section 2, we provide a brief introduction to the the concept of deep neural networks (NNs). In Section 3, we present the stellar parameters and model atmospheres used to train, validate, and test our code. Section 4 explains `cecilia`'s ML architecture, our training procedure, and our framework for estimating the elemental abundances of polluted WDs from their spectra. In Section 5, we test and explore `cecilia`'s functionality with synthetic and real spectroscopic observations. Section 6 discusses the strengths and weaknesses of `cecilia`, possible future work to enhance its retrieval accuracy, and its relevance in the context of massive data sets from wide-field astronomical surveys. Finally, we offer our conclusions in Section 7.

2 DEEP NEURAL NETWORKS

Since the 1950s (Turing 1950; McCarthy et al. 2006), the field of modern artificial intelligence (AI) has sought to create intelligent systems capable of emulating human skills, both in terms of human reasoning and behaviour. In recent years, the rapid development of technology has transformed this theoretical aspiration into a tangible reality, causing an important shift from model- to data-driven scientific analysis. The creation of faster and more powerful computers, coupled with the availability of extremely large and complex data sets (or 'Big Data'), has only fueled the growth of AI, making NNs increasingly relevant across a variety of disciplines, including astrophysics (Smith & Geach 2023).

To date, the task of developing computer programs that can learn, extract, and process meaningful insights from observations is typically referred to as ML (Samuel 1959). At the core of many ML algorithms are multilayered, fully connected, feed-forward NN. These networks are learning systems where information flows *forward* through several layers of artificial nodes called 'neurons'. A basic NN consists of an input layer, one or more hidden layers, and an output layer. The higher the number of hidden layers (or 'depth'), the more complex the ML architecture becomes (LeCun, Bengio & Hinton 2015). This is the case of `cecilia`, which consists of three NNs with 5–6 layers each. In general, deeper architectures are more capable of recognizing more intricate relationships in the data (Goodfellow, Bengio & Courville 2016).

In addition to the choice of ML architecture, the learning capabilities of a NN are also shaped by the way its neurons are interconnected with one another. This connection is regulated by three parameters known as 'weights', 'biases', and 'activation functions'. The weights and biases determine the importance of the neuronal connections, while the activation functions employ non-linear operations to transform the input data of a set of neurons into useful output

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

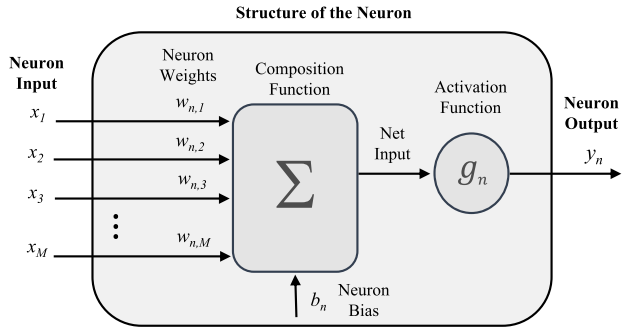


Figure 1. Parameters and operations associated to a single 'neuron', which represents the smallest unit of a NN.

parameters. In particular, for a generic neuron n, the activation function g_n performs the transformation

$$y_n = g_n\left(b_n + \sum_{m=1}^{M} w_{n,m} \cdot x_m\right), \tag{1}$$

where M denotes the total number of inputs fed into the neuron, m is a generic input index, x and y are the input and the output data, and $w_{n,m}$ and b_n represent the optimized weight and bias of the neuron, respectively (see Fig. 1). Typically, the activation function g_n is chosen by the user experimentally (e.g. see Section 4.1).

In the field of ML, the *training* of a NN is an iterative procedure designed to fine-tune the weights and biases of the neurons with the goal of minimizing the error (also known as 'loss' or 'cost' function) between the network's final predictions and an assumed ground truth. During this process, the NN is exposed to labeled observations in order to identify hidden patterns in the data and learn to generate predictions that closely resemble the expected outcome for a given set of input parameters. As the training proceeds, the predictive accuracy of the NN is gradually improved with a technique known as 'backpropagation'. This technique calculates the loss function at each iteration to adjust the model parameters of the network. A standard formulation of the loss function is the quadratic expression

$$\mathcal{J}_q = \frac{1}{2}\sum_{p=1}^{P}(y_{q,p} - \hat{y}_{q,p})^2 = \frac{1}{2}\sum_{p=1}^{P} e_{q,p}^2, \tag{2}$$

where q is the index of a generic instance,[2] P is the total number of output neurons, p is the index of a generic output neuron, $\hat{y}_{q,p}$ is the prediction of the p-th output neuron, $y_{q,p}$ is the known true output, and e denotes the residual error difference between $\hat{y}$ and y. For a given training data set, the aggregated loss function of all the instances Q can then be expressed as

$$\mathcal{J} = \sum_{q=1}^{Q} \mathcal{J}_q. \tag{3}$$

During training, the aggregated loss function is computed at every step in order to optimize the weights and biases of the network. These parameters are updated iteratively through the gradient of the loss function, $\nabla\mathcal{J}(w_{n,m})$. More specifically, the incremental change of a generic weight in the network, or $\Delta w_{n,m}$, is proportional to the derivative of the total loss function with respect to the current value of the weight,

$$\Delta w_{n,m} = w_{n,m}^{\rm new} - w_{n,m}^{\rm old} = -\eta\frac{\partial\mathcal{J}}{\partial w_{n,m}}. \tag{4}$$

In this expression, the partial derivative $\partial\mathcal{J}/\partial w_{n,m}$ dictates how much the network parameters should be changed in order to reach a minimum of the loss function, with the negative sign indicating that the direction of the 'descent' should be against the gradient and towards the minimum. The term η is the 'learning rate' of the network, i.e. a user-defined tuning parameter that provides stability to the learning process and controls the step size of the gradient descent.

Broadly speaking, the partial derivative $\partial\mathcal{J}/\partial w_{n,m}$ in equation (4) is proportional to the inputs of the neurons, the derivatives of their activation functions with respect to their weights, and the errors of the network. For the output layer, these factors are decoupled, which makes the computation of the derivative relatively straightforward. However, for the earlier (and deeper) layers, the term $\partial\mathcal{J}/\partial w_{n,m}$ depends on the derivative of the loss function for the outer layers, which in turn is proportional to the *product* of the derivatives of all the activation functions with respect to their weights. This product operation is implemented with the chain rule from the output to the initial layer, i.e. *backwards*, hence giving rise to the concept of *backpropagation*. We note that in many ML applications, including `cecilia`, the exact value of the gradient at each training iteration is not computed with the full training sample, but approximated with subsets (or 'batches') of the latter (e.g. Shallue & Vanderburg 2018).

3 DATA

In this section, we motivate the use of a NN spectral interpolator, describe the atmosphere models and synthetic stellar properties used to train, validate, and test `cecilia`, and discuss the normalization techniques used to improve the learning and predictive abilities of our code.

3.1 Generation of a training set: motivation

Current WD spectral modelling techniques rely on grids of stellar properties (or 'labels') with uniform step sizes (e.g. Coutu et al. 2019). Although this approach has worked well over the years, it can be computationally expensive for a large number of free model parameters. For example, using 13 stellar properties – as in this work (see Section 3.1.1 below) – and a relatively coarse grid of 10 points per label, conventional interpolators would require at least 10^{13} models to explore the parameter space of their study. Assuming, on average, that each model takes about 1 hour to generate on a single CPU core, it would take on on the order of 10^9 CPU-yr to compute the whole grid. Therefore, classical methods are prohibitively expensive.

To address this challenge, we have developed `cecilia`, a ML generative spectral interpolator capable of sampling from a high-dimensional label space based on *randomly drawn* training parameters, rather than a uniformly spaced grid of model points. With this ML approach, our pipeline only needs about 20 000 synthetic models to predict WD spectra, i.e. about 2×10^{-9}% of the models that classical methods would typically require for the same task.

3.1.1 Synthetic stellar labels

In this work, we have chosen to characterize He-rich polluted WDs with 13 stellar properties, namely, their effective temperature ($T_{\rm eff}$) and surface gravity ($\log g$), their hydrogen abundance, and 10 metal

[2]In this work, we define an 'instance' as a spectrum with its corresponding 13 stellar labels.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Table 1. `cecilia`'s ranges used to generate the training set of synthetic WD properties (or stellar 'labels'). All the atmospheric abundances are expressed as number abundance ratios in base 10 relative to helium.

Sampled from a random uniform distribution			
Label	Min. bound	Max. bound	Ref.
$T_{\rm eff}$ (K)	10 000	20 000	-
log g (cgs)	7	9	[1]
$\log_{10}$ (H/He)	−7	−3	[2]
$\log_{10}$ (Ca/He)	−12	−7	[2]
Sampled from a random gaussian distribution (equation 7)			
Label	μ_Z (dex)	σ_Z (dex)	Ref.
log (Mg/Ca)	1.26	2	[3]
log (Fe/Ca)	1.16	2	[3]
log (O/Ca)	2.35	2	[3]
log (Si/Ca)	1.17	2	[3]
log (Ti/Ca)	−1.39	2	[3]
log (Be/Ca)	−4.96	2	[3]
log (Cr/Ca)	−0.70	2	[3]
log (Mn/Ca)	−0.91	2	[3]
log (Ni/Ca)	−0.12	2	[3]

References: [1] Chandra et al. (2020), [2] Coutu et al. (2019), and [3] Asplund et al. (2009).

abundances for Ca, Mg, Fe, O, Si, Ti, Be, Cr, Mn, and Ni (see Table 1).[3] In our models, we have also considered 14 additional heavy elements (C, N, Li, Na, Al, P, S, Cl, Ar, K, Sc, V, Co, and Cu). However, none of these metals are fitted during `cecilia`'s optimization procedure.

To prepare `cecilia`'s training set of synthetic spectra, we generated 25 000 unique combinations of the 13 aforementioned stellar labels (i.e. $T_{\rm eff}$, log g, and 11 elemental abundances). In particular, for $T_{\rm eff}$, we generated our labels from a random uniform distribution spanning the range [10 000, 20 000] K. At higher effective temperatures, the presence of photospheric metal pollution may not be due to the accretion of exoplanetary material, but to the effects of radiative levitation pressure (Chayer, Fontaine & Wesemael 1995). For cooler WDs, there is less thermal energy to produce atomic transitions in the optical; as a result, there is a limited number of polluted systems exhibiting multiple heavy elements at once. With respect to log g, we followed Chandra et al. (2020) and sampled our values from a random uniform distribution between [7,9] cgs (with g in cgs units of cm s^{-2}). For $\log_{10}$ (H/He), we also drew our labels from a random uniform distribution between [−7,−3] dex, as in Coutu et al. (2019).

To generate our labels for the other elements, we used the following procedure. First, we drew our calcium abundances from a random uniform distribution between [−12,−7] dex:

$$\mathcal{U}_{\log(\rm Ca/He)} = \mathcal{U}(-12, -7). \tag{5}$$

For the remaining 10 heavy elements (i.e. Mg, Fe, O, Si, Ti, Be, Cr, Mn, and Ni), we generated a random Gaussian distribution centred around μ_Z with a standard deviation σ_Z,

$$\mathcal{N}_{\log(\rm Z/Ca)} = \mathcal{N}(\mu_Z, \sigma_Z), \tag{6}$$

where $\sigma_Z = 2$ dex and

$$\mu_Z = \log_{10}(\rm Z/He)_\odot - \log_{10}(\rm Ca/He)_\odot = \log_{10}(\rm Z/Ca)_\odot . \tag{7}$$

[3]In this work, all the abundances are expressed as logarithmic number abundance ratios (in base 10) relative to helium.

In the above expressions, Z is the metal under consideration, $\log_{10}(\rm Z/He)_\odot$ is the abundance of metal Z in the solar photosphere (Asplund et al. 2009), and $\log_{10}(\rm Ca/He)_\odot$ is the solar calcium abundance relative to helium (Asplund et al. 2009). We note that our choice of $\sigma_Z = 2$ dex was motivated by two factors: on the one hand, σ had to be large enough to allow `cecilia` to measure the true elemental abundances of a WD, even if the latter deviated from chondritic values; and on the other hand, we also wanted σ to be small enough to downweight unlikely regions of the parameter space in our training set, hence limiting the hypervolume of `cecilia`'s label parameter space.

As a final step, we generated the distribution $\mathcal{D}$ of abundance ratios log(Z/He) for our training set by adding

$$\mathcal{D} = \mathcal{N}_{\log(\rm Z/Ca)} + \mathcal{U}_{\log(\rm Ca/He)} \tag{8}$$

In Table 1, we list the values of μ_Z and σ_Z for Mg, Fe, O, Si, Ti, Be, Cr, Mn, and Ni, rather than the properties of the (non-Gaussian) distributions resulting from evaluating equation (8).

Finally, for the remaining 14 heavy elements (C, N, Li, Na, Al, P, S, Cl, Ar, K, Sc, V, Co, and Cu), we chose to scale their abundances such that their abundance ratios relative to calcium corresponded to the values of CI chondrites reported in in Lodders (2003). This type of bodies are primitive carbonaceous meteorites with a pristine composition unaltered by melting and differentiation processes. With the exception of their volatile components, including their lithium and noble gas abundances, their chemical makeup – mostly dominated by O, Mg, Si, and Fe– closely resembles that of the solar photosphere (Palme, Lodders & Jones 2014) as well as that of bulk Earth (e.g. Zuckerman et al. 2007; Xu et al. 2019; Doyle et al. 2023). Therefore, CI chondrites are often used as a proxy for the average bulk composition of non-volatile objects in the Solar System (Lodders 2003; Jura & Young 2014). In the field of polluted WDs, the assumption of chondritic ratios often constitutes a reasonable first-order approximation when there is no *a priori* knowledge of the stellar photospheric composition (Dufour et al. 2007; Coutu et al. 2019). In the context of this work, the use of chondritic abundances for the 14 'fixed' metals also allowed us to dramatically decrease the volume of the parameter space that `cecilia` had to learn during training.

3.1.2 *Synthetic atmosphere models*

Next, we used the 25 000 synthetic labels to generate their corresponding atmosphere models. To achieve this, we employed the code described in Dufour et al. (2007); Blouin, Dufour & Allard (2018a); Blouin et al. (2018b) and references therein. This code includes all the necessary physical principles to predict the emerging spectra of metal-polluted WDs and it has been used successfully to reproduce ultraviolet (UV) and optical observations across a wide range of effective temperatures and metal abundances (Dufour et al. 2012; Giammichele, Bergeron & Dufour 2012; Vanderburg et al. 2015; Melis & Dufour 2017; Xu et al. 2017; Coutu et al. 2019; Blouin et al. 2019a, b; Blouin 2020; Kaiser et al. 2021; Klein et al. 2021; Caron et al. 2023; Doyle et al. 2023). More specifically, for each model, the code calculates a 1D thermodynamic structure first, and this stratification is used to generate a synthetic spectrum with no instrumental broadening and sampled with a pixel spacing equal to $R \approx 50\,000$ (i.e. 55 000 wavelength points between 3000 and 9000 Å). Fig. 2 shows a selection of synthetic spectra for He-rich polluted WDs spanning different effective temperatures, surface gravities, and photospheric compositions.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

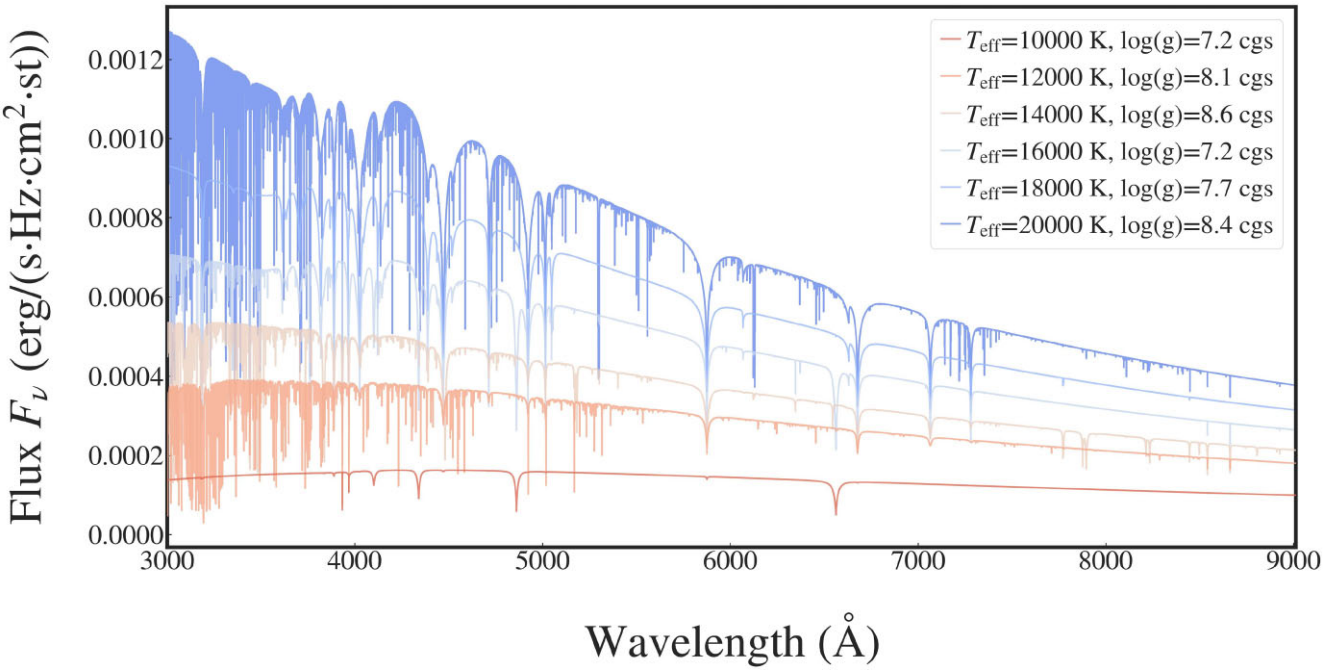


Figure 2. A selection of six atmosphere models for synthetic He-rich polluted WDs with effective temperatures between 10 000 and 20 000 K, and surface gravities between 7 and 9 cgs. Cool and hot WDs are shown, respectively, in red and blue.

For `cecilia`, we independently varied 13 parameters for each model (i.e. $T_{\rm eff}$, log g, and 11 elemental abundances; see Table 1), which we randomly selected for each calculation as described in Section 3.1.1. In total, we calculated 25 000 atmosphere models, but 2176 of them (8.7 per cent) had to be discarded due to computational issues, leaving 22 824 useful synthetic spectra for training, testing, and validation.[4] We note that exclusion of these models did not have any significantly impact `cecilia`'s predictive ability; this is due to our code's ML-based spectral interpolation approach, which facilitates the computation of spectral models in regions of the parameter space where there are no observations. We also recall that while the abundances of 10 metals (Ca, Mg, Fe, O, Si, Ti, Be, Cr, Mn, and Ni) were varied independently between different synthetic spectra, the remaining 14 heavy elements (C, N, Li, Na, Al, P, S, Cl, Ar, K, Sc, V, Co, and Cu) were also included in our models with their abundance ratios Z/Ca fixed to their nominal CI chondritic abundance from Lodders (2003).

3.2 Data normalization

As is standard procedure in ML, we then normalized the synthetic spectra and their corresponding stellar labels between 0 and 1 in order to improve `cecilia`'s learning abilities during training. For our synthetic spectra, we experimented with different scaling techniques and ultimately adopted a 'pixel-wise' (PW) normalization approach (also commonly known as *min-max* scaling in ML). For a given wavelength index λ_i, this approach performs the following linear transformation: first, it determines the minimum and maximum values (or 'pixels') of the synthetic flux across *all* the available spectra; and then, it uses these flux bounds to normalize the pixel at wavelength λ_i to a value between 0 and 1. Therefore, the normalization is conducted with respect to the flux axis, rather than along the wavelength dimension.

As illustrated in Fig. 3, the PW approach may sometimes cause the normalized spectra to appear distorted to the human eye. Nevertheless, it is a powerful scaling technique because it treats each pixel as an independent spectral 'feature', thus allowing `cecilia` to learn both the small- and large-scale variations in the synthetic flux. In other words, pixels associated to large variations become less prominent, while those linked to smaller changes become magnified. In addition to this advantage, the PW normalization differs from other ML scaling techniques (e.g. log-scaling) because it treats each pixel in the same way, therefore ensuring that `cecilia` does not have any preference for a specific absorption line.

4 METHODOLOGY

4.1 Design of `cecilia`'s ML architecture

The central problem in this paper lies in accurately determining the main properties – or stellar labels – of polluted WDs from their spectra. To achieve this goal, we have developed `cecilia`, an innovative ML pipeline capable of performing two tasks: (i) the generation of polluted WD spectra from a set of stellar labels, i.e. $\mathcal{S} = \beta_1(\mathcal{L})$, where $\mathcal{L}$ is the vector space of the WD labels (with dimension L), and β_1 is a generic non-linear function; and (ii) the prediction of polluted WD labels from their corresponding spectra, or $\mathcal{L} = \beta_2(\mathcal{S})$, where $\mathcal{S}$ is the vector space of the WD spectrum (with dimension S), and β_2 is a function different to β_1.

Developed with the open-source Python TENSORFLOW package (Abadi et al. 2015), `cecilia` is characterized by three deep NNs: an Autoencoder, a Fully Connected Neural Network (FCNN1), and a replica of the FCNN1 for fine-tuning purposes (FT FCNN2). The Autoencoder is a ML tool consisting of two symmetrical networks (Fig. 4): an *Encoder* and a *Decoder*. The *Encoder* identifies the

[4]The missing models were not clustered in any specific region of the parameter space. Therefore, they did not negatively affect the robustness of `cecilia`.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

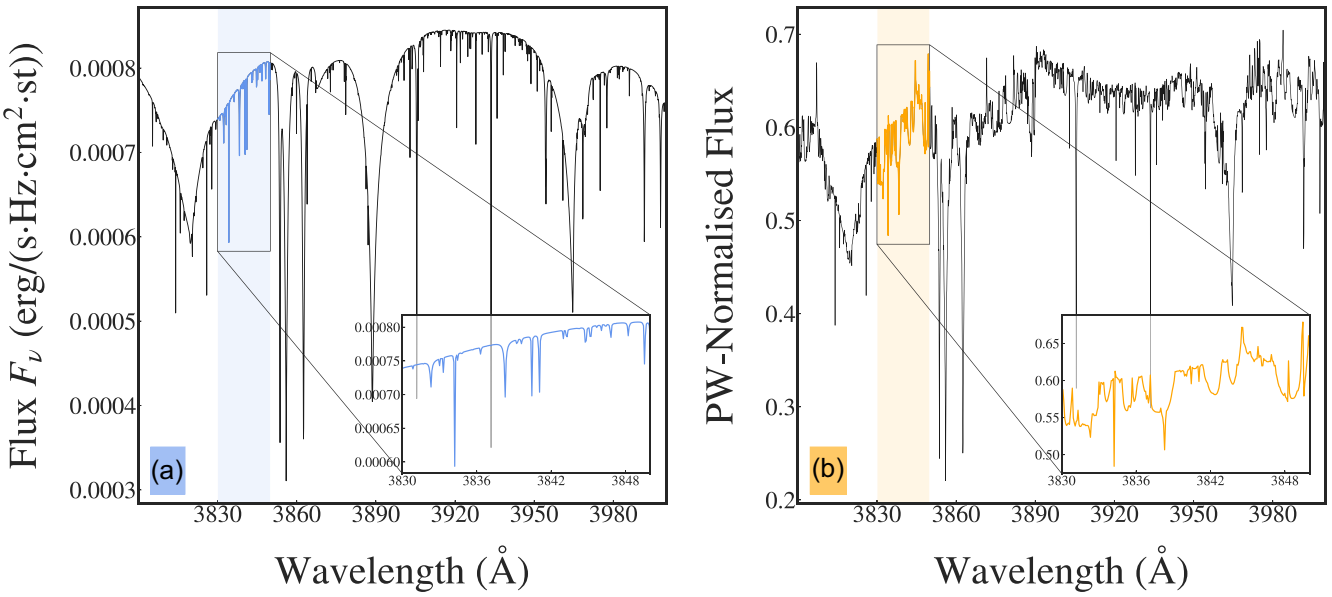


Figure 3. Effect of applying the PW normalization technique on a synthetic spectrum in the spectral window between 3800 and 4000 Å. This technique is designed to improve `cecilia`'s learning capabilities by amplifying any small variations in the stellar flux and attenuating the largest fluctuations. Panels (a) and (b) show, respectively, the raw and PW-normalized synthetic flux of the spectrum. The inset plots show the region between 3330 and 3350 Å.

main attributes of the input data and compresses the latter into a lower-dimensional representation. This operation takes place in the so-called 'hidden' or 'latent' space (denoted by $\mathcal{H}$), where $\mathcal{H}$ has a dimensionality equal to the number H of hidden/latent features used to encode the data (i.e. $\mathcal{H} \in \mathbb{R}^H$). The second part of the Autoencoder is the *Decoder*, which is a network designed to reconstruct the original input data based on the *Encoder*'s latent model while trying to minimizing the loss of information caused by the dimensionality reduction. We emphasize that the hidden features of the Autoencoder are fundamental properties of the ML model and are not intrinsically correlated to any stellar parameters. As a result, they have no astrophysical interpretation. In general, the choice of H constitutes a trade-off between the interpretability and the complexity of the ML model, characterized respectively by a low and a high H (Liang, Winn & Melchior 2023).

Following the Autoencoder, `cecilia` incorporates two deep NNs: the FCNN1 and the FT FCNN2. As shown in Figs 5 and 6, these networks consist of an input layer, six hidden layers, and an output layer. Unlike the Autoencoder, which is used to perform compression and dimensionality reduction, the FCNN1 and FT FCNN2 architectures are designed to teach `cecilia` how to interpolate WD spectra based on a set of 13 stellar labels. We provide a more detailed discussion about this subject in Section 4.2.

At the level of individual neurons, the choice of weights, biases, and activation functions can also shape the performance of a NN, as explained in Section 2. In this work, we followed common ML practice and set the weights of `cecilia` to random values before optimizing them during training. More specifically, we used the TENSORFLOW functions `he_uniform` and `glorot_uniform` to initialize the weights of all the initial layers and the final output layer, respectively. These two functions draw values from random uniform distributions in the range between [-limit, +limit]. If $\mathrm{fan_{in}}$ and $\mathrm{fan_{out}}$ are, respectively, the number of input and output connections of a neuron, the parameter 'limit' in TENSORFLOW is defined as

$$\mathrm{limit} = \begin{cases} \sqrt{6/\mathrm{fan_{in}}} & (\texttt{he_uniform}) \\ \sqrt{6/(\mathrm{fan_{in}} + \mathrm{fan_{out}})} & (\texttt{glorot_uniform}) \end{cases}. \tag{9}$$

With respect to `cecilia`'s activation functions, we opted for using a Rectified Linear Unit (ReLU; Jarrett et al. 2009) and a sigmoid function (Goodfellow, Bengio & Courville 2016), as specified in Table 2. These two activation functions regulate the flow of information by introducing non-linearities across the networks, normalizing the neurons' output parameters, and enabling the identification of hidden and complex patterns in the training sample. For an input value x, the ReLU function returns 0 if x is negative, or x if x is positive,

$$\mathrm{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases} = \max(0, x). \tag{10}$$

Therefore, its derivative is given by

$$\frac{d}{dx}\left[\mathrm{ReLU}(x)\right] = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}. \tag{11}$$

In contrast, the sigmoid activation function $\sigma(x)$ transforms the input value x into a value between 0 and 1 with

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{12}$$

As a consequence, its derivative approaches zero when $x \to \pm\infty$. In NNs with multiple hidden layers, the potential saturation of the sigmoid function towards zero values can lead to the so-called 'vanishing gradient problem' (Hanin 2018). This issue arises during backpropagation when the gradient of the loss function, which encapsulates the product of the derivatives of the activation functions with respect to their weights (see Section 2), becomes extremely small. Due to this 'vanishing' effect, the weights and biases of the initial layers may only be marginally updated, which can cause poor and slow model convergence. As evidenced by equation (11), ReLU functions are more robust to this problem because they have a more stable derivative (either 0 or 1). As a result, they often yield a better performance during training (Nair & Hinton 2010). In this work, we decided to employ ReLU activation functions for the initial layers of `cecilia`'s networks in order to avoid the vanishing gradient of the loss function. For the final layer of each network, we chose to

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

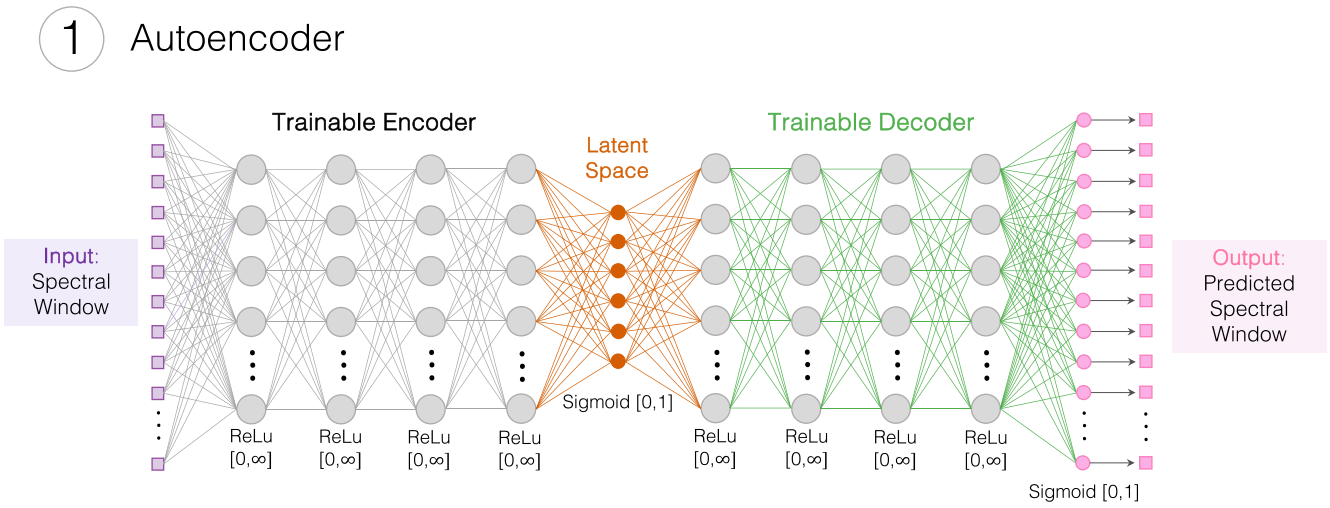


Figure 4. The first component of `cecilia`'s ML architecture is an Autoencoder, which consists of two networks: an Encoder (in grey) to compress input data into a small number of latent, hidden features (in orange); and a Decoder (in green) to reconstruct the encoded data from the latent space. In our work, the input and output data of the Autoencoder are spectral windows of 200 Å from our atmosphere models (in purple and pink, respectively).

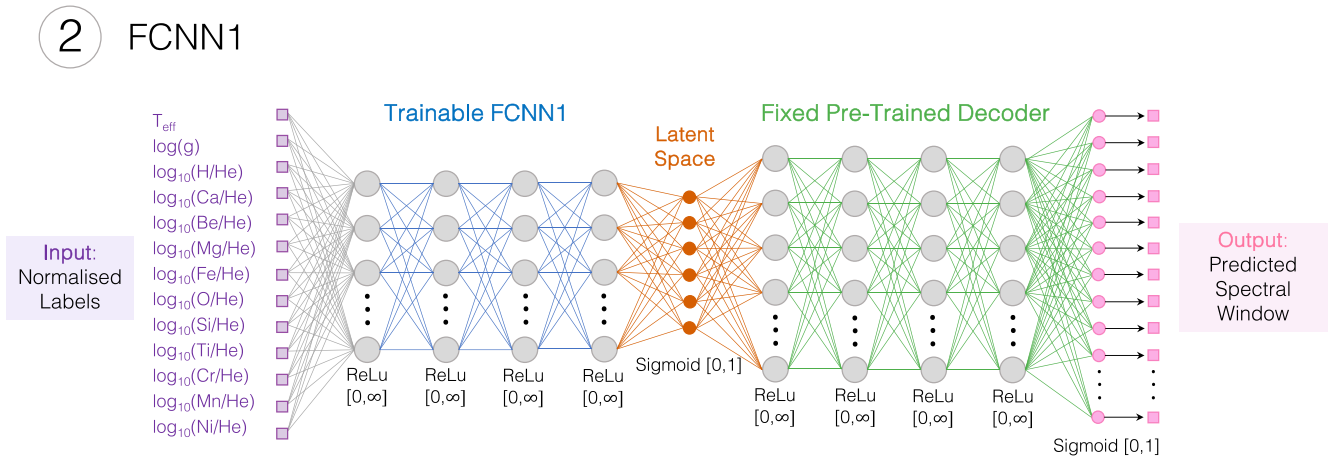


Figure 5. The second component of `cecilia`'s ML architecture is a feed-forward FCNN1 (in blue). This network leverages the trained Decoder in Fig. 4 (in green) to predict the spectrum of a polluted WD (in pink) from its corresponding stellar labels (in purple).

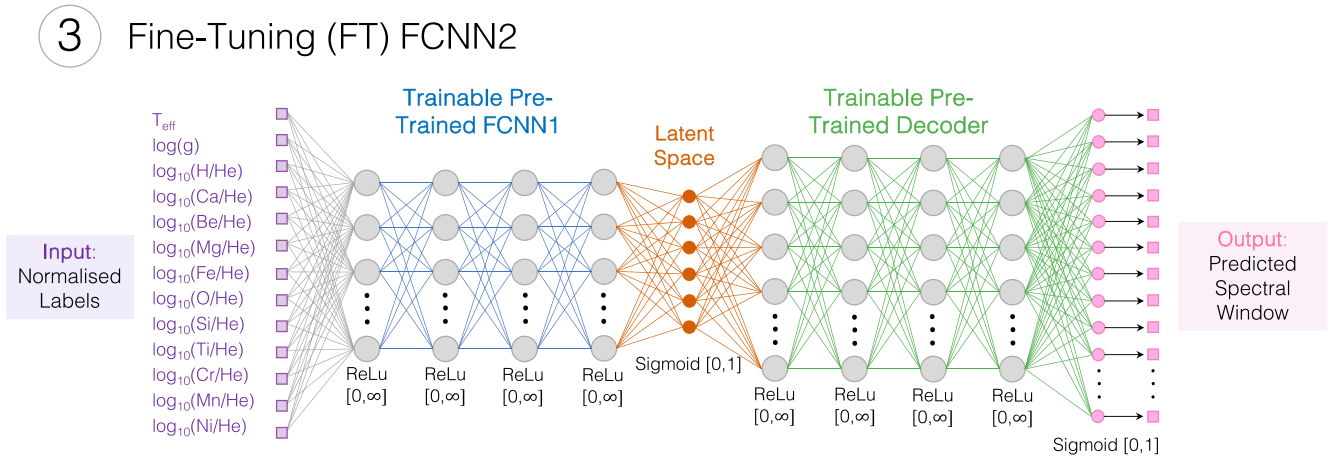


Figure 6. The third component of `cecilia`'s ML architecture is a replica of the FCNN1 aimed at fine-tuning the final predictions of our pipeline (FT FCNN2; in blue). Similarly to the FCNN1 Fig. 5, this network is connected to the trained Decoder (in green) to predict a metal-polluted spectrum (in pink) from its stellar labels (in purple).

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Table 2. Main properties of `cecilia`'s ML architecture. Its three NNs are trained sequentially, but only the final one (i.e. the FT FCNN2) is used to model the spectra of He-rich polluted WDs.

Parameter	Value
Autoencoder	
No. of layers in Encoder	5
No. of layers in Decoder	5
No. of neurons/layer	4000
No. of training epochs	20 000
No. of latent features	100
Learning rate	10^{-6}
Batch size training	64
Batch size validation	64
Activation functions	ReLu (initial four layers) sigmoid (output layer)
Weight Initializers	`he_uniform` (initial four layers) `glorot_uniform` (output layer)
Fully-Connected Neural Network (FCNN1)	
No. of total layers	6
No. of neurons/layer	2000
No. of training epochs	10 000
Learning rate	10^{-5}
Batch size training	64
Batch size validation	64
Activation functions	ReLu (initial five layers) sigmoid (output layer)
Weight initializers	`he_uniform` (initial five layers) `glorot_uniform` (output layer)
Fine-Tune Fully-Connected Neural Network (FT FCNN2)	
No. of total layers	6
No. of neurons/layer	2000
No. of training epochs	1000
Learning rate	10^{-4}
Batch size training	64
Batch size validation	64
Activation functions	ReLu (initial five layers) sigmoid (output layer)
Weight initializers	`he_uniform` (initial five layers) `glorot_uniform` (output layer)

use a sigmoid function with the goal of generating predicted values between 0 and 1.

In Table 2, we summarize the fundamental properties of `cecilia`'s Autoencoder, FCNN1, and FT FCNN2. These properties were determined with the Python `tensorboard` library, which is a built-in optimization module of TENSORFLOW designed to execute and compare several configurations of the same ML architecture. In particular, `tensorboard` performs a short training of each configuration to identify the one with the most optimal performance. During `cecilia`'s development phase, we used `tensorboard` to optimize the number of layers, the neurons per layer, the batch size (i.e. the number of training instances processed together during a given iteration), and the learning rate of the Autoencoder and the two FCNNs.

4.2 Network training

To train `cecilia`, we used a partition of MIT's Satori Supercomputer with one NVidia V100 GPU.[5] First, we randomly split our collection of synthetic labels and spectra into three data sets for training (70 per cent), testing (20 per cent), and validation (10 per cent) purposes. Next, we used the backpropagation technique to optimize the hyperparameters of `cecilia`'s NNs. To this end, we implemented the Adaptive Moment Estimation (Adam) optimization algorithm (Kingma & Ba 2017), which is an extension of the commonly used Stochastic Gradient Descent routine (Kiefer & Wolfowitz 1952). Finally, given that the synthetic spectra described in Section 3.1.2 represented a substantial amount of data (a total of about 10^9 flux pixels for 22 842 spectra), we chose to train our pipeline by dividing the synthetic spectra in 29 windows of 200 Å, rather than by considering the full wavelength range at once.[6] With this approach, the end-to-end training of `cecilia` took approximately three weeks.

For each spectral window of 200 Å, `cecilia`'s training consisted of three sequential learning phases: an initial one for the Autoencoder, another one for the FCNN1, and a final one for the FT FCNN2. In each phase, the networks employed the same data sets for training, validation, and testing purposes, with each data set consisting of unique combinations of the 13 randomly generated labels listed in Table 1 (i.e. $T_{\rm eff}$, log g, and 11 elemental abundances from H to Ni). It is important to clarify that the Autoencoder and the FCNN1 were *only* used to improve the learning of the fine-tuned NN. After concluding `cecilia`'s training, our code only uses the FT FCNN2 to generate spectral models of polluted WDs. Below, we summarize the three training phases in more detail.

(i) *Phase 1 (Fig. 4)*: First, we trained the Autoencoder in order to compress and reconstruct the normalized synthetic spectra based on only 100 latent features. At the end of this process, we used the trained Encoder to transform our full data base of 22 824 synthetic spectra into their latent space representations. If $\mathcal{S}$ and $\mathcal{H}$ represent, respectively, the vector spaces of the synthetic spectra and the hidden space (see Section 4.1), we can denote the operations of the trained Autoencoder as

$$\mathcal{S}\rightarrow\mathcal{H}\rightarrow\mathcal{S}. \tag{13}$$

(ii) *Phase 2 (Fig. 5)*: Next, we trained the first Fully-Connected Neural Network (FCNN1) to compress the normalized stellar labels into their latent space representations. We then appended the trained Decoder to the output layer of the FCNN1 to convert the latent features of the normalized labels into predictions of normalized synthetic spectra. If $\mathcal{L}$ is the vector space of the stellar labels, the FCNN1 and trained Decoder are thus in charge of performing the operation:

$$\mathcal{L}\rightarrow\mathcal{H}\rightarrow\mathcal{S}. \tag{14}$$

At this stage, we had two options for training the FCNN1 while keeping the trained Decoder fixed. On the one hand, we could directly determine the errors of the hidden layer with respect to the encoded spectra $\mathcal{H}$. This scenario required prior compression of the spectra with the Encoder and involved backpropagating the error through the FCNN1 to correct the weights and biases of the network. On the other hand, we could calculate the errors of the FCNN1's output layer with respect to the true synthetic spectra $\mathcal{S}$. This alternative option required backpropagating the errors through the fixed Decoder until reaching the hidden space, and then updating the weights and biases of the FCNN1 accordingly. In other words, the main difference between the two approaches was the location of the network used

[5] https://mit-satori.github.io/.

[6] We also evaluated the performance of `cecilia` with narrower and wider spectral windows. From this analysis, we found that 200 Å offered the highest computational efficiency and predictive accuracy.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

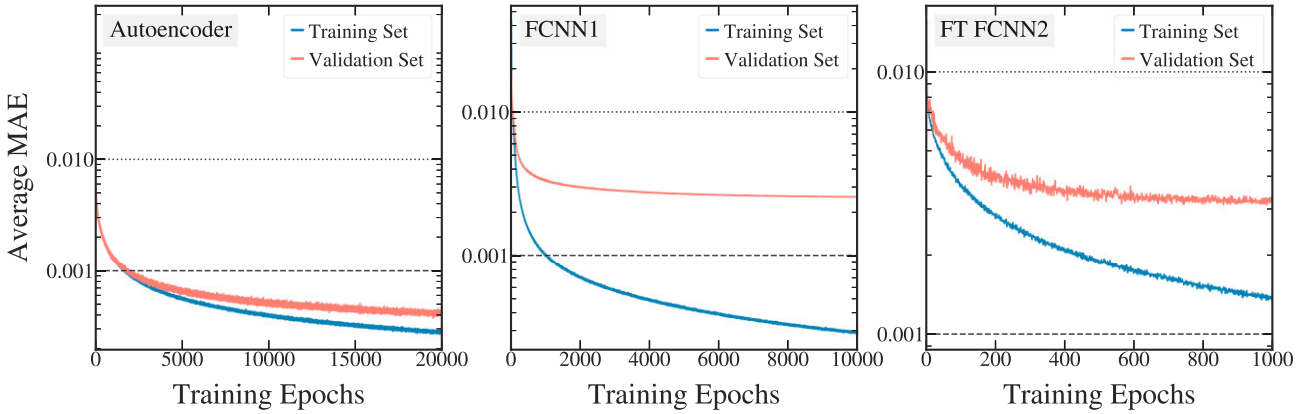


Figure 7. MAE statistic (or loss function) for the Autoencoder, FCNN1, and FT FCNN2, averaged for the 29 windows of 200 Å used to train `cecilia`. The training and validation sets are shown in blue and red. The dotted and dashed lines denote an average MAE of 0.01 and 0.001, respectively.

to calculate the loss function, with the first and second scenarios employing, respectively, the encoded synthetic spectra $\mathcal{H}$ and the normalized synthetic spectra $\mathcal{S}$. In this work, we decided to follow the first approach in order to (i) avoid unnecessary backpropagation steps while traversing the Decoder, and (ii) prevent gradient vanishing effects of the loss function, which could result in a slower and poorer performance.

(iii) *Phase 3 (Fig. 6)*: Lastly, we sought to improve and fine-tune `cecilia`'s final predictions by training a second Fully-Connected Neural Network (FT FCNN2) connected to the trained Decoder. This architecture replicates the structure of the FCNN1, but it differs from the latter because we allowed the parameters of the Decoder to be trainable. Similarly to equation (14), the joint FT FCNN2 and Trainable Decoder architecture performs the operation

$$\mathcal{L} \rightarrow \mathcal{H}^* \rightarrow \mathcal{S}, \tag{15}$$

where the latent space of this third network, represented by the term $\mathcal{H}^*$ in equation (15), is now different than those of the Autoencoder and the FCNN1 ($\mathcal{H}$).

Upon training `cecilia`'s Autoencoder and FCNNs, we calculated the error between the predictions of each network and their respective true synthetic models. To this end, we adopted the Mean Absolute Error (MAE) metric as our loss function – a popular choice in ML regression problems. In this work, we calculate the MAE of a spectral window as

$$\mathrm{MAE} = \frac{1}{QP} \sum_{q=1}^{Q} \sum_{p=1}^{P} \left| y_{q,p} - \hat{y}_{q,p} \right|, \tag{16}$$

where q and Q represent, respectively, the index and the total number of instances – where we recall that an instance is a synthetic spectrum with its corresponding stellar labels –, p and P denote the index and total number of output pixels (i.e. the number of flux points in a given spectral window), y and $\hat{y}$ correspond to the true and predicted synthetic flux. Fig. 7 shows our loss function for each network, with the MAE averaged for the 29 windows in the wavelength range between 3000 and 9000 Å. From this figure, we find that the training error rates of `cecilia`'s networks decrease with each epoch, as should be the case during training. This behaviour demonstrates the ability of `cecilia` to recognize abstract features in the training data and generalize its acquired knowledge to previously unseen observations. At the same time, Fig. 7 shows that none of `cecilia`'s networks are suffering from the dangers of underfitting or overfitting:[7] on the one hand, the absence of underfitting can be demonstrated with the convergence of the validation error, which becomes asymptotically horizontal, thus leaving no substantial margin for further generalization; on the other hand, the problem of overfitting can be discarded because the validation error does not increase over time. In Fig. 8, we also present an example of `cecilia`'s predictions for several 200 Å-windows of a synthetic spectrum.

A general note about our training process is that Phases 2 and 3 *recycle* the trained Decoder to improve the performance of the FCNN1 and the FT FCNN2. After experimenting with less sophisticated ML architectures, we found this approach resulted in more accurate predictions than a simpler NN designed to predict stellar labels directly from their spectra without any intermediary steps (i.e. $L = f(S)$). The use of a pre-trained model to improve the training of another network is known as 'Transfer Learning' and has been very successful in many scientific fields outside computational science (e.g. Tan et al. 2018).

4.3 Parameter estimation

After training, `cecilia` can use its FT FCNN2 architecture to rapidly (<1 s) generate a full high-resolution synthetic spectrum from 13 stellar labels ($T_{\rm eff}$, $\log g$, $\log_{10}$(H/He), and 10 metal abundances relative to helium). Exploiting the fast and automated ML interpolation capabilities of our pipeline, we have developed a new fitting framework to simultaneously and accurately measure the main physical and chemical properties of He-rich polluted WDs from their spectra. This section explains our methodology for achieving this goal using a step-by-step approach. In our discussion, we will employ the subscripts 'synth' and 'obs' to represent `cecilia`'s synthetic predictions and a real WD spectrum, respectively. With this

[7]The concept of *underfitting* refers to a NN that has not managed to learn the fundamental attributes of a training sample. This problem can arise from various reasons – e.g. an excessively simple ML model, a small training set, a short training duration – and it leads to poor predictions of new data sets. In contrast, *overfitting* describes a network that has memorised the underlying features of the training data too well but cannot make accurate predictions with a test data set. This behaviour can also happen due to multiple reasons, such as an overly complex ML model or an unnecessarily long training process. An optimal ML model represents a balance between generalization and specificity (Goodfellow, Bengio & Courville 2016).

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

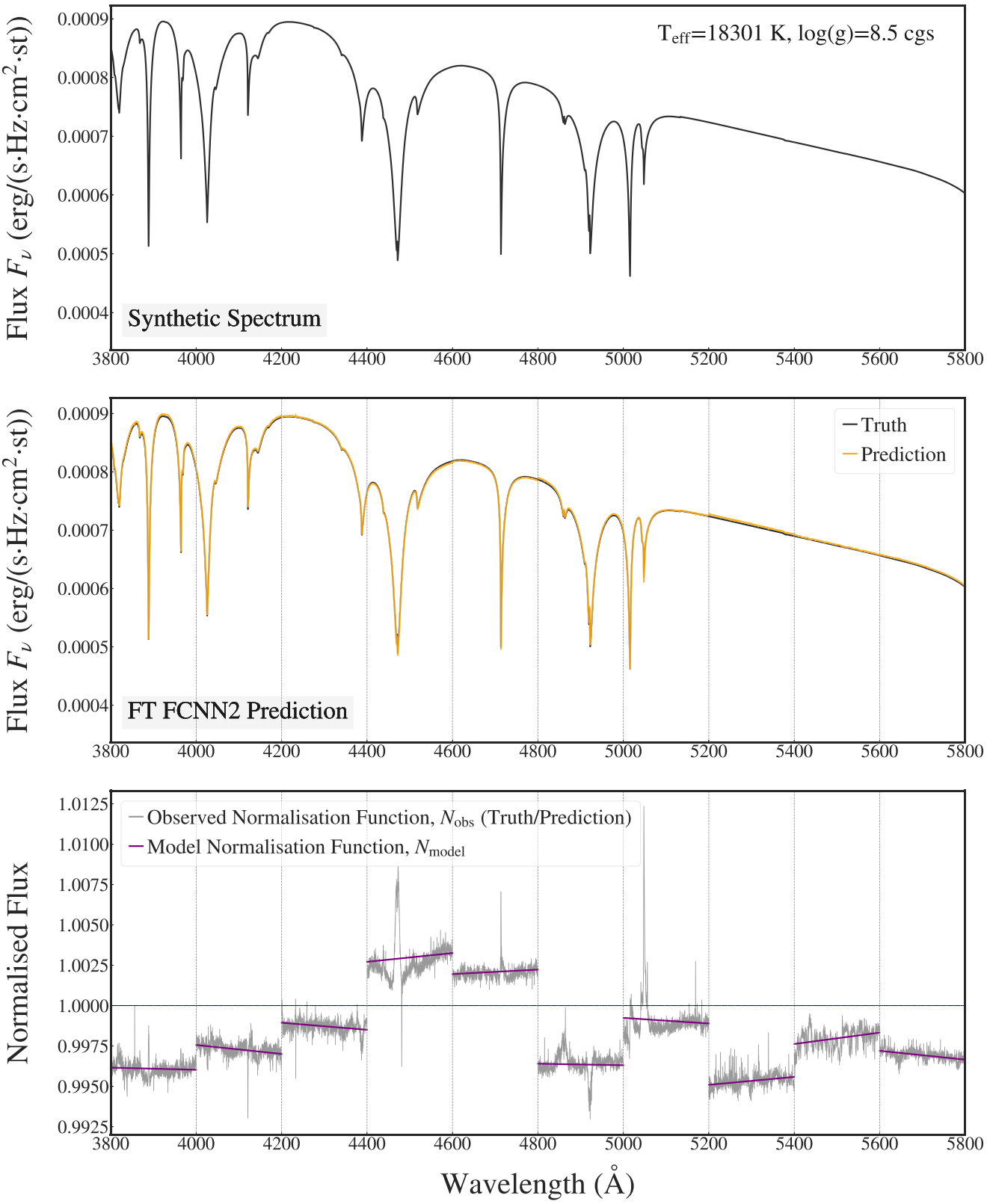


Figure 8. *Top Panel*: Synthetic spectrum corresponding to a WD with $T_{\rm eff}$=18,301 K and log g=8.5 cgs. To allow for a closer examination, we only show the wavelength range between 3800 and 5800 Å, which features 10 windows of 200 Å. *Middle Top Panel*: Denormalized `cecilia` FT FCNN2 prediction. *Bottom Top Panel*: Slope correction with a linear least-squares fit. In each panel, we use dashed vertical lines to indicate the start/end of a 200 Å spectral window.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

choice of notation, we will refer to their corresponding wavelength, flux, and flux error arrays as λ_X, f_X, and $f_{X,\,err}$, where X will indicate the nature of the data (i.e. synthetic 'synth', or observed 'obs'). Finally, we will denote their resolving power as R_X, where $R_X \equiv \lambda_X/\Delta\lambda_X$ (unitless).

Broadly speaking, our fitting framework can be divided into three main phases (see Fig. 9). First, `cecilia`'s trained FT FCNN2 architecture generates a preliminary atmosphere model based on the assumed astrophysical properties of the WD. Second, our pipeline optimizes this initial prediction with the non-linear least-squares Levenberg–Marquardt algorithm implemented by the fast Python `mpfit`[8] library (Moré 1978; Markwardt 2009). To conclude, we perform a global Bayesian fit with the differential evolution Markov Chain Monte Carlo (MCMC) sampler of the Python EDMCMC package (Ter Braak 2006; Vanderburg 2021).[9] More specifically, our optimization procedure consists of the following phases:

(i) **Step 1: Input spectrum and stellar parameters**: First, the user uploads the spectrum of their polluted WD (λ_{obs}, f_{obs}, $f_{obs,\,err}$) into `cecilia`, ensuring that the effective temperature and surface gravity of the star satisfy the ranges $10\,000 \leq T_{eff} \leq 20\,000$ K and $7 \leq \log g \leq 9$ cgs imposed in Section 3.1.1 (see Table 1). In parallel, the user provides a list of 14 WD properties to our pipeline, either complete with initial guesses, or lacking the latter if unknown. These properties consist of the 13 stellar labels used during `cecilia`'s training,[10] as well as an additional radial velocity shift (or RV thereafter) to account for the barycentric motion and gravitational redshift of the WD. If necessary, the RV term can also be used to handle potential wavelength calibration issues in the spectrum.

(ii) **Step 2: Initial guesses for stellar parameters**: Next, `cecilia` reviews the user's WD properties and performs the following operations based on the availability (or absence) of initial guesses for the 13 stellar labels and the RV term:

(a) *Scenario 1: The user has prior knowledge on all stellar parameters*: In this optimal situation, `cecilia` examines the user's initial guesses to confirm that they are within the ML bounds listed in Table 1. If the RV shift is not within the range given by $-500 \leq RV \leq 500$ km s^{-1}, our code automatically sets this parameter to 0 km s^{-1}. Alternatively, if any of the abundances are outside of their corresponding ML ranges, `cecilia` attempts to correct this problem by generating a normal distribution centered on the user's estimate with a scale of 0.1 dex, and subsequently adopting a random value from this distribution as an initial guess. If this modification is still insufficient, `cecilia`'s Gaussian estimate is then substituted by a chondritic abundance ratio based on the assumed $\log_{10}$(Ca/He) of the WD and the chondritic values of Asplund et al. (2009).

(b) *Scenario 2: The user has partial knowledge of some stellar parameters*: In this intermediate scenario, our code starts by verifying that the user's initial guesses for T_{eff}, log g, $\log_{10}$(H/He) and $\log_{10}$(Ca/He) are within `cecilia`'s allowed bounds in Table 1. If any of these parameters are unknown, `cecilia` uses the mean of their ML bounds as initial guesses for its optimization routine (i.e. $\bar{T}_{eff}$=15 000 K, $\bar{\log g}$=8 cgs, $\log(\bar{H}/He) = -5$, and $\log(\bar{Ca}/He) = -9.5$). For the remaining WD parameters, `cecilia` adopts a zero RV shift and assumes chondritic abundances for Mg, Fe, O, Si, Ti, Be, Cr, Mn, and Ni based on the estimated $\log_{10}$(Ca/He) of the WD and the chondritic ratios of Asplund et al. (2009).

(c) *Scenario 3: The user has no knowledge of the stellar parameters*: In this worst-case scenario, `cecilia` generates preliminary estimates of T_{eff}, log g, $\log_{10}$(H/He) and $\log_{10}$(Ca/He) by taking the mean of their corresponding ML bounds, as in Scenario 2 above. For the remaining WD properties, it also follows the same procedure described in Scenario 2, that is: it assumes a RV shift of 0 km s^{-1} and chondritic abundance ratios for those elements lacking a user-defined initial guess.

(iii) **Step 3: Definition of `cecilia`'s spectral model**: After loading the observed spectrum into `cecilia` together with (adjusted) preliminary estimates of the 14 WD parameters, the user is asked to choose the free and fixed parameters of their `cecilia` spectral model. If certain parameters are fixed, our code freezes their initial value to the user's guess or, if not provided, to `cecilia`'s estimated value.

(iv) **Step 4: Adjusting the format of the input data**: The last step before initiating our fitting procedure is to adjust the WD spectrum and its stellar labels to a format that `cecilia` can understand. To this end, our code automatically implements the following data-processing techniques:

(a) *Normalization of Initial Guess Labels*: First, it normalizes the guess stellar labels using the minimum and maximum values of the synthetic labels considered during training (see Section 3.2). Given that `cecilia` was trained in normalized space, this step is crucial to ensure that our code can effectively process and interpret the user's input data.

(b) *Check of Spectral Units*: Next, `cecilia` verifies that the observed stellar flux is expressed in terms of the calculated flux at the surface of the WD, i.e. F_ν, where $[F_\nu]$=erg/(s Hz cm^2 sr). Although this check is not of paramount importance due to the slope and offfset correction discussed in step (vi) below, F_ν is the unit of choice of `cecilia`'s atmosphere models (e.g. see Fig. 8) and is therefore the preferred notation of our pipeline. In reality, however, real spectra of WDs are typically not expressed in units of F_ν, but in f_λ, where $[f_\lambda]$=erg/(s cm^2 Å). As a consequence, our code integrates the possibility of transforming the observed stellar flux from f_λ to F_ν with

$$F_\nu = \left(\frac{f_\lambda \cdot w_{obs}^2}{c}\right) \cdot \left[\frac{1}{4}\left(\frac{d}{r}\right)^2\right], \tag{17}$$

where c is the speed of light (in [Å/s]), r is the radius of the WD, d is its distance from Earth, and the second term represents the inverse solid angle of the WD (in [1/sr]). In Section 5, we use the *Gaia* DR3 data base (Gaia Collaboration 2016, 2023) to determine the values of r and d.

(c) *Spectrum Cropping*: Lastly, our code examines the observed wavelength array to make sure that it falls within `cecilia`'s spectral coverage between 3000 and 9000 Å. Any regions outside these boundaries are automatically excluded from our fitting analysis (e.g. see Fig. 12).

(v) **Step 5: `cecilia`'s High-resolution spectral prediction**: Next, our code feeds the normalized stellar labels into `cecilia`'s trained FT FCNN2 architecture to quickly generate a normalized synthetic spectrum for the WD (λ_{synth}, f_{synth}). At this stage, `cecilia`'s

[8] https://github.com/segasai/astrolibpy/blob/master/mpfit/mpfit.py

[9] Differential evolution is a genetic algorithm designed to find an optimal distance and direction for a jumping distribution from one MCMC chain to another.

[10] A future improvement to `cecilia` will be to include a trained NN to estimate the effective temperature and surface gravity of the polluted WD from existing photometric observations (see Section 6). In the meantime, we assume that the user has good external constraints on T_{eff} and log g.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

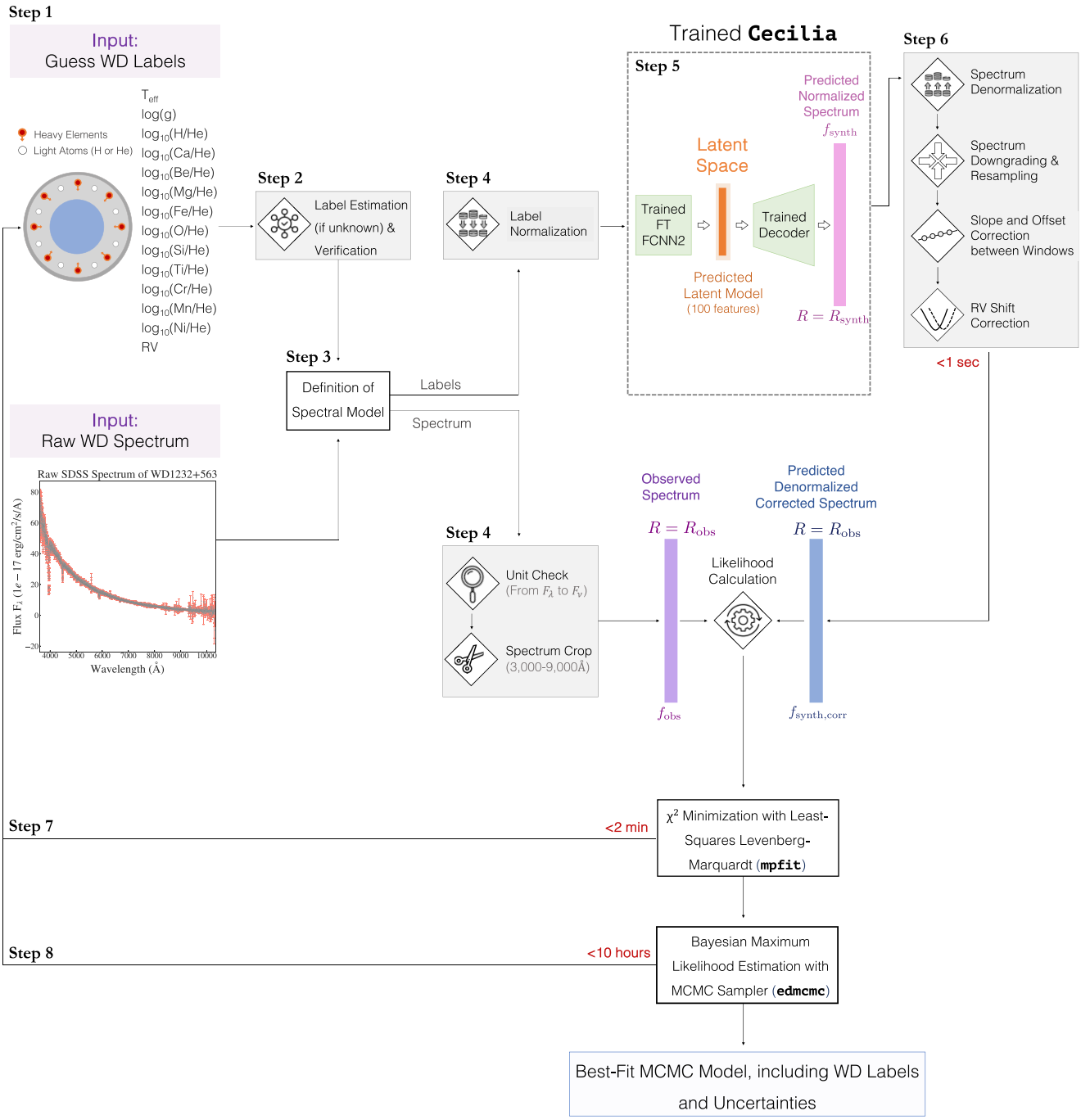


Figure 9. A schematic view of our fitting methodology using `cecilia` and a combination of frequentist and Bayesian statistical techniques. First, the user uploads the spectrum of their polluted WD onto `cecilia`, together with a list of 14 stellar properties [$T_{\rm eff}$, log g, 11 elemental abundances, and a RV shift; see step (i) in Section 4.3]. If these properties are unknown, `cecilia` estimates initial guesses as explained in step (ii); afterwards, it verifies that the labels are within `cecilia`'s ML bounds. Next, the user defines the free and fixed parameters of their spectral model (step (iii)). Then, our code adapts the input spectrum and stellar properties to `cecilia`'s format (step (iv)); this includes normalizing the guess labels and excluding any regions of the observed spectrum outside of `cecilia`'s coverage between 3000 and 9000 Å. After this data processing step, our code feeds the normalized labels to the trained FT FCNN2 network, which then takes <1 s to generate a normalized, high-resolution synthetic spectrum (step (v)). This prediction is subsequently denormalized and adjusted to the properties of the input spectrum (step (vi)) before being optimized with two techniques: a non-linear least-squares χ^2 minimization algorithm (step (vii)), and a Bayesian MCMC (step (viii)). Using 1 GPU from the MIT Satori Supercomputer, our fitting procedure produces a full spectroscopic solution in less than about 10 hours, including the stellar labels of the polluted WD with their corresponding uncertainties. Source of icons: *Flaticon*.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

synthetic spectrum shares the resolving power of the atmosphere models described in Section 3.1.2, i.e. $R_{\rm synth}\approx$50 000. Therefore, assuming that the resolving power of the observed WD spectrum is lower than that of our atmosphere models ($R_{\rm obs}\ll R_{\rm synth}$), `cecilia`'s prediction cannot be optimized yet.

(vi) **Step 6: Processing of `cecilia`'s High-resolution spectral rrediction**: To enable a meaningful comparison between the output of `cecilia`'s FT FCNN2 architecture and the observed input spectrum, our code performs the following automated and sequential operations:

(a) *Denormalization of `cecilia` 's synthetic prediction*: First, it denormalizes `cecilia`'s prediction with the minimum and maximum flux values of the synthetic models used during training. This transformation brings `cecilia`'s synthetic spectrum to the non-normalized space of the observed WD spectrum.

(b) *Smearing and resampling*: Secondly, our code smears and reasmples `cecilia`'s denormalized prediction to the resolving power ($R_{\rm obs}$) and wavelength grid ($\lambda_{\rm obs}$) of the input spectrum, respectively. To achieve this, it downgrades `cecilia`'s synthetic spectrum from $R_{\rm synth}$ to $R_{\rm obs}$ by convolution with a Gaussian smoothing function. Then, it resamples `cecilia`'s prediction by smearing the latter onto the wavelength grid of the observed spectrum using `Python`'s one-dimensional linear interpolator `numpy.interp`.

(c) *Slope and Offset Correction*: Thirdly, we correct for any spectral 'jumps' in `cecilia`'s prediction caused by the training of our ML pipeline in independent windows of 200 Å (e.g. see Fig. 8). For this task, we start by defining an *observed* normalization function $\mathcal{N}_{\rm obs}$ from the ratio between `cecilia`'s prediction and the observed flux,

$$\mathcal{N}_{\rm obs}{}^{i} = \frac{f^{i}_{\rm synth}}{f^{i}_{\rm obs}}, \tag{18}$$

where $f^{i}_{\rm synth}$ is `cecilia`'s synthetic flux at wavelength i, and $f^{i}_{\rm obs}$ is the observed stellar flux. Then, we use a weighted linear least-squares algorithm to fit $\mathcal{N}_{\rm obs}$ as a function of the observed wavelength array, $\lambda_{\rm obs}$. From this analysis, we obtain a *model* normalization function $\mathcal{N}_{\rm model}$ (e.g. see purple line in the second panel of Fig. 8), which we then multiply to `cecilia`'s prediction to minimize the jumps between consecutive windows. We note that `cecilia` can do a slope and offset correction with an n-th degree polynomial as well.

(d) *Radial velocity shift*: Finally, our code applies a radial velocity shift to `cecilia`'s denormalized, smeared, resampled, and slope-corrected prediction. In particular, if c is the speed of light, and $v_{\rm obs}$ is the assumed RV shift of the WD, the total wavelength shift of the spectral lines can be calculated with

$$\delta\lambda_{\rm synth} = \frac{\lambda_{\rm synth}\cdot v_{\rm obs}}{c} \equiv \lambda'_{\rm synth} - \lambda_{\rm synth}. \tag{19}$$

From the above, we determine the shifted wavelength array of `cecilia`'s synthetic prediction with

$$\lambda'_{\rm synth} = \lambda_{\rm synth}\cdot\left(1+\frac{v_{\rm obs}}{c}\right). \tag{20}$$

Using this new wavelength grid, we then perform a cubic spline interpolation to obtain the shifted flux and flux error arrays (equation 21). In the absence of flux points within a certain wavelength range, our function simply extrapolates the data. Therefore, for a shifted wavelength array $\lambda'_{\rm synth}$, the final products of our interpolation are:

$$\begin{aligned} f_{\rm synth}(\lambda_{\rm synth}) &\rightarrow f'_{\rm synth}(\lambda'_{\rm synth}) \\ f_{\rm err,synth}(\lambda_{\rm synth}) &\rightarrow f'_{\rm err,synth}(\lambda'_{\rm synth}) \end{aligned}. \tag{21}$$

In the remaining discussion, we will denote `cecilia`'s smeared, resampled, slope-corrected, and RV-shifted synthetic wavelength and flux as $w_{\rm synth,\,corr}$ and $f_{\rm synth,\,corr}$, respectively.

(vii) **Step 7: Initial `mpfit` χ^2 optimization**: After concluding steps (i)-(vi), our code initiates the first phase of our optimization procedure, namely: a preliminary non-linear least-squares fit implemented with the MPFIT package. The goal of MPFIT is to quickly (<2 min) identify good values for the model parameters which can then be used as initial guesses for our MCMC. To achieve this, MPFIT minimizes the chi-squared statistic (χ^2) of `cecilia`'s corrected prediction,

$$\chi^2 = \sum_{i=1}^{N_{\rm points}}\left[\frac{f^{i}_{\rm obs} - f^{i}_{\rm synth,corr}}{f^{i}_{\rm obs,err}}\right]^2, \tag{22}$$

where $N_{\rm points}$ is the total number of points in the observed input spectrum, $f^{i}_{\rm synth,corr}$ is `cecilia`'s corrected synthetic flux at wavelength i, $f^{i}_{\rm obs}$ is the observed stellar flux, and $f^{i}_{\rm obs,err}$ is the uncertainty of the latter.

As illustrated in Fig. 9, MPFIT computes the χ^2 metric in equation (22) as follows: first, it provides the normalized stellar labels to `cecilia`, which then invokes its trained FT FCNN2 architecture to rapidly generate a preliminary normalized, high-resolution ($R_{\rm synth}$), synthetic spectrum ($\lambda_{\rm synth}, f_{\rm synth}$). Next, `cecilia`'s prediction is adjusted with the techniques described in step (vi), which allows MPFIT to compare the observed input spectrum to `cecilia`'s corrected prediction. From this comparison, MPFIT calculates the χ^2 metric with equation (22) and repeats the entire optimization procedure until identifying a set of optimal model parameters. For these best-fitting labels, MPFIT also provides their covariance matrix, which can then be used to derive initial estimates of their uncertainties.

(viii) **Step 8: Final `MCMC` exploration**: The second phase of our fitting routine employs the best-fitting results of MPFIT to initialize an MCMC sampler and obtain more robust parameter uncertainties. In contrast to MPFIT, which uses frequentist (or 'classical') statistics to minimize the χ^2 value of `cecilia`'s predictions and determine best-fitting values with fixed error bars, an MCMC relies on Bayesian inference to maximize the likelihood $\mathscr{L}$ of the model parameters and evaluate their full posterior probability distributions. These 'posteriors' are proportional to the information contained in the observations as well as to any pre-conceived beliefs or assumptions about the model parameters (or 'priors'). Mathematically, the posterior probability (π) of observing the model parameters ($\boldsymbol{\theta}$) given an observed spectrum ($f_{\rm obs}$) is encoded in Bayes' Theorem

$$\pi\left(\boldsymbol{\theta}\,|\,f_{\rm obs}\right) \propto \mathscr{L}(f_{\rm obs}|\boldsymbol{\theta})\cdot\pi_0\left(\boldsymbol{\theta}\right), \tag{23}$$

where $\mathscr{L}$ is the likelihood function, and $\pi_0\left(\boldsymbol{\theta}\right)$ is the prior probability distribution given an array of model parameters $\boldsymbol{\theta}$. Assuming that the likelihood of `cecilia`'s model parameters can be modelled with a Gaussian distribution, our MCMC is designed to explore and map the multidimensional likelihood function, given by

$$\begin{aligned}\ln(\mathscr{L}) = &-\frac{1}{2}\sum_{i=1}^{N_{\rm points}}\left[\frac{f^{i}_{\rm obs} - f^{i}_{\rm synth,corr}}{f^{i}_{\rm obs,err}}\right]^2 \\ &+\sum_{j=1}\pi_{0,{\rm phot_j}} + \sum_{k=1}\pi_{0,{\rm chondr_k}}.\end{aligned} \tag{24}$$

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Above, we adopt the logarithmic scale for numerical and computational efficiency. In particular, the use of the log-likelihood is driven by three main factors: first, it simplifies and speeds up the calculation of derivatives; secondly, it penalizes unlikely fits by assigning them a very high χ^2 value and thus a very small log-likelihood; and thirdly, it does not entail any loss of information during the fit because both $\mathscr{L}$ and $\ln(\mathscr{L})$ have a maximum in the same location.

In equation (24), the last two terms represent the prior distributions in our `cecilia` model. The first term ($\sum_{j=1} \pi_{0,\mathrm{phot_j}}$) assumes that the user has reasonable estimates for the effective temperature and surface gravity of the WD from an external photometric fit. The second term ($\sum_{k=1} \pi_{0,\mathrm{chondr_k}}$) is optional and can be used to impose chondritic abundance priors on any metal of interest based on the assumed calcium abundance of the star and the chondritic values of Asplund et al. (2009). These physically motivated priors are useful to constrain the abundances of those heavy elements that are barely visible or undetectable in the observed WD spectrum.

After defining the prior distributions of our likelihood function, we execute our MCMC following the same structure presented in step (vii); this time, however, we do not minimize the χ^2 statistic of `cecilia`'s predictions, but maximize the log-likelihood function $\ln(\mathcal{L})$. To run our MCMC, we use an ensemble of n_{walkers} walkers and n_{draws} draws, discarding the first 20 per cent of the draws in a 'burn-in' phase to exclude non-stationary values.[11] With 1 Satori GPU and suitable combinations of these three MCMC hyperparameters, we find that `cecilia` can produce an optimal spectroscopic model in less than 10 h, with the time complexity of this process depending primarily on the the computational time associated to the predictions of the FT FCNN2, rather than the quality and/or resolution of the observed input spectrum. In addition to providing a best-fitting spectral solution, our MCMC also determines the conditional probabilities of the model parameters, also known as 'posterior probability distributions'. From the these posteriors, the MCMC computes the median values of each label, together with their 16th, 50th, and 84th percentiles. In this work, we report our MCMC results using the median statistic and its 1σ confidence interval.

At the end of the MCMC, we also examine the quality of `cecilia`'s fit with a variety of diagnostic figures and performance metrics. Among them, we generate scatterplot matrices (or 'corner plots') to visualize the two-dimensional 'marginal' posterior of a given pair of model parameters alongside their one-dimensional marginal histogram projection. For a parameter θ_i, the marginal posterior $\pi(\theta_i|f_{\mathrm{obs}})$ can be computed with

$$\pi\left(\theta_i \,|\, f_{\mathrm{obs}}\right) = \int \pi\left(\boldsymbol{\theta} \,|\, f_{\mathrm{obs}}\right) d\boldsymbol{\theta}_{j\neq i}, \tag{25}$$

where j iterates over all the parameters. Finally, we test the convergence of the chains using the Gelman–Rubin (GR) potential scale reduction factor $\hat{R}$ for each model parameter (Gelman & Rubin 1992). This metric compares the within-chain-variance with the scatter of the points between the chains. In this work, we assume convergence of a model parameter if its GR value is lower than 1.003 (Vats & Knudson 2018; Vehtari et al. 2021).

5 RESULTS

In this section, we investigate the performance of `cecilia` using two types of observations: noiseless synthetic models, and a real spectrum of a well-characterized polluted WD. By testing `cecilia` under ideal and realistic noisy conditions, we aim to understand its retrieval accuracy, its usage limits, and its suitability for the study of different WDs and astronomical data sets.

5.1 Sensitivity analysis with synthetic observations

One of the main disadvantages of ML techniques is the lack of their explainability, i.e. the difficulty of interpreting their predictions through a human lens. Despite notable progress in recent years to make ML models more transparent and easier to understand (Gunning et al. 2019; Vilone & Longo 2020; Saranya & Subhashini 2023), NNs are still perceived as 'black boxes', which can sometimes hinder our ability to trust their predictions and their decision-making process. In the context of `cecilia`, we have sought to address this problem by evaluating the baseline performance of its trained architecture under well-defined and optimal conditions.

To conduct our study, we selected 1000 random synthetic spectra from our test set; in other words, we only considered atmosphere models that `cecilia` had *never* seen during training. We then used `cecilia`'s FT FCNN2 architecture and the `mpfit` library to model the spectra. For this task, we fixed T_{eff} and $\log g$ to their true values,[12] and only fitted the 11 elemental abundances listed in Table 1 (i.e. from $\log_{10}(\mathrm{H/He})$ to $\log_{10}(\mathrm{Ni/He})$). To better understand the retrieval accuracy of `cecilia` under perfect conditions, we also decided to adopt the true stellar labels of each synthetic spectrum as our input guesses for our optimization routine. We then executed `mpfit` with 30 iterations, obtaining a best-fitting model in about 1 min per spectrum. For computational efficency purposes, we did not complement our χ^2 minimization analysis with a full MCMC, and instead treated the `mpfit`'s results as our final solutions.

After performing our fits, we determined the residuals $\vec{r}$ between their true labels and `cecilia`'s predictions. Our results are presented in Fig. 10 as a function of photospheric composition, with the colorbar indicating the true effective temperature of the synthetic spectra. Recognizing that the visibility of a metal is strongly dependent on T_{eff}, the general trend observed in Fig. 10 is that `cecilia` performs very well for synthetic WDs with high levels of metal pollution. As metal abundances decrease, the accuracy of our code is also lower because spectral signatures become less strong and harder for `cecilia` to mimic (typically less than roughly 1 per cent with respect to the flux continuum, except for a few outliers). In practice, this problem is mitigated by the fact that subtle, shallow lines are challenging to detect observationally as well.

In Fig. 10, we show all the independently varied elements during `cecilia`'s training except for Beryllium, which our code could not reliably constrain. This light metal has a relatively low abundance in the present-day solar photosphere and in meteoritic CI chondrites (Lodders 2003; Asplund et al. 2009), which could also be the case for most polluted WDs. With only two prominent resonance lines of Be in the near UV at 3,130.42 and 3,131.07 Å (Kramida et al. 2022), its abundance is challenging to determine unless there is a substantial amount of this element in the photosphere of the WD (e.g. Klein et al. 2021). Therefore, given the difficulty of `cecilia` to model Be, we recommend fixing its abundance to a chondritic ratio before using `cecilia` to fit the spectrum of a polluted WD.

[11] The walkers and draws represent, respectively, the number of chains in our MCMC, and the number of steps that are evaluated in each chain.

[12] We chose to keep T_{eff} and $\log g$ fixed because these parameters are usually constrained from photometric observations.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

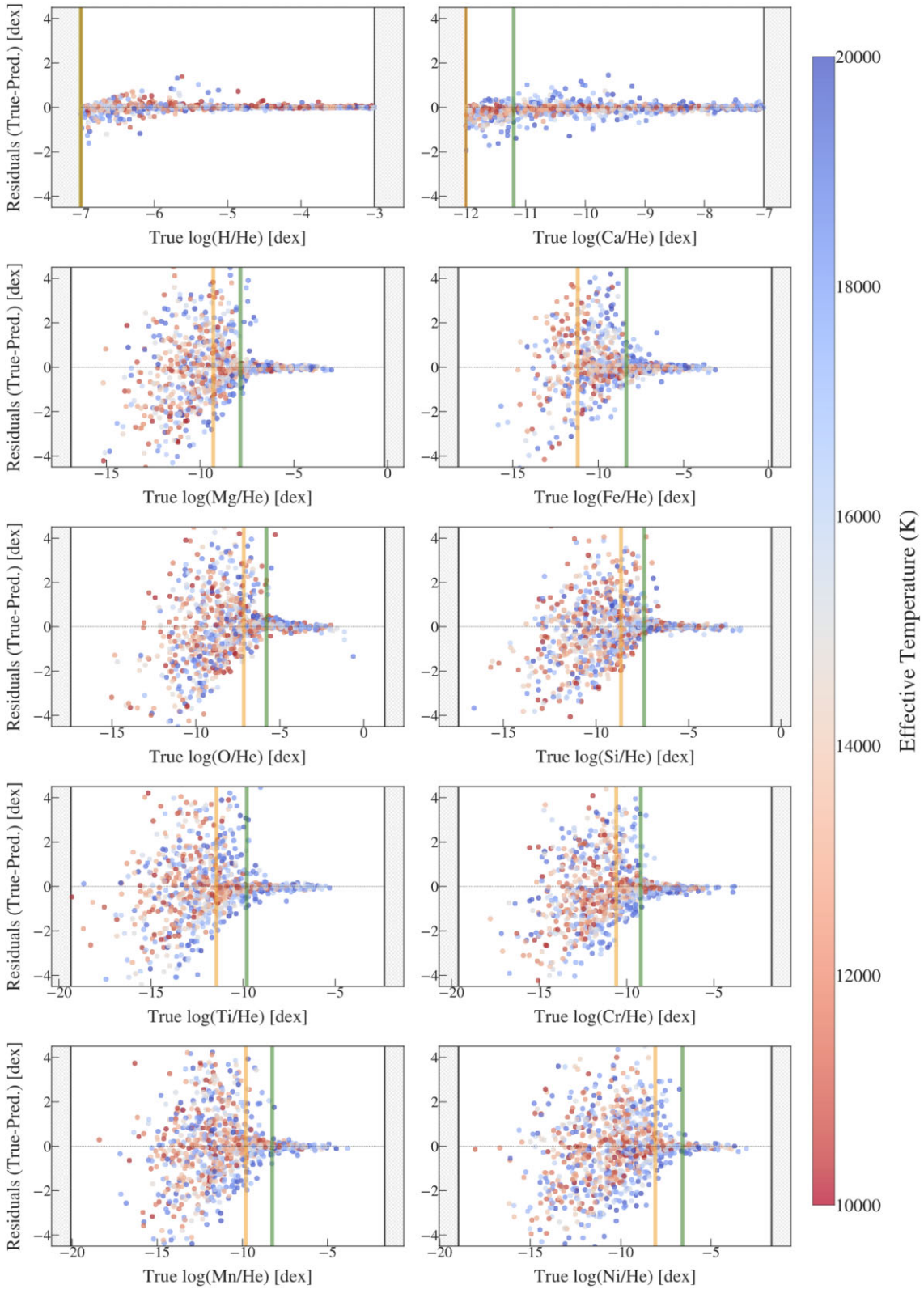


Figure 10. Residuals of `cecilia` as a function of true elemental abundance for 1000 noiseless atmosphere models from our test set (see Section 5.1). To conduct our analysis, we fixed $T_{\rm eff}$ and log g to their true values, and only fitted the 11 abundances considered in this work (i.e. from $\log_{10}$ (H/He) to $\log_{10}$ (Ni/He) in Table 1). The vertical lines in green and orange denote, respectively, the boundaries where `cecilia` achieves a predictive accuracy lower than 0.1 dex and 0.2 dex as calculated by equation (26) (see Table 3). The black lines represent the minimum and maximum ML bounds of `cecilia` (see Table 1), with the grey hatched regions symbolizing abundances outside of `cecilia`'s allowed parameter space. This figure does not show any results for $T_{\rm eff}$ and log g, which were fixed parameters during our fits. It also does not include our residuals for Be due to the poor performance of our code with this metal.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Table 3. Abundance accuracy of `cecilia` as determined from an analysis of 1000 noiseless synthetic spectra from the test set (see Section 5.1). We also show `cecilia`'s allowed ML bounds as well as the abundance ranges detected in He-rich polluted WDs with effective temperatures between 10 000 $\leq T_{\rm eff} \leq$ 20 000 K (source: default column in the Montreal White Dwarf Data base, MWDD; Dufour et al. 2016). The dagger symbol (†) indicates that `cecilia`'s performance is stable over the entire ML range of the parameter, while the star symbol (⋆) is used to highlight that `cecilia` cannot constrain Be to within 0.2 dex for synthetic data, hence suggesting a potentially worse performance for real, noisy observations.

	Accuracy bounds		`cecilia`'s ML bounds		MWDD	
Label	≤0.1 dex	≤0.2 dex	Min.	Max.	Min.	Max.
$\log_{10}$(H/He)	$\geq -7^{\dagger}$	$\geq -7^{\dagger}$	−7	−3	$<-6.5^{a}$	-2.9^{e}
$\log_{10}$(Ca/He)	≥ -11.20	$\ll -12^{\dagger}$	−12	−7	-10.8^{b}	-6.6^{f}
$\log_{10}$(Mg/He)	≥ -7.85	≥ -9.30	−16.89	−0.17	-8.5^{c}	-5.9^{g}
$\log_{10}$(Fe/He)	≥ -8.35	≥ -11.20	−18.20	0.18	-8.2^{c}	-5.6^{g}
$\log_{10}$(O/He)	≥ -5.80	≥ -7.15	−17.46	1.25	$<-6.6^{c}$	-5.1^{g}
$\log_{10}$(Si/He)	≥ -7.40	≥ -8.65	−17.42	−0.51	-8.0^{c}	-5.9^{g}
$\log_{10}$(Ti/He)	≥ -9.80	≥ -11.45	−19.35	−2.32	-9.5^{d}	-8.6^{g}
$\log_{10}$(Be/He)		$\gg -5.61^{\star}$	−23.85	−5.61	–	–
$\log_{10}$(Cr/He)	≥ -9.20	≥ -10.60	−19.61	−1.71	-9.3^{d}	$<-7.7^{h}$
$\log_{10}$(Mn/He)	≥ -8.25	≥ -9.80	−20.08	−1.65	-9.8^{d}	-8.6^{g}
$\log_{10}$(Ni/He)	≥ -6.60	≥ -8.10	−18.96	−1.66	-8.8^{d}	-7.0^{g}

References: [a] Genest-Beaulieu & Bergeron (2019), [b] Coutu et al. (2019), [c] Wolff, Koester & Liebert (2002), [d] Xu et al. (2013), [e] Koester et al. (2005), [f] Xu et al. (2016), [g] Fortin-Archambault, Dufour & Xu (2020), and [h] Desharnais et al. (2008). Last MWDD Access: November 2023.

In the remaining sections, we follow this approach to generate our results.

Using the residuals from Fig. 10, we then calculated the systematic errors associated to `cecilia`'s `mpfit` predictions. To this end, we estimated the retrieval accuracy of our pipeline from the ratio between the Median Absolute Deviation (MAD) of the residuals and the median of a standard Gaussian distribution,

$$\text{Accuracy} = \frac{\text{median}\left(|\text{r}_{\text{i}} - \text{median}(\vec{r}\,)|\right)}{0.67} \equiv \frac{\text{MAD}}{0.67}, \tag{26}$$

where the MAD provides a robust statistical metric against outliers. The green and orange vertical lines in Fig. 10 show the lowest possible abundance values that `cecilia` can retrieve with an average accuracy lower than 0.1 and 0.2 dex, respectively. These bounds are summarized in Table 3 and can be used to assess the usefulness and robustness of `cecilia` for the study of a real polluted WD.

In Fig. 11, we also illustrate `cecilia`'s estimated abundances as a function of synthetic effective temperature, with the colorbar representing the residuals of its predictions. From this figure, we find that `cecilia`'s performance is overall quite stable over the range between 10 000 and 20 000 K. However, its residuals are slightly higher for high-temperature WDs, which is consistent with the fact that hotter sources are more opaque, thus making it more difficult to detect weak signs of metal pollution (e.g. Zuckerman et al. 2010). It is also interesting to highlight that `cecilia`'s predictive power is particularly good for $\log_{10}$(H/He) and $\log_{10}$(Ca/He), as demonstrated by the low error in their predicted abundances. The level of accuracy for these two stellar labels is possibly due to the depth of their absorption lines compared to those of other metals (more than 1 per cent deep relative to the flux continuum, even for the lowest possible abundances of H and Ca at the maximum and minimum allowed values of $T_{\rm eff}$ and log g).

5.2 Performance with real observations

In this section, we test and validate `cecilia` against real spectroscopic observations of the He-rich, heavily polluted white dwarf WD 1232+563. This star was discovered by Eisenstein et al. (2006), who compiled a catalog of 19 712 spectroscopically confirmed WDs from the Sloan Digital Sky Survey (SDSS; York et al. 2000). More recently, it was studied by Coutu et al. (2019) and Xu et al. (2019) at low- and high-resolution: the former conducted a spectroscopic analysis of 1023 He-atmosphere polluted systems observed by SDSS, while the latter used data from the Keck HIRESb and ESI spectrographs to perform a detailed characterization of WD 1232+563. From these literature studies, the physical and chemical properties of WD 1232+563 appear to fall well within the allowed bounds of `cecilia` (see Table 4). Therefore, this WD constitutes an excellent candidate to understand the behaviour of our pipeline with real astrophysical observations.

To perform our sensitivity analysis, we choose to use `cecilia` to fit the low-resolution, flux-calibrated spectrum of WD 1232+563 from the SDSS DR18 data base (plate = 8232, fiber = 84; see Fig. 12).[13] Given that SDSS observations have a variable resolving power of R=1500 at 3800 Å and R=2500 Å at 9000 Å, we adopt a mean value of $R_{\rm obs}\approx$2000 in our fitting procedure. Moreover, to evaluate `cecilia`'s behaviour with unprocessed real data, we do not remove any problematic data points from the spectrum, such as bad flux points or instrumental artefacts.

As explained in Section 4.3, the first step in our fitting routine is to load the SDSS spectrum of WD 1232+563 into `cecilia`, together with reasonable estimates for the WD's effective temperature, surface gravity, elemental abundances (H, Ca, Mg, Fe, O, Si, Ti, Be, Cr, Mn, and Ni), and RV shift. To generate our list of stellar parameters, we adopt the $T_{\rm eff}$ and log g values obtained by Coutu et al. (2019) from existing photometric observations; the calcium photospheric abundance derived by the same authors from the Ca II H&K absorption lines at air wavelengths 3934 and 3369 Å (Kramida et al. 2022); and the elemental abundances of Xu et al. (2019) for the remaining heavy elements. We also assume a zero RV shift as our initial guess for the RV term in our spectral model. In our analysis, we choose to treat all 14 stellar properties as free model parameters.

To begin with, we execute our `mpfit` least-squares χ^2 minimization routine, restricting the label parameter space to `cecilia`'s ML bounds, requiring a maximum of 50 iterations, and using gradient

[13] https://skyserver.sdss.org/dr18/.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

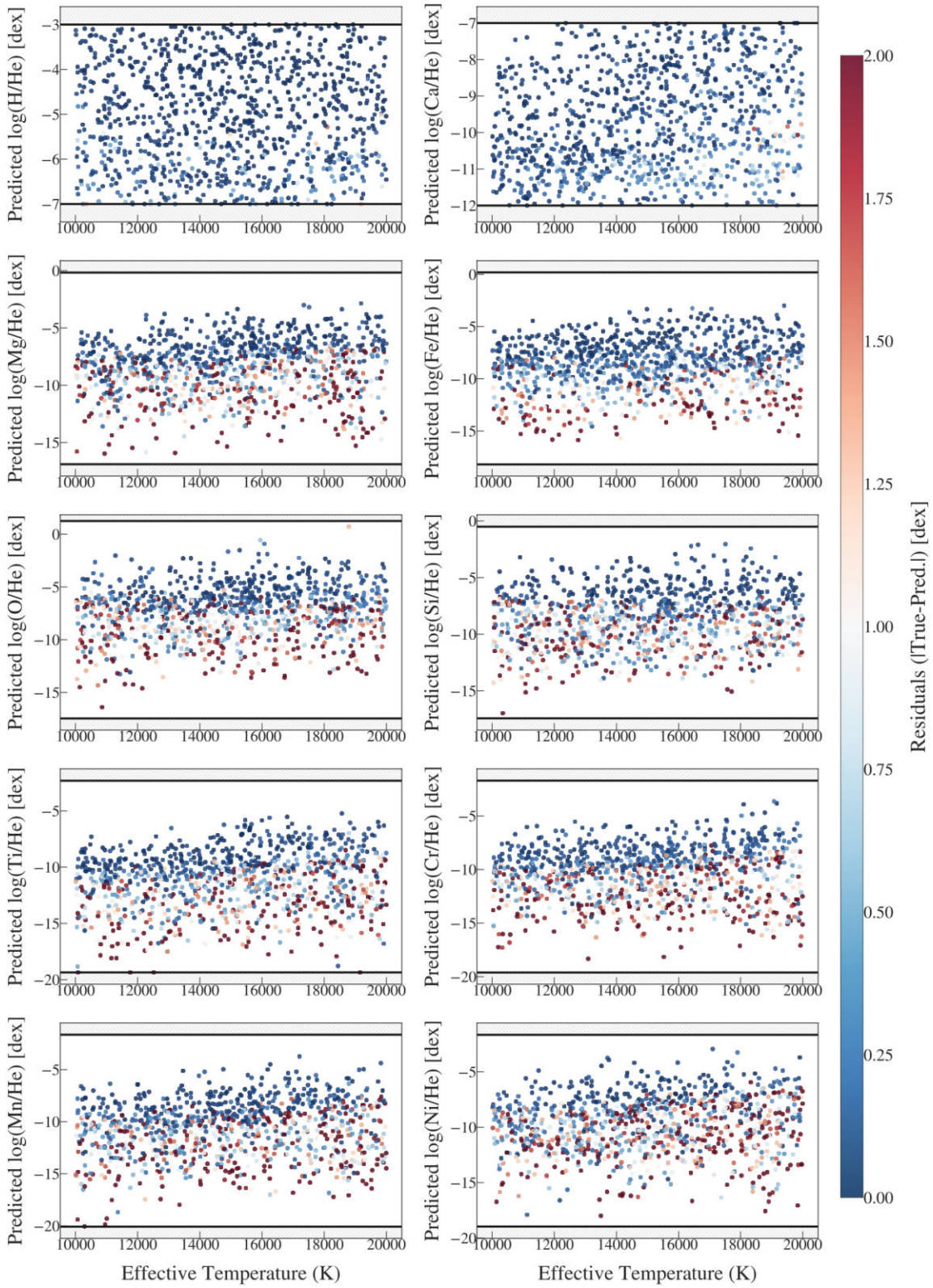


Figure 11. `cecilia`'s estimated elemental abundances as a function of synthetic effective temperature. The colourbar shows the residuals of its predictions, which are slightly higher for hotter WDs. Similarly to Fig. 10, the black lines denote the minimum and maximum allowed ML bounds of Cecilia (see Table 1), while the grey hatched areas represent photospheric compositions outside of `cecilia`'s range. From this figure, we find that `cecilia`'s predictions are particularly good for $\log_{10}$ (H/He) and $\log_{10}$ (Ca/He).

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Table 4. Main astrophysical properties of WD 1232+563.

Property	Value	Ref.
	Other target names	
Gaia DR3	1 571 584 539 980 588 544	[1]
SDSS	J123432.65+560643.1	[2]
GALEX	J123432.6+560642	[3]
	Astrometric properties	
R.A. (J2020; h:m:s)	12:34:32.676	[1]
Dec (J2020; d:m:s)	+56:06:43.034	[1]
Parallax (mas)	5.817 ± 0.095	[1]
Distance (pc)	171.92 ± 3.00	[1]
$\mu_{\rm R.A.}$ (mas y^{-1})	−72.884 ± 0.089	[1]
$\mu_{\rm Dec.}$ (mas y^{-1})	−0.223 ± 0.094	[1]
	Photometric properties	
Gaia $G_{\rm mag}$	18.05 ± 0.01	[1]
SDSS $g_{\rm mag}$	18.27 ± 0.01	[2]
Pan-STARRS $g_{\rm mag}$	17.95 ± 0.01	[4]
	Physical and chemical properties	
$T_{\rm eff}$ [K]	11,787 ± 423	[6]
log g [cgs]	8.30 ± 0.06	[6]
Mass [$M_\odot$]	0.77	[6]
Cooling Age [Gyr]	0.64	[6]
Atm. Composition	He	[5, 6]
Spectral Type	DBZA	[5, 6]

References: [1] Gaia Collaboration (2016, 2023), [2] York et al. (2000); Almeida et al. (2023), [3] Martin et al. (2005), [4] Kaiser et al. (2010), [5] Kleinman et al. (2013), and [6] Coutu et al. (2019).

step sizes[14] of 0.01 K for temperature, 0.01 cgs for surface gravity, 0.01 dex for all the elemental abundances, and 0.0001 km s^{-1} for the RV shift. With these hyperparameters, `mpfit` generates an optimal spectroscopic solution with its corresponding 14 stellar labels in less than 1.80 min. Next, we run a final MCMC with $n_{\rm walkers} = 50$ walkers, $n_{\rm draws} = 5000$ draws, and $n_{\rm burn} = 0.2 \cdot n_{\rm draws}$=1000 draws. We initialize the walkers in Gaussian balls with standard deviations of 0.01 K for $T_{\rm eff}$, 0.01 cgs for log g, 0.01 dex for the 11 elemental abundances, and 0.0001 km s^{-1} for the RV shift. We then define our likelihood function with equation (24), incorporating two types of prior distributions: first, we adopt weak photometric priors of 500 K and 0.1 cgs on $T_{\rm eff}$ and log g based on the uncertainties of these parameters in Xu et al. (2019); and second, we assume chondritic abundance priors for Be, Cr, Mn, and Ni, which are elements that we cannot visually and confidently detect in the SDSS spectrum.[15] For these chondritic labels, we impose a prior width of 0.5 for Be, Cr, and Ni, with a stricter width of 0.25 for Mn to ensure chain convergence.

With the exception of Be, Cr, Mn, and Ni, for which we use chondritic abundance ratios as initial guesses, we take the results of `mpfit` to initialize our MCMC sampler. This approach, combined with the `edmcmc` configuration mentioned above, generates a final spectral model in about 10 h with a converged GR statistic lower than 1.008 for all our model parameters. In Table 5, we also summarize the MCMC best-fitting labels of WD 1232+563, excluding the four undetected heavy elements with chondritic priors (i.e. Be, Cr, Mn, and Ni). We also report the systematic ($\sigma_{\rm sys,ML}$) and statistical errors ($\sigma_{\rm stat,\,MCMC}$) of each model parameter, obtained respectively from our sensitivity analysis of synthetic spectra (see Section 5.1) and our MCMC fit. The former measure the errors of `cecilia`'s predictions arising from the ML-based spectral interpolation process, while the latter represent the errors of the SDSS data and are hence observational in nature. Assuming Gaussian and uncorrelated errors, we can approximate the total uncertainty of a model parameter ($\sigma_{\rm tot}$) by adding these two sources of errors in quadrature (together with any other systematic errors, or $\sigma_{\rm other}$, coming from the atmosphere models, which we do not consider in this work),

$$\sigma_{\rm tot} = \sqrt{\sigma^2_{\rm stat,MCMC} + \sigma^2_{\rm sys,ML}(+\sigma^2_{\rm other})}. \quad (27)$$

Using the above expression, we calculate the total uncertainty of each stellar label and provide it in Table 5. As a conservative measure, we assume that the MCMC has underestimated the statistical error of those abundance parameters with $\sigma_{\rm tot}$<0.15 dex, where this threshold represents an approximate SDSS noise floor.[16] For these abundances (flagged with a star symbol in Table 5), we replace the total uncertainty calculated in this work by $\sigma_{\rm tot}$=0.15 dex.

In Fig. 13, we also show our `mpfit` and MCMC solutions as well as their corresponding 'Observed-Calculated' (O-C) flux residuals. To complement this figure, Fig. 14 presents the RV-shifted MCMC model for different regions of the SDSS spectrum with clear metallic absorption lines, together with two additional models obtained with the same $T_{\rm eff}$ and log g, but with the abundances of the detected elements changed by $3\sigma_{\rm stat,\,MCMC}$. Finally, Fig. 15 presents a corner plot of the posterior probability distributions of the model parameters as a measure of MCMC chain convergence.

Based on Table 5, and taking the properties of WD 1232+563 from Xu et al. (2019) as our assumed ground truth, our results demonstrate that `cecilia` can accurately retrieve the majority of parameters in our model. This includes $T_{\rm eff}$ and log g – for which we use weak photometric priors – as well as the elemental abundances of 6 heavy elements with visible absorption lines in the SDSS spectrum (Ca, Mg, Fe, O, Si, and Ti; see Fig. 14). Among these metals, a few of them (e.g. Si, Ti, and O) are weak detections, as evidenced by their small observational signatures and their high statistical uncertainties relative to those of other detected elements. For these elements, we might not have been able to claim a confident detection without knowledge of the higher-resolution and signal-to-noise spectra from Xu et al. (2019). However, given that our current knowledge of this star confirms their existence, it is promising to see that our pipeline can estimate abundance values consistent with those in the literature. In the future, `cecilia`'s detection of weak lines in low-resolution data can be a sign that a WD deserves spectroscopic follow-up at high resolution and high signal-to-noise.

With respect to those elements with no clear spectral signatures (Be, Cr, Mn, and Ni), `cecilia` can marginalize over the uncertainty in the true values of their abundances thanks to the use of chondritic priors in the likelihood function. These chondritic priors limit the values of the parameters explored in our fit, which in turn facilitates MCMC convergence. We note that the assumption of chondritic priors in our MCMC is effectively equivalent to the classical convention of including undetectable metals in atmosphere models with their abundances fixed to a certain value (e.g. chondritic levels).

Another interesting result from Table 5 is `cecilia`'s retrieved calcium abundance ($\log_{10}$ (Ca/He)≈-7.44), which is more consistent

[14] In the `mpfit` code, the 'step size' corresponds to the step size used to calculate gradients during the fitting process. It does not refer to the size of the parameter update at each iteration of the algorithm for a given model parameter.

[15] We note that we also considered using a chondritic prior on Ti, but we ended up discarding this option after realizing that `cecilia` could constrain this metal relatively well.

[16] Our assumed SDSS noise floor of 0.15 dex is a reasonable value based on the authors' experiences with WD spectral modelling.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

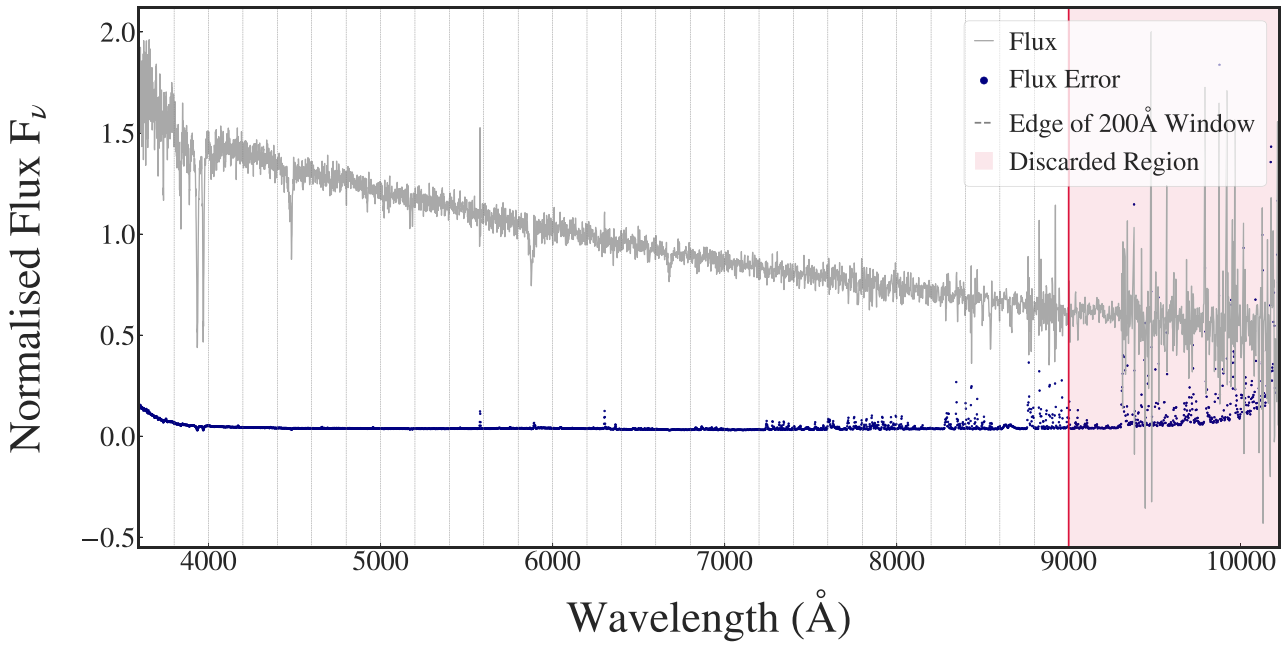


Figure 12. Low-resolution SDSS spectrum of the He-rich, heavily polluted WD WD 1232+563. The normalized F_ν flux and its corresponding observational error are presented in grey and blue, respectively. The vertical lines denote the spectral windows of 200 Å used to train `cecilia` (see Section 4.2). The red shaded area covers the wavelength region excluded in our optimization routine (see Section 5.2).

with that of Coutu et al. (2019; $\log_{10}(\mathrm{Ca/He})_{\mathrm{Coutu}} \approx$-7.41) than with the value reported by Xu et al. (2019) ($\log_{10}(\mathrm{Ca/He})_{\mathrm{Xu}} \approx$-7.69). This discrepancy is understandable given the different optimization strategies used in each paper. More specifically, Coutu et al. (2019) fitted the same SDSS spectrum of WD 1232+563 considered in this work, focusing exclusively on the Ca II H&K doublet and using a maximum likelihood approach similar to that of our code. In contrast, Xu et al. (2019) implemented a line-by-line χ^2 minimization technique of the Ca II H&K doublet and the near-infrared (IR) Ca II triplet in their high-resolution Keck spectrum. For this task, they individually fitted all the Ca lines and weighted them equally to obtain a mean calcium abundance, hence giving more importance to the IR triplet than `cecilia` does. Indeed, our code fits all the Ca lines simultaneously – in combination with the full spectrum – and subsequently weighs them based on their observed strengths, uncertainties, and signal-to-noise. As Fig. 14 shows, the Ca II H&K lines in the SDSS data are significantly stronger than the near-IR Ca II triplet at around 8498, 8542, and 8662 Å. Therefore, similarly to the study of Coutu et al. (2019), they dominate `cecilia`'s fitting procedure, leading to a best-fitting $\log_{10}(\mathrm{Ca/He})$ abundance somewhat higher than the result of Xu et al. (2019). In addition to the different fitting methodologies employed in each paper, it is also possible that the uncertainties on the SDSS flux, which are higher in the red region of the spectrum (likely due to telluric and/or instrumental effects), may have brought our Ca abundance closer to that of (Coutu et al. 2019) than to the result of (Xu et al. 2019).

Besides providing the main physico-chemical properties of WD 1232+563, `cecilia` generates the posterior probability distributions of its model parameters and shows their interdependence in the form of a corner plot (see Fig. 15). With this Bayesian approach, `cecilia` differs from conventional WD analysis techniques, which can only use a limited number of points in their interpolation grid to explore the relationship between two model parameters (e.g. Klein et al. 2011; Xu et al. 2019). In general terms, corner plots reveal possible correlations between two labels if their posterior distributions have a slanted and oval-like contour shape. For WD 1232+563, this seems to be the case for multiple stellar properties, such as T_{eff} and $\log g$, T_{eff} and $\log_{10}(\mathrm{Ca/He})$, $\log g$ and $\log_{10}(\mathrm{Mg/He})$, or $\log_{10}(\mathrm{Ca/He})$ and $\log_{10}(\mathrm{Mg/He})$. To qualitatively investigate the presence of these correlations, we computed the Pearson Coefficient (PC) for all our model parameters,

$$\mathrm{PC} = \frac{\sum(\pi_x - \mu_{\pi_x})(\pi_y - \mu_{\pi_y})}{\sqrt{\sum(\pi_x - \mu_{\pi_x})^2 \sum(\pi_y - \mu_{\pi_y})^2}} \subset [-1, 1], \tag{28}$$

where π_x and π_y are the posterior distributions of labels x and y, and μ_{π_x} and μ_{π_y} are the arithmetic means of their posteriors. A PC greater or lower than 0 denotes, respectively, a positive or negative correlation, while a PC=±1 indicates that the data can be perfectly modelled by a line.

In Fig. 16, we present the PCs for `cecilia`'s best-fitting MCMC model parameters. Assuming a strong positive correlation if PC≥0.5, our results suggest that certain stellar labels are linearly dependent on one another (e.g. T_{eff} and $\log g$, or $\log_{10}(\mathrm{Ca/He})$ and $\log_{10}(\mathrm{Mg/He})$), with an increase in one label resulting in an increase of its corresponding label pair. Based on the corner plot and the PCs alone, it is hard to determine whether the observed correlations can be generalized to the entire population of polluted WDs. Although such an investigation is outside the scope of this paper, it is possible that certain stellar properties are linearly dependent on one another due to observational or physical reasons. Observationally, some of the metal correlations may be caused by the presence of blended absorption lines. From a physical perspective, changes in T_{eff} and $\log g$ (a proxy for pressure) modify the temperature-pressure conditions at the stellar photosphere, which in turn control the strength and the width of the absorption lines. In addition, different values of T_{eff} and pressure result in different excitation, ionization, or dissociation states of the atoms in the stellar photosphere (e.g. Saha equation,

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Table 5. MCMC best-fitting stellar parameters for the SDSS spectrum of WD 1232+563, together with their uncertainties and assumed ground truths. The systematic ($\sigma_{\rm sys,ML}$) and statistical uncertainties ($\sigma_{\rm stat,\,MCMC}$) measure the errors associated to `cecilia`'s spectral interpolation (see Section 5.1) and to the MCMC fit, respectively. For abundance studies of exoplanetary material, we can approximate the total uncertainty of a model parameter ($\sigma_{\rm tot}$) by computing their quadrature sum (see equation 27). In this work, if the total error of an abundance is lower than an SDSS noise floor of 0.15 dex, we assume that the MCMC statistical errors are underestimated and impose $\sigma_{\rm tot} \approx 0.15$ dex (see star ⋆ symbol). The dagger symbol (†) flags the model parameters dominated by photometric priors ($T_{\rm eff}$, log g). In this table, we do not present the best-fitting abundances of the four heavy elements with chondritic priors (Be, Cr, Mn, and Ni); these metals were included in `cecilia`'s MCMC, but are not detected in the SDSS spectrum. Finally, we do not report the systematic errors for $T_{\rm eff}$ and log g because these two parameters were fixed during our sensitivity study of synthetic spectra (see Section 5.1).

	`cecilia`'s MCMC best-fitting model		Systematic[b] and total uncertainty		Assumed ground truth	
Label[a]	*Value*	$\sigma_{\rm stat,\,MCMC}$	$\sigma_{\rm sys,ML}$	$\sigma_{\rm tot}$	*Value*	*Ref.*
$T_{\rm eff}$† (K)	11777.80	$^{+86.33}_{-84.81}$	–	±85.57	11787 ± 423	[1]
log g† (cgs)	8.24	$^{+0.06}_{-0.06}$	–	±0.06	8.30 ± 0.06	[1]
RV[c] (km s^{-1})	36.09	$^{+5.41}_{-5.41}$	±10	±11.37	19.0 ± 2.0	[2]
$\log_{10}$ (H/He)	−5.92	$^{+0.15}_{-0.19}$	±0.03	±0.17	−5.52 ± 0.10 − 5.90 ± 0.15	[1][2]
$\log_{10}$ (Ca/He)	−7.44	$^{+0.05}_{-0.05}$	±0.07	±0.15⋆	−7.41 ± 0.06 − 7.69 ± 0.05	[1][2]
$\log_{10}$ (Mg/He)	−6.11	$^{+0.04}_{-0.04}$	±0.05	±0.15⋆	−6.09 ± 0.05	[2]
$\log_{10}$ (Fe/He)	−6.51	$^{+0.07}_{-0.08}$	±0.06	±0.15⋆	−6.45 ± 0.11	[2]
$\log_{10}$ (O/He)	−5.19	$^{+0.12}_{-0.13}$	±0.09	±0.15⋆	−5.14 ± 0.15	[2]
$\log_{10}$ (Si/He)	−6.45	$^{+0.10}_{-0.12}$	±0.07	±0.15⋆	−6.36 ± 0.13	[2]
$\log_{10}$ (Ti/He)	−8.92	$^{+0.11}_{-0.13}$	±0.08	±0.15⋆	−8.96 ± 0.11	[2]

References: [1] Coutu et al. (2019) and [2] Xu et al. (2019).
[a] All abundances are given as logarithmic number abundance ratios relative to Helium, i.e. $\log_{10}$(n(Z)/n(He)).
[b] We note that there is at least another source of systematic errors coming from the atmosphere models themselves; these errors should be of the order of 0.10 dex and are not considered in this work.
[c] We assume a SDSS systematic RV error of 10 km/s based on the estimates provided by e.g. Abazajian et al. (2009).

Boltzmann equation), which could also potentially affect the metal abundances of the WD.

6 DISCUSSION

6.1 Implications for white dwarf science

As demonstrated in Section 5, `cecilia` is the first pipeline capable of simultaneously estimating 10 elemental abundances with a retrieval accuracy of $\lesssim$0.1 dex. This performance compares favorably to that of conventional techniques, which in turn depend on a variety of factors, such as the main properties of the star (e.g. $T_{\rm eff}$, log g, level of metal contamination), the quality of the spectroscopic observations (e.g. resolution, signal-to-noise), or the constitutive physics of their model atmosphere codes. In particular, conventional methods can reach typical uncertainties of about 0.10–0.20 dex for He-rich polluted WDs, both across the UV and the optical range, and irrespective of their underlying atmosphere models (e.g. Zuckerman et al. 2007; Jura et al. 2012; Raddi et al. 2015; Wilson et al. 2015; Izquierdo et al. 2020; Klein et al. 2021; Doyle et al. 2023). Therefore, `cecilia`'s performance is close to – if not comparable to – that of classical approaches, only introducing small uncertainties due to the accumulation of errors in its ML predictions. With a promising performance and an automated architecture that eliminates the need for manual, time-consuming, and iterative work, `cecilia` stands as a promising tool for the study of polluted white dwarfs and the bulk composition of their accreted material.

With the advent of wide-field, multi-object astronomical surveys like SDSS-V (Kollmeier et al. 2017), the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration 2016; Cooper et al. 2023), or WEAVE (Dalton et al. 2014), the volume of spectroscopic observations is poised to grow exponentially (Smith & Geach 2023). These surveys are expected to acquire a large number of WD spectra, so they represent a unique opportunity to improve our fundamental knowledge of metal pollution. Unfortunately, as of today, the exploitation of these data sets remains a challenge for conventional techniques due to their time-consuming and human-supervised nature. With `cecilia`, we have developed a practical solution to characterize the spectra of polluted WDs in a fast, automated, and accurate way. For instance, using 1 GPU of the MIT Satori supercomputer and limiting our optimization procedure to `mpfit`, we estimate that `cecilia` would take about 14 and 17 d to process the 40 000 and 50 000 WD spectra in the cumulative DESI and WEAVE data bases, respectively. In reality, `cecilia`'s performance should be much faster than two weeks, as only a small fraction of these observations will exhibit signatures of metal pollution. To quickly identify these polluted spectra, `cecilia` could benefit from other efficient ML-based classificators, such as the deep NNs of Vincent, Bergeron & Dufour (2023) or the Random Forest approach of Garcia-Zamora, Torres & Rebassa-Mansergas (2023).

Another valuable point of comparison is the SDSS-V survey, which has monitored more 6000 unique *Gaia* WDs as of 2021, and will acquire a total of 100 000 spectra of these stars in the coming years (Kollmeier et al. 2017; Chandra et al. 2021). Assuming, unrealistically, that all these spectra will be metal-polluted, they would represent an intractable amount of data for traditional WD characterization methods. Nonetheless, with `mpfit` only, it would only take about a month to process all these observations. Therefore, in light of these capabilities, our ML-based pipeline is a powerful tool to harness the information content of massive spectroscopic data bases and scale up individual studies of polluted WDs, thus opening the door to a statistically-informed understanding of ancient extrasolar geochemistry. Ultimately, we aspire to use `cecilia` to bring the field of WD science into the era of big data, allowing

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

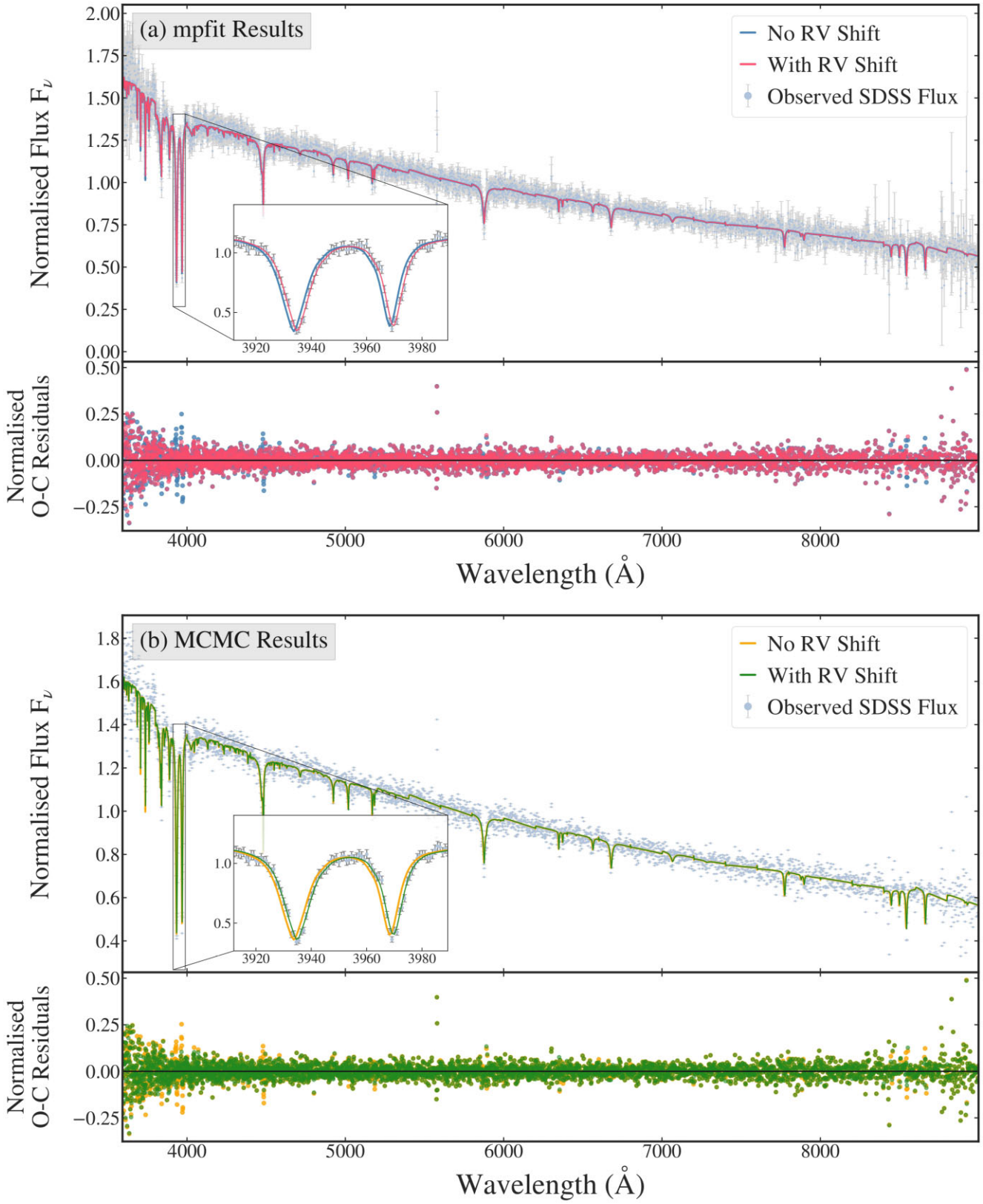


Figure 13. Normalized best-fitting models for the SDSS spectrum of WD 1232+563 (in light blue). *Panel (a)*: Best-fitting `mpfit` models using an RV shift of 0 and 94.42 km s^{-1} (in red and blue, respectively). The inset plot shows the spectral window between 3800 and 4000 Å, where there are two important Ca II lines at about 3934 and 3969 Å. The bottom plot shows the normalized flux residuals (Observed-Calculated; O-C) associated to each prediction. *Panel (b)*: Best-fitting `edmcmc` MCMC models using an RV shift of 0 and 94.74 km s^{-1} (in green and orange, respectively). Similarly to Panel (*a*), the inset plot shows the spectral window between 3800 and 4000 Å, while the bottom panel shows the normalized O-C residuals for each spectroscopic solution. We note that the MCMC value of the RV shift is different from that reported in Table 5 because it also accounts for the barycentric velocity and gravitational redshift correction.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

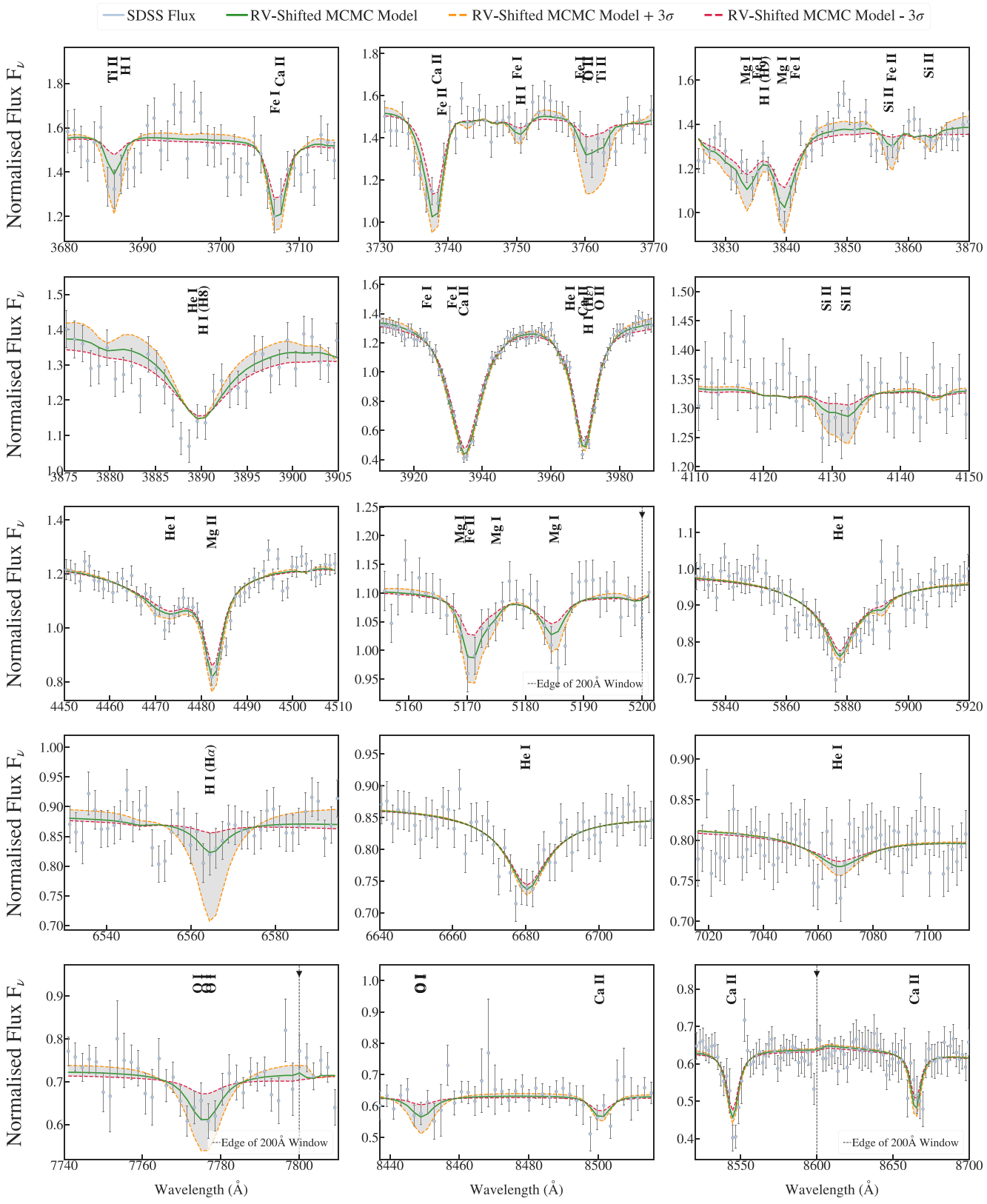


Figure 14. A selection of spectral windows illustrating our best-fitting, RV-shifted MCMC model (in green) for the raw SDSS spectrum of WD 1232+563 (in light blue with grey error bars). In each panel, we also show a grey shaded area bounded by an orange and red dashed line, which correspond, respectively, to a $\pm 3\sigma_{\text{stat, MCMC}}$ variation on the best-fitting model (using the same T_{eff} and $\log g$, and only changing the abundances of the detected elements by $3\sigma_{\text{stat, MCMC}}$). In black letters, we indicate some of the strongest spectral signatures of hydrogen, helium, and multiple heavy elements. Finally, we represent the start/end of a 200 Å spectral window with a dashed vertical line and a black inverted triangle symbol.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

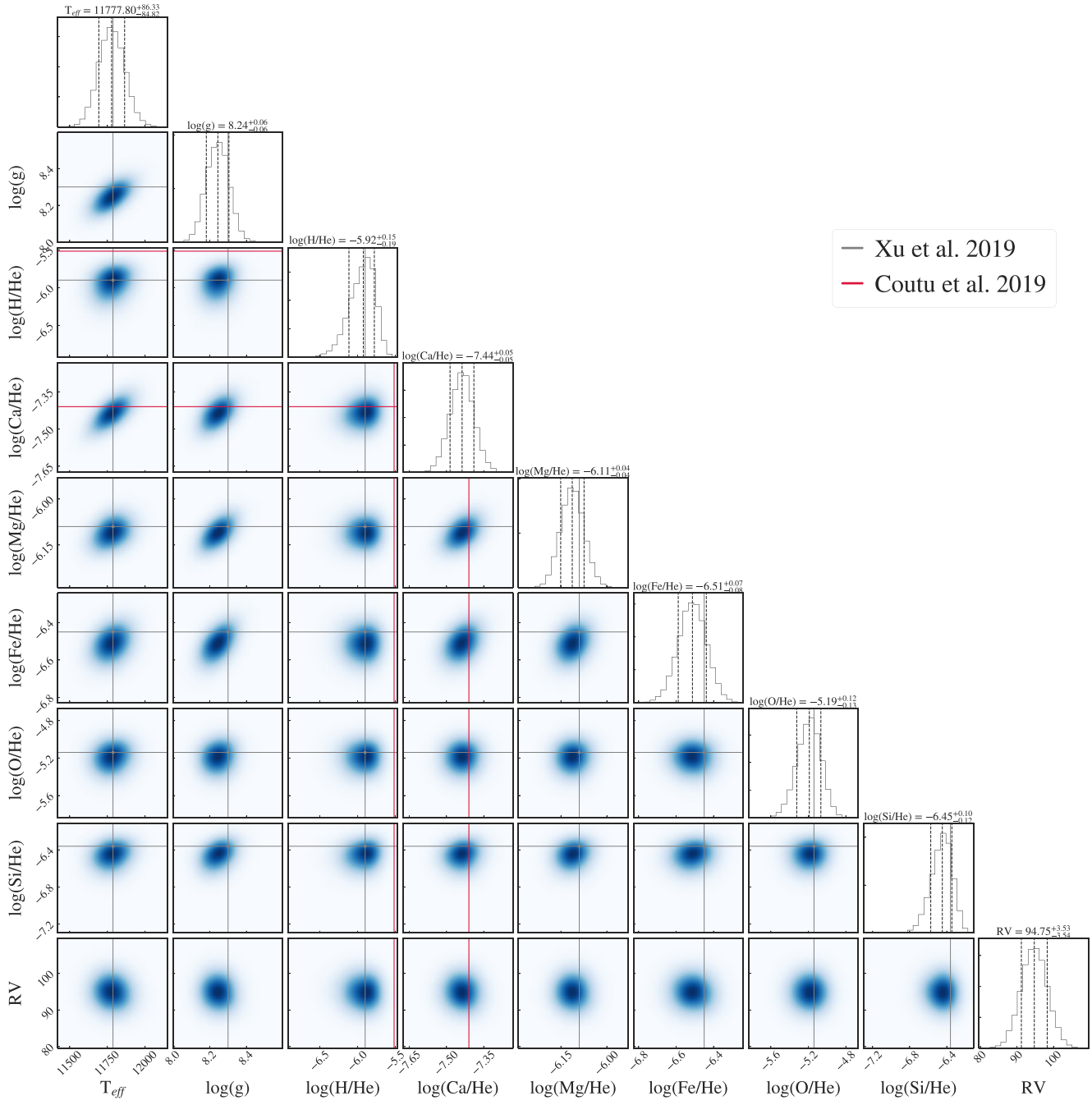


Figure 15. MCMC corner plot for the SDSS spectrum of WD 1232+563. The off-diagonal plots show the two-dimensional histograms of the posteriors of the model parameters, marginalized over the rest of parameter values. In contrast, the histograms along the diagonal illustrate their one-dimensional marginalized distributions, together with their corresponding 1σ confidence intervals. The grey and red lines show the best-fitting values of Xu et al. (2019) and Coutu et al. (2019), respectively, as defined in Table 5. In this figure, we exclude the four heavy elements with chondritic priors (i.e. Be, Cr, Mn, and Ni) and show the RV shift without subtracting the barycentric motion and gravitational redshift of the WD.

scientists to overcome the scalability problem associated to 'human-in-the-loop' conventional analysis techniques.

In addition to `cecilia`'s promising retrieval accuracy and speed, our pipeline offers a innovative Bayesisan framework to estimate the elemental abundances of polluted WDs. More specifically, our inference approach differs from that of classical methods because (i) it integrates prior knowledge about the spectra and the main properties of the WD, (ii) it provides robust parameter uncertainties based on their full posterior probability distributions, and (iii) it illustrates possible degeneracies between different model parameters thanks to the use of corner plots. Through these added capabilities, `cecilia` aims to provide a more detailed and nuanced perspective on the properties of WD pollutants and their parent bodies.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

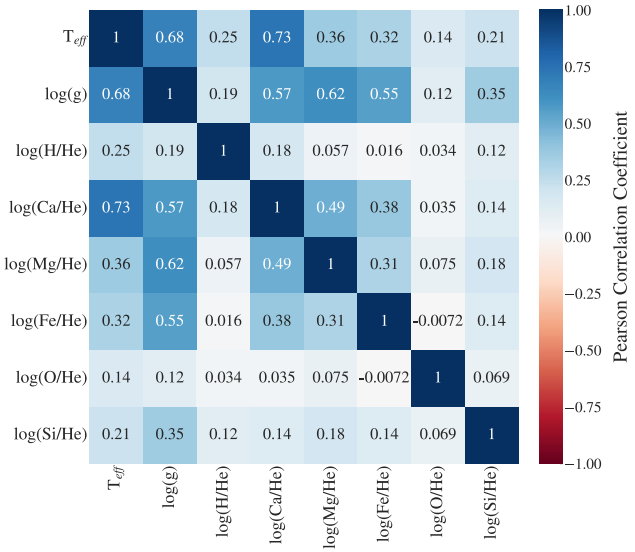


Figure 16. Pearson Correlation Coefficients (PC) for the stellar properties of WD 1232+563. Some label pairs (e.g. $T_{\rm eff}$ and log g, $T_{\rm eff}$ and Ca, H and Mg) exhibit a strong positive correlation (PC$\geq$0.5), as suggested by their slanted, oval-like corner plots in Fig. 15. In general, however, most labels have a weak or moderate correlation (PC$<$0.5).

6.2 Limitations and future work

The sensitivity analyses presented in Section 5 show that `cecilia` can effectively determine the main properties of He-rich polluted WDs, achieving an accuracy comparable to that of conventional techniques without the need for human supervision. This performance has the potential to unlock population-wide studies of metal pollution, but it has several limitations that deserve further investigation. In this section, we highlight the main caveats of our pipeline, discuss possible mitigation strategies, and propose new opportunities for future work.

6.2.1 General ML design choices

First, `cecilia`'s ML architecture can be improved along three important dimensions: versatility, speed, and accuracy. In terms of versatility, our work has revolved around He-dominated polluted WDs, which – for a given abundance level – exhibit stronger metallic lines than H-rich stars due to their lower photospheric opacity (Dufour et al. 2012; Klein et al. 2021; Saumon, Blouin & Tremblay 2022). Nevertheless, `cecilia`'s potential extends beyond He-dominated systems, as its ML architecture can always be retrained with the model atmospheres and astrophysical properties of other stellar objects.

With respect to speed, our full optimization procedure takes between 4 and 10 hours to provide a best-fitting spectroscopic solution, depending on the user's choice of MCMC hyperparameters (see Section 4.3). This behaviour makes `cecilia` an efficient pipeline capable of modelling several stars at the same time, while running uninterruptedly on a high-performance supercomputer cluster. The scalability of `cecilia` represents an improvement over the performance of classical methods, but it can still be optimized further in preparation for large-scale studies of polluted WDs. A possible solution to making `cecilia` more efficient would be to explore whether simpler and/or faster ML architectures can achieve similar performance.

Lastly, `cecilia`'s retrieval accuracy is better than 0.1 dex for 10 heavy elements, which places our pipeline at the level of many conventional techniques. Nevertheless, the errors accumulated in `cecilia`'s networks introduce a small amount of noise into its final synthetic prediction, which in turn challenges the modelling of very shallow lines and introduces limits below which `cecilia` performs poorly. As model atmospheres improve with our understanding of WD physics, larger training sets and/or better ML architectures might help improve `cecilia`'s predictive abilities, obtaining a better match between its best-fitting solutions and the observed metal abundances of a polluted WD. Regardless of whether we improve `cecilia`'s ML predictions, the errors are small enough that most absorption lines detected in the data would be well modelled by our code.

6.2.2 Training improvements

As explained in Section 3, the size and quality of a training data set can highly affect the performance of a NN. In this work, we have have trained `cecilia` with more than 22 000 synthetic labels and atmosphere models, assuming that WDs can be characterized with a set of 13 randomly varied independent parameters ($T_{\rm eff}$, log g, $\log_{10}$ (H/He), and 10 metal abundances). As we design future iterations of `cecilia`, a larger and more comprehensive training data set might help our pipeline to achieve a better level of precision. Another important update of `cecilia` would involve changing our training strategy to prevent distortions of the lines near the edges of the 200 Å spectral windows. A possible solution to this problem would be to re-train `cecilia` using a system of overlapping windows.

In relation to the stellar labels used in this work, we have identified three improvement opportunities that can be easily implemented in the future. To begin with, we can extend the abundance ranges listed in Table 1 – upwards and downwards – so that `cecilia` can gain sensitivity to both stronger and weaker levels of metal pollution. This is particularly relevant for Ca and Be. For instance, the inclusion of higher Ca abundances in our training set would allow us to validate our code against several heavily polluted WDs whose astrophysical properties are currently outside of `cecilia`'s allowed ML bounds (e.g. GD 40, Ton 345, GD 362, although the latter would also require extending `cecilia` to lower temperatures, as discussed below). Similarly, expanding the abundance ranges of Be to higher values would make our pipeline more sensitive to stronger Be lines. This would facilitate the analysis of several Be-polluted systems such as those described in Klein et al. (2021), which have a log(Be/Ca) ratio ($\approx$-2.50 dex) considerably higher than that of the Sun ($\approx$-4.96 dex, Asplund et al. 2009). Secondly, we can make our pipeline more useful by retraining it with more heavy elements. With this idea in mind, we aim to include at least three new elements: carbon (C), sodium (Na), and aluminium (Al). These species have been detected in the atmospheres of multiple WDs and can provide valuable insights into the volatile (C) and rocky (Na, Al) compositions of WD pollutants (Greenstein 1976; Holberg, Barstow & Sion 1998; see review by Klein et al. 2021). Finally, we can try to improve `cecilia`'s performance by exploring different ways to generate our stellar labels. This is especially important for the 14 metals with fixed chondritic abundances (C, N, Li, Na, Al, P, S, Cl, Ar, K, Sc, V, Co, and Cu), which might have introduced biases in our pipeline and caused `cecilia`'s spectral predictions to favor chondritic values. Although the results presented in Section 5 do not show any evidence of this behaviour, it will be important to investigate this potential problem in future versions of `cecilia`.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Regarding our atmosphere models, there are two possible changes that might make `cecilia` more useful and versatile. On the one hand, we can retrain our pipeline using a broader wavelength coverage. At present, `cecilia` can analyze spectroscopic observations in the optical range between 3000 and 9000 Å, but it is not prepared to model data in the UV ($\sim$900 $\leq \lambda \leq$3000 Å). The UV contains important spectral signatures of volatile species – e.g. oxygen, carbon, nitrogen, sulfur, phosphorous – and is thus a crucial region for understanding the volatile budget of a WD pollutant. From a practical point of view, the integration of the UV into our pipeline will also make `cecilia` particularly relevant for the study of UV observations from space-based facilities such as the Hubble Space Telescope.

Moreover, we can also retrain `cecilia` with synthetic spectra covering a broader range of effective temperatures. For example, it could be useful to include models of slightly cooler ($T_{\rm eff} < 10\,000$ K) and hotter WDs ($20\,000 \leq T_{\rm eff} \leq 25\,000$ K). At temperatures higher than $T_{\rm eff} > 25\,000$ K, the origin of metal pollution is unclear due to the outward effects of radiative levitation pressure, which can bring metals to the photosphere from the deep stellar interior (Chayer, Fontaine & Wesemael 1995). For cooler WDs, it is still possible to observe traces of metal pollution in their spectra (e.g. Hollands et al. 2017; Blouin 2020; Elms et al. 2022). While most of these cool systems only show a handful of metallic absorption lines (because of the insufficient thermal energy to populate atomic states responsible for many transitions in the optical), they also allow the detection of elements that cannot be easily observed in warmer objects where they are ionized (e.g. Na, K, Li).

6.3 Fitting analysis

In addition to exploring possible improvements to `cecilia`'s ML architecture and training data sets, we have also identified four opportunities to refine our fitting procedure. First, we can reconfigure our code to predict upper abundance limits when the absorption features of a metal are ambiguous or unobservable. This option is currently unavailable because our pipeline was trained in logarithmic space, which implies that `cecilia` can never predict a zero elemental abundance (i.e. the complete lack of an element). Introducing the possibility of quantifying upper limits would involve a reformulation of our code, but would constitute a valuable addition to `cecilia`.

A second enhancement to our pipeline would be to allow for a simultaneous spectroscopic and photometric fit of a polluted WD spectrum. Given that our code does not currently use photometric observations in its optimization routine, `cecilia` requires the user to have accurate estimates of the effective temperature and surface gravity of the star. These two parameters are also dependent on the amount of heavy elements in the stellar photosphere, so their prediction becomes an iterative problem.[17] Moreover, even if these two properties are well-defined, they might have been estimated with a set of atmosphere models different to those employed in this work, which could potentially introduce minor errors in `cecilia`'s predictions. Therefore, to ensure that our parameter estimation routine is entirely self-consistent, it would be beneficial to incorporate a simple NN capable of estimating values for $T_{\rm eff}$ and log g based on the metal abundances predicted by `cecilia` at each iteration.

Third, we hope to improve `cecilia`'s MCMC fitting procedure by introducing a jitter term in our log-likelihood function. This parameter would allow us to fit unknown sources of noise (e.g. instrumental noise, stellar activity), and could thus be particularly important to model weak absorption lines in real spectroscopic observations. Finally, given that MCMC pipelines can sometimes be quite sensitive to the initialization of the model parameters, we hope to make `cecilia` capable of calculating the RV shift of the WD when no estimates are provided by the user.

7 CONCLUSIONS

In this work, we have presented `cecilia`, the first ML pipeline designed to measure the main physical and chemical properties of intermediate-temperature, He-rich polluted WDs from their spectra. In particular, our code exploits the power of NN-based-interpolation to (i) rapidly generate synthetic spectral models of polluted WDs in high-dimensional space, (ii) map the latter to a set of 13 underlying stellar properties ($T_{\rm eff}$, log g, $\log_{10}$ (H/He), and 10 metal abundances); and (iii) generate a final spectral prediction through frequentist and Bayesian fitting techniques (see Sections 3 and 4).

With the development of `cecilia`, we have sought to tackle the dependence of conventional spectral analysis techniques on manual and iterative work, which in turn make them time-intensive, prone to human errors, and impossible to scale up to large samples of polluted WDs. As explained in Section 5, `cecilia`'s architecture speeds up and automates these conventional methods by providing preliminary metal abundances in less than 2 min and a final spectroscopic solution in less than 10 h. After testing the performance of our pipeline with noiseless synthetic data and with real spectroscopic observations of WD 1232+563, we have shown that `cecilia` can retrieve metal abundances with uncertainties better than 0.1 dex for up to 10 elements, with only beryllium proving hard to constrain. This level of accuracy is similar to that of state-of-the-art techniques, and it can only be improved in the future as we incorporate larger/better training data sets and integrate user feedback.

Acknowledging the main limitations of our pipeline (see Section 6), `cecilia` has the potential to enable population-wide studies of metal pollution without entailing any manual work. These studies will significantly expand the number of polluted WDs with well-characterized abundances, thus playing an important role in improving our knowledge of the geology and chemistry of extrasolar material. Unveiling the diversity of extrasolar compositions is especially relevant in the context of upcoming wide-field astronomical surveys like DESI, WEAVE, or SDSS-V, which will acquire a large volume of spectroscopic observations in the coming years. While conventional WD characterization techniques are too slow to process these massive data sets, our ML pipeline offers a fast, automated, and reliable solution to measuring trace elements in the atmospheres of polluted WDs.

Finally, as a feature extraction tool with a versatile ML architecture, `cecilia` can also be replicated and transferred to other disciplines beyond the realm of WD and of planetary science. As we step into the world of Big Data, where the ability to uncover meaningful insights from observations will increasingly become more important, `cecilia` exemplifies the immense potential of combining powerful ML tools with state-cutting-edge scientific knowledge and analysis techniques.

[17] We note that this issue also affects traditional WD characterization methods. As of today, several iterative approaches have been developed to mitigate it (e.g. Dufour, Bergeron & Fontaine 2005; Dufour et al. 2007; Coutu et al. 2019). These approaches perform an initial fit to the stellar photometry, followed by a spectroscopic fit. Afterwards, this optimization process is repeated multiple times – each time with a better set of initial parameters – until a best-fitting solution is found *both* for the existing photometric and spectroscopic observations.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

ACKNOWLEDGEMENTS

MBA thanks Vedant Chandra, Oriol Abril-Pla, Dr Kishalay De, Dr Ignasi Ribas, and Dr Amy Bonsor for useful discussions. MBA is supported by the MIT Department of the Earth, Atmospheric, and Planetary Sciences, NASA grants 80NSSC22K1067 and 80NSSC22K0848, and the MIT William Asbjornsen Albert Memorial Fellowship. SB is a Banting Postdoctoral Fellow and a CITA National Fellow, supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). SX is supported by NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

This work has made use of the Montreal White Dwarf Data base (Dufour et al. 2016).

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

This work has also made use of data from the Sloan Digital Sky Survey (SDSS), which is funded by the Alfred P. Sloan Foundation, the Heising-Simons Foundation, the National Science Foundation, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org. SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including the Carnegie Institution for Science, Chilean National Time Allocation Committee (CNTAC) ratified researchers, the Gotham Participation Group, Harvard University, Heidelberg University, The Johns Hopkins University, L'École polytechnique Fédérale de Lausanne (EPFL), Leibniz-Institut fur Astrophysik Potsdam (AIP), Max-Planck-Institut fur Astronomie (MPIA Heidelberg), Max-Planck-Institut fur Extraterrestrische Physik (MPE), Nanjing University, National Astronomical Observatories of China (NAOC), New Mexico State University, The Ohio State University, Pennsylvania State University, Smithsonian Astrophysical Observatory, Space Telescope Science Institute (STScI), the Stellar Astrophysics Participation Group, Universidad Nacional Autónoma de Mexico, University of Arizona, University of Colorado Boulder, University of Illinois at Urbana-Champaign, University of Toronto, University of Utah, University of Virginia, Yale University, and Yunnan University.

This work has employed the following open-source software packages: PYTHON (van Rossum 1995), NUMPY (Oliphant 2015), SCIPY (Virtanen et al. 2020), MATPLOTLIB (Hunter 2007), PANDAS (McKinney 2010), MPFIT (Markwardt 2009), EDMCMC (Vanderburg 2021), TENSORFLOW (Abadi et al. 2015), and CORNER (Foreman-Mackey 2016; Luger, Lustig-Yaeger & Agol 2017).

DATA AVAILABILITY

The SDSS spectrum of WD 1232+563 can be downloaded from the SDSS DR18 online data base.

REFERENCES

Abadi M. et al., 2016, Proc. 12th USENIX Conference on Operating Systems Design and Implementation. USENIX Association, Savannah, GA, p. 265
Abazajian K. N. et al., 2009, ApJS, 182, 543
Almeida A. et al., 2023, ApJS, 267, 44
Asplund M., Grevesse N., Sauval A. J., Scott P., 2009, ARA&A, 47, 481
Blouin S., 2020, MNRAS, 496, 1881
Blouin S., Dufour P., Allard N. F., 2018a, ApJ, 863, 184
Blouin S., Dufour P., Allard N. F., Kilic M., 2018b, ApJ, 867, 161
Blouin S., Dufour P., Allard N. F., Salim S., Rich R. M., Koopmans L. V. E., 2019a, ApJ, 872, 188
Blouin S., Dufour P., Thibeault C., Allard N. F., 2019b, ApJ, 878, 63
Campbell B., Walker G. A. H., Yang S., 1988, ApJ, 331, 902
Caron A., Bergeron P., Blouin S., Leggett S. K., 2023, MNRAS, 519, 4529
Chandra V., Hwang H.-C., Zakamska N. L., Budavári T., 2020, MNRAS, 497, 2688
Chandra V. et al., 2021, ApJ, 921, 160
Charbonneau D., Brown T. M., Latham D. W., Mayor M., 2000, ApJ, 529, L45
Chayer P., Fontaine G., Wesemael F., 1995, ApJS, 99, 189
Cooper A. P. et al., 2023, ApJ, 947, 37
Coutu S., Dufour P., Bergeron P., Blouin S., Loranger E., Allard N. F., Dunlap B. H., 2019, ApJ, 885, 74
DESI Collaboration, 2016, preprint (arXiv:1611.00037)
Dalton G. et al., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. SPIE, Bellingham, p. 91470L
Desharnais S., Wesemael F., Chayer P., Kruk J. W., Saffer R. A., 2008, ApJ, 672, 540
Dorn C., Khan A., Heng K., Connolly J. A. D., Alibert Y., Benz W., Tackley P., 2015, A&A, 577, A83
Doyle A. E. et al., 2023, ApJ, 950, 93
Dufour P., Bergeron P., Fontaine G., 2005, ApJ, 627, 404
Dufour P. et al., 2007, ApJ, 663, 1291
Dufour P., Kilic M., Fontaine G., Bergeron P., Melis C., Bochanski J., 2012, ApJ, 749, 6
Dufour P., Blouin S., Coutu S., Fortin-Archambault M., Thibeault C., Bergeron P., Fontaine G., 2016, in Tremblay P.-E., Gänsicke B., Marsh T., eds, ASP Conference Series, Vol. 509, 20th European White Dwarf Workshop. Astron. Soc. Pac, San Francisco, p. 3
Eisenstein D. J. et al., 2006, ApJS, 167, 40
Elms A. K., Tremblay P.-E., Gänsicke B. T., Koester D., Hollands M. A., Gentile Fusillo N. P., Cunningham T., Apps K., 2022, MNRAS, 517, 4557
Farihi J., Brinkworth C. S., Gänsicke B. T., Marsh T. R., Girven J., Hoard D. W., Klein B., Koester D., 2011, ApJ, 728, L8
Farihi J., Gaensicke B. T., Koester D., 2013, Science, 342, 218
Foreman-Mackey D., 2016, J. Open Source Softw., 1, 24
Fortin-Archambault M., Dufour P., Xu S., 2020, ApJ, 888, 47
van Rossum G., 1995, Technical Report, Python Tutorial, Technical Report CS-R9526. Amsterdam
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2023, A&A, 674, A1
Garcia-Zamora E. M., Torres S., Rebassa-Mansergas A., 2023, A&A, 679, A127
Gelman A., Rubin D. B., 1992, Stat. Sci., 7, 457
Genest-Beaulieu C., Bergeron P., 2019, ApJ, 882, 106
Giammichele N., Bergeron P., Dufour P., 2012, ApJS, 199, 29
Goodfellow I. J., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA
Greenstein J. L., 1976, ApJ, 207, L119
Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.-Z., 2019, Sci. Robotics, 4, eaay7120
Hanin B., 2018, Proc. 32nd International Conference on Neural Information Processing Systems. NIPS, Montrèal, Canada, p. 580
Holberg J. B., Barstow M. A., Sion E. M., 1998, ApJS, 119, 207
Hollands M. A., Koester D., Alekseev V., Herbert E. L., Gänsicke B. T., 2017, MNRAS, 467, 4970
Hoskin M. J. et al., 2020, MNRAS, 499, 171
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Izquierdo P., Toloza O., Gänsicke B. T., Rodríguez-Gil P., Farihi J., Koester D., Guo J., Redfield S., 2020, MNRAS, 501, 4276

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

Jarrett K., Kavukcuoglu K., Ranzato M., LeCun Y., 2009, in IEEE 12th International Conference on Computer Vision. p. 2146
Jura M., 2003, ApJ, 584, L91
Jura M., Young E. D., 2014, Annu. Rev. Earth Planet. Sci., 42, 45
Jura M., Xu S., Klein B., Koester D., Zuckerman B., 2012, ApJ, 750, 69
Kaiser N. et al., 2010, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Conf. Ser. Volume 7733, Ground-based and Airborne Telescopes III. SPIE, Bellingham, p. 77330E
Kaiser B. C., Clemens J. C., Blouin S., Dufour P., Hegedus R. J., Reding J. S., Bédard A., 2021, Science, 371, 168
Kiefer J., Wolfowitz J., 1952, Ann. Math. Stat., 23, 462
Kingma D. P., Ba J., 2017, Adam: A Method for Stochastic Optimization. preprint (arXiv:1412.6980)
Klein B., Jura M., Koester D., Zuckerman B., 2011, ApJ, 741, 64
Klein B. L., Doyle A. E., Zuckerman B., Dufour P., Blouin S., Melis C., Weinberger A. J., Young E. D., 2021, ApJ, 914, 61
Kleinman S. J. et al., 2013, ApJS, 204, 5
Koester D., 2009, A&A, 498, 517
Koester D., Napiwotzki R., Voss B., Homeier D., Reimers D., 2005, A&A, 439, 317
Koester D., Gänsicke B. T., Farihi J., 2014, A&A, 566, A34
Kollmeier J. A. et al., 2017, preprint (arXiv:1711.03234)
Kramida A., Ralchenko Yu., Reader J., NIST ASD Team, 2022, NIST Atomic Spectra Data base (ver. 5.10), available at: http://physics.nist.gov/asd
LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
Liang Y., Winn J. N., Melchior P., 2023, AJ, 167, 23
Lodders K., 2003, ApJ, 591, 1220
Luger R., Lustig-Yaeger J., Agol E., 2017, ApJ, 851, 94
Markwardt C. B., 2009, in Bohlender D. A., Durand D., Dowler P., eds, ASP Conf. Ser. Vol. 411, Astronomical Data Analysis Software and Systems XVIII. Astron. Soc. Pac., San Francisco, p.251
Martin D. C. et al., 2005, ApJ, 619, L1
McCarthy J., Minsky M. L., Rochester N., Shannon C. E., 2006, AI magazine, 27, 12
McKinney W., 2010, in van der Walt S., Millman J., eds, Proc. 9th Python in Science Conference. p. 51
Melis C., Dufour P., 2017, ApJ, 834, 1
Moré J. J., 1978, in Lecture Notes in Mathematics, Vol. 630. Springer-Verlag, Berlin, p. 105
Nair V., Hinton G. E., 2010, in Proc. 27th International Conference on International Conference on Machine Learning, ICML10. Omnipress, Madison, WI, p. 807
Ness M., Hogg D. W., Rix H. W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16
Oliphant T. E., 2015, Guide to NumPy, 2nd edn. CreateSpace Independent Publishing Platform, North Charleston, SC
Palme H., Lodders K., Jones A., 2014, in Davis A. M., ed., Planets, Asteriods, Comets and The Solar System, Vol. 2 of Treatise on Geochemistry, 2nd edn. Elsevier, p. 15
Paquette C., Pelletier C., Fontaine G., Michaud G., 1986, ApJS, 61, 197
Raddi R., Gänsicke B. T., Koester D., Farihi J., Hermes J. J., Scaringi S., Breedt E., Girven J., 2015, MNRAS, 450, 2083
Rogers L. A., Seager S., 2010, ApJ, 712, 974
Samuel A. L., 1959, IBM J. Res. Dev., 3, 210
Saranya A., Subhashini R., 2023, Decision Anal. J., 7, 100230
Saumon D., Blouin S., Tremblay P.-E., 2022, Phys. Rep., 988, 1
Seager S., Kuchner M., Hier-Majumder C. A., Militzer B., 2007, ApJ, 669, 1279
Shallue C. J., Vanderburg A., 2018, AJ, 155, 94
Smith M. J., Geach J. E., 2023, R. Soc. Open Sci., 10, 221454
Tan C., Sun F., Kong T., Zhang W., Yang C., Liu C., 2018, Artificial Neural Networks and Machine Learning – ICANN 2018. Springer Interna. Publ., Cham, p. 270
Ter Braak C. J. F., 2006, Stat. Comput., 16, 239
Ting Y.-S., Conroy C., Rix H.-W., Cargile P., 2019, ApJ, 879, 69
Turing A. M., 1950, Mind, 59, 433
Vanderburg A., 2021, avanderburg/edmcmc: v1.0.0. Zenodo, available at: https://zenodo.org/records/5599854
Vanderburg A. et al., 2015, Nature, 526, 546
Vats D., Knudson C., 2018, preprint (arXiv:1812.09384)
Vehtari A., Gelman A., Simpson D., Carpenter B., Bürkner P.-C., 2021, Bayesian Analysis, 16, 667
Veras D., 2021, Oxford Research Encyclopedia of Planetary Science. Oxford Univ. Press, p. 1
Vilone G., Longo L., 2020, preprint (arXiv:2006.00093)
Vincent O., Bergeron P., Dufour P., 2023, MNRAS, 521, 760
Virtanen P. et al., 2020, Nat. Meth., 17, 261
Wilson D. J., Gänsicke B. T., Koester D., Toloza O., Pala A. F., Breedt E., Parsons S. G., 2015, MNRAS, 451, 3237
Wolff B., Koester D., Liebert J., 2002, A&A, 385, 995
Xiang M. et al., 2022, A&A, 662, A66
Xu S., Jura M., Klein B., Koester D., Zuckerman B., 2013, ApJ, 766, 132
Xu S., Jura M., Dufour P., Zuckerman B., 2016, ApJ, 816, L22
Xu S., Zuckerman B., Dufour P., Young E. D., Klein B., Jura M., 2017, ApJ, 836, L7
Xu S., Dufour P., Klein B., Melis C., Monson N. N., Zuckerman B., Young E. D., Jura M. A., 2019, AJ, 158, 242
York D. G. et al., 2000, AJ, 120, 1579
Zeng L., Seager S., 2008, PASP, 120, 983
Zeng L., Sasselov D. D., Jacobsen S. B., 2016, ApJ, 819, 127
Zuckerman B., Koester D., Reid I. N., Hünsch M., 2003, ApJ, 596, 477
Zuckerman B., Koester D., Melis C., Hansen B. M., Jura M., 2007, ApJ, 671, 872
Zuckerman B., Melis C., Klein B., Koester D., Jura M., 2010, ApJ, 722, 725

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article/529/2/1688/7604626 by University of Wisconsin-Madison Libraries user on 09 May 2024

© 2024 The Author(s).
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.