Machines learn to infer stellar parameters just by looking at a large number of spectra

Nima Sedaghat, Martino Romaniello, Jonathan Carrick, François-Xavier Pineau

▶ **To cite this version:**

Nima Sedaghat, Martino Romaniello, Jonathan Carrick, François-Xavier Pineau. Machines learn to infer stellar parameters just by looking at a large number of spectra. Monthly Notices of the Royal Astronomical Society, 2021, 501 (4), pp.6026-6041. 10.1093/mnras/staa3540 . hal-03176264

HAL Id: hal-03176264
https://hal.science/hal-03176264

Submitted on 4 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **501,** 6026–6041 (2021) doi:10.1093/mnras/staa3540

Machines learn to infer stellar parameters just by looking at a large number of spectra

Nima Sedaghat [1]★ Martino Romaniello,[1] Jonathan E. Carrick[2] and François-Xavier Pineau[3]

[1] *European Southern Observatory (ESO), Karl-Schwarzschild-Str, D-85748 Garching, Germany*
[2] *Physics Department, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK*
[3] *CNRS, Observatoire astronomique de Strasbourg, Université de Strasbourg, UMR 7550, F-67000 Strasbourg, France*

Accepted 2020 November 9. Received 2020 November 9; in original form 2020 September 23

ABSTRACT

Machine learning has been widely applied to clearly defined problems of astronomy and astrophysics. However, deep learning and its conceptual differences to classical machine learning have been largely overlooked in these fields. The broad hypothesis behind our work is that letting the abundant real astrophysical data speak for itself, with minimal supervision and no labels, can reveal interesting patterns that may facilitate discovery of novel physical relationships. Here, as the first step, we seek to interpret the representations a deep convolutional neural network chooses to learn, and find correlations in them with current physical understanding. We train an encoder–decoder architecture on the self-supervised auxiliary task of reconstruction to allow it to learn general representations without bias towards any specific task. By exerting weak disentanglement at the *information bottleneck* of the network, we implicitly enforce interpretability in the learned features. We develop two independent statistical and information-theoretical methods for finding the number of learned *informative features*, as well as measuring their true correlation with astrophysical validation labels. As a case study, we apply this method to a data set of ∼270 000 stellar spectra, each of which comprising ∼300 000 dimensions. We find that the network clearly assigns specific nodes to estimate (notions of) parameters such as radial velocity and effective temperature without being asked to do so, all in a completely physics-agnostic process. This supports the first part of our hypothesis. Moreover, we find with high confidence that there are ∼4 more independently informative dimensions that do not show a direct correlation with our validation parameters, presenting potential room for future studies.

Key words: methods: data analysis – methods: numerical – techniques: spectroscopic.

1 INTRODUCTION

Big Data has already changed the way we do science in nearly all areas of research every day. Although *data-driven* methods have been around since almost the very beginning of the history of science, the meaning of the term has started to transform gradually; data are not used only to validate our analytical formulations and hypotheses any more, but have started taking more serious roles in defining the problem itself, and providing non-parametric solutions to it.

The rationale behind this reform is two-fold. First, huge amounts of new data are becoming available in many areas: from the ever increasing number of search-able images on the web to the petabytes-per-minute streams of data expected from future telescopes – e.g. see SKA (Quinn et al. 2015). Secondly, and perhaps more importantly, the scientific community has found, and is advancing, ways to handle such big volumes of data, thanks to advances in technology. At the core of these advances lies the recent revolution of techniques under the broad term of *machine learning*.

The number of machine-learning-based solutions to problems in astrophysics, astronomy, and cosmology has drastically increased in the past years, and providing a list of them is beyond the scope of this manuscript – we refer to Baron (2019) for a practical overview. We believe what particularly needs to be assessed, however, is the way learning has been utilized in these fields, and the potentials to broaden the horizons. Concretely speaking, the so-called revolution of the past two decades has been more about *deep learning* (DL, Raina, Madhavan & Ng 2009; Krizhevsky, Sutskever & Hinton 2012): a new family of methods *forked* out of classical machine learning (ML) – the latter has already been around since as early as 1980s (LeCun 1985). But most of the solutions used by our community have been plugin-style usages of classical ML, and the advantages deep learning brings upon have not found enough exposure.

Classical ML can be roughly modelled as a black box that implicitly learns how to connect input features (engineered by humans) to desired output. Deep learning, on the other hand, is a similar box, normally implemented as a neural network, with the additional capability to learn and decide what features are best to be used for the task at hand. The ability, also known as *representation learning* (Rumelhart, Hinton & Williams 1986; Bengio, Courville & Vincent 2013), is the key difference between the two methodologies – not the depth of the neural network.

Nevertheless, deep models have proven superiority in performance and accuracy over traditional methods in astronomy and astrophysics.

★ E-mail: nima.sedaghat@eso.org

© 2021 The Author(s)
Published by Oxford University Press on behalf of Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

Applications involving classification, detection, and regression have been extensively and successfully outsourced to neural networks in the past years, from redshift estimation (Vanzella et al. 2004) to morphological classification (Lukic & Brüggen 2016). Yet, there has been little work towards finding *how* a network is tackling a specific problem and indeed the interpretation of what the network has learned is still an open line of research, in all areas.

Unsupervised approaches have also been extensively studied, especially in the field of computer vision where deep learning was originally cultivated, e.g. see Bengio et al. (2013) for a review. Such methods have even been attempted in other fields of science too, including astronomy (e.g. see Baron & Poznanski 2017). However, they have often been used to either learn proper features for initialization of the *main* supervised task (e.g. Martinazzo, Espadoto & Hirata 2020), or simply as techniques for tasks such as dimensionality reduction (Hinton 2006), compression (Wulff 2020), and storage tractability.

In this work, we choose to take a fully unsupervised approach, without defining any specific tasks for the network. The idea is to attempt to interpret the representations by which the network decides to perceive and describe the data, and assess whether there are traces of (astro-)physical concepts in them.

The idea of 'distilling data into knowledge' in form of analytical expressions was introduced by Schmidt & Lipson 2009, and later adapted to astronomy (Graham et al. 2013) and cosmology (Krone-Martins, Ishida & De Souza 2014). Our work shares the same basic goal at the conceptual level: letting a machine learn from experimental data. However, we go beyond the constraints of analytical expressions and try to capture the knowledge in a non-parametric fashion, relying on the hierarchical feature learning capabilities of deep neural networks.

In the past years, there have been works lying at the cross-section of deep learning and the broad definition of the term *physics*. Most of such works implement *physics-guided* or *physics-informed* networks, where the network is explicitly or implicitly pre-fed with known physical laws (e.g. see Meng et al. 2020; Zhang, Liu & Sun 2020). Inspired by Hamiltonian mechanics, Greydanus, Dzamba & Yosinski (2019) and Choudhary et al. (2020) design Hamiltonian Neural Networks that learn to respect exact conservation laws. Raissi, Perdikaris & Karniadakis (2017) teach neural networks to solve tasks while respecting physical laws described by partial differential equations. Stewart & Ermon (2017) use prior knowledge to limit the space of possible learned mappings. Denil et al. (2017) use reinforcement learning to pursue physical experiments. Ehrhardt et al. (2017) use simulated motion sequences to teach a neural network to predict motion, where Sedaghat, Zolfaghari & Brox (2017) predict motion patterns in real videos. D'Agnolo & Wulzer (2019) and De Simone & Jacques (2019) use neural networks to detect discrepancy between reference models and actual (synthetic) data. However in all of them the flow of physics knowledge is, directly or indirectly, from human mind to the machine, whereas in this work, we focus on observing how the machines learn; i.e. the way Big Data enforces the machine to interpret it.

Fig. 1 outlines our implementation of the above idea. We use an archive of stellar spectra obtained using the HARPS (High Accuracy Radial-velocity Planet Searcher; Pepe et al. 2002; Santos et al. 2004; Romaniello et al. 2018) instrument, as an exemplar case for study, with easy access to a large number of samples.[1] We pass the data, as

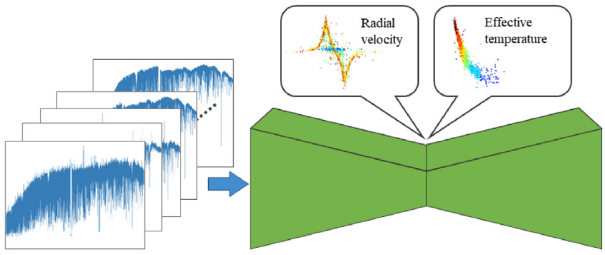

Figure 1. A large number of stellar spectra are passed through the *information bottleneck* of a deep convolutional autoencoder, in a fully unsupervised, physics-agnostic process. The network has zero information about the content of the numerical vectors it receives. We use techniques based on information maximization, to enforce learning of disentangled features, and find that the network learns representations for astrophysical parameters such as *radial velocity* and *effective temperature*, without being asked to do so.

a set of 1D[2] numerical arrays, through the *information bottleneck*[3] of a deep convolutional autoencoder, seeking a low-dimensional yet informative representation of the data (Tishby et al. 1999). The process is fully unsupervised and the network is completely agnostic of the type of the content it is seeing. The only constraint we apply during training is enforcing disentanglement of the learned representations (Bengio et al. 2013), based on maximization of the mutual information (MI; Cover 1991) between latent representations and the main signal. This, however, is the key component of our implementation, as we need to tune the disentanglement weight to a lower-than-standard level, for the method to work.

We crack open the trained network, and surprisingly find that clear traces of physical concepts, such as the effective temperature of stars and radial velocity are captured by the network. In other words, the network learns to identify and map such physical features to individual dedicated latent nodes. Such correlations are identified by seeking MI between the latent nodes and astrophysical validation labels we manage to collect from published catalogues (through the VizieR interface, Ochsenbein, Bauer & Marcout 2000), for a subset of HARPS object.

In parallel, we define a purely statistical informativeness measure and run it on the latent nodes to find probable candidates for analysis. Although the weight we put on disentanglement affects the results, we find in a reasonable setting that six nodes (out of 128) supposedly capture a noticeable amount of information. Interestingly, the two *physical nodes* we already identified are among the 6, leaving the remaining 4 open for future studies. As scientifically surprising as the identified physical nodes are, the remaining 4 are potentially even more important in the context of the long-term goal of our studies, as they may open doors for us to learn *new* patterns/correlations from data.

Our implementation is based on autoencoders (Vincent et al. 2010): the de-facto framework for unsupervised approaches in deep learning. The image generating capability of convolutional encoder–decoder architectures has also been utilized in for tasks such as transient detection (Sedaghat & Mahabal 2018) and de-

[1]We henceforth refer to the data set itself as HARPS.

[2]The term 1D here is used the way it is used in the signal processing literature, to differentiate vectors from 2D arrays, a.k.a. matrices, and higher dimensionalities. Otherwise, from a computer scientific point of view each spectrum in our case has a dimensionality of ∼300 000.

[3]We use the term 'information bottleneck' in a loose manner for both the exact theory of Tishby, Pereira & Bialek (1999), as well as the architectural bottleneck formed where the encoder and decoder of an autoencoder meet.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

blending (Boucaud et al. 2020). However, we move from the deterministic version to Variational AutoEncoders (VAE; Kingma & Welling 2014), where statistical analysis is made possible. VAEs and their extensions have been widely used to achieve (or enforce) interpretability in latent representations – e.g. see Bengio et al. (2013), Higgins et al. (2017), Chen et al. (2018), Zhao, Song & Ermon (2018), Tschannen, Bachem & Lucic (2018), and Crescimanna & Graham (2020). A comprehensive tutorial on VAEs can be found in Doersch (2016). Information-theoretic extensions to VAEs have also been studied recently by e.g. Crescimanna & Graham (2020) and Rezaabad & Vishwanath (2020).

Perhaps the closest to our implementation is the parallel work of Iten et al. (2020) where a β-VAE is used to look for traces of physics in latent representations. However, in that work only 'toy examples' based on simulations are tried, with a rather shallow non-convolutional network. This makes the work orthogonal to our long-term goal of 'learning from data': simulations are created based on simplistic mathematical models we already know. Hence, they can teach us, at best, the things we already know.

Moreover, for a network to be able to learn *semantics* from data, it needs to be (a) presented with huge amounts of *real* data, to avoid overfitting and falling in the covariate shift trap (Sugiyama & Kawanabe 2012), and (b) at the same time sophisticated and deep enough to learn *useful* representations.

There have also been a few attempts towards finding physical parameters in spectra based on typical dimensionality reduction methods such as principal component analysis (PCA; Jolliffe & Cadima 2016). However, PCA provides a linear decomposition of data and hence, as expected, does not yield the desired one-to-one mapping between the principal components and physical features – e.g. see Bailer-Jones, Irwin & Von Hippel (1998). We illustrate such an effect on our data set in Appendix C.

Our contributions

(i) To the best of our knowledge, this is the first work to allow deep convolutional neural networks to learn to infer (astro-)physical parameters just by looking at *real* data, with *zero supervision*.

(ii) We provide methods based on MI and statistics, to track true correlation between learned representations and physical parameters, as well as autodiscovery of the potentially informative latent dimensions.

(iii) We identify but leave open, cues for doing science with potentially new patterns that neural networks discover in data.

Section 2 presents the basic deterministic convolutional autoencoder we start our study with. Section 3 explains how we enforce interpretability of the learned representations via disentanglement. Section 4 details the specifications of the data set. In Section 5, we briefly look at reconstruction results. Finally, in Section 6, we analyse the learned latent representations and assess traces of physics in them.

2 A DETERMINISTIC CONVOLUTIONAL AUTOENCODER

Although the final implementation of the proposed method involves treatment of the input and the latent representation as statistical variables, in this section, we start by detailing the architecture of a *deterministic* deep convolutional autoencoder (Vincent et al. 2010) and training details. This allows us to clarify the migration from a traditional *fully connected* autoencoder to a convolutional one, as well as to briefly illustrate that even the deterministic variant is capable of learning useful information from Big Data.

2.1 Architecture

We design an autoencoder composed of a combination of convolutional, up-convolutional and fully connected layers (Fig. 2). A fully detailed illustration of the network architecture is presented in Appendix A. There are 15 convolutional layers in the *encoder* part that transform the input spectrum, x, down to 512 vectors of length 20 (in case of HARPS). The vectors are then transformed to a single vector of scalars, called *code*, using a fully connected layer. The code, also referred to as the *latent representation* throughout this article, contains the most compressed version of the input spectrum throughout the network. The dimensionality of this vector is chosen based on the desired compression rate. We experiment with different code sizes, from 2 to 128. On the other side of the bottleneck, a second fully connected layer transforms the code back to a similar set of 512 vectors. Then a set of up-convolutional layers take them step-by-step up to the same dimensionality as the original input (327 680 for HARPS).

2.2 Reconstruction loss

E_ϕ and D_θ represent the deterministic encoder and decoder, respectively, where ϕ and θ are the learn-able parameters of the network. We aim for pixel-level accuracy in the reconstructed spectrum and so choose to minimize the per-pixel L1 loss function:

$$\mathcal{L}_{\mathrm{AE}}(\theta, \phi) = \mathbb{E}_{\mathrm{data}} \left[||x - D_\theta(E_\phi(x))||_1^1 \right], \tag{1}$$

which is empirically computed as

$$\mathcal{L}_{\mathrm{AE}} = \frac{\sum_{i \in \mathcal{M}} |x_i - \hat{x}_i|}{n}, \tag{2}$$

where x and $\hat{x}$ are the input and reconstructed spectra, respectively, i is the pixel index and n is the total number of pixels.

Set $\mathcal{M}$ represents a mask, constant over all the spectra in the data set, which masks out the three *information gaps* in the beginning, middle, and end of each HARPS spectrum (Pepe et al. 2002). This is a safe procedure, because these are just instrumental artifacts that bear no meaning for the astrophysical interpretation of the spectra.[4]

2.3 Median normalization

For stability of the training process, we want the input samples not to feed extremely different value ranges into the input of the network. Thus, without loss of generality, we normalize the spectra in the data set according to

$$x = \frac{\mathring{x}}{\underset{i \in \mathcal{M}}{\mathrm{median}}\{\mathring{x}_i\}} \tag{3}$$

in which $\mathring{x}$ is the original input spectrum before normalization.

Our initial experiments show that a deterministic autoencoder not only can compress and reconstruct the whole data sets with as few as eight nodes at the bottleneck and with a high quality, but also can grasp a degree of understanding about the underlying signal sources. This is reflected in the way the network treats the telluric

[4]The location of such artifacts is not exactly fixed across different spectra. Therefore, we chose to use a single constant mask to cover all of them, at the cost of losing a small fraction of informative pixels from each spectrum.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

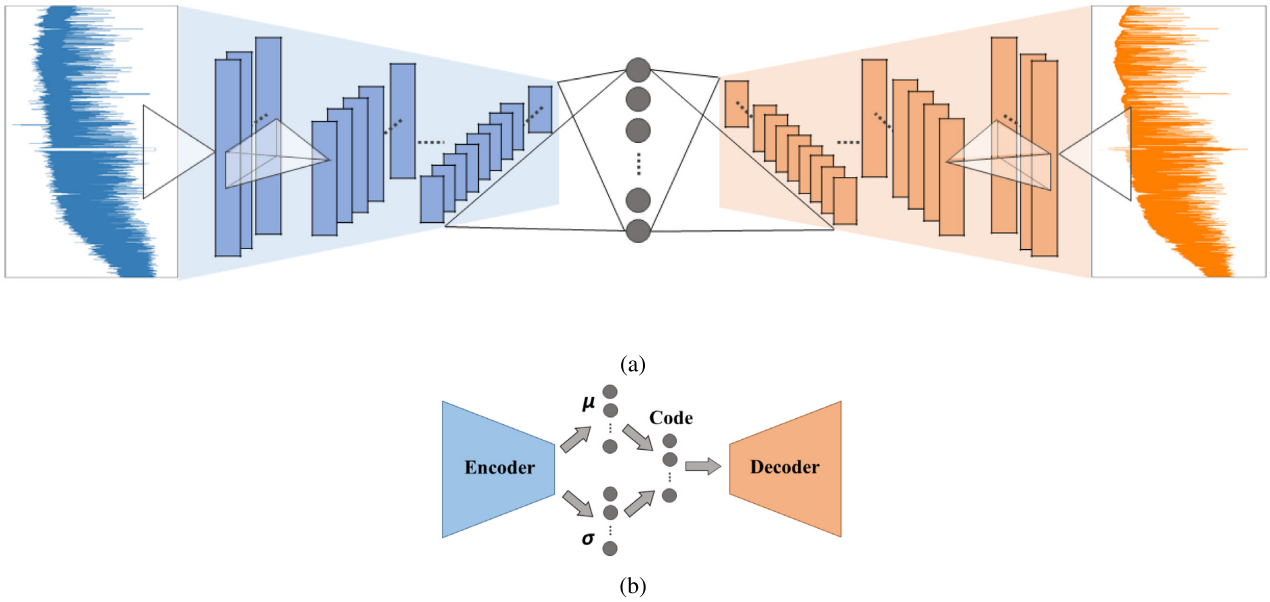

Figure 2. Brief architecture of the deterministic autoencoder on top, with the schematic variational counterpart of it at the bottom. In the VAE version, the code is not directly connected to the encoder, but is drawn from the learnable parameters of the normal distribution: *reparametrization trick* (Kingma & Welling 2014).

lines differently to other (stellar) lines. Details of this part of the study will be published in a future article.

3 ENFORCING INTERPRETABILITY

Learning disentangled representations for composing factors of observed phenomena is key to interpretability (Bengio et al. 2013).

Although our deterministic autoencoder proves to be capable of learning interesting aspects of the observations, the de facto methods of enforcing disentanglement in deep autoencoders are built on top of the VAE-based family of methods, and are done by regularization of the variational autoencoder objective, one way or another (Tschannen et al. 2018).

We convert our classic autoencoder to a VAE, as seen in Fig. 2, where the deterministic code is replaced by a probabilistic one and each element of it is drawn from a normal distribution defined by a pair of learnable parameters: mean (μ) and standard deviation (σ).

In the most basic form of a VAE, the objective is of the form:

$$\mathcal{L}_{\rm VAE}(\theta, \phi) = \mathcal{L}_{\rm reconst}(\theta, \phi) + \mathbb{E}_{\rm data}[D_{\rm KL}(q_\phi(z|x)||p_\theta(z))], \quad (4)$$

where z is the latent variable, $p_\theta(z)$ is the prior distribution on the latent space. $q_\phi(z|x)$ is the approximation of the posterior, learned by the encoder and $D_{\rm KL}$ represents the Kullback–Liebler divergence (Kullback & Leibler 1951).

Higgins et al. (2017) introduce β-VAE in which more disentanglement is enforced by increasing the weight (λ) of the second term:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{\rm reconst}(\theta, \phi) + \lambda\mathbb{E}_{\rm data}[D_{\rm KL}(q_\phi(z|x)||p_\theta(z))] \quad (5)$$

which from another perspective, pushes for maximizing the MI between z and x – e.g. see Burgess et al. (2018). We follow the same formulation for enforcing disentanglement in our implementation. However, we find that pushing for too much disentanglement by setting λ to too high a value, even values close to 1 as suggested by Kingma & Welling (2014) and Higgins et al. (2017), results in too much loss of reconstruction quality, rendering it against the main goal of this work. We assess this trade-off between disentanglement and reconstruction quality in the upcoming sections and find $\lambda = 0.3$ a reasonable choice for the current task.

4 DATA SET

The data set is built from observations using the HARPS instrument, a fibre-fed high-resolution echelle spectrograph dedicated to the discovery of exoplanets (Mayor et al. 2003). The spectrograph has a resolving power of 115 000 and covers the spectral range 378–691 nm. We use the ∼270 000 HARPS fully reduced spectra available in the ESO Science Archive[5] in our investigations.

The data set consists primarily of stellar spectra, although has an extended diversity due to the presence of Solar system objects such as Jupiter and its Galilean moons, and asteroids. Although these objects are potential contaminants, we decide to leave them in the data set, to keep the degree of supervision close to zero. We only had to remove unusable spectra: the ones containing undefined or unrealistic flux values, reflecting instrumental errors.

The spectra are homogenized by trimming down to the same minimum (3785 Å) and maximum (6910 Å) wavelengths, and then zero-padded either side to the reach the same number of pixels. We chose this length to be $327\,680 = 2^{18} + 2^{16}$ – reasonably close to a power of 2 for computational purposes. With the same resolution (0.01 Å), the wavelengths in the spectra are therefore represented by the index of the flux vector. The result is a one-dimensional input for the network to train on.

[5]The retrieval form to access these spectra is at http://archive.eso.org/wdb/wdb/adp/phase3_main/form

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

4.1 Imbalanced Observations

Any data set can potentially have different numbers of observations (instances) for different objects. An extreme example in the case of HARPS is HD128621 (α Cen B) for which there are ∼20 000 instances in the data set, whereas many other objects have been observed only once.

Just like in any other data-driven method, ignoring this effect, which is quite similar to a *selection function*, would allow dominant objects to inject bias and prevent the learned features from being representative of the whole data set. But in order to stay fully unsupervised we take two parallel approaches and compare the results: First we implement a *visibility balancing* technique in which visibility weights are incorporated during training, set to be inversely proportional to the occurrence frequency of each object in the data set. Then we also run the same experiments ignoring the imbalance.

As we will see in the upcoming sections, the major physical concepts that are captured by the network remain consistent across the two experiments. However, as expected, some other nodes start to learn features influenced by the dominant (class of) objects.

Also, in some of the test experiments, we are interested in looking at each object only once. We extract a 'unique' list of objects for this purpose, in which multiple observations of each object are discarded and simply the first one is picked. We extract the number of occurrences only based on the 'target-name' field in the database. While the target names in HARPS are not 100 per cent reliable, we decide to accept the error as it can only influence the results in a negative way, and does not introduce any kind of false hope. In the 272 376 spectra queried from the data base at the start of the work,[6] we get 7653 unique target names.

5 RECONSTRUCTION RESULTS

5.1 Deterministic autoencoder

Theoretically, the quality of the reconstructed spectra should heavily depend on the size of the bottleneck, as it reflects the amount of preserved information.

Reconstructions with various bottleneck sizes are displayed in Fig. 3. Interestingly, with a bottleneck as low as eight dimensions, we already get a very good reconstruction of most of the spectra.

With only two latent dimensions, the network tends to preserve only the overall shape of the spectrum. Conversely, the higher the number of bottleneck dimensions is, the more accurately the output follows fine features of the input. A detailed analysis of this behaviour is beyond the scope of this paper and will be provided in an upcoming article.

5.2 With disentangled features

Fig. 3 also depicts reconstruction examples with disentangled features. As expected, disentanglement comes at the cost of losing reconstruction quality. Hence, to obtain a high degree of reconstruction quality and disentanglement at the same time, the bottleneck needs to have a higher number of dimensions.

5.3 Training set versus validatation set

We split HARPS into training and validation subsets simply based on the index, after being sorted on the 'ADP ID' field. The field presents just a unique identifier and does not have any meaningful correlation with real-world features, such as observation time or object type, and is therefore safe for the purpose.

The split has been used to monitor the training process and avoid overfitting. We also investigated possible differences in reconstruction quality across the two subsets subjectively, and found no meaningful difference.

[6]We make the subset available to public.

6 THE PHYSICS THE NETWORK LEARNS TO INFER

The main objective is not for the network to reconstruct the input with a high accuracy, but rather to learn a minimal *useful* representation of the spectra. In this section, we try to interpret the learned features, and seek to find traces of physical semantics. We pursue *ablation study* by cracking open the network and analysing the statistical behaviour of the latent nodes.

To this end, we forward-pass an ensemble of spectra half-way through the network and store the ensemble of latent representations, to form a $n \times d$ matrix of *codes*. This compact matrix, in practice, contains the whole ensemble, in a compressed format, and suffices for all statistical analyses. We use the unique subset introduced in Section 4.1 for this purpose, since dominant objects in the data set, like α Cen-B with ∼20 000 instances, would bias and occlude our analyses otherwise.

6.1 Informative dimensions

Our very first analysis is to find out how many *informative* features the network really has learned. To this end, we utilize median absolute deviation (*MAD*), as a robust measure of statistical dispersion as an initial score of informativeness. The score for the ith latent node (Z^i) is computed as

$$\mathrm{MAD}^i = \underset{j}{\mathrm{median}} \left(|Z^i_j - \tilde{Z}^i| \right), \tag{6}$$

where j iterates over samples (spectra) and $\tilde{Z}^i = \underset{j}{\mathrm{median}}(Z^i)$.

Although such a dispersion measure is, by definition, tied to the diversity of the underlying data set, still any *important* property of the samples should show enough variability across different samples – or else it contains close to zero *information* for our purpose, hence deemed unimportant.

There is one degree of freedom (hyper parameter) which seems to affect the number of informative nodes: the disentanglement weight (λ) of equation (5). In Fig. 4, we see that, lower levels of disentanglement simply result in too many *significant* dimensions, which cannot be called informative anymore, as disentanglement is not really happening. Fig. 5 depicts how two significant dimensions may still be highly correlated – evidence that the disentanglement has failed.

Too much disentanglement, on the other hand, results in fewer significant dimensions, which may seem as a good outcome in the first look. However, our experiments show that reconstruction quality decays so much that fine details are discarded and the few learned features are all centred around the overall shape of the spectra – Fig. 6. This trade-off is a well-studied characteristic of unsupervised disentanglement methods – e.g. see Burgess et al. (2018). Networks with other bottle-neck dimensionalities follow the same trend, although narrower bottlenecks inherently tend to (have to) discard fine details.

We find that a disentanglement weight λ of around 0.3 provides a reasonable trade-off, where no two significant dimensions show

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

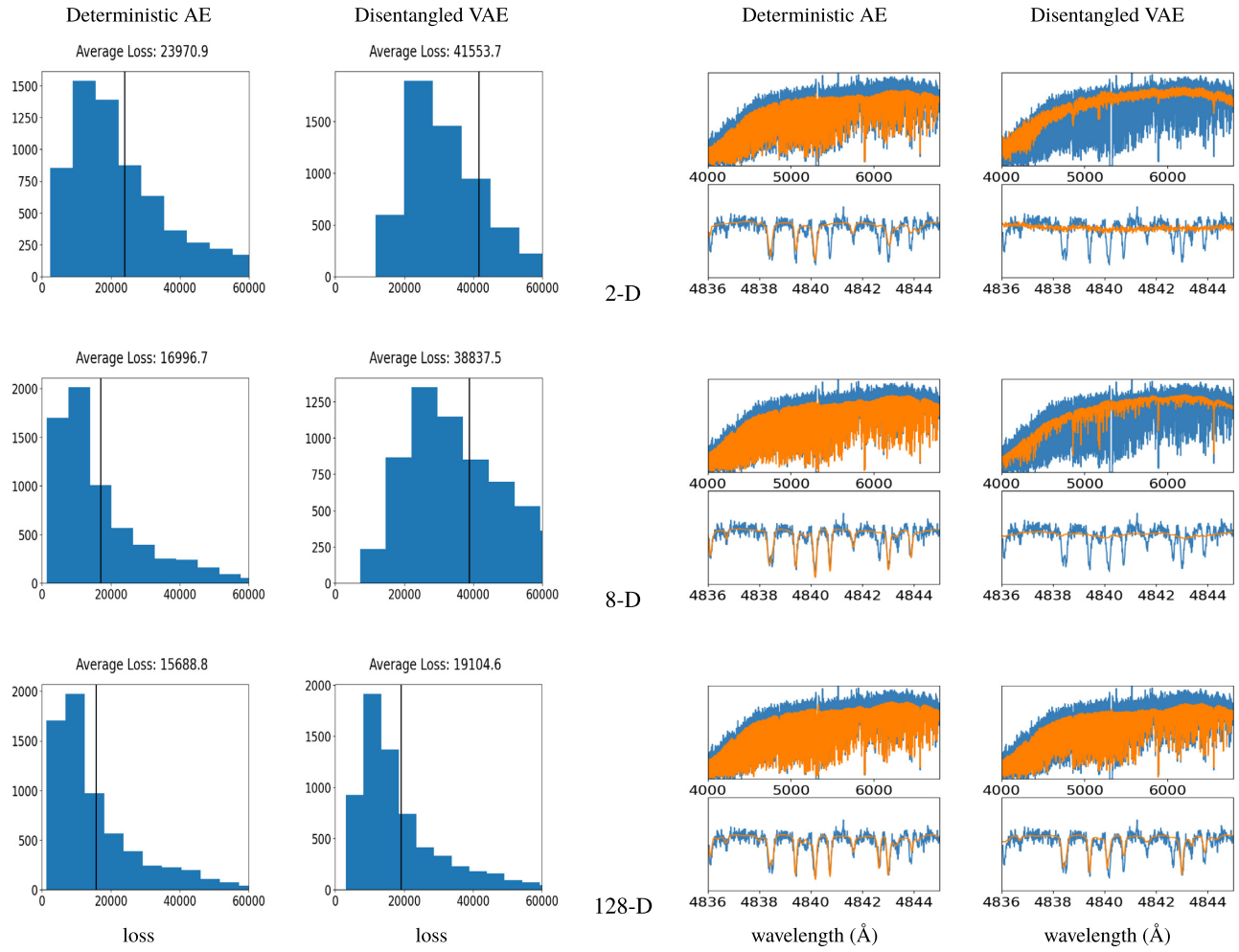

Figure 3. Illustration of the effects of two major factors on reconstruction quality: latent space dimensionality and disentanglement. The left two columns illustrate reconstruction loss over the whole spectra, while on the right the same effects are depicted, in two different zoom levels, on an exemplar single spectrum: Input (blue) and reconstructed version (orange) are overplotted. Comparing the results of the deterministic autoencoder, and that of the disentangled variational autoencoder, we can clearly see the sacrifice in reconstruction quality, that occurs for the sake of disentanglement. On the other hand, as we increase the number of latent dimensions (top-down direction in the figure), reconstruction quality for fine details is enhanced.

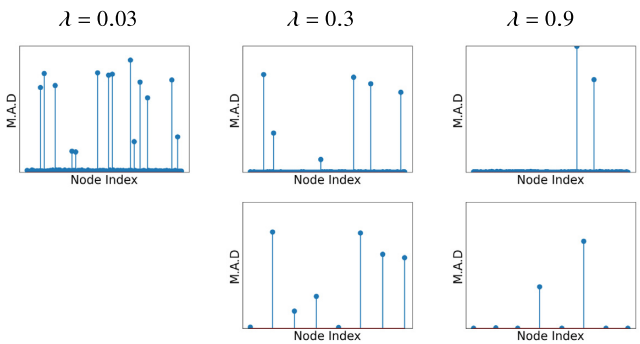

Figure 4. M.A.D. values for 128-d network on the top and 8-d network on the bottom. From left to right, the disentanglement weight (λ) is increased. Too low weights result in leak of information among different dimensions, while too high values cause loss of details which causes better disentanglement, yet less useful features. Interestingly the 128-d and 8-d networks agree on the number of informative features at $\lambda = 0.3$.

significant correlation – i.e. good disentanglement. Interestingly, we find exactly six informative latent dimensions in two different networks with latent dimensionalities of 8 and 128.

In the next section, we take an information theoretic approach towards detection of traces of physics in latent features, which is completely independent of the informativeness indicator of this section. But as we move forward we find a reassuring harmony between the two methods.

6.2 Mutual information – with known physics

So far we have identified the dimensions which, from a purely statistical point of view, seem to have captured *significant* features of the stars. Now we seek to interpret the learned features and find specific traces of physics. The search is conducted over all the latent features, to avoid any bias from the statistical scores of previous section.

Assuming we have access to a large number of known (astro-)physical parameters, we seek *MI* between them and the latent features the network has learned. Pearson correlation is too limited

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

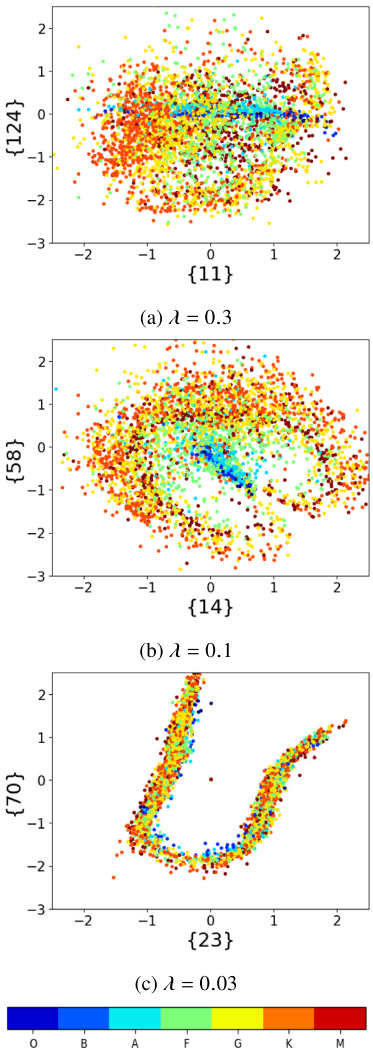

Figure 5. Scatter plots illustrating mutual behaviour of pairs of latent dimensions. On the top, there is little to no significant correlation between the two. In contrast, the bottom two plots show clear correlation between exemplar dimension pairs, in networks where λ has been too low, which is a strong hint for failure of disentanglement. In such cases, a high M.A.D. does not directly translate to possession of exclusive information. Contrary to intuition, the less structured the plots are, the more successful the disentanglement has been. Different colours show different spectral classes and are used for illustration purposes only.

as it can only capture linear dependence with Gaussian noise, while 'MI is able to quantify the strength of dependencies without regard to the specific functional form of those dependencies' (Kinney & Atwal 2014).

MI of two jointly discrete random variables is defined as (Cover 1991)

$$I(X;Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{(X,Y)}(x,y)\log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\,p_Y(y)}\right). \quad (7)$$

A more intuitive formulation is given by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (8)$$

and defines MI as the amount of uncertainty lost in one of the variables by knowing the other one. In equation (8), $H(.)$ is the *Shannon Entropy* (Shannon 2001).

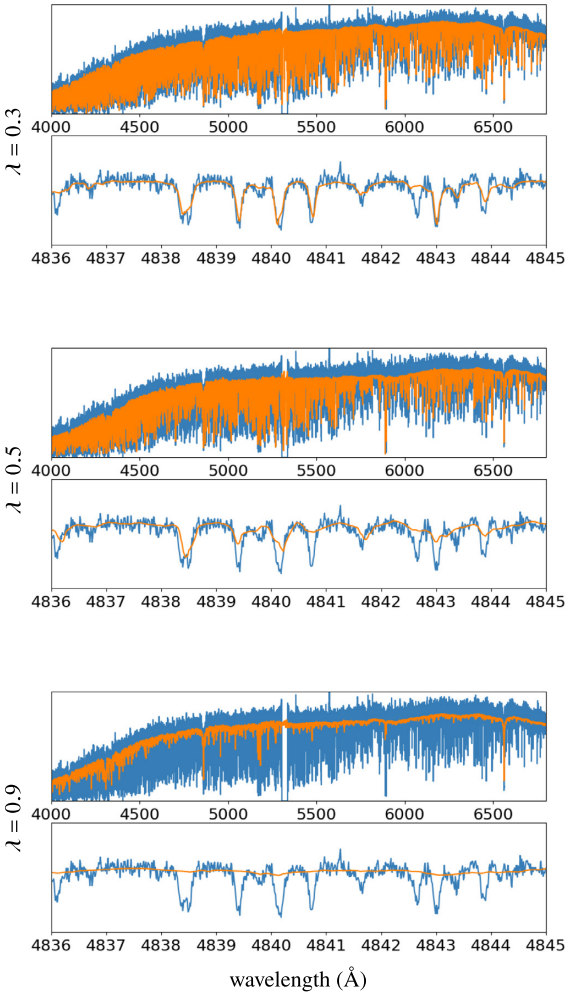

Figure 6. From top to bottom, the effect of too much disentanglement enforcement is visualized. The network loses the ability to preserve details, i.e. narrow lines, and starts focusing on the overall shape only. In such a case, although the significant dimensions learn disentangled representations, the captured concepts are too simplistic and not much useful.

Given a number of data points, it is often difficult to obtain an accurate estimate of the MI of the underlying random variables, as it involves estimation of the underlying joint distribution. For the task at hand, however, we are not much interested in the exact value of the MI, as it is a relative indicator when considering all latent dimensions.

We use joint histograms to simply approximate the joint density. Still, the estimated MI's turn out to be quite sensitive to the chosen number of bins. Therefore, to have a simple, yet robust indicator, we provide a two-step workaround: (a) sigma-clipping at 5σ, and (b) multiscale (scan at various bin resolutions).

We extracted some of the known astrophysical features, for a portion of our data set, from SIMBAD (Wenger et al. 2000), TIC Stassun et al. (2019), and observation-time parameters:

(i) effective temperature ($T_{\rm eff}$)
(ii) surface gravity [log(g)]
(iii) metallicity ([M/H])
(iv) radial velocity
(v) airmass
(vi) signal-to-noise ratio (SNR)

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

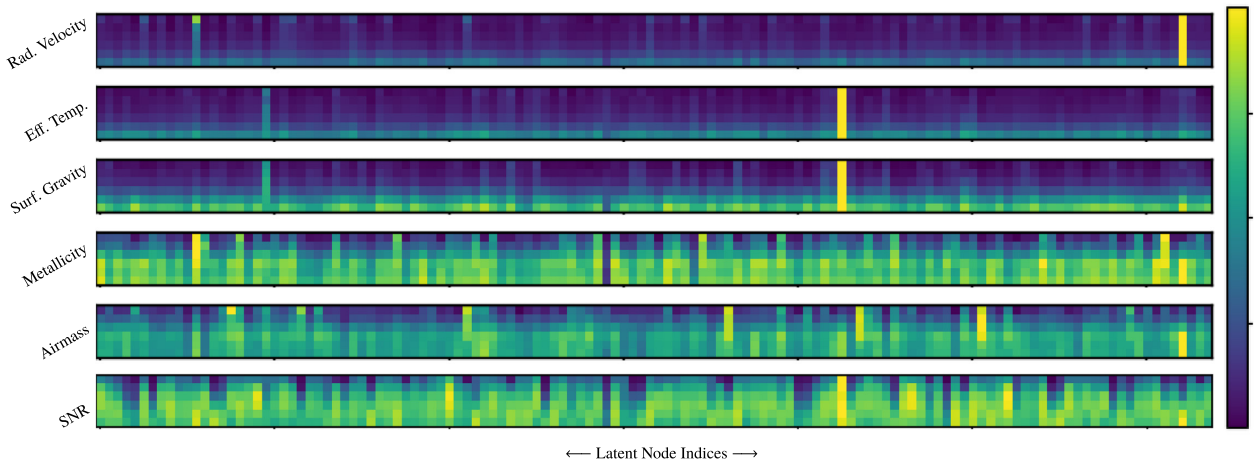

Figure 7. Correlation indicators based on MI at different scales. The depicted matrix at each row shows different scales (binning configurations) along the vertical axis and different nodes are sitting horizontally. Each row of each indicator, representing a single scale, is normalized by max. For radial velocity, effective temperature, and surface gravity, individual nodes stand out, while for metallicity, airmass, and SNR, that is not the case.

Steps of the process are detailed in Appendix B.[7]

We construct MI indicators as explained above, to seek traces of these intrinsic astrophysical stellar parameters in all dimensions of our networks. Results for the 128-dimensional network are illustrated in Fig. 7. Clear signs of strong correlation are seen for radial velocity at dimension {124}, and $T_{\rm eff}$, $\log(g)$ at dimension {85}. No clear dimension stands out for [M/H] airmass and signal-to-noise ratio (SNR).

The two detected 'physical dimensions' have already been identified by the purely statistical indicator of the previous section, which increases the reliability of the finding. Visualization of the direct relationship between latent features and their corresponding validation labels in Figs 8 and 9, shows that the network has clearly grasped a direct notion of these physical concepts.

6.2.1 Analysis

Node {85} shows correlation with both effective temperature and surface gravity. Its correlation with the effective temperature is clear, monotonic and tight, providing close to a one-to-one mapping from node values to temperatures – Fig. 8, top row.

The reason surface gravity is captured with the same dimension becomes clearer after plotting the scatter of the two physical parameters (not the node values) against each other – bottom row of Fig. 8. It turns out that the input data set presents a biased view when it comes to temperature and gravity, in that it does not sample uniformly the general underlying stellar population. Concretely speaking, in the objects the network has seen, temperature and surface gravity are more or less strongly correlated. From an information theoretic point of view, surface gravity does not provide much exclusive information, and a big fraction of the information in it is shared with effective temperature. In other words, the network does not need to dedicate an independent node to store information about this physical parameter, when it can obtain most of what it needs from another node – especially under disentanglement pressure. Of course, the network needs to store the exclusive part of the information about this parameter, which is reflected in the scattered points in the plot, somewhere. That place is most likely in one of the discarded nodes.

Node {124} has captured information on the stars' radial velocity. The correlation is shown visually in Fig. 9. The plot shows that the network has automatically learned a model for hypothetical, reference, zero-velocity spectra, since it has formed a symmetric mapping around it. The mapping is of course not a bijective function. It is also worth noting that for colder stars the correlation is quite tight and progressively loosens for hotter stars, until it essentially vanishes at the highest temperature available in our data set. We speculate that the increasing sparseness of absorption features with increasing temperature is responsible for the observed behaviour.

The spectral absorption from the Earth atmosphere as parametrized by the airmass affects the large-scale shape of the spectra, a prominent feature that could be expected to be picked out by the network. The same could be expected for metallicity. A posteriori, however, this does not seem to be the case since neither of these parameters are significantly correlated with any of the dimensions, as gauged by the MI results, which may look puzzling at first glance. This may be, however, related to the fact that HARPS has a relatively narrow wavelength range, mostly bluewards of most telluric features. HARPS is mostly an exoplanet hunter, and those are mostly looked at around solar-like or cooler stars, and our sample is strongly biased against containing early-type stars. This can be seen in Fig. 8, where it is also clear that our data set is mostly comprising main-sequence stars. It also covers a limited range in metallicity, while the optimized `New Short Term Scheduler` used by most HARPS visitors implies that most targets are observed at the best (i.e. lowest) airmass possible. It is therefore not surprising that the algorithm could not find a correlation with metallicity and airmass.

One may also expect SNR to be captured by the network as an independent feature, since it plays a role in forming the appearance of an spectrum. This is, however, not the case and comes as little surprise; the noise is uncorrelated with any other type of information in the data set and by definition does not contain any *pattern* across different spectra to be learned. Thus, for a model to capture and reconstruct pixel-accurate noise, it would need to assign one parameter per pixel per spectrum – i.e. memorize the noise. This advantageous limitation is a well-known feature of even the simplest classical autoencoders, such that denoising autoencoders have been

[7]We re-emphasize that the learning process has been a fully unsupervised one and such labels have been merely used post-training for validation purposes only.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

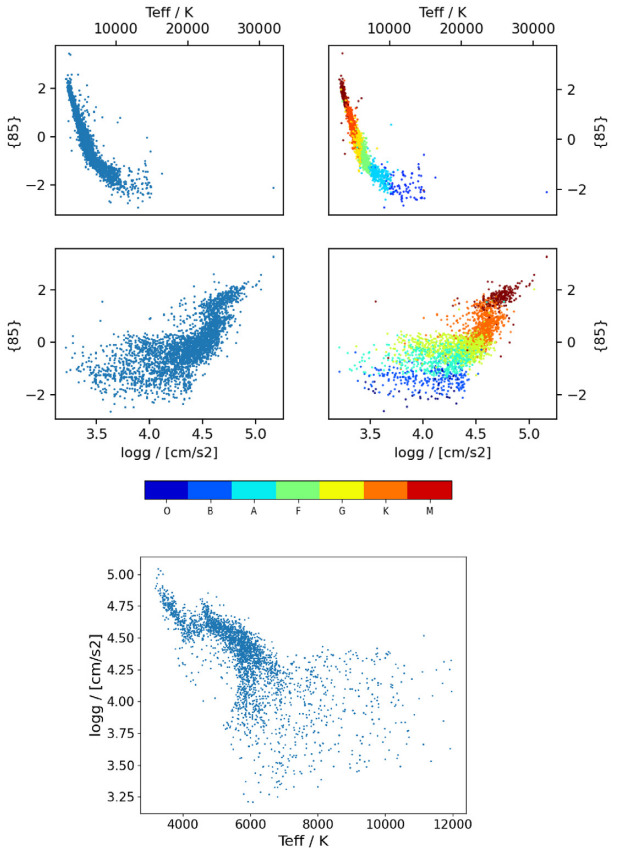

Figure 8. Node {85} shows a good correlation with effective temperature – top row. The tightness of the structure reflects the strength of the MI. The same node shows a not-so-strong correlation with surface gravity – middle row. Plotting log(g) versus $T_{\rm eff}$ in the bottom row reveals the reason. Please refer to the main text for a detailed analysis. It is also useful to note that our sample is very biased towards main sequence stars, with the log g only varying between ∼3.5 and 5.

among the first ones to be used (Vincent et al. 2010). Such behaviour is of course seen in many other methods used for dimensionality reduction, such as PCA – e.g. see Bailer-Jones et al. (1998).

6.3 Latent space traversal

Although we run out of available physical labels or/and automatically detected correlations, we go further and pursue deeper investigation based on a method known as *latent space traversal*. We start by forward-passing single spectra half-way through the network, just the way we did in the beginning of this section, to encode the spectrum into its latent representation. Then by perturbing (or traversing, in extreme cases) the code and generating the corresponding spectrum, we can have synthetic spectra which are different to the (reconstructed version of) the original spectrum as a result of the change in the code. So, singling out dimensions of the latent space allows for analysis of the effects of specific dimensions on the generated spectra, hopefully equal to interpretable features.

To this end, we create an interface with sliders which allow for traversal over different dimensions and visualization of the effects on the fly – Fig. 10. In the following, we list the significant findings.

Node {**11**} seems to be, partly, related to the rotation of the star, which is another parameter that is known to affect the spectra – a higher rotation will broaden the lines, making them less deep. Varying the value of this node does not affect the shape of the continuum,

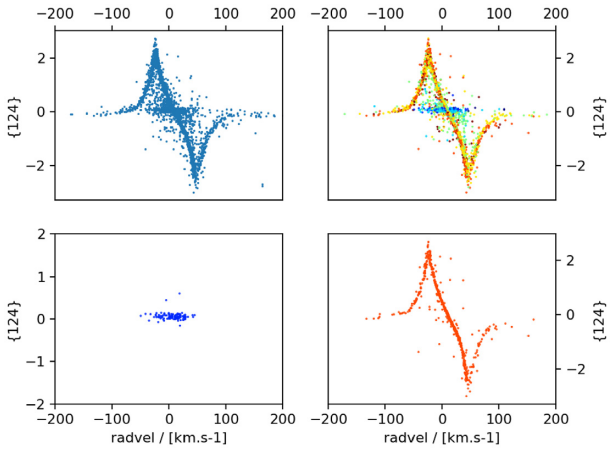

Figure 9. Node {124} learns a clear understanding of a notion of radial velocity – top row. The symmetric shape, and the fact that the network has automatically gained an understanding of *zero velocity* as a reference point are notable observations. Different temperatures have apparently been treated differently, as also detailed in the bottom row.

but only the depth of the lines. Thus, an increased value of the node corresponds to much broader lines and this is clearly an effect of increasing rotational velocities (or macroturbulence in general). Above a given threshold, however, the situation is more complex: for solar-like stars, the match seems to be done only on stars that have quite a large radial velocity shift. We have not yet found a physical reason for this. For early-type stars, the lines do not become broader either, but instead the Balmer lines clearly become narrower. This is likely an effect of the gravity of the star.

Node {**19**} is only affecting a subset of our sample, namely only the coolest stars. It has indeed no effect on solar-like stars or early-type stars, but only affects stars that have a value of node {85} above about 0.85, that is, stars cooler than ∼4500 K. For these stars, this node is clearly linked with the luminosity of the star. This node is thus also physical, but confined to a subset of the stars, only the coolest ones – see Fig. 10 for an illustration.

Node {**58**} has, similarly to above, no apparent effect on the spectra of solar-like or early-type stars, but only manifests itself for even cooler stars than node {19}, those that are characterized with a value of node {85} above about 1.2. However, we could not find as yet a clear explanation of the effect at play when varying the value of node {58}, and we defer a detailed analysis to further work.

Node {**99**} is contrarily to node {11} affecting the continuum of the star, more than the lines themselves. It is also, unlike the previous two nodes, not really affecting solar-like and cooler stars, but has only a visible effect on stars hotter than the Sun. From a phenomenological point of view, this node appears to be looking at the inflexion point of the continuum and whether the spectrum is thereby concave or convex. Thus, for very negative values (e.g. −4.5), there is a depression in the spectrum around 5800 Å, which disappears at about −1.7, while for positive values, there is a maximum around 5300 Å. The clear physical explanation of this apparent phenomenological node is hard to find, but a first investigation indicates that it may be related to the presence of a disc (such as around Be stars) or a companion. Further studies are needed.

6.4 Discarding observation frequencies

Using the non-balanced data set, we obtain five significant nodes, two of which correspond exactly to the major captured physical features: {85} and {124}.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

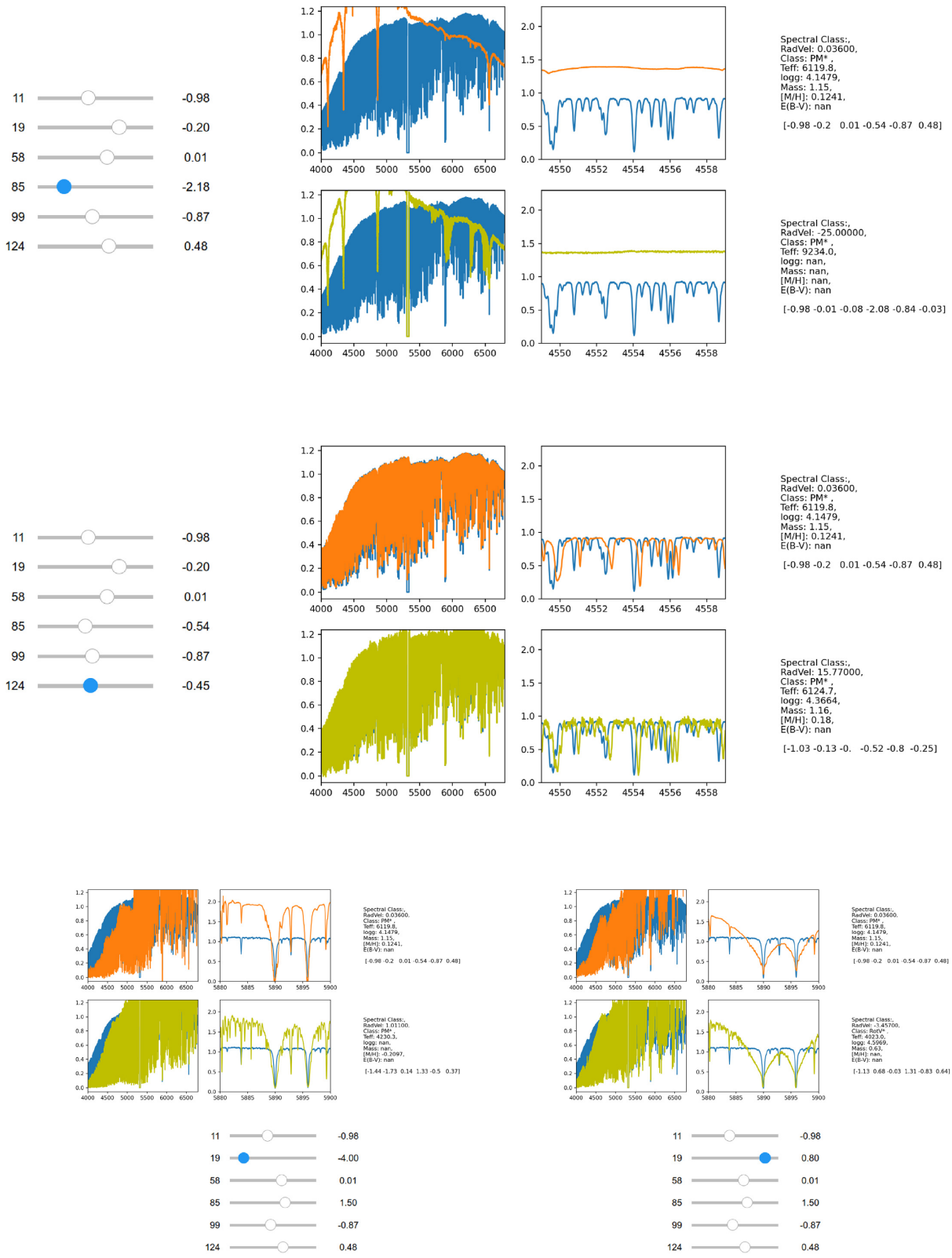

Figure 10. Our interface for latent space traversal, showing three different experiments. All experiments share the same randomly chosen 'reference' star, shown in blue. This reference is encoded by the network, the obtained code is slightly modified using the sliders, and decoded to generate the orange spectrum. This resulting spectrum is usually an imaginary one and thus, we illustrate the closest real object to it in green. This closest object is searched for in the learned latent space. From top to bottom, we show experiments for the effects of {85} (effective temperature), {124} (radial velocity), and {19}, respectively. For the latter, which applies only to cool stars, we had to 'move' the base spectrum to a late-type star, using {85}.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

Two nodes represent features that are also seen as in the balanced set: {11} and {88} (the latter, corresponding to node {99} of the balanced set).

Representations captured in nodes {19} and {58} of the balanced net are clearly not present any more, as we cannot spot any node specifically representing only the coolest stars. The effect of the remaining node, {99}, is not clear cut. It seems that for the hottest stars (>10 000 K), it is partially sensitive to the gravity of the stars: the lowest values of this node correspond to white dwarfs (i.e. high gravity), while the highest values correspond to solar-like stars. It has no apparent effect for A/F/G stars, nor for M stars, but there is an effect on K stars as well. We could not identify the physical nor phenomenological criteria that would correlate with this node.

7 CONCLUSION

We implemented the idea of 'letting the data speak for itself' in action, in the context of an astrophysical application, where we let a deep convolutional neural network look at stellar spectra and learn from them without any predefined objectives in mind. We showed that the network 'chose to' learn how to extract and capture specific physical parameters of stars, among other unidentified ones, as their canonical features. The importance of the finding is in network's answer to 'what is important to learn?', and should not be confused with the relatively trivial problem of training a network for estimation of those parameters.

Specifically, our purely statistical measure revealed that 6 out of 128 latent nodes of our network stand out as informative ones. We also developed an information-theoretic indicator to track true/non-linear correlations between the learned features and a set of known astrophysical parameters. We found that two latent nodes, which interestingly turned out to be among the six informative ones, have clearly learned a notion of radial-velocity and effective temperature.

The automatic method did not indicate correlations between the remaining significant dimensions and the validation labels we had at hand. This does not necessarily indicate a false alarm on those nodes. They may have captured known physical parameters for which we do not have labels yet, or the existing labels might have not been quite reliable to reveal weaker correlations.

Also, it is quite possible that the other nodes have not captured direct representations of familiar physical parameters, but rather other complex (or even simpler) features. Artificial neural networks do not have to think like humans! For example, We spot nodes which capture variations of specific absorption lines. They may have captured fine features of chemical abundances – something that is not formulated in classical astronomy, with this level of granularity. We believe such features that are not directly interpretable are interesting for follow up studies, since understanding the reasons behind a network's *decision* to prioritize more complex/simpler features, or higher level relationships between basic features, may help advance our physical understanding of the underlying target – stars in this case.

We continued with latent space traversal and found traces of rotation, luminosity, presence of a disc or a companion, in the unidentified nodes, some affecting only a subset of our sample (either the coolest stars, or the hottest). The latter correlations were, however, not as clear as the previous ones and were decided to be left for future studies. We make the interface available to public for this purpose.

As mentioned earlier, our data set for this case study is very specific, due to the particularities of HARPS usage. It is to be expected that in more generic samples, other features, e.g. luminosity or metallicity, may come out more easily. In general, the concepts the network learns to capture, are dependent on the biases in the data set.

ACKNOWLEDGEMENTS

This work is in part supported by the ESCAPE project (the European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures) that has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement no. 824064. We also acknowledge support for our research by funding from the Science and Technology Facilities Council. Lastly, we thank Michael F. Sterzik, Mark Allen, Henri M. J. Boffin, and Felix Stoehr for their help in preparation of the manuscript.

DATA AVAILABILITY

We release the code for the convolutional neural network, the list of IDs of the spectra used for training and validation and the physical validation labels on https://www.eso.org/~nsedagha/universe. We also make the 'sliders' interface freely accessible to the community to facilitate study and discovery of new relationships with the introduced framework.

REFERENCES

Bailer-Jones C. A., Irwin M., Von Hippel T., 1998, MNRAS, 298, 361
Baron D., 2019, preprint (arXiv:1904.07248)
Baron D., Poznanski D., 2017, MNRAS, 465, 4530
Bengio Y., Courville A., Vincent P., 2013, IEEE Trans. Pattern Anal. Mach. Intell., 35, 1798
Boucaud A. et al., 2020, MNRAS, 491, 2481
Burgess C. P., Higgins I., Pal A., Matthey L., Watters N., Desjardins G., Lerchner A., 2018, preprint (arXiv:1804.03599)
Chen R. T. Q., Li X., Grosse R. B., Duvenaud D. K., 2018, in Bengio S., Wallach H., Larochelle H., Grauman K., Cesa-Bianchi N., Garnett R., eds, Advances in Neural Information Processing Systems, Vol. 31. Curran Associates, Inc., p. 2610
Choudhary A., Lindner J. F., Holliday E. G., Miller S. T., Sinha S., Ditto W. L., 2020, Phys. Rev. E, 101, 062207
Cover T. M., 1991, Elements of Information Theory. Wiley, New York
Crescimanna V., Graham B., 2020, International Joint Conference on Neural Networks. IEEE, Glasgow, United Kingdom, p. 1–8
D'Agnolo R. T., Wulzer A., 2019, Phys. Rev. D, 99, 015014
Denil M., Agrawal P., Kulkarni T. D., Erez T., Battaglia P., de Freitas N., 2017, Proceedings of the International Conference on Learning Representation (ICLR). OpenReview.net, Toulon, France
De Simone A., Jacques T., 2019, Eur. Phys. J. C, 79, 289
Doersch C., 2016, preprint (arXiv:1606.05908)
Ehrhardt S., Monszpart A., Mitra N. J., Vedaldi A., 2017, preprint (arXiv:1703.00247)
Graham M. J., Djorgovski S., Mahabal A. A., Donalek C., Drake A. J., 2013, MNRAS, 431, 2371
Greydanus S., Dzamba M., Yosinski J., 2019, in Wallach H., Larochelle H., Beygelzimer A., Alché-Buc F. d., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems 32. Curran Associates, Inc., p. 15379
Higgins I., Matthey L., Pal A., Burgess C., Glorot X., Botvinick M., Mohamed S., Lerchner A., 2017, Proceedings of the International Conference on Learning Representation (ICLR). OpenReview.net, Toulon, France
Hinton G. E., 2006, Science, 313, 504
Iten R., Metger T., Wilming H., del Rio L., Renner R., 2020, Phys. Rev. Lett., 124, 010508
Jolliffe I. T., Cadima J., 2016, Phil. Trans. R. Soc. A, 374, 20150202
Kingma D. P., Welling M., 2014, Proceedings of the International Conference on Learning Representation (ICLR). OpenReview.net, Banff, AB, Canada
Kinney J. B., Atwal G. S., 2014, Proc. Natl. Acad. Sci., 111, 3354

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

Krizhevsky A., Sutskever I., Hinton G. E., 2012, Proceedings of the Advances in Neural Information Processing Systems (NIPS). Curran Associates Inc., Red Hook, NY, p. 1097
Krone-Martins A., Ishida E. E. D. O., De Souza R., 2014, MNRAS, 443, L34
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
LeCun Y., 1985, In Proceedings of Cognitiva 85, Une Procedure D'apprentissage Pour Reseau a Seuil Asymmetrique (A learning Scheme for Asymmetric Threshold Networks). Paris, France, p. 599
Lukic V., Brüggen M., 2016, Proc. IAU, 12, 217
Martinazzo A., Espadoto M., Hirata N. S., 2020, preprint (arXiv:2004.11336)
Mayor M. et al., 2003, The Messenger, 114, 20
Meng X., Li Z., Zhang D., Karniadakis G. E., 2020, Comput. Methods Appl. Mech. Eng., 370, 113250
Ochsenbein F., Bauer P., Marcout J., 2000, A&AS, 143, 23
Pepe F. et al., 2002, The Messenger, 110, 9
Perryman M. A. C. et al., 1997, A&A, 323, 49
Quinn P., Axelrod T., Bird I., Dodson R., Szalay A., Wicenec A., 2015, preprint (arXiv:1501.05367)
Raina R., Madhavan A., Ng A. Y., 2009, in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, p. 873
Raissi M., Perdikaris P., Karniadakis G. E., 2017, preprint (arXiv:1711.10561)
Rezaabad A. L., Vishwanath S., 2020, IEEE International Symposium on Information Theory (ISIT). IEEE, Los Angeles, CA, p. 2729
Romaniello M. et al., 2018, in Peck A. B., Seaman R. L., Benn C. R., eds, Proc. SPIE 10704, Observatory Operations: Strategies, Processes, and Systems VII. SPIE, Austin, Texas, p. 1070416
Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533
Santos N. et al., 2004, A&A, 426, L19
Schmidt M., Lipson H., 2009, Science, 324, 81
Sedaghat N., Mahabal A., 2018, MNRAS, 476, 5365
Sedaghat N., Zolfaghari M., Brox T., 2017, preprint (arXiv:1612.03777)
Shannon C. E., 2001, ACM SIGMOBILE Mobile Comput. Commun. Rev., 5, 3
Stassun K. G. et al., 2019, AJ, 158, 138
Stewart R., Ermon S., 2017, Thirty-First AAAI Conference on Artificial Intelligence. The AAAI Press, Palo Alto, California
Sugiyama M., Kawanabe M., 2012, Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation. MIT Press
Taylor M. B., 2005, in Patrick L., Shopbell M. C., Britton R. E., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29
Tishby N., Pereira F. C., Bialek W., 1999, Proc. 37th Allerton Conf. on Communication, Controland Computing, 368
Tschannen M., Bachem O., Lucic M., 2018, preprint (arXiv:1812.05069)
Vanzella E. et al., 2004, A&A, 423, 761
Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol P.-A., Bottou L., 2010, J. Mach. Learn. Res., 11, 3371
Wenger M. et al., 2000, A&AS, 143, 9
Wulff E., 2020, Deep Autoencoders for Compression in High Energy Physics. Available at: https://lup.lub.lu.se/student-papers/search/publication/9004751
Zhang R., Liu Y., Sun H., 2020, Eng. Struct., 215, 110704
Zhao S., Song J., Ermon S., 2018, preprint (arXiv:1706.02262)

APPENDIX A: NETWORK ARCHITECTURE

Fig. A1 illustrates details of the deterministic autoencoder. The VAE version follows the exact same architecture, and differs only around the bottleneck, as illustrated in Fig. 2.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

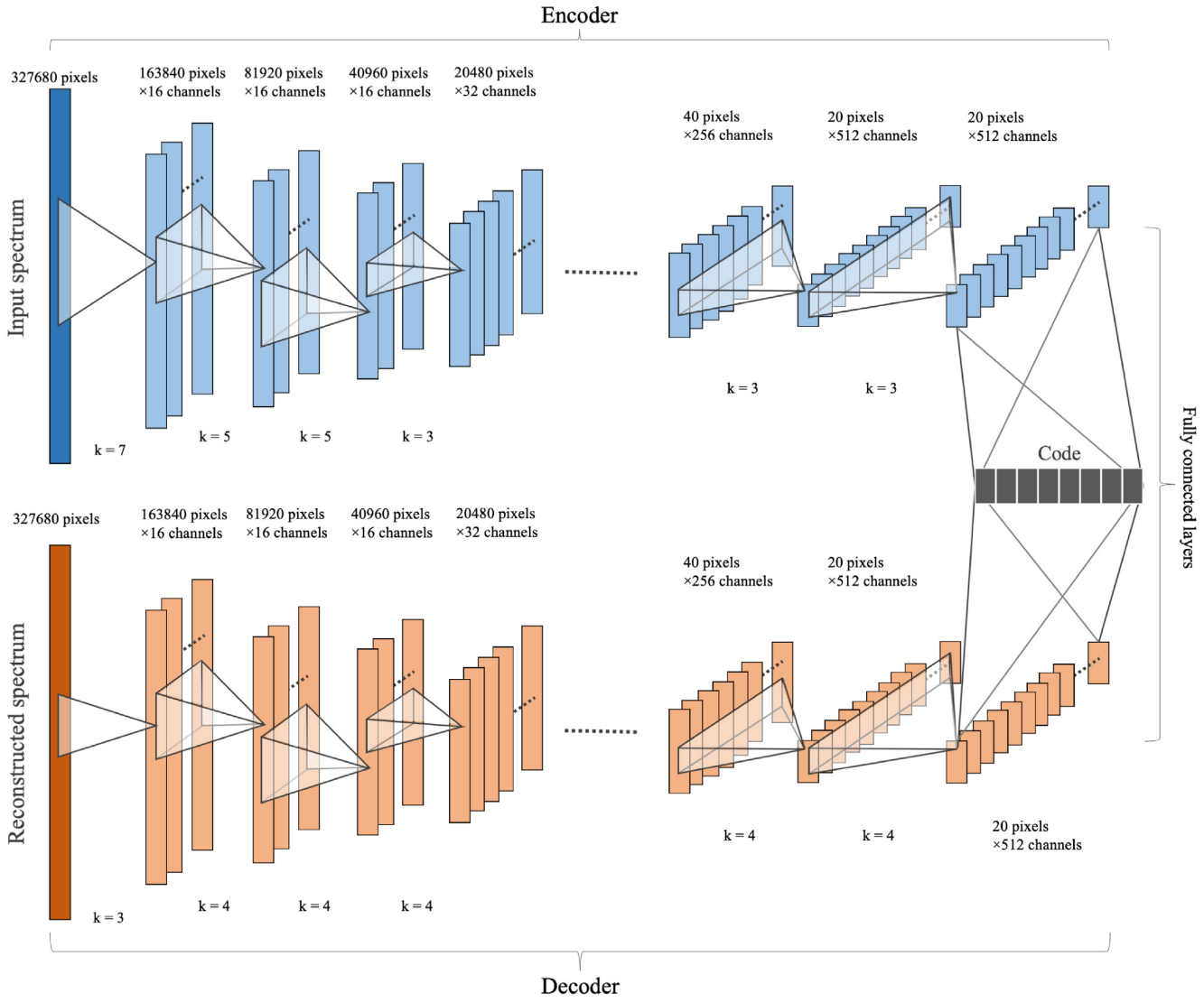

Figure A1. Detailed architecture of the deterministic autoencoder. Due to lack of space, not all the layers have been visualized. The missing information can be extracted from the released source code.

APPENDIX B: RETRIEVING VALIDATION LABELS

To collect a large number of existing physical labels from the literature, we use both SIMBAD (Wenger et al. 2000) and the TESS Input Catalogue (TIC; Stassun et al. 2019), a 'compiled catalogue of stellar parameters'.

B1 Cross-matching process

We produce metadata of our HARPS subset as a reference table. Each row of the table contains information regarding a single spectrum, including its position on the sky. The table contains possibly many observations of the same star, as discussed in Section 4. Moreover, the position accuracy is quite low and the photometric counterpart may be located at a distance of tens of arcseconds from the input table position.

In the first step, we consider each entry of the table as an independent object. We load the reference table into TOPCAT (Taylor 2005) and perform a cross-match with the remote VizieR version of the TIC table using the CDS Xmatch service from TOPCAT. We perform a simple cone-search cross-match returning all TIC objects in a radius of 40 arcsec around each of the input table positions. This results in more than 5 million matches (5610 122 exactly): the TIC being deep, we get a large number of spurious association in such a large search box.

Plotting angular separation versus magnitude V versus distance of the TIC stars plot (Fig. B1), good matches seem to be separable from the spurious ones based on the $V_{\rm mag}$. It complies with the prior knowledge that most HARPS objects have a magnitude lower than 12–13 (the green points on the plot). This magnitude corresponds more or less to the limiting magnitude in the *Hipparcos* and Tycho catalogue (HIP Perryman et al. 1997). We thus decided to filter TIC data to keep only objects having an observation in HIP. With this

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

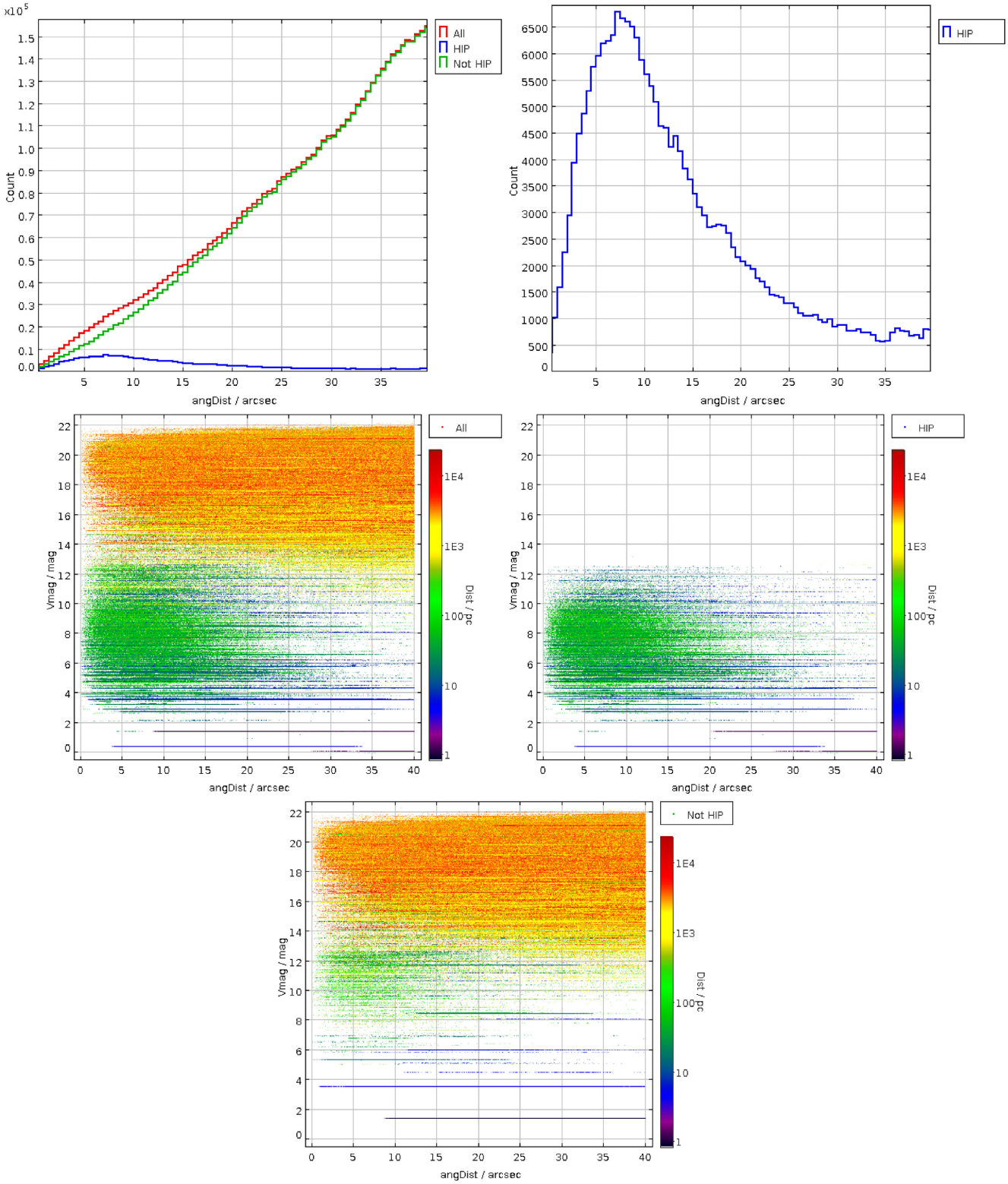

Figure B1. HARPS versus TIC cross-match. Top left: angular separations (in arcsecond) for all cross-matches (red), the HIP selection (blue), and its complement (green). Top right: re-scaled histogram of the angular separations for the HIP selection. Bottom: angular separation (in arcsecond) versus magnitude *V* versus distance of the TIC stars. The left-hand, centre, and right-hand panels show all matches, the selected HIP matches, and its complement, respectively.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

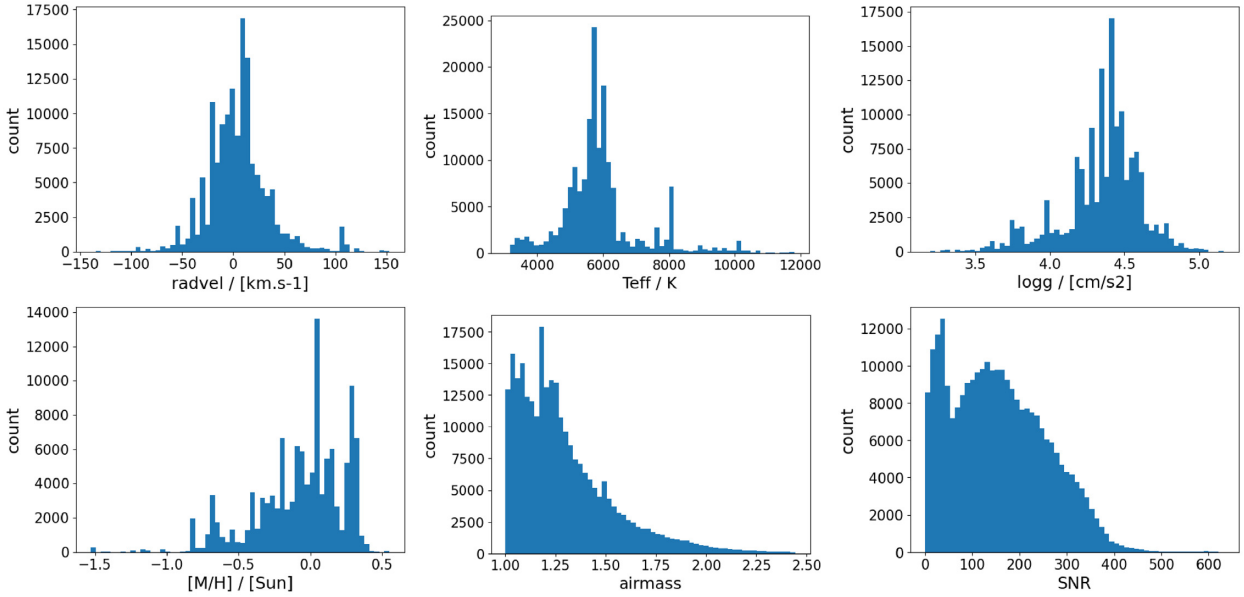

Figure B2. Distributions of successfully acquired labels. It should be noted, however, that these only represent a subset of HARPS spectra, and do not necessarily represent the exact distributions of the parameters in HARPS.

single selection criteria, we put a (loose) constraint on magnitudes and ensure a better homogeneity of the selected sample, since all objects have been observed at least in the *Hipparcos* catalogue. This leaves us with 209 183 (4 per cent) associations. With this selection, we probably miss good matches in a mag range of 10–12.5, the green points on the right part of the plot.

Initially, the histogram of the angular separations of all matches is dominated by spurious matches (almost only the linear – Poisson – component is visible). But selecting only HIP objects in the TIC catalogue, the histogram is now dominated by good matches (the linear component being quite low).

After this first selection, we still get HARPS entries associated with multiple TIC objects. As we favoured reliability over completeness, we removed those objects resorting to an internal match in TOPCAT. We get an output 185 662 HARPS spectra associated each with a single TIC entry.

To add SIMBAD 'labels', we finally cross-match our results with SIMBAD, keeping the closest match in a 2 arcsec radius around the TIC object positions. 5000 HARPS objects are lost at this last step.

In the end, we get 179 389 matches with

(i) 151 743 radial velocities
(ii) 120 440 metallicity
(iii) 145 372 mass
(iv) 145 372 log(g)
(v) 167 728 $T_{\rm eff}$.

The cross matching and the resulting labels are by no means complete. The labels may not be quite accurate either. This, however, suffices for our validation experiments as we seek only the overall possible patterns and correlations, not the exact values. Figure B2 illustrates distributions of the acquired labels.

APPENDIX C: PRINCIPAL COMPONENT ANALYSIS

In Fig. C1, we depict how our MI indicators would work on the first 128 principal components of our data set. As expected, being essentially a linear transformation, PCA should not be expected to result in any sort of 'smart' disentanglement of features.

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023

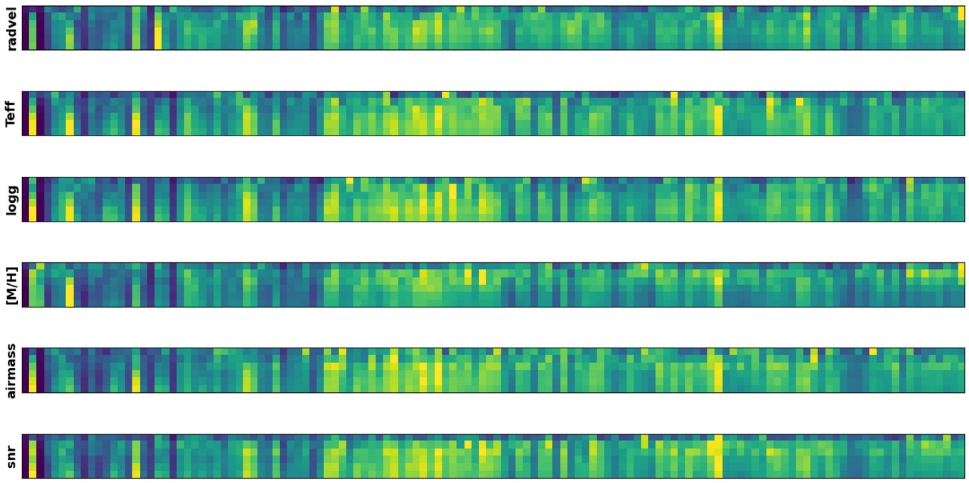

Figure C1. MI indicators for detection of traces of physical parameters in the first 128 components of PCA. As expected, no clear traces of individual parameters can be seen; in other words, information about each physical parameter is spread over many dimensions.

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

Downloaded from https://academic.oup.com/mnras/article/501/4/6026/6121645 by CNRS user on 04 May 2023