Deep Learning Analyses of Synthetic Spectral Libraries With an Application to the Gaia-ESO Database

by

Spencer Bialek
B.Sc., University of Victoria, 2017

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Physics and Astronomy

Deep Learning Analyses of Synthetic Spectral Libraries With an Application to the Gaia-ESO Database

by

Spencer Bialek
B.Sc., University of Victoria, 2017

Supervisory Committee

Dr. Kim Venn, Supervisor
(Department of Physics and Astronomy)

Dr. Sébastien Fabbro, Co-Supervisor
(Department of Physics and Astronomy)

ABSTRACT

In the era of stellar spectroscopic surveys, synthetic spectral libraries will form the basis for the derivation of the stellar parameters and chemical abundances. In this thesis, four popular synthetic grids (INTRIGOSS, FERRE, AMBRE, and PHOENIX) are used in a deep learning prediction framework ("StarNet"), and compared in an application to observational optical spectra from the Gaia-ESO survey. The stellar parameters for temperature, surface gravity, metallicity, radial velocity, rotational velocity, and [α/Fe] are determined simultaneously for FGK type dwarfs and giants. StarNet was modified from its application to SDSS APOGEE infrared spectra, not only to optical wavelengths, but also to mitigate the differences in the sampling between the synthetic grids and the observed spectra, and by augmenting the grids with realistic observational signatures, in an attempt to incorporate both modelling and statistical uncertainties as part of the training. When applied to spectra from the Gaia-ESO spectroscopic survey and the Gaia-ESO benchmark stars, the INTRIGOSS-trained StarNet showed the best results with the least scatter. Training with the FERRE synthetic grid produces similarly accurate predictions (followed closely by the AMBRE grid), but over a wider range in stellar parameters and spectroscopic wavelengths. This is an exciting and encouraging result for the direct application of synthetic spectra to the analysis of the planned spectroscopic surveys in the coming decade (WEAVE, 4MOST, PFS, and MSE). In the future, improvements in the underlying physics that generates these synthetic grids can be incorporated for consistent high precision stellar parameters and chemical abundances from machine learning and other sophisticated data analysis tools.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My family, for their unrelenting, incredible support. I could not have made this journey without you, you were with me every step of the way.

My supervisor, Kim Venn, for always believing in me, challenging me to help me grow, motivating me and helping me feel excited about my work, achieving the delicate balance of giving me independence and assisting me when necessary, and for being a wonderful friend and mentor.

My co-supervisor, Sébastien Fabbro, for the stimulating discussions, all the help (and there was a lot) with diagnosing my Linux and computing issues, for supporting me in improving my coding and research skills, for the delicious food, and for being a great friend through it all.

My partner, Katelyn Bunn, for gib food and gib drink, for helping me celebrate the highs, and more importantly for being a dependable and loving partner, always. You are a fantastic human to have in my life <3 Thanks for flying across the country to be with me, and thanks for bringing some incredible cats with you. Look out for your name in the acknowledgements of my Ph.D dissertation, after many more adventures. You are my Number One.

My pals on the fourth floor of Elliott, for making this journey bearable, fun, and the best path I ever could have chosen.

My dear friends, of which there are far too many to name. Sharing my passion with you and the intrigue and warm reception I get in return have helped keep my spark alive. My office (kitchen) is always open. I love you all!

Someone once told me that time was a predator that stalked us all our lives. I rather believe that time is a companion who goes with us on the journey and reminds us to cherish every moment, because it will never come again. What we leave behind is not as important as how we've lived. After all Number One, we're only mortal.

Jean-Luc Picard

Chapter 1

Introduction

The bulk of this thesis includes the development of a novel data analysis and processing pipeline, based on machine learning, to study the properties of hundreds of thousands of stars in our galaxy, the Milky Way. It is necessary to begin by providing the context for why astronomers care about understanding the characteristics of stars in great detail, how this task has been historically completed, what challenges arise in the modern era of big data collection, and finally, what machine learning is and how it can be utilized to solve these unique challenges.

1.1 Analyzing the light from the stars within our Galaxy

Through physical processes operating on an enormous range of physical scales and time scales, the gas, dust, and stars within galaxies have evolved in complex ways throughout the history of our universe. Our observations of the regularity of galaxies today betrays this complexity, and it is an ongoing challenge in astrophysics to explain how such ordered properties can emerge from such complex physics. The interstellar material and stars we observe today encode information about their evolution, and thus knowledge of the properties and evolution of galaxies can be acquired through a careful examination of the light emanating from the objects within them (Freeman & Bland-Hawthorn, 2002).

Our unique place within the Milky Way galaxy offers us an opportunity to record the light from a huge number of its stars. Finding meaningful relationships between the stars which exist today and our galaxy's formation and history depends on the quality and amount of information we can decode from starlight. A rich avenue for this task is through the transformation of the light from a star into its constituent wavelengths, forming a stellar

spectrum. A spectrum encapsulates fundamental physical parameters of a star, including its kinematics, chemistry, and age, and thus high quality spectra, collected in spectroscopic surveys, are desired by astronomers studying our galaxy.

1.1.1 Stellar Spectroscopic Surveys

Astronomers have been collecting spectra of hundreds of thousands of stars in the Milky Way galaxy for several years now, beginning with surveys like the Sloan Digital Sky Survey (SDSS) Sloan Extension for Galactic Understanding and Exploration (SEGUE), in which spectra of over 200,000 unique stars were collected for investigating the structure of the Milky Way (SEGUE-1; Yanny et al. 2009a) and spectra of over 100,000 unique stars occupying the *in situ* galactic halo were collected for better understanding the formation of the outer halo (SEGUE-2). SEGUE helped to uncover the rich kinematic and chemical substructures in the halo and thick disc. Other surveys like the LAMOST Experiment for Galactic Understanding and Exploration (LEGUE; Deng et al. 2012) and the SDSS Apache Point Observatory Galactic Evolution Experiment (APOGEE), which have collected spectra for ∼6 million and ∼0.4 million stars, respectively, have helped in sampling all the major components of the Milky Way, providing detailed chemical abundances and kinematics.

Exciting new surveys of our sky, seeking to systematically observe millions of stars in fine detail, are currently being planned at optical and infrared (IR) wavelengths over the next decade. Many of these will be "blind surveys", wherein astronomers record the light from as many stars in as many regions of the galaxy as possible – providing a global map of our galaxy that is contiguous and densely sampled – so they can find groups of stars which are chemo-dynamically similar but dispersed (indicating the disruption of a dwarf galaxy through an ancient merger with our galaxy or other cluster of stars, e.g. Helmi et al. 2018), or stars that are chemically peculiar and rare, e.g. carbon-enhanced and extremely metal-poor stars (thought to be remnants of the first generation of stars, e.g. Starkenburg et al. 2017).

SDSS-V (Kollmeier et al., 2017) will observe 5 million unique stars, helping to form a massive spectroscopic census of the stars in the disk and bulge which will contain detailed information of ages, kinematics, and chemical abundances as a function of three-dimensional position in our sky (see Figure 1.1 for coverage). ESO's 4MOST (de Jong et al., 2012) is a similarly ambitious project, but will additionally focus on high galactic latitudes, collecting spectra of ∼1.5 million stars in the Galactic halo. The gaps that 4MOST will miss in the northern hemisphere will be filled in by WHT Enhanced Area Velocity

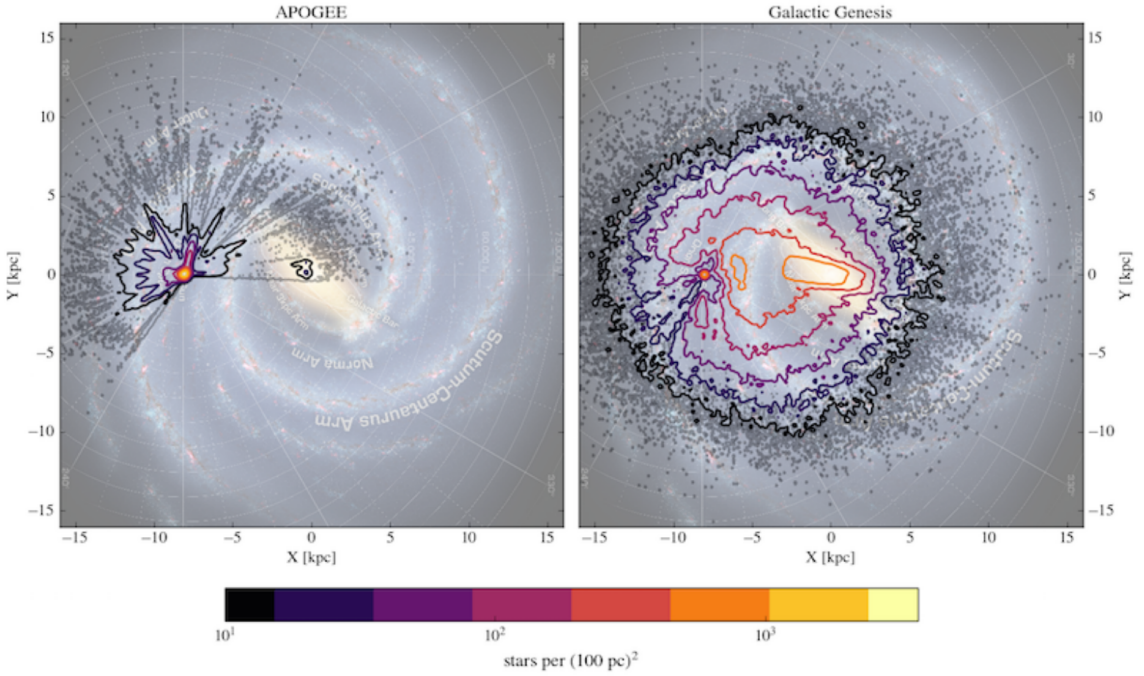

Figure 1.1: How the in-plane target density will evolve with SDSS-V: contours showing the surface density of the APOGEE DR14 catalog (left) and SDSS-V's Galactic Genesis Survey (GGS; right). The contours contain stars within 500 pc of the midplane, summing to 1.5x10^5 in APOGEE DR14 and 3.6x10^6 stars in GGS.

Explorer (WEAVE; Dalton et al. 2012) and its Galactic Archaeology survey, which will target faint stars in the outer disk and Galactic halo. Surveys like these will collect an enormous amount of valuable data for astronomers and will help them address long-lived questions of our Galaxy like its hierarchical accretion history, its formation mechanisms, the properties and characteristic parameters of its dark matter halo, the origin, structure, and dynamics of its disk (including radial migration, the bar and spiral arms, and its vertical structure) and how the Milky Way fits into a cosmological context.

Answering these important questions will require exquisite precision in the derived properties of stars. Determining how to acquire the necessary information from a stellar spectrum in an efficient, accurate, and precise way, has been an ongoing challenge in modern astrophysics.

1.1.2 Processing the spectra

Traditionally, a telescope armed with a spectrograph would observe one star at a time, collecting one spectrum in a single integration. Once a collection of spectra was recorded,

the astronomer would then, one at a time, laboriously analyze the spectra and derive properties of each – a very inefficient process. The process of deriving the stellar parameters usually involved a by-hand comparison of the observed spectrum to synthetic models of spectra (e.g. by using MOOG software; Sneden et al. 1997). Synthetic spectra are still used as the basis for more modern automated methods.

The creation of synthetic models of stellar spectra was a project started several decades ago (e.g., Kurucz, 1970) and requires a detailed understanding of the physics involved in stellar atmospheres, in particular the stellar photosphere: atomic and quantum theories dictate the excitation and ionization states of atoms (as a function of temperature and pressure, via the Saha-Boltzmann equations) and the probability that particular wavelengths of light will be absorbed by those species of atoms and molecules as light propagates from the inner regions of a star to its photosphere (via solutions to the radiative transfer equation). The atomic and molecular data used in these equations is, surprisingly, still incomplete and continuously being improved upon (e.g. see Kurucz, 2014; Franchini et al., 2018).

1.2 Machine Learning

To maximize the scientific impact of spectroscopic surveys, astronomers are starting to develop the necessary data processing backbones to tease out as much useful information from the stars as possible. The requirements for these backbones, which will be novel due to both the massive amounts of data being collected and the level of precision and accuracy needed, are uniquely met by the careful implementation of machine learning methods. Indeed, there have been a number of recently published methods, e.g. “The Payne” (Ting et al., 2019), “The Cannon” (Ness et al., 2015a; Casey et al., 2016a), “AstroNN” (Leung & Bovy, 2018), and our application "StarNet" (Fabbro et al., 2018), all of which rely on analytic and machine learning algorithms to derive the fundamental stellar properties from the spectra of stars.

In supervised machine learning methods, the task given to the machine is to minimize the discrepancy between the predictions of the model and the desired or known outputs of the data. It is analogous to teaching a child how to classify objects, by repeatedly telling the child what the desired output is and correcting the child if they are incorrect. In the case of a child, one might ideally strengthen the learning by using positive reinforcement (e.g. “Good job Farbod, that is indeed a cat!"), but for a machine, the learning is typically strengthened with negative reinforcement in the form of a *loss function*: the output of the loss function is relatively large if the prediction is incorrect, and relatively small if the prediction is correct,

so the machine will adjust the model parameters to incur as small of a punishment, or *loss*, as possible. This process of course necessitates that the data being used is already labeled, i.e. the output is known beforehand.

1.2.1 Neural networks

Neural networks (NNs), a popular type of machine learning algorithm, have a history of use in astrophysics going back more than 20 years (Von Hippel et al., 1994). In Bailer-Jones et al. (1997) and Bailer-Jones (2000), a neural network was applied to synthetic stellar spectra to predict the effective temperature $T_{\rm eff}$, surface gravity $\log g$, and metallicity [Fe/H]. Machine learning methods were also used in one of the SEGUE pipelines (Lee et al., 2008), where two NNs were trained: one on synthetic spectra and the other on previous SEGUE parameters. These earlier applications were quite limited in their use since they required an expertise in machine learning and used flawed algorithms.

More recently, dramatic improvements have occurred in the usability and performance of algorithms implemented in machine learning and NN software, including proper initialization (Glorot & Bengio, 2010), advanced activation functions (Nair & Hinton, 2010), better solvers (Kingma & Ba, 2014), and the development of high-level user-friendly codes (e.g. Keras; Chollet (2015)). Combined with the use of Graphic Processing Units (GPU) for high performance computing and the availability of large data sets, this has led to the successful implementation of more complex NN architectures which have proven to be pivotal in difficult image recognition tasks and natural language processing.

One such example of a more complex NN architecture is the *convolutional* NN (CNN), created by Krizhevsky et al. (2012) for an annual image classification competition, ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The CNN, with their particular architecture now referred to as *AlexNet*, outperformed the competitors of ILSVRC 2012 by a margin of more than 10% in accuracy, leading to the widespread use and further development of CNNs in research.

CNNs and many deep learning methods learn patterns between nearby pixels on ascending levels of abstraction to produce outputs of interest, using thousands to millions of computations at each level. In CNNs, at each of these levels, referred to as layers, this is done by processing the image by convolving it with a number of filters. The resulting maps are then fed to the following layer as an input. After a number of these layers, the output of the last layer is interpreted as the output of the network. The values of the filters, also called network weights, are learned through a process known as training, where pairs of

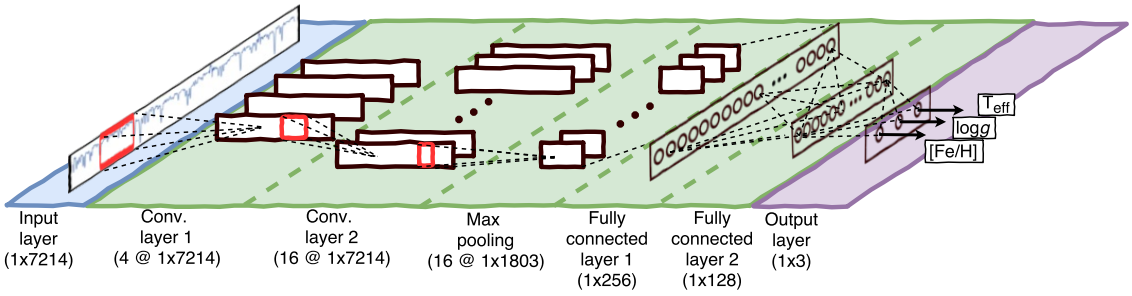

Figure 1.2: The StarNet CNN model composed of seven layers. The first layer is solely the input data; followed by two convolutional layers with four and 16 filters, then a max pooling layer with a window length of four units. A flattening operation allows the output of the max pooling layer to be followed by three fully connected layers with 256, 128, and three nodes. The final layer is the output layer.

correct input-output examples are shown to the network. Given enough training examples, these networks can make accurate predictions on previously unseen examples using these learned parameters.

I helped to develop a CNN, called StarNet (see Figure 1.2 for a schematic), used in the prediction of fundamental stellar parameters from stellar spectra.

The details of StarNet

Fundamentally, a NN is a function which transforms an input to a desired output. The function is composed of many parameters, or nodes, arranged in layers – input and output layers, with hidden layers in between – which form a highly non-linear combination of the input features. Instead of being an exact function, it is approximated and tuned based on data, placing it in the realm of machine learning: the internal parameters of a NN are adjusted to accomplish the particular task given to it. Each node is parameterized as a linear function of weights, $\mathbf{w}$, applied to an input, $\mathbf{x}$, with an additional bias value, $\mathbf{b}$. A node is then activated by a nonlinear function, g, giving an output of a node to be written as:

$$h(\mathbf{x}) = g(\mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathbf{b})$$

Common activation functions include the sigmoid function and the Rectified Linear Unit (ReLU):

$$g(z) = \max(0, z)$$

which allow the network to adapt to non-linear problems (Chen et al., 1990).

In a traditional sequential NN architecture, each node is connected to every node from the previous layer as well as every node in the following layer, thereafter referred to as *fully connected* layers. At hidden layer, l, the output, $\mathbf{h}^{(l)}$, is a vector valued function of the previous layer, $\mathbf{h}^{(l-1)}$, and is given by:

$$\mathbf{h}^{(l)} = \mathbf{g}(\mathbf{w}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})$$

The first layer is simply the input $\mathbf{h}^{(0)}(\mathbf{x}) = \mathbf{x}$, in our case: the spectra.

The next two layers of StarNet are *convolutional layers*, which are more adapted to higher dimensional inputs by leveraging local connectivity in the previous layer. In convolutional layers, the weights are applied as filters. The filter slides across the previous layer taking the dot product between the filter weights and sections of the input. For a given filter covering a section, s, this operation can be summarized as:

$$\mathbf{h}_s^{(l)} = \mathbf{g}(\mathbf{w_s}^{(l)} \otimes \mathbf{h}^{(l-1)} + \mathbf{b_s})$$

These filters allow for the extraction of features in the input and learn which features to extract through training. After the convolutional layers in StarNet, we use a max pooling layer. A max pooling layer is a non-linear down-sampling technique typically used in CNNs to decrease the number of free parameters and to extract the strongest features from a convolutional layer. In a max pooling layer, a window moves along the feature map generated by each of the filters in the previous convolutional layer - in strides of length equal to the length of the window - extracting the maximum value from each sub-region. These pools of maxima are then passed on to the following layer. The next two layers in StarNet are fully-connected layers.

The combination of all those layers allows for the formation of non-linear combinations of an input vector, $\mathbf{x}$, to produce an output vector prediction, $f(\mathbf{x}; \mathbf{w}, \mathbf{b})$. For each training sample of spectra, $\mathbf{x_t}$, and corresponding known stellar parameters, $\mathbf{y_t}$, the NN model weights and biases are estimated by minimizing the *empirical risk* that computes the *loss* between the predictions and targets for a batch of T training samples, often supplemented with a regularizing function. We ended up adopting a mean-squared-error loss function without regularization for StarNet, such that the StarNet empirical risk to be minimized reads:

$$\arg\min_{\mathbf{w},\mathbf{b}} \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{y_t} - f(\mathbf{x_t}; \mathbf{w}, \mathbf{b})\right)^2.$$

The minimization is performed with a stochastic gradient descent (SGD) algorithm. SGD algorithms require the computation of the loss function gradients with respect to the weights, and make adjustments to those weights iteratively until reaching a minimum. In our case, the optimization is performed using the ADAM optimizer (Kingma & Ba, 2014), an SGD variant using adaptive estimates of the gradient moments to adjust learning rates. Initially, the weights of the model are randomly set and therefore the predictions will be quite poor. Computing the gradients is operated backwards through each sequential layer, a process referred as *back-propagation*, and is the computationally expensive part of the training.

In the case of StarNet, a cross-validation set was used to test the model following every iteration to evaluate whether or not the model had improved; if improvements were not made after a given number of iterations, the training was stopped. This minimum may differ depending on the complexity of the model architecture as well as various hyper-parameters. Following each iteration, the cross-validation set is sent through a single forward propagation where the outputs are predicted and compared against the target values. This set is not used for training, but only to ensure that the model is not over-fitting to the training set. Over-fitting occurs when a model learns a function that may be able to compute the outputs for the training set very well, but can not generalize that function to be applied to a test set that is not included in the training. A cross-validation set is used as a type of middle-ground between the training and test set, and ensures that over-fitting does not occur. If the cross-validation predictions do not improve after several iterations, the training will be stopped. Using a cross-validation set to avoid over-fitting and tuning hyper-parameters is common practice in machine learning applications (Gurney, 1997).

In Fabbro et al. (2018), we showed that the stellar parameters (temperature, gravity, and metallicity) for the entire SDSS-III APOGEE spectral database can be determined with similar precision and accuracy as the APOGEE pipeline, in only a few seconds, using StarNet. Ultimately, we showed that machine learning algorithms are excellent tools in the analysis of stellar spectra, and the point of my M.Sc. thesis is to extend the utility of our methods to spectra from any spectroscopic survey.

1.3 Agenda

The following is the outline of this MSc research as presented in this thesis;

Chapter 1, this section, an introduction to stellar spectroscopy, spectroscopic surveys,

machine learning techniques, and neural networks.

Chapter 2, my research project, as described in my submitted paper to MNRAS on the *"Deep Learning Analyses of Synthetic Spectral Libraries With an Application to the Gaia-ESO Database"*.

Chapter 3, a summary of my MSc work, including a Table of my conference presentations on this work, and my future plans to extend this research as a PhD thesis.

Chapter 2

Deep Learning Analyses of Synthetic Spectral Libraries With an Application to the Gaia-ESO Database

The following is the paper that I have lead and submitted to the Monthly Notices of the Royal Astronomical Society (MN-19-4054-MJ). Its contents reflect the research component of my M.Sc. degree. I am the first author, and I wrote nearly the whole paper, including all of the data augmentation, computational developments, and visualizations, with assistance from my supervisor, Prof. Kim Venn, and co-supervisor, Dr. Sébastien Fabbro.

2.1 Abstract

In the era of stellar spectroscopic surveys, synthetic spectral libraries will form the basis for the derivation of the stellar parameters and chemical abundances. In this paper, four popular synthetic grids (INTRIGOSS, FERRE, AMBRE, and PHOENIX) are used in our deep learning prediction framework (StarNet), and compared in an application to optical spectra from the Gaia-ESO survey. The stellar parameters for temperature, surface gravity, metallicity, radial velocity, rotational velocity, and [α/Fe] are determined simultaneously for FGK type dwarfs and giants. StarNet was modified to mitigate the differences in the sampling between the synthetic grids and the observed spectra, by augmenting the grids with realistic observational signatures, in an attempt to incorporate both modelling and statistical uncertainties as part of the training. When applied to spectra from the Gaia-ESO spectroscopic survey and the Gaia-ESO benchmark stars, the INTRIGOSS-

trained StarNet showed the best results with the least scatter. Training with the FERRE synthetic grid produces similarly accurate predictions (followed closely by the AMBRE grid), but over a wider range in stellar parameters and spectroscopic wavelengths . In the future, improvements in the underlying physics that generates these synthetic grids will be necessary for consistent high precision stellar parameters and chemical abundances from machine learning and other sophisticated data analysis tools.

2.2 Introduction

Astronomy has entered an era of spectroscopic surveys. The first large scale spectroscopic surveys, pioneering new methods to efficiently observe and determine spectroscopic parameters, include the Sloan Digital Sky Survey (SDSS) Sloan Extension for Galactic Understanding and Exploration (SEGUE) surveys of over 200,000 stars (Yanny et al., 2009b; Lee et al., 2011) and the RAdial Velocity Experiment (RAVE) survey of nearly 1 million stars (Steinmetz et al., 2006). Since then, the SDSS Baryon Oscillation Spectroscopic Survey (BOSS) has gathered medium resolution spectra for another ~250,000 stars (Abolfathi et al., 2018), and the Large Sky Area Multi-Object Fibre Spectoscopic Telescope (LAMOST) has collected spectra for ~6 million stars (Cui et al., 2012; Zhang et al., 2019). In addition, high resolution spectroscopic surveys have begun to provide precise radial velocities, stellar parameters, and exciting results in chemical abundances for over 400,000 stars, e.g., SDSS APOGEE (Holtzman et al., 2018; Zasowski et al., 2019), and GALAH (Buder et al., 2018). Deeper optical high resolution spectroscopic surveys will soon begin at the 4-metre telescopes, including INT/WEAVE (Dalton et al., 2018) and ESO/4MOST (de Jong et al., 2019), and at the 8-metre telescopes, e.g., Subaru/PFS (Tamura et al., 2018).

To prepare for this era of large data sets, methods to consistently and efficiently analyse stellar spectra are being explored, particularly with sophisticated data analysis algorithms, e.g., “The Cannon" (Ness et al., 2015b; Buder et al., 2018; Zasowski et al., 2019), “The Payne" (Ting et al., 2019; Xiang et al., 2019), and “Matisse" (Recio-Blanco et al., 2006; Kordopatis et al., 2013). We have also been exploring the application of “StarNet", a convolutional neural network (Fabbro et al., 2018). StarNet was found to reproduce the stellar parameters of benchmark stars at least as well as traditional methods, and it could predict the stellar parameters for the entire APOGEE spectral data set within minutes. Furthermore, StarNet was the first application that could be trained either from data with a priori known stellar labels (data-driven mode) or from a synthetic spectral grid (synthetic mode). Leung & Bovy (2018) improved on the data-driven StarNet implementation by

modifying the neural network architecture to track individual abundances, the capability to train on missing or noisy stellar labels, and to estimate prediction uncertainties.

Machine learning methods have now been shown to exceed the performance of traditional methods for spectroscopic analysis, both in terms of time and quality. Machine learning applications are highly versatile, and are an active line of research well beyond astronomical applications, providing a symbiosis where astronomical datasets can both help validate new techniques and also benefit from new analysis methods, e.g., new and clever techniques are being developed to examine the propagation of errors within neural networks (Lakshminarayanan et al., 2017) and generative methods can be used to identify missing physics (O'Briain et al., *in prep*.).

In this paper, we examine the impacts of training StarNet with a variety of publicly available high resolution, optical synthetic stellar grids. These include INTRIGOSS (Franchini et al., 2018), AMBRE (de Laverny et al., 2012), PHOENIX (Husser et al., 2013), and FERRE (Allende Prieto et al., 2018). These grids of synthetic spectra have been generated using independent model atmospheres and radiative transfer codes (all 1D and in LTE), with a range of atomic and molecular opacities required to describe the stellar photosphere. We also considered exploring other available synthetic grids, but found the wavelength coverage too small (e.g., non-LTE grids from M. Kovalev and M. Bergemann, private communications) or that the stellar parameter range was too small (e.g., optical regions of the APOGEE ASSET grid, by S. Mészàros, private communications).

We describe our continuum normalization scheme and the upgrades to StarNet in Section 2, including a new deep ensembling method that provides estimates of uncertainties in the stellar labels. In Section 3, a description is provided for the data preparation and augmentation of the synthetic grids for training StarNet, which is then used to assess the synthetic gaps. In Section 4, we address the sources of biases in our methods, and provide a validation of StarNet's predicted uncertainties. In Section 5, FLAMES-UVES spectra from the Gaia-ESO Survey provide a test for StarNet's performance on observational spectra when trained on the INTRIGOSS grid. In Section 6, we discuss the results of StarNet trained on the other synthetic grids, extending our analysis to larger wavelength and parameter ranges, and the utility in and caveats with training a neural network on synthetic spectra. We end with concluding remarks in Section 7.

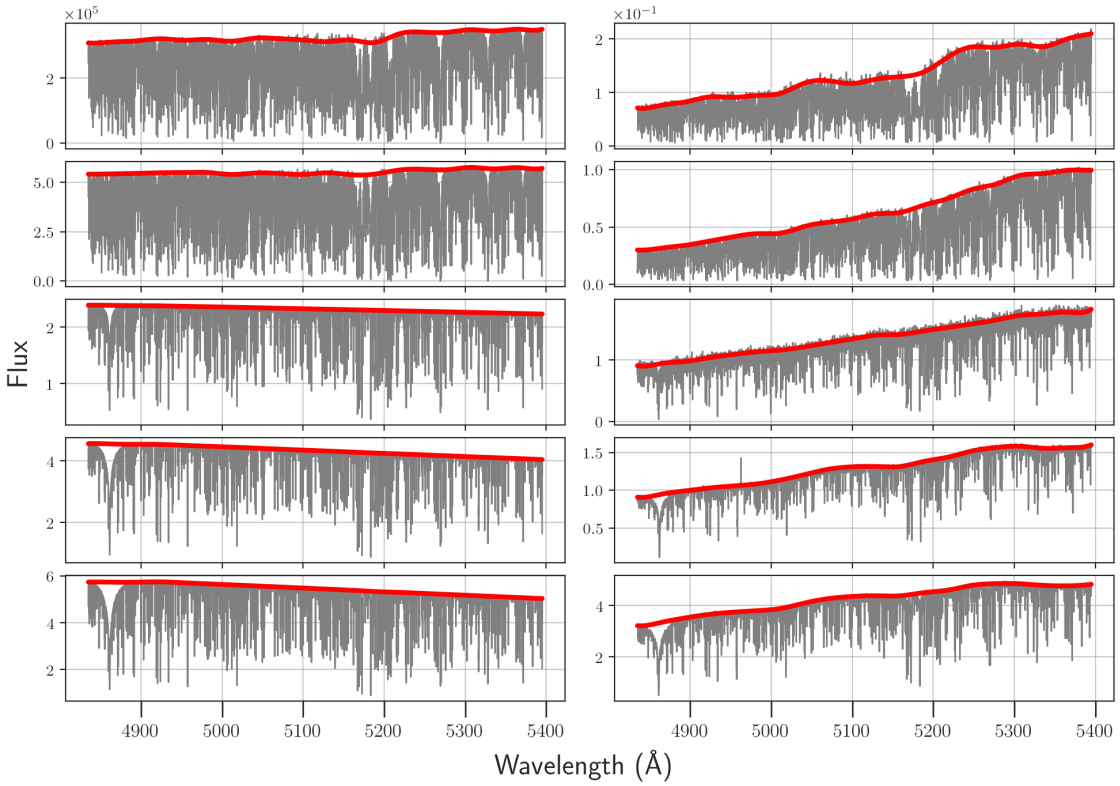

Figure 2.1: The results of our continuum fitting procedure for a sample of FLAMES-UVES spectra (right column) and closest matching INTRIGOSS spectra (left column). The red line indicates the estimated continuum. The complex, somewhat cyclical shape of the FLAMES-UVES spectra eludes simple fits of polynomials.

2.3 Methods

2.3.1 Analysis with neural networks

Only a brief description of neural networks is provided here in order to establish the terminology used throughout this paper. For a more complete description of StarNet and our machine learning methodology used, see Fabbro et al. (2018).

Fundamentally, a neural network (NN) is a function which transforms an input to a desired output. The function is composed of many parameters, arranged in layers, which form a highly non-linear combination of the input features, allowing for complex mappings to be represented accurately. StarNet is a *convolutional* NN, in which a series of learned filters, followed by a series of learned inter-connected nodes, transform a stellar spectrum to a prediction of associated stellar parameters.

To ensure the NN does not over- or under-fit the data, typically the full data set is split into a training, validation, and test set. The training set is used to directly influence the parameters of the NN, and the validation set is used to periodically check the performance of the NN on a separate data set. Both of these sets are utilized during the training of the NN, in which data is iteratively sent through the NN, the parameters of the NN are nudged in a direction which minimizes the output of the *loss function* (for regression problems, the loss is typically the residual between the prediction and expected output), and in this study, the training is stopped when performance on the validation set ceases to improve. Since both the training and validation sets influence the final trained NN, the test set is used to quantify the final performance for an independent data set.

For a training set of 90,000 spectra, each with ~40,000 flux values, the training time for StarNet rarely exceeds three hours using a single Tesla V100 GPU. With a final trained model, predictions for a set of thousands of spectra can take a matter of seconds.

2.3.2 Modifications to StarNet

Uncertainty Predictions

To derive predictive uncertainties we have adapted the method of *deep ensembling*, in which an ensemble of StarNet NNs with different initialization are trained, as outlined in Lakshminarayanan et al. (2017). Each NN predicts the mean and variance which, after averaging, is associated to the predictive uncertainty of each stellar parameter. This simple scheme has been shown to have good coverage in a variety of applications (Ovadia et al.,

2019) and it is easy to implement, as only two modifications to an existing NN are required:

1. Instead of the mean squared error being used as a loss function, a proper scoring rule which includes the variance, σ_θ^2, is used. In this case, the negative log-likelihood criterion is minimized:
$$-\log p_\theta(y_n|\mathbf{x}_n) = \frac{\log \sigma_\theta^2(\mathbf{x})}{2} + \frac{(y - \mu_\theta(\mathbf{x}))^2}{2\sigma_\theta^2(\mathbf{x})} \tag{2.1}$$
where $\mathbf{x}$ and $\mathbf{y}$ are respectively the inputs and targets, and $\mu_\theta(\mathbf{x})$ is the predicted mean (note that this is the mean of one model's prediction, since we are treating the target values as samples from a Gaussian distribution)

2. The last layer of the NN is changed such that – in addition to its regular linear output, $\mu_\theta(\mathbf{x})$, needed for a regression problem – it outputs another linear value, $\sigma_\theta(x)$, needed for determining the variance of its predictions.

Once the ensemble of NNs is trained, the final prediction, $\mu_*(\mathbf{x})$, and final variance, $\sigma_*^2(x)$, can be obtained by combining the outputs from each model as you would for a mixture of uniformly-weighted Gaussian distributions. Explicitly, $\mu_*(\mathbf{x})$ is given by the average of the predicted means of each NN, and the final variance is determined via the following equation:

$$\sigma_*^2(x) = M^{-1} \sum_{m=1}^{M} (\sigma_{\theta_m}^2 + (\mu_{\theta_m}^2(x) - \mu_*^2(x)) \tag{2.2}$$

where M is the number of NNs used in the ensemble, typically 5-10. In this study, 7 NNs were used.

The method of deep ensembling is a powerful upgrade to the StarNet architecture for its ability to quantify how closely the spectra in a test set resemble the spectra used to train the model. The uncertainty not only covers the finite sample training size, but also some of the out-of-distribution uncertainties, and the ensembling of models captures some of the NN model uncertainty by averaging over several models. In contrast with the Monte-Carlo dropout method for uncertainty predictions, it does not perturb the network architecture as much (Ovadia et al., 2019). Furthermore, since each model can be trained in parallel, an ensemble of networks takes no longer to train than one model.

2.3.3 Augmenting and pre-processing the data

Synthetic and observed spectra typically have vastly different shapes due to instrumental effects and other signatures that uniquely affect the observed spectra. Special care is required to ensure both sets of spectra are standardized to minimize this *synthetic gap*. There are several steps involved in this process, including both pre-processing the spectra (matching the resolution of the spectra, re-sampling the spectra to a common wavelength grid, and removing the continuum) and augmenting the spectra (adding noise, effects of rotational and radial velocity, and zeroing flux values to mimic bad pixels). Augmenting data is a popular method used in machine learning experiments, serving a dual purpose of increasing both the robustness of the NN to variations existing in reality (which are not necessarily represented in a vanilla training set) and the *size of a training dataset*: spectral grids usually contain several thousand templates, however typically more data is required for training a deep NN that can make accurate predictions.

With all of this in mind, the synthetic spectra used for training StarNet were adapted for application to VLT/UVES spectra, by having the following modifications applied (in order):

1. *Resolution matching*: spectra were convolved to a resolution of R ~ 47,000, the resolution of the UVES spectra
2. *Rotational velocity*: randomly chosen with the constraint $0 < v_{rot} < 70$ km/s
3. *Radial velocity*: randomly chosen with the constraint $|v_{rad}| < 200$km/s
4. *Sampling matching*: the wavelength grid was re-sampled onto the UVES wavelength grid
5. *Noise*: Gaussian (white) noise with a standard deviation, σ, randomly chosen under the constraint $\sigma < 7\%$ median flux value, corresponding to S/N > 14. Note: a more accurate noise model would likely improve results, but white noise was found to be sufficient for this study.
6. *Continuum removal*: using the method described in Section 2.3.3
7. *Zeroing flux values*: a maximum of 10% of a synthetic spectrum is randomly given a flux value of zero
8. *Masking tellurics*: all telluric lines[1] are given a value of zero.

[1]Telluric lines from the Keck-MAKEE pipeline, available online at https://tinyurl.com/y4f5flpx

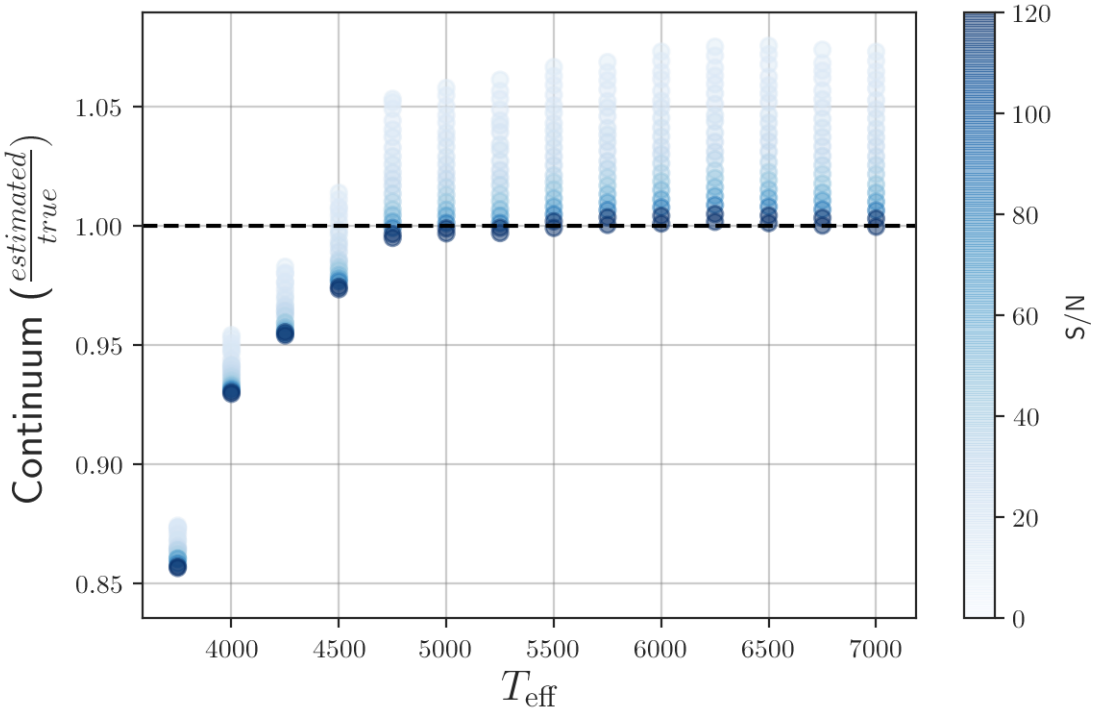

Figure 2.2: The systematic bias in the asymmetric sigma clipping method for the continuum estimation. Each INTRIGOSS spectrum was modified by varying the Gaussian noise, estimating the continuum, and averaging the offset from the true continuum. The median offsets shown here for all INTRIGOSS spectra were derived in bins of noise and temperature. At the lowest temperatures, most of the spectrum lies below the true continuum due strong absorption features.

All of the modifications up to and including the continuum removal [(i)-(vi) above] were pre-computed in parallel before training. The last two items were applied to the spectra during training.

Continuum removal

Special attention is required for good estimates of the stellar continuum in a spectroscopic analysis. Any method used for estimating the continuum should be invariant to both the shape and the signal-to-noise (S/N) of the spectrum to prevent the introduction of noise-dependent biases into the parameter estimations.

Several methods involve polynomial fits, with some groups selecting high order polynomial fits to the entire spectrum, and others fitting a lower order polynomial to a set of identified 'continuum pixels' (Casey et al., 2016b). Other popular methods involve splitting

the spectrum into short segments of equal length and estimating the continuum of each segment (e.g., García Pérez et al., 2016; Ness et al., 2015b). The segment methods perform well in cases where the spectral shape varies significantly over the wavelength range, possibly due to different detectors.

In this paper, a method based on segmenting the spectra was adopted: with each segment of 10 Angstroms, the known strong absorption features are masked, then iteratively the median is found and flux points are rejected above and below when discrepant by 2 and 0.5 standard deviations, respectively, until convergence is achieved. This 'asymmetric sigma clipping' more aggressively rejects absorption features in order to find the true continuum. Once the continuum has been estimated in each segment, a cubic spline is fit to the segments. Figure 2.1 shows the ability of this method to fit both the complex shape of VLT/UVES spectra and the synthetic INTRIGOSS spectra.

A known caveat with the asymmetric sigma clipping method is its noise dependent bias: as the noise levels increase in a spectrum, the found continuum is pushed further towards the 'noise ceiling', and thus the estimated continuum is above the true continuum. Figure 2.2 shows this bias as a function of temperature. It can be seen that in all cases the estimated continuum for a set of synthetic spectra, where the true continuum is known a priori, is higher for a noisy spectrum. Also shown is the trend of spectra with lower temperatures to have a continuum estimate well below the true continuum. This is expected since the majority of a low temperature spectrum lies below the continuum (due to extensive line blanketing), but this is not a problem here since this trend exists in both the synthetic and observed spectra.

Section 2.5.1 shows how this noise-dependent bias is minimized by simply adding noise at training time, forcing the network to learn the bias correction.

Other continuum estimation techniques were experimented with, e.g. Gaussian smoothing normalization (Ho et al., 2017), but they were found to affect the synthetic spectra differently than the observed spectra and led to more discrepant results.

2.4 Synthetic Spectral Grids

There are numerous grids of synthetic spectra available online ((for a summary, see Martins & Coelho, 2017), each differing in their spectral parameter and wavelength samplings, and generated from different radiative transfer codes, atomic and molecular line lists, model stellar atmospheres, and comparisons or corrections to observed spectra. These differences have significant impacts on the synthetic spectra, making comparisons between

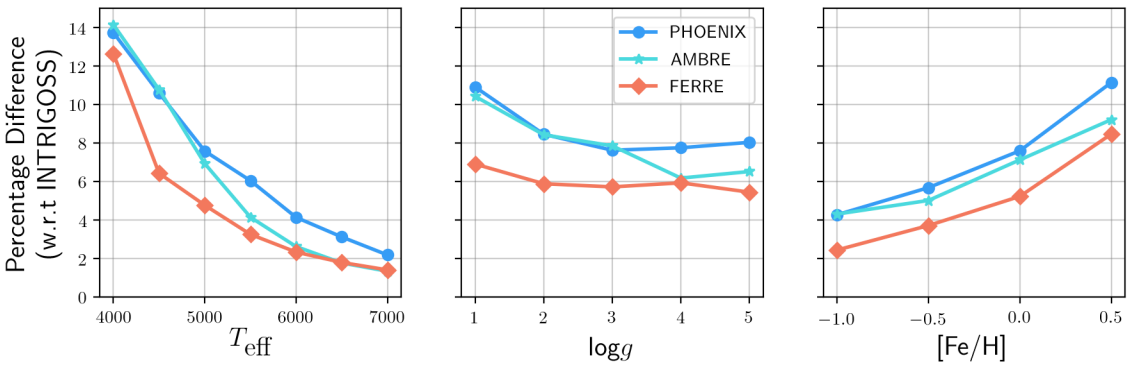

Figure 2.3: The differences in synthetic spectra when compared to INTRIGOSS, as a function of the three main stellar parameters. For each INTRIGOSS spectrum, spectra with matching parameters from the PHOENIX, AMBRE, and FERRE grids were collected, and the percentage difference between the spectra was calculated. Finally, the average difference across all matched spectra in bins of temperature, surface gravity, and metallicity were determined.

Table 2.1: The parameter space covered by and sampling of the synthetic spectra grids used in this study.

	$T_{\rm eff}$ (K)			log$g$ (dex)			[Fe/H] (dex)			[α/M] (dex)			$v_{\rm micro}$ (km/s)		
	Min.	Max.	Step	Min.	Max.	Step	Min.	Max.	Step	Min.	Max.	Step	Min.	Max.	Step
INTRIGOSS	3750	7000	250	0.5	5.0	0.5	-1.0	0.5	0.25	-0.25	0.5	0.25	1	2	1
FERRE	3500	6000	500	0	5.0	1	-5.0	0.5	0.5	–	–	–	1.5	1.5	–
	5500	8000	500	1.0	5.0	1	-5.0	0.5	0.5	–	–	–	1.5	1.5	–
AMBRE	2500	8000	250	-0.5	5.5	0.5	-5.0	1.0	0.25	-0.4	0.4	0.2	1	2	1
PHOENIX	2300	7000	100	0	6.0	0.5	-4.0 -2.0	-2.0 1.0	1.0 0.5	-0.2	1.2	0.2	0	4	f($T_{\rm eff}$)
	7000	15000	200	0	6.0	0.5	-4.0 -2.0	-2.0 1.0	1.0 0.5	–	–	–	0	4	f($T_{\rm eff}$)

grids inconsistent. With each new grid produced, the quality of the synthetic spectra increases by focusing on the atomic data in the line lists (e.g., see Kurucz, 2011), which already include information for many millions of spectral features. To train a machine learning algorithm, it is necessary to carefully consider which grid of synthetic spectra is best to use in a particular spectroscopic analysis.

2.4.1 The synthetic grids used in this study

The synthetic spectra used in this analysis include the high spectral resolution grids INTRIGOSS, AMBRE, FERRE, and PHOENIX. When StarNet is trained and tested on these grids, they are pre-processed and augmented according to Section 2.3.3, unless otherwise noted.

The parameter space covered by the grids is summarized in Table 2.1, and a brief description of each grid follows:

1. INTRIGOSS: created by Franchini et al. (2018), this grid is a set of high resolution synthetic spectra specifically created for the analysis of F, G, and K type stars in the Gaia-ESO survey. The synthetic spectra were tuned by direct comparison to Gaia-ESO spectra, and in some cases the line list was modified to better match absorption features in the observed spectra without identifying which atom or molecule was the source of the feature. The INTRIGOSS spectra allow the stellar parameters $T_{\rm eff}$, log$g$, [Fe/H], [α/M], and $v_{\rm micro}$ to vary within relatively small ranges (see Table 2.1) and span the wavelength range 483-540 nm only. Although this wavelength range is only a subset of the entire wavelength range of the UVES spectra (480-680 nm, in three settings), it contains important features such as Hβ, the Mgb lines, and numerous metal lines.

2. FERRE: this newer grid represents a huge wavelength (120-6500 nm) and parameter range (3500 $\geq T_{\rm eff} \geq$ 30,000 K, 0 $\geq \log g \geq$ 5, -5 $\geq$ [Fe/H] $\geq$ 1), using the newest sources of atomic and molecular data from the literature to model B to early-M type stars at varying resolutions (R $\sim$ 10,000, 100,000, 300,000). Although not specifically tuned to spectra from any particular survey, the spectra do reproduce the main absorption features when compared to HST UV-optical and APOGEE IR spectra (Allende Prieto et al., 2018). FERRE appears to be the largest general purpose grid of synthetic spectra created to date, though the FERRE authors caution that the grid is, in some ways, already outdated. The full FERRE grid is split into 5 sub-grids with increasing ranges of temperature, and only the first two are used in this study (see Table 2.1).

3. AMBRE: a high resolution (R > 150,000) grid of optical spectra (300-1200 nm) modeling F, G, K, and M type stars, with 4 stellar parameters over a relatively large extent (2500 $\geq T_{\rm eff} \geq$ 8000 K, -0.5 $\geq \log g \geq$ 5.5, -5 $\geq$ [M/H] $\geq$ 1, -0.4 $\geq$ [α/M] $\geq$ 0.4). Although it was created several years ago (de Laverny et al., 2012), and thus uses outdated atomic data, it has been used recently, for example, in accurately predicting stellar parameters for Gaia-ESO UVES spectra (Worley et al., 2016).

4. PHOENIX: this grid was created as a resource for very high resolution (R > 100,000) stellar spectra spanning ultra-violet to infrared wavelengths (50-5000 nm); Husser et al. (2013) use it to analyse MUSE integral field spectra of stars in the metal-poor globular cluster NGC 6397. It spans a large parameter space (2300 $\geq T_{\rm eff} \geq$ 12,000 K, 0 $\geq \log g \geq$ 6, -4 $\geq$ [M/H] $\geq$ 1, -0.2 $\geq$ [α/M] $\geq$ 1.2). It has also been used recently for machine learning applications, e.g., of LAMOST data (Wang et al., 2019).

Since the INTRIGOSS grid was created specifically for the Gaia-ESO survey and includes a carefully crafted line list and comparisons to both UVES spectra and other synthetic grids, it was chosen as the baseline for our exploration of the impact of the various synthetic grids, and as the primary grid for our analyses of the FLAMES-UVES spectra.

2.4.2 Comparisons of synthetic grids

To perform a comparison of the synthetic spectral grids, INTRIGOSS was chosen as the baseline. For each INTRIGOSS spectrum, spectra with matching stellar parameters from each grid were selected (if none were found, the INTRIGOSS spectrum was skipped), and the residual of the flux values of each spectrum with respect to the INTRIGOSS spectrum was calculated and converted to a percentage difference. The average percentage difference

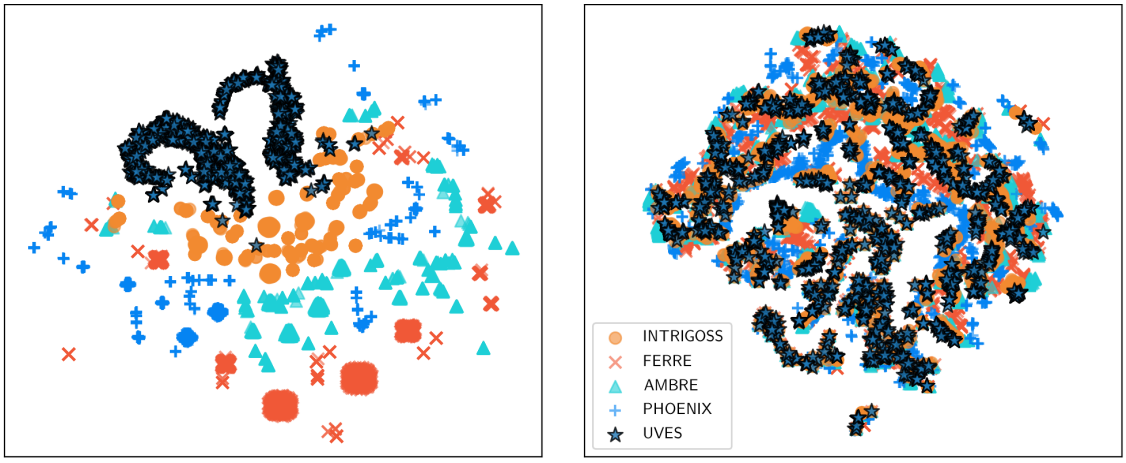

Figure 2.4: t-SNE plots to visualize any synthetic gaps between the four synthetic spectral grids used in this analysis (INTRIGOSS, FERRE, PHOENIX, and AMBRE) and the observed Gaia-ESO UVES spectra. Left panel is the raw, non-augmented synthetic data; right panel shows augmented synthetic spectra. For each UVES spectrum, the synthetic spectrum from each grid with closest matching parameters to the associated GES iDR4 values was collected. Clearly there is significant overlap, with one another and especially with the UVES spectra, when the synthetic spectra are augmented.

was then determined in bins of temperature, surface gravity, and metallicity. As shown in Figure 2.3, the differences in the spectra are more pronounced at lower temperatures and higher metallicities, i.e., in the grid regions that would be the most sensitive to line blanketing. The FERRE spectra are the most closely matched to the INTRIGOSS spectra, over the widest range in stellar parameters, whereas the PHOENIX spectra are the most dissimilar.

To qualitatively assess how closely the synthetic spectral grids match the Gaia-ESO FLAMES-UVES spectra (discussed further in Section 2.6), a t-SNE² test was carried out to compare the closest matching spectra from each grid to each UVES spectrum. As seen in Figure 2.4, there is a distinct difference between the raw observed and synthetic spectra; the *synthetic gap*. However, when the data is augmented as described in Section 2.3.3 then the synthetic gap is significantly narrowed: the augmented synthetic spectra occupy the same compressed low-dimensional space as the observed UVES spectra.

2.5 Training StarNet with INTRIGOSS

For our first application, StarNet has been trained using the augmented INTRIGOSS spectra, and is referred to as `"StarNet-INTRIGOSS"`. The grid of 7,616 INTRIGOSS spectra were split into a *reference set* (6,093 spectra) and a *test set* (1,523 spectra), an 80/20 split. These two datasets were then pre-processed and augmented (as described in Section 2.3.3) to create datasets several times their size: the 6,093 reference spectra were turned into an *augmented reference set* of 100,000 spectra (no further improvements were seen with a larger training sample) and the 1,523 test spectra were turned into an *augmented test set* of 10,000 spectra.

The augmented reference set was then split into a training set (90,000 spectra) and a validation set (10,000 spectra), a 90/10 split. These steps help to mitigate over-fitting during training (further discussed below).

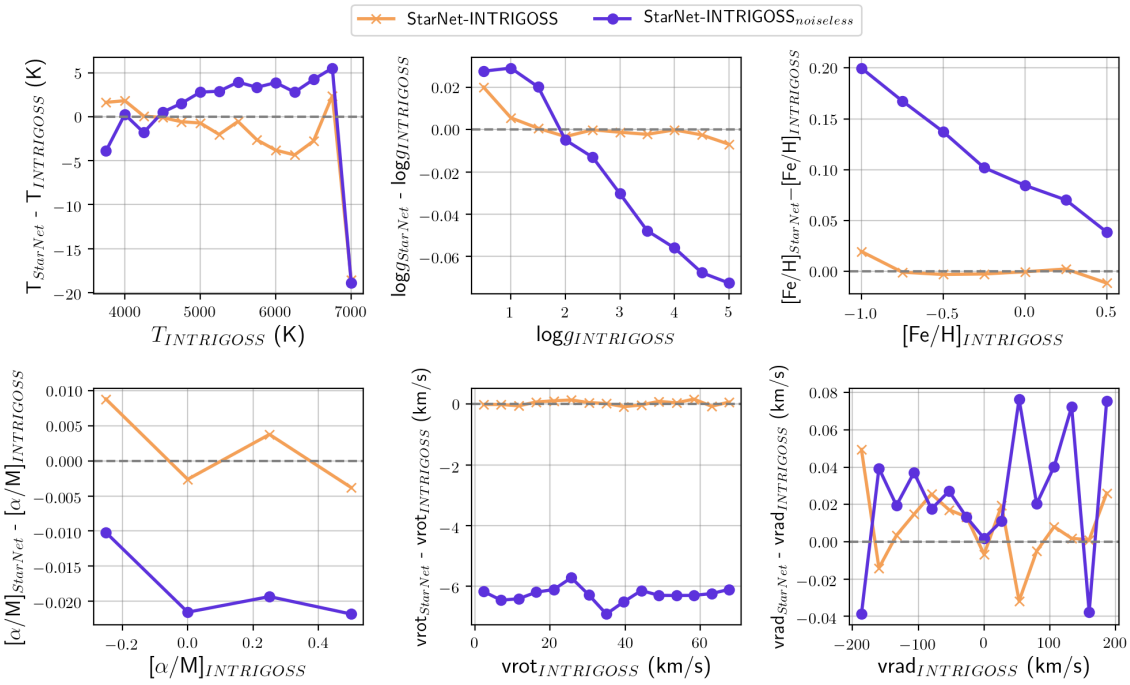

Figure 2.5: Residual plots to show noise-dependent biases from the asymmetric sigma clipping continuum removal in the stellar parameter estimations. Two versions of StarNet were trained: one model, `StarNet-INTRIGOSS` (orange), was trained on 90,000 INTRIGOSS spectra augmented as outlined in Section 2.3.3, and the other, `StarNet-INTRIGOSS`$_{\text{noiseless}}$ (purple), was trained identically except without the addition of noise to the synthetic spectra prior to continuum removal. Each was tested on 10,000 noisy INTRIGOSS spectra, the median residual at each grid point was calculated, and the results for all spectra with S/N < 80 are shown here. The discrepancies are the most pronounced at lower metallicities, higher surface gravities, and across all rotational velocities.

2.5.1 Addressing method-dependent biases: testing with INTRIGOSS spectra

The performance of `StarNet-INTRIGOSS` is assessed here using the INTRIGOSS synthetic spectra themselves to first explore the limitations and systematic biases inherent in the method. This is because we know the spectral properties (stellar parameters and continuum) a priori, and we can investigate and mitigate errors or degeneracies before predicting on real spectra. In addition, we want to ensure StarNet does not over-fit to the training data, which would result in both poor interpolation between the synthetic grid points and poor predictions of observed spectra. Both of these issues are discussed below.

Noise-dependent biases in continuum fitting

As discussed in Section 2.3.3, the asymmetric sigma-clipping continuum removal method has a known noise-dependent bias. Figure 2.2 illustrates this, where the estimated continuum for low S/N spectra can be discrepant by several percent above the true continuum (with an exception at lower temperatures where the stronger absorption features cause much of the spectrum to lie below the continuum). If the estimated continuum is significantly higher than the true continuum, the resulting continuum-normalized spectra will contain artificially lowered flux values. This would lead to deeper absorption features which could mimic a lower temperature or higher metallicity than the true value.

To assess the impact of continuum fitting due to noise, `StarNet-INTRIGOSS` was trained with noiseless synthetic spectra and with Gaussian noise added (augmentation step (v) in Section 2.3.3). Both of these trained models were tested on a set of 10,000 augmented (noisy) INTRIGOSS spectra, and the predictions for both models on all spectra with S/N < 100 are shown in Figure 2.5. As expected, there are clear biases for all stellar parameters when StarNet is trained on noiseless spectra, with more prominent discrepancies at low metallicities, high surface gravities, and across all rotational velocities. These biases are reduced when trained with noisy spectra; the network is capable of learning how to successfully manage the effects of noise during the training process.

By adding noise to the spectra before the continuum removal step in the pre-processing stage, the NN can compensate for noise-dependent bias. Although this bias dependence is smooth, and it can be corrected in other ways and in other methodologies, the NN

[2]T-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. It is often used to visualize high-level representations learned by a NN.

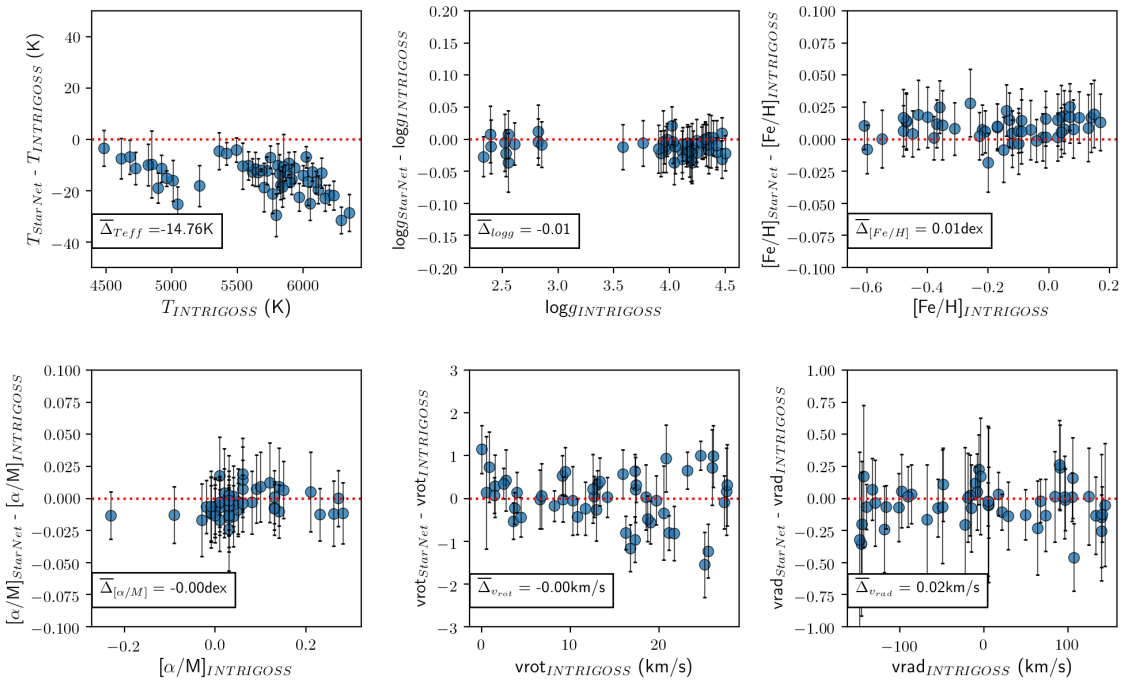

Figure 2.6: The residuals between truth values and predictions from `StarNet-INTRIGOSS` on the intra-mesh INTRIGOSS spectra. No significant biases or erroneous trends are found. The minor offsets in temperature are discussed in the text.

compensates for it automatically. Furthermore, the flexibility of the NN means that it has the potential to handle even more complex bias dependencies (e.g., persistence in some of the early APOGEE spectra; see Jahandar et al. 2017).

Testing for over and under-fitting with intra-grid synthetic spectra

Along with the published INTRIGOSS grid of spectra, a set of 50 spectra at intra-grid locations was provided by the INTRIGOSS team for testing the ability of a chosen methodology to interpolate between grid points. These intra-grid spectra also provide an excellent test set to confirm that the model for `StarNet-INTRIGOSS` did not over-fit to the training set nor result in other systematic biases in its predictions.

The predictions from `StarNet-INTRIGOSS` on the 50 intra-grid spectra are shown in Figure 2.6. The results are excellent, with no signs of under or over-fitting from the training set. The slight offset of temperature is unexpected, but we note that it is very small, ranging from 1-2 σ (the uncertainty propagated by the NN itself, ~10-20K). We also have no information on how the intra-grid spectra were selected and generated, and therefore do not consider this result to be significant. We also note that the intra-grid spectra do not extend

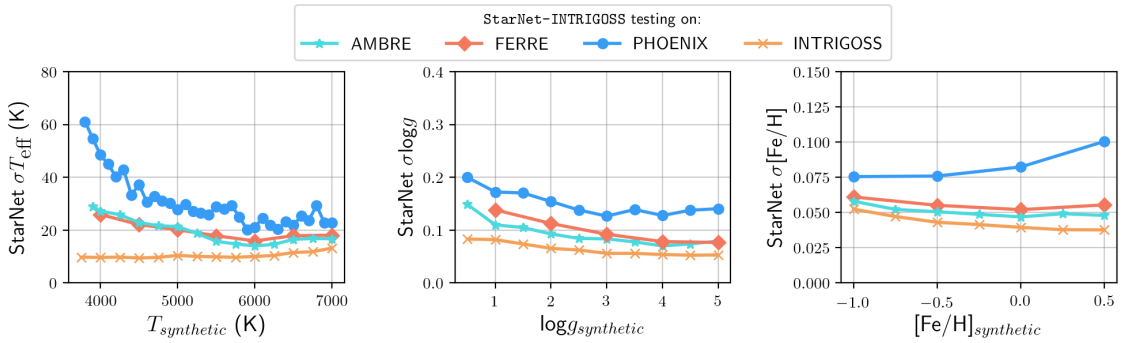

Figure 2.7: The uncertainties in the predictions of `StarNet-INTRIGOSS` for the three main stellar parameters. The test sets are augmented INTRIGOSS, AMBRE, FERRE, and PHOENIX spectra (limited to the INTRIGOSS parameter range), and the median uncertainty in bins of temperature, surface gravity, and metallicity, were calculated. In general, the uncertainties grow w.r.t INTRIGOSS based on how dissimilar the spectra are (see Figure 2.3 for these trends), especially pronounced at lower temperatures, lower surface gravities, and higher metallicities.

below $T_{\rm eff} = 4500$ K or $\log g < 2.4$, so we cannot confidently evaluate our stellar parameter predictions in those ranges.

Interestingly, the predictions for the radial velocity, $v_{\rm rad}$, are excellent and do not show significant bias, and have typical uncertainties below 0.5 km s^{-1}. This is somewhat surprising, given that convolutional NNs with pooling layers are built to be invariant to small translations.

2.5.2 Testing StarNet-INTRIGOSS with other synthetic spectral grids

To explore the accuracies and uncertainty estimates from the deep ensembling method, the predictions of `StarNet-INTRIGOSS` are compared between the INTRIGOSS, FERRE, AMBRE, and PHOENIX grids. These grids have been previously examined by Franchini et al. (2018) in their comparison of seven synthetic grids (see their Figure 7), and in our percentage difference analysis and t-SNE comparisons in Section 2.4.2 (Figs. 2.3 and 2.4). Both analyses show that FERRE is the most similar to INTRIGOSS, while PHOENIX is the least similar.

The validity of the deep ensembling method can be further verified by examining the predictions from within the parameter space used for training, and also beyond those boundaries. As a first test, `StarNet-INTRIGOSS` is used to predict stellar parameters

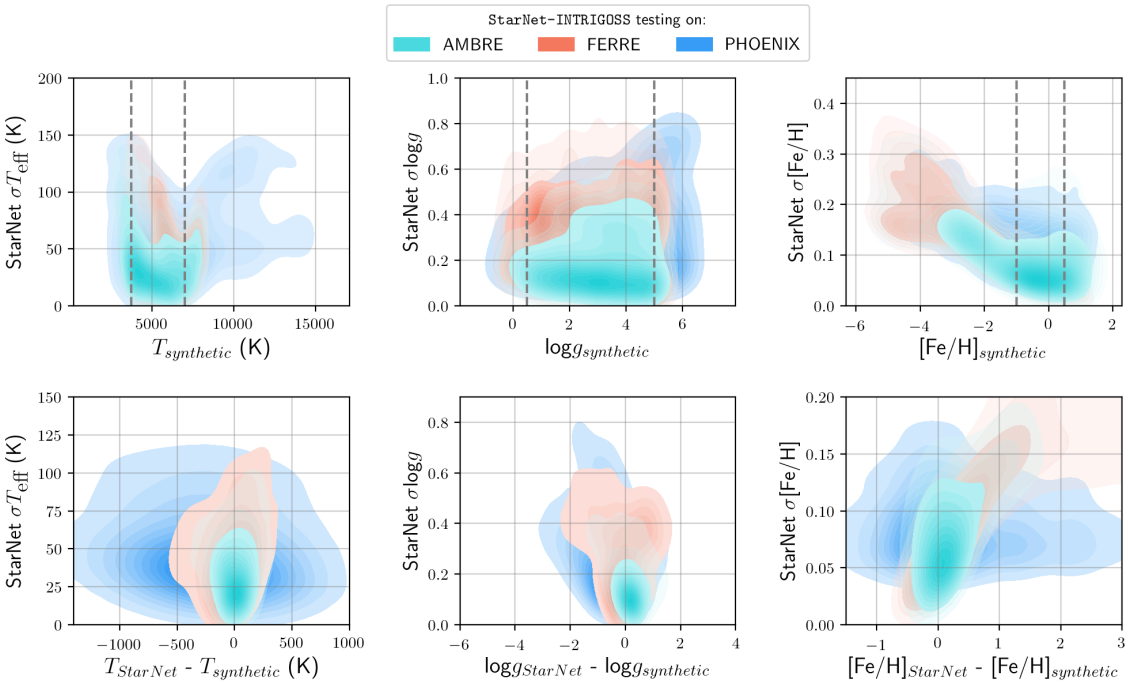

Figure 2.8: The uncertainties in the predictions of `StarNet-INTRIGOSS` for the three main stellar parameters. The test sets are augmented AMBRE, FERRE, and PHOENIX spectra (spanning their entire parameter ranges). The first row shows the uncertainties as a function of the specified parameter, whereas the second row shows the uncertainties as a function of the residual between `StarNet-INTRIGOSS` predictions and truth values of the specified parameter. The grey dashed lines correspond to the limits of the INTRIGOSS grid. As expected, the uncertainties grow both when StarNet predicts outside the ranges of the INTRIGOSS spectra it was trained on, and as the residuals increase.

for test sets of 3,000 augmented INTRIGOSS, AMBRE, FERRE, and PHOENIX spectra which span the *same parameter space*; the uncertainties are summarized in Figure 2.7. The uncertainties increase relative to the predictions from the INTRIGOSS spectra at lower temperatures, lower surface gravities, and higher metallicities, i.e., where the synthetic grids were previously shown to deviate the most (see Figure 2.3). Similarly, the uncertainties in the predictions from the PHOENIX grid are the largest, consistent with the known larger differences between the INTRIGOSS and PHOENIX spectra.

To test the uncertainties in the predictions in a parameter space beyond the training data set, `StarNet-INTRIGOSS` was applied to spectra from the *full* parameter ranges in the AMBRE, FERRE, and PHOENIX grids. Each extend to higher and lower temperatures, and much lower metallicities; the results are shown in Figure 2.8. As expected, the uncertainties tend to increase when predicting outside of the parameter ranges used for training, as well as when the predictions become more discrepant from their true values.

2.6 An application to Gaia-ESO FLAMES-UVES spectra

The Gaia-ESO public spectroscopic survey ((GES, Gilmore et al., 2012) is a large survey with the goal of exploring all components of the MW in a complementary way to Gaia. Along with the observed spectral database, an official Gaia-ESO Survey Internal Data Release (GES iDR) is available, containing stellar parameters derived as the weighted average of the results from a set of working groups (each using different methods). The fourth data release (GES iDR4) is used in this study as a comparison for our StarNet predictions (Pancino et al., 2017).

The GES is carried out using FLAMES at the VLT (Pasquini et al., 2002) to obtain high-quality medium-resolution Giraffe spectra for 10^5 stars and high-resolution UVES spectra for ~5000 stars. Currently, a dataset of 2308 FLAMES-UVES spectra is available, spanning field and cluster stars from the bulge, halo, thick disc and thin disc. The S/N distribution of these stars is shown in Figure 2.9, where the majority of the stars have S/N < 100.

In addition, the Gaia-ESO survey includes a set of 34 *benchmark spectra* of well-known bright stars (Blanco-Cuaresma et al., 2014), available online[3], to be used as a reference. Their parameters $T_{\rm eff}$ and log$g$ were determined independent of spectroscopy, using angular diameter measurements and bolometric fluxes (Heiter et al., 2015), and [Fe/H]

[3]ftp://obsftp.unige.ch/pub/sblancoc/Gaia$_B enchmark_S tars_L ibrary$/

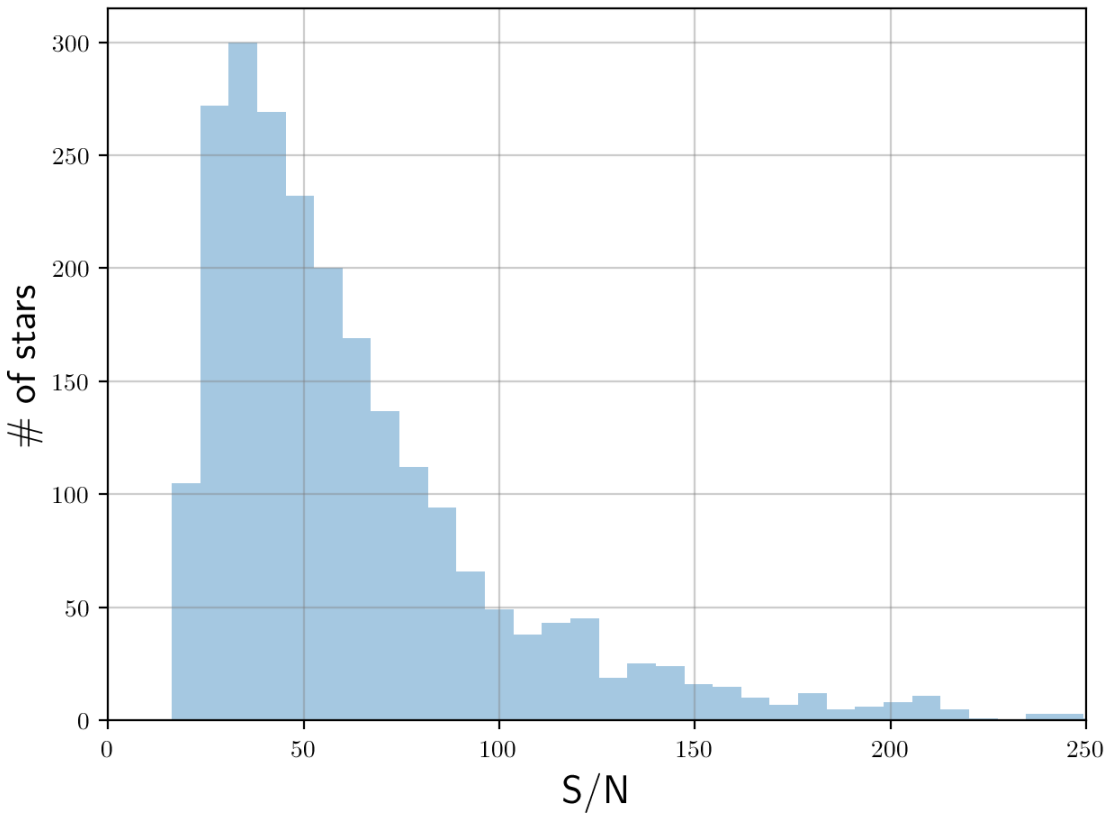

Figure 2.9: The S/N distribution of the Gaia-ESO FLAMES-UVES spectra.

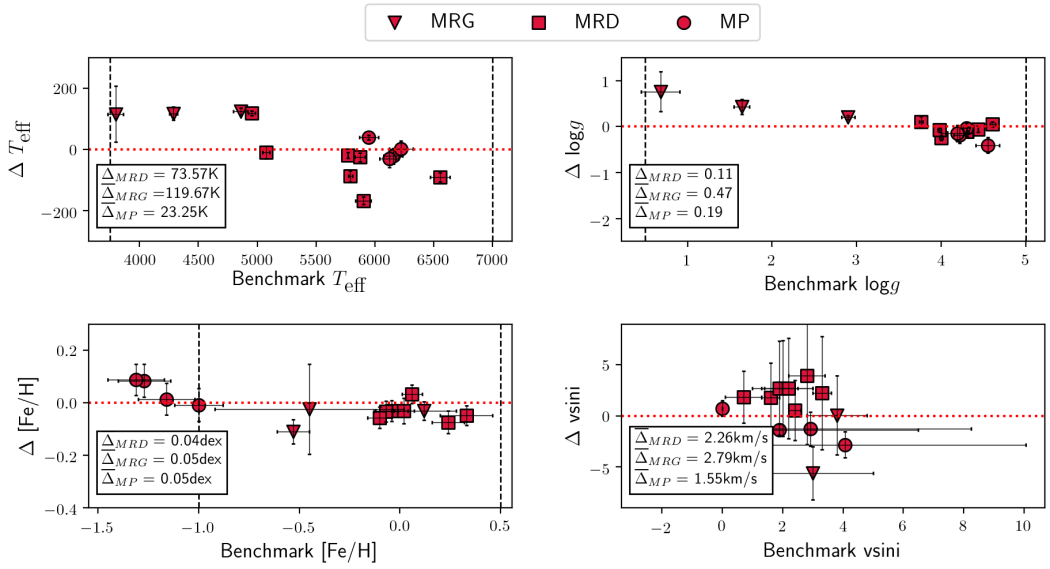

Figure 2.10: `StarNet-INTRIGOSS` was used to predict stellar parameters for the Gaia-ESO benchmark stars, and the residuals between predictions and published values are shown here. The stars were split into metal-poor (MP) stars, metal-rich giants (MRGs) and metal-rich dwarfs (MRDs), following the procedure in R. Smiljanic et al. (2014). The average quadratic difference, $\overline{\Delta}$, between StarNet's predictions and benchmark values is used to evaluate the accuracy of the predictions.

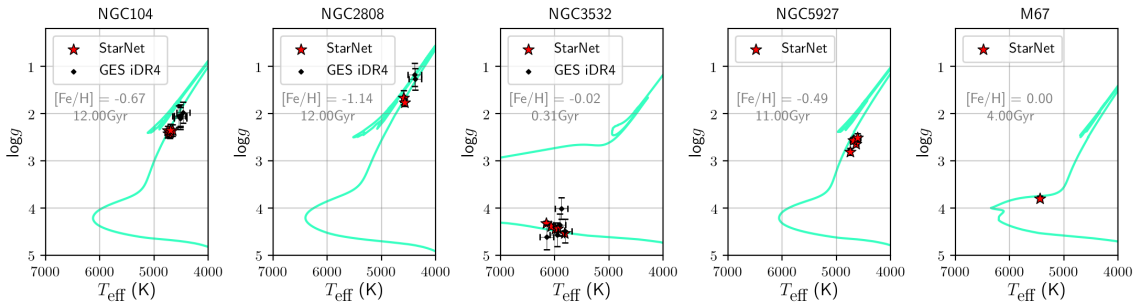

Figure 2.11: `StarNet-INTRIGOSS` predictions of log$g$ and $T_{\rm eff}$ compared with theoretical MIST isochrones with the ages and metallicities shown in light grey text. The cluster metallicities and ages were retrieved from the online updated catalog of Harris (2010) and the WEBDA database. Also plotted are the GES iDR4 stellar parameters for the same stars (except NGC5927 and M67 for which none could be found).

was determined from these values (Jofré et al., 2014).

The Gaia-ESO survey has also observed several calibration clusters, including the globular clusters M 15, NGC 104, NGC 1851, NGC 2808, NGC 4372, NGC 4833, NGC 5927, and NGC 6752, and the open clusters M 67, NGC 3532, and NGC 6705. Some of these clusters have metallicities much lower than the INTRIGOSS metallicity grid ([Fe/H] $\geq$ -1), so they were removed from this analysis. This leaves five clusters for testing `StarNet-INTRIGOSS`, including NGC 104, NGC 2808, NGC 3532, NGC 5927, and M 67.

As a first test, we will examine the abilities of `StarNet-INTRIGOSS` to predict stellar parameters for the GES benchmark stars. This will be followed by testing its predictions for stars in the calibration clusters. Finally, we test the predictions made on the entire sample of FLAMES-UVES spectra in the Gaia-ESO survey.

2.6.1 StarNet-INTRIGOSS predictions for the GES benchmark stars

Following the procedure in Smiljanic et al. (2014), the benchmark stars were separated into three groups in order to assess the accuracy in different regions of parameter space:

1. Metal-rich dwarf (MRD): [Fe/H] > -1.00 and log$g$ > 3.5
2. Metal-rich giant (MRG): [Fe/H] > -1.00 and log$g$ $\leq$ 3.5
3. Metal-poor (MP): [Fe/H] $\leq$ -1.00

Shown in Figure 2.10 are the results of `StarNet-INTRIGOSS` predictions on seven MRDs, three MRGs, and four MP stars from the set of benchmarks. The metric for

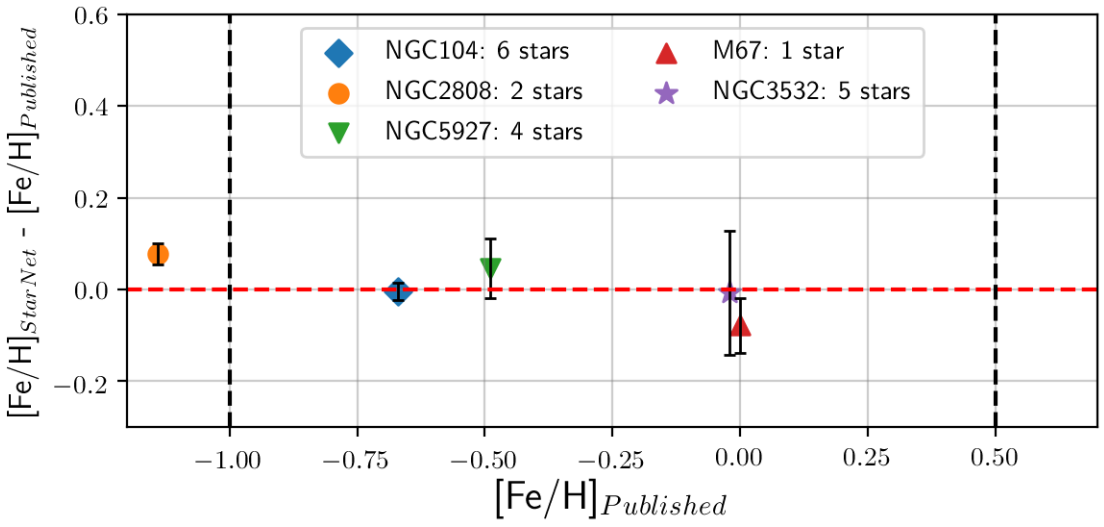

Figure 2.12: Average residuals of `StarNet-INTRIGOSS` metallicities for a sample of calibration clusters. The error bars indicate the standard deviation on the residual (except for M67, containing only one star, which shows the StarNet uncertainty). Literature values were retrieved from the online updated catalog of Harris (2010) and the WEBDA database. The vertical dashed lines correspond to the metallicity limits of the INTRIGOSS grid

evaluating performance, as in Smiljanic et al. (2014), is the average quadratic difference, $\overline{\Delta}$, between the predictions and benchmark values, and is small for all three groups of stars ($\overline{\Delta}_{T_{\rm eff}}$< 120 K, $\overline{\Delta}_{\log g}$< 0.47, and $\overline{\Delta}_{\rm [Fe/H]}$< 0.05). Additionally, there exists no significant deviation for any parameters, with the exception of larger results for both log$g$ and $T_{\rm eff}$ for the MRGs, and an increasing trend at lower [Fe/H] for the MP stars. We note that the MP stars lay outside the metallicity range of the spectra used for training, so this is not surprising. The benchmark uncertainties for $v$sini are so large that it is difficult to determine if `StarNet-INTRIGOSS` produced accurate predictions. It is also interesting that in most cases when the published uncertainties for the benchmark parameters are large, so too are the predicted uncertainties of our deep ensembling method.

Altogether the results obtained through tests on the Gaia benchmark stars provide a convincing validation that our method works well across the range of parameters *for high S/N spectra*. However, we caution that these comparisons are against a statistically small sample (the benchmark stars) and that systematic errors could potentially appear in larger samples.

2.6.2 StarNet-INTRIGOSS predictions for the GES calibration clusters

StarNet can predict stellar parameters for FLAMES-UVES spectra, but are those predictions physically realistic? A common method used to assess the fidelity of astrophysical parameters is to compare them to a theoretical understanding of stellar evolution.

The predictions of $T_{\rm eff}$ and log$g$ of the stars in each cluster from `StarNet-INTRIGOSS` were compared to the MESA Isochrones and Stellar Tracks (MIST, Choi et al. (2016)), generated by adopting the published metallicities and ages for each cluster from the Harris catalogue (Harris, 2010); see Fig. 2.11. While StarNet appears to predict both higher surface gravities and temperatures for giants than the GES iDR4 values, they remain physically consistent when compared to the isochrones, and are more constrained. It is important to keep in mind that the predictions from StarNet are uncalibrated and that StarNet recovers the stellar parameters $T_{\rm eff}$, log$g$, and [Fe/H] for both dwarfs and giants in a physically consistent manner.

As a further check to ensure physically consistent stellar parameters, the metallicity predictions of `StarNet-INTRIGOSS` were compared directly to the literature values for each cluster. Figure 2.12 shows the average StarNet [Fe/H] predictions for the stars in each calibrating cluster, with error bars derived from the standard deviation of the predictions. In

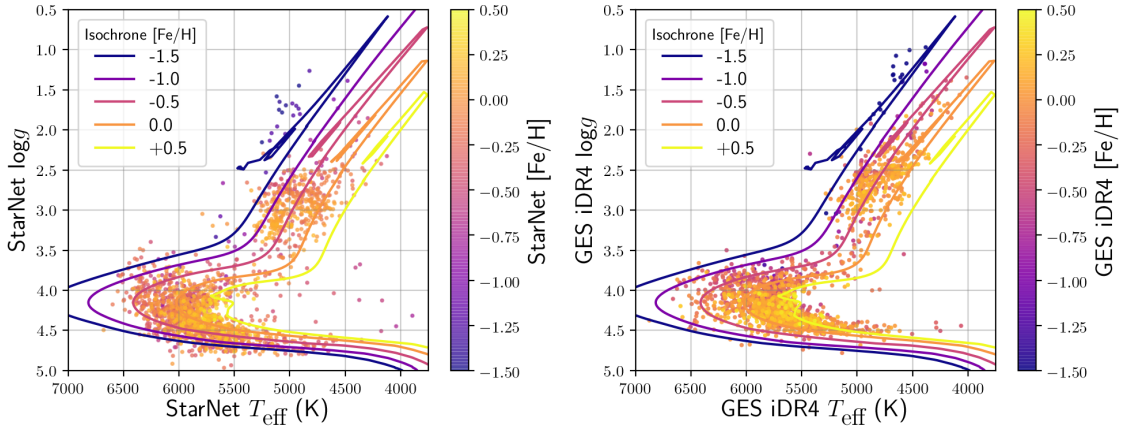

Figure 2.13: HR diagrams showing the physical consistency of `StarNet-INTRIGOSS` predictions for $T_{\rm eff}$, log$g$, and [Fe/H] on the test set of FLAMES-UVES spectra. Overlaying the predictions are MIST isochrones with an age of 8 Gyr and the metallicities shown. The figure on the left shows the predictions of `StarNet-INTRIGOSS` and the figure on the right shows the published GES iDR4 values.

the parameter space that INTRIGOSS was trained on (shown by the vertical dashed lines), the metallicity predictions show excellent agreement with cluster values, and even NGC 2808, which is just outside the trained parameter range, is well predicted.

2.6.3 StarNet-INTRIGOSS predictions for the entire Gaia-ESO Survey (GES iDR4)

The entire sample of FLAMES-UVES spectra was examined with `StarNet-INTRIGOSS`, with a few cuts made to produce the final sample for predictions: stars were removed if they had `NaN` values for any parameter in the GES iDR4 catalog and if the uncertainties produced by StarNet for any parameter were abnormally large ($\sigma T_{\rm eff}$ > 65K, σ[Fe/H] > 0.50, σlog$g$ > 0.80, $\sigma v_{\rm rot}$ > 3km/s, $\sigma v_{\rm rad}$ > 5km/s), decreasing the sample size from 2308 to 2200. The $T_{\rm eff}$−logg plots for the final sample are shown in Figure 2.13, where the left panel shows the predictions made by `StarNet-INTRIGOSS`, and the right panel shows the GES iDR4 catalog values. MIST isochrones for age = 8 Gyr and varying metallicities are overlaid for clarity.

In general `StarNet-INTRIGOSS` finds slightly larger values for both $T_{\rm eff}$ and log$g$, especially for giants, as seen for the benchmarks in Figure 2.10. However, we note that our results are from only a narrow window of the spectrum (483-540 nm), whereas the full GES

Table 2.2: StarNet was separately trained on sets of 90,000 augmented spectra from the INTRIGOSS, FERRE, AMBRE, and PHOENIX grids. The results of each trained model when predicting on the Gaia-ESO benchmark stars are shown here.

	MRDs			MRGs			MPs		
	$\overline{\Delta}T_{\rm eff}$ (K)	$\overline{\Delta}\log g$	$\overline{\Delta}$[Fe/H]	$\overline{\Delta}T_{\rm eff}$ (K)	$\overline{\Delta}\log g$	$\overline{\Delta}$[Fe/H]	$\overline{\Delta}T_{\rm eff}$ (K)	$\overline{\Delta}\log g$	$\overline{\Delta}$[Fe/H]
`StarNet-INTRIGOSS`	74	0.11	0.04	120	0.47	0.05	23	0.19	0.05
`StarNet-FERRE`	77	0.19	0.22	64	0.12	0.29	63	0.17	0.19
`StarNet-AMBRE`	125	0.12	0.31	100	0.17	0.36	105	0.30	0.06
`StarNet-PHOENIX`	152	0.34	0.39	184	0.25	0.36	79	0.11	0.17

iDR4 analyses are from the full UVES spectral region (480-680 nm). We also note that the working groups who contributed to the GES iDR4 were using a different set of synthetic spectra than INTRIGOSS.

Final predictions for the stellar parameters $T_{\rm eff}$, $\log g$, [Fe/H], and $v_{\rm rad}$ for the FLAMES-UVES spectra are compared with the GES iDR4 values in Figure 2.14. The difference in the predictions for $T_{\rm eff}$ and $\log g$ both show a slight decrease with increasing values when compared with the GES iDR4 results, and is likely a result of the INTRIGOSS spectra themselves (there is no negative slope when StarNet is trained on FERRE or AMBRE spectra, see Section 2.7.1). The predictions for [Fe/H] and $v_{\rm rad}$ are in excellent agreement. To the best of our knowledge, this is the first time that a NN has been able to accurately predict radial velocities on real spectra.

Finally, the uncertainties on the stellar parameter predictions from `StarNet-INTRIGOSS` for the full sample of FLAMES-UVES spectra are shown in Figure 2.15. In the $T_{\rm eff}$ and $\log g$ plots there are two distinct populations of stars corresponding to dwarfs and giants, where main sequence stars tend to dominate the sample, and where the uncertainties are moderately larger for the giants. The [Fe/H] and [α/Fe] uncertainties are very small, suggesting that the INTRIGOSS spectra model the absorption features quite well. Finally, the $v_{\rm rad}$ uncertainties are comparable to the typical error in the GES iDR4 (0.4 km s^{-1}).

2.7 Discussion

2.7.1 Exploring StarNet trained on other synthetic grids

In Fig. 2.8, the differences between the four synthetic grids in this paper were compared according to the predictive `StarNet-INTRIGOSS` values and uncertainties. We showed that the differences between AMBRE and FERRE, with respect to INTRIGOSS, were

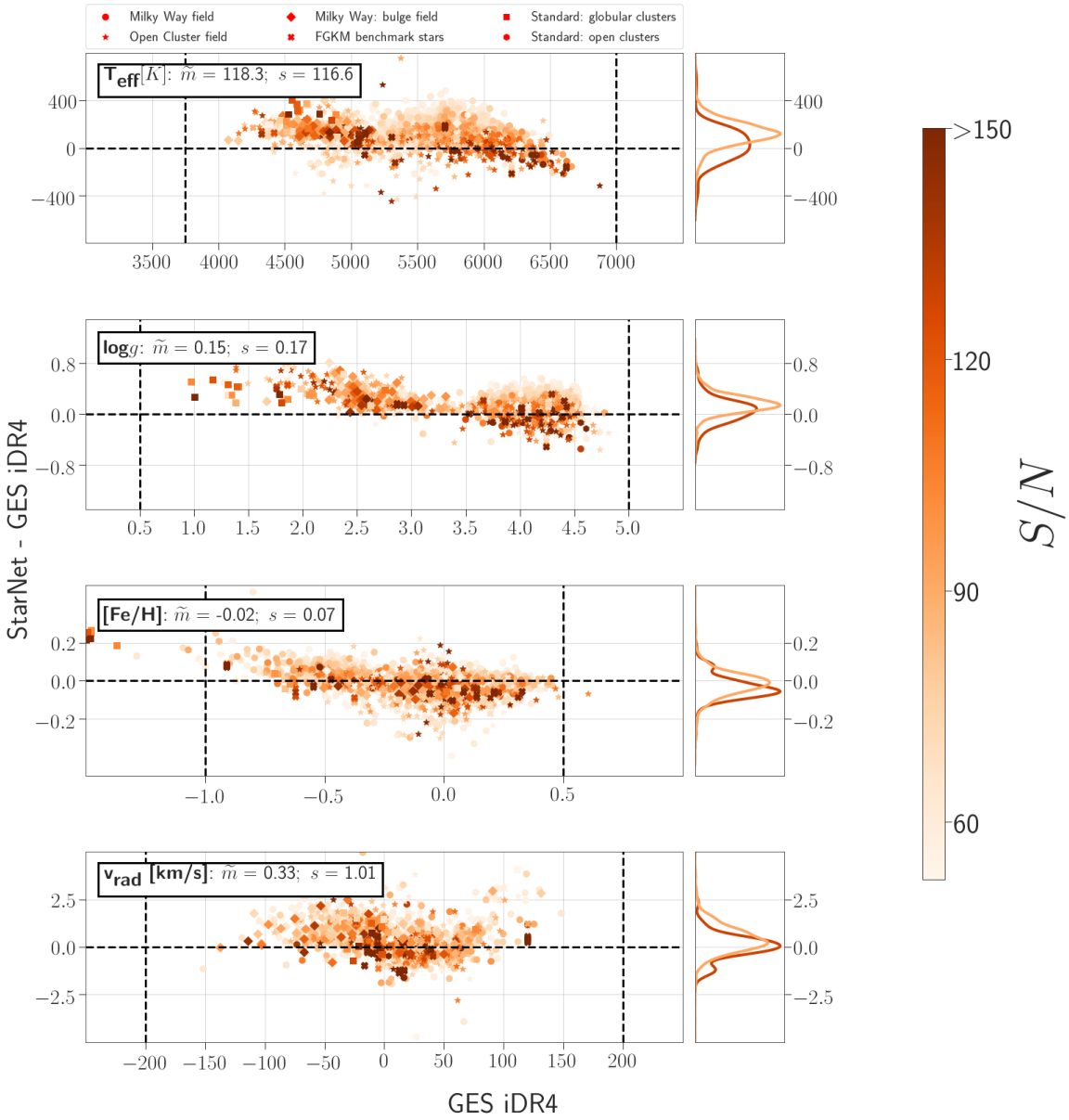

Figure 2.14: StarNet was trained on 100,000 augmented INTRIGOSS spectra and tested on 2200 FLAMES-UVES spectra, using parameters from the GES iDR4. In the histogram plots, the dark red and light red lines correspond to distributions of stars with S/N >150 and <100, respectively

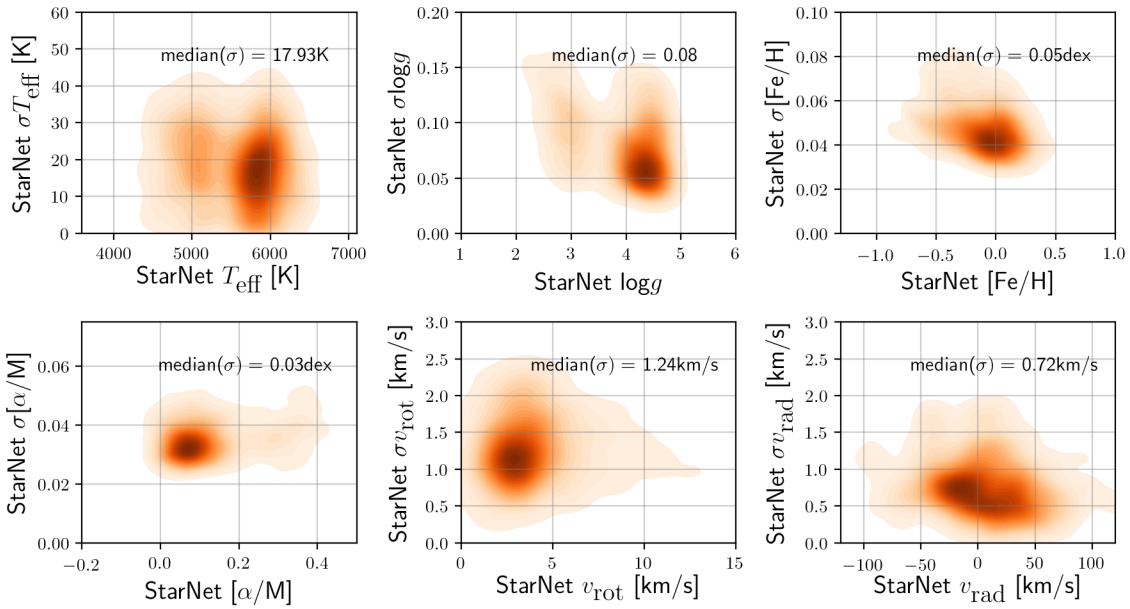

Figure 2.15: `StarNet-INTRIGOSS` was tested on the Gaia-ESO FLAMES-UVES spectra and shown here are density plots for the uncertainties of StarNet's predictions

relatively small (both in terms of the uncertainties and the residual of predictions), while the differences between PHOENIX and INTRIGOSS were relatively large. Therefore, if StarNet is trained on the AMBRE or FERRE grids, one might expect similar results when predicting stellar parameters from the FLAMES-UVES spectra; however, more discrepant results may be expected when trained on the PHOENIX spectra.

To test this assumption, the same StarNet architecture was separately trained on 90,000 augmented AMBRE, FERRE, and PHOENIX spectra (spanning the same parameter space as INTRIGOSS). These will be referred to as `StarNet-AMBRE`, `StarNet-FERRE`, and `StarNet-PHOENIX`. The results for the predictions on the GES benchmark stars for each trained StarNet model are summarized in Table 2.7.1. Overall, training with INTRIGOSS spectra gives better results than training with the other grids, while training with PHOENIX spectra give the most discrepant results. `StarNet-INTRIGOSS` tests especially well on metallicity (which can be accounted for by their highly tuned line list), where the maximum absolute deviation is 0.05, as opposed to 0.29 - 0.39 for the other models (corresponding to a consistent *under*-prediction in all models).

An apparent shortcoming of `StarNet-INTRIGOSS` seems to be in the predictions of temperature and surface gravity for the metal-rich giants: whereas `StarNet-FERRE` and `StarNet-AMBRE` have consistent residuals for $T_{\rm eff}$ and log$g$ values for all three groups of stars, `StarNet-INTRIGOSS` appears to systematically over-predict those values for the MRGs. One possibility is the offsets are due to the improved line list, carried out specifically for the cool giants, in the INTRIGOSS analysis. As described by Franchini et al. (2018), the INTRIGOSS spectra were computed with atomic and molecular line lists built by tuning oscillator strengths in order to reproduce a set of high-resolution reference spectra, namely the Solar spectrum and the GES spectra of five cool giants with high SNR (>100). It is very likely that these improvements impact the comparisons with the GES benchmark star parameters for these MR giant stars.

2.7.2 Recommendations: beyond INTRIGOSS

While training with the INTRIGOSS grid has yielded the most precise results for the GES UVES spectra, the grid is currently limited to significantly smaller temperature, metallicity, and wavelength regimes than FERRE, AMBRE, and PHOENIX. It is inherently less versatile for very metal-poor stars, and for the analysis of observed spectra which do not lay in its very narrow wavelength range. However, the promising results from the INTRIGOSS grid, as shown in this study, do suggest that further work to extend the wavelength coverage of

INTRIGOSS is warranted.

For applications outside of the INTRIGOSS parameter and/or wavelength regimes, we have found that StarNet can be trained on any of the other sets of synthetic grids. However, the FERRE and AMBRE trained StarNet predictions are marginally more precise than those when trained with the PHOENIX grid. Overall, and for general purposes, we recommend StarNet trained with the FERRE grid, `StarNet-FERRE`, combining both good precision and large parameter extent. Applications for `StarNet-FERRE` can include the analysis of optical spectral archives, such as for CFHT ESPaDOnS (Donati et al. (2006), and Gemini GRACES (Chene et al. (2014), for precision stellar parameters. The flexibility of `StarNet-FERRE` also means that it can be trained for lower resolution spectral archives as well, e.g., the SDSS BOSS database (Dawson et al. (2016) or ESO Xshooter library (Vernet et al. (2011). Unfortunately, the current StarNet pipeline requires retraining for each new observational data set and/or for each new synthetic grid library. In the future, this could be accelerated by using transfer learning techniques, e.g., training a very large NN that would cover most cases and would be tuned to specific data sets or spectral parameters.

2.7.3 Caveats for ML applications

One of the main advantages of the CNN with deep ensembling method developed in this paper is its adaptability to any spectroscopic survey and any grid of synthetic spectra, and its ability to predict a consistent set of stellar parameters across surveys, with the same calibration data set. The precision in the method depends on the quality of the synthetic spectra, how closely they match the observed spectra, and how well the model can learn a representation of the synthetic spectra: ideally all synthetic grids would include intra-grid spectra for assessing the interpolation accuracy, but a simple train/validation/test split would suffice.

As opposed to training a NN on observed spectra, training on a grid of synthetic spectra has the added benefit of not needing to worry about correlations between stellar parameters being picked up in the training process. For example, when the bulk of a training set of observed spectra has a Mg-Al correlation, then a NN is more likely to falsely assign a Mg-Al correlation to globular cluster stars even if they are known *a priori* to be anti-correlated (e.g., see the discussion by Leung & Bovy, 2019). This problem can be mitigated with domain knowledge, e.g. by windowing the spectra according to spectral features from a particular element, or through an extensive (though potentially prohibitive) array of chemical abundances in the synthetic spectral grids.

There is also the problem of finding rare stars (e.g. carbon-enhanced metal-poor stars, ultra metal-poor stars, stars captured from nearby dwarf satellites, or r-process rich stars, and even spectroscopic binaries; see Venn et al. 2019; Monty et al. 2019; Arentsen et al. 2019; Sakari et al. 2018; Kielty et al. 2017). If a training set does not include a significant proportion of peculiar stars, then predictions on these rare populations will suffer. In machine learning applications, the training set is often the limiting factor, so special care is required to account for out-of-distribution samples. For data-driven methods, this problem is even more difficult to address (tiny sample sizes); however, for synthetic grids, spectra of rare stars can be added *a posteriori* and the NN re-trained.

In cases where the sample size of a spectroscopic survey is low (in the hundreds or low thousands of spectra), then it *might* be infeasible to acquire a trained NN which produces accurate results, since the size of the training set could be a limiting factor. This problem is overcome by synthetic spectra: the only limits to the size of a synthetic training set are storage space and the computing time required to produce the spectra.

To extend this analysis to predictions of chemical abundances, spectra could be produced within the parameter range of an existing grid, but not aligned with the grid points (see Ting et al., 2019). Indeed, producing spectra in a grid is quite a rigid and perhaps out-dated strategy as there will inevitably be multiple realizations of the same stellar parameter, resulting in an over abundance of spectra needed for a NN analysis. It is much more economical to produce spectra with randomly varying parameters, especially when considering extending grids to > 10 dimensions.

Training on synthetic spectra allows for a complete model and analysis pipeline to be created before the first light is collected for a spectroscopic survey, meaning as spectra are collected from a telescope their parameters (even radial velocities), along with uncertainties, can be derived in real time. Because our method derives uncertainties, we can also in real-time assess the accuracy of predictions, providing valuable feedback needed to determine how long a star should be observed to achieve a certain level of accuracy.

2.8 Conclusions

In this paper, we have presented an updated version of our StarNet convolutional neural network used for the precision analysis of high-resolution stellar spectra. The main update has been the implementation of deep ensembling to estimate realistic uncertainties in the predicted stellar parameters. In addition:

- StarNet has been trained successfully on four independent grids of high-resolution synthetic spectra (INTRIGOSS, FERRE, AMBRE, and PHOENIX), highlighting its versatility.

- Data augmentation is necessary to overcome the synthetic gap, such that different synthetic grids overlap with one another, as well as with observational data from the Gaia-ESO FLAMES-UVES spectroscopic survey.

- Data pre-processing included resolution matching (R=47,000), sampling matching (put onto the UVES wavelength grid), and continuum normalization that was consistent between synthetic and observed spectra. The spectra were augmented with a range of rotational and radial velocities, Gaussian noise, and random zero flux values to mimic bad pixels. Finally, regions of known telluric lines were masked in the synthetic and observational data.

- Augmenting the training data with noise *before* the asymmetric sigma-clipping continuum estimation step was necessary to decrease the biases in predictions.

- Once trained, StarNet was shown to predict stellar parameters for ~2300 FLAMES-UVES optical spectra with high precision compared with traditional methods, and within seconds.

- The precision in StarNet's predictions for FLAMES-UVES spectra, when compared to Gaia-ESO benchmark stars and calibration clusters, is best when StarNet is trained on the INTRIGOSS grid, as expected since this grid has been specifically tuned for the Gaia-ESO survey in this wavelength region.

- When StarNet is trained with the FERRE synthetic spectral grid, the precision in the results are also excellent (closely followed in precision by training with the AMBRE grid). Due to the limited stellar parameter range and wavelength coverage of INTRIGOSS, we suggest that the FERRE grid is currently the best choice for general purpose machine learning applications of high resolution optical spectra.

For the near future, we plan to train StarNet for the analysis of optical spectra from Canadian observational facilities, such as CFHT ESPaDOnS and Gemini GRACES, and to prepare for observational data from the upcoming Gemini GHOST spectrograph. We are also developing new tools for detailed chemical abundances. Our codes are publicly available and simple to adapt to any set of synthetic spectra.

Chapter 3

Summary and Future Plans

3.1 Summary

Spectroscopic surveys coming online in the next several years have unique needs, including a high level of precision in derived properties and analysis techniques that can handle massive amounts of data. These needs can be met by the implementation of machine learning methods, and in this thesis I have developed a deep learning framework for the analysis of stellar spectra using *synthetic* spectra as a training set. The methods can be applied to *any* grid of synthetic spectra, and are thus applicable to any spectroscopic survey, no matter the size.

3.2 Conference Presentations

Early stages and results from this work have been presented at several conferences, as listed here:

Name of event	Date (DD/MM/YYYY)	Type of presentation	Title of work
NTCO AGM (Victoria BC)	20/11/2017	Poster	Deep neural networks in the analysis of stellar spectra
CASCA 2018 (Victoria BC)	24/05/2018	Poster	Deep neural networks in the analysis of stellar spectra
Pristine AGM (Victoria, BC)	09/25/2018	Talk	StarNet in the optical
NTCO AGM (Quebec City, QC)	20/11/2018	Poster	Deep learning of synthetic spectra in all wavelength regimes
ARCNet Seminar (Victoria, BC)	13/12/2018	Talk	The continuing voyages of StarNet: Deep learning of optical spectra
CASCA 2019 (Montreal, QC)	19/06/2019	Poster	Deep learning of synthetic spectra in all wavelength regimes

3.3 Future Plans

I plan to continue to work in this field and develop the StarNet application as part of my PhD research. In particular, and in contrast to my MSc work, the research will focus on the determination of several chemical abundances needed to reveal a full understanding of the Galaxy by finding the most interesting and important populations of stars, e.g. disrupted star clusters and dwarf galaxies, r-process enhanced stars, spectroscopic binaries, and carbon-enhanced metal poor stars. I also propose to make StarNet more flexible for the upcoming era of multi-dimensional stellar spectroscopy, i.e., so that stellar parameters can be determined homogeneously from any wavelength regime or spectral resolution. This project will minimize systematic differences between the "big data" surveys over the next decade, helping in the precision necessary to unravel the formation and accretion history of the Milky Way.

The **first stage** will be using spectral synthesis codes to generate a multi-element abundance dataset of synthetic spectra covering optical and IR wavelengths, the first of its kind. The dataset will consist of spectra of varying temperature, surface gravity, metallicity, rotational and radial velocities, and several chemical abundances. As it stands, all publicly available synthetic spectra exist in rigid "grids" with a minimal number of dimensions; at the sampling required for traditional spectra processing pipelines, additional dimensions incur a heavy cost of requiring exponentially more synthesized spectra, meaning these grids normally do not include chemical abundances as a parameter. I seek to disrupt this grid paradigm, as machine learning does not require it: instead, the full parameter space (with some constraints) will be randomly sampled and a NN will learn the optimal interpolation. Although this approach will still require upwards of millions of spectra to be generated (as opposed to tens of thousands for the existing grids), the computing and storage infrastructure available to Canadian astronomers through Compute Canada will easily allow it.

The **second stage** will involve training StarNet on this dataset and subsequently testing StarNet on spectra from existing surveys, such as APOGEE and Gaia-ESO, to assess the overall performance. It is in this stage that a newer NN architecture will be tested, combining elements of The Payne (which generates spectra from stellar properties) and StarNet (which predicts stellar properties from spectra) in an invertible network (Ardizzone et al., 2018). Such a network can simultaneously provide stellar properties and a best-fit synthetic stellar spectrum for an observed spectrum, allowing for a richer analysis.

The **third** stage will be integrating StarNet with new and upcoming optical surveys using spectrographs like ESPaDOnS, GRACES, and GHOST, to derive precise chemical abun-

dances for hundreds of thousands of stars; our research group already has ESPaDOnS and GRACES spectra from the current ongoing CFHT and Gemini programs for spectroscopy of new metal-poor stars found in the Pristine survey (Aguado et al., 2019, Venn et al. 2019). Completion of this stage will make StarNet an ideal application for the analysis of spectra from massive-scale surveys like SDSS-V, DESI, WEAVE, 4MOST, PFS, and ultimately the Canadian-led MSE (Mcconnachie et al., 2016). Combining the spectroscopic radial velocities and chemical abundances derived by StarNet with proper motions and parallaxes from Gaia (Brown et al., 2018) enables a full chemo-dynamic characterization of the Galaxy.

Throughout all of these stages, the algorithms developed will be open sourced so that members of the community can use and expand upon the work. The opportunities for collaborations with researchers worldwide will help direct which surveys to focus our efforts on, with the ultimate goal being the most comprehensive scientific analysis possible. This could include additional observational information from photometric indices and time domain variability (e.g., from the Large Synoptic Survey Telescope), and this work will be carried out when the Gaia DR3 data is released, giving us fainter stars, binary and variability information, and higher precision chemo-dynamics.

Research Significance: Machine learning algorithms, and in particular NNs, are ubiquitous in our society, used in applications ranging from tumor identification in CT scans, language recognition and translation, to self-driving vehicles. The reason for the widespread use of NNs is that they are excellent tools for making precise and accurate predictions given an appropriate dataset. Ultimately, I envision implementing StarNet directly into observatory operations such that a data processing pipeline could provide science right off the telescope. These applications will contribute to the most precise map of our Galaxy, the formation history of its various stellar populations (thin disk, thick disk, and bulge), the accretion history of the metal-poor halo, and a better idea of the nature of dark matter. Understanding our host Galaxy through this galactic archaeology gives us a better insight into both our own home and the formation processes in other galaxies, deepening our sense of place in the universe.

Bibliography

Abolfathi B., Aguado D. S., Aguilar G., Allende Prieto C., Almeida A., Ananna T. T., Anders F. e. a., 2018, http://dx.doi.org/10.3847/1538-4365/aa9e8a , https://ui.adsabs.harvard.edu/abs/2018ApJS..235...42A 235, 42

Allende Prieto C., Koesterke L., Hubeny I., Bautista M. A., Barklem P. S., Nahar S. N., 2018, http://dx.doi.org/10.1051/0004-6361/201732484 , https://ui.adsabs.harvard.edu/abs/2018AA...618A..25A 618, A25

Arentsen A., Starkenburg E., Shetrone M. D., Venn K. A., Depagne É., McConnachie A. W., 2019, http://dx.doi.org/10.1051/0004-6361/201834146 , https://ui.adsabs.harvard.edu/abs/2019AA...621A.108A 621, A108

Blanco-Cuaresma S., Soubiran C., Jofré P., Heiter U., 2014, http://dx.doi.org/10.1051/0004-6361/201323153 , https://ui.adsabs.harvard.edu/abs/2014AA...566A..98B 566, A98

Buder S., et al., 2018, http://dx.doi.org/10.1093/mnras/sty1281 , https://ui.adsabs.harvard.edu/abs/2018MNRAS.478.4513B 478, 4513

Casey A. R., Hogg D. W., Ness M., Rix H.-W., Ho A. Q., Gilmore G., 2016a, arXiv preprint arXiv:1603.03040

Casey A. R., Hogg D. W., Ness M., Rix H.-W., Ho A. Q. Y., Gilmore G., 2016b, arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2016arXiv160303040C p. arXiv:1603.03040

Chen S., Billings S., Grant P., 1990, International journal of control, 51, 1191

Chene A.-N., et al., 2014, in . p. 915147 (http://arxiv.org/abs/1409.7448 arXiv:1409.7448), http://dx.doi.org/10.1117/12.2057417 doi:10.1117/12.2057417

Choi J., Dotter A., Conroy C., Cantiello M., Paxton B., Johnson B. D., 2016, The Astrophysical Journal, 823, 102

Chollet F., 2015, keras, https://github.com/fchollet/keras

Cui X.-Q., et al., 2012, http://dx.doi.org/10.1088/1674-4527/12/9/003 Research in Astronomy and Astrophysics, https://ui.adsabs.harvard.edu/abs/2012RAA....12.1197C 12, 1197

Dalton G., et al., 2012, in Ground-based and Airborne Instrumentation for Astronomy IV. p. 84460P

Dalton G., et al., 2018, in . p. 107021B, http://dx.doi.org/10.1117/12.2312031 doi:10.1117/12.2312031

Dawson K. S., et al., 2016, http://dx.doi.org/10.3847/0004-6256/151/2/44 , https://ui.adsabs.harvard.edu/abs/2016AJ....151...44D 151, 44

Deng L.-C., et al., 2012, Research in Astronomy and Astrophysics, 12, 735

Donati J. F., Catala C., Landstreet J. D., Petit P., 2006, in Casini R., Lites B. W., eds, Astronomical Society of the Pacific Conference Series Vol. 358, Solar Polarization 4. p. 362

Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, http://dx.doi.org/10.1093/mnras/stx3298 , https://ui.adsabs.harvard.edu/abs/2018MNRAS.475.2978F 475, 2978

Franchini M., et al., 2018, The Astrophysical Journal, 862, 146

Freeman K., Bland-Hawthorn J., 2002, Annual Review of Astronomy and Astrophysics, 40, 487

García Pérez A. E., Allende Prieto C., Holtzman J. A., Shetrone M., Mészáros S., Bizyaev D. e. a., 2016, http://dx.doi.org/10.3847/0004-6256/151/6/144 , https://ui.adsabs.harvard.edu/abs/2016AJ....151..144G 151, 144

Gilmore G., et al., 2012, The Messenger, https://ui.adsabs.harvard.edu/abs/2012Msngr.147...25G 147, 25

Glorot X., Bengio Y., 2010, in Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp 249–256

Gurney K., 1997, An introduction to neural networks. CRC press

Harris W. E., 2010, arXiv preprint arXiv:1012.3224

Heiter U., Jofré P., Gustafsson B., Korn A. J., Soubiran C., Thévenin F., 2015, Astronomy & Astrophysics, 582, A49

Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G., 2018, Nature, 563, 85

Ho A. Y., et al., 2017, The Astrophysical Journal, 836, 5

Holtzman J. A., et al., 2018, http://dx.doi.org/10.3847/1538-3881/aad4f9 , https://ui.adsabs.harvard.edu/abs/2018AJ....156..125H 156, 125

Husser T. O., Wende-von Berg S., Dreizler S., Homeier D., Reiners A., Barman T., Hauschildt P. H., 2013, http://dx.doi.org/10.1051/0004-6361/201219058 , https://ui.adsabs.harvard.edu/abs/2013AA...553A...6H 553, A6

Jahandar F., et al., 2017, http://dx.doi.org/10.1093/mnras/stx1592 , https://ui.adsabs.harvard.edu/abs/2017MNRAS.470.4782J 470, 4782

Jofré P., et al., 2014, Astronomy & Astrophysics, 564, A133

Kielty C. L., Venn K. A., Loewen N. B., Shetrone M. D., Placco V. M., Jahandar F., Mészáros S., Martell S. L., 2017, http://dx.doi.org/10.1093/mnras/stx1594 , https://ui.adsabs.harvard.edu/abs/2017MNRAS.471..404K 471, 404

Kingma D. P., Ba J., 2014, arXiv preprint arXiv:1412.6980

Kollmeier J. A., et al., 2017, arXiv preprint arXiv:1711.03234

Kordopatis G., et al., 2013, http://dx.doi.org/10.1088/0004-6256/146/5/134 , https://ui.adsabs.harvard.edu/abs/2013AJ....146..134K 146, 134

Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Advances in neural information processing systems. pp 1097–1105

Kurucz R. L., 1970, SAO Special report, 309

Kurucz R. L., 2011, Canadian Journal of Physics, 89, 417

Kurucz R., 2014, in , Determination of Atmospheric Parameters of B-, A-, F-and G-Type Stars. Springer, pp 63–73

Lakshminarayanan B., Pritzel A., Blundell C., 2017, in Advances in Neural Information Processing Systems. pp 6402–6413

Lee Y. S., et al., 2008, http://dx.doi.org/10.1088/0004-6256/136/5/2022 , https://ui.adsabs.harvard.edu/abs/2008AJ....136.2022L 136, 2022

Lee Y. S., et al., 2011, http://dx.doi.org/10.1088/0004-6256/141/3/90 , https://ui.adsabs.harvard.edu/abs/2011AJ....141...90L 141, 90

Leung H. W., Bovy J., 2018, Monthly Notices of the Royal Astronomical Society, 483, 3255

Leung H. W., Bovy J., 2019, http://dx.doi.org/10.1093/mnras/sty3217 , https://ui.adsabs.harvard.edu/abs/2019MNRAS.483.3255L 483, 3255

Martins L., Coelho P., 2017, Canadian Journal of Physics, 95, 840

Monty S., Venn K. A., Lane J. M. M., Lokhorst D., Yong D., 2019, arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2019arXiv190911969M p. arXiv:1909.11969

Nair V., Hinton G. E., 2010, in Proceedings of the 27th international conference on machine learning (ICML-10). pp 807–814

Ness M., Hogg D. W., Rix H.-W., Ho A. Y., Zasowski G., 2015a, The Astrophysical Journal, 808, 16

Ness M., Hogg D. W., Rix H. W., Ho A. Y. Q., Zasowski G., 2015b, http://dx.doi.org/10.1088/0004-637X/808/1/16 , https://ui.adsabs.harvard.edu/abs/2015ApJ...808...16N 808, 16

Ovadia Y., et al., 2019, arXiv preprint arXiv:1906.02530

Pancino E., et al., 2017, http://dx.doi.org/10.1051/0004-6361/201629450 , https://ui.adsabs.harvard.edu/abs/2017AA...598A...5P 598, A5

Pasquini L., et al., 2002, The Messenger, 110, 1

Recio-Blanco A., Bijaoui A., de Laverny P., 2006, http://dx.doi.org/10.1111/j.1365-2966.2006.10455.x , https://ui.adsabs.harvard.edu/abs/2006MNRAS.370..141R 370, 141

Sakari C. M., et al., 2018, http://dx.doi.org/10.3847/1538-4357/aae9df , https://ui.adsabs.harvard.edu/abs/2018ApJ...868..110S 868, 110

Smiljanic R., et al., 2014, Astronomy & astrophysics, 570, A122

Sneden C., Kraft R. P., Shetrone M. D., Smith G. H., Langer G., Prosser C. F., 1997, The Astronomical Journal, 114, 1964

Starkenburg E., et al., 2017, Monthly Notices of the Royal Astronomical Society, 471, 2587

Steinmetz M., et al., 2006, http://dx.doi.org/10.1086/506564 , https://ui.adsabs.harvard.edu/abs/2006AJ....132.1645S 132, 1645

Tamura N., et al., 2018, in . p. 107021C, http://dx.doi.org/10.1117/12.2311871 doi:10.1117/12.2311871

Ting Y.-S., Conroy C., Rix H.-W., Cargile P., 2019, The Astrophysical Journal, 879, 69

Venn K., et al., 2019, arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2019arXiv191006340V p. arXiv:1910.06340

Vernet J., et al., 2011, http://dx.doi.org/10.1051/0004-6361/201117752 , https://ui.adsabs.harvard.edu/abs/2011AA...536A.105V 536, A105

Wang R., et al., 2019, http://dx.doi.org/10.1088/1538-3873/aaf25f , https://ui.adsabs.harvard.edu/abs/2019PASP..131b4505W 131, 024505

Worley C., de Laverny P., Recio-Blanco A., Hill V., Bijaoui A., 2016, Astronomy & Astrophysics, 591, A81

Xiang M., et al., 2019, arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2019arXiv190809727X p. arXiv:1908.09727

Yanny B., et al., 2009b, http://dx.doi.org/10.1088/0004-6256/137/5/4377 , https://ui.adsabs.harvard.edu/abs/2009AJ....137.4377Y 137, 4377

Yanny B., et al., 2009a, The Astronomical Journal, 137, 4377

Zasowski G., et al., 2019, http://dx.doi.org/10.3847/1538-4357/aaeff4 , https://ui.adsabs.harvard.edu/abs/2019ApJ...870..138Z 870, 138

Zhang X., Zhao G., Yang C. Q., Wang Q. X., Zuo W. B., 2019, http://dx.doi.org/10.1088/1538-3873/ab2687 , https://ui.adsabs.harvard.edu/abs/2019PASP..131i4202Z 131, 094202

de Jong R. S., Chiappini C., Schnurr O., 2012, in EPJ Web of Conferences. p. 09004

de Jong R. S., et al., 2019, http://dx.doi.org/10.18727/0722-6691/5117 The Messenger, https://ui.adsabs.harvard.edu/abs/2019Msngr.175....3D 175, 3

de Laverny P., Recio-Blanco A., Worley C. C., Plez B., 2012, http://dx.doi.org/10.1051/0004-6361/201219330 , https://ui.adsabs.harvard.edu/abs/2012AA...544A.126D 544, A126