Extending the Reach of *Gaia* with Masked Stellar Autoencoders

by

Aydan McKay
B.Sc., University of Victoria, 2022

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

MASTER OF SCIENCE

in the Department of Physics and Astronomy

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Xʷsepsəm/Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Extending the Reach of *Gaia* with Masked Stellar Autoencoders

by

Aydan McKay
B.Sc., University of Victoria, 2022

Supervisory Committee

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

Dr. Kim Venn, Co-Supervisor
(Department of Physics & Astronomy)

Abstract

I present the Masked Stellar Autoencoder, a new data-driven holistic stellar model for Galactic archaeology. The MSA is trained using the complete *Gaia* DR3 XP spectra catalogue by implementing a self-supervised masking algorithm to enforce the learning of the relationships within the data itself. Photometry from six additional surveys spanning optical and infrared wavelengths are integrated into the dataset, making the model robust to missing spectroscopic and photometric data. This allows the embeddings to retain accuracy beyond the depth of the XP spectra. The model was first pretrained on the ∼220 million stars from *Gaia* DR3 with photometry for the purpose of reconstructing the information. I then demonstrate the informative embeddings produced by this astronomical foundation model with the predictive task of deriving atmospheric parameters and stellar ages using high-resolution spectroscopic surveys (APOGEE, GALAH). The model achieved mean absolute errors of 92 K in $T_{\rm eff}$, 0.08 dex in log g, and 0.09 dex in [Fe/H], demonstrating its competitive position with XGBoost and transformer-based models trained with APOGEE labels. Furthermore, the model achieved mean absolute errors of 0.05 dex in [α/Fe] and 1.3 Gyr in age, with only marginal increases in metrics when missing XP spectra. The MSA also predicts errors for the stellar parameters, which were shown to be largely representative of the predicted values, with slight underconfidence in the width of the asymmetric errors. The change in the accuracy of the predictions with pretraining dataset size was examined, and the model was leveraged to predict stellar parameters for a subset of open clusters and dwarf galaxies Leo I and Fornax. These estimates displayed a potential improvement in parallax measurements at higher distances and crowded regions. This model effectively bridges the gap between spectroscopic and photometric samples within a single, consistent framework, poised to improve with the inclusion of additional photometric surveys and upcoming *Gaia* releases.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank Dr. Sébastien Fabbro, for his encouragement, wealth of knowledge, and continuous support during my research. I will forever be grateful for the opportunity that you and Kim gave me as an honours student, and the continuous oppurtunities you've provided for me during my degree since. I would also like to thank Dr. Kim Venn for her expertise and wonderful research group, of which I am honoured to have been a part.

I would like to thank all my friends, but in particular Scott, Daniel, Simon, Noah, Nicola, Celina, and Linzhi, for the support throughout my undergraduate and graduate studies. You are all a large part of who I am today, from the countless hours we spent studying together, to those we spent laughing and celebrating together. Your friendships are invaluable to me, and I look forward to sharing many more memories in the future.

I would like to thank my family, Tammy, Todd, Liam, and Nathan. Your belief in me, and your unquestionable support throughout all the forks in the road have helped me reach where I am now.

Most of all, I would like to thank my fiancée, Jenna. This work would not have been possible without your love and your support. You are my inspiration, my best friend, and the love of my life. I can't wait for this next stage of our lives.

Dedication

This work is dedicated to my family, by blood and by choice.

Chapter 1

Introduction

As the most accessible example of a large star-forming galaxy, the Milky Way serves as a primary astrophysical laboratory for studying the fundamental mechanisms driving galaxy formation and evolution over cosmic time. The chemodynamical structure evidenced through stars and Galactic stellar populations trace the history of both galaxies and the universe (Freeman & Bland-Hawthorn, 2002). The field of Galactic archaeology employs stellar chemical abundances, kinematics, and ages (where attainable) to disentangle this evolution history, revealing merger and accretion events through stellar fossils in the components of the Galaxy (thin and thick disk, bulge, and in particular, the halo) (Ivezić et al., 2012; Frebel & Norris, 2015; Bland-Hawthorn & Gerhard, 2016). The stellar halo of the Milky Way consists of an old population of stars which formed at very early times relative to the Galaxy tracing large amounts of substructure. This includes stellar streams left in the wake of dwarf galaxies that reveal the conditions under which the Local Group was formed (Helmi, 2008). The cosmological applications of Galactic archaeology leverage the properties of the Milky Way and its stellar populations to test our models of the early universe from devolving the history of galactic and stellar evolution back through time (Freeman & Bland-Hawthorn, 2002). Therefore, for Galactic archaeology, deriving precise stellar ages is paramount. To understand and characterise stellar populations formed at the earliest times, proxies for stellar ages are often required due to intrinsic difficulties in measuring ages for field stars (Soderblom, 2010).

Proxies for age exist from the spectroscopic content of a stellar atmosphere. The abundance of iron to hydrogen (metallicity, [Fe/H]) in a star traces the history of star formation and iron enrichment through Type-Ia supernovae, whereas the ratio of α elements to iron ([α/Fe]) correlates with the history of Type II/core-collapse supernovae, dating the population based on the evolution of the main sequence. The [C/N] abundance from red giants also acts as a proxy for age as it correlates with mass, which can be used to derive age through stellar modelling of the mass-dependent dredge-up process in red giant stars (Masseron &

Gilmore, 2015; Martig et al., 2016). In circumstances for which ages are calculable for individual stars and stellar populations, their derivations are model-dependent and require assumptions based on the stellar physics used. These methods include isochrone fitting of the main sequence for star clusters (see VandenBerg et al., 2013; Hunt & Reffert, 2024, for examples), the isochrone projection method for individual stars (Jørgensen & Lindegren, 2005), and gyrochronology, the spin down of low-mass stars (Skumanich, 1972; Barnes, 2003). Additionally, asteroseismology, the analysis of fluctuations and oscillations in stellar atmospheres (Cunha et al., 2007), can reveal stellar ages for red giants. However, a recent study showed the limitation of deriving ages using asteroseismology with respect to metal-poor stars (Lindsay et al., 2025). Using asteroseismic datasets combined with atmospheric parameters (e.g. the APOKASC project, Pinsonneault et al., 2014, 2018, 2024), ages have also been predicted using data-driven models that extract the ages of red giant stars embedded in high-resolution stellar spectra (Mackereth et al., 2019; Leung et al., 2023), made possible only with large and accurate datasets.

The study of Galactic and stellar archaeology has flourished in recent years with increasing Galactic dataset sizes such as *Gaia* (Gaia Collaboration 2016), the DECam Local Volume Exploration Survey (DELVE; Drlica-Wagner et al., 2021), and the *Pristine* Survey (Starkenburg et al., 2017). These datasets contain multiple *modalities* of data, including radial velocities, photometric, spectroscopic, and astrometric observations, which when combined reveal otherwise obscured, features such as streams and dwarf galaxy remnants. Discoveries with these datasets have led to fundamental shifts in our knowledge of the Local Group's history. One such example includes the *Gaia*-Sausage-Enceladus (GS/E) feature, revealed through proper motions, radial velocities of globular clusters, and elemental abundances of millions of stars in the *Gaia* dataset, which acts as evidence of the last major merger with the Milky Way (Belokurov et al., 2018; Helmi et al., 2018). *Gaia* has also been used to confirm the existence of the "Splash", a metal-rich component of the solar neighbourhood exhibiting halo-like properties (e.g. Bonaca et al., 2017; Haywood et al., 2018), and reassign the discovery's designation as part of the thick disc, using kinematics, abundances, and isochrone ages computed for *Gaia* stars by Sanders & Das (2018) (Belokurov et al., 2020). Belokurov & Kravtsov (2022) use the proper motions of stars measured by *Gaia* to distinguish a pre-Milky Way disc in situ stellar population as the progenitor of star formation and the Galaxy (for a full review on Galactic archaeology with *Gaia* see Deason & Belokurov, 2024). Many other examples of substructure in the Milky Way have been discovered through large surveys such as *Gaia* and DELVE including streams (e.g. Thomas et al., 2020; Martin et al., 2022) and ultra-faint dwarf galaxies (e.g. Torrealba et al., 2019; Cerny et al., 2023; Smith et al., 2024),

which challenge the currently understood mechanisms of cosmology and stellar evolution.

Milky Way-based projects including DELVE and *Gaia* have data on billions of stars, contributing the greater understanding of the Galaxy and its components, as they represent some of the smallest building blocks. As these collaborations and surveys have greatly increased the amount of existing raw data, automated algorithms have replaced manual boutique pipelines. However, the extraction of physics and fundamental knowledge from the data has been slower to transition to automation, despite the overwhelming amount of data, as the pertinent information hidden in the data requires a general understanding of the data diversity, and its complexity and inter-connectivity to reveal, which lies outside the scope and ability of traditional data reduction pipelines. To parse through a galaxy's worth of data, machine learning (ML) methods have shown to be well suited to the complex and increasingly large datasets. When ML algorithms are given a large mixture of high-quality measurements, they have shown the ability to extract and parameterize catalogues to a degree similar to that of experts in a fraction of the time (Pearson et al., 2018), while the performance of the models improve with increasing dataset size (Kaplan et al., 2020).

The *Gaia* third data release (hereafter DR3, Gaia Collaboration et al., 2023) comprised ∼219 million stars (De Angeli et al., 2023) with low-resolution spectroscopy (R ∼ 30-100 Carrasco et al., 2021), which has been of particular interest in the field of astronomical ML. A popular application of these low-resolution spectra (BP/RP or XP spectra, Section 2.1) has been stellar atmospheric parameter prediction. Andrae et al. (2023) employed a decision tree-based model to fit the spectral information and re-derive the atmospheric parameters of effective temperature, $T_{\rm eff}$, surface gravity, log g, and [Fe/H] for 175 million stars (requiring good parallax measurements accounting for the discrepancy). [α/Fe] has been predicted via similar algorithms (e.g. Gavel et al., 2021; Guiglion et al., 2024; Hattori, 2024), along with metallicity for very metal-poor stars (VMPs, [Fe/H] < -2) (Yao et al., 2024), and carbon enhancement (carbon-enhanced metal-poor stars, CEMPs Lucey et al., 2023). Furthermore, with *Gaia* XP spectra, stellar abundance estimation (Ardern-Arentsen et al., 2025; Fallows & Sanders, 2024; Laroche & Speagle, 2025; Li et al., 2024; Kane et al., 2025) and stellar classification (García-Zamora et al., 2023; Vincent et al., 2024; Kao et al., 2024; Viscasillas Vázquez et al., 2024) algorithms have been facilitated through neural networks and machine learning methods. Beyond classification, machine learning techniques have been leveraged for dust extinction determinations (Zhang et al., 2023; Zhao et al., 2024) and age prediction using spectral-derived abundances (Almannaei et al., 2024).

Despite their widespread application, supervised algorithms have few diverse labels to train with. To train a supervised algorithm on specifically finding VMP stars, the dataset

must be curated to deal with the label imbalance. Since such few examples exist of labelled VMP stars relative to the greater population, a model can learn only to predict more common metallicities while maintaining an extremely high accuracy for the entire sample. These supervised learning algorithms are also limited by requiring homogeneity in the data, as they struggle with mixed labels and cross-domain generalization, the ability of the algorithm to perform well with different types of data or datasets. In particular, these models have difficulty integrating data from different modalities (e.g., catalogues and spectra). Additionally, predictive biases can arise when models are trained on small or non-uniformly distributed datasets, because the true properties of rare or scientifically interesting objects may fall outside the parameter space defined by the training sample (imbalance).

With increasing dataset size, the field of *deep learning* (LeCun et al., 2015) has grown in popularity, with larger models being constructed that implement vastly more complex and numerous non-linear connections. *Foundation models* have emerged as a powerful approach, designed with the purpose of learning generalizable representations from large and diverse datasets through *self-supervised learning* (SSL) and/or generative modelling. This methodology learns representations of the data from the data itself, with one example being through predicting missing or masked parts of the input/real data (features), extracting meaningful embedded data representations without requiring explicit labels. The labels refer to categorized or annotated data (e.g. stellar features, chemical abundances, distance, etc.), that require several steps to acquire, including spectroscopic surveys, stellar pipelines, and ultimately, funding. Unlike traditional supervised models that rely on fully labelled data in the hundreds of thousands for each individual task, foundation models leverage this self-supervised pre-training to develop feature-rich embeddings that can be adjusted, or *fine-tuned* on a variety of different tasks with minimal data and further training.

The adaptability of foundation models has made them particularly useful in modern astronomy with increasingly complex and heterogeneous datasets such as the Dark Energy Spectroscopic Instrument (DESI Dey et al., 2019) and *Gaia*. By capturing high-dimensional astrophysical patterns, foundation models can be exploited in fields such as stellar population analysis, stellar parameter predictions, and spectral analysis, offering a solution that is *scalable*, maintaining efficiency as dataset size and model capacity increase. Furthermore, surveys with few sources in common can be integrated through these models, resulting in an extended domain over which astronomical foundation models can be trained and tested, thus improving their robustness across different observational conditions. One such example combines DESI galaxy spectra (DESI Collaboration et al., 2024) and images in an organized embedded space, such that physical information about the galaxies can be extracted, while

also being usable for searches across both spectral and photometric modalities (Parker et al., 2024). Specific Galactic archaeology foundation models have also been conceived, performing well in both discriminative and derivation tasks, as well as in generative tasks creating spectra from spectral products, using a subset of *Gaia* spectroscopic data (Leung & Bovy, 2024).

The Masked Stellar Autoencoder (MSA) has been developed as a foundation model using the information in *Gaia* DR3 data to capture the underlying properties of a given star with self-supervision. The architecture of the MSA enables the reconstruction of masked or missing information in the features, through the learning of patterns in the low-resolution spectra, which allow the primary goal of the MSA to be realised: Predicting key stellar/spectral properties ($T_{\rm eff}$, log g, [M/H], [α/Fe], stellar age (τ_*), and parallax (ϖ)) using multiple photometric, spectroscopic, astrometric, and asteroseismic catalogues. The usefulness of the low-resolution spectra and its application in supervised machine learning networks have been shown; however, the spectra and the photometric magnitudes are not complete, as many stars have no published spectra due to a magnitude cutoff of $G < 17.65$ (De Angeli et al., 2023), far too bright for most stars in the deep stellar halo. With SSL, the MSA can be used to predict stellar labels with features beyond the limiting magnitude of a given survey as it is fundamentally trained to handle "missing data". This secondary goal of extending the depth of the XP spectra through reconstructing missing information allows for a more complete view of the Galaxy's stellar populations, filling in gaps in the structure and history of the Milky Way, and holds the promise of enabling stellar properties prediction beyond *Gaia*'s reach. By connecting the XP spectra with stellar features from high-resolution spectra, the MSA has the ability to reconcile systematic differences between surveys, reducing selection biases in spectral parameter pipelines.

I present the MSA, a data-driven model of stars, learning from spectra and photometric surveys. I demonstrate its informative embeddings by deriving stellar parameters and ages for several Galactic clusters and the dwarf galaxies Leo I and Fornax, while discussing the limitations of the model. In Chapter 2, I explain the datasets and reduction implemented for the two-stage training of the algorithm, while in Chapter 3, the deep learning paradigms for training on sparsely labelled data and the architectures used in creating the model are discussed. In Chapter 4, the reconstructive ability of masked learning is shown and the results for the downstream stellar parameter regression task are presented. Chapter 5 contains the examination and scrutinization of the model with scaling, the applicability of the MSA to other spectroscopic datasets, and the predictions for the galactic test subjects. The results are summarized in Chapter 6.

Chapter 2

Datasets

To derive robust stellar properties for a vast number of stars, I produced a comprehensive dataset from numerous stellar catalogues. *Gaia* DR3 was chosen as the reference catalogue as the wealth of low-resolution spectra were used in the pre-training portion of the algorithm, and with the added benefit of most new catalogue releases containing a cross match with *Gaia* DR3. In preparing the pre-training dataset, multiple photometric catalogues were compiled, containing wide and narrow bands which range from ultraviolet to infrared wavelengths and cover the full sky. In the same vein, the fine-tuning dataset was composed of multiple spectroscopic surveys, covering the 5 parameters previously discussed with the exception of parallax included in the *Gaia* catalogue.

2.1 Stellar Features and the *Gaia* DR3 BP/RP Low-Resolution Spectra

The European Space Agency (ESA) *Gaia* mission (Gaia Collaboration 2016) was launched in 2013 with the primary science goals of obtaining distances and proper motions to give full 6 dimensional view of the Milky Way, with positional information, distance, and their associated velocities in each direction. This was achieved with state-of-the-art precision, due to the 106 (Prusti et al., 2016) charge-coupled devices (CCD) facilitating the collection area of almost 1 billion pixels, creating galactic maps of unprecedented accuracy. The focal plane composed of the CCDs is shared by the 3 mounted instruments on the telescope: The radial velocity spectrograph (RVS), the blue and red (spectro-)photometric (BP/RP, or collectively, XP) instruments, and the astrometric instrument. Starting with the third data release, the stellar features included in the data were derived from both the medium-resolution RVS and XP instruments. The observations from the spectro-photometric instruments are aperture-free and cover a wavelength range from approximately 330nm to 1050nm with spectral resolution, R$\sim 30-100$ (Carrasco et al., 2021). The BP spectra cover the wavelengths

from 330nm to 680 nm, while the RP spectra cover the wavelength range of 640nm to 1050nm. The *Gaia* DR3 provided approximately 220 million XP sources (De Angeli et al., 2023). For the multi-epoch spectra that were included in the *Gaia* DR3 catalogue, the brightness cutoff for inclusion of a given spectrum was G $<$ 17.65 mag, bolstered by the additional requirement of 15 transits of the focal plane of the telescope (De Angeli et al., 2023). Additional sources were included beyond the magnitude cutoff for calibration and specific scientific applications, being selected dwarf stars, quasars, and galaxies. Spectra for sources observed for few transits or with a lower apparent magnitude will be included in upcoming *Gaia* data release 4.

2.1.1 The Representation of the XP Spectra

This section briefly covers the description of the flux and data model used for internally calibrating the low-resolution spectra explained in Carrasco et al. (2021) and an extensive mathematical summary Weiler et al. (2023).

The XP spectra are represented by a linear combination of basis functions (Carrasco et al., 2021) which describe the spectra observed. This differs from the usual method of tabulating spectral data as flux measurements on a wavelength grid. As listed above, a requirement of the XP spectra to be included was a minimum number of transits by the *Gaia* telescope. Through the multiple passes, a mean spectra is constructed, for both the BP and RP spectra, which is represented by the linear combination of 55 Hermite functions in both BP and RP, individually. This approach was necessary to consistently combine the tens to hundreds of individual epoch spectra for each star, which were all taken under slightly different instrumental conditions, into a single, high signal-to-noise product.

Hermite functions have multiple uses in physics and mathematics. They are the eigenfunctions of both the continuous Fourier transform and the quantum harmonic oscillator, related to Laguerre polynomials. The nth Hermite function in the linear combination is denoted as $\varphi_n(u)$ where n, the order of the Hermite function, increases linearly from 0. The recurrence relation for increasing n is:

$$u\varphi_n(u) = \sqrt{\frac{n}{2}}\varphi_{n-1}(u) + \sqrt{\frac{n+1}{2}}\varphi_{n+1}(u) \tag{2.1}$$

Where

$$\varphi_0(u) = \pi^{-1/4}\mathrm{e}^{-\frac{1}{2}u^2}$$

$$\varphi_1(u) = \pi^{-1/4}\sqrt{2}u\mathrm{e}^{-\frac{1}{2}u^2}$$

for which the Hermite function of any order can be computed for any argument, u. A linear combination of Hermite functions, as required to define for the expression of the XP spectra, can be written as such:

$$\phi(u) = \sum_{n=0}^{N} c_n \varphi_n(u)$$

Which for the sake of consistency with Carrasco et al. (2021), I will rewrite as Equation 2.2

$$h_{s\mu}(u) = \sum_{n=0}^{N} b_{sn} \cdot \varphi_n(u) \tag{2.2}$$

With $h_{s\mu}(u)$ denoting the mean spectrum of a singular source s and mean internal instrument, μ, and b_{sn} being the coefficients of the spectrum for $N+1$ basis Hermite functions. u holds the position of the pseudo-wavelength, which denotes coordinates in wavelength space corresponding to a mean CCD pixel scale.

Putting this in context with epoch spectra of celestial sources, another set of equations needs to be defined. Typically, the spectrum, $h_s(u)$, of a source is related to the spectral photon distribution (SPD; $S(\lambda)$) through a transformation given by:

$$h_s(u) = \int_0^{\infty} L(u,\lambda) R(\lambda) S(\lambda) d\lambda \tag{2.3}$$

Where $L(u,\lambda)$ is the line spread function (LSF, the integrated point spread function of a source along the line) that also encompasses the dispersion function relating u and λ in this framework, and $R(\lambda)$ is the response function, which describes the response to different features in the spectrum by the instrument. Carrasco et al. (2021) combine $L(u,\lambda)$ and $R(\lambda)$ into one integral kernel $I(u,\lambda)$ for simplicity.

As the XP spectra are required to be internally calibrated, the actual parameter spaces of R and L are of the form

$$I(u,\lambda,\omega_1,...,\omega_\Omega) = L(u(\lambda,\omega_1,...,\omega_\Omega),\lambda)\ R(u,\lambda,\omega_1,...,\omega_\Omega) \tag{2.4}$$

with a set of Ω parameters $\omega_1,...,\omega_\Omega$ to describe the influence the instrument has on the observed spectrum, for which this dependence is needed to be calibrated out. A calibration unit κ, for which within the given interval there are no discontinuities in measurements and the instrument has a smooth variation in the ω parameters, is introduced by the authors to rewrite $I(u,\lambda)$ as $I_\kappa(u,\lambda)$, the kernel for the κth calibration unit. For the internal calibration process, every calibration unit must have a linear transformation with the mean internal

instrument, μ, for which the spectrum produced has been assumed and defined as the above Equation 2.2. The transformation from the mean instrument spectrum of a given source to the spectrum of the source in the calibration unit κ is then given by

$$h_{s\kappa}(u) = \int_0^\infty A_\kappa(u, u')h_{s\mu}(u')du' \tag{2.5}$$

Where A_κ is the transformation kernel. Approximating the transformation kernel as localised and inserting Equation 2.2 into Equation 2.5 with a Riemann sum transformation, Carrasco et al. (2021) obtain the *Gaia* DR3 XP spectra formulation of

$$h_{s\kappa}(u_i) = \sum_{n=0}^{N} b_{sn} \sum_{j=-J}^{J} A_\kappa(u_i, u_{i+j})\ \varphi_n(u_{i+j})\ . \tag{2.6}$$

From the above equation, it may be seen that the XP coefficients, b_{sn}, are the data from the *Gaia* DR3 low-resolution catalogue that contain the information on the unique spectrum of an individual source. The rest of the parameters involved with Equation 2.6 act similar to a mapping of the coefficients to the sampled spectrum. By consequence, the majority of neural networks trained on the low resolution spectra contain only the coefficients as part of the features fed to these algorithms, as the observed information has been condensed to just the vector of coefficients (see list of applied algorithms in Section 1).

The final note on the low-resolution spectra is that a transformation was applied to reorder the basis functions such that those containing the largest amount of spectral information were first (Carrasco et al., 2021).

2.1.2 Astrophysical Information Content of XP Coefficients

XP spectra are only as useful as the spectral features they can encode, and what can be extracted from the coefficients. Before the release of *Gaia* DR3, the information encoded in the data was already demonstrated through theoretical models of the XP spectra. Allende Prieto (2016) showed that common stellar parameters including $T_{\rm eff}$, log g, and [Fe/H] were obtainable from the XP spectra with low standard deviations, enabling the detection metal-poor stars in the XP spectra sample. Witten et al. (2022) built upon this result, to assess the ability to detect CEMP stars from synthetic XP spectra. The main findings in detecting carbon in the XP spectra showed that with decreasing metallicity and increasing effective temperature, the error in the carbon abundance measurement grew, resulting in difficulty accurately detecting carbon enhancement in hyper metal-poor ([Fe/H] $\sim$ -5)

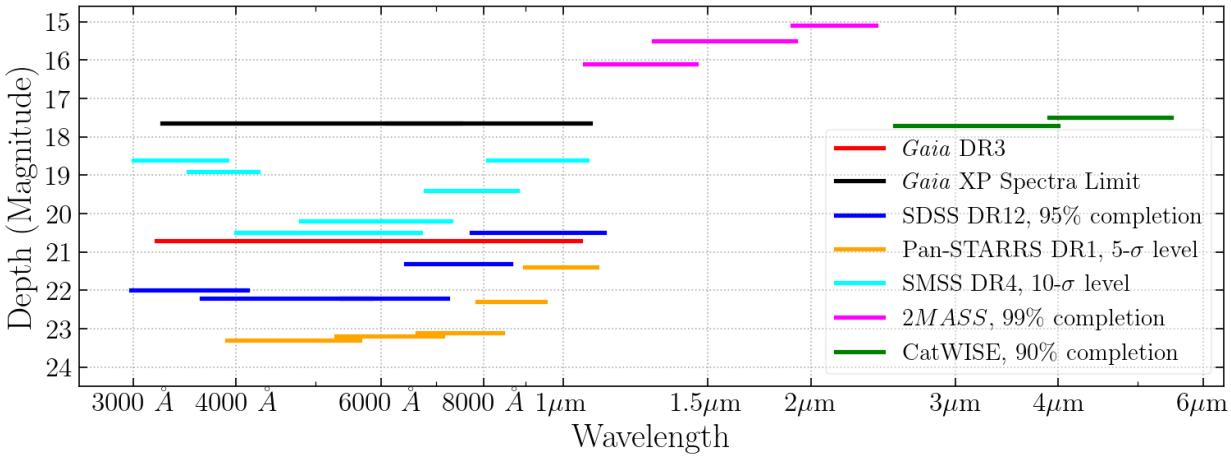

Figure 2.1: The coverage of the different photometric surveys included in the pre-training dataset. The bars represent the effective wavelength range covered by each band in a survey, coloured by survey. The quoted depth is given as the catalogue limit for *Gaia* DR3, and at the reported completeness and confidence levels for each survey in their respective magnitude systems or in AB magnitudes. The magnitudes are not converted to a universal magnitude system for the MSA, thus are not changed for this plot.

stars. The same conclusion was found for detecting [α/Fe] from the XP spectra, with cool ($T_{\rm eff} \lesssim 5000$ K), high-metallicity ([Fe/H] = 0) stars having the lowest errors.

Using real data, Fallows & Sanders (2024) found that their neural network model predictions correlated with features in high-resolution APOGEE spectra corresponding to Fe, CN, and C_2. This implies that the low-resolution XP spectra are sensitive to the abundances of C, N, and Fe, with O-abundances also being indirectly traced via the CNO cycle. They also find that [α/Fe] is not directly measured but rather inferred by the model through correlations with other elements like Mg and Si. From the abundances that can be reliably extracted, stellar ages can be inferred for certain populations, such as red giants, using abundance ratios like [C/N] as a chemical clock (e.g., Masseron & Gilmore, 2015). These are the key astrophysical features the pre-trained model aims to extract from the data.

2.2 Photometric Datasets

To supplement the low-resolution XP spectra, I incorporate broad-band photometry from several all-sky surveys. This ancillary data serves two purposes. First, it extends the wavelength coverage from the ultraviolet to the mid-infrared, allowing for a more complete construction of each star's spectral energy distribution. Second, in cases where XP spectra are unavailable or of low quality, this photometry in the same spectral range as the XP spectra

provides constraints for estimating stellar properties.

The reference catalogue, *Gaia* DR3, contains photometric magnitudes for $\sim$ 1.5 billion stars in 3 bands (Gaia Collaboration et al., 2023), G (330 nm - 1050 nm), G_{BP} (330 nm - 680 nm), and G_{RP} (620 nm - 1050 nm), that exist for all stars with XP spectra and to a depth of $\sim$ 21 mag in G. But, the XP spectra released in DR3 are limited at $<$ 17.85 mag in the *Gaia* G-band, except for a catalogue of spectroscopically interesting sources such as approximately 17,000 galaxies, $\sim 100,000$ quasars, and candidate white and ultra-cool dwarfs (De Angeli et al., 2023). This catalogue contains a total of 219,197,643 sources.

In addition to photometry, I included the *Gaia* astrometric measurements (parallax and proper motions) as input features, to aid in resolving the interstellar dust that obscures stars and dims their apparent magnitudes. Absolute magnitudes, which are important in analytic stellar parameter derivation, may be better inferred by the model by injecting the positional information. The inclusion of astrometric data To construct the pre-training set, I began with all 219,197,643 sources with published XP spectra and applied minimal quality cuts: A source was required to have a parallax measurement, not be flagged as a potential galactic or QSO candidate in the *Gaia* catalogues, and have the `has_xp_continuous` flag set to true. This results in a primary dataset of 218,131,884 sources with XP coefficients. The photometric data from other surveys represents a subset of these sources.

2.2.1 Infrared Dataets

Infrared photometry is particularly valuable as longer wavelength observations are less susceptible to the effects of interstellar extinction (reddening) by dust. The dust grains are more permeable to longer wavelengths, and the inclusion of these bands is to incite the neural network to implicitly learn the reddening of a source from training with spectral features and both infrared and optical magnitudes. The Two Micron All Sky Survey ($2MASS$, Skrutskie et al., 2006) includes coverage from both the northern and southern hemispheres. The survey consists of 3 photometric bands being the J (1080 nm - 1410 nm), H (1480 nm - 1820 nm), and the K_s band (1950 nm - 2360 nm), encompassing over 470 million sources in the point source catalogue. The depth across the filters is inconsistent, resulting in some underpopulated magnitudes in the dataset, with the depths of the bands being $J < 16.1$, $H < 15.5$, and $K_s < 15.1$ mag at the 99% completeness level. The cross match of this catalogue had been previously computed by the *Gaia* Data Processing and Analysis Consortium (DPAC), and the only selection cut I applied was the removal of the source if it exists in the $2MASS$ extended source catalogue.

To further improve the discrimination of reddening through combination of infrared and optical magnitudes, the CatWISE catalogue (Eisenhardt et al., 2020; Marocco et al., 2021) has been included in the feature list. CatWISE comprises two infrared bands being the 3.4 μm $W1$ band (2.8 μm - 3.9 μm) and the 4.6 μm $W2$ band (3.9 μm - 5.3 μm) which reach depths of 17.7 mag and 17.5 mag at the 90% completeness level, respectively. Being another all sky survey and expanding further into the infrared wavelengths adds insurance to the rationale for the inclusion of $2MASS$. For this catalogue, the cross match between the XP spectra and CatWISE resulted in 174,922,161 sources in the pre-training dataset, while cross match with $2MASS$ resulted in 207,409,237 sources. The wavelength coverage and depth of all the surveys is shown in Figure 2.1.

Extinction and dust obscuration can greatly influence spectral feature predictions, particularly in the plane of the Milky Way. To train the model on handling of dust obscuration, the full line-of-sight maps of reddening computed by Schlegel et al. (1998) as provided by the `dustmaps` package (Green, 2018) are added as features to the dataset. This particular dust map was chosen as it is full sky, independent of distance, and does not assume a galactic structure prior, giving reddening values for every star in *Gaia*.

2.2.2 Optical Datasets

As most observations are performed from ground-based observatories for a myriad of reasons, the coverage of the sky changes wildly between catalogues. Therefore, a mix of surveys from both the northern and the southern hemispheres was gathered to complete the full dataset. The Sloan Digital Sky Survey (SDSS, Alam et al., 2015) when it was first commissioned was the largest survey of its kind in astronomy. The northern sky coverage consists of 5 bands spanning the optical range of u, g, r, i, and z. The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS, Chambers et al., 2016), another northern sky survey, provides a larger footprint covering a range of ∼30,000 square degrees compared to SDSS with 14,555 unique square degrees. The Pan-STARRS DR1 catalogue contains coverage in 5 photometric bands: PS1 g, r, i, z, and y, of which many share the same name and wavelength coverages as the SDSS filter transmissions, but differ slightly at the edges of the photometric filters. The comparison of all the different photometric bands with respect to their optical depths and overlap is shown in Figure 2.1. The SkyMapper Southern Survey (SMSS, Wolf et al., 2018; Onken et al., 2019) fourth data release (Onken et al., 2024) is also included in the dataset. Representing the sole source of photometry used for the entire southern sky, SMSS fills a lot of the missing g, r, i, and z bands for stars observed in *Gaia* but are not

visible from the northern hemisphere. SMSS also has unique u and v filters, which together span a typical range for the u band, that are separated around the Balmer jump in stellar spectra.

Some minor cuts were made on the photometric datasets to ensure mostly stars were being matched with *Gaia* DR3, while keeping the largest amount of data and diversity possible from the available catalogues. The SDSS photometry catalogue provides flags for each photometric band, for which those measurements were kept if the bit flags for saturation, bad pixels, edge detections, and "not a star" were all set to 0, along with dropping any measurements that were returned as -9999 in the table. The Pan-STARRS DR1 catalogue was filtered via the `obj_info_flag` which removed sources that were saturated, and had cosmic rays or artifacts. The SMSS DR4 dataset had more flags with constraints recommended by the authors (Onken et al., 2024). The stars were kept given the flag `class_star` > 0.9 when the r-band magnitude was < 19 and `class_star` > 0.5 when $r \geq 19$, with the restriction on the full catalogue of `flags` < 4, `nimaflags` < 5, and `gaia_dr3_dist1` < 2", taking the closest cross match with *Gaia* DR3. The breakdown of the cross-matches with the *Gaia* XP spectra are given in Table 2.1.

Table 2.1: The break down of the stars included in the pre-training data set.

Sources and Conditions	$N_{\rm stars}$	Ref.
Sources with *Gaia* XP coefficients	219,197,643	1
Stars With *Gaia* parallaxes	218,131,884	
With *2MASS* magnitudes	207,409,237	2
With CatWISE magnitudes	174,922,161	3
With Sky-Mapper DR4 magnitudes	141,310,053	4
With Pan-STARRS1 DR2 magnitudes	129,043,651	5
With SDSS DR9 magnitudes	20,770,505	6
Total stars in pre-training set	218,131,884	

References: (1) De Angeli et al. (2023); (2) Skrutskie et al. (2006); (3) Marocco et al. (2021); (4) Onken et al. (2024); (5) Chambers et al. (2016); (6) Alam et al. (2015).

2.3 Stellar Parameters Datasets for Training and Validation

To train and validate my method for deriving stellar parameters, I require a "ground-truth", or labelled dataset. This consists of stars with parameters derived often from high-resolution

Table 2.2: The sources of the spectroscopically derived labels used in the fine-tuning data set, broken down by number of individual stellar feature by catalogue. The totals for each stellar feature are based off the number of stars with XP coefficients.

Catalogue	N_{stars} with XP	T_{eff}	$\log g$	Fe/H	$[\alpha/\text{Fe}]$	Ref.
GALAH	851,894	851,894	851,894	851,894	850,721	1
APOGEE	493,075	493,075	493,075	467,223	467,047	2
RAVE	396,442	396,442	396,442	396,442	388,852	3
VMP stars	291	291	291	291	-	4
Total unique rows:	1,741,702					
Union of stars	1,645,698					

References: (1) Buder et al. (2024); (2) Majewski et al. (2017); (3) Steinmetz et al. (2020); (4) Li et al. (2022).

spectroscopic observations, which are generally considered the gold standard in abundance analysis. This process, often referred to as fine-tuning in a machine learning context, calibrates the model to predict physical parameters from the low-resolution XP spectra and ancillary photometry.

2.3.1 Spectroscopic Labels

The core of the training and validation sample is built from several major public high-resolution spectroscopic surveys. These provide precise measurements of T_{eff}, $\log g$, [Fe/H], and $[\alpha/\text{Fe}]$. To ensure the model performs well on chemically peculiar or rare objects, I augment this core sample with a smaller, specialized catalogue, that is a collection of very metal-poor stars. A summary of all spectroscopic datasets used for training and validation is provided in Table 2.2.

Stars from SDSS DR17 with APOGEE spectra (Abdurro'uf et al., 2022) were selected for the spectral properties of T_{eff}, $\log g$, [Fe/H], and $[\alpha/\text{Fe}]$. The largest restriction on the inclusion of a given star with APOGEE spectra was the existence of XP spectra for the source. This led to stars being excluded if beyond the limiting magnitude of 17.85 in the *Gaia* G-band, and if missing XP continuous spectra. Following the procedure of Andrae et al. (2023), 291 stars from Li et al. (2022) were included to push beyond the [Fe/H] ~ -2.5 dex floor that exists for the APOGEE spectral pipeline and to provide additional measurements for T_{eff}, $\log g$, and [Fe/H].

The fourth data release of the Galactic Archaeology with HERMES (GALAH De Silva et al., 2015) dataset was also chosen for the fine-tuning dataset, injecting 851,894 more

sources into the fine-tuning catalogue. This larger dataset provides measurements of $T_{\rm eff}$, log g, [Fe/H], and [α/Fe] as APOGEE does, with a further inclusion of ages for 849,596 stars determined through isochrone fitting. Stars were selected using the criteria that the first two validation flags corresponding to having no problems and that the value is not an upper limit were set to 0 for the detection. The final made when curating all three datasets were to ensure that the *Gaia* DR3 source ID was not duplicated, where it and its duplicates were removed from the dataset. The fine-tuning dataset is summarized in Table 2.2.

In addition to the fine-tuning dataset curated for the MSA, its ability to predict in different bases, being to other spectroscopic datasets, is tested (Section 5.2). The Rave DR6 (Steinmetz et al., 2020) dataset was chosen for this purpose, having measurements for each of the five stellar parameters being predicted. Similarly to the ages for GALAH DR4, the RAVE DR6 ages used for testing its applicability to other were determined via Bayesian isochrone fitting using the BDASP pipeline (McMillan et al., 2018). This pipeline also output values for $T_{\rm eff}$, log g, and [M/H], with [α/M] coming from their CNN trained on stellar spectra.

Included in the rationale for the inclusion of APOGEE in the fine-tuning dataset is the need for many stellar ages for the prediction algorithm to train with. APOGEE contains a large amount of red giant stars, on which many of the methods that exist for determining stellar ages for individual stars rely, such as asteroseismology or stellar models using the [C/N] abundance. A large amount of age labels come from the `astroNN` models developed by Mackereth et al. (2019) and Leung et al. (2023), which have spectroscopic age predictions derived from APOGEE spectra and the *Kepler* asteroseismic dataset. The Leung et al. (2023) ages were cut according to the authors' recommendations: having an error in the age estimate $< 40\%$, and having both `STARFLAG` and `ASPCAPFLAG` set to 0. This dataset was made up entirely of red giants, for a total of 39,662 stars. The ages from Mackereth et al. (2019) filled in the rest of the rows for apogee stars, accounting for an additional 481,036, stars. Shown in Table 2.3, the total number of stellar ages across the datasets used as labels are shown, totalling 1,727,074 stellar age labels in the fine-tuning dataset.

Table 2.3: The different catalogues contributing to the ages included in the fine-tuning data set. The number of stars reflects those from the data set with a *Gaia* source ID and with XP spectra coefficients.

Catalogue	N_{stars}	References
GALAH	849,596	Buder et al. (2024)
RAVE	396,442	Steinmetz et al. (2020)
AstroNN-1	441,374	Mackereth et al. (2019)
AstroNN-2	39,662	Leung et al. (2023)
Total unique ages:	1,727,074	
Total unique stars	1,631,423	

Chapter 3

Constructing the Masked Stellar Autoencoder

The MSA is a deep learning model specifically designed to leverage high-dimensional astronomical data and sparsely available labels for applications in Galactic Archaeology. The construction of the model is intrinsically linked to its two-stage training methodology. This process begins with self-supervised pre-training, which builds a foundational understanding of the complex relationships within the data, followed by a fine-tuning stage that adapts the model for specific predictive tasks. This section details the deep learning paradigms that motivate the architecture of the model, a masked reconstructive residual autoencoder. This design enables the creation of powerful and informative embeddings from which stellar parameters can be accurately derived, even from highly entangled and correlated datasets that combine photometric and low-resolution spectroscopic information. The goal is to produce a fairly general model for stars in *Gaia* DR3 that can be tuned for a variety of astronomical applications.

3.1 Self-Supervised and Reconstruction Learning

Supervised learning methods learn the connection between labels and features through a series of interconnected layers and neurons and while powerful, depend fundamentally on the availability of large, completely and accurately labelled datasets. In astronomy, this dependency presents several challenges. The sheer volume of modern surveys makes manual labelling infeasible, while automated pipelines for deriving stellar atmospheric parameters can introduce systematic biases. Because these pipelines are often optimized for the bulk of the stellar population, they can under-perform on, or even mischaracterize, rare objects, such as very metal-poor (VMP) or carbon-enhanced metal-poor (CEMP) stars. Furthermore, these models often struggle with out-of-distribution (OOD) objects and exhibit poor generalization when applied to data from different instruments or surveys, which may have

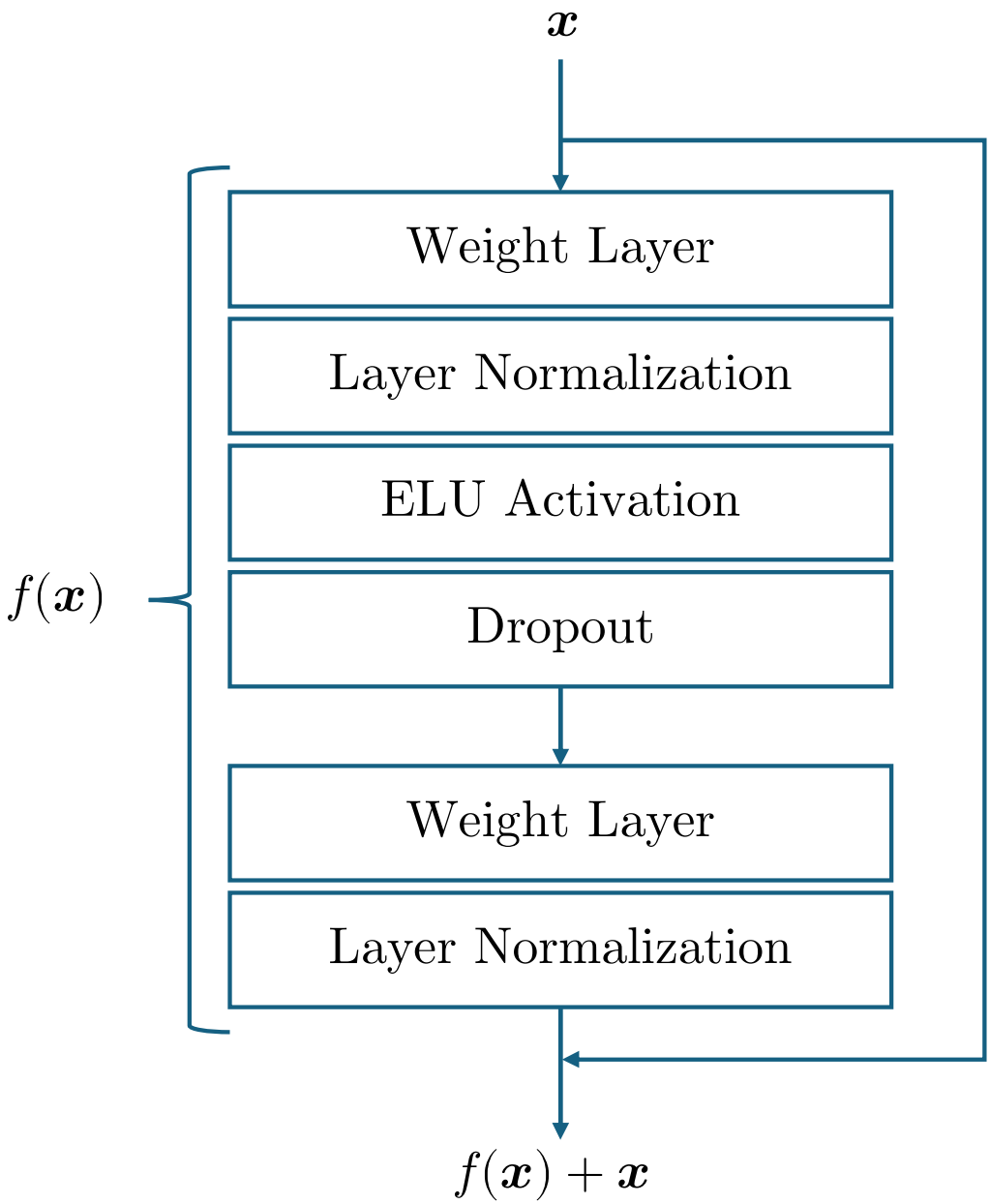

Figure 3.1: The conceptual formation of a residual block and skip-connection, with the inclusion of the specific layers used in the MSA. The input x is passed downwards following the path of both arrows, such that after passing through the residual block $f(x)$, the value passed to the next residual block is $f(x) + x$.

unique reduction and analysis pipelines (Allende Prieto, 2016). As an example, APOGEE is often used for training astronomical ML models, but may result in predictions biased against

metal-poor stars from other spectroscopic datasets. In other words, models trained on one survey are often not directly transferable to another, limiting their utility.

To circumvent these aforementioned limitations, the MSA adopts a self-supervised learning (SSL) paradigm. SSL enables a model to learn meaningful representations from the intrinsic structure of the data itself, without reliance on external labels. This is typically achieved through a two-stage process:

- Pre-training - the model is first trained on a large, unlabelled or partially labelled dataset using a pretext task. These tasks are designed to help the model understand correlations and relationships of different features with respect to the data itself. For instance, the model might be tasked with predicting a deliberately corrupted or masked portion of its own input. By solving this pretext task, the model develops a rich, generalized feature representation that captures the essential physics encoded in the data.

- Fine-Tuning - Once pre-trained, the representations learned by the model are adapted for specific downstream tasks, such as predicting stellar parameters. This stage uses a much smaller, labelled dataset to tune a simple prediction head attached to the pre-trained encoder, leveraging the powerful foundation built in the first stage.

The choice of pretext task is critical. Contrastive methods (e.g., Chen et al. 2020; Radford et al. 2021) utilise different realizations of the same input from applying distinct augmentations, such as adding noise or resampling within the errors of the data, and enforcing that the algorithm predicts that the realizations are the same despite the changes made. Although contrastive learning can yield very strong representations on smaller datasets, we employ a reconstructive method, which can learn more general representations at massive scales by forcing the model to understand the data internal structure, not just what makes examples different. Rather than the emphasis on the robustness to noise of contrastive methods, the reconstructive methodology involves either masking or dimensionality reduction to train the model in inferring the missing or compressed data. For reconstructive tasks, portions of the input data, being tiles for images or words for language models, are masked upon input to the encoder and filled-in using the decoder. The loss can then be calculated using the unmasked input data, forcing the latent space to learn the most useful information from the partial data on input.

3.1.1 Autoencoders

Autoencoders are types of neural networks suited for unsupervised learning. They are composed of an encoder, a latent space, and a decoder, which often function to compress and decompress the input data. Compressing and decompressing the features results in the learning of embeddings which extract the pertinent information to reconstruct or replicate the input feature vector, which have applications in both feature learning for sparse datasets and generative tasks (Bengio et al., 2013).

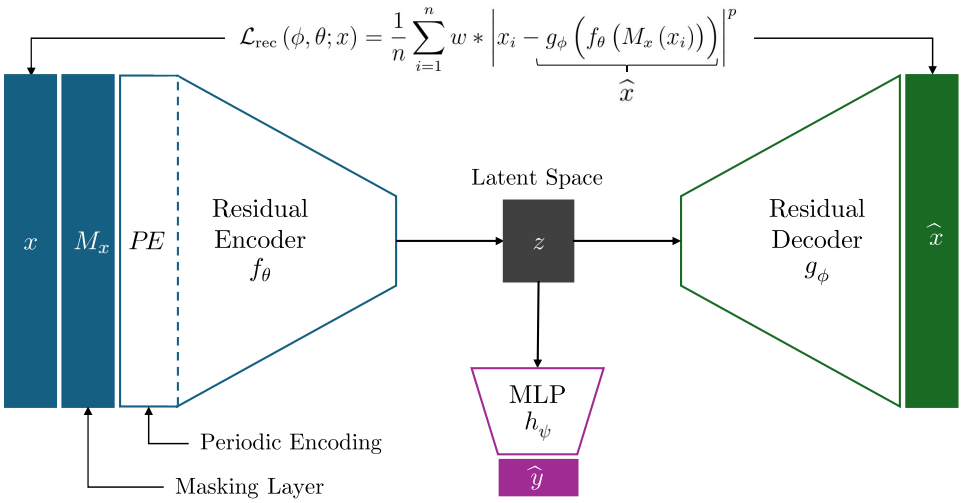

Figure 3.2: A high-level model of the MSA with both the pre-training and fine-tuning components shown. The features are represented by x beginning on the left of the diagram. The pre-training and reconstructive penalty term when fine-tuning is the horizontal path through the latent space, with the latent vector denoted as z, while the regression to stellar label predictions, denoted by $\hat{y}$, is the path downward from z through the MLP. An example of the reconstructive loss term is shown connecting the inputs x to the outputs of the model $\hat{x}$

For our reconstructive pretext task, we employ a specific variant known as a masked autoencoder. This approach builds upon the concept of denoising autoencoders from Vincent et al. (2008). In our setup, a portion of the input features are deliberately hidden or "masked" before being passed to the encoder. The network is then tasked with reconstructing the complete, original data, including the masked values. This forces the model to learn the inherent correlations and dependencies within the data to accurately infer the missing information.

The dimensionality of the latent space is a critical hyperparameter that must be carefully tuned. It needs to be large enough to retain the information necessary for accurate recon-

struction, but small enough to prevent the model from simply learning an identity function and to encourage generalization. After the self-supervised pre-training phase is complete, the decoder is typically discarded. In our case, the decoder is kept and we fine-tune using multi-task regression and reconstruction to help with generalisation. For downstream tasks, in our case prediction of stellar properties, a separate prediction head, such as a multi-layer perceptron (MLP), is attached to the latent space representation and fine-tuned on a smaller, labelled dataset to predict specific target values.

3.1.2 Pre-training Architecture of the Model

The core of the MSA is a deep autoencoder with a symmetric encoder-decoder structure. The architecture is constructed from a series of residual blocks, inspired by Residual Neural Networks (ResNets) from computer vision (He et al., 2016). The framework of these networks uses a modular block structure, for which the input is passed through multiple weight layers. The output of the residual block is then summed with the original input, via a skip connection, which has been shown to combat the issues of vanishing gradients and high complexity in the data and network, improving both efficiency and accuracy (Borawar & Kaur, 2023). The MSA encoder consists of 8 residual blocks of decreasing dimensionality, compressing the input into a latent space representation, while the decoder symmetrically mirrors this architecture to reconstruct the original features.

Each residual block within the MSA (see Figure 3.1) is a sequence of layers designed to learn complex data transformations. It includes two linear (fully-connected) layers, with Layer Normalization (Lei Ba et al., 2016) applied after each layer to stabilize training. An Exponential Linear Unit (ELU) serves as the non-linear activation function, and a dropout layer is included for regularization.

As outlined in Figure 3.2, input features undergo a two-step preprocessing procedure before being fed to the encoder:

1. Feature Masking (`mask`): A multi-level masking strategy is applied. First, any features that are genuinely missing from the source catalogues are padded with a placeholder value (-9999). Second, for the self-supervised pretext task, a fraction of the existing data is strategically masked. A core set of features comprising the Gaia G, BP, RP, and WISE W1, W2 magnitudes are never masked, preserving the basic SED. The 110 XP coefficients are masked as a single group with a 90% probability to simulate the magnitude limit of the XP spectra sample. All other ancillary features (photometry, astrometry) are masked individually with a 60% probability to reflect the heterogeneous

sky coverage of their respective surveys.

2. Periodic Encoding (`periodic`): Following the masking step, all features are transformed using a periodic encoding layer. This technique has been shown to significantly boost the performance of MLP-based architectures on tabular data, making them competitive with gradient-boosted decision trees and transformers (Gorishniy et al., 2022). Each feature is mapped to a higher-dimensional space using a series of sine and cosine functions with trainable frequencies, as given by:

$$\begin{aligned} f_i(x) &= \texttt{periodic}(x) \\ &= \texttt{concat}[\sin(v), \cos(v)], v = [2\varpi c_1 x, ..., 2\varpi c_k x] \end{aligned}$$

where c_i represents the trainable frequency parameters. This allows the model to more easily learn periodic and non-linear relationships in the data.

The pre-training objective is to minimize the reconstruction error between the decoder's output ($\hat{x}$) and the original, unmasked input (x). The general form of the loss $\mathcal{L}_{\text{rec}}$ is given by:

$$\mathcal{L}_{\text{rec}} = \frac{1}{n}\sum_{i=1}^{n} w_i * \left| x_i - g_\phi\Big(f_\theta\big(\texttt{mask}(x_i)\big)\Big)\right|^p \tag{3.1}$$

where f_θ and g_ϕ are the encoder and decoder, p is the exponent (1 for L1 loss, 2 for L2 loss), and w represents optional weights. Through hyperparameter tuning, we found that an unweighted Mean Absolute Error (MAE, or L1) loss performed best, simplifying the objective function to $p = 1$, and $w_i = 1$.

3.1.3 Fine-Tuning Scheme for the MSA

Once the self-supervised pre-training is complete, the model is adapted for specific downstream prediction tasks through a process of fine-tuning. For this stage, a 5-layer MLP, serving as a prediction head, is attached to the latent space output of the pre-trained encoder, as illustrated in Figure 3.2. Unlike some transfer learning approaches, the weights of the encoder are not frozen during this phase; instead, the entire network is trained end-to-end, from the input layers of the encoder to the output of the prediction head.

Our fine-tuning employs a multi-task, multi-target learning strategy. We simultaneously predict all six target labels ($\mathbf{y} = \{T_{\text{eff}}, \log\, g, [\text{Fe/H}], [\alpha/\text{Fe}], \tau_*, \varpi\}$), as these physical parameters are often degenerate and co-dependent. Training them together allows the model

to leverage these intrinsic correlations, leading to more robust and physically consistent predictions.

To provide robust uncertainty estimates, the prediction head is trained using quantile regression. Instead of predicting a single mean value, the model outputs three quantiles for each label: the 16th, 50th (median), and 84th percentiles. This is achieved by minimizing the quantile loss, or pinball loss, function:

$$\mathcal{L}_{\text{pred}} = \frac{1}{\sum_i w_i} \sum_{i=1}^{N_L} w_i \sum_{q=1}^{Q} \sum_{j=1}^{M} \max\left(\tau_q\big(y_{ij} - h_\psi^{jq}(x_i)\big), (\tau_q - 1)\big(y_{ij} - h_\psi^{jq}(x_i)\big)\right) \tag{3.2}$$

where N_L is the number of labelled samples, y_{ij} is the true value for the j-th label of the i-th sample, and $h_\psi^{jq}(x_i)$ is the model prediction via the prediction head h_ψ for the corresponding τ_q quantile.

To prevent the model from forgetting the general representations learned during pre-training (a phenomenon known as "catastrophic forgetting"), we include the original reconstruction task as part of the fine-tuning objective. This process effectively helps for imbalanced data (Liu et al., 2021). The total loss function is a weighted sum of the prediction loss and the reconstruction loss, ending in an optimization for the model to be:

$$\theta, \phi, \psi = \underset{\theta,\phi,\psi}{\operatorname{argmin}}\, \lambda_1 \mathcal{L}_{\text{pred}}(\theta, \phi) + \lambda_2 \mathcal{L}_{\text{rec}}(\psi) \tag{3.3}$$

where $\mathcal{L}_{\text{pred}}$ and $\mathcal{L}_{\text{rec}}$ are the prediction (Eq. 3.2) and reconstruction (Eq. 3.1) losses, respectively. Based on hyperparameter tuning, we set the weights to $\lambda_1 = 0.8$ and $\lambda_2 = 0.2$, prioritizing the accuracy of the downstream prediction task while still regularizing the model through reconstruction.

Finally, to further refine the predictions and quantify the uncertainty of the model itself, we employ deep ensemble learning (Lakshminarayanan et al., 2017). The final stellar parameters are derived by averaging the median predictions from 20 models, each fine-tuned with a different random seed for weight initialization and data augmentation. The final uncertainty incorporates both the quantile range from individual models and the standard deviation across the ensemble.

Chapter 4

Training Results

The results from the two-stage training process are presented below, with the reconstructive pretext task discussed in Section 4.1, and the results from the stellar parameter prediction described in Section 4.2.

4.1 Pre-Training Results

The model was pre-trained for 80 epochs on the tasks of XP coefficients and photometric magnitudes reconstruction. The residuals for the reconstruction of the first 3 coefficients in BP and RP are shown in Figure 4.1. The large dynamic range of the coefficients (varies by $\sim 10^2 - 10^6$, as shown in Figure 4.1) required an appropriate scaling to help gradient-based optimization of the neural network parameters. A median and interquartile range scaling method was used for the magnitudes and XP coefficients, which minimized the compression of the lower magnitude regime relative to other scaling methods while preserving the distribution of the data. For an accurate representation of the model's abilities, two million stars from the pre-train validation dataset were used in creating Figure 4.1. The reconstructions for the first coefficients in BP and RP show that the model performs well at reconstructing the 0th order Hermite polynomial coefficient. However, the reconstructions for the higher order polynomials show an extreme variation at low magnitudes. This may result from either their massive dynamic range, or the difficulty in extracting information from the coefficients, which themselves are a compressed representation of the spectra. Despite the large spread, the predictions are not biased, showing an even dispersion about 0.

The reconstructions of the photometric magnitudes in the pre-training dataset are shown in Figure 4.2 as a function of their magnitude. The MSA is shown to be accurately reconstructing the majority of the magnitudes from the optical surveys, with only slight errors in reconstruction towards the saturation and faint limits of the photometric surveys. For

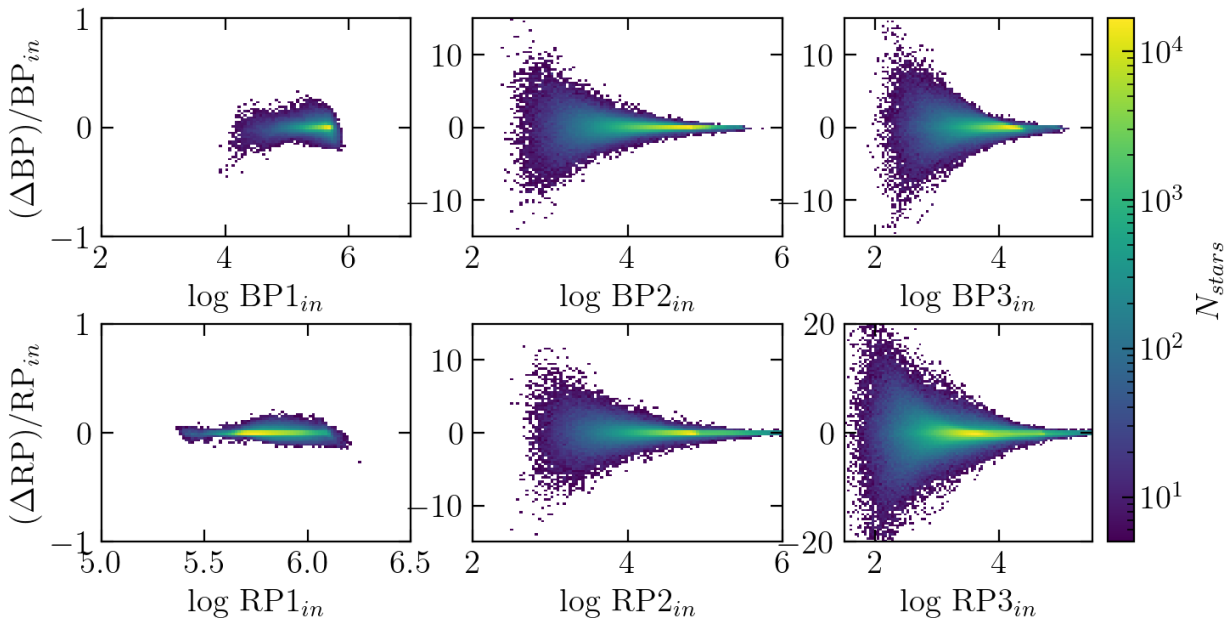

Figure 4.1: Shown are the residuals for the reconstructions for the first three coefficients in both BP and RP by the pre-trained MSA. The x-axis of each plot is the logarithm of the observed coefficient by *Gaia* (BP_{in}), while the y-axis corresponds to the difference between the predicted and observed coefficient, divided by the observed coefficient. The BP1 and RP1 denote the first BP and RP coefficient of fifty-five, with increasing number denoting increasing Hermite polynomial order. The 2D histogram of each plot shares a common colour bar, plotted on the left.

the brighter stars, a bias is shown to exist for $2MASS$ magnitudes, but was observed to be attenuating with training epochs. This bias was investigated further and is shown due to be a result of the poor fits for a minority of stars contaminating the dataset. These stars likely either have poor cross matches with *Gaia* DR3, being $2MASS$ magnitudes for extra-galactic sources rather than stars, or have poor quality flags, which were not filtered from the dataset to include more information (See Figure A.1 for the reconstructions of the $2MASS$ magnitudes).

4.2 Fine-Tuning Results

The prediction head for stellar parameters as outlined in Section 3.1.3 was fine-tuned for 100 epochs, still with masked features and maintaining the same masking ratio as for pre-training. The residuals of the regression targets $\mathbf{y}$ for the test dataset are shown in Figures 4.3-4.4. The performance of the model on several metrics for each label is shown in Table 4.1, separated

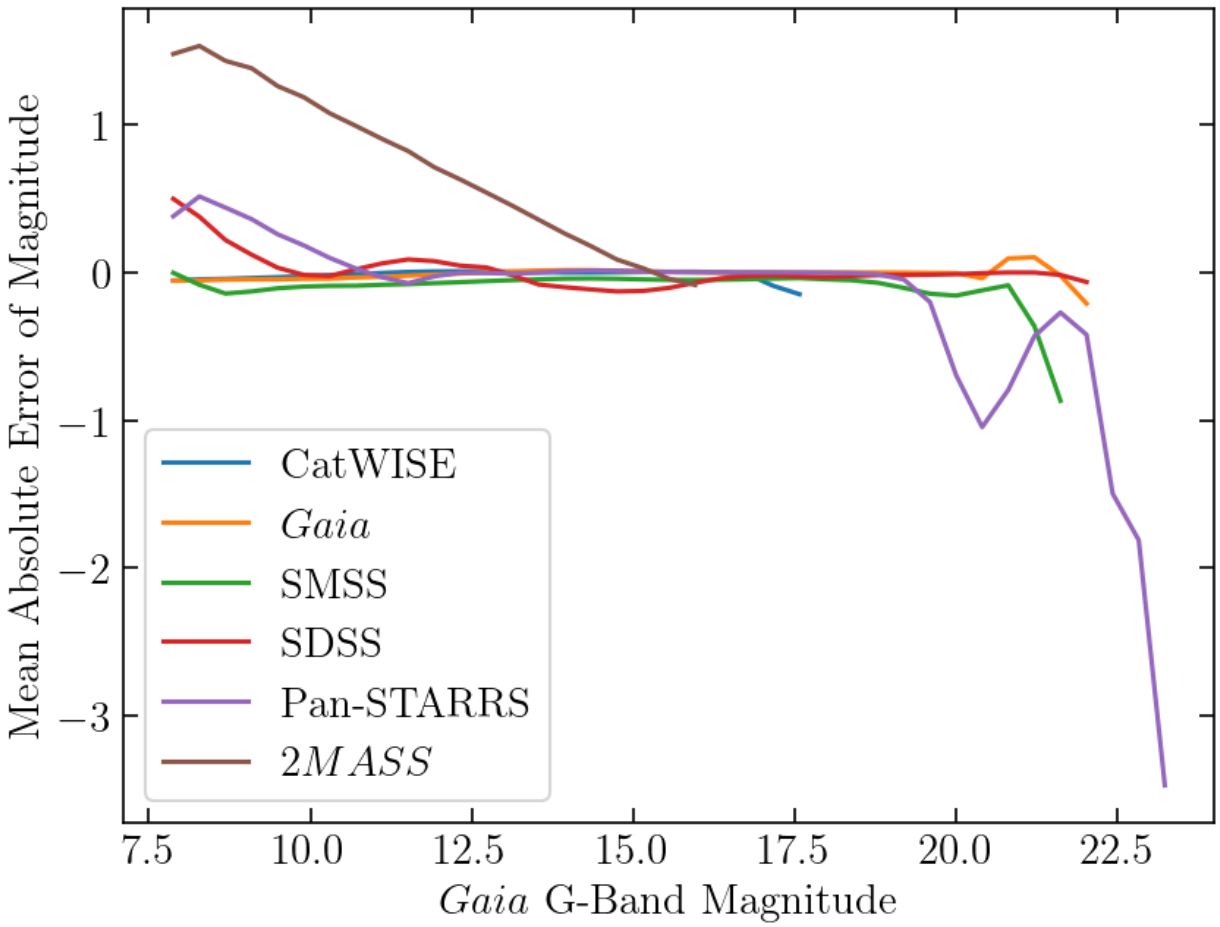

Figure 4.2: Shown are the residuals for the reconstructions for the masked magnitudes per bin per survey by the pre-trained MSA. All surveys are plotted as a function of the star's G-band magnitude and average by bin, which was chosen to be 0.4 mag in width. The magnitudes reach beyond the limiting magnitude of the XP spectra of 17.65 due to the special stellar types included (Section 2.2).

by predictions based on including and excluding the XP coefficients in the features passed to the MSA and prediction head (MSA+P). The mean absolute errors for the six stellar parameters were shown to be 92 K in $T_{\rm eff}$, 0.08 dex in log g, 0.09 dex in [Fe/H], 0.05 dex in [α/Fe], 1.3 Gyr in τ_*, and 0.04 in log ϖ.

To compare with the reported accuracies from Andrae et al. (2023) and Leung & Bovy (2024), the normalized median absolute deviations (NMADs) are calculated. This metric is an indicator of how tightly clustered the residuals are and is more robust to outliers than MAE or RMSE. The MSA shows NMADs of ($T_{\rm eff}$, log g, [Fe/H]) = (62.0 K, 0.058 dex, 0.076 dex), as presented in Table 4.1. The normalization constant (= 1.4826) is multiplied with

the MAD such that the NMAD is the standard deviation for a Gaussian distribution, so long that the data is distributed as a Gaussian. Thus, this value can be divided to produce MADs for comparison with the Andrae et al. (2023) reported values, giving in MADs of 41.8 K, 0.039 dex, and 0.051 dex for the MSA predicted $T_{\rm eff}$, $\log g$, and [Fe/H], respectively. The Andrae et al. (2023) XGBoost model achieved MADs of 24.8 K for $T_{\rm eff}$, 0.039 dex for $\log g$, and 0.044 dex for [Fe/H], with respect to their training data, while the Leung & Bovy (2024) transformer-based model report NMADs of 47.17 K in $T_{\rm eff}$, 0.11 dex in $\log g$, and 0.07 dex in [Fe/H]. The MSA+P is shown be competitive with the two models improving with respect to $\log g$, but having a larger NMAD in $T_{\rm eff}$. This may be expected from the residual plots in Figures 4.4-4.4, which show the issues of the model in regressing hot stars. This subset of stars is missing metallicity and [α/Fe] measurements and is often ignored by supervised algorithms using APOGEE labels due to incompatibility in training, which obscures potential limitations in the data from supervised algorithms. With respect to the XGBoost model, the authors may have under reported errors since the metrics were computed with the training set rather than the unseen test dataset.

Table 4.1: The performance of the MSA and prediction head on several key metrics, being the root mean square error of the residuals, the standard deviation, the mean absolute error, the normalized median absolute deviation, and the R^2 coefficient of determination. The subscript M denotes whether the XP spectra are fully masked before passing through the MSA, while if missing, reflects that the full XP coefficients have been fed to the model for predicting the given label and calculating the metric.

Full XP	RMSE	σ	MAE	NMAD	R^2
$T_{\rm eff}$ [K]	330.5	330.4	92.3	62.0	0.910
$\log g$	0.145	0.145	0.075	0.058	0.982
[Fe/H]	0.166	0.166	0.090	0.076	0.807
[α/Fe]	0.072	0.072	0.045	0.042	0.622
τ_* [Gyr]	1.997	1.997	1.297	1.175	0.639
$\log \varpi$	0.095	0.095	0.037	0.023	0.961
Masked XP	RMSE_M	σ_M	MAE_M	NMAD_M	${R^2}_M$
$T_{\rm eff}$ [K]	357.7	357.6	108.3	74.4	0.894
$\log g$	0.141	0.141	0.071	0.052	0.984
[Fe/H]	0.193	0.193	0.114	0.101	0.741
[α/Fe]	0.080	0.080	0.052	0.050	0.531
τ_* [Gyr]	2.107	2.106	1.398	1.317	0.599
$\log \varpi$	0.134	0.134	0.030	0.024	0.997

The [α/Fe] abundance residuals reflect the limitations expected of the informative content

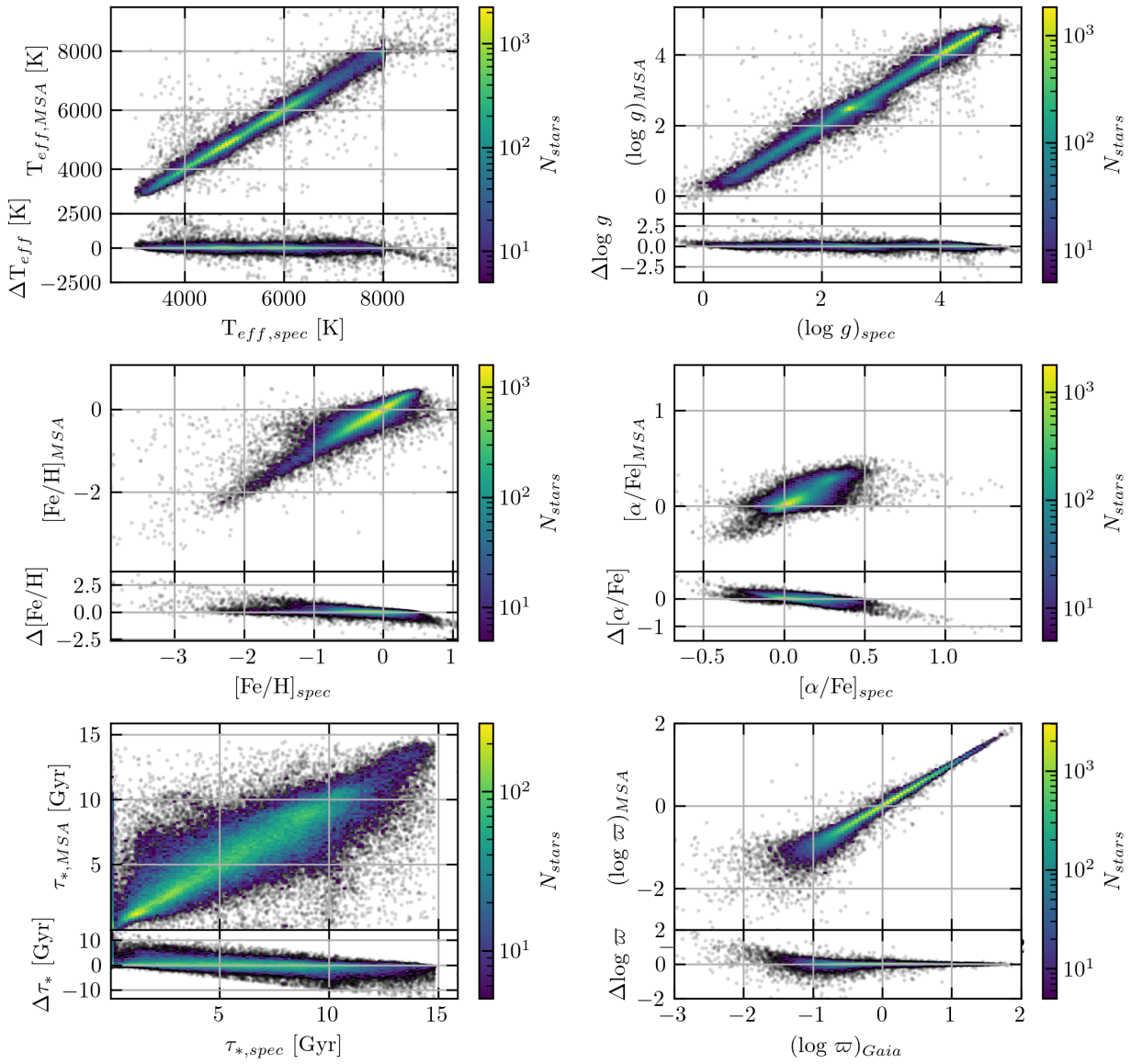

Figure 4.3: Test dataset residuals after fine-tuning the best fit model with ensemble learning. For each label, both the predictions (above) and the residuals (below) are plotted versus the spectroscopic/catalogue label using a 2 dimensional histogram, coloured by density. From left-to-right and top-to-bottom, the labels are $T_{\rm eff}$, log g, [Fe/H], [α/Fe], τ_*, and log ϖ. The axes are equal for the top figures for each label.

of the XP spectra, as primarily stars with solar [α/Fe] are fit accurately. The RMSE matches closer the mean absolute error, and the coefficient of determination shows a worse agreement between the data, as with the ages predicted by the MSA+P. However, shown in Figure 4.4 the model performs only marginally worse when having the entirety of the XP coefficients

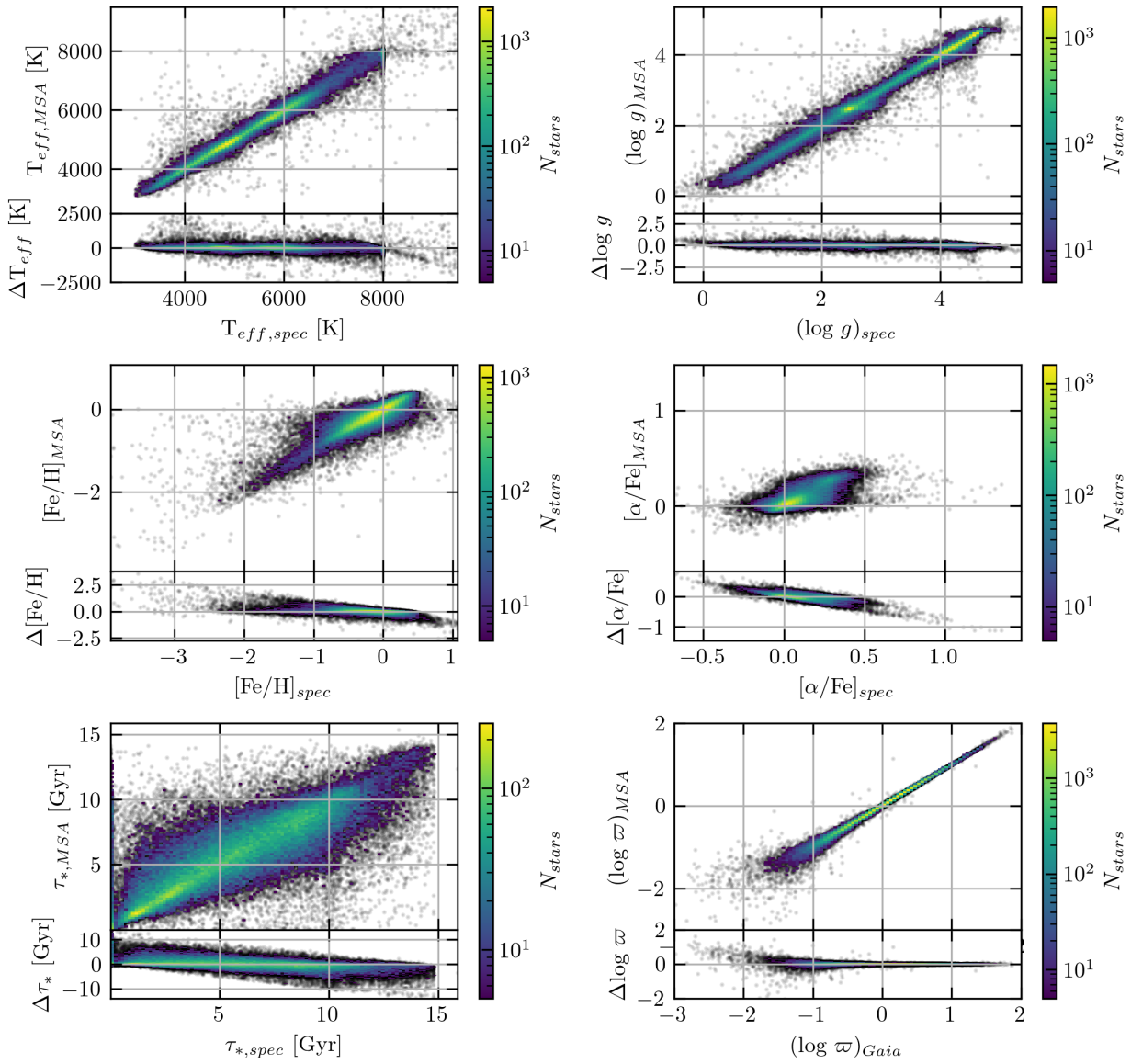

Figure 4.4: The same as Figure 4.3 with no spectroscopic information fed to the model as part of the features, keeping the same test set.

removed from the feature list, with a RMSE increase for [α/Fe] of 11.1% and for τ_* of only 5.5%, showing a robustness to scarsely populated data.

The structure of the latent space reveals the inferential ability of the model, shown in Figure 4.5 using the t-SNE prejection method (Van der Maaten & Hinton, 2008). Before fine-tuning the stellar labels, the latent space appears organized, but not discretized as the t-SNE is mostly continuous, showing regions where similar stars are being grouped together, with respect to τ_* in Figure 4.5. The fine-tuning appears to further segregate stars, but the

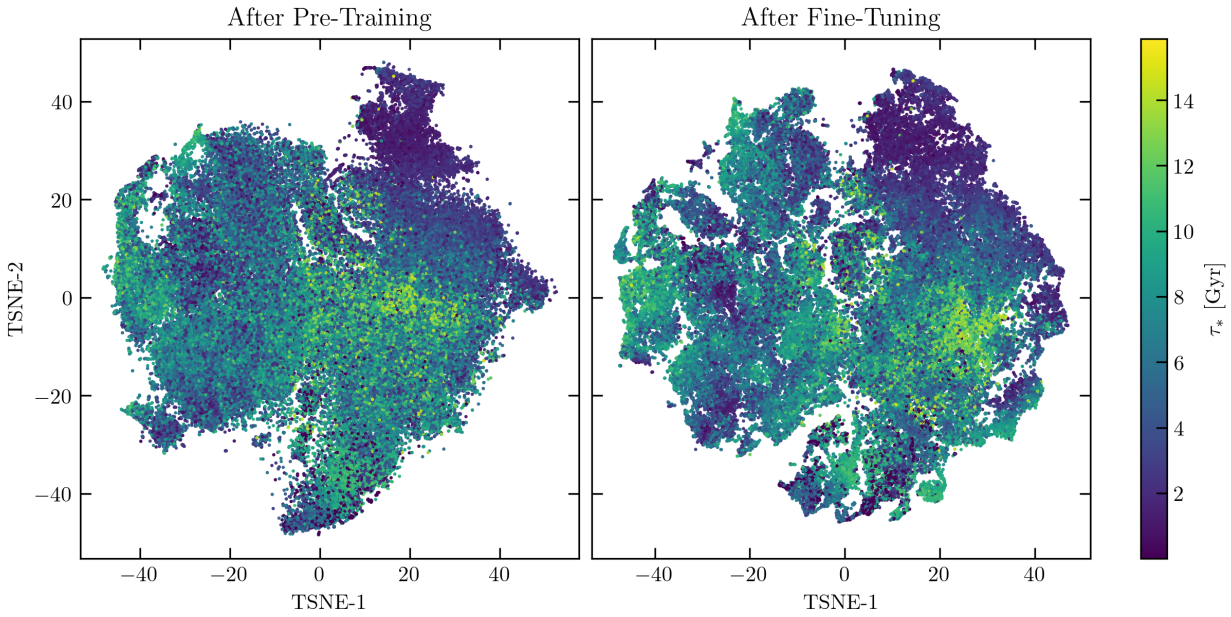

Figure 4.5: t-SNEs of the latent space of the MSA coloured by derived stellar age before (left) and after (right) fine-tuning the algorithm with the weights of the autoencoder unfrozen and adjusted to the fine-tuning dataset. The test set from the fine-tuning labels are used for plotting, totalling 88,788 stars with 87,840 age labels.

same overall trends are mostly undisturbed, as the embeddings are already informative for stellar label prediction.

Self-consistency in the predictions was measured by plotting the Kiel diagram of the predictions, the [α/Fe]-[Fe/H] chemistry plot, and the age-metallicity relationship (AMR) between the data, shown in Figure 4.6. The Kiel diagram (top row) is a stellar diagnostic plot of log g versus T_{eff} commonly used to trace stellar evolution with spectroscopic quantities derived from stellar atmospheres. The positions of stars on this diagram have physical importance, relating to their spectral type and stage of evolution. The predictions from the MSA+P show similar ordering in all three axes (including [Fe/H]), with a slight bias toward the mean of the labels, compressing the intrinsic scatter of the data. [α/Fe]-[Fe/H] is shown in the middle row, which traces chemical enrichment of stellar populations. It is a useful tool in separating distinct populations, such as the thin and thick disks of the Milky Way (Hayden et al., 2015). It can be shown from these plots that while the scatter in the predicted abundances gets compressed, the distinct sequences separating the low-α thin disk and high-α thick disk are preserved. This shows the ability of the model to separate the types of stars despite the low resolution of the input features with respect to metal absorption lines. The bottom row of Figure 4.6 shows the AMR of the stars in the fine-tuning dataset and the predictions from the model. This diagram is used to describe the chemical enrichment of a population of stars over time, tracing the star formation history of a group of stars. Scatter in the diagram reflects the dynamics of the environment, including radial migration and merger events. While the stars in the plot are not shown to have lower metallicities at older epochs as one may expect, the predictions do populate the diagram similarly to the labelled dataset.

To evaluate the quality of the predicted lower and upper quantiles of the stellar parameters, the z-score is computed. The distribution of the z-scores are expected to follow normal distribution asymptotically, with a perfect model M having $M \sim \mathcal{N}(0, 1)$. As the amount of data increases, the mean and standard deviation of the z-score can be used as a validation test for the errors. A distribution with a mean greater than 0 signifies predictions that are too low, with the inverse signifying a model that over-predicts. Additionally, a standard deviation greater than 1 signifies predicted uncertainty intervals that are too narrow, or an underconfident model, and smaller standard deviations are correlated with overconfidence. The z-score requires symmetric uncertainties, which are in conflict with the potentially asymmetric 16th and 84th percentiles predicted by the model. Upon inspection of the errors, the percentiles showed an approximately symmetric distribution in general, allowing for the simple average of the deviations to be used in place of the percentiles, with

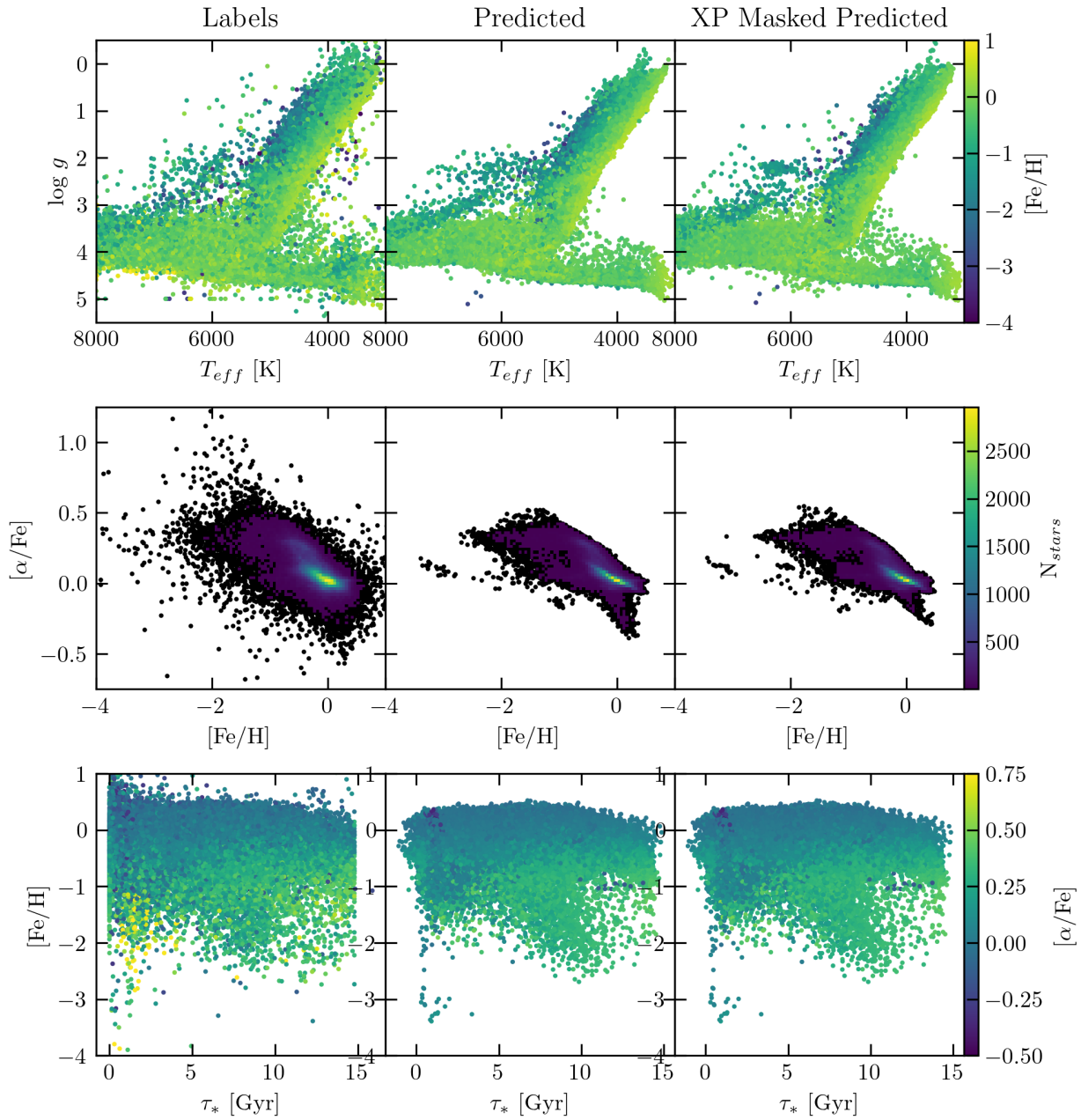

Figure 4.6: The self-consistency checks for the labels and predicted values of the test dataset. *Left*: The Kiel diagram displaying the stars in log g-$T_{\rm eff}$ space, coloured by their metallicities. *Center*: The [α/Fe]-[Fe/H] plot showing the populations of stars formed in environments enriched through separate sequences. *Right*: The age-metallicity relation coloured by [α/Fe], demonstrating the grouping of stars according to age and metallicity, important for star formation histories.

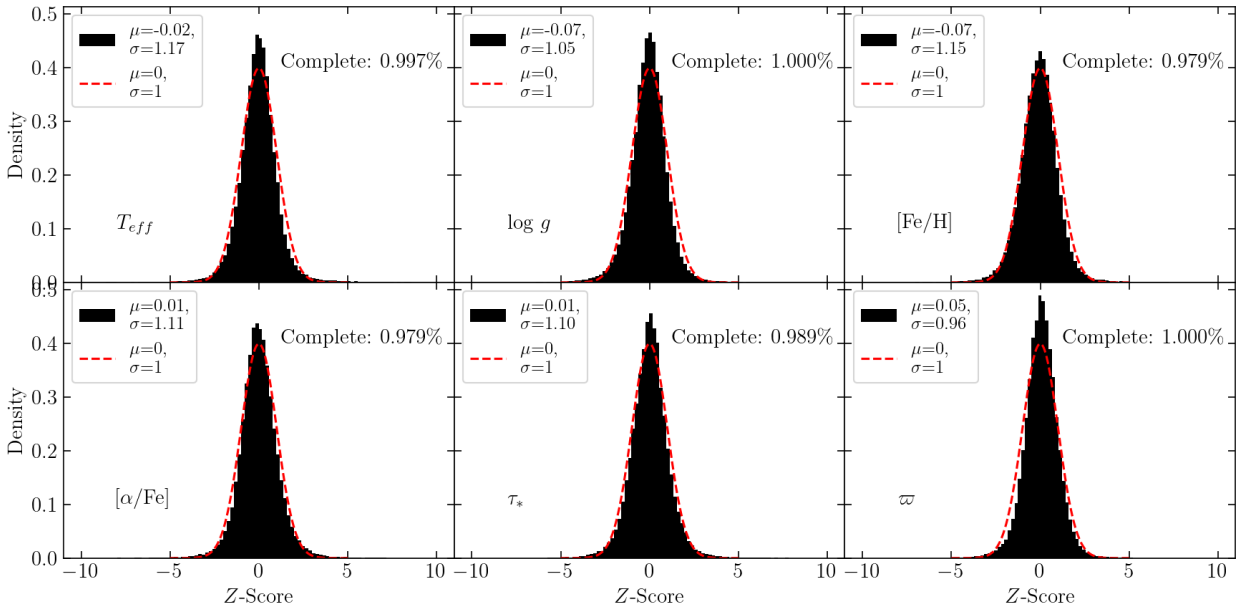

Figure 4.7: Z-scores of the labels to determine the is the errors predicted by the model are over or under represented by the model. The stellar parameters are labelled in the lower left of the plots, with the completeness, the percentage of stars used when computing the mean and standard deviation of the z-score, given in the top left. The black bins are the z-scores computed using the average of the deviations, with the red-dashed line representing a normal distribution

the z-score distributions given in Figure 4.7. The means of the z-score distributions are shown to be consistently near 0, while the standard deviations are greater than 1 for all parameters except parallax. The larger standard deviations demonstrate that the model is underconfident and underestimates the uncertainty intervals. This conclusion may, in part, result from averaging of the 1σ deviations at the edges of the distribution of labels, as well as from a few outliers at the far tails of the z-score distributions.

Chapter 5

Discussion

5.1 Scaling the Datasets

To probe how the algorithm's training benefited from having all 220M sources, the pre-training was also performed with subsets of the data, being datasets with 1M and 10M stars, randomly sampled from the *Gaia* DR3 XP spectra catalogue. After, the same fine-tuning procedure was performed, holding all hyperparameters the same. The metrics computed on the test dataset for the 3 models is shown in Table 5.1, revealing a small increase in accuracy with increasing dataset size. However, the increase in scale from 10M sources to 220M did not show a significant improvement for the model when predicting with full XP coefficients.

The scaling of the data had the most noticeable effects for predicting stellar parameters with missing XP coefficients. The stellar parameters with the greatest weight of importance for this research, [Fe/H] and τ_*, both show visible improvements in distribution and metrics, shown alongside $T_{\rm eff}$ in Figure 5.1 and Table 5.1 (The differences in the stellar parameters of $\log g$, [α/Fe], and ϖ are minimal but are included in Table 5.1 for completeness). These upgrades in accuracy act as another example of the robustness of the model with respect to imbalance, as these informative embeddings are valid for stars with minimal coverage in photometric bands. By increasing the photometric surveys included in the dataset, the embeddings may display a measureable capacity to adapt to rare and OOD stars, which will be the subject of future tests with the MSA.

To further investigate the effect of scaling the pre-training dataset size, I plotted a t-SNE of the latent space for each model. Figure 5.2 reveals the latent space organization of the models, showing a large difference in the visual separation of stars with similar ages between the 1M model and the 10M and 220M models, which may correlate with the information encoded in the embeddings. This observation may be linked to the increase in accuracy shown for predictions with masked spectra.

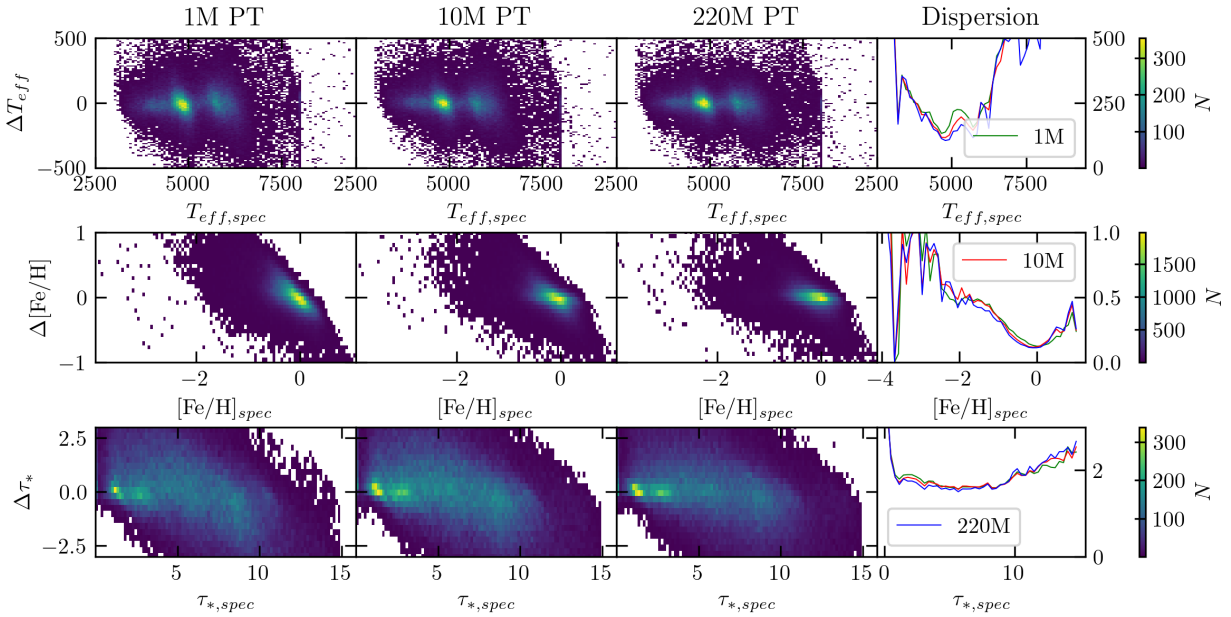

Figure 5.1: Residuals per label of the MSA+P for 3 different sizes of pre-training dataset. The x- and y-axes are equivalent for the first three panels in each row, with the last column having a separate y-axis scale labelled on the left side equal to the positive half of the other plots.

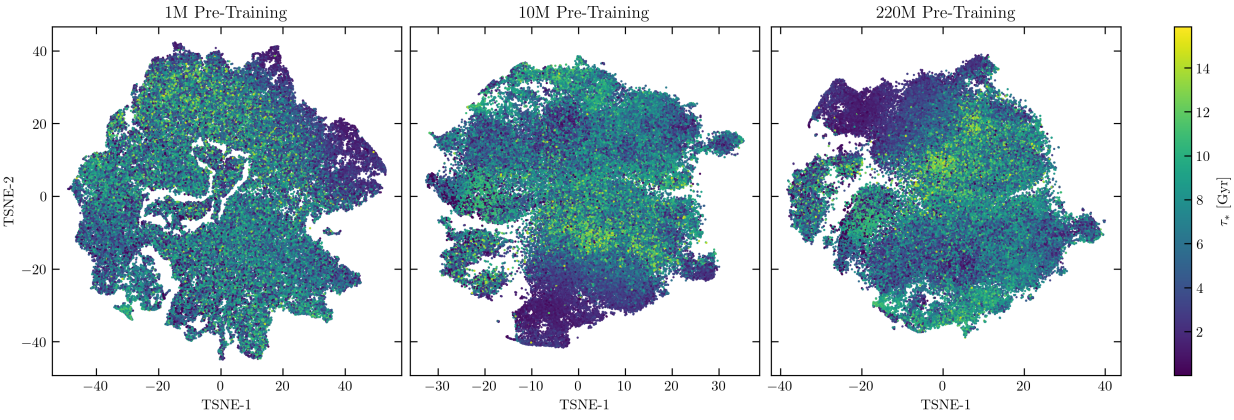

Figure 5.2: t-SNE of the latent space for the pre-trained MSA with 3 different sizes of pre-trainined dataset.

Table 5.1: The metrics for every pre-trained model after fine-tuning the same prediction head to the pre-trained autoencoder. The top half of the table contains the predictions for the entirety of the observed features as they exist, while the bottom half of the table had all XP spectral coefficients masked before passing to the periodic encoding layer of the MSA.

Size:		1M		10M		220M	
Label	Unit	RMSE	MAE	RMSE	MAE	RMSE	MAE
$T_{\rm eff}$	[K]	339.3	104.7	338.1	97.0	330.5	92.3
$\log g$		0.165	0.092	0.162	0.089	0.145	0.075
[Fe/H]		0.182	0.102	0.171	0.093	0.166	0.090
[α/Fe]		0.076	0.049	0.073	0.046	0.072	0.045
τ_*	[Gyr]	2.097	1.402	2.050	1.376	1.997	1.297
$\log \pi$		0.127	0.054	0.114	0.050	0.095	0.037
-	-	RMSE_M	MAE_M	RMSE_M	MAE_M	RMSE_M	MAE_M
$T_{\rm eff}$	[K]	415.1	147.8	390.7	131.9	357.7	108.3
$\log g$		0.165	0.088	0.154	0.079	0.141	0.072
[Fe/H]		0.261	0.177	0.223	0.140	0.193	0.114
[α/Fe]		0.092	0.063	0.085	0.057	0.080	0.053
τ_*	[Gyr]	2.363	1.678	2.220	1.534	2.107	1.398
$\log \pi$		0.149	0.050	0.146	0.035	0.135	0.030

An important aspect of the model is the ordering of the predictions of the continuous variables with respect to the observed labels. If the errors are large for a particular prediction, its stellar parameters relative to the stars predictions can still mark it as an interesting candidate for spectroscopic follow-up observations. To quantify whether the predicted parameters are ordered similarly to the labels, I measured the change in the Spearman's rank correlation coefficient to determine how well the monotonicity was preserved between the labels and the predictions. In Figure 5.3, one can observe certain trends emerging with increasing pre-training dataset size. With 1.0 equal to a perfect ordering with respect to the labels, all models show improvement with scaling the data, with the largest changes in [Fe/H], [α/Fe], and τ_*, as there was the most to improve with the correlation. The $T_{\rm eff}$ also improves by ~ 0.01 when masked, being a significant change from an initial ρ of 0.97.

5.2 Applicability to Heterogeneous Spectroscopic Datasets

While experimenting with different spectroscopic datasets to obtain labels for the fine-tuning algorithm, it was found that increasing the fine-tuning dataset diversity inhibited convergence in training rather than accelerating it. The impact of heterogeneity in the spectroscopic

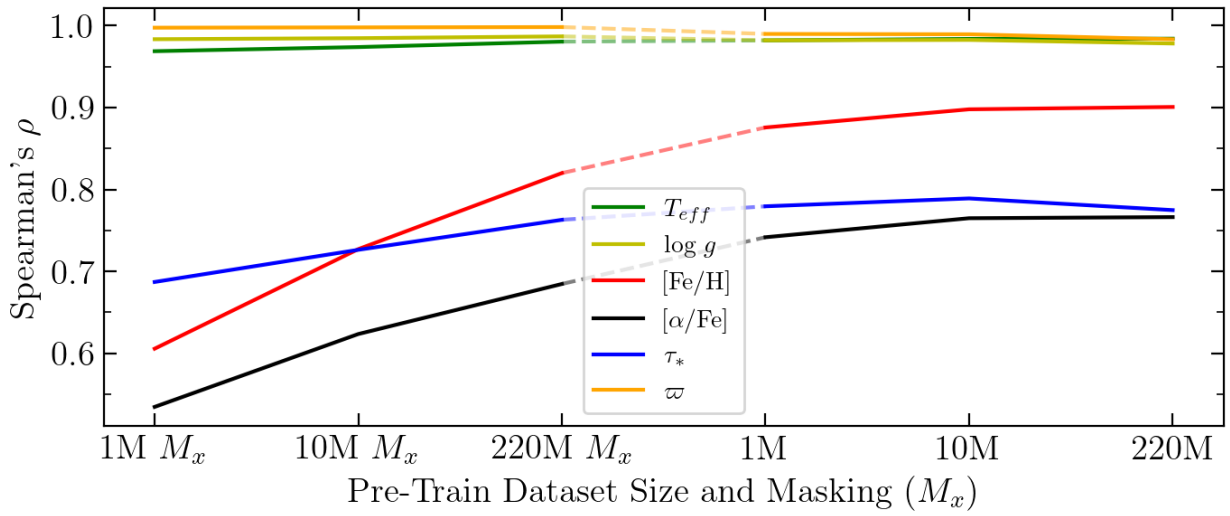

Figure 5.3: The change in Spearman's rank correlation coefficient with respect to differing pre-train dataset size and masking scheme. M_x denotes that the XP spectra were fully masked when passing through the model to generate the predictions for that particular data point and pre-train size. The hashed line connects the plot between masked and unmasked XP predictions.

dataset pipelines resulted in early plateauing gradients and predictions towards the mean of the six labels. This led to training the predictor of the MSA on individual spectroscopic datasets to draw comparisons between the spectroscopic datasets themselves and with other state-of-the-art models working towards the common goal of stellar parameter derivation.

5.2.1 Fine-Tuning on APOGEE and GALAH Individually

First, I split the fine-tuning dataset described in Chapter 2 into its respective components. The VMP stars from Li et al. (2022) are used in both datasets to fill in the metal-poor regime of the data, specifically with respect to APOGEE. The model is then fine-tuned on the datasets for 100 epochs, but without ensembling the model over multiple random seeds. The self-consistency plots for the labels and predictions are shown in Figures 5.4-5.5, with the MAE and RMSE metrics computed for the full distributions presented in Table 5.2. Additional figures showing the residuals of the algorithm are located in Appendix A,

A deviation from the labels is shown in the predicted Kiel diagrams, with a lengthening along the $T_{\rm eff}$ axis towards hotter stars. The metrics for $T_{\rm eff}$ are also elevated relative to the MSA+P and other spectroscopic datasets in this section. This trend likely appears in the predictions from this model due to the numerous hot stars included in the fine-tuning dataset which are usually omitted in supervised networks as a result of missing [Fe/H] label

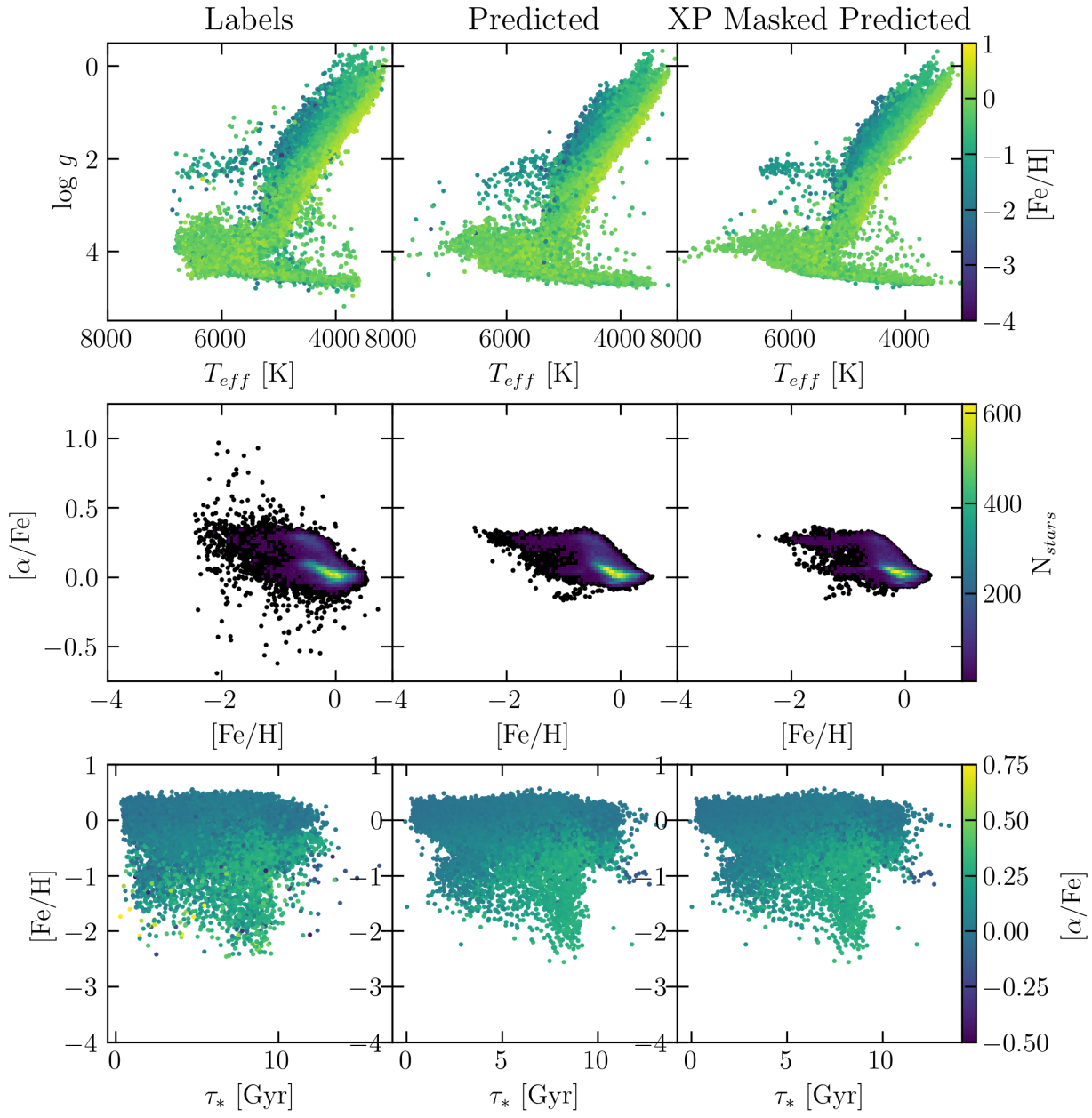

Figure 5.4: The same as Figure 4.6, for the MSA and prediction head trained solely on APOGEE and VMPs from Li et al. (2022).

derivations (these stars can be seen in the top right of the upper left panel in Figures A.2-A.3). Also, shown in the Kiel diagrams are the tendency for the model to predict values increasingly clustered together with decreasing spectroscopic information. This may shed light on the disagreements shown in Tables 4.1 and 5.2 with respect to better metrics for predictions from masked XP coefficients. This tendency to cluster the sources together appears to follow the observed behaviour of the predictions "collapsing towards the mean", in which data with limited coverage or information tends to be grouped with the mean of other similar stars, improving the scores on metrics penalizing large outliers. This has similarly been observed in the Kiel diagrams for GALAH (Figure 5.5), which harbours lower metallicity stars that are being predicted as low metallicity, but are varying with temperature.

The predicted [α/Fe]-[Fe/H] plots both show better preservation of the shape of the label distribution with respect to the middle panels of Figure 4.6. However, a tighter grouping of the predicted stellar parameters remains seen in the distributions. The predictions for the GALAH dataset reveal better residuals in the metal-poor regime, possibly due to increased sources from the VMP catalogue that represent a larger portion of the dataset and in-distribution stars. Finally, the AMR plots are shown for both APOGEE and GALAH in the lower panels of Figures 5.4-5.5. The labels of both datasets show similar distributions, with the tracers of the different age derivation methods appearing in the lower left plots. The left-most APOGEE AMR shows a large density of solar metallicity and solar [α/Fe] in the dataset, spanning from young to old stars. The plateau in age from the astroNN model (Mackereth et al., 2019) can be seen at approximately 10 Gyr, with some older ages from Leung et al. (2023). For GALAH, the isochrone ages for very young stars can be seen by the hard line at 0 Gyr, which is attempted to be reproduced in the predicted values leading to negative predictions for τ_*. These stars also represent a large amount of the metal-poor sources, with some metal-poor stars at older ages, but with less density (leading to potential problems in Section 5.3). The metal-poor stars at older ages are better reconstructed by the GALAH-fine-tuned model, but a significant artifact from the young stars in GALAH is shown to be present in the AMR.

5.2.2 Fine-tuning with RAVE DR6

RAVE DR6 was selected as the external catalogue for the MSA and prediction head. The pre-processing of this table consisted only of removing duplicates, and combining the derivations from the BDASP pipeline of $T_{\rm eff}$, log g, [M/H], and τ_*, with the [α/M] estimate from the convolutional neural network (CNN) trained on the RAVE spectra. Overall metallicity is

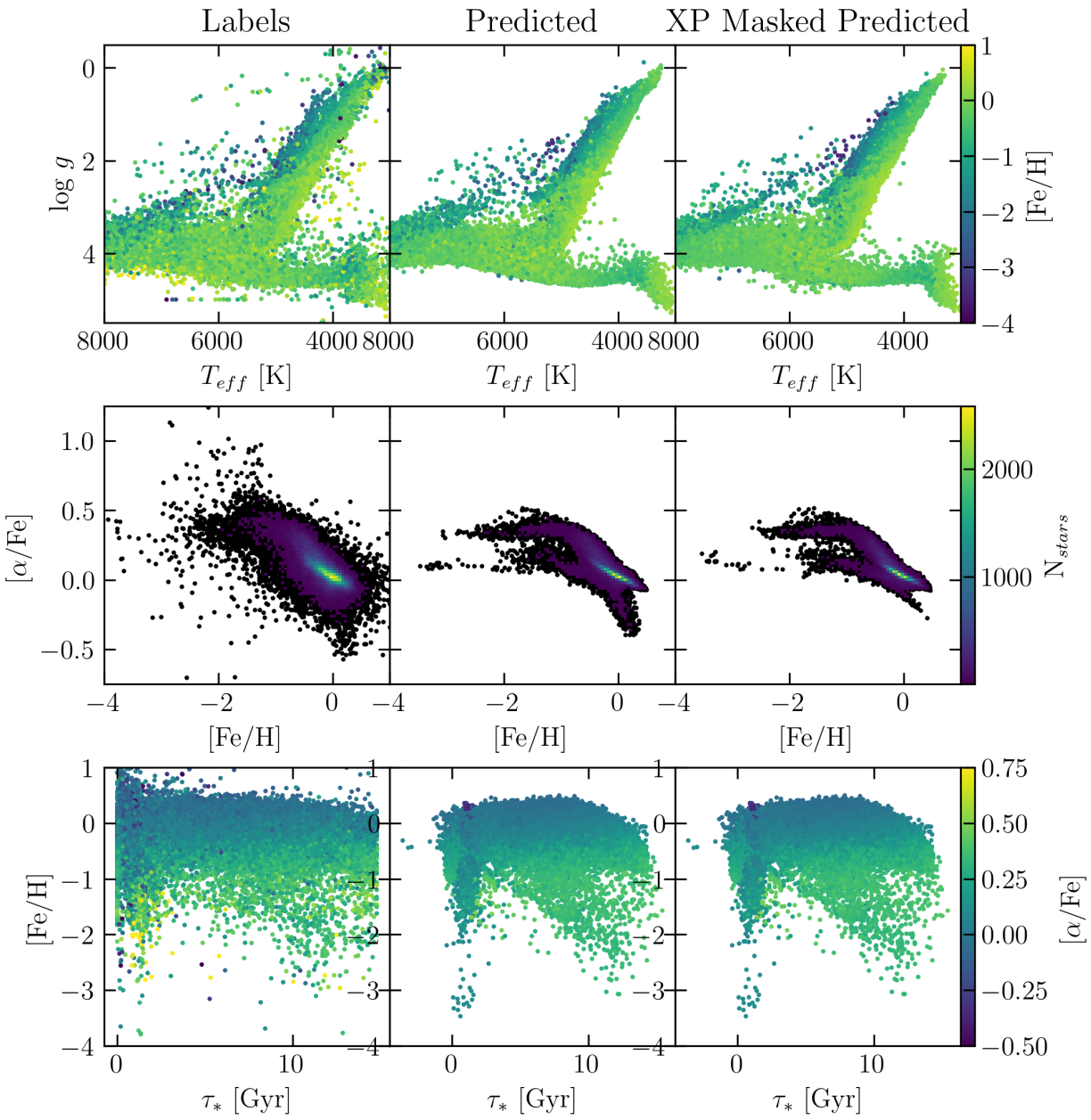

Figure 5.5: The same as Figures 4.6 and 5.4, but with GALAH and VMPs.

given in lieu of [Fe/H] in this dataset, which will result in a higher metallicity label relative to the other datasets. The VMP stars are omitted from this subset. The self-consistency plot for this dataset is shown in Figure 5.6 and the metrics are given in Table 5.2. Additional figures of the residuals of the model are given in Appendix A. Figure 5.6 confirms the trend seen for APOGEE and GALAH of reducing the intrinsic scatter of the labels in the predictions. The RAVE DR6 catalogue contains less variation in $T_{\rm eff}$ and $\log g$, resulting in better metrics and a greater similarity in Kiel diagram predictions. From the bottom left panel of Figure 5.6, a cluster of stars with ages on the order of Myr can be seen as with the GALAH dataset. As a result, the predicted ages replicate Figure 5.5, with some erroneous predictions of negative ages.

From the models trained on the different spectroscopic datasets, it is shown that the embeddings are tunable to multiple catalogues while replicating the distributions of the atmospheric labels. However, the MSA+P fine-tuned on the concatenated dataset displayed better generalisation to the multiple surveys, by learning the connections between the embeddings and a large set of in-distribution stars. When comparing the fine-tuned algorithms, no one model trained on an individual spectroscopic dataset obtained better metrics than the MSA+P for every label. This is reaffirmed by the computed MADs for the individually trained APOGEE and GALAH datasets. For APOGEE, the MAD was determined to be 44.5 K in $T_{\rm eff}$, 0.093 dex in $\log g$, and 0.057 dex in [Fe/H], while with GALAH, the MAD was determined to be 46.9 K in $T_{\rm eff}$, 0.062 dex in $\log g$, and 0.055 dex in [Fe/H]. These measurements are shown to be higher for the individual spectroscopic datasets than for the MSA+P, which likely resulted from the increased scatter due the smaller portion of the parameter space covered by the individual datasets.

5.3 Applications in the Near-Universe

To gauge the model's ability to predict labels for out-of-distribution stars, I predict full stellar properties for a few clusters and dwarf galaxies. The clusters were selected from the Hunt & Reffert (2024) catalogue, based on the criteria that (1) they had stars with magnitudes in *Gaia* DR3 and at least a few stars with XP coefficients, (2) those stars had membership probabilities of greater than 0.99, and (3) they had a reported cluster age of greater than 1 Gyr. This final selection criterion was based on poor performance of the model on stellar ages for young stars (Figure 4.3-4.4). The final cluster dataset comprised a total of 84 open clusters with the number of member stars varying on the order of 10$\sim$100. The self-consistency in the open cluster set is plotted the same as in Section 4.2, shown

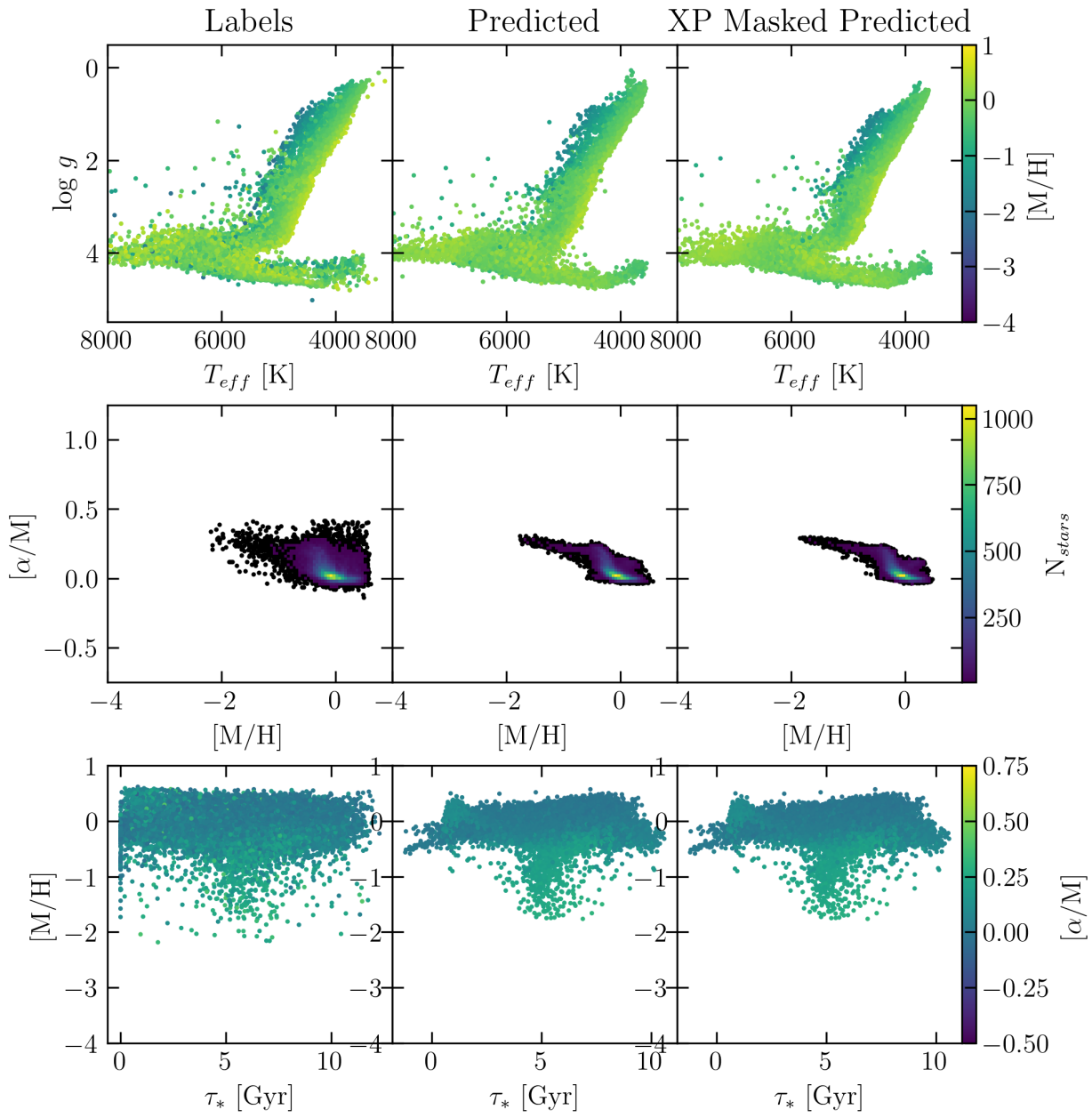

Figure 5.6: The same as Figures 5.4-5.5, for the MSA on the external RAVE DR6 dataset. Provided in the RAVE DR6 catalogue from the BDASP pipeline are the overall metallicities [M/H], differing slightly from the Fe-abundances used as labels in the other spectroscopic datasets, reflected in all nine panels.

Table 5.2: The performance of the MSA and prediction head on the individual spectroscopic datasets of APOGEE, GALAH, RAVE, and the Li et al. (2022) VMPs for both masked and unmasked XP coefficients.

Dataset		APOGEE		GALAH		RAVE	
Label	Unit	RMSE	MAE	RMSE	MAE	RMSE	MAE
$T_{\rm eff}$	[K]	443.0	114.3	160.1	80.4	139.4	86.1
$\log g$		0.238	0.139	0.140	0.092	0.144	0.097
[Fe/H]		0.139	0.084	0.178	0.096	0.197	0.129
[α/Fe]		0.054	0.033	0.076	0.050	0.043	0.027
τ_*	[Gyr]	1.262	0.902	2.098	1.411	1.701	1.260
$\log \pi$		0.225	0.126	0.102	0.064	0.091	0.052
-	-	RMSE_M	MAE_M	RMSE_M	MAE_M	RMSE_M	MAE_M
$T_{\rm eff}$	[K]	531.8	137.8	188.8	100.8	119.5	75.2
$\log g$		0.202	0.116	0.089	0.051	0.109	0.065
[Fe/H]		0.195	0.129	0.214	0.130	0.200	0.132
[α/Fe]		0.067	0.044	0.087	0.059	0.047	0.030
τ_*	[Gyr]	1.543	1.126	2.134	1.418	1.595	1.148
$\log \pi$		0.622	0.072	0.204	0.040	0.178	0.053

in Figure 5.7. It can be seen that some systematics and biases in the model are apparent in the predictions, particularly with respect to the third panel. The Kiel diagram shows a general metallicity near 0 dex with alignment of metallicity in the red giant branch similar to expectations from Figure 4.6.

However, the [α/Fe]-[Fe/H] plot shows a large [α/Fe] enhancement, with the AMR panel showing this trend for all ages in the open cluster subset. Open clusters situated in the disk of the Milky Way are expected to have solar to sub-solar metallicities and α-abundances, with some variation in galactocentric radius (e.g. Donor et al., 2020). For older open clusters, a small enhancement in [α/Fe] ($\sim$0.1 dex), particularly with low metallicity, can be seen (Yong et al., 2012). As discussed in Section 2.1.2, the chemical information derivable from the XP spectra consists solely of C, N, and Fe lines, resulting in a reliance on inferring [α/Fe] from the information of the photometric surveys and from the labels of the fine-tuning dataset. Other sources of error for the measurement have been shown using analytical models to extract the α-abundance from the XP spectra (Witten et al., 2022), which showed success only for cool, solar-metallicity stars.

To measure the precision of the cluster age predictions, the stars were grouped by cluster membership to give one combined estimate. To obtain the grouped age per cluster, the mode

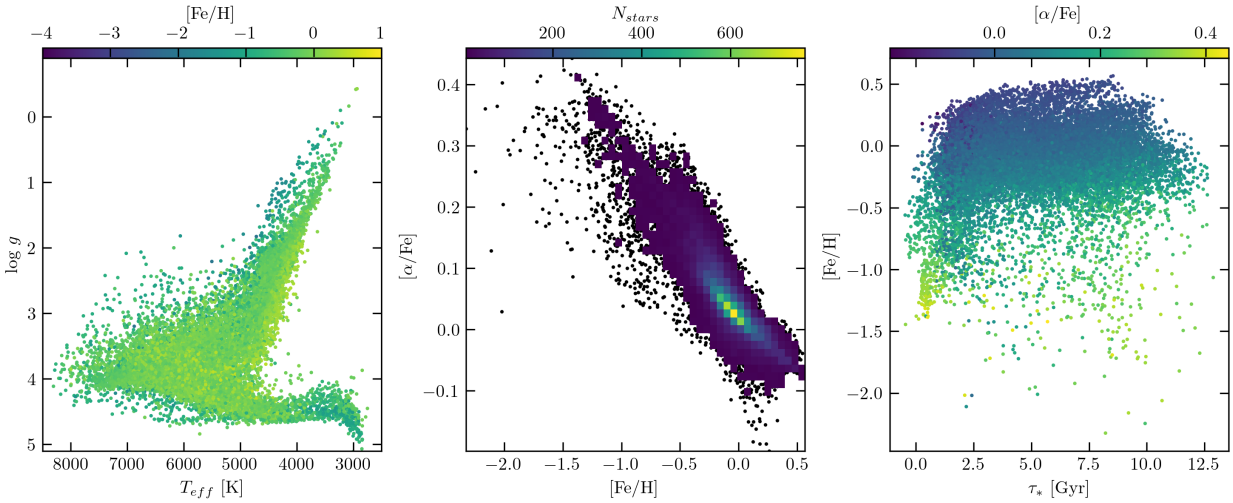

Figure 5.7: The self consistency plots as with Figure 4.6, consisting solely of the predictions for the open cluster dataset. No data was masked, nor were there labels other than age included in this dataset.

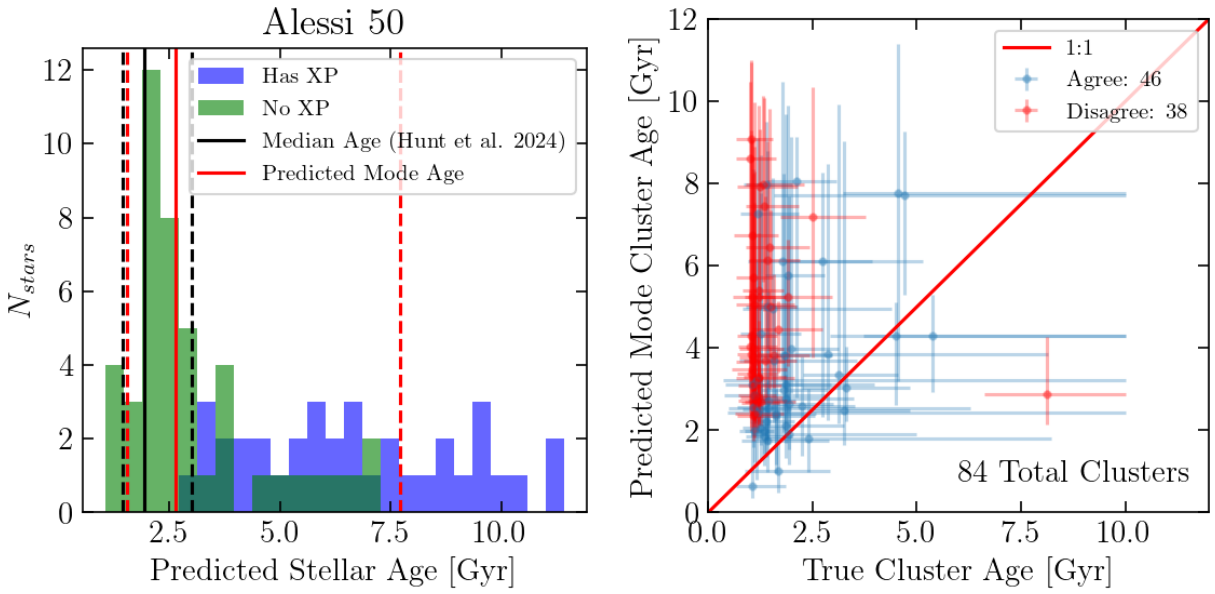

Figure 5.8: These plots show the predicted age for a cluster Alessi 50 (left), and the consistency between predicted and measured ages for clusters in the open cluster dataset (right). The dashed lines represent the 16th and 84th percentiles from the errors, which for the predicted ages, are the mode from the predicted error distributions. The left-hand plot shows the mode of the predicted ages for a given cluster against the literature age, with errors from the 16th and 84th percentiles of both predictions and measured.

of the predicted ages was chosen rather than the mean, as the distributions of stellar age often had long tails which skewed the estimate, such as for the cluster Alessi 50 shown in Figure 5.8. For the open cluster dataset, the MSA+P obtained an agreement with the ages from Hunt & Reffert (2024) of 54.7%.

Of particular interest, the age bins most associated with the peak of the distribution had no XP coefficients within their features when fed to the MSA. Upon inspection, nothing with respect to the magnitudes included and omitted was of note, leading to the speculation that the embeddings were able to group the various magnitudes that were available with the younger stars in the latent space. This investigation will be followed up in a future paper (McKay et al. *in prep.*).

Two dwarf galaxies were selected as test subjects for the MSA, appealing to the strengths of the model. These galaxies are the Leo I dwarf spheroidal, chosen as it has experienced a relatively recent (∼1-7Gyr ago) star-forming epoch (Lanfranchi & Matteucci, 2010), and the Fornax dwarf spheroidal, which hosts a large number of stars with intermediate ages (Buonanno et al., 1999; de Boer et al., 2012). Other dwarfs were also investigated, though

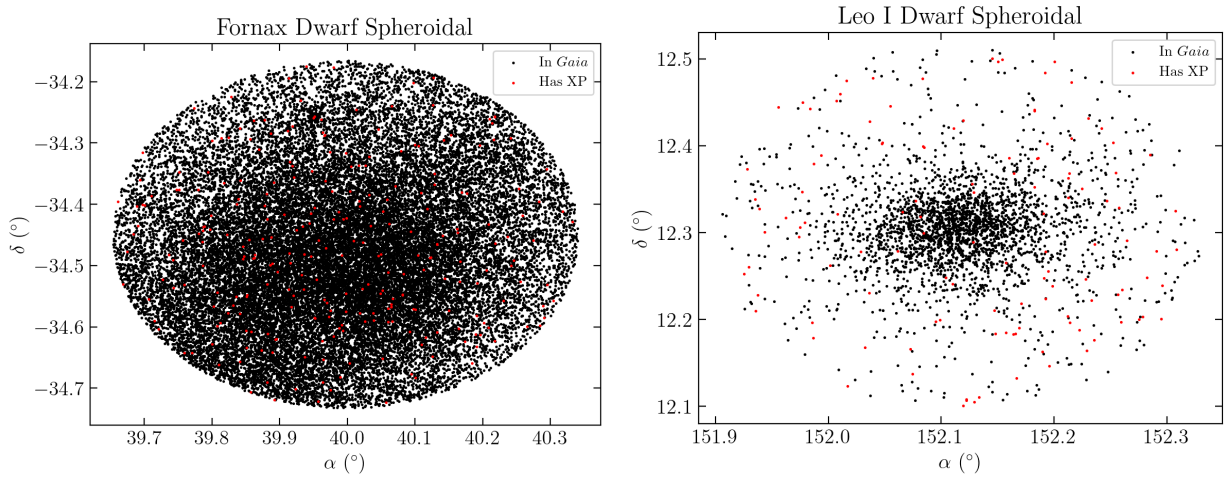

Figure 5.9: The stars included in the *Gaia* Archive cone search for dwarf spheroidals Fornax (left), and Leo I (right), plotted in RA and declination. The black dots represent all stars within the *Gaia* DR3 main source catalogue, with the red dots denoting those with XP continuous spectra released in DR3.

resulted in worse predictions with respect to age and metallicity, possibly due to being mostly out-of-distribution with the fine-tuning dataset. The systems were selected from their pre-computed `target_id` available in the *Gaia* Archive and a cone search based on extents listed in the NASA/IPAC Extragalactic Database (Figure 5.9). No other cuts were made on the data, as the potential inclusion of anomalous points acts as another verification of the ability of the MSA+P to detect outliers. Like with the open clusters, in Figures 5.10-5.11, I plot the Kiel diagram and AMR of the dwarf galaxies to verify self-consistency in the predictions and to compare to literature values of the stellar parameters.

Unlike with the previous self-consistency plots, I show the MSA+P parallax predictions compared to the *Gaia* DR3 measurements, shown in the middle panel of Figure 5.10 for Leo I. The distribution of predicted parallaxes is consistent with the positive tail of the distribution of *Gaia* parallaxes. The same is replicated for Fornax in the middle panel of Figure 5.11. Negative parallaxes make up approximately 24% of the measurements in DR3, and often arise for stars with small parallaxes situated in crowded regions, where small angular separations between sources results in poor measurements. A zero-point correction should be made to the raw parallaxes, but this itself only shifts the parallaxes on the order of ~ 10 μas, which does not account for the large negative values seen in the labels for both dwarfs. In comparison, the MSA+P predicts positive parallaxes consistently, while having a mode consistent with the parallax label distribution.

The Kiel diagrams and AMRs for Leo I and Fornax show the limitations of the model and

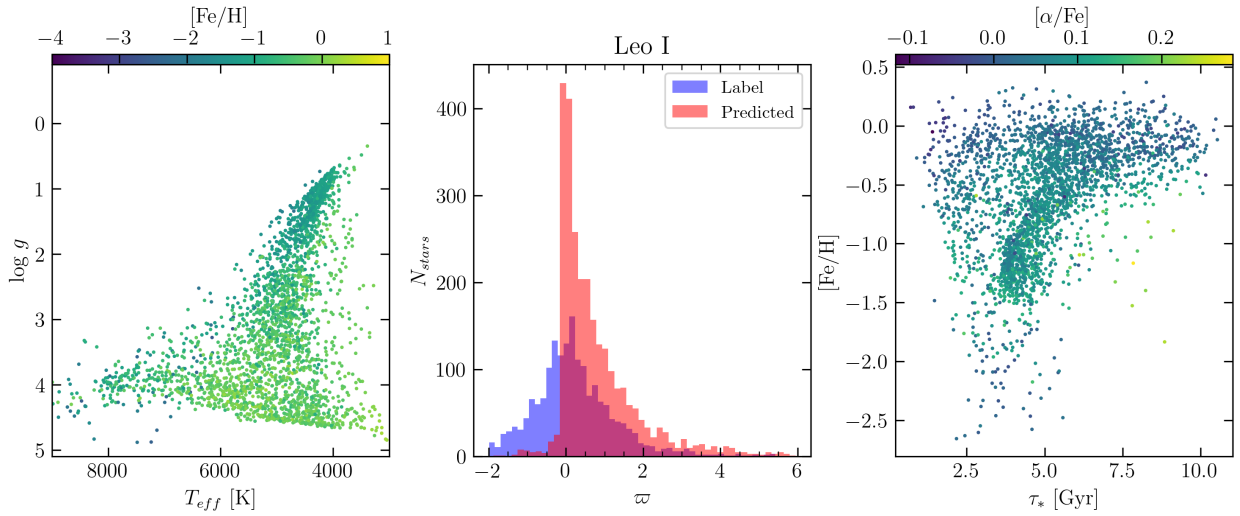

Figure 5.10: The self-consistency checks for Leo I. *Left*: A Kiel diagram of predictions in $T_{\rm eff}$, log g, and [Fe/H]. *Center*: The predictions for the parallax of stars in Leo I (red) compared to the *Gaia* DR3 values (blue). *Right*: The age-metallicity relation coloured by [α/Fe].

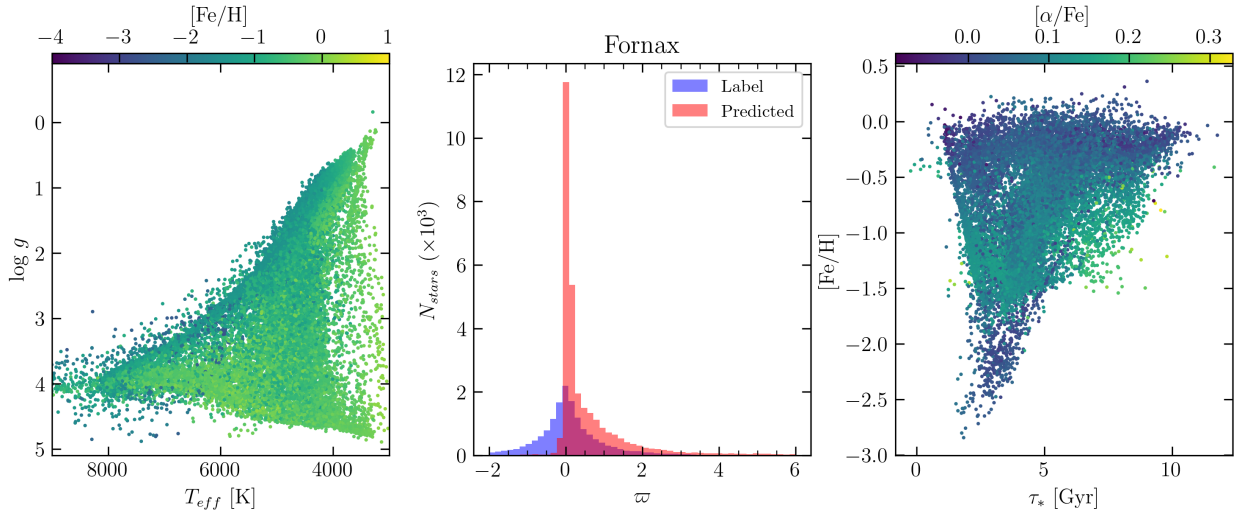

Figure 5.11: The same as Figure 5.10 for the Fornax dwarf Spheroidal.

where improvements will be focused for future iterations of the MSA. The Kiel diagram for Leo I (Figure 5.10) is shown to have a distinguishable main sequence and red giant branch, with a tight clustering of stars at the top of the distribution and some scatter, as expected. However, the metallicity distribution is shown to be near-solar and centered in the range -1 dex <[Fe/H]< 0 dex, which is much more metal-rich than the expected range of -3 dex <[Fe/H]< -1 dex (Lanfranchi & Matteucci, 2010). It is shown from the predicted AMR of Leo I that the majority of the star have ages in agreement with the galaxy's recent boost

in star formation. However, the distribution in metallicities shows a much more metal-rich population, with the oldest stars having the higher metallicities. The same trends are shown for Fornax in Figure 5.11. Though the ordering of the red giant branch with respect to metallicity is shown to be correct, the distribution in metallicities is higher than expected, as confirmed by the AMR, with higher metallicities for the oldest stars.

These limitations in the model can likely be mitigated through further curation of the spectroscopic dataset and more input data for the pre-training task. The parameters the algorithm was attempting to predict represent the very edges of the labelled dataset domain, stars in crowded regions and with few XP coefficients. With increasing training epochs, the biases in the model were shown to attenuate, and the MSA reconstructed more accurate magnitudes and coefficients with more pretraining data. As the model already performs well with predicting spectral parameters with missing coefficients for in-distribution stars, improving the predictions for these test stellar populations will hopefully follow by including more photometric catalogues readily available.

Chapter 6

Conclusions

In this thesis, I have presented the Masked Stellar Autoencoder, a self-supervised deep learning model designed to learn meaningful representations of photometric magnitudes and *Gaia* XP spectra to extend the reach of *Gaia*. To accomplish this feat, I trained the encoder-decoder model on the $\sim$ 220 million sources with XP coefficients and parallaxes in *Gaia* DR3, including photometric magnitudes from six separate surveys spanning optical and infrared wavelengths. I demonstrate the model's ability to reconstruct the spectral coefficients and photometric magnitudes from its latent embeddings, then further show its applicability to Galactic archaeology by training a simple prediction head on its embedded data for stellar parameter regression. The model showed itself to be competitive with other models trained on the same data, while demonstrating its ability to predict stellar parameters when missing the XP information with minimal increases in prediction error. This aspect of the model, resulting from the reconstructive pretext task, indicates potential for autoencoder models to derive accurate stellar parameter predictions for the full sky with only a fraction of measured high-resolution spectra and vast, but incomplete, photometry.

With respect to stellar parameter regression, the model achieved mean absolute errors of 92 K in $T_{\rm eff}$, 0.08 dex in log g, and 0.09 dex in [Fe/H] for the mixed GALAH and APOGEE dataset, which are competitive with XGBoost and transformer-based models trained on APOGEE data. Furthermore, the MSA+P achieved MAEs of 0.05 dex in [α/Fe], 1.3 Gyr in τ_*, and 0.04 in log ϖ. The model showed the capacity to predict accurate stellar features even when missing XP coefficients, which will motivate the injection of many more pre-existing photometric surveys into the pretraining dataset to increase its predictive power for rare stellar objects in future iterations. Errors are predicted by the model, and were shown to only be marginally underconfident. The embeddings were shown to be tunable to multiple spectroscopic datasets, replicating the Kiel diagrams, [α/Fe]-[Fe/H] plots, and AMRs with less scatter in the predictions than the labels. This demonstrates the ability of the algorithm

to cluster and group like stars together, despite a small decrease in prediction accuracy. The model was tested on stellar populations including clusters and dwarf galaxies, revealing the strengths of the embeddings with missing XP information while also showing the limitations from the fine-tuning sample distribution. Surprisingly, the model was shown to predict dwarf galaxy parallaxes accurately with respect to the *Gaia* observed measurements, with the re-derivation of non-physical negative parallaxes in the dataset.

This work demonstrates a powerful tool for creating large, homogeneous catalogues of stellar parameters. The pre-trained embeddings of the MSA have much applicability yet to be explored, including searching for similar or anomalous stars using the latent representations of the data. This first iteration of a foundation model trained on the full set of available *Gaia* XP coefficients showed robustness against incomplete and imbalanced data which, via self-supervised learning, will increase in precision with additional photometric surveys and upcoming *Gaia* data releases.

Bibliography

Abdurro'uf, Accetta, K., Aerts, C., et al. 2022, The Astrophysical Journal Supplement Series, 259, 35, doi: 10.3847/1538-4365/ac4414

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, The Astrophysical Journal Supplement Series, 219, 12, doi: 10.1088/0067-0049/219/1/12

Allende Prieto, C. 2016, Astronomy & Astrophysics, 595, A129, doi: 10.1051/0004-6361/201628789

Almannaei, A. S., Kawata, D., Ciuca, I., et al. 2024, Towards Galactic Archaeology with Inferred Ages of Giant Stars From Gaia Spectra, doi: 10.48550/arXiv.2412.09040

Andrae, R., Rix, H.-W., & Chandra, V. 2023, The Astrophysical Journal Supplement Series, 267, 8, doi: 10.3847/1538-4365/acd53e

Ardern-Arentsen, A., Kane, S. G., Belokurov, V., et al. 2025, Monthly Notices of the Royal Astronomical Society, 537, 1984, doi: 10.1093/mnras/staf096

Barnes, S. A. 2003, The Astrophysical Journal, 586, 464, doi: 10.1086/367639

Belokurov, V., Erkal, D., Evans, N. W., Koposov, S. E., & Deason, A. J. 2018, Monthly Notices of the Royal Astronomical Society, 478, 611, doi: 10.1093/mnras/sty982

Belokurov, V., & Kravtsov, A. 2022, Monthly Notices of the Royal Astronomical Society, 514, 689, doi: 10.1093/mnras/stac1267

Belokurov, V., Sanders, J. L., Fattahi, A., et al. 2020, Monthly Notices of the Royal Astronomical Society, 494, 3880, doi: 10.1093/mnras/staa876

Bengio, Y., Yao, L., Alain, G., & Vincent, P. 2013, in Advances in Neural Information Processing Systems, Vol. 26 (Curran Associates, Inc.)

Bland-Hawthorn, J., & Gerhard, O. 2016, Annual Review of Astronomy and Astrophysics, 54, 529, doi: 10.1146/annurev-astro-081915-023441

Bonaca, A., Conroy, C., Wetzel, A., Hopkins, P. F., & Kereš, D. 2017, The Astrophysical Journal, 845, 101, doi: 10.3847/1538-4357/aa7d0c

Borawar, L., & Kaur, R. 2023, in Lecture Notes in Networks and Systems, Vol. 600, Proceedings of International Conference on Recent Trends in Computing (Singapore: Springer Nature Singapore), 235–247

Buder, S., Kos, J., Wang, E. X., et al. 2024, The GALAH Survey: Data Release 4, arXiv, doi: 10.48550/arXiv.2409.19858

Buonanno, R., Corsi, C. E., Castellani, M., et al. 1999, The Astronomical Journal, 118, 1671, doi: 10.1086/301034

Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, Astronomy & Astrophysics, 652, A86, doi: 10.1051/0004-6361/202141249

Cerny, W., Martínez-Vázquez, C. E., Drlica-Wagner, A., et al. 2023, The Astrophysical Journal, 953, 1, doi: 10.3847/1538-4357/acdd78

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, The Pan-STARRS1 Surveys

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, in Proceedings of the 37th International Conference on Machine Learning (PMLR), 1597–1607

Cunha, M. S., Aerts, C., Christensen-Dalsgaard, J., et al. 2007, Astronomy and Astrophysics Review, 14, 217, doi: 10.1007/s00159-007-0007-0

De Angeli, F., Weiler, M., Montegriffo, P., et al. 2023, Astronomy & Astrophysics, 674, A2, doi: 10.1051/0004-6361/202243680

de Boer, T. J. L., Tolstoy, E., Hill, V., et al. 2012, Astronomy & Astrophysics, 544, A73, doi: 10.1051/0004-6361/201219547

De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, Monthly Notices of the Royal Astronomical Society, 449, 2604, doi: 10.1093/mnras/stv327

Deason, A. J., & Belokurov, V. 2024, New Astronomy Reviews, 99, 101706, doi: 10.1016/j.newar.2024.101706

DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2024, The Astronomical Journal, 168, 58, doi: 10.3847/1538-3881/ad3217

Dey, A., Schlegel, D. J., Lang, D., et al. 2019, The Astronomical Journal, 157, 168, doi: 10.3847/1538-3881/ab089d

Donor, J., Frinchaboy, P. M., Cunha, K., et al. 2020, The Astronomical Journal, 159, 199, doi: 10.3847/1538-3881/ab77bc

Drlica-Wagner, A., Carlin, J. L., Nidever, D. L., et al. 2021, The Astrophysical Journal Supplement Series, 256, 2, doi: 10.3847/1538-4365/ac079d

Eisenhardt, P. R. M., Marocco, F., Fowler, J. W., et al. 2020, The Astrophysical Journal Supplement Series, 247, 69, doi: 10.3847/1538-4365/ab7f2a

Fallows, C. P., & Sanders, J. L. 2024, Monthly Notices of the Royal Astronomical Society, 531, 2126, doi: 10.1093/mnras/stae1303

Frebel, A., & Norris, J. E. 2015, Annual Review of Astronomy and Astrophysics, 53, 631, doi: 10.1146/annurev-astro-082214-122423

Freeman, K., & Bland-Hawthorn, J. 2002, Annual Review of Astronomy and Astrophysics, 40, 487, doi: 10.1146/annurev.astro.40.060401.093840

Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023, Astronomy and Astrophysics, 674, A1, doi: 10.1051/0004-6361/202243940

García-Zamora, E. M., Torres, S., & Rebassa-Mansergas, A. 2023, Astronomy and Astrophysics, 679, A127, doi: 10.1051/0004-6361/202347601

Gavel, A., Andrae, R., Fouesneau, M., Korn, A. J., & Sordo, R. 2021, Astronomy & Astrophysics, 656, A93, doi: 10.1051/0004-6361/202141589

Gorishniy, Y., Rubachev, I., & Babenko, A. 2022, in Advances in Neural Information Processing Systems

Green, G. M. 2018, Journal of Open Source Software, 3, 695, doi: 10.21105/joss.00695

Guiglion, G., Nepal, S., Chiappini, C., et al. 2024, Astronomy and Astrophysics, 682, A9, doi: 10.1051/0004-6361/202347122

Hattori, K. 2024, Metallicity and α-Abundance for 48 Million Stars in Low-Extinction Regions in the Milky Way, doi: 10.48550/arXiv.2404.01269

Hayden, M. R., Bovy, J., Holtzman, J. A., et al. 2015, The Astrophysical Journal, 808, 132, doi: 10.1088/0004-637X/808/2/132

Haywood, M., Di Matteo, P., Lehnert, M. D., et al. 2018, The Astrophysical Journal, 863, 113, doi: 10.3847/1538-4357/aad235

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778, doi: 10.1109/CVPR.2016.90

Helmi, A. 2008, The Astronomy and Astrophysics Review, 15, 145, doi: 10.1007/s00159-008-0009-6

Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Nature, 563, 85, doi: 10.1038/s41586-018-0625-x

Hunt, E. L., & Reffert, S. 2024, Astronomy and Astrophysics, 686, A42, doi: 10.1051/0004-6361/202348662

Ivezić, Z., Beers, T. C., & Jurić, M. 2012, Annual Review of Astronomy and Astrophysics, 50, 251, doi: 10.1146/annurev-astro-081811-125504

Jørgensen, B. R., & Lindegren, L. 2005, Astronomy & Astrophysics, 436, 127, doi: 10.1051/0004-6361:20042185

Kane, S. G., Belokurov, V., Cranmer, M., et al. 2025, Monthly Notices of the Royal Astronomical Society, 536, 2507, doi: 10.1093/mnras/stae2689

Kao, M. L., Hawkins, K., Rogers, L. K., et al. 2024, The Astrophysical Journal, 970, 181, doi: 10.3847/1538-4357/ad5d6e

Kaplan, J., McCandlish, S., Henighan, T., et al. 2020, Scaling Laws for Neural Language Models, arXiv, doi: 10.48550/arXiv.2001.08361

Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, in Advances in Neural Information Processing Systems, Vol. 30 (Curran Associates, Inc.)

Lanfranchi, G. A., & Matteucci, F. 2010, Astronomy & Astrophysics, 512, A85, doi: 10.1051/0004-6361/200913045

Laroche, A., & Speagle, J. S. 2025, The Astrophysical Journal, 979, 5, doi: 10.3847/1538-4357/ad9607

LeCun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436, doi: 10.1038/nature14539

Lei Ba, J., Kiros, J. R., & Hinton, G. E. 2016, Layer Normalization, arXiv, doi: 10.48550/arXiv.1607.06450

Leung, H. W., & Bovy, J. 2024, Monthly Notices of the Royal Astronomical Society, 527, 1494, doi: 10.1093/mnras/stad3015

Leung, H. W., Bovy, J., Mackereth, J. T., & Miglio, A. 2023, Monthly Notices of the Royal Astronomical Society, 522, 4577, doi: 10.1093/mnras/stad1272

Li, H., Aoki, W., Matsuno, T., et al. 2022, The Astrophysical Journal, 931, 147, doi: 10.3847/1538-4357/ac6514

Li, J., Wong, K. W. K., Hogg, D. W., Rix, H.-W., & Chandra, V. 2024, The Astrophysical Journal Supplement Series, 272, 2, doi: 10.3847/1538-4365/ad2b4d

Lindsay, C. J., Hon, M., Ong, J. M. J., et al. 2025, arXiv e-prints, arXiv:2507.01091, doi: 10.48550/arXiv.2507.01091

Liu, H., HaoChen, J. Z., Gaidon, A., & Ma, T. 2021, in International Conference on Learning Representations

Lucey, M., Al Kharusi, N., Hawkins, K., et al. 2023, Monthly Notices of the Royal Astronomical Society, 523, 4049, doi: 10.1093/mnras/stad1675

Mackereth, J. T., Bovy, J., Leung, H. W., et al. 2019, Monthly Notices of the Royal Astronomical Society, 489, 176, doi: 10.1093/mnras/stz1521

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, The Astronomical Journal, 154, 94, doi: 10.3847/1538-3881/aa784d

Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2021, The Astrophysical Journal Supplement Series, 253, 8, doi: 10.3847/1538-4365/abd805

Martig, M., Fouesneau, M., Rix, H.-W., et al. 2016, Monthly Notices of the Royal Astronomical Society, 456, 3655, doi: 10.1093/mnras/stv2830

Martin, N. F., Venn, K. A., Aguado, D. S., et al. 2022, Nature, 601, 45, doi: 10.1038/s41586-021-04162-2

Masseron, T., & Gilmore, G. 2015, Monthly Notices of the Royal Astronomical Society, 453, 1855, doi: 10.1093/mnras/stv1731

McMillan, P. J., Kordopatis, G., Kunder, A., et al. 2018, Monthly Notices of the Royal Astronomical Society, 477, 5279, doi: 10.1093/mnras/sty990

Onken, C. A., Wolf, C., Bessell, M. S., et al. 2024, Publications of the Astronomical Society of Australia, 41, e061, doi: 10.1017/pasa.2024.53

—. 2019, Publications of the Astronomical Society of Australia, 36, e033, doi: 10.1017/pasa.2019.27

Parker, L., Lanusse, F., Golkar, S., et al. 2024, Monthly Notices of the Royal Astronomical Society, 531, 4990, doi: 10.1093/mnras/stae1450

Pearson, K. A., Palafox, L., & Griffith, C. A. 2018, Monthly Notices of the Royal Astronomical Society, 474, 478, doi: 10.1093/mnras/stx2761

Pinsonneault, M. H., Elsworth, Y., Epstein, C., et al. 2014, The Astrophysical Journal Supplement Series, 215, 19, doi: 10.1088/0067-0049/215/2/19

Pinsonneault, M. H., Elsworth, Y. P., Tayar, J., et al. 2018, The Astrophysical Journal Supplement Series, 239, 32, doi: 10.3847/1538-4365/aaebfd

Pinsonneault, M. H., Zinn, J. C., Tayar, J., et al. 2024, APOKASC-3: The Third Joint Spectroscopic and Asteroseismic Catalog for Evolved Stars in the Kepler Fields, doi: 10.48550/arXiv.2410.00102

Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., et al. 2016, Astronomy & Astrophysics, 595, A1, doi: 10.1051/0004-6361/201629272

Radford, A., Kim, J. W., Hallacy, C., et al. 2021, in Proceedings of the 38th International Conference on Machine Learning (PMLR), 8748–8763

Sanders, J. L., & Das, P. 2018, Monthly Notices of the Royal Astronomical Society, 481, 4093, doi: 10.1093/mnras/sty2490

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, The Astrophysical Journal, 500, 525, doi: 10.1086/305772

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, The Astronomical Journal, 131, 1163, doi: 10.1086/498708

Skumanich, A. 1972, The Astrophysical Journal, 171, 565, doi: 10.1086/151310

Smith, S. E. T., Cerny, W., Hayes, C. R., et al. 2024, The Astrophysical Journal, 961, 92, doi: 10.3847/1538-4357/ad0d9f

Soderblom, D. R. 2010, Annual Review of Astronomy and Astrophysics, 48, 581, doi: 10.1146/annurev-astro-081309-130806

Starkenburg, E., Martin, N., Youakim, K., et al. 2017, Monthly Notices of the Royal Astronomical Society, 471, 2587, doi: 10.1093/mnras/stx1068

Steinmetz, M., Guiglion, G., McMillan, P. J., et al. 2020, The Astronomical Journal, 160, 83, doi: 10.3847/1538-3881/ab9ab8

Thomas, G. F., Jensen, J., McConnachie, A., et al. 2020, The Astrophysical Journal, 902, 89, doi: 10.3847/1538-4357/abb6f7

Torrealba, G., Belokurov, V., Koposov, S. E., et al. 2019, Monthly Notices of the Royal Astronomical Society, 488, 2743, doi: 10.1093/mnras/stz1624

Van der Maaten, L., & Hinton, G. 2008, Journal of machine learning research, 9

VandenBerg, D. A., Brogaard, K., Leaman, R., & Casagrande, L. 2013, The Astrophysical Journal, 775, 134, doi: 10.1088/0004-637X/775/2/134

Vincent, O., Barstow, M. A., Jordan, S., et al. 2024, Astronomy and Astrophysics, 682, A5, doi: 10.1051/0004-6361/202347694

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. 2008, in Proceedings of the 25th International Conference on Machine Learning, ICML '08 (New York, NY, USA: Association for Computing Machinery), 1096–1103, doi: 10.1145/1390156.1390294

Viscasillas Vázquez, C., Solano, E., Ulla, A., et al. 2024, Astronomy and Astrophysics, 691, A223, doi: 10.1051/0004-6361/202451247

Weiler, M., Carrasco, J. M., Fabricius, C., & Jordi, C. 2023, Astronomy & Astrophysics, 671, A52, doi: 10.1051/0004-6361/202244764

Witten, C. E. C., Aguado, D. S., Sanders, J. L., et al. 2022, Monthly Notices of the Royal Astronomical Society, 516, 3254, doi: 10.1093/mnras/stac2273

Wolf, C., Onken, C. A., Luvaul, L. C., et al. 2018, Publications of the Astronomical Society of Australia, 35, e010, doi: 10.1017/pasa.2018.5

Yao, Y., Ji, A. P., Koposov, S. E., & Limberg, G. 2024, Monthly Notices of the Royal Astronomical Society, 527, 10937, doi: 10.1093/mnras/stad3775

Yong, D., Carney, B. W., & Friel, E. D. 2012, The Astronomical Journal, 144, 95, doi: 10.1088/0004-6256/144/4/95

Zhang, X., Green, G. M., & Rix, H.-W. 2023, Monthly Notices of the Royal Astronomical Society, 524, 1855, doi: 10.1093/mnras/stad1941

Zhao, H., Wang, S., Jiang, B., et al. 2024, The Astrophysical Journal, 974, 138, doi: 10.3847/1538-4357/ad6d64

Appendix A

Additional Figures

Present in this section, are all the figures associated with the application of the MSA and predictor to separate spectroscopic datasets, not shown in the main text. Details are found in Section 5.2.

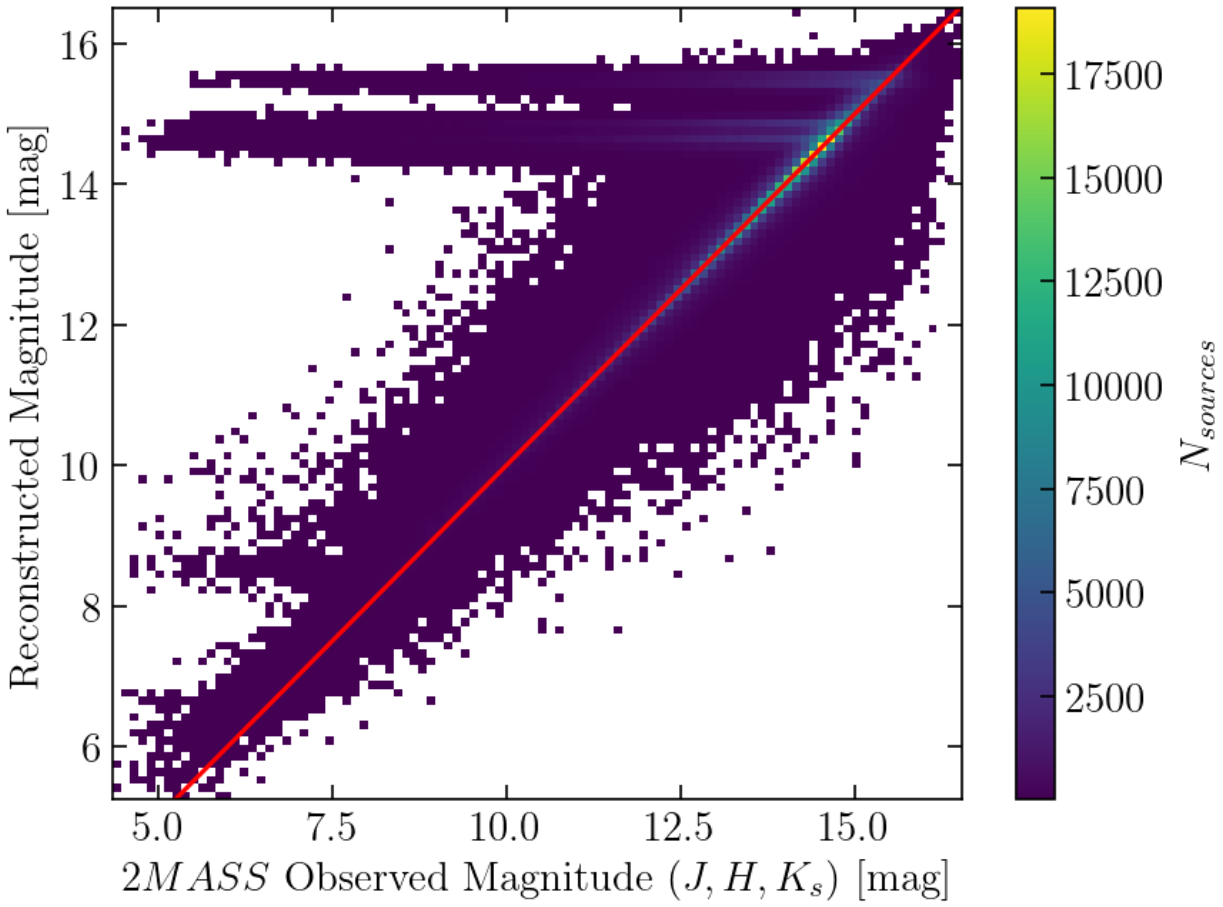

Figure A.1: The reconstructed $2MASS$ magnitudes versus the input magnitudes after pretraining the MSA for 80 epochs. All three reconstructed magnitudes (J, H, K_s) are plotted if the magnitude exists in the pretraining dataset, as the trend was apparent for all bands.

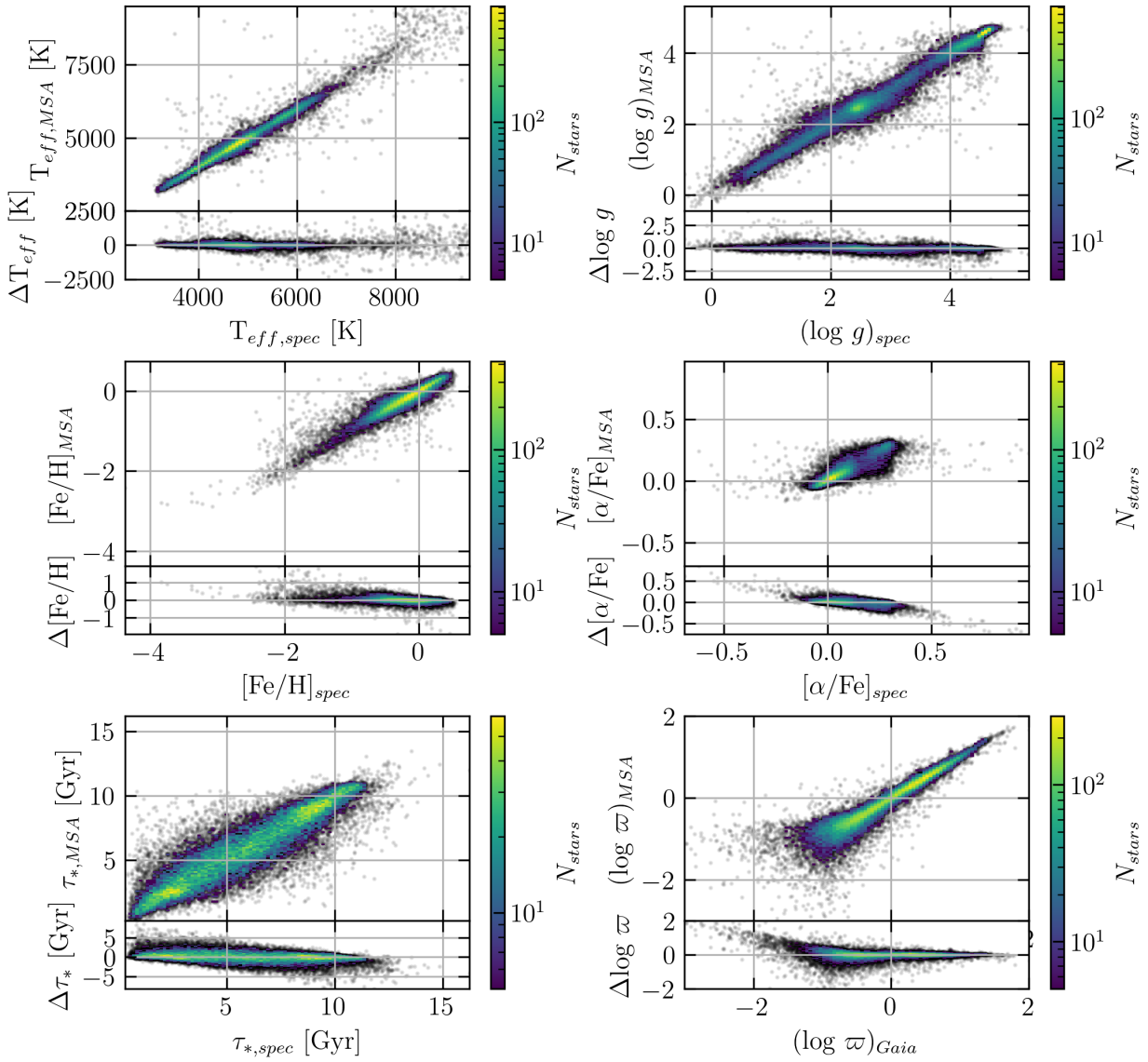

Figure A.2: The test dataset residuals after fine-tuning the model on solely APOGEE stellar labels and VMPs.

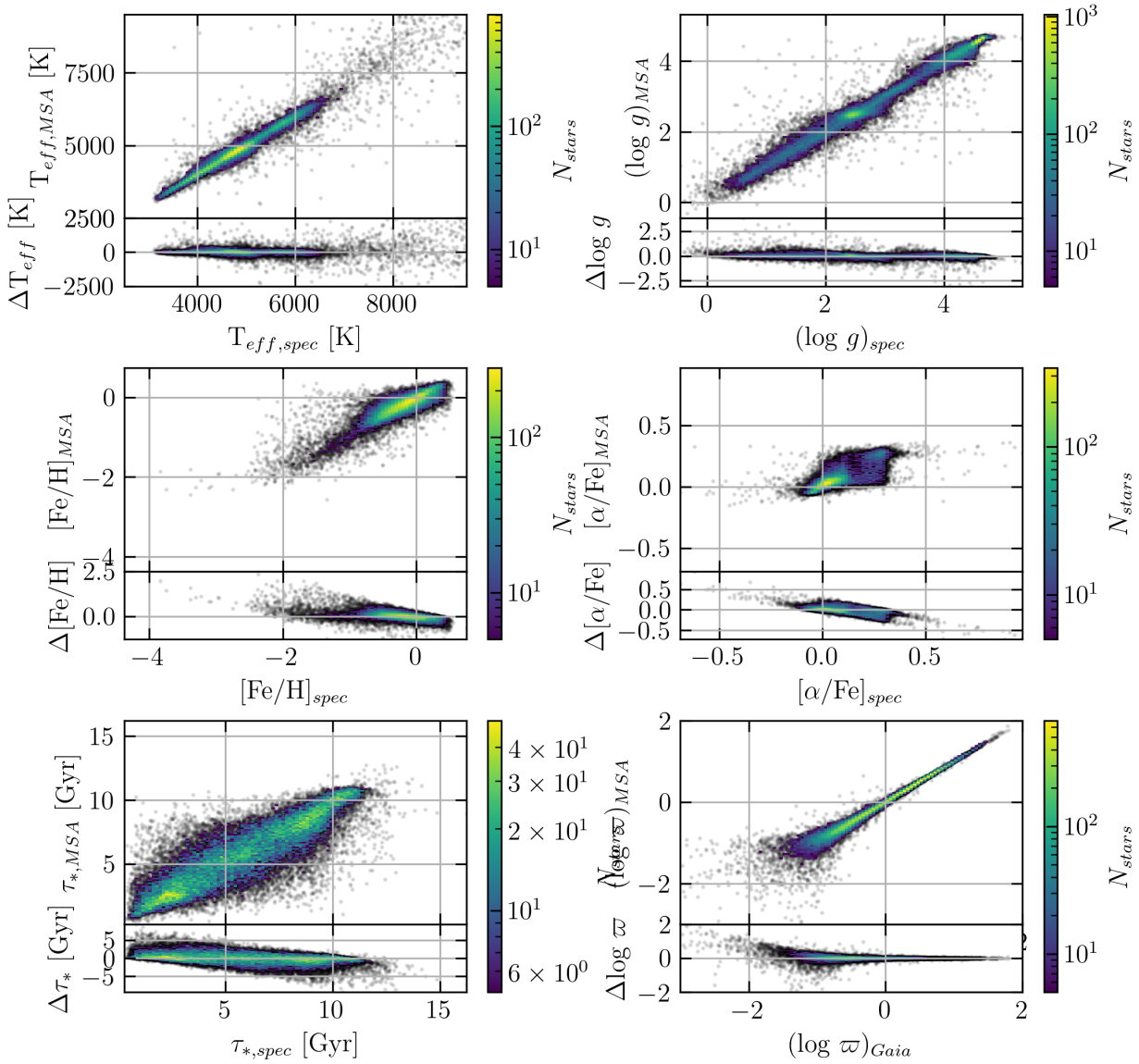

Figure A.3: The same as Figure A.2, with the XP coefficients masked upon input.

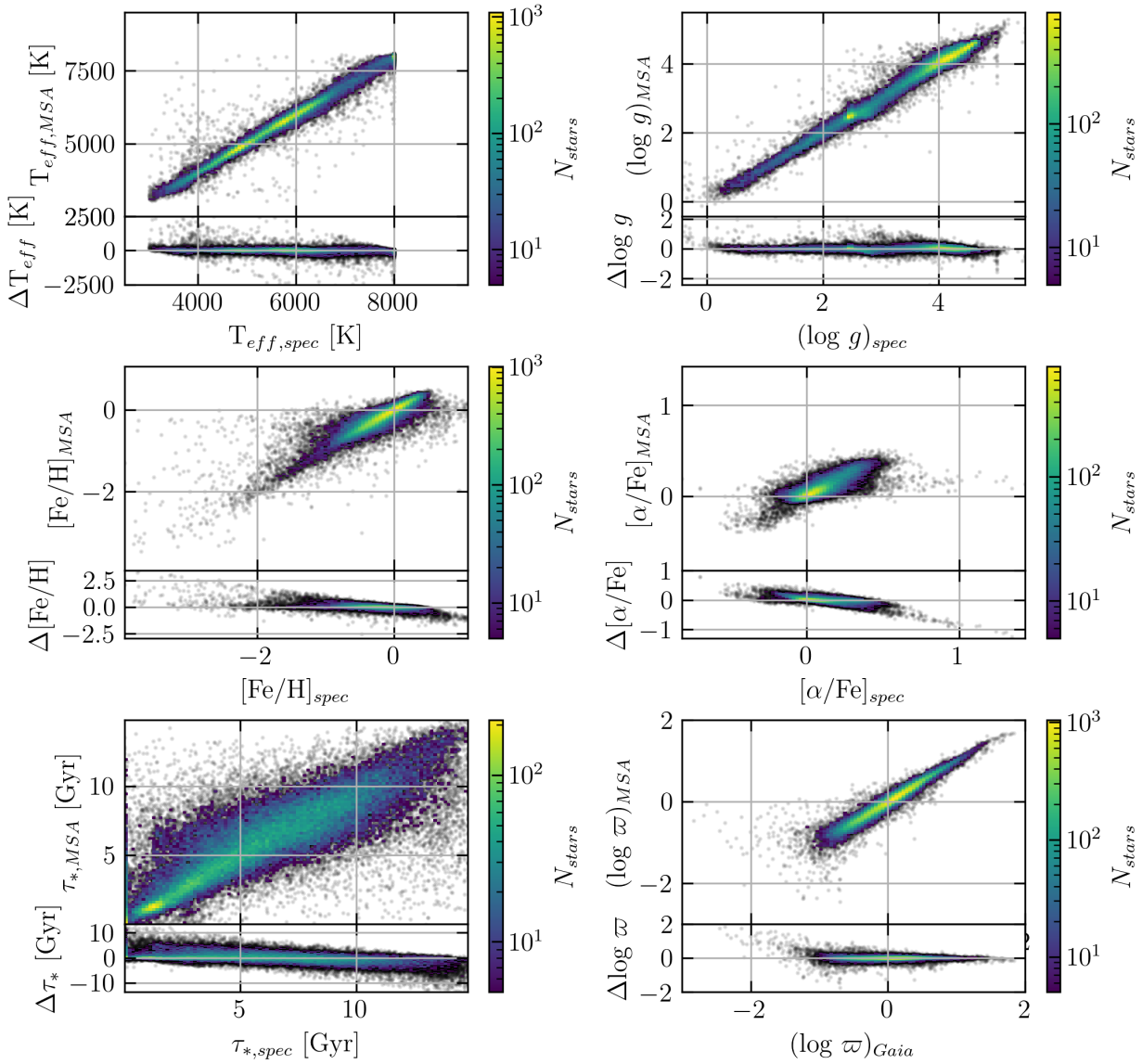

Figure A.4: The test dataset residuals after fine-tuning the model on solely GALAH DR4 stellar labels and VMPs.

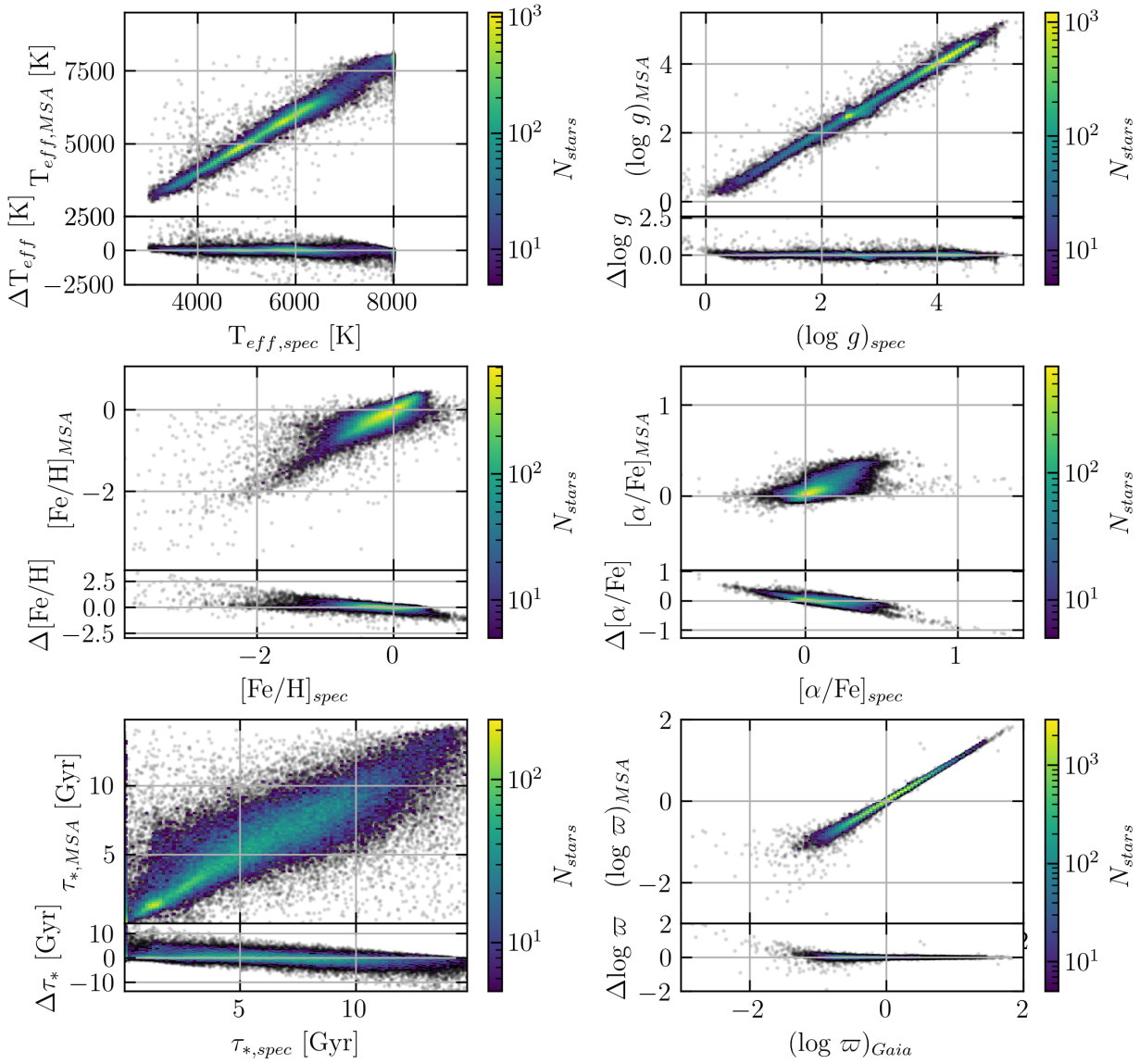

Figure A.5: The same as Figure A.4, with the XP coefficients masked upon input.

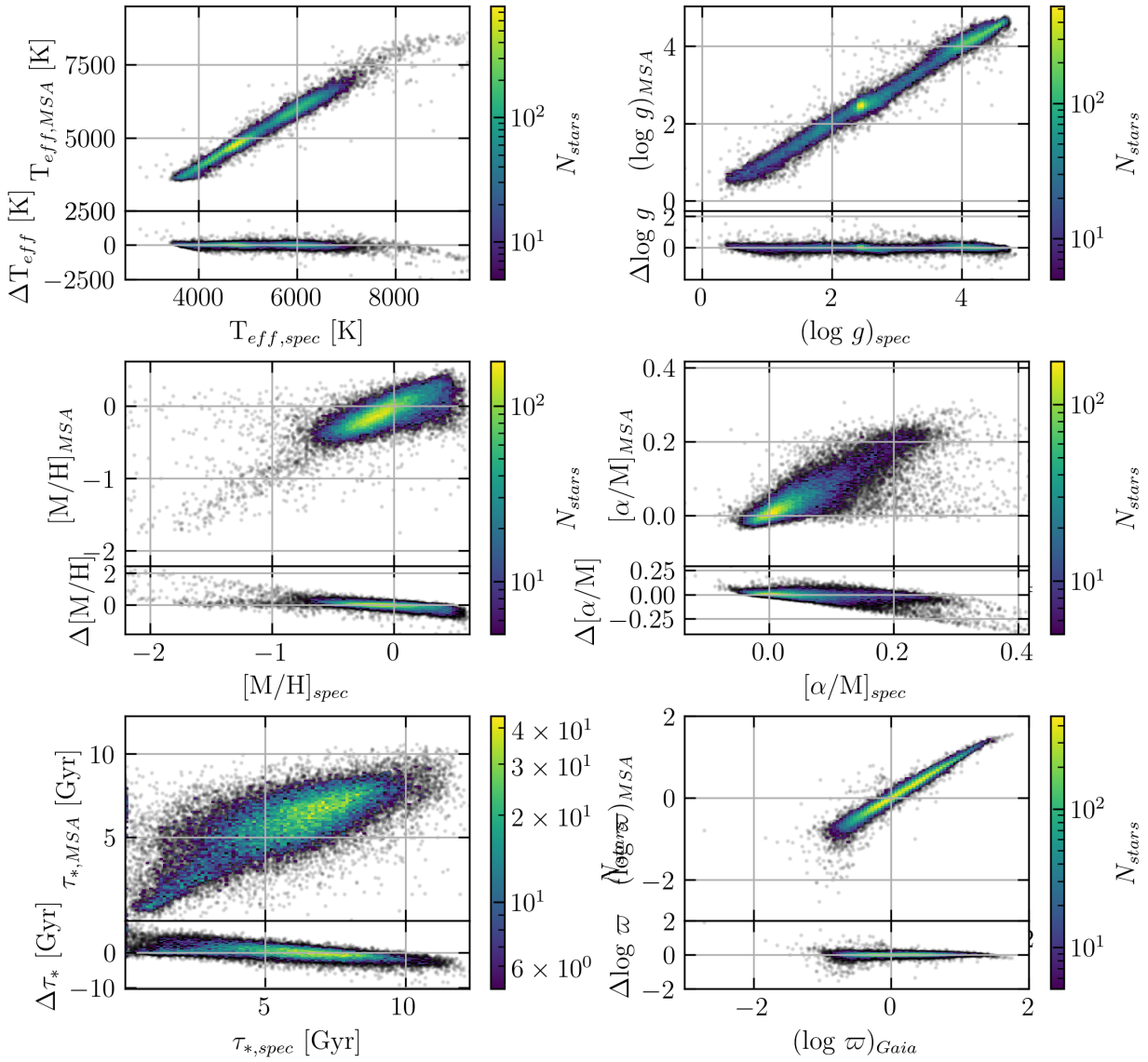

Figure A.6: The test dataset residuals after fine-tuning the model on solely RAVE DR6 spectroscopic labels.

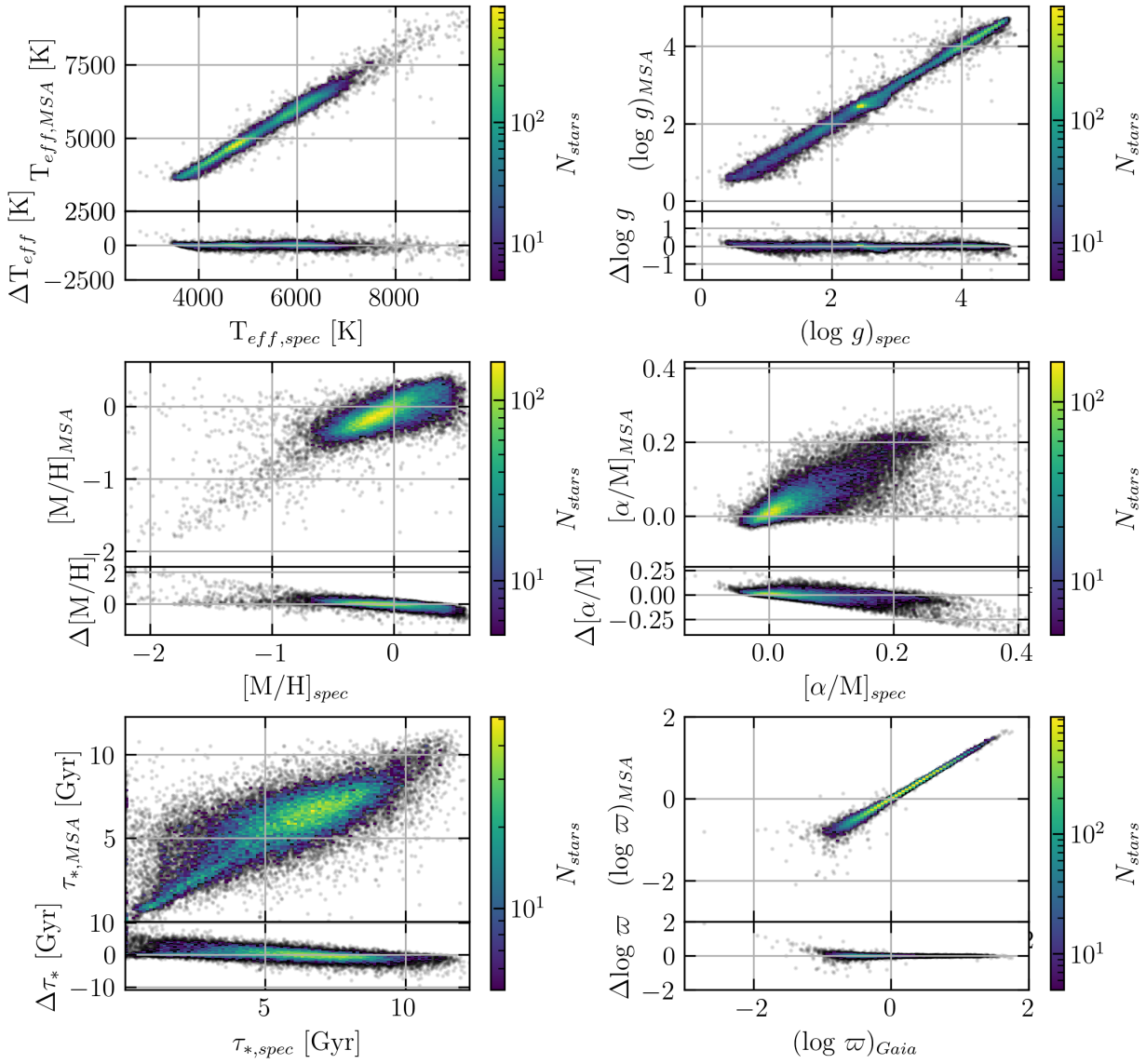

Figure A.7: The same as Figure A.6, with the XP coefficients masked upon input.