SpectraFM: Tuning into Stellar Foundation Models

Nolan Koblischke*, **Jo Bovy**
David A. Dunlap Department of Astronomy and Astrophysics
University of Toronto
50 St. George Street, Toronto, Ontario, M5S 3H4, Canada
*nolan.koblischke@astro.utoronto.ca

Abstract

Machine learning models in astrophysics are often limited in scope and cannot adapt to data from new instruments or tasks. We introduce SpectraFM, a Transformer-based foundation model architecture that can be pre-trained on stellar spectra from any wavelength range and instrument. SpectraFM excels in generalization by combining flexibility with knowledge transfer from pre-training, allowing it to outperform traditional machine learning methods, especially in scenarios with limited training data. Our model is pre-trained on approximately 90k examples of synthetic spectra to predict the chemical abundances (Fe, Mg, O), temperature, and specific gravity of stars. We then fine-tune the model on real spectra to adapt it to observational data before fine-tuning it further on a restricted 100-star training set in a different wavelength range to predict iron abundance. Despite a small iron-rich training set of real spectra, transfer learning from the synthetic spectra pre-training enables the model to perform well on iron-poor stars. In contrast, a neural network trained from scratch fails at this task. We investigate the Transformer's attention mechanism and find that the wavelengths receiving attention carry physical information about chemical composition. By leveraging the knowledge from pre-training and its ability to handle non-spectra inputs, SpectraFM reduces the need for large training datasets and enables cross-instrument and cross-domain research. Its adaptability makes it well-suited for tackling emerging challenges in astrophysics, like extracting insights from multi-modal datasets.

1 Introduction

In many scientific fields, a rapid increase in data availability has led to the wide-spread use of machine learning methods [Smith and Geach, 2023]. The machine learning models used in research are often restricted to a single task, e.g. to predict the iron content of stars from spectra, and only make accurate predictions on data from the instrument that collected its training set. Foundation models, pre-trained on multiple tasks and data from a variety of sources, have unlocked a versatility that was lacking with 'classical' machine learning algorithms [Bommasani et al., 2022]. Pre-trained foundation models can be released to scientists and used out-of-the-box, or fine-tuned for specific research tasks, in the hopes of leveraging transfer learning to outperform a machine learning algorithm trained from scratch [McCabe et al., 2023]. They are particularly useful when applied to 1) datasets too small to train an effective machine learning model [Walmsley et al., 2024] and 2) cross-instrument and cross-domain datasets that require a synergistic analysis [Parker et al., 2024].

Both situations appear frequently in astronomy. For example, spectroscopic observations from the James Webb Space Telescope (JWST) with accurate stellar property labels are insufficient in number for a training set, O($\sim$dozens), and may not represent the required diversity of stellar properties to train a useful machine learning model. However, if a foundation model were pre-trained on a much

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

larger dataset of stellar spectra, from other telescopes and synthetic sources, and then fine-tuned on the JWST dataset, it could transfer its knowledge to this new task.

Training on synthetic data for application to real data often leads to poor predictions due to the ‘synthetic gap’ - the differences between the simplified synthetic spectra and the complex observed spectra [Fabbro et al., 2018, O’Briain et al., 2021]. The synthetic gap stems from physical assumptions and idealized instruments in synthetic models, leading to systematic biases and reduced prediction accuracy when applied to real-world data. We investigate a potential remedy: fine-tuning the model on a small but well-characterized set of real spectra, allowing the foundation model to adjust to these characteristics of real observational data.

The second situation appears when the same source (e.g. a star or galaxy) has been observed by multiple instruments and in different modalities. APOGEE (Apache Point Observatory Galactic Evolution Experiment) and the Gaia space telescope are two large-scale astronomical surveys [Majewski et al., 2017, Gaia Collaboration et al., 2023]. Gaia Data Release 3 has released photometric observations, positions, and motions for more than a billion stars. APOGEE Data Release 17 provides high-resolution infrared spectra for over 650,000 stars, capturing crucial atomic lines necessary for determining the abundances of elements in these stars, vital information for stellar and galactic astrophysics. Recently, machine learning models have combined APOGEE spectra observations with Gaia observations to extract more information about stellar properties than what can be done with either alone [Cantat-Gaudin et al., 2024, Laroche and Speagle, 2024].

Leung and Bovy [2023], hereafter LB23, developed a proof-of-concept Transformer-based foundation model for stars observed by Gaia and other sources. The single model trained by LB23 can predict stellar properties like temperature, gravity, and chemical composition from low-resolution Gaia spectra or other properties, generate synthetic spectra from stellar parameters, and reconstruct missing spectral regions, which demonstrates a versatility in handling astronomical data.

Furthermore, work is underway to create a single database that encompasses dozens of surveys and modalities with the intention of developing an astronomy-wide foundation model [Lanusse et al., 2024]. However, the prototype from LB23 cannot be scaled up to accomplish this feat, since it only accepts tabular data and lacks the ability to work with other modalities, like images, time-series measurements, and high-resolution spectra. We adapt the LB23 foundation model to work with stellar spectra from any wavelength, not just the low-resolution Gaia spectra it was trained on. Extracting information from other wavelength ranges and higher resolutions is critically important for stellar astrophysics research. This involved scaling up the model to interpret spectra of a hundred-fold larger size and using a wavelength encoding scheme adapted from the positional encodings in language models. Unlocking this important modality brings us closer to a foundation model that can be applied to any instrument in stellar astrophysics research.

2 Transformer Foundation Model

Unlike regular neural networks, Transformers are flexible in terms of input size and missing inputs [Vaswani et al., 2017]. This flexibility is advantageous for developing a foundation model, as it allows the model to be adapted for a variety of tasks. LB23 developed a Transformer encoder-decoder model and trained it with 118 unique inputs, consisting of various stellar properties and observations, with a maximum input size of $N = 64$. They used a custom non-linear embedding process to encode tabular inputs:

$$y_x = f(w_x \cdot M_x) + w_{b,x} \tag{1}$$

with M_x as the value of the property ‘token’, x. $w_x, w_{b,x}$ are learnable weights unique to x, and f is a non-linear activation function. The architecture is structured such that you can provide the encoder whatever information you have about the star and then request any property by requesting vector w_x from the decoder. Our approach extends upon LB23 by introducing a wavelength encoding mechanism tailored for stellar spectroscopy with the capability to accept input spectra from across any range of wavelengths.

Spectroscopic observations consist of pixels of flux in specific wavelength bins, capturing the intensity of light within those wavelengths. There has been prior work using deep learning for APOGEE spectra to predict stellar properties and elemental abundances, notably with the `AstroNN` framework [Leung and Bovy, 2018]. It is a Convolutional Neural Network designed to work exclusively with APOGEE data and would not be useful when applied to other instruments like JWST or other wavelength ranges. In our experiments, we compare SpectraFM to `AstroNN`. Unlike `AstroNN`'s fixed input sizes and spectral regions, SpectraFM's Transformer architecture allows for generalization and is designed for multi-instrument analysis.

In order to develop a model that can work with spectra from any instrument, we input every pixel in the spectra as an individual token rather than pre-processing the spectra with a fixed input size linear layer or PCA like other approaches that process spectra with Transformers [Zhang et al., 2024b,a]. Each spectra token shares the embedding weights w_x, $w_{x,b}$, but are differentiated by a wavelength embedding added on to Equation 1 that consists of:

$$\text{Wavelength Positional Encoding:} \quad \text{PE}(\hat{\lambda}, k) = \begin{cases} \sin\left(\frac{1000\cdot\hat{\lambda}}{10000^{k/d_{\text{model}}}}\right), & \text{if } k \text{ is even} \\ \cos\left(\frac{1000\cdot\hat{\lambda}}{10000^{k/d_{\text{model}}}}\right), & \text{if } k \text{ is odd} \end{cases} \tag{2}$$

$$\hat{\lambda} = \frac{\lambda - \lambda_{\text{min}}}{\lambda_{\text{max}} - \lambda_{\text{min}}}, \quad k \in [0, d_{\text{model}} - 1] \tag{3}$$

where $d_{\text{model}} = 256$ embedding dimensions, and $\lambda_{\text{min}} = 15,000$ Å, $\lambda_{\text{max}} = 17,000$ Å were chosen for infrared spectra ranges like the APOGEE dataset we use for this prototype, though can be expanded to a larger wavelength range in a future model. This embedding scheme was inspired by Różański et al. [2023] which used it in constructing a Transformer-based model for generating synthetic spectra. This allows for inputting spectra pixels of any wavelength into our model, such that it can trained on any instrument and wavelength range.

While LB23 focused on low-resolution Gaia spectra, SpectraFM's wavelength encoding mechanism pushes the model to recognize that spectra from different instruments should contain similar information about the star at the same wavelengths. In this study we demonstrate generalization from synthetic to real spectra, but this encoding is designed to support cross-instrument generalization as well.

SpectraFM is an encoder-decoder Transformer model with approximately 8 million trainable parameters. Figure 1 illustrates its architecture. The encoder contains two Transformer blocks interspersed with dense layers. The decoder receives the encoder output along with a request, encoded as the vector w_x representing the desired stellar property. The decoder consists of three Transformer blocks and dense layers which result in the final prediction and a predicted uncertainty.

The loss function, adapted from LB23 and also used in `AstroNN`, is a mean-squared loss that combines the uncertainties in both the training data and the model's predictions. This approach allows the model to estimate a predictive uncertainty that reflects how confident it is in each prediction. The objective function L is defined as:

$$L(y, \hat{y}) = \frac{(\hat{y} - y)^2}{2e^s} + \frac{s}{2},$$

where y and $\hat{y}$ represent the ground truth and predicted values, respectively. The term $s = \ln\left(\sigma_{\text{data}}^2 + \sigma_{\text{pred}}^2\right)$ includes both the known uncertainty in the data, σ_{data}, and the predictive uncertainty of the model, σ_{pred}, which is learned during training. Minimizing L not only improves model prediction accuracy but also improves the model's ability to produce a confidence measure for each prediction that accounts for variations that neither label uncertainties nor model predictions fully capture.

3 Dataset and Training

In astrophysics, machine learning solutions often face challenges on tasks with limited availability of labeled data. For instance, while spectroscopic observations from a new telescope might be abundant, only a few dozen stars might have accurately measured properties like iron abundances ([Fe/H]). To address this, some approaches have trained machine learning models on synthetic spectra generated from theoretical simulations [Bialek et al., 2020, Fabbro et al., 2018]. However, the synthetic gap often leads to inaccurate predictions. Our solution is to fine-tune on a small dataset of real spectra only after extensively pre-training on synthetic spectra.

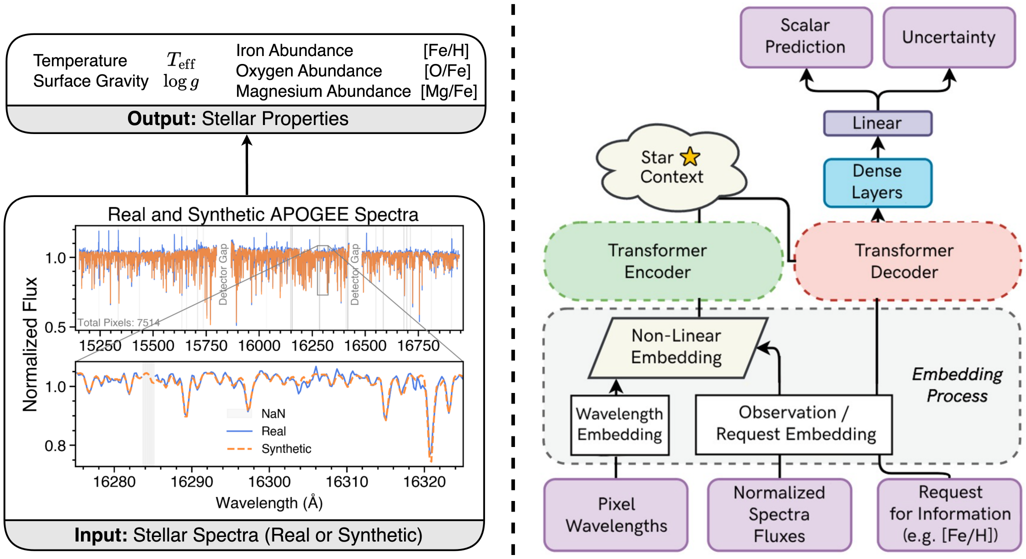

Figure 1: Overview of our Transformer-based foundation model architecture for stellar spectroscopy. *Left:* A comparison between the two possible inputs for a star: the real APOGEE infrared spectra and the synthetic best-fit spectra as generated by ASPCAP assuming simplified physics and without the observational issues that cause the noisy or NaN pixels in the real spectra. Also shown is the outputs requested from the model for each star that are important for astrophysical studies: the surface temperature and gravity, abundance of iron compared to hydrogen and abundance of oxygen and magnesium compared to iron. *Right:* The encoder processes the spectral pixels as individual tokens, embedding both flux and wavelength. This allows for input spectra from any wavelength range and instrument. The encoded spectral data forms a context about the star, which the decoder uses, along with specific output requests, to generate predictions. The model also estimates a prediction confidence. Diagram adapted from LB23.

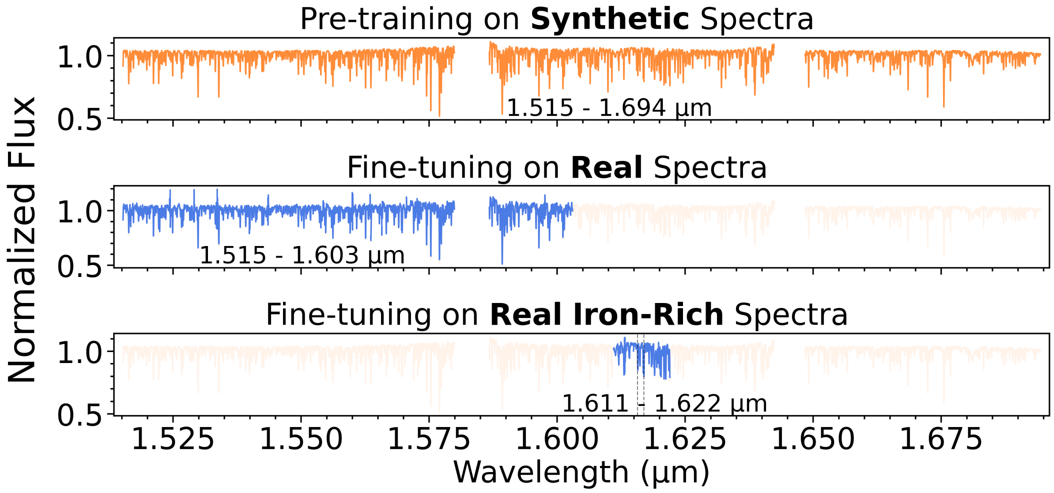

Figure 2: Wavelength regions used for each training and fine-tuning stage for this study. *Top:* Synthetic spectra used for initial pre-training. *Middle:* Real spectra used in the first fine-tuning step for comparison to `AstroNN`. *Bottom:* Real spectra used for the fine-tuning stage focusing on iron abundance prediction in a region not seen in the second step. This interval includes two prominent Fe lines and uses a limited training set of only 100 iron-rich stars to investigate transfer-learning.

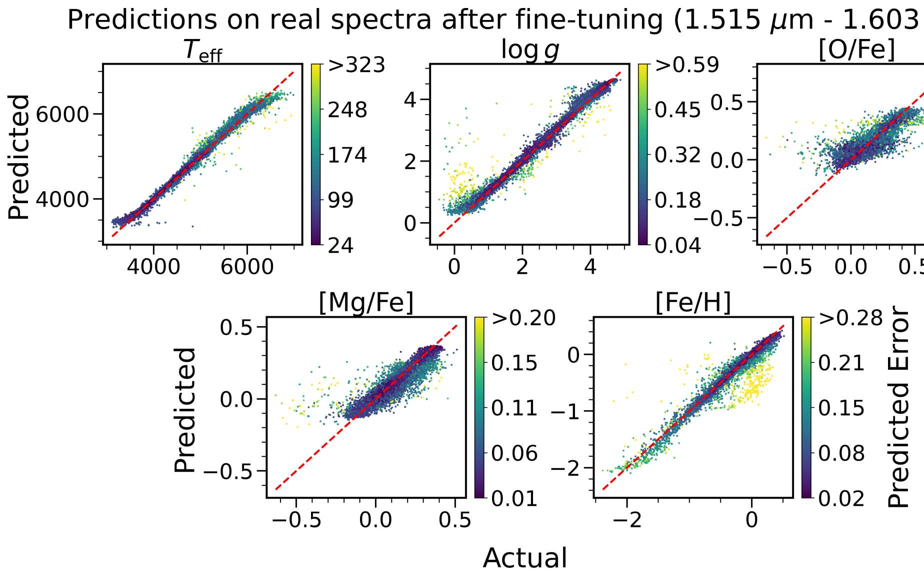

Figure 3: Stellar properties and chemical abundance predictions from real spectra using our model fine-tuned on real spectra. The point colors represent the prediction uncertainty learned by the base model. Our Transformer-based foundation model performs similarly to traditional deep learning methods like `AstroNN` [Leung and Bovy, 2018]

The synthetic spectra used in this study come from simulations by ASPCAP (APOGEE Stellar Parameters and Abundances Pipeline) [Pérez et al., 2016]. ASPCAP generates spectra under the assumption of 1D Local Thermodynamic Equilibrium (LTE), which simplifies the radiative transfer calculations in stellar atmospheres. For each real stellar spectra in APOGEE, ASPCAP releases the best-fit synthetic spectra. This synthetic spectra dataset thus has the same distribution of stellar properties and is split into a train/test set with its real spectra pairs to prevent test set leakage. The synthetic spectra are much cleaner than real observations, lacking instrumental noise and observational effects like cosmic rays, atmospheric effects and detector issues. Figure 1 highlights the differences which lead to the 'synthetic gap' that arises when training models only on synthetic spectra, which we attempt to resolve with fine-tuning on a small dataset of real spectra.

We select the following stellar properties for training: temperature ($T_{\rm eff}$), specific gravity (log g), [Fe/H], [O/Fe], and [Mg/Fe]. $T_{\rm eff}$ and log g are important physical properties for classifying stars. [Fe/H], [O/Fe], and [Mg/Fe] abundances are essential for understanding the chemical and dynamical evolution of galaxies and were listed as top-priority elements for APOGEE to detect. Since Transformer models can work effectively with missing data, we do not need to remove stars from the training set with one or more missing properties. For more details about our data refinement pipeline see Section B.

We pre-train our model on synthetic APOGEE spectra to predict combinations of $T_{\rm eff}$, log g, [Fe/H], [O/Fe], and [Mg/Fe]. Our training set consists of ∼90k synthetic stars that matches the distribution of stars seen in APOGEE. Pre-training occurs on 4x Nvidia A100 GPUs over 325 epochs, taking approximately 21 hours. The learning rate starts at 10^{-4} and varies throughout training according to a cyclic scheme known as Cosine Annealing with Warm Restarts, a common method that accelerates convergence [Loshchilov and Hutter, 2017]. Specifically, we set the initial learning rate to 10^{-4}, the minimum learning rate to 10^{-10}, and the restart length to 50 epochs. We use the Adam optimizer [Kingma and Ba, 2017] for optimization. Our model is implemented using PyTorch [Paszke et al., 2019]. At this point, to compare to `AstroNN` we fine-tune on a real spectra dataset of the same size but restrict it to the first half of the wavelength range to leave the second half for further tests focusing on transfer learning on an unseen wavelength range and fine-tuning on small datasets. The specific wavelength regions used at each step of training is illustrated in Figure 2. The maximum input size of our model is limited to 512 pixels due to computational constraints. All predictions from inputs

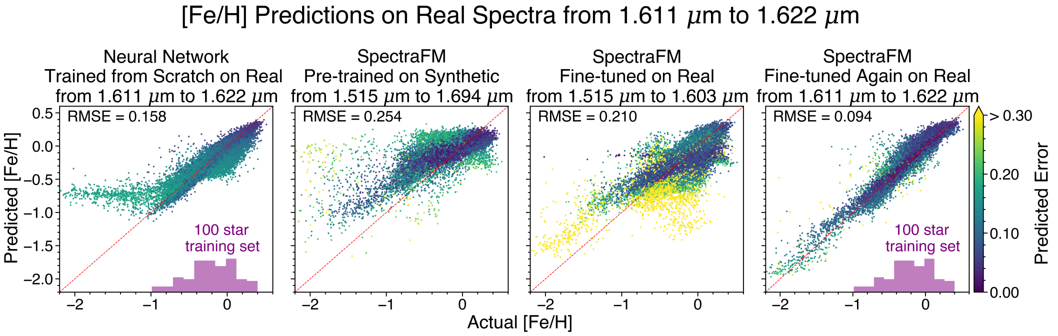

Figure 4: The [Fe/H] predictions from 512 pixels of real APOGEE spectra around two iron lines from the second half of APOGEE spectra (1.611 μm - 1.622 μm). *Left*: a basic neural network trained on real APOGEE spectra in the target wavelength range with a dataset of only 100 iron-rich stars; *Center-left*: the SpectraFM base model, pre-trained on synthetic spectra from approximately 90k stars; *Center-right:* the SpectraFM base model fine-tuned on real APOGEE spectra from only the first half of the spectra which demonstrates increased performance even though it is not trained on real spectra in the target wavelength range; *Right*: SpectraFM fine-tuned again on real APOGEE spectra in the target wavelength range with a dataset of only 100 iron-rich stars. Only SpectraFM, pre-trained on synthetic spectra and fine-tuned on 100 iron-rich real spectra, is able to generalize to iron-poor spectra for this task. This demonstrates that a pre-trained model is a better starting point than training from scratch for this small dataset task.

larger than 512 pixels are averages of predictions from 512 pixel chunks. For reference, the entire APOGEE spectra contain 7514 pixels.

Small training sets limit the accuracy of a machine learning algorithm, especially if the training set does not encompass the necessary label distribution. To mimic this scenario, we fine-tune again on a dataset limited to 100 iron-rich stars ([Fe/H] > -1) with a focus on [Fe/H] prediction. We choose the chunk of 512 pixels around two Fe lines in the second half of the spectra (1.611μm-1.622μm). For comparison, we train a fully connected neural network (FCNN) with 5 million trainable parameters, three hidden layers and dropout fraction of 10% to reduce overfitting.

We suspect that any neural network trained solely on this limited dataset, e.g. convolutional neural networks like AstroNN or larger FCNNs, will fail in generalizing to iron-poor stars ([Fe/H] < -1), due to the out-of-distribution nature of the task. In this test, there is no exposure to iron-poor stars during training. SpectraFM benefits from pre-training on synthetic spectra, which enables it to transfer knowledge to real spectra and generalize beyond the fine-tuning training distribution. If our fine-tuned model accurately predicts [Fe/H] in the iron-poor range, this indicates that knowledge transferred from the synthetic pre-train and the fine-tuning on real spectra from a different wavelength range. We can compare the [Fe/H] prediction accuracy at each stage to see how knowledge is gained.

4 Results & Discussion

The prediction accuracy of our fine-tuned model on the first half of real spectra as seen in Figure 3 and Table 1 is similar to that of previous machine learning methods like `AstroNN` [Leung and Bovy, 2018]. Training only on the synthetic spectra is insufficient to handle real spectra due to the synthetic gap (middle column of Table 1), while fine-tuning allows the model to make accurate predictions. We make a selection of $1.0 < \log g < 3.5$ on our test set to match the `AstroNN` training sample for a fair comparison.

4.1 Transfer learning from synthetic to real data

To investigate the extent to which the pre-training and the fine-tuning steps help the model generalize to previously unseen data regimes, we compare the results from our pre-trained and fine-tuned models to those from the basic neural network trained from scratch on the 100 iron-rich stars in Figure 4. The from-scratch neural network fails to predict [Fe/H] in the iron-poor range and SpectraFM pre-trained

Scatter	AstroNN	SpectraFM Pre-trained Synthetic	SpectraFM Fine-tuned Real
T_{eff}	10 K	183 K	19 K
log g	0.037 dex	0.326 dex	0.056 dex
[O/Fe]	0.020 dex	0.038 dex	0.021 dex
[Mg/Fe]	0.015 dex	0.034 dex	0.019 dex
[Fe/H]	0.011 dex	0.048 dex	0.019 dex

Table 1: Comparison of prediction scatter for a selection of the test set between AstroNN [Leung and Bovy, 2018] and our foundation model after fine-tuning on real spectra. The scatter is the median absolute deviation: median $(|y_{\mathrm{true},i} - y_{\mathrm{pred},i}|)$. Our foundation model is only fine-tuned on the first half of the APOGEE spectra while AstroNN was trained on the full spectra. Furthermore, AstroNN likely included some of these stars in its training set. These stars satisfy $1.0 <$ log $g <$ 3.5 for a fair comparison since AstroNN only trained on stars in this log g range. The AstroNN [O/Fe] and [Mg/Fe] are found from [X/Fe] = [X/H] - [Fe/H] since AstroNN does not directly predict [X/Fe].

on only synthetic spectra also performs poorly due to the simulations that generated the synthetic spectra not accurately modeling every physical process behind the real spectra (like in the middle column of Table 1). We observe an increase in [Fe/H] accuracy from $\sigma_{\mathrm{RMSE}} = 0.254$ to 0.210 by fine-tuning on real spectra in an entirely different wavelength range, which already outperforms the neural network trained from scratch in the iron-poor range ($\sigma_{\mathrm{RMSE,[Fe/H]}<-1.0} = 0.763$ vs. 0.510). Fine-tuning again, but only using observed spectra of 100 iron-rich stars, which are the only real, observed spectra used for training in this wavelength region, leads to strong performance even in the iron-poor region ($\sigma_{\mathrm{RMSE,[Fe/H]}<-1.0} = 0.232$) and is the best overall estimator ($\sigma_{\mathrm{RMSE}} = 0.094$). Skipping the first fine-tuning step and directly fine-tuning on the 100 iron-rich stars leads to similar performance: $\sigma_{\mathrm{RMSE,[Fe/H]}<-1.0} = 0.256$ and $\sigma_{\mathrm{RMSE}} = 0.100$. Therefore, although fine-tuning on a different wavelength range with real spectra increased performance on this task compared to the base model, the knowledge required to achieve a high accuracy came from the 100 star fine-tune in the same wavelength range. Even though the model never sees iron-poor real spectra in the target wavelength range during training, it can make accurate predictions in this region, which demonstrates a new ability unlocked by generalizing knowledge from other tasks.

We found that freezing the parameters in the layers closer to the output of the last fine-tuned model, i.e. the Decoder and last layer of Encoder, led to stronger performance in the iron-poor region. This suggests that the layers closer to the output retain information about translating high-level spectral features to iron abundance from the synthetic pre-training. Meanwhile, the layers closer to the input require adjustment to recognize features in real spectra. This approach allows the model to leverage the pre-trained knowledge effectively without overfitting to the small real dataset.

The ability to accurately predict [Fe/H] in the iron-poor range suggests that the synthetic-to-real knowledge transfer is successful and highlights the advantage of starting with a pre-trained foundation model, which requires only minimal adjustments to perform well across diverse conditions, such as new telescopes like JWST.

4.2 Attention

A major concern in using machine learning for scientific research arises from the limited understanding of how a prediction is made and where the model sources the information to make that prediction. The attention mechanism behind Transformers can help us understand what information the model has learned to use [Vaswani et al., 2017]. We look at the relative values of attention scores that the Encoder transformer block of our base model assigns to each input pixel when making predictions for T_{eff}, log g, [Fe/H], [O/Fe], [Mg/Fe] (see Appendix A for more info on attention scores). We average the scores for giants and dwarfs, selected based on specific gravity (log g < 3.0 for giants and log g > 4.0 for dwarfs). The results of this can be seen in Figure 5.

Stellar spectra contain many narrow absorption lines with physically meaningful information, so we invert the attention to easily compare the attention scores to the spectra and determine whether regions of high attention correspond to spectral features. The shape and depth of spectral lines contain information about the abundance of certain chemical elements along with the overall properties of the star, like T_{eff} and log g. The source of our simulated spectra, ASPCAP, identifies windows where the synthetic spectra are uniquely sensitive to the abundance of a given element [Jönsson et al., 2020]. We find that many of the wavelengths the Transformer attends to hold physical significance.

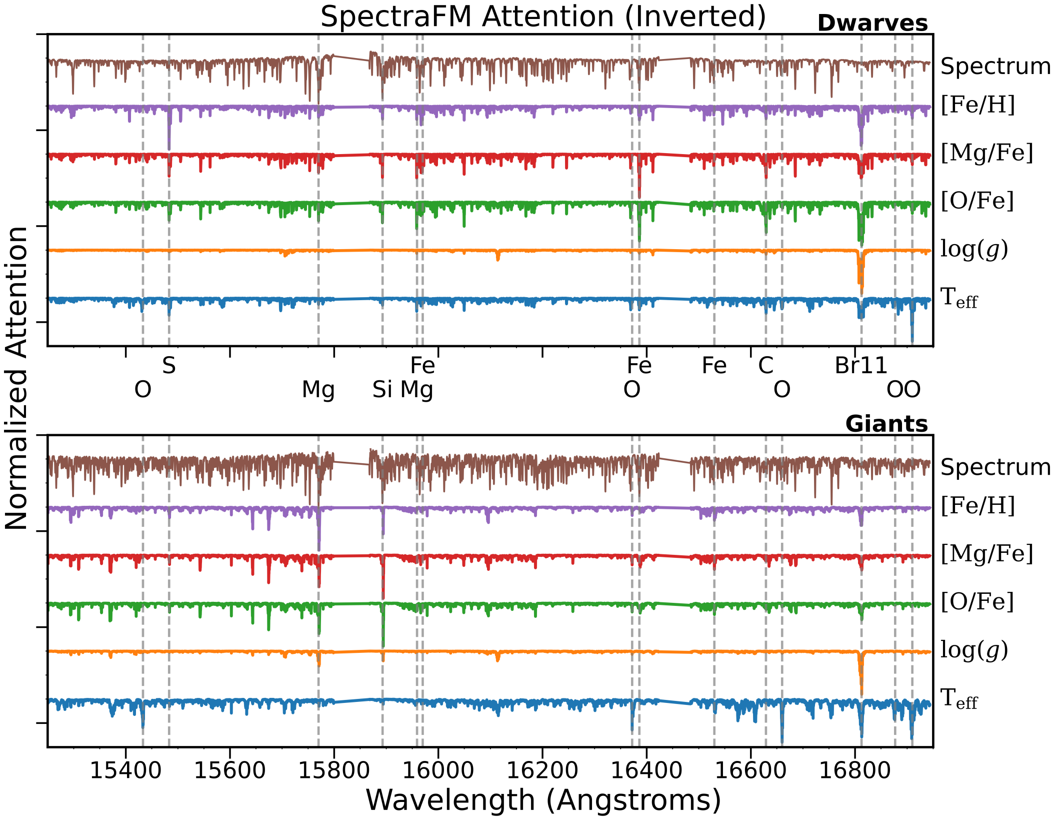

Figure 5: Inverted attention maps for synthetic spectra from dwarf (top) and giant (bottom) stars as analyzed by SpectraFM. Each row corresponds to a different stellar property: effective temperature (T_{eff}), surface gravity (log g), oxygen-to-iron ratio ([O/Fe]), magnesium-to-iron ratio ([Mg/Fe]), and iron abundance ([Fe/H]). The attention scores, normalized and averaged across stars, reveal the specific wavelength regions the model focuses on for each prediction. Spectral lines and wavelengths that contain information about certain elements are marked with dashed lines as determined by ASPCAP [Pérez et al., 2016]. The corresponding average spectrum for each stellar type is also plotted, to show how the attention aligns with physically meaningful spectral features. Br11 is a Hydrogen Brackett line [Campbell et al., 2022] known to be sensitive to log g. The attention maps have been vertically shifted for clarity. For further details on how attention scores are calculated and averaged, see Appendix A.

In particular, the attention for every property strongly focuses on the Br11 hydrogen line associated with a Brackett transition around 16813Å [Campbell et al., 2022] known to be highly sensitive to the surface gravity log g. Because determining the abundance of a chemical element from a stellar absorption line requires knowledge of the overall properties T_{eff} and log g, we expect the Br11 hydrogen line to be important for determining all elements and this is exactly what we see in Figure 5. The [Fe/H] prediction looks at lines associated with iron at 16530Å and 16386Å, with the latter catching the attention of all other property predictions as well. T_{eff} pays attention to O lines at 16660Å and 16910Å. The [Mg/Fe] and [O/Fe] attentions look similar, which is expected since Mg, O, Si, and Ca are all grouped as alpha elements, which form through the fusion of helium nuclei and are dispersed by core-collapse supernovae and so their abundances are usually correlated. Each of [Mg/Fe] and [O/Fe] pay strong attention to the Mg line at 15570Å, Si line at 15893Å, O line at 16372Å and the Fe line at 16386Å, which is necessary to determine their abundance relative to Fe.

The attention mechanism effectively identifies and focuses on spectroscopically significant regions that correspond to chemical element transitions, demonstrating the model's ability to learn physically meaningful features for accurate predictions. Previous machine learning methods like `AstroNN` employed masking techniques to mitigate the risk of the model learning spurious correlations between different elements, which could introduce bias and compromise prediction accuracy. By selectively masking specific regions during training, `AstroNN` focused on the relevant features for predicting a given element. A Transformer-based spectra foundation model has the advantage of a variable-length input and the ability to investigate attention. So future research could train the model to predict elements only based on their relevant features in the spectra, and then investigate the attention to ensure it is not basing its predictions off of correlations with other features that might not universally hold true. Furthermore, examining the model's attention when predicting a property may unveil hidden relationships not previously recognized, offering a new method for discovery.

5 Conclusion

This work presents a Transformer-based foundation model for stellar spectroscopy. The model, pre-trained on synthetic spectra and fine-tuned with limited real data, outperforms traditional neural networks by bringing out-of-distribution tasks inside the distribution with pre-training. Its attention mechanism targets key physical features in the spectra, ensuring predictions are physically grounded.

Our results show that for new tasks in astrophysics, fine-tuning a foundation model will likely lead to better results than a basic neural network. For example, if a James Webb Space Telescope data release contained a limited number of stars with measured abundances, our results suggests that fine-tuning our foundation model on this training set would lead to a highly accurate model that could then be used to get abundances for other stars. Our model could be fine-tuned to predict other properties as well, for example stellar ages, mass, and spectra-photometric distances [Leung and Bovy, 2019, Leung et al., 2023b]. Our understanding of the Galaxy's evolutionary history, like the formation of the bar, disk, and stellar halo, along with our understanding of globular clusters and dwarf galaxies, rely heavily on measuring these properties to high accuracy [e.g., Leung et al., 2023a].

Future directions include integrating diverse datasets to enhance cross-instrument and cross-domain generalization. We plan to exploit our model's flexibility by pre-training on all major stellar spectroscopic surveys such as LAMOST DR9 (10 million stars, 370-900 nm) [Liu et al., 2020], GALAH DR3 (588k stars, optical and infrared bands) [Buder et al., 2021], and Gaia DR3 low-resolution spectra (220 million stars, 330-1050 nm) [De Angeli et al., 2022], each with differing resolutions. A wider variety of training data should increase performance on all tasks due to knowledge generalization.

Few-shot learning is also an area of interest, especially for its applications for analyzing rare stellar types and datasets from new instruments that are too small for fine-tuning. Moving to a decoder-only approach with positional encoding would enable such ability.

This work lays the groundwork for developing comprehensive astronomical foundation models, which could greatly assist in data analysis in large-scale surveys.

For reproducibility and transparency, we open-source our code, training scripts, and models for SpectraFM at `https://github.com/NolanKoblischke/SpectraFM_NeurIPS_FM4Science`.

Acknowledgments and Disclosure of Funding

We thank Henry W. Leung for useful discussion. NK and JB acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC; funding reference number RGPIN-2020-04712). NK is partially funded through the NSERC Canada Graduate Scholarship - Master's.

This research was enabled in part by support provided by Compute Ontario (`https://www.computeontario.ca`) and the Digital Research Alliance of Canada (`alliancecan.ca`).

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is `https://www.sdss.org/`.

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

References

Spencer Bialek, Sébastien Fabbro, Kim A Venn, Nripesh Kumar, Teaghan O'Briain, and Kwang Moo Yi. Assessing the performance of LTE and NLTE synthetic stellar spectra in a machine learning framework. *Monthly Notices of the Royal Astronomical Society*, 498(3):3817–3834, October 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa2582. URL `https://doi.org/10.1093/mnras/staa2582`.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL `http://arxiv.org/abs/2108.07258`. arXiv:2108.07258 [cs].

Sven Buder, Sanjib Sharma, Janez Kos, Anish M. Amarsi, Thomas Nordlander, Karin Lind, Sarah L. Martell, Martin Asplund, Joss Bland-Hawthorn, Andrew R. Casey, Gayandhi M. de Silva, Valentina D'Orazi, Ken C. Freeman, Michael R. Hayden, Geraint F. Lewis, Jane Lin, Katharine J. Schlesinger, Jeffrey D. Simpson, Dennis Stello, Daniel B. Zucker, Tomaž Zwitter, Kevin L. Beeson, Tobias Buck, Luca Casagrande, Jake T. Clark, Klemen Čotar, Gary S. da Costa, Richard de Grijs, Diane Feuillet, Jonathan Horner, Prajwal R. Kafle, Shourya Khanna, Chiaki Kobayashi, Fan Liu, Benjamin T. Montet, Govind Nandakumar, David M. Nataf, Melissa K. Ness, Lorenzo Spina, Thor Tepper-García, Yuan-Sen Ting, Gregor Traven, Rok Vogrinčič, Robert A. Wittenmyer, Rosemary F. G. Wyse, Maruša Žerjal, and Galah Collaboration. The GALAH+ survey: Third data release. *Monthly Notices of the Royal Astronomical Society*, 506:150–201, September 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab1242. URL `https://ui.adsabs.harvard.edu/abs/2021MNRAS.506..150B`. Publisher: OUP ADS Bibcode: 2021MNRAS.506..150B.

Hunter Campbell, Elliott Khilfeh, Kevin R. Covey, Marina Kounkel, Richard Ballantyne, Sabrina Corey, Carlos G. Román-Zúñiga, Jesús Hernández, Ezequiel Manzo Martínez, Karla Peña Ramírez, Alexandre Roman-Lopes, Keivan G. Stassun, Guy S. Stringfellow, Jura Borissova, S. Drew Chojnowski, Valeria Ramírez-Preciado, Jinyoung Serena Kim, Javier Serna, Amelia M. Stutz, Ricardo López-Valdivia, Genaro Suárez, Jason E. Ybarra, Penélope Longa-Peña, and José G. Fernández-Trincado. Pre-main-sequence Brackett Emitters in the APOGEE DR17 Catalog: Line Strengths and Physical Properties of Accretion Columns. *The Astrophysical Journal*, 942(1):22, December 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/aca324. URL `https://dx.doi.org/10.3847/1538-4357/aca324`. Publisher: The American Astronomical Society.

Tristan Cantat-Gaudin, Morgan Fouesneau, Hans-Walter Rix, Anthony G. A. Brown, Ronald Drimmel, Alfred Castro-Ginard, Shourya Khanna, Vasily Belokurov, and Andrew R. Casey. Uniting Gaia and APOGEE to unveil the cosmic chemistry of the Milky Way disc. *Astronomy & Astrophysics*, 683:A128, March 2024. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202348018. URL `https://www.aanda.org/articles/aa/abs/2024/03/aa48018-23/aa48018-23.html`. Publisher: EDP Sciences.

F. De Angeli, M. Weiler, P. Montegriffo, D. W. Evans, M. Riello, R. Andrae, J. M. Carrasco, G. Busso, P. W. Burgess, C. Cacciari, M. Davidson, D. L. Harrison, S. T. Hodgkin, C. Jordi, P. J. Osborne, E. Pancino, G. Altavilla, M. A. Barstow, C. a. L. Bailer-Jones, M. Bellazzini, A. G. A. Brown, M. Castellani, S. Cowell, L. Delchambre, F. De Luise, C. Diener, C. Fabricius, M. Fouesneau, Y. Fremat, G. Gilmore, G. Giuffrida, N. C. Hambly, S. Hidalgo, G. Holland, Z. Kostrzewa-Rutkowska, F. van Leeuwen, A. Lobel, S. Marinoni, N. Miller, C. Pagani, L. Palaversa, A. M. Piersimoni, L. Pulone, S. Ragaini, M. Rainer, P. J. Richards, G. T. Rixon, D. Ruz-Mieres, N. Sanna, L. M. Sarro, N. Rowell, R. Sordo, N. A. Walton, and A. Yoldas. Gaia Data Release 3: Processing and validation of BP/RP low-resolution spectral data, June 2022. URL `https://arxiv.org/abs/2206.06143v1`.

S. Fabbro, K. A. Venn, T. O'Briain, S. Bialek, C. L. Kielty, F. Jahandar, and S. Monty. An application of deep learning in the analysis of stellar spectra. *Monthly Notices of the Royal Astronomical Society*, 475:2978–2993, April 2018. ISSN 0035-8711. doi: 10.1093/mnras/stx3298. URL `https://ui.adsabs.harvard.edu/abs/2018MNRAS.475.2978F`. Publisher: OUP ADS Bibcode: 2018MNRAS.475.2978F.

Gaia Collaboration, A. Vallenari, A. G. A. Brown, T. Prusti, J. H. J. de Bruijne, F. Arenou, C. Babusiaux, M. Biermann, O. L. Creevey, C. Ducourant, D. W. Evans, L. Eyer, R. Guerra, A. Hutton, C. Jordi, S. A. Klioner, U. L. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, C. Soubiran, P. Tanga, N. A. Walton, C. A. L. Bailer-Jones, U. Bastian, R. Drimmel, F. Jansen, D. Katz, M. G. Lattanzi, F. van Leeuwen, J. Bakker, C. Cacciari, J. Castañeda, F. De Angeli, C. Fabricius, M. Fouesneau, Y. Frémat, L. Galluccio, A. Guerrier, U. Heiter, E. Masana, R. Messineo, N. Mowlavi, C. Nicolas, K. Nienartowicz, F. Pailler, P. Panuzzo, F. Riclet, W. Roux, G. M. Seabroke, R. Sordo, F. Thévenin, G. Gracia-Abril, J. Portell, D. Teyssier, M. Altmann, R. Andrae, M. Audard, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, P. W. Burgess, D. Busonero, G. Busso, H. Cánovas, B. Carry, A. Cellino, N. Cheek, G. Clementini, Y. Damerdji, M. Davidson, P. de Teodoro, M. Nuñez Campos, L. Delchambre, A. Dell'Oro, P. Esquej, J. Fernández-Hernández, E. Fraile, D. Garabato, P. García-Lario, E. Gosset, R. Haigron, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, J. Hernández, D. Hestroffer, S. T. Hodgkin, B. Holl, K. Janßen, G. Jevardat de Fombelle, S. Jordan, A. Krone-Martins, A. C. Lanzafame, W. Löffler, O. Marchal, P. M. Marrese, A. Moitinho, K. Muinonen, P. Osborne, E. Pancino, T. Pauwels, A. Recio-Blanco, C. Reylé, M. Riello, L. Rimoldini, T. Roegiers, J. Rybizki, L. M. Sarro, C. Siopis, M. Smith, A. Sozzetti, E. Utrilla, M. van Leeuwen, U. Abbas, P. Ábrahám, A. Abreu Aramburu, C. Aerts, J. J. Aguado, M. Ajaj, F. Aldea-Montero, G. Altavilla, M. A. Álvarez, J. Alves, F. Anders, R. I. Anderson, E. Anglada Varela, T. Antoja, D. Baines, S. G. Baker, L. Balaguer-Núñez, E. Balbinot, Z. Balog, C. Barache, D. Barbato, M. Barros, M. A. Barstow, S. Bartolomé, J. L. Bassilana, N. Bauchet, U. Becciani, M. Bellazzini, A. Berihuete, M. Bernet, S. Bertone, L. Bianchi, A. Binnenfeld, S. Blanco-Cuaresma, A. Blazere, T. Boch, A. Bombrun, D. Bossini, S. Bouquillon, A. Bragaglia, L. Bramante, E. Breedt, A. Bressan, N. Brouillet, E. Brugaletta, B. Bucciarelli, A. Burlacu, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, T. Cantat-Gaudin, R. Carballo, T. Carlucci, M. I. Carnerero, J. M. Carrasco, L. Casamiquela, M. Castellani, A. Castro-Ginard, L. Chaoul, P. Charlot, L. Chemin, V. Chiaramida, A. Chiavassa, N. Chornay, G. Comoretto, G. Contursi, W. J. Cooper, T. Cornez, S. Cowell, F. Crifo, M. Cropper, M. Crosta, C. Crowley, C. Dafonte, A. Dapergolas, M. David, P. David, P. de Laverny, F. De Luise, R. De March, J. De Ridder, R. de Souza, A. de Torres, E. F. del Peloso, E. del Pozo, M. Delbo, A. Delgado, J. B. Delisle, C. Demouchy, T. E. Dharmawardena, P. Di Matteo, S. Diakite, C. Diener, E. Distefano, C. Dolding, B. Edvardsson, H. Enke, C. Fabre, M. Fabrizio, S. Faigler, G. Fedorets, P. Fernique, A. Fienga, F. Figueras, Y. Fournier, C. Fouron, F. Fragkoudi, M. Gai, A. Garcia-Gutierrez, M. Garcia-Reinaldos, M. García-Torres, A. Garofalo, A. Gavel, P. Gavras, E. Gerlach, R. Geyer, P. Giacobbe, G. Gilmore, S. Girona, G. Giuffrida, R. Gomel, A. Gomez, J. González-Núñez, I. González-Santamaría, J. J. González-Vidal, M. Granvik, P. Guillout, J. Guiraud, R. Gutiérrez-Sánchez, L. P. Guy, D. Hatzidimitriou, M. Hauser, M. Haywood, A. Helmer, A. Helmi, M. H. Sarmiento, S. L. Hidalgo, T. Hilger, N. Hładczuk, D. Hobbs, G. Holland, H. E. Huckle, K. Jardine, G. Jasniewicz, A. Jean-Antoine Piccolo, Ó. Jiménez-Arranz, A. Jorissen, J. Juaristi Campillo, F. Julbe, L. Karbevska, P. Kervella, S. Khanna, M. Kontizas, G. Kordopatis, A. J. Korn, Á. Kóspál, Z. Kostrzewa-Rutkowska, K. Kruszyńska, M. Kun, P. Laizeau, S. Lambert, A. F. Lanza, Y. Lasne, J. F. Le Campion, Y. Lebreton, T. Lebzelter, S. Leccia, N. Leclerc, I. Lecoeur-Taibi, S. Liao, E. L. Licata, H. E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, A. Lorca, C. Loup, P. Madrero Pardo, A. Magdaleno Romeo, S. Managau, R. G. Mann, M. Manteiga, J. M. Marchant, M. Marconi, J. Marcos, M. M. S. Marcos Santos, D. Marín Pina, S. Marinoni, F. Marocco, D. J. Marshall, L. Martin Polo, J. M. Martín-Fleitas, G. Marton, N. Mary, A. Masip, D. Massari, A. Mastrobuono-Battisti, T. Mazeh, P. J. McMillan, S. Messina, D. Michalik, N. R. Millar, A. Mints, D. Molina, R. Molinaro, L. Molnár, G. Monari, M. Monguió, P. Montegriffo, A. Montero, R. Mor, A. Mora, R. Morbidelli, T. Morel, D. Morris, T. Muraveva, C. P. Murphy, I. Musella, Z. Nagy, L. Noval, F. Ocaña, A. Ogden, C. Ordenovic, J. O. Osinde, C. Pagani, I. Pagano, L. Palaversa, P. A. Palicio, L. Pallas-Quintela, A. Panahi, S. Payne-Wardenaar, X. Peñalosa Esteller, A. Penttilä, B. Pichon, A. M. Piersimoni, F. X. Pineau, E. Plachy, G. Plum, E. Poggio, A. Prša, L. Pulone, E. Racero, S. Ragaini, M. Rainer, C. M. Raiteri, N. Rambaux, P. Ramos, M. Ramos-Lerate, P. Re Fiorentin, S. Regibo, P. J. Richards, C. Rios Diaz, V. Ripepi, A. Riva, H. W. Rix, G. Rixon, N. Robichon, A. C. Robin, C. Robin, M. Roelens, H. R. O. Rogues, L. Rohrbasser, M. Romero-Gómez, N. Rowell, F. Royer, D. Ruz Mieres, K. A. Rybicki, G. Sadowski, A. Sáez Núñez, A. Sagristà Sellés, J. Sahlmann, E. Salguero, N. Samaras, V. Sanchez Gimenez, N. Sanna, R. Santoveña, M. Sarasso, M. Schultheis, E. Sciacca, M. Segol, J. C. Segovia, D. Ségransan, D. Semeux, S. Shahaf, H. I. Siddiqui, A. Siebert, L. Siltala, A. Silvelo, E. Slezak, I. Slezak, R. L. Smart, O. N. Snaith, E. Solano, F. Solitro, D. Souami, J. Souchay, A. Spagna, L. Spina, F. Spoto, I. A. Steele, H. Steidelmüller, C. A. Stephenson, M. Süveges, J. Surdej, L. Szabados, E. Szegedi-Elek, F. Taris, M. B. Taylor, R. Teixeira, L. Tolomei, N. Tonello, F. Torra, J. Torra, G. Torralba Elipe, M. Trabucchi, A. T. Tsounis, C. Turon, A. Ulla, N. Unger, M. V. Vaillant, E. van Dillen, W. van Reeven, O. Vanel, A. Vecchiato, Y. Viala, D. Vicente, S. Voutsinas, M. Weiler, T. Wevers, Ł. Wyrzykowski, A. Yoldas, P. Yvard, H. Zhao, J. Zorec, S. Zucker, and T. Zwitter. Gaia Data Release 3. Summary of the content and survey properties. *Astronomy and Astrophysics*, 674:A1, June 2023. ISSN 0004-6361. doi:

10.1051/0004-6361/202243940. URL https://ui.adsabs.harvard.edu/abs/2023A&A...674A...1G. ADS Bibcode: 2023A&A...674A...1G.

Henrik Jönsson, Jon A. Holtzman, Carlos Allende Prieto, Katia Cunha, D. A. García-Hernández, Sten Hasselquist, Thomas Masseron, Yeisson Osorio, Matthew Shetrone, Verne Smith, Guy S. Stringfellow, Dmitry Bizyaev, Bengt Edvardsson, Steven R. Majewski, Szabolcs Mészáros, Diogo Souto, Olga Zamora, Rachael L. Beaton, Jo Bovy, John Donor, Marc H. Pinsonneault, Vijith Jacob Poovelil, and Jennifer Sobeck. APOGEE Data and Spectral Analysis from SDSS Data Release 16: Seven Years of Observations Including First Results from APOGEE-South. *The Astronomical Journal*, 160(3):120, August 2020. ISSN 1538-3881. doi: 10.3847/1538-3881/aba592. URL http://arxiv.org/abs/2007.05537. arXiv:2007.05537 [astro-ph].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Francois Lanusse, Liam Parker, Micah Bowles, and et al. Multimodaluniverse: Enabling large-scale machine learning with 70tbs of astronomical scientific data. https://github.com/MultimodalUniverse/MultimodalUniverse, 2024. Accessed: 2024-09-02.

Alexander Laroche and Joshua S. Speagle. Closing the stellar labels gap: Stellar label independent evidence for [α/M] information in $\textit{Gaia}$ BP/RP spectra, April 2024. URL https://arxiv.org/abs/2404.07316v1.

Henry W. Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, November 2018. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/sty3217. URL http://arxiv.org/abs/1808.04428. arXiv:1808.04428 [astro-ph].

Henry W Leung and Jo Bovy. Simultaneous calibration of spectro-photometric distances and the Gaia DR2 parallax zero-point offset with deep learning. *Monthly Notices of the Royal Astronomical Society*, 489 (2):2079–2096, October 2019. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stz2245. URL https://academic.oup.com/mnras/article/489/2/2079/5549526.

Henry W. Leung and Jo Bovy. Towards an astronomical foundation model for stars with a Transformer-based model, September 2023. URL http://arxiv.org/abs/2308.10944. arXiv:2308.10944 [astro-ph].

Henry W Leung, Jo Bovy, J Ted Mackereth, Jason A S Hunt, Richard R Lane, and John C Wilson. A measurement of the distance to the Galactic centre using the kinematics of bar stars. *Monthly Notices of the Royal Astronomical Society*, 519(1):948–960, February 2023a. ISSN 0035-8711. doi: 10.1093/mnras/stac3529. URL https://doi.org/10.1093/mnras/stac3529.

Henry W Leung, Jo Bovy, J Ted Mackereth, and Andrea Miglio. A variational encoder–decoder approach to precise spectroscopic age estimation for large Galactic surveys. *Monthly Notices of the Royal Astronomical Society*, 522(3):4577–4597, May 2023b. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stad1272. URL https://academic.oup.com/mnras/article/522/3/4577/7146226.

Chao Liu, Jianning Fu, Jianrong Shi, Hong Wu, Zhanwen Han, Li Chen, Subo Dong, Yongheng Zhao, Jian-Jun Chen, Haotong Zhang, Zhong-Rui Bai, Xuefei Chen, Wenyuan Cui, Bing Du, Chih-Hao Hsia, Deng-Kai Jiang, Jinliang Hou, Wen Hou, Haining Li, Jiao Li, Lifang Li, Jiaming Liu, Jifeng Liu, A.-Li Luo, Juan-Juan Ren, Hai-Jun Tian, Hao Tian, Jia-Xin Wang, Chao-Jian Wu, Ji-Wei Xie, Hong-Liang Yan, Fan Yang, Jincheng Yu, Bo Zhang, Huawei Zhang, Li-Yun Zhang, Wei Zhang, Gang Zhao, Jing Zhong, Weikai Zong, and Fang Zuo. LAMOST Medium-Resolution Spectroscopic Survey (LAMOST-MRS): Scientific goals and survey plan, May 2020. URL http://arxiv.org/abs/2005.07210. arXiv:2005.07210 [astro-ph].

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts, May 2017. URL http://arxiv.org/abs/1608.03983. arXiv:1608.03983 [cs, math].

Steven R. Majewski, Ricardo P. Schiavon, Peter M. Frinchaboy, Carlos Allende Prieto, Robert Barkhouser, Dmitry Bizyaev, Basil Blank, Sophia Brunner, Adam Burton, Ricardo Carrera, S. Drew Chojnowski, Katia Cunha, Courtney Epstein, Greg Fitzgerald, Ana E. Garcia Perez, Fred R. Hearty, Chuck Henderson, Jon A. Holtzman, Jennifer A. Johnson, Charles R. Lam, James E. Lawler, Paul Maseman, Szabolcs Meszaros, Matthew Nelson, Duy Coung Nguyen, David L. Nidever, Marc Pinsonneault, Matthew Shetrone, Stephen Smee, Verne V. Smith, Todd Stolberg, Michael F. Skrutskie, Eric Walker, John C. Wilson, Gail Zasowski, Friedrich Anders, Sarbani Basu, Stephane Beland, Michael R. Blanton, Jo Bovy, Joel R. Brownstein, Joleen Carlberg, William Chaplin, Cristina Chiappini, Daniel J. Eisenstein, Yvonne Elsworth, Diane Feuillet, Scott W. Fleming, Jessica Galbraith-Frew, Rafael A. Garcia, D. Anibal Garcia-Hernandez, Bruce A. Gillespie, Leo Girardi, James E. Gunn, Sten Hasselquist, Michael R. Hayden, Saskia Hekker, Inese Ivans, Karen Kinemuchi, Mark Klaene, Suvrath Mahadevan, Savita Mathur, Benoit Mosser, Demitri Muna, Jeffrey A. Munn, Robert C. Nichol, Robert W. O'Connell, A. C. Robin, Helio Rocha-Pinto, Matthias Schultheis, Aldo M.

Serenelli, Neville Shane, Victor Silva Aguirre, Jennifer S. Sobeck, Benjamin Thompson, Nicholas W. Troup, David H. Weinberg, and Olga Zamora. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *The Astronomical Journal*, 154(3):94, September 2017. ISSN 0004-6256, 1538-3881. doi: 10.3847/1538-3881/aa784d. URL http://arxiv.org/abs/1509.05420. arXiv:1509.05420 [astro-ph].

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple Physics Pretraining for Physical Surrogate Models, October 2023. URL http://arxiv.org/abs/2310.02994. arXiv:2310.02994 [cs, stat].

Teaghan O'Briain, Yuan-Sen Ting, Sébastien Fabbro, Kwang M. Yi, Kim Venn, and Spencer Bialek. Cycle-StarNet: Bridging the Gap between Theory and Data by Leveraging Large Data Sets. *The Astrophysical Journal*, 906(2):130, January 2021. ISSN 0004-637X. doi: 10.3847/1538-4357/abca96. URL https://dx.doi.org/10.3847/1538-4357/abca96. Publisher: The American Astronomical Society.

Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Ruben Ohana, Mariel Pettee, Bruno Regaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. AstroCLIP: A Cross-Modal Foundation Model for Galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, June 2024. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stae1450. URL http://arxiv.org/abs/2310.03024. arXiv:2310.03024 [astro-ph].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019. URL http://arxiv.org/abs/1912.01703. arXiv:1912.01703 [cs, stat].

Ana E. García Pérez, Carlos Allende Prieto, Jon A. Holtzman, Matthew Shetrone, Szabolcs Mészáros, Dmitry Bizyaev, Ricardo Carrera, Katia Cunha, D. A. García-Hernández, Jennifer A. Johnson, Steven R. Majewski, David L. Nidever, Ricardo P. Schiavon, Neville Shane, Verne V. Smith, Jennifer Sobeck, Nicholas Troup, Olga Zamora, Jo Bovy, Daniel J. Eisenstein, Diane Feuillet, Peter M. Frinchaboy, Michael R. Hayden, Fred R. Hearty, Duy C. Nguyen, Robert W. O'Connell, Marc H. Pinsonneault, David H. Weinberg, John C. Wilson, and Gail Zasowski. ASPCAP: The Apogee Stellar Parameter and Chemical Abundances Pipeline. *The Astronomical Journal*, 151(6):144, May 2016. ISSN 1538-3881. doi: 10.3847/0004-6256/151/6/144. URL http://arxiv.org/abs/1510.07635. arXiv:1510.07635 [astro-ph].

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do Vision Transformers See Like Convolutional Neural Networks?, March 2022. URL http://arxiv.org/abs/2108.08810. arXiv:2108.08810 [cs, stat].

Tomasz Różański, Yuan-Sen Ting, and Maja Jabłońska. Toward a Spectral Foundation Model: An Attention-Based Approach with Domain-Inspired Fine-Tuning and Wavelength Parameterization, June 2023. URL http://arxiv.org/abs/2306.15703. arXiv:2306.15703 [astro-ph].

Michael J. Smith and James E. Geach. Astronomia ex machina: a history, primer and outlook on neural networks in astronomy. *Royal Society Open Science*, 10:221454, May 2023. doi: 10.1098/rsos.221454. URL https://ui.adsabs.harvard.edu/abs/2023RSOS...1021454S. ADS Bibcode: 2023RSOS...1021454S.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, June 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv170603762V. Publication Title: arXiv e-prints ADS Bibcode: 2017arXiv170603762V.

Mike Walmsley, Micah Bowles, Anna M. M. Scaife, Jason Shingirai Makechemu, Alexander J. Gordon, Annette M. N. Ferguson, Robert G. Mann, James Pearson, Jürgen J. Popp, Jo Bovy, Josh Speagle, Hugh Dickinson, Lucy Fortson, Tobias Géron, Sandor Kruk, Chris J. Lintott, Kameswara Mantha, Devina Mohan, David O'Ryan, and Inigo V. Slijepevic. Scaling Laws for Galaxy Images, April 2024. URL http://arxiv.org/abs/2404.02973. arXiv:2404.02973 [astro-ph].

Gemma Zhang, Thomas Helfer, Alexander T. Gagliano, Siddharth Mishra-Sharma, and V. Ashley Villar. Maven: A Multimodal Foundation Model for Supernova Science, August 2024a. URL https://ui.adsabs.harvard.edu/abs/2024arXiv240816829Z. Publication Title: arXiv e-prints ADS Bibcode: 2024arXiv240816829Z.

Mengmeng Zhang, Fan Wu, Yude Bu, Shanshan Li, Zhenping Yi, Meng Liu, and Xiaoming Kong. SPT: Spectral transformer for age and mass estimations of red giant stars. *Astronomy and Astrophysics*, 683:A163, March 2024b. ISSN 0004-6361. doi: 10.1051/0004-6361/202347994. URL https://ui.adsabs.harvard.edu/abs/2024A&A...683A.163Z. ADS Bibcode: 2024A&A...683A.163Z.

A Attention Mechanism in Transformers

The attention mechanism enables the model to focus on the relevant spectral features for each prediction. In the model, the self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent the query, key, and value matrices derived from the input tokens (spectral pixels) [Vaswani et al., 2017]. The attention scores α_{ij} between query Q_i and key K_j are defined as:

$$\alpha_{ij} = \frac{\exp\left(\frac{Q_i \cdot K_j}{\sqrt{d_k}}\right)}{\sum_k \exp\left(\frac{Q_i \cdot K_k}{\sqrt{d_k}}\right)}$$

These normalized scores determine the relative importance of each key-value pair. We present attention scores from the second layer of the encoder, as these higher layers capture more complex patterns rather than simple structures [Raghu et al., 2022]. This is analogous to convolutional neural networks, where deeper layers detect meaningful features (e.g. the shape of spectral lines) while lower layers identify simpler structures (e.g., edges or peaks). Our attention analysis in Section 4.2 aims to determine if the model identifies physically relevant wavelengths and the information stored in spectral lines at these wavelengths. The attention scores are averaged across all attention heads and across all stars within each category (dwarfs or giants) to smooth out individual variations and emphasize the key regions that the model consistently focuses on.

B Data Refinement for APOGEE DR17 Spectra

To prepare a high-quality training dataset, we applied a series of selections to the APOGEE DR17 spectroscopic dataset [Majewski et al., 2017]. First, we selected stars with a signal-to-noise ratio above 200, ensuring reliable measurements. Next, we removed stars flagged for quality issues and observational problems. We then eliminated binaries by filtering stars with high radial velocity scatter ($v_{\rm scatter} < 1$km/s) since this often indicates the source is actually a binary star. For stars with multiple observations, we deduplicated the data by selecting the highest-SNR observation for each star. These steps reduced the dataset from 733,901 spectra to 128,762. Our synthetic spectra dataset is sourced from the best-fit synthetic spectra for each of these stars found by ASPCAP [Pérez et al., 2016]. We divided the refined set into 70% for training, 20% for testing, and 10% for training validation. Stars with NaN labels remained in the dataset, as our loss function and Transformer model handles missing data effectively during training.