Deep Multimodal Representation Learning for Stellar Spectra

Christian Schwarz, Tobias Buck

Interdisciplinary Center for Scientific Computing, Heidelberg University,
Im Neuenheimer Feld 205, D-69120 Heidelberg
`christian.schwarz@stud.uni-heidelberg.de,`
`tobias.buck@iwr.uni-heidelberg.de`

Abstract

Recently, contrastive learning (CL), a technique most prominently used in natural language and computer vision, has been used to train informative representation spaces for galaxy spectra and images in a self-supervised manner. Following this idea, we implement CL for stars in the Milky Way, for which recent astronomical surveys have produced a huge amount of heterogeneous data. Specifically, we investigate Gaia XP coefficients and RVS spectra. Thus, the methods presented in this work lay the foundation for aggregating the knowledge implicitly contained in the multimodal data to enable downstream tasks like cross-modal generation or fused stellar parameter estimation. We find that CL results in a highly structured representation space that exhibits explicit physical meaning. Using this representation space to perform cross-modal generation and stellar label regression results in excellent performance with high-quality generated samples as well as accurate and precise label predictions.

1 Motivation and Related Work

Over the past few decades, large-scale astronomical surveys of Milky Way stars have proliferated, including missions like Gaia [13], APOGEE [25], GALAH [7, 8], and LAMOST [38], with upcoming efforts like 4MOST [10]. Each survey typically develops its own pipeline, requiring post-hoc calibration to reconcile discrepancies across different data sets [e.g. 30]. This highlights the need for techniques capable of integrating data from diverse observational setups to maximize scientific output from heterogeneous datasets. Recently, machine learning (ML) has enabled cross-survey analysis by harmonizing diverse datasets. For instance, AspGap [23] aligns Gaia XP data with APOGEE, enabling the use of the APOGEE pipeline on Gaia observations. Similarly, ASTROCLIP [20] generates low-dimensional vector representations of multi-band images and spectra from the Dark Energy Spectroscopic Instrument (DESI), effectively mapping objects to shared latent spaces for cross-modal analysis. Moreover, [21] leverage a Transformer-based model trained on Gaia XP spectra and APOGEE stellar parameters to facilitate information transfer across different observation sets, exemplifying the emerging trend of foundation models in astronomy. Recent works in astrophysics also reflect the growing adoption of multimodal ML methods [37, 17, 14, 24, 34, 1, 28], with foundation models drawing inspiration from advances in natural language processing and computer vision [31, 21]. The objective of this work is to explore how deep representation learning can be applied to multimodal stellar spectra to: (i) generate informative representations from varied stellar spectra observations, and (ii) evaluate these representations through three tasks: stellar type classification, regression of stellar parameters, and cross-survey data generation. We also emphasize the scalability of these methods to a range of modalities. To achieve this, we focus on Contrastive Learning (CL) due to its demonstrated effectiveness in downstream tasks across natural language and computer vision, as evidenced by the success of cross-modal models like CLIP [29]. CL has also shown promise in galaxy surveys [20], is supported by strong theoretical foundations for generating robust representations [33], and is well-suited for scalable multimodal learning.

2 Method: Multimodal Machine Learning and Contrastive Learning

Theoretical considerations attribute multimodal approaches with some advantages relative to unimodal algorithms, e.g. better training with less data [19]. A common understanding of a *modality* is to associate it with one specific sensory input, that has been acquired in one specific way, as opposed to any other. Thus, *multimodality* describes a research problem or dataset that incorporates multiple different modalities. Currently, the most prevalent form of multimodality includes natural language and vision, while historically it's beginnings lie in audio-visual speech recognition [36]. Typical tasks for multimodal ML are: representation, translation, alignment, fusion or co-learning [4], see section A.1 in the appendix for more explanation on these terms.

Contrastive Learning Implementation   We implement our model as a coordinated representation by employing one encoder per modality and coordinating the representations via a similarity-based loss (InfoNCE [32] / NT-Xent [9]). Following CLIP [29], we implement CL as in ALIGN [18] such that the loss function takes the form:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_i^N log\frac{exp(x_i^T y_i/\tau)}{\sum_{j=1}^N exp(x_i^T y_j/\tau)} - \frac{1}{N}\sum_i^N log\frac{exp(y_i^T x_i/\tau)}{\sum_{j=1}^N exp(y_i^T x_j/\tau)} \tag{1}$$

where x_i are L2-normalized embeddings of one modality, y_i are L2-normalized embeddings of the other modality and τ is a temperature hyper-parameter. The second term switches the role of the modalities in the first term. As eq. 1 implies, the similarity between samples is calculated as cosine similarity. The temperature hyper-parameter was introduced in [9] to help the model learn from hard negatives - negatives closer to the anchor than the anchor positive. Working with more than two modalities commonly entails a linear combination of pairwise contrastive losses. In detail, for every training batch (16384 spectra), the data is run through the modality specific encoders and L2-normalized, positioning the embeddings on a hypersphere. From here, a similarity matrix is computed, which calculates the dot-product of every cross-modality instance combination scaled by the temperature τ from eq. 1. The diagonal elements correspond to positive pairings; every other element to a negative pairing. The loss (eq. 1) can be calculated by applying multi-class softmax cross-entropy (also called categorical cross-entropy) to every row, where only the diagonal entries (positive pairs) are considered as correct class predictions. To calculate the symmetrizing second term of eq. 1, the same calculation is repeated with a transposed similarity matrix and both loss terms are added for the total loss. We use the LAMB optimizer [35] because of its higher performance capabilities on bigger batches. At the start of every epoch, the training set is shuffled to avoid that any data instance would always only be compared to a fixed subset (the corresponding batch) of other instances. We choose a fixed temperature of $\tau = 0.1$ as in [20] (but see also [29] for a trainable temperature). Note, the results presented here depend on the exact temperature value chosen. We have explored different temperature values and found $\tau = 0.1$ to perform best. For the latent size, the physical dimensionality is taken into account–eight stellar parameters and up to 20 abundances– which we round to the next highest power of two, i.e. 32. The hyper-parameters for CL training are summarized in the appendix 1.

Network Architectures   Two different data formats are encountered – a spectrum and spectral coefficients. The physics of spectra suggests strong correlations between adjacent bins. RVS spectra are encoded by convolutional neural networks (CNNs) while spectral coefficients are encoded by a 1-layer MLP. The *1-layer MLP* consists of an input layer, one hidden layer and an output layer, the biases of the linear combinations are set to zero and the non-linear function is Leaky-ReLu (0.01). Additionally, to combat overfitting, Dropout is used between the input and hidden layer with dropout probability $p_{\mathrm{dpo}} = 0.2$. Dropout is only active in the training phase, in the evaluation phase no output is masked. Entries, which are not masked out are scaled by $1/(1 - p_{\mathrm{dpo}})$ to keep the expected output for training the same as for testing. After an extensive hyper-parameter search we choose for the XP 1-Layer MLP a hidden layer size of 1024.

Our *Convolutional Neural Network* is adopted from the RVS-CNN by [15], specifically only the RVS processing part, which is then followed by a custom feed-forward network. Again, all layers are implemented without bias terms and the beginning and end of the layer inputs are padded by zeros, to allow the kernel to center on the positions of the edge elements. Convolutions are implemented in parallel, which in turn increases the channel dimensionality. Dimensionality reduction is achieved by max-pooling layers. Every layer, except the output and max pooling layers, implements Leaky-ReLu (0.01) as activation. The final feed-forward network implements layer-wise dropout with a dropout rate of 20%. For a summary of the encoder architectures see Fig. 4 in the appendix.

Our *cross-modal generation decoder* follows [27] which implement a MLP decoder for galaxy spectra, which we transfer to the RVS spectra. Key feature is an activation function of the form

$$f(x) = \left[\gamma + \frac{1 - \gamma}{(1 + \mathrm{e}^{-\beta \odot x})}\right] \odot x \qquad (2)$$

where γ and β are trainable parameters [2]. This activation is able to cover smooth features for small β and sharp changes in gradient for $\beta \to \infty$ which supports easier modeling of spectral lines. For XP coefficients we use a simple MLP.

Code and Hardware All code is implemented in the python library `JAX` [6]. Neural networks are build with the `Flax` [16] package. Training routines, like optimizer and learning rate scheduling, are supported by `Optax`, while saving and loading network parameters is done by `Orbax` [11]. Datasets are handled by the Huggingface `datasets` package [22]. Training and inference takes place on a single Nvidia A100 with 40 GB of VRAM.

Dataset We use Gaia DR3 RVS spectra and Gaia XP coefficients of 841,300 instances from [15]. This dataset is ideally suited for the task of exploring representation learning as the RVS spectra represent a relatively small dataset of high quality and high information content while the XP coefficients encompass a much larger dataset of low quality data. Being able to transfer labels and find a common representation space for these two data modalities nicely showcases the benefit of representation learning for stellar spectra analysis. Furthermore, for most of these instances there are spectral types available from Gaia's Extended Stellar Parametrizer for Emission-Line Stars (ESP-ELS) [12], as well as stellar parameters from GSP-Spec [30] and GSP-Phot [3].

For an additional 44,780 samples, APOGEE labels are available. The RVS spectra have been z-scale normalized to facilitate training the neural networks. Additionally, all BP & RP coefficients have been divided by the first coefficient, like in [15]. Afterwards, the first coefficients of BP & RP have been log-scaled to bring the distribution closer to a normal distribution, and then z-scaled as well. All other XP entries are left in their unnormalized state, as their absolute magnitudes contain information on their relevance. The 841,300 instances without APOGEE labels are used as the training set. While, the 44,780 instances with APOGEE labels constitute the validation set, to facilitate downstream regression tasks for validation.

3   Results

Structured Embedding Space For a qualitative inspection of the information content and structuredness of the embedding space we apply UMAP [26] to generate a 2-d visualization of the 32-d embedding space. UMAP parameters are listed in the appendix **??**. We color the resulting UMAP embedding for the XP training data by spectral type (Fig. 1) and stellar parameters T_{eff}, $\log g$ and the abundance of all metals ($[M/H]$) and α-elements ($[\alpha/M]$, Fig. 5 in the appendix). We have further analysed the RVS training set embeddings (see Fig. 5 in the appendix) and find the same results as for XP training data which we show here. All projections colored by spectral type show large homogeneously colored regions. This implies that in the original embedding space, objects with the same type share a common neighborhood. Moreover, some classes continuously transition into others, like K & M, F & G, B & A. These pairings also form more
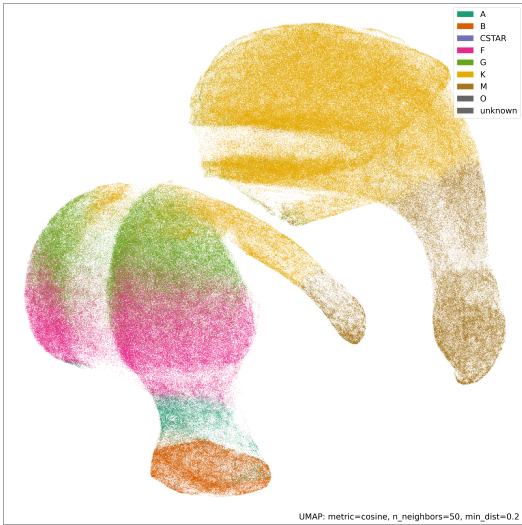

Figure 1: UMAP visualization of the XP training set embeddings with parameters `metric=cosine`, `n_neighbors=50`, `min_dist=0.2`; colored by spectral type.

separated clusters from another. Thus leading to separations from G to K and A to F. Additionally, there is some kind of duplication, in that there are two big clusters with a similar color sequence and subclusters. Overall, it is noteworthy that the transitions between classes coincide with the physically informed stellar sequence: OBAFGKM. An inspection of the projections colored by stellar
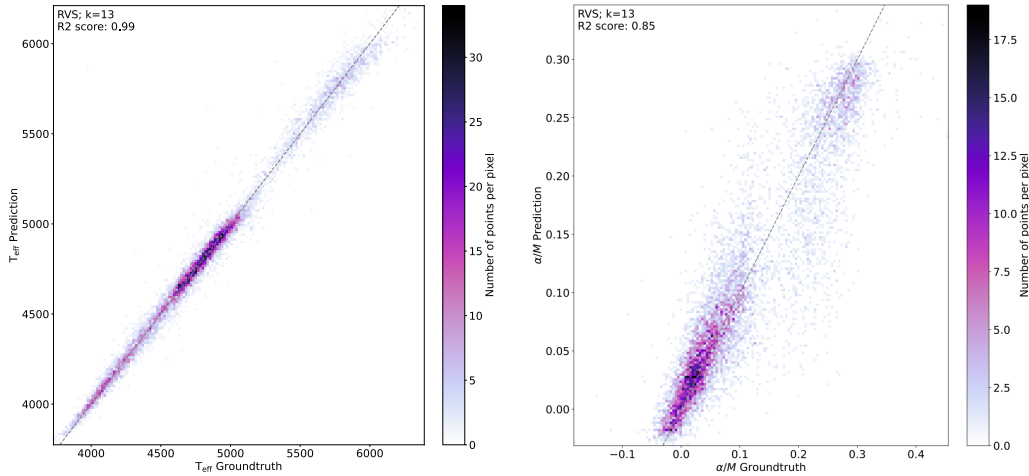
Figure 2: Prediction-vs-ground-truth plots for k-NN predictions (k=13) from RVS and XP for T_{eff} and $[\alpha/M]$; grey line corresponds to the one-to-one line.

parameters (see Fig. 6 in the appendix) shows continuous transitions between parameter values. This is most notable for the effective temperature T_{eff}, the surface gravity $\log g$ and the abundance of α-elements $[\alpha/M]$. Still, also for $[M/H]$ homogeneous color patches are visible, if slightly more interspersed in some regions for the training set. Additionally, we find two separate patches where $T_{\text{eff}} \sim 4500$ K (spectral class M) is similar but $\log g$ is different (2 vs. 4) which might point to a physical separation of dwarfs and giants. All together, the UMAP projections indicate, that the embedding space encodes physically meaningful information.

Zero-shot regression with k-Nearest Neighbours Next, zero-shot regression onto stellar parameters is performed and the R^2-score, also called coefficient of determination, for each of the stellar parameters is calculated. In Fig. 2 we show results for T_{eff} and $[\alpha/M]$ for $k = 13$. The effective temperature achieves the highest score with 0.9874, while the α-element abundance performs worst with a R^2-score of 0.8488. We find that T_{eff} shows only a small spread over the whole temperature domain, while $[\alpha/M]$ exhibits larger errors between 0.1 and 0.2 dex. In general a high R^2-score from a k-NN algorithm was again to be expected, since the visual inspection in Fig. 6 also indicated homogeneous neighbourhoods with respect to stellar parameters.

Cross-modal retrieval The four closest entries based on the cosine similarity are retrieved to perform cross-modal retrieval. In the appendix, Fig. 7 shows RVS spectra retrieved by XP coefficients, and Fig. 8 XP coefficients retrieved by RVS spectra. Note, that in the case of spectra the retrieved neighbors show variation in the noisy plateau area. Meanwhile, in the area of spectral lines the retrieved entries are mostly placed directly behind the ground truth. For the XP coefficients the variation is higher, with some neighbors in proximity of the ground truth, but not all.

Cross-modal generation Using the cross-modal decoder, we investigate the performance of cross-modal generation of the more complicated modality (RVS) from the simpler one (XP). Example RVS spectra generated from XP coefficient embeddings together with the ground truth spectra and the respective residual are shown in Fig. 3. This analysis shows that the residuals are on the magnitude of the measurement error and that key features of the spectra such as strong and weak absorption lines are well reproduced.

4 Summary and Outlook

In this work, we have used contrastive learning, a state-of-the-art multimodal ML algorithm, to generate informative representations of stellar spectra from multi-modal data, namely Gaia RVS and XP spectra. We map the raw data into a shared representation space using a CNN for RVS and a 1-layer MLP for XP. We find that multi-modal learning creates a highly structured latent embedding of the stellar spectra that aligns well with fundamental stellar parameters. We show the information content of this embedding space for downstream tasks such as regression, classification and cross-modal retrieval via simple k-Nearest Neighbor search. Moreover,
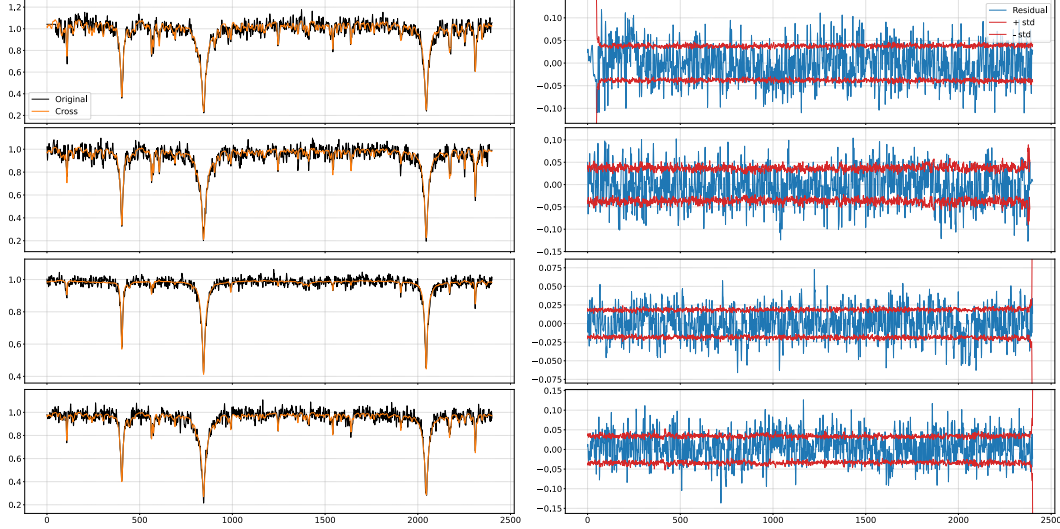
Figure 3: Four samples of cross-modal generation with fusion scheme Single Fusion; on the right side the residuals are plotted against the ground truth measurement error.

we explore cross-modal generation via decoder networks to generate RVS spectra from XP coefficients and vice-versa. For all tasks we find excellent performance which highlights the benefit of contrastive learning for stellar spectra analysis from multiple different surveys. All our code is publicly available at `https://codeberg.org/cschwarz/ba_multispecs`, as well as the dataset at `https://huggingface.co/datasets/christianschwarz/deep-multimodal-representation-learning-for-stellar-spectra` for full reproducibility.

In general, taking on more modalities and leveraging on the large body of stellar spectral surveys is the obvious next direction of this project. Another avenue, would be to train not only observed spectra, but also on synthetic ones. This could also be combined with pre-training an encoder on synthetic spectra, to then only fine-tune the last layers for the real spectra. Regarding the encoder networks themselves, several other networks might be explored. Specifically, the XP coefficients should be better processed by an attention-based architecture. This is speculated because the spectral coefficients don't share stronger correlations with their neighbors. However, spectral data can also be processed by attention-layers, after CNNs have reduced the dimensionality of the input. Lastly, the cross-modal generation of any data is restricted by the capabilities of the decoder. Recently, the design which showed the best performance for the generation task are diffusion models. For those, Contrastive Learning lays the foundation, in that the learned embeddings are used as conditions for generation. Furthermore, diffusion models often employ a pre-trained autoencoder. Since in the context of this work contrastive learning, as well as autoencoder training routines were implemented, the way towards diffusion models in multimodal astrophysics has been paved. Finally, to extend this work to probabilistic representation spaces, entails expanding the similarity metric to compare probabilistic embeddings.

Broader impact statement

The authors are not aware of any immediate ethical or societal implications of this work. This work purely aims to aid scientific research and proposes to apply multi-modal contrastive learning techniques to stellar spectra to learn about fundamental physics.

Acknowledgments and Disclosure of Funding

References

[1] R. Ait-Ouahmed, S. Arnouts, J. Pasquet, M. Treyer, and E. Bertin. Multimodality for improved CNN photometric redshifts. 2023. doi: 10.48550/arXiv.2310.02185. URL `http://arxiv.org/abs/2310.02185`.

[2] Justin Alsing, Hiranya Peiris, Joel Leja, ChangHoon Hahn, Rita Tojeiro, Daniel Mortlock, Boris Leistedt, Benjamin D. Johnson, and Charlie Conroy. SPECULATOR: Emulating Stellar Population Synthesis for Fast and Accurate Galaxy Spectra and Photometry. *The Astrophysical Journal Supplement Series*, 249(1):5, 2020. ISSN 0067-0049. doi: 10.3847/1538-4365/ab917f. URL `https://dx.doi.org/10.3847/1538-4365/ab917f`.

[3] R. Andrae, M. Fouesneau, R. Sordo, C. a. L. Bailer-Jones, T. E. Dharmawardena, J. Rybizki, F. De Angeli, H. E. P. Lindstrøm, D. J. Marshall, R. Drimmel, A. J. Korn, C. Soubiran, N. Brouillet, L. Casamiquela, H.-W. Rix, A. Abreu Aramburu, M. A. Álvarez, J. Bakker, I. Bellas-Velidis, A. Bijaoui, E. Brugaletta, A. Burlacu, R. Carballo, L. Chaoul, A. Chiavassa, G. Contursi, W. J. Cooper, O. L. Creevey, C. Dafonte, A. Dapergolas, prefix=de useprefix=false family=Laverny, given=P., L. Delchambre, C. Demouchy, B. Edvardsson, Y. Frémat, D. Garabato, P. García-Lario, M. García-Torres, A. Gavel, A. Gomez, I. González-Santamaría, D. Hatzidimitriou, U. Heiter, A. Jean-Antoine Piccolo, M. Kontizas, G. Kordopatis, A. C. Lanzafame, Y. Lebreton, E. L. Licata, E. Livanou, A. Lobel, A. Lorca, A. Magdaleno Romeo, M. Manteiga, F. Marocco, N. Mary, C. Nicolas, C. Ordenovic, F. Pailler, P. A. Palicio, L. Pallas-Quintela, C. Panem, B. Pichon, E. Poggio, A. Recio-Blanco, F. Riclet, C. Robin, R. Santoveña, L. M. Sarro, M. S. Schultheis, M. Segol, A. Silvelo, I. Slezak, R. L. Smart, M. Süveges, F. Thévenin, G. Torralba Elipe, A. Ulla, E. Utrilla, A. Vallenari, prefix=van useprefix=false family=Dillen, given=E., H. Zhao, and J. Zorec. Gaia Data Release 3 - Analysis of the Gaia BP/RP spectra using the General Stellar Parameterizer from Photometry. 674: A27. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243462. URL `https://www.aanda.org/articles/aa/abs/2023/06/aa43462-22/aa43462-22.html`.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2798607. URL `https://ieeexplore.ieee.org/abstract/document/8269806`.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *arXiv e-prints*, art. arXiv:1206.5538, June 2012. doi: 10.48550/arXiv.1206.5538.

[6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

[7] S. Buder, K. Lind, M. K. Ness, M. Asplund, L. Duong, J. Lin, J. Kos, L. Casagrande, A. R. Casey, J. Bland-Hawthorn, G. M. de Silva, V. D'Orazi, K. C. Freeman, S. L. Martell, K. J. Schlesinger, S. Sharma, J. D. Simpson, D. B. Zucker, T. Zwitter, K. Čotar, A. Dotter, M. R. Hayden, E. A. Hyde, P. R. Kafle, G. F. Lewis, D. M. Nataf, T. Nordlander, W. Reid, H.-W. Rix, Á. Skúladóttir, D. Stello, Y.-S. Ting, G. Traven, R. F. G. Wyse, and Galah Collaboration. The GALAH survey: An abundance, age, and kinematic inventory of the solar neighbourhood made with TGAS. *A&A*, 624:A19, April 2019. doi: 10.1051/0004-6361/201833218.

[8] 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, Maruša Žerjal,

and Galah Collaboration. The GALAH+ survey: Third data release. *MNRAS*, 506(1):150–201, September 2021. doi: 10.1093/mnras/stab1242.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. 2020. doi: 10.48550/arXiv.2002.05709. URL http://arxiv.org/abs/2002.05709.

[10] R. S. de Jong, O. Agertz, A. A. Berbel, J. Aird, D. A. Alexander, A. Amarsi, F. Anders, R. Andrae, B. Ansarinejad, W. Ansorge, P. Antilogus, H. Anwand-Heerwart, A. Arentsen, A. Arnadottir, M. Asplund, M. Auger, N. Azais, D. Baade, G. Baker, S. Baker, E. Balbinot, I. K. Baldry, M. Banerji, S. Barden, P. Barklem, E. Barthélémy-Mazot, C. Battistini, S. Bauer, C. P. M. Bell, O. Bellido-Tirado, S. Bellstedt, V. Belokurov, T. Bensby, M. Bergemann, J. M. Bestenlehner, R. Bielby, M. Bilicki, C. Blake, J. Bland-Hawthorn, C. Boeche, W. Boland, T. Boller, S. Bongard, A. Bongiorno, P. Bonifacio, D. Boudon, D. Brooks, M. J. I. Brown, R. Brown, M. Brüggen, J. Brynnel, J. Brzeski, T. Buchert, P. Buschkamp, E. Caffau, P. Caillier, J. Carrick, L. Casagrande, S. Case, A. Casey, I. Cesarini, G. Cescutti, D. Chapuis, C. Chiappini, M. Childress, N. Christlieb, R. Church, M. R. L. Cioni, M. Cluver, M. Colless, T. Collett, J. Comparat, A. Cooper, W. Couch, F. Courbin, S. Croom, D. Croton, E. Daguisé, G. Dalton, L. J. M. Davies, T. Davis, P. de Laverny, A. Deason, F. Dionies, K. Disseau, P. Doel, D. Döscher, S. P. Driver, T. Dwelly, D. Eckert, A. Edge, B. Edvardsson, D. E. Youssoufi, A. Elhaddad, H. Enke, G. Erfanianfar, T. Farrell, T. Fechner, C. Feiz, S. Feltzing, I. Ferreras, D. Feuerstein, D. Feuillet, A. Finoguenov, D. Ford, S. Fotopoulou, M. Fouesneau, C. Frenk, S. Frey, W. Gaessler, S. Geier, N. Gentile Fusillo, O. Gerhard, T. Giannantonio, D. Giannone, B. Gibson, P. Gillingham, C. González-Fernández, E. Gonzalez-Solares, S. Gottloeber, A. Gould, E. K. Grebel, A. Gueguen, G. Guiglion, M. Haehnelt, T. Hahn, C. J. Hansen, H. Hartman, K. Hauptner, K. Hawkins, D. Haynes, R. Haynes, U. Heiter, A. Helmi, C. H. Aguayo, P. Hewett, S. Hinton, D. Hobbs, S. Hoenig, D. Hofman, I. Hook, J. Hopgood, A. Hopkins, A. Hourihane, L. Howes, C. Howlett, T. Huet, M. Irwin, O. Iwert, P. Jablonka, T. Jahn, K. Jahnke, A. Jarno, S. Jin, P. Jofre, D. Johl, D. Jones, H. Jönsson, C. Jordan, I. Karovicova, A. Khalatyan, A. Kelz, R. Kennicutt, D. King, F. Kitaura, J. Klar, U. Klauser, J. P. Kneib, A. Koch, S. Koposov, G. Kordopatis, A. Korn, J. Kosmalski, R. Kotak, M. Kovalev, K. Kreckel, Y. Kripak, M. Krumpe, K. Kuijken, A. Kunder, I. Kushniruk, M. I. Lam, G. Lamer, F. Laurent, J. Lawrence, M. Lehmitz, B. Lemasle, J. Lewis, B. Li, C. Lidman, K. Lind, J. Liske, J. L. Lizon, J. Loveday, H. G. Ludwig, R. M. McDermid, K. Maguire, V. Mainieri, S. Mali, H. Mandel, K. Mandel, L. Mannering, S. Martell, D. Martinez Delgado, G. Matijevic, H. McGregor, R. McMahon, P. McMillan, O. Mena, A. Merloni, M. J. Meyer, C. Michel, G. Micheva, J. E. Migniau, I. Minchev, G. Monari, R. Muller, D. Murphy, D. Muthukrishna, K. Nandra, R. Navarro, M. Ness, V. Nichani, R. Nichol, H. Nicklas, F. Niederhofer, P. Norberg, D. Obreschkow, S. Oliver, M. Owers, N. Pai, S. Pankratow, D. Parkinson, J. Paschke, R. Paterson, A. Pecontal, I. Parry, D. Phillips, A. Pillepich, L. Pinard, J. Pirard, N. Piskunov, V. Plank, D. Plüschke, E. Pons, P. Popesso, C. Power, J. Pragt, A. Pramskiy, D. Pryer, M. Quattri, A. B. d. A. Queiroz, A. Quirrenbach, S. Rahurkar, A. Raichoor, S. Ramstedt, A. Rau, A. Recio-Blanco, R. Reiss, F. Renaud, Y. Revaz, P. Rhode, J. Richard, A. D. Richter, H. W. Rix, A. S. G. Robotham, R. Roelfsema, M. Romaniello, D. Rosario, F. Rothmaier, B. Roukema, G. Ruchti, G. Rupprecht, J. Rybizki, N. Ryde, A. Saar, E. Sadler, M. Sahlén, M. Salvato, B. Sassolas, W. Saunders, A. Saviauk, L. Sbordone, T. Schmidt, O. Schnurr, R. D. Scholz, A. Schwope, W. Seifert, T. Shanks, A. Sheinis, T. Sivov, Á. Skúladóttir, S. Smartt, S. Smedley, G. Smith, R. Smith, J. Sorce, L. Spitler, E. Starkenburg, M. Steinmetz, I. Stilz, J. Storm, M. Sullivan, W. Sutherland, E. Swann, A. Tamone, E. N. Taylor, J. Teillon, E. Tempel, R. ter Horst, W. F. Thi, E. Tolstoy, S. Trager, G. Traven, P. E. Tremblay, L. Tresse, M. Valentini, R. van de Weygaert, M. van den Ancker, J. Veljanoski, S. Venkatesan, L. Wagner, K. Wagner, C. J. Walcher, L. Waller, N. Walton, L. Wang, R. Winkler, L. Wisotzki, C. C. Worley, G. Worseck, M. Xiang, W. Xu, D. Yong, C. Zhao, J. Zheng, F. Zscheyge, and D. Zucker. 4MOST: Project overview and information for the First Call for Proposals. *The Messenger*, 175: 3–11, March 2019. doi: 10.18727/0722-6691/5117.

[11] DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John

Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/google-deepmind.

[12] M. Fouesneau, Y. Frémat, R. Andrae, A. J. Korn, C. Soubiran, G. Kordopatis, A. Vallenari, U. Heiter, O. L. Creevey, L. M. Sarro, P. de Laverny, A. C. Lanzafame, A. Lobel, R. Sordo, J. Rybizki, I. Slezak, M. A. Álvarez, R. Drimmel, D. Garabato, L. Delchambre, C. a. L. Bailer-Jones, D. Hatzidimitriou, A. Lorca, Y. Le Fustec, F. Pailler, N. Mary, C. Robin, E. Utrilla, A. Abreu Aramburu, J. Bakker, I. Bellas-Velidis, A. Bijaoui, R. Blomme, J.-C. Bouret, N. Brouillet, E. Brugaletta, A. Burlacu, R. Carballo, L. Casamiquela, L. Chaoul, A. Chiavassa, G. Contursi, W. J. Cooper, C. Dafonte, C. Demouchy, T. E. Dharmawardena, P. García-Lario, M. García-Torres, A. Gomez, I. González-Santamaría, A. Jean-Antoine Piccolo, M. Kontizas, Y. Lebreton, E. L. Licata, H. E. P. Lindstrøm, E. Livanou, A. Magdaleno Romeo, M. Manteiga, F. Marocco, C. Martayan, D. J. Marshall, C. Nicolas, C. Ordenovic, P. A. Palicio, L. Pallas-Quintela, B. Pichon, E. Poggio, A. Recio-Blanco, F. Riclet, R. Santoveña, M. S. Schultheis, M. Segol, A. Silvelo, R. L. Smart, M. Süveges, F. Thévenin, G. Torralba Elipe, A. Ulla, E. van Dillen, H. Zhao, and J. Zorec. Gaia Data Release 3 - Apsis. II. Stellar parameters. *Astronomy & Astrophysics*, 674: A28, 2023. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243919. URL https://www.aanda.org/articles/aa/abs/2023/06/aa43919-22/aa43919-22.html.

[13] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, C. Babusiaux, M. Biermann, O. L. Creevey, D. W. Evans, L. Eyer, A. Hutton, F. Jansen, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, C. Soubiran, N. A. Walton, F. Arenou, C. A. L. Bailer-Jones, U. Bastian, M. Cropper, R. Drimmel, D. Katz, M. G. Lattanzi, F. van Leeuwen, J. Bakker, C. Cacciari, J. Castañeda, F. De Angeli, C. Ducourant, C. Fabricius, M. Fouesneau, Y. Frémat, R. Guerra, A. Guerrier, J. Guiraud, A. Jean-Antoine Piccolo, E. Masana, R. Messineo, N. Mowlavi, C. Nicolas, K. Nienartowicz, F. Pailler, P. Panuzzo, F. Riclet, W. Roux, G. M. Seabroke, R. Sordo, P. Tanga, F. Thévenin, G. Gracia-Abril, J. Portell, D. Teyssier, M. Altmann, R. Andrae, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, E. Brugaletta, P. W. Burgess, G. Busso, B. Carry, A. Cellino, N. Cheek, G. Clementini, Y. Damerdji, M. Davidson, L. Delchambre, A. Dell'Oro, J. Fernández-Hernández, L. Galluccio, P. García-Lario, M. Garcia-Reinaldos, J. González-Núñez, E. Gosset, R. Haigron, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, D. Hatzidimitriou, U. Heiter, 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, A. Lorca, M. Manteiga, O. Marchal, P. M. Marrese, A. Moitinho, A. Mora, K. Muinonen, P. Osborne, E. Pancino, T. Pauwels, J. M. Petit, A. Recio-Blanco, P. J. Richards, M. Riello, L. Rimoldini, A. C. Robin, T. Roegiers, J. Rybizki, L. M. Sarro, C. Siopis, M. Smith, A. Sozzetti, A. Ulla, E. Utrilla, M. van Leeuwen, W. van Reeven, U. Abbas, A. Abreu Aramburu, S. Accart, C. Aerts, J. J. Aguado, M. Ajaj, G. Altavilla, M. A. Álvarez, J. Álvarez Cid-Fuentes, J. Alves, R. I. Anderson, E. Anglada Varela, T. Antoja, M. Audard, 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, A. Baudesson-Stella, U. Becciani, M. Bellazzini, M. Bernet, S. Bertone, L. Bianchi, S. Blanco-Cuaresma, T. Boch, A. Bombrun, D. Bossini, S. Bouquillon, A. Bragaglia, L. Bramante, E. Breedt, A. Bressan, N. Brouillet, B. Bucciarelli, A. Burlacu, D. Busonero, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, H. Cánovas, T. Cantat-Gaudin, R. Carballo, T. Carlucci, M. I. Carnerero, J. M. Carrasco, L. Casamiquela, M. Castellani, A. Castro-Ginard, P. Castro Sampol, L. Chaoul, P. Charlot, L. Chemin, A. Chiavassa, M. R. L. Cioni, G. Comoretto, W. J. Cooper, T. Cornez, S. Cowell, F. Crifo, 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, P. de Teodoro, A. de Torres, E. F. del Peloso, E. del Pozo, M. Delbo, A. Delgado, H. E. Delgado, J. B. Delisle, P. Di Matteo, S. Diakite, C. Diener, E. Distefano, C. Dolding, D. Eappachen, B. Edvardsson, H. Enke, P. Esquej, C. Fabre, M. Fabrizio, S. Faigler, G. Fedorets, P. Fernique, A. Fienga, F. Figueras, C. Fouron, F. Fragkoudi, E. Fraile, F. Franke, M. Gai, D. Garabato, A. Garcia-Gutierrez, M. García-Torres, A. Garofalo, P. Gavras, E. Gerlach, R. Geyer, P. Giacobbe, G. Gilmore, S. Girona, G. Giuffrida, R. Gomel, A. Gomez, I. Gonzalez-Santamaria, J. J. González-Vidal, M. Granvik, R. Gutiérrez-Sánchez, L. P. Guy, M. Hauser, M. Haywood, A. Helmi, S. L. Hidalgo, T. Hilger, N. Hładczuk, D. Hobbs, G. Holland, H. E. Huckle, G. Jasniewicz, P. G. Jonker, J. Juaristi Campillo, F. Julbe, L. Karbevska, P. Kervella, S. Khanna, A. Kochoska, M. Kontizas, G. Kordopatis, A. J. Korn,

Z. Kostrzewa-Rutkowska, K. Kruszyńska, S. Lambert, A. F. Lanza, Y. Lasne, J. F. Le Campion, Y. Le Fustec, Y. Lebreton, T. Lebzelter, S. Leccia, N. Leclerc, I. Lecoeur-Taibi, S. Liao, E. Licata, E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, P. Madrero Pardo, S. Managau, R. G. Mann, J. M. Marchant, M. Marconi, M. M. S. Marcos Santos, S. Marinoni, F. Marocco, D. J. Marshall, L. Martin Polo, J. M. Martín-Fleitas, 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, P. Montegriffo, R. Mor, R. Morbidelli, T. Morel, D. Morris, A. F. Mulone, D. Munoz, T. Muraveva, C. P. Murphy, I. Musella, L. Noval, C. Ordénovic, G. Orrù, J. Osinde, C. Pagani, I. Pagano, L. Palaversa, P. A. Palicio, A. Panahi, M. Pawlak, X. Peñalosa Esteller, A. Penttilä, A. M. Piersimoni, F. X. Pineau, E. Plachy, G. Plum, E. Poggio, E. Poretti, E. Poujoulet, 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, C. Reylé, V. Ripepi, A. Riva, G. Rixon, N. Robichon, C. Robin, M. Roelens, L. Rohrbasser, M. Romero-Gómez, N. Rowell, F. Royer, K. A. Rybicki, G. Sadowski, A. Sagristà Sellés, J. Sahlmann, J. Salgado, 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, E. Slezak, R. L. Smart, E. Solano, F. Solitro, D. Souami, J. Souchay, A. Spagna, F. Spoto, I. A. Steele, H. Steidelmüller, C. A. Stephenson, M. Süveges, L. Szabados, E. Szegedi-Elek, F. Taris, G. Tauran, M. B. Taylor, R. Teixeira, W. Thuillot, N. Tonello, F. Torra, J. Torra, C. Turon, N. Unger, M. Vaillant, E. van Dillen, 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, C. Zurbach, and T. Zwitter. Gaia Early Data Release 3. Summary of the contents and survey properties. *A&A*, 649:A1, May 2021. doi: 10.1051/0004-6361/202039657.

[14] Jialin Gao, Jianyu Chen, Jiaqi Wei, Bin Jiang, and A.-Li Luo. Deep Multimodal Networks for M-type Star Classification with Paired Spectrum and Photometric Image. *Publications of the Astronomical Society of the Pacific*, 135(1046):044503, 2023. ISSN 1538-3873. doi: 10.1088/1538-3873/acc7ca. URL `https://dx.doi.org/10.1088/1538-3873/acc7ca`.

[15] G. Guiglion, S. Nepal, C. Chiappini, S. Khoperskov, G. Traven, A. B. A. Queiroz, M. Steinmetz, M. Valentini, Y. Fournier, A. Vallenari, K. Youakim, M. Bergemann, S. Mészáros, S. Lucatello, R. Sordo, S. Fabbro, I. Minchev, G. Tautvaišienė, Š. Mikolaitis, and J. Montalbán. Beyond Gaia DR3: Tracing the [α/M] - [M/H] bimodality from the inner to the outer Milky Way disc with Gaia-RVS and convolutional neural networks. *Astronomy and Astrophysics*, 682:A9, 2024. ISSN 0004-6361. doi: 10.1051/0004-6361/202347122. URL `https://ui.adsabs.harvard.edu/abs/2024A&A...682A...9G`.

[16] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2023. URL `http://github.com/google/flax`.

[17] Shuxin Hong, Zhiqiang Zou, A-Li Luo, Xiao Kong, Wenyu Yang, and Yanli Chen. PhotoRedshift-MML: A multimodal machine learning method for estimating photometric redshifts of quasars. *Monthly Notices of the Royal Astronomical Society*, 518(4):5049–5058, 2023. ISSN 0035-8711. doi: 10.1093/mnras/stac3259. URL `https://doi.org/10.1093/mnras/stac3259`.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. 2021. doi: 10.48550/arXiv.2102.05918. URL `http://arxiv.org/abs/2102.05918`.

[19] Ari Karchmer. On Stronger Computational Separations Between Multimodal and Unimodal Machine Learning. 2024. URL `http://arxiv.org/abs/2404.02254`.

[20] Francois Lanusse, Liam Parker, Siavash Golkar, 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: Cross-Modal Pre-Training for Astronomical Foundation Models. 2023. URL `https://arxiv.org/abs/2310.03024v1`.

[21] Henry W Leung and Jo Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, 2024. ISSN 0035-8711. doi: 10.1093/mnras/stad3015. URL `https://doi.org/10.1093/mnras/stad3015`.

[22] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.emnlp-demo.21`.

[23] Jiadong Li, Kaze W. K. Wong, David W. Hogg, Hans-Walter Rix, and Vedant Chandra. AspGap: Augmented Stellar Parameters and Abundances for 23 million RGB stars from Gaia XP low-resolution spectra. 2023. doi: 10.48550/arXiv.2309.14294. URL `http://arxiv.org/abs/2309.14294`.

[24] Yi Liu, Jing Jin, Hongyang Zhao, Xujie He, and Yanan Guo. MFPIM: A Deep Learning Model Based on Multimodal Fusion Technology for Pulsar Identification. *The Astrophysical Journal*, 954(1):86, 2023. ISSN 0004-637X. doi: 10.3847/1538-4357/acd9c8. URL `https://dx.doi.org/10.3847/1538-4357/acd9c8`.

[25] S. R. Majewski, APOGEE Team, and APOGEE-2 Team. The Apache Point Observatory Galactic Evolution Experiment (APOGEE) and its successor, APOGEE-2. *Astronomische Nachrichten*, 337(8-9):863, Sep 2016. doi: 10.1002/asna.201612387.

[26] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020. doi: 10.48550/arXiv.1802.03426. URL `http://arxiv.org/abs/1802.03426`.

[27] Peter Melchior, Yan Liang, ChangHoon Hahn, and Andy Goulding. Autoencoding Galaxy Spectra I: Architecture. *The Astronomical Journal*, 166(2):74, 2023. ISSN 0004-6256, 1538-3881. doi: 10.3847/1538-3881/ace0ff. URL `http://arxiv.org/abs/2211.07890`.

[28] Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. PAPERCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models. 2024. doi: 10.48550/arXiv.2403.08851. URL `http://arxiv.org/abs/2403.08851`.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. 2021. doi: 10.48550/arXiv.2103.00020. URL `http://arxiv.org/abs/2103.00020`.

[30] A. Recio-Blanco, P. de Laverny, P. A. Palicio, G. Kordopatis, M. A. Álvarez, M. Schultheis, G. Contursi, H. Zhao, G. Torralba Elipe, C. Ordenovic, M. Manteiga, C. Dafonte, I. Oreshina-Slezak, A. Bijaoui, Y. Fremat, G. Seabroke, F. Pailler, E. Spitoni, E. Poggio, O. L. Creevey, A. Abreu Aramburu, S. Accart, R. Andrae, C. A. L. Bailer-Jones, I. Bellas-Velidis, N. Brouillet, E. Brugaletta, A. Burlacu, R. Carballo, L. Casamiquela, A. Chiavassa, W. J. Cooper, A. Dapergolas, L. Delchambre, T. E. Dharmawardena, R. Drimmel, B. Edvardsson, M. Fouesneau, D. Garabato, P. Garcia-Lario, M. Garcia-Torres, A. Gavel, A. Gomez, I. Gonzalez-Santamaria, D. Hatzidimitriou, U. Heiter, A. Jean-Antoine Piccolo, M. Kontizas, A. J. Korn, A. C. Lanzafame, Y. Lebreton, Y. Le Fustec, E. L. Licata, H. E. P. Lindstrom, E. Livanou, A. Lobel, A. Lorca, A. Magdaleno Romeo, F. Marocco, D. J. Marshall, N. Mary, C. Nicolas, L. Pallas-Quintela, C. Panem, B. Pichon, F. Riclet, C. Robin, J. Rybizki, R. Santovena, A. Silvelo, R. L. Smart, L. M. Sarro, R. Sordo, C. Soubiran, M. Suvege, A. Ulla, A. Vallenari, J. Zorec, E. Utrilla, and J. Bakker. Gaia Data Release 3: Analysis of RVS spectra using the General Stellar Parametriser from spectroscopy. *Astronomy & Astrophysics*, 674:A29, 2023. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243750. URL `http://arxiv.org/abs/2206.05541`.

[31] 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. 2023. URL `https://arxiv.org/abs/2306.15703v1`.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. 2019. doi: 10.48550/arXiv.1807.03748. URL `http://arxiv.org/abs/1807.03748`.

[33] Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. 2022. doi: 10.48550/arXiv.2005.10242. URL `http://arxiv.org/abs/2005.10242`.

[34] Jiaqi Wei, Bin Jiang, and Yanxia Zhang. Identification of Blue Horizontal Branch Stars with Multimodal Fusion. *Publications of the Astronomical Society of the Pacific*, 135(1050):084501, 2023. ISSN 1538-3873. doi: 10.1088/1538-3873/acea43. URL `https://dx.doi.org/10.1088/1538-3873/acea43`.

[35] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. 2020. doi: 10.48550/arXiv.1904.00962. URL `http://arxiv.org/abs/1904.00962`.

[36] B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, 1989. ISSN 1558-1896. doi: 10.1109/35.41402. URL `https://ieeexplore.ieee.org/abstract/document/41402`.

[37] Mingru Zhang, Junping Gao, A-Li Luo, Xia Jiang, Liwen Zhang, Kuang Wu, and Bo Qiu. A multimodal celestial object classification network based on 2D spectrum and photometric image. *RAS Techniques and Instruments*, 2(1):408–419, 2023. ISSN 2752-8200. doi: 10.1093/rasti/rzad026. URL `https://doi.org/10.1093/rasti/rzad026`.

[38] Gang Zhao, Yongheng Zhao, Yaoquan Chu, Yipeng Jing, and Licai Deng. LAMOST Spectral Survey. *arXiv e-prints*, art. arXiv:1206.3569, June 2012. doi: 10.48550/arXiv.1206.3569.

A Appendix / supplemental material

A.1 Multi-modal learning taxonomy [4]

Representation Any vector or tensor representing an entity, is referred to as a representation (equivalently embedding) [5]. For a multimodal problem the challenge lies in exploiting the complementarity and redundancy of multiple modalities which is complicated by the heterogeneity of the multimodal data.

Translation Given an entity in one modality the task is to generate the same entity in a different modality. The challenge arises from the possible existence of multiple mapping outputs that are still consistent/correct.

Alignment The task is to find direct relations and correspondences between sub-components of several modalities, by some kind of similarity definition.

Fusion This task entails integrating data from multiple modalities with the goal of predicting an outcome measure, e.g. performing classification or regression. The challenge is posed by each modality's varying predictive power or noise regarding the outcome. Also, at least one modality might be missing at inference time. Fusion benefits from multimodality, since predictions become more robust, complementary information are utilized and the model gains redundancy regarding sensory absence.

Co-Learning Here, the modeling of a (resource poor) modality is aided by exploiting knowledge from another (resource rich) modality. Limited resources might entail: lack of annotated data, noisy input, and unreliable labels. "Learning" points to the helper modality only being used during training, not testing time.

A.2 Hyperparameters

Table 1: Contrastive Learning hyperparameters

Hyperparameter	Value
Batch size	16384
Training Epochs	4 x 100 (4 runs, keeping model parameters but resetting optimizer)
Temperature	0.1
LAMB weight decay	$5 \cdot 10^{-5}$
LAMB β_1	0.9
LAMB β_2	0.999
LAMB ϵ	10^{-6}
Learning rate schedule	10 cosine decay cycles, with 1 linear warmup per cycle to peak value
Learning rate peak	$5 \cdot 10^{-3}; 5 \cdot 10^{-4}; 5 \cdot 10^{-5}; 5 \cdot 10^{-6}$

Table 2: Cross-modal generation hyperparameters

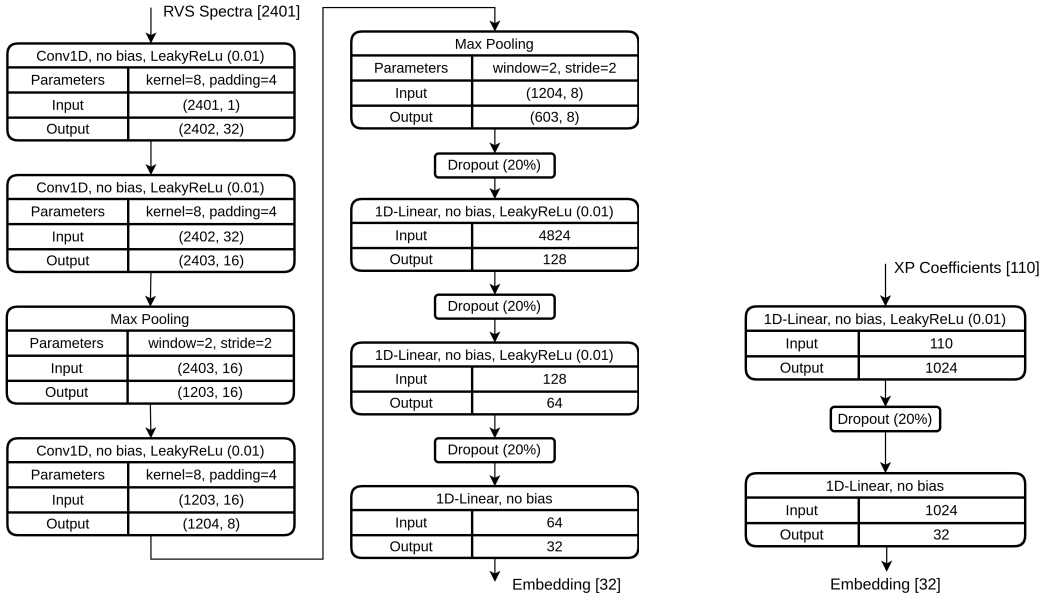| Hyperparameter	Value
Batch size	16384
Training Epochs	100
LAMB weight decay	0.0
LAMB β_1	0.9
LAMB β_2	0.999
LAMB ϵ	10^{-6}
Learning rate schedule	1 cosine decay cycle, with 1% linear warmup to peak value
Learning rate peak	$5 \cdot 10^{-4}$

Figure 4: Network architecture of RVS spectra encoder (left), based on [15], and XP coefficient encoder (right)
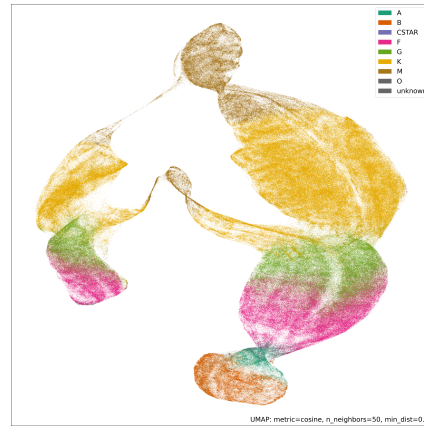
B   Additional Figures

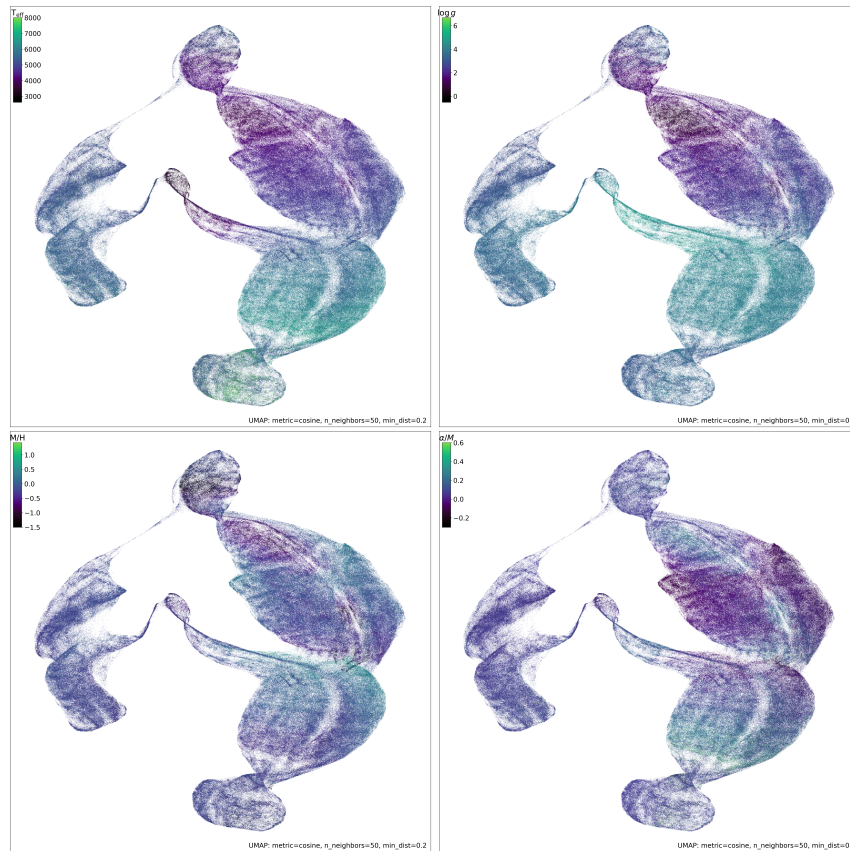Figure 5: UMAP visualization of the RVS training set embeddings with parameters `metric=cosine`, `n_neighbors=50`, `min_dist=0.2`; colored by spectral type.

Figure 6: UMAP visualization of RVS training set embeddings with parameters `metric=cosine`, `n_neighbors=50`, `min_dist=0.2`; colored by stellar parameters; T_{eff} upper left, $\log g$ upper right, the abundance of all metals ($[M/H]$) lower left and α-elements $[\alpha/M]$ lower right.
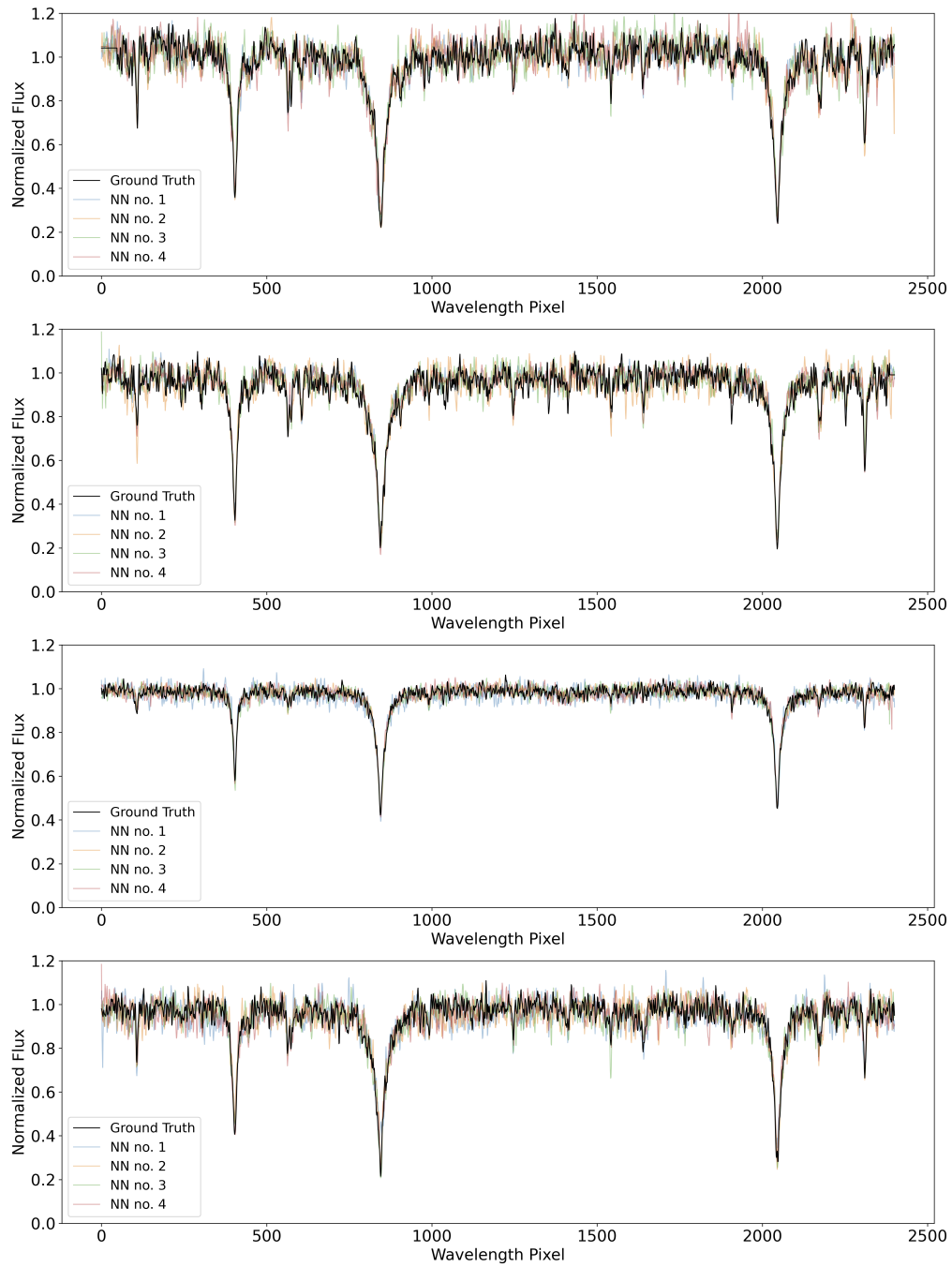
Figure 7: Similarity lookup from XP coefficients to RVS spectra for four example coefficients.
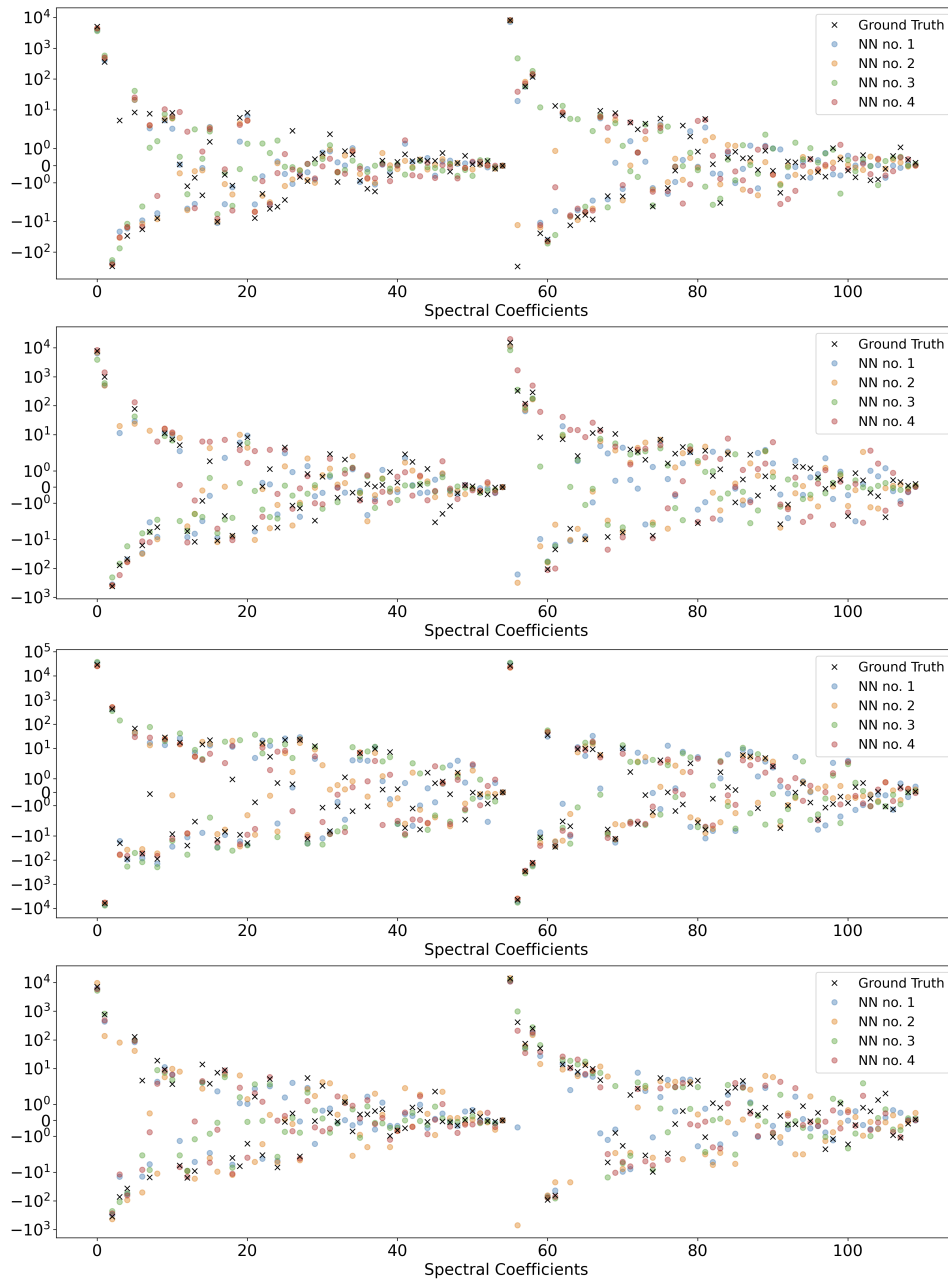
Figure 8: Similarity lookup from RVS spectra to XP coefficients for four example spectra; the scale is a symmetric logarithmic scale with linear scaling around zero.