Scaling Laws for Transformer-Based Stellar Spectral Emulation

Tomasz Różański [1] **Yuan-Sen Ting** [2 3]

Abstract

Modern astronomical surveys will deliver spectra for over 10^7 stars, requiring efficient methods to extract stellar properties from these observations. Neural network emulators have become essential for this task, yet their accuracy limitations currently introduce significant systematic errors. We demonstrate that Transformer-based stellar spectral emulators exhibit predictable scaling laws similar to large language models, providing a quantitative framework for achieving any desired emulation precision. Using TransformerPayne, we establish power-law relationships between emulation error and three key dimensions: training data size, model parameters, and computational resources. These scaling laws enable practitioners to determine the resources needed to reduce emulation errors below a certain target threshold, paving the way for robust spectral emulation for massive spectroscopic surveys.

1. Introduction

The next generation of spectroscopic surveys, including 4MOST (de Jong et al., 2019), WEAVE (Jin et al., 2024), DESI, and PFS, building upon APOGEE, LAMOST, Gaia-ESO, and GALAH, will deliver spectra for tens of millions of stars. Extracting stellar properties from these spectra requires solving a high-dimensional inverse problem: determining dozens of parameters (effective temperature, surface gravity, and elemental abundances) from observed flux measurements.

Traditional approaches face a computational bottleneck. Direct spectral synthesis, even under simplified 1D-LTE assumptions, requires minutes per spectrum, making real-time analysis of millions of spectra computationally prohibitive, as inference requires dozens of forward model evaluations. This has driven the development of neural network emulators that learn to approximate the forward mapping from stellar parameters to spectra, enabling rapid inference once trained.

Current emulators, however, suffer from accuracy limitations that introduce non-negligible systematic errors into stellar parameter estimates. Models like *The Cannon* (Ness et al., 2015) and *The Payne* (Ting et al., 2019) typically achieve $\sim$1% precision in flux space, insufficient when spectroscopic observations routinely achieve signal-to-noise ratios exceeding 100. This emulation error becomes a dominant source of uncertainty in the final parameter estimates, especially for low metallicity stars and elements with just a few lines present in spectra (Sandford et al., 2020).

The concept of neural scaling laws has recently emerged in astronomy, with predictable scaling demonstrated for stellar light curves (Pan et al., 2024) and galaxy images (Walmsley et al., 2024; Smith et al., 2024). However, these laws have not been established for spectroscopy, despite its fundamental role in astronomy.

In this work, we demonstrate that Transformer-based emulators can achieve arbitrary precision through predictable scaling laws. By adapting TransformerPayne (Różański et al., 2025) with Maximum Update Parametrization (Yang et al., 2022), we show that emulation error follows power-law relationships with model size, training data, and computational resources. These quantitative scaling laws, similar to those governing large language models (Kaplan et al., 2020; Hoffmann et al., 2022), provide concrete design rules for building emulators that meet a required precision threshold.

The established scaling law enables practitioners to calculate exactly how much computational investment is needed to achieve their scientific requirements, transforming emulator design from empirical trial-and-error to principled engineering.

2. Related Work

No previous work has explored scaling laws for spectral emulation, though many studies have applied deep learning to stellar spectroscopy. Spectroscopy differs from typ-

[1]Research School of Astronomy & Astrophysics, Australian National University, Cotter Rd., Weston, ACT 2611, Australia [2]Department of Astronomy, The Ohio State University, Columbus, USA [3]Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH 43210, USA. Correspondence to: Tomasz Różański <Tomasz.Rozanski1@anu.edu.au>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ical machine learning domains: unlike language or audio, stellar spectra have fixed wavelength-transition correspondences that break under translation. This introduces unique challenges for model design beyond simpler architecture choices like convolutional or recurrent neural networks. Recently, Transformer models have gained popularity in spectral studies due to their inductive biases being better suited for spectroscopy, where spectral information is often dispersed across long-range wavelength dependencies (Koblischke & Bovy, 2024; Leung & Bovy, 2024; Różański et al., 2025; Zhao et al., 2025).

Prior deep learning work in spectroscopy includes spectral pre-processing (Różański et al., 2022), stellar parameter prediction (Leung & Bovy, 2019), dimensionality reduction (Portillo et al., 2020), and contrastive learning combining spectra with photometry (Buck & Schwarz, 2024; Rizhko & Bloom, 2025). Others have addressed the synthetic-observational gap through domain adaptation (O'Briain et al., 2021), cross-survey adaptation (Leung & Bovy, 2024; Zhao et al., 2025), and cross-wavelength-range adaptation (Koblischke & Bovy, 2024).

3. Methodology

3.1. TransformerPayne Architecture and Training Dataset

We build upon TransformerPayne (Różański et al., 2025), a Transformer-based emulator that leverages attention mechanisms to capture long-range correlations in stellar spectra. To systematically investigate scaling behavior, we adopt the same model architecture while exploring how performance improves as we scale three key dimensions: model width (embedding dimension d), depth (number of Transformer blocks N), and the number of parameter tokens t. The total parameter count scales approximately as $P \approx (t + 12N)d^2$, with width being the dominant factor. This parametric relationship allows us to construct models ranging from 69k to 30.9M parameters, spanning over two orders of magnitude, to map the scaling landscape.

For direct comparison with the original TransformerPayne work, we utilize the same grid of 100,000 synthetic spectra generated using Kurucz ATLAS9 atmospheric models (Kurucz, 1979) and the Synthe spectral synthesis code (Kurucz, 2005). The grid spans effective temperatures from 4000 to 6000 K, surface gravities $\log g$ from 4 to 5, with individual elemental abundances varying from -2 to $+1$ dex relative to solar. Each continuum-normalized spectrum covers 4000–5000 Å at resolution $R = 10^5$, sampled at 22,315 wavelengths. This controlled dataset enables reproducible scaling experiments while maintaining sufficient complexity to represent realistic stellar spectra.

3.2. Scaling Experiment Design

To establish quantitative scaling laws, we systematically vary three dimensions:

Model Size: We train 10 architectures with parameter counts approximately doubling at each step: 69k, 119k, 273k, 469k, 1.05M, 1.86M, 4.14M, 8.10M, 13.7M, and 30.9M. This geometric progression ensures even coverage on a logarithmic scale, essential for fitting power laws.

Dataset Size: We create seven subsets containing 100, 300, 1k, 3k, 10k, 30k, and 100k spectra. This explores a critical practical constraint, high-fidelity spectral synthesis remains computationally expensive, making data efficiency a key consideration for real applications.

Training Compute: We vary training duration from 13k to 2.3M steps using a checkpoint reuse strategy. A single long training run generates frequent checkpoints, from which we launch parallel runs with different decay schedules. This approach reduces computational cost by approximately $\times 10$ while densely sampling the compute axis.

From our 100,000 spectra dataset, we reserve 1,024 for validation, consistent with the original TransformerPayne setup. For each configuration, we track validation MSE throughout training and report the minimum achieved. This experimental grid, encompassing over 300 individual training runs, enables precise characterization of scaling behavior along each dimension.

3.3. Maximum Update Parametrization

A challenge in scaling neural networks is maintaining fair comparisons across model sizes. Optimal hyperparameters typically shift with scale, larger models often require smaller learning rates, different initialization scales, and adjusted regularization. Exhaustive hyperparameter searches for each model size would be computationally prohibitive, potentially requiring hundreds of training runs per configuration.

Recent advances in large language models have solved this challenge through Maximum Update Parametrization (μP) (Yang et al., 2022). This framework provides principled rules for how hyperparameters should scale with model width, enabling optimal settings discovered on small models to transfer directly to larger scales. For a weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$, μP prescribes initialization scale $\sigma \propto \frac{1}{\sqrt{n}} \min\left(1, \sqrt{\frac{m}{n}}\right)$ and learning rate $\eta \propto \frac{1}{n}$.

We validate our μP implementation through systematic experiments varying model width from 64 to 512 dimensions while keeping other architectural parameters fixed. Under standard parametrization, the optimal learning rate decreases by a factor of $\times 3.2$ across this range (from 3×10^{-3} to 4×10^{-4}). With μP, the optimal learning rate remains

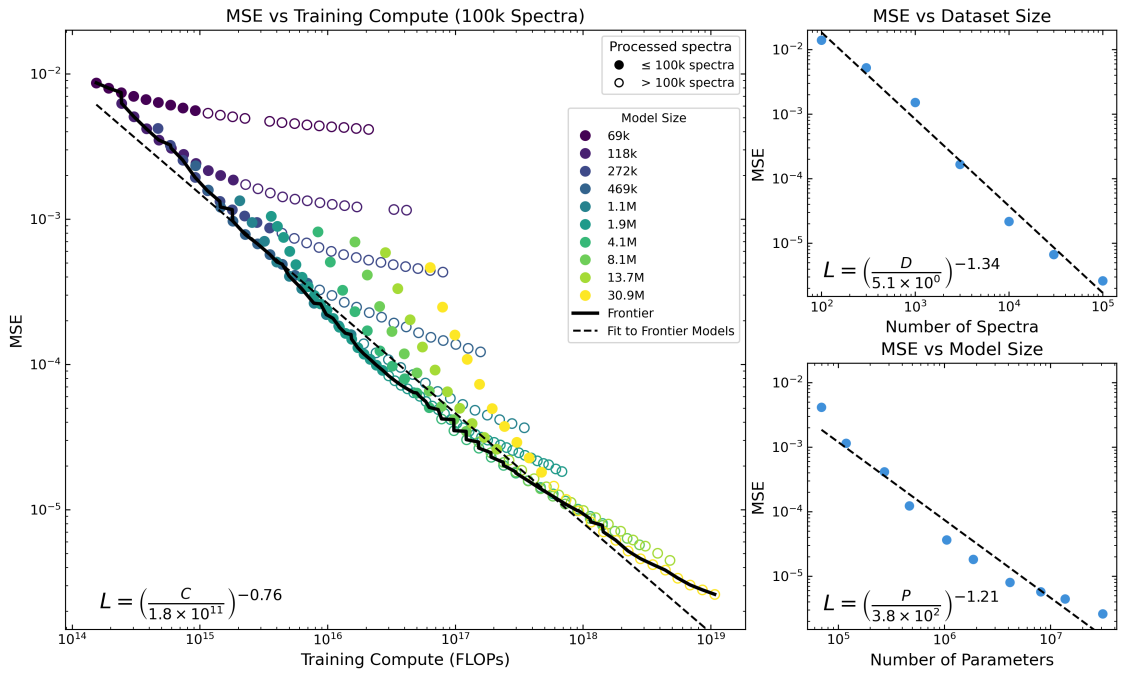

Figure 1. Scaling behavior of TransformerPayne emulators across three regimes. **Left**: Compute scaling showing validation MSE versus training FLOPs for all models. Colors indicate model size (69k–30.9M parameters); filled/open circles distinguish training within/beyond one epoch. The black line traces the compute-optimal frontier with power law exponent -0.76. **Upper-right**: Data scaling with exponent -1.34, demonstrating 22-fold error reduction per 10-fold data increase. **Lower-right**: Parameter scaling with exponent -1.21, showing 16-fold improvement per 10-fold parameter increase. Slight flattening at large model sizes indicates data limitations.

stable at approximately 2×10^{-3} across all widths. This stability enables us to perform hyperparameter optimization once on a small 128-dimensional proxy model, then confidently apply these settings to models up to 30.9M parameters. Beyond learning rate, we find that other key hyperparameters, including number of tokens, head dimensionality, and (to smaller extend) depth, also transfer reliably under μP, making large-scale experimentation tractable.

4. Results

4.1. Power-Law Scaling Across Three Regimes

Our experiments reveal that stellar spectral emulation follows predictable power-law relationships across all three scaling dimensions. Figure 1 presents the scaling behavior, demonstrating that improvements in emulation accuracy can be systematically achieved through increased resources.

To isolate scaling behavior along each dimension, we adopt the methodology from neural scaling law literature (Kaplan et al., 2020): for each regime, we fix one dimension and report the best performance achieved across all combinations of the other two. In the data-limited regime (upper-right panel), we show the minimum MSE achieved for each dataset size, optimized over all model architectures and training durations tested. This reveals the limit of what can be learned from a given amount of data, regardless of computational constraints.

The data scaling follows $\mathcal{L}(D) \propto D^{-1.34}$, where the steep exponent indicates that each 10-fold increase in training data reduces emulation error by approximately 22-fold. This relationship holds consistently across three orders of magnitude, from 100 to 100,000 spectra, suggesting that spectral emulation remains strongly data-hungry even at our largest scales.

Similarly, the parameter-limited regime (lower-right panel) shows the best performance achieved by each model size when given sufficient data and training time. The scaling $\mathcal{L}(P) \propto P^{-1.21}$ reveals that a 10-fold increase in parameters yields 16-fold error reduction. Notably, the largest models show slight deviation from the power law, indicating data saturation - these models have extracted most learnable patterns from our 100,000 spectra dataset.

The compute-limited regime (left panel) captures the most realistic scenario: finite computational budgets that must be allocated between model size and training duration. Each point represents a complete training run, but the compute frontier (black line) traces only the best-performing model at each compute level. This frontier follows $\mathcal{L}(C) \propto C^{-0.76}$,

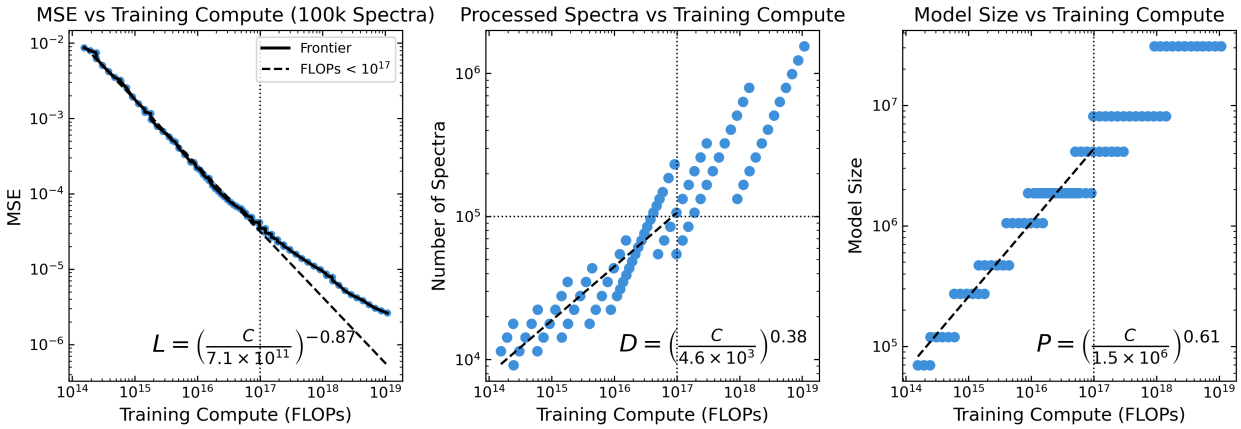

Figure 2. Optimal resource allocation along the compute frontier. **Left**: Compute-optimal models achieving lowest MSE at each FLOP budget, with power law fit $\mathcal{L} \propto C^{-0.87}$ in the single-epoch regime (left of vertical line). **Middle**: Optimal dataset size scales as $D \propto C^{0.38}$, a 10-fold compute increase requires 2.4-fold more data. Horizontal line at 100k spectra indicates our dataset limit. **Right**: Optimal model size scales as $P \propto C^{0.61}$, models should grow 4.1-fold per 10-fold compute increase. These relationships provide quantitative guidance for resource allocation under computational constraints.

with the smaller exponent reflecting the tension between model capacity and optimization time.

4.2. Compute-Optimal Scaling Strategy

While understanding individual scaling dimensions provides theoretical insight, practical applications require allocating fixed computational budgets optimally. The challenge is that compute, model size, and data size are interdependent, a larger model requires more FLOPs per training step, leaving fewer steps within a fixed budget. Similarly, processing more data requires either more training steps or larger batches. Figure 2 resolves this complexity by analyzing the compute frontier models.

To construct this frontier, we first identify the best-performing model at each compute level from our experiments. These frontier models represent optimal trade-offs, they achieved lower MSE than any other model using the same computational budget. By analyzing how model size and dataset size vary along this frontier, we can extract practical scaling recipes. The left panel shows these frontier models achieve better scaling (exponent -0.87) than the full dataset (exponent -0.76), confirming they represent truly optimal configurations.

The key insight emerges from the middle and right panels: optimal scaling does not simply mean "make everything bigger proportionally." When compute increases 10-fold, the frontier models show that parameters should grow by $\times 4.1$ (following $P \propto C^{0.61}$) while dataset size grows by only $\times 2.4$ (following $D \propto C^{0.38}$). This asymmetric scaling, models growing faster than data requirements, reflects that larger models can extract more information from the same data, making model capacity the better investment as compute scales.

Nonetheless, the frontier analysis reveals a limit in our current experimental setup. The vertical line at $\sim 10^{17}$ FLOPs marks where frontier models exhaust our 100,000 spectra dataset and begin reprocessing data. Below this threshold, models complete training within one epoch, achieving ideal scaling with exponent -0.87. Beyond it, repeated data provides diminishing returns, degrading the exponent to -0.76. This transition demonstrates that scaling laws, while robust, ultimately depend on data availability, highlighting the need for larger spectral grids to sustain efficient scaling.

5. Conclusions and Broader Impact

We have demonstrated that Transformer-based stellar spectral emulators exhibit predictable scaling laws analogous to those governing large language models. These quantitative relationships, power laws connecting emulation error to model parameters, training data, and computational resources, transform emulator design from empirical trial-and-error into systematic planning for next-generation spectroscopic surveys.

Our key contributions are threefold: (1) establishing that stellar spectral emulation follows universal scaling principles with measurable power-law exponents, (2) demonstrating that Maximum Update Parametrization enables stable scaling across two orders of magnitude in model size, and (3) providing quantitative recipes for optimal resource allocation under computational constraints. Together, these findings offer actionable guidance: data provides the highest leverage (scaling exponent -1.34), models and data must

scale asymmetrically (as $C^{0.61}$ and $C^{0.38}$ respectively), and simply training longer has diminishing returns (compute scaling of only -0.76).

For upcoming surveys, achieving systematic-free analysis may require emulation precision of 10^{-8} MSE or better. Our framework predicts this necessitates approximately 10^{21} FLOPs (150 GPU-days on modern hardware), 1.5B parameter models, and 3.56M high-fidelity training spectra. While substantial, these requirements are within reach of current computational facilities and provide concrete targets for preparatory efforts.

Our scaling laws characterize interpolation performance within the training domain—a critical first step for any machine learning application. While domain transfer (generalizing to different stellar models or real observations) represents an important future direction, establishing these baseline scaling relationships was essential. Our results provide the quantitative foundation necessary for investigating emergent capabilities like few-shot learning (Brown et al., 2020; Zhang et al., 2024) in stellar spectral emulation, which we expect to manifest as models scale according to the principles we have established.

Future work can build upon these relationships to explore domain transfer, investigate emergent capabilities at larger scales, and ultimately develop true foundation models for stellar spectroscopy. Models that achieve both the accuracy and generalization needed to serve as universal tools across diverse astronomical applications.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. *arXiv e-prints*, art. arXiv:2005.14165, May 2020. doi: 10.48550/arXiv.2005.14165.

Buck, T. and Schwarz, C. Deep Multimodal Representation Learning for Stellar Spectra. *arXiv e-prints*, art. arXiv:2410.16081, October 2024. doi: 10.48550/arXiv.2410.16081.

de Jong, R. S., Agertz, O., Berbel, A. A., et al. 4MOST: Project overview and information for the First Call for Proposals. *The Messenger*, 175:3–11, March 2019. doi: 10.18727/0722-6691/5117.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models. *arXiv e-prints*, art. arXiv:2203.15556, March 2022. doi: 10.48550/arXiv.2203.15556.

Jin, S., Trager, S. C., Dalton, G. B., et al. The wide-field, multiplexed, spectroscopic facility WEAVE: Survey design, overview, and simulated implementation. *Monthly Notices of the Royal Astronomical Society*, 530(3):2688–2730, May 2024. doi: 10.1093/mnras/stad557.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models. *arXiv e-prints*, art. arXiv:2001.08361, January 2020. doi: 10.48550/arXiv.2001.08361.

Koblischke, N. and Bovy, J. SpectraFM: Tuning into Stellar Foundation Models. *arXiv e-prints*, art. arXiv:2411.04750, November 2024. doi: 10.48550/arXiv.2411.04750.

Kurucz, R. L. Model atmospheres for G, F, A, B, and O stars. *The Astrophysical Journal Supplement Series*, 40:1–340, May 1979. doi: 10.1086/190589.

Kurucz, R. L. ATLAS12, SYNTHE, ATLAS9, WIDTH9, et cetera. *Memorie della Societa Astronomica Italiana Supplementi*, 8:14, 2005.

Leung, H. W. and Bovy, J. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, 483(3):3255–3277, March 2019. doi: 10.1093/mnras/sty3217.

Leung, H. W. and Bovy, J. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, January 2024. doi: 10.1093/mnras/stad3015.

Ness, M., Hogg, D. W., Rix, H. W., Ho, A. Y. Q., and Zasowski, G. The Cannon: A data-driven approach to Stellar Label Determination. *The Astrophysical Journal*, 808(1):16, July 2015. doi: 10.1088/0004-637X/808/1/16.

O'Briain, T., Ting, Y.-S., Fabbro, S., Yi, K. M., Venn, K., and Bialek, S. Cycle-StarNet: Bridging the Gap between Theory and Data by Leveraging Large Data Sets. *The Astrophysical Journal*, 906(2):130, January 2021. doi: 10.3847/1538-4357/abca96.

Pan, J.-S., Ting, Y.-S., Huang, Y., Yu, J., and Liu, J.-F. The Scaling Law in Stellar Light Curves. *arXiv e-prints*, art. arXiv:2405.17156, May 2024. doi: 10.48550/arXiv.2405.17156.

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., and Connolly, A. J. Dimensionality Reduction of SDSS Spectra with Variational Autoencoders. *The Astronomical Journal*, 160 (1):45, July 2020. doi: 10.3847/1538-3881/ab9644.

Rizhko, M. and Bloom, J. S. AstroM^3: A Self-supervised Multimodal Model for Astronomy. *The Astronomical Journal*, 170(1):28, July 2025. doi: 10.3847/1538-3881/adcbad.

Różański, T., Niemczura, E., Lemiesz, J., Posiłek, N., and Różański, P. SUPPNet: Neural network for stellar spectrum normalisation. *Astronomy and Astrophysics*, 659: A199, March 2022. doi: 10.1051/0004-6361/202141480.

Różański, T., Ting, Y.-S., and Jabłońska, M. Transformerpayne: Enhancing spectral emulation accuracy and data efficiency by capturing long-range correlations. *The Astrophysical Journal*, 980(1):66, feb 2025. doi: 10.3847/1538-4357/ad9b99. URL https://dx.doi.org/10.3847/1538-4357/ad9b99.

Sandford, N. R., Weisz, D. R., and Ting, Y.-S. Forecasting chemical abundance precision for extragalactic stellar archaeology. *The Astrophysical Journal Supplement Series*, 249(2):24, jul 2020. doi: 10.3847/1538-4365/ab9cb0. URL https://dx.doi.org/10.3847/1538-4365/ab9cb0.

Smith, M. J., Roberts, R. J., Angeloudi, E., and Huertas-Company, M. AstroPT: Scaling Large Observation Models for Astronomy. *arXiv e-prints*, art. arXiv:2405.14930, May 2024. doi: 10.48550/arXiv.2405.14930.

Ting, Y.-S., Conroy, C., Rix, H.-W., and Cargile, P. The Payne: Self-consistent ab initio Fitting of Stellar Spectra. *The Astrophysical Journal*, 879(2):69, July 2019. doi: 10.3847/1538-4357/ab2331.

Walmsley, M., Bowles, M., Scaife, A. M. M., Shingirai Makechemu, J., Gordon, A. J., Ferguson, A. M. N., Mann, R. G., Pearson, J., Popp, J. J., Bovy, J., Speagle, J., Dickinson, H., Fortson, L., Géron, T., Kruk, S., Lintott, C. J., Mantha, K., Mohan, D., O'Ryan, D., and Slijepevic, I. V. Scaling Laws for Galaxy Images. *arXiv e-prints*, art. arXiv:2404.02973, April 2024. doi: 10.48550/arXiv.2404.02973.

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *arXiv e-prints*, art. arXiv:2203.03466, March 2022. doi: 10.48550/arXiv.2203.03466.

Zhang, B., Liu, Z., Cherry, C., and Firat, O. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. *arXiv e-prints*, art. arXiv:2402.17193, February 2024. doi: 10.48550/arXiv.2402.17193.

Zhao, X., Huang, Y., Xue, G., Kong, X., Liu, J., Tang, X., Beers, T. C., Ting, Y.-S., and Luo, A.-L. SpecCLIP: Aligning and Translating Spectroscopic Measurements for Stars. *arXiv e-prints*, art. arXiv:2507.01939, July 2025. doi: 10.48550/arXiv.2507.01939.