universe

MDPI

Article

KAN-Payne: A Controlled Evaluation of Kolmogorov–Arnold Networks for Stellar Spectral Emulation and Label Recovery

Shuo Zhang [1] **and Rui Wang** [2,3,*]

[1] Department of Astronomy, Tsinghua University, Beijing 100084, China; zhangs0808@tsinghua.edu.cn
[2] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wangrui@nao.cas.cn

Abstract

Kolmogorov–Arnold networks (KANs) replace the fixed activations of a multilayer perceptron (MLP) with learnable univariate edge functions. We evaluate KANs as Payne-style label-to-flux emulators using the public 1000-spectrum Kurucz grid released with The Payne. Two capacity-matched KAN–MLP pairs (approximately 2.3 and 18–19 million parameters) were tuned separately, trained for the same 120,000-update budget across three seeds, and evaluated on a fixed 200-spectrum holdout. At the smaller capacity, the KAN did not outperform the tuned MLP. At the larger capacity, KAN-L achieved the lowest flux errors among the models trained from scratch: its mean absolute error was 35% below the best small model, all three large-MLP layouts remained less accurate, and its residual tails were lighter. In injection–recovery tests, KAN-L improved the recovery of effective temperature, surface gravity, and metallicity at all the tested signal-to-noise ratios relative to the small matched MLP; the best large-MLP control was not included in this test. The learned KAN response shapes were reproducible across seeds. These gains required substantially greater training and inference costs. In an APOGEE Data Release 17 deployment, surface-gravity offsets decreased by approximately 0.09 dex for all the tested emulators when calibrated APOGEE Stellar Parameter and Chemical Abundances Pipeline (ASPCAP) values were replaced by the raw spectroscopic scale, indicating a common reference-scale contribution. KANs are therefore not efficient drop-in replacements for MLP emulators, but the large KAN tested here improved spectral emulation accuracy and residual-tail robustness.

Keywords: stellar spectroscopy; APOGEE; stellar parameters; Kolmogorov–Arnold networks; spectral emulators; machine learning

Check for updates

Academic Editor: Sergio Cristallo

Received: 21 June 2026
Revised: 17 August 2026
Accepted: 21 August 2026
Published: 23 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Stellar spectra connect large Galactic surveys to stellar astrophysics. Effective temperature, surface gravity, metallicity, and elemental abundances enter distance and age inference, chemical-tagging analyses, selection-function modeling, and chemo-dynamical studies of the Milky Way. The scientific return of surveys such as APOGEE, LAMOST, GALAH, and RAVE therefore depends on the reliability and homogeneity of the labels inferred from their spectra [1–5].

Many production pipelines compare observed spectra with synthetic models and optimize atmospheric labels against a spectral residual. APOGEE's ASPCAP pipeline performs pseudo-continuum normalization, fits global parameters against synthetic grids

with FERRE, and determines individual abundances in selected wavelength windows [6,7]. Related pipelines operate at the resolutions of LAMOST, RAVE, and GALAH [4,5,8,9]. A complementary family of methods learn spectrum–label relations from reference stars [10–15]. These models are fast but inherit the coverage and calibration systematics of their reference sets.

The Payne occupies a middle ground: a neural emulator is trained on synthetic spectra and used as a differentiable forward model during observed-spectrum fitting [16]. This retains a connection to a physical grid while avoiding repeated high-dimensional interpolation. Direct synthesis has become much faster since The Payne was introduced. Korg enables efficient one-dimensional local-thermodynamic-equilibrium synthesis [17]; Payne Zero reports an APOGEE-band spectrum in roughly one second on a modern GPU [18]; and pyKurucz provides a maintained implementation of the Kurucz atmosphere and synthesis stack [19]. Emulators nevertheless remain useful when the same forward model must be evaluated repeatedly for survey-scale samples or posterior sampling and when expensive physics such as three-dimensional convection or non-LTE line formation is available only through precomputed grids. TransformerPayne has shown that wavelength-conditioned attention can improve accuracy and data efficiency on large synthetic grids [20]. The present work instead examines the vector-output family used by The Payne.

KANs replace fixed node activations with learnable univariate edge functions, commonly parameterized by B-splines [21,22]. Their suitability relative to MLPs is task dependent. Controlled comparisons have found MLP advantages across many standard tasks at matched capacity, with KAN gains concentrated in structured function-fitting problems [23]; a stellar-classification study attributes an apparent KAN robustness advantage largely to implicit regularization [24]; and KAN performance can degrade in noisy function-fitting settings [25]. The closest astronomical spectral precedent is the 21cmKAN emulator, which emulates a low-dimensional global 21 cm signal and uses its learned activations for sensitivity analysis [26]. These results motivate a controlled test rather than an assumption of KAN superiority.

This study evaluates KANs as Payne-style stellar spectral emulators under matched parameter-count and optimization budgets. We compare two capacity-matched KAN–MLP pairs and two additional large-MLP architectures in terms of flux accuracy, residual tails, and computational cost. Known-truth injection–recovery experiments examine how emulator errors propagate into stellar-label estimates, while response-function diagnostics assess the reproducibility of the learned KAN functions. Selected emulators are further applied to APOGEE DR17 spectra. At the Payne-sized scale, the KAN does not outperform a tuned MLP. At the larger scale, KAN-L achieves the lowest flux errors and lightest residual tails among the models trained from scratch, although with substantially higher training and inference costs. Section 2 describes the synthetic and observed data; Section 3 presents the models and experimental design; Section 4 reports the results; Section 5 discusses their implications and limitations; and Section 6 summarizes the conclusions.

2. Data

2.1. The Payne Synthetic Spectral Grid and Evaluation Split

The emulator benchmark uses the Kurucz APOGEE-like synthetic spectra released in the public repository for The Payne [16]. The underlying model atmospheres and synthesis belong to the ATLAS12/SYNTHE family [27–32]. The public file contains 1000 continuum-normalized spectra. After loading the released wavelength and mask arrays, the working grid contains 7214 pixels from 15,168.13 to 16,936.75 Å. Each spectrum has 25 labels: $T_{\rm eff}$, $\log g$, microturbulence $v_{\rm turb}$, 20 elemental abundances [X/H], the isotope ratio $^{12}C/^{13}C$, and macroturbulence $v_{\rm macro}$.

For each training split, label ℓ_i is scaled as follows

$$\widetilde{\ell}_i = \frac{\ell_i - \ell_{i,\min}}{\ell_{i,\max} - \ell_{i,\min}} - 0.5, \quad (1)$$

so that the training range maps approximately to $[-0.5, 0.5]$. The public pixel mask removes 875 pixels (12.1%), leaving 6339 unmasked pixels for the primary metrics. The synthetic spectra are broadened with the combined APOGEE line-spread function used by The Payne. Fiber-dependent LSF variations are not represented in the grid; in observed-spectrum fits, residual broadening differences can be partly absorbed by $v_{\rm macro}$.

The 200 spectra distributed as the repository validation set are used as a fixed holdout for the revised experiments. The remaining 800 spectra are split independently for each training seed into 720 training and 80 validation spectra. Hyperparameter and checkpoint selection use only the 80-spectrum validation subset. Because the fixed 200 spectra were examined in the earlier version of this project, they are a holdout for the present training protocol rather than a prospectively unseen test set for the project as a whole. In particular, the two supplementary large-MLP shape controls were selected after earlier results on these 200 spectra had been seen.

2.2. Label-Space Coverage

The public sample is sparse in the nominal 25-dimensional label space. To quantify its structure, we calculate the covariance eigenvalues λ_k of the scaled labels and the participation ratio

$$d_{\rm eff} = \frac{(\sum_k \lambda_k)^2}{\sum_k \lambda_k^2}. \quad (2)$$

For the 800 released training spectra, $d_{\rm eff} = 2.86$; 6, 11, and 20 principal components are required to explain 90%, 95%, and 99% of the variance, respectively. Thus the released grid occupies a low-dimensional physical subspace rather than a uniform hypercube, consistent with the adaptive sampling motivation of The Payne [16]. The six structural labels $T_{\rm eff}$, $\log g$, $v_{\rm turb}$, [Fe/H], ^{12}C/^{13}C, and $v_{\rm macro}$ each cover at least 97% of their public ranges. Fifteen elemental-abundance dimensions underfill their public ranges by 5–32%, with the largest shortfall for [C/H]. Our conclusions therefore apply most directly to the well-covered structural parameters and strong-line abundances in this released data-limited regime.

2.3. APOGEE DR17 Sample and ASPCAP Labels

The observed-spectrum analysis uses APOGEE DR17 `aspcapStar` products and the official DR17 synthetic-grid `allStar` catalogue. APOGEE observes near-infrared H-band spectra at a resolving power of approximately 22,500 over 1.51–1.70 μm [1,33]. DR17 contains the completed APOGEE-2 data set [2]. ASPCAP supplies the calibrated atmospheric parameters and abundances used for comparison, together with the raw spectroscopic `FPARAM` values [6,34].

We select red giants with finite ASPCAP labels, no severe ASPCAP or stellar warning flags, at least two visits, APO or LCO spectra, $3500 < T_{\rm eff} < 5500$ K, $0 < \log g < 3.8$, $-2.0 <$ [Fe/H] < 0.5, and `SNREV`> 100. From the stars passing these cuts, 10,000 are drawn with random seed 42. Their median calibrated labels are $T_{\rm eff} = 4658$ K, $\log g = 2.39$, and [Fe/H] $= -0.23$; the median signal-to-noise ratio estimated from the normalized flux and uncertainty arrays is 137.

2.4. Observed-Spectrum Preprocessing

We use combined rest-frame `aspcapStar` spectra. To improve consistency with the Payne synthetic spectra, the APOGEE pseudo-normalized flux is interpolated onto the

7214-pixel Payne grid and renormalized with a fourth-order Chebyshev polynomial fitted separately to the blue, green, and red detector segments (pixel ranges 0–2920, 2920–5320, and 5320–7214). The fit uses a fixed set of 576 continuum pixels, distributed as 233, 192, and 151 pixels across the three detector segments and weighted by inverse flux uncertainty. Across the public synthetic grid these pixels have median fluxes in the range 0.9965–1.0071 and spectrum-to-spectrum standard deviations below 0.01.

A floor of 0.005 in normalized-flux units is added to the flux uncertainties before fitting. The final fitting mask combines APOGEE bad pixels, non-finite values, non-positive uncertainties, and the public Payne mask. The selected sample retains a median of 6339 usable pixels per star. Figure 1 illustrates the continuum treatment.

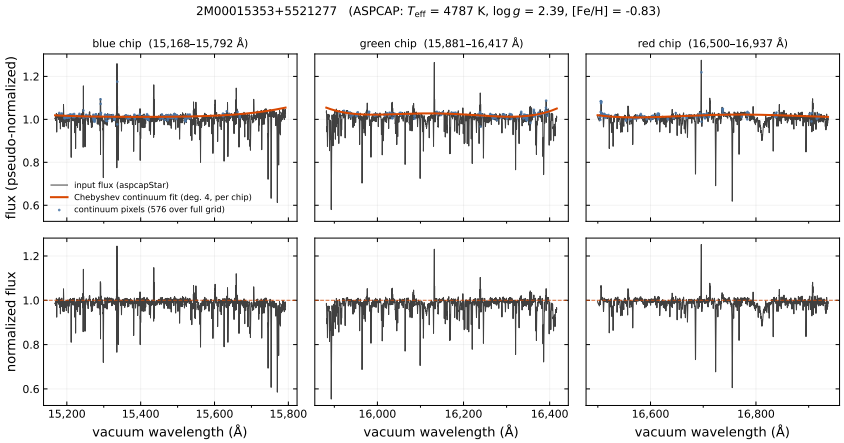


Figure 1. Continuum treatment of the APOGEE DR17 input spectra for a representative star. The upper panels show the `aspcapStar` pseudo-normalized flux, the chip-wise fourth-order Chebyshev continuum, and the 576 continuum pixels. The lower panels show the Payne-style normalized flux used in the fitting analysis.

3. Models and Experimental Design

3.1. MLP Baselines and the Public Payne Reference

The public pretrained Payne network is a $25 \to 300 \to 300 \to 7214$ MLP with LeakyReLU(0.01) activations. A NumPy forward pass reproduces its released predictions. On the fixed 200-spectrum holdout it has good-pixel MAE and RMSE values of 17.44 and 68.31, respectively, in units of 10^{-4} in continuum-normalized flux. Its exact training membership and overlap with the released 1000-spectrum file are not documented, and it was trained under a different protocol and on a larger sample. We therefore use it only as an external non-budget-matched reference. Its good-pixel MAE is 11.19 on the released 800 training spectra and 17.44 on the holdout, reinforcing the need to keep it separate from the controlled from-scratch comparison.

The primary MLP models use GELU activations. MLP-P has hidden widths 300×300 and 2,269,514 trainable parameters. MLP-L has hidden widths 300×2400 and 18,051,014 parameters. A matched LeakyReLU control for MLP-P yields a test MAE of 44.56 ± 0.76 compared with 31.83 ± 1.35 for GELU, so the selected activation strengthens the MLP baseline relative to the earlier configuration.

3.2. KAN Emulators

A KAN layer maps an input representation $\boldsymbol{h}^{(l)}$ to the next layer by assigning a trainable univariate function to every edge,

$$h_j^{(l+1)} = \sum_i \phi_{ji}^{(l)}\left(h_i^{(l)}\right), \tag{3}$$

whereas an MLP layer composes an affine transformation with a fixed scalar nonlinearity,

$$\boldsymbol{h}^{(l+1)} = \sigma\Big(\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l)} + \boldsymbol{b}^{(l)}\Big). \tag{4}$$

For stellar spectra, the relevant one-dimensional functions are responses to labels or hidden coordinates, not functions of wavelength. The hypothesis tested here is therefore that learnable label-response shapes may represent nonlinear temperature, gravity, abundance, and broadening dependencies more effectively than fixed activations.

We use the `efficient-kan` implementation. Each edge combines a cubic B-spline with a GELU base activation. The configuration is `grid_size = 5`, `spline_order = 3`, base grid range $[-1, 1]$, per-edge spline scalers, initialization-noise scale 0.1, and unit base and spline weight scales. The extended knot span reaches approximately $[-2.2, 2.2]$. KAN-P has widths $25 \to 64 \to 32 \to 7214$ and 2,344,960 trainable parameters. KAN-L has widths $25 \to 64 \to 256 \to 7214$ and 18,647,680 parameters. Both produce the full spectrum in one forward pass and remain differentiable with respect to all the input labels.

3.3. Capacity-Matched Architecture Set

The comparison separates architecture from capacity through two matched pairs. At the Payne-sized scale, the parameter counts of MLP-P and KAN-P agree to within 3.5%, as do those of MLP-L and KAN-L at the large scale. Two additional large-MLP layouts test sensitivity to width allocation: MLP-L-bal uses 2048×2048 hidden layers and 19,031,086 parameters, while MLP-L-deep uses three hidden layers of width 1739 and 18,649,294 parameters. These controls are not an exhaustive MLP architecture search, but they determine whether the large-KAN result depends only on comparison with one wide–shallow MLP.

The primary comparison fixes parameter count and optimizer updates rather than theoretical FLOPs. Because spline evaluation contributes to the computational cost of the tested KAN parameterization, we report measured wall-clock time, memory, latency, and fit cost without attempting a FLOP-matched comparison.

3.4. Training, Tuning, and Evaluation Protocol

All the primary models use AdamW, batch size 64, and mean-squared error over the full 7214-pixel output,

$$\mathcal{L} = \frac{1}{N_{\text{batch}} N_{\text{pix}}} \sum_{n,p} \Big(f_{n,p}^{\text{pred}} - f_{n,p}^{\text{true}}\Big)^2. \tag{5}$$

The learning rate follows a cosine schedule with 5% linear warm-up, and every formal run contains 120,000 optimizer updates. Three seeds are used for every primary architecture; each changes the initialization and the 720/80 train–validation split, while the 200-spectrum holdout is fixed. Masked pixels remain in the vector-output training loss to constrain the full predicted spectrum but are excluded from the primary metrics because the reference synthesis is unreliable at those pixels and they are not used in downstream fitting. We report both good-pixel and all-pixel metrics.

Learning rate and weight decay are tuned separately for each primary architecture using validation MAE only. The search covers learning rates 10^{-3}, 3×10^{-4}, and 10^{-4}, together with zero or modest weight decay. Within this grid, every primary architecture selects a learning rate of 10^{-3} and zero weight decay. Additional MLP weight-decay controls at the selected learning rate produce only small changes—nominal improvements of at most 2.05×10^{-4}, below our pre-specified 2σ criterion—and even the best weight-decay run remains approximately 45% above KAN-L (29.80 versus 20.60 ± 0.25); these larger weight-decay arms were run only for the two MLPs, so whether weight decay would

improve the KANs was not tested. The tuning also identifies the learning rate as the dominant cause of the weak legacy baseline: the legacy MLP result (57.71) is reproduced by the $(3 \times 10^{-4}, 0)$ configuration at 20,000 updates (57.62), whereas the selected $(10^{-3}, 0)$ configuration yields a lower MAE of 44.30 at the same update budget. The primary test statistics are MAE, RMSE, and the 99th percentile of the absolute residual distribution, Q_{99}, reported as the mean and standard deviation across seeds.

3.5. Injection–Recovery Experiment

To determine whether flux accuracy translates into scientific inference, the 200 holdout spectra are fitted at per-pixel S/N values of ∞, 100, 50, and 20. Two Gaussian-noise realizations are used at each finite S/N, giving 400 fits per model, seed, and finite-S/N cell; the noiseless experiment contains 200 fits. Every fit begins at the midpoint of the scaled label box and optimizes all 25 labels for 600 Adam steps at learning rate 0.03 using the unmasked pixels. No true-label or catalogue-informed initialization is used.

For each label and S/N, we report the median recovered-minus-true bias and the robust scatter 1.4826 MAD. We also record the fraction of solutions within 1% of a label-box boundary and the fraction with absolute error above 10% of the public label range. The four primary models are tested with all three training seeds. The balanced and deep large-MLP shape controls were not included in this experiment; consequently, the recovery comparison tests KAN-L against MLP-P, KAN-P, and the wide–shallow MLP-L, not against MLP-L-deep.

A separate basin-consistency experiment fits 50 noiseless holdout spectra from five initial points: the box midpoint and four fixed random starts.

3.6. Computational Profiling

Training wall-clock time and peak allocated GPU memory are measured on a single NVIDIA A100. Inference latency is measured from the seed-0 checkpoint after ten warm-up calls and 50 timed repeats for GPU batch sizes 1 and 1024 and for single-thread CPU batch size 1. We also time a complete $25 \rightarrow 7214$ autograd Jacobian and a 600-step full-pixel spectral fit. These measurements characterize practical implementation cost; they are not intended as hardware-independent complexity estimates.

3.7. Response-Function Diagnostics

We use three diagnostics to test whether the learned KAN functions are reproducibly inspectable. First, the highest-attribution first-layer edges leaving $T_{\rm eff}$, $\log g$, and [Fe/H] are extracted from all three KAN-L seeds. Hidden units are matched and curve signs aligned before computing pairwise Pearson shape correlations on a common input grid. Second, whole-layer label attributions are computed by propagating activation standard deviations, following the KAN 2.0 framework [22], and compared between seeds with Spearman rank correlations. Third, emulator Jacobians $\partial F_\lambda / \partial \ell_i$ at the median training label vector are compared with those of the public pretrained Payne model. This is a cross-model concordance test rather than a measurement against physical derivative truth. A targeted diagnostic examines the known threshold-like $^{12}C/^{13}C$ response near 15,332.7 Å [16].

We additionally measure the fraction of activations outside the spline knot span and test a narrower base grid $[-0.5, 0.5]$ at fixed grid size, spline order, parameter count, and training budget. Extrapolation scans displace one label at a time beyond the training box; these scans assess boundedness, not physical accuracy.

3.8. APOGEE Fitting and Start Dependence

The primary APOGEE deployment uses the seed-0 checkpoints of KAN-L, MLP-P, and MLP-L-deep. For star n, labels are optimized by minimizing

$$\chi^2_{\mathrm{like}}(\boldsymbol{\ell}) = \frac{1}{N_{\mathrm{pix},n}} \sum_{p \in \mathcal{P}_n} \left[\frac{f_p^{\mathrm{obs}} - f_p^{\mathrm{emu}}(\boldsymbol{\ell})}{\sigma_p} \right]^2, \tag{6}$$

where $\mathcal{P}_n$ is the valid-pixel set for that star and $N_{\mathrm{pix},n} = |\mathcal{P}_n|$ is the number of valid pixels for star n. Because the continuum-normalized pixels have correlated errors and an imposed error floor, this quantity is reduced-χ^2-like rather than a calibrated reduced χ^2.

Each fit uses the same 2048 unmasked pixels, sampled without replacement with random seed 42, and runs 600 Adam steps at learning rate 0.03. Final residuals are evaluated over all unmasked pixels. The main fits begin from ASPCAP-informed labels clipped to the public training range. A 200-star subset stratified in T_{eff}, $\log g$, [Fe/H], and S/N is fitted from five starting points (ASPCAP, midpoint, and three random starts) to assess multimodality. This subset was verified to be distribution-consistent with the full 10,000-star sample in T_{eff}, $\log g$, [Fe/H], and S/N using Kolmogorov–Smirnov tests at the 0.01 significance level. A 1200-step trace on the same subset tests whether the fixed 600-step budget affects population-level results.

For the primary population comparison, each model must retain at least 5000 valid pixels, remain within the label box, and avoid the upper 1% tails in the fit objective and spectral MAE. The intersection of the three model-specific samples contains 9129 stars. The results from the original-submission checkpoints and the local wavelength-conditioned reference model are provided separately in Appendix A; they are not used to rank the revised primary architectures.

4. Results

4.1. Emulation Accuracy and Computational Cost

Table 1 summarizes the computational cost, and Table 2 gives the fixed-holdout accuracy. All the primary checkpoints reach their validation floor near the end of the matched 120,000-update schedule; the training curves are shown schematically in Figure 2.

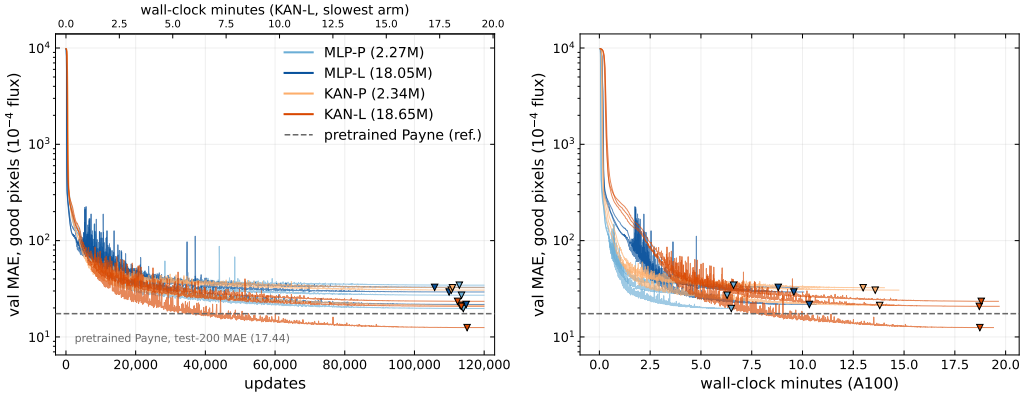


Figure 2. Validation MAE as a function of optimizer updates under the matched protocol. Each architecture is shown for all three seeds. The public pretrained Payne value is included only as an external reference. Downward triangles mark the validation-selected checkpoint for each seed; their partial overlap reflects similar selected update counts across seeds.

Table 1. Training and inference cost of the four budget-matched emulators under the new 720/80/200 protocol (120,000 updates, three seeds), with the two supplementary MLP shape controls (Section 5.2) below the rule. Wall time is the seed mean ± s.d. on a single NVIDIA A100; peak GPU memory is the peak allocated CUDA memory during training (identical across seeds). Latencies are medians over 50 repeats (10 warm-up) on the seed-0 checkpoints: GPU at batch size 1 and 1024 and single-thread CPU at batch size 1. The Jacobian column times one full 25 → 7214 `torch.autograd` Jacobian (median of 20). The $t_{\mathrm{fit},600}$ column is the per-star wall-clock of one full-pixel 600-step Adam fit on GPU (seed-0 checkpoints; median [16th–84th percentile] over the first 20 frozen test spectra) [1,2,3].

Architecture	Parameters	Updates	Wall Time (s)	Peak GPU (MiB)	GPU $b = 1$ (ms)	GPU $b = 1024$ (ms)	CPU $b = 1$ (ms)	Jacobian (s)	$t_{\mathrm{fit},600}$ (s)
MLP-P	2,269,514	120,000	407.6 ± 9.6	65.5	0.084	0.35	0.51	1.02	0.455 [0.375–0.572]
KAN-P	2,344,960	120,000	868.6 ± 18.6	74.7	1.386	2.29	2.44	10.08	2.124 [1.808–2.626]
KAN-L	18,647,680	120,000	1175.4 ± 7.4	419.9	1.122	3.16	28.83	9.91	1.770 [1.675–2.368]
MLP-L	18,051,014	120,000	611.6 ± 21.5	367.3	0.102	2.29	4.37	0.86	0.353 [0.351–0.360]
MLP-L-bal [3]	19,031,086	120,000	798.0 ± 18.6	384.5	—	—	—	—	—
MLP-L-deep [3]	18,649,294	120,000	780.9 ± 37.9	382.6	—	—	—	—	—

[1] All architectures use GELU hidden activations and were tuned per architecture; within the tested grid, all four primary architectures selected the same configuration (learning rate 10^{-3}, no weight decay), which the shape controls inherit. The legacy runs used LeakyReLU and are reported separately. Parameters = trainable parameters only. [2] Two latency entries run against the naive parameter-count expectation and are kernel-shape effects, not noise: KAN-P is slower than KAN-L at batch size 1 (1.386 vs. 1.122 ms) because per-layer kernel-launch overhead dominates at batch 1 and KAN-L's wider layers vectorize more efficiently; for the same reason the wide MLP-L completes the full Jacobian faster than MLP-P (0.86 vs. 1.02 s). [3] The supplementary shape controls MLP-L-bal (2048 × 2048) and MLP-L-deep (1739 × 1739 × 1739) were not profiled for inference latency or per-fit timing (—).

Table 2. Emulation accuracy on the frozen 200-spectrum test set in units of 10^{-4} in continuum-normalized flux: mean ± s.d. over three seeds. The good-pixel metrics ($\mathrm{MAE}_{\mathrm{good}}$, $\mathrm{RMSE}_{\mathrm{good}}$, and $\mathrm{Q99}_{\mathrm{good}}$) are computed over the 6339 unmasked (good) pixels of the 7214-pixel grid; the all-pixel metrics ($\mathrm{MAE}_{\mathrm{all}}$ etc.) use all 7214 pixels.

	Good Pixels			All Pixels		
Model	**MAE**	**RMSE**	**Q99**	**MAE**	**RMSE**	**Q99**
MLP-P	31.83 ± 1.35	79.67 ± 5.27	342.03 ± 21.18	33.44 ± 1.43	83.43 ± 5.52	359.07 ± 21.93
KAN-P	33.57 ± 1.64	78.81 ± 5.58	345.58 ± 30.33	35.49 ± 1.60	82.95 ± 5.19	362.39 ± 29.94
KAN-L	20.60 ± 0.25	53.49 ± 0.89	235.66 ± 7.20	21.63 ± 0.23	55.83 ± 0.88	245.12 ± 7.36
MLP-L	33.51 ± 1.35	80.27 ± 4.17	364.84 ± 27.81	35.26 ± 1.39	84.43 ± 4.09	381.48 ± 25.69
The Payne (pretrained) [1]	17.44	68.31	246.61	18.29	70.89	254.45

[1] External non-budget-matched reference, not trained under our split; its training-set overlap with the public file cannot be established, and the audit is consistent with substantial overlap: its good-pixel MAE is 11.19 on the 800 grid training spectra vs. 17.44 on test-200, whereas the four budget-matched models above see only the 720 training spectra. MLP baselines use GELU hidden activations.

At the Payne-sized scale, KAN-P does not outperform MLP-P. Their MAEs are 33.57 ± 1.64 and 31.83 ± 1.35, respectively, and the difference is within the combined seed scatter. The large KAN behaves differently: increasing capacity from KAN-P to KAN-L reduces the MAE by 39% and places KAN-L 35% below the best small model. The same wide–shallow scaling of the MLP does not improve the error, but the balanced and deeper layouts do: MLP-L-bal and MLP-L-deep reach MAEs of 28.16 ± 1.48 and 26.78 ± 0.84. Thus MLP performance at fixed 18–19-million-parameter capacity depends strongly on shape, yet the best of the three tested layouts remains about 30% above KAN-L in MAE and has RMSE and Q_{99} values of 71.09 and 307.5 compared with 53.49 and 235.7 for KAN-L.

Relative to the public pretrained Payne network, KAN-L improves only the tail-sensitive metrics considered here. The public model retains the lowest bulk MAE, whereas KAN-L is the only model trained from scratch with lower RMSE and Q_{99}. The advantage of KAN-L over the public model is therefore concentrated in the residual tail rather than in the typical residual.

These accuracy improvements are accompanied by higher computational costs. Relative to MLP-P, KAN-L requires 2.9 times the training time, approximately 13 times the single-spectrum GPU latency, 57 times the single-thread CPU latency, and 3.9 times the cost of a full 600-step fit. This cost difference decreases at large GPU batch sizes because the wider KAN kernels vectorize more effectively. Figure 3 summarizes the trade-off between model capacity, emulation accuracy, and training cost.

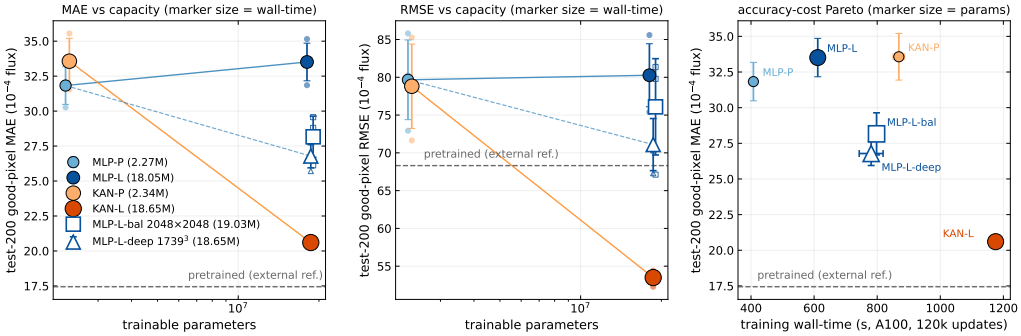


Figure 3. Accuracy as a function of model capacity and training cost. The left and central panels show the holdout MAE and RMSE as functions of trainable parameter count, with marker area proportional to the A100 training wall-clock time. The right panel shows the holdout MAE as a function of training wall-clock time, with marker area proportional to parameter count. Large markers and error bars denote the means and standard deviations across three seeds, while faint markers denote individual seeds. The horizontal dashed line indicates the public pretrained Payne reference. Blue markers denote MLP architectures and orange markers denote KAN architectures; marker shapes distinguish the individual layouts as indicated in the legend. KAN-L has the lowest MAE and RMSE among the models trained from scratch, whereas the public reference retains the lowest MAE overall.

4.2. Residual Tails and Line-Sensitive Pixels

The difference between MAE and RMSE motivates a direct examination of residual tails. KAN-L has the lightest extreme tail among the five models shown in Figure 4. Its pooled complementary cumulative distribution crosses the public reference at an absolute residual of 208.9×10^{-4}; the three seed-specific crossover values are 184, 229, and 219. Below this threshold the public reference has fewer exceedances, consistent with its lower bulk MAE. Above it KAN-L has fewer exceedances. The probabilities $P(|r| > 10^{-2})$ are 0.0403 and 0.0286 for KAN-L and the public reference, whereas $P(|r| > 2 \times 10^{-2})$ is 0.0135 and 0.0132. More pronounced differences appear in the extremes: the largest per-star MAE is 191 for KAN-L and 460 for the public reference, and the largest single-pixel residual is 1692 versus 2745 in the same 10^{-4} units.

Pixels were also binned by line depth and by the standard deviation of their synthetic-grid fluxes, a proxy for label sensitivity. KAN-L has a lower Q_{99} than the public reference in all four line-depth quartiles and in three of four sensitivity quartiles; the exception is the most label-sensitive quartile, where the values are 366.5 and 329.5. The tail advantage is therefore broad but not universal, and it is not confined solely to the strongest lines.

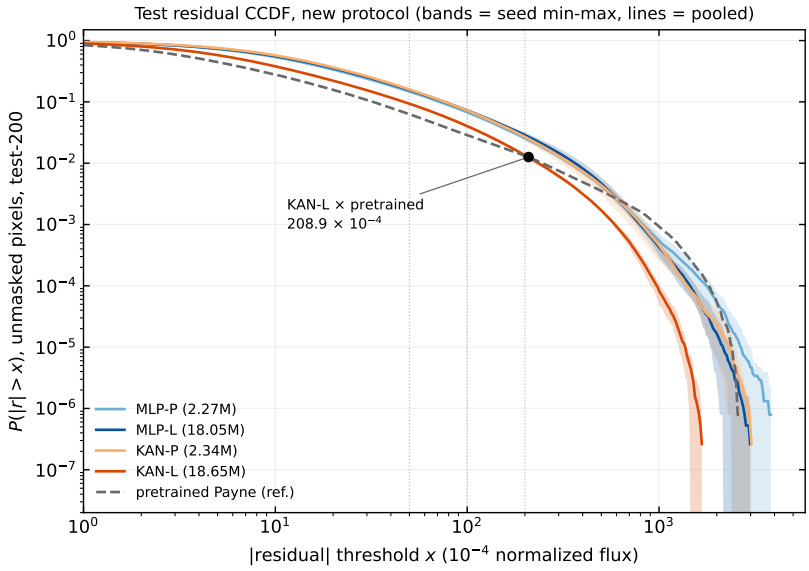


Figure 4. Complementary cumulative distribution of absolute holdout residuals. KAN-L has the lightest extreme tail and crosses the public pretrained reference at 208.9×10^{-4}.

4.3. Recovery of Known Labels

Table 3 gives the known-truth injection–recovery results for the principal stellar parameters and [Mg/H]. KAN-L has lower three-seed mean scatter than MLP-P for $T_{\rm eff}$, $\log g$, and [Fe/H] at every S/N, with KAN-L/MLP-P scatter ratios between 0.69 and 0.96. At S/N $= \infty$, the scatters are 26.3 K, 0.037 dex, and 0.021 dex for KAN-L compared with 37.9 K, 0.051 dex, and 0.029 dex for MLP-P. KAN-L is nominally the lowest of the four tested models in all 12 parameter–S/N cells, although its $T_{\rm eff}$ result at S/N $= 20$ is statistically indistinguishable from MLP-L within the seed scatter.

Table 3. Injection–recovery of known labels on the frozen 200-spectrum test set over the three training-seed checkpoints of every emulator. Each cell reports the median bias / robust scatter (1.4826 MAD) of recovered minus injected labels as the three-seed mean, with the seed standard deviation of the scatter (population s.d., ddof = 0) after the ± sign; the 200 test stars and the frozen noise realizations are identical across seeds. S/N $= \infty$ fits the 200 noiseless spectra; each finite S/N combines two noise realizations (400 fits per seed).

Label	S/N	MLP-P	KAN-P	KAN-L	MLP-L
$T_{\rm eff}$ (K)	∞	+4.2 / 37.9 ± 1.0	+21.5 / 107.5 ± 12.5	+2.4 / 26.3 ± 0.6	+5.6 / 34.5 ± 3.3
	100	+4.9 / 49.7 ± 2.5	+32.7 / 126.7 ± 14.2	+6.2 / 42.2 ± 1.5	+6.9 / 47.6 ± 2.1
	50	+1.0 / 63.3 ± 0.8	+39.4 / 148.5 ± 13.2	+5.6 / 59.7 ± 3.5	+5.0 / 64.2 ± 6.4
	20	−2.1 / 134.5 ± 8.9	+65.0 / 221.6 ± 17.4	+13.0 / 125.7 ± 4.3	+10.4 / 126.1 ± 9.9
$\log g$ (dex)	∞	+0.005 / 0.051 ± 0.003	+0.009 / 0.142 ± 0.007	+0.001 / 0.037 ± 0.003	+0.004 / 0.051 ± 0.008
	100	+0.006 / 0.070 ± 0.002	+0.010 / 0.168 ± 0.016	+0.002 / 0.057 ± 0.004	+0.006 / 0.066 ± 0.006
	50	+0.001 / 0.093 ± 0.001	+0.009 / 0.198 ± 0.010	+0.000 / 0.080 ± 0.008	+0.007 / 0.094 ± 0.004
	20	−0.008 / 0.185 ± 0.004	+0.003 / 0.299 ± 0.028	+0.011 / 0.178 ± 0.007	+0.008 / 0.184 ± 0.017
[Fe/H] (dex)	∞	+0.001 / 0.029 ± 0.001	−0.007 / 0.073 ± 0.004	+0.001 / 0.021 ± 0.002	+0.002 / 0.032 ± 0.002
	100	−0.002 / 0.038 ± 0.002	−0.003 / 0.085 ± 0.006	+0.001 / 0.032 ± 0.001	+0.002 / 0.038 ± 0.003
	50	+0.001 / 0.051 ± 0.003	−0.003 / 0.105 ± 0.007	+0.002 / 0.046 ± 0.002	+0.003 / 0.049 ± 0.003
	20	−0.007 / 0.094 ± 0.001	−0.002 / 0.157 ± 0.012	−0.004 / 0.085 ± 0.005	−0.004 / 0.097 ± 0.005
[Mg/H] (dex)	∞	+0.003 / 0.030 ± 0.001	−0.011 / 0.083 ± 0.012	+0.004 / 0.023 ± 0.002	+0.006 / 0.029 ± 0.002
	100	+0.001 / 0.038 ± 0.002	−0.014 / 0.101 ± 0.012	+0.004 / 0.036 ± 0.003	+0.004 / 0.038 ± 0.003
	50	−0.000 / 0.054 ± 0.002	−0.011 / 0.125 ± 0.016	+0.003 / 0.052 ± 0.003	+0.004 / 0.058 ± 0.004
	20	−0.007 / 0.116 ± 0.008	−0.020 / 0.188 ± 0.007	+0.005 / 0.105 ± 0.004	−0.001 / 0.113 ± 0.007

Bias and scatter are in the physical units of each label (column 1). Bias is the three-seed mean of the per-seed median bias.

KAN-P is consistently the weakest recovery model and has a systematic warm $T_{\rm eff}$ bias that grows from +21.5 K in the noiseless case to +65.0 K at S/N = 20. The three-seed mean biases of the other models remain within 13 K in $T_{\rm eff}$, 0.011 dex in $\log g$, and 0.008 dex in [Fe/H].

Across all 25 labels, KAN-L outperforms MLP-P for 20 labels in the noiseless experiment, 13 at S/N = 100, and nine at S/N = 50 and 20. Thus the advantage becomes less general as observational noise dominates the emulator error. The isotope ratio ^{12}C/^{13}C is poorly recovered by every model, with catastrophic-error fractions of 52–72%; $v_{\rm macro}$ also develops boundary-hit rates of up to 11% at S/N = 20. These failures reflect weak information in much of the released label domain and should not be used to rank the architectures.

The recovery test does not include MLP-L-deep, the best large-MLP layout in the flux benchmark. It therefore establishes a KAN-L advantage over MLP-P and the wide–shallow MLP-L but does not establish an advantage over the best large MLP identified in Section 4.1. The principal known-truth recovery results are summarized in Figure 5.

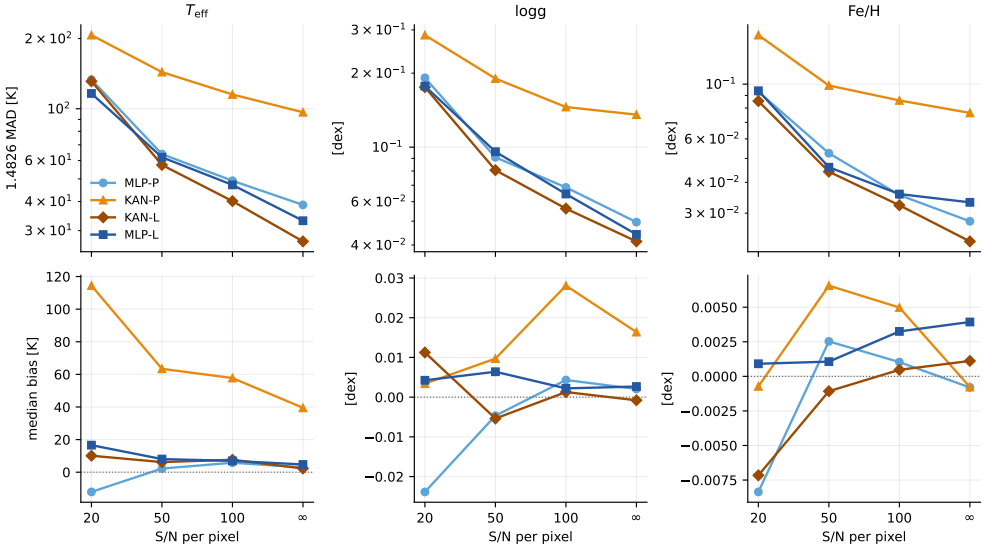


Figure 5. Known-truth injection–recovery. KAN-L reduces the robust scatter of the three principal stellar parameters relative to MLP-P at every tested S/N, while KAN-P is consistently worse and develops a warm temperature bias.

4.4. Learned Response Functions

The highest-attribution first-layer edges leaving $T_{\rm eff}$, $\log g$, and [Fe/H] have reproducible shapes across the three KAN-L seeds. After hidden-unit matching and sign alignment, pairwise absolute Pearson correlations span 0.970–0.998. Whole-layer label-attribution rankings have Spearman correlations of 0.959–0.982. Function amplitudes are less stable and can vary by up to a factor of approximately 1.7, so the reproducible object is the response shape and ranking rather than the absolute scale. The mean attribution shares place $T_{\rm eff}$ first (0.190), followed by $\log g$ (0.103), [Fe/H] (0.067), $v_{\rm macro}$, and [C/H].

KAN-L Jacobians also show slightly higher concordance with the public Payne reference than MLP-P for the four tested labels. The pixel-wise correlations are 0.9931 versus 0.9837 for $T_{\rm eff}$, 0.9934 versus 0.9931 for $\log g$, 0.9915 versus 0.9830 for [Fe/H], and 0.9812 versus 0.9569 for [Mg/H]. The mean correlations are 0.9898 and 0.9792. Because the comparison is between emulators, these values indicate a stable shared response pattern rather than physical derivative truth.

At the ^{12}C/^{13}C-sensitive pixel near 15,332.7 Å, KAN-L concentrates 86% of its total response variation below an isotope ratio of 50 compared with 98% for the public Payne

reference and 64% for MLP-P. The threshold location is therefore reproduced qualitatively, but the saturation contrast is much weaker than in the reference.

All the layer inputs remain within the spline knot span on the training, validation, and holdout spectra. The labels occupy only about 2.5 of the 11 inter-knot intervals. Narrowing the base grid from $[-1, 1]$ to $[-0.5, 0.5]$ at otherwise fixed architecture increases the effective knot density but worsens the MAE and Q_{99} to 23.34 ± 0.60 and 272.9 ± 9.3 compared with 20.60 ± 0.25 and 235.7 ± 7.2 for the primary KAN-L. The coarse effective basis is therefore not the limiting factor in this experiment.

Beyond the knot span, the B-spline component vanishes and each edge reduces to its GELU base branch. Extrapolation scans remain smooth and bounded, and KAN-L drifts less than MLP-P when labels are displaced 1.5–2 half-ranges beyond the training box. These outputs contain no learned spline information and should not be interpreted as physically accurate extrapolations. Figure 6 shows the highest-attribution first-layer KAN edge functions for $T_{\rm eff}$, $\log g$, and $[\mathrm{Fe/H}]$ across the three training seeds.

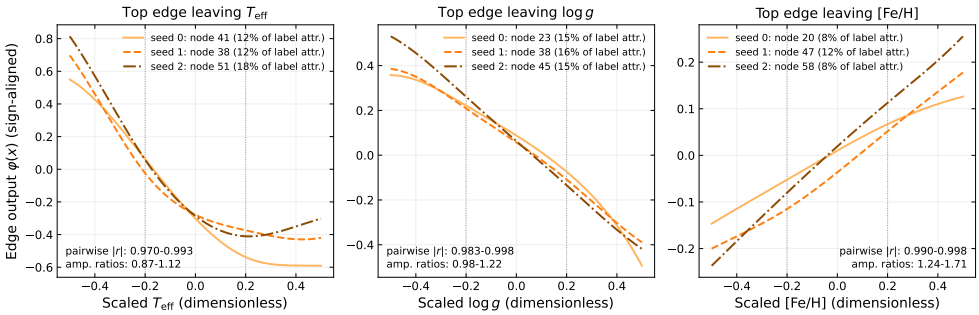


Figure 6. Seed stability of learned KAN edge functions. Response shapes are reproducible after hidden-unit matching and sign alignment, while absolute amplitudes vary more substantially.

4.5. APOGEE DR17 Consistency and Reference-Scale Offsets

On the common 9129-star quality sample, the median spectral MAEs of KAN-L, MLP-P, and MLP-L-deep are 147.1, 147.6, and 147.3 in units of 10^{-4} in continuum-normalized flux. The three emulators therefore fit the observed spectra to similar residual levels; differences in their recovered labels reflect the forward-model geometry and degeneracies rather than large differences in the final spectral mismatch.

Table 4 compares the inferred labels with calibrated ASPCAP `PARAM` values and the raw spectroscopic `FPARAM` scale. Relative to calibrated ASPCAP, KAN-L has the smallest robust scatter in $T_{\rm eff}$ (91.3 K versus 100.2 and 106.8 K) and $\log g$ (0.170 dex versus 0.193 and 0.203 dex). Against the raw scale, KAN-L and MLP-P have statistically indistinguishable $T_{\rm eff}$ scatter (89.3 and 89.7 K) and nearly identical $[\mathrm{Fe/H}]$ scatter (0.0479 and 0.0482 dex). The apparent KAN advantage in $[\mathrm{Fe/H}]$ in the legacy run (scatter ratio 0.600) does not persist with the tuned MLP baseline: the MLP-P scatter decreases from 0.0880 to 0.0482 dex (45%), whereas the KAN scatter decreases from 0.0528 to 0.0479 dex (9%). Relative to MLP-L-deep, KAN-L retains lower scatter in all the reported label–scale comparisons.

The median calibrated-scale surface-gravity offsets are +0.143, +0.127, and +0.158 dex for KAN-L, MLP-P, and MLP-L-deep. Against the raw spectroscopic scale, they fall to +0.050, +0.035, and +0.068 dex. The decrease of 0.089–0.093 dex is nearly model independent and matches ASPCAP's own -0.099 dex calibrated-minus-raw shift. Thus approximately 45–85% of the apparent calibrated-scale offsets, depending on the model, arises from the reference-scale convention rather than from the emulator. A residual raw-scale offset remains and cannot be uniquely assigned among emulator error, the synthetic spectral model, continuum placement, LSF mismatch, or label calibration. An independent forward-

fitting analysis reports a +0.17 dex spectroscopic-minus-asteroseismic log g systematic [18], matching the sign and size of the calibrated-scale offsets found here.

Table 4. APOGEE DR17 label residuals (fitted minus ASPCAP) for the new-protocol deployment (KAN-L, MLP-P, and MLP-L-deep seed-0 checkpoints; one training seed per architecture) on the common 9,129-star quality-controlled sample against the calibrated (PARAM) and raw spectroscopic (FPARAM) ASPCAP scales. Median and robust scatter (1.4826 MAD) are in the units of column 1. Paired star-level bootstrap (10,000 resamples) gives KAN-L/MLP-P scatter ratios of $T_{\rm eff}$ calibrated 0.911 [0.887, 0.939]; $T_{\rm eff}$ raw 0.995 [0.969, 1.027]; log g calibrated 0.881 [0.857, 0.905]; log g raw 0.876 [0.854, 0.898]; [Fe/H] 0.994 [0.965, 1.027]; the intervals cover star-sampling noise only, not training-seed scatter [1].

		vs. PARAM (Calibrated)		vs. FPARAM (Raw)	
Label	**Model**	**Median**	**1.4826 MAD**	**Median**	**1.4826 MAD**
$T_{\rm eff}$ (K)	KAN-L	−47.6	91.3	+39.5	89.3
	MLP-P	−38.3	100.2	+51.5	89.7
	MLP-L-deep	−58.7	106.8	+28.9	97.8
log g (dex)	KAN-L	+0.143	0.170	+0.050	0.190
	MLP-P	+0.127	0.193	+0.035	0.217
	MLP-L-deep	+0.158	0.203	+0.068	0.225
[Fe/H] (dex) [2]	KAN-L	+0.027	0.048	–	–
	MLP-P	+0.013	0.048	–	–
	MLP-L-deep	+0.030	0.065	–	–

[1] ASPCAP's own calibration (PARAM minus FPARAM) shifts the same stars by a median of +85.9 K in $T_{\rm eff}$ and −0.099 dex in log g. [2] No DR17 calibration exists for [M/H] or [Fe/H]; the metallicity comparison is already on the raw scale.

The binned residuals show an additional structural difference. On the raw scale, the KAN-L median log g residual is nearly flat across the red-clump interval, with a red-clump minus lower-giant-branch difference of +0.021 dex. The MLP-P residual changes by +0.154 dex over the same comparison and follows the ASPCAP calibration curve more closely. This does not establish external accuracy, but it suggests that the KAN-L solution is less aligned with the imposed calibrated-gravity trend.

The five-start experiment confirms that the 25-label objective is multimodal for both architectures. Across the 200-star subset, the median max–min spreads are 85.8 K, 0.112 dex, and 0.046 dex for KAN-L and 62.8 K, 0.109 dex, and 0.026 dex for MLP-P in $T_{\rm eff}$, log g, and [Fe/H]. For 10.5–11.0% of stars, the best-of-five solution differs from the ASPCAP-initialized solution by more than 1% of at least one atmospheric-label range. Population-level comparisons are therefore stable under a common initialization protocol, but a single-start solution is not guaranteed to be the global optimum for an individual star. The synthetic basin-consistency experiment gives a similar result. Across the 50 noiseless holdout spectra, all five starts agree on the full 25-label solution for only 0–6% of the spectra depending on the model, and KAN-L does not exhibit greater basin consistency than MLP-P.

Extending the fits from 600 to 1200 steps changes the median objective by only -1.6×10^{-5} for KAN-L and -1.7×10^{-5} for MLP-P. The median $|\Delta T_{\rm eff}|$ is approximately 1.2–1.3 K, much smaller than the start-to-start spread and the model-to-ASPCAP scatter. The fixed 600-step budget is adequate for the reported population statistics, although 15.5% of KAN-L and 6.0% of MLP-P synthetic fits can still improve their objective by more than 1% with a further 600 steps. The fitted-minus-ASPCAP residual distributions are shown in Figure 7, and the corresponding calibrated-versus-raw surface-gravity trends are shown in Figure 8.

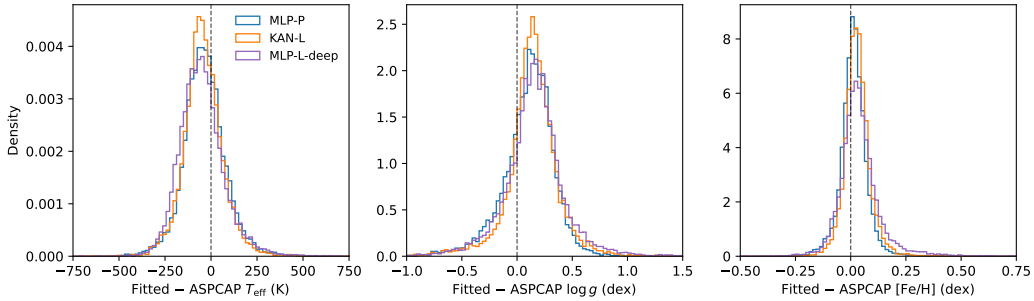


Figure 7. APOGEE DR17 comparison for the primary revised-protocol emulators: distributions of fitted-minus-ASPCAP residuals in $T_{\rm eff}$, $\log g$, and $[\mathrm{Fe/H}]$ (calibrated scale) on the common 9129-star quality sample. The three models achieve similar spectral residuals but different label distributions relative to ASPCAP.

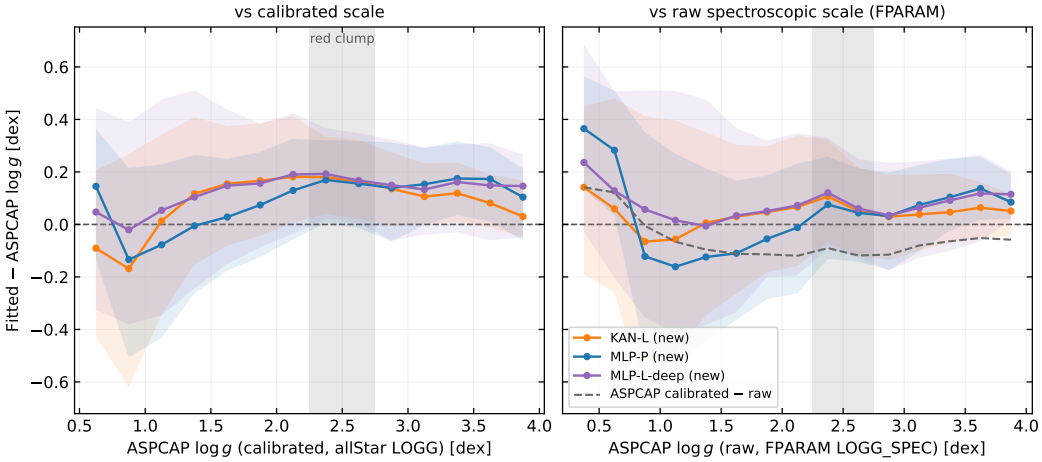


Figure 8. Decomposition of the APOGEE $\log g$ offset across calibrated and raw reference scales. The nearly common 0.09-dex shift across models follows ASPCAP's own calibration. Colored bands show the 16th–84th percentile residual range within each $\log g$ bin, and the vertical gray band marks the red-clump interval.

5. Discussion

5.1. Emulation Performance and Computational Cost

The controlled comparison provides no evidence for a general KAN advantage. At approximately 2.3 million parameters, KAN-P is no more accurate than MLP-P and is substantially slower. Improved performance is observed only at the larger scale. Relative to the small MLP and the wide–shallow large MLP, KAN-L reduces the bulk flux error and the extreme residual tail and improves the known-truth recovery of the principal stellar parameters. Its learned response shapes are also stable across seeds. Relative to MLP-P, however, KAN-L increases the training wall-clock time by a factor of 2.9 and the single-spectrum latency by more than an order of magnitude.

The preferred architecture therefore depends on the capacity and deployment requirements. For a compact Payne-like emulator or a CPU application in which latency is critical, the tuned MLP provides greater computational efficiency. For applications based on a fixed synthetic grid, KAN-L provides lower errors among the architectures tested here when accuracy and residual-tail control take priority and batched GPU inference is feasible.

5.2. Capacity, Architecture, and Training-Set Size

The large-model result is not simply a parameter-count effect. MLP-L, MLP-L-bal, and MLP-L-deep have similar parameter counts but substantially different errors, demonstrating that MLP shape matters. MLP-L-deep closes part of the gap to KAN-L, but its MAE remains about 30% higher. Conversely, the failure of KAN-P shows that edge splines do not automatically provide parameter efficiency.

Within the tested hyperparameter grid, explicit MLP weight decay does not materially improve the result. The largest observed single-seed gain is small relative to the KAN-L–MLP gap, and the best regularized large MLP remains well above KAN-L. This does not exclude other regularization strategies or a larger MLP search; it shows only that the tested KAN advantage is not reproduced by the weight-decay controls used here.

Table 5 reports two additional axes. Replacing the KAN-L training loss with L1 leaves MAE nearly unchanged but increases RMSE and Q_{99}, so MSE is the more tail-conservative objective. Training-set-size experiments at 60,000 updates show similar empirical slopes for MLP-P and KAN-P and a steeper slope for KAN-L. KAN-L is statistically tied with the small models at $N = 100$ and develops its advantage as the sample grows. The fitted exponents describe an exploratory trend within the public grid rather than a converged scaling law.

Table 5. Ablations. (a) Loss ablation for KAN-L (three seeds each; only the training loss differs): test-200 metrics in units of 10^{-4} in continuum-normalized flux, with the L1−MSE difference and the number of seed pairs whose difference shares the sign of the mean difference. (b) Data-scaling arm: [2] power-law fits of good-pixel MAE vs. training-set size $N \in \{100, 200, 400, 720\}$ (three seeds per point), read as an exploratory scaling trend.

(a) Training-loss ablation (KAN-L) [1]				
Metric	MSE (main)	L1	L1−MSE	Seeds
MAE (good)	20.60 ± 0.25	20.92 ± 0.20	+1.6 %	2/3
RMSE (good)	53.49 ± 0.89	59.83 ± 3.23	+11.8 %	3/3
Q99 (good)	235.66 ± 7.20	261.15 ± 9.83	+10.8 %	3/3
MAE (all)	21.63 ± 0.23	22.11 ± 0.21	+2.2 %	3/3
RMSE (all)	55.83 ± 0.88	62.79 ± 3.44	+12.5 %	3/3
Q99 (all)	245.12 ± 7.36	273.10 ± 11.04	+11.4 %	3/3

(b) Data-scaling exponents (60,000-update arm)					
Architecture	Exponent α	R^2	MAE@$N = 100$	MAE@$N = 720$	Ratio
MLP-P	-0.378 ± 0.032	0.986	71.5	34.9	2.05
KAN-P	-0.376 ± 0.027	0.990	75.2	36.1	2.08
KAN-L	-0.567 ± 0.053	0.983	73.7	23.8	3.10

[1] MSE is the tail-conservative choice: L1 leaves MAE statistically unchanged but inflates RMSE and Q99. The legacy KAN run (L1 loss) reached MAE 20.89 on the same 200 stars, consistent with the new-protocol L1 arm (20.92). [2] The scaling arm uses a half budget of 60,000 updates; its values are not directly comparable to Table 2, and we read the fitted exponents as an exploratory scaling trend rather than a converged scaling law. Exponent ± standard error and R^2 are from the seed-mean fit of $\log_{10}$ MAE vs. $\log_{10} N$; ratio = MAE@$N = 100$/MAE@$N = 720$. At $N = 100$ the three architectures are statistically tied.

The corresponding training-set-size trends are shown in Figure 9.

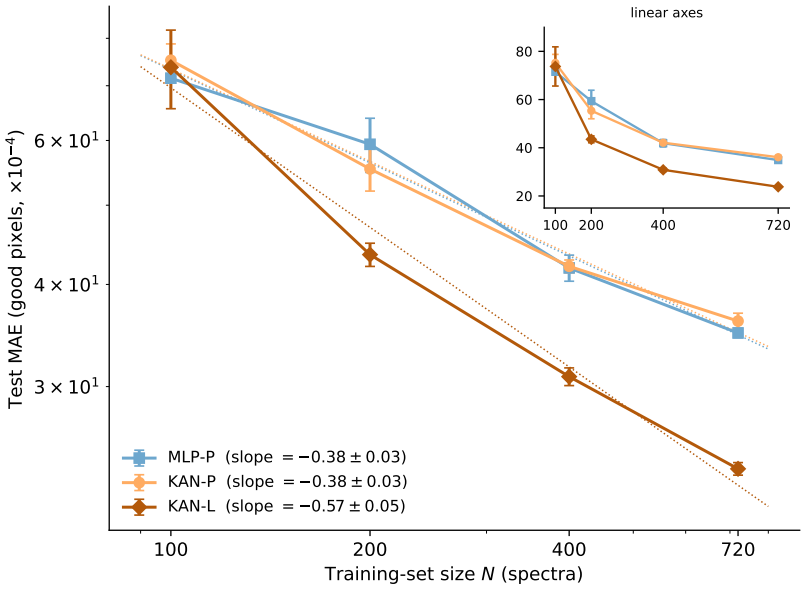


Figure 9. Exploratory data-efficiency experiment. KAN-L is tied with the small models at $N = 100$ and improves more steeply as the public training sample grows. Blue squares denote MLP-P, light-orange circles denote KAN-P, and dark-orange diamonds denote KAN-L; dotted lines show the fitted power-law trends.

5.3. Residual-Tail Errors and Stellar-Label Recovery

Average flux error can obscure the scientific impact of a small number of catastrophic pixels. In a gradient-based fit, a residual aligned with a label-sensitive direction can shift the optimum even when the global MAE is small. KAN-L's lower RMSE, Q_{99}, worst-star error, and worst-pixel error therefore provide information not contained in the MAE alone. The injection–recovery experiment confirms that the lower tail error translates into improved recovery of the main atmospheric labels in the tested high- and moderate-S/N regimes.

The mechanism is not fully isolated. The results are consistent with the hypothesis that learnable spline responses represent local label nonlinearities more effectively than fixed activations, but the present experiment evaluates the chosen KAN parameterization as a whole. It does not include an MLP with learnable B-spline activations, so it cannot separate the contribution of the spline basis from the edge-function topology. The Jacobian comparison likewise establishes cross-model concordance, not physical derivative truth.

5.4. Limitations and Deployment Scope

The main limitations are as follows.

5.4.1. Label-Space Coverage

The conclusions are conditional on the public 1000-spectrum grid. Several abundance dimensions underfill their released ranges, and $^{12}C/^{13}C$ is not reliably recovered at this training density. A larger grid generated with the same atmosphere code and custom line list would enable evaluation at higher training density. Incorporating spectra based on a different line list into the present benchmark would confound differences in the spectral models with emulator error.

5.4.2. Recovery Comparison

MLP-L-deep, the best large-MLP layout in the flux benchmark, was not included in the injection–recovery experiment. The current known-truth result therefore does not determine whether KAN-L also improves label recovery relative to that model.

5.4.3. Optimization and Uncertainties

The 25-label objective is multimodal for all the tested emulators. The population statistics are comparable because the same initialization and fitting budget are applied to every model, but a single-start fit is not guaranteed to find the global optimum for an individual star. The flux residuals, injection–recovery scatter, start dependence, and agreement with ASPCAP are distinct diagnostics. None is a calibrated per-star uncertainty. Production inference would require multi-start optimization or posterior sampling, repeat observations, and calibration against independent reference data.

5.4.4. Observed-Spectrum Systematics

The continuum-pixel mask is reconstructed from the public grid rather than taken from the original binary product. Fiber-dependent APOGEE LSF variation is not modeled explicitly and can be partly absorbed by v_{macro}. The remaining raw-scale $\log g$ offsets require external calibration against asteroseismic or cluster samples regardless of emulator architecture.

5.4.5. Implementation Cost

The models fit within consumer GPU memory, and batched GPU inference is fast. Single-spectrum and CPU latency are the principal deployment limitations for the KAN. Because direct synthesis methods are becoming competitive for small samples and relatively simple physical models, emulators provide the greatest computational benefit for repeated inference over large samples or for interpolation within grids generated with more expensive physical models.

6. Conclusions

We have presented a controlled evaluation of Kolmogorov–Arnold networks as Payne-style stellar spectral emulators on the public Kurucz grid released with The Payne. Parameter count, optimizer updates, validation-based hyperparameter selection, and training seeds are controlled explicitly.

At approximately 2.3 million parameters, the KAN does not outperform a tuned GELU MLP. At approximately 18.6 million parameters, KAN-L achieves a holdout MAE of 20.60 ± 0.25, 35% below the best small model; the best of three tested large-MLP layouts remains about 30% higher. KAN-L also has the lightest extreme residual tail and is the only model trained from scratch whose RMSE and Q_{99} are lower than those of the public pretrained Payne reference, while the reference retains the best bulk MAE.

In synthetic fits with known labels, KAN-L reduces the robust scatter of T_{eff}, $\log g$, and [Fe/H] relative to MLP-P at every tested S/N, although the best large-MLP shape was not included in this experiment. The learned KAN edge-function shapes and label-attribution rankings are reproducible across seeds, supporting a limited claim of inspectability. The computational penalties remain substantial: relative to MLP-P, KAN-L requires 2.9 times the training time and 13–57 times the single-spectrum GPU–CPU latency.

In APOGEE DR17, KAN-L gives the smallest calibrated-scale scatter in T_{eff} and $\log g$ among the three revised-protocol models, whereas [Fe/H] is statistically tied with MLP-P. The positive calibrated-scale $\log g$ offsets of all the models decrease by approximately 0.09 dex on the raw spectroscopic scale, matching ASPCAP's own calibration shift; the remaining offsets cannot be assigned uniquely to the emulator.

These results do not support KANs as cheaper or universally better replacements for MLP spectral emulators. They show instead that, for the public low-data benchmark studied here, a sufficiently large KAN converts additional capacity into flux accuracy and

residual-tail robustness more effectively than the tested MLP layouts at a cost that matters for deployment.

Author Contributions: Conceptualization, R.W.; methodology, S.Z. and R.W.; software, R.W.; validation, S.Z.; draft preparation, S.Z.; project administration, R.W.; funding acquisition, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, grant number 2025YFF0511001, and by the National Astronomical Observatories, Chinese Academy of Sciences, under Project No. E4ZR050901.

Data Availability Statement: The APOGEE DR17 spectra and ASPCAP catalogue are available from the SDSS Science Archive Server. The Payne synthetic spectra are available from the public repository for The Payne. Analysis code is available at https://github.com/wangruibistu/kan_payne (accessed on 7 August 2026). The reproducibility package, including data splits, model outputs, injection–recovery results, continuum masks, APOGEE fit summaries, and figure/table source arrays, is published as a Zenodo record at https://doi.org/10.5281/zenodo.20782050; this DOI resolves to the latest version of the archive.

Acknowledgments: We thank the two anonymous referees for constructive comments that substantially improved the experimental design and presentation. This work used public data from SDSS-IV APOGEE and public synthetic spectra and software from The Payne and TransformerPayne projects. Numerical experiments were performed on local NVIDIA A100 resources. During the preparation and revision of this manuscript, the authors used ChatGPT (GPT-5.6 Pro, GPT-5.6 Sol; OpenAI) and Claude (Fable 5, Opus 5; Anthropic) to assist with literature organization, review of the experimental design and protocol, drafting and debugging of analysis code, LaTeX preparation, and language and structural editing of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Local Wavelength-Conditioned Reference Model

Appendix A.1. Architecture and Training Protocol

For completeness, we document the wavelength-conditioned reference model used in the original analysis. The label vector is encoded into 16 learned tokens of width 128. A wavelength, scaled to $[-1, 1]$, is expanded with sinusoidal features, and four cross-attention blocks with four heads update a wavelength query against the label tokens [35]. A regression head predicts the scalar flux at the requested wavelength. During training, 256 wavelength pixels are sampled per spectrum per step; the full 7214-pixel spectrum is evaluated in chunks.

The model contains 620,929 trainable parameters. Under the original 800/200 split, L1 loss, AdamW learning rate 10^{-4}, batch size 32, and 10,000 epochs, it reaches a good-pixel MAE of 35.18 and RMSE of 86.94. These numbers should not be interpreted as a controlled comparison with the published TransformerPayne architecture [20]: the local model is much smaller, receives less pixel-level supervision per update, and was trained for a much smaller effective budget.

Appendix A.2. Legacy APOGEE Deployment

The original 10,000-star deployment used the local wavelength-conditioned model together with the original MLP and KAN checkpoints. On the common 9117-star sample, their residuals relative to calibrated and raw ASPCAP scales are listed in Table A1. These results are retained to document the original deployment and the reference-scale decomposition, not to rank the revised primary architectures.

Table A1. APOGEE DR17 label residuals (fitted minus ASPCAP) on the common 9117-star quality-controlled sample compared against the calibrated (PARAM) and raw spectroscopic (FPARAM) ASPCAP scales. Median and robust scatter (1.4826 MAD) are in the units of column 1. Legacy-protocol fits [1].

		vs. PARAM (Calibrated)		vs. FPARAM (Raw)	
Label	**Model**	**Median**	**1.4826 MAD**	**Median**	**1.4826 MAD**
$T_{\rm eff}$ (K)	Payne-MLP	−62.9	124.8	+30.0	127.6
	KAN-Payne	−17.4	112.9	+71.7	102.6
	TransformerPayne	−47.1	136.4	+37.8	122.5
$\log g$ (dex)	Payne-MLP	+0.105	0.222	+0.016	0.256
	KAN-Payne	+0.197	0.199	+0.108	0.208
	TransformerPayne	+0.173	0.193	+0.081	0.211

[1] ASPCAP's own calibration (PARAM minus FPARAM) shifts the same stars by a median of +88.1 K in $T_{\rm eff}$ and −0.096 dex in $\log g$; part of each apparent offset is therefore a scale choice, not a model error.

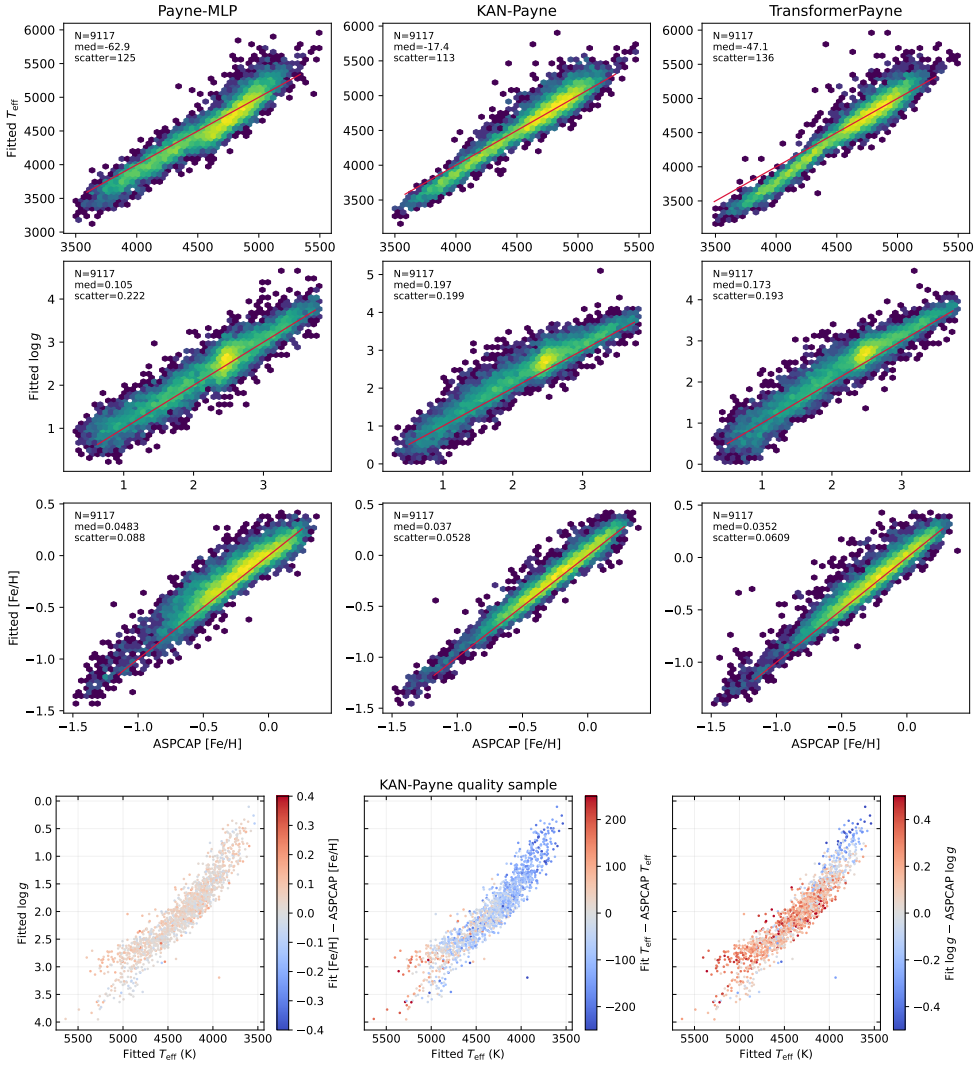


Figure A1. Legacy-protocol APOGEE diagnostics: fitted versus ASPCAP atmospheric labels for the three legacy emulators on the 9117-star common sample (**top**) and Kiel-diagram residual diagnostics for the legacy KAN-Payne fits (**bottom**). In the upper panels, color encodes the local source density from purple (lower density) to yellow (higher density), and the red diagonal line marks the one-to-one relation. These plots document the original deployment and are not part of the controlled revised benchmark.

Appendix A.3. Start Dependence

A 50-star start-dependence test was performed for the wavelength-conditioned reference model. Midpoint and random starts coincide for most stars, but 6 of 50 stars reach a distinct solution differing by more than 1% of at least one principal label range; the 95th-percentile $|\Delta T_{\mathrm{eff}}|$ is 148 K. The result is consistent with the multimodality observed for the vector-output emulators.

References

1. Majewski, S.R.; Schiavon, R.P.; Frinchaboy, P.M.; Allende Prieto, C.; Barkhouser, R.; Bizyaev, D.; Blank, B.; Brunner, S.; Burton, A.; Carrera, R.; et al. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *Astron. J.* **2017**, *154*, 94. [CrossRef]
2. Abdurro'uf; Accetta, K.; Aerts, C.; Silva Aguirre, V.; Ahumada, R.; Ajgaonkar, N.; Filiz Ak, N.; Alam, S.; Allende Prieto, C.; Almeida, A.; et al. The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data. *Astrophys. J. Suppl. Ser.* **2022**, *259*, 35. [CrossRef]
3. Luo, A.L.; Zhao, Y.H.; Zhao, G.; Deng, L.C.; Liu, X.W.; Jing, Y.P.; Wang, G.; Zhang, H.T.; Shi, J.R.; Cui, X.Q.; et al. The First Data Release (DR1) of the LAMOST Regular Survey. *Res. Astron. Astrophys.* **2015**, *15*, 1095–1124. [CrossRef]
4. Buder, S. et al. [GALAH Collaboration]. The GALAH+ Survey: Third Data Release. *Mon. Not. R. Astron. Soc.* **2021**, *506*, 150–201. [CrossRef]
5. Kordopatis, G.; Gilmore, G.; Steinmetz, M.; Boeche, C.; Seabroke, G.M.; Siebert, A.; Zwitter, T.; Binney, J.; de Laverny, P.; Recio-Blanco, A.; et al. The Radial Velocity Experiment (RAVE): Fourth Data Release. *Astron. J.* **2013**, *146*, 134. [CrossRef]
6. García Pérez, A.E.; Allende Prieto, C.; Holtzman, J.A.; Shetrone, M.; Mészáros, S.; Bizyaev, D.; Carrera, R.; Cunha, K.; García-Hernández, D.A.; Johnson, J.A.; et al. ASPCAP: The APOGEE Stellar Parameter and Chemical Abundances Pipeline. *Astron. J.* **2016**, *151*, 144. [CrossRef]
7. Allende Prieto, C.; Beers, T.C.; Wilhelm, R.; Newberg, H.J.; Rockosi, C.M.; Yanny, B.; Lee, Y.S. A Spectroscopic Study of the Ancient Milky Way: F- and G-Type Stars in the Third Data Release of the Sloan Digital Sky Survey. *Astrophys. J.* **2006**, *636*, 804–820. [CrossRef]
8. Wu, Y.; Luo, A.L.; Li, H.N.; Shi, J.R.; Prugniel, P.; Liang, Y.C.; Zhao, Y.H.; Zhang, J.N.; Bai, Z.R.; Wei, P.; et al. Automatic Determination of Stellar Atmospheric Parameters and Construction of Stellar Spectral Templates of the Guoshoujing Telescope (LAMOST). *Res. Astron. Astrophys.* **2011**, *11*, 924–946. [CrossRef]
9. Xiang, M.S.; Liu, X.W.; Yuan, H.B.; Huang, Y.; Huo, Z.Y.; Zhang, H.W.; Chen, B.Q.; Zhang, H.H.; Sun, N.C.; Wang, C.; et al. The LAMOST Stellar Parameter Pipeline at Peking University – LSP3. *Mon. Not. R. Astron. Soc.* **2015**, *448*, 822–854. [CrossRef]
10. Ness, M.; Hogg, D.W.; Rix, H.W.; Ho, A.Y.Q.; Zasowski, G. The Cannon: A Data-driven Approach to Stellar Label Determination. *Astrophys. J.* **2015**, *808*, 16. [CrossRef]
11. Ho, A.Y.Q.; Ness, M.K.; Hogg, D.W.; Rix, H.W.; Liu, C.; Yang, F.; Zhang, Y.; Hou, Y.; Wang, Y. Label Transfer from APOGEE to LAMOST: Precise Stellar Parameters for 450,000 LAMOST Giants. *Astrophys. J.* **2017**, *836*, 5. [CrossRef]
12. Fabbro, S.; Venn, K.A.; O'Briain, T.; Bialek, S.; Kielty, C.L.; Jahandar, F.; Monty, S. An Application of Deep Learning in the Analysis of Stellar Spectra. *Mon. Not. R. Astron. Soc.* **2018**, *475*, 2978–2993. [CrossRef]
13. Leung, H.W.; Bovy, J. Deep Learning of Multi-element Abundances from High-resolution Spectroscopic Data. *Mon. Not. R. Astron. Soc.* **2019**, *483*, 3255–3277. [CrossRef]
14. Zhang, B.; Liu, C.; Deng, L.C. Deriving the Stellar Labels of LAMOST Spectra with the Stellar LAbel Machine (SLAM). *Astrophys. J. Suppl. Ser.* **2020**, *246*, 9. [CrossRef]
15. Xiang, M.; Ting, Y.S.; Rix, H.W.; Sandford, N.; Buder, S.; Lind, K.; Liu, X.W.; Shi, J.R.; Zhang, H.W. Abundance Estimates for 16 Elements in 6 Million Stars from LAMOST DR5 Low-Resolution Spectra. *Astrophys. J. Suppl. Ser.* **2019**, *245*, 34. [CrossRef]
16. Ting, Y.S.; Conroy, C.; Rix, H.W.; Cargile, P. The Payne: Self-consistent ab initio Fitting of Stellar Spectra. *Astrophys. J.* **2019**, *879*, 69. [CrossRef]
17. Wheeler, A.J.; Abruzzo, M.W.; Casey, A.R.; Ness, M.K. Korg: A Modern 1D LTE Spectral Synthesis Package. *Astron. J.* **2023**, *165*, 68. [CrossRef]
18. Ting, Y.S.; Kim, E.M. The Payne Zero Project I: Stellar Spectra from Physical Models in Seconds. *arXiv* **2026**, arXiv:2607.24141. [CrossRef]
19. Kim, E.M.; Ting, Y.S. pyKurucz: A Pure Python Reimplementation of Kurucz ATLAS12 and SYNTHE for Stellar Spectrum Synthesis. *arXiv* **2026**, arXiv:2603.11693. [CrossRef]
20. Różański, T.; Ting, Y.S.; Jabłońska, M. TransformerPayne: Enhancing Spectral Emulation Accuracy and Data Efficiency by Capturing Long-range Correlations. *Astrophys. J.* **2025**, *980*, 66. [CrossRef]
21. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. KAN: Kolmogorov–Arnold Networks. *arXiv* **2024**, arXiv:2404.19756. [CrossRef]

22. Liu, Z.; Ma, P.; Wang, Y.; Matusik, W.; Tegmark, M. KAN 2.0: Kolmogorov–Arnold Networks Meet Science. *arXiv* **2024**, arXiv:2408.10205. [CrossRef]
23. Yu, R.; Yu, W.; Wang, X. KAN or MLP: A Fairer Comparison. *arXiv* **2024**, arXiv:2407.16674. [CrossRef]
24. Sestak, K. Kolmogorov–Arnold Networks as Implicit Regularizers: Noise Robustness and Interpretability for Stellar Classification. *arXiv* **2026**, arXiv:2605.29039. [CrossRef]
25. Shen, H.; Zeng, C.; Wang, J.; Wang, Q. Reduced Effectiveness of Kolmogorov–Arnold Networks on Functions with Noise. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2025; pp. 1–5. [CrossRef]
26. Dorigo Jones, J.; Reyes, B.; Rapetti, D.; Bahauddin, S.M.; Burns, J.O.; Barker, D.W. The Wrath of KAN: Enabling Fast, Accurate, and Transparent Emulation of the Global 21 cm Cosmology Signal. *Astrophys. J.* **2025**, *991*, 152. [CrossRef]
27. Kurucz, R.L. *Atlas: A Computer Program for Calculating Model Stellar Atmospheres*; SAO Special Report 309; Smithsonian Astrophysical Observatory: Cambridge, MA, USA, 1970.
28. Kurucz, R.L.; Avrett, E.H. *Solar Spectrum Synthesis. I. A Sample Atlas from 224 to 300 nm*; SAO Special Report 391; Smithsonian Astrophysical Observatory: Cambridge, MA, USA, 1981.
29. Kurucz, R.L. *ATLAS9 Stellar Atmosphere Programs and 2 km/s Grid*; Kurucz CD-ROM No. 13; Smithsonian Astrophysical Observatory: Cambridge, MA, USA, 1993.
30. Castelli, F.; Kurucz, R.L. New Grids of ATLAS9 Model Atmospheres. *arXiv* **2004**, arXiv:astro-ph/0405087. [CrossRef]
31. Kurucz, R.L. ATLAS12, SYNTHE, ATLAS9, WIDTH9, et cetera. *Mem. Della Soc. Astron. Ital. Suppl.* **2005**, *8*, 14.
32. Kurucz, R.L. ATLAS12: Opacity Sampling Model Atmosphere Program. Astrophysics Source Code Library, 2013, Record ascl:1303.024. Available online: https://ascl.net/1303.024 (accessed on 20 August 2026).
33. Wilson, J.C.; Hearty, F.R.; Skrutskie, M.F.; Majewski, S.R.; Holtzman, J.A.; Eisenstein, D.; Gunn, J.; Blank, B.; Henderson, C.; Smee, S.; et al. The Apache Point Observatory Galactic Evolution Experiment (APOGEE) Spectrographs. *Publ. Astron. Soc. Pac.* **2019**, *131*, 055001. [CrossRef]
34. Jönsson, H.; Holtzman, J.A.; Allende Prieto, C.; Cunha, K.; García-Hernández, D.A.; Hasselquist, S.; Masseron, T.; Osorio, Y.; Shetrone, M.; Smith, V.; et al. APOGEE Data and Spectral Analysis from SDSS Data Release 16: Seven Years of Observations Including First Results from APOGEE-South. *Astron. J.* **2020**, *160*, 120. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: New York, NY, USA, 2017; pp. 5998–6008.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.