TransformerPayne: Enhancing Spectral Emulation Accuracy and Data Efficiency by Capturing Long-Range Correlations

https://arxiv.org/abs/2407.05751

Tomasz Różański, Yuan-Sen Ting, Maja Jabłońska

Australian National University

ML4Astro2, Catania, 10.07.2024

Background and motivation

- Inferring parameters of stellar atmospheres, including precise individual abundances, becomes computationally prohibitive when considering large surveys like 4MOST or WEAVE, which will collect millions of stellar spectra,
- One method to amortize the cost of inference is by developing precise and fast emulators that can replace spectral synthesis in pipelines,
- The current state-of-the-art emulator, The Payne (Ting et al., 2018), tends to saturate in prediction accuracy even as the size of the training dataset increases.

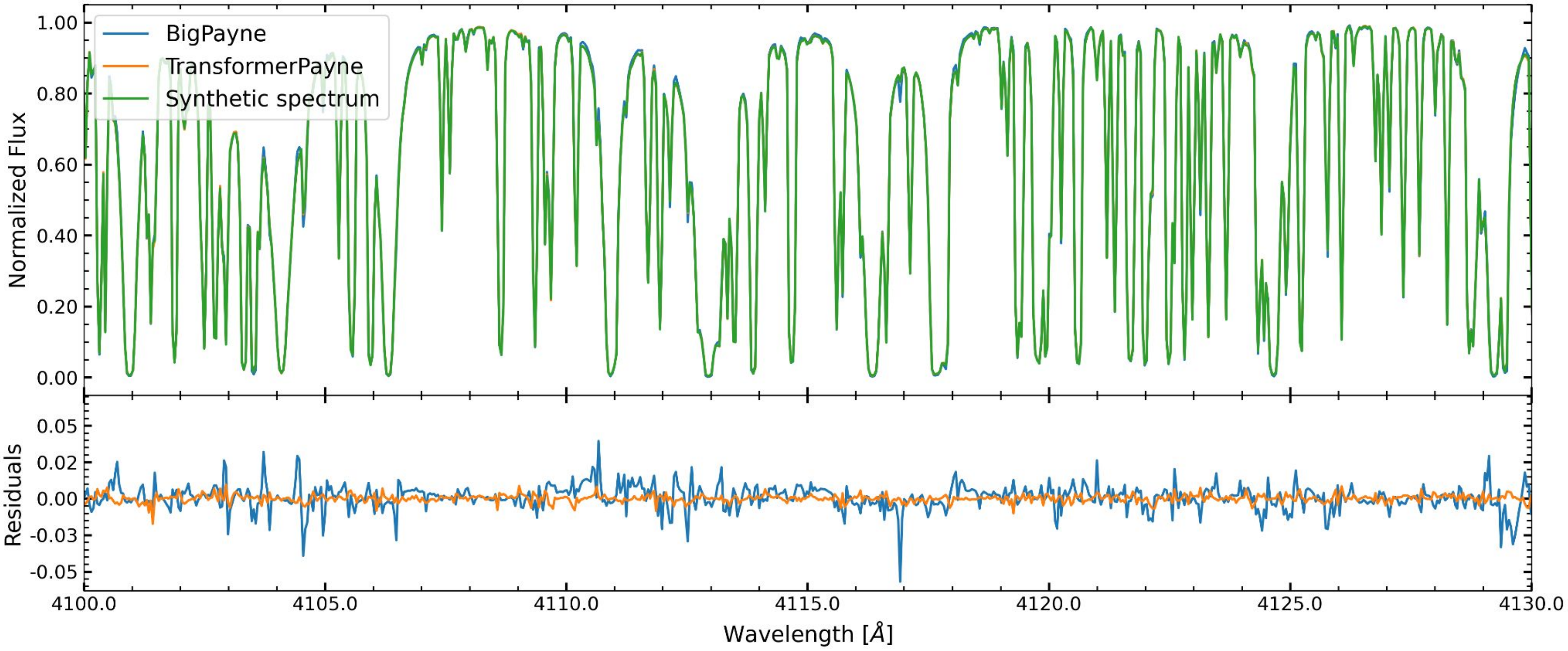
Goals

- Developing more precise spectrum emulation and reducing the number of spectra required to train an emulator with equivalent accuracy,
- Addressing the limitations of the small scale inherent in the original The Payne emulator,
- Developing a new architecture capable of efficiently handling correlations across widely separated wavelengths, such as those associated with spectral lines of the same elements in stellar spectra,
- Experimenting with fine-tuning as a way to increase data efficiency of training spectra emulators.

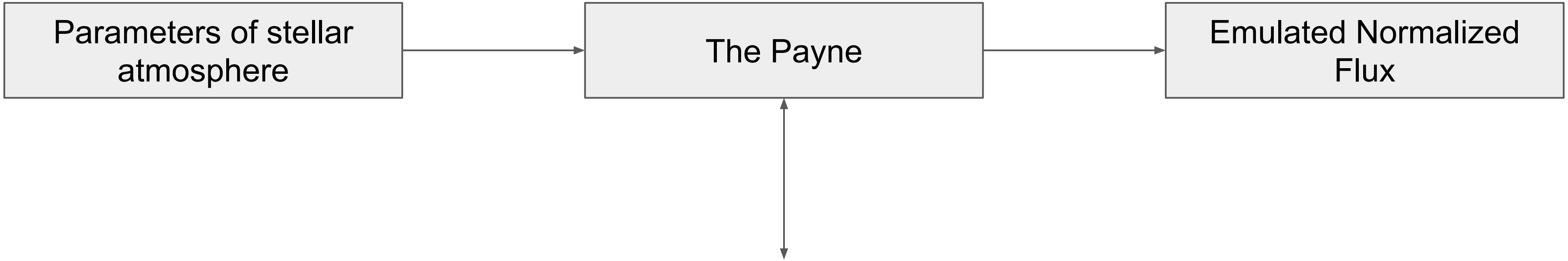
Pre-training and training datasets

Grid definition	Pre-training	Training
Effective Temperature	5000 K	$[4000, 6000]$ K
Surface Gravity	4.5	$[4.0, 5.0]$
# Training Spectra	Up to 100000	
Wavelength range	$[4000, 5000]$ Å	
# Wavelengths	22315	
Microturbulence, ξ	0 km/s	
Helium Abundance	$[0, 0.1568]$	
Other Abundances, $[X/H]^*$	$[-2, 1]$	

Emulation quality overview

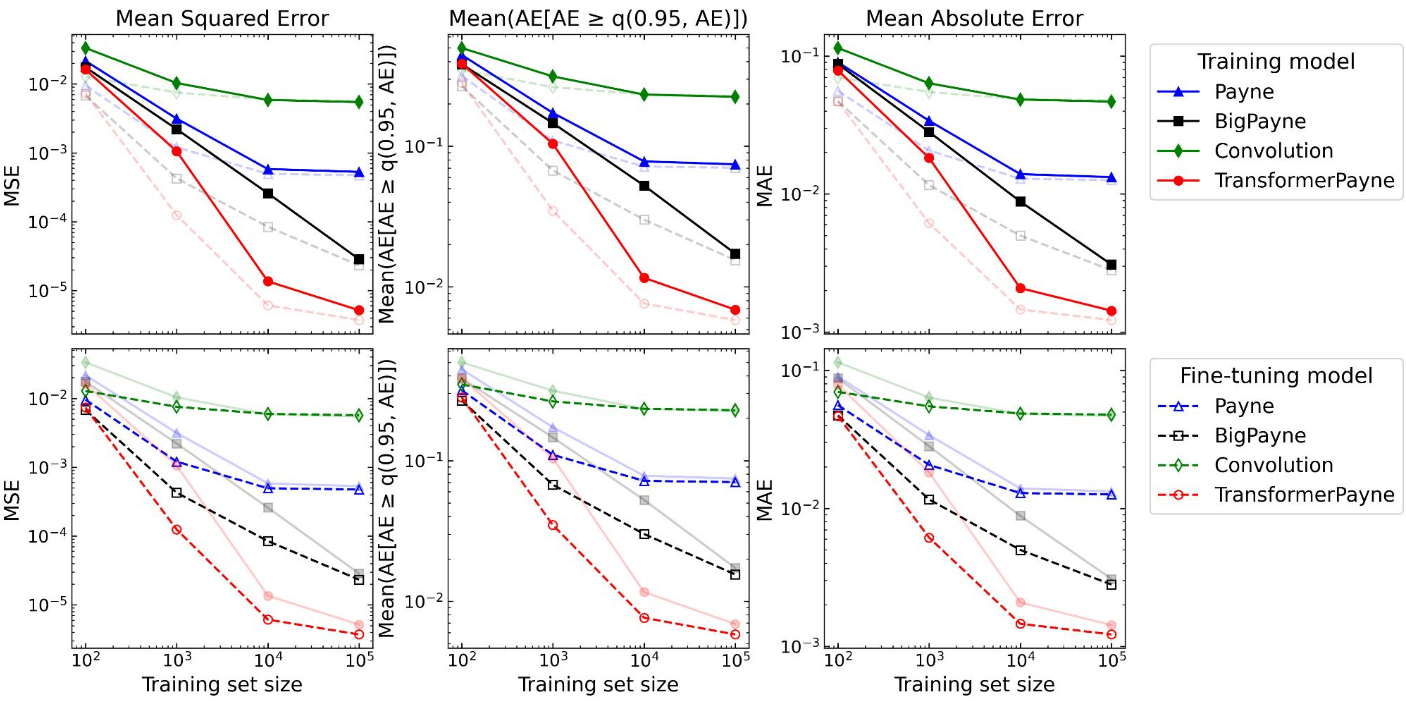


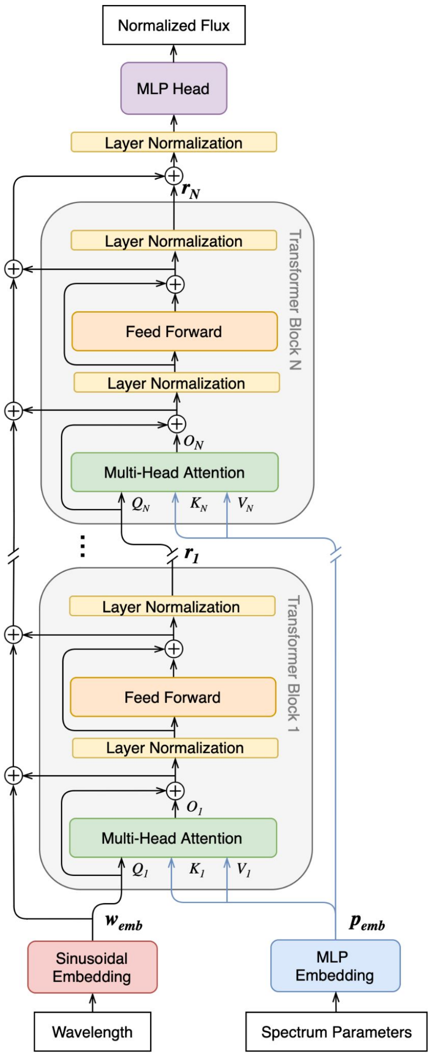
The Payne emulator



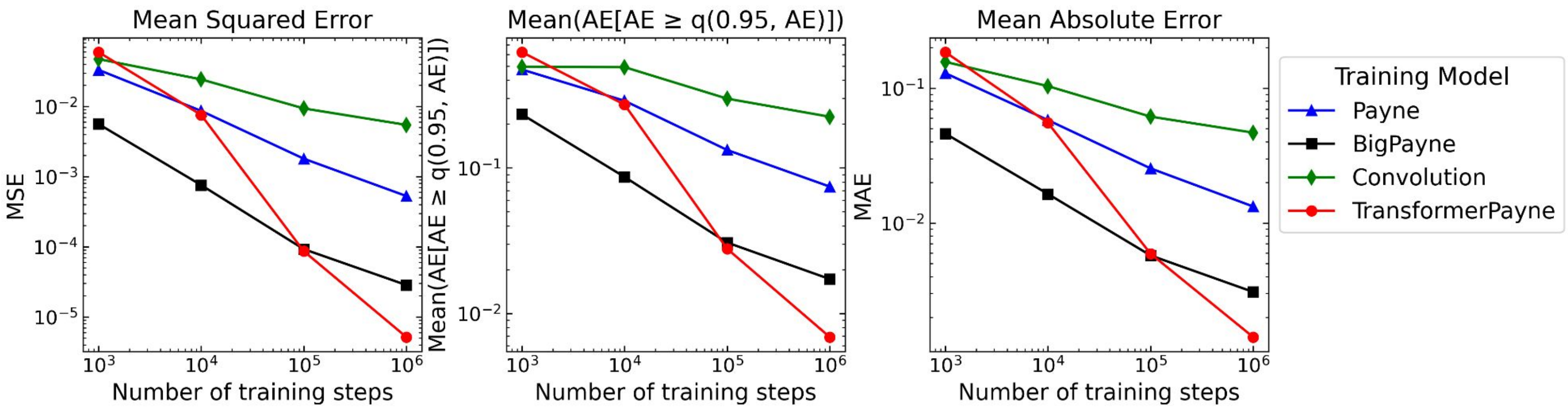
$$\vec{f} = \mathbf{W_3}\mathrm{gelu}(\mathbf{W_2}\mathrm{gelu}(\mathbf{W_1}\vec{p} + \vec{b}_1) + \vec{b}_2) + \vec{b}_3$$

TransformerPayne emulator





TransformerPayne emulator

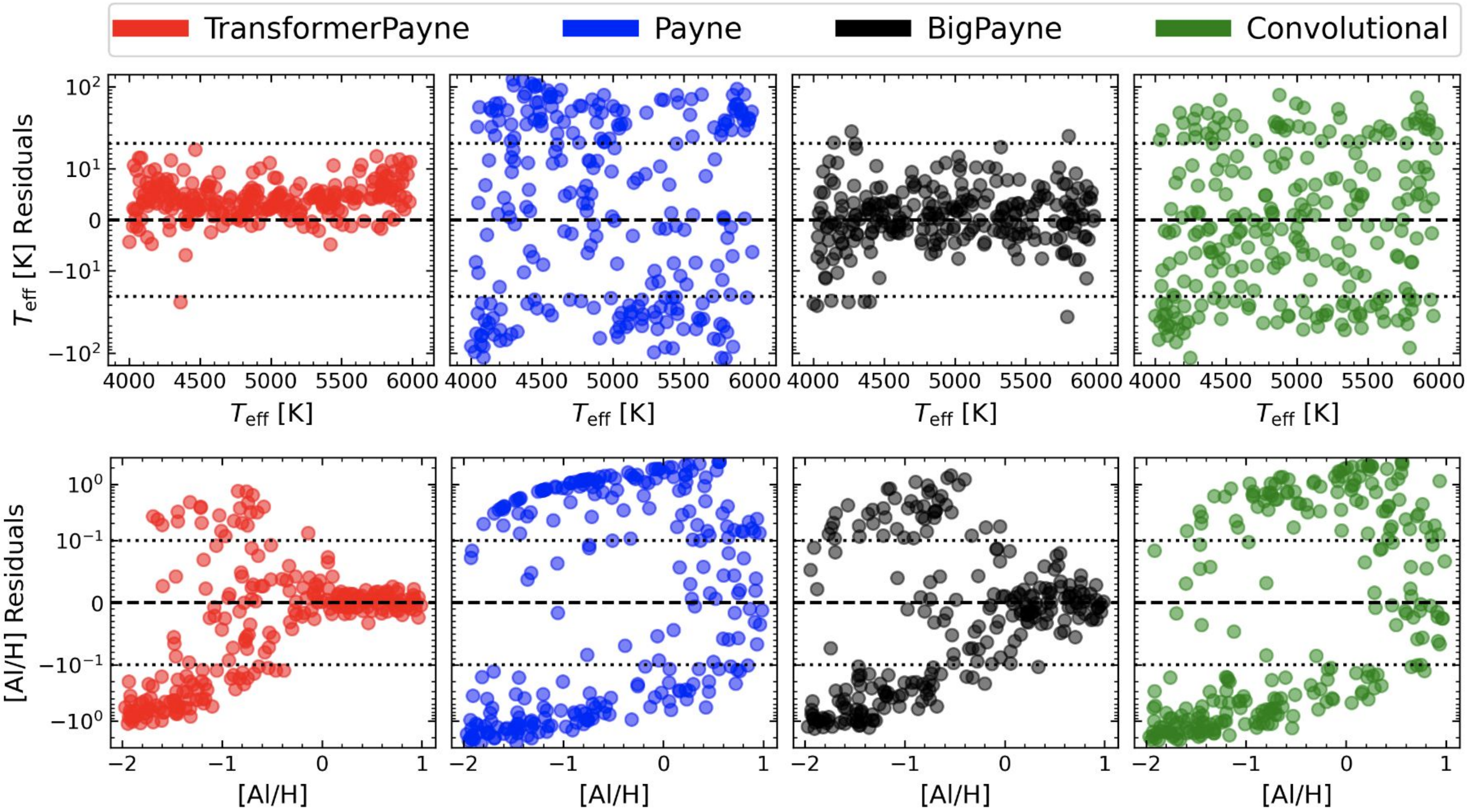


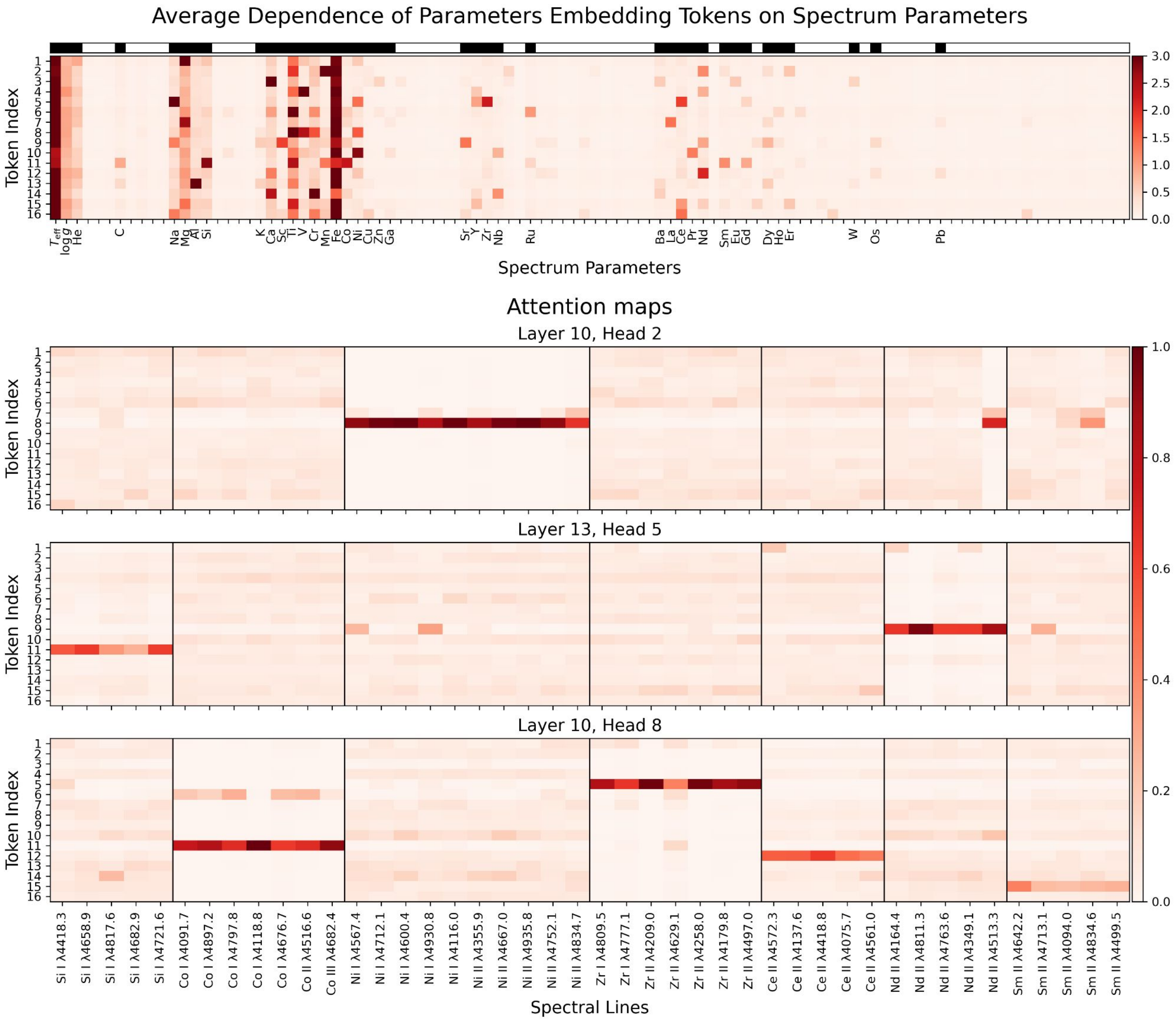
Parameters inference

Parameter	TransformerPayne (TP)	The Payne (P)	BigPayne (BP)	Convolutional (C)
T_{eff} [K]	**3.7027**	43.3008	6.3656	24.3971
$\log g$	**0.0050**	0.0632	0.0080	0.0330
$N_{\text{He}}/N_{\text{tot}}$	**0.0029**	0.0213	0.0034	0.0136

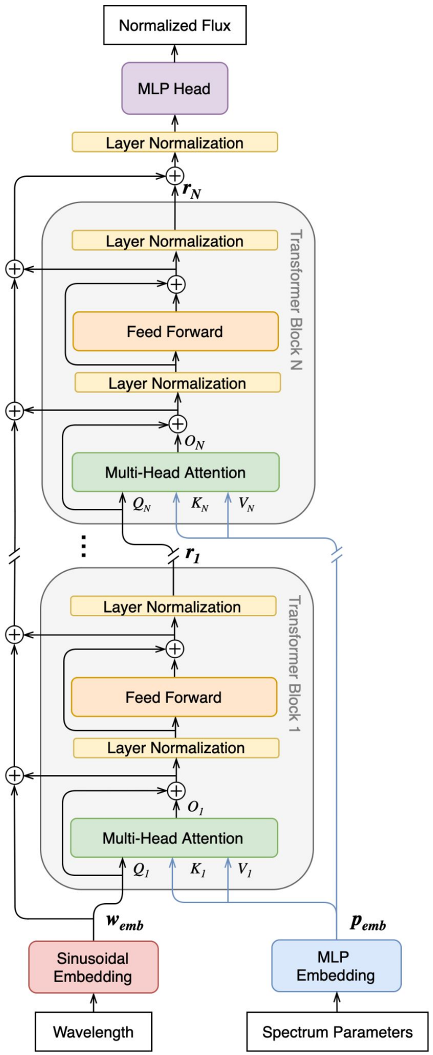
	$-2.0 \leq$ [X/H] < -1.0				$-1.0 \leq$ [X/H] < 0.0				$0.0 \leq$ [X/H] ≤ 1.0			
[X/H]	TP	P	BP	C	TP	P	BP	C	TP	P	BP	C
C	**0.4933**	0.6964	0.5955	0.7195	**0.3920**	0.6510	0.4557	0.6436	**0.4629**	0.7459	0.5566	0.8036
Na	**0.0313**	0.2824	0.0493	0.1618	**0.0076**	0.1678	0.0167	0.1484	**0.0070**	0.0592	0.0102	0.0324
Mg	**0.0058**	0.1444	0.0142	0.0689	**0.0052**	0.0759	0.0109	0.0300	**0.0068**	0.0599	0.0081	0.0317
Al	**0.3713**	0.9599	0.5278	0.7450	**0.1884**	0.9245	0.4038	0.7979	**0.0161**	0.7982	0.0332	0.7952
Si	**0.0096**	0.2231	0.0244	0.1295	**0.0089**	0.1482	0.0223	0.0922	**0.0077**	0.0996	0.0133	0.0624
K	**0.2222**	0.5804	0.2996	0.5176	**0.0208**	0.3401	0.0423	0.3774	**0.0230**	0.4331	0.0415	0.5008
Ca	**0.0057**	0.0829	0.0092	0.0306	**0.0054**	0.0498	0.0077	0.0204	**0.0052**	0.0406	0.0064	0.0210
Sc	**0.0105**	0.1320	0.0183	0.0549	**0.0090**	0.0577	0.0103	0.0368	0.0090	0.0504	**0.0088**	0.0301
Ti	**0.0052**	0.0579	0.0073	0.0231	**0.0050**	0.0466	0.0062	0.0236	**0.0053**	0.0331	0.0055	0.0176
V	**0.0064**	0.1347	0.0275	0.0766	**0.0055**	0.0529	0.0069	0.0260	**0.0055**	0.0359	0.0071	0.0231
Cr	**0.0057**	0.0734	0.0098	0.0355	**0.0045**	0.0425	0.0068	0.0176	**0.0051**	0.0337	0.0057	0.0181
Mn	**0.0068**	0.1227	0.0142	0.0319	**0.0053**	0.0445	0.0072	0.0211	**0.0059**	0.0390	0.0074	0.0191
Fe	**0.0042**	0.0516	0.0060	0.0192	**0.0038**	0.0338	0.0050	0.0175	**0.0045**	0.0333	0.0073	0.0204

Parameters inference





Average Dependence of Parameters Embedding Tokens on Spectrum Parameters
Token Index
Spectrum Parameters
Teff
logg
He
C
Na
Mg
Al
Si
K
Ca
Sc
Ti
V
Cr
Mn
Fe
Co
Ni
Cu
Zn
Ga
Sr
Y
Zr
Nb
Ru
Ba
La
Ce
Pr
Nd
Sm
Eu
Gd
Dy
Ho
Er
W
Os
Pb
Attention maps
Layer 10, Head 2
Layer 13, Head 5
Layer 10, Head 8
Spectral Lines
Si I λ4418.3
Si I λ4658.9
Si I λ4817.6
Si I λ4682.9
Si I λ4721.6
Co I λ4091.7
Co I λ4897.2
Co I λ4797.8
Co I λ4118.8
Co I λ4676.7
Co II λ4516.6
Co III λ4682.4
Ni I λ4567.4
Ni I λ4712.1
Ni I λ4600.4
Ni I λ4930.8
Ni I λ4116.0
Ni II λ4355.9
Ni II λ4667.0
Ni II λ4935.8
Ni II λ4752.1
Ni II λ4834.7
Zr I λ4809.5
Zr I λ4777.1
Zr II λ4209.0
Zr II λ4629.1
Zr II λ4258.0
Zr II λ4179.8
Zr II λ4497.0
Ce II λ4572.3
Ce II λ4137.6
Ce II λ4418.8
Ce II λ4075.7
Ce II λ4561.0
Nd II λ4164.4
Nd II λ4811.3
Nd II λ4763.6
Nd II λ4349.1
Nd II λ4513.3
Sm II λ4642.2
Sm II λ4713.1
Sm II λ4094.0
Sm II λ4834.6
Sm II λ4499.5



Normalized Flux
MLP Head
Layer Normalization
Transformer Block N
Feed Forward
Multi-Head Attention
Transformer Block 1
Sinusoidal Embedding
MLP Embedding
Wavelength
Spectrum Parameters

Summary

- Scaling The Payne emulator by increasing the number of layers and the number of neurons in each layer prevents it from saturating as the number of training spectra increases,
- Parametrizing the emulator explicitly as a function of wavelength and incorporating flexible attention blocks leads to more precise and data-efficient emulation,
- Fine-tuning serves as a method to enhance data efficiency in stellar spectrum emulation,
- An interpretability study reveals that the TransformerPayne emulator naturally learns physically relevant features without direct supervision.

https://arxiv.org/abs/2407.05751