机器学习在恒星光谱分析中的应用

向茂盛
国家天文台

内容提纲

（一）什么是恒星参数测定?

（二）为什么恒星参数测定需要机器学习?

（三）机器学习在恒星参数测定中的应用示例

（四）机器学习光谱分析面临的问题

（五）总结与展望

内容提纲

（一）什么是恒星参数测定？

（二）为什么恒星参数测定需要机器学习？

（三）机器学习在恒星参数测定中的应用示例

（四）机器学习光谱分析面临的问题

（五）总结与展望

1.1，什么是恒星参数测定？

- P(θ | O) : 给定观测数据O，求解决定参数θ

1.1，什么是恒星参数测定？

- P(θ | O) : 给定观测数据O，求解决定参数θ
- 需要参数θ与数据O之间的模型O=f(θ)+ε，如恒星大气模型+辐射转移
  - $P(\theta \mid O) \propto L(O|\theta)P(\theta)$

1.1，什么是恒星参数测定？

- P(**θ** | **O**) : 给定观测数据**O**，求解决定参数**θ**
- 需要参数**θ**与数据**O**之间的模型**O=f(θ)+**ε，如恒星大气模型**+**辐射转移
  - P(**θ** | **O**) ∝ L(**O**|**θ**)P(**θ**)

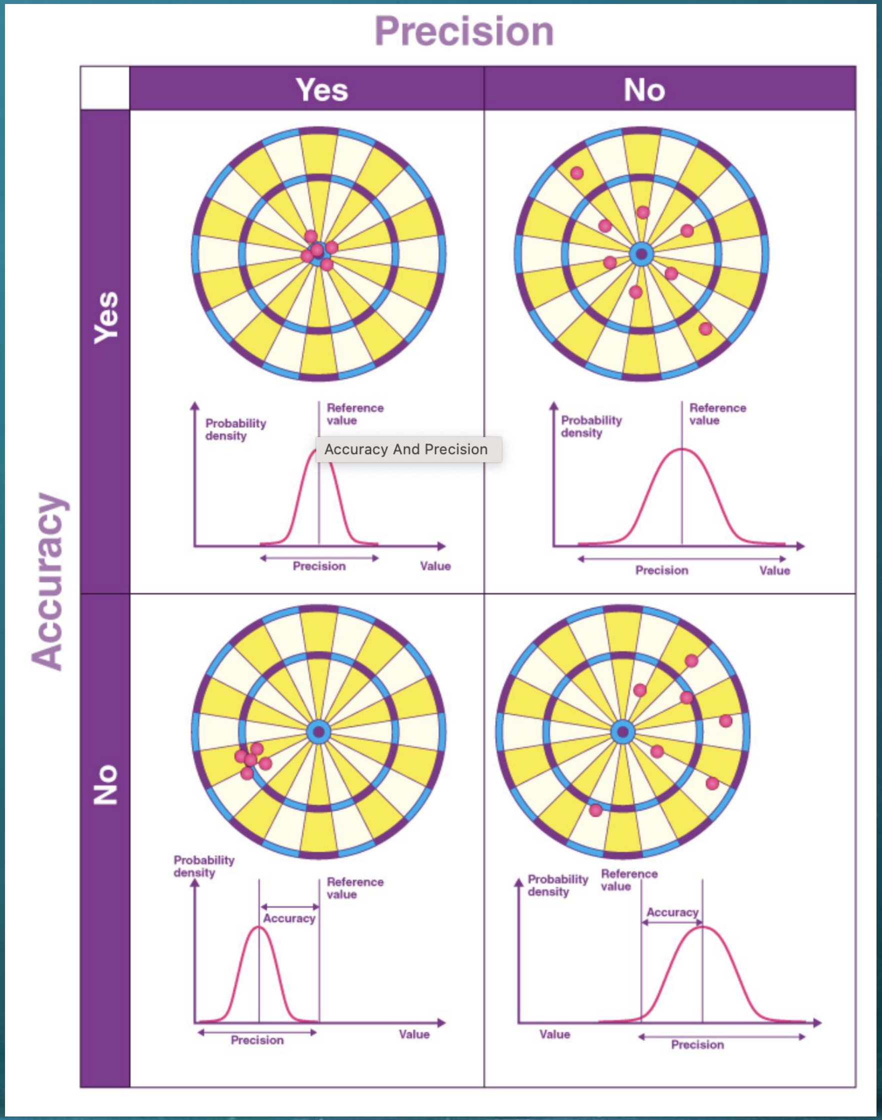
恒星参数测定是指：**给定一组观测数据O（光谱/测光），寻找一组恒星参数组合*θ*，使得模型f(*θ*)与观测匹配最佳**

1.2，参数质量评价指标

- 精度(precision)
- 准确度(accuracy)

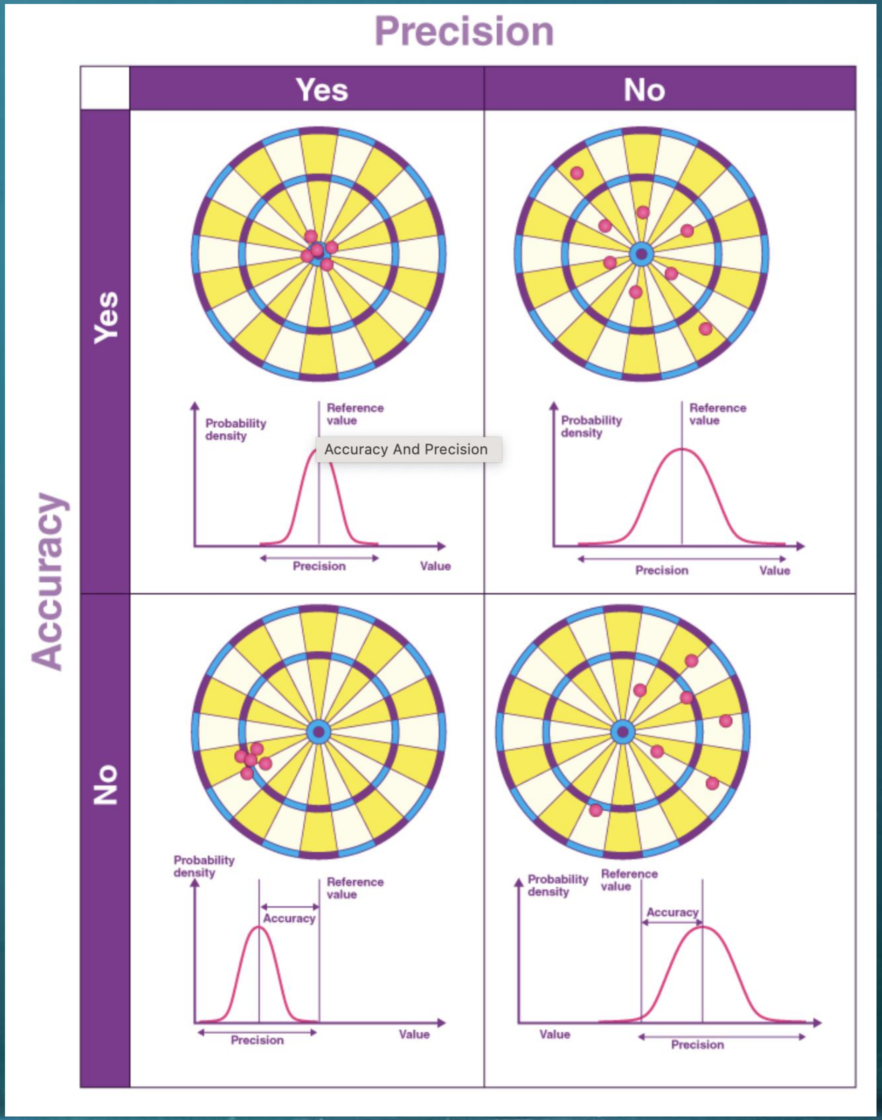
1.2，参数质量评价指标

- ◆ 精度(precision): 测量随机误差
- ◆ 准确度(accuracy): 方法系统误差



1.2，参数质量评价指标

- 精度(precision): 测量随机误差
- 准确度(accuracy): 方法系统误差
  - 模型/方法误差，定标误差
  - 非随机，可重复
  - 可随参数变化（非简单零点效应）



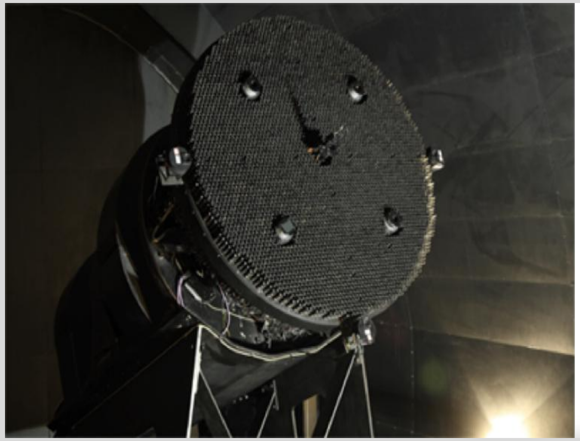
1.2，参数质量评价指标

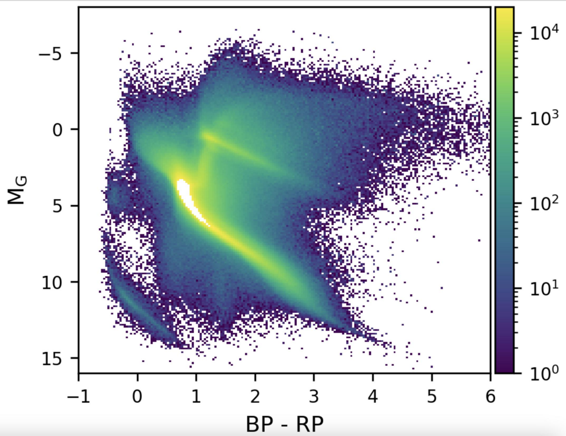
- ◆ 精度(precision): 测量随机误差
- ◆ 准确度(accuracy): 方法系统误差
  - 模型/方法误差，定标误差
  - 非随机，可重复
  - 可随参数变化（非简单零点效应)

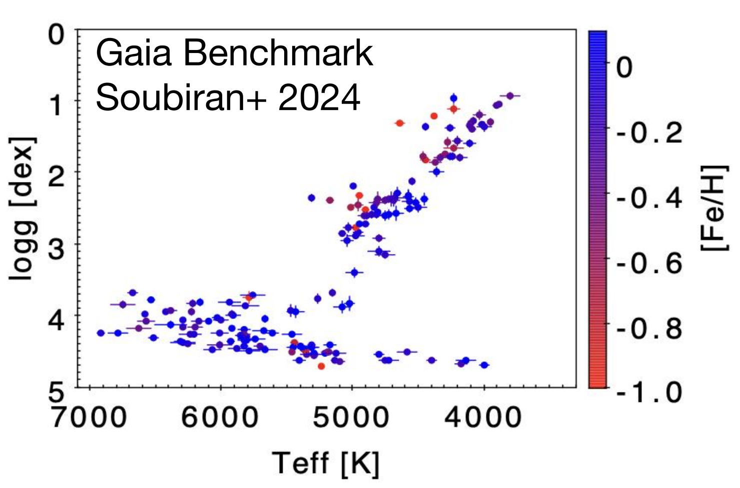
➤系统误差无法避免

➤ 精准定标的困难：标准星的欠缺

- 样本小，维度低，不均匀等



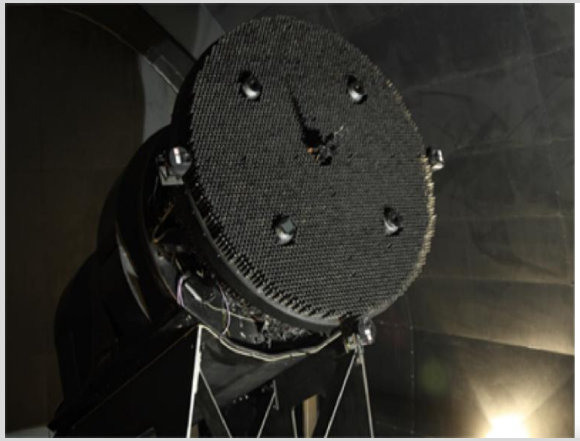
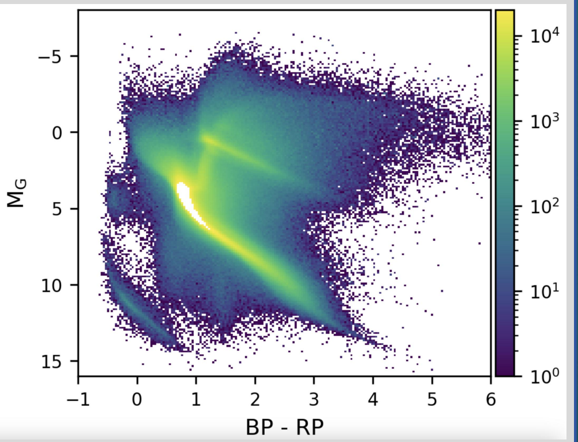


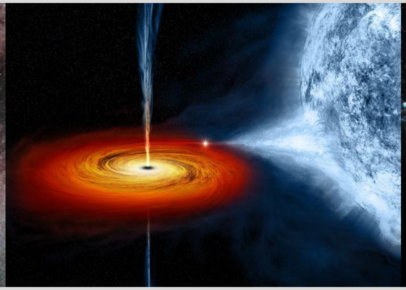


1.2， 参数质量评价指标

- 精度(precision): 测量随机误差
- 准确度(accuracy): 方法系统误差
    - 模型/方法误差， 定标误差
    - 非随机， 可重复
    - 可随参数变化 （非简单零点效应）

➢对巡天大样本分析， 精度/内部一致性通常优先于绝对定标


较差温度/颜色
-星际尘埃

较差速度
-双星/黑洞

较差年龄
-银河演化

内容提纲

（一）什么是恒星参数测定?

（二）为什么恒星参数测定需要机器学习?

（三）机器学习在恒星参数测定中的应用示例

（四）机器学习光谱分析面临的问题

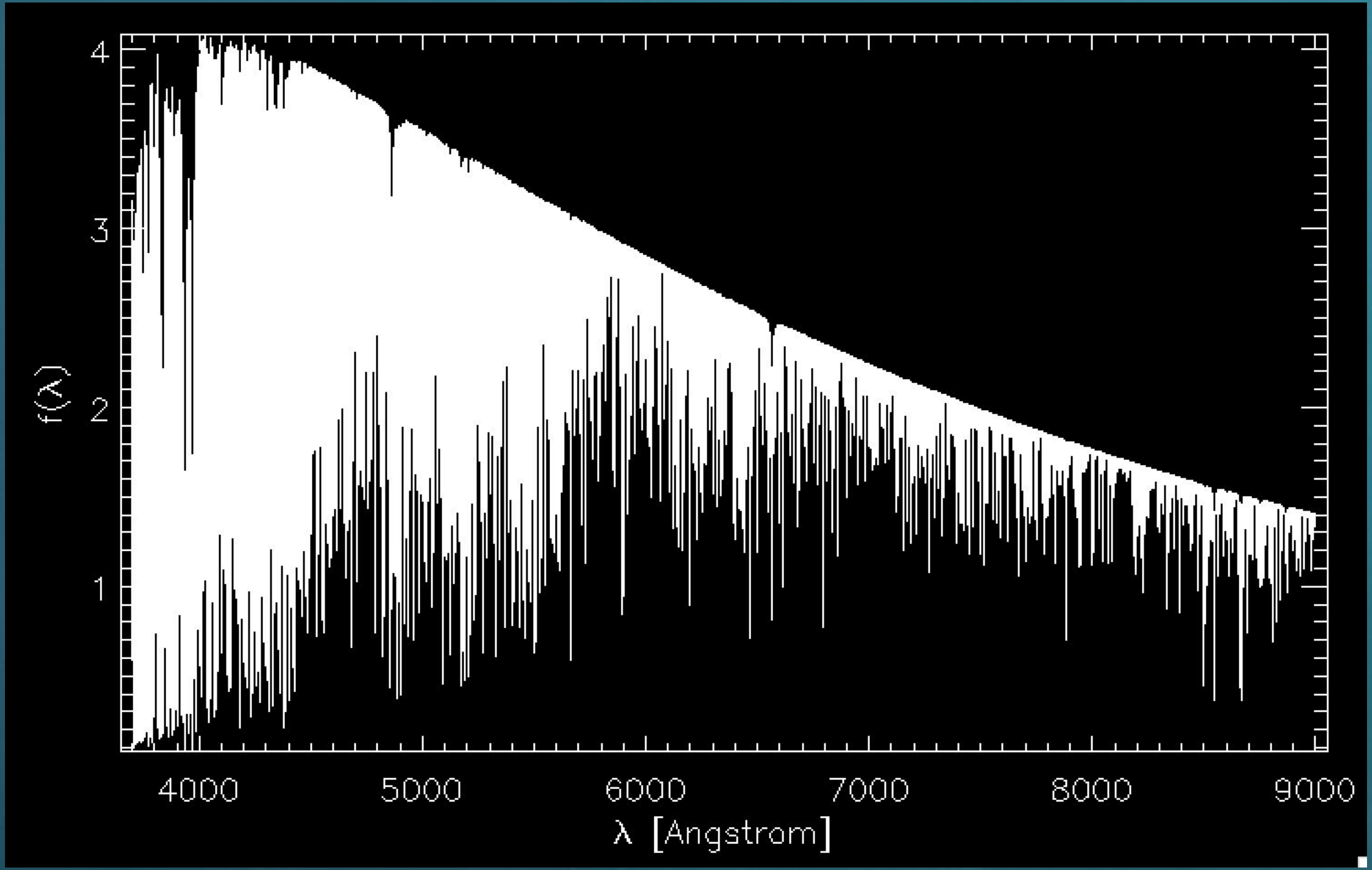
（五）总结与展望

2.1， 光谱的物理信息含量

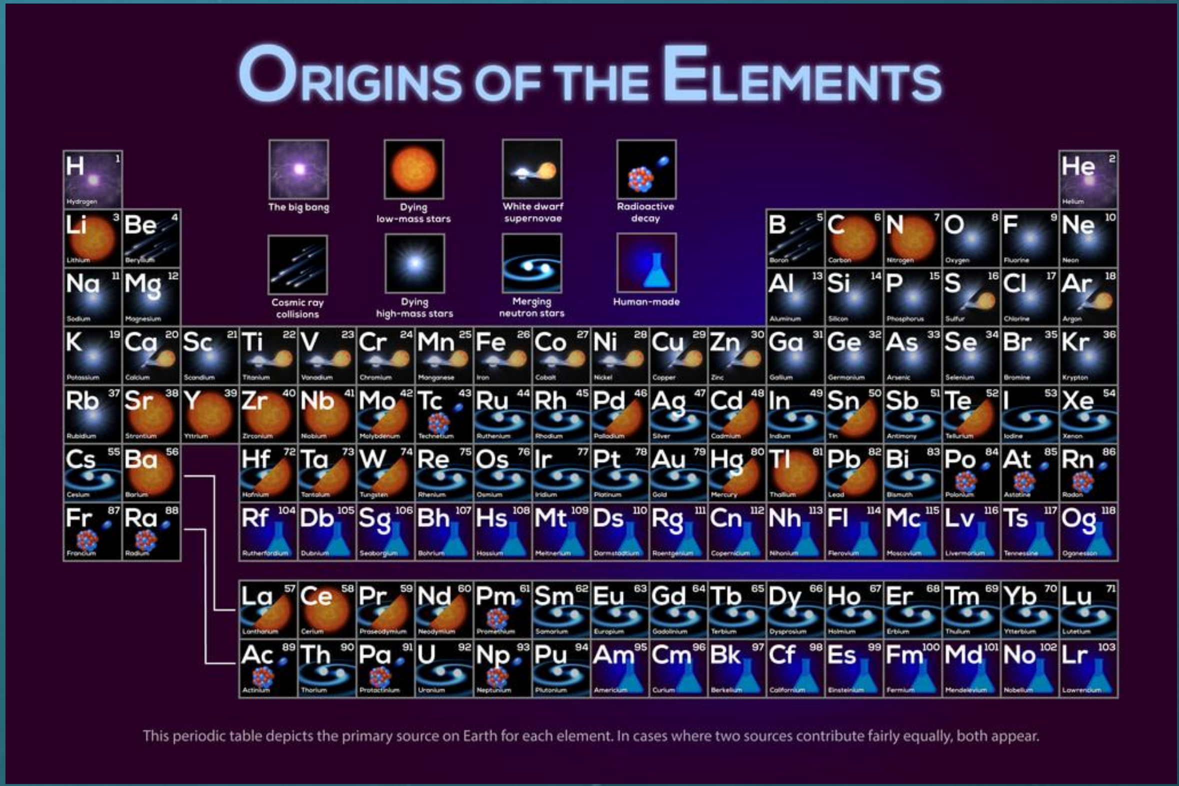
◆ 给定仪器和观测时间，光谱的信息含量取决于

- 波长范围
- 分辨率
- 信噪比

2.1，光谱的物理信息含量

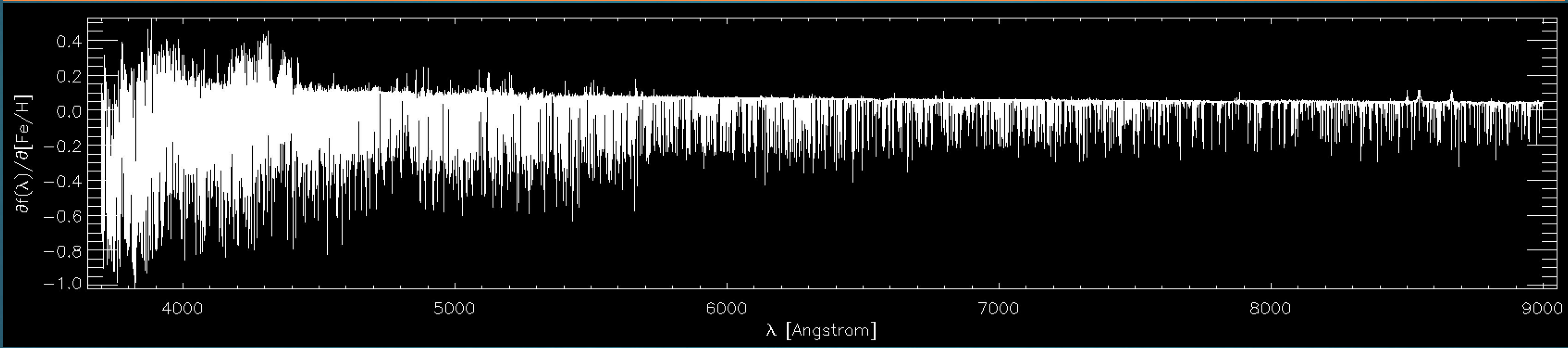


数百万条谱线

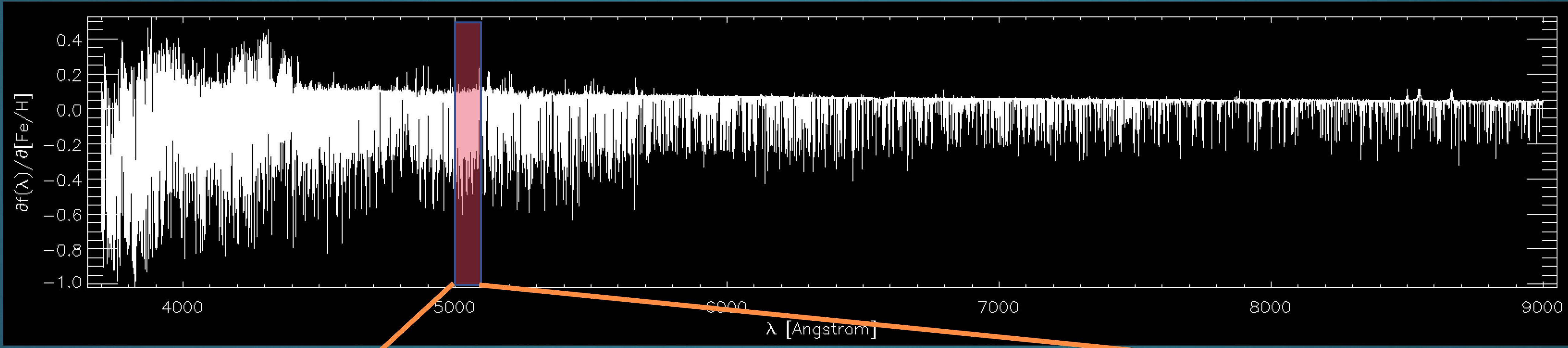


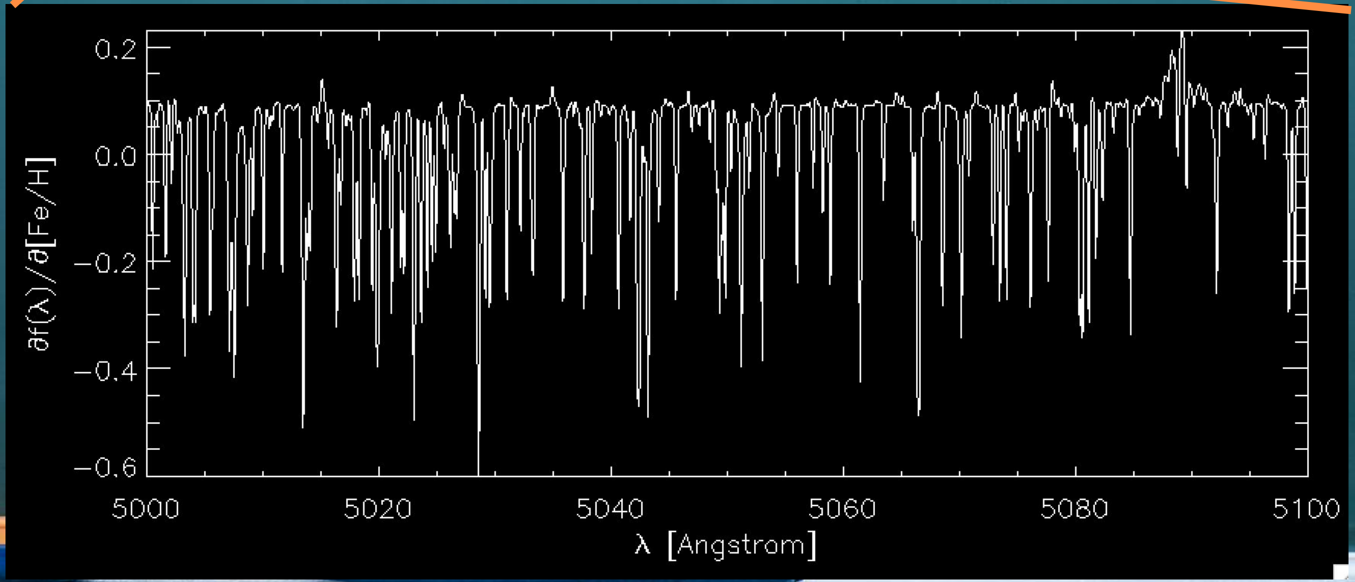
~100种元素

2.1，光谱的物理信息含量

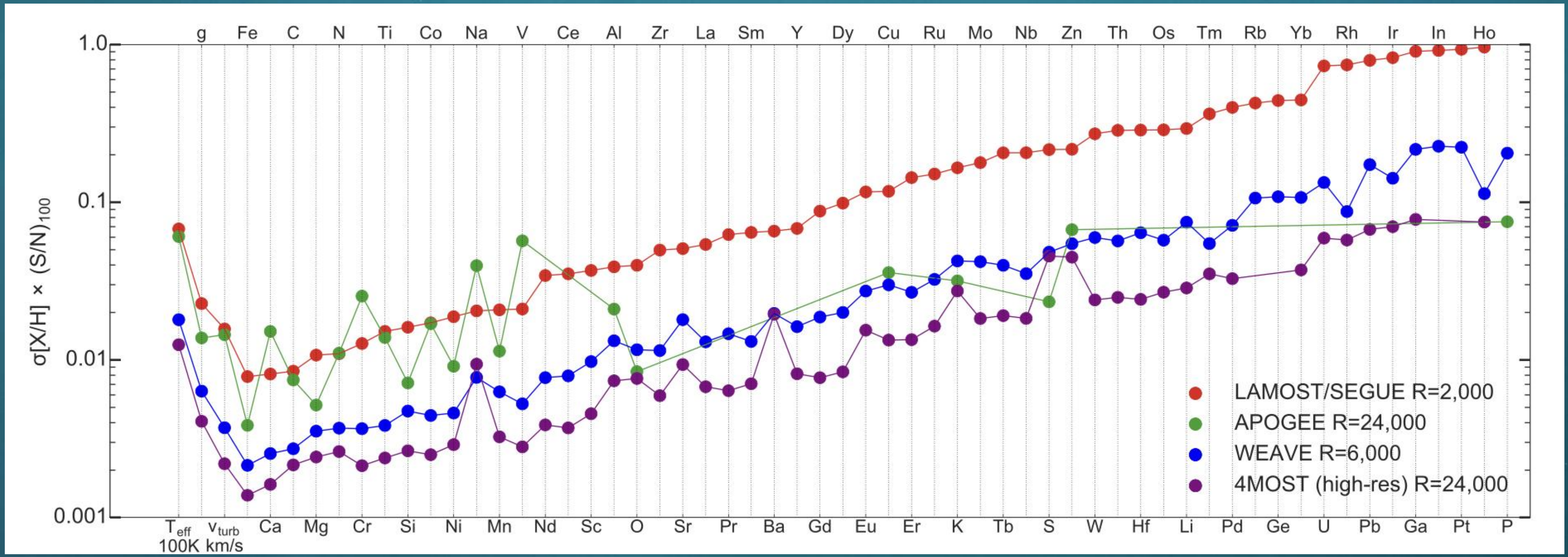


2.1，光谱的物理信息含量





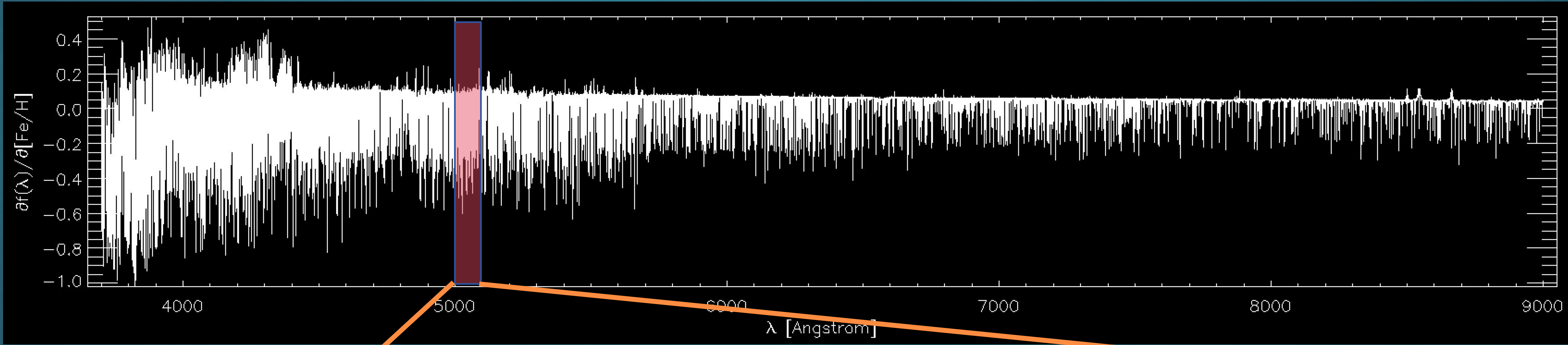
2.1，光谱的物理信息含量

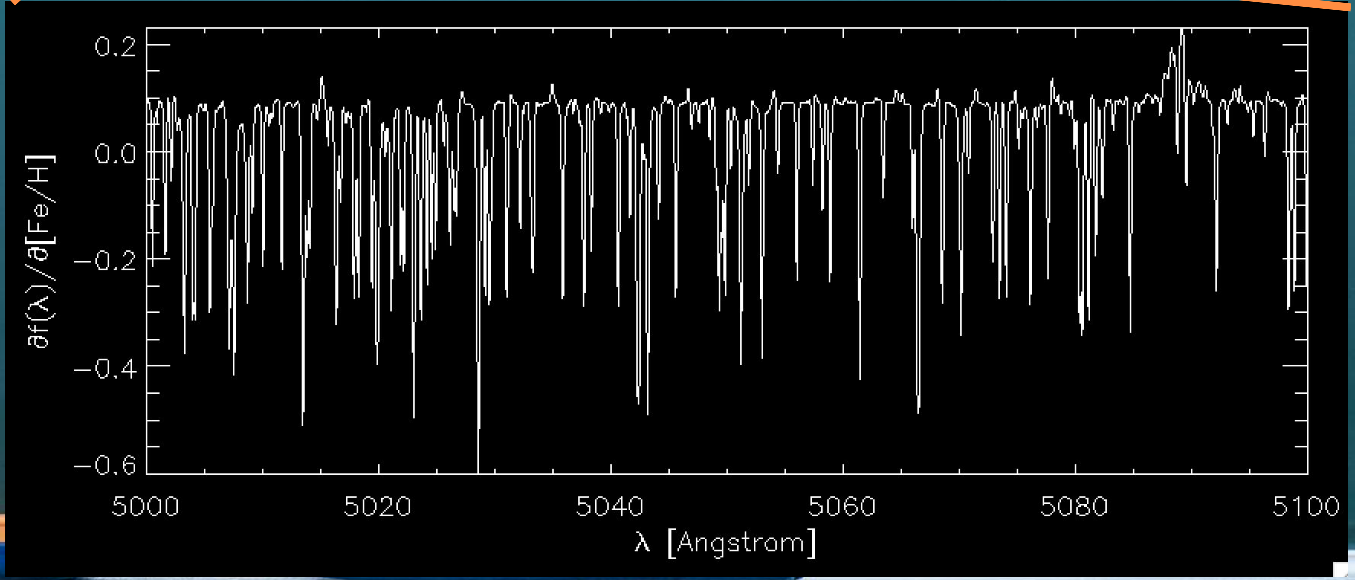


Ting et al. 2017, ApJ, 843

理想情形：仅考虑光子噪声误差

2.1，光谱的物理信息含量





仅少量谱线具有精确原子物理参数(loggf)定标
--传统方法仅用部分信息

2.2，机器学习的优势-大数据

- 大数据分析（光谱巡天）
  - >10^8 spectra
  - GALAH, SDSS-V, WEAVE, 4MOST; LAMOST, DESI，PFS, Gaia

GALAH
R=28000

WEAVE
R=20000/5000

4MOST
R=20000/6500

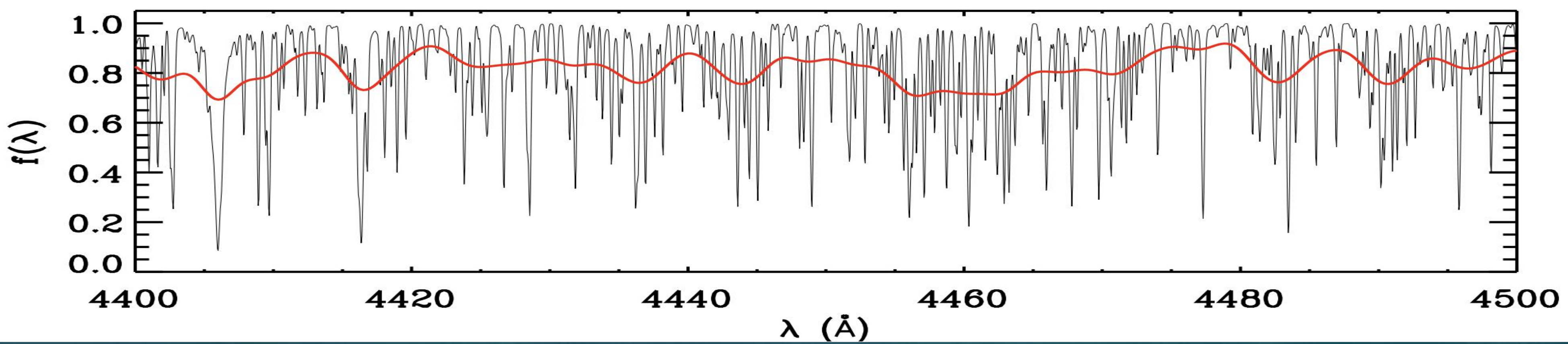
SDSS/APOGEE
R=22,500

LAMOST
R=7500/1800

2.2，机器学习的优势-多参数高维度建模

◆ 恒星光谱参数测定是多参数求解问题

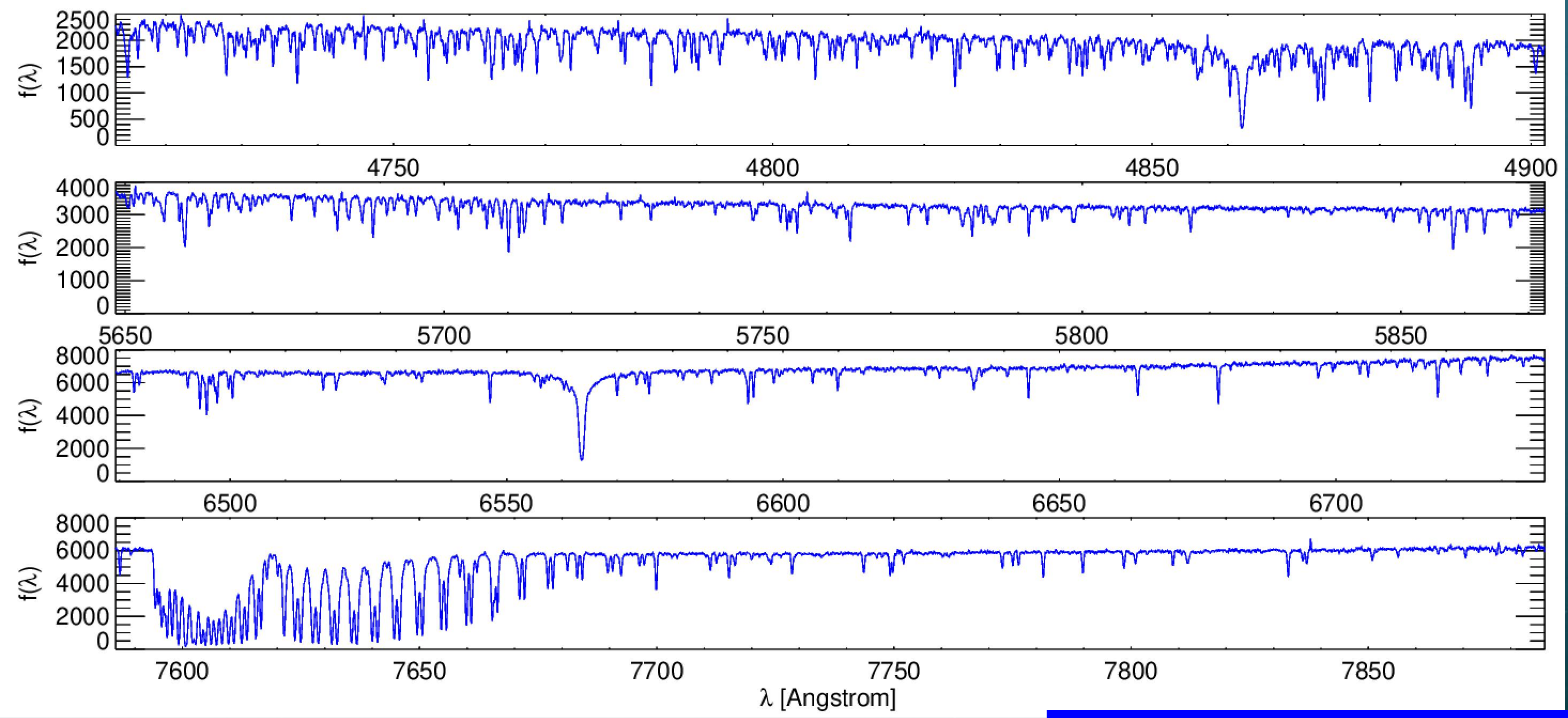
- 谱线混叠（内禀+仪器展宽）
- 高维度光谱建模：Teff, logg, v_{mic}, [X/H]，X ~ 100种元素



2.2，机器学习的优势-利用全光谱信息

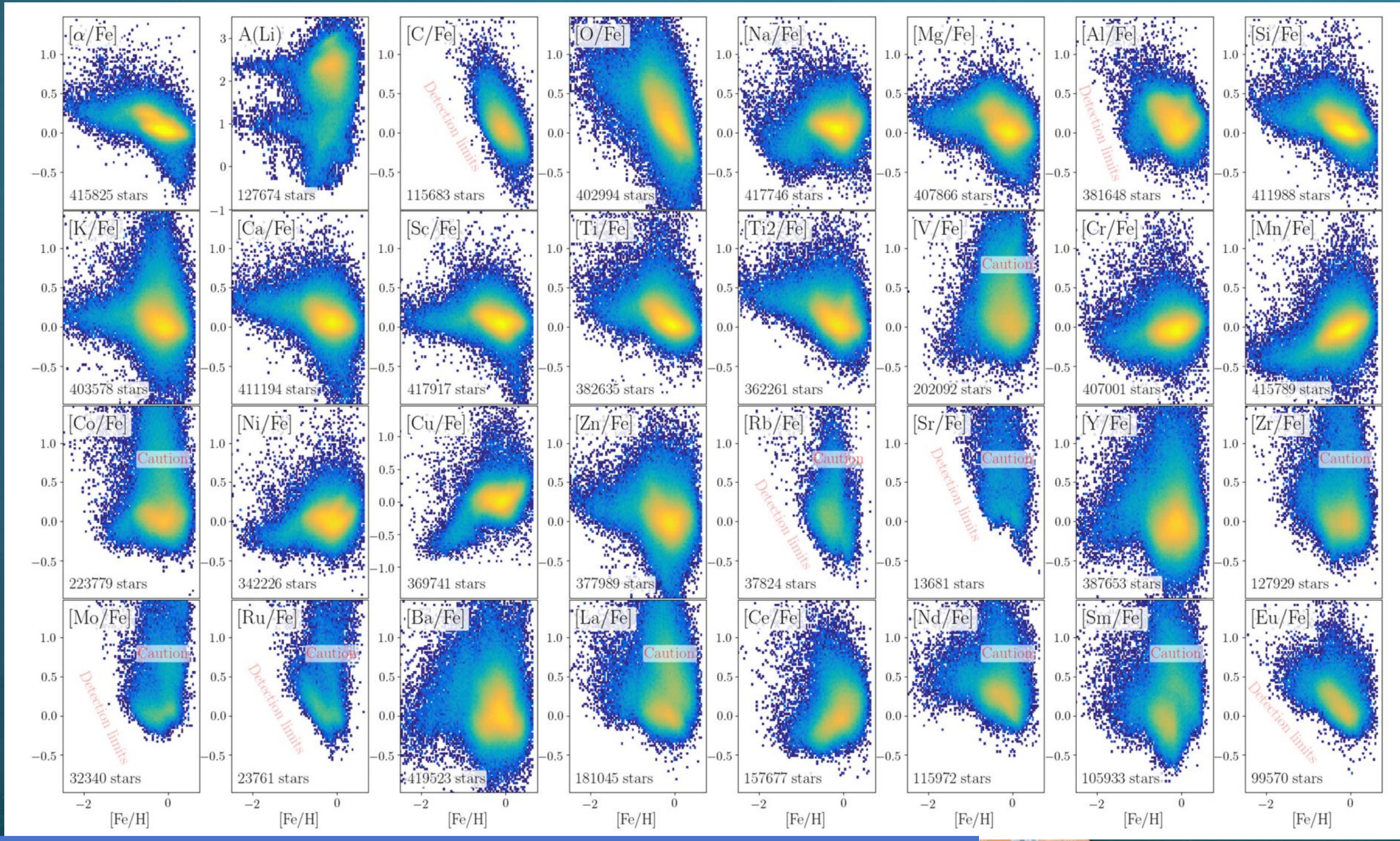
◆ 利用全光谱信息

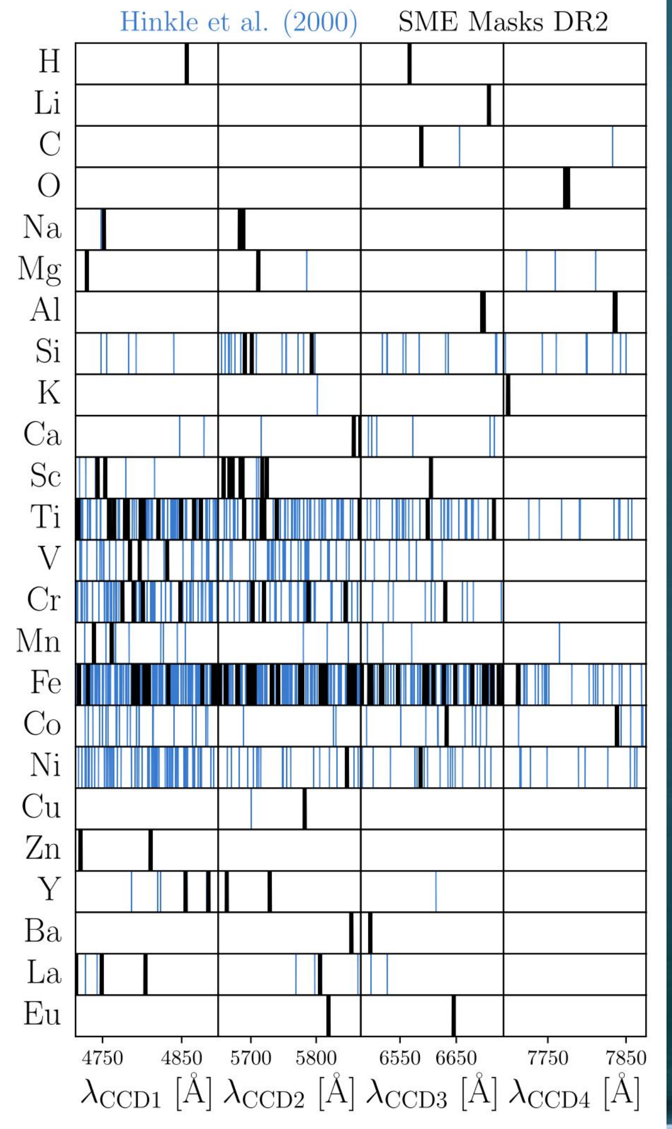
2.2，机器学习的优势-利用全光谱信息



GALAH光谱(R~28,000)

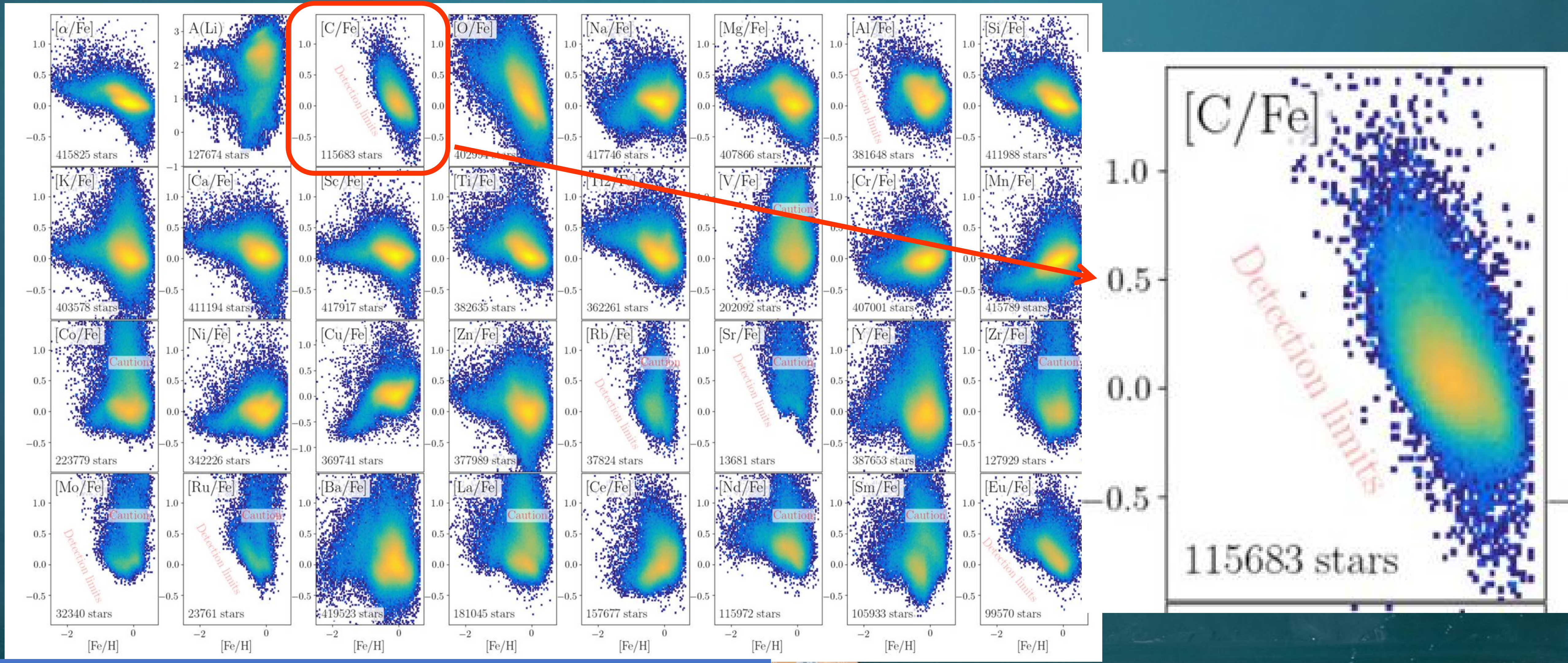
2.2，机器学习的优势-利用全光谱信息





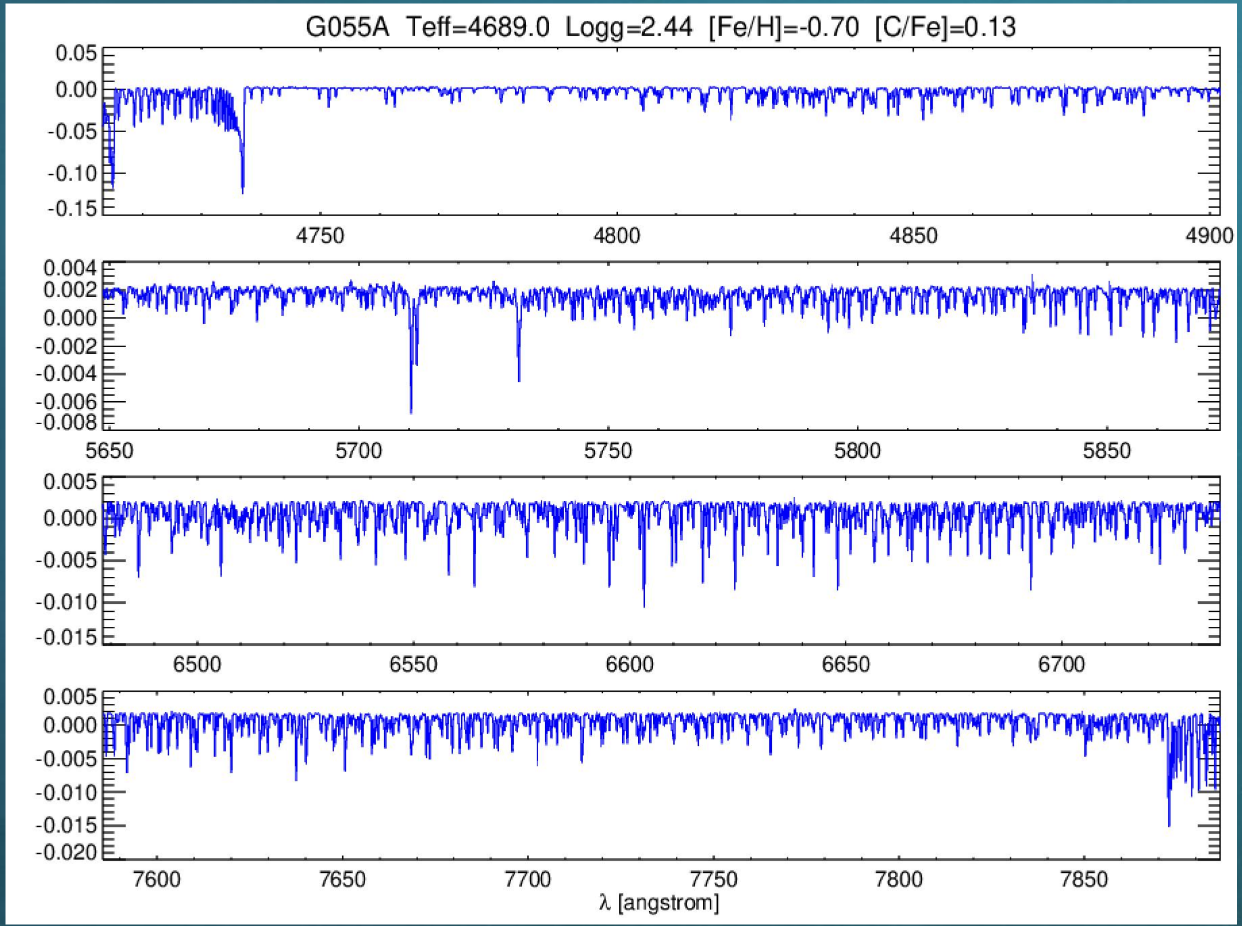
GALAH DR3；Buder+ 2021, MNRAS

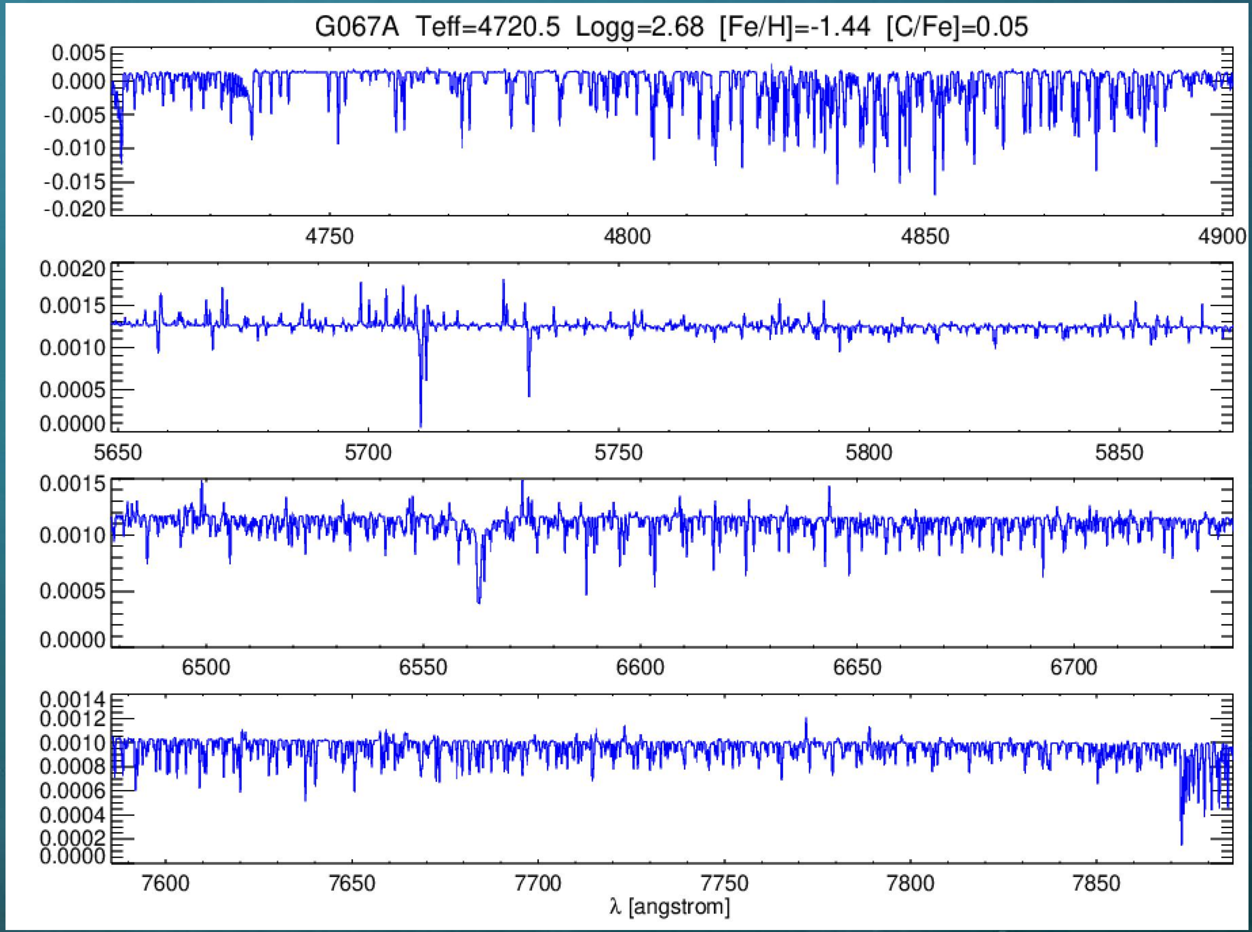
2.2，机器学习的优势-利用全光谱信息



GALAH DR3；Buder+ 2021, MNRAS

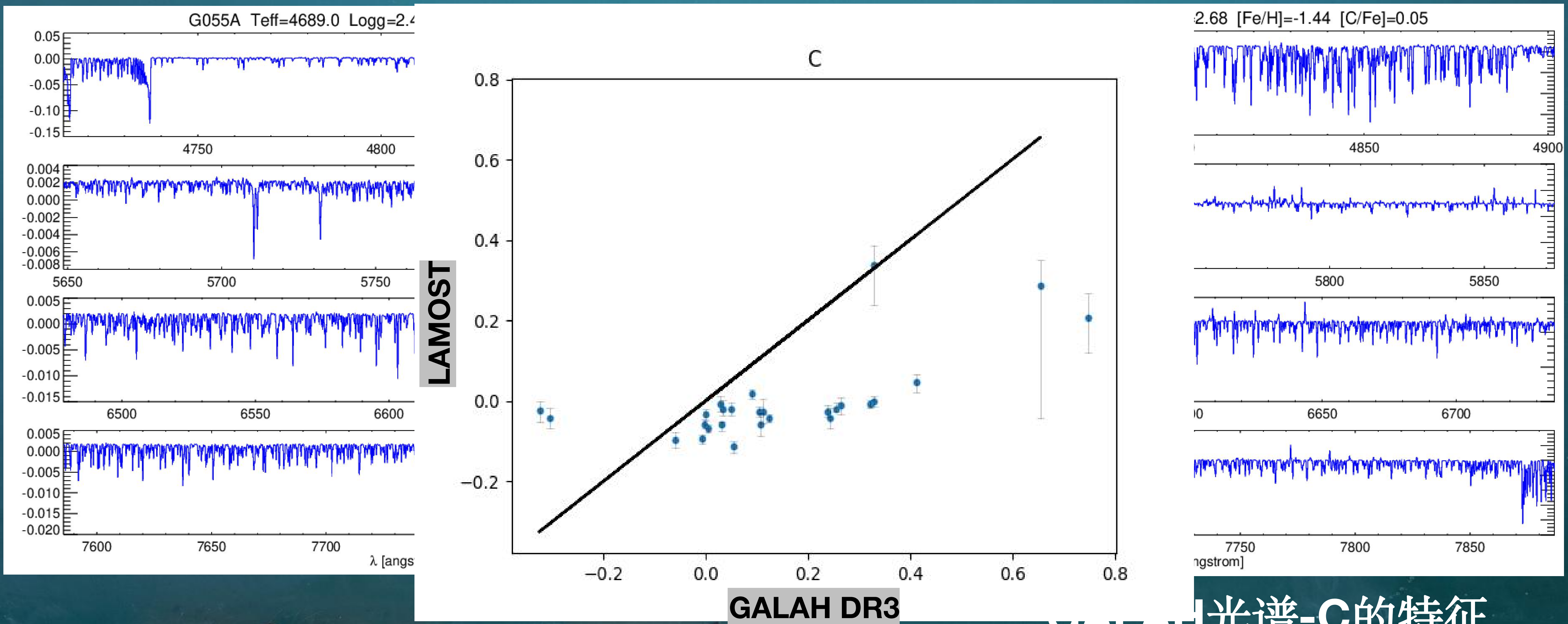
2.2，机器学习的优势-利用全光谱信息





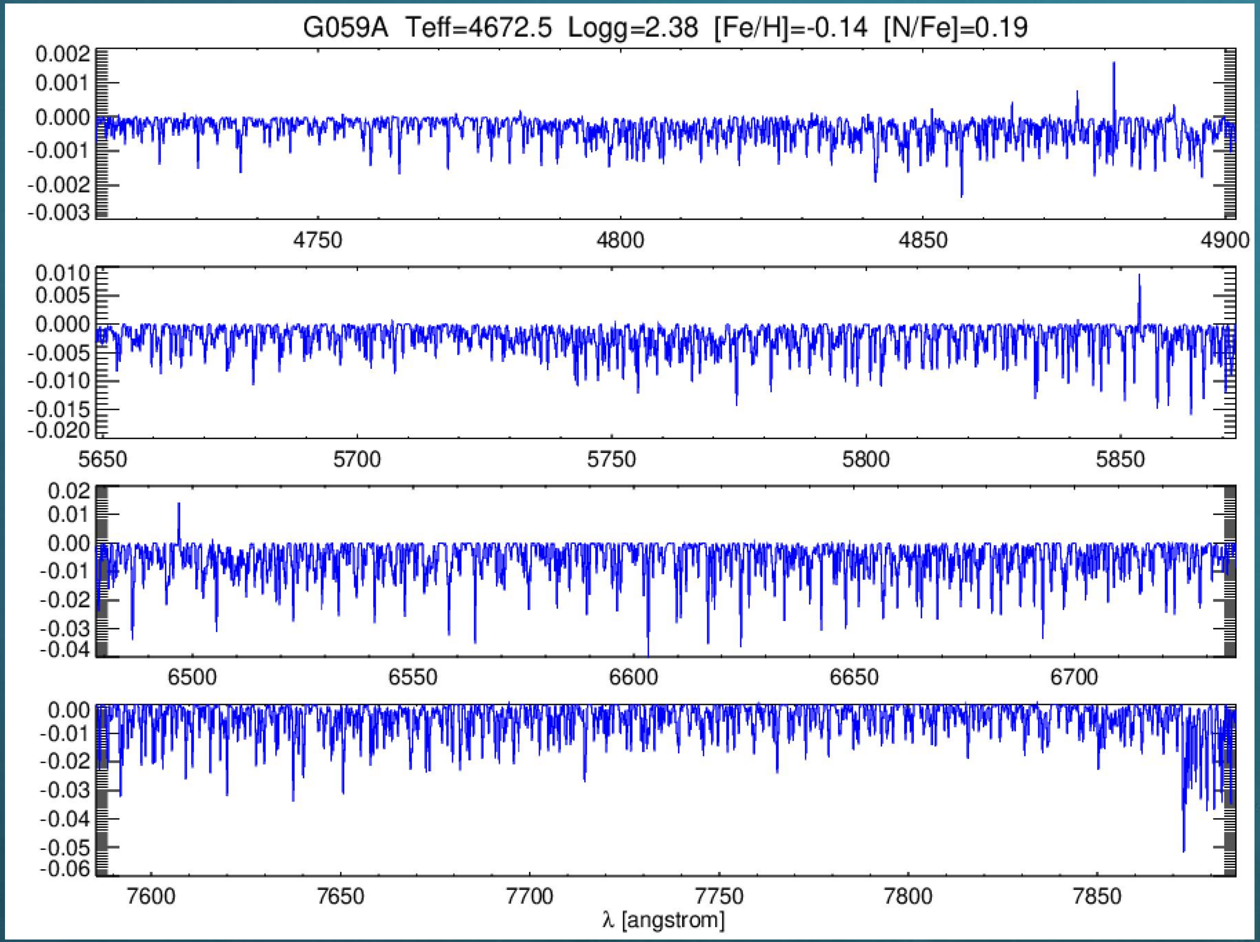
GALAH光谱-C的特征

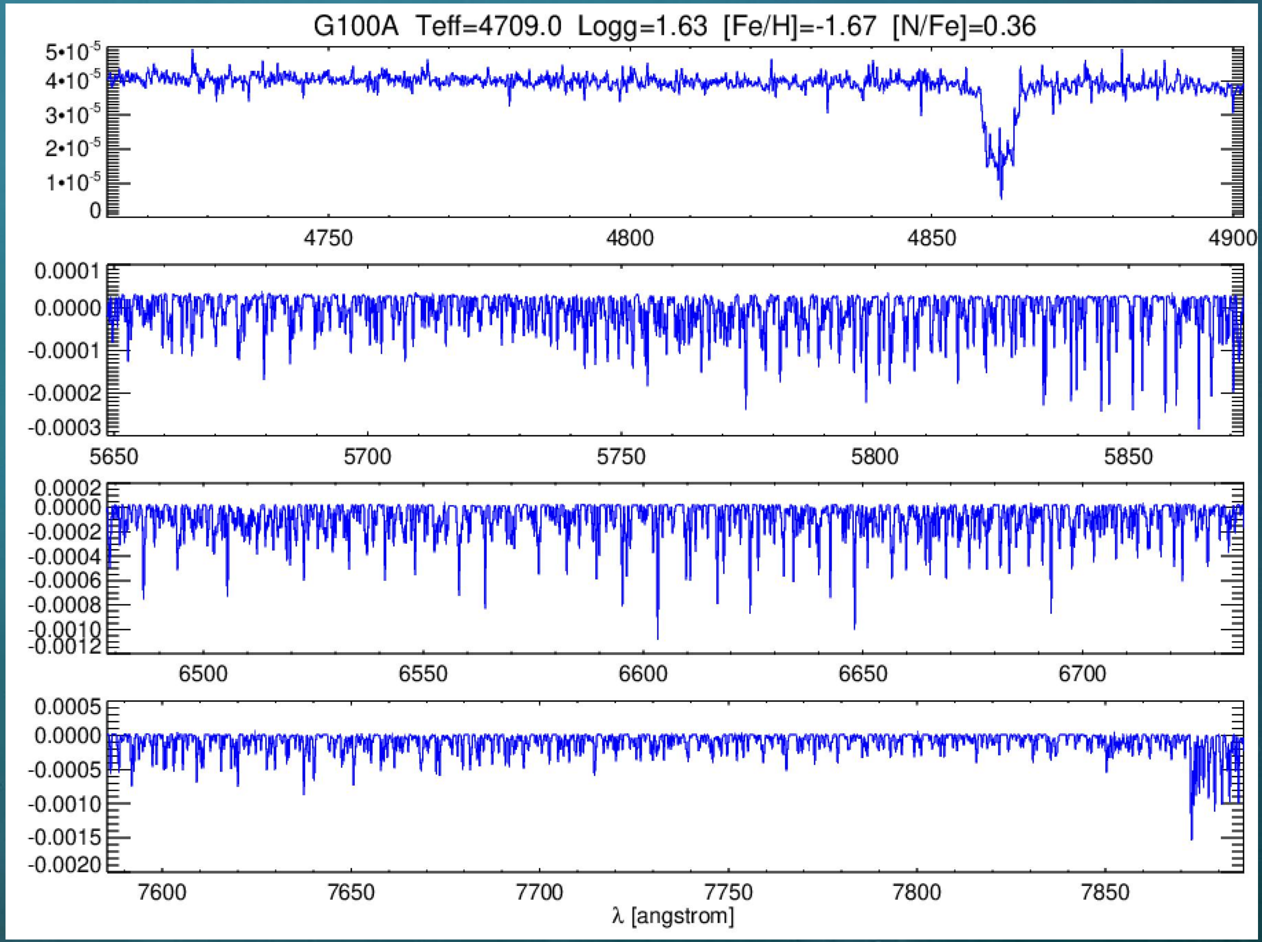
2.2，机器学习的优势-利用全光谱信息



GALAH光谱-C的特征

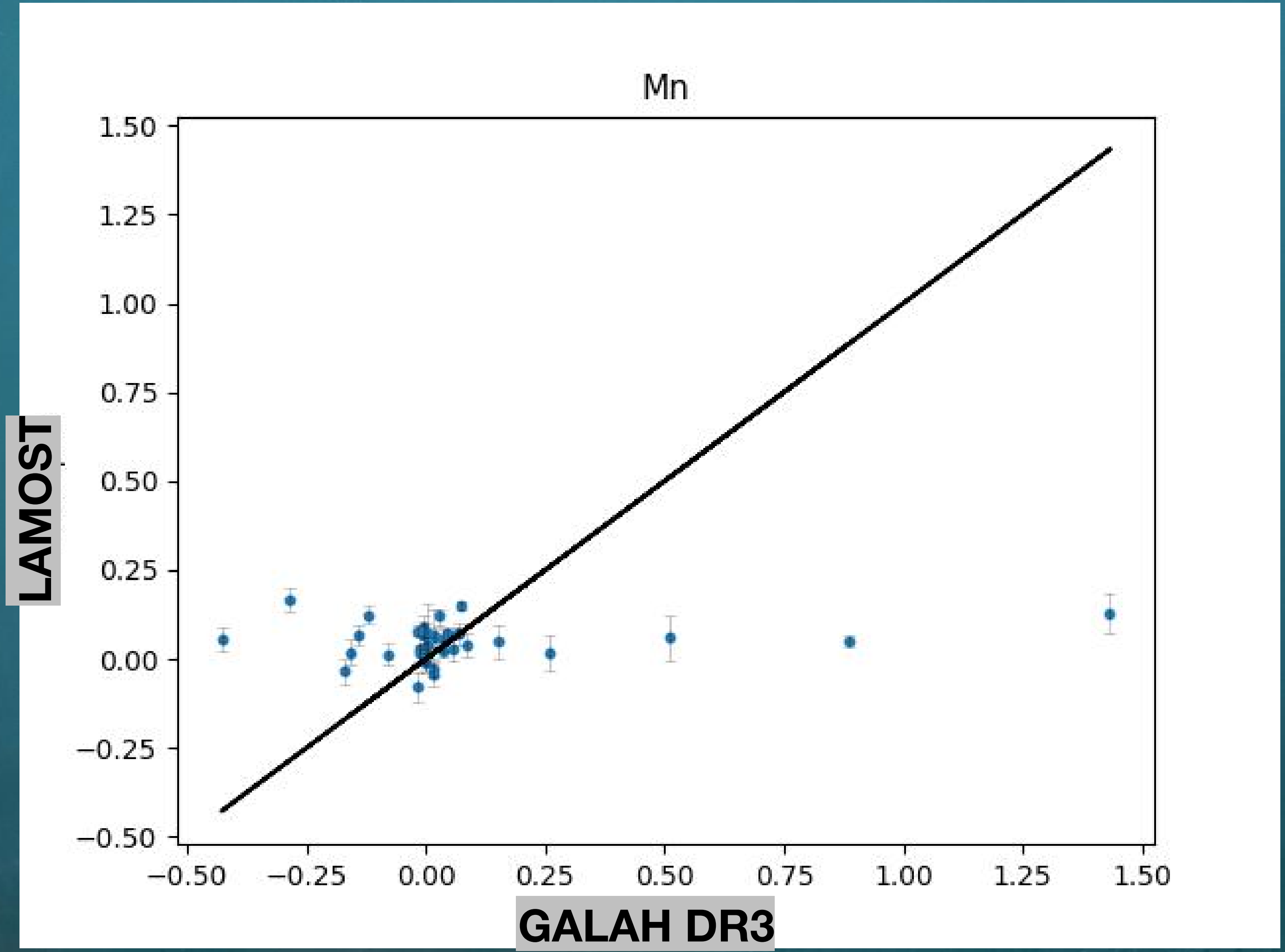
2.2，机器学习的优势-利用全光谱信息



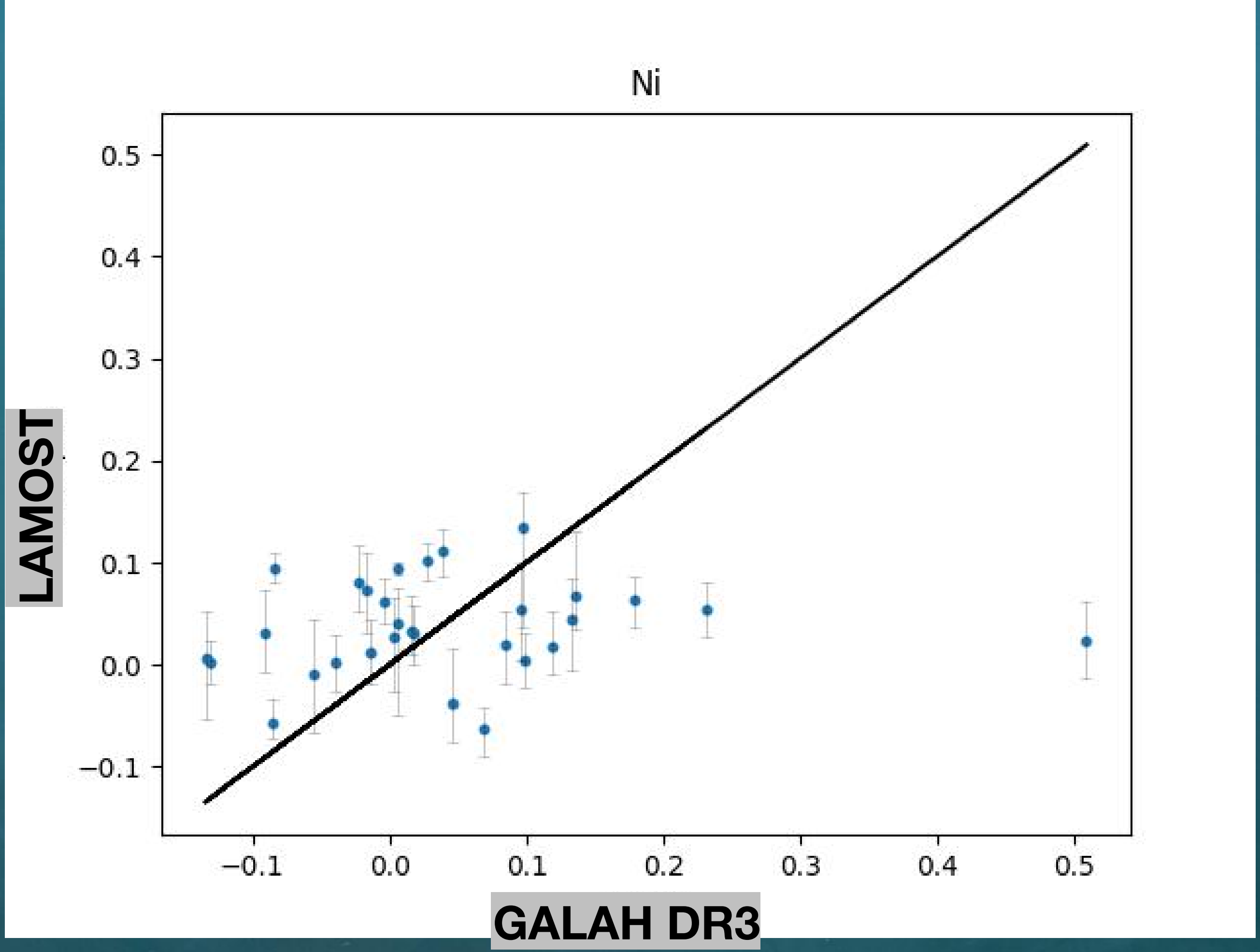


GALAH光谱-N的特征

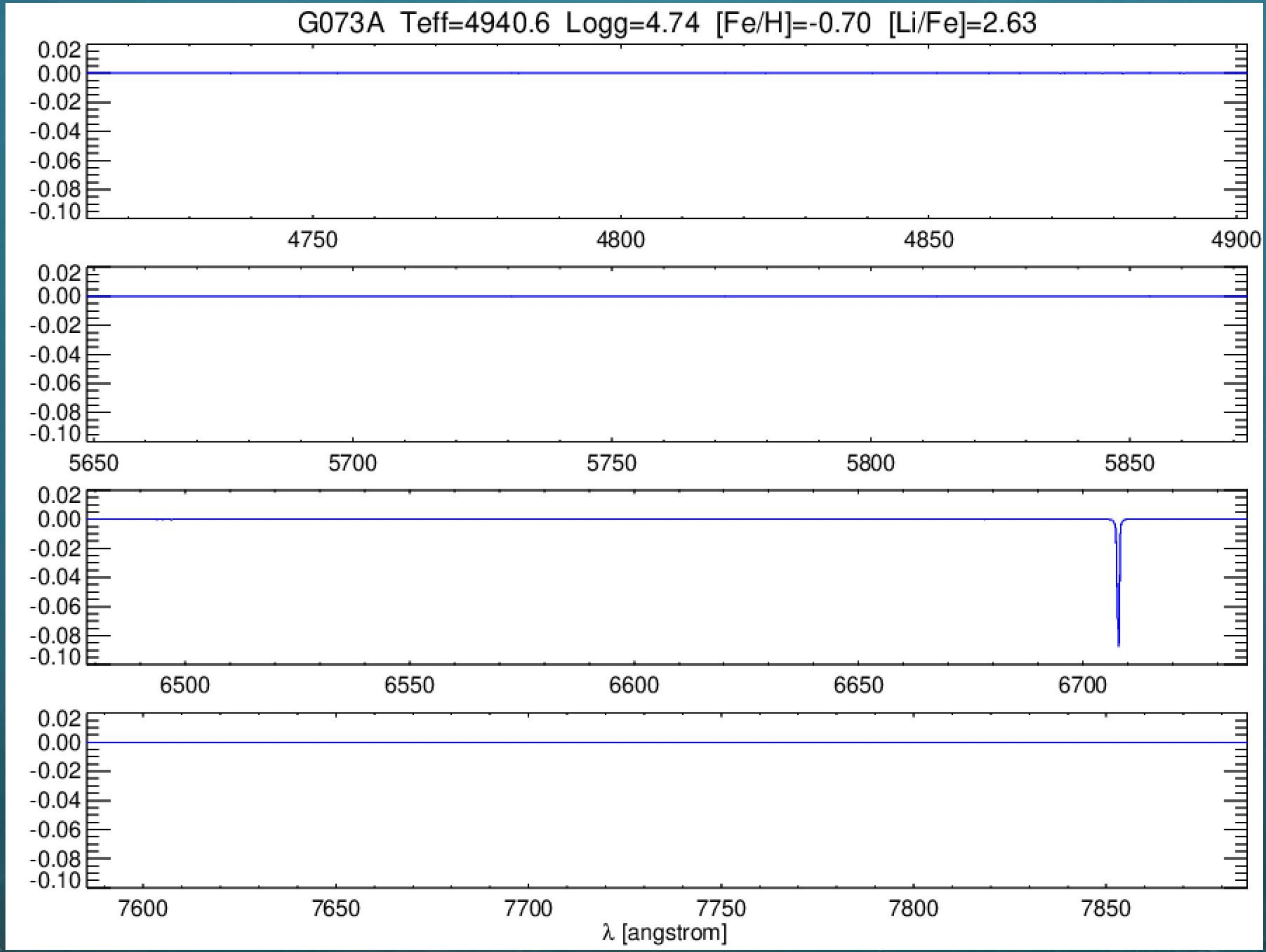
2.2，机器学习的优势-利用全光谱信息



2.2，机器学习的优势-利用全光谱信息



2.2，机器学习的优势-利用全光谱信息



内容提纲

（一）什么是恒星参数测定？

（二）为什么恒星参数测定需要机器学习？

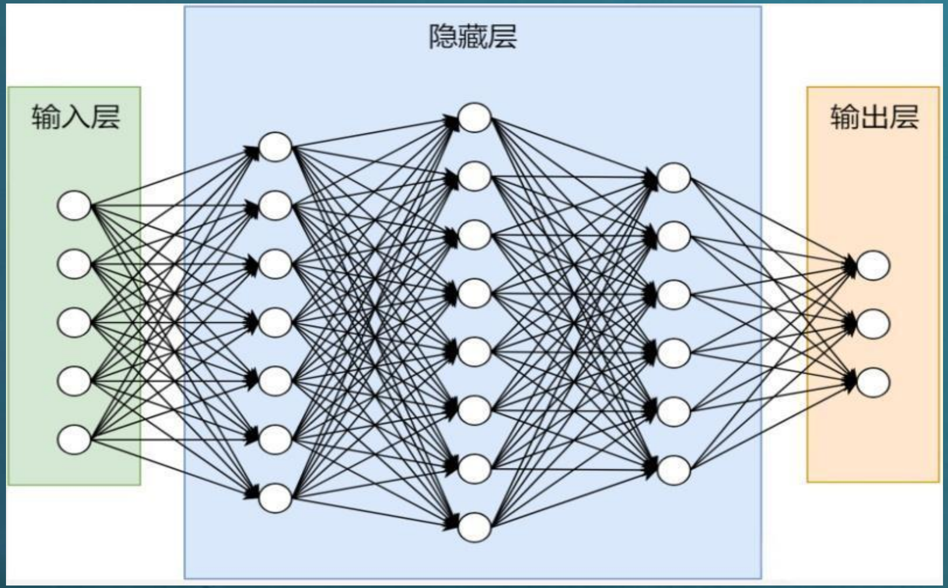
（三）机器学习在恒星参数测定中的应用

（四）机器学习光谱分析面临的问题

（五）总结与展望

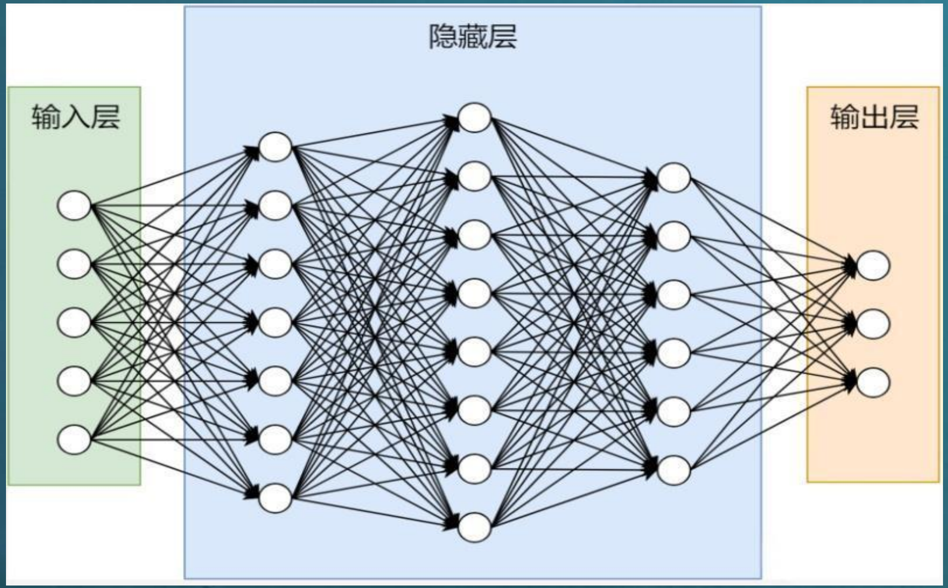
3.1，机器学习方法概述

光谱 ——→ 参数



3.1，机器学习方法概述

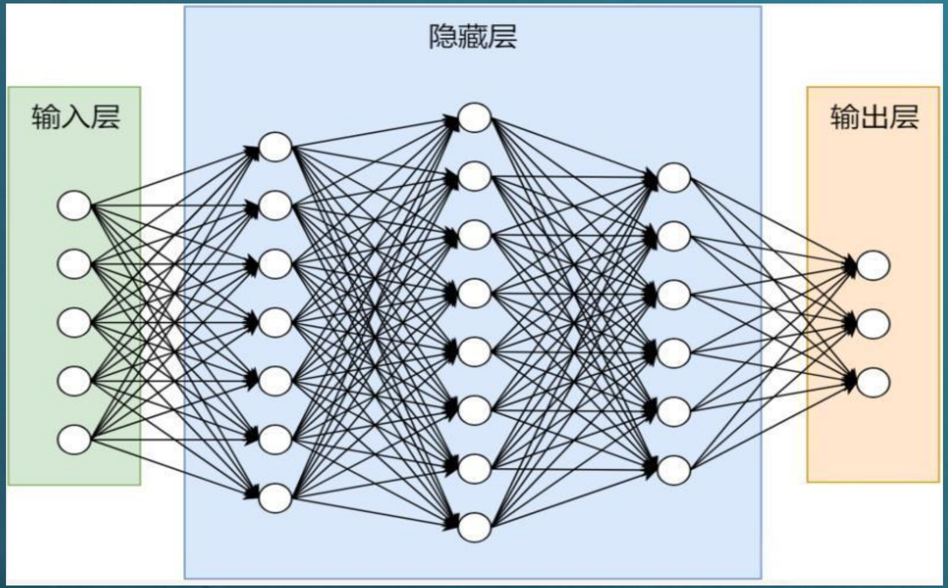
光谱 ⟶ 参数



- ◆ 主要计算在于训练过程，预测较简单

3.1，机器学习方法概述

光谱 ⟶ 参数



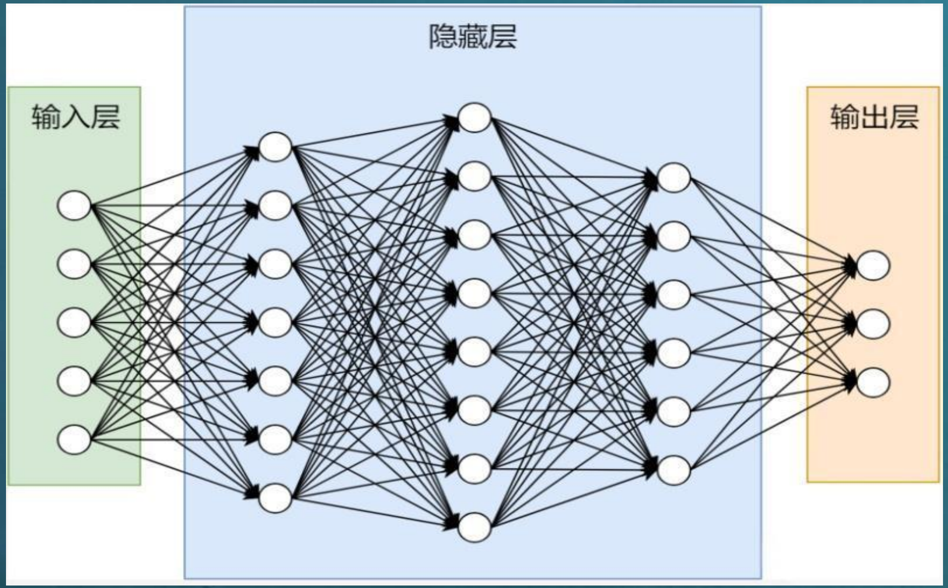
◆ 主要计算在于训练过程，预测较简单

◆ 黑箱模式

- 过程不够透明（可能基于统计相关性，物理可解释性较差)

3.1，机器学习方法概述

光谱 ——→ 参数



◆ 主要计算在于训练过程，预测较简单

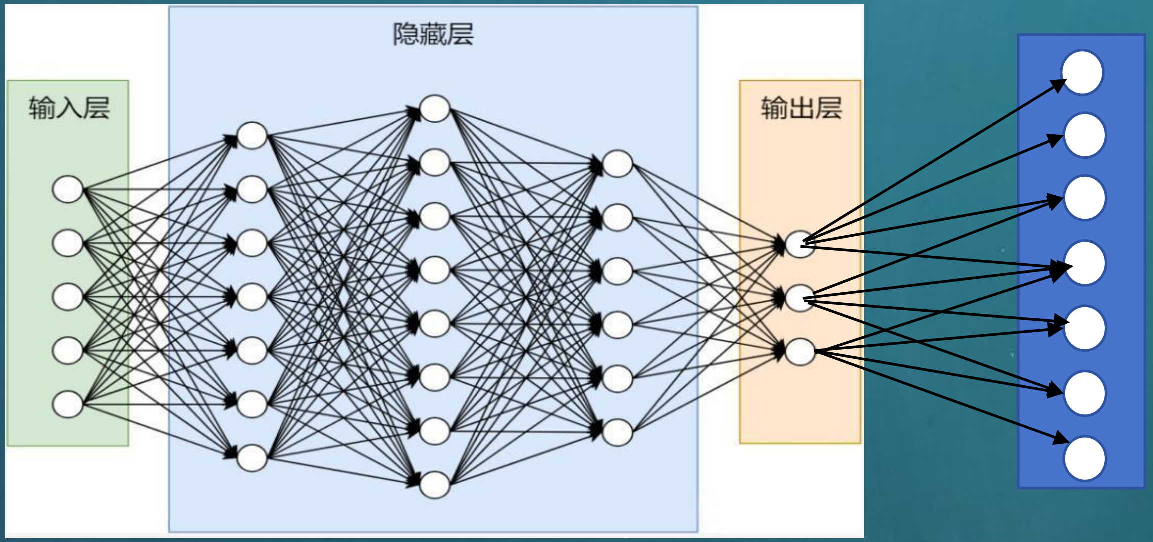
◆ 黑箱模式

- 过程不够透明（可能基于统计相关性，物理可解释性较差）

思考：利用天体物理相关性由A谱线推断B丰度？

3.1，机器学习方法概述

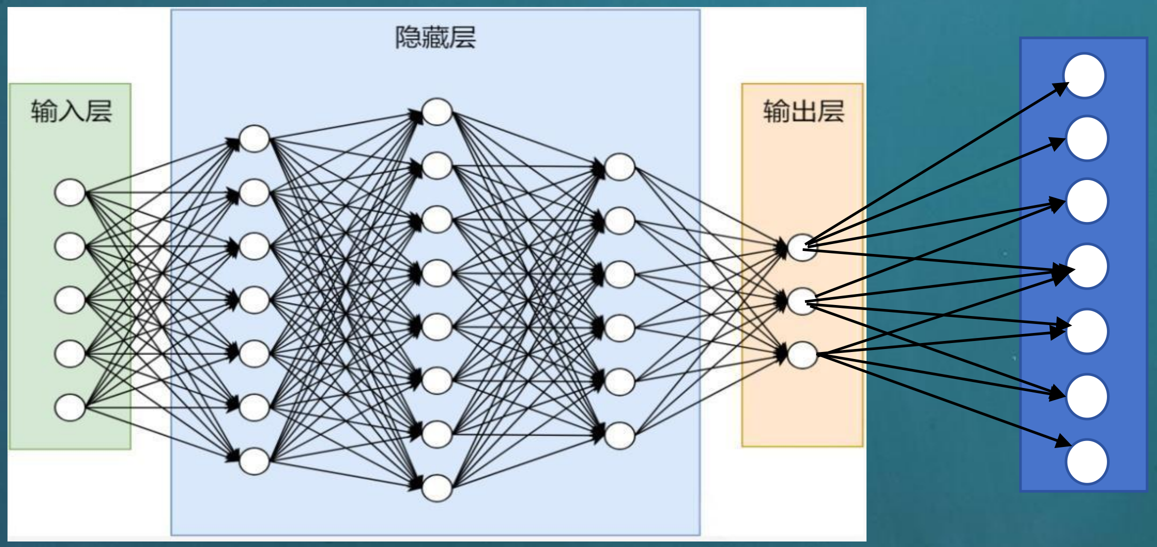
参数 → 光谱 → 参数



◆ 机器学习正向建模 + 模型匹配/参数拟合

3.1，机器学习方法概述

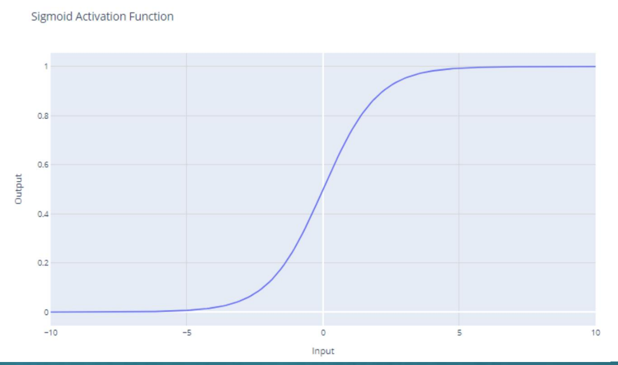
参数 ⟶ 光谱 ⟶ 参数

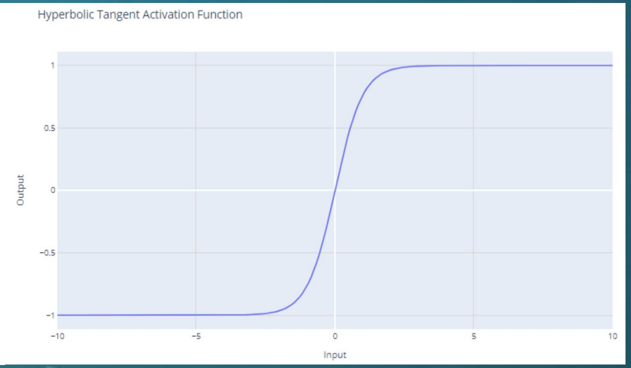


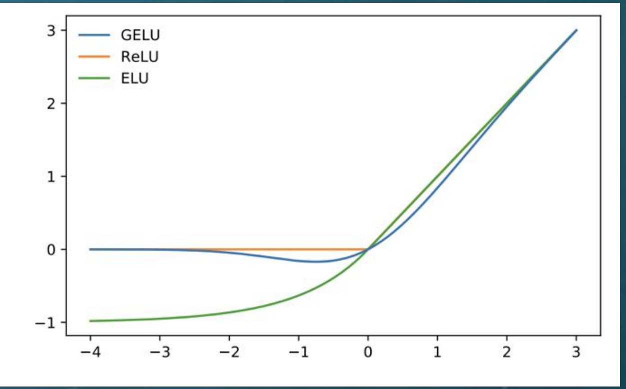
- 机器学习正向建模 + 模型匹配/参数拟合
- 白箱模式
  - 过程更透明，物理可解释
  - 获取参数完整信息（协方差）
- 需要多参数训练样本，参数求解较耗时

3.2，机器学习方法示例

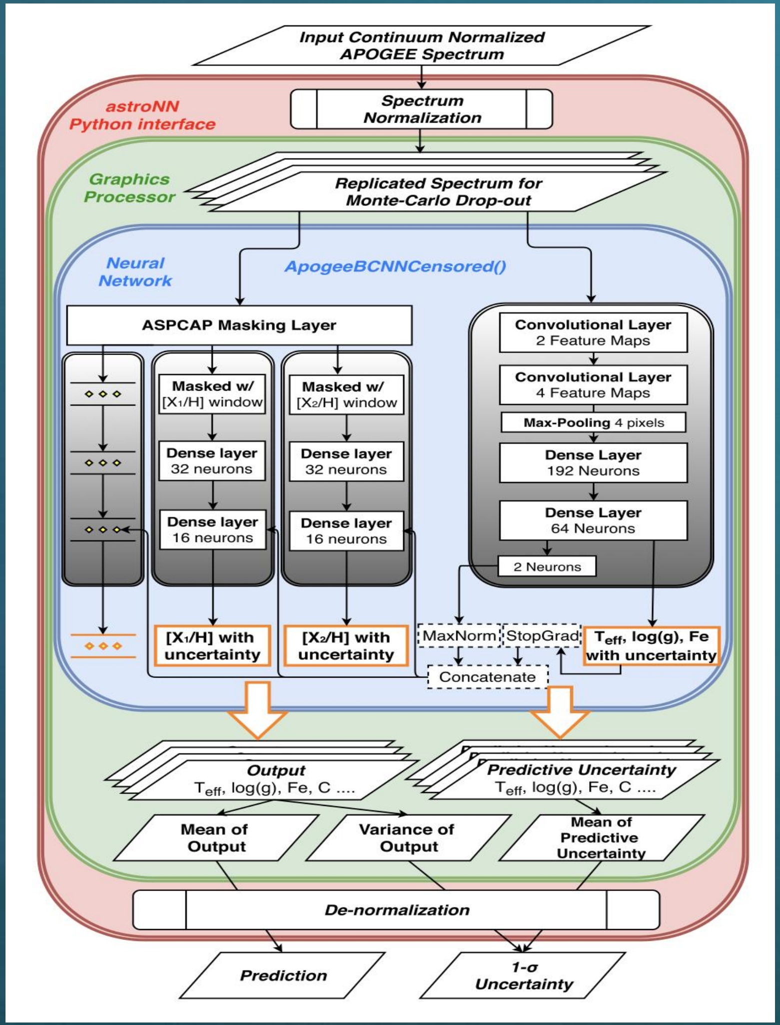
```
# define model
model = torch.nn.Sequential(
        torch.nn.Linear(num_label, 100),
        torch.nn.LeakyReLU(),
        torch.nn.Linear(100, 100),
        torch.nn.LeakyReLU(),
        torch.nn.Linear(100,num_pix) )

# define optimizer
learning_rate = 0.001
optimizer = torch.optim.Adam(model.parameters(), lr=learning_rate)

# train the model
#divide training set into batches
... ...
optimizer.zero_grad()
loss.backward(retain_graph=False)
optimizer.step()
```

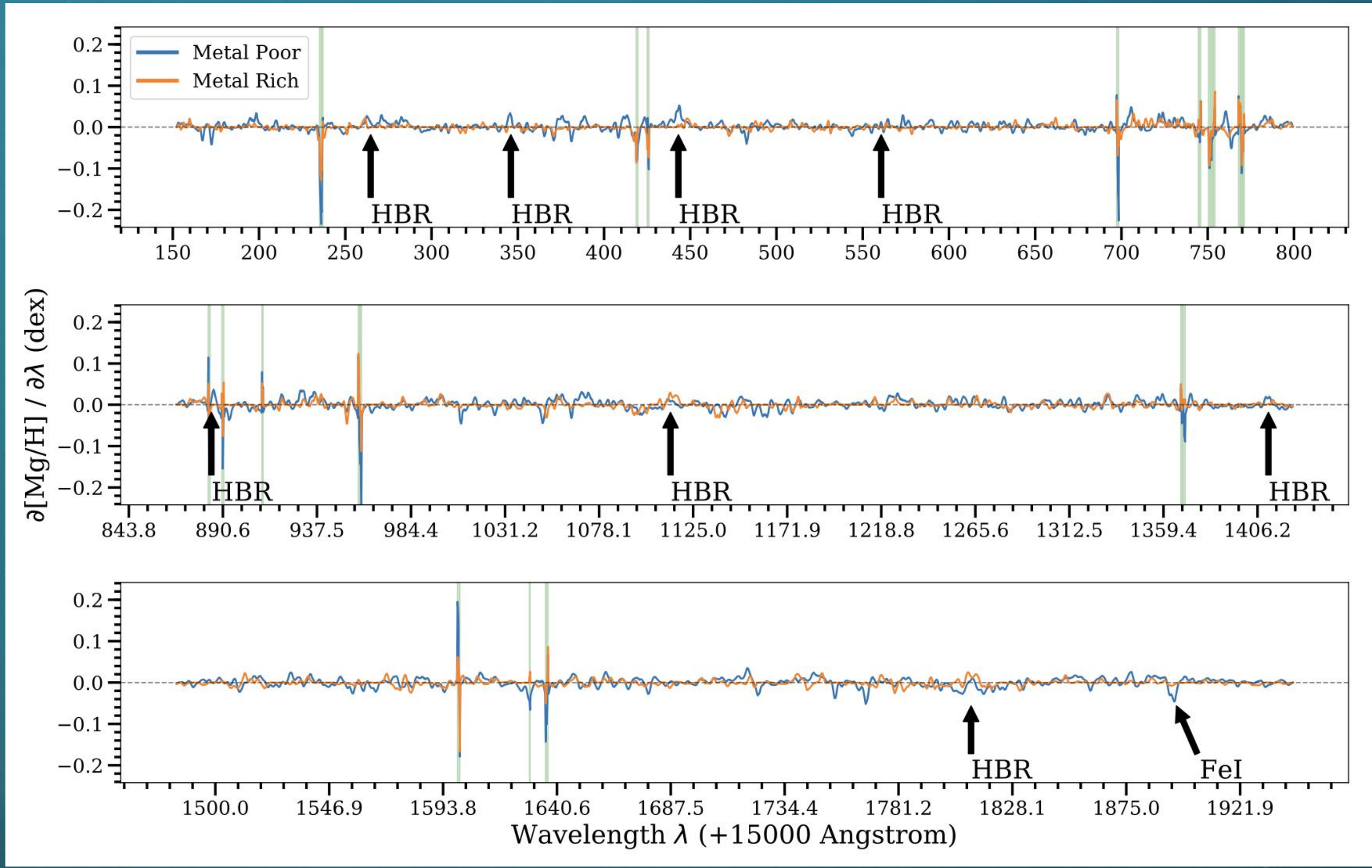






3.2，机器学习示例- AstroNN





Leung & Bovy, 2019, MNRAS, 483, 3255

https://github.com/henrysky/astroNN

3.2，机器学习方法示例-Cannon

The Astrophysical Journal, 808:16 (21pp), 2015 July 20
doi:10.1088/0004-637X/808/1/16

THE CANNON: A DATA-DRIVEN APPROACH TO STELLAR LABEL DETERMINATION

M. Ness[1], David W. Hogg[1,2,3], H.-W. Rix[1], Anna. Y. Q. Ho[1], and G. Zasowski[4,5]
[1] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany; ness@mpia.de
[2] Center for Cosmology and Particle Physics, Department of Phyics, New York University, 4 Washington Pl., Room 424, New York, NY 10003, USA
[3] Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, NY 10003, USA
[4] Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
Received 2015 January 16; accepted 2015 June 3; published 2015 July 14

ABSTRACT

New spectroscopic surveys offer the promise of stellar parameters and abundances ("stellar labels") for hundreds of thousands of stars; this poses a formidable spectral modeling challenge. In many cases, there is a subset of *reference objects* for which the stellar labels are known with high(er) fidelity. We take advantage of this with *The Cannon*, a new data-driven approach for determining stellar labels from spectroscopic data. *The Cannon* learns from the "known" labels of reference stars how the continuum-normalized spectra depend on these labels by fitting a flexible model at each wavelength; then, *The Cannon* uses this model to derive labels for the remaining survey stars. We illustrate *The Cannon* by training the model on only 542 stars in 19 clusters as reference objects, with $T_{\rm eff}$, $\log\ g$, and [Fe/H] as the labels, and then applying it to the spectra of 55,000 stars from *APOGEE* DR10. *The Cannon* is very accurate. Its stellar labels compare well to the stars for which *APOGEE* pipeline (*ASPCAP*) labels are provided in DR10, with rms differences that are basically identical to the stated *ASPCAP* uncertainties. Beyond the reference labels, *The Cannon* makes no use of stellar models nor any line-list, but needs a set of reference objects that span label-space. *The Cannon* performs well at lower signal-to-noise, as it delivers comparably good labels even at one-ninth the *APOGEE* observing time. We discuss the limitations of *The Cannon* and its future potential, particularly, to bring different spectroscopic surveys onto a consistent scale of stellar labels.

Key words: methods: data analysis – methods: statistical – stars: abundances – stars: fundamental parameters – surveys – techniques: spectroscopic

$$f_{n\lambda} = \boldsymbol{\theta}_\lambda^T \cdot \boldsymbol{\ell}_n + \text{noise}$$

$$\ln p\left(f_{n\lambda} \middle| \boldsymbol{\theta}_\lambda^T, \boldsymbol{\ell}_n, s_\lambda^2\right) = -\frac{1}{2}\frac{\left[f_{n\lambda} - \boldsymbol{\theta}_\lambda^T \cdot \boldsymbol{\ell}_n\right]^2}{s_\lambda^2 + \sigma_{n\lambda}^2} - \frac{1}{2}\ln\left(s_\lambda^2 + \sigma_{n\lambda}^2\right).$$

数据驱动

Ness+ 2015, ApJ, 808, 16

https://github.com/annayqho/TheCannon

3.2，机器学习方法示例- The Payne

The Payne: Self-consistent ab initio Fitting of Stellar Spectra

Yuan-Sen Ting (丁源森)[1,2,3,4,5,7], Charlie Conroy[5], Hans-Walter Rix[6], and Phillip Cargile[5]

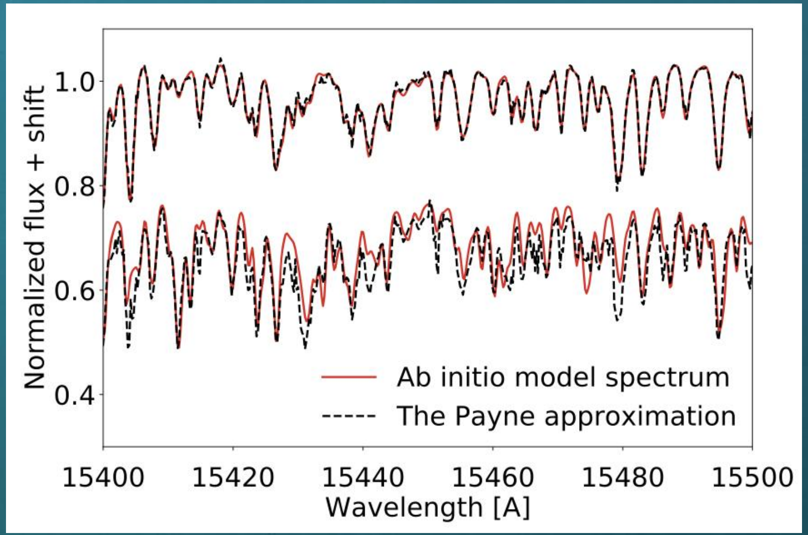
[1] Institute for Advanced Study, Princeton, NJ 08540, USA
[2] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[3] Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA
[4] Research School of Astronomy and Astrophysics, Australian National University, Cotter Road, ACT 2611, Canberra, Australia
[5] Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[6] Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany

Received 2018 April 3; revised 2019 May 9; accepted 2019 May 11; published 2019 July 8

Abstract

We present *The Payne*, a general method for the precise and simultaneous determination of numerous stellar labels from observed spectra, based on fitting physical spectral models. *The Payne* combines a number of important methodological aspects: it exploits the information from much of the available spectral range; it fits all labels (stellar parameters and elemental abundances) simultaneously; it uses spectral models, where the structure of the atmosphere and the radiative transport are consistently calculated to reflect the stellar labels. At its core *The Payne* has an approach to accurate and precise interpolation and prediction of the spectrum in high-dimensional label space that is flexible and robust, yet based on only a moderate number of ab initio models ($\mathcal{O}(1000)$ for 25 labels). With a simple neural-net-like functional form and a suitable choice of training labels, this interpolation yields a spectral flux prediction good to 10^{-3} rms across a wide range of $T_{\rm eff}$ and $\log g$ (including dwarfs and giants). We illustrate the power of this approach by applying it to the APOGEE DR14 data set, drawing on Kurucz models with recently improved line lists: without recalibration, we obtain physically sensible stellar parameters as well as 15 elemental abundances that appear to be more precise than the published APOGEE DR14 values. In short, *The Payne* is an approach that for the first time combines all these key ingredients, necessary for progress toward optimal modeling of survey spectra; and it leads to both precise and accurate estimates of stellar labels, based on physical models and without "recalibration." Both the codes and catalog are made publicly available online.

$$f_\lambda = w \cdot \sigma(\tilde{w}^i_\lambda \sigma(w^k_{\lambda i}\ell_k + b_{\lambda i}) + \tilde{b}) + \bar{f}_\lambda,$$



模型驱动

Ting+ 2017, ApJ, 879, 69

https://github.com/tingyuansen/The_Payne/

3.2，机器学习方法示例- HotPayne

A&A 662, A66 (2022)
https://doi.org/10.1051/0004-6361/202141570

Astronomy & Astrophysics

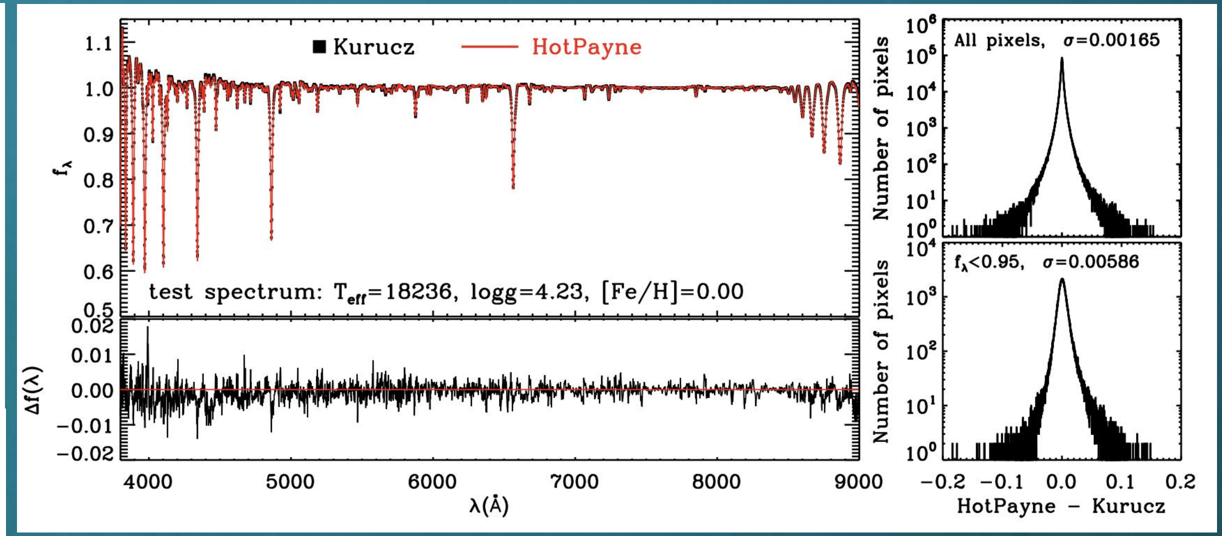
Stellar labels for hot stars from low-resolution spectra

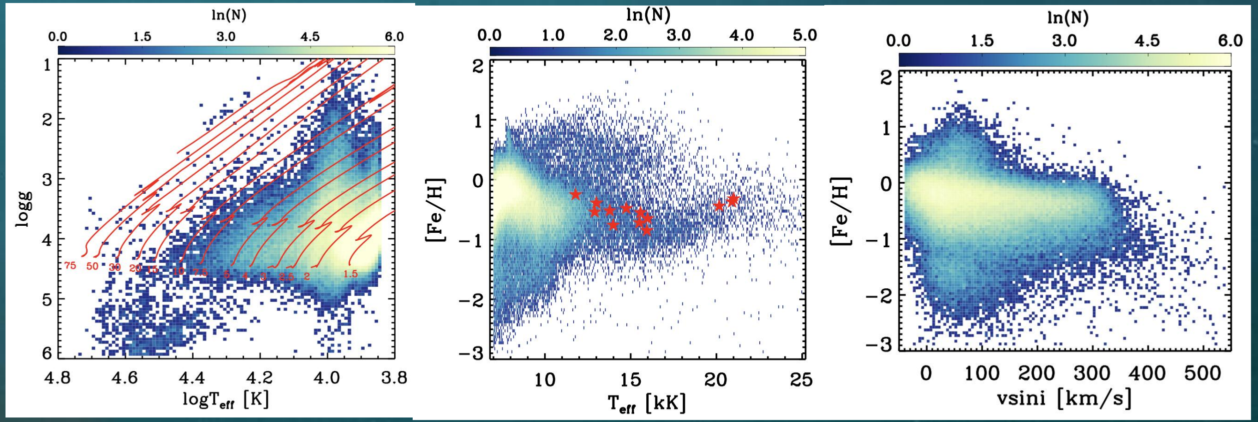
I. The HotPayne method and results for 330 000 stars from LAMOST DR6★

Maosheng Xiang[1], Hans-Walter Rix[1], Yuan-Sen Ting[2,3,4,5,★★], Rolf-Peter Kudritzki[6,7], Charlie Conroy[8], Eleonora Zari[1], Jian-Rong Shi[9,10], Norbert Przybilla[11], Maria Ramirez-Tannus[1], Andrew Tkachenko[12], Sarah Gebruers[12], and Xiao-Wei Liu[13]

ABSTRACT

We set out to determine stellar labels from low-resolution survey spectra of hot stars, specifically OBA stars with $T_{\rm eff} \gtrsim 7500$ K. This fills a gap in the scientific analysis of large spectroscopic stellar surveys such as LAMOST, which offers spectra for millions of stars at $R \sim 1800$ and covers 3800 Å $\le \lambda \le$ 9000 Å. We first explore the theoretical information content of such spectra to determine stellar labels via the Cramér-Rao bound. We show that in the limit of perfect model spectra and observed spectra with signal-to-noise ratio ~50−100, precise estimates are possible for a wide range of stellar labels: not only the effective temperature, $T_{\rm eff}$, surface gravity, $\log g$, and projected rotation velocity, $v\sin i$, but also the micro-turbulence velocity, $v_{\rm mic}$, helium abundance, $N_{\rm He}/N_{\rm tot}$, and the elemental abundances [C/H], [N/H], [O/H], [Si/H], [S/H], and [Fe/H]. Our analysis illustrates that the temperature regime of $T_{\rm eff} \sim 9500$ K is challenging as the dominant Balmer and Paschen line strengths vary little with $T_{\rm eff}$. We implement the simultaneous fitting of these 11 stellar labels to LAMOST hot-star spectra using the Payne approach, drawing on Kurucz's ATLAS12/SYNTHE local thermodynamic equilibrium spectra as the underlying models. We then obtain stellar parameter estimates for a sample of about 330 000 hot stars with LAMOST spectra, an increase by about two orders of magnitude in sample size. Among them, about 260 000 have good *Gaia* parallaxes ($\omega/\sigma_\omega > 5$), and their luminosities imply that ≳95% of them are luminous stars, mostly on the main sequence; the rest are evolved lower luminosity stars, such as hot subdwarfs and white dwarfs. We show that the fidelity of the results, particularly for the abundance estimates, is limited by the systematics of the underlying models as they do not account for nonlocal thermodynamic equilibrium effects. Finally, we show the detailed distribution of $v\sin i$ of stars with 8000−15 000 K, illustrating that it extends to a sharp cutoff at the critical rotation velocity, $v_{\rm crit}$, across a wide range of temperatures.





Xiang+ 2022, A&A, 662, A66

3.2，机器学习方法示例- TransformerPayne

TransformerPayne: enhancing spectral emulation accuracy and data efficiency by capturing long-range correlations

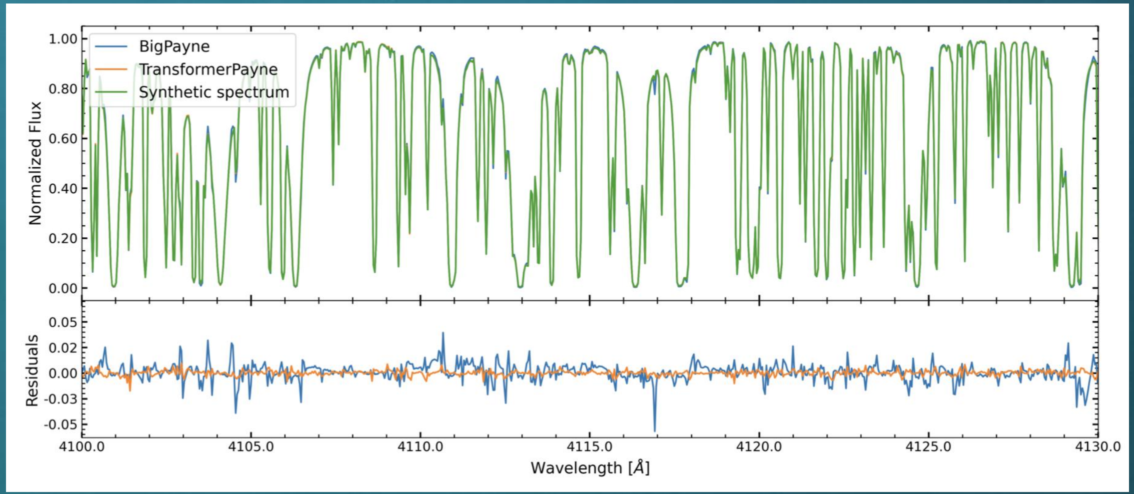
Tomasz Różański[1,2], Yuan-Sen Ting[1,3,4,5], and Maja Jabłońska[1]

[1] Research School of Astronomy & Astrophysics, The Australian National University, Cotter Rd., Weston, ACT 2611, Australia
e-mail: tomasz.rozanski1@anu.edu.au
[2] Astronomical Institute, University of Wrocław, Kopernika 11, 51-622 Wrocław, Poland
[3] School of Computing, The Australian National University, Acton, ACT 2601, Australia
[4] Department of Astronomy, The Ohio State University, Columbus, OH 45701, USA
[5] Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH 43210, USA

Received 11 July 2024 / Accepted date

ABSTRACT

Context. Stellar spectra emulators often rely on large grids and tend to reach a plateau in emulation accuracy, leading to significant systematic errors when inferring stellar properties.
Aims. Our study explores the use of Transformer models to capture long-range information in spectra, comparing their performance to The Payne emulator (a fully connected multilayer perceptron), an expanded version of The Payne, and a convolutional-based emulator.
Methods. We develop the TransformerPayne neural network architecture, which leverages the attention mechanism to efficiently capture long-range correlations in stellar spectra. We adopt two grids of synthetic spectra and compare emulators using residuals of emulation and by inference of spectral parameters for synthetic spectra from validation dataset.
Results. The TransformerPayne emulator outperforms all other tested emulators, achieving a mean absolute error (MAE) in emulation of approximately 0.0015, when trained on the full grid. The largest improvements with respect to other emulators, measured using MAE, are for grids containing between 1000 and 10 000 spectra, and vary between 2 and 5 times when comparing to the large version of The Payne. Fine-tuning enables up to a tenfold reduction in the size of the training grids when comparing to version trained from scratch. We also investigated the attention maps of TransformerPayne emulator, finding that they encode interpretable features shared across many lines of chosen elements. We show that although scaling to a much larger network can improve The Payne emulator significantly, decreasing the MAE of emulation from 0.012 to 0.003 when trained on the full training dataset, the TransformerPayne consistently emulates spectra with the smallest MAE. Convolutional-based architectures saturate with a MAE around 0.05.
Conclusions. Appropriate inductive biases in the TransformerPayne architecture result in improved accuracy, data efficiency, and interpretability over existing methods.



Róz˙an´ski+ 2024, arXiv

3.2，机器学习方法示例- Data-driven Payne

The Astrophysical Journal Letters, 849:L9 (6pp), 2017 November 1 https://doi.org/10.3847/2041-8213/aa921c

CrossMark

Measuring 14 Elemental Abundances with $R = 1800$ LAMOST Spectra

Yuan-Sen Ting (丁源森)[1,2,3,4,5], Hans-Walter Rix[2], Charlie Conroy[6], Anna Y. Q. Ho[7], and Jane Lin[1]
[1] Research School of Astronomy and Astrophysics, Mount Stromlo Observatory, Cotter Road, Weston Creek, ACT 2611, Australia
[2] Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany
[3] Institute for Advanced Study, Princeton, NJ 08540, USA
[4] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[5] Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA
[6] Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[7] Cahill Center for Astrophysics, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA
Received 2017 August 5; revised 2017 October 6; accepted 2017 October 8; published 2017 October 23

Abstract

The LAMOST survey has acquired low-resolution spectra ($R = 1800$) for 5 million stars across the Milky Way, far more than any current stellar survey at a corresponding or higher spectral resolution. It is often assumed that only very few elemental abundances can be measured from such low-resolution spectra, limiting their utility for Galactic archaeology studies. However, Ting et al. used ab initio models to argue that low-resolution spectra should enable precision measurements of many elemental abundances, at least in theory. Here, we verify this claim in practice by measuring the relative abundances of 14 elements from LAMOST spectra with a precision of $\lesssim 0.1$ dex for objects with $S/N_{LAMOST} \gtrsim 30$ (per pixel). We employ a spectral modeling method in which a data-driven model is combined with priors that the model gradient spectra should resemble ab initio spectral models. This approach assures that the data-driven abundance determinations draw on physically sensible features in the spectrum in their predictions and do not just exploit astrophysical correlations among abundances. Our analysis is constrained to the number of elemental abundances measured in the APOGEE survey, which is the source of the training labels. Obtaining high quality/resolution spectra for a subset of LAMOST stars to measure more elemental abundances as training labels and then applying this method to the full LAMOST catalog will provide a sample with more than 20 elemental abundances, which is an order of magnitude larger than current high-resolution surveys, substantially increasing the sample size for Galactic archaeology.

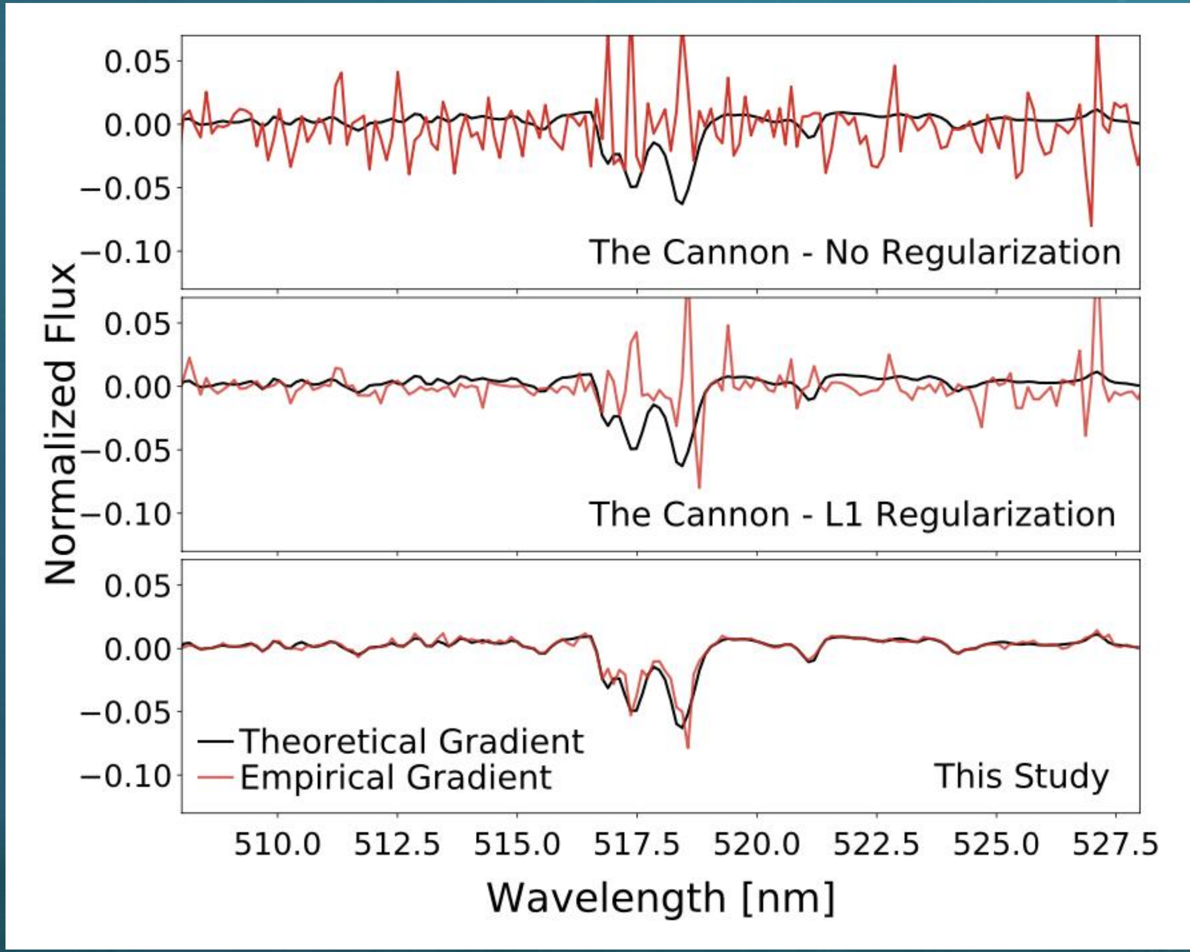
$$f_\lambda = w \cdot \sigma(\tilde{w}^i_\lambda \sigma(w^k_{\lambda i}\ell_k + b_{\lambda i}) + \tilde{b}) + \bar{f}_\lambda,$$

$$\mathcal{L}(\{f_{\rm obs}\}|\boldsymbol{w}, \boldsymbol{b}, \{\boldsymbol{\ell}_{\rm obs}\}) = \frac{1}{N_S}\sum_{i=1}^{N_S}\frac{(f(\lambda|\boldsymbol{\ell}_{\rm obs,i}) - f_{\rm obs,i}(\lambda))^2}{\sigma^2_{\rm obs,i}(\lambda)} + D_{\rm scale}\cdot\frac{1}{N_l}\sum_{j=1}^{N_l} \times \log\frac{|f'(\lambda|\boldsymbol{\ell}_{\rm ref}) - f'_{\rm ab\ initio}(\lambda)|}{|f'_{\rm ab\ initio}(\lambda)|},$$

数据+物理模型联合驱动

Ting+ 2017, ApJ, 808, 16

3.2，机器学习方法示例- Data-Drive Payne

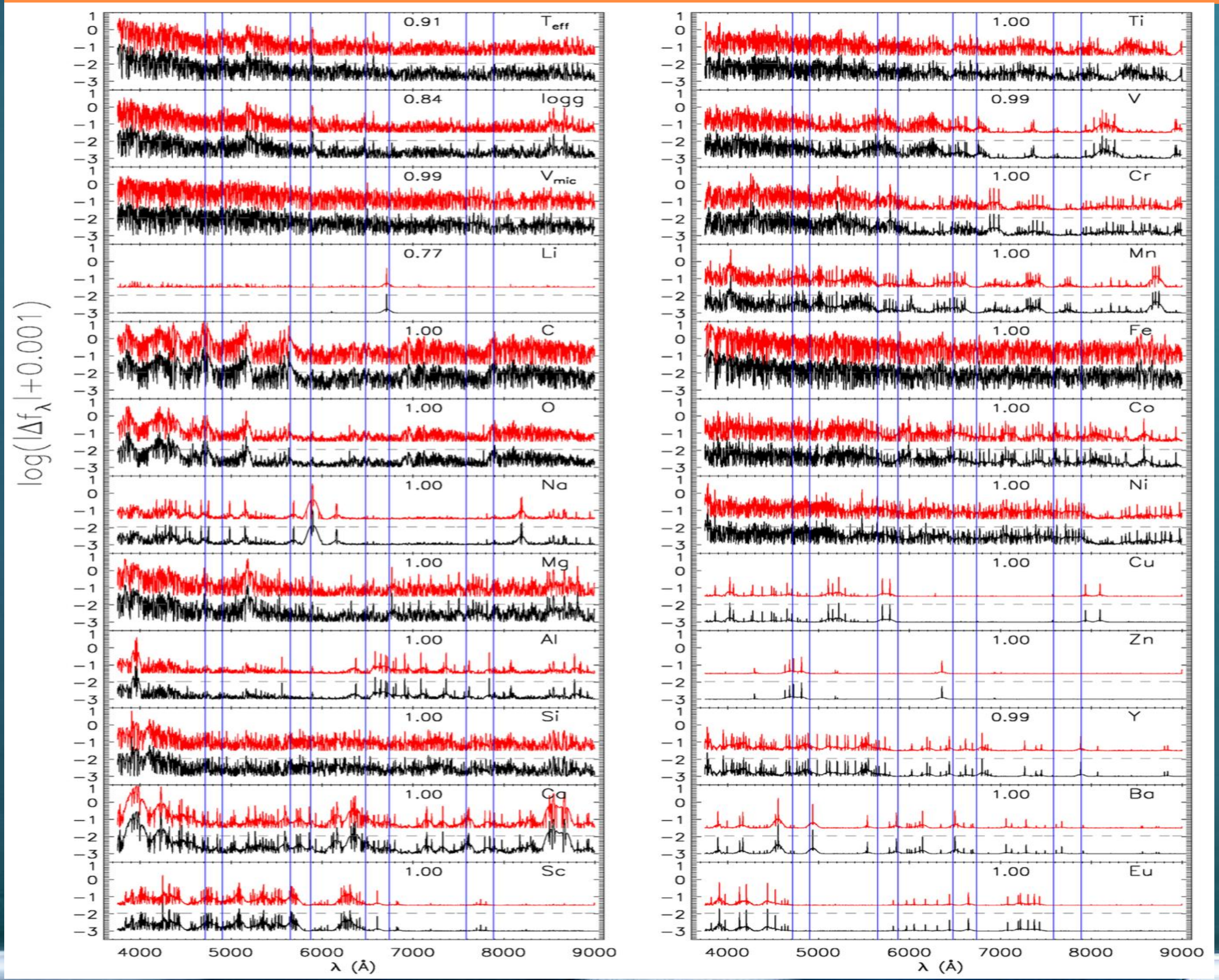


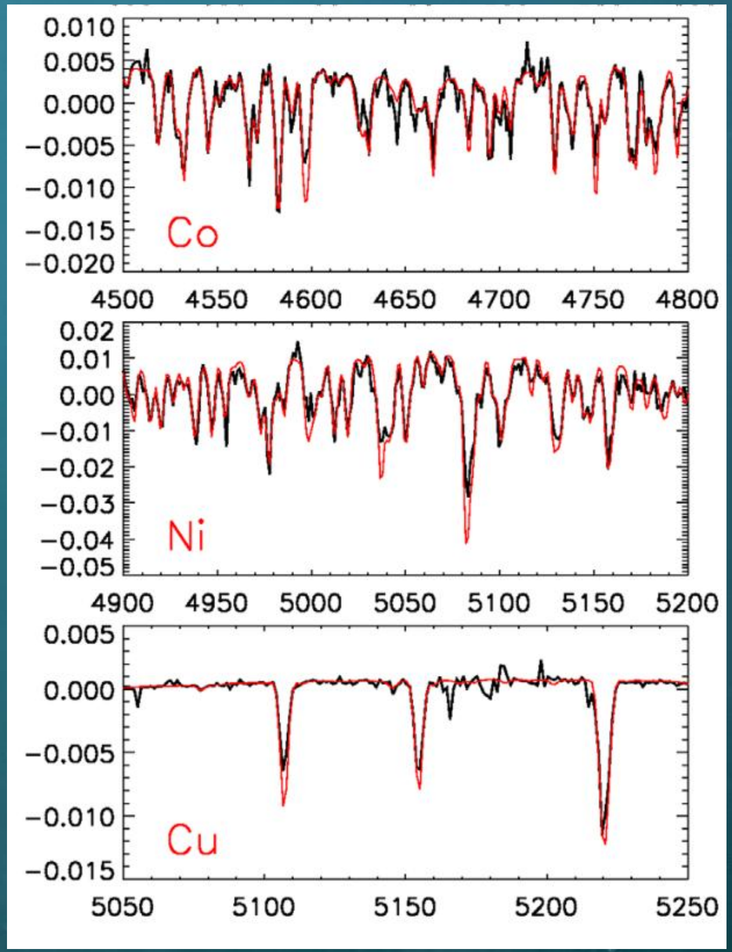
Ting+ 2017, ApJ, 808, 16

数据驱动（Cannon）

数据+物理模型联合驱动（DD-Payne）

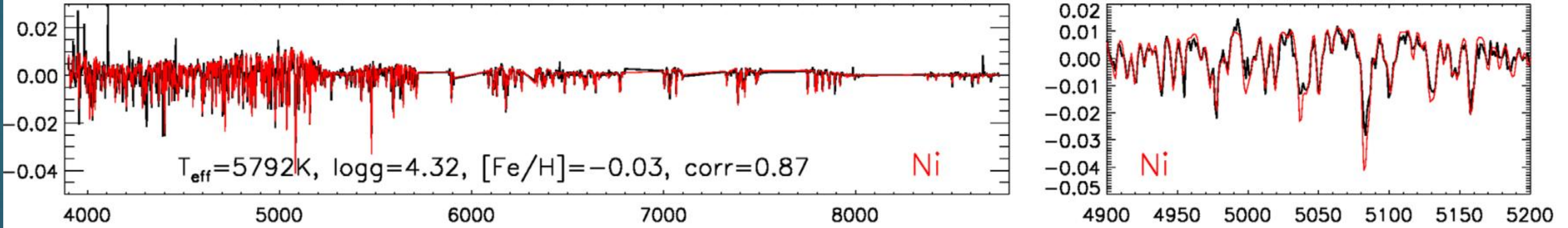
3.3，机器学习应用-LAMOST光谱分析

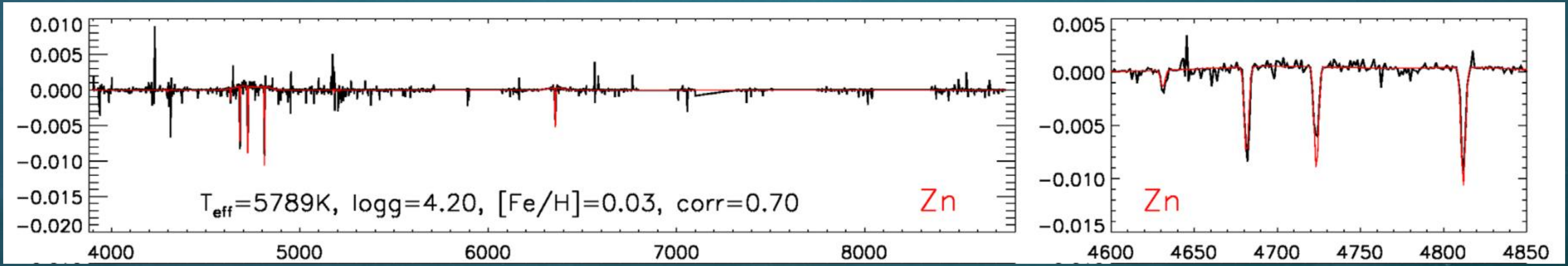




DD-Payne:
Xiang+ 2019, ApJS, 245, 34

3.3，机器学习应用-LAMOST光谱分析





Xiang+ 2019, ApJS, 245, 34

3.3，机器学习应用-DESI光谱分析

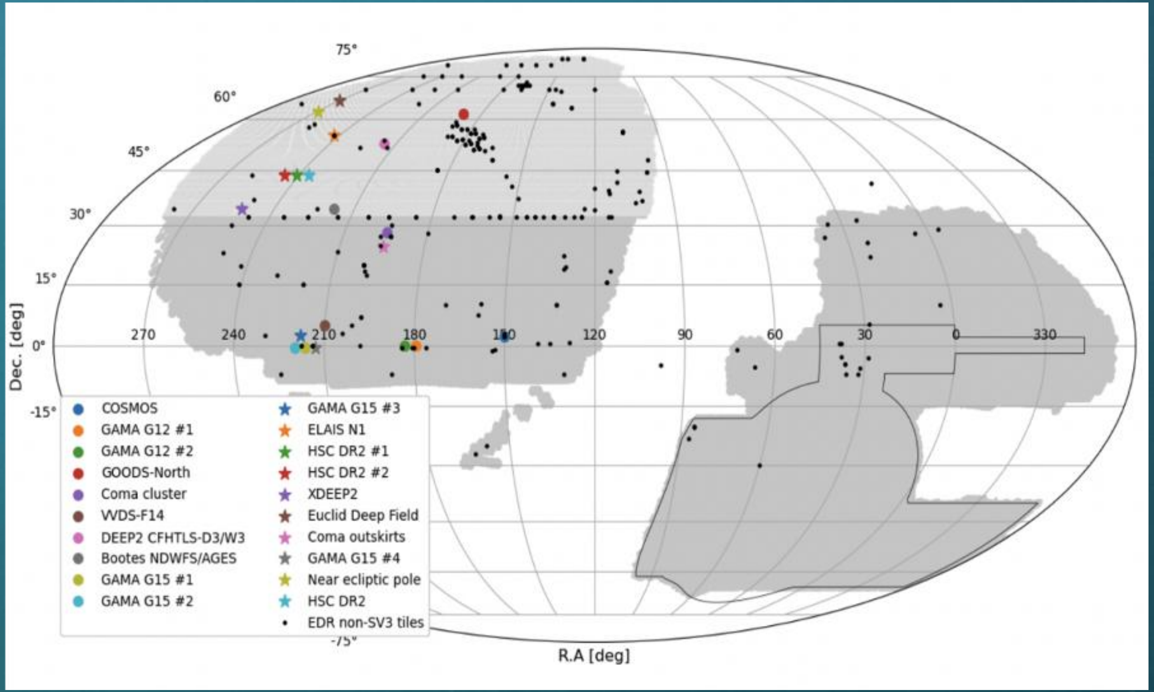
DESI: R=2000-5000, λλ3700-9500埃

Determining Stellar Elemental Abundances from DESI Spectra with the Data-Driven Payne

Meng Zhang,[1] Maosheng Xiang,[1,2] Yuan-Sen Ting,[3,4,5,6] Jiahui Wang,[7,1] Haining Li,[1] Hu Zou,[1,7] Jundan Nie,[1] Lanya Mou,[1,7] Tianmin Wu,[1,7] Yaqian Wu,[1] and Jifeng Liu[1,7,2,8]

ABSTRACT

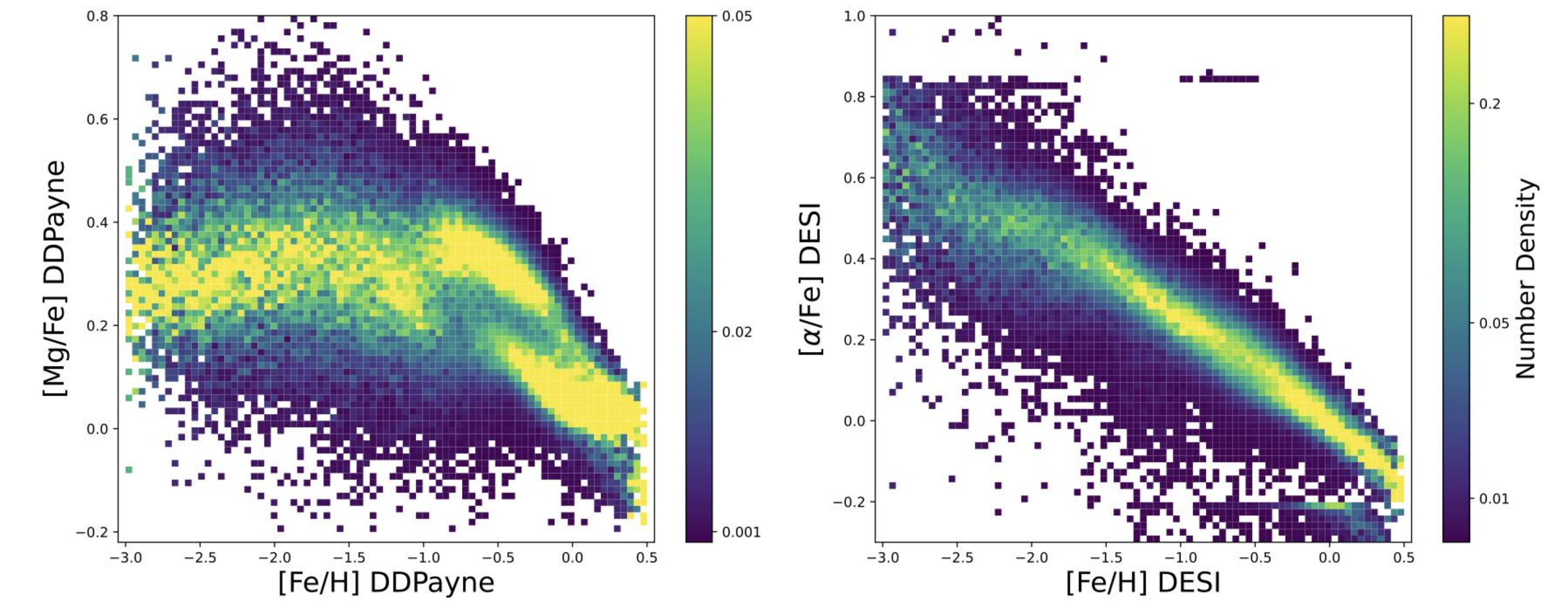
Stellar abundances for a large number of stars are key information for the study of Galactic formation history. Large spectroscopic surveys such as DESI and LAMOST take median-to-low resolution ($R \lesssim 5000$) spectra in the full optical wavelength range for millions of stars. However, line blending effect in these spectra causes great challenges for the elemental abundances determination. Here we employ the DD-Payne, a data-driven method regularised by differential spectra from stellar physical models, to the DESI EDR spectra for stellar abundance determination. Our implementation delivers 15 labels, including effective temperature $T_{\rm eff}$, surface gravity $\log\ g$, microturbulence velocity $v_{\rm mic}$, and abundances for 12 individual elements, namely C, N, O, Mg, Al, Si, Ca, Ti, Cr, Mn, Fe, Ni. Given a spectral signal-to-noise ratio of 100 per pixel, internal precision of the label estimates are about 20 K for $T_{\rm eff}$, 0.05 dex for $\log\ g$, and 0.05 dex for most elemental abundances. These results are agree with theoretical limits from the Crámer-Rao bound calculation within a factor of two. The Gaia-Enceladus-Sausage that contributes the majority of the accreted halo stars are discernible from the disk and in-situ halo populations in the resultant [Mg/Fe]-[Fe/H] and [Al/Fe]-[Fe/H] abundance spaces. We also provide distance and orbital parameters for the sample stars, which spread a distance out to ∼100 kpc. The DESI sample has a significant higher fraction of distant (or metal-poor) stars than other existed spectroscopic surveys, making it a powerful data set to study the Galactic outskirts. The catalog is publicly available.



Zhang+ 2024, ApJS, in press; arXiv2402.06242

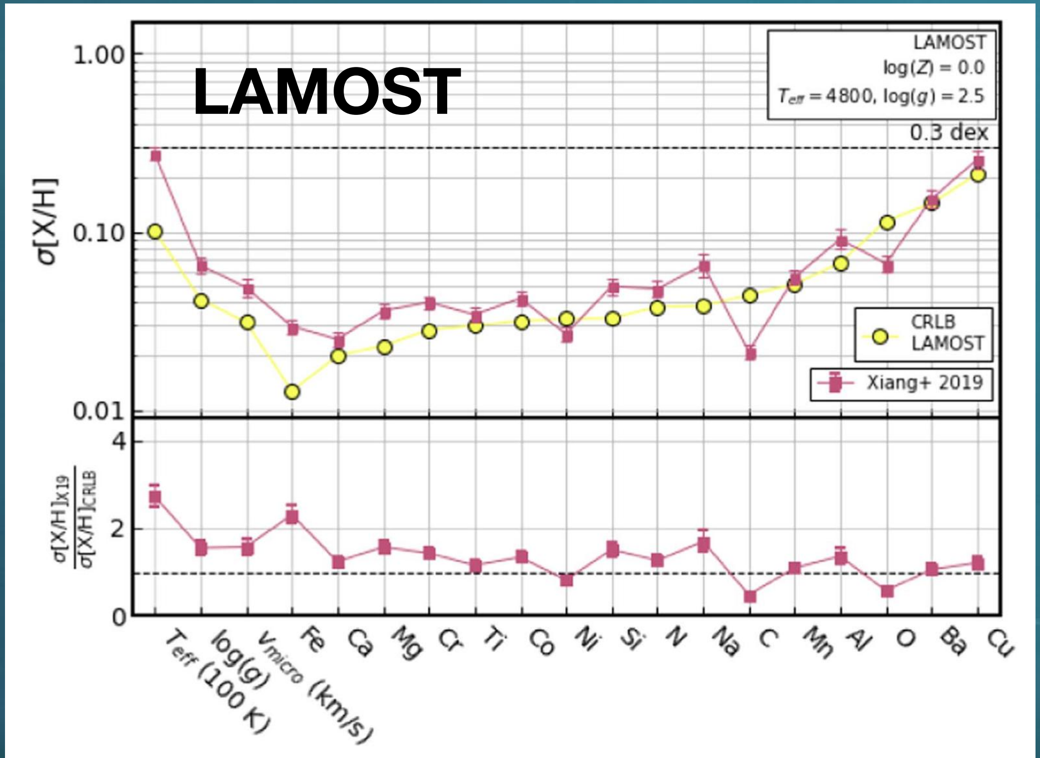
3.3，机器学习应用-DESI光谱分析

DESI DD-Payne **DESI Official**

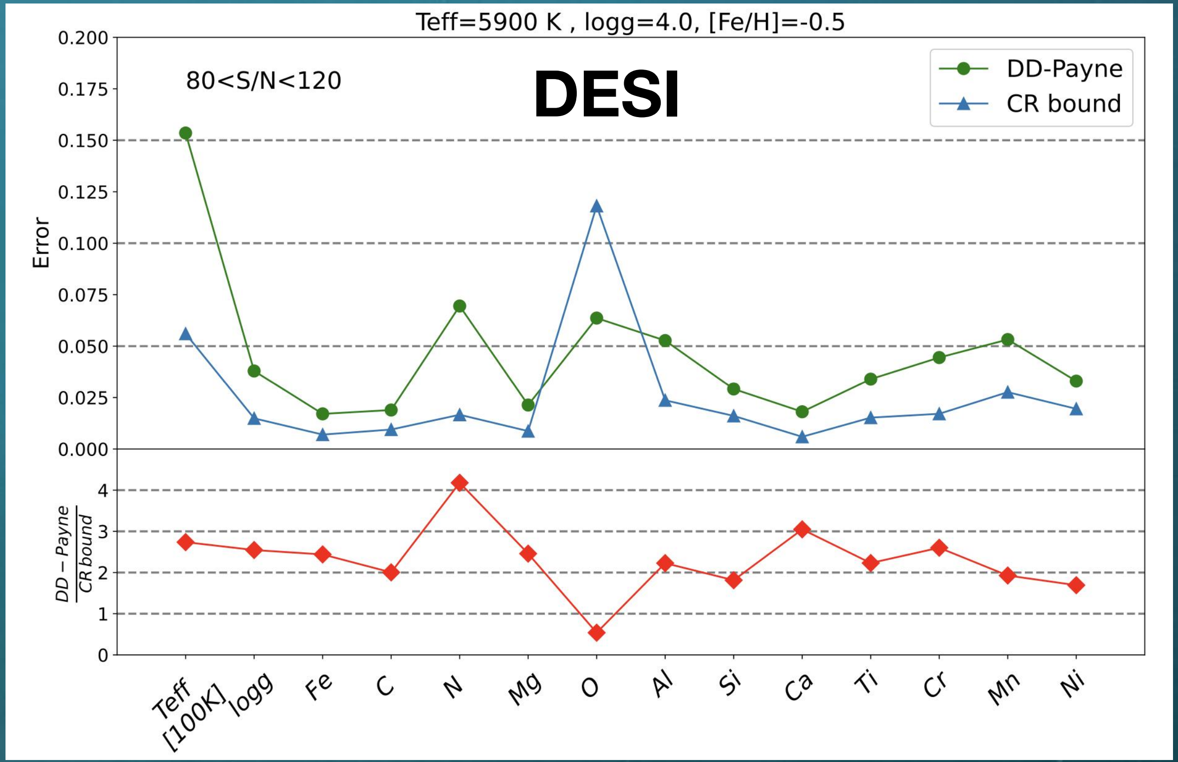


Zhang+ 2024, ApJS, in press; arXiv2402.06242

3.3，机器学习应用-LAMOST/DESI参数精度



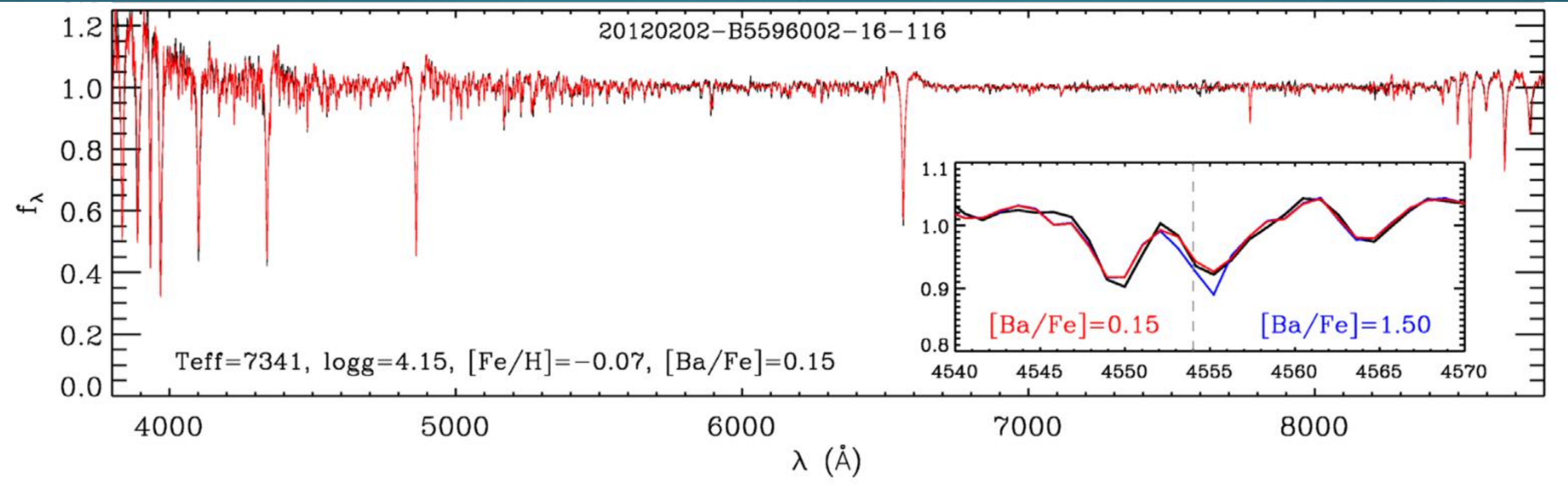
Sandford+ 2020, ApJS, 249, 24



Zhang+ 2024, ApJS, in press

3.3，机器学习应用-光谱标准化

◆ 利用卷积平滑滤波取代传统连续谱
- ✓ 模型与目标谱一致归一化
- ✓ 对噪声不敏感



内容提纲

（一）什么是恒星参数测定？

（二）为什么恒星参数测定需要机器学习？

（三）机器学习在恒星参数测定中的应用

（四）机器学习光谱分析面临的问题

（五）总结与展望

4， 机器学习光谱分析面临的问题

- 算法的物理可解释性
- 高维度建模和预测
- 训练样本稀疏
- 训练样本标签缺失
- 多模态数据
- 光谱噪声优化解决方案
- 参数基准与定标
- 机器学习 =>(AI =>)新物理

4， 机器学习光谱分析面临的问题

- ◆ 算法的物理可解释性 -- 数据+物理模型双驱动
- ◆ 高维度建模和预测
- ◆ 训练样本稀疏
- ◆ 训练样本标签缺失
- ◆ 多模态数据
- ◆ 光谱噪声优化解决方案
- ◆ 参数基准与定标
- ◆ 机器学习 =>(AI =>)新物理

4， 机器学习光谱分析面临的问题

- 算法的物理可解释性
- 高维度建模和预测 -- 需要优化算法+强大的算力
- 训练样本稀疏
- 训练样本标签缺失
- 多模态数据
- 光谱噪声优化解决方案
- 参数基准与定标
- 机器学习 =>(AI =>)新物理

4， 机器学习光谱分析面临的问题

- 算法的物理可解释性
- 高维度建模和预测
- 训练样本稀疏 -- 样本构建，算法优化
- 训练样本标签缺失
- 多模态数据
- 光谱噪声优化解决方案
- 参数基准与定标
- 机器学习 =>(AI =>)新物理

4， 机器学习光谱分析面临的问题

- 算法的物理可解释性
- 高维度建模和预测
- 训练样本稀疏
- 训练样本标签缺失
- 多模态数据 -- 光谱 + 测光 + 天测 + 星震 + ...
- 光谱噪声优化解决方案
- 参数基准与定标
- 机器学习 =>(AI =>)新物理

4， 机器学习光谱分析面临的问题

- ◆ 算法的物理可解释性
- ◆ 高维度建模和预测
- ◆ 训练样本稀疏
- ◆ 训练样本标签缺失
- ◆ 多模态数据
- ◆ 光谱噪声优化解决方案 -- 低信噪比光谱优化算法
- ◆ 参数基准与定标
- ◆ 机器学习 =>(AI =>)新物理

4， 机器学习光谱分析面临的问题

- 算法的物理可解释性
- 高维度建模和预测
- 训练样本稀疏
- 训练样本标签缺失
- 多模态数据
- 光谱噪声优化解决方案
- 参数基准与定标 -- 仍缺乏适用于巡天大数据的理想定标星
- 机器学习 =>(AI =>)新物理

4，机器学习光谱分析面临的问题

- ◆ 算法的物理可解释性
- ◆ 高维度建模和预测
- ◆ 训练样本稀疏
- ◆ 训练样本标签缺失
- ◆ 多模态数据
- ◆ 光谱噪声优化解决方案
- ◆ 参数基准与定标
- ◆ 机器学习 =>(AI =>)新物理

4，机器学习光谱分析面临的问题

- 算法的物理可解释性
- 高维度建模和预测
- 训练样本稀疏
- 训练样本标签缺失
- 多模态数据
- 光谱噪声优化解决方案
- 参数基准与定标
- 机器学习 =>(AI =>)新物理

天文学家
起决定性作用!

5，总结与展望

- 什么是恒星参数测定：给定一组光谱，寻找一组恒星参数组合*θ*，使得模型**f(*θ*)**与观测最佳匹配 **-- forward modelling**；其本质是高维度多参数求解问题
- 机器学习是充分利用光谱信息、高维度参数求解、大样本光谱分析的必要选择
- 机器学习在巡天数据分析上取得了成功（**LAMOST, DESI**）；并将继续发挥关键作用**(SDSS-V, GALAH, 4MOST, WEAVE, etc.)**
- 机器学习光谱分析面临的瓶颈是传统光谱天文学家及各位同学的新机遇

5，总结与展望

◆ 如何做好机器学习?

- 定性问题，选取/提出合适的算法
- 优质的训练/监督样本（干净的数据）
- 反复测试、检验、调整

5，总结与展望

◆ 如何做好机器学习？

● 定性问题，选取/提出合适的算法

● 优质的训练/监督样本（干净的数据）

● 反复测试、检验、调整

我的模型最优化了吗？

谢谢！