

基于神经网络算法的恒星化学元素丰度 估计与贫金属恒星的搜寻

卜育德 山东大学（威海）数学与统计学院

合作者：陈玉琴，罗阿里，谈克峰，李海宁，赵景昆，
苑海波，衣振萍，王瑞，梁俊超等

CONTENTS

目录

01

研究背景

Review of work

02

化学元素丰度估计

Estimation of abundance

03

贫金属星的搜寻

Search for metal-poor stars

04

总结与展望

Summary and work plan

一 研究背景

- 恒星大气物理参数与化学元素丰度对研究恒星的演化和银河系的演化与结构有着非常重要的意义。
- 贫金属星是银河系演化的化石，提供银河系早期的演化信息，搜寻贫金属恒星具有非常重要的价值。
- CSST可以提供海量的恒星光谱和测光数据，为银河系研究提供前所未有的机会。

二 基于残差神经网络的化学元素丰度估计

- 恒星的化学元素包括Mg, C, N, O等元素。
- 恒星的化学元素丰度估计主要使用高分辨率光谱，光谱数量少。
- LAMOST光谱提供海量的恒星化学元素丰度，为研究银河系结构和演化提供了素材。
- CSST发布低分辨光谱，获取化学元素丰度信息非常困难。
- 机器学习方法提供了一种可能。

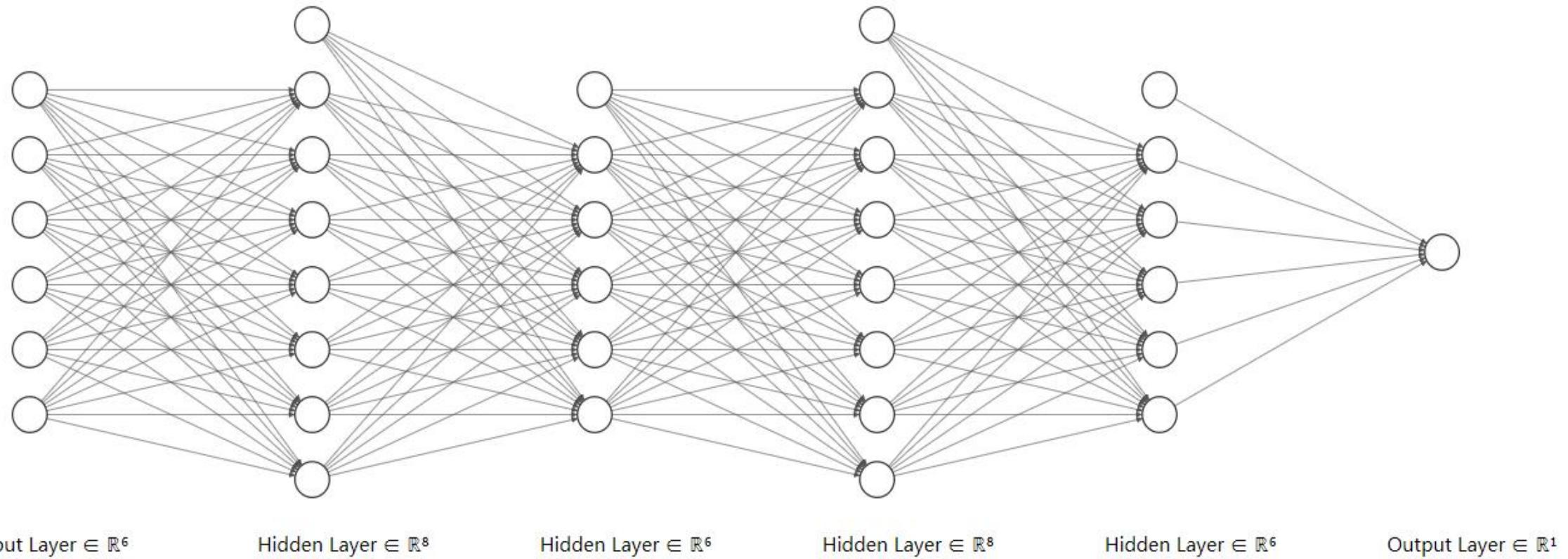
研究综述

- Lee et al. (2011) 使用SSPP 估计SDSS 光谱的元素丰度，光谱信噪比大于20时，元素丰度误差为0.1 dex；光谱信噪比大于50时，元素丰度误差为0.07 dex。
- Bu et al. (2016) 研究了一种基于ELM模型的Mg元素丰度估计方法，该研究证实了ELM算法能够精确确定MILES光谱的Mg元素的丰度。该模型推广到LAMOST低分辨率光谱取得了较高精度，信噪比大于50时，估计误差为0.06 dex。
- Ho et al. (2017) 使用The Cannon建立了450000条LAMOST低分辨率光谱与元素丰度的回归模型，误差为0.05 dex。
- Leung et al. (2018) 使用卷积神经网络(CNN)估计了APOGEE光谱的18种元素丰度
- Xiang et al. (2019) 使用 DD-Payne 估计了600万 LAMOST 恒星光谱的16种化学元素丰度

残差神经网络

- 残差神经网络是一种新的神经网络结构
- 深度学习层数越深，会出现学习能力退化问题
- 残差神经网络通过添加快捷连接，解决学习退化问题

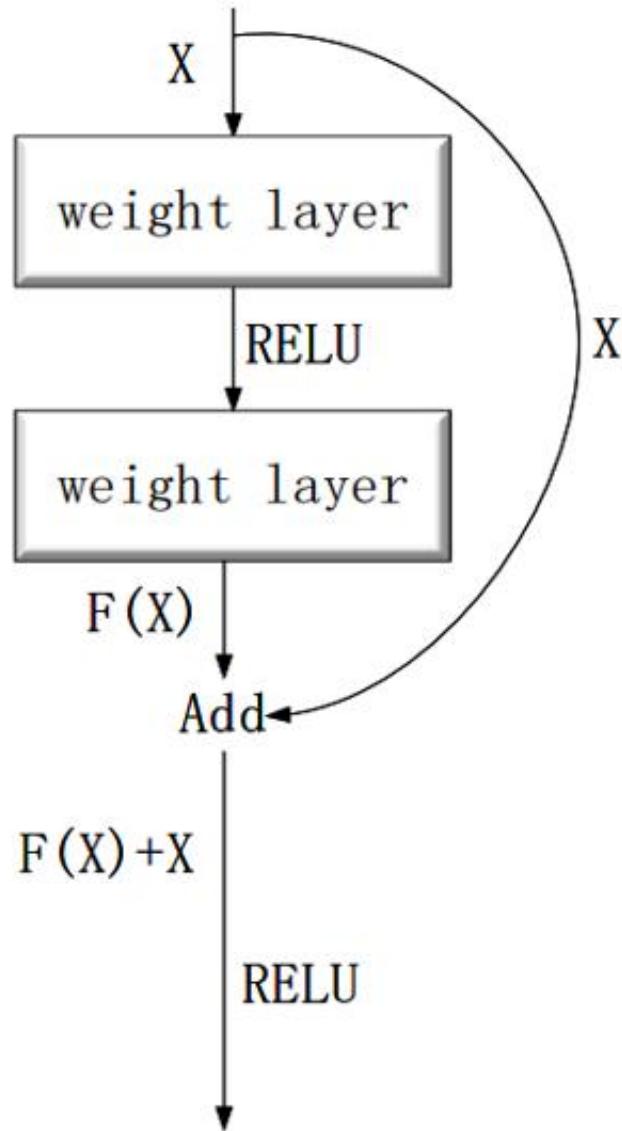
传统神经网络结构图



优点：比起机器学习模型更适应大数据时代预测任务，随着隐藏层不同以及深度增加模型性能增强。

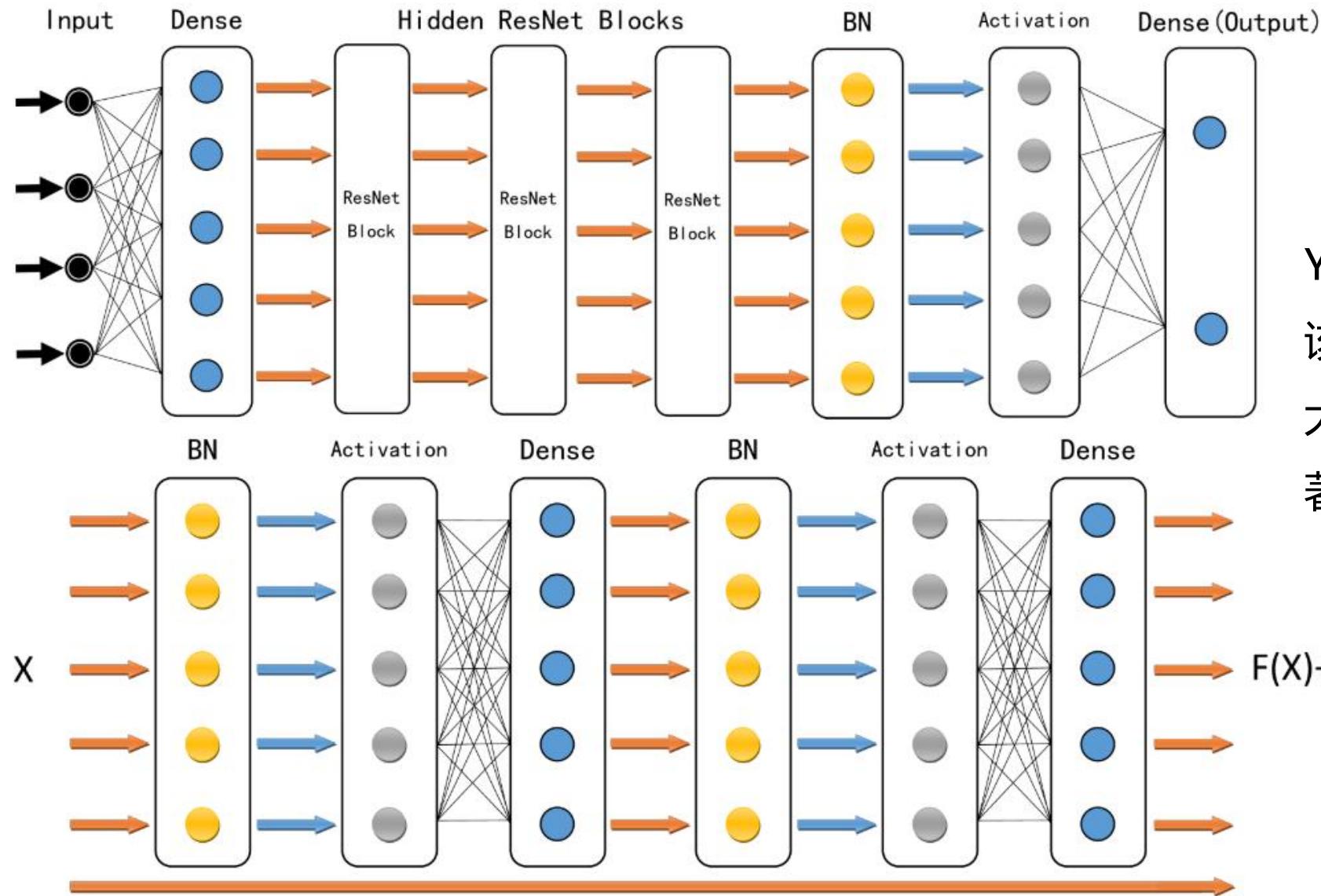
缺点：隐藏层加深带来模型退化问题，即训练集与验证集预测精度降低。

残差块



- 为了克服传统神经网络隐藏层加深带来的模型退化问题，何凯明等提出左图残差块。通过加入恒等变换 X 连接浅层与深层，使得深层网络可以学习到浅层任务。
- 其中每一个残差块对应神经网络一个新的隐藏层。

残差神经网络结构



Yang et al.(2020)使用该模型估计DA白矮星大气参数，参数精度显著提升。

2.2 实验数据

光谱数据：来源于LAMOST DR8低分辨率光谱；

元素丰度：来源于APOGEE DR16。

实验数据中包含[C/Fe]、[Mg/Fe]、[N/Fe]的光谱分别为25158条、

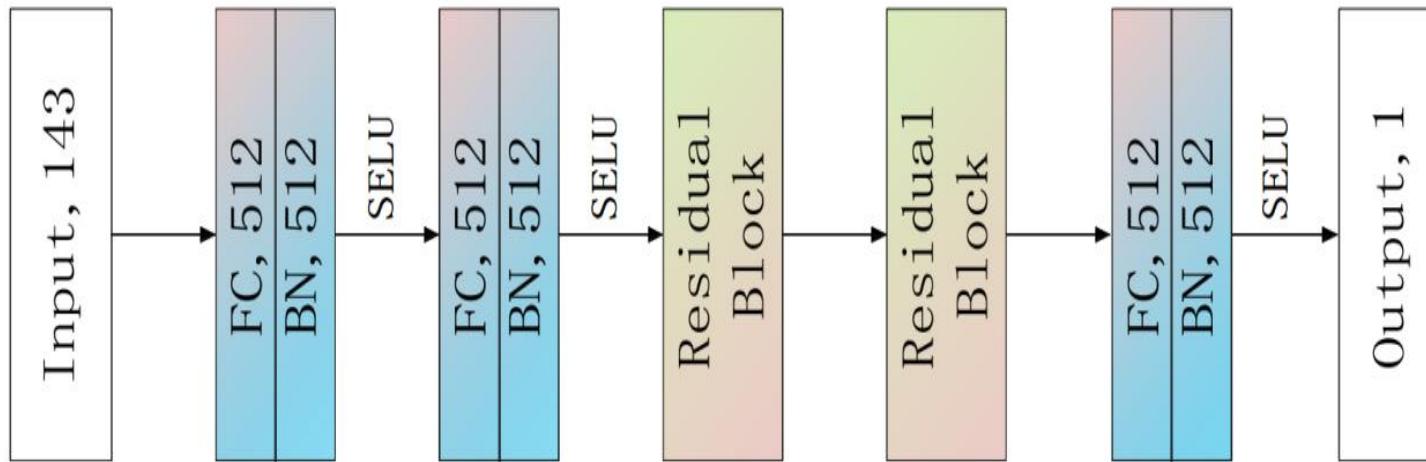
25219条、24196条

2.3 实验过程

- LAMOST光谱数据预处理：剔除异常值、光谱样本归一化处理。
- 模拟CSST光谱：将LAMOST低分辨光谱分辨率降为200。
- 构造主成分特征：根据测光数据实验结果，同样使用主成分特征缓解模型不稳定性问题。
- 模型训练：调整超参数，得到最优模型。
- 模型对比：对比ANN+残差块、ANN, CNN等模型关于[C/Fe], [Mg/Fe], [N/Fe]的预测精度，为CSST低分辨光谱提供高性能模型。

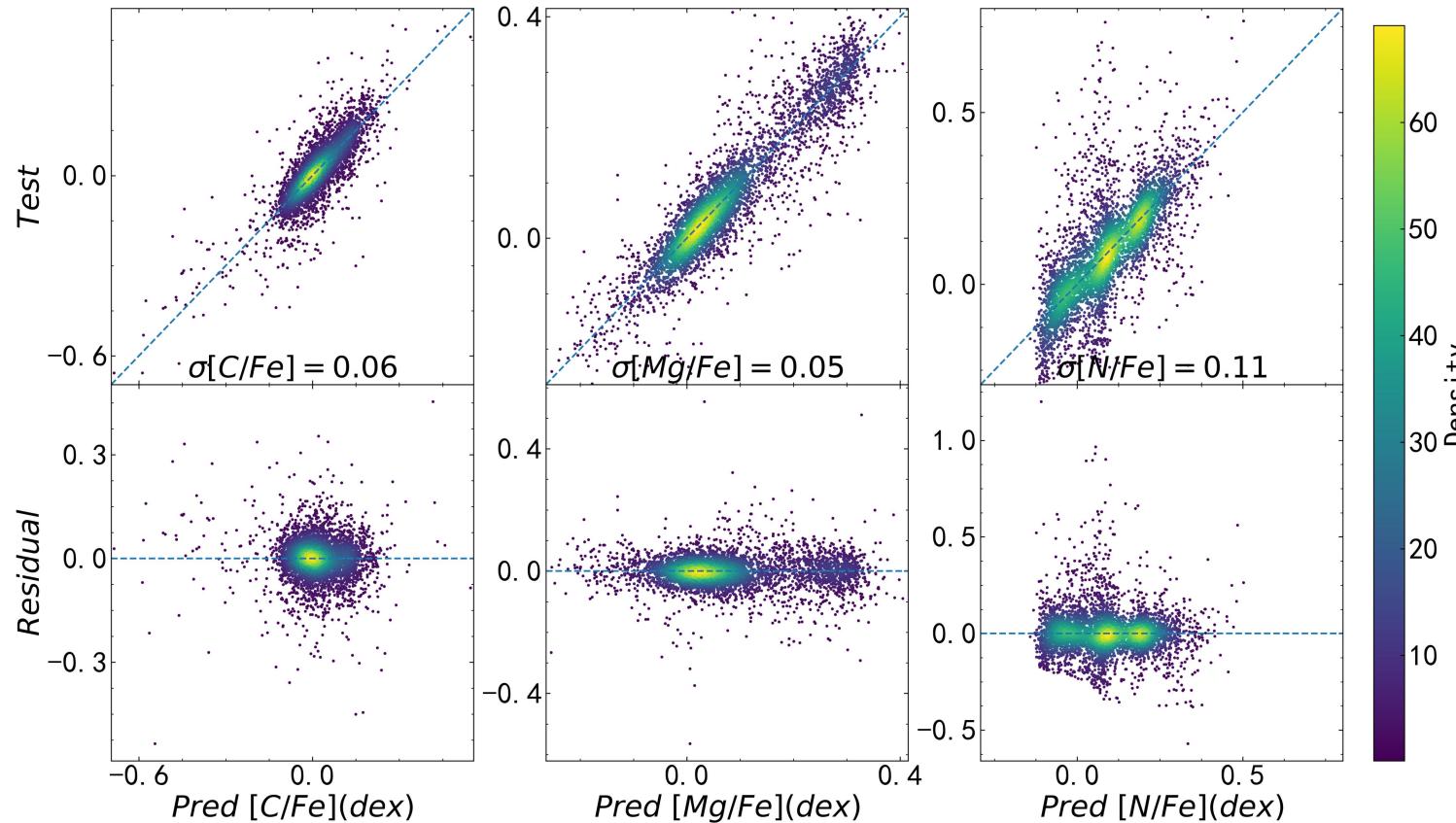
网络结构

- 输入层：输入层为CSST光谱数据，光谱维度为143，即每一条样本包含143个变量。
- 残差块：每个残差块结构前图所示，残差块中的每个全连接层包含512个神经元。
- 全连接层：每个全连接层包含512个神经元，使用SELU激活函数。



2.4 实验结果与分析

1. ANN+残差块预测[C/Fe], [Mg/Fe], [N/Fe]



[C/Fe], [Mg/Fe], [N/Fe]预测精度接近高分辨光谱，可以弥补测光数据不足以估计该参数的缺点。



2.4 实验结果与分析

► 模型对比

模型	[C/Fe] (dex)	[Mg/Fe] (dex)	[N/Fe] (dex)
ANN	0.068	0.056	0.121
StarNet	0.072	0.064	0.131
二维卷积神经网络	0.071	0.065	0.122
ANN+残差块	0.056	0.049	0.112

► 结论

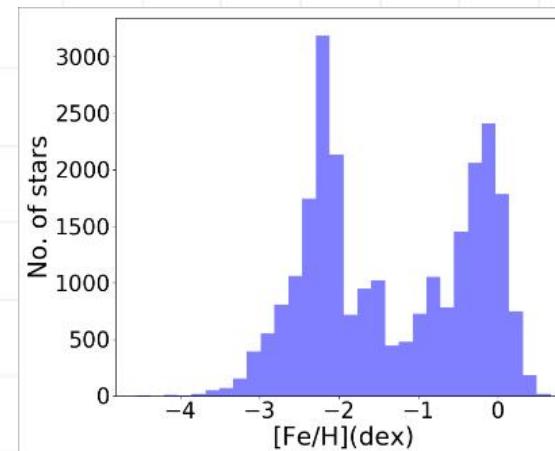
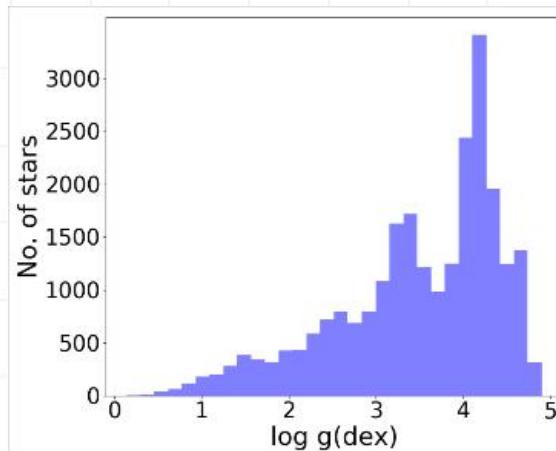
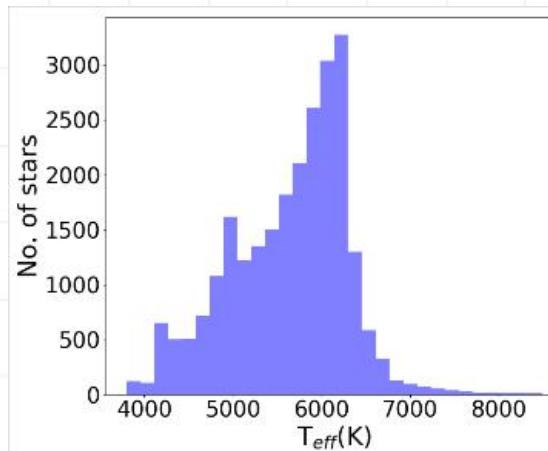
- 低分辨光谱可以以较高精度预测[C/Fe], [Mg/Fe], [N/Fe]等元素丰度。
- ANN+残差块适用于低分辨率光谱的回归分析，且性能高于其余神经网络。

三. 贫金属星的参数估计与搜寻

- VMP (Very metal-poor star) 星是指 $[Fe/H] < -2.0$ 的贫金属星恒星；
- VMP恒星金属丰度低，谱线特征弱，目前大多数的恒星参数估计方法在确定VMP恒星的大气参数时效果并不理想；
- 本研究提出一种新的卷积神经网络模型，可以对LAMOST恒星尤其是VMP恒星的大气物理参数及其碳元素丰度进行精确估计；
- 可以正确识别VMP恒星，并给出VMP星的碳元素丰度。

➤ 数据

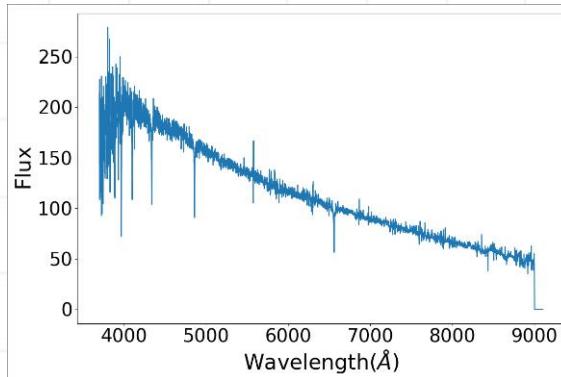
- 实验中使用的光谱数据来自于LAMOST，共包含25000条恒星光谱。
- 10008条VMP恒星光谱来自于李海宁在2018年给出的VMP星表，她们用光谱匹配和线指数的方法对这些恒星光谱的参数进行了测量校正。
- 14992条恒星光谱是从LAMOST DR7中随机选取的，恒星的参数通过LAMOST pipeline确定的。
- 恒星大气参数的覆盖范围： $3802\text{K} \leq \text{Teff} \leq 8482\text{K}$, $0.14\text{dex} \leq \log g \leq 4.90\text{dex}$ 和 $-4.55\text{dex} \leq [\text{Fe}/\text{H}] \leq 0.67\text{dex}$ 。



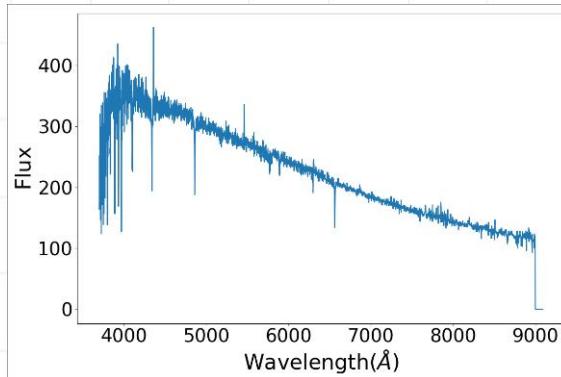
3.1 实验数据

数据介绍

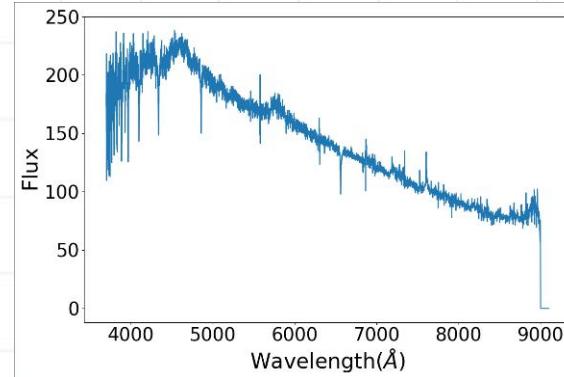
➤ VMP恒星光谱和正常光谱比较



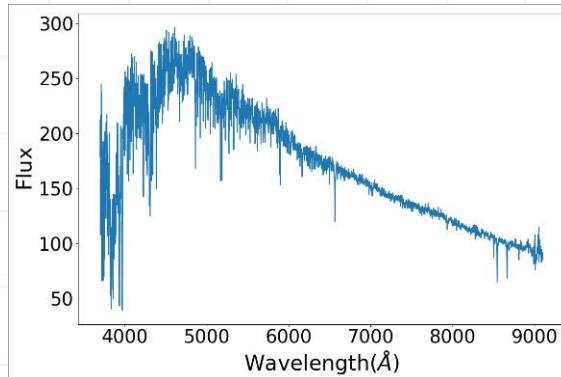
(a) $[\text{Fe}/\text{H}] = -2.63 \text{dex}$



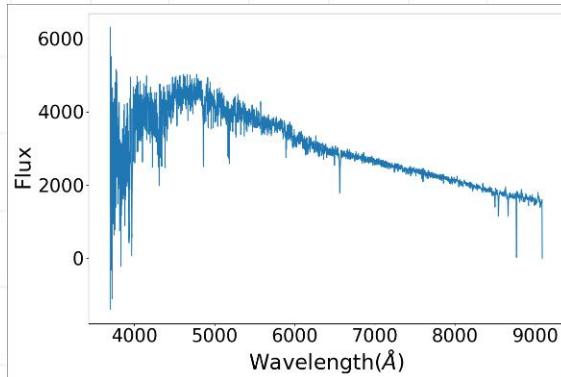
(b) $[\text{Fe}/\text{H}] = -2.99 \text{dex}$



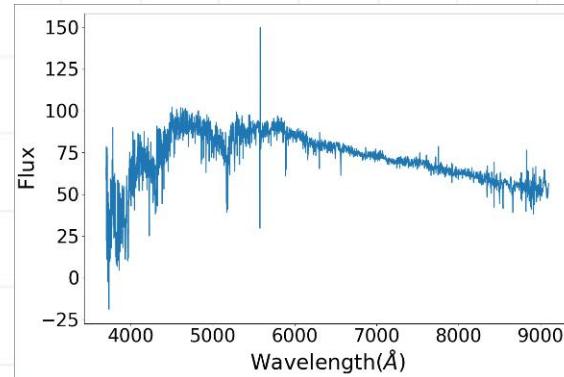
(c) $[\text{Fe}/\text{H}] = -3.33 \text{dex}$



(d) $[\text{Fe}/\text{H}] = 0.18 \text{dex}$



(e) $[\text{Fe}/\text{H}] = 0.14 \text{dex}$



(f) $[\text{Fe}/\text{H}] = -0.61 \text{dex}$

3.2 数据预处理

01

不同光谱的波长范围有可能是不同的，需要将光谱的波长规定在同一范围内。使用线性插值法将原始光谱数据插值到波长范围为4000-8095 Å以获得4096维光谱流量数据。

02

奇异样本数据的存在可能会引起训练时间增长，甚至导致模型无法收敛。因此，在进行训练之前需要对光谱数据进行归一化，来统一光谱流量的范围。

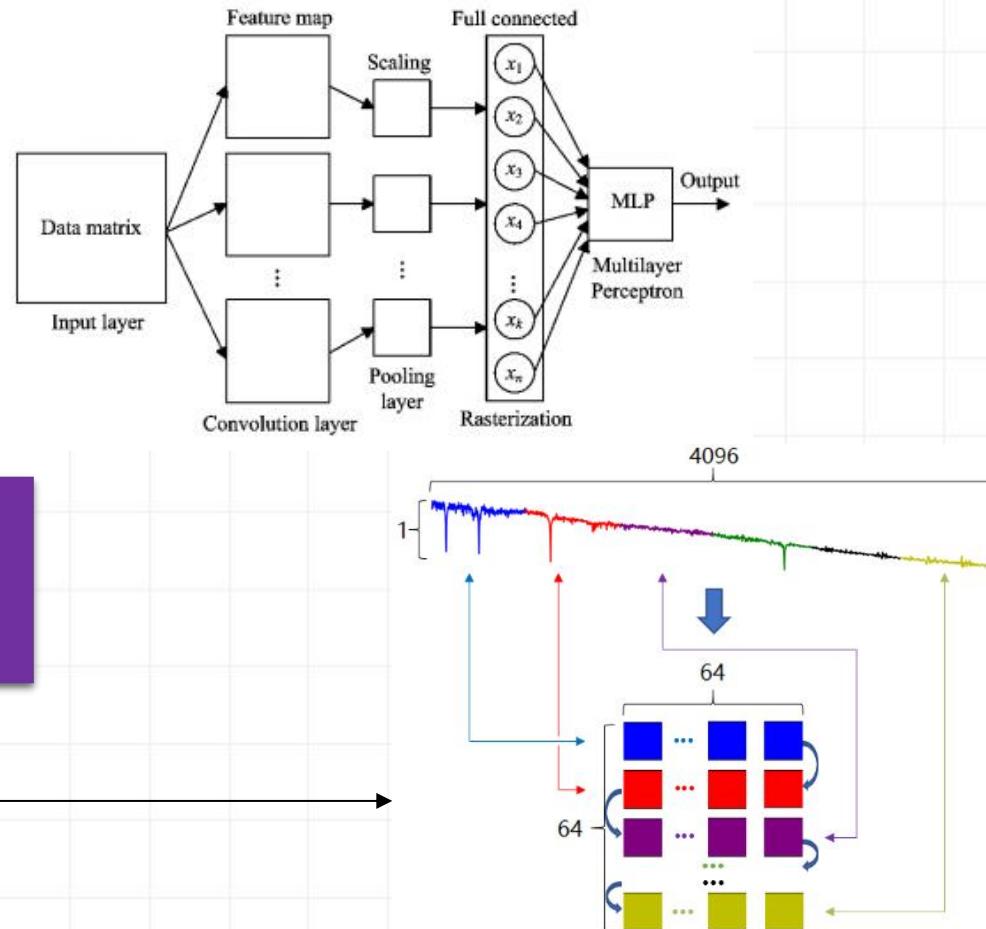
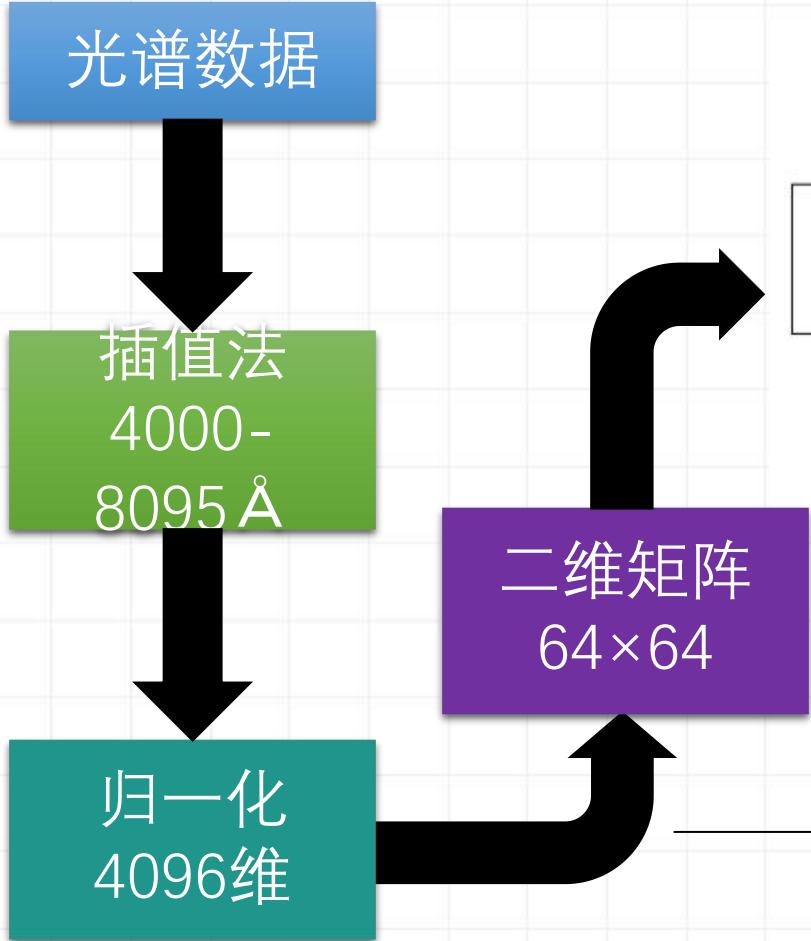
$$\hat{X}^i = \frac{X^i - X_{\min}^i}{X_{\max}^i - X_{\min}^i}$$

03

卷积神经网络对二维图像特征的提取能力非常强，因此我们将长度为4096的一维光谱数据通过“S”型折叠转化为 64×64 的二维光谱矩阵，让CNN像处理二维图像一样去处理光谱数据。

3.2

数据预处理



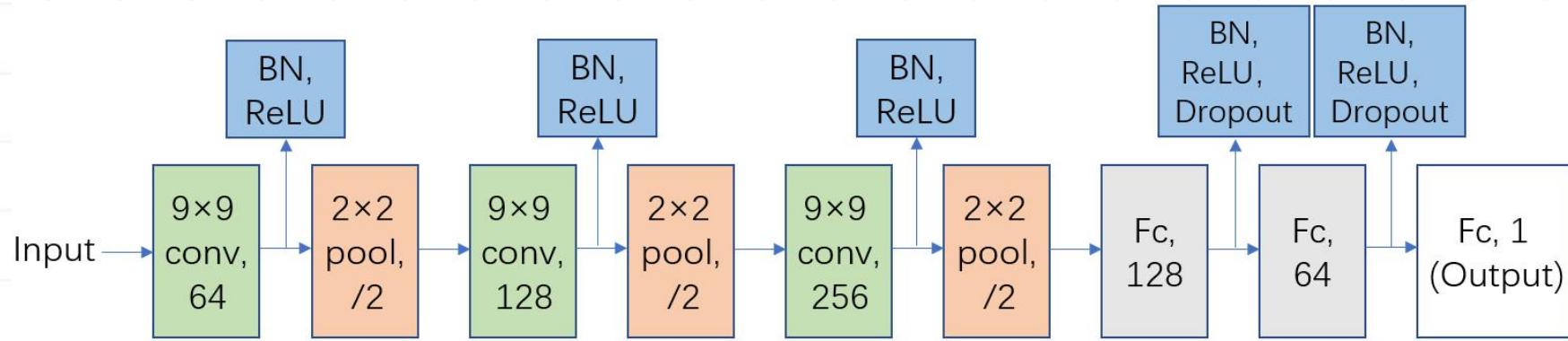


表1 CNN模型的网络结构

名称	超参数	层数
卷积层	9×9, 64	3
	9×9, 128	
	9×9, 256	
池化层	2×2	3
	2×2	
	2×2	
全连接层	128	2
	64	

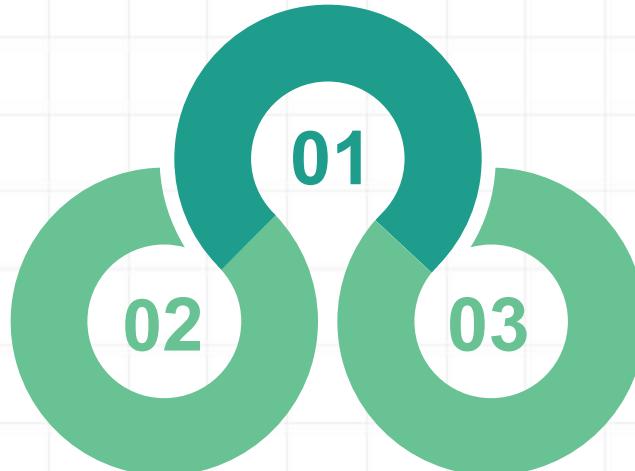
假设测试集的数量为 N , y_{test}^i 是测试集中第*i*个参数的目标值, Y_{test}^i 是CNN对应 y_{test}^i 的回归结果。我们在测试集上用以下方法来评价模型的预测性能。

平均绝对误差(MAE)可以避免误差相互抵消的问题, 因而可以准确反映实际预测误差的大小。将CNN模型的损失函数也设置为MAE。

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_{test}^i - Y_{test}^i|$$

$$SD = \sqrt{\frac{1}{N} \sum_{i=1}^N (E^i - \bar{E}^i)^2}$$

标准差(SD)可以反映CNN估计结果的离散程度。其中 $E^i = y_{test}^i - Y_{test}^i$, \bar{E}^i 是 E^i 的平均值。



E^i 的中位数可以很好的反应数据中的异常值。我们用M来表示 E^i 的中位数。

3.4 实验结果及分析

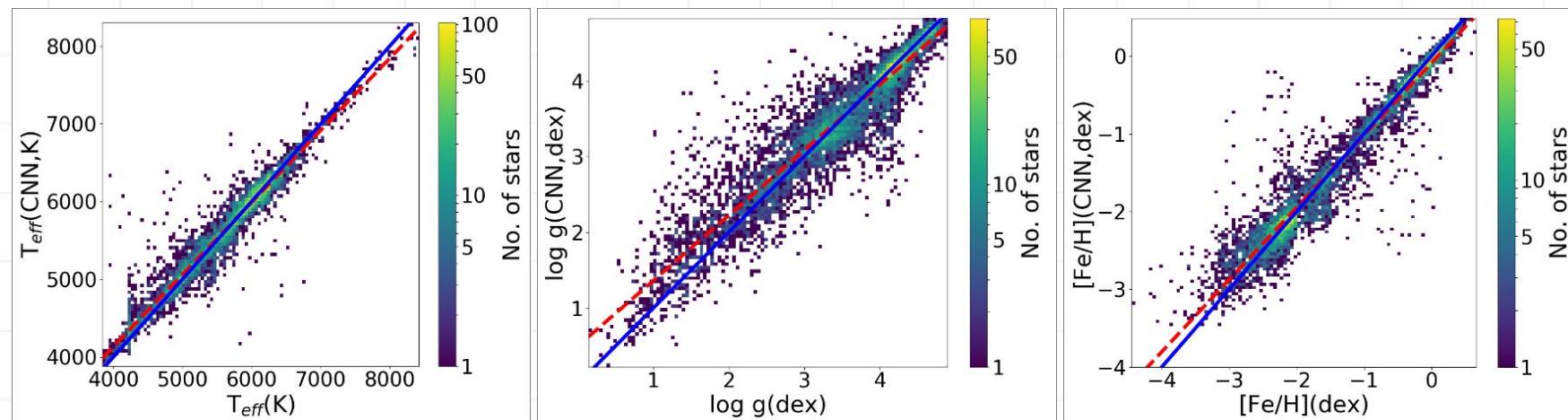
Teff、 $\log g$ 和[Fe/H]的估计

➤ 实验结果

本文使用二维卷积神经网络模型估计恒星大气参数，根据模型在验证集上的表现去调整网络构造与网络参数。按照7:3的比例将数据集随机划分为训练集(17500)和测试集(7500)

表1 CNN模型估计Teff、 $\log g$ 和[Fe/H]的结果

参数	MAE	SD	M
Teff(K)	93.03	159.86	-6.68
$\log g(\text{dex})$	0.21	0.33	-0.0018
[Fe/H](dex)	0.14	0.24	0.011



➤ 模型对比

为了表明CNN模型估计恒星大气参数的准确性和稳定性，本文同其它应用广泛的回归方法进行了比较。

- 支持向量回归(SVR)和随机森林回归(RFR)以及ANN
- 神经网络模型StarNet和数据驱动模型The Cannon。

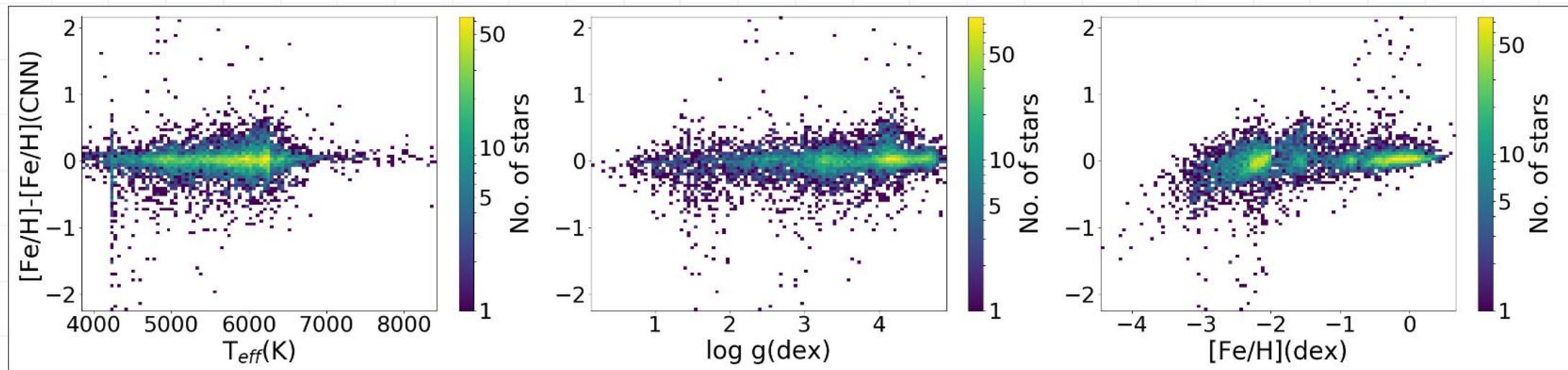
表2 CNN模型与其它模型的对比结果

方法	Teff(K)			log g(dex)			[Fe/H](dex)		
	MA	SD	M	MA	SD	M	MA	SD	M
SVR	172.69	269.15	-0.35	0.40	0.54	0.042	0.36	0.51	-0.019
RFR	146.20	229.52	9.44	0.35	0.51	0.045	0.30	0.45	0.021
ANN	165.99	247.18	48.76	0.27	0.43	0.00041	0.18	0.30	0.0027
StarNet	115.80	187.08	-8.61	0.25	0.39	-0.012	0.20	0.31	0.018
Cannon	151.80	291.03	1.01	0.33	0.58	0.030	0.23	0.45	0.050
CNN	93.03	159.86	-6.68	0.21	0.33	-0.0019	0.14	0.24	0.011

3.5 实验结果及分析

VMP恒星的识别

- 在估计出恒星的金属丰度之后，可以利用金属丰度来识别VMP恒星。
- 在测试集中包含3026条VMP恒星光谱和4474条正常恒星光谱。
- 由于VMP星光谱金属丰度低，光谱特征弱，在参数估计中没有正常恒星估计精度高。本文中的CNN模型也是如此，模型会把较低的金属丰度估计的比较高，把比较高的金属丰度估计的比较低。从下图最右侧子图中也可以证实这个结论。



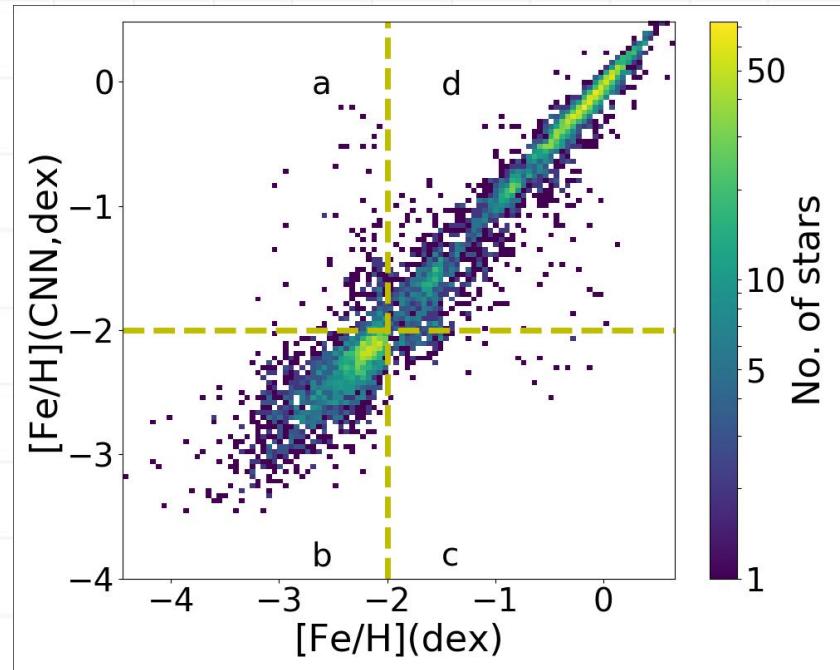
CNN模型在VMP星与正常恒星上的误差比较

表3 CNN模型在VMP恒星和正常恒星上的结果

参数	MAE		SD		M	
	VMP星	正常星	VMP星	正常星	VMP星	正常星
Teff	118.91	75.53	179.05	143.82	-19.05	-2.74
log g	0.30	0.15	0.41	0.25	-0.041	-0.0053
[Fe/H]	0.18	0.11	0.27	0.19	0.043	0.026

3.5 实验结果及分析

VMP恒星的识别



- CNN将167颗正常恒星错误地识别为VMP恒星，错误率为3.73%。
- 将2761颗VMP恒星正确识别，正确率为91.24%。

由此可见，CNN模型不仅可以准确估计恒星物理参数，还可以从大量恒星光谱中搜寻出VMP恒星，并且识别正确率非常高。

表4 CNN型在不同区域对VMP恒星和正常恒星的识别

[Fe/H] (CNN,dex)	[Fe/H](dex)	
	[Fe/H]<-2.0	[Fe/H]>-2.0
[Fe/H]<-2.0	2761	167
[Fe/H]>-2.0	265	4307

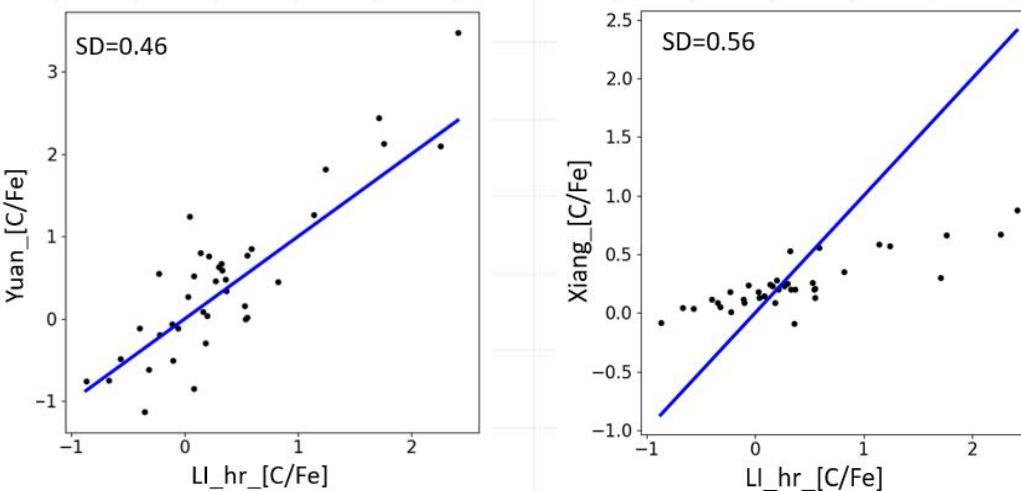
➤ [C/Fe]标签来源

原始的10008条VMP恒星光谱中并没有提供全部恒星的[C/Fe]，只提供了57条高分辨率光谱的[C/Fe]。Yuan提供了一个带有[C/Fe]参数的星表(catalog)，星表中恒星参数是使用SSPP (SEGUE Stellar Parameter Pipeline) 进行测量的。Xiang也提供了一个包含800万条恒星光谱的星表，里面含有部分[C/Fe]标签。

1 使用原始的10008条VMP恒星光谱先与SSPP提供的星表进行交叉得到8117条光谱及其碳丰度。

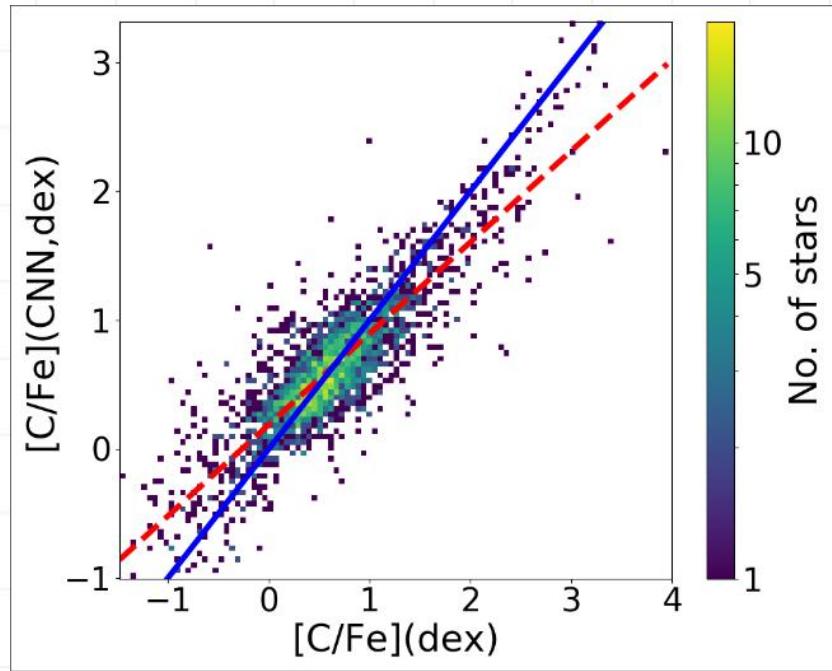
2 将8117条光谱与Xiang提供的星表交叉，得到两者公共的恒星光谱及其对应的[C/Fe]标签。

3 将两者的[C/Fe]标签分别与高分辨率光谱的[C/Fe]进行比较。

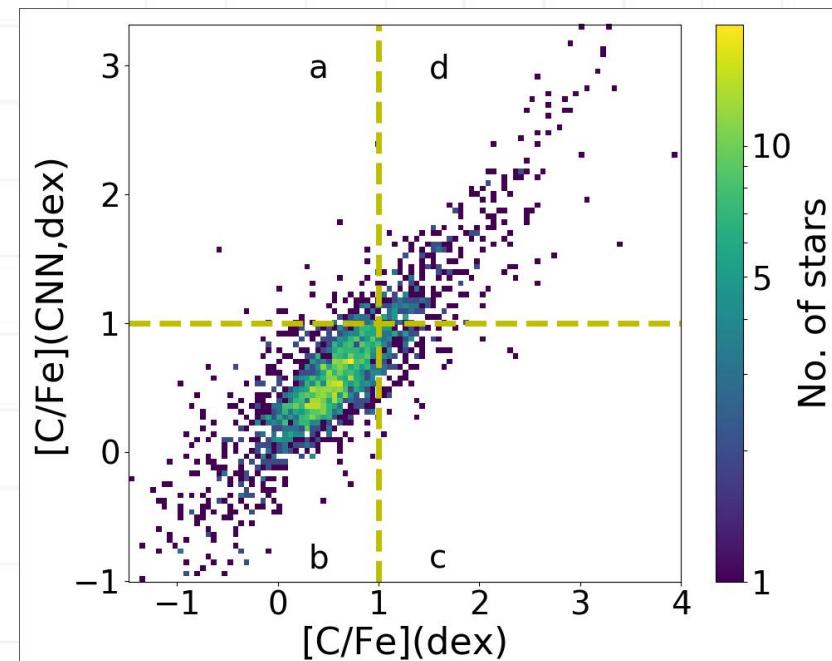


可以看出Yuan提供的[C/Fe]标签与李海宁提供的高分辨率光谱的[C/Fe]相关性更好，本文最终也选择使用Yuan提供的经过SSPP测量的[C/Fe]标签。

➤ 实验结果



MAE=0.22dex, SD=0.30dex, M=-0.0028dex



CEMP恒星的识别

在VMP恒星中，许多恒星的碳丰度[C/Fe]的值很大，甚至比太阳大几个数量级，这一大类 $[C/Fe] > 1.0 \text{dex}$ 的恒星，称作碳增丰的贫金属(Carbon-Enhanced Metal-Poor, CEMP)星。

CNN模型在不同区域对CEMP恒星和正常VMP恒星的识别

[C/Fe]	[C/Fe]<1.0	[C/Fe]>1.0
[C/Fe] (CNN)	1878	164
	83	311

3.5 实验结果及分析

碳丰度[C/Fe]的估计

➤ CNN估计[C/Fe]选取的特征

在天文研究中，天文学家不仅关心提取的碳元素丰度，还关心哪些波段或者谱线与碳元素丰度具有较高的相关性。了解与碳元素丰度相关的特征将有助于天文学家更好地研究碳元素丰度的变化规律，也有助于后续高分辨率光谱的研究。

本研究用恒星光谱CH G带的流量特征进行了训练测试。具体地，使用对应G波段的波长范围为4000-5500 Å 的光谱流量，同时将光谱其它波段对应的流量特征作为模型的输入值全部设置为0。

表6 CNN模型分别在CH G带和全光谱上估计[C/Fe]的结果

参数	输入	MAE	SD	M
[C/Fe]	CH G-band	0.23	0.33	0.0070
	Full spectra	0.22	0.30	- 0.0028

3.5 实验结果及分析

Li_Teff	CNN_Teff	Li_log g	CNN_log g	Li_[Fe/H]	CNN_[Fe/H]	Li_[C/Fe]_hr	SSPP_[C/Fe]	CNN_[C/Fe]
E	E	E	E	E	E	E	E	E
6.250000E+003	6.254957E+003	3.520000E+000	3.513815E+000	-2.320000E+000	-2.292852E+000	NULL	1.311000E+000	1.182249E+000
6.000000E+003	6.021433E+003	3.240000E+000	3.736258E+000	-2.280000E+000	-2.112926E+000	NULL	7.430000E-001	5.889318E-001
5.933400E+003	5.961624E+003	3.880000E+000	3.844056E+000	-2.120000E+000	-2.155648E+000	NULL	NULL	NULL
5.947000E+003	5.946590E+003	4.090000E+000	4.082838E+000	-2.250000E+000	-2.164052E+000	NULL	6.000000E-002	2.387209E-001
5.872300E+003	5.867523E+003	4.090000E+000	4.099563E+000	-2.180000E+000	-2.207411E+000	NULL	7.830000E-001	7.620608E-001
5.987800E+003	6.013337E+003	3.180000E+000	3.193212E+000	-2.120000E+000	-2.078598E+000	NULL	6.690000E-001	5.920415E-001
6.000000E+003	6.042685E+003	3.050000E+000	3.230338E+000	-2.630000E+000	-2.490194E+000	NULL	8.010000E-001	6.825827E-001
6.309900E+003	6.209735E+003	3.410000E+000	3.500566E+000	-2.240000E+000	-2.363336E+000	NULL	1.390000E-001	2.244511E-001

CNN_Teff	CNN_log g	CNN_[Fe/H]
E	E	E
5.395661E+003	3.026678E+000	-2.027941E+000
6.806663E+003	2.988560E+000	-2.023829E+000
5.563365E+003	3.488029E+000	-2.240676E+000
5.131706E+003	3.543460E+000	-2.006736E+000
7.001592E+003	3.651490E+000	-2.126982E+000
7.638874E+003	4.092719E+000	-2.618127E+000
7.606139E+003	3.689544E+000	-2.537038E+000
7.475098E+003	3.968615E+000	-2.325271E+000

本研究已将CNN模型对10008条VMP恒星光谱参数的估计结果和从LAMOST DR7中搜寻到的32937颗VMP恒星候选体分别做成了fits文件放到了GitHub上，感兴趣的读者可以访问<https://github.com/xiejhhh/VMP>。

3.6 对空间恒星光谱的预研究 MARCS理论光谱仿真空间光谱

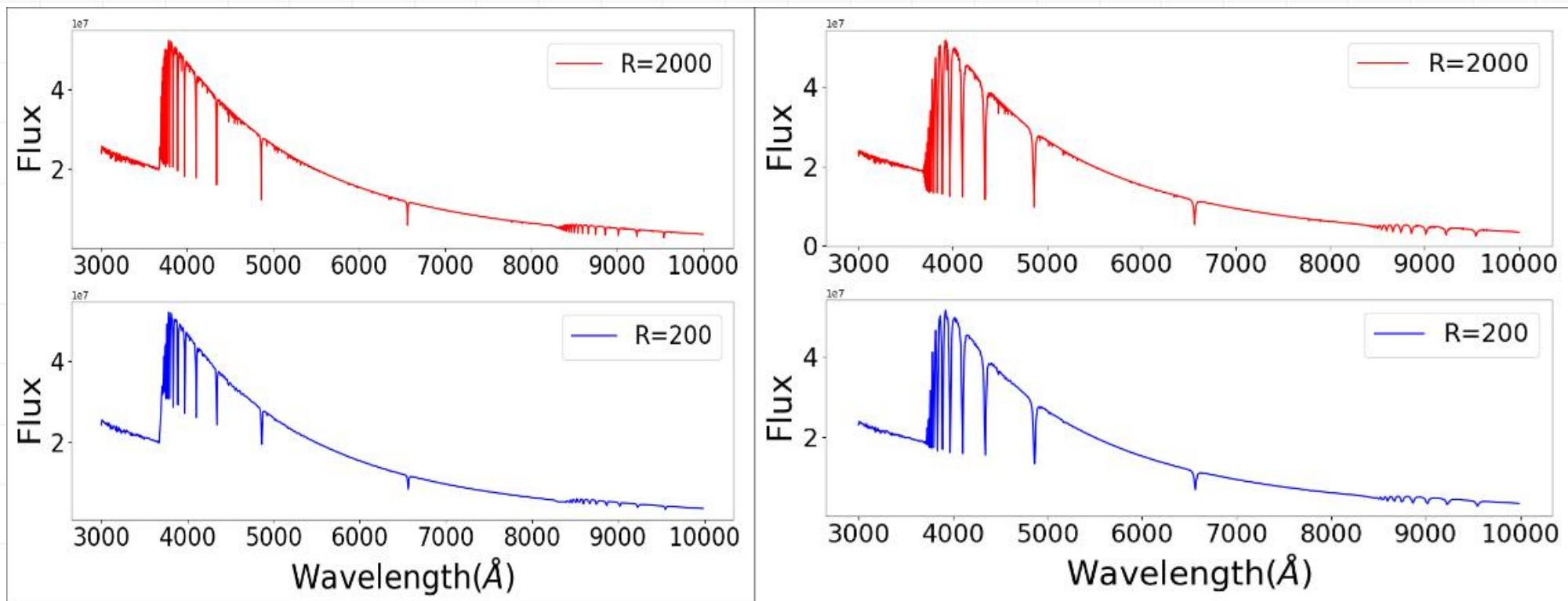
➤ MARCS理论光谱

理论光谱具有更明显的特征和更少的噪声，可以做到更宽的波长覆盖区间和更均匀的参数覆盖范围。

MARCS提供了R=20000和R=2000两种分辨率的理论光谱，每条光谱的波长范围为3000-10000 Å，共有9644条光谱。

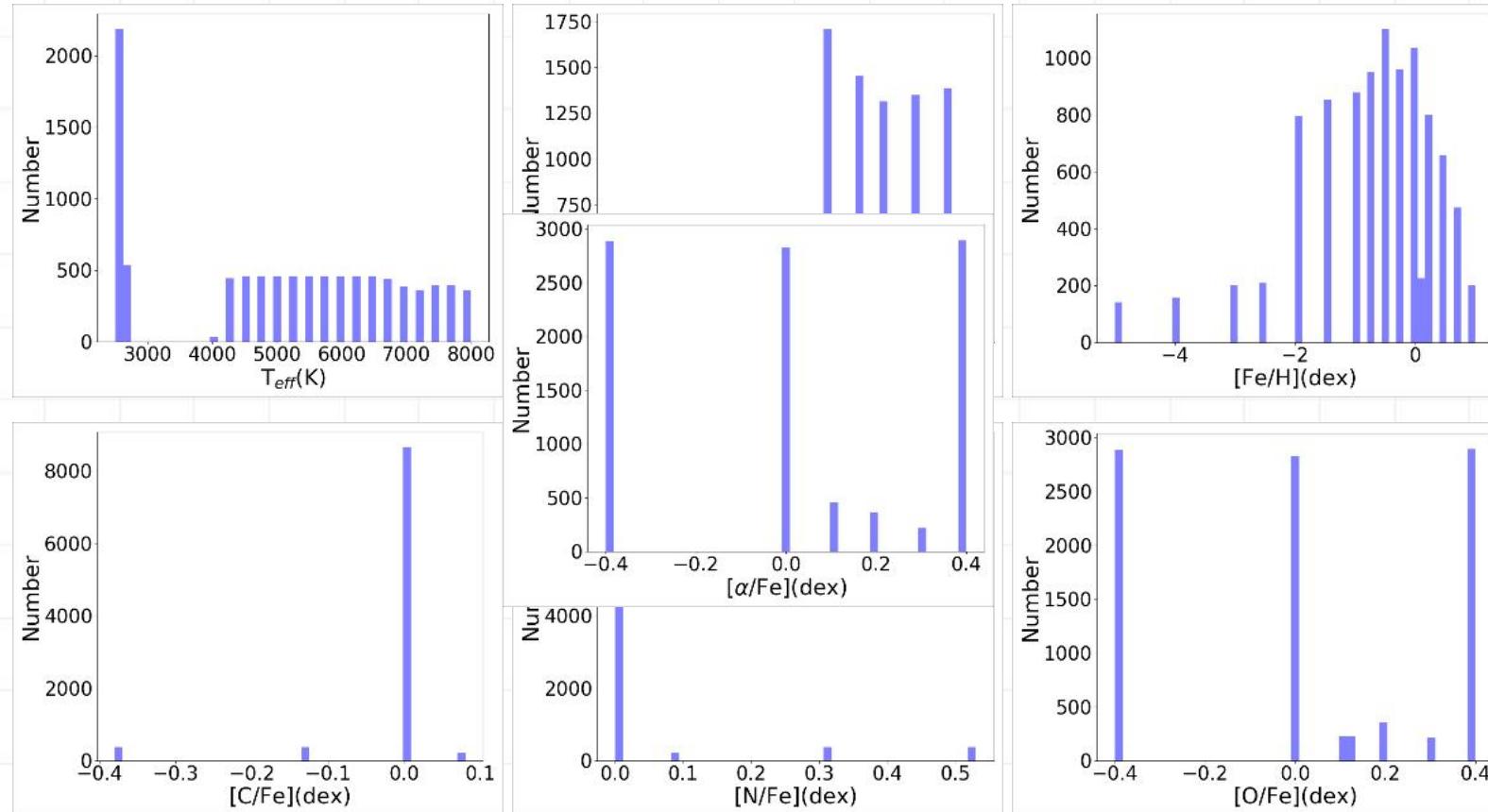
表8 MARCS模板库中光谱的Teff、 $\log g$ 和[Fe/H]的范围及步长

参数	范围	步长
Teff	2500-4000K	100K
	4000-8000K	250K
$\log g$	-0.5-5.5dex	0.5dex
[Fe/H]	-5-1.0dex	0.25dex



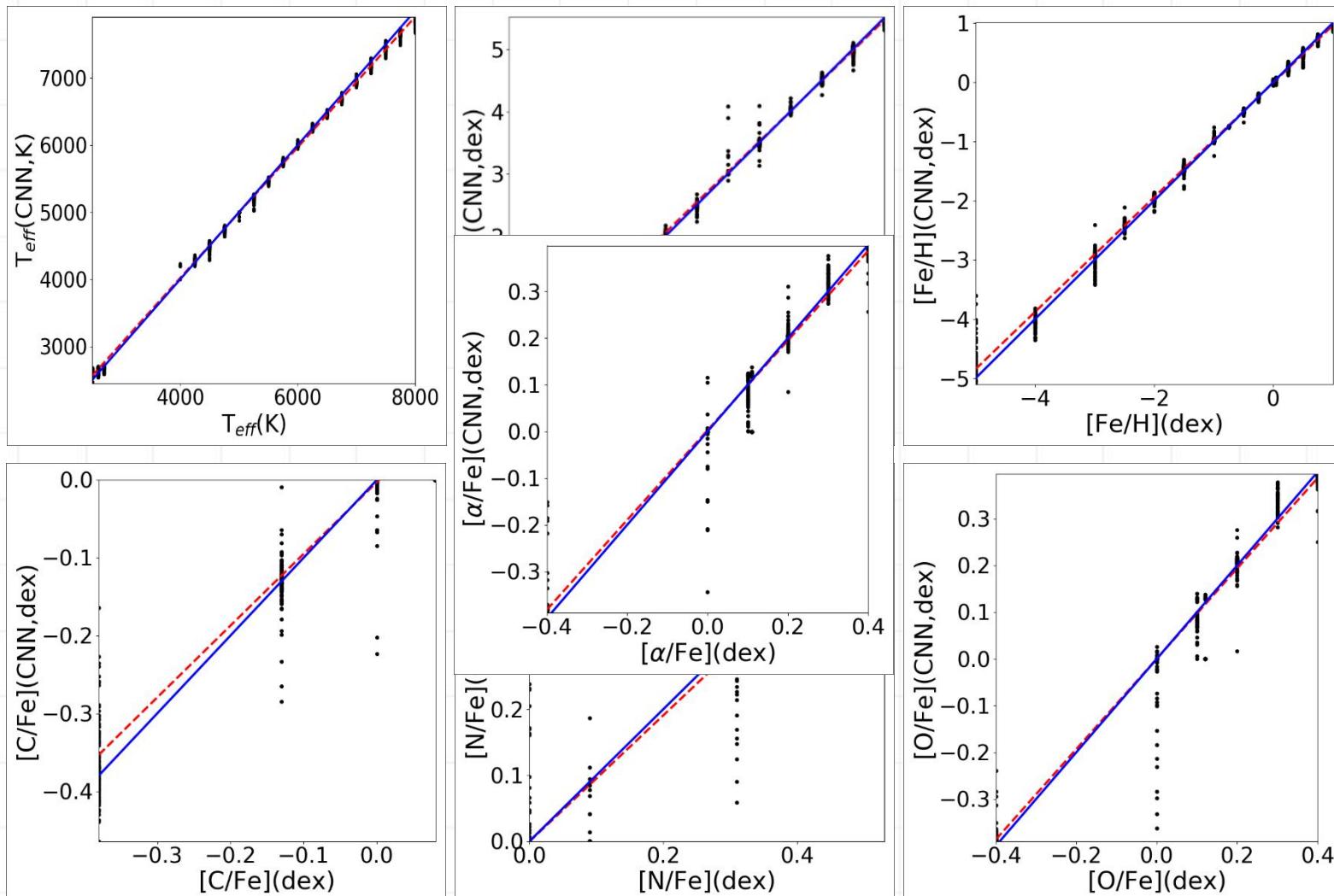
3.6 对低分辨率恒星光谱的研究 MARCS理论光谱仿真空间光谱

➤ 参数分布



3.6 对低分辨率恒星光谱的研究 MARCS理论光谱仿真空间光谱

➤ 实验结果



3.6 对低分辨率恒星光谱的研究 MARCS理论光谱仿真空间光谱

➤ 实验结果

表9 CNN模型估计Teff、 $\log g$ 和[Fe/H]的结果

参数	MAE	SD	M
Teff(K)	52.05	69.45	7.32
$\log(g)$ (dex)	0.041	0.070	-0.0072
[Fe/H](dex)	0.038	0.087	-0.0056
[C/Fe](dex)	0.0043	0.018	0.000010
[N/Fe](dex)	0.0049	0.023	0.000033
[O/Fe](dex)	0.13	0.026	0.00011
[α /Fe](dex)	0.014	0.025	0.00068

实验结果表明，模型利用仿真光谱可以精确估计Teff、 $\log g$ 、[Fe/H]、[C/Fe]、[N/Fe]、[O/Fe]和[α /Fe]等恒星参数。

四 总结与展望

总结



展望

- 提出了一种基于残差神经网络算法估计低分辨率恒星光谱化学元素丰度的算法模型。
- 构建了卷积神经网络模型估计贫金属恒星的大气参数，并且对VMP恒星的识别正确率达到91.24%。
- 验证了CNN模型在低分辨恒星仿真光谱上的表现，证实该模型可以用于VMP恒星的参数估计与搜寻
- 空间望远镜光谱数据是使用现有的光谱数据进行仿真的，未来要寻找真实的空间望远镜观测光谱来训练模型，从而进一步提高模型的性能。
- 使用模型筛选出32937颗极贫金属(VMP)星候选体需要后续的证认。
- 研究光谱加入紫外波段后的效果。同时，将算法融入CSST数据处理pipeline



谢谢大家

Any question?