





Template AASTeXv7 Article with Examples*

TUNDRA NORTH AMERICA ^{1,†} FORREST SUR AMÉRICA ^{2,‡} S. AFRICA,^{3,4} RIVER EUROPE,⁵ ASIA MOUNTAIN ^{6,§}
CORAL AUSTRALIA ⁷ AND PENGUIN ANTARCTICA⁸

THE TERRA MATER COLLABORATION

¹ *University of Saskatchewan*

² *Universidad de Chile, Department of Astronomy*

³ *South African Astronomical Observatory*

⁴ *University of Cape Town, Department of Astronomy*

⁵ *University of Heidelberg*

⁶ *Tata Institute of Fundamental Research, Department of Astronomy*

⁷ *James Cook University, Department of Physics*

⁸ *Amundsen–Scott South Pole Station*

摘要

This example manuscript is intended to serve as a tutorial and template for authors to use when writing their own AAS Journal articles. The manuscript includes a history of AASTeX and documents the new features in the previous versions as well as the new features in version 7. This manuscript includes many figure and table examples to illustrate these new features. Information on features not explicitly mentioned in the article can be viewed in the manuscript comments or more extensive online documentation. Authors are welcome replace the text, tables, figures, and bibliography with their own and submit the resulting manuscript to the AAS Journals peer review system. The first lesson in the tutorial is to remind authors that the AAS Journals, the Astrophysical Journal (ApJ), the Astrophysical Journal Letters (ApJL), the Astronomical Journal (AJ), and the Planetary Science Journal (PSJ) all have a 250 word limit for the abstract. The limit is 150 for RNAAS manuscripts. If you exceed this length the Editorial office will ask you to shorten it. This abstract has 189 words.

Keywords: Galaxies (573) — Cosmology (343) — High Energy astrophysics (739) — Interstellar medium (847) — Stellar astronomy (1583) — Solar physics (1476)

1. INTRODUCTION

CEMP (Carbon-Enhanced Metal-Poor) 恒星是一类贫金属、富碳的老年恒星，其金属丰度 ([Fe/H]) 通常低于太阳值的千分之一，而碳丰度却显著增强。它们被认为是研究宇宙早期化学演化和恒星形成过程的关键天

Email: fakeemail1@google.com

* Footnotes can be added to titles

† Kitt Peak National Observatory

‡ Las Campanas Observatory

§ Astrosat Post-Doctoral Fellow

体。根据碳 enrich 水平和重元素特征, CEMP 恒星被进一步划分为多个子类, 如 CEMP-s (s-process 元素丰富)、CEMP-r (r-process 元素丰富) 和 CEMP-no (氮和氧也贫化)。这些子类反映了不同的起源机制。

CEMP-s 恒星约占 CEMP 恒星的 80%, 多位于双星系统中, 其碳和 s-process 元素 enrichment 可能源自伴星——一颗已经演化的 AGB (渐近巨星支) 恒星通过质量转移污染了原初气体 [abia2002]。这一模型得到了观测到的双星轨道周期和自转速度的支持。

CEMP-no 恒星则更为古老, 可能形成于首批恒星 (Population III) 爆炸后的遗迹气体中。这些气体富含碳但缺乏 heavier metals, 因此形成的第二代恒星继承了这种成分 [christlieb2002]。例如, HE 1327-2326 是已知最 metal-poor 的 CEMP-no 恒星之一, 其大气参数和化学组成为此假说提供了直接证据 [ito2009]。

CEMP-r 恒星含有较高的 r-process 元素 (如铕 Eu), 这表明它们的形成环境受到了中子星并合或磁驱动超新星等事件的影响 [hansen2011]。然而, 这类恒星较为罕见, 其确切起源仍需更多观测数据验证。

CEMP-no 恒星多见于极低 $[\text{Fe}/\text{H}]$ (< -3), 如 HE 1327-2326 ($[\text{Fe}/\text{H}] = -5.7$, [ito2009]) 和 SMSS J0313-6708 ($[\text{Fe}/\text{H}] < -7.1$, [keller2014]), 这表明它们可能形成于首批恒星 (Population III) 超新星爆炸后富含碳但贫 metal 的气体中。相比之下, CEMP-s 恒星的 $[\text{Fe}/\text{H}]$ 相对较高 (-2.5 至 -3.0), 这可能与双星系统中 AGB 伴星的质量转移有关, 该过程需足够的时间积累, 故对应稍高的 metallicity ([lucatello2005])。

CEMP-no 恒星的 $[\text{C}/\text{Fe}]$ 通常超过 $+1.0$, 且在极低 $[\text{Fe}/\text{H}]$ 下保持高位, 暗示其碳源可能来自高能爆发现象 (如 faint supernovae)。而 CEMP-s 恒星的 $[\text{C}/\text{Fe}]$ 虽同样高, 但常伴随 s-process 元素 (如 Ba), 其形成与 AGB 星的慢中子俘获过程相关, 且在 $[\text{Fe}/\text{H}] > -3$ 时更为常见 ([abia2002], [bisterzo2011])。

CEMP-no 恒星常表现为低 $[\text{N}/\text{Fe}]$ 和高 $[\text{C}/\text{N}]$, 与 Population III 恒星的混合模型一致 ([norris2013])。CEMP-s 恒星则多有 $[\text{Ba}/\text{Fe}] > +1.0$, 且 $[\text{Ba}/\text{Eu}]$ 接近太阳能值, 指示 s-process 主导 ([masseron2010])。少数 CEMP-r 恒星同时具有 r-process 元素 (如 Eu) enrichment, 可能由中子星并合或磁驱动超新星引起 ([hansen2011])。

观测数据显示, 当 $[\text{Fe}/\text{H}] < -3$ 时, CEMP 恒星比例显著上升 (约 20%), 其中 CEMP-no 占主导 ([yoon2016])。这一趋势符合理论预测: 早期宇宙中低 metallicity 气体更易在碳增强条件下形成低质量恒星 ([frebel2015])。

CEMP 恒星的研究不仅有助于理解早期宇宙的化学增丰历史, 还为恒星形成和核合成理论提供了宝贵约束。未来高分辨率光谱观测和大样本统计将进一步揭示各类 CEMP 恒星的形成机制及其宇宙学意义。

我们对恒星的所有认知, 从化学成分到物理状态, 几乎都源于对其光谱的精细解读。理论上, 恒星光谱由两部分组成: 一个由其炽热表面发出的、平滑的 **连续谱** 作为背景; 以及叠加其上的、由恒星外层大气中各种元素吸收特定波长光子而形成的 **吸收线**。这些吸收线如同一张独特的 “条形码”, 其位置、深度和形状精确地揭示了恒星的化学丰度 (如 $[\text{Fe}/\text{H}]$ 和 $[\text{C}/\text{Fe}]$)、温度、压力乃至自转速度。然而, 从观测数据中准确提取这些微弱的物理信号并非易事, 需要仔细区分真实的吸收线与宇宙线、探测器噪声等干扰, 这是恒星光谱分析中的关键挑战。

CEMP (Carbon-Enhanced Metal-Poor) 恒星作为宇宙早期化学演化的关键探针, 其发现与研究离不开大型天文巡天项目的推动。LAMOST (郭守敬望远镜) 凭借其高效的光纤光谱观测能力, 已成为发现和表征 CEMP 恒星的重要平台。自运行以来, LAMOST 已释放数百万条恒星光谱数据, 极大拓展了 CEMP 恒星样本量。例如, Li et al. (2015) [li2015] 利用 LAMOST 数据鉴定出逾千颗 CEMP 候选星, 显著提升了该类天体的统计基础。此外, 结合 Gaia 的距离和自行数据及 SDSS 的测光信息, 研究人员得以进一步解析 CEMP 恒星的大气参数和 kinematic 特性, 深入探讨其起源机制 (Yuan et al., 2020) [yuan2020]。未来, LAMOST 将持续助力 CEMP 恒星的系统性研究, 为理解首批恒星核合成及银河系形成提供宝贵线索。

2. 相关工作和方法

在利用大规模巡天数据搜寻如 CEMP 星等特殊恒星时，模板匹配法是一种被广泛应用的经典技术。其核心思想是通过将观测光谱与一个覆盖了广阔参数空间 (T_{eff} , $\log g$, $[\text{Fe}/\text{H}]$, $[\text{C}/\text{H}]$ 等) 的参考光谱库进行比对，通过寻找最佳匹配来推断前者的物理参数。针对 LAMOST 巡天数据，该方法的具体实施包含以下三个关键环节：

2.1. 模板匹配法

2.1.1. 参考光谱库的构建

模板库的完备性与精度是决定该方法成败的基石。一个为 CEMP 星搜索优化的模板库需满足以下要求：

- **理论光谱的生成：**使用如 MOOG 等合成光谱代码，结合 MARCS 等先进的恒星大气模型，生成覆盖 CEMP 星典型参数范围的理论光谱。在生成过程中，必须精细考虑关键分子带（如 CH G-band, C_2 Swan bands）的贡献，并模拟 LAMOST 的仪器展宽和星际消光效应。
- **参数空间的覆盖：**模板库必须在 T_{eff} (4000-6500 K), $\log g$ (1.0-4.5 dex), $[\text{Fe}/\text{H}]$ (-4.0 to 0.0 dex), 和 $[\text{C}/\text{Fe}]$ (0.0 to +3.0 dex) 等维度上进行精细网格化采样，以确保对各类恒星都有精确的模板对应。
- **经验光谱的补充：**为修正理论模型在极低金属丰度区域可能存在的系统偏差，通常会引入来自高分辨率观测（如 SDSS/SEGUE, HET/HRS）的真实 CEMP 星光谱作为经验模板，以提升模板库的可靠性。

2.1.2. 光谱匹配与参数优化

此环节的目标是在模板库中为每一条观测光谱找到最佳匹配。对于经过归一化等预处理的 LAMOST 低分辨率光谱，该过程通常通过 χ^2 最小化来实现。具体而言，即求解一个非线性优化问题，其目标函数如下：

$$\min_{\theta, a, v_{\text{rad}}} \sum_i \left[\frac{f_{\text{obs}}(\lambda_i) - a \cdot T(\lambda_i(1 + v_{\text{rad}}/c); \theta)}{\sigma_i} \right]^2$$
 其中， f_{obs} 是观测流量， $T(\lambda; \theta)$ 是参数为 θ 的模板光谱， a 和 v_{rad} 分别是归一化尺度因子和视向速度， σ_i 是噪声。此优化过程通常采用 Levenberg-Marquardt 等高效算法并行求解。参数的最终误差则常通过 MCMC 等贝叶斯方法进行估计。

2.1.3. 结果的后处理与验证

为确保结果的可靠性，必须对初步匹配结果进行严格的后处理与验证：

- **关键谱线检验：**对所有候选体，计算关键的碳分子带（如 CH G-band）的等值宽度或线指数，并设置信噪比阈值，以剔除因噪声导致的伪信号。
- **人工目视检查：**对所有高置信度的候选体进行人工检查，以排除宇宙射线、光纤交叉污染等仪器效应的干扰。
- **外部交叉验证：**将得到的结果与来自其他高分辨率巡天（如 APOGEE）的参数测量结果进行比对，以评估系统误差和测量精度。

尽管模板匹配法在处理低信噪比数据时表现稳健，并能同时给出多个恒星参数，但其性能高度依赖于模板库的完备性，且在面对千万量级的巡天数据时，其计算成本高昂，难以满足自动化、高效率分析的需求。

随着 LAMOST 等大规模光谱巡天项目的深入，数据量呈爆炸式增长，传统的模板匹配法在效率和可扩展性上逐渐暴露瓶颈。与此同时，以支持向量机、随机森林为代表的传统机器学习方法，虽在一定程度上提升了自动化水平，但其性能高度依赖于人工设计的特征（如等值宽度、颜色指数等），在面对高维、非线性的光谱数据时，其建模能力和泛化性均受到限制。

近年来，数据驱动的深度学习方式为恒星光谱分析带来了范式上的革新。这类方法能够直接从原始光谱数据中自动学习层次化的特征表示，极大地降低了对特征工程的依赖，并在参数估计的精度和效率上取得了突破。在恒星参数估计领域，已有多种深度学习模型被成功应用：

2.1.4. *The Cannon*: 数据驱动的光谱建模基石

由 Ness 等人提出的 **The Cannon** [ness2015cannon] 是该领域的里程碑式工作。它并非一个物理模型，而是一个数据驱动的生成模型。其核心思想是：在存在一个高精度“训练集”（即一组同时拥有光谱和精确参数的恒星）的前提下，可以学习到一个普适的函数，该函数将任意恒星的光谱流量与它的物理参数（“标签”）精确地联系起来。*The Cannon* 通常使用一个简单的多项式模型来拟合这种关系。在训练阶段，模型从参考样本中学习光谱流量与恒星标签之间的映射关系；在应用阶段，对于只有光谱的未知天体，模型能够以极高的速度反解出其最可能的标签。该方法以其高效、稳健以及在低信噪比数据上的优异表现，被广泛应用于 APOGEE 等大型巡天项目中。

2.1.5. *SPCANet*: 融合主成分分析的混合模型

SPCANet [wang2020spcanet] 是一种结合了主成分分析（PCA）与深度神经网络（DNN）的混合模型。它首先利用 PCA 对高维光谱数据进行降维，提取出最主要的成分作为特征，此举能有效去除噪声并减少信息冗余。随后，这些低维特征被送入一个深度神经网络，通过非线性回归建模，学习从光谱主成分到恒星物理参数的复杂映射。SPCANet 结合了 PCA 的线性和 DNN 的非线性建模能力，在处理低信噪比光谱时表现出很强的鲁棒性。

2.1.6. *SLAM*: 端到端的卷积神经网络方法

SLAM (Stellar LAbel Machine) [zhang2020slam] 是专为大规模光谱巡天设计的端到端深度学习模型。它采用卷积神经网络（CNN）架构，能够直接从原始的一维光谱像素中自动学习相关特征，无需任何人工干预。通过多层卷积与池化操作，SLAM 能够有效捕捉光谱中的局部和全局结构信息（如吸收线的轮廓、相对强度等），并最终通过全连接层回归得到恒星参数。SLAM 的端到端特性使其具备极高的自动化处理能力和对数据噪声、采样不均等问题的鲁棒性，是处理海量光谱数据的理想选择。

2.1.7. *SpecCLIP*: 基于对比学习的光谱基础模型

近年来，天文学界开始借鉴自然语言处理和计算机视觉领域的成功经验，发展面向海量多模态数据的“基础模型”（Foundation Model）。**SpecCLIP** [li2023, zhang2024] 正是这一趋势下的前沿代表，它将恒星光谱视为一种结构化的“语言”，旨在学习到一种能够泛化至多种下游任务的普适性光谱表示。SpecCLIP 的核心思想源于强大的‘CLIP’（Contrastive Language-Image Pre-training）框架。它创新性地采用 **对比学习** 策略，将来自不同巡天项目、具有不同特征（如分辨率、波长覆盖）的光谱数据进行对齐。例如，将 LAMOST 的低分辨率光谱与 Gaia BP/RP (XP) 的光谱在同一个高维“嵌入空间”（Embedding Space）中进行匹配。其目标是让来自同一颗恒星的不同光谱在该空间中的表示尽可能接近，而来自不同恒星的光谱则相互远离。这种自监督的学习方式使得模型能够从海量的无标签数据中学习光谱的内在物理结构。作为一个多模态模型，SpecCLIP 能够自然地融合光谱信息与 Gaia 提供的高精度天体测量数据（如视差、自行）。这种跨模态学习带来了显著优势：一方面，精确的距离信息可以有效打破恒星参数（如 T_{eff} 和 $\log g$ ）之间的简并性，从而大幅提升参数估计的精度；另一方面，统一的嵌入空间使得在不同巡天数据之间进行交叉定标和信息迁移成为可能。SpecCLIP 的应用远不止于高精度的参数估计。其强大的表示学习能力使其在 **异常光谱检测**、**相似天体搜寻** 以及 **跨模态光谱生成**（例如，从 Gaia XP 光谱预测 LAMOST 光谱）等任务中也展现出巨大潜力，为天文学研究提供了全新的探索工具。

2.2. 研究动机与方法革新

下一代光谱巡天引发的数据挑战与科学机遇。以 LAMOST 和 SDSS-V 为代表的光谱巡天项目，正将天文学带入一个前所未有的“太字节” (Petabyte) 时代。其产生的数据量已逾千万量级，并仍在持续增长。这一数据的“洪流”为在银河系中系统性地搜寻像 CEMP 星这样的稀有天体提供了空前的机遇，但同时也对传统分析方法构成了严峻挑战。如前所述，经典的模板匹配法（第??节）在这一新形势下面临着三个难以逾越的瓶颈：

- **计算可扩展性瓶颈：**模板匹配本质上是基于 χ^2 最小化的网格搜索，其计算成本随参数空间维度（如考虑 C, N, α 等多种元素丰度）呈指数级增长，无法满足海量数据的实时处理需求。
- **低信噪比下的稳健性问题：**对于 LAMOST 产生的大量低信噪比 ($S/N < 15$) 光谱，关键的吸收线特征（如 CH 分子带）极易被噪声淹没，导致模板匹配法的参数测定，特别是碳丰度的误差急剧增大，可靠性降低。
- **模板库的不完备性与泛化难题：**任何理论或经验模板库都难以完全覆盖所有可能的恒星参数，尤其是在极端物理区域（如 $[Fe/H] < -3.5$ 或 $[C/Fe] > +2.0$ ）。这使得模板匹配法在搜寻最稀有、最极端的 CEMP 星时，其探测效率和可靠性都受到根本性限制。

深度学习：应对挑战的系统性解决方案。近年来，深度学习的迅猛发展为突破上述瓶颈提供了全新的、系统性的解决思路。它不再依赖于固定的模板库，而是直接从数据中学习光谱与物理参数之间的复杂映射关系。具体而言：

- **应对可扩展性瓶颈：**以卷积神经网络 (CNN) 为代表的端到端模型 [ting2019]，能够直接从原始光谱像素中自动提取特征并进行回归，完全绕开了计算昂贵的网格搜索，处理速度相比传统方法可提升数个数量级。
- **应对低信噪比挑战：**CNN 的层次化特征提取能力使其对噪声具有更强的鲁棒性。同时，贝叶斯神经网络等技术的发展 [bayescnn2022]，使得模型在给出参数预测的同时，还能提供可靠的不确定性量化，这对于评估低信噪比数据的分析结果至关重要。
- **应对泛化能力难题：**生成对抗网络 (GAN) 等生成式模型 [zhang2021_gan]，能够学习真实光谱的数据分布，并据此生成海量的、物理真实的合成光谱，从而有效扩充和补完训练集在极端参数区域的覆盖，极大提升了模型在稀有天体搜寻任务中的泛化能力。

正是基于以上挑战与机遇，本研究旨在开发一种专为大规模巡天数据中的 CEMP 星高效搜寻而设计的新型深度学习框架。我们的方法将融合上述先进技术，以期在保证高精度的前提下，实现对海量光谱数据快速、稳健的自动化分析，从而为揭示银河系早期的化学演化历史提供一个前所未有的大样本。

2.3. 研究方法概述

CEMP (Carbon-Enhanced Metal-Poor) 恒星是研究早期宇宙化学演化和第一代恒星形成的关键探针。其分类标准随着观测技术的进步和理论模型的发展而不断演进 [beers1992, aoki2007, norris2013]。本研究采纳当前广泛接受的严格标准（详见表1），基于大规模巡天数据，利用数据驱动的方法系统性地搜寻 CEMP 恒星候选体。

我们的研究流程遵循一个三阶段框架，旨在从海量低分辨率光谱数据中高效、准确地识别目标。具体步骤如下：

1. **构建训练样本库：**实验数据源于 LAMOST DR11 发布的低分辨率光谱。为给后续深度学习模型提供精确的标签，我们首先将 LAMOST 星表与多个高分辨率光谱巡天（如 APOGEE DR17、LAMOST-Subaru 等）

的星表进行交叉匹配。匹配过程基于天球坐标（赤经 RA, 赤纬 Dec），容许的最大位置误差为 3 角秒，最终建立一个包含低分辨率光谱和高精度恒星物理参数（有效温度 T_{eff} 、表面重力 $\log g$ 、金属丰度 $[\text{Fe}/\text{H}]$ 、碳丰度 $[\text{C}/\text{H}]$ ）的综合数据集。

2. **恒星参数预测：**利用构建好的数据集，我们训练一个深度学习模型。该模型以预处理后的低分辨率光谱作为输入，学习光谱特征与恒星物理参数之间的复杂非线性关系，并最终用于预测海量未知恒星的 T_{eff} 、 $\log g$ 、 $[\text{Fe}/\text{H}]$ 和 $[\text{C}/\text{H}]$ 。

3. **CEMP 候选体验证与筛选：**在获得大规模恒星的预测参数后，我们依据表1中定义的 CEMP 恒星物理标准，对预测结果进行严格筛选，最终确定一个高置信度的 CEMP 恒星候选体列表。

这一流程整合了大规模巡天数据与先进的机器学习技术，为在银河系中系统性地发掘稀有的贫金属恒星提供了有效途径。

表 1. 恒星分类参数标准

恒星类别	参数标准
CEMP（碳增强金属贫星）	$[\text{Fe}/\text{H}] < -2.0$ 且 $[\text{C}/\text{Fe}] > 0.7$
EMP（极贫金属星）	$[\text{Fe}/\text{H}] < -3$
VMP（甚贫金属星）	$-2.5 \leq [\text{Fe}/\text{H}] < -1.5$

2.4. 数据筛选标准

为了确保后续恒星参数预测的准确性和可靠性，我们对从 LAMOST DR11 中提取的光谱数据实施了一系列严格的质量控制标准。这些标准旨在剔除低质量、存在问题的观测数据，保留最适合进行科学分析的高置信度光谱。筛选过程主要依据 LAMOST 数据处理流水线提供的质量评估参数，具体标准和理由如下表 2 所示。

这些筛选标准分为两个层次。第一梯队标准（信噪比和视向速度误差）是保证光谱基础质量的“硬性”门槛，直接关系到光谱特征的真实性和基础参数（如红移）的可靠性。任何不满足这些条件的观测都会被直接舍弃。第二梯队标准（恒星大气参数误差）则提供了进一步的质量控制，用于剔除那些虽然通过了基础筛选，但其物理参数测量结果由流水线判定为置信度不高的光谱。

通过这套多层次的筛选流程，我们构建了一个高质量的光谱样本库，为后续深度学习模型的训练和 CEMP 恒星的搜寻奠定了坚实的基础。

2.5. 波长范围选择依据

本研究针对 CEMP 恒星搜索任务，精心选择了 3800-8800 Å 的宽波段光谱数据。这一选择旨在全面覆盖对恒星大气物理参数（ T_{eff} , $\log g$, $[\text{Fe}/\text{H}]$, $[\text{C}/\text{H}]$ ）敏感的各类关键光谱特征，从而确保后续参数估计的精度与可靠性。该波段的选择依据与各科学参数的测量需求密切相关。

该波长范围不仅受益于地球大气较高的透过率，更重要的是其内部包含了从蓝端到近红外的一系列重要诊断特征：

- **有效温度 (T_{eff}) 测量：**光谱的整体形状，特别是蓝端的巴尔末跳变（Balmer Jump，约 4000 Å 附近）和氢的巴尔末线系（如 $\text{H}\beta$, $\text{H}\gamma$ ），对恒星有效温度高度敏感。这些特征能够有效反映恒星大气的整体热辐射特性。

表 2. 光谱数据质量筛选标准

参数	筛选标准	理由
第一梯队：基础质量控制		
SNR	SNRG > 20 或 SNRI > 20 (精细丰度分析: SNRI > 100)	确保光谱特征的真实性，避免被噪声淹没。 g 和 i 波段信息丰富。
RV_err	RV_err < 20 km/s	保证红移校正的准确性。
第二梯队：进一步质量保证		
Teff_err	< 100 K	
logg_err	< 0.1 dex	反映流水线对大气参数建模的置信度，剔除不确定性大的光谱。
FeH_err	< 0.1 dex	

- **表面重力 ($\log g$) 诊断:** 多个谱线特征可用于约束表面重力。例如，Ca II H&K 线 (3933, 3968 Å) 的线翼宽度、Mg I b 三线系 (约 5170 Å) 以及近红外的 Ca II 三线系 (8498, 8542, 8662 Å) 都对恒星表面重力十分敏感，是确定恒星演化阶段（如主序星、巨星）的重要诊断工具。
- **碳丰度 ([C/H]) 指标:** CH 分子的 G 带 (约 4300 Å) 是测量恒星碳丰度的最关键特征之一。对于 CEMP 恒星的识别，该区域的高质量光谱数据至关重要。
- **金属丰度 ([Fe/H]) 指标:** 在 5000-6000 Å 区域分布着密集的 Fe I 和 Fe II 谱线，它们是精确测定恒星铁丰度的基础。此外，其他元素的谱线（如 Mg, Na）也为研究详细的化学丰度提供了信息。

综合来看，选择 3800-8800 Å 的光谱数据，能够覆盖所有目标科学参数的关键诊断特征，为 CEMP 恒星的有效识别和详细物理参数测量提供全面且互补的信息。

2.6. 训练数据集构建与预处理

深度学习模型的性能高度依赖于训练数据的质量。为了构建一个适用于 CEMP 恒星参数估计的高质量参考数据集，我们整合了来自不同巡天项目的数据，并设计了一套系统性的数据预处理流程。首先，我们构建用于模型训练的标签集。本研究使用的低分辨率光谱主要来源于 LAMOST 巡天。为了获得这些光谱对应的精确恒星物理参数（即标签），我们将 LAMOST 星表与多个高分辨率光谱巡天（如 APOGEE DR17、LAMOST-Subaru、SAGA 等）的星表进行交叉匹配。这一匹配过程确保了我们的训练样本同时拥有低分辨率光谱和高精度的参数标签 (T_{eff} , $\log g$, [Fe/H], [C/H])。在构建了交叉匹配样本库之后，我们进一步对数据集进行筛选与优化以构造最终的训练集。第一步，为保证数据质量，我们移除了信噪比 (SNR) 低于 50 的光谱，确保了训练数据的可靠性。第二步，针对恒星金属丰度 ([Fe/H]) 存在的长尾分布问题，我们采用了随机均值采样方法。该方法通过平衡不同金属丰度区间的样本数量，有效缓解了因贫金属星样本稀少导致的数据不平衡问题，避免模型性能受到影响。完成数据集的构造后，我们执行以下数据预处理步骤，旨在消除仪器效应和观测噪声，并将数据转换为适用于深度学习模型的标准化格式：

1. **加载 FITS 文件 (step1_extraction.py):** 预处理的第一步是从 FITS 格式文件中加载恒星光谱数据。我们将观测到的输入光谱序列表示为 $F_{\text{obs}}(\lambda_{\text{obs}})$ ，其中 F_{obs} 是流量， λ_{obs} 是对应的观测波长。

2. **红移校正 (step2_redshift_correction.py)**: 利用已知的目标天体红移值 z , 通过多普勒效应公式将观测波长 λ_{obs} 转换到静止波长坐标系 λ_{rest} :

$$\lambda_{\text{rest}} = \frac{\lambda_{\text{obs}}}{1 + z} \quad (1)$$

该过程可表示为

$$F_{\text{obs}}(\lambda_{\text{obs}}) \rightarrow F(\lambda_{\text{rest}}).$$

为简洁起见, 后续步骤中的光谱 $F(\lambda)$ 均指代静止系光谱。

3. **光谱截断**: 为了专注于包含最多恒星物理信息的波段, 我们对光谱 $F(\lambda)$ 进行了截断, 选取波长范围为 $[3800, 5700]$ Å 和 $[5900, 8000]$ Å。

4. **小波变换去噪 (step3_noise_removal.py)**: 恒星光谱 $F(\lambda)$ 中普遍存在随机噪声 ϵ 。我们采用小波变换 (Wavelet Transform, WT) 技术对其进行去噪处理。该过程可以形式化地描述为: 首先对光谱信号进行小波分解得到一系列系数 w , 然后对系数进行阈值处理 (thresholding) 得到新的系数 w' , 最后通过逆小波变换 (Inverse Wavelet Transform, IWT) 重构出去噪后的光谱 $F_{\text{denoised}}(\lambda)$ 。整个流程可以表示为:

$$F(\lambda) \xrightarrow{\text{WT}} w \xrightarrow{\text{threshold}} w' \xrightarrow{\text{IWT}} F_{\text{denoised}}(\lambda) \quad (2)$$

5. **光谱归一化 (step4_normalization.py)**: 为了消除各种与距离和仪器相关的因素, 需要对光谱进行归一化。此步骤的核心是拟合出去噪后光谱 $F_{\text{denoised}}(\lambda)$ 的连续谱 (continuum), 记作 $C(\lambda)$ 。然后, 通过以下公式计算归一化光谱:

$$F_{\text{norm}}(\lambda) = \frac{F_{\text{denoised}}(\lambda)}{C(\lambda)} \quad (3)$$

此步骤的产出是成对的: **连续谱** $C(\lambda)$ 和 **归一化光谱** $F_{\text{norm}}(\lambda)$ (?)。

6. **等间距重采样 (step5_resample.py)**: 原始光谱的波长采样点 $\{\lambda_i\}$ 通常是不均匀的。为了便于后续模型

输入，我们通过插值方法（如线性或样条插值），将归一化光谱 $F_{\text{norm}}(\lambda)$ 从其原始的波长网格 $\{\lambda_i\}$ 重采样

到一个统一的、等间距的新波长网格 $\{\lambda'_j\}$ 上，得到最终的光谱序列 $F_{\text{final}}(\lambda'_j)$ 。

7. 数据集划分 (`step6_split_dataset.py`): 最后，将预处理完成的光谱数据集划分为训练集、验证集和测试集。

经过以上流程，我们最终获得了一个标准化的、清洁的、可直接用于模型训练的数据集。该数据集中的每一个样本都是一个代表恒星光谱特征的向量，并附有高精度的物理参数标签，为后续精准预测恒星参数提供了高质量的输入。

3. 异构特征流网络 (HFS-NET)

为了更全面地从恒星光谱中提取多模态信息，我们提出了一种新颖的端到端深度学习架构，名为异构特征流网络 (Heterogeneous Feature Stream Network, HFS-Net)。该模型的核心思想是，针对物理来源与信号特性完全不同的连续谱和归一化谱，设计两个独立的、架构异构的特征提取分支，然后对提取到的深层特征进行高效融合，最终通过一个共享的预测头实现对多个恒星物理参数的同步回归。该模型的所有组件均采用模块化设计，可通过配置文件灵活替换与组合。

3.1. 模型总体架构

本模型的总体架构如图 1 所示。数据处理流程遵循“分离-并行处理-融合-预测”的原则。首先，通过??中描述的预处理方法将原始光谱分离为连续谱和归一化谱。随后，这两个信号被送入两个并行的、专门设计的特征提取分支：

- **连续谱分支**，一个基于频域的混合专家网络 (MoE)，旨在捕捉光谱的全局、低频特性。
- **归一化谱分支**，一个基于多尺度卷积金字塔的网络，旨在捕捉吸收线的局部、高频细节。

两个分支提取的特征在经过维度对齐后，被一个融合模块进行合并。最终，融合后的特征向量被送入一个多任务预测头，以回归得到四个关键的恒星物理参数 (T_{eff} , $\log g$, $[\text{Fe}/\text{H}]$, $[\text{C}/\text{Fe}]$)。

3.2. 核心模块详解

3.2.1. 连续谱分支：基于频域的混合专家网络

设计动机：连续谱的形状复杂，其缓变的趋势中可能混合了由恒星自身物理过程、星际红化及仪器响应等多种因素叠加而成的宽带特征。单一的卷积结构难以完全解耦这些多样化的模式。我们采用混合专家网络 (Mixture-of-Experts, MoE)，旨在通过“分而治之”的策略，让不同的“专家”子网络学习处理不同的特征模式，从而在不显著增加总计算量的前提下，提升模型的容量和表达能力。

实现细节：该分支 (`CustomMoEBranch`) 首先通过快速傅里叶变换 (FFT) 将连续谱信号从波长域转换到频域。然后，一个轻量级的门控网络 (Gating Network) 会根据输入的频域特征，动态地、稀疏地选择激活多个专家网络中的 Top-k 个。每个专家网络都是一个独立的轻量级金字塔卷积模块，可以专注于学习和处理特定的频域模式（如特定的频率成分或形状）。

3.2.2. 归一化谱分支：多尺度金字塔卷积网络

设计动机：归一化谱中的吸收线形态各异，有的因恒星高速自转或高压环境而变得宽阔平缓，有的则因低温环境而显得尖锐狭窄。为了能同时、且鲁棒地捕捉这些不同形态的谱线特征，该分支 (`MultiScalePyramidBranch`) 采用了多尺度并行卷积的设计。



图 1. 异构特征流网络 (HFS-Net) 的整体架构图。

表 3. HFS-Net 模型关键模块的超参数配置。

模块 (Module)	参数 (Parameter)	值 (Value)
连续谱分支 (CustomMoEBranch)	FFT: n_fft	512
	FFT: hop_length	128
	MoE: num_experts	4
	MoE: k (top-k)	2
	专家网络结构	3 层金字塔 CNN
归一化谱分支 (MultiScalePyramidBranch)	金字塔层数	3
	输出通道数	[16, 32, 64]
	并行卷积核尺寸	[3, 5, 7]
融合模块	融合策略	add (逐元素相加)
预测头 (MultiTaskHead)	卷积金字塔层数	4
	FFN 层维度	[1024, 256, 128]
	Dropout	0.3

实现细节: 在金字塔结构的每一层, 都并列着多个使用不同尺寸卷积核 (例如, 3x1, 5x1, 7x1) 的卷积层。小卷积核擅长定位精确的、孤立的窄线轮廓, 而大卷积核则能更好地整合宽线翼部或由多个邻近谱线混合 (blending) 而成的复合特征。通过将这些不同尺度的特征图进行拼接, 模型能够构建一个对谱线形态变化不敏感的、信息丰富的层级化特征表示。

3.2.3. 特征融合与多任务预测

特征融合: 两个分支的输出特征图在经过自适应池化层对齐长度, 并通过 1x1 卷积对齐通道数后, 通过简单的逐元素相加 (`strategy: 'add'`) 进行融合。这种融合策略简洁高效, 它假设两个分支学习到的特征是相互对齐且可以线性叠加的, 共同构成一个统一的、描述恒星状态的特征空间。

多任务预测头: 融合后的特征被送入一个共享的多任务预测头 (`MultiTaskHead`)。该模块首先通过一个内部的卷积金字塔做最终的特征抽象, 然后由一个多层前馈网络 (FFN) 回归出所有目标参数。将所有参数的预测放在同一个头中进行, 是一种典型的多任务学习范式, 它使得模型在学习预测一个参数时, 可以利用到与其他参数共享的、潜在的物理关联信息 (例如, T_{eff} 和 $\log g$ 之间存在的物理关联), 从而可能提升所有任务的整体预测精度。

3.3. 训练策略

我们采用均方误差 (Mean Squared Error, MSE) 作为模型的损失函数, 并使用 AdamW 优化器进行端到端的模型训练。学习率调度器采用余弦退火 (Cosine Annealing) 策略。所有超参数均在配置文件 `flexiblefusionnet.yaml` 中定义。

4. RESULTS

This is the results section.

5. DISCUSSION

This is the discussion section.

6. CONCLUSION

This is the conclusion section.

7. ACKNOWLEDGMENTS

This is the acknowledgments section.

8. APPENDIX

This is the appendix section.