# Stellar Karaoke：通过速度白化从恒星光谱中深度盲分离地球大气效应

Nima Sedaghat,[1,2★] Brianna M. Smart,[1†] J. Bryce Kalmbach,[1†] Erin L. Howard[1] and Hamidreza Amindavar[3]
[1]DiRAC 研究所和华盛顿大学天文学系，美国华盛顿州西雅图 [2]*The Raw Data Speaks*倡议 [3]阿米尔卡比尔理工大学电气工程系，伊朗德黑兰

接收于 2023 年 8 月 t 22。收稿日期：2023年8月22日；原稿日期：2023年1月 17日

arXiv:2301.00313v2 [astro-ph.SR] 7 Nov 2023

**摘要**
我们报告了一项研究，探讨如何利用天文大数据和深度神经网络，帮助我们发现并揭示潜在现象的新见解。在利用天文大数据进行无监督知识提取的实验中，我们意外发现，深度卷积自编码器往往会滤除恒星光谱中的地球大气吸收线。通过进一步实验，我们发现，只有当光谱处于重心参考系时，网络才会自动识别恒星成分与地球大气成分这两种成分之间的统计独立性，并滤除后者。我们利用这一发现，将其转化为一种概念验证方法，以完全无监督的方式去除恒星光谱中的地球大气吸收线：通过为观测光谱施加随机的虚拟径向速度，提高不同观测之间地球大气吸收线的熵。这项技术使光谱中的大气成分实现了一种非标准形式的“白化”，从而消除它们在多次观测中的相关性。我们处理了超过25万条来自高精度径向速度行星搜索（HARPS）的光谱，并通过与已知地球大气吸收线数据库进行定性和定量评估，表明大多数地球大气吸收线都能被成功滤除。我们的方法“Stellar Karaoke”完全不需要观测时间、地点或大气分子分布等参数的先验知识，并能在毫秒内处理每条光谱。我们还在斯隆数字巡天（SDSS）数据上进行训练和测试，发现由于分辨率较低，性能显著下降。最后，我们探讨了未来基于所提出方法开发工具的方向。

**关键词：** 技术：光谱学 – 大气效应 – 方法：数据分析 – 方法：统计学 al

## 1 引言

本文将介绍并详细阐述一种方法，该方法结合了一种新颖的信号处理技术与完全无监督的[1] 深度神经网络，无需手动建模和调参，即可从观测光谱中去除地球大气吸收线的非线性时变影响——图1。具体而言，我们展示了如何通过考察庞大的光谱数据集，揭示其特征的新见解，从而为现有的难题提供创新的解决方案。

在我们探索从天文大数据中无监督提取知识的实验中，我们训练了深度神经网络，对大量存档恒星光谱进行非线性降维，并偶然发现，网络在重建光谱时往往会舍弃某些特定区域（图1）：地球大气吸收线。进一步实验发现，只有当光谱处于重心参考系时，网络才会表现出这种行为。“重心参考系”是以天体系统质心为中心的参考系，可为高精度天文测量提供稳定的参考

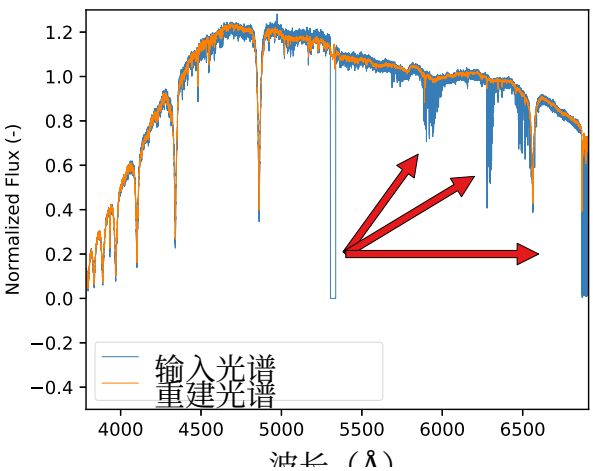


**图 1.** 我们利用大型数据集中的恒星光谱统计特性，将其输入卷积自编码器，并以极少的工作量剔除地球大气吸收线。图中的三个红色箭头标出了自编码器识别为地球大气吸收线、因而未纳入重建光谱的谱线。

★ 电子邮件：nimaseda@uw.edu（NS）
† 这些作者对本研究的贡献相同[1] 一种无需真实值数据的训练方案——概述参见James 等人(2023)。

© 2023 作者

（Soffel 等，2003年；Bohlin 等，2014年）。相较之下，“地心参考系”以观测者在地球上的位置为中心。后者是光谱观测表征所采用的自然坐标系，但它没有考虑地球绕太阳系质心运动的影响，因此可能导致恒星光谱测量出现变异或误差，通常需要将其转换到重心参考系。

我们进一步的实验表明，通过观察大量此类变换，网络能够自动识别两个成分——恒星成分与地球大气成分——之间的统计独立性，并决定舍弃后者。我们利用这一发现，将其转化为一种概念验证方法，以完全无监督的方式从整组恒星光谱中去除地球大气吸收线：在原始地心参考系中，为观测光谱施加一个随机的虚拟径向速度，以提高地球大气吸收线在不同观测之间的熵。这种技术使光谱中的大气成分以一种非标准形式进行“白化”，从而使其在多次观测之间去相关——这类似于转换至重心参考系所产生的效果，但完全无需与观测相关的参数。

### 1.1 大气效应对恒星光谱的影响

在恒星光谱的地基观测中，天体源的光谱到达探测器时已发生改变，这种情况并不少见。光谱可能受到空间中多种现象的影响（红移/蓝移、发射、吸收等）。光子最终到达地球时，在穿过地球大气并由望远镜收集的过程中，光谱还会经历额外的变化。望远镜效应通常已有充分表征，且受到严格约束（CCD噪声、条纹效应、镜面缺陷等）。然而，其他效应通常更难建模和校正。

尤其是，光线穿过存在于地球大气中的地球大气分子，会在观测光谱中产生吸收线。这些吸收线被称为地球大气吸收线，是光子与分子之间由于一系列电子、旋转和振动跃迁相互作用的结果（Rybicki & Light- man 1986）。这些相互作用会持续造成观测光谱的污染和信息损失——例如参见Bolton 等 (2012)、Artigau 等 (2014）。负责吸收的大多数分子包括 $O_2$、$H_2O$、$CO_2$、CO、$CH_4$、COH、$O_3$、$N_2O$，以及夏帕伊臭氧吸收带（Catanzaro，1997年；Griffin，2005年；Moehler 等，2014年；Smette 等，2015年）。这些吸收特征中有许多位于电磁波谱的近红外和紫外部分，而臭氧、氧气和水在光学波长处也会产生较弱的吸收特征。由于分子发射光子，光谱中还会增加额外的发射特征。通过观测天空中的无源区域，可以相对直接地处理这些发射特征。在“空白”区域观测到的发射线提供了一个易于获取的发射去除模板。吸收特征则更加复杂。不同的地球大气吸收线可能随气团质量呈线性或非线性变化，并且会受到观测时大气条件的影响（Bailey 等，2007年）。

大多数传统的地球大气吸收线去除方法都将每条光谱视为一个独立的整体，而不把观测光谱集作为一个整体来处理——例如，参见 Hrudková & Harmanec (2005年)。这类方法中的一类会借助标准星去除大气吸收线。它们通过观测光谱特征相对较少的标准星，并将其用作大气模板来实现这一点。这通常通过观测 A0V 型或 G-型星来完成，不过这种方法主要用于近红外和红外波长（Vacca 等，2003年；Artigau 等，2014年）。使用 G 型星时，可将太阳的高分辨率傅里叶变换光谱用作模板，以校正恒星的本征谱线（Livingston& Wallace 1991；Hase 等，2010年）。

理想情况下，参考星应位于目标星附近，并在相近的时间进行观测，以尽可能准确地测量大气条件。随后，用目标天体源的光谱除以地球大气模板。然而，如果某个观测时段内没有可用的此类标准星，或观测条件导致光谱质量较差，这种方法便会受到限制。大气吸收也可能在短短几分钟内迅速变化（Moehler 等，2014年），因此需要投入大量时间和资源来观测模板星。相除过程需要迭代进行，并调整波长和强度缩放，以匹配大气效应。

另一种常见方法是对地球大气建模，并创建能够精确模拟大气吸收的合成光谱。这些模型利用数值模型求解地球大气的辐射传输方程，例如 Allart 等(2022年。这些方法依赖于夜间观测所获得的大气条件测量。通常会使用 `molecfit` (Smette 等，2015年；Kausch 等，2015年) 和 `telfit` (Gullikson 等，2014年) 等程序来完成这项工作。辐射传输代码会获取观测时的温度、压力和湿度，并利用分子参数数据库拟合地球大气吸收。尽管这种技术相当有效，但当光谱中本征特征过多、连续谱很少或不存在、信噪比较低，或在高水蒸气含量条件下观测气团质量较大时，其表现可能不佳（参见Alonso-Floriano 等(2019年）；Shulyak 等(2019年）。此外，目前实现的此类技术（例如 `molecfit`）运行速度相当缓慢，以至于在当今的计算机上，对单条光谱拟合一个模型可能需要数分钟。McCloat 等(2021年；Nicholls 等(2017) 报告了 `molecfit`在其特定实验中遇到的其他类型的模型拟合问题。有关上述方法的比较，请参见 Ulmer-Moll 等(2019年)。

### 1.2 降维

天文学领域近期的一些研究通过考察一组示例，对光谱进行隐式建模，从而有望消除手动显式建模的需要。这类方法通常被宽泛地称为数据驱动方法，在过去几十年中取得了良好的发展势头。尤其是在天文学中，以数据驱动的方式研究光谱，主要应用于降维这一广泛背景下。

可以说，天文学中最受欢迎的降维方法是主成分分析（PCA——综述参见 Jolliffe 和 Cadima (2016年），多年来一直应用于天文光谱。 Connolly 等人 (1995) 使用该方法，仅凭前两个主成分便对星系光谱进行分类；而 Bailer-Jones 等人 (1998年) 表明，PCA 可将恒星光谱压缩 30 倍以上，并对数据集中的异常光谱和非恒星光谱进行分类。此外，Bailer-Jones 等人(1998年) 发现，压缩不仅去除了噪声，还消除了光谱中的伪特征，例如底片扫描过程中观测板上出现的尘埃。一些研究，例如 Artigau 等 (2014)，甚至尝试使用主成分分析分解，从高精度径向速度行星搜索项目（HARPS；Mayor 等人 2003）数据集中去除少数特定天体的地球大气吸收线，并展示了径向速度测量精度的提升

然而，他们的方法并非仅针对恒星光谱进行处理，而且仍需观测地球大气标准星。我们的方法只需要一组输入恒星光谱，无需任何额外信息。

此外，关于 PCA 适用于此类用例的一个非常重要但经常被忽视的问题是，其本质为线性，这使得它从定义上无法对非线性现象进行建模。例如，对于地球大气吸收线，仅基于 PCA 的方法可能无法在繁杂的光谱区域中区分地球大气吸收线和恒星谱线。这种局限性早在 Bailer-Jones 等人 (1998年) 的研究中就已得到充分观察。在本文中，我们展示了为何必须转换到更复杂的特征空间，才能完成此类任务。

### 1.3 自编码器

有一类密切相关的神经网络，称为编码器-解码器网络（图 2）。压缩部分将输入压缩为通常是低维的表示，也称为“潜在表示”，或简称为“编码”。扩展部分随后用于将这一低维表示解码为所需的输出。用于编码（和解码）的变换会在训练过程中根据具体任务进行学习。这类网络的一个特殊情况是**自编码器**，它是一种无监督网络，无需带标签的输入数据即可训练以重构输入。如 Hinton 和 Salakhutdinov (2006) 所述，这些网络可用于降低输入数据的维度，是 PCA 的一种非线性推广。Wang 等人 (2016年) 展示了使用自编码器进行降维的综合实验，并利用合成图像和真实图像研究了不同潜维度的影响。

在Yang 和 Li（2015年）中，作者使用经典的（非卷积）自编码器，将斯隆数字巡天（SDSS）第七次数据发布（Abazajian 等人 2009年）的 3000 维光谱转换为低维特征，随后利用这些特征估计大气参数。Wang 等人（2016年）使用半经典（全连接解码器）自编码器，从天文光谱中学习特征。他们将基于所学习特征的光谱分类与 PCA 和局部线性嵌入（LLE，一种替代性的非线性降维方法）进行比较。结果发现，在一个包含 F、G 和 K 型恒星的数据集中对光谱进行分类时，他们的自编码器方法表现最佳。

仅使用自编码器的一个问题，尤其是在（天体）物理学应用中，是概念映射到降维后的信号——潜在空间时会相互纠缠。变分自编-码器（VAE；Kingma 和 Welling 2014）已被证明能在一定程度上减轻这一问题。Portillo 等人（2020年）使用 VAE 重建斯隆数字巡天的星系光谱，并将其重新采样为包含 1000 个分量的输入光谱。他们发现，当将潜在空间的维度限制为 2、4、6 或 10 个分量（≥ 99% 压缩）时，与采用相同分量数的 PCA 和非负矩阵分解（NMF）相比，VAE 和传统自编码器重建斯隆数字巡天星系光谱的重建误差都更低。随后，作者证明，不同星系类别位于潜在空间的不同区域，从而能够实现分类。

经典神经网络（包括自编码器）仅由全连接层构成（Hinton 2006）。然而，一方面，随着深度神经网络的出现，我们认识到，更深的网络对于学习更有用的特征至关重要（Szegedy etal. 2015）；另一方面，全连接网络难以扩展到更深的网络和高维数据（Bengioet al. 2009）。因此，一个足够深的仅基于全连接层的自编码器无法处理现代光谱信号：例如，对于 HARPS，像素数为 ~ 300,000 时，一个仅由 6 层组成的全连接自编码器就包含超过 65 亿个可训练参数，即使只输入一条光谱，也超出了当今普通图形处理器的内存容量。附录 B 给出了此类架构的示例。因此，恒星光谱通常会被下采样至可处理的分辨率：Portillo 等人（2020）、Wang 等（2016b）和 Yang 和 Li（2015）分别使用了 1000、2000 和 3000 个像素。这会导致原始高分辨率光谱中的高分辨率信息含量遭受不可挽回的损失，我们也将在一项实验中展示这一点。为解决这一问题，我们还采用卷积层和上卷积层，将数据变换至所需的特征空间；这一思路借鉴自计算机视觉，也使网络获得一定程度的平移不变性（Fukushima（1980））。通过这种方式，我们成功训练出一个深达 32 层、可训练参数少于 600 万的自编码器。借助当今的硬件，这样的模型已可处理，并能让我们进一步认识恒星光谱。对卷积层的技术细节、优势及超参数选择的进一步探讨超出了本文的研究范围；不过，感兴趣的读者可参阅本世纪前二十年间的计算机视觉文献（包括上述作品），以深入了解从经典机器学习向深度学习的这一转变。

请注意，尽管卷积神经网络近年来已广泛应用于天文数据，但其研究重点大多是从二维图像中提取信息：从光度红移回归（Hoyle（2016）；Pasquet等（2019））和星系形态分类（Dieleman等（2015）；Huertas-Company等（2015）；Ghosh等（2020））到用于暂现源探测的图像生成编码器-解码器（Sedaghat和Mahabal（2018））。然而，天文学中很少使用研究光谱所必需的一维卷积网络。Sharma等（2020年）实现了这样一种网络，用于对斯隆数字巡天的恒星光谱进行分类，并将其方法与全连接（前馈）实现进行比较。

### 1.4 盲源分离

地球大气吸收线本质上来自不同于恒星谱线的天体源。因此，消除这些吸收线属于“源分离”问题；这一课题主要在数字信号处理领域中得到研究。该问题的一般定义假设存在多个传感器和多个信号，即一个多输入多输出（MIMO）系统。然而，在当前问题中，我们只有一个输出，因此任务稍微困难一些。此外，经典的源分离方法大多考虑来自不同天体源的信号线性混合，而我们的情况远非如此简单。由于每个观测光谱中的径向速度具有随机性，而且我们不对两种信号混合后所产生的影响性质作任何假设，因此这可以视为一项“盲”源分离（BSS）任务。

近期的盲源分离方法通常以独立成分分析（Comon 1992）及其扩展为核心——例如参见 Choi 等人 (2005年)；Pal 等人 (2013 年 ) 的综述。Särelä 等人 Särelä 等人 (2005 年 ) 提出的去噪源分离（DSS）是另一种方法，它将源分离任务视为去噪过程，因此与我们方法所采用的基本概念相似。不过，DSS 旨在对混合矩阵（混合机制）进行统计估计，而我们的方法则通过对白化不需要的成分、使其更像噪声，来直接恢复目标信号。

近期许多源分离方法，尤其是针对音频/音乐信号的方法，都将编码器-解码器神经网络用作其方法的核心组件（Kameoka 等人 2018年；Seki 等人 2019年；Grais 和 Plumbley 2017；Liu 和 Yang 2018）。但这些方法主要依赖显式监督，因此需要合成真实值。Neri 等人 (2021年) 训练了一种变分编码器-解码器网络，利用变分自编码器的解缠结特性，在输出中生成分离后的来源。上述音频处理方法都在频谱图图像上训练二维神经网络。Wisdom 等人 (2020年) 提出了一种半监督音频源分离方法，可直接处理时域信号。尽管他们同时实现了监督和自监督[2]网络，但作者表明，后者需要由前者进行正则化才能正常工作。最近两篇论文中的无监督部分都基于一个强假设：输入是其组成天体源的线性混合。我们的网络是一种简单的自编码器，不作线性假设，并且经证明能够在后文所述的复杂非线性条件下良好运行。在此背景下，我们的 Karaoke 网络将恒星光谱视为音乐，将对抗效应视为需要从中去除的人声。

### 1.5 复合光谱

早在1994年，就出现了另一条旨在解缠结复合光谱的研究路线，当时 Simon 和 Sturm (1994年) 尝试分解双星光谱中的各个组成成分。近年来，`wobble` (Bedell 等（2019）) 等工具也将恒星和地球大气成分建模为两个独立成分，并尝试将这些参数模型拟合到观测光谱上。然而，这类方法依赖于观测时间参数的可用性，其中最主要的是日心系径向速度（BERV）。更重要的是，这些方法需要在同一个数据集中对同一颗恒星进行多次观测，以便为单个天体获取足够多样的日心系径向速度值，而通常并非有很多天体都能满足这一条件。这些模型也针对特定天体，即每个模型都是为单个天体准备（即优化）的，只用于处理该天体的光谱。

相反，我们的方法不需要光谱本身以外的任何外部信息：只需一组原始标量向量。尽管为不同日心系径向速度值的使用提供理论依据的物理机制与我们的白化技术相近，但我们的方法依赖于多个不同天体的光谱之间的多样性，同时也捕捉其中的共性，因此不受特定天体的限制——无论是在训练阶段，还是应用于光谱时都如此。我们还在实验部分将其性能与 wobble 进行了比较。

在所用网络架构以及数据驱动的无监督概念方面，与我们工作最接近的或许是Sedaghat 等人（2021年），他们以完全无监督的方式使用卷积变分自编码器，从大量 HARPS 光谱中提取知识。我们的工作沿用并使用了完全相同的神经网络架构。

### 我们的贡献

- 我们将大量超高维光谱作为一个整体进行处理，使其统计特性自然显现，从而可以将其视为随机过程。
- 我们将恒星光谱中的地球大气分量建模为独立恒星，并为其施加虚拟径向速度，以实现统计白化／去相关。
- 我们采用卷积自编码器，使其自动充当源分离工具，以完全无监督的方式去除地球大气吸收线：无需显式建模，也无需进行任何标注。
- 我们的方法与目标无关，并能在不到一秒的时间内处理新的观测。

## 2 问题表述

我们旨在从任意观测信号 $x$ 中去除有害的大气效应。在我们的案例中，观测信号是恒星光谱，因此是波长 $\lambda$ 的函数，我们将其记为 $x(\lambda)$。

在传感器捕获恒星光谱之前，我们使用以下公式对影响它的各种现象进行建模：

$$x(\lambda) = \Big(s(\lambda)t(\lambda)\Big) * h(\lambda) + n(\lambda) \tag{1}$$

其中 $s$ 是恒星光谱，包含相对恒星速度和星际介质造成的视线效应。$t$ 是一个虚构信号，用来表示影响该光谱的地球大气吸收线，也称为大气透射光谱（Rudolf 等（2016））。

我们用 $h$ 表示我们所称的观测传递函数。该函数用于建模信号在传感过程中经历的变化，包括但不限于仪器轮廓（Rudolf 等，2016年）。$n$ 用于建模传递函数中未建模的加性噪声，$h$。

我们处理一个由大量 $N$ 观测组成的集合，这些观测大多来自不同的天体源。我们使用下标 $i$ 来区分不同的观测：

$$x_i(\lambda) = \Big(s_i(\lambda)t_i(\lambda)\Big) * h(\lambda) + n_i(\lambda) \qquad i \in \{1, 2, \ldots, N\} \tag{2}$$

请注意，由于我们同时处理多个天体，光谱 $s$ 取决于 $i$。大气条件也随时间变化，这一点反映在 $t$ 对 $i$ 的依赖关系中。噪声 $n$ 也是如此。不失一般性，为简化起见，我们假设 $h$ 在整个观测集合中保持不变。

我们的目标是消除地球大气吸收线对观测信号的影响。在此模型中，这相当于对每次观测提取 $s_i(\lambda) * h(\lambda)$——请注意，消除观测效应$h$并非本研究的目标。

我们还将目标天体的径向速度 $v$ 对观测光谱 $s$ 的影响建模为：

$$\begin{aligned} s_i(\lambda) &= \mathcal{V}\{\mathring{s}_i(\lambda), v_i\} \\ &= \mathring{s}_i\Big(\lambda\big(1 - \frac{v_i}{c}\big)\Big) \end{aligned} \tag{3}$$

其中 $\mathring{s}_i$ 表示径向速度为零时的观测光谱，也就是不存在多普勒频移的情况。下文将 $\mathring{s}_i$ 称为**静态光谱**。$\mathrm{V}\{\cdot,\cdot\}$算子表示径向速度的非线性[3] 效应，

[2] 一种依赖真实值数据的训练方案，而这些数据可“免费”获得：标签会从未标注输入数据中自动生成。

[3] 该算子对其第一个自变量呈线性，但对第二个自变量呈非线性：$\mathrm{V}\{s, v_1 + v_2\} \neq \mathrm{V}\{s, v_1\} + \mathrm{V}\{s, v_2\}$

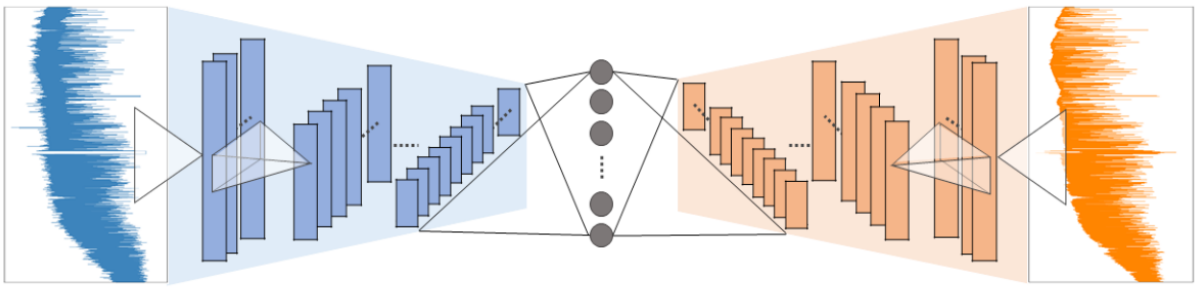

**图 2.** 一个典型的变分自编码器经过训练，可用于重建恒星光谱；当压缩程度足够高且卷积核数量保持较少时，它能够从输入中分解出具有物理意义的成分。为简化起见，此处省略了瓶颈层的变分实现。左右两侧的主要收缩和扩展部分由（上）卷积层组成，而瓶颈层则需要全连接层。蓝色表示输入光谱，橙色表示重建结果。该图最初发表于 Sedaghat 等人（2021年）。

该效应对光谱的影响已在上式第二行展开，其中 $v_i$表示径向速度， $c$ 表示光速。 $v_i$ 和 $c$ 的物理单位可以任意选择，只要二者保持一致即可。

$t_i$ 另一方面，根据定义，它们不依赖于 $v_i$，因此可以建模为：

$$t_i(\lambda) = \mathcal{V}\left\{\mathring{t}_i(\lambda), 0\right\} \tag{4}$$

### 2.1 连续与离散

上述所有信号在被探测器感测之前都具有连续的性质。探测器感测过程包含采样这一步骤，将光谱转换为一系列实值样本。因此，我们在实验中使用的数据是 $x_i$（$\lambda$）的离散表示，即 $X_i^l$——我们使用 $l$ 作为采样像素的离散索引。

不过，出于以下原因，保留连续表征的表述是稳妥且更清晰的。首先，目前为止建模的信号都是采样前信号的组成部分，$x_i(\lambda)$，因此连续模型成立。其次，我们的方法中唯一明确沿波长轴修改信号的部分，即$\mathcal{V}$ 算子（第3.1节），通过对信号的插值版本重新网格化来完成工作；实际上，它将离散信号转换回连续版本，应用变换，再次采样回离散空间。因此，该算子在连续空间中的数学定义是有效的。

### 2.2 信号作为随机过程

从天空中任意选取一个天体并感测其信号 $x_i$，该信号在传播过程中会受到许多非确定性现象的影响；这一过程可以视为随机过程的一次实现。换句话说，对于各种 $i$， $x_i$的集合或其离散表示 $X_i$，构成了随机过程 $\{X\}$[4] 的实现集合。请注意，在此特定应用中，随机过程的索引集是从波长 $\lambda$ 中采样得到的——这有点违背直觉，因为在典型应用中，索引集通常与时间有关。

因此，每个 $X^l$ （如第2.1节所定义）都表示一个随机变量，其结果为$X_i^l$。类似地，我们可以将$\{S\}$、$\{\mathring{S}\}$ 和$\{T\}$ 建模为同一类型的离散索引随机过程，分别对应 $s_i$、 $\mathring{s}_i$ 和 $t_i$。 $\{S\}$ 于是成为生成各种干净光谱$S_i$的生成器，而$\{\mathring{S}\}$ 则生成静态 $\mathring{S}_i$ 光谱。正如我们将在下一节中看到的，对信号采取这种非确定性观点，有助于我们更清晰地解释各组成部分的统计运算和性质。

## 3 方法

该方法由两个关键步骤组成：

a) 提高地球大气成分在不同观测中的熵，并且，

b) 将光谱转换到一个恒星成分与地球大气成分可分离的空间中，然后去除其中非主导的成分。

**中位数归一化** 作为预处理步骤，我们根据数据集中每条光谱的中位数对其进行归一化，以减轻天体源距离不同所造成的影响，否则会导致样本间通量范围出现极端变化。

### 3.1 速度白化

在默认条件下，光谱观测处于所谓的地心参考系中；如果存在地球大气吸收线，它们的波长（或像素位置）都相同。不过，从统计学角度来看，这意味着：

为简化起见，我们假设所有观测到的天体都具有相同的静态光谱， $\mathring{s}_i(\lambda)$。因此得到的光谱， $s_i(\lambda)$ 将仅因其径向速度而有所不同， $v_i$，根据定义，这些速度被建模为沿波长轴的缩放变换（公式（3））。然而，由于 $v_i$ 是随机变量的取值，

[4] 在此语境中，随机一词与恒星光谱中的结构并不矛盾。结构编码在.....的基本参数中

随机变量的组成集合，例如期望值和协方差矩阵。

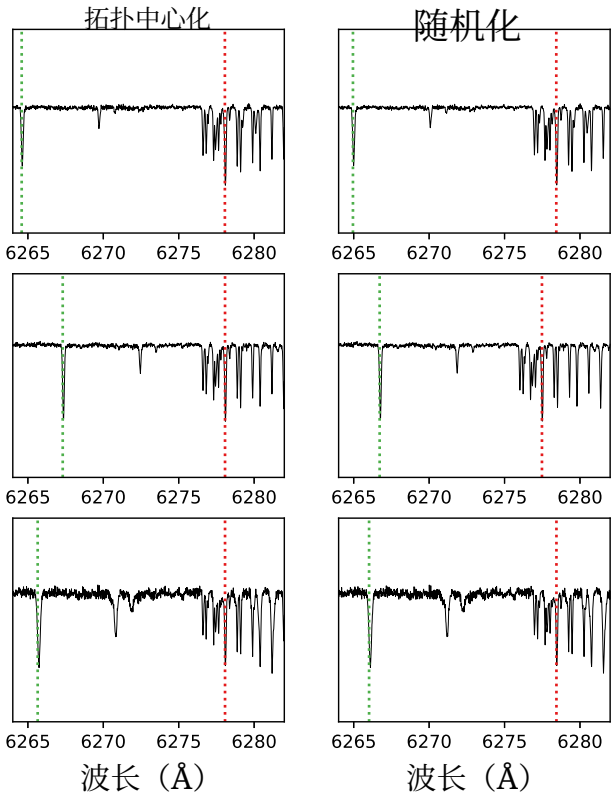


**图3.** 左侧展示了一组示例光谱。由于径向速度不同，恒星谱线并未对齐。但地球大气吸收线彼此对齐，尽管由于其固有的时间依赖性，形状可能各不相同。右侧展示的是同一组光谱经过速度随机化后的结果。此时地球大气吸收线也不再对齐，但其排列模式与恒星谱线不同。红色和绿色的垂直虚线分别标示示例地球大气吸收线和恒星谱线的位置。标记位置右侧还能看到更多地球大气吸收线。

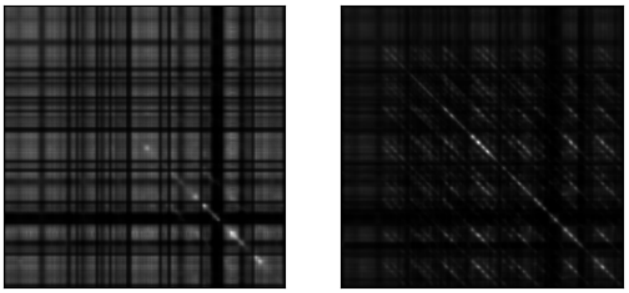

**图4.** 信号集协方差矩阵在白化前后的直观示意图。左侧展示了部分观测中一个由90个样本组成的子集在 [6275, 6285] 埃区域内的协方差矩阵。右侧展示了经过 速度白化后的相同矩阵，其中 $v_i$ 是从均匀分布中采样得到的： $V \sim U(-30$ 千米/秒，30 千米/秒)。现在协方差矩阵更接近单位矩阵，证实实现了一定程度的白化/去相关。

$s_i(\lambda)$ 由于由此产生的随机位移，仍会高度不相关。请注意，每个随机变量都定义在特定像素上，跨多个观测取值，如第2.2节所述。

$t_i$ 另一方面，它们由一组强度各异、但都出现在预先指定波长处的吸收线构成。这一特性使它们在多次观测之间高度相关，也容易被任何模型拟合——图3，左栏。

为降低地球大气信号各次实现之间现有的相关性，我们通过使用模拟径向速度对其进行随机化，来增加 $t_i$ 的熵，

$$t'_i(\lambda) = \mathcal{V}\left\{\mathring{t}_i(\lambda), v'_i\right\} \tag{5}$$

其中每个 $v_{ni}$ 都是一个随机速度，取自

$$V' : \mathbb{R} \to \mathbb{R}, \tag{6}$$

一个实值随机变量生成器，其分布任意——在我们的实验中选择的是均匀分布。根据方程（3），只需收缩或扩展波长轴即可实现上述效果。但请注意，这纯粹是一种人为现象。尽管我们选择使其产生与恒星“真实”径向速度相同的效果，但它对地球大气吸收线并没有特别的意义——将观测光谱变换到重心参考系时，也会产生类似的确定性效果。

然而在实际中，我们只能获取观测信号 $x$，而无法获取其形成成分。因此，所提出的随机化无法仅直接应用于 $t$，整个观测信号都会受到影响。但下文将表明，它仍能达到预期效果：

$$\lambda \longrightarrow \lambda'_i = \lambda(1 - \frac{v'_i}{c}) \tag{7}$$

$$\begin{aligned} x'_i(\lambda) &= \mathcal{V}\left\{x_i(\lambda), v'_i\right\} \\ &= \mathcal{V}\left\{\Big(s_i(\lambda)t_i(\lambda)\Big) * h(\lambda) + n_i(\lambda), v'_i\right\} \\ &= \mathcal{V}\left\{\Big(s_i(\lambda)t_i(\lambda)\Big) * h(\lambda), v'_i\right\} + \mathcal{V}\left\{n_i(\lambda), v'_i\right\} \end{aligned} \tag{8}$$

根据附录 A中给出的证明，可得：

$$\begin{aligned} x'_i(\lambda) &= (1 - \frac{v'_i}{c})\mathcal{V}\left\{s_i(\lambda)t_i(\lambda), v'_i\right\} * \mathcal{V}\left\{h(\lambda), v'_i\right\} + \mathcal{V}\left\{n_i(\lambda), v'_i\right\} \\ &= (1 - \frac{v'_i}{c})\Big(s_i(\lambda'_i)t_i(\lambda'_i)\Big) * h(\lambda'_i) + n_i(\lambda'_i). \end{aligned} \tag{9}$$

现在，由于对于每个 $x_{ni}$， $\left(1 - \frac{v'_i}{c}\right)$ 都是一个常数系数，并且在传递到方法的下一步之前会被归一化消除，因此可以看出，模拟随机速度已从 $x_i$ 转移到了 $t_i$。换言之，我们通过修改 $x_i$，在 $t_i$ 中实现了所需的速度随机化。现在可以将经过调整的信号中的两个主要成分重写为：

$$\begin{aligned} &\text{Telluric:} && \text{Stellar:} \\ &\mathcal{V}\left\{\mathring{t}_i(\lambda), v'_i\right\} && \mathcal{V}\left\{\mathring{s}_i(\lambda), v_i + v'_i\right\} \end{aligned} \tag{10}$$

因此，在检视不同观测数据时，我们实现了两个几乎彼此独立移动的分量。图 3展示了这一效果的实际表现。

从统计学角度来看，通过上述

速度随机化实现的效果，是对地球大气随机过程进行一定程度的“白化”，$\{T\}$。更准确地说，我们降低了每一对$(T^l,\ T^m)$之间的互相关，使地球大气成分的统计行为趋近于白噪声（Li & Zhang 1998；Eldar & Oppenheim 2003；Kessy et al. 2018）。再次注意，$(T^l,\ T^m)$是我们试图去相关的随机变量，它们各自定义在光谱的单个像素上，其取值会随着不同观测而变化。图4展示了上述白化过程的可视化结果。

### 3.2 深度特征空间

我们借鉴并采用了Sedaghat 等人（2021年）——图2中介绍的一维卷积自编码器的完全相同架构。自编码器的卷积层将输入光谱降维至网络瓶颈层预先设定数量的“潜变量”。这种输入的低维表示也称为“编码”，是输入光谱最压缩的形式。该向量的维度根据不同实验所需的压缩率来确定。不过需要注意的是，在我们网络的变分自编码器版本中（本研究的大多数实验均采用该版本），潜在节点以概率方式实现，每个节点都用一对标量建模：正态分布的均值和标准差。在瓶颈层的另一侧，解码器接收压缩后的编码，并逐步将其扩展至与原始输入相同的维度（$2^{18}+2^{16}$适用于 HARPS）。我们采用原始研究中介绍的相同逐像素 L1 端到端损失函数，以实现可接受的像素级精度。

这种典型架构已证明能够将输入转换到一个空间中，使信号中的类噪声成分易于分离。正如我们的实验所示，上一节提出的新型统计技术在学习到的特征空间中，将地球大气成分推得离恒星特征更远、离噪声更近，从而使其像噪声一样易于剔除。

当然，我们需要通过较高的压缩率以及瓶颈层处的潜变量损失来限制网络的重建能力，以免网络过于强大，能够同时拟合这两个独立成分。

在 HARPS 数据上训练一个典型网络 100 个观测时期（即训练阶段中网络遍历所有训练样本的次数），使用一块 Nvidia V100[5]，大约需要 60 小时。当验证损失趋于平稳时，我们便停止训练。

## 4 数据

我们使用来自两个公开数据集的资料：HARPS 和斯隆数字巡天（SDSS）。每个数据集都需要一个从头开始训练的专属网络[6]。我们的主要实验基于 HARPS 开展，而 SDSS 则作为一个较难的数据集用于比较。

与典型的机器学习实现不同，我们的方法源自档案数据探索，从技术上说并不需要推广到未见过的数据：如有需要，可以在一整套光谱上对其进行训练，目的是处理同一套数据；这套数据已在训练阶段使用过。对于这一特定应用，由于没有针对地球大气吸收线剔除这一高层任务的目标，网络不会针对该任务过拟合。不过，我们确实希望该方法能够推广到未见过的数据。因此，对于每个数据集，我们都会留出一小部分样本，不将其纳入训练集，用于验证，并确保网络没有对较低层次的重建任务过拟合——也就是说，没有记住这些光谱。训练集-验证集划分与我们的编码一并发布。

[5] 一张配备 16GB 图形处理器内存的显卡。[6] 我们宽泛地使用“数据集”一词，指由特定仪器（如 HARPS）或巡天项目（如 SDSS）采集的所有光谱。因此，我们在 HARPS 数据上训练的模型，也适用于 HARPS 在训练集采集日期之后拍摄的任何新恒星光谱。

#### 4.1 HARPS

HARPS 是安装在 3.6米拉西拉望远镜上的仪器。它是一台光纤馈入式高分辨率阶梯光栅光谱仪，专用于发现系外行星，分辨率为 $R = 115{,}000$，覆盖 378–691 nm 的光谱范围[7]。所用数据已经过仪器校正、探测器校正和天空背景扣除。

我们下载的数据集最初包含共计 272376 条光谱；自动移除损坏文件、未定义值（NaN：非数值）和类似噪声的光谱后，数量减少至 267361 条“稳定”光谱。该数据集主要由恒星光谱构成。不过，其中也有一些随机、不规则目标天体的光谱，例如 SUN、MOON 等；我们特意允许它们进入训练集，以提高所学习特征的鲁棒性。在将所有光谱输入网络之前，我们对其进行了裁剪、重新网格化和均一化，处理方式与 Sedaghat 等人 (2021年) 相同；最终得到 327680 个像素，分辨率均为 0.01 埃。

请注意，HARPS 光谱经过的数据处理流程经过专门设置，会在光谱存入档案库之前自动将其转换到日心参考系并重新网格化。最初采集的地心参考系版本也未予保留。因此，我们不得不将其转换回地心参考系，以用于我们的实验。尽管这涉及我们所提出方法的核心概念，但事实证明这是一个安全的操作：转换到日心参考系会给径向速度带来额外的随机性，这与我们的方法完全兼容。事实上，我们的方法正是受到初期实验中观察到上述现象的启发。

#### 4.2 斯隆数字巡天

我们的第二个数据集由 SEGUE（Yanny 等，2009年）和 SEGUE-2 光谱巡天（Rockosi 等，2022年）的光谱组成，这两个巡天项目均属于规模更大的斯隆数字巡天（SDSS）。SEGUE 巡天均使用位于阿帕奇角天文台的斯隆基金会2.5米望远镜（Gunn 等，2006年），并配备两台最初的斯隆数字巡天光纤光谱仪（Smee 等，2013年）。SDSS 光谱仪的分辨率为 ~ 1800，两台仪器总共配有 640 根光纤（每台 320 根），这些光纤会插入每次观测使用的铝制“插塞板”。每块观测板上有 32 个插孔留作空白天空观测，另有 16 个用于观测视场中的光谱光度标准星。每台光谱仪均设有红通道和蓝通道，分别通过独立的电荷耦合器件采集数据；蓝通道的波长范围约为 3800 埃至 6100 埃，红通道的波长范围约为 5900 埃至 9200 埃。最初的 SEGUE 巡天项目观测了距离、颜色和金属丰度各异的银河系恒星，而 SEGUE-2 的目标则集中于银河系晕中的恒星。

在我们的实验中，我们使用了 ~660,000 条来自其中一台光谱仪红光电荷耦合器件的未校准光谱（在斯隆数字巡天中标记为“r1”）

[7] http://archive.eso.org/wdb/wdb/adp/phase3_main/form

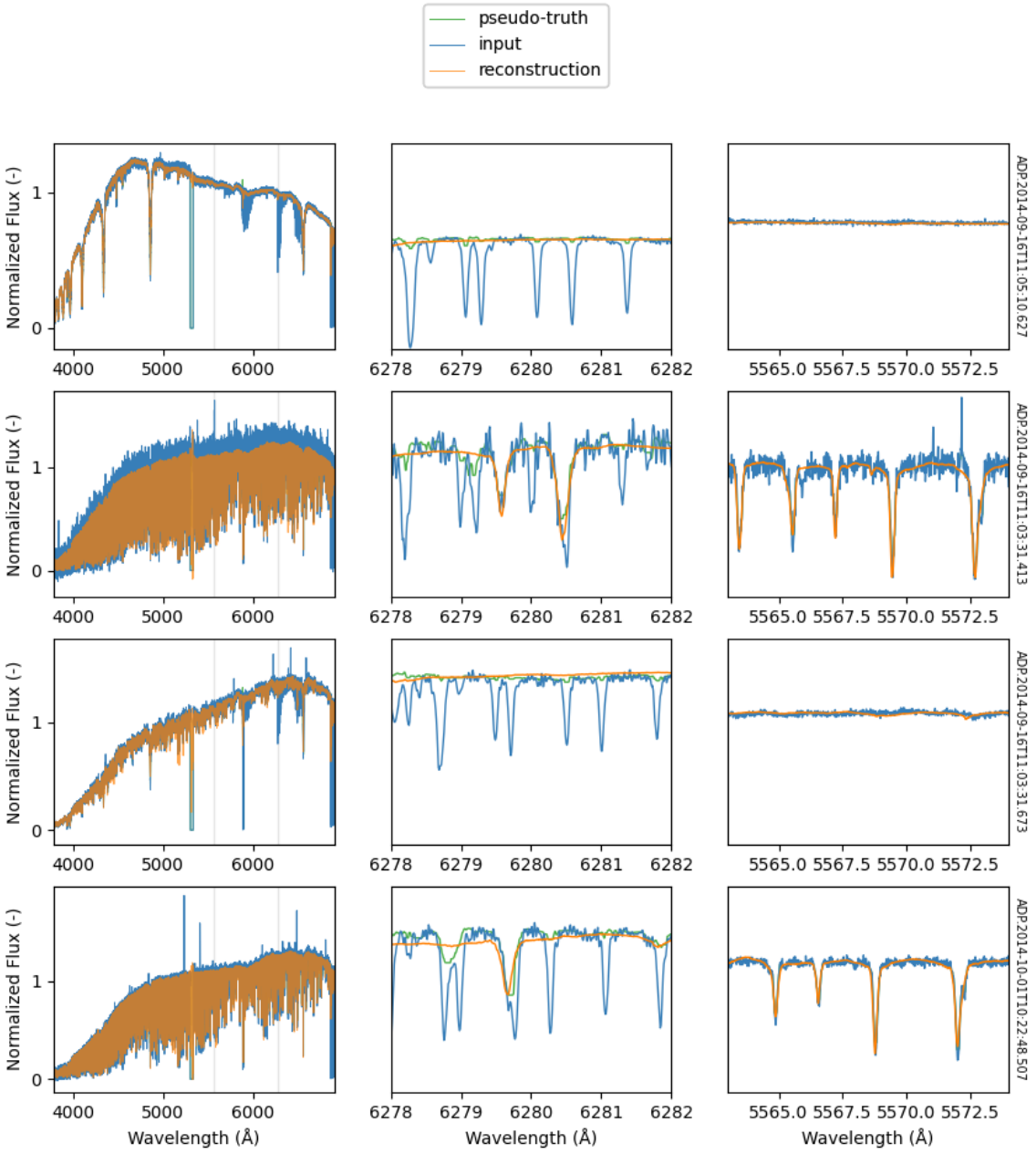


**图 5**。 从一组具有代表性的 HARPS 光谱中选取样例，对结果进行定性说明。每一行展示一条光谱，其 HARPS编号标注在右侧；各列则聚焦于不同的关注区域（左侧以灰色区域标出）。解读提示：重建结果（橙色）应贴合伪真值（绿色——定义见第 5.2 节）。最左列涵盖输入网络的完整光谱，展示了网络对光谱不同特征（连续谱、噪声水平等）的鲁棒性。较不复杂的样例中，可以轻易辨认出 $H\alpha$ 等主要恒星谱线。中间列放大了一个可能较复杂的区域：若存在狭窄的恒星谱线，它们可能会与形状相似的地球大气吸收线重叠。例如，第二行和第四行就展示了这类情况：网络去除了地球大气成分，同时很好地保留了恒星部分，因此可以排除这样的假设：网络只是通过移动平均来去除狭窄谱线。第三列聚焦于这样一个区域：某些光谱中有恒星谱线，而另一些光谱中没有。由此可以排除网络可能已经“记住”谱线位置的假设。

随 SDSS 第 17 次数据发布一同提供的数据档案库（Abdurro'uf 等，2022年），可通过 DR17 FITS 网站访问[8]。未校准光谱由每个天体源的多次 10–30 分钟曝光组成。由于 SEGUE 巡天对每个天体源进行了多次成像以生成合并光谱，因此我们的数据集中同一天体源有多条光谱。对于每块观测板，我们会排除那些有意指向空白天空区域、并在数据中标记为“SKY”的光纤。最后，与 HARPS 的情况一样，我们会先对光谱进行均一化，再将其输入网络。不过，由于此处的分辨率较低，每条光谱最终有 2048 个像素，统一分辨率为 1.66 埃。

所用的未校准光谱在 SDSS 数据模型中标记为“spFrame”[9]，根据 SDSS 光谱处理流程的详细说明（Stoughton 等，2002年），这些光谱经过平场校正，但未进行流量定标。流量定标光谱（数据模型中的“spFluxCalib”）包含由光谱处理流程根据每块观测板观测集中的光谱光度标准星计算得出的地球大气吸收。这些定标结果被用作比较网络结果的“伪真值”。

## 5 实验与结果

我们基于以下（超）参数的多种组合训练并评估多个网络：

- 潜空间维度
- 速度随机化程度
- 变分损失的使用

HARPS 数据集是我们的主要研究对象，我们的主要实验（包括超参数扫描和受控测试）均仅在该数据集上进行。SDSS 数据集仅用于对该方法进行对比测试，以评估其在极端条件下的表现——即分辨率非常低的情况。我们还在第 `wobble`（Bedell 等（2019））节中将我们的结果与名为 5.6 的工具进行比较。

此外，除非另有说明，网络的潜维度均为 128；随机速度取自均匀分布，其设定大致参考了地球在重心坐标系中的切向速度：

$$V \sim U(-30\,\mathrm{km/s}, 30\,\mathrm{km/s}), \tag{11}$$

测试集为 HARPS 光谱[10]的一个固定子集，用于计算定量指标的感兴趣区域（ROI）为 $[5800, 6910]$ Å，该区域内恒星和地球大气吸收线的密度都很高。

### 5.1 主要结果

图 5 展示了几个示例，说明该方法如何剔除地球大气吸收线并保留恒星特征。该示例
这些光谱和区域经过专门挑选，用来排除网络如何拒绝地球大气吸收线的潜在简单假设。换言之，结果证实网络已经学会了对各组成部分的语义表征，并在该特征空间中分离各个天体源；而不是例如仅仅“漏掉”较窄的谱线（简单平均、低通滤波）。在图注中

[8] https://data.sdss.org/sas/dr17/
[9] https://data.sdss.org/datamodel/
[10] 100个选择天体的光谱

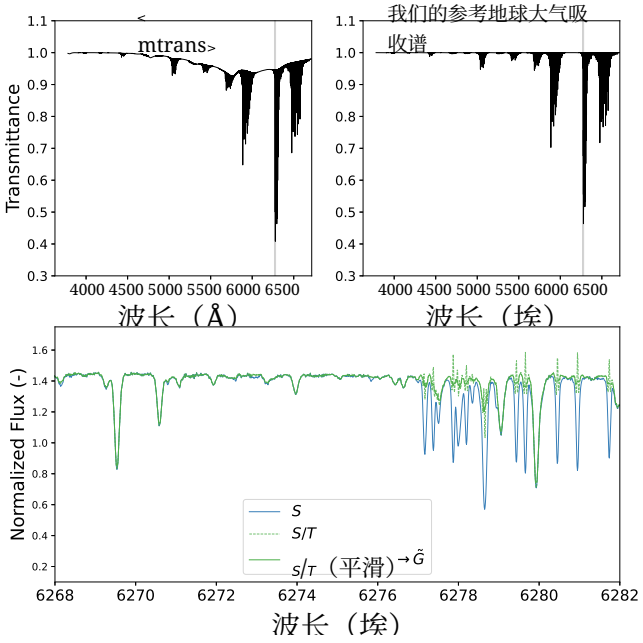


**图6.** 左上图展示了飞马座51的一次观测所对应的 mtrans 信号（HARPS ADP ID：ADP.2014-10-06T10:07:52.453）。右上图为同一信号去除缓慢变化的连续吸收分量后的结果，该结果构成了我们的参考地球大气吸收谱（T）。下图展示了生成伪真值光谱的三个不同步骤：输入光谱（S）、将其除以地球大气吸收谱后的结果（S/T），以及最终的平滑版本。

图 5还详细阐述并排除了其他几个类似的假设。本图示中使用的伪真值将在下一小节中说明。

### 5.2 定量评估

各种（超）参数，例如降维过程中的压缩程度或速度随机化程度，都可能影响网络去除地球大气吸收线的能力。地球大气吸收线去除与恒星重建之间也往往存在取舍。

因此，我们提出以下相互关联的度量指标，用于同时量化这两个方面：

$$Q^t = \frac{\sum\left(M^t\left|\hat{S}-\tilde{G}\right|\right)}{N^t} \tag{12}$$

$$Q^s = \frac{\sum\left(M^s\left|\hat{S}-\tilde{G}\right|\right)}{N^s} \tag{13}$$

其中，$Q^t$ 可作为地球大气吸收线剔除效果的指标，而 $Q^s$则反映恒星重建效果。由于二者都是距离度量，数值越低越好。$\hat{S}$ 表示重建光谱（即网络的直接输出），$M^t$ 是一个二值掩码；它是一个与输入大小相同的向量，在所有包含已知 地球大气吸收线的像素处取值为 1，其他位置均为 0。$M^s$ 是地球大气吸收线掩码的补集，只需将其取反即可得到：$M^s = 1 - M^t$

**˜伪真实基准**， $G$是我们定量评估的关键组成部分。对于上述公式，理想的真实基准应是一条在除地球大气吸收线之外的所有方面都与输入光谱完全匹配的光谱：红移、发射、吸收、CCD噪声和镜面缺陷均相同，但不含 地球大气吸收线。数据集中的每个天体都需要这样一条理想光谱。对于真实 光谱而言，几乎不可能获得这样一组信号。还要注意，仅提供地球大气吸收线的波长列表并不足以实现这一目的，因为按照上述公式，我们计算的是逐像素距离。因此，我们基于 HITRAN 提供的谱线数据库定义伪真实基准，作为一种近似理想的真实基准（Rothman 等，2009年）。我们利用`molecfit` 软件包提供的工具访问数据库，并将每条光谱的观测参数以及估算的分子密度纳入考量。具体而言，`mtrans` 组件由`molecfit` 生成，其中封装了针对当前观测所实现的大气透射光谱——此类光谱示例见图6。我们通过高通滤波器处理信号，去除缓慢变化的连续谱吸收分量（Rosenkranz，1998年；Baranov 等，2008年），[11] 以获得地球大气吸收线的参考轮廓。接下来，为得到 $\tilde{G}$，我们先用输入光谱除以得到的地球大气吸收线参考轮廓，再对结果应用中值平滑核。这样做是为了去除噪声，以及除法运算不稳定导致谱线线心处出现的尖峰。

一个导致 $\tilde{G}$ 的过程示例见图6。

**归一化因子**、 $N^t$ & $N^s$ 在公式（12）和（13）中进行选择，使得即使执行跨区域或跨数据集比较时，度量指标仍具有意义。具体而言， $N^t$ 的获得方式是将重建信号替换为输入，并计算一个未归一化的距离度量：

$$N^t = \sum \left( M^t \left| S - \tilde{G} \right| \right). \tag{14}$$

其基本思想是，最糟糕的结果是完全没有去除地球大气吸收线，即使用未经处理的输入信号。当然，这并不是一个十分准确的假设，因为存在一些不规则情况，可能会使度量指标超过 1。对于此类情况，我们简单地将该值截断为 $[0, 1]$。对于 $N^s$，我们计算连续谱的估计值 $\tilde{C}$，即 $G$，以表示所有恒星谱线都被剔除时的最坏情况，并将其代入未归一化的度量指标公式[12]：

$$N^s = \sum \left( M^s \left| \hat{C} - \tilde{G} \right| \right). \tag{15}$$

这样，归一化后的度量指标大多不受所考虑感兴趣区域（ROI）长度或其中谱线密度的影响。不过请注意，如前所述，这些度量指标并非旨在适用于无线谱线区域等不规则选区。在本小节的所有公式中，为简洁起见，表示观测光谱的下标$i$在 $S$、$\hat{S}$、 $\tilde{G}$ 、$\hat{C}$、 $M^s$ 和$M^t$中均省略，求和则针对每条光谱的所有像素进行。

[11] 这段“连续谱归一化”编码代码片段可在本项目公开提供的代码仓库中找到。[12] 这段“连续谱估计”代码片段可在本项目公开提供的代码仓库中找到。

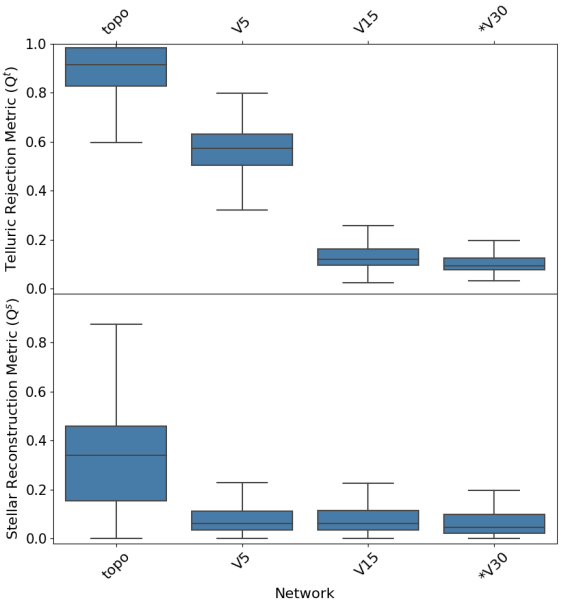


**图 7.** 白化程度（速度随机化）与大气吸收线剔除质量之间相关性的示意图。最左侧项目（“topo”）的随机化设为零，右侧项目的随机化程度依次增加。数字表示随机速度的极值（$\mathcal{N}$，见公式（16）），单位为 km/s：例如，V15 表示 $\mathcal{N}$ = 15 km/s，其他情况依此类推。最右侧项目以 ⋆标示，是我们的主网络，其大气吸收线剔除质量也已趋于稳定。请注意，各随机化网络的恒星重建质量没有显著差异。“topo”的恒星重建质量方差较大，这或许是因为网络的部分学习能力用于学习重建地球大气吸收线。

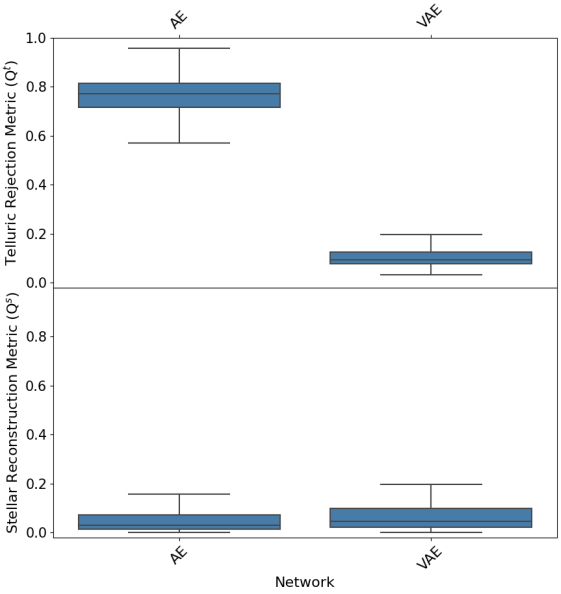


**图 8.** 确定性自编码器（AE，左）与其变分版本（VAE，右）的定量比较。VAE 中潜变量损失的权重设为 0.1。

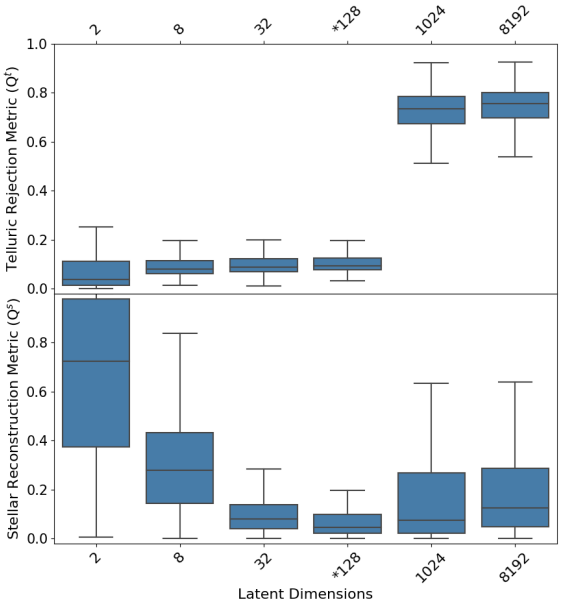


**图 9.** HARPS 上恒星谱线重建质量与地球大气吸收线剔除效果的关系，比较编码（即潜在表示）的不同维度数。两幅图中，数值越低越好。潜维度达到 1024 及以上的网络显然已经开始学习重建地球大气吸收线。综合考虑 $Q^t$ 和 $Q^s$，我们选择了大致处于最佳点的主网络（*128）。

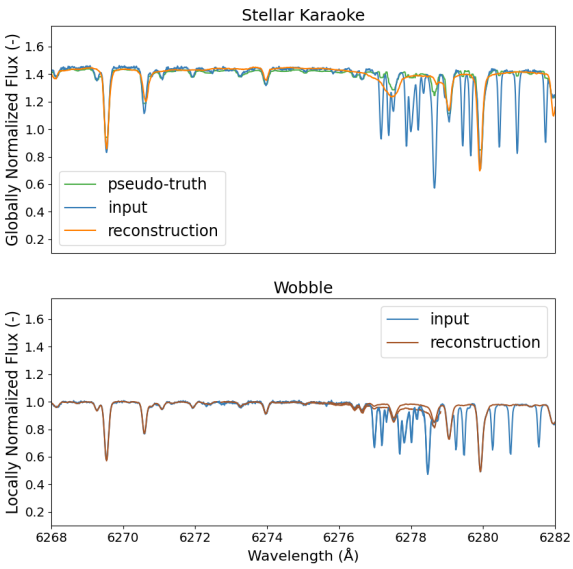


**图 10.** 将飞马座51的光谱观测分别输入 wobble 和 Stellar Karaoke，并对结果进行比较。绿色部分为本方法使用的伪真值。图中所示的感兴趣区域与我们其他可视化中使用的区域相同，同时包含恒星和地球大气吸收线。为此，我们需要合并 wobble 的两个独立输出（光谱阶）：第 62 和第 63 阶。

### 5.3 速度白化的影响

为展示白化步骤的效果，我们在受控条件下训练并比较多个网络。我们首先使用拓扑中心参考系中的光谱训练一个网络，这相当于零速度随机化。随后逐步提高随机化程度，继续训练网络。具体而言，我们将 $V$，式（6）定义为服从均匀分布的随机变量：

$$V' \sim U(-\mathcal{N}, \mathcal{N}) \tag{16}$$

其中 $\mathcal{N}$ 表示允许的随机速度绝对值上限。我们在三个不同网络中分三步增大 $\mathcal{N}$，同时保持其他配置不变。图 7 显示，随机化程度与地球大气吸收线剔除效果之间存在明显的相关性。

### 5.4 变分自编码器与确定性自编码器

确定性（即非变分）自编码器通常能提供更好的重建质量，但代价是会学到纠缠特征——例如，参见 Sedaghat 等，2021年。本研究的重点并非研究解纠缠特征。但通过实验我们发现，VAE 组件能充当网络的正则化器，防止网络学习拟合较深的地球大气吸收线——也就是说，防止网络将其记忆下来。图 8 比较了确定性自编码器（AE）与对应的 VAE。与预期一致，AE 的重建质量略好，但在地球大气吸收线剔除方面则明显不如 VAE。

### 5.5 压缩的影响

可以合理地预期，而且我们的实验也证实了这一点：随着网络信息瓶颈处的维度降低，网络在压缩过程中保留多个并发特征的能力会减弱。这会导致重建光谱中的细节被剔除得更多，表面上看似地球大气吸收线剔除效果有所改善——$Q^t$有所降低。不过需要注意的是，恒星重建质量也可能同时变差，因此需要同时考量这两个度量指标。

图 9 比较了潜在空间中具有不同维度的多次运行结果。请注意，随着压缩程度增加（潜在维度降低），恒星重建质量开始下降。另一方面，随着潜在节点数量增加，网络开始变得过于强大，能够同时重建两个组成部分，从而失去源分离能力。

### 5.6 与wobble的比较

用于拟合 `wobble` （Bedell 等（2019））观测光谱的模型包含两个主要组成部分：恒星与地球大气。从这个意义上说，它是与 Stellar Karaoke 进行比较的最佳候选方法之一。然而，它具有天体特异性，即利用同一天体的多次观测来优化专门针对该天体的模型。这使得可用的 `wobble` 结果仅限于少数天体，无法在大量光谱上进行具有统计意义的比较。此外，例如对于 HARPS，`wobble` 会针对对应于不同阶梯光栅级次的小波长范围拟合 72 个不同的模型。

这些因素使得难以对结果进行同类比较。尽管如此，在图10 中，我们展示了两种方法在飞马座51上的视觉对比[13]。无法进行定量比较

[13] 据我们所知，要精确对应 HARPS 编号与 wobble 所用的光谱并非易事。在这项比较中，我们选取了 ADP.2014-10-06T10:07:52.453，它只是飞马座51的一次观测。同样，我们使用了 wobble 已发表的预计算结果中的第一个“观测时期”。

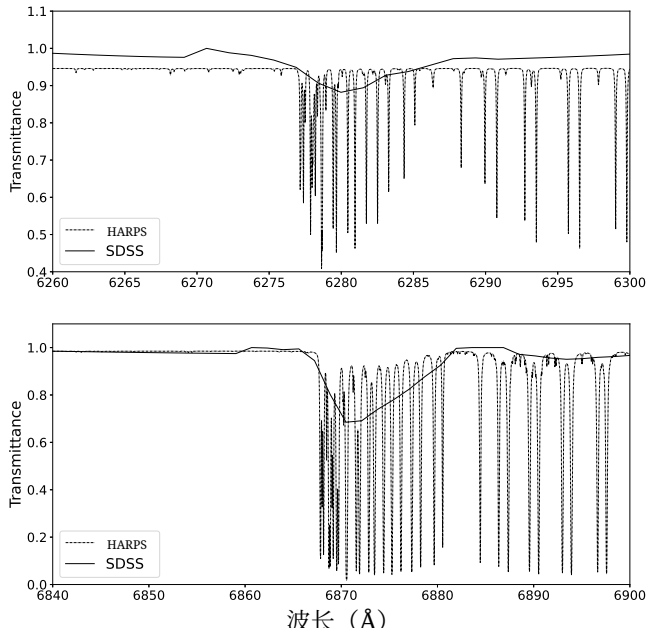


**图11.** 在高分辨率数据集（HARPS）和分辨率较低的数据集（斯隆数字巡天）中，对地球大气吸收特征进行了比较。两幅图展示了来自两个数据集中地球大气吸收特征丰富的重叠区域的地球大气吸收参考谱的两个不同区段。低分辨率光谱丢失了这些特征中的大部分“信息”。在斯隆数字巡天中，地球大气吸收谱的获取方式与 HARPS 相同，不同之处在于我们直接使用巡天项目提供的流量定标，而不是通过`molecfit` 软件包提取。详情请参见第4.2 节。

这是可行的，因为 wobble 会对通量值进行逐阶归一化并执行连续谱归一化，而我们的方法采用全局中位数归一化。不过，从定性来看，与伪真值相比，Stellar Karaoke 和 wobble 都去除了大部分地球大气吸收线，例如 6278 埃至 6281 埃之间的吸收线。在较浅的吸收线处，例如 6271 埃至 6276 埃之间，wobble 比 Stellar Karaoke 更能贴合输入光谱。在 6278 埃附近的区域，可以看到 wobble 中两个光谱阶的重叠[14]。

除了上述方法上的差异之外，在假设有多次观测和相关外部信息可用的情况下，`wobble` 拟合一个天体的模型约需 60 分钟（Bedell 等（2019）），而 Stellar Karaoke 只需训练一次，之后便可对任何新光谱进行推理，在单个中央处理器上仅需 50毫秒（若在常见的 Nvidia 图形处理器上运行，则需 4 毫秒）。

### 5.7 HARPS 与斯隆数字巡天：高分辨率与低分辨率

图12 展示了我们的方法在 SDSS 光谱上的训练和应用结果。该数据集的像素分辨率极低（仅有 2K 个像素，间隔为 1.66 埃；相比之下，HARPS 有 300K 个像素，分辨率为 0.01 埃），因此在实际应用中，这对我们的方法构成了压力测试，因为许多地球大气吸收线彼此融合，呈现为宽阔的伪影——参见图 11 进行比较。因此，Stellar Karaoke 无法找到并学习到对地球大气吸收线非常有意义的表征：即使恒星谱线得以恢复，且部分地球大气吸收线被剔除，网络仍会学习并重建较显著的地球大气伪影（即不会将其剔除）。由于第5.2 节介绍的量化指标主要围绕地球大气“吸收线”这一概念设计，因此在这种情况下，这些指标的数值并无意义，我们认为定性比较已足以得出结论。

### 5.8 与`molecfit`的比较

`molecfit` 是一种广泛使用的工具，用于对光谱观测中的大气吸收进行建模和校正，尤其适用于 HARPS 等高分辨率观测（Smette 等，2015年；Kausch 等，2015年）。由于我们使用了 `molecfit` 软件包提供的一个组件来访问 HITRAN 数据库以进行验证，因此此处不直接比较 `molecfit` 的结果。不过，我们列出了一些方法上的差异，以突出我们的方法相较于 `molecfit` 的优势。

molecfit 通过将合成透射光谱拟合到观测光谱来运行，因此需要预先了解若干观测时的参数。这些参数包括大气压、温度、湿度、气团质量、观测地点的纬度和经度、海拔、观测时间和日期，以及仪器配置。因此，molecfit 校正的准确性与这些参数的精度密切相关，参数值中的任何不确定性都会传递到校正过程中。该工具从全球数据同化系统（GDAS）网站获取两个大气廓线，这两个廓线对应的时间分别位于观测时间之前和之后。随后，将这些廓线插值至观测时间，并与标准廓线和当地气象数据合并，从而为辐射传输代码生成输入廓线。molecfit 使用辐射传输代码模拟大气发射和透射光谱，并依赖分子光谱数据库。在我们的实验中，molecfit 的“拟合时间”平均为每个输入光谱 3.36 CPU 分钟——这不包括访问数据库等所花费的时间。

相比之下，针对特定仪器的 Stellar Karaoke 网络只需将由标量值组成的原始向量作为输入光谱，并能在不到一秒的时间内完成处理[15]。

## 6 结论与未来方向

我们介绍了一项研究：该研究借鉴大数据视角看待恒星光谱，利用恒星径向速度与其观测光谱中的地球大气吸收线之间的统计独立性，并通过一种新颖技术强化这种独立性，再利用完全无监督的卷积神经网络去除不需要的部分。

我们认为这项成果体现在两个方面：首先，我们报告了一项跨学科科学发现，展示了如何将一个科学领域（数据科学）的理念引入另一个科学领域（天文学），从而带来新的见解并开辟新的道路。其次，与传统的地球大气谱线去除工具相比，我们提出了一种方法，或其原型，为进一步开发提供了颇具前景的方向。

从实际角度来看，Stellar Karaoke 的优势在于能够识别地球大气吸收线，而无需任何外部信息；无论是现有方法所需的观测时间参数，还是借助已知地球大气吸收线列表进行手动模型初始化，

[14]请参阅 https://github.com/megbedell/wobble/blob/fe32e885eba4704ff4f24345f4f99337d7d81a34/docs/quickstart.rst 了解当前合并不同光谱阶时的限制。

[15] 例如，本文提出的 HARPS 模型已经完成训练，可以去除任何新出现、此前未见过的 HARPS 光谱中的地球大气吸收线。

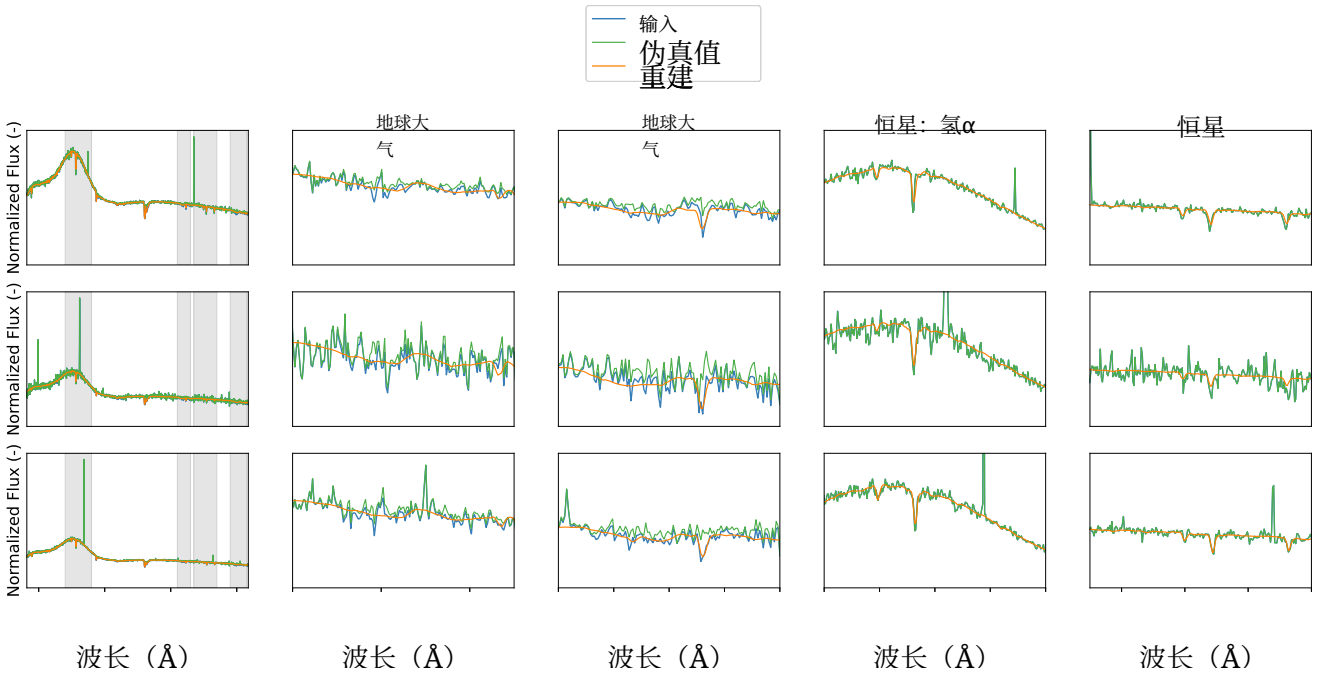


**图12**。 低分辨率数据集的测试结果：斯隆数字巡天（SDSS）。三行中分别展示了三个随机样本。左侧显示完整光谱，其余各列则放大了波长轴上感兴趣的不同区域，每个区域主要包含列顶部所标示的主导特征；这些区域在最左列中以灰色标出。即使在信噪比较差的条件下，恒星特征仍得以保留，但由于分辨率远低于 HARPS，部分较宽的地球大气吸收线未能完全去除。这里看到的尖峰是校准轮廓未能捕捉到的伪影，因此也出现在伪真值中。

又或是在 molecfit 及类似方法中使用的地球大气分子分布。实际上，训练我们的模型只需将大量光谱输入网络，其余工作交给网络完成。这有望降低地球大气吸收线拟合的复杂性，并填补大气和视线测量值缺失或不准确所造成的空白。

Stellar Karaoke 的运行速度极快。只需训练一个网络，就足以处理整个数据集，以及今后加入其中的任何新光谱。训练完成后，它处理每条光谱只需不到一秒。考虑到目前 molecfit 和 wobble 的处理时间分别为数分钟和数十分钟，这一优势尤为显著；而如今的数据集包含数十万条光谱。为凸显这一点，我们使用 Stellar Karaoke 处理了全部约 26 万条公开的 HARPS 数据，并将结果公开；据我们所知，这是前所未有的举措，可能是因为过去所需的时间和计算资源过多。

尽管如此，当前版本仍只是一个新型研究成果的演示，展现了完全无监督方法的优势。若要让 Stellar Karaoke 发展成为适用于各种应用的即用型工具，还需要进一步改进。值得注意的是，由于这项研究采用无监督方式，网络目前选择了舍弃地球大气成分并重建恒星成分。这当然是由这两种现象的统计特性所致（恒星谱线中的显著性更高，等等）。不过，为了使其成为一种更稳健、对所涉及超参数不那么敏感的工具，可以在训练过程中引入一种“提示”，引导网络关注所需的成分——例如，分解后各成分之间的互信息。

另一个值得探索的方向是离群值的情况。与任何其他方法一样，在某些情况下，该方法的表现会不尽如人意——要么完全失效，要么效果明显不佳。在这种情况下，适度的监督（例如手动将稀有样本加入训练集）可能会有所帮助。

**致谢**

本文展示的部分实验是在 NS 于2019年至2021年参与 ESCAPE 项目期间，使用欧洲南方天文台提供的计算服务器完成的。此外，作者1至4感谢华盛顿大学天文学系 DiRAC 研究所提供的支持。DiRAC 研究所的经费来自查尔斯和丽莎·西蒙尼艺术与科学基金以及华盛顿研究基金会的慷慨捐赠。

**数据可用性**

我们已发布速度白化和卷积神经网络的编码，以及来自两个数据集的光谱标识符列表，详见 https://github.com/NimSed/stellar-karaoke。方法的示例实现和输出，以及未来可能推出的版本，将公开发布在项目网页上：https://www.rawdataspeaks.com/projects/stellar-karaoke。

**参考文献**

Abazajian K. N., 等，2009年，《天文学杂志增刊》，182, 543

Abdurro'uf 等，2022年，《天文学杂志增刊》，259, 35 Allart R. 等，2022年，《天文学与天体物理学》，666, A196 Alonso-Floriano F. J. 等，2019年，《天文学与天体物理学》，629, A110 Artigau É. 等，2014年，载于《天文台运行：策略、流程与系统 V》，第 914905 页 Bailer-Jones C. A., Irwin M., Von Hippel T., 1998年，《皇家天文学会月刊》，298, 361 Bailey J., Simpson A., Crisp D., 2007年，《太平洋天文学会出版物》，119, 228 Baranov Y. I., Lafferty W. J., Ma Q., Tipping R., 2008年，《定量光谱学与辐射传输杂志》，109, 2291 Bedell M., Hogg D. W., Foreman-Mackey D., Montet B. T., Luger R., 2019年，《天文学杂志》，158, 164 Bengio Y. 等，2009年，《机器学习基础与趋势》，2, 1 Bohlin R. C., Gordon K. D., Tremblay P.-E., 2014年，《太平洋天文学会出版物》，126, 711 Bolton A. S. 等，2012年，《天文学杂志》，144, 144 Catanzaro G., 1997年，《天体物理学与空间科学》，257, 161 Choi S., Cichocki A., Park H.-M., Lee S.-Y., 2005年，《神经信息处理：通讯与评论》，6, 1 Comon P., 1992年，独立成分分析 Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D., 1995年，《天文学杂志》，110, 1071 Dieleman S., Willett K. W., Dambre J., 2015年，MNRAS，450, 1441 Eldar Y. C., Oppenheim A. V., 2003年，《IEEE 信息论汇刊》，49, 1846 Fukushima K., 1980年，《生物控制论》，36, 193 Ghosh A., Urry C. M., Wang Z., Schawinski K., Turp D., Powell M. C., 2020年，《天体物理学杂志》，895, 112 Grais E. M., Plumbley M. D., 2017年，载于 2017 年 IEEE 全球信号与信息处理大会（GlobalSIP），第 1265–1269 页 Griffin R. E., 2005年，《太平洋天文学会出版物》，117, 885 Gullikson K., Dodson-Robinson S., Kraus A., 2014年，《天文学杂志》，148, 53 Gunn J. E. 等，2006年，《天文学杂志》，131, 2332 Hase F., Wallace L., McLeod S. D., Harrison J. J., Bernath P. F., 2010年，《定量光谱学与辐射传输杂志》，111, 521 Hinton G. E., 2006年，《科学》，313, 504 Hinton G. E., Salakhutdinov R. R., 2006年，《科学》，313, 504 Hoyle B., 2016年，《天文学与计算》，16, 34 Hrudková M., Harmanec P., 2005年，《天文学与天体物理学》，437, 765 Huertas-Company M. 等，2015年，《天文学杂志增刊》，221, 8 James G., Witten D., Hastie T., Tibshirani R., Taylor J., 2023，在《统计学习导论：Python 应用》中，施普林格，第 503–556 页 Jolliffe I. T., Cadima J., 2016年，《皇家学会哲学汇刊 A 辑：数学、物理与工程科学》，374, 20150202 Kameoka H., Li L., Inoue S., Makino S., 2018年，arXiv 预印本 arXiv:1808.00892 Kausch W. 等，2015年，《天文学与天体物理学》，576, A78 Kessy A., Lewin A., Strimmer K., 2018年，《美国统计学家》，72, 309 Kingma D. P., Welling M., 2014年，arXiv:1312.6114 [cs，统计]Li G., Zhang J., 1998年，Sankhy¯a：《印度统计学杂志》A 辑，第 119–133 页 Liu J.-Y., Yang Y.-H., 2018年，载于 2018 年第 17 届 IEEE 机器学习与应用国际会议（ICMLA），第 773–778 页 Livingston W., Wallace L., 1991年，《红外太阳光谱图集》，涵盖 1850 至 9000 cm-1（1.1 至 5.4 微米）Mayor M. 等，2003年，《信使》，114, 20 McCloat S., von Essen C., Fieber-Beyer S., 2021年，《天文学杂志》，162, 132 Moehler S. 等，2014年，《天文学与天体物理学》，568, A9 Neri J., Badeau R., Depalle P., 2021年，载于 2021 年第 29 届欧洲信号处理会议（EUSIPCO），第 311–315 页 Nicholls C. P. 等，2017年，《天文学与天体物理学》，598, A79 Pal M., Roy R., Basu J., Bepari M. S., 2013年，载于 2013 年东方 COCOSDA 国际会议暨 2013 年亚洲口语研究与评估会议（O-COCOSDA/CASLRE），第 1–5 页 Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019年，《天文学与天体物理学》，621, A26 Portillo S. K. N., Parejko J. K., Vergara J. R., Connolly A. J., 2020年，《天文学杂志》，160，45；Rockosi C. M. 等，2022年，《天体物理学杂志增刊》，259，60；Rosenkranz P. W., 1998年，《无线电科学》，33，919；Rothman L. S. 等，2009年，《定量光谱学与辐射传输杂志》，110，533；Rudolf N., Günther H., Schneider P., Schmitt J., 2016年，《天文学与天体物理学》，585，A113；Rybicki G. B., Lightman A. P., 1986年，《天体物理学中的辐射过程》；Särelä J., Valpola H., Jordan M., 2005年，《机器学习研究杂志》，6；Sedaghat N., Mahabal A., 2018年，《皇家天文学会月刊》，476，5365；Sedaghat N., Romaniello M., Carrick J. E., Pineau F.-X., 2021年，《皇家天文学会月刊》，501，6026；Seki S., Kameoka H., Li L., Toda T., Takeda K., 2019年，《IEEE Access》，7，168104；Sharma K., Kembhavi A., Kembhavi A., Sivarani T., Abraham S., Vaghmare K., 2020年，MNRAS，491，2280；Shulyak D. 等，2019年，《天文学与天体物理学》，626，A86；Simon K., Sturm E., 1994年，《天文学与天体物理学》，281，286；Smee S. A. 等，2013年，《天文学杂志》，146，32；Smette A. 等，2015年，《天文学与天体物理学》，576，A77；Soffel M. 等，2003年，《天文学杂志》，126，2687；Stoughton C. 等，2002年，《天文学杂志》，123，485；Szegedy C. 等，2015年，载于 IEEE计算机视觉与模式识别会议论文集，第1–9页；Ulmer-Moll S., Figueira P., Neal J. J., Santos N. C., Bonnefoy M., 2019年，《天文学与天体物理学》，621，A79；Vacca W. D., Cushing M. C., Rayner J. T., 2003年，《太平洋天文学会出版物》，115，389；Wang Y., Yao H., Zhao S., 2016年，《神经计算》，184，232；Wang Y., Yao H., Zhao S., 2016年，《神经计算》，184，232；Wang K., Guo P., Luo A.-L., 2016年，《皇家天文学会月刊》，465，4311；Wisdom S., Tzinis E., Erdogan H., Weiss R., Wilson K., Hershey J., 2020年，《神经信息处理系统进展》，33，3846；Yang T., Li X., 2015年，《皇家天文学会月刊》，452，158；Yanny B. 等，2009年，《天文学杂志》，137，4377

## 附录 A：波长变换与卷积

这里展示了波长轴的缩放如何分配到观测光谱的各个组成部分中——这与式（8）、式（9）有关。

$$x(\lambda) = z(\lambda) * h(\lambda) = \int_w z(w)h(\lambda - w)dw \tag{A1}$$

$$x(\lambda_i') = x(a\lambda) = \int_w z(w)h(a\lambda - w)dw \tag{A2}$$

$$\text{Let } w = au, \tag{A3}$$

$$x(\lambda_i') = a\int_u z(au)h(a\lambda - au)du \tag{A4}$$

$$\text{Let } z(au) = z'(u),\ h(au) = h'(u) \tag{A5}$$

$$
\begin{aligned}
\Longrightarrow x(\lambda_i') &= a \int_u z'(u) h'(\lambda - u) du \\
&= a\, z'(\lambda) * h'(\lambda) \\
&= a\, z(a\lambda) * h(a\lambda) \\
&= a\, z(\lambda_i') * h(\lambda_i')
\end{aligned}
\tag{A6}
$$

使用傅里叶变换也可以用更简短的方式展示同样的内容。然而，为避免混淆在两个不同域中使用的术语频率和波长，这里我们采用卷积算子的直接展开。

**附录 B：全连接架构**

下面展示了一个示例全连接自编码器的可训练参数详情——这是一个与第 1.3 节相关的假设网络，并非 Stellar Karaoke 所使用的网络。即使仅用三个步骤降低 HARPS 光谱的维度，可训练参数的数量也会迅速超过 60 亿；与卷积对应模型相比，差异十分显著。

```
==============================================
层（类型：深度索引）输出形状 参数数量
==============================================
AE1D_FC                  --            --
├-序列模块:      1-1      [1, 1, 128]  --
|   ├-线性层：2-1          [1, 1, 10000] 3,276,810,000
|   ├-ReLU激活函数：2-2      [1, 1, 10000] --
|   ├-线性层：2-3          [1, 1, 1000]  10,001,000
|   ├-ReLU激活函数：2-4      [1, 1, 1000]  --
|   ├-线性层：2-5          [1, 1, 128]   128,128
|   ├-ReLU激活函数：2-6      [1, 1, 128]   --
├-序列模块：1-2           [1, 1, 327680]--
|   ├-线性层：2-7          [1, 1, 1000]  129,000
|   ├-ReLU激活函数：2-8      [1, 1, 1000]  --
|   ├-线性层：2-9          [1, 1, 10000] 10,010,000
|   ├-ReLU激活函数：2-10     [1, 1, 10000] --
|   ├-线性层: 2-11         [1, 1, 327680]3,277,127,680
|   ├-ReLU激活函数: 2-12     [1, 1, 327680]--
==============================================
总参数量: 6,574,205,808 可训练参数
量: 6,574,205,808
```

本文由作者准备的 TEX/LaTEX 文件排版。