arXiv:2305.03761v1 [astro-ph.GA] 5 May 2023

银河系中的弱监督异常检测

M. Pettee, [a] **S. Thanvantri,** [b] **B. Nachman,** [a] **D. Shih,** [c]**M. R. Buckley,** [c] **和 J. H. Collins**[d]

[a]美国加利福尼亚州伯克利市劳伦斯伯克利国家实验室[b]美国加利福尼亚州伯克利市加利福尼亚大学伯克利分校[c]美国新泽西州新不伦瑞克市罗格斯大学[d]美国加利福尼亚州门洛帕克市 *SLAC* 国家加速器实验室[e]美国加利福尼亚州桑尼维尔市博世北美研究院电子邮箱：mpettee@lbl.gov，sowmya.thanvantri@berkeley.edu，bpnachman@lbl.gov，dshih@physics.rutgers.edu，mbuckley@physics.rutgers.edu，jcollins@slac.stanford.edu

摘要。 大规模天体物理学数据集为采用新的机器学习技术识别可能被传统搜索忽略的感兴趣区域提供了机会。为此，我们使用无标签分类（CWoLa）这一弱监督异常检测方法，在盖亚卫星观测到的超过十亿颗银河系恒星中识别冷恒星流。CWoLa 无需使用带标签的恒星流，也不需要了解天体物理学原理。相反，我们训练一个分类器来区分混合样本，而这些样本中信号和背景样本的比例是未知的。这种计算开销较低的策略能够在数据中检测出模拟星流以及已知的 GD-1 星流。该技术最初是为高能对撞机物理学设计的，可能广泛适用于天体物理学，以及其他致力于识别局部异常的领域。

目录

1 引言 **1**
1.1 研究动机 1.2 相关工作 1.3 CWoLa：无 1
标签分类 1.4 论文结构 2
2
3

2 盖亚数据集2.1 数据预处 **3**
理 5

3 方法 **5**
3.1 无标签分类（CWoLa） 3.2 定义信号和边带区域 3.3 5
神经网络架构与训练流程 3.4 模型评估 7
9
10

4 结果 **11**
4.1 GD-1恒星流识别 114.2 GD-1星流标注的扩展方向 14

5 讨论 **15**

6 结论 **16**6.1 复现这些结果 17

A 模拟恒星流 24B 逐块性能 27C 潜在的GD-1候选成员 28

1 引言

1.1 动机

我们的家园星系——银河系——的历史，始终伴随着从宇宙各处汇聚而来的恒星、气体和暗物质不断聚集。这些吸积事件包括与其他星系的并合，也包括恒星以较小尺度形成的引力束缚群体，例如球状星团。尽管许多此类碰撞发生在遥远的过去，较近期碰撞留下的残余仍包含关键信息，可揭示银河系的并合历史 [1–6], 、其背后的引力势 [7–12], 以及暗物质含量 [13–20]。

自 1971 年以来，天文学家观测到了一类名为恒星流的恒星集合：它们是环绕银河系中心运行的纤细带状弧形结构 [21]。这些动力学上较冷的恒星集合被认为是银河系的引力潮汐力扰乱并扭曲附近低质量前身天体——矮星系或球状星团——

直至这些恒星不再受引力自束缚。由于起源相同，这些恒星往往具有许多共同特征，从本动速度到年龄皆是如此。

截至目前，在银河系中已识别出近100条恒星流 [22]。由于其稀疏的密度和宽广的角跨度，发现和研究它们颇具挑战。更深入地了解已知流的结构，并发现更多流，对于加深我们对银河系的认识至关重要。一份涵盖高精度流测量结果的详尽目录，将极大改善我们对银河系大尺度和小尺度结构特征的估计，包括冷暗物质的贡献。

1.2 相关工作

识别恒星流的传统方法是寻找在各种指标上相似的恒星群组：颜色和星等 [23, 24]，速度 [25–27]，或沿着穿过天空的大圆路径分布的位置 [28, 29]。近年来，一种名为 Streamfinder [30]，的自动化技术利用位置和运动学信息，在恒星流候选体周围的多维相空间中构建称为“超管道”的体积区域，然后通过迭代优化一个类似于对数似然的统计量，以确定给定超管道的最佳参数。尽管该算法已成功发现多条低密度恒星流，但它基于天体物理学原理做出了若干假设。例如，它根据特定的等时线选择来确定恒星流的距离，并假设特定的银河系引力势来计算恒星流的轨道。数据挖掘和聚类技术（例如 DBSCAN）也已应用于 盖亚 数据，以搜索恒星流[31]。

机器学习技术也已应用于搜寻恒星流。最近，一种名为 Via Machinae [32, 33]，的无监督机器学习技术 [34] 基于 ANODE 方法开发，该方法最初是为高能粒子物理学搜寻而设计的；Via Machinae 被用于在盖亚 第二次数据发布目录中自动发现恒星流。Via Machinae 将用于异常检测的归一化流密度估计技术 [35] 与寻线算法相结合，无需对等时线或恒星流轨道做出详细假设，即可识别恒星流候选体。

1.3 CWoLa：无标签分类

机器学习已广泛且成功地应用于基础物理学领域，用于分类和描述各种物理现象，其范围从亚原子尺度延伸至宇宙学尺度。这些技术擅长在不对数据集分布施加任何先验假设的情况下识别复杂模式。处理标签部分不准确或不完整的真实数据时，弱监督机器学习方法尤其有帮助。

高能粒子物理学和天体物理学领域都致力于从庞大且高维的数据集中识别局部特征，即集中在相空间有限区域内的数据过密区。与模型无关的异常检测——也就是识别偏离数据集典型特征的这些局部特征——能够以无偏的方式高效筛选这些大型数据集，并助力潜在发现。

在本文中，我们展示了无标签分类 (CWoLa[1]) [36]，这一弱监督机器学习技术在天体物理学中的首次应用。该技术基于一种简单的、

[1]注：CWoLa 的发音是“koala”。

轻量级神经网络分类器。与 ANODE [34], 和 Via Machinae [32, 33], 中使用的神经网络架构类似，CWoLa 最初是为在高能粒子物理学数据集中识别粒子而设计的。因此，CWoLa 已被用作一种颇具前景、与模型无关的异常检测方法，用于搜寻局部特征，例如大型强子对撞机（LHC）中新基本粒子的潜在信号 [37——39],但此前尚未应用于天体物理学数据集。

这项分析使用与 Via Machinae 相同的数据集，并遵循相同的总体框架。我们沿一个自行坐标定义的区域进行扫描，并训练神经网络，使用五个输入变量为这些区域中的恒星分配异常分数：两个角位置坐标、一个自行坐标（即未用于扫描的那个）、星等和颜色。在两项分析中，我们都沿自行方向扫描，因为我们预计恒星流在运动学上较冷，因此相较于背景恒星，在这一特征方向上的分布较为局部；不过，我们并不要求这些恒星流位于特定轨道或恒星等时线上。分配异常分数后，我们会查看每个扫描窗口中异常分数最高的恒星子集，对它们进行聚类，并应用一些基准筛选进一步筛选。我们将这些技术应用于一个名为 GD-1 的已知恒星流以及模拟恒星流，以此作为基准。

尽管存在这些总体相似之处，本分析在某些方面也不同于 Via Machi-nae 。最显著的区别在于，用于分配异常分数的神经网络模型选择以及训练后的聚类方法均有所不同。这些替代性方法选择旨在通过计算开销更低的框架，在异常检测中实现相似的性能。第 5 节将详细讨论这些分析之间的相似之处、差异以及潜在的优缺点。

1.4 概述

本文组织如下。首先，在第 2节，我们介绍 盖亚 数据集，以及如何处理该数据集以用于我们的异常检测研究。接着，在第 3节，我们说明应用 CWoLa 进行异常检测的方法，以及我们在 CWoLa 处理 盖亚 数据时采用的具体实现，包括如何定义信号区域和旁带区域，以及神经网络模型架构和训练流程。第 CWoLa 节列出了将其应用于已知恒星流 GD-1 的结果。最后，在第 4节，我们总结并讨论 5CWoLa在助力未来发现恒星流方面的潜在用途，以及本研究为实现这一目标可以采取的后续步骤。

2 盖亚数据集

盖亚 盖亚 星表是内容丰富的天文数据发布成果，绘制了银河系中恒星的分布图 [40]。盖亚 盖亚 卫星于2013年发射，其数据目录在整个运行期间分阶段发布，直至2025年。本研究分析的数据来自盖亚的第二次数据发布（DR2），收录了盖亚从2014年7月25日至5月 23, 2016 [41]的观测数据。盖亚 DR2包含约13亿颗恒星的位置、自行和测光信息，约占银河系恒星总群体的1%。本分析开展期间，盖亚 又发布了两个数据版本：早期第三次数据发布（eDR3）和第三次数据发布（DR3）。如果使用eDR3或DR3数据扩展本分析，由于测量误差减小，我们可能会看到进一步的改进。eDR3和DR3的其他变化

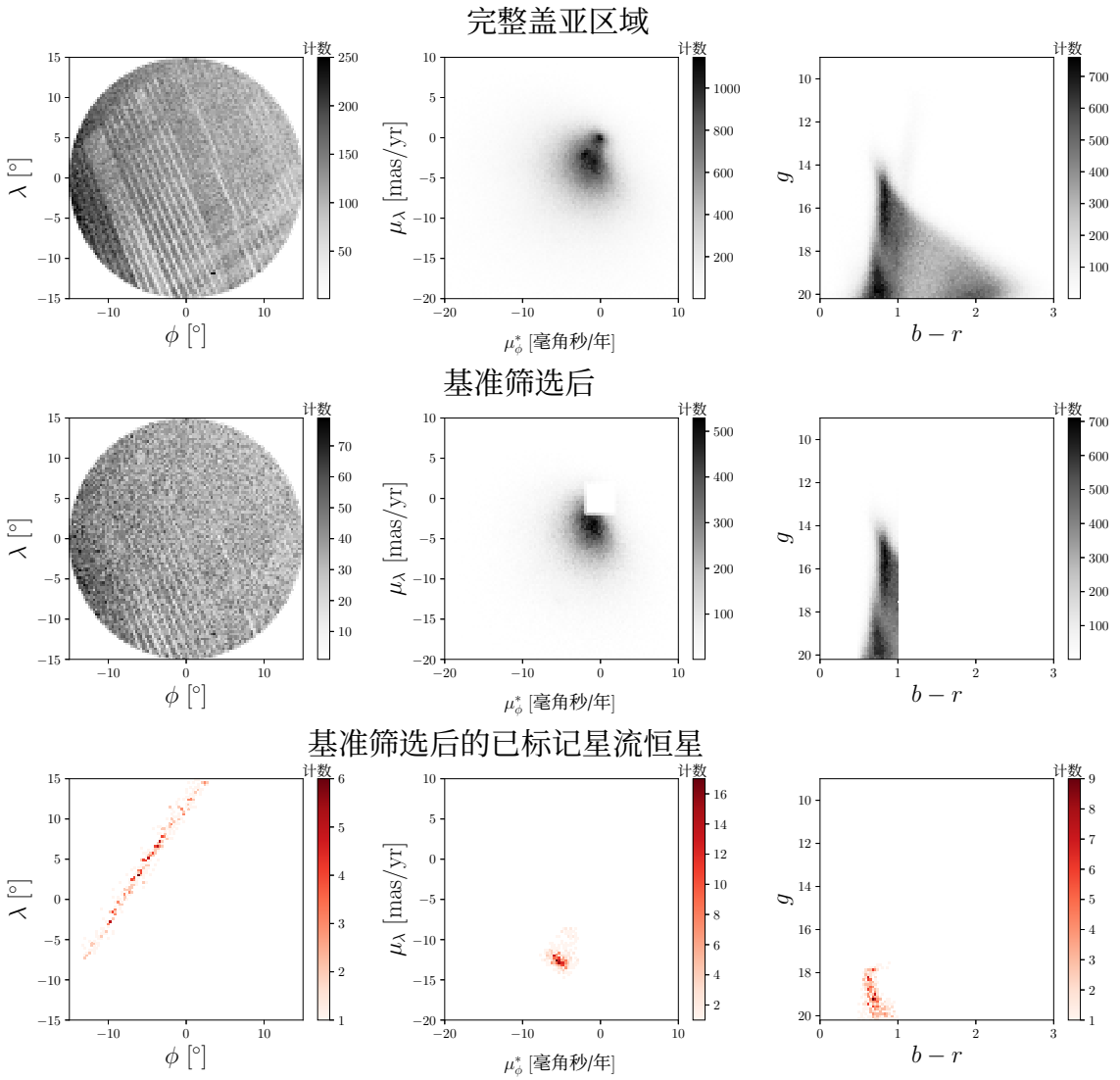


图 1。图示为天空中包含一些 GD-1 恒星的单个天区内，本分析所使用的六个特征的二维直方图。该天区以 ($l = 207.0$, $b = 50.2$) 为中心。第一行显示未应用任何筛选的完整天区。第二行显示应用基准筛选后的天区： $g < 20.2$ 以减少条纹效应； $|\mu_\lambda| > 2$ 毫角秒/年或 $|\mu_\phi^*| > 2$ 毫角秒/年，以移除距离过远的恒星；以及 $0.5 \leq b - r \leq 1$ 以专注于识别冷恒星流。第三行展示应用基准筛选后 GD-1星流的六个特征。

还包括针对少部分恒星改进的距离和径向速度测量，但本分析未考虑这些变量。

虽然盖亚 DR2 也包含视差信息和平均径向速度，但在本分析中，这些变量并未被视为 CWoLa 的输入变量。我们排除视差，因为对于距离如此遥远、也是我们最感兴趣的星流成员而言，与其他可观测量相比，视差并不是那么可靠的特征。不过，我们确实使用视差对盖亚 数据进行筛选，将恒星的最大视差限制为 1，也就是说，恒星距离至少为 1 千秒差距。平均径向速度，即沿着两者之间轴线的运动。

地球与每颗恒星之间的距离，也在 盖亚 数据集中针对部分恒星进行了测量，但 盖亚 DR2 仅包含约 700 万颗恒星的径向速度数据，占整个恒星目录的比例不到百分之一。因此，为了最大限度地利用可用的训练统计数据，我们省略了径向速度这一参数。

与 [32], 一样，我们使用在 2006 [42], 中发现的恒星流 GD-1，作为评估 CWoLa 作为一种星流搜寻技术性能的主要候选对象。GD-1 很可能是一个受潮汐力扰动而瓦解的球状星团的残骸，主要由贫金属恒星组成，总数约为 2×10^4 $M_\odot$ [43]颗。它距离太阳约 10 千秒差距，距离银河系中心约 15 千秒差距。与银河系中的其他恒星流相比，GD-1 尤其狭窄 [42], 、动力学上较冷 [44], 且明亮 [15] 。它还具有多种已知的物理特征，包括间隙和波纹 [45], 、分支（“分叉”）和过密区（“团块”） [16], ，甚至还有环绕其周围的恒星“茧状结构” [46]。因此，对其中一些密度扰动进行忠实重建，可以作为衡量 CWoLa输出在物理层面是否有效的另一项指标。

2.1 数据预处理

我们使用与参考文献 [32] 中相同的盖亚恒星数据集。按照相同的数据处理方法，我们在盖亚数据集中选取一系列半径为 15°、彼此重叠的 21 个圆形“天区”来训练模型。虽然 DR2 中自然的角位置坐标是赤经（α）和赤纬（δ），但我们使用旋转并居中的坐标 ϕ 和 λ （以及旋转后的自行坐标μ_ϕ 和 μ_λ），使每个天区都具有欧几里得距离度量，且中心位于（α_0, δ_0） = （0°，0°）。这一变换使用 Astropy [47–49]完成。每个天区的中心位置也记录在 [32]中。

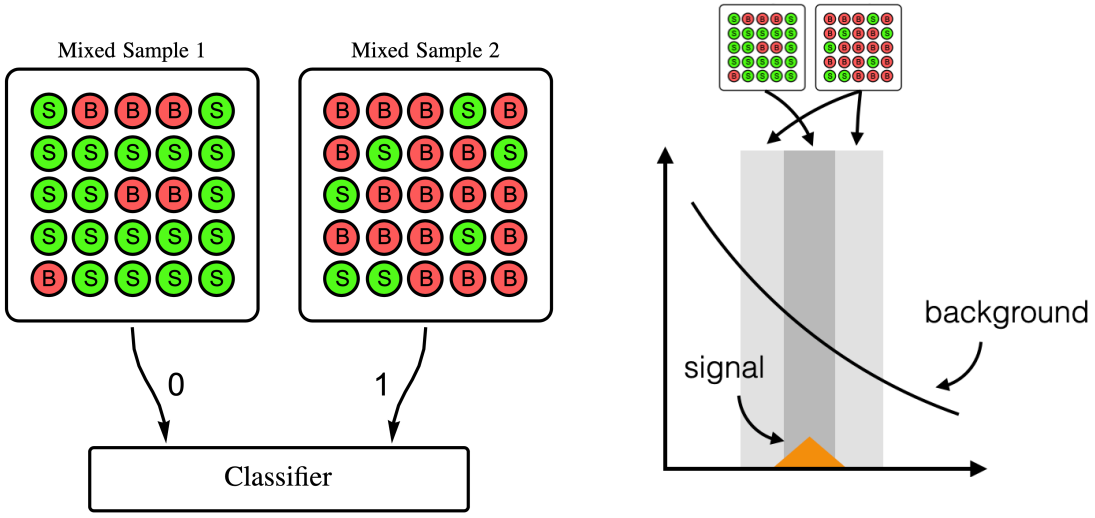
除了这两个经过旋转并居中的角位置外，我们还考虑数据集中每颗恒星的另外四个特征：两个角自行（$\mu_{\phi*}$，其中$\mu_{\phi*} \equiv \mu_\phi \cos$（$\lambda$），以及 μ_λ）、颜色（$b-r$，其中 b表示蓝光度计的亮度，而 r 表示红光度计的亮度）、以及星等（g）。图 1展示了一个此类天区中分析所用六个相关变量（ϕ、 λ、 μ_{ϕ^*}、 μ_λ、 $b-r$和 g）的分布。这些天区覆盖了一个形状不规则的区域，大致延伸于 $\alpha \in [-250°, -100°]$ 和 $\delta \in [-10°, 80°]$之间。

经过这些筛选，在本分析所考虑的约 800 万颗恒星中，有 1,957 颗根据文献 [50, 51] 开发的星表被标记为可能属于 GD-1星流。此标记依据的是位置、自行以及颜色和星等等时线上的筛选条件。虽然这些标签不能被视为完全准确或完整，但可作为评估我们模型效能的有用参考。

3 方法

3.1 无标签分类（CWoLa）

CWoLa 是一种弱监督机器学习技术，旨在发现数据集中至少沿一个维度局部聚集的异常特征。它最初是为高能粒子物理学领域的应用而设计的，在该领域中，信号和背景比例未知的粒子类别混合数据十分常见。它通过沿着某个局部维度进行扫描来检测异常——例如粒子物理事件末态的不变质量——并学习区分不同数据类别的混合样本，而这些样本中各类别的确切



(a) 信号富集（M_1）与信号贫化（M_2）　　(b) 局部异常

图 2。上图展示了信号富集组和信号贫化组。标记为“S”的绿色数据点代表信号事件，而标记为“B”的红色数据点代表背景事件。信号区域和旁带区域的选取方式使得，与周围的旁带区域相比，位于中央信号区域内的信号事件（以橙色显示）更多。

比例无需已知。只需学习区分信号比例较高和较低的区域，即“信号”区域与“旁带”区域，CWoLa 就能有效地指示异常事件的模式。

如图 M_1 所示，考虑一个信号富集混合样本 M_2和一个信号贫化混合样本2a。“信号”指感兴趣的对象类别——此处指局部异常特征（例如恒星流）的成员，也就是我们希望检测的对象；而“背景”则指不属于该异常的对象。两种混合样本都包含信号事件和背景事件，但与信号贫化混合样本相比，信号富集混合样本中的信号事件比例显著更高（即 $f_1 > f_2$，其中 f_i 表示每种混合样本中信号事件所占的比例）。我们利用这样一个事实（证明见 [36]中的定理 1）：在 M_1 和 M_2 之间区分事件而训练得到的最优分类器，与通过完全监督的方式训练、用于区分信号事件和背景事件的最优分类器相同。重要的是，无需知道每种混合样本中信号所占的确切比例（f_1 和 f_2），这一结论依然成立。该定理依据奈曼－皮尔逊引理 [52]，该引理指出，最优分类器 $h(\vec{x})$ 是由信号和背景的概率分布构造出的似然比的任意单调函数 $p_S(\vec{x})$ 和 $p_B(\vec{x})$，针对输入变量 $\vec{x}$。

我们可以应用 CWoLa 这一不依赖模型、由数据驱动的异常检测技术[37, 38]，通过识别数据集中可能包含局部异常的某个特征来实现，如图 2b所示。然后，我们训练一个全监督分类器，以区分来自“信号区域”和周围“旁带区域”的事件，这些区域由该特征的取值范围定义。如果不存在异常，作为分类器输入的辅助变量应与用于定义信号区域和旁带区域的特征不相关。如果存在异常，且其主要位于信号区域内，那么我们预计分类器会给异常事件排出更高的排名。接着，我们可以在一定的取值范围内滑动信号窗口和旁带窗口，重复这一过程。对于每一种信号

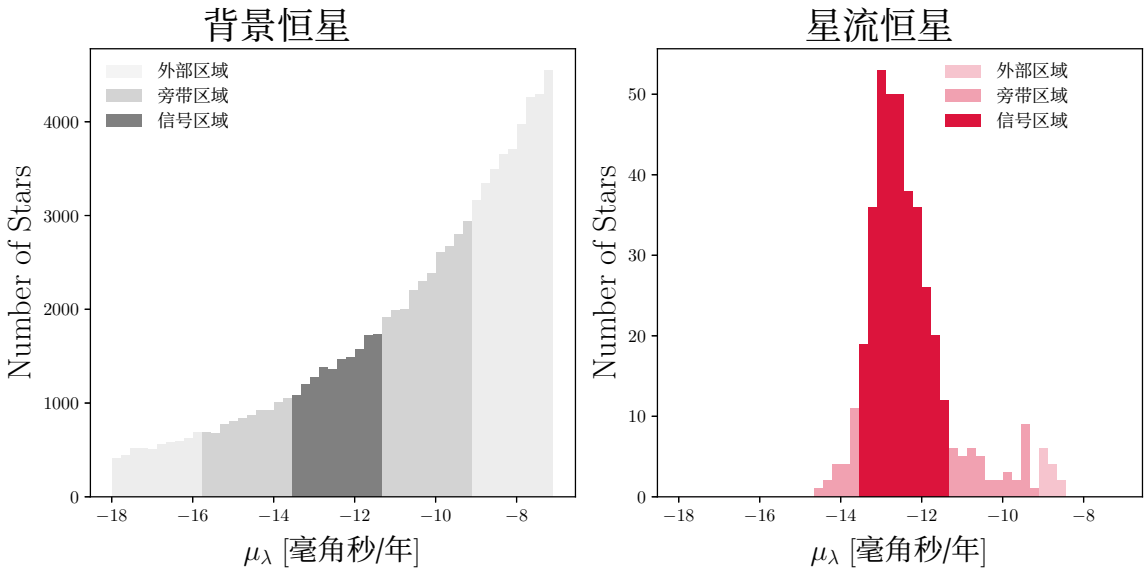


图 3。与同一天区中的背景恒星相比，与恒星流 GD-1 相关的恒星在 μ_λ 空间中高度集中，这些数据来自图 盖亚 数据，见图 1。信号区域在每个图中以最深色区域表示，其定义为取星流恒星的中位数 $\pm 1\sigma$ 值的 μ_λ，本例中该值为 $[-13.6, -11.4]$。旁带区域的定义为取星流中位 $\pm 3\sigma$ 数 μ_λ 值的，不包括信号区域： $[-15.8, -13.6)$ & $(-11.4, -9.3]$.

对于旁带，我们对分类器输出分数设置阈值（即测试集中的前 N 个事件或排名最靠前的百分位），只保留分数最高的事件。出现异常时，这些分数最高的事件将具有更高的事件信噪比。

该 CWoLa 异常搜索过程有两个关键假设。首先，它要求异常在我们搜索的维度上是局部集中的。由于恒星流在运动学上较冷，因此它们在自行的两个坐标上都相对集中。我们选择 μ_λ 作为主要坐标，用于定义 DR2 每个天区的信号区域和旁带区域。图 μ_λ展示了 GD-1 一个示例天区中的高度集中分布直方图，见图 4。其次，它假定信号区域与旁带区域中的背景事件和信号事件无法区分。如图 4所示，对一个数据天区而言，这是合理的假设。

3.2 定义信号与边带区域

对于本研究考察的 DR2 中每个天区，我们构建一个信号区域，使其包含尽可能多的可用星流恒星；同时在其邻近处构建旁带区域，确保信号区域和旁带区域中的背景恒星尽可能难以区分。理想情况下，两个区域中的恒星应具有相似的特征：信号区域中的背景恒星应与旁带区域中的背景恒星十分相似，星流恒星也遵循同样的原则。属于恒星流的恒星仅占每个天区内恒星总数的一小部分，因此在我们的研究中，每个信号区域仍将以未标记为属于该星流的背景恒星为主。不过，每个信号区域的信噪比应高于旁带区域。

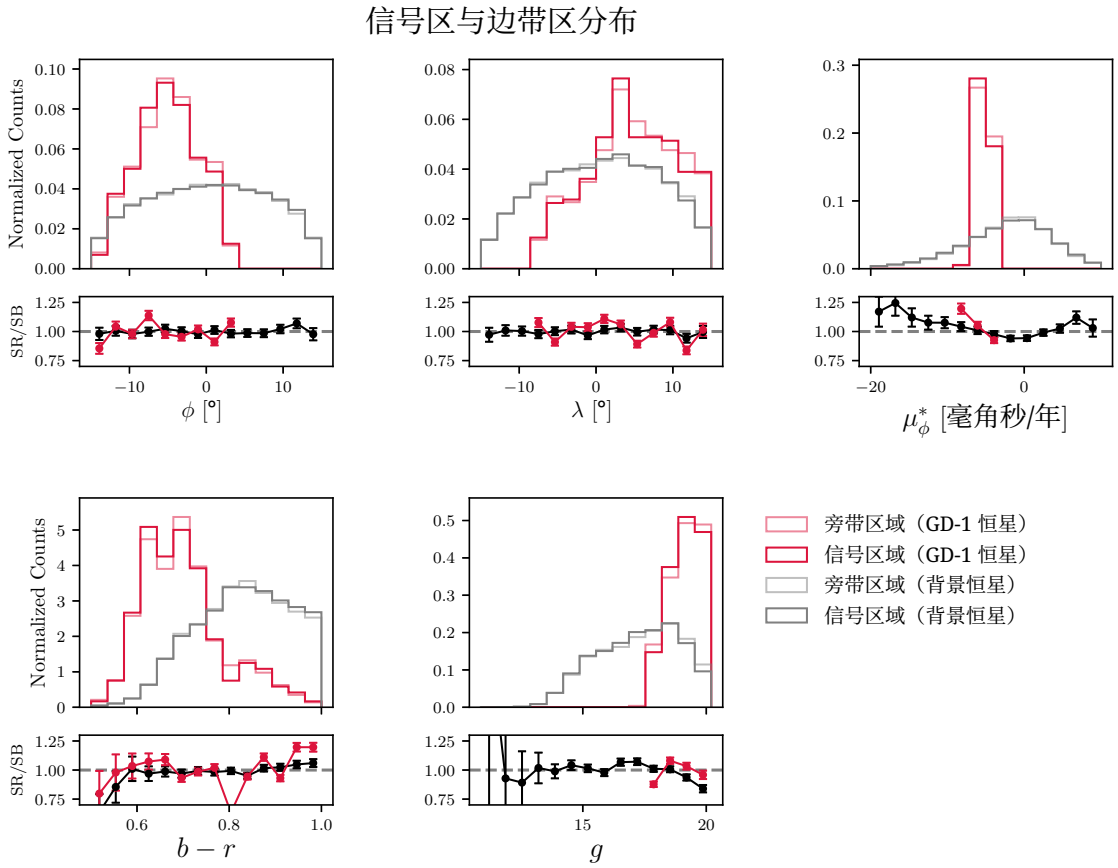


图 4。比较了 GD-1 恒星（红色）和背景恒星（灰色）在信号区域和旁带区域中五个神经网络输入变量的分布。此处所示的天区与图 1中的示例天区相同。对于星流恒星和背景恒星而言，这五个变量在信号区域和旁带区域中的分布大致无法区分。

理想情况下，信号区域应根据一系列 μ_λ 值来定义，这些值涵盖了大部分星流恒星。通常，定义这类区域有许多有效方法；在某些情况下，最佳定义可能取决于模型。

对于这一概念验证结果，我们根据已知星流在自行上的集中位置，选取了理想化的信号区域和旁带区域界限。不过，必须注意的是，在进行全面的异常搜索时，可以扫描一系列 μ_λ 数值；这意味着，即使选取了不同的信号区域和旁带区域，仍应能探测到星流。

在此情况下，我们将每个天区中的信号区域定义为该天区 GD-1 恒星中位数 μ_λ 一个标准差以内的区域。随后，将每个天区中的旁带区域定义为落在中位数 $[-3\sigma, -\sigma]$ 或 $[\sigma, 3\sigma]$ 以内的恒星。鉴于信号区域涵盖了星流的大部分区域，旁带区域中的星流恒星将显著减少，并且会如预期般信号耗尽。实际上，在 μ_λ 的 21 个天区中，信号区域的平均跨度为 2.34 ± 0.36 毫角秒/年，而旁带区域的平均跨度为 7.02 ± 1.09 毫角秒/年。

3.3 神经网络架构和训练流程

我们使用CWoLa 结合在 Keras 中构建的神经网络 [53]以及 TensorFlow 后端[54]。该模型由 3 个隐藏全连接层组成，每层包含 256 个节点，并采用 ReLU 激活函数 [55]。每个全连接层之后均使用丢弃操作，丢弃率为 20% [56]。这些层之后是一个仅含单个节点并采用 Sigmoid 激活函数的最终输出层。层大小、批量大小和 k折（如下文所述）的超参数值通过使用 Optuna 进行优化来确定 [57]。

对于我们在 GD-1 搜寻中考察的 21 个 盖亚 天区，我们训练了一系列分类器，将标记为属于信号区域的恒星与标记为属于旁带区域的恒星区分开来。正是这一特性使 CWoLa 被称为“弱监督”：与完全无监督网络相比，它利用的信息略多一些，因为我们预期最优信号区域中的 GD-1 恒星比例会高于旁带区域；但它无法获取实际的 GD-1 标签。该训练流程与其他 CWoLa 搜寻的流程 [37, 38], 高度一致，具体如下：

1. **k-折：** 我们采用分层 k 折（$k = 5$），将给定天区中的所有恒星随机分为五个部分，即“折”。每一折的划分都确保其中标记为“信号”与“旁带”的恒星比例与总体比例一致。第一折（占所有恒星的 20%）留作测试集。第二折（另占所有恒星的 20%）用作验证集。其余 60% 的恒星用于训练。

2. **训练：**接下来，使用上述架构的分类器在训练集上进行训练，批量大小为 10,000，最多训练 100 轮；不过，设置耐心值为 30 的提前停止通常会在达到这一上限之前很久就终止训练。由于整个数据集中带标签的星流恒星数量较少，因此需要较大的批量大小——例如，GD-1 尾端的一个天区中，星流恒星群体仅占 0.15%。因此，较大的批量大小有助于确保网络在训练过程中每次都能看到不止少数几颗星流恒星。我们使用二元交叉熵损失函数和 Adam 优化器 [58]。使用验证集监控验证损失，以实现提前停止。

3. **重复：**分类器训练再重复两次，每次都随机初始化可训练参数。在三次独立训练中，保存验证损失最低的模型权重。

4. **轮换验证集：**除测试集保持不变外，使用其余每一折作为验证集，重复这一过程。对于每种配置，除用于测试集和验证集的折之外，其余三折用于训练。

5. **在测试集上评估：** 使用验证集的四种 k 折选项训练出的最佳模型，均在测试集上进行评估。测试集中每颗恒星的最终 CWoLa 得分定义为四个得分的平均值。

6. **合并测试集：** 重复整个过程，依次将五个可能的 k 折用作测试集。随后将这些测试集合并为一个数据集，使天区中的每颗恒星恰好进入测试集一次。

3.4 模型评估

训练 CWoLa 方法中固有的神经网络分类器后，会对每个天区应用一系列基准筛选，进一步优化结果，以获得尽可能高的信噪比。所使用的基准筛选与 [32, 33]中的几乎完全相同：

- $g < 20.2$，以确保盖亚 卫星的接收率一致
- $|\mu_\lambda| > 2$ 毫角秒/年或 $|\mu_\phi^*| > 2$ 毫角秒/年，以剔除非常遥远的恒星。这些恒星主要集中在自行接近零的区域，因此在整个天区中的分布并不均匀
- $0.5 \leq b - r \leq 1$，以在颜色空间中分离出古老且低金属丰度的恒星流。

然而，与 Via Machinae 不同，我们在训练后无需应用截断，将天区半径限制在 15° $\rightarrow$ 10°，因为 CWoLa 不会受到 ANODE 实现中出现的边缘效应影响。

Via Machinae 采用一种复杂的线段搜索策略，使用经过修改的霍夫变换 [59] ，在识别出的异常恒星中搜索线状结构，然后将这些线段组合成一个整体的恒星流候选体。我们则在自行空间中使用 k-均值聚类 [60] ($k = 2$)，以较低的计算开销取得了类似结果。将恒星分为两个簇后，我们选取恒星数量最多的簇，并剔除另一簇中的恒星。这样做是因为我们预计恒星流在运动学上较冷，因此其组成恒星的速度应在速度空间中密集聚集。这种聚类策略可能最适合作为发现后的工具；在污染恒星数量较多、且这些恒星不属于任何星流的情况下，效果可能不佳。在这些情况下，选择较大的$k > 2$ 或采用线段搜索技术可能是更好的选择。

根据这些基准筛选，我们将分类器应用于合并测试集中相当于整个天区的恒星，以评估模型性能。我们按输出分数从高到低排序，其中较高的分数表示模型认为这些恒星更有可能属于信号区域，而非旁带区域。我们选取神经网络输出分数排名最高的 $N = 250$颗恒星进行评估。

250 这一数量是通过同时优化纯度（排名最高的恒星中与已标记的GD-1恒星重叠的比例）和完备度（CWoLa排名最高的恒星中涵盖的已标记GD-1恒星比例）选定的。不过，原则上也可以筛选出其他绝对数量或相对比例的排名靠前恒星，但建议不超过每个天区平均430颗已标记的星流恒星。

值得强调的是，这种模型评估方法需要真实标签。如果缺少可靠的星流标签，或者要发现一条新的星流，我们就必须采用其他方法来评估模型性能，更不用说模型实现本身也需要调整策略。我们将在第 5 节进一步讨论这一点。

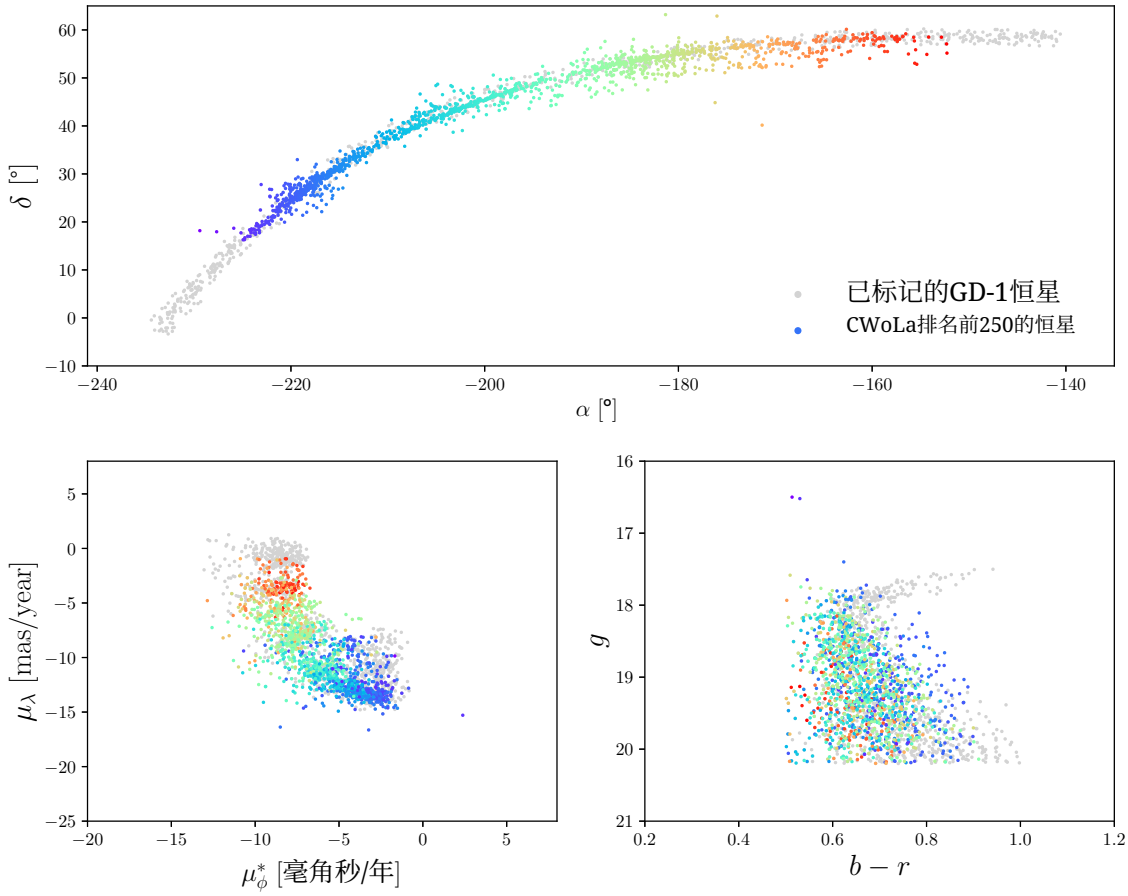


图5。CWoLa方法在对应于GD-1的角范围内、各重叠天区中识别出的恒星全貌。浅灰色点表示 GD-1 恒星的真实标签 [51], 而每个天区中由CWoLa识别出的排名前250的恒星以彩色点表示。颜色的选择与每颗恒星的 α 值相对应。

4 结果

在查看真实的 盖亚 数据之前，我们还评估了 CWoLa 应用于 100 条随机选取的模拟恒星流时的性能。如附录 A 所示，仅经过两轮 CWoLa，就有 96% 的星流被识别出来且纯度非零，其中 69% 的星流纯度超过 50%。

4.1 GD-1 星流识别

将 CWoLa 技术应用于 盖亚DR2 的 21 个天区所得的综合结果如图5所示。各个天区的结果详见附录 B。在这 21 个天区中，共有 1,498 颗独特的 GD-1 恒星通过了我们的基准筛选。每个 CWoLa 天区中排名前 $N = 250$颗恒星合并后，共识别出 1,360 颗独特的恒星。其中，760 颗属于已标注的GD-1恒星集合 [51]。因此，在整个星流范围内，我们的纯度达到 56%，完备度达到 51%。在优化研究中，我们发现，选取每个天区中排名前 $N = 25$ 颗恒星时，星流纯度的峰值稳定在 78%；但这种选择的 N完备度仅为 13%。相反，选择 $N = 300$ 会使纯度降至仅 30%，但完备度更高，达到 54%。

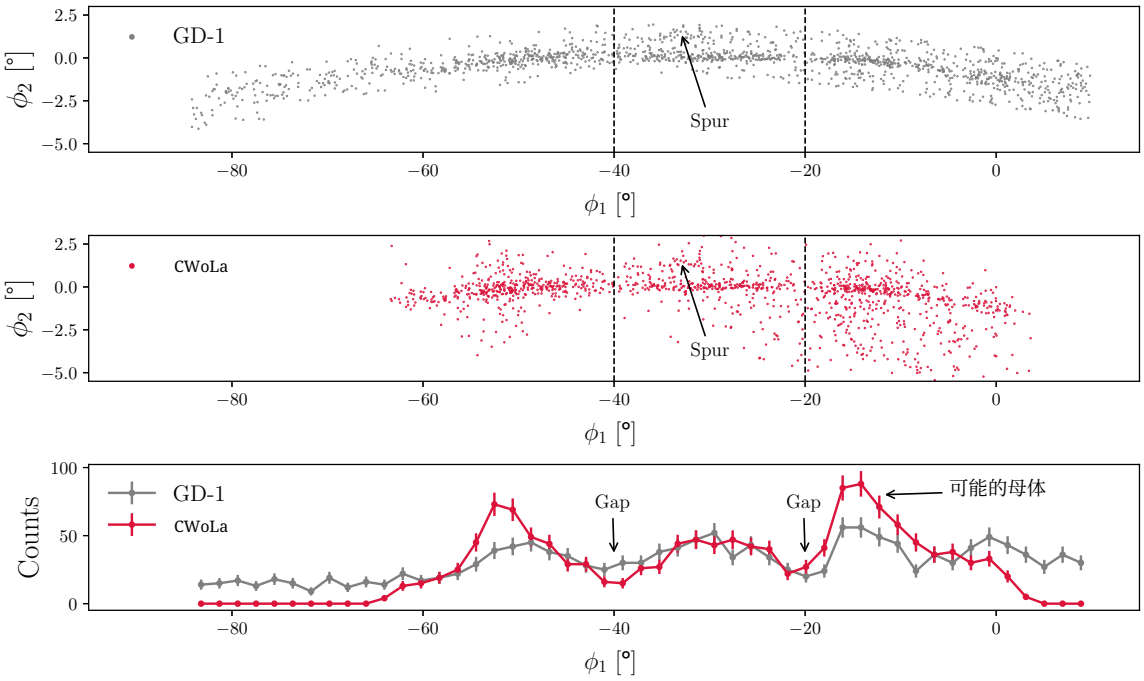


图 6。CWoLA识别出的所有天区中的恒星，与 [51] 中流对齐坐标下的 ϕ_1 和 ϕ_2所示的 CWoLA 识别出了 GD-1 特有的若干密度扰动：靠近$\phi_1 = -40°$ 和 $\phi_1 = -20°$的两个恒星稀疏分布的“间隙”；靠近 $\phi_1 = -35°$的一个分支（或称“分叉”）；以及靠近 $\phi_1 = -15°$的一个恒星过密区（或称“团块”）。

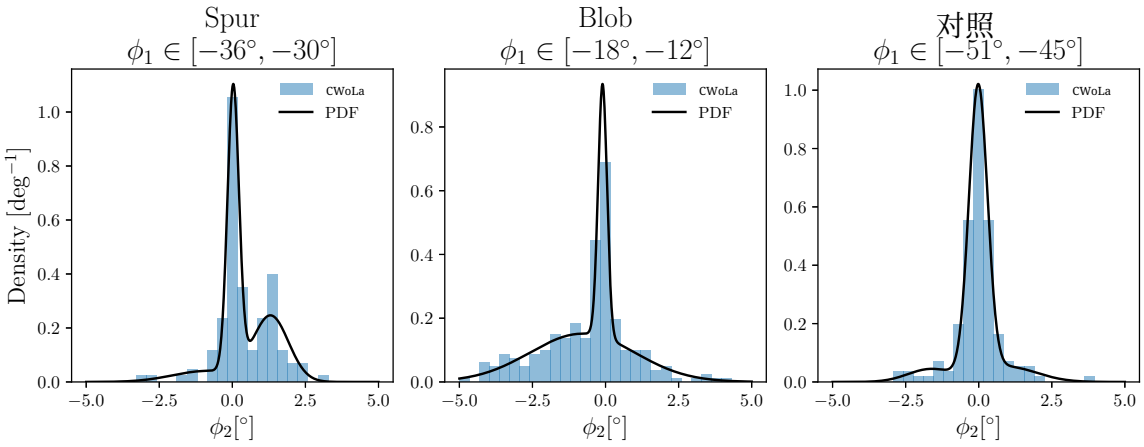
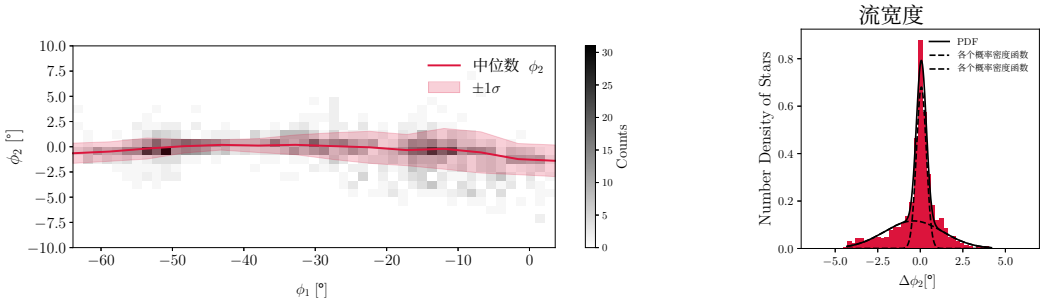


图 7。选取三组 CWoLa识别的恒星（“团块”、“分支”和一个对照区域），并使用三组分高斯混合模型进行拟合，以突出显示附加特征（如果存在）的运动学特征。在每种情况下，GD-1星流对应于以 $\phi_2 = 0°$ 附近为中心的主要窄峰。在前两幅图中，我们清楚地看到分别代表每个特征的第二个峰。



(a) 按 ϕ_1 分箱后 ϕ_2 的中位数对应的 CWoLa识别的恒星　　(b) CWoLa 恒星的宽度

图 8。CWoLa识别的恒星的宽度通过首先计算 ϕ_2 中 10 个 ϕ_1 分箱的星流位置中位数来确定（图 (a) 中叠加的红线）。随后将 ϕ_2 坐标减去这些中位数，从而得到图 (b) 中的直方图。在图 (b) 中，我们使用双组分高斯混合模型展示两个独立的高斯组分，其 $\sigma \approx 0.3^\circ$ 对应 GD-1 的核心，而$\sigma \approx 1.7^\circ$ 对应“茧状结构”。这似乎支持了图 6 中观察到的总体趋势，即 [46]。

GD-1 的大部分区域相当狭窄，平均角宽度约为 0.5°[46], ，且恒星分布密集，在 $\phi_1 = -60°$至 $\phi_1 = 0°$之间的每个 5° 分箱中约有 100 颗恒星。在该区域内，CWoLa 能够可靠地识别星流恒星。GD-1 的尾端（$\phi_1 \leq -60°$ 和 $\phi_1 \geq 0°$）恒星分布较为稀疏，每个分箱中的恒星数量约为星流主体平均数量的一半，并且在 μ_λ 中的局域性较弱，这意味着使用 CWoLa 更难识别该区域内的星流恒星。由于这些区域靠近银河盘或存在邻近尘埃，其中的一些恒星也可能因此被排除在 21 个片区之外。这些区域还往往包含自行较小的恒星，这意味着星流恒星更容易被遥远的背景恒星淹没。

我们还可以在旋转位置坐标组 ϕ_1 和 ϕ_2 [43]经过设计、与星流主体方向对齐的坐标系中分析这些结果，如图6所示。从这一视角可以看出，CWoLa 已经识别出 GD-1 星流所特有的几种密度扰动：靠近 $\phi_1 = -40°$ 和 $\phi_1 = -20°$附近的两个恒星稀疏分布的“间隙”；靠近 $\phi_1 = -35°$附近的一个分支，或称“分支结构”；以及靠近 $\phi_1 = -15°$附近的一个恒星过密区，或称“团块”。我们可以通过考察不同 ϕ_2 范围内的 ϕ_1 直方图，对这些特征的识别进行更定量的验证，如图7所示，这与 [50], 中的图5类似。本研究尤其重点关注“分支”和“团块”，分别对每个特征附近的恒星直方图进行拟合，并同时纳入第三个对照区域，采用三组分高斯混合模型，假设数据由背景、GD-1星流以及相应特征（“团块”或“分支”）组成。

如上所述，GD-1 中的欠密度区域（即“间隙”）通常出现在 $\phi_1 \approx -40°$ 和 $\phi_1 \approx -20°$ 附近。在 $\phi_1 \approx -3°$ [44] 附近还发现了第三个欠密度区域。我们的结果并不与这第三个间隙相矛盾，因为该区域内 CWoLa识别的恒星密度确实较低，与另外两个“间隙”处的密度相当；但由于该区域离 CWoLa识别的恒星所能延伸到的最远处如此之近，目前尚不清楚这种欠密度究竟是星流的特征，还是星流两端天区中恒星纯度降低的反映。

关于超密区域， [44] 报告了四个超密区域，峰值分别位于 $\phi_1 \approx -48°$（星流中密度最高的区域）、$\phi_1 \approx -27°$、 $\phi_1 \approx -10°$ 和 $\phi_1 \approx +2°$。虽然 CWoLa识别的恒星未能可靠地覆盖 $\phi_1 = 0°$ 以上的区域，但其余三个位于 $\phi_1 < 0°$ 的峰值也出现在 CWoLa识别的恒星中。通过拟合

CWoLA识别的恒星，并使用由三个高斯分布组成的混合模型，我们可以提取出近似的过密峰，分别位于 $\phi_1 \approx -51°$、 $\phi_1 \approx -30°$和 $\phi_1 \approx -11°$。值得注意的是，CWoLA 检测到了一个显著的 $\phi_1 \approx -51°$ 峰，与 [44], 报告的最高密度峰一致，尽管从 [51]标记的恒星中，这个峰并不那么明显。

GD-1 的另一项已报告特征是，围绕星流密度高得多的核心，还存在一层宽度约为 1 的较宽“茧状结构”，其中包含恒星° [46]。为探究这一特征，我们首先在较宽的 5° 区间内计算 ϕ_1 的中位数，以找出星流的平滑轨迹。随后，根据各自的中位数 ϕ_2 ，将每颗经 CWoLa 识别的恒星的位置移至对应的中位数 ϕ_1 位置（见图 8a）。以这条轨迹为中心对星流进行校准后，我们按照 3σ 进行筛选 [46], ，然后绘制平移后的 ϕ_2 的直方图（见图 8b）。用双分量高斯混合模型拟合该分布，得到一个标准差为 $\sigma \approx 0.3°$ 的窄峰（即星流核心），以及另一个更宽的峰，其标准差为 $\sigma \approx 1.7°$。这似乎支持了这样的观察：星流中心核心周围存在由更多弥散恒星构成的“茧状结构”。

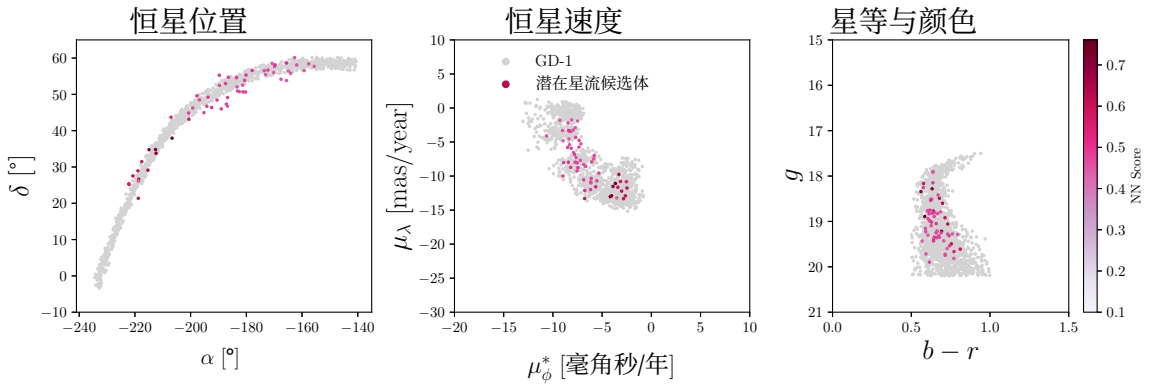


图 9。从 CWoLa 识别出的排名靠前的恒星中，筛选出与最近的已标记恒星之间具有最小 10% 的五维欧几里得距离的子集 d ，发现了另外 60 个可能属于 GD-1 的恒星候选体，这些候选体可能被遗漏在 GD-1 的真实标签标注中。

4.2 朝着扩充GD-1星流标注迈进

除了在没有标签的情况下识别冷恒星流外，CWoLa 技术还有助于改进用于标示哪些恒星属于特定星流的标注系统。例如，CWoLa 可以识别出一些很有希望的恒星候选体：它们在 [51], 中未被标注为 GD-1成员，但其性质与已标注的恒星高度相符。

如第4节所示，CWoLa 在全部 21 个天区中识别出 1,360 颗不重复的恒星，其中 760 颗（56%）属于已标注的GD-1恒星集合。我们还可以进一步筛选剩余的 600 颗未标记恒星，找出其中与测试集中各自最近的已标注GD-1恒星之间的欧几里得距离最小的个体恒星子集 s ，距离沿着作为 d CWoLa 输入的 5 个维度计算，这些维度经过独立标准化，使其具有 s' $\mu = 0$ 和 $\sigma = 1$: （ϕ、 λ、 $\mu_{\phi*}$、颜色（$c \equiv b - r$）和星等（g）：

$$d = \sqrt{(\phi - \phi')^2 + (\lambda - \lambda')^2 + (\mu_{\phi*} - \mu'_{\phi*})^2 + (c - c')^2 + (g - g')^2}. \tag{4.1}$$

筛选出距离值最小的 10% 的恒星 d 后，可发现另外 60 颗恒星，如图 9所示，并在附录 C中明确列出。这些恒星在这六个维度上似乎与已标记的GD-1恒星相符，值得进一步研究，以判断它们是否为潜在的GD-1成员候选体。未来的工作将对这些候选GD-1星流恒星与其他更精确的GD-1星流星表进行详细交叉核查。

5 讨论

显然，CWoLa 成功识别了 GD-1 的大部分区域，这不仅体现在纯度和完备度上，也体现在对该星流特有物理特征和密度扰动的忠实重建上。此外，CWoLa 在识别模拟星流方面也非常有效。

尽管本文的分析与Via Machinae在许多核心策略要素和所用数据集方面相同，但二者的主要区别在于：如何为恒星赋予异常分数、如何将天空划分为若干子区域以进行扫描，以及训练后如何对恒星进行聚类。CWoLa 采用一种相对简单、轻量且易于训练的基于神经网络的分类器，而非归一化流模型，以解决同一异常检测问题。在同一示例恒星流 GD-1 上应用时，CWoLa 能够以低得多的计算开销识别出纯度相当的恒星。我们总共找到 760 颗已标记的恒星，纯度为 56%；而Via Machinae的首次迭代 [32] 找到了 738 颗恒星，纯度为 49%。Via Machinae的最新迭代 [33], 增加了额外的基准筛选，并对两个自行变量进行了扩展扫描，使检出的恒星增至 820 颗，纯度达到 65%。

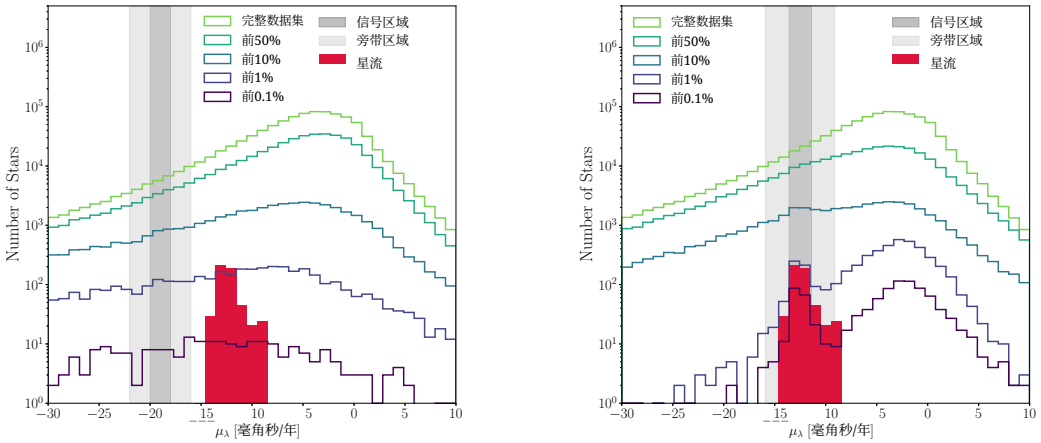
值得强调的是，我们的方法并未像 Via Machinae 那样使用任何线条检测或预聚类算法——这里的异常恒星只是合并在一起，并通过 k-means 聚类进行筛选。这种轻量级聚类策略在发现之后的研究中尤其有用，因为此时我们希望完善星流成员目录。这些技术之间的另一个重要区别是，CWoLa 使用宽度各异的信号区域和边带区域，并根据信号的位置为每个天区选择这些区域；而 Via Machinae 则在由正交自行坐标定义的感兴趣区域内进行搜索，区域宽度固定为 6 毫角秒/年，中心间隔为 1 毫角秒/年。这些实现所采用的基准筛选也略有不同：Via Machinae 将每个天区限制在位置空间中最内侧的 10° 圆形区域内，以避免边缘效应；而 CWoLa 不会出现这些效应，因此我们不施加这一筛选条件。

训练 ANODE 中使用的归一化流等模型可能耗时较长，而在 NVIDIA A40 GPU 上，完成 GD-1 的完整 CWoLa 训练流程，每个天区只需 15 分钟。对每个天区进行处理非常适合并行计算，可根据 GPU 资源轻松同时运行，也可在 CPU 上使用多进程处理。使用单块 GPU 处理全部 21 个片区总共约需 5 小时，因此研究人员完全可以优化信号区域和边带区域的定义，并结合多个变量扫描的结果。

通过在训练之前 对样本应用基准筛选，还能进一步缩短训练时间——将训练时间大约减半，同时使总体纯度达到 44%。训练时间提升 100% 时，这项技术仅使识别出的恒星最终纯度降低约 20%，因此对于在大范围相空间内进行粗粒度扫描，尤其是一个很有价值的选择。因此，CWoLa 和 Via Machinae 可视为在不同情境或计算资源限制下用于星流探测的互补工具。

如果 CWoLa 可用于识别已知流，那么它也可能被用于寻找新的未发现流。将 CWoLa 扩展到完整的 盖亚 数据集、以寻找新的恒星流时，还会面临一些额外挑战。检测到的异常不一定能保证是恒星流，因为 CWoLa 可能会识别任何局部异常的

特征。由于缺乏某个流的真实标签，我们还需要重新评估性能指标——例如，应使用标准的异常检测方法评估流：在自行直方图上进行一系列筛选（例如，依据神经网络评分的某个百分位数范围，或根据旁带区域中的本底率人工设定的阈值）。如果不存在异常，这些逐渐变严格的筛选会减少样本统计量，但不会显著改变直方图形状，如图 10a 所示。然而，如果存在异常且被 CWoLa 识别出来，随着对相关分布施加越来越严格的筛选，将会出现新的形状，如图 10b 所示。此外，如果事先 CWoLa不知道异常的大致宽度，可能还需要针对信号区域和旁带区域宽度的不同选择，多次运行 *CWoLa* 。



(a) 当信号区域不包含星流时，随着筛选百分比增大，自行空间中任何位置都不会出现明显的凸起。

(b) 当信号区域包含星流时，在星流所在的同一区域会出现明显的凸起。

图 10。演示了一次扫描，其中异常的位置事先未知。

搜索新的恒星流需要扫描数据集中自行值的整个范围，因为我们无法预知新星流可能位于何处。通过在 μ_λ 上以粗略的滑动窗口方式应用 CWoLa，可以识别出感兴趣区域。随后可以对这些区域进行更精细的扫描，直至识别出异常数据点。将多个天区合并以获得总体结果时，我们或许还需要使用寻线算法来识别尺度更大的类星流结构，例如 Via Machinae [32] 中使用的改进版霍夫变换。

6 结论

我们展示了“CWoLa 搜寻”这一新应用。这是一种基于弱监督机器学习分类器 CWoLa 的异常检测技术，旨在以模型无关的方式检测局部异常。研究表明，CWoLa 易于训练、计算效率极高，而且最重要的是，能够有效识别异常，包括恒星流 GD-1 和数十条具有高纯度的模拟星流。GD-1

CWoLa 识别出的候选恒星呈现出多个独立研究在该星流中发现的相同密度扰动和物理特征（“分支”、“团块”、“茧状结构”、间隙和超密区域）。神经网络的输出分数也为哪些恒星可能在更正式的 GD-1 标记方案中被意外遗漏提供了线索，并指出了若干有望成为候选者的恒星。本研究成功应用 CWoLa，表明它有很大潜力提升已知星流成员识别的信噪比，并有望揭示银河晕中此前尚未探测到的星流。CWoLa 作为一种弱监督异常检测技术，在高能物理以外也具有广泛适用性，而且还有望应用于更多基础科学领域。

6.1 重现这些结果

本文中各个图表的复现说明代码库位于 https://github.com/hep-lbdl/GaiaCWoLa。完整复现本文图表所需的数据集（已应用 CWoLa ）均已公开发布 [61]。覆盖 GD-1 的全部 21 个片区也已公开发布 [62]。

致谢

M.P. 感谢 Shirley Ho、Adrian Price-Whelan、David Spergel 和 Joshua Bloom 的宝贵交流，并衷心感谢 Flatiron Institute Center for Computational Astrophysics 的热情接待。S.T. 感谢 Vanessa Boehm 在使用 Optuna 进行超参数优化方面提供的帮助。本研究使用了欧洲空间局（ESA）任务 盖亚 (https://www.cosmos.esa.int/gaia) 的数据，该数据由 盖亚 数据处理与分析联盟（DPAC，
https://www.cosmos.esa.int/web/gaia/dpac/consortium）处理。DPAC 的经费由各国机构提供，特别是参与 盖亚 多边协议的机构。B.N.、M. P. 和 S. T. 获得美国能源部科学办公室通过合同编号 DE-AC02-05CH11231 提供的支持。

参考文献

[1] Kathryn V. Johnston。《通过瓦解的卫星星系构建银河系晕》。发表于《天体物理学杂志》495.1（1998年3月），第297页。doi：10.1086/305273。网址：https://dx.doi.org/10.1086/305273.[2] Amina Helmi 和 Simon D. M. White。《构建星系的恒星晕》。发表于《皇家天文学会月刊》307.3（1999年8月），第495–517页。issn：0035-8711。doi：10.1046/j.1365-8711.1999.02616.x。电子预印本：https://academic.oup.com/mnras/article-pdf/307/3/495/2842682/307-3-495.pdf。网址：https://doi.org/10.1046/j.1365-8711.1999.02616.x.[3] Amina Helmi 等。《导致银河系内层恒星晕和厚盘形成的并合》。发表于《自然》563.7729（2018年10月），第85–88页。doi：10.1038/s41586-018-0625-x。网址：https://doi.org/10.1038%2Fs41586-018-0625-x.[4] V. Belokurov 等。《星流之野：人马座及其同伴》。发表于《天体物理学杂志》642.2（2006年4月），第L137–L140页。doi：10.1086/504797。网址：https://doi.org/10.1086%2F504797。

[5] V Belokurov 等。《盘与恒星晕的共同形成》。发表于《皇家天文学会月刊》478.1（2018年6月），第611–619页。doi：10.1093/mnras/sty982。网址：https://doi.org/10.1093%2Fmnras%2Fsty982.[6] Khyati Malhan 等。《矮星系星流填充银河系内层晕的证据》。发表于《天体物理学杂志》920.1（2021年10月），第51页。doi：10.3847/1538-4357/ac1675。网址：https://doi.org/10.3847%2F1538-4357%2Fac1675.[7] David R. Law 和 Steven R. Majewski。《人马座矮星系：三轴银河系晕中演化的模型》。发表于《天体物理学杂志》714.1（2010年4月），第229页。doi：10.1088/0004-637X/714/1/229。网址：https://dx.doi.org/10.1088/0004-637X/714/1/229.[8] Andy Eyre 和 James Binney。《确定潮汐星流所描绘的轨道》。发表于《皇家天文学会月刊》400.1（2009年11月），第548–560页。issn：0035-8711。doi：10.1111/j.1365-2966.2009.15494.x。电子预印本：h
tt
ps://academic.oup.com/mnras/article-pdf/400/1/548/18701250/mnras0400-0548.pdf。网址：https://doi.org/10.1111/j.1365-2966.2009.15494.x.[9] Walter Dehnen 等。《银河潮汐对球状星团 Palomar 5 瓦解过程的建模》。发表于《天文学杂志》127.5（2004年5月），第2753页。doi：10.1086/383214。网址：https://dx.doi.org/10.1086/383214.[10] Harshil Kamdar、Charlie Conroy 和 Yuan-Sen Ting。《银河盘中的恒星流：预测寿命及其在测量银河系引力势中的用途》。2021年。arXiv：2106.02050 [astro-ph.GA].[11] Stella Reino 等。《来自合并恒星流数据的作用量空间聚类对银河系引力势的约束》。发表于《皇家天文学会月刊》502.3（2021年2月），第4170–4193页。issn：0035-8711。doi：10.1093/mnras/stab304。电子预印本：https://academic.oup.com/mnras/article-pdf/502/3/4170/36338711/stab304.pdf。网址：https://doi.org/10.1093/mnras/stab304.[12] Jacob Nibauer 等。《利用恒星流和机器学习绘制银河系加速度图》。发表于《天体物理学杂志》940.1（2022年11月），第22页。doi：10.3847/1538-4357/ac93ee。网址：https://doi.org/10.3847%2F1538-4357%2Fac93ee.[13] R. G. Carlberg、C. J. Grillmair 和 Nathan Hetherington。《Pal 5 恒星流的间隙》。发表于《天体物理学杂志》760.1（2012年11月），第75页。doi：10.1088/0004-637x/760/1/75。网址：https://doi.org/10.1088%5C%2F0004-637x%5C%2F760%5C%2F1%5C%2F75.[14] Jason L. Sanders、Jo Bovy 和 Denis Erkal。《星流与子晕相互作用的动力学》。发表于《皇家天文学会月刊》457.4（2016年1月），第3817–3835页。doi：10.1093/mnras/stw232。网址：https://doi.org/10.1093%2Fmnras%2Fstw232.[15] Denis Erkal 等。《恒星流中由子晕引发的间隙数量与大小》。发表于《皇家天文学会月刊》463.1（2016年8月），第102–119页。issn：0035-8711。doi：10.1093/mnras/stw1957。电子预印本：https://academic.oup.com/mnras/article-pdf/463/1/102/18470778/stw1957.pdf。网址：https://doi.org/10.1093/mnras/stw1957。

[16] Ana Bonaca 等. “GD-1中的分支与间隙：银河系晕中暗物质子结构的动力学证据”。载于：《天体物理学杂志》880.1（2019年7月），第38页。doi: 10.3847/1538-4357/ab2873。url: https://doi.org/10.3847%2F1538- 4357%2Fab2873.
[17] Nilanjan Banik 和 Jo Bovy. “重子物质与暗物质子结构对Pal 5星流的影响”。载于：《皇家天文学会月刊》484.2（2019年1月），第2009–2020页。doi: 10.1093/mnras/stz142。url: https://doi.org/10.1093% 2Fmnras%2Fstz142.[18] Ana Bonaca 等. “GD-1恒星流的高分辨率光谱观测将扰动体定位在靠近人马座轨道平面的位置”。载于：《天体物理学杂志》892.2（2020年4月），第L37页。doi: 10.3847/2041-8213/ab800c。url: https://doi.org/ 10.3847%5C%2F2041-8213%5C% 2Fab800c.[19] Chris W Purcell、Andrew R Zentner 和 Mei-Yu Wang. “银河系与人马座星流模拟中的暗物质直接探测率”。载于：《宇宙学与天体粒子物理学杂志》2012.08（2012年8月），第027–027页。doi: 10.1088/1475- 7516/2012/08/027。url: https://doi.org/10.1088%2F1475- 7516%2F2012%2F08% 2F027.[20] Lina Necib 等. “FIRElight之下：银河系局部暗物质速度分布的恒星示踪体”。载于：《天体物理学杂志》883.1（2019年9月），第27页。doi: 10.3847/1538-4357/ab3afc。url: https://doi.org/10.3847%2F1538- 4357%2Fab3afc.[21] O. J. Eggen. “大角星星群”。载于：《太平洋天文学会出版物》83（1971年6月），第271页。doi: 10.1086/129120。url: https://doi.org/10. 1086/129120.[22] Cecilia Mateu. “galstreams：银河系恒星流足迹与轨迹库”。载于：《皇家天文学会月刊》520.4（2023年1月），第5225–5258页。doi: 10 . 1093 / mnras / stad321。url: https : / / doi . org / 10 . 1093 % 2Fmnras%2Fstad321.[23] Constance M. Rockosi 等. “对球状星团Palomar 5潮汐尾的匹配滤波分析”。载于：《天文学杂志》124.1（2002年7月），第349页。doi: 10.1086/340957。url: https://dx.doi.org/10.1086/340957.[24] Eduardo Balbinot 等. “NGC 2298的潮汐尾”。载于：《皇家天文学会月刊》416.1（2011年8月），第393–402页。issn: 0035-8711。doi: 10.1111/ j.1365- 2966.2011.19044.x。eprint: https://academic.oup.com/mnras/article-pdf/416/1/393/4045323/mnras0416-0393.pdf。url: https://doi.org/10.1111/j.1365-2966.2011.19044.x.[25] M. E. K. Williams 等. “RAVE中宝瓶座星流的初现”。载于：《天体物理学杂志》728.2（2011年1月），第102页。doi: 10.1088/0004-637x/728/2/102。url: https://doi.org/10.1088%2F0004-637x%2F728%2F2%2F102.[26] M. I. Arifyanto 和 B. Fuchs. “附近矮星次巨星相空间分布中的精细结构”。载于：《天文学与天体物理学》449.2（2006年3月），第533–538页。doi: 10 . 1051 / 0004 - 6361 : 20054355。url: https : / / doi . org / 10 . 1051 % 2F0004 - 6361%3A20054355.

[27] Sonia Duffau 等. “QUEST RR Lyrae变星的光谱观测：新发现的室女座恒星流”。载于：《天体物理学杂志》636.2（2005年12月），第L97–L100页。doi: 10.1086/500130。url: https://doi.org/10.1086%2F500130.[28] Kathryn V. Johnston、Lars Hernquist 和 Michael Bolte. “晕中远古吸积事件的化石印记”。载于：《天体物理学杂志》465（1996年7月），第278页。doi: 10.1086/177418。url: https://doi.org/10.1086%2F177418.[29] Cecilia Mateu、Justin I Read 和 Daisuke Kawata. “银河系内盘中的十四条候选RR Lyrae恒星流”。载于：《皇家天文学会月刊》474.3（2017年11月），第4112–4129页。doi: 10.1093/mnras/stx2937。url: https://doi. org/10.1093%2Fmnras%2Fstx2937.[30] Khyati Malhan 和 Rodrigo A Ibata. “STREAMFINDER——I. 一种检测恒星流的新算法”。载于：《皇家天文学会月刊》477.3（2018年4月），第4063–4076页。doi: 10.1093/mnras/sty912。url: https://doi.org/10. 1093%5C%2Fmnras%5C%2Fsty912.[31] Nicholas W Borsato、Sarah L Martell 和 Jeffrey D Simpson. “利用数据挖掘技术识别盖亚DR2中的恒星流”。载于：《皇家天文学会月刊》492.1（2019年12月），第1370–1384页。doi: 10.1093/mnras/stz3479。url: https://doi.org/10.1093%5C%2Fmnras%5C%2Fstz3479.[32] David Shih 等. “Via Machinae：利用无监督机器学习搜寻恒星流”。载于：《皇家天文学会月刊》509.4（2021年11月），第5992–6007页。doi: 10.1093/mnras/stab3372。url: https://doi.org/10. 1093%5C%2Fmnras%5C%2Fstab3372.[33] David Shih、Matthew R. Buckley 和 Lina Necib. Via Machinae 2.0：对盖亚DR2中恒星流进行全天空、模型无关的搜寻。2023年。arXiv: 2303.01529 [astro-ph.GA]

[34] Benjamin Nachman 和 David Shih。《基于密度估计的异常检测》。载于：《物理评论D》101 (2020)，第 075042 页。doi: 10.1103/PhysRevD.101.075042。arXiv: 2001.04990 [hep-ph].[35] Danilo Jimenez Rezende 和 Shakir Mohamed。《基于归一化流的变分推断》。2015年。doi: 10.48550/ARXIV.1505.05770。url: https://arxiv.org/abs/1505. 05770。[36] Eric M. Metodiev、Benjamin Nachman 和 Jesse Thaler。《无标签分类：从高能物理混合样本中学习》。载于：《高能物理学杂志》2017.10（2017年10月）。doi: 10.1007/jhep10(2017)174。url: https://doi.org/ 10.1007%5C%2Fjhep10%5C%282017%5C%29174。[37] Jack H. Collins、Kiel Howe 和 Benjamin Nachman。《利用机器学习拓展新共振态搜索》。载于：《物理评论D》99.1（2019年1月）。doi: 10. 1103/physrevd.99.014038。url: https://doi.org/10.1103%5C%2Fphysrevd.99. 014038。[38] Jack H. Collins、Kiel Howe 和 Benjamin Nachman。《利用机器学习检测共振新物理中的异常》。载于：《物理评论快报》121.24（2018年），第 241803 页。doi: 10.1103/PhysRevLett.121.241803。arXiv: 1805.02664 [hep-ph].[39] Georges Aad 等。《利用弱监督在$\sqrt{s} = 13$ TeVpp 碰撞中使用 ATLAS探测器搜寻双喷注共振》。载于：《物理评论快报》125.13（2020年），第 131801 页。doi: 10.1103/PhysRevLett.125.131801。arXiv: 2005.02983 [hep-ex]。

[40] T. Prusti 等。《盖亚任务》。载于：《天文学与天体物理学》595（2016年11月），A1。doi: 10.1051/0004- 6361/201629272。url: https://doi.org/10.1051/0004-6361/201629272。[41] A. G. A. Brown 等。《盖亚数据发布第2版》。载于：《天文学与天体物理学》616（2018年8月），A1。doi: 10.1051/0004- 6361/201833051。url: https://doi.org/10.1051% 5C%2F0004-6361%5C%2F201833051。[42] C. J. Grillmair 和 O. Dionatos。《在斯隆数字巡天中发现一条63°冷恒星流》。载于：《天体物理学杂志》643.1（2006年5月），第 L17–L20 页。doi: 10.1086/505111。url: https://doi.org/10.1086%2F505111。[43] Sergey E. Koposov、Hans-Walter Rix 和 David W. Hogg。《利用 GD-1恒星流的六维相空间图约束银河系引力势》。载于：《天体物理学杂志》712.1（2010年2月），第 260–273 页。doi: 10. 1088 /0004 - 637x/712/1/260。url: https://doi.org/10.1088/0004-637x/712/1/260。[44] T J L de Boer、D Erkal 和 M Gieles。《进一步探究 GD-1 中的分支、团块、波纹和间隙》。载于：《皇家天文学会月刊》494.4（2020年4月），第 5315–5332 页。doi: 10 . 1093 / mnras / staa917。url: https : / / doi . org / 10 . 1093%5C%2Fmnras%5C%2Fstaa917。[45] T J L de Boer 等。《深入探究 GD1星流：密度变化与波纹》。载于：《皇家天文学会月刊》477.2（2018年3月），第 1893–1902 页。issn: 0035-8711。doi: 10 . 1093 / mnras / sty677。eprint: https : / / academic.oup.com/mnras/article-pdf/477/2/1893/25009845/sty677.pdf。url: https://doi.org/10.1093/mnras/sty677。[46] Khyati Malhan 等。《茧中蝴蝶：理解球状星团星流的起源与形态——以 GD-1 为例》。载于：《天体物理学杂志》881.2（2019年8月），第 106 页。doi: 10.3847/1538-4357/ab2e07。url: https://doi.org/10. 3847%2F1538-4357%2Fab2e07。[47] Astropy Collaboration 等。《Astropy：用于天文学的社区 Python 软件包》。载于：《天文学与天体物理学》558，A33（2013年10月），A33。doi: 10 . 1051 / 0004 - 6361/201322068。arXiv: 1307.6212 [astro-ph.IM].[48] Astropy Collaboration 等。《Astropy 项目：构建开放科学项目及 v2.0 核心软件包的现状》。载于：《天文学杂志》156.3，123（2018年9月），第 123 页。doi: 10.3847/1538-3881/aabc4f。arXiv: 1801.02634 [astro-ph.IM].[49] Astropy Collaboration 等。《Astropy 项目：维持并发展一个社区

面向……的开源项目及核心软件包的最新主要版本（v5.0）”。载于：《天体物理学杂志》935.2, 167（2022年8月），第167页。DOI：10.3847/1538-4357/ac7c74。arXiv：2206.14220 [astro-ph.IM]。[50] Adrian M. Price-Whelan 和 Ana Bonaca。《另辟蹊径：盖亚揭示主流之外的 GD-1 恒星》。载于：《天体物理学杂志》863.2（2018年8月），第 L20 页。DOI：10.3847/2041-8213/aad7b5。网址：https://doi.org/10.3847%5C%2F2041-8213%5C%2Faad7b5。[51] Adrian M. Price-Whelan 和 Ana Bonaca。*GD-1*星流周边区域的 *Gaia*数据、*Pan-STARRS* 测光数据和星流筛选掩模。版本 1.0。Zenodo，2018年6月。DOI：10.5281/zenodo.1295543。网址：https://doi.org/10.5281/zenodo.1295543。

[52] Jerzy Neyman 和 Egon Sharpe Pearson。《关于统计假设最有效检验问题》。载于：伦敦皇家学会哲学汇刊。A辑，数学或物理学论文集（1933年）。网址：https://doi.org/10.1098/rsta.1933.0009。[53] François Chollet 等。Keras。https://keras.io。2015年。[54] Martín Abadi 等。TensorFlow：异构系统上的大规模机器学习。软件可从 tensorflow.org 获取。2015年。网址：https://www.tensorflow. org/。[55] Abien Fred Agarap。《使用修正线性单元（ReLU）进行深度学习》。载于：arXiv 预印本 arXiv:1803.08375（2018年）。[56] Nitish Srivastava 等。《Dropout：一种防止神经网络过拟合的简单方法》。载于：《机器学习研究杂志》15.56（2014年），第1929–1958页。网址：http://jmlr.org/papers/v15/srivastava14a.html。[57] Takuya Akiba 等。Optuna：新一代超参数优化框架。2019年。doi：10.48550/ARXIV.1907.10902。网址：https://arxiv.org/abs/1907. 10902。[58] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。2014年。doi：10.48550/ARXIV.1412.6980。网址：https://arxiv.org/abs/1412.6980。[59] Richard O. Duda 和 Peter E. Hart。《使用霍夫变换检测图像中的直线和曲线》。载于：Commun. ACM 15.1（1972年1月），第11–15页。ISSN：0001- 0782。doi：10.1145/361237.361242。网址：https://doi.org/10.1145/361237.361242。[60] S. Lloyd。《脉冲编码调制（PCM）中的最小二乘量化》。载于：《IEEE信息论汇刊》28.2（1982年），第129–137页。doi：10.1109/TIT.1982.1056489。[61] Mariel Pettee 等。使用 CWoLa 在银河系中寻找恒星流。2023年5月。doi：10.5281/zenodo.7897840。[62] Matthew R Buckley、David Shih 和 Lina Necib。《盖亚 DR2 恒星流测试》。载于：（2023年5月）。doi：10.5281/zenodo.7897936。[63] Adrian M. Price-Whelan。《Gala：用于星系动力学的 Python 软件包》。载于：《开放源代码软件杂志》2.18（2017年10月）。doi：10.21105/joss.00388。网址：https://doi.org/10.21105%2Fjoss.00388。[64] Paul J. McMillan。《银河系的质量分布与引力势》。载于：《皇家天文学会月刊》465.1（2016年10月），第76–94页。doi：10 . 1093 / mnras / stw2759。网址：https : / / doi . org / 10 . 1093 % 2Fmnras % 2Fstw2759。[65] Aaron Dotter。《MESA等时线与恒星演化轨迹（MIST）0：恒星等时线的构建方法》。载于：《天体物理学杂志增刊》222.1, 8（2016年1月），第8页。doi：10.3847/0067-0049/222/1/8。arXiv：1601.05144 [astro-ph.SR]。[66] Jieun Choi 等。《MESA等时线与恒星演化轨迹（MIST）。I. 太阳丰度标度模型》。载于：《天体物理学杂志》823.2, 102（2016年6月），第102页。doi：10.3847/0004-637X/ 823/2/102。arXiv：1604.08592 [astro-ph.SR]。[67] Bill Paxton 等。《恒星天体物理实验模块（MESA）》。载于：《天体物理学杂志增刊》192.1, 3（2011年1月），第3页。doi：10.1088/0067-0049/ 192/1/3。arXiv：1009.1622 [astro-ph.SR]。

[68] Bill Paxton 等人。《恒星天体物理实验模块（MESA）：行星、振荡、自转与大质量恒星》。载于：《天体物理学杂志增刊》208.1, 4（2013年9月），第 4 页。doi: 10 . 1088 / 0067 - 0049 / 208 / 1 / 4. arXiv: 1301 . 0319[astro-ph.SR].[69] Bill Paxton 等人。《恒星天体物理实验模块（MESA）：双星、脉动与爆炸》。载于：《天体物理学杂志增刊》220.1, 15（2015年9月），第 15 页。doi: 10.1088/0067-0049/220/1/15. arXiv: 1506. 03146 [astro-ph.SR].[70] Bill Paxton 等人。《恒星天体物理实验模块（MESA）：对流边界、元素扩散与大质量恒星爆炸》。载于：《天体物理学杂志增刊》234.2, 34（2018年2月），第 34 页。doi: 10 . 3847 / 1538 - 4365 / aaa5a8. arXiv: 1710.08424 [astro-ph.SR].[71] Anthony Brown。PyGaia。https://github.com/agabrown/PyGaia。2013年。

模拟恒星流

我们针对恒星流发现的 CWoLa 实现首先在模拟恒星流上进行了测试。这些模拟星流在自行坐标 μ_λ 中通常高度局部化，这意味着与 GD-1 这样的真实恒星流相比，使用我们的方法可能更容易发现它们。

这些恒星流是使用 Gala [63] Python 软件包模拟的：该软件包让模拟球状星团中的恒星沿轨道穿过模拟的银河系引力势演化 [64], ，并将星流中心随机置于天空中，距离地球随机选取为 5 至 20 千秒差距。星流组成恒星的恒星属性根据 MIST [65–70] 等时线生成，假设 $[\mathrm{Fe/H}] = -1.6$ ，年龄为 100 亿年。使用 PyGaia [71] 软件包为合成恒星流恒星添加了与 Gaia DR2 数据集相符的观测误差。

一个示例模拟星流仅包含模拟天区中 886,677 颗恒星中的 1,161 颗（0.13%），其在角位置空间中的分布如图 11所示。模拟星流以独立天区的形式呈现，因此 CWoLA 一次仅应用于一个模拟天区。模拟天区未应用任何基准切割。

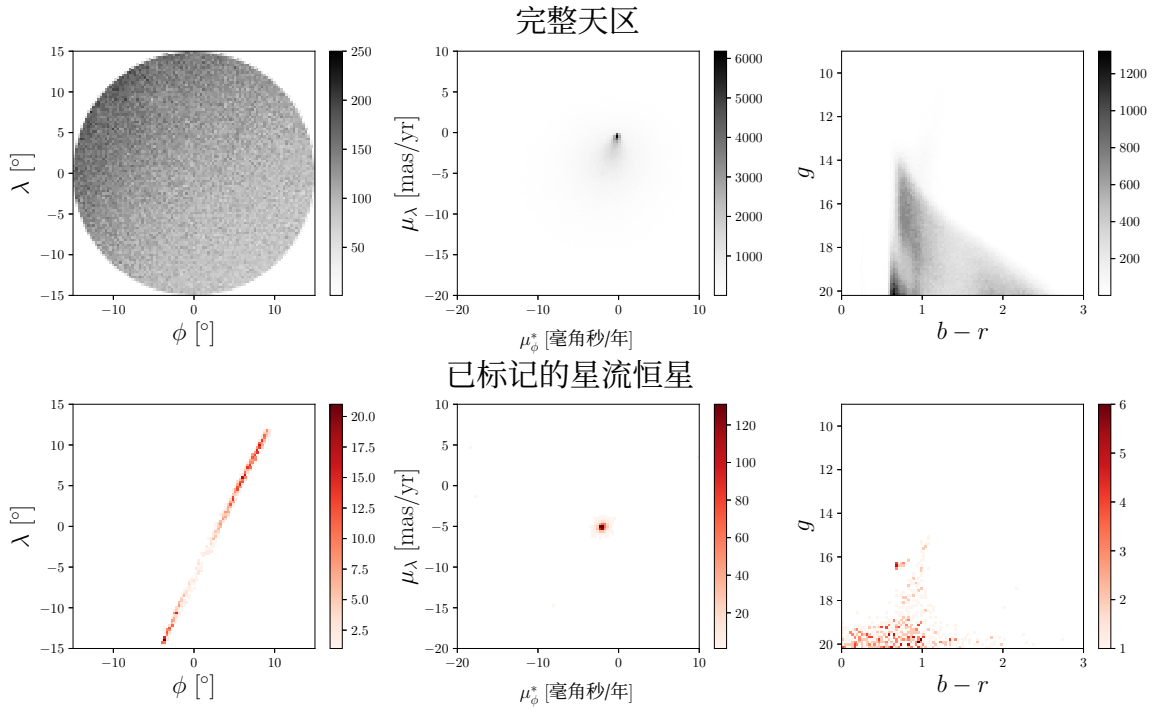


图 11。模拟天区及其中所包含的模拟星流在位置、速度和颜色空间中的分布。背景恒星和模拟星流恒星都高度集中于速度空间，但星流恒星的自行峰值相较于背景恒星，位于距离 $(\mu^*_\phi, \mu_\lambda) = (0,0)$ 更远的位置。

作为概念验证，我们利用对恒星流位置的先验知识，选取理想化的信号区域和旁带区域界限：旁带定义为围绕星流中位数值 μ_λ 、总宽度为 $\frac{\sigma}{4}$ 的窗口，而旁带定义为

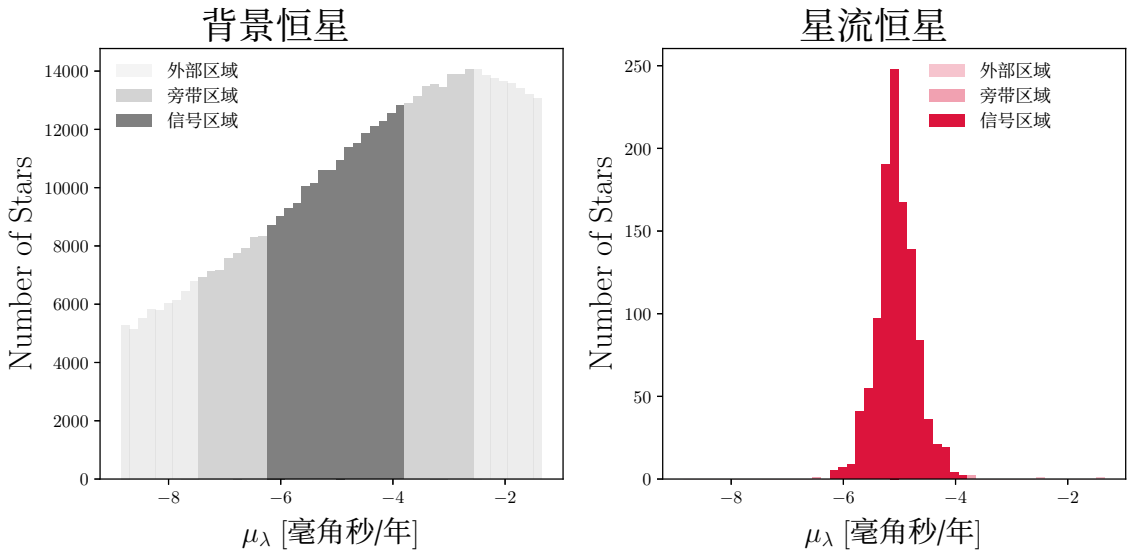
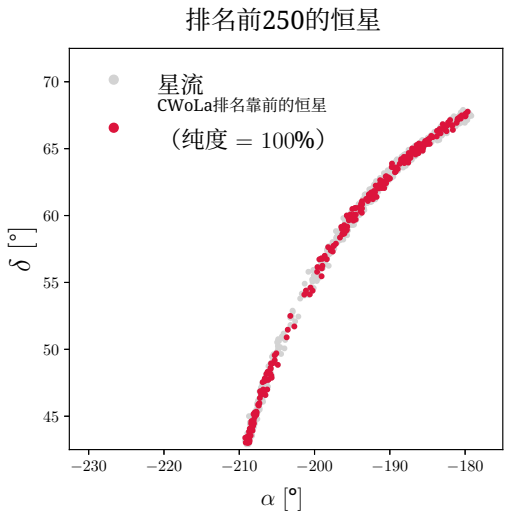


图 12。与盖亚数据集中的典型星流相比，模拟星流在角速度空间中的分布要集中得多。因此，信号区域和旁带区域被定义在星流中位数 μ_λ 周围一个窄得多的带状区域内。信号区域定义在 $\pm\sigma/4$ 内，或 $[-6.3, -3.8]$ 内；而旁带区域定义为 $\pm\sigma/2$，但不包括信号区域： $[-7.6, -6.3)$ & $(-3.8,\ 2.6]$。星流恒星几乎全部位于信号区域内。

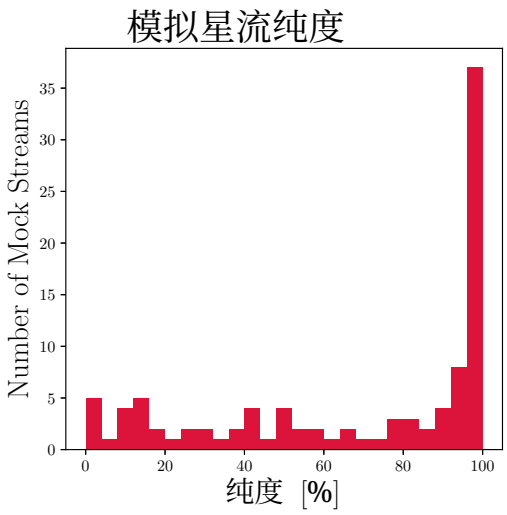
围绕信号区域、总宽度为 $\frac{\sigma}{2}$ 的附加窗口。这些针对背景恒星和星流恒星的信号区域与旁带区域绘于图12。

我们训练 CWoLa 来区分来自这些信号区域和旁带区域的事件，然后按照 CWoLa 的分类器输出分数排序，选出排名前 250 的恒星。如图 13a所示，为该天区选出的排名前 250 的恒星中，100% 都属于该天区中真实标注的星流群体。

将该技术应用于100条随机抽样的模拟星流后，有76%的星流被识别出来，且纯度为 > 0%；其中75%的星流被识别出来，且具有较高的纯度（定义为纯度大于50%）。然而，这些纯度为零的案例中，有相当一部分是沿 μ_λ方向分布更宽的星流，因此可以通过额外扫描，并选用不同的信号区域和旁带区域来进一步改善结果。若增加一次扫描，并采用更宽的信号区域和旁带区域定义（信号区域 $= \pm\sigma$ 和旁带区域 $= \pm 3\sigma$ $-$ 信号区域），则96%的星流都能以非零纯度被识别出来，其中69%具有较高的纯度。在这100条星流中，CWoLa识别结果的中位数纯度为86%。图13b显示，绝大多数模拟星流都以较高的纯度被识别出来。



(a) 一条样本模拟星流中排名前250的恒星，按CWoLa分类器输出分数排序，全部属于真实标注的恒星群体。



(b) 经过两轮 CWoLa 搜索后，绝大多数模拟星流恒星都以较高的纯度被识别出来。

图 13。在多个模拟星流中，CWoLa性能随纯度的变化。

B 逐块性能

图 14 展示了将 CWoLa 应用于 GD-1 时按天区划分的结果。21 个片区中的每个片区都分别进行 CWoLa 方法的单独应用，包括基准切割和 k-均值聚类。这些结果汇总于图 5中。CWoLa 在几乎所有天区中都实现了较高的纯度，但星流最左侧和最右侧边缘附近的天区例外，这些天区中的星流恒星相对较少（接近 $\alpha = -230°$和 $\alpha = -150°$）。

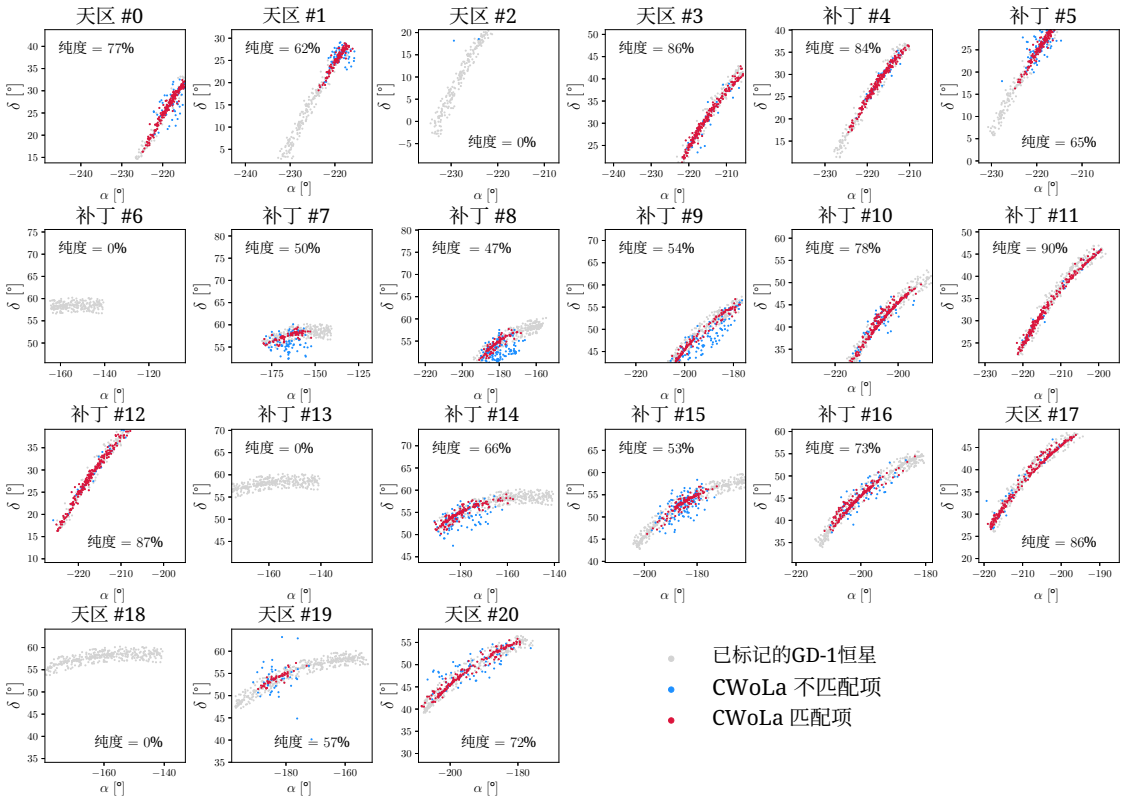


图14。从盖亚数据集中 GD-1 各天区中识别出的前 250 颗恒星可以看出，除了星流尾部最远处的天区外，CWoLa 能够在所有天区中以较高的纯度有效识别 GD-1 恒星。

C 潜在 GD-1 候选成员

下表详细列出了由 CWoLa 识别出的未标记恒星子集，这些恒星与已标记的 GD-1 集合中的恒星之间的五维距离处于最小的 10% 范围内 d （见公式 4.1），并按照 CWoLa 的神经网络（NN）分类器得分降序排列。

索引	天区	α	δ	μ_ϕ^*	μ_λ	$b-r$	g	d	神经网络得分
0	3	153.343750	37.944386	-3.648079	-11.087369	0.585669	18.889666	0.199219	0.760671
1	3	147.407532	34.807056	-4.174263	-13.030541	0.632448	18.280497	0.161793	0.710143
2	3	147.581451	33.771442	-4.031364	-12.893708	0.692181	19.217056	0.200405	0.703576
3	3	145.067444	34.774712	-3.889829	-11.528839	0.560968	18.343340	0.208592	0.687666
4	0	144.668640	29.123871	-3.294825	-9.761044	0.698603	18.599613	0.157535	0.628063
5	0	139.284058	27.534752	-2.474001	-11.759004	0.678160	18.488266	0.202521	0.608509
6	0	141.217377	21.388325	-3.016653	-12.211497	0.731615	19.056505	0.207587	0.607377
7	0	141.366760	26.641132	-2.638692	-12.904817	0.754404	19.493120	0.169499	0.603854
8	0	137.850464	25.221642	-2.823553	-13.327346	0.713997	18.919308	0.181071	0.600161
9	0	141.072662	28.944424	-3.398272	-11.644784	0.810871	19.610243	0.077491	0.590473
10	10	159.375610	43.155155	-6.771805	-13.304443	0.577917	18.163719	0.193080	0.582036
11	0	142.465027	31.453848	-3.129723	-10.795928	0.638067	17.908943	0.162005	0.582026
12	0	137.821594	25.365973	-2.537415	-10.836835	0.771456	19.665091	0.186863	0.547747
13	10	152.977966	43.714725	-5.651521	-11.580858	0.640087	18.764666	0.173924	0.541810
14	14	190.253159	58.102585	-8.347640	-3.359662	0.640982	19.427656	0.206074	0.513133
15	14	194.429871	58.634842	-8.420991	-7.967748	0.623426	18.534977	0.183633	0.506830
16	15	182.109512	56.585564	-6.765204	-12.116959	0.627577	19.014071	0.191147	0.505615
17	16	162.275238	49.602287	-7.460149	-8.779883	0.658072	19.046940	0.198527	0.505006
18	16	166.262054	49.207397	-8.196286	-8.721247	0.743465	19.535479	0.206786	0.498103
19	9	167.548462	46.377983	-6.176190	-11.992228	0.646498	19.345844	0.088779	0.497648
20	14	179.886322	55.335400	-8.907069	-8.033201	0.637199	18.149529	0.158184	0.492758
21	15	172.536072	52.980537	-7.740779	-7.965540	0.740067	19.315273	0.179650	0.487783
22	15	170.305725	55.276653	-8.972501	-9.987629	0.618397	18.751921	0.121189	0.484890
23	8	178.793396	50.679008	-7.688651	-4.308014	0.688902	19.250017	0.209937	0.484879
24	19	179.387024	53.861988	-7.553021	-9.258564	0.633465	18.817064	0.177797	0.484814
25	14	173.568787	54.718918	-8.506504	-5.805184	0.657850	19.249598	0.127450	0.484371
26	9	169.494263	48.418388	-6.111476	-10.342608	0.673502	19.317274	0.128495	0.484024
27	8	187.757996	56.693752	-7.339619	-4.749725	0.576080	18.252359	0.151545	0.484008
28	9	165.833618	46.725655	-5.527166	-10.702893	0.613531	18.830116	0.150776	0.483593
29	9	171.009125	45.997330	-5.812016	-6.994314	0.617441	18.986839	0.169053	0.483200
30	9	173.235535	49.141151	-5.340712	-9.560485	0.606213	18.940699	0.100418	0.479875
31	8	179.718201	50.653923	-7.778558	-6.846837	0.664310	19.358950	0.182553	0.477425
32	9	159.310638	44.860550	-6.879270	-10.555117	0.646944	18.803308	0.209683	0.477115
33	9	164.777161	44.976212	-6.129719	-11.014638	0.603947	18.769825	0.195550	0.476689
34	9	176.500488	54.726501	-7.533120	-8.043803	0.614882	19.896984	0.196763	0.475622
35	19	177.458374	50.769260	-7.356768	-5.745450	0.645157	18.890451	0.145692	0.475277
36	9	172.817383	47.330818	-6.053680	-7.922386	0.642633	19.386862	0.202310	0.474907
37	9	163.159119	48.486298	-6.168635	-10.865490	0.666342	18.826414	0.182255	0.472844
38	15	170.374420	52.475494	-7.904741	-6.552203	0.715446	19.745758	0.196269	0.472336
39	9	160.771454	46.261520	-5.855455	-7.474462	0.591482	19.225233	0.171725	0.464586
40	7	186.439041	56.541378	-10.644468	-4.102360	0.592939	19.715214	0.175948	0.463402
41	8	178.696075	52.531986	-7.652631	-7.286445	0.669975	18.414928	0.178466	0.462668
42	7	190.129303	57.698757	-8.130330	-5.936471	0.590197	19.154623	0.140599	0.462299
43	8	176.834106	50.947758	-7.213428	-7.684932	0.678570	19.351543	0.182765	0.461110
44	9	173.211731	46.760998	-6.690597	-8.577130	0.730625	19.276901	0.171330	0.460547
45	7	202.413101	58.419315	-8.423603	-3.303266	0.626270	19.095215	0.075321	0.460429
46	7	198.092728	58.096100	-7.937617	-3.228269	0.629757	18.509893	0.170395	0.459476
47	8	176.415344	54.566105	-7.152437	-5.265400	0.794621	19.286375	0.115260	0.458131
48	7	192.434601	54.163914	-9.025725	-3.314333	0.714384	19.720222	0.199648	0.457166
49	7	197.343521	60.127289	-8.470557	-2.185798	0.705921	19.428307	0.190847	0.456135
50	8	176.661682	52.064625	-6.882434	-8.300988	0.686596	18.789845	0.134871	0.454971
51	7	194.420959	53.819744	-7.492790	-1.907870	0.605728	19.301640	0.188744	0.452361
52	7	195.111771	56.396278	-8.144449	-3.451988	0.727108	19.597523	0.174442	0.452323
53	7	204.361725	57.704208	-8.309439	-5.412242	0.568638	19.430788	0.205758	0.452242
54	7	195.939850	55.778358	-7.729141	-2.890564	0.624128	18.510710	0.162695	0.451727
55	7	189.208954	58.567673	-8.094695	-1.741354	0.663921	19.207617	0.192408	0.449935
56	7	186.898468	55.441170	-8.953770	-4.824331	0.760971	19.275148	0.133778	0.448510
57	7	189.780121	56.009251	-9.288628	-1.816544	0.769598	19.816544	0.209066	0.447054
58	7	200.713043	56.548923	-8.371297	-4.555718	0.603445	18.767347	0.129104	0.446953
59	1	141.269653	26.119368	-3.506178	-13.209450	0.637341	17.898008	0.194991	0.444487