Draft version April 18, 2023年 使用 LaTeX 双栏 样式,
采用 AASTeX63 排版

arXiv:2304.07348v1 [astro-ph.IM] 14 Apr 2023

利用暗能量探索者和机器学习提升霍比-埃伯利望远镜暗能量实验

Lindsay R. House [1] * Karl Gebhardt [1] Keely Finkelstein [1] Erin Mentuch Cooper [1 2] Dustin Davis [1],,,
Robin Ciardullo [3 4] Daniel J Farrow [5] Steven L. Finkelstein [1] Caryl Gronwall [3 4] Donghui Jeong [3]
,,, ,,, ,,,L. Clifton Johnson [6 7] Chenxu Liu [8 1] Benjamin P. Thomas [1] and Gregory Zeimann[9] ,,,,

[1] *Department of Astronomy, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA*
[2] *McDonald Observatory, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA*
[3] *Department of Astronomy & Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA*
[4] *Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA*
[5] *Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), University of Hull, Cottingham Road, Kingston-upon-Hull HU6 7RX, UK* [6] *Adler Planetarium, 1300 S. DuSable Lake Shore Dr., Chicago, IL 60605, USA*
[7] *Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) and Department of Physics and Astronomy, Northwestern University, 1800 Sherman Ave., Evanston, IL 60201, USA* [8] *South-Western Institute for Astronomy Research, Yunnan University, Kunming, Yunnan, 650500, People's Republic of China* [9] *Hobby Eberly Telescope, The University of Texas at Austin, Austin, TX, 78712, USA*

摘要

我们介绍一项分析，利用公民科学活动改进霍比-埃伯利望远镜暗能量实验（HETDEX）的宇宙学测量。HETDEX 的目标是在 $z = 2.4$ 测量哈勃膨胀率 $H(z)$ 和角直径距离 $D_A(z)$，二者的准确率均达到百分比级别。这一准确率主要取决于探测到的莱曼-α 发射体（LAE）总数、由噪声造成的假阳性率，以及 [OII] 发射星系造成的污染。本文介绍公民科学项目暗能量探索者，旨在增加莱曼α发射体的数量，减少由噪声导致的假阳性以及 [OII] 星系造成的污染。初步分析表明，公民科学是一种高效且有效的工具，适用于最需要人眼进行的分类，尤其是在与无监督机器学习相结合时。公民科学活动中影响最大的三个方面是：1）识别探测结果中的具体问题；2）提供一个在信噪比阈值以上、经目视识别率达100%的纯净样本；3）为机器学习工作提供标签。自2022年年底以来，暗能量探索者 已汇集来自全球85个不同国家的11,000名志愿者所作出的超过350万条分类结果。通过纳入暗能量探索者 的结果，我们预计在 $z = 2.4$"处， $D_A(z)$ 和 $H(z)$ 参数的准确率将提高 $10 - 30\%$。尽管其首要目标是改进 HETDEX，暗能量探索者 已证明，它是推动科学进步、提升全球科学可及性的独特而强大的工具。

关键词：暗能量、宇宙学、公民科学

1. 引言

超新星观测发现，宇宙正在加速膨胀（Riess et al.1998；Perlmutter 1999；Riess et al. 2021），这一发现已被大量后续宇宙学观测所证实（Colless 等，2003；Dawson 等，2013；Tegmark 等，2004；普朗克合作组等，2020；DES 合作组等，2021）。学界仍在努力寻求对这种加速膨胀的理论理解（Albrecht 等，2009），而宇宙学常数被视为罪魁祸首（Weinberg，1989）。目前已有一些实验（DES 合作组等，2021；DESI 合作组，2016；Laureijs 等，2011）旨在测量——

通讯作者：Lindsay R. House lindsay.r.house@gmail.com
* 美国国家科学基金会研究生研究员

这种加速膨胀，以提高准确率并延长观测时间基线。两方面的改进都有助于缩小可用于解释加速膨胀（即暗能量）的物理模型范围。

最初的实验在 $z < 1$测量宇宙晚期的加速膨胀。尽管计划中的实验（如 DESI 和 Euclid）将显著提高膨胀率的准确率，但其红移范围仅会小幅扩大（DESI 合作组 2016；Laureijs 等 2011）。目前，不同红移下的 $H(z)$和 $D_A(z)$的不确定度为 1.8%–3%（总结见 DESI 合作组（2016））。重子振荡光谱巡天（BOSS）及其扩展项目 eBOSS 以 1.8% 的准确率测量 $D_{\rm A}$，在 $z = 0.5$处（Bautista 等 2020；Gil-Mar´ın 等 2020），在$z = 0.7$ 处的准确率为 2.0%（de Mattia 等 2021），在 $z = 2.3$ 处为 ~ 3%（du Mas des Bourboux 等 2020）；暗能量巡天（DES）在 $z = 0.84$处的不确定度为 2.7%（DES 合作组等 2021）。预计包括暗能量光谱仪（DESI）和 Euclid 在内的各种任务，将在 $z \sim 1$ 处达到 ∼0.5% 的精度（DESI 合作组 2016；Laureijs 等 2011）。

为了探测暗能量性质随宇宙时间的演化，必须尽可能覆盖较大的红移范围。霍比-埃伯利望远镜暗能量实验（HETDEX）旨在以 $1.9 < z < 3.5$的准确率研究膨胀速率，与最佳低z 实验的准确率相当（Gebhardt 等人，2021年）。HETDEX 将测定至少一百万个莱曼α发射体（LAE）星系的红移，这些星系来自α 。我们的方案是使用一台由 74 个积分视场单元组成、并连接至 156 台光谱仪的仪器，在 350–550 n m 的光谱范围内观测 540 平方度的区域。
$1.9 < z < 3.5$
HETDEX 将利用这些数百万个莱曼α发射体星系测量并分析星系功率谱和星系关联函数的完整形状，包括宇宙的重子声学振荡（BAO），$1.9 < z < 3.5$ 以确定该时期的暗能量密度。沿视线平行和垂直方向的聚集测量结果可通过阿尔科克–帕钦斯基检验，对哈勃参数、 $H(z)$ 以及角直径距离、 $D_{\rm A}(z$提供约束（Alcock & Paczy´nski 1979；S´anchez 等人，2014年）。角度平均后的聚集测量结果可确定平均距离、 $D_{\rm V}(z)$，其

$$D_{\rm V}(z) \propto \frac{D_{\rm A}(z)^2}{H(z)}. \qquad (1)$$

根据预期的测量莱曼α发射体样本体积和密度，HETDEX 将测定哈勃膨胀率、 $H(z)$ 以及角直径距离、 $D_{\rm A}(z)$，两者在 .附近的准确率均为 0.8%，这将是该时期最准确的哈勃膨胀率测量（Gebhardt 等人，2021年）。这些准确率意味着体积平均距离的总体准确率将$D_{\rm V}$低于 0.7%。

HETDEX 得出的宇宙学限制的准确率主要取决于 LAE 源的数量、由噪声导致的假阳性（FP），以及 [O II] 污染。我们优化了检测与分类算法（Davis 等人，2023年；Gebhardtet al. 2021；Mentuch Cooper etal.2023；Farrow 等人，2021b；Leung 等人，2017年），使这些参数都达到规定要求。本文介绍了我们如何利用人眼的模式识别能力，进一步改进项目算法，并有望拓展至新的领域，例如更低的信噪比和较差的发射线拟合。

公民科学是科学家与公众合作，共同实现更大科学目标的一种方式。为了利用人眼识别模式的能力，我们创建了一个公民科学项目，暗能量探索者[1],以改进 HETDEX 数据产品。

其他大型巡天也将公民科学贡献纳入分析流程，例如，LIGO 的 Gravity Spy 项目帮助剔除噪声（Bahaa- dini 等人，2018年）、TESS 的行星猎人（Eisner 等人，2021年）以及卡特琳娜外太阳系巡天（Drake 等人，2014年）对候选天体进行评估，从而提高完备度。*Galaxy Zoo* 是一个成熟的项目，旨在对斯隆数字化巡天中的星系进行分类，并为大型巡天中的公民科学奠定了基础（Lintott 等人，2008年；Land 等人，2008年）。Galaxy Zoo 最初只是一个简单的网站，如今已发展壮大，并创建了一个主平台——*Zooniverse*[2]，，它现已成为全球最大的公民科学平台。*Zooniverse* 汇集了数十个不同学科的公民科学项目，这些项目已促成数百篇论文发表。暗能量探索者 便是其中之一。***Zooniverse*** 的 使命是“促成那些原本无法实现或不切实际的研究”。秉承这一使命，暗能量探索者 旨在提高 HETDEX 的准确率，同时让参与者率先观察天文源，并教会他们对霍比-埃伯利望远镜观测到的数百万个源进行分类。

[1] https://www.zooniverse.org/projects/erinmc/dark-energy-explorers[2] https://www.zooniverse.org/

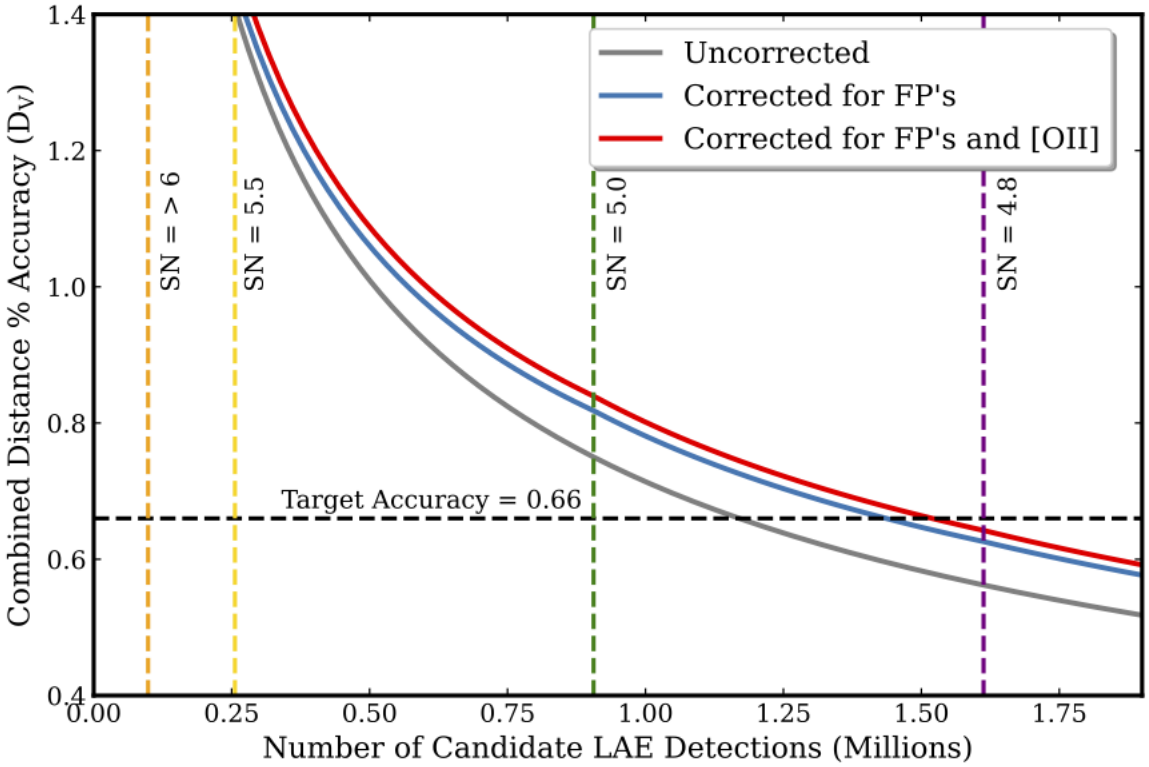


图1. 根据 HDR2 结果的建模，该图展示了 D_V 的准确率随探测到的莱曼α发射体数量和信噪比的变化关系。主要关系是， D_V 与莱曼α发射体数量的平方根成反比；灰色线表示未对假阳性率或 [O II] 污染进行校正时的情况。该图表明，假阳性（FP）会如何影响宇宙学参数 D_V 的准确率（蓝色线）；若进一步计入 [OII] 污染，则如红色线所示。黑色水平线代表整个 HETDEX 巡天的目标。对于给定信噪比限值所能达到的精度，由灰色、蓝色和红色曲线与表示不同信噪比阈值的垂直虚线的交点给出。请注意，由于存在污染，蓝色和红色曲线的准确率较低。这里假设星系偏置为 $b_{\rm LAE} = 2$。为获得 D_V 的准确率所需的信噪比阈值和/或莱曼α发射体数量，决定了 *Dark Energy Explorers* 公民科学活动的目标。

HETDEX 不会预先筛选源，这与其他大尺度巡天截然不同。我们使用 74 个积分视场单元，对 540 平方度的天空进行铺测，填充因子为 1/4.5。这些单元连接至 156 台光谱仪，覆盖 350–550 nm 的波长范围，分辨本领为 750–950。随后，软件会搜索每个光谱和空间分辨单元中的发射线，包括莱曼α发射体发出的发射线。在整个观测期间，HETDEX 将获取约十亿条光谱和一万亿个分辨单元（空间和光谱）。我们预计能从这万亿个分辨单元中找到约 1.3 百万个莱曼α发射体、92万个 [O II]发射体，以及许多恒星、流星、小行星等。从一万亿个分辨单元缩减到一百万个莱曼α发射体，取决于实验所采用的信噪比（S/N）阈值。在较低的 S/N下，能够检测到更多发射线，但代价是伪影导致的假阳性也会增多（Mentuch Cooper 等，2023）。HET-DEX 正在探索如何利用全部一万亿个分辨单元，这需要对像素级缺陷进行精准控制。

HETDEX面临的一项根本挑战，是筛查数据并从中区分出莱曼α发射体

与其他探测结果，包括假阳性、 [O II]发射星系、流星轨迹及其他类似谱线的特征。HETDEX已优化其算法，以区分各种源，但人眼也能提供一些优势，从而改进这项巡天。小型研究团队不可能对数百万个源进行目视审核，但我们意识到，如果能够开展此类审核，就能显著改进对 $H(z)$和 $D_A(z)$的测量。我们的目标是利用公民科学对数百万个HETDEX源进行分类，并帮助保持较低的污染率。

目视审查有两个重要方面。首先，我们利用公民科学帮助识别由非高斯噪声造成的源，从而降低假阳性率，并生成更干净的机器学习训练集。其次，我们能够探索那些更难通过算法分类的情形，例如信噪比较低和/或单条发射线拟合的卡方值较高的探测结果（见第 3.2 节）。例如，活动星系核（AGN）具有多种发射线轮廓，人眼很容易辨别，但对于使用单条未分辨发射线的通用算法而言，可能难以识别。通过纳入这些额外的源，我们提高了膨胀率测量的准确率。目前的结果表明，HETDEX 莱曼-α发射体样本最多可增加 50%，从而使距离估计提高 $\sim$ 20%（见图 1）。我们的目标是将目视审查与机器学习结合起来，而实现这一目标的第一步是为机器学习模型创建标签。创建这些标签来训练机器，为 暗能量探索者 和 HETDEX 创造了丰富的机遇。

本文讨论了我们如何利用 暗能量探索者 帮助 HETDEX 达到既定指标。在第2节中，我们将讨论如何利用从 暗能量探索者 收集的数据提高距离测量的准确率。第3 节聚焦于 暗能量探索者 界面，以及我们如何培训公众成为 HETDEX 天文学家。第4 节介绍 暗能量探索者 的初步结果。第5节将展示我们如何利用机器学习，将暗能量探索者 纳入 HETDEX 数据库；第6节将讨论结论。

2. 提高 HETDEX 距离测量的准确率

在整个巡天期间，HETDEX 将获取约十亿条光谱和一万亿个分辨单元（Gebhardt 等人，2021年；Mentuch Cooper 等人 2021年）。HETDEX 完成后，我们预计将获得超过一百万个红移介于 $1.88 < z < 3.52$之间的远距离莱曼α发射体红移，以及超过一百万个附近 [O II] 星系的红移，其 $z < 0.5$。这些星系将用于宇宙学分析。我们的要求是：由噪声导致的假阳性率低于 $< 10\%$，由 [OII] 发射体造成的污染率低于 $< 2\%$，且莱曼α发射体总数超过一百万。如Mentuch Cooper 等人(2023年）和 Davis 等人(2023年）所示，我们几乎没有误差余地地达到了这些指标。我们的目标是将信噪比要求降低，并提高相对于高斯拟合测得的卡方值。这样可以增加莱曼α发射体的数量，同时维持较低的假阳性率和污染率。为保持这些比率处于较低水平，我们采用目视审查和机器学习。

HETDEX HDR2 包含 ~ 50,,000 个发射线轮廓拟合良好（信噪比 > 5.5 且 $\chi^2 < 3$）并被分类为莱曼α发射体的天体（Mentuch Cooper 等，2023）。该目录来自覆盖 25 平方度[2] 天空的连续光纤光谱观测。阈值之外无疑还有许多源；例如，具有宽发射线的活动星系核会偏离单线拟合，导致卡方值较高。我们预计，通过稳健的目视审查和机器学习工作，可从目录中额外提取 $\sim$ 10%，同时保持较低的假阳性率和污染率。

图1 量化了主宇宙学参数 D_V 的信噪比截断、卡方阈值、假阳性率和污染率之间的权衡。对于这幅图，我们需要假定莱曼α发射体星系偏置，因为综合距离准确率随偏置线性提升（偏置值越高，准确率越好）。我们正在准确测量这一偏置（Farrow 等，准备中），而在这幅图中，我们假定 $b_{LAE} = 2.0$。对于我们能够控制的属性（信噪比、假阳性率、卡方、 [OII] 污染），我们根据Davis 等人（2023年）和 Mentuch Cooper 等人（2023年）的分析估算这些比率。如图 1 所示， D_V 的准确率取决于这三个因素。黑色水平线表示整个 HETDEX 巡天的目标值。主要关系是， D_V 与莱曼α发射体数量的平方根成反比；灰线表示未对假阳性率或 [OII]污染进行校正时的情况。垂直线表示图例中所示的不同信噪比截断。假阳性带来的影响会降低距离准确率（蓝线），进一步计入 [OII]污染后，准确率还会进一步降低（红线）。 D_V的不确定性随假阳性率线性变化，因为假阳性既会影响真实莱曼α发射体的数量，也会增加

白噪声。 D_V 的不确定性随 [O II] 污染呈二次变化，因为污染会影响莱曼α发射体的数量，并将近邻星系的聚集功率叠加到莱曼-α 发射体功率谱上。Farrow 等（2021a）表明，将 [O II] 污染限制在 2% 以内，不会显著影响科学要求。在图1中，我们假定 [O II] 污染率为 0.013，该数值基于 Davis 等人（2023）的研究。假阳性率来自对莱曼α发射体确认率的测量，因此我们的假阳性率将低于确认率。Gebhardt 等人（2021）和 Mentuch Cooper 等人（2023）给出的初步假阳性率估计分别为 $<$0.24（对应 $S/N = 4.8$）、 $<$0.19（对应 $S/N = 5$）、 $<$0.05（对应 $S/N = 5.5$）和 $<$0.01（对应 $S/N > 6$），这些数值均为上限。

灰色、蓝色和红色曲线与黑色水平虚线的交点给出了 D_V 的不确定性。红线的 D_V 不确定性最大，因为它考虑了污染和假阳性，我们在最终预测中采用的正是这条线。随后可以读出达到 D_V 所需准确率所要求的信噪比阈值，这也说明了我们开展 暗能量探索者 公民科学活动的目标。

鉴于主要修正源于假阳性和 [OII] 污染，我们将公民科学活动重点放在这些问题上。如下文所述，假阳性率涉及对真实源和虚假源的目视区分，而 [OII] 污染则涉及对远处星系和近邻星系的区分。

3. 开发暗能量探索者

截至2022年底，HETDEX 数据收集工作已完成50%，对2019年首次内部数据发布的分析表明，当前主要需求是对源目录进行仔细的目视审查（Gebhardt 等人，2021年）。为解决这一问题，我们设计、创建并于2021年2月下旬推出了全球公民科学项目，暗能量探索者。

暗能量探索者的目标是解决一个难题：数据量过大，无法由人数有限的内部团队完成分类。*HETDEX* 合作组成员最初进行了目视审查并提供了源分类结果，但耗时过长，无法对整个数据集完成分类。在 *Zooniverse* 上创建项目，促成了来自世界各地数千名志愿者的参与；该平台拥有数百万名注册志愿者，用户只需通过互联网，就能在任何智能手机、平板电脑或台式电脑上访问该项目。除了帮助招募志愿者外，Zooniverse 团队还从项目创建阶段开始提供支持，直至项目启动乃至后续阶段。除了实现更大的科学目标外，项目教程（详见第 3.1节）还旨在让任何人都能参与其中，即使没有科学背景也不例外。

我们为暗能量探索者设计了两种不同的工作流程：“近处与遥远星系”和“在噪声海洋中寻找信号”。第一种工作流程旨在区分 [O II] 发射天体中位于 $z < 0.45$ 的（“近邻星系”）与 $1.88 < z < 3.52$ 莱曼α发射体（“遥远星系”）。“近处与遥远”星系工作流程针对 [O II] 污染问题，并帮助我们优化判别算法。因此，目标天体都具有 $S/N > 5$，包括候选莱曼α发射体、活动星系核、 [OII] 星系和恒星。第二种较新的工作流程“在噪声海洋中寻找信号”旨在处理假阳性。由于假阳性是由噪声造成的，因此更难识别。因此，该工作流程重点对 $S/N > 6$ 候选莱曼α发射体进行分类。本文通过机器学习工作详细研究的正是这一子集，而较低信噪比的情况将留待未来工作探索。

3.1. 教程与指南

暗能量探索者 可通过 *Zooniverse*应用程序或网站访问。参与者可以创建账户，以保存分类结果和数据，也可以选择匿名参与。暗能量探索者 项目可通过其网址直接访问，也可在 Zooniverse 项目页面中选择；该项目列在“太空”或“物理学”类别下。随后，志愿者可以在项目主页选择一种工作流程：“近处与遥远星系”或“在噪声海洋中寻找信号”。选择工作流程后，参与者将通过教程学习如何对 HETDEX 数据进行分类（第 3.2节和图 2中将进一步讨论）。

暗能量探索者 的基础是简单易懂的教程。该教程通过将分类流程简化为易于理解、没有专业术语的任务，让公众成员无需任何天文学或一般科学背景，也能成为业余 HETDEX 天文学家。参与者阅读教程后，会获得若干标准，以便在工作流程中从两个二元选项中进行选择。我们选择将分类结果限制为二选一，以最大限度地提高分类速度。例如，在提供左滑／右滑分类选项的移动设备上，一位暗能量探索者 不到一分钟就能对二十个源进行分类。这确保用户能够在移动

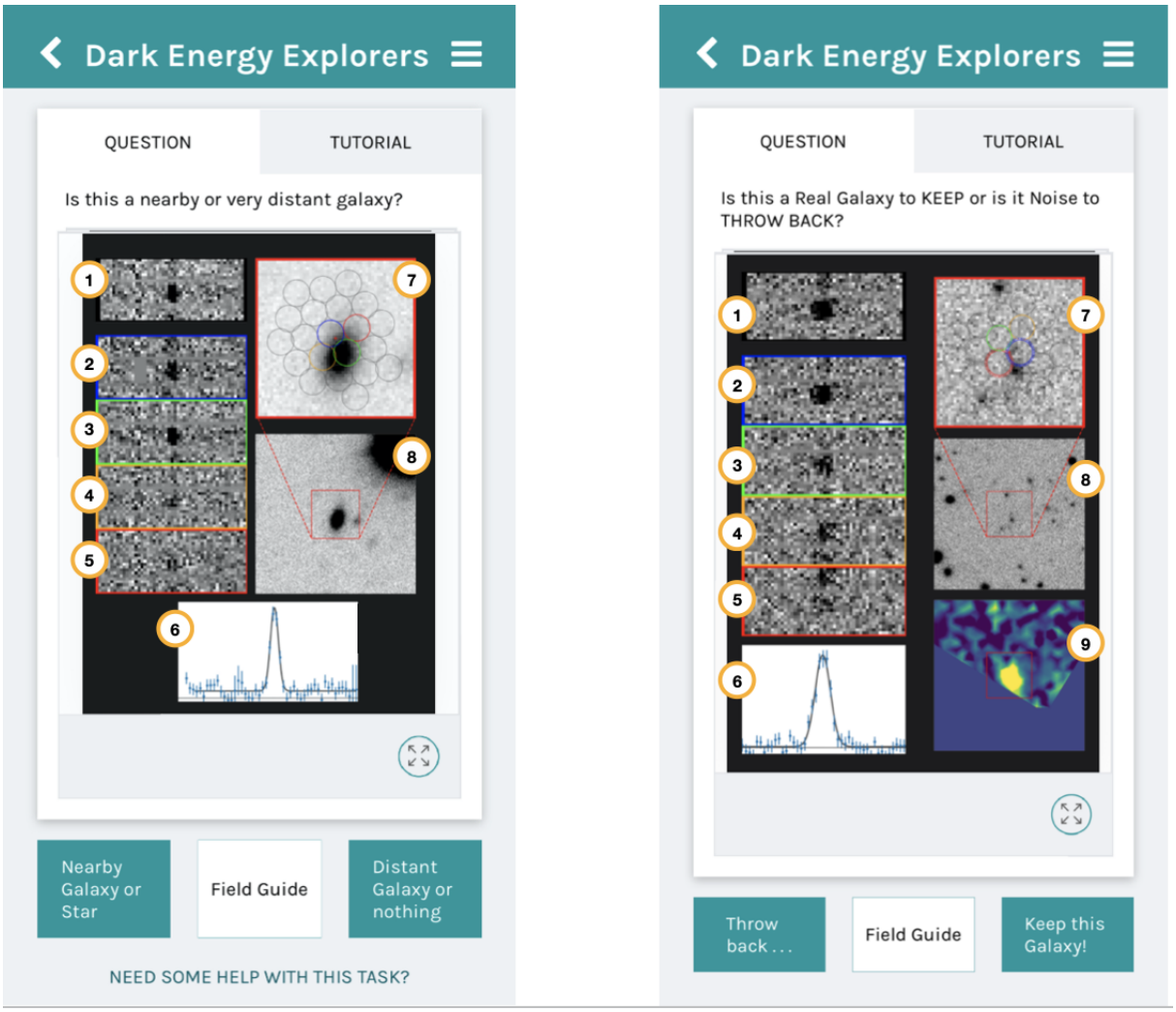


图 2. 上图展示了 *Dark Energy Explorers* 的移动版。在任一工作流程中（左侧的“近处与遥远星系”或右侧的“在噪声海洋中寻找信号”），参与者都会看到来自霍比-埃伯利望远镜（HET）的 HETDEX 探测结果的各种可视化，并通过二选一的方式对天文源进行分类。教程入口位于右上角，现场指南入口位于底部中央。图像摘要：(1)–(5) 二维光纤截取图，(6) 一维发射线拟合，(7) 带光纤位置的成像邮票图（放大视图，$9'' \times 9''$），(8) 成像邮票图（缩小视图，$30'' \times 30''$），(9) *Lineflux Map*。图像 (9) 仅包含在“在噪声海洋中寻找信号”工作流程中。

设备上轻松使用滑动功能，通过简单选项避免多重或嵌套选择。

“近处还是远处”工作流程中，用户需要考虑的主要标准包括：

1. 天体的相对大小，
2. 发射线的强度，以及
3. 发射线在至少一条光纤光谱中的表现。

例如，近处的星系在图像中通常表现为一个巨大、明亮的分辨源，而遥远的星系通常较小且暗淡。参见图3进行比较。

用户在“在噪声海洋中捕捞信号”工作流程中需要考虑的主要标准是：

1. 所采集数据的质量，
2. 发射线的强度，以及 3. 发射线在至少一条光纤光谱中的表现。

参见图4 进行比较。除了教程之外，参与者还可以在主分类页面上使用简短易懂的现场指南。该指南会快速提醒他们两种工作流程的选择标准。图 2 展示了暗能量探索者 的移动版本，参与者可以在其中

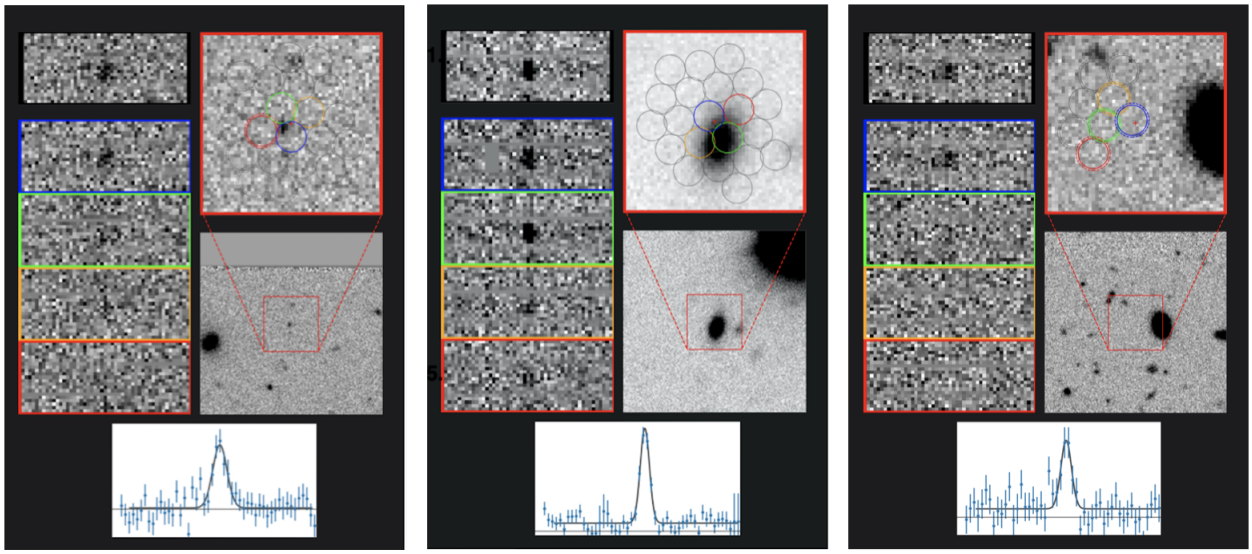


图 3. 上方展示了来自 *Dark Energy Explorers* “近处与远处”工作流的“迷你图”示例。从左至右：一个遥远的莱曼-α 发射星系，一个邻近的 [OII] 发射体，以及一个棘手的案例，可能需要更多信息。

在分类过程中，随时访问现场指南，并选择返回完整教程。

3.2. 界面和数据输入

我们结合多张图像和光谱数据，制作了我们称之为“迷你图”的图像。我们称它们为“迷你图”，是因为它们是更完整、复杂的 Emission Line eXplorer (ELiXer)（Davis 等，2021 年）的精简紧凑版。这些图像旨在保持小巧，并兼容台式设备和移动设备。由于我们希望该工作流能在各种屏幕分辨率和宽高比的移动设备上使用，因此决定让可视化内容保持简洁紧凑。此外，我们希望没有科学背景的人也能轻松访问该网站；因此，图像中不含术语和数字，以免分散参与者的注意力或让他们望而却步。

“mini”由“近处与遥远星系”工作流程的面板 (1) - (8)，以及“在噪声海洋中寻找信号”工作流程的图像 (1) - (9) 组成。请参见图 2。这些图像的详细定义如下：

(1) - (5) 二维光纤截取图：沿光谱方向，在探测线中心 $\pm 40^{\circ}$A 范围内、横跨探测器 ± 1 根光纤的五幅截取图。光谱图像已扣除天空背景。图像 (1)（黑色高亮）是所有贡献光纤的加权和。下方各行的图像 (2) - (5)（蓝色、绿色、橙色、红色）是按与源位置的距离排序的四根光纤。

(6) - 一维谱线拟合：最终得到的一维光谱及发射线拟合。

(7)含光纤位置的邮票图：所有参与此次探测的光纤覆盖范围，叠加显示在某个区域的深度地基成像上；该区域的 $9'' \times 9''$ 中心位于发射线信噪比最大值对应的空间位置。出于内部分类目的，我们提供了多个成像星表，详情见Davis 等人，(2023年)。不过，在当前工作流程中，我们仅显示由 HSC 昴星团战略计划（HSC-SSP）通过昴星团望远镜上的 Hyper Suprime-Cam 获取的辅助 r波段成像，以及 HSC HETDEX 巡天（HSC-DEX）获取的成像（Aihara 等人，2018年）。这样可以减少由图像质量和通量灵敏度不均一造成的混淆。四根彩色光纤与图像 (1)–(5) 左侧标出的颜色相对应。外圈为虚线的光纤位于探测器边缘。探测的点扩散函数加权中心以红色十字标出。

(8) -成像邮票图：从辅助成像数据中截取的邮票图，显示以发射线空间位置为中心、大小为 $30'' \times 30''$ 的区域。这是图像 (7) 的缩小版本，用于显示可能的目录对应源或附近源。

(9)- 线通量图：以发射线中心为基准，对 $\pm 3\sigma$ 区域进行波长折叠后得到的通量强度图。由于该区域恰好延伸至探测器边缘之外，因此这幅图的下半部分为空白。有关这些小图的生成方式，详情请参阅 Gebhardt 等人(2021年) 和Davis 等人(2023年)。

这些望远镜图像展示了天空中一个随机区域的探测，中心是一个天体（即星系、恒星，

从样本数据库中随机选取的活动星系核）。暗能量探索者 避免在教程与现场指南中使用数字和天文学术语，让任何人都能轻松接受培训，成为 HETDEX 志愿者。

3.3. 教育与参与

虽然 暗能量探索者 的主要关注点是推动 HETDEX 的科学进步，但其其他主要目标还包括教育、科普和公众参与。为此，我们设计了一项功能，让参与者在对星系进行分类时也能学习相关知识。Galaxy Zoo 等其他项目也采用了这种方法，以持续吸引参与者并延长他们参与项目的时间（Lintott 等人，2008年；Land 等人，2008年）。借鉴 Galaxy Zoo 项目的经验，暗能量探索者 旨在为参与者提供与彼此及 HETDEX 团队交流的途径，以便进一步了解相关科学（Raddick 等，2019）。沿用这些方法，我们也为暗能量探索者的常驻用户设定了类似的目标。每当个人完成 50 个源的分类，便会收到一条弹出消息，引导他们进一步了解 HETDEX 背后的科学、暗能量和天文学常识。这种“升级”激励机制既能持续吸引公众参与项目，也能让他们了解天体物理学相关主题，并增强他们作为科学家的信心。

为了与参与者互动并解答他们可能提出的问题，暗能量探索者 设有一个“讨论”板，供参与者发表评论和／或提问。这些“讨论”板由 HETDEX 团队监控，参与者可以借此与科学家交流或提出常见问题。迄今为止，已有 500 多名参与者参与讨论，并发布了 5000 多条评论和图片。讨论板为参与者提供了一个空间，让他们查看每个工作流程中已分类源的诸多示例。HETDEX 团队成员可以回复评论，确认或否定分类结果，以解答这些问题，其他参与者也可以查看讨论内容。

4. 暗能量探索者的结果

4.1. 分类与用户统计

暗能量探索者 于2021年2月上线，此后已收集到来自全球85个不同国家的11,000名志愿者所作的约350万条分类结果。工作流程分类及 HET-DEX 数据发布版本中的筛选结果见表 1 。这些结果来自“近处与远处”工作流程中的300多万条分类结果——流程。这些分类结果涉及 ∼209,000 个被分类为莱曼α发射体或 [OII] 星系的源。“近处与远处”工作流程最初上线，因此累计分类结果更多。“在噪声海洋中寻找信号”于2021年10月上线，共有14万条分类结果，产生了 ∼14,000 个已完成的研究对象，这些对象是机器学习工作的重点（见第5节）。

自项目启动以来，约 60% 的用户已对 50 多个天体进行分类，并收到弹出提示（见第 3.3 节）。在这些用户中，三分之二持续参与该项目，完成了 100 多项分类。换言之，超过 40% 的参与者至少完成了 100 项分类。

通过目视审查，我们承认人为偏差的存在，同时也致力于确保人工分类的可靠性。为了为每个源提供可靠的条目，我们必须对一定数量参与者的结果取平均值。为每个对象选取的分类结果数量，需要在分类效率与借助统计方法减少错误分类之间取得平衡。我们发现，10 个分类结果所得数值与更大的样本量同样一致，同时还允许对更多对象进行分类。二元分类的优势在于，可以通过简单地对各项工作流程中的单项分类结果取中位数或平均值，轻松得出分类统计量。在这项工作中，我们对十项人工二元分类结果取平均值，以提高编制准确目录的可信度（Santos-Fernandez 和 Mengersen（2021））。最终产暗能量探索者 概率（DEE 概率），表示根据相应工作流程通过目视审查判定该源为遥远星系或真实信号的可能性。

我们注意到， N 分类结果的标准差也可能具有应用价值，因为它能指出公民科学家对某个目标存在严重分歧的情况。HETDEX 团队的内部成员会对这些特殊情况进行目视跟进。因此，暗能量探索者的测量结果减少了 HETDEX 内部小团队必须通过目视检查的源的数量，并提供了一种有效方法，可在 HETDEX 巡天中识别罕见且独特的源。在“在噪声海洋中寻找信号”的 140,000 个分类结果中，标准差大于 0.5 的占 0%。

4.2. 近邻与遥远结果

来自“近处与远处”工作流的公民科学家数据，正通过比较目视审查得出的 DEE 概率与模型概率，帮助验证当前研究（Davis 等人，2023年）。我们获得了超过三

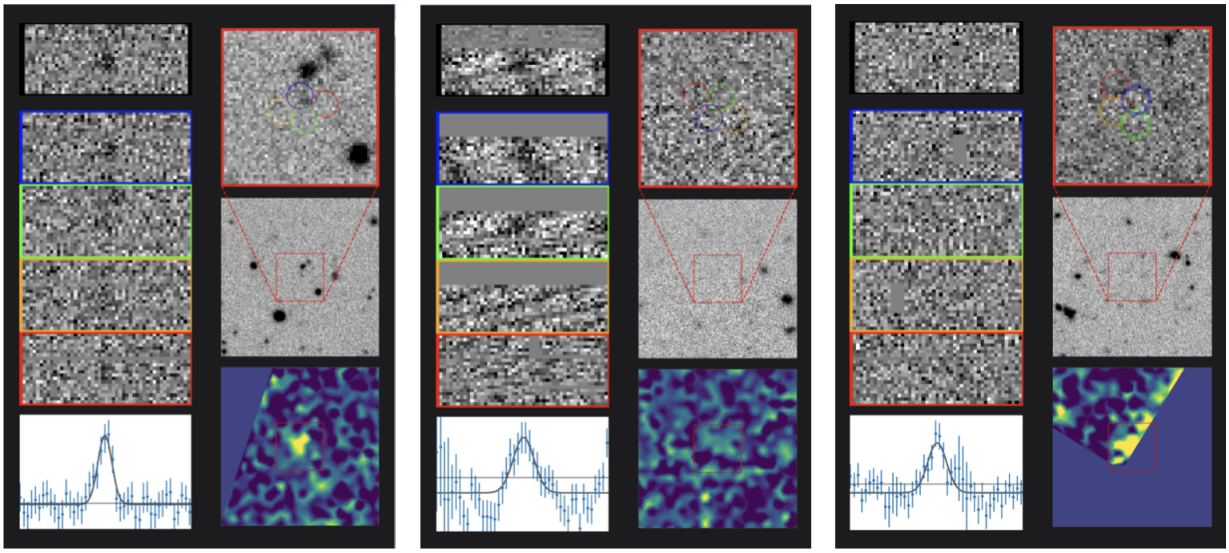


图4. 上面展示了 *Dark Energy Explorers* “在噪声海洋中寻找信号”工作流中的一些“迷你图”示例。从左到右依次为：保留（真实星系/发射线）、退回（错误探测），以及一个可能需要更多信息才能判断的棘手案例。

百万条分类结果，来自暗能量探索者，由此对209,000 个源完成分类，因为每个源至少由 10 位不同的参与者查看。计算出 DEE 概率后，我们便能比较 [O II] 和莱曼α发射体的分类结果。两者不一致之处让我们得以深入研究这些源，并确定分歧的根源。在某些情况下，暗能量探索者参与者给出的分类结果是正确的，我们会利用这些情况改进模型（Davis 等人，2023年）。暗能量探索者给出的分类结果与 EliXer 模型的吻合度超过 92%。暗能量探索者取得的成功，大幅减少了 HETDEX 团队手动检查源所花费的时间（Davis 等人，2023年）。这是因为我们采用“近处与远处”工作流，为问题区域进行目视审查、调整算法提供了一种途径。得益于暗能量探索者的努力，我们如今可以依靠Davis 等人（2023年）提供的 [OII] 污染率（见图1）。公民科学家帮助确认了 [O II]发射源，因此我们已停用这一工作流，转而专注于“在噪声海洋中寻找信号”所采用的机器学习方法。

4.3. 在噪声海洋中寻找信号：结果

HETDEX 的原始数据产品包含许多像素级事件，例如电荷陷阱、热像素以及校准随时间发生的变化。由于算法往往难以识别这些不Gebhardt 等人，2021年完善之处（通常需要人工查找（Mentuch Cooper 等，2023）。人眼善于识别这些特征，尤其是在高S/N分辨率下，因为它们往往会产生相关残差，或在二维电荷耦合器件（CCD）成像中形成明显特征。因此，暗能量探索者的主要目标是从目录中进一步剔除虚假源。本文着重关注假阳性，因此 DEE 概率为 1.0 表示目标是真实的，概率为 0.0 则表示这是一次虚假探测。因此，DEE 概率仅适用于“在噪声海洋中寻找信号”工作流程。

自 2022年年底以来，我们已为“在噪声海洋中寻找信号”工作流完成了超过 14 万次分类，从总计 6 万个输入中筛选出 1.4 万个源。图5 展示了参与者（即暗能量探索者）相较于一位 HETDEX 团队成员，在不同 DEE 概率下的表现（即准确率）。图中数据按 0.1 的概率区间分组，每组包含 200 个源。这项分析旨在确定 DEE 概率阈值，以便准确剔除假检测，同时避免剔除过多真实源。目前，暗能量探索者主要希望排除明显的假阳性。同样，在计算 DEE 概率时，假源和真实源分别对应 0 和 1。因此，我们关注暗能量探索者的分类结果在源很可能是伪影（即概率低于 0.4）时的准确率。

表 1。 *Dark Energy Explorers workflows*

工作流名称	启动 date	父级 样本	信噪比截断	Down 选择	源 完整	总计 分类结果
近处与远处	2021年2月	500,000	5.5	209,000	114,000 莱曼α发射体 95,000 [OII]	3.1 M
在噪声海洋中寻找信号	2021年10月	700,000	6	60,000	$14,000^a$	140,000

[a] 持续数据收集

图表显示，当目标是真实源的概率较低时，暗能量探索者最常与 HETDEX 天文学家意见一致。下文将介绍如何汇总暗能量探索者给出的概率，并将其对伪影的分类结果与机器学习工作结合起来。

5. 结合暗能量探索者的结果与机器学习

主要目标是利用暗能量探索者改进 HETDEX 星表。由于 HETDEX 是一个历时多年的项目，使用了大量探测器，因此存在许多仪器和校准问题，可能产生模仿真实源的特征。次要目标是将分类结果用作无监督学习工作的训练集。我们从这两个方面分析公民科学的输出。

为进一步确保从数据中去除假阳性，我们在此探讨将机器学习纳入整体分类的做法。有几个方面值得考虑。首先，我们可以将 DEE 分类用作机器学习的标签。其次，我们可以利用机器学习为暗能量探索者 训练提供参考。第三，我们可以将暗能量探索者 分类结果与机器学习结合起来，以更有力地支持最终分类。例如，将无监督学习与暗能量探索者 概率相结合，使我们能够对完整的源样本应用阈值，作为从目录中移除虚假探测的额外方法。

参与者共进行了约 140,000 次分类，因为每个源至少由 10 名不同的参与者检查。在这 14,000 个已分类的探测结果中， ∼有 2,000 个的 DEE 概率低于 0.1，因此最有可能是假阳性。此外， ∼ 4,000 个源的 DEE 概率小于或等于 0.3。从图5 可见， ≤ 0.3假阳性（DEE 概率 ≤ 0.3）的判断一致率暗能量探索者保持在 92% 以上。当 DEE 概率介于 $0.3 - 0.7$之间时，准确率也会下降。如图4 ，我们必须进一步查阅信息（例如 EliXer），并请 HETDEX 团队收集更多信息，以便对该源进行分类。

为了减少 HETDEX 中的假阳性，我们将把分析重点放在假阳性上。我们使用无监督学习来映射暗能量探索者的概率值与光谱特征之间的关系。我们通过 t 分布随机邻域嵌入 t-SNE（van der Maaten& Hinton 2008）来实现这一点，该方法来自 `scikit-learn` Python 软件包（Pedregosa 等， 2012）。

我们分析了完整光谱的 t-SNE 输入，以及 $\pm$ 50 $^{\circ}$ 以检测到的发射特征为中心的区域。随后，t-SNE 将高维数据集转换为这些数据点在低维空间中的表示。在这里，我们将数据降至二维，因此 n- 分量设为 2。接下来设置的参数是困惑度。根据最新研究，当困惑度设为数据点数量的平方根，或对于大型数据集固定为 50 时，可获得最佳结果（van der Maaten2015）。

高维数据集，例如我们的光谱数据（完整光谱有 1036 个维度），在分析时往往会面临挑战（van der Maaten 和 Hinton， 2008；Valentina Tardugno Poleo，准备中）。在本次分析中，我们只使用发射线周围一个较小的波长范围（$\pm$ 50 $^{\circ}$A）。我们设置这一阈值，是为了聚焦于发射线本身，并使其在 t-SNE 可视化中保持居中。我们尝试过使用完整光谱，但在这种情况下，t-SNE 往往会依据红移进行区分。由于我们的主要关注点是真实发射还是假阳性，目前我们会移除红移信息。未来的工作将继续探索如何使用完整光谱。例如，应用主成分分析（PCA）或自编码器，或许有助于在可视化中移除红移信息，并进一步区分不同的源。因此，将数据集的维度从完整光谱的 1036 维缩减到发射线周围截取区域的 100 维，可以避免这些复杂问题，并简化数据可视化。进行这种降维后，得到的

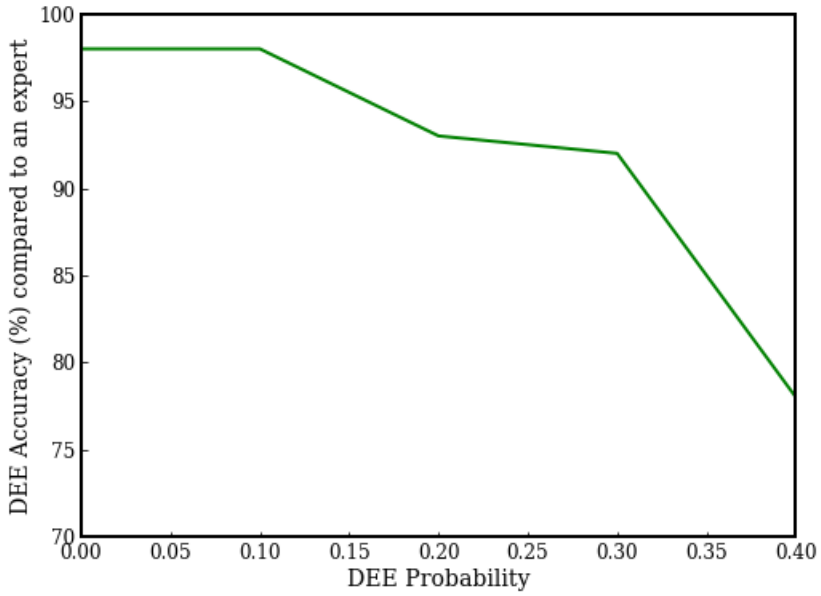


图 5. 上图显示了在“在噪声海洋中寻找信号”工作流程中， *Dark Energy Explorers* 参与者相较于 HETDEX 专家的准确率百分比。这一准确率体现了 *Dark Energy Explorers* 参与者与 HETDEX 团队成员相比的表现。DEE 概率是每个源的 10 个分类结果的平均值，表示某个源是假阳性（0）还是真实探测（1）的概率。本文中的此图及分析着重于识别假阳性或 DEE 概率 < 0.4。

结果具有更好的区分能力，本文将讨论这些结果。

在图 6 中，我们展示了将暗能量探索者 参与者与 t-SNE 机器学习技术相结合的结果，这种方法能够最大限度地去除伪源，同时将真实探测结果的去除比例降至最低。图 6 展示了 t-SNE 机器学习算法的可视化表示。彩色点是根据暗能量探-索者 概率（$\sim$ 14,000）生成的标签，黑色点则表示源目录（ > 5.5 ,000 个源）中的源总数。白色叠加线条显示了 t-SNE空间中的源密度轮廓。这些结果表明，结合公民科学目视审查与 t-SNE 可以得出更一致的结果。

筛查所有 暗能量探索者 概率低于 0.1 的源后，我们发现结果的一致率为 98%，如图 5所示。HETDEX 团队成员会审核所有这些源，然后将审核结果与 暗能量探索者 给出的概率进行比较，以确定准确率。

我们在大量探测器和分辨单元中看到的问题种类繁多，因此很难量化所有类别的假阳性，这就是我们首先借助 HETDEX 团队成员的原因。最终，t-SNE 可视化将与监督学习算法相结合，以便对区域进行分组，形成经过充分训练的分类方案。本文的目标是确定假阳性，并将其从样本中剔除，同时尽可能少地移除真实探测。我们的下一个目标是利用这些经过审核的假阳性来改进分类结果和筛选配置。

我们发现，暗能量探索者和 t-SNE 的最高准确率均出现在 t-SNE 的区域x 取值 > 3 和 t-SNE y 取值 < -1 ，使用图6中的可视化结果。所选的 t-SNE 区域移除真实探测结果的比例也最低，该区域内没有任何源的 DEE 概率达到 > 0.9。换言之，结合使用这些方法，不会移除任何经目视审查判定为真实的源。此外，该区域内仅有 0.08% 的源的 DEE 概率达到 > 0.6；根据暗能量探索者的判断，这些源属于真实探测结果。由于专业天文学家已目视确认这些源是真实的，我们将利用 DEE 概率确保它们保留在目录中，这意味着不会移除任何真实探测结

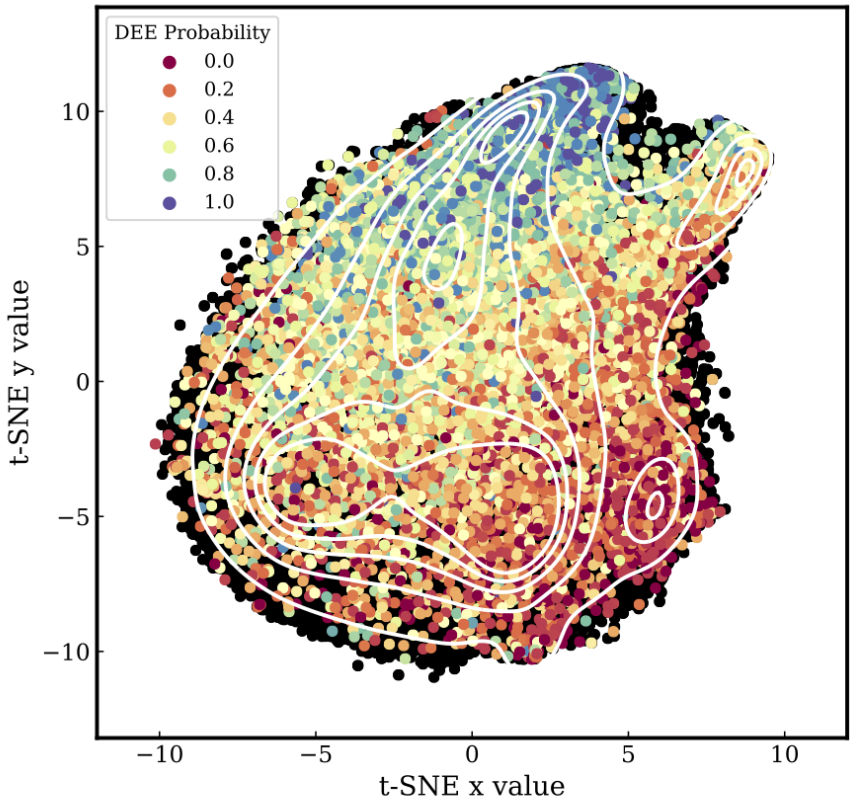


图 6. 上图展示了 t-SNE 机器学习算法的可视化表示。彩色点是根据 *Dark Energy Explorers* 概率（∼ 14,000）创建的标签，叠加在黑色点上；黑色点表示源目录中信噪比 > 5.5 的源总数（∼ 120,000 个源）。白色轮廓线显示了 t-SNE空间中的源密度。

果。这项技术展示了如何结合机器学习与公民科学，在将真实探测结果保留在目录中的同时，有针对性地移除虚假探测结果。

来自 暗能量探索者 的这些结果使我们能够使用 t-SNE，并扩展至包含 ∼ 120,000 个源的完整目录；这些源在图 6中以黑色点表示。通过应用选定的 t-SNE 空间区域，我们可以移除源，同时确保真实探测结果仍保留在样本中，因为我们已从 暗能量探索者获得了目视确认。这项工作从目录中移除了 7,871 个虚假探测。

我们可以进一步运用这种清理技术，查看当前目录中的其他源位于何处。这提供了一种高效的方法，无需经过人工全面审查即可从目录中移除源。总之，公民科学与机器学习协同运用，可以提升科学目标，同时为公众创造独特的教育机会。

6. 结论

自2022年年底以来，暗能量探索者 已收集了来自全球85个不同国家的11,000名志愿者所作的大约400万条分类结果。本次活动的另一项目标是由暗能量探索者 对所有信噪比降至 > 4.8 或以下的源进行目视审查，覆盖整个 HETDEX 星表；我们预计该星表将包含约三百万个源。对所有源以更低的信噪比重复这一流程，将结合算法、数据可视化和人工审核，提供可靠的分类结果。虽然这一数量远超我们目前已处理的规模，但我们将探索如何吸引更多参与者、减少每个目标所需的个人分类数量，以及采用有针对性的工作流程。

到目前为止，暗能量探索者 已减少了我们团队的目视审查工作，使这种细致程度的研究成为可能；若没有这项帮助，以我们团队有限的人手，研究原本无法开展。这个项目为研究暗能量以及对需要目视审查的大型数据集进行分类提供了一种创新方法。

通过使用 暗能量探索者 我们预计将宇宙学参数的约束改进∼ 10 至 30%。例如，这使我们能够将信噪比从> 5.2 提升至信噪比 > 4.8，从而使宇宙学参数 D_V 的精度提高 30%，这一动机见图 1。除了 暗能量探索者 是推动科学进步的独特而强大的工具之外，该项目还正在提升全球科学可及性。

7. 致谢

如果没有志愿者的悉心付出与奉献，暗能量探索者的研究结果就不会如此可靠且实用。我们对他们的工作深表感激。这项工作影响深远，也鼓舞人心。

LH 感谢美国国家科学基金会研究生研究奖学金计划（NSF GRFP）DGE 2137420 和美国国家航空航天局（NASA）21-CSSFP21-0009 的支持。KG 感谢美国国家科学基金会 NSF-2008793 和美国国家航空航天局（NASA）21-CSSFP21-0009 的支持。

暗能量探索者已获认可，成为美国国家航空航天局公民科学的官方合作伙伴。本出版物使用的数据由 Zooniverse.org 平台生成。该平台的开发得到了慷慨资助，包括谷歌颁发的全球影响力奖，以及阿尔弗雷德·P·斯隆基金会提供的资助。

HETDEX由德克萨斯大学奥斯汀分校麦克唐纳天文台和天文学系牵头，参与机构包括慕尼黑路德维希-马克西米利安大学、马克斯·普朗克地外物理研究所（MPE）、莱布尼茨天体物理研究所波茨坦分部（AIP）、得克萨斯农工大学、宾夕法尼亚州立大学、哥廷根天体物理研究所、牛津大学、马克斯·普朗克天体物理研究所（MPA）、东京大学和密苏里科技大学。除各机构的支持外，HETDEX还获得美国国家科学基金会（资助编号AST-0926815）、得克萨斯州、美国空军（AFRL FA9451-04-2-0355）的资助，以及私人和基金会的大力支持。

霍比-埃伯利望远镜（HET）是德克萨斯大学奥斯汀分校、宾夕法尼亚州立大学、慕尼黑路德维希-马克西米利安大学和哥廷根乔治-奥古斯特大学共同合作的项目。HET以其主要资助者 William P. Hobby 和 Robert E. Eberly 的名字命名。引力与宇宙研究所由宾夕法尼亚州立大学埃伯利理学院和高级研究副校长办公室提供支持。

作者感谢德克萨斯大学奥斯汀分校的德克萨斯高级计算中心（TACC）[3] 提供高性能计算、可视化和存储资源，为本文报告的研究成果作出了贡献。

参考文献

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, 日本天文学会出版物, 70, S4, doi: 10.1093/pasj/psx066Albrecht, A., Amendola,L., Bernstein, G., et al. 2009,doi: 10.48550/ARXIV.0901.0721Alcock, C., & Paczy´nski, B.1979, 281, 358,doi: 10.1038/281358a0Bahaadini, S., Noroozi, V., Rohani, N., et al. 2018, 信息科学, 444, 172, doi: https://doi.org/10.1016/j.ins.2018.02.068Bautista, J.E., Paviot, R., Vargas Maga˜na, M., et al.2020, 500,736, doi:10.1093/mnras/staa2800Colless, M., Peterson, B. A., Jackson, C., et al.2003, arXiv 预印本, astro, doi: 10.48550/arXiv.astro-ph/0306581Davis, D., Gebhardt, K., Mentuch Cooper, E., et al. 2021, 《天体物理学杂志》, 920, 122, doi:10.3847/1538-4357/ac1598

—. 2023年，arXiv 预印本，arXiv:2301.01799。https://arxiv.org/abs/2301.01799Dawson, K. S., Schlegel, D. J., Ahn, C. P., 等。2013年，《天文学杂志》，145，10，doi：10.1088/0004-6256/145/1/10de Mattia, A., Ruhlmann-Kleider, V., Raichoor, A., 等。2021年，《皇家天文学会月刊》，501，5616，doi：10.1093/mnras/staa3891DES 合作组，Abbott, T. M. C., & Aguena, M. 2021年，arXiv:2101.05765 DESI 合作组。2016年，arXiv:1611.00036 Drake, A. J., Graham, M. J., Djorgovski, S. G., 等。2014年，《天体物理学杂志增刊系列》，213，9，doi：10.1088/0067-0049/213/1/9du Mas des Bourboux, H., Rich, J., Font-Ribera, A., 等。2020年，《天体物理学杂志》，901，153，doi：10.3847/1538-4357/abb085Eisner, N.L., Barrag´an, O., Lintott, C., 等。2021年，《皇家天文学会月刊》，501，4669，doi：10.1093/mnras/staa3739

[3] http://www.tacc.utexas.edu

Farrow,D., S´anchez, A., Ciardullo, R., 等。2021a年，《皇家天文学会月刊》，507，doi：10.1093/mnras/stab1986Farrow, D. J., S´a nchez, A. G., Ciardullo, R., 等。2021b年，《皇家天文学会月刊》，507，3187，doi：10.1093/mnras/stab1986Gebhardt, K., Mentuch Cooper, E., Ciardullo, R., 等。2021年，《天体物理学杂志》，923，217，doi：10.3847/1538-4357/ac2e03Gil-Mar´ın, H., Bautista, J. E., Paviot, R., 等。2020年，498，2492，doi：10.1093/mnras/staa2455Land, K., Slosar, A., Lintott, C., 等。2008年，388，1686，doi：10.1111/j.1365-2966.2008.13490.xLaureijs, R., Amiaux, J., Arduini, S., 等。2011年，Euclid定义研究报告 Leung, A. S., Acquaviva, V., Gawiser, E., 等。2017年，《天体物理学杂志》，843，130，doi：10.3847/1538-4357/aa71afLintott, C. J., Schawinski, K., Slosar, A., 等。2008年，389，1179，doi：10.1111/j.1365-2966.2008.13689.xMentuch Cooper, E., Gebhardt, K., Davis, D., 等。2023年，arXiv 预印本，arXiv:2301.01826。https://arxiv.org/abs/2301.01826 Pedregosa, F., Varoquaux, G., Gramfort, A., 等. 2012年，计算机科学研究报告，abs/1201.0490 Perlmutter, S. 1999年，《天文学杂志》，517 普朗克合作组，Aghanim, N., Akrami, Y., 等. 2020年，《天文学与天体物理学》，641，A6，doi：10.1051/0004-6361/201833910Raddick, M. J., Prather, E. E., & Wallace, C. S. 2019年，28，636，doi：10.1177/0963662519840222Riess, A. G., Casertan, S., Yuan, W., & Bowers, J. B. 2021年，《天文学杂志通讯》，908 Riess, A. G., Filippenko, V., A., Challis, P., & Clocchiatti, A. 1998年，《天文学杂志》，116，1009 Santos-Fernandez, E., & Mengersen, K. 2021年，《生态学与进化方法》，12，1533，doi：https://doi.org/10.1111/2041-210X.13623S´anchez, A. G., Montesano, F., Kazin, E. A., 等. 2014年，440，2692，doi：10.1093/mnras/stu342Tegmark, M., Strauss, M. A., Blanton, M. R., 等. 2004年，《物理评论D》，69，103501，doi：10.1103/PhysRevD.69.103501van der Maaten, L. 2015年，《机器学习研究杂志》，15，3221 van der Maaten, L., & Hinton, G. 2008年，《机器学习研究杂志》，9，2579 Weinberg, S. 1989年，《现代物理评论》，61，1，doi：10.1103/RevModPhys.61.1