从决策轨迹推断显微镜实验目标

Kamyar Barakati[1,*]、Boris N. Slautin[1], Aditya Raghavan[1], Babita Babita[2], Surya Prakash Reddy[1], Jordan Marshall[1], Munireach Nannory[3], Stephen Puplampu[3], Alla Slautina[1], Faith M. Hines[1], Newsha Javanmardi[1], 和 Sergei V. Kalinin[1,2,†]

[1] 田纳西大学材料科学与工程系，美国田纳西州诺克斯维尔[2] 田纳西大学电气工程与计算机科学系，美国田纳西州诺克斯维尔[3] 田纳西大学土木与环境工程系，美国田纳西州诺克斯维尔，邮编 37996

摘要

现代显微镜采集、分析数据并据此采取行动的速度，远远快于人类操作员决定下一步测量内容的速度。要在自动化实验中利用这一速度，仪器就必须遵循操作者的科学逻辑；将这种逻辑编码为奖励函数，并使其成为操作流程的一部分，是一种实现方式。在这里，我们探讨能否从显微镜操作员的决策历史中学习此类奖励函数。我们提出了多个框架，将在不同测量位置之间记录的选择作为比较证据，以反映操作者的偏好；并比较了几种重建奖励的算法，包括基于具有物理意义的图像描述符构建的可解释线性模型、基于相同描述符的非线性函数、先进行无监督图像表示再开展偏好建模的方法，以及偏好训练的深度核启发模型。分析得出的结果是在样品上定义的奖励函数空间场，可用于操控仪器。我们还提出一种评估这些方法的基准测试方式，依据保留决策以及基于已知目标生成的决策历史进行评估。随后，我们考察决策历史是否包含实验过程中不断变化的优先事项。隐马尔可夫模型为反复出现的选择策略分配不同的奖励，而连续状态模型则追踪平滑变化的偏好权重。在真实显微成像候选集上生成的受控序列表明，这两种模型都能恢复隐藏切换和渐变漂移。这些直接从记录的决策中学习得到的奖励函数，为未来的自动化显微镜实验提供了候选目标。

* kbarakat@vols.utk.edu

† sergei2@utk.edu

1. 引言

现代科学仪器采集、分析数据并据此采取行动的速度，远远快于人类操作员做出决策的速度。扫描探针显微镜和电子显微镜可以重新定位、采集图像并进行光谱学测量，不同测量技术所需的采集时间各不相同；而人类操作员可能需要几分钟才能决定下一步在哪里测量以及为何在那里测量。缩小这一差距是实现自动化实验和自主实验的关键要求；近年来，自驾实验室 [1–6]，对扫描探针显微镜的机器学习控制 [7–10] ，以及最近用于规划和执行多步骤实验的智能体工作流 [11–17]，都推动了相关领域的快速发展。

然而，自动化实验必须明确其寻找目标。无论决策是通过贝叶斯优化、强化学习，还是（隐式地）由语言模型智能体做出，最终都由奖励函数驱动，该函数为可能的测量结果 [18–21] 赋予价值。对于仪器优化，可以使用针对具体目标的测量质量指标来定义奖励。贝叶斯优化已被用于扫描探针显微术，以调节成像参数 [22,23] ，并被用于电子显微术，以优化像差校正器设置 [24–27]。

在我们此前的工作中，奖励也基于物理学和领域知识进行构建，倾向于采用具有物理可解释性的分析、发现结构–性能关系，以及结合多个图像特征的多目标优化 [28–36]。当科学目标能够预先明确写出时，这类奖励十分有效。然而，经验丰富的操作者所掌握的大量专业知识是隐性的：实验人员通常会根据观察到的对比度，在看似相似的区域之间进行区分，却不会制定明确的判据。

在这里，我们展示如何从人类决策本身构建奖励函数。操作者在显微镜上的每个实验阶段都会留下决策历史：在每一步中，操作者都从可用的备选项中选择一项测量。在比较判断和离散选择理论中，这类选择体现了有关潜在效用 [37–40], 的信息；逆强化学习和偏好学习则提供了从行为 [41–51]中推断该效用的方法。如果将备选项与所做选择一并记录，操作者决策所隐含的奖励就成为一个可通过数据求解的逆问题，而由此得到的函数可以传输至仪器，并以机器速度执行。

我们首先建立从记录的备选项中学习奖励的问题，并定义如何评估重建奖励（第 2 节）。第 3 节重建平稳奖励，依次从基于可解释图像描述符的线性模型，拓展到非线性函数和

学习到的图像表示。第 4 节考察操作者在实验过程中优先事项发生变化的决策历史，并通过离散潜在策略和连续演变的偏好对其进行建模。第 5 节讨论重建的奖励揭示了什么、它们的局限性，以及如何利用它们驱动自动化实验。

2. 从记录的备选项中学习奖励

作为示例，我们考察位于 $PbTiO3$ 衬底上的$KTaO3$ 薄膜中的铁电畴结构。这种材料的铁弹性 a/c 畴图案、畴壁和极化翻转已通过压电响应力显微镜（PFM） [54–60]进行了广泛研究。我们分析一幅存档的256 × 256像素 PFM 图像，其中包含畴图案、边界和非均匀纹理。所有记录的决策历史和受控模拟均基于该图像及其相关的候选图像块。在典型的压电响应力显微工作流程中，操作者首先获取样品图像，然后确定需要进一步检查的位置，例如进行高分辨率成像或光谱学分析。

在这里，我们关注对这类现有图像及其记录的选择历史进行回顾性解释。该记录可能代表仪器操作期间做出的选择，也可能代表随后对存档内容进行的评估；只要保留所提供的备选项和选定项，逆问题就是相同的。重要的是，在记录的序列过程中，潜在偏好不必始终保持不变。随着科学问题的发展，科学家可能最初优先关注清晰的畴壁，随后寻找具有有序条纹的区域，之后又重新考察另一种形貌。我们要探究的是：一个奖励是否能够描述完整记录，或者能否从同一系列观测到的选择中识别出不断变化的偏好。

为了研究这一问题，我们考虑这样一种情形：在每一步中，操作者从图像的 20 个候选图像片段中选择一个。我们采用这种形式，因为它符合人类在实践中的能力：立即比较少量备选项既快速又可靠，而对大型图像中的数千个重叠图像块进行排名或评分则不现实，而且容易受到视觉显著特征的影响 [52]。同时，对于固定的奖励函数，20 个随机提出的候选项中最佳的一个接近奖励分布的顶部：平均而言，它优于图像中约 95% 的所有图像块，从而使反复选择能够集中于高价值区域（补充材料第 S2 节）。将备选项与选择结果一并记录至关重要，因为胜出项及其 19 个竞争项能告诉我们哪些特征优于哪些特征，而仅凭选定坐标列表无法提供这些信息。由此得到的奖励是一个

该协议中对操作者偏好的操作性描述，而非物理定律或心理模型 [42,53]。

利用这幅图像和选择协议，我们生成了六条决策历史，每条都包含从由 20 个候选图像块组成的集合中连续作出的 100 次选择：其中两条来自人类参与者（P01 和 P02），两条来自人为定义的启发式方法（HDH 1 和 HDH 2），另外两条来自基于 GPT 的语言模型评估器（GPT 1 和 GPT 2）。对于第 4 节中的受控测试，我们还针对相同的候选集，生成了来自具有已知依赖图像的奖励的模拟决策者的新获胜项标签。

设决策 t 时记录的备选项为 $C\mathrm{t} = \{x\mathrm{t}1, \ldots, x\mathrm{t}20\}$，并设 $a\mathrm{t}$ 表示所选图像块。我们拟合一个函数 $R(x)$，为每个候选项分配一个相对效用。条件式*SoftMax*选择模型将这些分数映射为选择概率 [38–40,61,62]；逆温度因子 β 描述了概率集中于高分备选项的程度。似然是在完整记录的候选集中，对观测到的获胜项进行评估的。由于观测结果约束的是相对效用而非绝对效用，因此奖励只能在假定的选择模型所允许的变换下确定。高度相关的图像描述符和未观测情境会带来更多歧义，详见下文和补充材料第 S2 节。

$$P(a_t = i \mid C_t) = \frac{exp[\beta R(\boldsymbol{x}_{t,i})]}{\sum_{j=1}^{20} exp[\beta R(\boldsymbol{x}_{t,j})]} \tag{1}$$

通过评估在拟合时留出的决策，可以判断重建奖励对操作者行为的描述效果。最简单的问题是：模型是否将实际胜者排在首位（Top-1 准确率），或排在前三个选项之中（前 3 准确率）；随机选择的模型得分分别为 5% 和 15%。更精细的指标是模型赋予实际胜者的概率，以负对数似然（NLL）概括；随机选择时该值为 $\ln 20 \approx 3.0$，而自信地选错的模型会比谨慎的模型受到更重的惩罚。除了这些数值，我们还可以探究模型学到了什么：它会奖励哪些图像特征、拟合系数有多稳定、选中和拒绝的图像块在学习到的表征中如何分布，以及整幅图像中哪些位置的奖励较高。对于偏好随时间变化的模型，我们还会考察操作者使用了多少种不同的策略、策略切换发生在何时，以及模型在根据每次选择更新之前能否预测该选择。

对于 T 保留决策，两个主要指标是

$$\mathrm{NLL} = -\frac{1}{T}\sum_{t=1}^{T} \ln P\left(a_t \mid C_t\right) \tag{2}$$

And

$$\mathrm{Acc}_k = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left[r_t(a_t) \le k\right] \quad (3)$$

其中 $r\mathrm{t}(a\mathrm{t})$ 是记录的胜出选项在 20 个预测分数中的排名，$\mathbb{1}[\cdot]$ 是指示函数。

逆问题所使用的信息汇总于**图1**。存档图像提供候选集，记录的胜出选项及其备选项共同约束奖励，而冻结后的奖励可投射回整幅图像（**图1a**）。单条决策历史中的选择集中在少数几个领域区域（**图1b**）。每次决策都会提供20个图像块，其中一个会被选中（**图1c**）。16种图像描述符中，有几种呈现强相关性，尤其是强度统计量与梯度-拉普拉斯组（**图1d**）。针对同一幅图像，不同的决策历史会产生明显不同的空间模式，这促使我们为每条历史学习一个奖励，而不是假设存在一种通用的图像判据（**图1e-j**）。

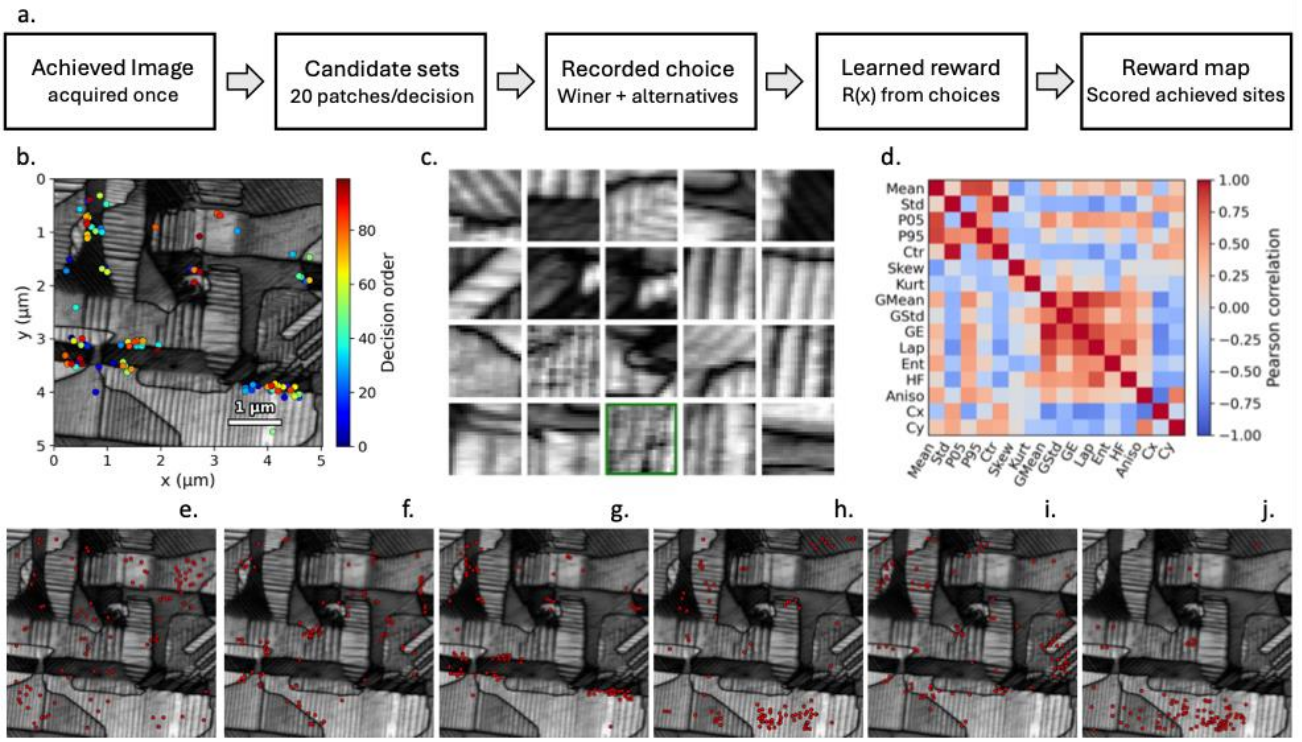


图1。基于显微镜决策轨迹的回顾性偏好重建。（a）工作流程：存档图像提供有限的候选集，记录的选择及其备选项共同约束推断出的奖励，并将冻结的奖励投影到同一图像上。（b）记录的选择的空间分布，按决策顺序进行颜色编码。（c）一组记录的20个候选图像块；获选者以轮廓线标出。（d）实验决策轨迹中16个图像描述符之间的皮尔逊相关性。（e–j）100次记录的选择的空间分布，分别对应（e）GPT 1、（f）GPT 2、（g）HDH 1、（h）HDH 2、（i）P01和（j）P02。红色标记表示原始PFM图像上的选定位置。

重建的奖励可用于三个层面。首先，它应能预测操作者做出的、但未用于拟合的选择。其次，冻结后的奖励可为图像中的所有其他图像块或新图像中的图像块打分，从而生成可供检查的奖励图。第三，它可用于选择仪器上的新测量。在本文中，我们对第一个层面进行定量测试，展示第二个层面的示例，并为第三个层面奠定基础。

我们比较了几种重构奖励的方式，而不是假定更复杂的模型必然更好。每种方法由两个选择决定：如何表示一个图像块，即使用物理可解释描述符，还是使用从图像中学习到的特征；以及使用哪种函数将这种表征转换为奖励，从加权和到非线性函数和核函数。随后，我们考察奖励本身是否会在一段决策历史中发生变化：用有限状态集描述反复出现的策略，并用持续更新的权重描述渐进变化。

3. 在静态偏好假设下重构奖励

本节从每段决策历史中重构一个单一且与时间无关的奖励。我们首先研究基于物理可解释图像描述符的线性奖励（**第 3.1 节**），这种方法能够用熟悉的图像特征直接解读推断出的偏好。随后，我们保留相同的描述符，但允许使用非线性加性函数和核函数（**第 3.2 节**）；最后，我们以学习到的图像表征取代手工设计的描述符，先使用无监督变分自编码器学习，再使用根据选择本身训练的卷积模型学习（**第 3.3 节**）。每个模型都使用公式（2）和（3）中的指标在保留决策上进行评估，并以存档图像上的空间奖励场形式进行可视化。

3.1. I 可解释的图像描述符：图像块的哪些特征可以解释选择 n?

一个直观的起点是通过熟悉的图像特征来表达图像块的效用。局部对比度和强度变化、梯度和拉普拉斯响应、边界含量、各向异性以及空间频率统计量，都能为铁性畴形貌提供可测量的描述。最初提出的物理引导公式引入了六个描述符，包括对比度、边界强度、取向、周期性、特征尺度和局域性，并使用排序间隔优化参数化评分 [63]。在本分析中，我们使用 16 个图像统计量，并测试一个精简的六项集合。该精简模型并不等同于先前基于物理引导的六特征奖励：其描述符、拟合目标和数据划分均不相同。因此，我们比较两者的表现，而不将它们视为同一个模型。

每个候选图像块都表示为一组图像特征：一个定长描述符向量，根据像素强度计算得出，不考虑该图像块在图像中的位置 [64,65]。这些描述符使用训练候选项的均值 μm 和标准差 s m 进行标准化，

$$\phi(x) = [\phi_1(x), \dots, \phi_{16}(x)]^{\mathrm{T}}, \qquad \tilde{\phi}_m(x) = \frac{\phi_m(x) - \mu_m}{s_m} \tag{4}$$

对于像素 $p \in x$ 处强度为 $I(p)$ 的图像块，具有代表性的描述符包括平均梯度幅值和平均绝对拉普拉斯，

$$\phi_{\mathrm{grad}} = \frac{1}{|x|}\sum_{p\in x}\|\nabla I(p)\|, \qquad \phi_{\mathrm{lap}} = \frac{1}{|x|}\sum_{p\in x}|\nabla^2 I(p)| \tag{5}$$

直方图熵（其中 hβ 表示强度区间 b 中像素所占的比例），以及高于截止空间频率 k_c, 的傅里叶功率所占比例，

$$\phi_H = -\sum_b h_b \ln h_b, \qquad \phi_{\mathrm{HF}} = \frac{\sum_{|k|>k_c}|\hat{I}(k)|^2}{\sum_k |\hat{I}(k)|^2} \tag{6}$$

以及结构张量 J 的各向异性，其特征值为 $\lambda 1 \geq \lambda 2$ [66],

$$\phi_{\mathrm{aniso}} = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}, \qquad J = \sum_{p\in x} \nabla I(p) \nabla I(p)^{\mathrm{T}} \tag{7}$$

对于各向同性纹理，该值为零；对于取向单一的条纹畴，该值则趋近于一。其余描述符包括强度统计量、梯度统计量，以及水平和垂直方向的最近邻相关性。其定义和实现细节总结于补充材料第 S3 节。

对于标准化后的描述符，线性奖励是公式（8）中给出的加权和。这种线性方法可以直接解释 [67–69]：每个系数都表示某一图像特征属性在多大程度上吸引或排斥决策者。正系数表示，该描述符取值较大的图像块更有可能被选中；负系数则表示相反的情况。由于若干描述符衡量的是相关属性，例如梯度和拉普拉斯响应，它们之间的权重可以有不同分配方式，而对选择结果的预测效果同样好。因此，解读单个系数时，应结合选定图像块本身，以及它们胜过的候选图像块一并考量。

$$R(x) = \boldsymbol{w}^{\mathrm{T}} \tilde{\boldsymbol{\varphi}}(x) = \sum_{m=1}^{16} w_m \tilde{\varphi}_m(x) \tag{8}$$

通过最小化正则化选择似然，根据训练决策估计系数，

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left[-\sum_{t\in train} \ln P(a_t | C_t; \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \right] \tag{9}$$

其中，P 由公式（1）给出，λ 设定$L2$ 惩罚项的强度。系数为正表示，在控制其他描述符后，较大的数值与选定图像块相伴出现。

图2展示了描述符模型如何从图像测量值推导出预测选择。对于一个代表性图像块，标准化描述符值（**图2a**）与拟合系数相结合，得到各项对单一效用函数的贡献（**图2b**）；此处强度离散度占主导地位。16 描述符模型的系数矩阵在不同轨迹之间差异显著（**图2c**）：启发式算法驱动的历史显示，所选图像描述符的贡献相对较强；而 GPT 生成的历史则呈现更为分散的系数。留出决策的 Top-1 和 Top-3 准确率分别介于随机水平至 90% 和 100% 之间（**图2d, e**）；对于重建效果良好的轨迹，几乎每个留出的胜出选项在其 20 个备选项中排名第一或第二（**图2f**）。两个参数向量可能相差很大，却仍能对观测到的备选项给出相似排名，因此应综合解读这些诊断结果，而不能只看系数图。

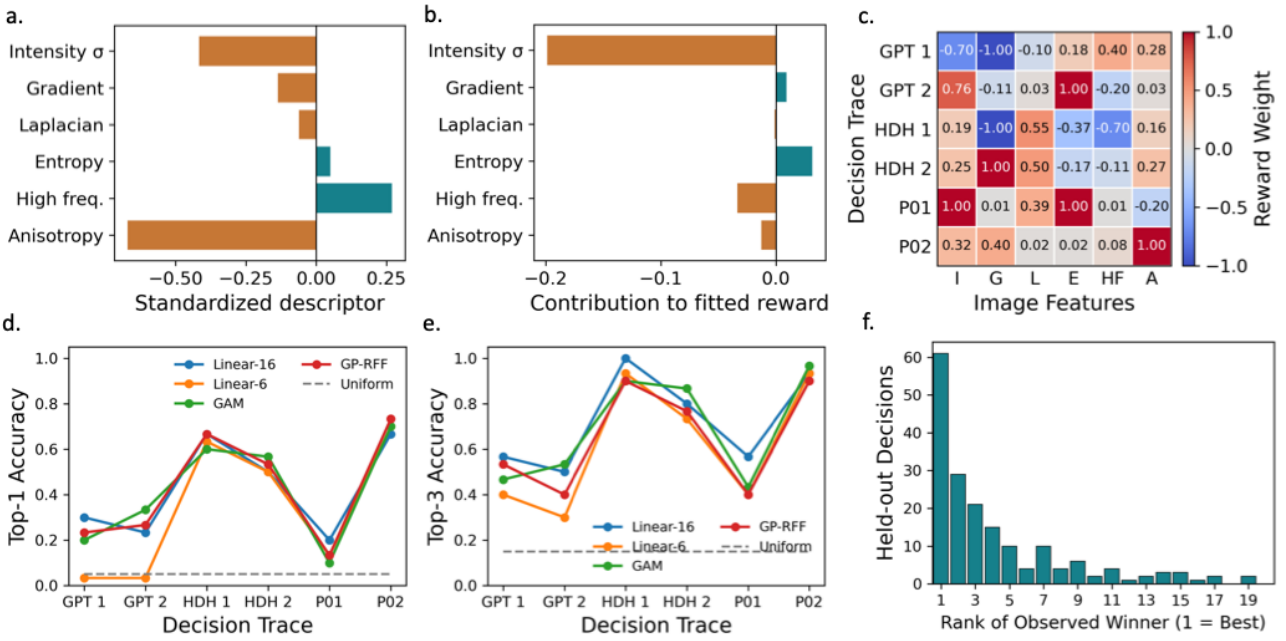


图2。基于描述符的奖励的解释与测试。(a) 一个代表性图像块的六种图像描述符的标准化值。(b) 它们对线性效用的贡献（描述符 × 拟合系数）。(c) 六条记录的决策历史（GPT 1、GPT 2、HDH 1、HDH 2、P01 和 P02）对应的 16 描述符线性模型的标准化系数。(d, e) 各轨迹和模型的留出 Top-1 (d) 和 Top-3 (e) 准确率；虚线表示均匀选择（5% 和 15%）。(f) 某条代表性轨迹的留出决策中，记录的胜出选项在其 20 个备选项中的排名。

最终的可恢复性高度依赖轨迹。根据图像描述符重构启发式选择规则的准确度最高，这与生成这些规则时依赖可见图像块统计特征相一致。对于其他选择历史，尤其是某些 GPT 生成的选择，相同特征仍难以重构。低分可能有多种解释：

表征遗漏了相关的图像特征属性；选择具有概率性或受到未记录的上下文影响；底层评估者改变了目标；或者有限的训练集未能充分约束模型。数据本身无法自动区分这些可能性。同样，图像内部的高准确率并不能证明人类科学目标已被唯一恢复。它表明，在特定的候选生成协议下，模型能够预测与某一特定记录选择过程相一致的结果。

所选指标会影响我们对变化的解读。对于许多决策，模型可能将实际胜出者排在第二位，却仍捕捉到了有意义的偏好；而另一个模型可能将大多数胜出者排在首位，却在其余预测失败时，给错误图像块赋予极不合理的高概率。因此，所选项与最佳未选项之间的边际分布，以及每个胜出者的条件排名，都能补充 Top-1 准确率。我们还会检查模型收敛情况、不同轨迹间拟合系数的变化，以及结果对留出决策的敏感程度。这些检查对于更灵活的函数尤为重要，因为表面上平滑的奖励场，也可能来自无法准确预测已记录考量集合的模型。

3.2. 同一组图像描述符上的备选函数

没有物理上的理由认为实验偏好必然是图像统计量的线性函数。与对比度极低或饱和相比，中等对比度可能更受偏好；特定的畴周期性可能仅在狭窄范围内具有信息价值；而梯度的含义，则取决于它属于有序边界还是噪声。因此，我们在保持候选集和逆向选择目标不变的同时，比较定义在同一组十六个图像描述符上的若干函数。这有助于区分建模灵活性的两个来源：图像的描述方式，以及偏好对该描述的数学依赖关系。这样也能避免将学习得到的图像编码器表面上的优势自动归因于非线性奖励头。

广义加性模型（GAM）用平滑单变量函数替代单个线性项，能够表示非单调依赖关系，同时保留特征层面的可解释性。对核奖励进行有限随机傅里叶特征近似，可以表示描述符之间的交互，但这并不是精确的高斯过程偏好后验分布。迭代潜在效用方法通过交替估计与各个候选决策相关的效用，并更新系统性奖励预测器；它是一种正则化后验众数构造，而非精确的期望最大化。这些是关于有效偏好形式的备选假设，并非预期会按单调递增顺序改进的层级结构。它们都不会

仅仅因为某个图像块被赋予了较高效用，就预测其未测量的偏振或光谱响应；若没有共同的识别约定，也无法比较不同拟合结果的数值尺度。按照公式（8）的记号，加性和核备选方案将线性评分替换为

$$R_{\mathrm{GAM}}(x) = \sum_m f_m\left(\tilde{\phi}_m(x)\right), \qquad f_m(u) = \sum_b c_{mb}\, B_b(u) \tag{10}$$

$$R_{\mathrm{RFF}}(x) = \beta^{\mathrm{T}}\mathrm{z}\left(\tilde{\phi}(x)\right), \qquad \mathrm{z}(u) = \sqrt{\frac{2}{D}}\cos(\Omega u + \mathrm{b}) \tag{11}$$

其中，B_b 是样条基函数，其系数为 c_{mb}；D 的各行 Ω 从逆长度尺度为 l^{-1} 的高斯分布中抽取，且相位 b 在 $[0, 2\pi)$, 上服从均匀分布，因此 $z(u)^{\mathrm{T}} z(u')$ 可近似径向基函数核 [70–72]。两个模型均以公式（9）中的目标函数进行拟合。

六轨迹基准测试对每个模型分别进行拟合，并在原始的种子化 80/20 决策级留出集上进行评估。**图 3**将 GPT 1 的记录选择与全部五个模型的奖励场并列展示（**图 3a–f**），并报告每条轨迹及每个模型在留出集上的 Top-1 准确率（**图 3g**），以及该准确率与选择似然的关系（**图 3h**）。整幅图像的映射是通过在已采集的母图像中的重叠图像块上评估冻结的奖励而得出的；它们并不代表额外的测量。在每次决策可用的二十个备选项中进行预测，是检验回顾性学习的主要方式。相比之下，空间场所揭示的是：如果提供这些图像块，拟合后的模型会认为哪些结构有价值；即使模型的留出集准确率相近，其空间场也可能存在显著差异。

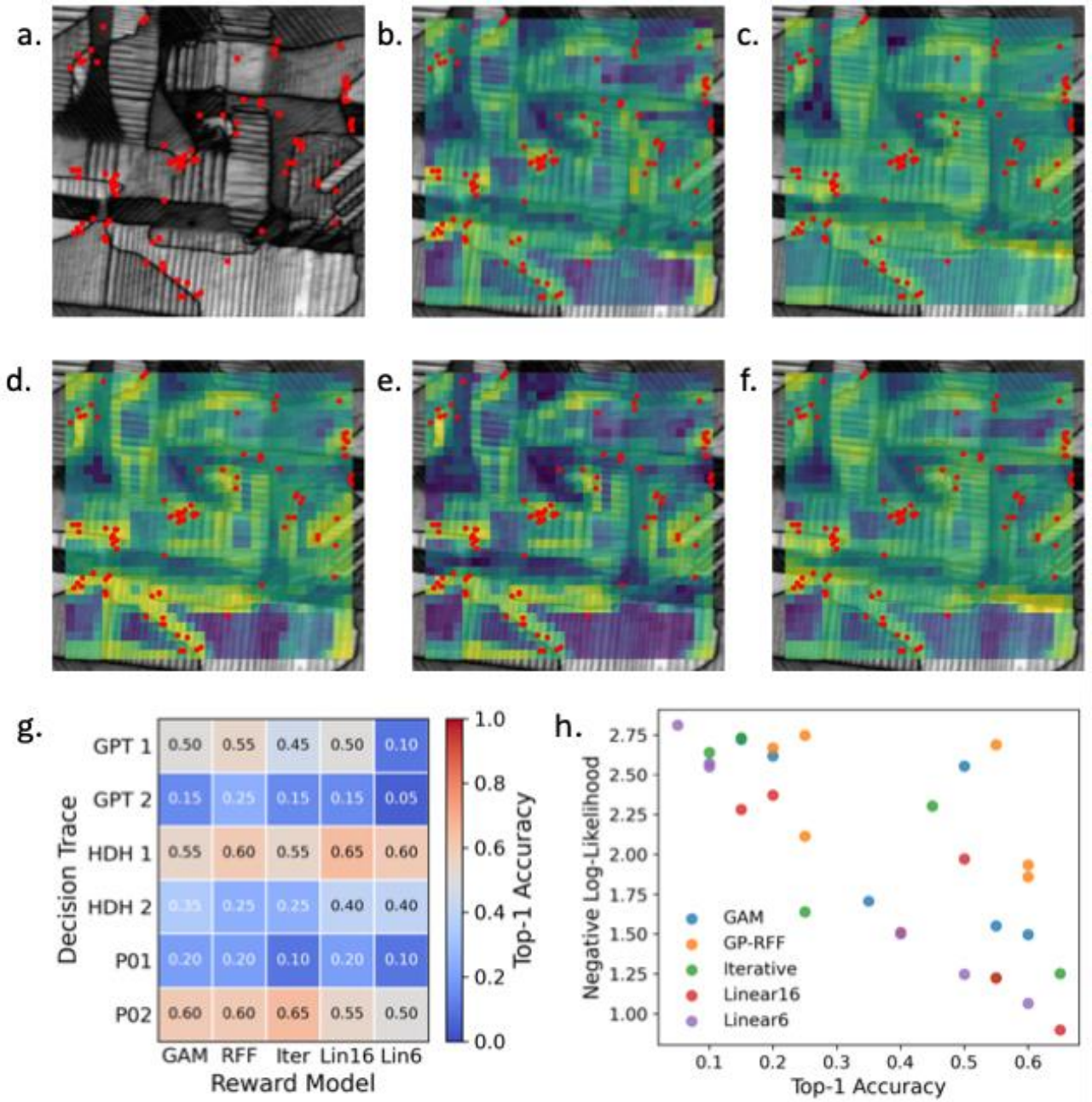


图 3。 平稳奖励模型的空间重建与预测性能。 (a) GPT 1 在原始PFM图像上的记录选择位置（红色标记）。(b–f) 使用同一张存档图像重建的空间奖励图：(b) 16 描述符线性模型（Linear16），(c) 六描述符线性模型（Linear6），(d) 广义加性模型（GAM），(e) 随机傅里叶特征核模型（GP-RFF），以及 (f) 迭代奖励模型。每张奖励图均采用独立归一化的色标显示，较暖的颜色表示推断出的奖励较高。红色标记表示记录的选择，奖励场在有效图块采样区域内进行评估。(g) 五种奖励模型在六条决策轨迹上的留出集 Top-1 准确率：GPT 1、GPT 2、HDH 1、HDH 2、P01 和 P02。(h) 所有 30 个轨迹–模型组合的留出集 Top-1 准确率与负对数似然之间的关系，展示了排序准确率与概率预测性能之间的差异。

在六条记录的决策历史中，十六描述符线性奖励模型的留出集 Top-1 准确率平均为 40.8%，与 GP-RFF（40.8%）和 GAM（39.2%）相当，

迭代模型和六描述符线性模型的准确率分别为 35.8% 和 29.2%（图 3g；补充表 S3）。不同轨迹之间的表现差异显著，反映了记录到的选择策略各不相同。这些平均值不应被视为对模型架构的最终排名：每条轨迹仅包含 20 个留出决策，因此只要有一个预测发生变化，准确率就会相差 5 个百分点；而且，所有轨迹中的候选图像块都来自同一张存档图像。尽管如此，结果表明，相对简单且具有物理可解释性的图像描述符能够还原相当一部分已观察到的选择偏好，而更灵活的非线性奖励函数并不能始终提升预测性能。

Top-1 准确率和似然可能会偏向不同的模型（**图 3h**）。Top-1 只关注得分最高的候选项是否为实际胜出者，而似然还会对模型为未排在首位的胜出者分配合理概率给予肯定，并对模型过度自信地作出错误判断予以惩罚。训练曲线、已选与已拒绝图像块的得分分布，以及实际胜出者的排名，有助于判断较差的结果究竟源于图像信息缺失、校准不佳还是过拟合。即使似然不同，奖励图往往也会突出相似的畴特征，因此奖励图最适合用于解释模型，而留出选择则用于检验模型。该分析生成一幅稠密空间奖励图：图像中的每个图像块都会获得一个得分，并且可以直接根据决策者偏好的特征来解读这幅图，例如清晰边界、周期性条纹或对比度变化。同一幅图还可作为有效的奖励函数，用于选择新的采样位置。不过，必须认识到它的局限性。相邻图像块高度重叠，因此一个连续的高奖励区域并不代表一组相互独立且富有信息的位置；而且，在作出某次选择时，操作者从未看到大多数高得分位置。因此，这幅图可以补充对留出选择的检验，但不能取而代之。固定的奖励还可以对存档图像中的所有其他图像块重新排名，从而展示自动化系统将如何为它们排列优先级。这是对所学目标的投影，而非对实际执行过的实验的记录；如何将此类奖励转化为采样策略，将在第 5 节和补充第 S6 节中讨论。

3.3. 学习图像表征，而不是指定所有描述符

手工设计的描述符具有透明性，但显微镜图像包含连接处、大小不同的畴以及长程纹理，简短的统计量列表可能无法捕捉这些特征。例如，即使不同区域的图像统计量几乎等价，人类操作员也可能更偏好具有弯曲畴壁而非笔直畴壁的区域。另一种方法是让神经网络学习图像表示。这引出了一个简单的问题：通过重构图像学习到的表征

是否足以预测选择，还是必须根据这些选择本身训练网络？下面我们将测试这两种方案：第一种使用变分自编码器，第二种使用经过偏好训练的卷积模型。

在变分自编码器（VAE）[73,104–106], 中，编码器通过图像重建目标和潜在空间正则化进行学习。所选候选项标签不会监督这一表征。随后在潜在空间中进行偏好推断。对于包含六条轨迹的数据集，我们将八维变分自编码器与高斯过程偏好模型结合使用（**图 4j-l**）。另一次单独执行的探索性运行使用 VAE 坐标和有限随机特征效用近似，不应将其描述为对完整高斯过程后验的复现。这一区别在科学上很重要：较低的重建误差并不能保证某个潜在方向能够捕捉操作者感兴趣的领域结构；而良好的奖励预测也不要求忠实重构每一个像素。这两个目标衡量的是图像所保留信息的不同方面。

编码器和解码器通过最大化证据下界 [73],进行训练，

$$L_{\mathrm{VAE}} = E_{q(z|x)}[\ln p\,(x|z)] - \mathrm{KL}[q(z|x)\|p(z)] \tag{12}$$

随后将编码器的均值作为公式（1）中选择模型的输入。

偏好训练卷积模型针对记录的获胜选项，联合优化图像表示和奖励函数。在所实现的深度核学习启发式构造中，卷积编码器将 20 × 20 像素的图像块映射到两个潜在坐标，有限随机傅里叶特征头则为其分配效用。选择似然通过训练表征，使其能够区分选定图像块与其实际竞争项。尽管该架构类似于深度核学习 [74,107–110], ，但不应将所实现的有限特征模型等同于完全贝叶斯高斯过程后验：任何最终层海森矩阵或拉普拉斯近似都以拟合后的编码器为条件，且不会传播所有图像编码器不确定性。因此，我们考察内部激活、潜在空间几何、奖励分布和留出选择，而不提出训练过程无法支持的贝叶斯不确定性主张。架构和预处理的详细信息见补充材料第 S4 节。

经过偏好训练的奖励由带有参数 ϑ 的卷积编码器 g 和公式（11）的随机特征头 [74], 组成

$$R_{\mathrm{DKL}}(x) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}\big(g_\theta(x)\big), \qquad g_\theta\colon \mathbb{R}^{20\times 20} \to \mathbb{R}^2 \tag{13}$$

并通过最小化训练决策上的式（2）中的选择负对数似然，联合拟合 θ 和 β；拉普拉斯近似 [75,76] 仅应用于 β 。

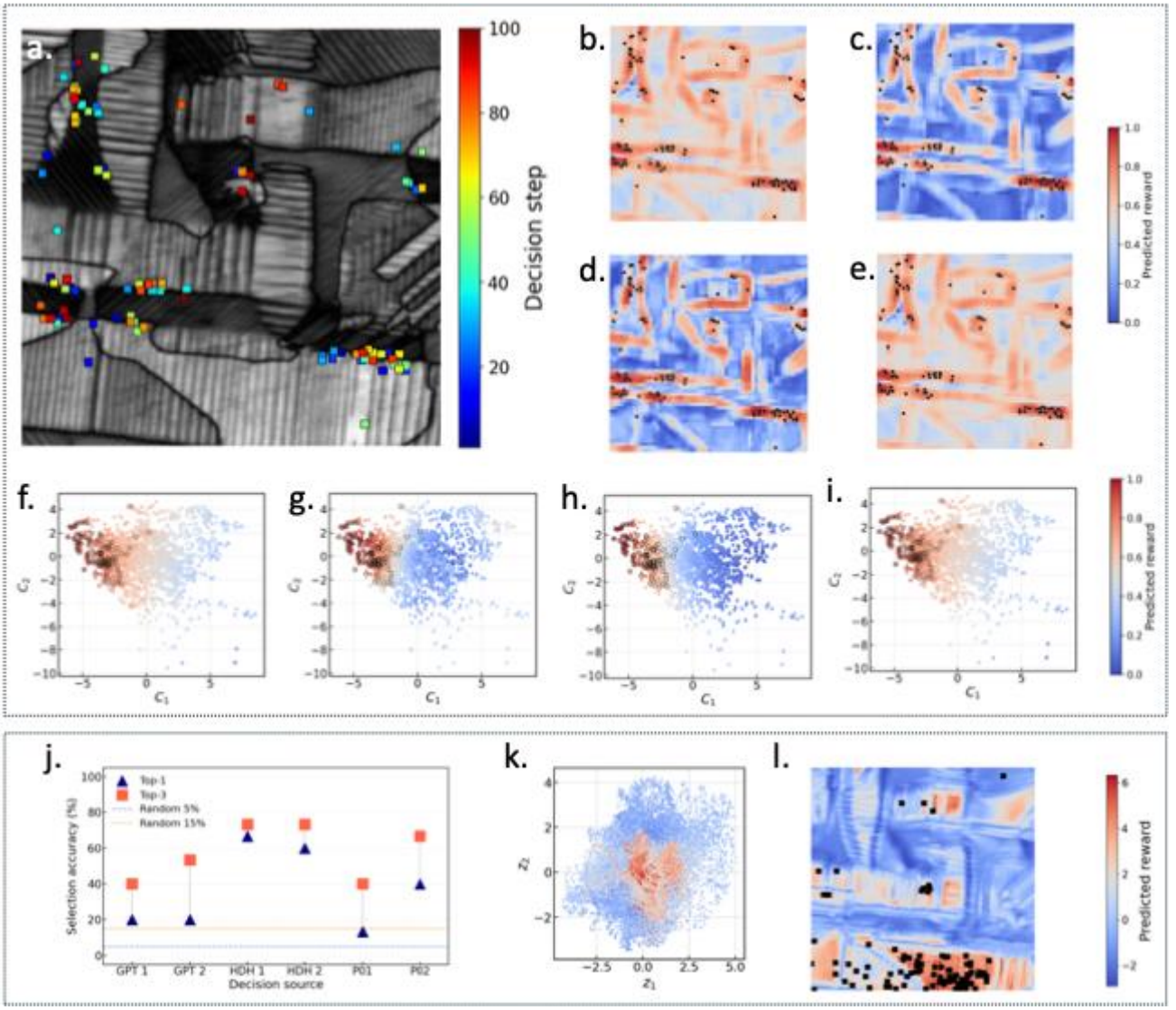


图 4. 使用偏好训练的深度核启发模型和变分自编码器重构奖励–高斯分布过程（VAE–GP）表征。上部（a–i）：决策偏好的深度核启发式重构。（a）原始PFM图像，以及来自 HDH 1 的记录选择位置，按决策顺序进行颜色编码。（b–e）为四个具有代表性的决策历史重构的空间奖励图，黑色标记表示记录的选择位置。（f–i）对应的奖励景观投影到二维学习特征空间，其中色标表示预测奖励，带轮廓的标记表示记录的选择位置。（e、i）对应于 P02。每个模型的奖励值均进行独立归一化。下部（j–l）：VAE–GP 奖励重构。（j）六条决策轨迹（GPT 1、GPT 2、HDH 1、HDH 2、P01 和 P02）上的留出 Top-1 和 Top-3 选择准确率，虚线分别表示 5% 和 15% 的随机选择基线。（k）P02 的预测奖励景观，以 VAE潜在表示中的二维空间进行可视化。（l）对应的空间奖励图投影到原始PFM图像上，黑色方块标记记录的选择位置。

图 4 将平稳奖励重建扩展到学习到的图像表示，

比较深度核学习（DKL）和变分自编码器–高斯过程（VAE–GP）

六条决策轨迹中的方法。上方各面板（**图 4a–i**）展示了使用深度核启发模型进行的偏好重建。图 4a 显示了 HDH 1 的记录选择，而图（**4b–e**）和图（**4f–i**）则展示了将同一模型架构分别拟合到四条具有代表性的决策历史后得到的空间奖励景观和潜在奖励景观。记录的选择集中在显著的域边界附近，以及局部强度变化较大的区域（**图 4a**）。重建得到的空间奖励图（**图 4b–e** ）显示，存档的 PFM 图像中存在不同的偏好模式，其中若干包含记录选择的区域具有较高奖励。这些区域在空间范围和强度上的差异表明，独立拟合的模型捕捉到了记录的决策历史中不同的偏好。投影到二维学习特征空间后，还能进一步看出各模型如何在候选图像表示中分布奖励（**图 4f–i**）。记录的选择仅分布在该空间中相对有限的区域，预测奖励较高的区域与其分布部分重叠。这些投影以直观的方式展示了学习到的偏好景观。对于偏好训练模型，这两个坐标构成了用于奖励预测的学习到的表征；而 VAE–GP 结果仅显示八维 VAE 表示中的两个坐标。

下方各面板（**图 4j–l**）考察了 VAE–GP 方法，其中高斯过程偏好模型 [50,51] 基于变分自编码器学习到的图像表示进行运算。在六段决策历史中，对于两条启发式驱动轨迹，该模型在约 73% 的留出决策中将记录的胜出选项列入排名前三的候选项；对于参与者 P01 和 P02，这一比例分别为 40% 和 67%；对于 GPT 1 和 GPT 2，则分别为 40% 和 53%（**图 4j**）。这些结果表明，学习到的图像表示能够恢复有意义的选择偏好，但不同决策来源的预测性能差异显著。对于参与者 P02，重建的奖励景观在二维 VAE 潜在空间中平滑变化（**图 4k**）。将这一奖励投影到存档的 PFM 图像上后可以发现，在域对比度明显的区域附近，偏好较高，其中包括记录的选择聚集的区域（**图 4l**）。然而，由于 VAE 表示是通过图像重建而非候选选择监督学习得到的，视觉上显著的特征未必对应于驱动已记录决策的特征。相比之下，基于偏好训练的 DKL 方法利用观察到的选择直接优化图像表示，因此可能会突出与潜在决策策略更相关的特征。

对于 HDH 1，偏好训练模型将 20 个留出获胜项中的 14 个正确排在首位（70%），而对应的线性描述符模型为 20 个中的 13 个（65%）。然而，线性

模型取得了更低的负对数似然（0.90，而非 1.19）。因此，偏好训练表征仅使该轨迹的 Top-1 准确率略有提升，却没有改善概率预测。由此，非线性表征在计数上有所改善，但却降低了分配给观测到的选择的概率。这使得该示例具有信息价值，而不是一个对深度模型普遍有利的例子。若要确立某种表征的普遍优越性，还需要采用可比的监督预算、特征预处理、重复测试划分以及独立图像评估。在归档基准中，无论采用何种表征，六条轨迹中的一些仍然难以处理，这表明仅仅扩展网络无法解决被遗漏的背景信息和非平稳性问题。

这两个潜在坐标还可以通过比视觉上吸引人的散点图更具物理意义的方式进行检验。将其与简单的图像统计量进行相关性分析，可以揭示学习得到的坐标轴主要编码的是强度、边界、周期性还是其他测量属性。在示例中（**图 5c**），与强度相关的描述符解释了潜在变化中的很大一部分，这表明即使是深度偏好模型，也可能重新发现相对简单的图像规律。这些相关性有助于我们理解潜在坐标所代表的内容，但无法说明决策者为何偏好这些特征。要进行更直接的解释，可以在近似保持图像结构不变的同时，改变某一项特征——对比度、边界锐度、取向或空间周期——并测量拟合效用的敏感度。在将潜在方向解释为科学概念之前，这类反事实扰动仍是一项重要检验。

图 5 总结了偏好训练模型。对于一条具有代表性的启发式轨迹（HDH 1），所选候选项位于学习得到的二维嵌入的一端（图 5a），拟合效用朝该端递增（**图 5b**），且第一个潜在坐标与平均强度及强度百分位数正相关，而第二个坐标的相关方向相反（**图 5c**）。将同一架构分别应用于六条记录的历史后，可在共同图像上得到各异的空间效用场（**图 5d、e**）。由此得到的图谱因 GPT生成的、由启发式算法驱动的以及人类参与者的历史记录而有所不同，反映了记录下来的选择偏好存在差异。各列中的拟合模型和色标并不相同，因此这组图比较的是空间结构，而非奖励幅度。

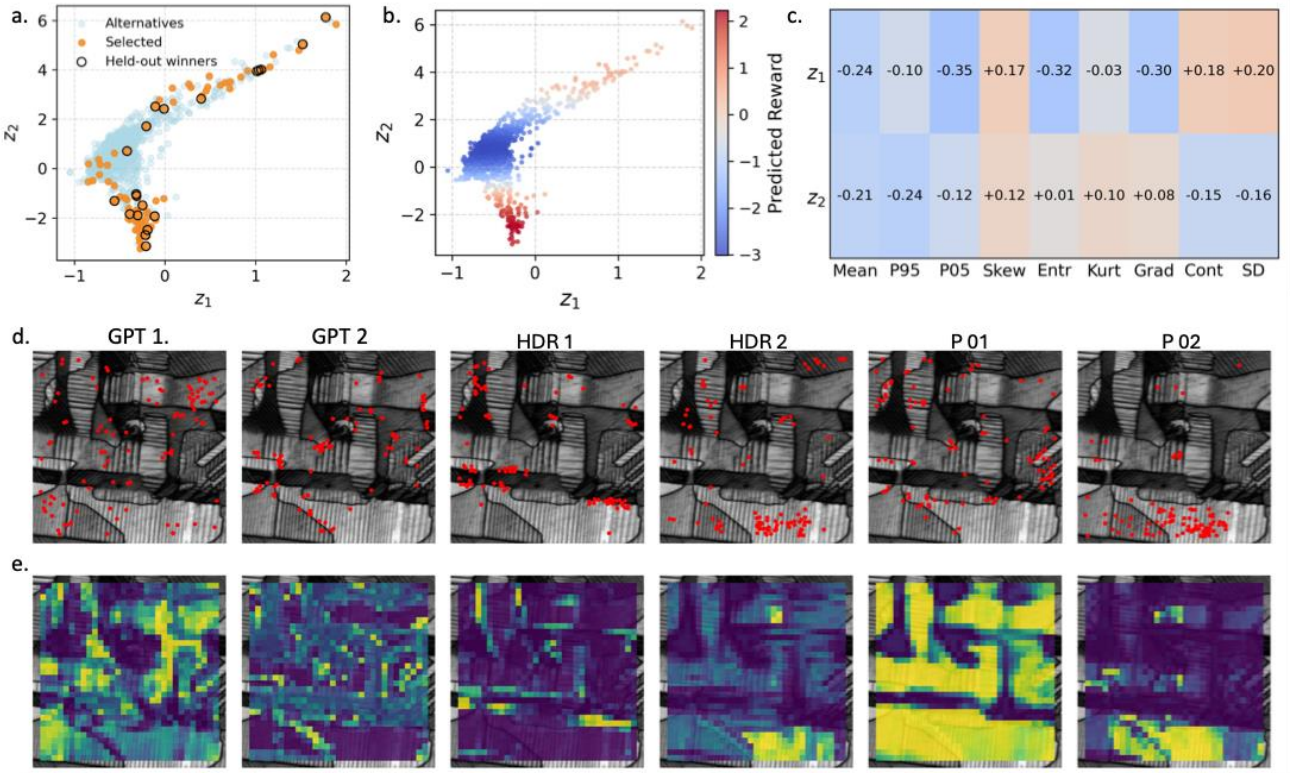


图 5。 用于奖励重建的偏好训练的深度核学习（DKL）。 (a) HDH 1 学得的二维嵌入，显示所选候选项（橙色）、备选项（浅蓝色）和留出获胜项（黑色圆圈）。(b) 同一潜在空间中的预测奖励。(c) 潜在坐标与九种图像描述符之间的斯皮尔曼相关系数。(d) GPT 1、GPT 2、HDH 1、HDH 2、P01 和 P02 的记录选择位置。(e) 独立训练的 DKL 模型所对应的空间奖励图，每张图均使用各自的色标显示。

总之，这些实验展示了如何利用预定义的形态学描述符或学习得到的图像表示，重建并探究一种稳定的有效偏好。主要发现是存在异质性：简单特征能够有效描述多种选择过程，而表达能力更强的函数则提供了其他描述方式，却未能带来一致的预测收益。与记录的历史记录吻合度较低可能有多种原因：重要的图像特征可能缺失，选择可能不一致，相关背景信息可能未被记录，或者决策者的优先事项可能已经改变。时间分析使用相同的记录候选集考察了最后一种可能性。

4. 从同一记录实验中学习变化的奖励

在人类实验过程中，人类决策很少始终如一。人类操作员的典型实验阶段会经历几个阶段：首先确定仪器的基线性能，接着解读样品中统计显著的特征，然后关注那些出于好奇而引人注意的异常特征，之后还可能探索

由特定科学问题决定的特征。偏好也可能因其他原因而变化：操作者会根据已获取的信息进行学习，在探索新区域和返回已知区域之间交替[77,78], ，或者只是变得疲倦；这类变化可能是渐进的，也可能在个别时点突然发生[79]。在行为研究中，离散策略之间的切换以及决策权重的渐变漂移都有充分记录[80–82]。用单一平稳奖励拟合这类历史，会将若干不同偏好取平均。本节通过两种互补的描述方式对此类变化进行显式建模：一种是隐马尔可夫模型，其中潜在状态从一小组反复出现的奖励中进行选择（第 4.1 节和第 4.2 节）；另一种是连续状态模型，其中奖励权重随每次决策逐渐变化（第 4.3 节）。按照验证行为模型的既定做法[83], ，我们在已知变化的受控序列以及原始决策历史上对这两种模型进行测试。

4.1. 决策历史可能反映多个目标

一系列显微镜选择可能反映多项科学优先事项，尽管每次决策的形式都相同。实验人员可能先定位畴壁，接着寻找排列有序的条纹以建立参考形貌，之后再考察边界，检验某种解释是否具有普遍性。根据完整决策历史拟合的平稳奖励会对这些状态取平均；即使每种状态各自都遵循简单的依赖图像的规则，这种平均也可能导致预测不佳。因此，我们将平稳性视为有待检验的假设，而非记录轨迹的固有属性。现在，我们要探究：用少数几种反复出现的偏好来描述这段历史是否更合适，还是用随时间逐渐变化的奖励来描述更合适。

隐马尔可夫表述将每个未观测状态 zt 与奖励 Rk(x) 关联起来，因此也将其与决策 t 时所提供候选项上的特定条件选择分布关联起来。马尔可夫转移矩阵规定了这些状态的持续性，以及模型返回先前状态的可能性。状态并未预先标记；模型会根据选择学习其奖励函数和转移概率。潜在状态是关于选择行为的一种操作性假设，并非对注意力或心理意图的测量。只有当这些状态能对可比的候选集作出不同预测，并且对状态变化进行建模能比平稳描述带来可复现的优势时，这一假设才有意义。

$$P(a_t = i \mid C_t, z_t = k) = \frac{\exp[\beta R_k(x_{t,i})]}{\sum_j \exp[\beta R_k(x_{t,j})]}, \qquad R_k(x) = \mathbf{w}_k^{\mathrm{T}} \tilde{\boldsymbol{\phi}}(x) \tag{14}$$

其中，

$$P(z_t = k \mid z_{t-1} = j) = A_{jk} \tag{15}$$

隐马尔可夫模型需要比绘制后验状态概率图更多的验证。不同的系数向量可能得出实质上相同的候选项排名，而灵活的模型可能会将一种稳定策略拆分为多个冗余状态。因此，我们会比较状态特定选择分布、排名最高的候选项、转移持久性，以及与不同状态数量对应的留出似然。状态标签可以互换，因此在根据受控真实值计算恢复准确率之前，必须先对齐这些标签。离线前向–后向平滑和维特比解码可以概括完整的历史序列，但会使用后续观测；它们无法衡量提前预测能力。在按时间顺序测试中，模型必须根据每次选择之前可用的信息预测获胜项，然后才能更新其状态概率。

按时间顺序进行预测时，会通过转移矩阵传播经过滤的状态分布 πt1，并混合状态特定选择模型 [84,85],

$$P(a_t \mid C_t, H_{t-1}) = \sum_k \pi_{t|t-1}(k) P(a_t \mid C_t, z_t = k), \qquad \pi_{t|t-1}(k) = \sum_j \pi_{t-1}(j) A_{jk} \tag{16}$$

随后，观测到的胜出项会更新状态分布，

$$\pi_t(k) = \frac{\pi_{t|t-1}(k) P(a_t|C_t, z_t=k)}{\sum_j \pi_{t|t-1}(j) P(a_t|C_t, z_t=j)} \tag{17}$$

参数通过广义期望最大化进行估计，在前向–后向状态责任度与公式（9）的加权拟合 [86,87] 之间交替。随后，维特比解码会识别训练决策 [88] 中最可能的状态序列。持久性先验、显式时长模型和时间扩展选项为这一形式提供了自然的扩展 [89–91]。

为了检验状态恢复，而不是依赖主观解读，我们基于实际的显微成像候选集构造受控序列，同时根据已知的依赖图像的奖励生成新的胜出项标签。教师系数、状态身份和转移时间对推断程序保持隐藏，仅在评估时揭示。我们检验固定奖励、一次隐藏切换、在两种熟悉的奖励之间反复切换、渐变漂移，以及训练后出现新策略这几种情况。这些条件将三个问题区分开来：奖励固定时，模型是否会臆造变化；模型是否能恢复反复出现的策略；以及偏好朝全新的方向发生质变时，其状态空间能否适应。这些是对逆向方法的检验，而非人类动机发生变化的观察证据。

4.2. 发现反复出现的策略及其切换

图 6 和图 7 考察了一段构造的 200 次决策序列，该序列由 GPT 1 和 GPT 2 各自记录的 100 次决策拼接而成。拼接后在第 100 次决策之后形成了一个已知的来源边界，无需额外的显微成像数据。双状态模型（**图 6a**）提供了

一种以两个潜在奖励机制描述这段构造历史的方法。将滤波状态概率与第 100 次决策处已知的拼接边界进行比较；该边界仅在拟合后显示（**图 6b**）。两种状态的主要差异体现在各向异性系数和方向相关系数上，这些系数的符号相反（**图 6c**）。它们的空间效用强调不同的领域区域；稳态图与状态 0 的图相似，终端连续图则与状态 1 的图相似（**图 6d**）。这种构造结合了不同的生成历史，因此检验的是来源判别，而非同一决策者内部的变化。

更严格的预设切换测试采用单一的候选生成方案，其中包含两个已知奖励函数和未公开的转移时间。对于在第64、123、204和243个决策处引入的重复变化，拟合的隐马尔可夫模型在按时间顺序留出的片段上取得了61.9%的 Top-1 准确率，而平稳模型和持续更新模型分别为31.0%和44.0%。这支持了一个更为狭义的结论：在受控条件下，只要选择足够具有信息量，就可以恢复离散的潜在奖励状态。但这并不能证明原实验人员在这些时间点改变了实验目标。受控的平稳、重复切换和连续漂移对比见补充材料第 S5 节。

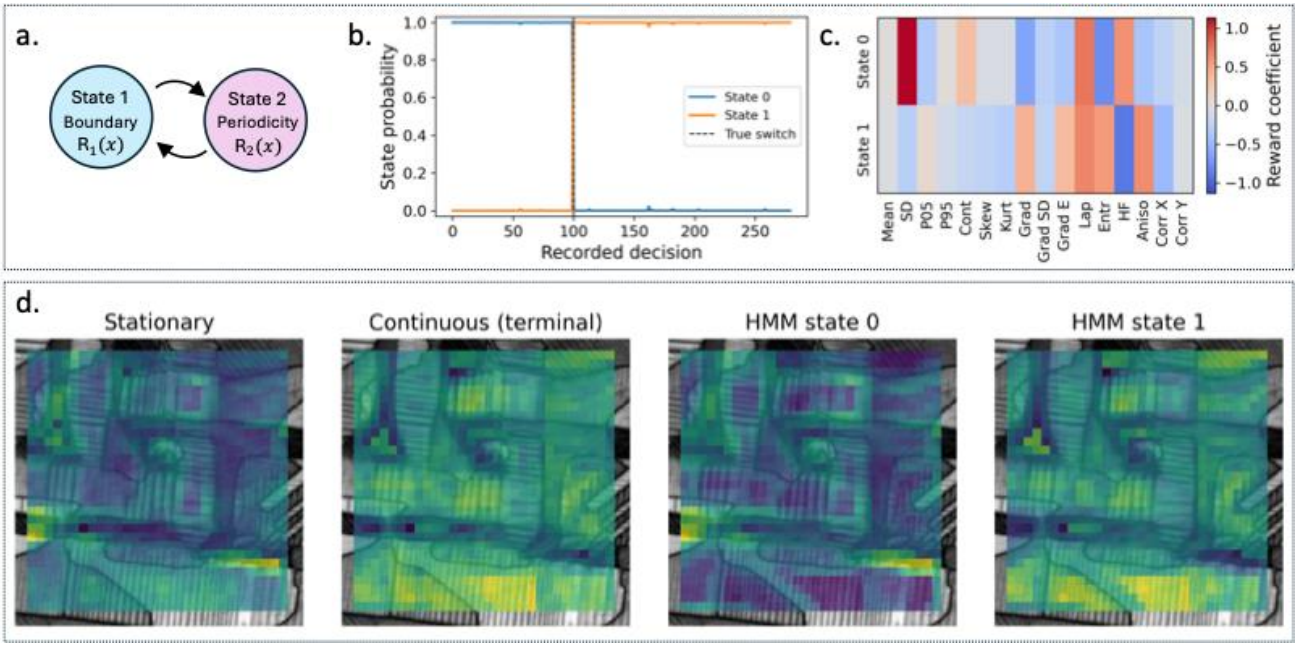


图 6. 构造的双来源决策历史中的潜在偏好状态。(a) 双状态模型：隐状态决定支配每个已记录选择的奖励。(b) 200 次决策的滤波状态概率；虚线标示决策 100 处的拼接边界，该图在拟合后显示。(c) 16 个图像描述符的特定状态系数。(d) 平稳模型、终端持续更新模型和两个隐马尔可夫模型状态的空间效用。

图7通过考察使用验证似然选出的三状态隐马尔可夫模型，扩展了对同一串联 GPT 1–GPT 2 序列的分析。滤波概率显示，边界之前存在两个交替出现的状态，而边界之后则只有一个持续存在的状态（图7a）；训练前缀的维特比路径重现了这一结构（图7b）。各状态之间差异最大的是峰度、均值梯度和各向异性系数（图7c）。在按时间顺序划分的测试窗口（第140–199个决策）中，隐马尔可夫模型和持续更新模型累积的负对数似然显著低于平稳模型（图7d）。由于该窗口完全位于边界之后，这一优势反映的是对第二个来源的适应，而不是对未知切换的预判；预设切换实验针对的正是这种情况。不同状态之间的空间图仅存在轻微差异（图7e）。

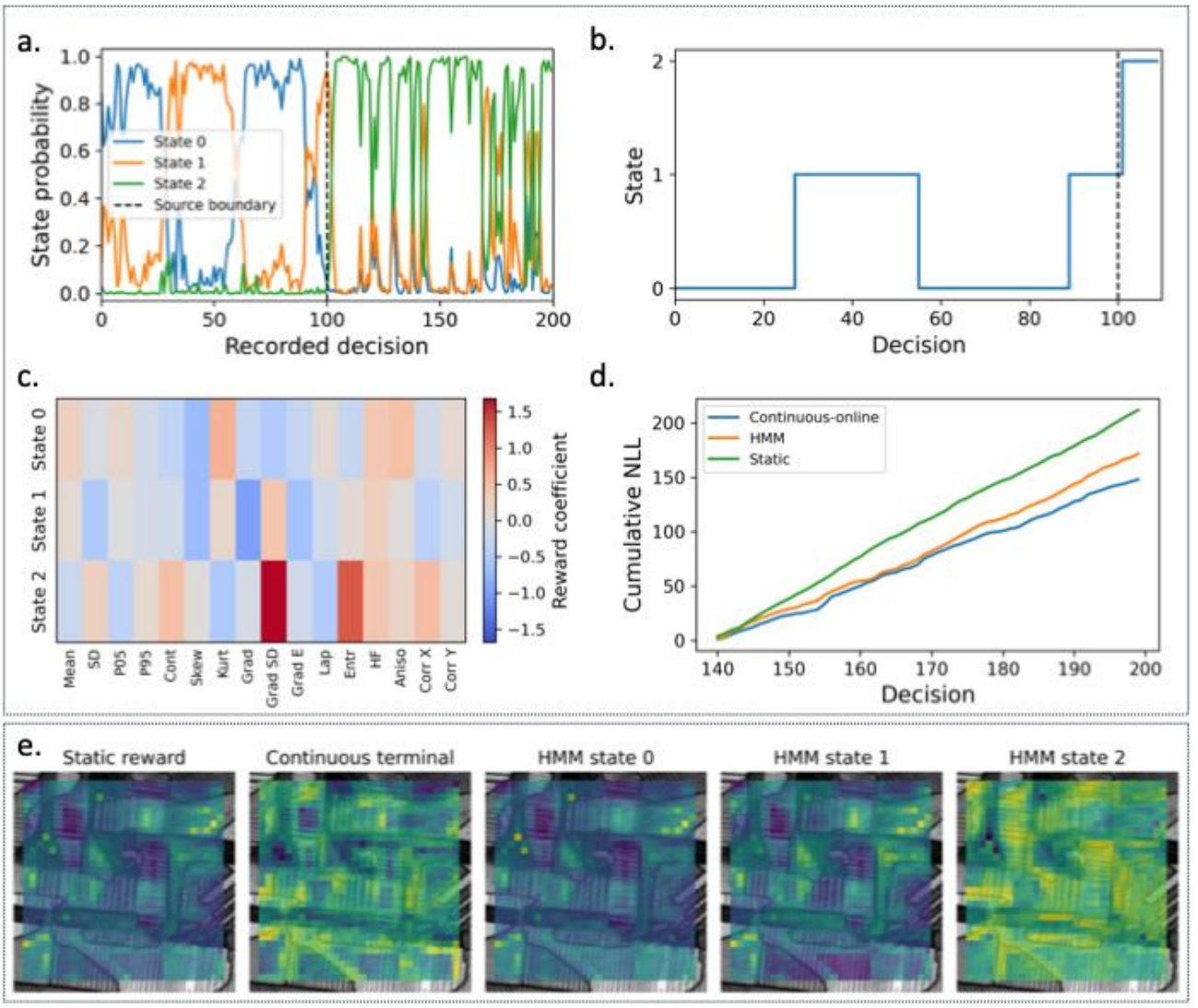


图7。 对两条 GPT 生成的决策轨迹进行串联后建立的时序模型。(a) 三个隐马尔可夫模型状态的滤波概率；虚线标示决策 100 处已知的边界，该结果在拟合后显示。(b) 训练前缀上最可能的状态路径（Viterbi）。(c) 状态特异性描述符系数。(d) 平稳模型、隐马尔可夫模型和连续模型的累积留出选择负对数似然

在按时间顺序划分的测试窗口上更新模型。（e）平稳模型、终端连续模型和三个隐马尔可夫模型状态的空间效用。

阴性对照与成功的切换演示同样重要。在对原始轨迹进行探索性拟合时，两个数值状态向量经常选择同一个首选候选项，选择分布差异较小，并且在近似状态数判据下未能优于单状态模型。因此，仅增加参数并不能证明存在额外的科学策略。相比之下，受控的平稳教师模型应当只产生一个偏好状态，或几乎没有证据表明存在有意义的切换。我们还考察了中断之后，推断出的策略是否会以相近的奖励排名再次出现，因为这是用马尔可夫表示来描述重复出现的目标这一独特应用场景。

此前未见过的策略条件揭示了第二个局限。仅使用历史记录第一段拟合的隐马尔可夫模型，无法立即表示训练中从未出现过的奖励。它可能会将此类决策分配给最不矛盾的现有状态，也可能要在收集到更多证据后才进行调整。连续演化的奖励有时响应更快，但代价是难以表示突发的重复变化，或过度追踪单个噪声选择。这些是相互竞争的解释性假设，而非必然只有一个胜出的普遍竞争。它们的相对效用取决于生成奖励之间的差异、转移频率、选择次数以及候选集所包含的信息量。

4.3. 渐进变化：离散策略的替代方案

科学优先事项也可能在没有发生离散转变的情况下改变。随着操作者对样本逐渐熟悉，他们可能逐渐倾向于不同的空间频率、扩大可接受的对比度范围，或将注意力从孤立边界转向扩展模式。我们使用一个依据平滑随机过程演变的奖励向量来模拟这一替代情形，每次新的选择都会更新当前估计。先前的估计用于预测下一次选择，而新观察到的胜出选项则在之后才被纳入。过程噪声设置控制适应性与稳定性之间的权衡。较小的值表示偏好几乎保持不变；较大的值则允许偏好快速变化，但容易受到噪声影响。完整的滤波方程和时间顺序评估方案见补充材料第 S5 节。

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, q\mathbf{I}), \qquad R_t(x) = \mathbf{w}_t^{\mathrm{T}} \tilde{\boldsymbol{\varphi}}(x) \tag{18}$$

采用后验分布的高斯近似时，每个观测到的选择都会按如下方式更新均值和协方差：

$\mathbf{w}_t = \mathbf{w}_{t-1} + \mathbf{\Sigma}_t\big[\tilde{\boldsymbol{\varphi}}\big(x_{t,\ a_t}\big) - \overline{\boldsymbol{\varphi}}_t\big], \quad \mathbf{\Sigma}_t^{-1} = (\mathbf{\Sigma}_{t-1} + q\mathbf{I})^{-1} + \sum\ p_{t,\ j} \quad _j\big(\tilde{\boldsymbol{\varphi}}_{t,\ j} - \overline{\boldsymbol{\varphi}}_t\big)\big(\tilde{\boldsymbol{\varphi}}_{t,\ j} - \overline{\boldsymbol{\varphi}}_t\big)^{\mathrm{T}}$ (19)，其中 pt，j 是候选项 j 的预测概率，而 $\bar{\phi} = \Sigma$j pt，j φt，j是预期描述符向量。这是 softmax 观测的 $Kalman$ 滤波器 [82,92]类比：更新会根据当前不确定性进行缩放，使权重朝着观测到的胜出项的描述符移动，而 q 则决定稳定性与适应性之间的权衡。

图8对比了受控非平稳序列中的离散状态描述和连续状态描述。HMM通过为不同的潜在状态分配概率来表示不断变化的偏好（**图8a**），而连续状态模型则允许奖励系数随着连续的决策逐步变化（**图8b**）。虚线标示训练窗口和验证窗口的结束位置（决策154和196）。图8c展示了持续更新模型在存档图像上的终端奖励图，说明如何在空间上读取拟合系数向量。

受控比较涵盖平稳选择、一次突变、重复切换、连续漂移，以及训练后引入新偏好这几种情况。在保存的平滑漂移构造中，在线连续更新的留出集 Top-1 准确率达到 61.9%，相比之下，隐马尔可夫模型为 50.0%，平稳模型为 4.8%。在重复切换构造中，隐马尔可夫模型的 Top-1 准确率为 61.9%，相比之下，平稳模型为 31.0%，连续更新为 44.0%。这些差异反映了对所测试生成过程和所选超参数的观察结果，并不能构成证明某种方法本质上更优的定理。这些结果表明，非平稳性的类型是一项科学假设，应当加以检验，而非想当然地认定。

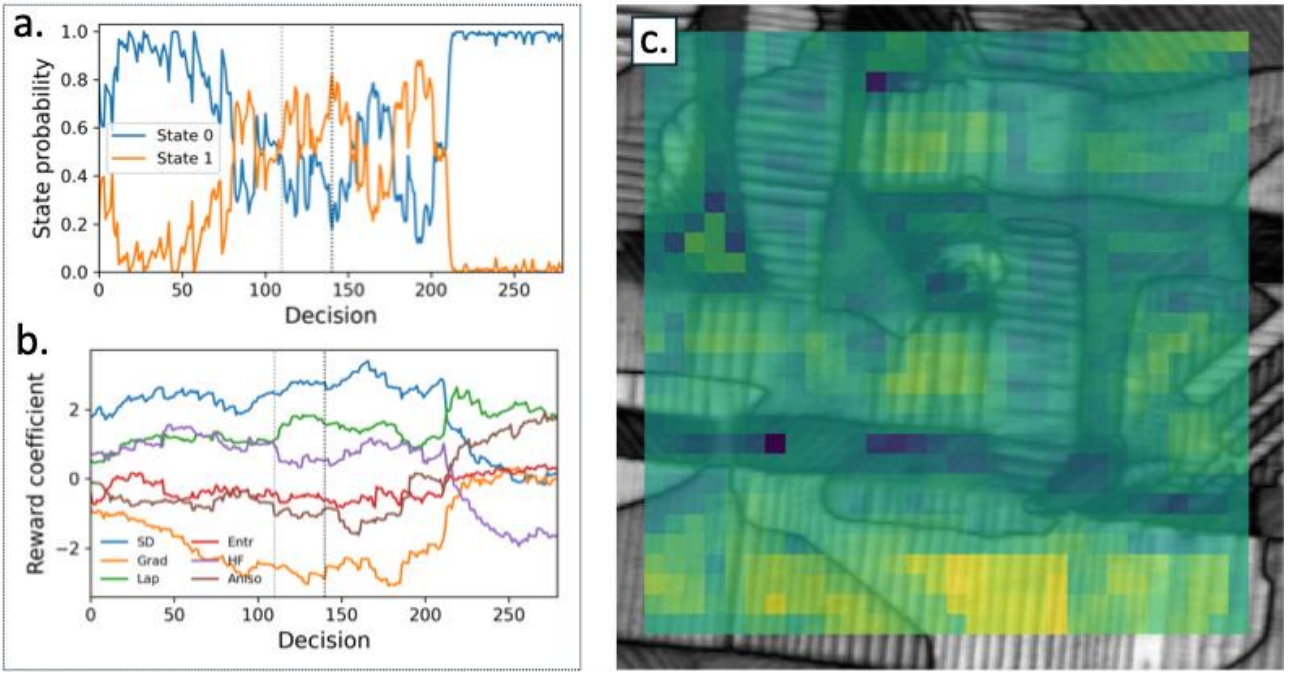


图8。 偏好变化的离散状态和连续状态描述。(a) 在受控非平稳序列中，每次观察到选择后拟合的隐马尔可夫模型状态的在线滤波概率。(b) 同一序列中持续更新模型的奖励系数。(a) 和 (b) 中的虚线标示训练和验证窗口的结束位置。(c) 在存档图像上持续更新模型的终端空间奖励图。

六条未经修改的决策历史分别由两条 GPT生成的历史、两条由启发式算法驱动的历史和两条人类参与者历史记录组成，可用于单独检验时间建模是否能改善回顾性预测。在保存的按时间顺序进行的评估中，平稳模型、隐马尔可夫模型和连续更新的平均留出集 Top-1 准确率分别为 40.0%、40.0% 和 45.0%，相应的平均负对数似然值分别为 2.10、2.10 和 1.85。相比程序化切换案例，这些提升要小得多。若干历史记录用一个有效状态就足以描述，而部分拟合的多状态模型中，各状态之间的行为区分并不明显。这些观察结果并不能证实人类研究者的动机自发发生了变化。六条记录的历史中有四条是由基于 GPT 的评估器或人为定义的启发式方法生成的，而且本研究并非旨在前瞻性调查人类动机的变化。

同一种表面上的决策变化可能有多种解释。候选群体可能发生变化，图像质量可能有所不同，任务说明可能改变，或者评估者可能获取了记录的图像块描述符未能呈现的信息。条件建模能够考虑所提供的备选项，但仅凭图像选择无法识别遗漏的情境变量。因此，未来的回顾性数据集应保留指令、仪器模式、较早的观察结果，以及在存在此类记录时独立标注的任务变更。此类记录将

有助于比较奖励变化、在固定奖励下对样本的信念发生变化，以及可采纳动作发生变化这几种情况。本研究建立了相关方法并验证了可控条件下的可恢复性，但并未实现唯一的心理重构。

5. 讨论

5.1. 决策历史揭示了什么

分析表明，记录的决策历史是实验目标的一种可用来源。当保留每一步提供的备选项时，根据部分决策历史重构的奖励，其对留出选择的预测准确性显著高于随机猜测。对于某些启发式驱动的历史，模型表现尤其出色，这表明可以从记录的决策中恢复明确的依赖图像的选择规则。可解释描述符使推断偏好变得清晰易读：拟合系数能够确定哪些图像特征（例如强度统计量、梯度或各向异性）区分了选定图像块及其竞争项。学习得到的表征提供了互补的描述；在偏好训练形式下，其潜在轴能够与同样熟悉的统计量建立联系。在样品上进行投影后，每个模型都会产生一个空间奖励场，将价值集中于决策者偏好的领域区域。在时间设置下，隐马尔可夫模型在编程设定的重复切换条件下提高了预测性能，而持续更新模型在所测试的渐进漂移条件下表现更佳。总体而言，这些结果既为从决策轨迹中学习操作者奖励提供了框架，也提供了基准测试。

这些结果也有明确的局限。该逆问题学习的是一种由可观测选择行为定义的有效奖励。它既不是对某种物理属性的直接测量，也不是对科学意图的唯一可恢复报告。在条件逻辑斯蒂模型下，对所有候选项效用加上同一个常数，不会改变选择概率；当选择温度也作为自由参数时，奖励尺度同样会与其混淆。高度相关的特征可能对应许多几乎等效的系数向量，而任何取决于未记录上下文的偏好，都无法通过纯图像模型辨识。因此，不同的奖励可能会对观测到的候选集作出相似预测，却为从未提供过的图像块赋予不同分数。实际问题并非是否恢复了某种单一的“真实人类奖励”，而在于该模型能否对某一特定决策规则提供稳定、可检视且可迁移的解释。只有当这一主张经受住留出选择、改变候选集、独立视野以及对所得科学观测的独立评估时，它才更有说服力。

平稳比较表明，有两个选择会塑造重建奖励：哪些图像属性纳入模型，以及用什么函数将它们转化为偏好。基于可识别描述符的线性奖励是一个透明的起点，通常也能很好地捕捉启发式策略；而非线性模型和学习模型则提供了其他描述方式，但在当前基准测试中，平均而言并未提升预测效果。这反映的是本文研究的任务和数据，并不意味着显微镜操作员的思维方式是线性的。对于不断变化的偏好，拟合权重的变化并不必然意味着目标发生了变化，因为相关描述符之间可以相互转移权重，而不改变人们偏好的图像块 [41,42,93]。受控序列通过将推断出的变化与已知变化进行比较来处理这一问题；若要确认真实操作员历史记录中的优先事项发生了变化，则需要有关操作者意图的独立记录，例如指令或标注。

5.2. 从重建奖励到仪器可能采取的行动

仅有奖励并不能决定仪器应如何行动。奖励对候选位置进行排名，而采集策略还要决定考虑哪些候选项、每次测量的成本、是否可以重新访问某个位置，以及如何避免冗余 [94]。若仅在图像上选择预测奖励最高的位置，往往会在某个有利区域内选出许多重叠图像块，因此，所提出的方法必须与能够保持多样性的策略相结合，例如抽取规模有限的候选集、加入覆盖均衡项，或使用不确定性估计。本文中的所有奖励图均基于已采集的图像计算；它们展示了模型会在哪些位置赋予价值，而利用这些奖励图来驱动新的测量仍有待验证。补充材料第 S2 和 S6 节概述了各种策略选项及其假设。

将推断出的偏好转化为自主科学效用，离不开独立终点。用模型自身的奖励来评估该模型选出的位置，会形成循环论证。更有力的检验方式是固定奖励函数，将其应用于新的考量集合或不同的样品，并通过盲法专家判断或可物理测量的响应来评估由此产生的选择。如果目的是最大化信息增益，就必须明确科学未知量及其观测模型；仅有空间离散度并不等同于信息论增益。可信的部署研究应在匹配候选生成成本和选择预算的前提下，对比均匀提议、全局贪心选择、最优 K 选一采样，以及明确的奖励加覆盖策略。这类比较仍属于拟议中的验证实验，并非当前显微镜档案得出的结果。

5.3. 记录决策改变了从实验室中能够学到的内容

实验室应记录哪些内容，才能实现奖励学习？仅有图像和光谱，即使按照 FAIR 原则 [95], 存档，也不足以满足要求。决策轨迹应保留曾考虑的备选项、其身份或坐标、最终选中的备选项、时间顺序、仪器条件，以及操作者可获取的指令或先验信息。选择集合尤为重要：若缺少这些信息，就无法判断某个位置被选中是因为操作者偏好它，还是因为其他选项不可用。结合仪器和样品的轨迹，这些信息有助于重构反复出现的操作目标，并发现操作者似乎重视的内容是否发生变化。此外，这些信息还可用于检验：观察到的变化究竟源于奖励变化，还是物理知识或行动可行性的变化。本研究展示了在保留考量集合时能够提取哪些信息，同时也指出了因背景信息缺失以及对单一样本反复成像而产生的局限。

偏好学习、专家比较、逆强化学习和人类引导的自主显微镜技术，为从选择中推断效用提供了大量先前研究 [33–38,41– 49,53,96–99]。我们的贡献并非发明条件选择建模或隐马尔可夫推断，而是提出了一种受实验启发的表述：将存档的显微镜考量集合用作数据，以比较可解释的奖励表征与学习得到的奖励表征；并且可以利用同一份已存储的决策历史，探究偏好是稳定的、反复出现的还是随时间漂移的。对照实验分别考察预测一致性、对预设目标变化的还原能力，以及最终的科学价值。在本研究中，后者仍未得到验证。这些学习得到的目标能否推广至不同的材料、成像模态、操作者和科学问题，仍是一个需要通过实证解决的挑战，而非模型表达能力必然带来的结果。

6. 结论

显微镜操作员的决策历史包含有关人类逻辑和决策的隐含信息：哪些区域被认为值得测量、相对于哪些备选项，以及这些判断在实验过程中如何演变。这些信息很少被记录下来，然而它们恰恰是自动化仪器代表操作者采取行动所需要的。

当每次选择时呈现的备选项都得到保留，记录下来的显微镜技术决策便包含可用于推断实际实验目标的信息。我们使用同一张存档图像，评估了基于可解释描述符、非线性函数和学习到的图像表示的平稳奖励。它们的预测价值在不同决策历史之间差异显著

，而增加模型复杂度并不能保证改善重构。空间投影使拟合效用可供检查，但并不能证明遵循这些效用会产生独立科学信息。在受控图像序列中，隐马尔可夫推断找回了隐藏的重复出现的奖励策略；对于经过测试的平滑漂移过程，持续更新的奖励则更为适用。在未经修改的决策历史中，时间模型带来的改进有限，因此无法推断人类动机会自发变化。综合来看，这些分析确立了一条严谨的路径，可从比较选择推导出定量实验偏好：重构奖励，在未见过的决策上检验奖励，审视其物理含义和可辨识性，并通过独立的科学终点评估任何拟议用途。

这为实现能够继承人类专业知识的自动化仪器开辟了一条切实可行的路径。从操作者的决策中学习得到的奖励函数可以固定下来，并由仪器以机器速度执行，用于评估样品中的候选区域、确定测量优先级，并作为贝叶斯优化或强化学习策略的目标。由于奖励是明确且可检查的，操作者可以在部署前验证仪器重视的内容、加以修正，并将其与基于物理的奖励结合起来。在运行期间监测推断的偏好状态，还能让仪器识别操作者的优先事项何时发生变化，并据此进行调整。因此，应将决策历史及其备选项一并记录，作为仪器操作的常规环节；这样，每次手动操作会话也能为后续自动化实验提供训练数据。

数据与代码可用性

分析代码整理在一个 Jupyter 笔记本中，涵盖数据加载、平稳奖励重建、图像表征学习和时序偏好推断。分析使用了一张存档的 PFM 图像（扫描_主_0000.ibw）和六个决策历史文件（gpt_1.pkl、gpt_2.pkl、HDH_1.pkl、HDH_2.pkl、P01.pkl 和 P02.pkl）。分析笔记本及复现所报告结果所需的数据可在[GitHub 仓库]中获取。

致谢与利益冲突

本研究由美国国家科学基金会通过 ATHENA 可编程云实验室测试平台（资助编号：2607469）提供支持。所有作者均通过黑客马拉松和黑客马拉松后分析机制参与了代码库开发，并参与了决策轨迹创建。

方法摘要

回顾性数据包含针对一幅 256 × 256 像素图像的六段100次决策历史；每次决策记录二十个 20 × 20 像素候选图像块及获胜者索引。这些历史包括两段GPT生成的轨迹、两段人为定义的启发式方法轨迹，以及两段人类参与者轨迹。标准化的十六描述符特征向量表征局部强度、梯度、拉普拉斯响应、频率成分、熵、各向异性和方向相关性；六描述符子集则提供了一个简化基线。线性模型、加性样条模型、有限随机特征核模型和迭代模型均通过正则化条件选择优化进行拟合，并在完整的留出决策集上进行评估。主要的平稳结果采用设定随机种子的80/20随机决策划分。基于偏好的图像模型采用卷积编码器、二维潜在空间和有限随机特征效用头，而先前独立的VAE分析则在进行选择建模之前学习图像表征。时间实验采用按时间顺序划分的训练、验证和测试窗口。在受控测试中，获胜选项由真实候选集上已知的、依赖图像的教师奖励生成；教师参数和变点均未提供给任何拟合程序。隐马尔可夫模型推断采用状态依赖的选择发射和转移概率，并使用验证似然选择状态数量。连续推断在每次观察到新的获胜选项后，更新正则化偏好向量。完整协议、局限性和代码沿革见补充信息。模型使用Python以及Scikit-learn、SciPy和PyTorch [100–102]实现；本文评估的六描述符基线使用与十六描述符模型相同的正则化条件选择似然进行拟合。它不同于此前介绍的六描述符排序边际公式 [63]。

参考文献

1 Canty, R. B. & Abolhasani, M. 自主驾驶实验室的过去、现在与未来。Nature Reviews Chemistry 10, 523–537（2026年）。https://doi.org/10.1038/s41570-026-00847-2。2 Volk, A. A. 等。AlphaFlow：利用强化学习引导的自驱动流体实验室实现多步化学的自主发现与优化。Nature Communications 14, 1403（2023年）。3 Burger, B. 等。一名移动机器人化学家。Nature 583, 237-241（2020年）。4 MacLeod, B. P. 等。用于加速发现薄膜材料的自动驾驶实验室。Science Advances 6, eaaz8867（2020年）。5 Szymanski, N. J. 等。用于加速合成新型材料的自主实验室。Nature 624, 86-91（2023年）。6 Abolhasani, M. & Kumacheva, E. 自主驾驶实验室在化学与材料科学中的兴起。Nature Synthesis 2, 483-492（2023年）。7 Krull, A.、Hirsch, P.、Rother, C.、Schiffrin, A. & Krull, C. 人工智能驱动的扫描探针显微术。Communications Physics 3, 54（2020年）。https://doi.org/10.1038/s42005-020-0317- 3。8 Rashidi, M. & Wolkow, R. A. 通过机器学习实现原位探针尖端调理的自主扫描探针显微术。ACS Nano 12, 5185-5189（2018年）。9 Chen, I.-J. 等。通过深度强化学习实现精确的原子操控。Nature Communications 13, 7499（2022年）。10 Leinen, P. 等。通过强化学习实现自主机器人纳米制造。Science Advances 6, eabb6987（2020年）。11 Feng, B. 等。用于科学发现的智能体人工智能系统。Nature Reviews Methods Primers 6, 81（2026年）。12 Chen, Z. 等。一名智能体人工智能X射线科学家。Nature Machine Intelligence 8, 1075–1086（2026年）。https://doi.org/10.1038/s42256-026-01261-5。13 Jamali, V.、Aghazadeh, A. & Kacher, J. 思维显微镜：智能体人工智能与电子显微术的未来。npj Computational Materials 12, 149（2026年）。https://doi.org/10.1038/s41524- 026-02077-y。14 Liu, Y. 等。人类-智能体发现可重构面内铁电超畴控制。Digital Discovery（2026年）。https://doi.org/10.1039/D6DD00706F。15 Boiko, D. A.、MacKnight, R.、Kline, B. & Gomes, G. 利用大语言模型开展自主化学研究。Nature 624, 570-578（2023年）。

16 Bran, A. M. 等。利用化学工具增强大语言模型。Nature Machine Intelligence 6，525–535（2024年）。https://doi.org/10.1038/s42256-024-00832-8。17 Yao, S. 等。ReAct：在语言模型中协同推理与行动。载于 International Conference on Learning Representations（2023年）。18 Kalinin, S. V. 等。将机器学习带入现实世界：我们只需要奖励。APL Machine Learning 4，010901（2026年）。https://doi.org/10.1063/5.0311536。19 Kalinin, S. V. 等。电子显微术和扫描探针显微术中的自动化与自主实验。ACS nano 15，12604-12627（2021年）。20 Shahriari, B.、Swersky, K.、Wang, Z.、Adams, R. P. 和 de Freitas, N. 让人类脱离闭环：贝叶斯优化综述。Proceedings of the IEEE 104，148-175（2016年）。21 Kusne, A. G. 等。通过贝叶斯主动学习开展动态闭环材料发现。Nature Communications 11，5966（2020年）。22 Liu, Y. 等。基于机器学习的奖励驱动扫描探针显微术调谐：迈向全自动显微术。ACS nano 19，19659-19669（2025年）。23 Liu, Y. 和 Kalinin, S. V. 帕累托前沿的力量：平衡扫描探针显微术中自适应实验的不确定奖励。arXiv预印本 arXiv:2504.06525（2025年）。24 Pratiush, U.、Houston, A.、Liu, R.、Duscher, G. 和 Kalinin, S. V. 迈向自优化电子显微镜：通过物理感知多目标贝叶斯优化实现像差系数的稳健调谐。Microscopy and Microanalysis 32，ozag071（2026年）。https://doi.org/10.1093/mam/ozag071。25 Ma, D. 等。像差校正的发射度最小化 I：在未知像差系数的情况下进行电子显微镜像差校正。Ultramicroscopy 273，114137（2025年）。26 Ma, D. 等。像差校正的发射度最小化 II：电子显微镜的物理信息驱动贝叶斯优化。Ultramicroscopy 273，114138（2025年）。27 Pattison, A. J. 等。BEACON——使用贝叶斯优化进行扫描透射电子显微术的自动像差校正。npj Computational Materials 11，274（2025年）。28 Barakati, K. 等。用于从原子分辨成像数据中对相和铁性变体进行无监督可解释分析的奖励驱动工作流程。Advanced Materials 37，2418927（2025年）。29 Barakati, K.、Yuan, H.、Goyal, A. 和 Kalinin, S. V. 显微镜技术中的基于物理的奖励驱动图像分析。Digital Discovery 3，2061-2069（2024年）。30 Barakati, K.、Liu, Y.、Pratiush, U.、Slautin, B. N. 和 Kalinin, S. V. 显微镜技术中的基于奖励的图像分析。npj Computational Materials（2026年）。https://doi.org/10.1038/s41524-026-02235-2。

31 Barakati, K.、Slautin, B. N.、Pratiush, U.、Funakubo, H. 和 Kalinin, S. V. PATHFINDER：结构空间和谱空间中的多目标发现。arXiv预印本 arXiv:2604.04194（2026年）。32 Barakati, K. 等。具有稳健特征袋多目标贝叶斯优化的自主探针显微术：纳米尺度结构–性能权衡的帕累托前沿映射。arXiv预印本 arXiv:2601.05528（2026年）。33 Liu, Y. 等。组合铁电材料库的自动扫描探针显微术：高斯过程引导的探索与噪声感知实验规划。Journal of Applied Physics 140，084101（2026年）。https://doi.org/10.1063/5.0341018。34 Slautin, B. N. 等。计算物理模型与原子力显微术实验的贝叶斯协同导航：自主测量组合材料库。ACS nano 20，23740-23747（2026年）。35 Vasudevan, R. K. 等。扫描探针显微术和光谱学中的自主实验：选择铁电体中极化动力学的探索位置。ACS nano 15，11253-11262（2021年）。36 Liu, Y. 等。通过主动学习实验发现铁电材料中的结构–性能关系。Nature Machine Intelligence 4，341-350（2022年）。37 Thurstone, L. L. 比较判断定律。Psychological Review 34，273-286（1927年）。38 McFadden, D. 定性选择行为的条件逻辑回归分析。载于 Frontiers in Econometrics（编者：Zarembka, P.），105-142（Academic Press，1974年）。39 Luce, R. D. 个体选择行为：理论分析。（Courier Corporation，2005年）。40 Bradley, R. A. 和 Terry, M. E. 不完整区组设计的排名分析：I. 成对比较方法。Biometrika 39，324-345（1952年）。41 Ng, A. Y. 和 Russell, S. J. 逆强化学习算法。载于第17届国际机器学习会议论文集，663-670（2000年）。42 Armstrong, S. 和 Mindermann, S. 奥卡姆剃刀不足以推断非理性智能体的偏好。载于神经信息处理系统进展31（2018年）。43 Abbeel, P. 和 Ng, A. Y. 通过逆强化学习进行学徒学习。载于第21届国际机器学习会议论文集，1-8（2004年）。44 Ziebart, B. D.、Maas, A.、Bagnell, J. A. 和 Dey, A. K. 最大熵逆强化学习。载于第23届AAAI人工智能大会论文集，1433-1438（2008年）。45 Arora, S. 和 Doshi, P. 逆强化学习综述：挑战、方法与进展。Artificial Intelligence 297，103500（2021年）。

46 Ross, S., Gordon, G. & Bagnell, D. 将模仿学习和结构化预测归约为无悔在线学习。载于《第14届国际人工智能与统计会议》627-635页（2011年）。47 Christiano, P. F. 等。基于人类偏好的深度强化学习。载于《神经信息处理系统进展》第30卷（2017年）。48 Ouyang, L. 等。通过人类反馈训练遵循指令的语言模型。载于《神经信息处理系统进展》第35卷，27730-27744页（2022年）。49 Biyik, E. & Sadigh, D. 奖励函数的批量主动偏好学习。《第2届机器人学习会议》, PMLR 87, 519–528页（2018年）。https://proceedings.mlr.press/v87/biyik18a.html。50 Chu, W. & Ghahramani, Z. 使用高斯过程进行偏好学习。《第22届国际机器学习会议》，137–144页（2005年）。https://doi.org/10.1145/1102351.1102369。51 González, J., Dai, Z., Damianou, A. & Lawrence, N. D. 偏好贝叶斯优化。载于《第34届国际机器学习会议》1282-1291页（2017年）。52 Itti, L., Koch, C. & Niebur, E. 用于快速场景分析的基于显著性的视觉注意模型。《IEEE模式分析与机器智能汇刊》20, 1254-1259页（1998年）。53 Kim, K., Garg, S., Shiragur, K. & Ermon, S. 逆强化学习中的奖励识别。《机器学习研究会议录》139, 5496–5505页（2021年）。54 Kwak, B. S. 等。外延铁电异质结构中的畴形成与应变弛豫。《物理评论B》49, 14865–14879页（1994年）。https://doi.org/10.1103/PhysRevB.49.14865。55 Scott, J. 等。铁电-铁弹薄膜中的超畴动力学：开关、阻塞与弛豫。《应用物理评论》4, 041104页（2017年）。https://doi.org/10.1063/1.5005994。56 Gruverman, A., Alexe, M. & Meier, D. 压电响应力显微镜与纳米铁性现象。Nature Communications 10, 1661页（2019年）。https://doi.org/10.1038/s41467-019-09650- 8。57 Pertsev, N. A., Zembilgotov, A. G. & Tagantsev, A. K. 力学边界条件对外延铁电薄膜相图的影响。《物理评论快报》80, 1988-1991页（1998年）。58 Gong, F.-H. 等。应变超薄PbTiO3薄膜中依赖厚度的极性畴演化。《ACS应用材料与界面》14, 9724-9733页（2022年）。59 Kelley, K. P. 等。通过动态压电响应力谱对压电响应弛豫机制进行阐释的张量分解。npj Computational Materials 6, 113页（2020年）。

60 Kelley, K. P. 等。压电响应力显微镜中的动态操控：利用大机电响应产生非平衡相。ACS Nano 14, 10569-10577页（2020年）。61 Plackett, R. L. 排列分析。《英国皇家统计学会杂志：C辑（应用统计）》24, 193-202页（1975年）。62 Train, K. E. 《离散选择模拟方法》。第2版（剑桥大学出版社，2009年）。63 Joachims, T. 利用点击流数据优化搜索引擎。载于《第8届ACM SIGKDD知识发现与数据挖掘国际会议》133-142页（2002年）。64 Csurka, G., Dance, C. R., Fan, L., Willamowski, J. & Bray, C. 使用关键点袋进行视觉分类。载于《计算机视觉统计学习研讨会》，ECCV 1-22页（2004年）。65 Haralick, R. M., Shanmugam, K. & Dinstein, I. 用于图像分类的纹理特征。《IEEE系统、人类与控制论汇刊》SMC-3, 610-621页（1973年）。66 Bigün, J., Granlund, G. H. & Wiklund, J. 多维方向估计及其在纹理分析和光流中的应用。《IEEE模式分析与机器智能汇刊》13, 775-790页（1991年）。67 Castelvecchi, D. 我们能打开人工智能的黑箱吗？《自然》新闻538, 20页（2016年）。68 Azodi, C. B., Tang, J. & Shiu, S.-H. 打开黑箱：面向遗传学家的可解释机器学习。《遗传学趋势》36, 442-455页（2020年）。69 Rudin, C. 停止解释用于高风险决策的黑箱机器学习模型，改用可解释模型。《Nature Machine Intelligence》1, 206-215页（2019年）。70 Hastie, T. & Tibshirani, R. 广义加性模型。《统计科学》1, 297-310页（1986年）。71 Rahimi, A. & Recht, B. 面向大规模核机器的随机特征。载于《神经信息处理系统进展》第20卷，1177-1184页（2007年）。72 Rasmussen, C. E. & Williams, C. K. I. 《机器学习中的高斯过程》。（麻省理工学院出版社，2006年）。73 Kingma, D. P. & Welling, M. 自编码变分贝叶斯。载于《International Conference on Learning Representations》（2014年）。74 Wilson, A. G., Hu, Z., Salakhutdinov, R. & Xing, E. P. 深度核学习。载于《人工智能与统计国际会议》370-378页（2016年）。75 MacKay, D. J. C. 反向传播网络的实用贝叶斯框架。《神经计算》4, 448-472页（1992年）。76 Daxberger, E. 等。拉普拉斯再现：轻松实现贝叶斯深度学习。载于《神经信息处理系统进展》第34卷，20089-20103页（2021年）。

77 Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. 人类探索性决策的皮层基底。《自然》441，876-879（2006年）。78 Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. 人类利用定向探索和随机探索来解决探索-利用困境。《实验心理学杂志：总论》143，2074-2081（2014年）。79 Adams, R. P. & MacKay, D. J. C. 贝叶斯在线变点检测。arXiv预印本 arXiv:0710.3742（2007年）。80 Ashwood, Z. C. 等。小鼠在知觉决策过程中交替采用离散策略。《自然-神经科学》25，201-212（2022年）。81 Calhoun, A. J., Pillow, J. W. & Murthy, M. 无监督识别塑造自然行为的内部状态。《自然-神经科学》22，2040-2049（2019年）。82 Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D. & Pillow, J. W. 从感觉决策实验中提取行为动力学。《神经元》109，597-610（2021年）。83 Wilson, R. C. & Collins, A. G. E. 行为数据计算建模的十条简单规则。《生命科学》8，e49547（2019年）。84 Bengio, Y. & Frasconi, P. 一种输入输出隐马尔可夫模型架构。载于《神经信息处理系统进展》7，427-434（1995年）。85 Rabiner, L. R. 隐马尔可夫模型及其在语音识别中若干应用的教程。《Proceedings of the IEEE》77，257-286（1989年）。86 Baum, L. E., Petrie, T., Soules, G. & Weiss, N. 对马尔可夫链概率函数进行统计分析时出现的一种最大化技术。《数学统计年鉴》41，164-171（1970年）。87 Dempster, A. P., Laird, N. M. & Rubin, D. B. 通过EM算法从不完整数据中进行最大似然估计。《皇家统计学会杂志：B辑》39，1-38（1977年）。88 Viterbi, A. 卷积码的误差界与渐近最优译码算法。《IEEE信息论汇刊》13，260-269（1967年）。89 Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. 黏性HDP-HMM及其在说话人日志分段中的应用。《应用统计年鉴》5，1020-1056（2011年）。90 Yu, S.-Z. 隐半马尔可夫模型。《Artificial Intelligence》174，215-243（2010年）。91 Sutton, R. S., Precup, D. & Singh, S. 介于马尔可夫决策过程和半马尔可夫决策过程之间：强化学习中时间抽象的框架。《Artificial Intelligence》112，181-211（1999年）。92 Kalman, R. E. 线性滤波与预测问题的新方法。《基础工程杂志》

Engineering 82, 35-45（1960）。

93 Skalse, J. M. V., Farrugia-Roberts, M., Russell, S., Abate, A. & Gleave, A. 策略优化中的不变性与奖励学习中的部分可识别性。载于《第40届国际机器学习会议论文集》32033-32058（2023年）。94 Levine, S., Kumar, A., Tucker, G. & Fu, J. 离线强化学习：教程、综述及开放问题展望。arXiv预印本 arXiv:2005.01643（2020年）。95 Wilkinson, M. D. 等。《科学数据管理与数据管护的FAIR指导原则》。Scientific Data 3, 160018（2016年）。96 Biswas, A. 等。一种动态贝叶斯优化的主动推荐系统，用于好奇心驱动、部分由人工参与的自动化实验。npj Computational Materials 10, 29（2024年）。DOI: 10.1038/s41524-023-01191-5。97 Bulanadi, R. 等。超越标量目标：由专家反馈驱动的自主实验，助力纳米尺度科学发现。ACS Nano 20, 27774–27783（2026年）。https://doi.org/10.1021/acsnano.6c09449。98 Burke, M. 等。利用概率时间排序，从探索性示范中学习奖励。Autonomous Robots 47, 733–751（2023年）。DOI: 10.1007/s10514-023-10120-w。99 Guo, W., Agrawal, K. K., Grover, A., Muthukumar, V. K. & Pananjady, A. 从探索型示范者中学习：多臂老虎机的最优奖励估计。机器学习研究会议录 151, 6357–6386（2022年）。100 Pedregosa, F. 等。Scikit-learn：Python中的机器学习。机器学习研究期刊 12, 2825-2830（2011年）。101 Virtanen, P. 等。SciPy 1.0：用于科学计算的Python基础算法。Nature Methods 17, 261-272（2020年）。102 Paszke, A. 等。PyTorch：一种命令式、高性能深度学习库。载于《神经信息处理系统进展》第32卷，8024-8035（2019年）。103 Storn, R. & Price, K. 差分进化：一种简单高效的启发式算法，用于连续空间中的全局优化。Journal of Global Optimization 11, 341-359（1997年）。104 Rezende, D. J., Mohamed, S. & Wierstra, D. 深度生成模型中的随机反向传播与近似推断。第31届国际机器学习会议论文集，PMLR 32(2), 1278–1286（2014年）。https://proceedings.mlr.press/v32/rezende14.html。105 Kalinin, S. V., Steffes, J. J., Liu, Y., Huey, B. D. & Ziatdinov, M. A. 通过无监督机器学习，在动态压电响应力显微镜中解析铁电畴壁的几何形态与演化路径。Nanotechnology 33, 055707（2021年）。https://doi.org/10.1088/1361- 6528/ac2f5b。

106 Valleti, M., Ziatdinov, M., Liu, Y. & Kalinin, S. V. 从简约表示中揭示物理学与化学：通过不变变分自编码器进行图像分析。npj Computational Materials 10, 183（2024年）。https://doi.org/10.1038/s41524-024-01250-5. 107 Liu, Y., Vasudevan, R. K., Kelley, K. P., Funakubo, H., Ziatdinov, M. & Kalinin, S. V. 在多模态成像中学习正确的通道：压电响应力显微镜中的自动化实验。npj Computational Materials 9, 34（2023年）。https://doi.org/10.1038/s41524-023-00985- x. 108 Roccapriore, K. M., Kalinin, S. V. & Ziatdinov, M. 通过扫描透射电子显微术中的自主实验发现纳米等离激元系统中的物理学。《先进科学》9, 2203422（2022年）。https://doi.org/10.1002/advs.202203422. 109 Slautin, B. N., Pratiush, U., Lupascu, D. C., Ziatdinov, M. A. & Kalinin, S. V. 通过深度核学习–变分自编码器模型的潜在空间设计，整合预测与生成能力。npj Computational Materials 12, 231（2026年）。https://doi.org/10.1038/s41524-026-02197-5. 110 Valleti, M., Vasudevan, R. K., Ziatdinov, M. A. & Kalinin, S. V. 深度核方法表现更优：从卡片到过程优化。《机器学习：科学与技术》5, 015012（2024年）。https://doi.org/10.1088/2632-2153/ad1a4f.