arXiv:2611.32990v1 [q-bio.QM] 24 Nov 2026

# 细胞和细胞核形状可预测受限运动

Malcolm Mehrabian-Terlexis, [1], * Jacqueline Janssen, [1], * Paul Sitoleux, [1], 2, * Thierry Mora, [1], † Aleksandra M. Walczak, [1], † 和 C´ecile Sykes[1], †

[1] *Laboratoire de Physique de l'École Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France*

[2] *Laboratoire Jean Perrin, Sorbonne Université, CNRS, F-75005 Paris France*（日期: 11月 24th 2026）

癌症转移依赖于细胞增强的迁移能力，使其即使穿过小于自身细胞核的狭窄通道，也能高效迁移。细胞和细胞核形态都可能影响这一过程。然而，这些形态与细胞穿越狭窄通道可能性之间的关系仍不清楚。在此，我们将微流控实验与机器学习相结合，以预测MDA-MB-231乳腺癌细胞是否能够穿越狭窄通道。细胞在垂直受限的方框内迁移，并可通过穿越狭窄通道进入相邻方框，这一过程需要细胞核发生变形。我们定义了一组源自细胞和细胞核轮廓的12个几何特征，发现这些特征的平均值在统计学上并不能表明细胞是否会穿越狭窄通道。为克服这一局限，我们利用轨迹训练分类器，并评估其预测未来穿越事件的能力。我们发现，细胞面积和细胞圆度携带了大部分预测信息，且圆度较低的细胞更有可能穿越狭窄通道。在不重新训练的情况下，这些模型还能够预测细胞穿越更窄的狭窄通道以及在黏附性更强的基底上的穿越情况。穿越事件发生前约30分钟的细胞形态特征，对于预测随后是否发生穿越事件最具信息量。总之，我们的结果表明，从简单的轮廓描述符中读取的细胞形状包含一种可迁移但依赖条件的、预示即将发生穿越的特征。

## 引言

癌症的严重程度与转移密切相关。在转移过程中，癌细胞会朝血管迁移，并通过挤压穿过受限空间，侵入其他器官 [1, 2]。这一过程已在上皮-间质转化的背景下得到广泛研究；在此过程中，细胞的表型会从静止的上皮状态转变为运动性极强的间质状态 [3, 4]。微流控技术的最新进展使研究人员能够利用精确设计的几何结构，在复杂环境中研究单细胞运动 [5, 6]，并实现对几何结构的精确控制 [7–10]。此外，微流控研究表明，侵袭过程中细胞核的变形可能是一个限速步骤 [11, 12]。尽管细胞和细胞核形态是影响侵袭动力学的重要变量，但它们能否预测受限迁移效率仍不明确。

根据局部几何结构，迁移细胞可以采用多种运动策略。间充质型迁移涉及形成富含肌动蛋白的突起，呈片状伪足或丝状伪足的形态，并通过黏着斑与周围环境结合 [13]。另一种策略是变形虫式运动，细胞与周围环境的黏附较弱，并通过肌动蛋白-肌球蛋白收缩性产生称为膜泡的突起，以此移动 [14, 15]。单个癌细胞通常会同时采用这两种策略，或在两者之间交替切换 [16, 17]。与间充质表型相比，变形虫式细胞的形状更圆。这些基于细胞骨架动态组织的迁移模式在不同物种间高度保守，无论其是否具有细胞核（真核细胞） [18]，还是不具有细胞核（原核细胞）。虽然细胞形态能够很好地指示细胞迁移所采用的策略，但细胞在受限空间中迁移时，细胞核所起的作用可能并不像最初认为的那样被动。事实上，细胞核在基因调控中发挥重要作用，也会影响生物力学 [12, 19, 20]。作为细胞中最坚硬的细胞器，细胞核在细胞穿过受限空间时必须发生变形，直接影响迁移动力学 [21, 22]。细胞核可塑性可能使中性粒细胞等运动能力很强的健康细胞能够穿行于各种空间 [23]。同样，转移性癌细胞的细胞核也可能具有很强的可变形性，从而提高其在受限条件下的迁移效率 [24]。因此，细胞形态和细胞核形态有望成为反映细胞状态的重要信息来源，因而是深入理解受限条件下运动机制的基础。尽管已有若干研究采用一些形态描述符对细胞迁移进行理论建模 [25–28], ，但要全面描述

细胞和细胞核轮廓动力学可能是判断细胞通过狭窄处能力的信息来源。然而，这种方法需要大量难以通过实验测量的参数。

机器学习方法非常适合处理复杂的多变量数据，以执行预测任务。机器学习已广泛应用于基因组数据，根据转录组学数据 [29],和体外实验 [30, 31]预测转移。最近，机器学习已被用于

* 这些作者作出了同等贡献（共同第一作者）。
† 这些作者作出了同等贡献（共同末位作者）。

根据原始单细胞轨迹 [32, 33], 对细胞迁移类型进行分类, 以及根据轮廓衍生特征对在细胞外基质中迁移的细胞形态类型进行分类 [34]。在另一项研究中，研究人员利用细胞和线粒体的荧光图像提取特征，以预测细胞在 5μ 米的垂直限域下迁移时的速度 [35]。在此，我们观察了在一种微流控装置中运动的细胞及其细胞核的形态；该装置由通过狭窄通道连接的扁平方框构成。研究人员选取了一组特征，通过细胞及细胞核的轮廓来描述其形态。随后利用这组数据训练机器学习模型，该模型能够成功预测穿越概率。该模型仅基于有限数量的特征，其中细胞圆度和面积是预测能力最强的特征。该模型成功迁移至其他实验条件：一方面是更狭窄的通道，另一方面是在微流控装置表面涂覆纤连蛋白以增强细胞黏附的条件。我们进一步扩展该模型，使其能够预测某个细胞通过的时间，并揭示模型预测能力最强的时间尺度为 30 分钟。

## 结果

### 轨迹与通过

我们使用细胞核经 Hoechst-33342染料染色的 MDA-MB-231细胞。细胞在微流控装置内高为 5μm 的 PDMS-玻璃腔室中自发迁移。腔室由十字阵列构成，形成边长为 $50 \times 50\mu\mathrm{m}$ 的$^2$ 方形盒，方形盒之间由宽度为 15μm 或 5μm 的狭窄通道分隔（**图 1(a-d)**）。细胞在方形盒内自发迁移，并可能通过这些狭窄通道从一个方形盒移动到相邻方形盒。细胞核半径平均为 7.5μm。因此，与通过 5μm 狭窄通道的细胞核相比，通过 15μm 狭窄通道的细胞核变形程度更大（示例见**图 1(e,f)**的插图）。图像采集持续 22 小时，每 5 分钟采集一张图像。我们使用经过专门训练的分割模型，分别从明场图像和荧光图像中获取细胞和细胞核轮廓。

为判断细胞是否会从一个方框通过到相邻方框，我们使用细胞核的位置作为替代指标。细胞核的位置由其质心（COM）定义：$\mathbf{r}_{\mathrm{nuc}}(t) = (X_{\mathrm{nuc}}(t)\ Y_{\mathrm{nuc}}(t))$ 在由细胞核轮廓界定的表面积上计算，并对整个表面赋予相同权重。我们记录细胞核质心的轨迹，以下称为“轨迹”。若发生以下任一事件，轨迹便会截断：(1) 细胞间接触；(2) 实验结束（22 小时后）；(3) 细胞离开视野；(4) 细胞死亡开始；(5) 细胞开始分裂。

鉴于方框具有正方形对称性，我们将坐标系的原点定义为方框中心即任意给定细胞在开始观察时所处的位置。我们认为所有轨迹均从同一个初始方框开始（**图 1(e,f)**）。通过时间定义为细胞核质心离开初始方框的时刻，且只有当细胞核质心连续至少 15 分钟处于方框外时，才将其计为一次有效通过。通过时间的累积概率分布绘于**图 1(g)**。我们观察到，在给定时间内， 15μm 狭窄通道中的穿越概率高于 5μm 狭窄通道。此外，细胞更容易通过较大的 15μm 狭窄通道：在 102 条轨迹中发生了 56 次穿越事件，而在 5μm 狭窄通道中，131 条轨迹中仅发生了 30 次。

### 定义几何特征

以下基于形状的分析需要同时提取细胞和细胞核轮廓，因此必须对每条轨迹的每一帧进行分割和人工校订。因此，我们仅对每种条件下进行的四次实验中的两次进行分析；这两次实验的选择不受穿越结果影响，且均提取了细胞和细胞核轮廓。该子集包含 15μm 条件下的 56 条轨迹（31 次穿越）和 5μm 条件下的 59 条轨迹（8 次穿越）。从此处开始的所有分析均针对这一数据子集。

为评估哪些因素决定细胞是否可能通过，我们定义了 12 个特征，用以反映细胞和细胞核形态，以及它们的相对位置和与运动方向的取向一致性。每个特征均根据细胞和细胞核轮廓获得，并沿轨迹按时间逐个快照计算。用于定义这些特征的几何符号示意图见（**图 2(a)**）。用于描述细胞和细胞核形态的 12 个特征如下。描述细胞和细胞核形状时，我们采用细胞面积、细胞核面积、细胞周长、细胞核周长、细胞圆度、细胞核圆度、细胞长宽比和细胞核长宽比，其数学表达式见**表 1**。为描述细胞核与细胞质心的相对位置，以及它们轨迹的方向，我们将间距 定义为细胞质心与细胞核质心之间的距离，即向量$\mathbf{s}$ 的模（**图 2(a)**）。点积是分离向量$\mathbf{s}$ 与细胞核速度$\mathbf{v}$的标量积；取向一致性 是分离向量$\mathbf{s}$ 与细胞核速度$\mathbf{v}$所成角的余弦。此外，前导投影 是速度在分离向量$\mathbf{s}$上的投影。

### 特征值的分布

对于 5 或 15μm 限制通道条件下轨迹的每个快照，都会计算每个特征的值，并将其归类为属于通过轨迹或未通过轨迹。4 个特

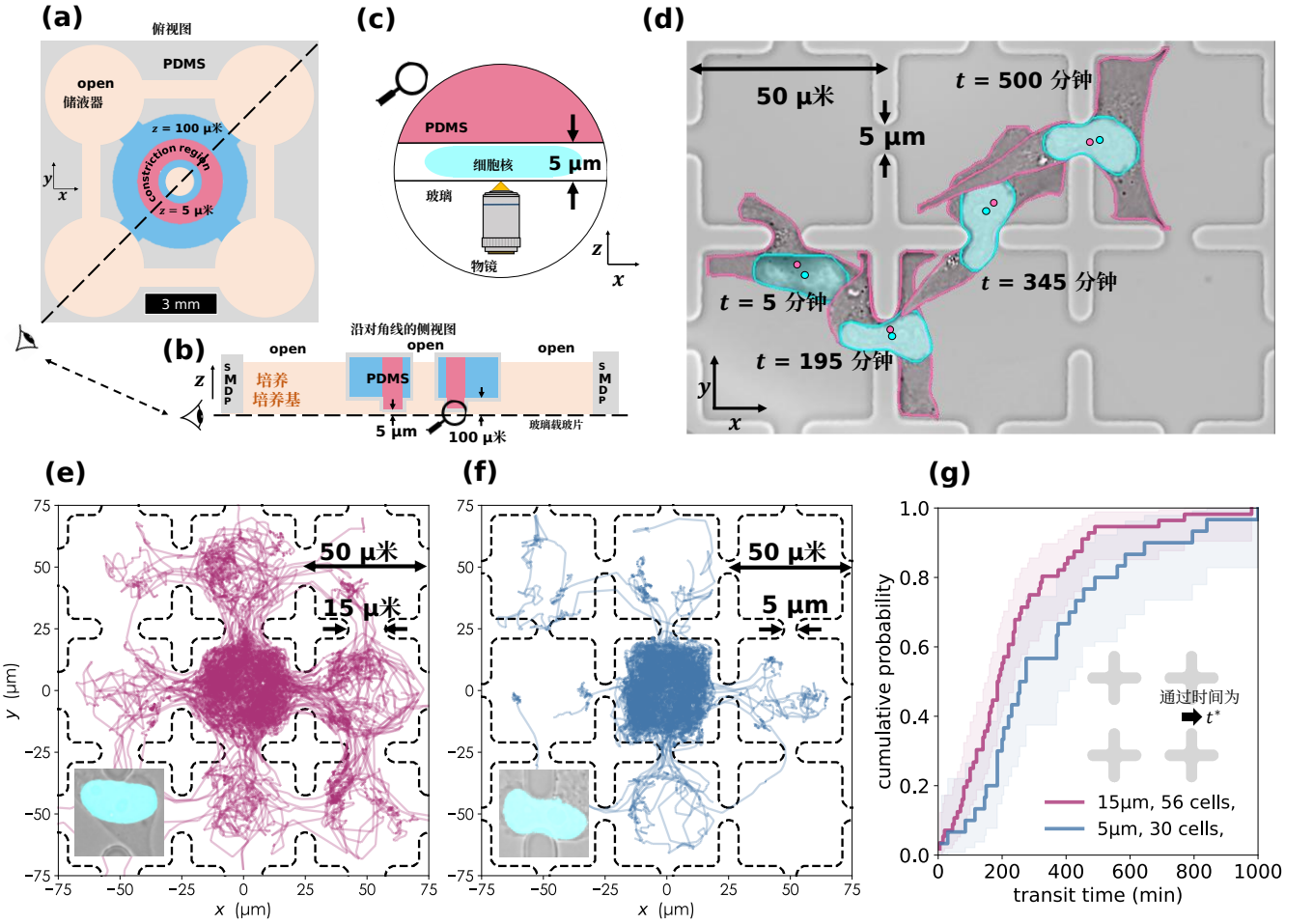


图 1. **实验装置、轨迹和通过时间 (a)**PDMS-玻璃迁移装置的俯视比例示意图。灰色区域表示 PDMS 与玻璃粘合，橙色的开放储液器处于开放状态，培养基可流入。蓝色区域的高度为 $z = 100$μ米，细胞在粉色狭窄区域迁移，其中 $z = 5$μ米。(b) 侧视示意图，未按 $z$ 轴比例绘制。(c) 侧视近景示意图（未按比例绘制）。在狭窄区域中，细胞核在 PDMS 与玻璃之间的 5μ米空间内发生变形。(d) MDA-MB-231 细胞在狭窄区域迁移的示例（明场）。细胞通过 5μ米狭窄通道，在 50μ米和 ×50μ米方框之间迁移。图中叠加了经 Hoechst-33342 染色的细胞核分割图像。细胞和细胞核轮廓分别以粉色和青色绘出，同时标出了各自的质心。(e) 15μ米狭窄图案中的轨迹（102 个细胞），以及(f) 5μ米狭窄图案中的轨迹（131 个细胞）。叠加了细胞核荧光图像的明场图像显示，细胞核在 5μ米狭窄通道中会变形，而在 15μ米狭窄通道中则不会。(g) 细胞在 15 米和 5μ米狭窄图案中的首次通过时间累积概率分布。通过时间是指细胞核质心离开初始方框的时刻。只有当细胞核质心至少 15 分钟持续处于初始方框之外时，该通过时间才有效。

征在每种实验条件下所有轨迹中的值分布示例见**图 2(b-e)**。完整的特征值分布见补充信息（**图 S1**）。在所有限制通道条件下，无论轨迹是否通过，特征值都呈现广泛分布。对于大多数特征，通过轨迹和未通过轨迹的分布在统计学上无法区分。仅有的例外是 15μm 限制通道条件下的细胞圆度，以及 5μm 限制通道条件下的细胞周长；在这两种情况下，通过细胞与未通过细胞之间均观察到显著差异（**图 2(b)**和**图 S1(a)**）。在缺乏明确

尽管大多数特征与通过可能性的相关性较弱，但形态特征和动态特征之间复杂的相互相关性中，可能蕴含着有用的信息。为进一步探究，我们转而采用机器学习方法。

### 基线特征和动力学特征分类任务

除上述 12 项基础特征 外，我们还为每条轨迹中以 5 分钟为间隔采集的每个快照计算一组

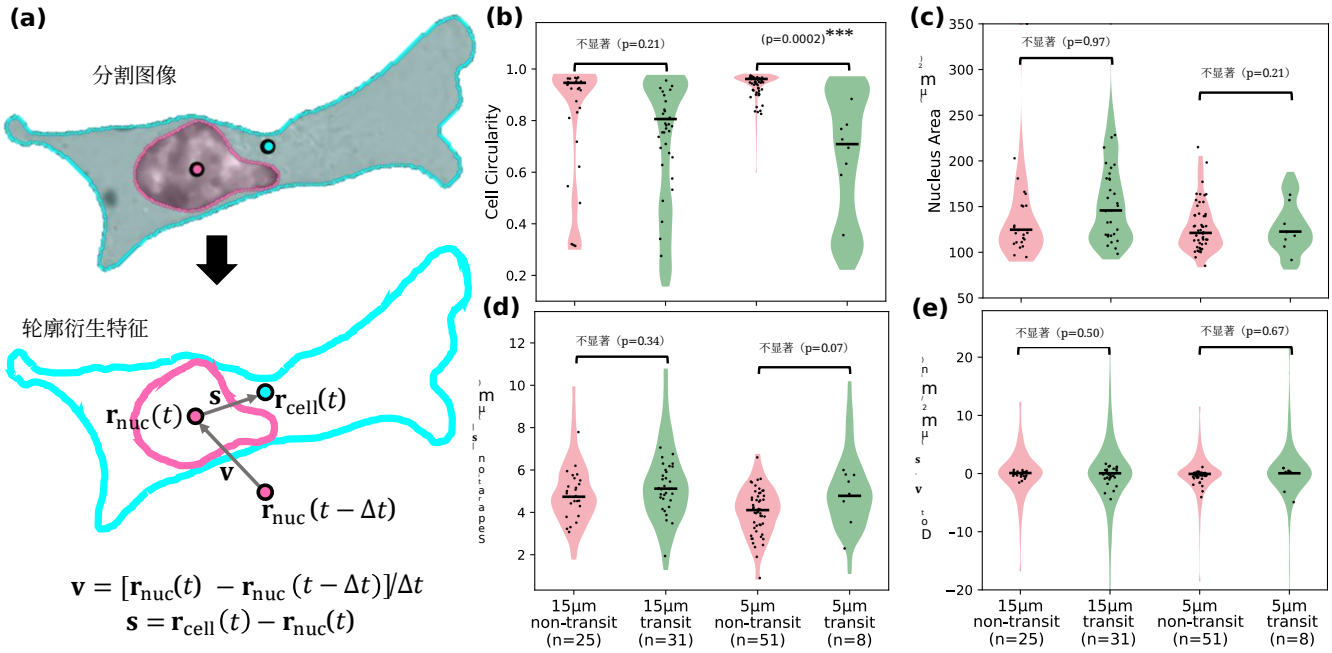


图2。**轮廓派生特征及其分布示例（a）**用于获取细胞（青色）和细胞核（洋红色）轮廓的分割图像示意图，并标示细胞和细胞核质心、速度向量$\mathbf{v}$ 以及细胞-细胞核分离向量$\mathbf{s}$。**（b-e）**在15和 5μμm狭窄图案中，通过狭窄通道的（通过，红色）和未通过的（未通过，绿色）细胞每个快照的轮廓衍生特征分布。**（b）** *cell circularity*，**（c）** *nucleus area* **（d）***separation* 定义为 $|\mathbf{s}|$，以及**（e）** *dot* 特征，定义为$\mathbf{v}\cdot\mathbf{s}$。红色和绿色小提琴图面积表示所有时间帧快照中的数据，而黑色点表示各条单独轨迹的时间平均值。统计显著性基于对时间平均值进行曼-惠特尼U检验所得的 $p$ 值报告，以避免跟踪时间较长的细胞使数据产生偏差。

动力学特征，其计算方式为基础特征的后向有限差分。此外，我们还纳入了细胞和细胞核速度。总共，我们收集了 26 项特征，并用它们预测轨迹是否以及何时穿越狭窄通道。轨迹从测量开始延续至结束，有时会包含多次穿越连续狭窄通道的过程。我们将轨迹截取为子轨迹，每条子轨迹均由一次穿越狭窄通道的过程定义。随后，子轨迹包含转变前从 $t_{\min} = 5$ 到 $t_{\max} = 120$ 分钟的快照，且彼此不重叠。我们通过调整时间窗口，并考察通过类别与未通过类别的类别平衡（使数据集中两个类别的样本数量相等），来确定这一选择：选择更大的 $t_{\max}$ 会纳入更多穿越事件，并扰乱类别平衡。不过，我们发现，不同的时间窗口选择不会对结果产生定性影响。实验中所有轨迹对应的子轨迹集合，就是我们所说的数据集。

为了解哪些特征决定通过狭窄通道的概率，我们定义了一个二元分类任务：根据单个快照预测一条轨迹是否会在之后发生穿越事件。利用在 15μm 狭窄通道条件下获得的全部特征值，我们训练一个支持向量分类器，并采用径向基函数核（SVC-RBF）来捕捉数据中的潜在非线性关系，同时将其性能与逻辑回归（LR）进行比较。

### 训练和特征选择设置

模型采用精确率-召回率曲线下面积（PR-AUC）进行评估。由于 5μm 轨迹中穿越事件数量较少（对于我们选定的区间和筛选流程，穿越事件仅占通过轨迹的 ∼ 14%），因此我们仅使用15μm 狭窄通道条件下的数据集进行训练（穿越轨迹占 ∼ 55%）。设置方案如**图3(a)**所示。

为了确定预测能力最强的特征，我们比较使用不同特征集训练的模型性能。模型选择需要将数据集划分为训练集（轨迹的50%）、验证集（25%，用于比较模型性能），以及测试集（25%，用于在未用于训练或模型比较的留出数据上评估最佳模型的性能）。SVC-RBF 和 LR 模型还具有用于避免过拟合的正则化参数（有关其设置方法，请参见方法和**表3**）。

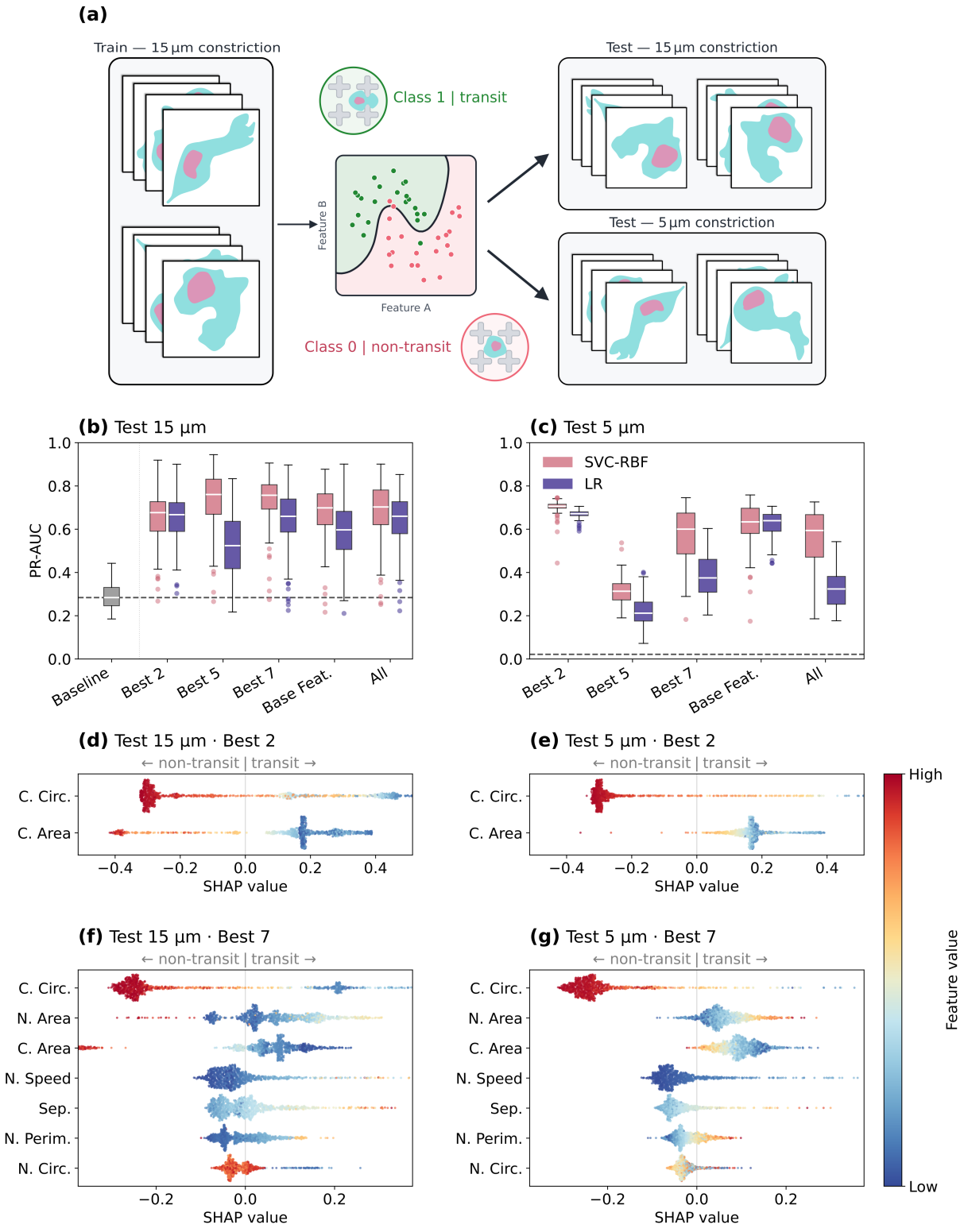


图 3. **分类训练的设置与结果。(a)** 训练设置：我们首先使用 15µm 狭窄通道中的轨迹进行训练（细胞为蓝色，细胞核为粉色），并在 15µm 数据以及 5µm 狭窄通道数据上进行测试（下图）。**(b)** 15µm 测试集的精确率-召回率曲线下面积箱线图，比较四种不同特征集（**表 1**）。比较采用径向基函数核的支持向量分类（SVC-RBF，粉色箱体）与逻辑回归（LR，紫色箱体）。基线（灰色箱体、虚线）为随机分类器在 15µm 测试集上的结果。**(c)** 迁移学习在 5µm 测试集上的精确率-召回率曲线下面积箱线图，比较四种不同特征集（**表 1**）。比较 SVC-RBF（粉色箱体）与 LR（紫色箱体）的结果。基线（灰色虚线）为随机分类器在测试集上的结果。精确率-召回率曲线下面积的分布源于在**(b)**中训练的模型结果分布。 *best 2* 特征集的沙普利值蜂群图：**(d)** 15µm 狭窄通道、**(e)** 5µm 狭窄通道数据。 *best 7* 特征集的沙普利值蜂群图：**(f)** 15µm 狭窄通道、**(g)** 5µm 狭窄通道数据。色彩映射表示特征值，红色为高、蓝色为低（在第 5 至第 95 百分位数之间线性插值）。两个测试集中，较高的细胞圆度都预示着非穿越轨迹，而较小的细胞面积则是轨迹会穿越的指标。

模型性能通过训练集和验证集（合计占 75%）的 3 折平均精确率-召回率曲线下面积进行评估。待比较的模型空间涵盖所有可能的特征集，因此规模非常庞大。我们采用贪心搜索寻找预测能力最强的特征集。首先随机生成一个特征随机排序列表，然后从列表顶部开始逐个添加特征，直到添加某个特征后精确率-召回率曲线下面积的提升不足 0.01。我们对许多种随机排序列表重复这一过程。随后按特征集大小对模型分组，并根据模型在验证折上的平均性能进行排序。为避免折特异性偏差，我们通过在 3 折划分的 5 种随机选择上重新计算精确率-召回率曲线下面积，对每种特征集大小排名前三的模型进行进一步评估。

通过贪心搜索，我们得到三种候选特征集，分别称为最佳 *2*、最佳 *5* 和最佳 *7* 特征。最佳 *2* 特征集由细胞圆度 和细胞面积组成。这些特征也包含在最佳 *7* 特征集中，此外还有间距、细胞核速度、细胞核圆度、细胞核面积 和细胞核周长。最佳 *5*特征包括细胞长宽比、细胞纵横比变化速率、细胞圆度、细胞核面积 和细胞核速度范数。请注意，细胞核速度 和细胞纵横比变化速率 是我们的搜索程序为这些特征集选出的仅有的动力学特征。

### 使用 $15\mu m$ 狭窄通道数据获得的模型测试

测试集精确率-召回率曲线下面积会因数据划分方式不同而有显著差异。因此，我们在 100 次随机训练/测试划分上重复评估，并以箱线图展示由此得到的测试集精确率-召回率曲线下面积分布。每个箱体的跨度反映了不同划分之间的变异性，**图3(b)**。

对于 15μ 毫米数据的三种特征集，最佳 *5* 和 最佳 *7* 特征的 SVC-RBF 精确率-召回率曲线下面积平均值最高，最佳 *5* 特征为 0.731，最佳 *7* 特征为 0.733，置信区间 (CI) 分别为 $[0.56, 0.91]$ 和 $[0.59, 0.88]$ 95%，**表S1**。对于 SVC-RBF 和 LR，最佳 *7* 特征的表现优于包含 26 个特征的完整特征集（全部），该完整特征集包括动力学特征和基础特征（各特征集概览见**表2** ）。所有模型的性能中位数仍比基线高 30%，见**表S1**。

与逻辑回归相比，径向基函数支持向量分类器的平均性能更佳，但在最佳 *2* 特征集上二者表现相当。对于所有模型，我们发现，不同特征集都包含有关未来通过狭窄通道的潜在穿越的重要信息；即使采用简单的线性逻辑回归模型，最佳 *2* 特征集也已捕捉到有关穿越子集与未通过子集的相关信号（**图3(b)**）。尽管贪心搜索算法将特征集扩展到了七个特征，但排名最前的两个特征——细胞圆度和细胞面积——捕捉到了穿越事件的主要信息。

### 预测 $5\mu m$ 狭窄通道中轨迹的穿越事件

我们发现，将在 15μm 条件下训练的模型应用于 5μm 条件时，所有模型的结果均优于基线（**图 3(c)**，随机基线为 0.02），这证实了模型成功实现了泛化。由于与 15μm 数据集相比，这种狭窄通道条件下的穿越事件较少，因此基线值很低。我们注意到，与其他特征集相比，最佳 *5* 特征集的表现显著下降，其 SVC-RBF 和逻辑回归的均值表现均低于其他所有特征/模型组合（见**表 S1**）。**表S2** 中给出的模型间精确率-召回率曲线下面积配对差值表明，最佳 *2* 特征集的表现显著优于最佳 *5* 特征集（从置信区间来看；均值差距为 $+0.39$）。总体而言，对于大多数特征集，支持向量分类器的精确率-召回率曲线下面积均值高于逻辑回归，这表明核函数的非线性对于描述 5μm 数据集十分重要。我们得出结论，对于 5μm 狭窄通道数据，最佳 *2* 特征集表现最佳。

我们注意到，在测试 15μm 数据上评估的精确率-召回率曲线下面积具有较高的方差（可从箱线图的分布范围和误差条中看出），而在 5μm 测试数据上评估时，误差条明显更小。**图 3(b, c)** 中的箱线图在分布范围上存在差异，原因有二：训练集和测试集的抽样不同，以及训练后的模型之间存在差异。在5μm 狭窄通道条件下，我们通过让使用不同划分进行训练的模型共享同一个测试集，来区分这两种影响。**图 3(c)** 中的精确率-召回率曲线下面积的剩余变化完全源于训练后的模型之间的差异。

这些结果使我们得出结论：细胞圆度和细胞面积 能够很好地泛化到两种狭窄通道尺寸。此外，最佳 *5* 特征的性能下降表明，其中一些未包含在最佳 *2* 和最佳 *7* 特征集中的特征（例如细胞长宽比）在 5 μm 狭窄通道设置下与在 15 μm 设置下的表现不同。

### 使用 Shapley值评估特征重要性

我们现在知道哪些特征集可以预测转移概率，但还不知道各个特征值如何影响预测。为此，我们使用 SHAP 软件包 [36, 37], 计算 Shapley值，以评估各个特征及其取值（高或低）对分类任务（预测轨迹通过还是未通过）的影响。Shapley值评估某个特征对穿越轨迹预测概率的平均影响。我们将每个 Shapley值

与相应的特征值联系起来（色彩映射：高 =红色，低 = 蓝色），以分析 15μm 和 5μm 狭窄通道配置（**图 3(d-g)**）。每个点代表测试集中的一个快照，其水平位置表示该快照的特征预测方向：右侧表示预测倾向于通过，左侧表示倾向于未通过，0 表示没有影响。

对于 15μm狭窄通道设置，细胞圆度和面积 对于判断细胞是否可能通过最具信息量，无论是最佳 *2*（**图 3(d)**）还是最佳 *7* 特征集（**图 3(f)**）均是如此。圆形的大细胞比细长的小细胞更不可能通过。此外，与非常大的细胞类似，非常大的细胞核也不太可能通过，而较小的细胞和细胞核更可能通过。然而，细胞核面积 与通过之间的关系是非单调的，中等大小的细胞核也很可能通过。较低的细胞核速度、较小的细胞核周长、较高的细胞核圆度 以及较低的间距 与非穿越细胞仅呈弱关联。然而，较大的细胞核速度 和较大的间距 值与穿越细胞强烈相关。总体而言，该模型预测，具有中等大小且快速移动细胞核的细长小细胞很可能通过。圆形的大细胞若细胞核移动缓慢，则表明其为未通过。

对于 5μm 狭窄通道设置，最佳 *2* 和最佳 *7* 特征集对特征影响细胞行为的排序方式基本相同于 15μm设置，这表明模型使用相同的逻辑来区分类别（**图 3(e, g)**）。最大的差异出现在对细胞核运动速度、面积 和细胞核周长 的预测模式上（**图 3(g)**）。对于 5μm 数据，细胞核速度 始终较低，而细胞核面积 和周长 呈单调变化，这与 15μm 的结果相反。对于 5μm 数据的狭窄通道设置，大细胞核预示着通过；而对于15μm设置，中等大小的细胞核预示着通过，但较大的细胞核则强烈预示着未通过。在**图 S2** 中，我们展示了特征值的分布，这证实了：在 15μm 数据中（以及下文将讨论的 15μm纤连蛋白数据中），较高的细胞核速度 与通过的轨迹相关；而在 5μm 数据中，几乎所有通过的轨迹都具有较低的速度。这些结果证实，正如我们在讨论（**图 3(c)**）中的精确率-召回率曲线下面积时所指出的，细胞核运动速度特征无法在两种限制通道条件之间泛化。

### 预测纤连蛋白包被的 15$\mu m$宽狭窄通道中的穿越事件

纤连蛋白是一种促进细胞与周围环境黏附的蛋白质，长期以来一直被研究其在癌症转移和肿瘤发生中的作用[38, 39]。黏附细胞以间充质模式迁移，因此形态更为细长 [40]。为了了解这些模型能否适用于增强黏附并诱发形态变化的条件，我们还在 15μm 狭窄通道条件下，对纤连蛋白包被的微流控装置中的轨迹进行模型测试。与未包被的基底相比，纤连蛋白包被的狭窄通道中的细胞具有更大的细胞面积和更低的圆度 **图 4(b,c)**。

所有模型的表现均优于基线（0.42），其中最佳 *2* 特征集的 SVC-RBF/逻辑回归均值精确率-召回率曲线下面积最高，分别为 0.70/0.71，$[0.45, 0.90]$/$[0.47, 0.90]$ 95% 置信区间（**图 4(d)**）。尽管较大特征集的置信区间与基线重叠，但其均值均高于 0.53，这证实了在未处理基底上训练的模型能够泛化到纤连蛋白数据。有趣的是，对于所有特征集，逻辑回归的表现都优于支持向量分类器，这表明非线性核函数无助于区分类别，反而可能导致过拟合。与 5μm 狭窄通道数据的结果类似，我们再次发现，最佳 *5* 特征集的平均表现不如其他特征集。这一观察结果表明，某些单独特征，尤其是细胞长宽比，可能无法很好地泛化到不同的狭窄通道和黏附条件。

对最佳 *2* 和最佳 *7* 特征集的 Shapley值分析显示，其结果与未包被实验及 5μm 数据中观察到的结果相似（**图 4(e, f )**）。总体而言，模型将细胞细长且细胞核较小的细胞判定为穿越细胞，尤其是在间距 和细胞核速度 较高时（**图 4(f)**）。我们观察到，细胞核速度的表现与使用 15μm 训练的模型相同。尽管黏附增强会显著影响各项特征的数值，包括细胞和细胞核的几何形态与动力学（**图 S2**），但使用 15μm 训练的模型仍保有预测能力，并能成功泛化至增强黏附数据。这些结果表明，穿越概率取决于细胞和细胞核几何形态与动力学之间复杂的非线性关系。

### 通过时间的预测

确定了能够区分穿越轨迹与未穿越轨迹的特征后，我们现在评估能否预测轨迹通过狭窄通道的时间。接下来，我们将从最佳 *7* 特征集开始分析，其他所有特征集的结果见**图 S3**。

我们将通过时间任务重新表述为一组由分割点$t_c$参数化的二元分类任务：对于在时间 $t$记录的快照，如果其所在轨迹在时间 $t^* \geq t$发生通过，则 $\Delta t = t^* - t$ 表示距离通过的剩余时间。如果 $0 \leq \Delta t \leq t_c$，则将该快照标记为近通过；如果 $\Delta t > t_c$，则标记为远凌日。我们考察在 $t_c$取何值时，轨迹的特征对通过时间的预测能力最强。我们将分割点 $t_c$ 在$[5, 120]$ 分钟之间变化（步长为5分钟），并使用前一次二元分类分析中的特征集训练SVC-RBF和逻辑回归分类器（见方法）。

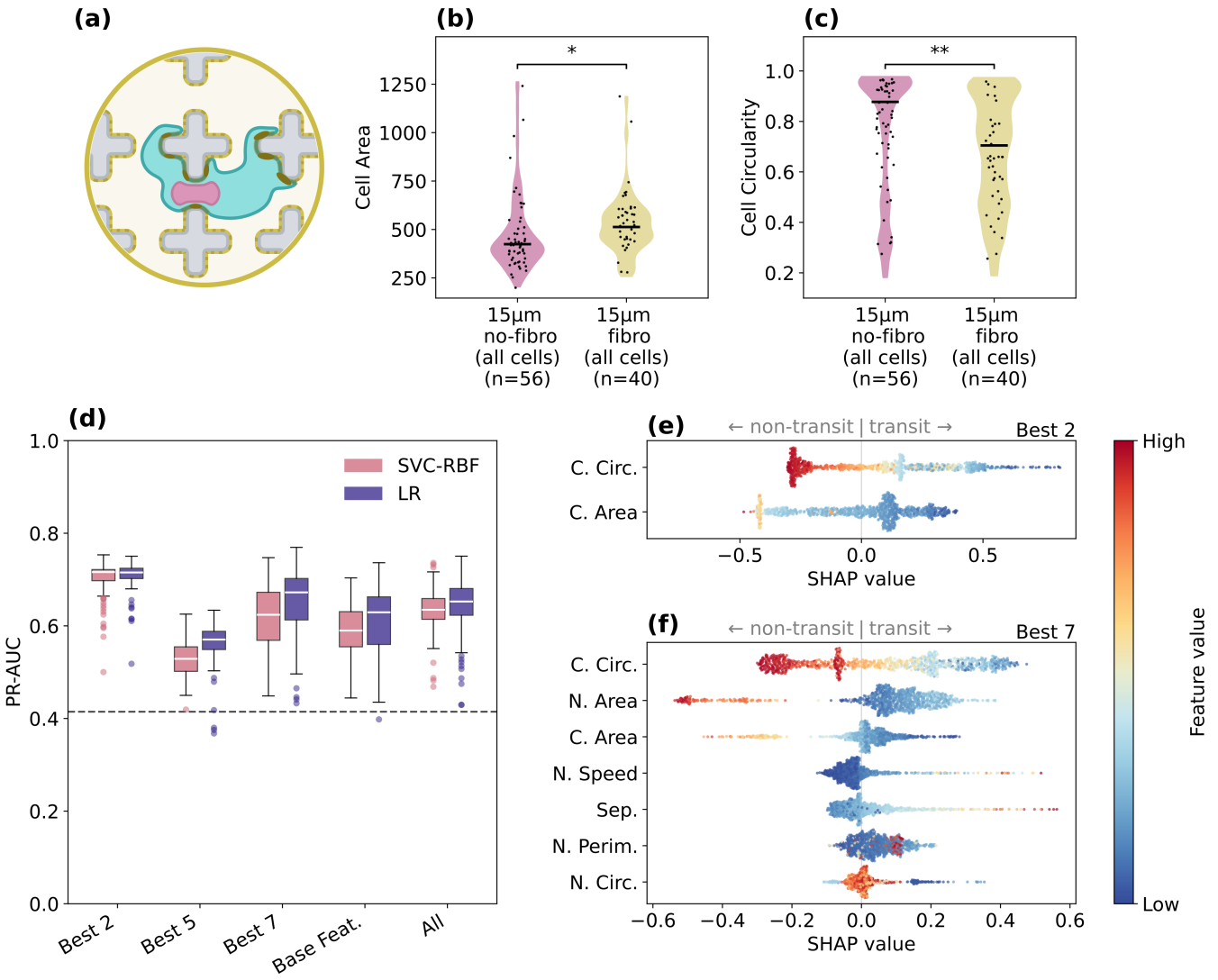


图4。**在增强黏附条件（纤连蛋白包被）下，通过** 15µ米**狭窄通道的轨迹测试集上的预测结果**。（**a**）纤连蛋白包被会导致轨迹更易黏附于狭窄通道的晶格柱上。（**b**）比较未包被晶格与纤连蛋白包被晶格中轨迹的 *cell area* 和（**c**） *cell circularity* 的小提琴图。统计显著性根据对时间平均值进行曼-惠特尼U检验所得的 $p$ 值报告。（**d**）在 15µ米纤连蛋白测试集上，针对分布外预测，采用五种不同特征集（表1）时的精确率-召回率曲线下面积箱线图。比较采用径向基函数核的支持向量分类（SVC-RBF，粉色）与逻辑回归（LR，紫色）的结果。（**e**）针对 15µ米狭窄通道和纤连蛋白包被的 *best 2* 及（**f**） *best 7* 特征的Shapley值蜂群图。色彩映射表示特征值，红色表示高值，蓝色表示低值（在5 – 95百分位数之间进行线性插值）。

我们使用来自 15µ米狭窄通道的数据训练模型，仅将轨迹限定为发生通过的轨迹，其中一些轨迹包含多个穿越事件。为确保模型始终在留出数据上进行测试，我们使用发生通过的轨迹中的3折进行训练，并在剩余1折上进行测试（对应于75%/25%的训练/测试集划分）。这一过程在折之间随机选择分割点的情况下重复100次。此外，为补偿不同时间分割 $t_c$选择下近通过类别与远凌日类别之间的类别不平衡，每个训练观测的类别 $a$ 根据 $w_a = N_{\text{train}}/\ (2N_{\text{train},\ a})$ 进行加权，其中$a$ = 近通过 或 *distant*。特征值仅使用根据训练数据计算的统计量进行标准化，使其达到零均值和单位方差。

通过真实标签与预测标签之间的互信息 $I(y;\hat{y})$ 来量化近通过和远凌日类别之间的可辨识性，其中标签是分配给每个快照的类别，见**图 5(a)**。互信息受真实标签的熵 $H(y)$ 的限制，该熵反映了在没有任何信息的情况下猜测标签的难度。如果所有数据

都属于同一类别，则互信息为零；当两个类别出现的频率相等时，互信息最大。对于这两个类别（近通过和远凌日），$H(y) = -p\log_2 p - (1-p)\log_2(1-p)$，其中 $p$ 是近通过轨迹所占的比例，并且会随 $t_c$变化。当 $p = 0.5$ 为$t_c = 45$ 分钟时，$H(y)$达到 1 比特的最大值。对于这个 $t_c$，两个类别完全平衡（见**图 S4**）。有关互信息计算的详细信息，请参阅方法。

为将分类性能与类别平衡的影响区分开来，我们报告归一化互信息 $U = I(y;\hat{y})/H(y)$，**图 5(b)**。作为基线，我们使用一个零模型：对于每个 $t_c$，在将通过时间标记为近通过/远凌日之前，先随机打乱训练集中的通过时间。随后使用这些打乱后的标签重新训练模型，并根据测试细胞的真实标签对其预测结果进行评分。我们训练径向基函数支持向量分类器（蓝色）、逻辑回归（红色）以及零模型（虚线；分别以蓝色和红色表示这两种分类器）。**图 5(a,b)** 中的实线表示 100 次交叉验证重复的均值，阴影带表示最佳 7 特征集的标准差。

对于径向基函数支持向量分类器和逻辑回归这两种分类器，互信息及其归一化版本随 $t_c$ 呈非单调变化，并在 $t_c = 30$ 分钟时达到最大值。两种分类器得到的曲线几乎完全相同，且始终高于零模型。这些分类器在整个 $t_c$ 范围内都能识别出持续存在的转变前信号，其在 30 分钟左右最为显著。由于归一化互信息对类别平衡进行了校正，而且 30 分钟处的峰值不同于类别平衡达到最大值时的 $t_c \approx 45$ 分钟，因此我们得出结论：该最大值并非仅仅是类别平衡时标签熵较高所造成的假象。

对于其他特征集，归一化互信息在约 $t_c$ = 30-35 分钟时同样最高（**图 S3**）。这些结果表明，在穿越事件发生前的大部分两小时内，细胞和细胞核特征都携带有关通过时间的信息，其中最清晰的形态学特征很可能出现在通过前约半小时。

### 在增强黏附条件和更窄的狭窄通道上测试模型

与前述分析类似，我们在 5μm 和 15μm 纤连蛋白包被的狭窄通道中的穿越轨迹上，测试 **图 5(a)**（最佳 7）中的支持向量分类器-RBF 模型，以验证所学习的形态学特征能否推广到训练几何结构之外。我们不进行任何重新训练，直接将分类器应用于 5μm 的轨迹以及 15 μm 纤连蛋白包被的轨迹。两种条件下的全部轨迹均用于测试集，因此无需进行交叉验证。每种测试条件的特征均使用其自身的均值和标准差进行标准化。

在 5μm 和 15μm 纤连蛋白包被的狭窄通道上测试的 SVC-RBF 分类器（最佳 7）的结果如**图5(c-d)** 所示（黄色和绿色曲线分别代表相应结果，蓝色曲线代表在 15μm 条件下的 SVC-RBF，供参考）。与对 15μm 轨迹的分析类似，基线是由 15μm 训练集得到的零模型，并在 5μm 和 15μm 纤连蛋白测试集上评估（黄色和绿色虚线分别代表相应结果）。在 5μm 狭窄通道条件下，互信息波动，并在 $t_c = 40-70$分钟内降至零模型水平，表明在此范围内，预测结果不包含有关穿越事件的信息。相比之下，在 15μm 狭窄通道轨迹上训练的 SVC-RBF 分类器，在使用 15μm 纤连蛋白包被狭窄通道的所有可用数据进行测试时，表现出良好的泛化能力。对于所有$t_c$，归一化互信息均高于零模型水平，其变化趋势与未包被的15μm 数据相似，平均值也相近（0.016 vs 0.017）。归一化互信息没有单一明确的显著峰值，但在大部分 $t_c > 30$ 分钟内都保持较高水平。我们得出结论，最佳 7 特征集对于纤连蛋白包被狭窄通道中的轨迹，在穿越前半小时以上，包含最多的通过时间信息。总之，这些结果表明，在 15μm 狭窄通道轨迹上学习到的转变前信号可以迁移到 15μm 纤连蛋白包被数据，但无法迁移到 5μm 实验设置。

## 讨论

我们展示了一种微流控装置，可测量细胞核穿过狭窄通道时发生穿越事件的细胞的二维投影轨迹。这种几何设计的优势在于，细胞进入狭窄通道之前，可以在受限程度相对较低的区域测量细胞和细胞核轮廓。在我们定义的基础特征集中，只有细胞圆度 和细胞周长 在穿越与未穿越的轨迹之间呈现出统计上可区分的分布（见**图 2**和**图 S1**中的特征值分布）。

为将分析扩展到单个特征分布的比较之外，我们引入了一种结合多种细胞和细胞核特征的机器学习方法。机器学习分类器表明，单独来看，穿越细胞与非穿越细胞之间分布差异不显著的特征，例如细胞面积、间距 和细胞核速度，在组合使用时确实有助于预测；而且，对穿越的预测效果显著优于随机基准。此外，Shapley值会按重要性对特征排序，并揭示特征值及其非单调效应如何改变通过概率。对于 15μm狭窄通道数据，模型将细长的细胞以及体积较大、移动较快的细胞核视为穿越的信号。相比之下，圆形、体积较大的细胞若其细胞核移动缓慢，则表明细胞未穿越。

圆形形态是变形虫式迁移模式的特征，而伸长则是间充质迁移的特征，在这种迁移中，细胞与基底的黏附更强。尽管已有研究报告称

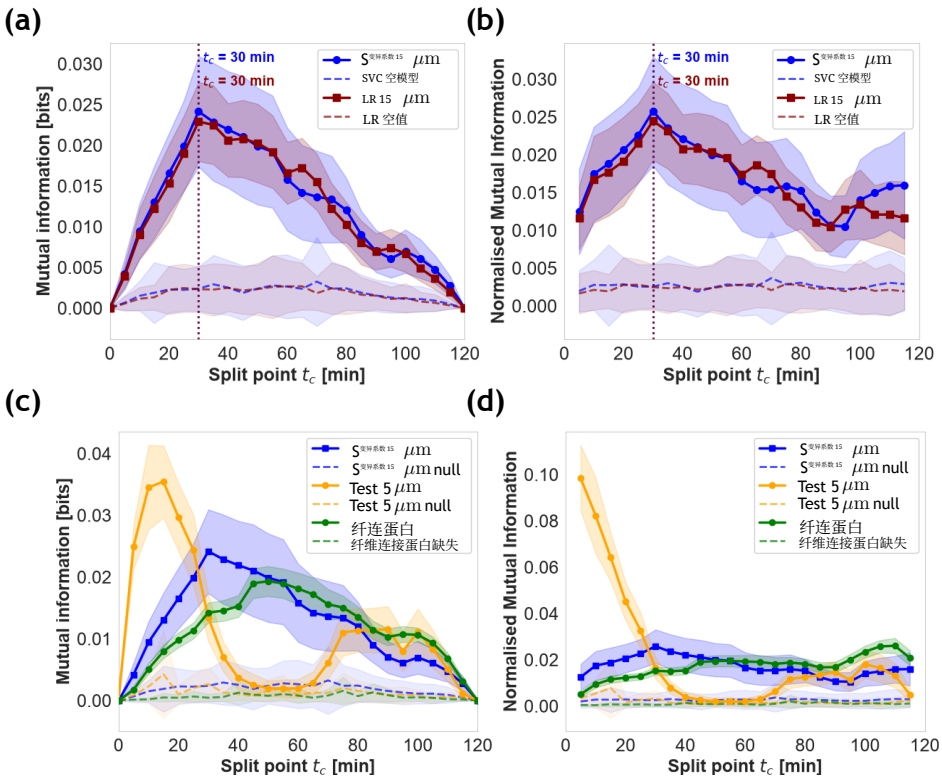


图 5。**用于预测通过时间的互信息**。使用 *best 7* 特征集，在 15μm狭窄通道上训练支持向量分类器径向基函数核（SVC，蓝色）和逻辑回归（LR，红色）分类器。（**a**）在15μm狭窄通道轨迹上测试的分类器中，真实类别与预测类别之间的互信息 $I$ 随分割点 $t_c$ 的变化。实线表示100次重复的均值，阴影区域表示标准差。虚线表示零模型，即在随机打乱通过时间标签的 15μm数据上训练的同一分类器。（**b**）归一化互信息。在（**a**）和（**b**）中，两个分类器的均值均在 $t_c = 30$ 分钟处达到最高（点线）。（**d**）SVC-RBF的互信息，（**e**）归一化互信息：在 15μm狭窄通道上训练后，无需重新训练，分别在 5μm狭窄通道（黄色）和涂有纤连蛋白的 15μm狭窄通道轨迹（绿色）上进行测试。作为比较，蓝色表示（**a-b**）中的 15μm结果。虚线表示相应的零模型（对于 5μm和纤连蛋白：20次重复的均值和标准差）。

变形虫样迁移可在受限环境中带来更高的迁移速度 [41, 42]，我们的结果表明，在我们研究的几何结构中，这种策略在通过狭窄通道时可能效率较低。

对于轨迹是否会通过 15μm 狭窄通道的分类任务，模型在另外两种截然不同的实验条件下也能很好地泛化，例如更狭窄的通道，以及涂有黏附性纤连蛋白的表面。请注意，在 15μm 狭窄通道条件下，细胞圆度 是通过与未通过类别之间分布差异最显著的唯一特征；然而，模型能够找出其他可预测通过或未通过的特征。同样，在 5μm 条件下，我们的分析发现了更多具有预测性的特征，这些特征与细胞周长在几何上的关联仅为部分相关；而细胞周长是实验上区分通过与未通过细胞的唯一差异最显著的特征值。尽管与未涂层条件相比，涂层显著改变了特征值的分布（见 **图 S3**），模型仍保留了其预测其中细胞圆度和细胞面积最为重要，这与数值分布存在显著差异的结果一致。总的来说，我们发现，在 15μm狭窄通道条件下训练的模型能够检测出携带另外两种条件下通过概率信息的其他特征。该模型能够很好地迁移到 5μm数据，但与15μm狭窄通道数据相比，沙普利值存在一些差异。在 5μm数据中，细胞核速度几乎总是较低，因此，在 15μm下观察到的快速细胞核与通过之间的关联并未出现，而细胞核面积和细胞核周长则呈单调作用，数值越大越有利于通过。这可能反映了活性细胞骨架的作用：它拥有更多表面来抓住细胞核，推动其通过更小的狭窄通道。由于在狭窄通道中速度达到最高值 [43]，，因此数据集中的快照极少对应于处在狭窄通道中的细胞核。因此，较大的细胞核尺寸并不是实验装置几何结构造成的伪影，而是细胞通过狭窄通道迁移能力的真实特

通过狭窄通道迁移的细胞能力特征。

为预测通过时间，我们将任务重新表述为二元分类：判断一条轨迹是否会在给定时间间隔内通过。在通过前的两小时内，支持向量分类器（径向基函数）和逻辑回归都能以高于随机基准的水平区分近通过 和远凌日 时间点，这表明细胞和细胞核的形态与动力学中存在持续的通过前特征。关于通过时间的信息最有可能在通过前约半小时达到最高。基于 15μm 狭窄通道条件训练、使用最佳 7 个特征的径向基函数支持向量分类器，能够泛化至纤连蛋白包被的 15μm 数据，在几乎所有 $t_c$上均取得高于随机基准的表现。与未包被条件不同，归一化互信息在约 30 分钟处没有明显峰值，而是在通过前约 30 分钟以上的所有 $t_c$ 中均有所升高。这表明，在纤连蛋白包被实验装置中，特征在通过前超过 30 分钟的一段时间内就已显示出通过的特征。相比之下，在更窄的 5μm 狭窄通道中，归一化互信息在 $t_c = 40$-70 分钟时降至零水平，意味着这一范围内的预测不包含有效信息。或许由于细胞必须在更窄的狭窄通道中发生变形，在 15μm 数据集上学到的预测性特征无法以相同方式标示即将发生的通过。

以往的研究根据从细胞轮廓中提取的几何特征，将细胞外基质中的MDA-MB-231细胞划分为不同的形态表型 [34]。作者将“驻留时间”量化为细胞从一种形态状态转变为另一种形态状态之前所经历的时间。对于某些形态表型，这一驻留时间约为30–60分钟。这表明细胞改变形态所需的时间尺度。我们观察到的峰值最可能出现的位置——$t_c \approx 30$ 分钟——与这一时间尺度一致。大约30分钟后，细胞可能已经改变了形态，因此其在时间 $t$ 时的特征对预测之后是否通过的作用会减弱。

总之，我们的结果表明，通过概率以复杂的非线性方式取决于细胞和细胞核的几何形态及动态变化。相比单独考察某个特征，综合考察一组特征能够捕捉这种依赖关系，并成功预测通过情况。

除了此处使用的几何特征外，我们的框架还可作为一种系统方法，用于确定哪些细胞特性决定了细胞通过受限空间迁移的能力。任何能够在活细胞中进行时间序列测量的量，都可以作为输入特征添加进去；随后，分类器会根据其预测通过能力的效果对其进行排序。这一点很有用，因为目前许多可能决定迁移能力的因素都是单独研究的，因此它们对受限迁移的相对重要性尚不清楚。例如，染色质凝缩 [44], 核支架蛋白的含量 [45], 凝聚体 [46], 高尔基体的定位 [47], 以及 Rac1 和 Cdc42 等信号蛋白的活性；这些信号蛋白会影响细胞形态和迁移 [48]。同时对其中若干因素进行成像并将其纳入特征，便可直接定量比较它们的预测能力。

## 方法

细胞培养。MDA-MB-231细胞在含10%体积胎牛血清（胎牛血清–Gibco）的DMEM培养基（杜氏改良伊格尔培养基–Gibco）中培养，于37°C下置于含5% $CO_2$的加湿培养箱中。细胞每周传代两次，每次按1:4稀释。细胞传代次数不超过14次。

微流控器件组装。将聚二甲基硅氧烷（PDMS）与交联剂（道康宁Sylgard 184弹性体套件）按10:1的比例混合，并在真空干燥器中脱气30分钟以去除气泡。将混合物直接倒在硅片上（由Taha Messlmani在皮埃尔-吉尔·德热纳研究所通过光刻制备），并置于 60°C烘箱中固化6小时。用手术刀从聚二甲基硅氧烷上切下各个图案，并分别使用3毫米和0.5毫米活检打孔器打孔，制作用于培养基储液池和中央抽吸口的开口。将玻璃盖玻片置于0.1 M盐酸溶液中浸泡过夜，然后用去离子$H_2O$冲洗，再用异丙醇冲洗，并用Kim擦拭纸擦干。将聚二甲基硅氧烷图案和盖玻片放入等离子清洗机中处理1.5分钟，然后将图案粘贴到玻璃上。将组装好的装置置于 100°C热板上20分钟，以完成粘接。组装完成后不超过2天即使用这些装置进行迁移实验。

细胞迁移实验。 器件先用70%乙醇和30% $H_2O$（按体积计）的混合液冲洗两次，再用磷酸盐缓冲盐水（PBS - Gibco）冲洗两次，最后用细胞培养基（DMEM培养基 - Gibco）冲洗两次，以进行灭菌。将浓度为 $10^6$/mL的细胞悬液加入培养基中，并注入器件。使用 25μL汉密尔顿注射器，从中央抽吸口轻轻抽取 5μL体积的液体，将细胞吸入狭窄图案区域（见**图1a**）。移除培养基，并换入含有 1μg/mL Hoechst 33342的培养基，孵育20分钟，最后用常规培养基冲洗两次。孵育5小时后，将培养基更换为不含酚红、含15 mM HEPES（Gibco）、100 units/mL青霉素和 100 μg/m:链霉素（Life Technologies）的DMEM培养基。进行纤连蛋白实验时，需在前一天进行包被。纤连蛋白包被时，将组装好的器件置于等离子清洗机中处理5分钟，然后用70%乙醇、30% $H_2O$冲洗两次，再用PBS冲洗一次。每个器件中加入 100μL PBS中的纤连蛋白（100 μg/mL），并置于用封口膜密封的培养皿中，在 37°C下孵育2小时。随后将培养皿转移至 4°C，放置至次日早晨。然后按常规方式装载器件，但省略两次乙醇冲洗和第一次PBS冲洗步骤。

显微成像。 所有延时成像均使用Nikon Ti-E落射荧光显微镜进行，该显微镜配备sCMOS相机（2048 ORCA Flash 4.0 V2，Hamamatsu或Prime BSI, Teledyne）、60倍Nikon油镜、Nikon完美聚焦系统，以及设定为 37°C、5% $CO_2$, 并带加湿功能的环境控制器。使用明场和落射荧光成像，采用360nm LED激发细胞核。每5分钟采集一帧，持续22小时。

图像处理。 图像通过一系列自定义 Python脚本进行处理。完成漂移校正后，使用 YOLO-SAM 模型 [49] 检测细胞和细胞核轮廓，该模型通过使用人工标注的实验轮廓重新训练以适配我们的数据。所有坐标均经过平移，使原点位于细胞最初占据的盒子中心。边界像素值用于定义细胞和细胞核轮廓。随后通过样条拟合对轮廓进行进一步处理，并对轮廓进行下采样，使点之间的间距始终为 $1\mu$m。

特征定义。 12 个特征的定义见**表 1**。特征 1 至 8 直接基于细胞和细胞核轮廓，其公式对于细胞和细胞核完全相同。因此，以下以具有 $n$ 个边界点的任意轮廓为例介绍计算 $\omega_j (j = 1, \ldots\ldots, n)$。轮廓内部包含坐标为 $(x_i, y_i)$ 的 $N$ 像素。其余特征（9-12）由分离向量 $\mathbf{v} = \mathbf{r}_{\text{cell}} - \mathbf{r}_{\text{nuc}}$ 和速度向量 $[\mathbf{r}_{\text{cell}}(t) - \mathbf{r}_{\text{nuc}}(t - \Delta t)]/\Delta t$推导得到。

模型推断。 模型训练和推断使用 scikit-learn 软件包 [50] 进行。我们比较了两种不同模型的结果：采用径向基函数核 $K(\mathbf{x},\mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ 的支持向量分类器（SVC-RBF），以及逻辑回归（LR）。核函数参数 $\gamma$ 采用中位数启发式方法进行选择，即 $\gamma = 1/\left(2\,\text{median}_{i<j}\|\mathbf{z}_i - \mathbf{z}_j\|^2\right)$，其中 $\mathbf{z}_i$ 为标准化特征向量，中位数取自数据集中所有观测值对。对于这两种分类器，正则化参数 $C$通过交叉验证进行调优，优化指标为精确率-召回率曲线下面积（PR-AUC），其计算使用 sklearn.metrics.average precision score。

特征选择。 我们通过对特征的排列进行贪心遍历来构建特征集。第一步：生成特征列表的一个排列。使用第一个特征训练模型。然后使用第一个和第二个特征训练模型。如果分数提升大于 $\varepsilon = 0.01$，则继续迭代；否则终止，并在此处停止循环。这里的分数是留出集 PR-AUC 的平均值，采用三折交叉验证并重复 5 次划分计算（75 − 25% 训练-测试集划分）。通过大量划分/重复实验对特征集进行评估。

在特征选择的每一步中，都会针对 $C \in \{0.01, 0.1, 1\}$ 训练并评估模型。每个训练集都与其表现最佳的 $C$相关联，该 $C$定义为在3折交叉验证集上的最佳平均值。

在方法部分，我们介绍如何选择核函数参数 $\gamma$ 和正则化参数 $C$，它们用于控制模型对训练数据的拟合程度（见**表3**）。较小的 $C$ 会得到更简单、正则化程度更高的模型，能够容忍一些训练误差；而较大的 $C$ 会使模型更贴合训练数据，但也有过拟合的风险。

可解释性。 使用沙普利值[36]估计特征重要性。设 $U$ 为参与者的全集，价值函数 $v : U \mapsto \mathbb{R}_+$ 为 $S \in U$分配总收益 $v(S)$之和，即参与者联盟$S$ 通过合作所能获得的收益。沙普利值 $\phi_i$ 为每个参与者 $i \in N$ 分配其对收益的贡献，该贡献是在所有可能的联盟中对 $i$参与情况进行平均得到的：

$$\begin{aligned}\phi_i(v) &= \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right) \\ &= \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} \left(v(S \cup \{i\}) - v(S)\right).\end{aligned}$$

我们使用SHAP软件包[37]中的 shap.KernelExplainer计算Shapley值。

筛选时间点。 我们删除每条轨迹最后的 $t_{\max}$ 分钟，因为这些子轨迹可能会被错误地关联到未观测到的转换。此外，对于穿越细胞，我们删除部分时间点，以确保每个通过前时间窗口仅包含一次通过事件。

通过时间分类。 我们分析了在 $15\,\mu$m 狭窄通道中发生过至少一次过渡的 $N_{\text{cells}} = 29$条轨迹。对于每次过渡，我们提取其发生前 $T \in \{5, 10, \ldots, 120\}$ 分钟内记录的所有观测（帧），并要求期间同一细胞没有发生其他过渡。总计有 $N = 1446$ 个快照。根据特征集的选择，每个快照 $i$ 由一个特征向量 $\mathbf{x}_i$ 描述，该向量包含细胞的形态和取向一致性特征（正文中采用最佳 7 项）、其过渡时间 $T_i$，以及所属轨迹的标识 $c(i)$。

为了预测近通过 和远凌日 快照，我们扫描以 $t_c$ =5、10、.....、115 分钟为标记的时间区间。在 $t_c = 0$ 和 $t_c = 120$ 分钟时，互信息设为零，因为此时仅存在一个类别（近通过/远凌日）。我们使用两个分类器，均采用 scikit-learn 中的实现：带径向基函数核的支持向量分类器 $K(\mathbf{x},\mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$（径向基函数支持向量分类器），以及经 $L_2$ 正则化的逻辑回归（LR）。对于径向基函数支持向量分类器， $\gamma$ 根据模型推断中所述的中位数启发式方法设定，并针对 $15\mu$m 数据集上的每个特征集计算一次。此过程不使用任何通过时间信息。

与凌日/非凌日分类不同，前者通过在精确率-召回率曲线下面积上进行交叉验证来调优 $C$，而这里我们对两个分类器都使用了固定的正则化参数 $C = 1$ 。凌日时间分类任务在数据、标签和类别平衡方面有所不同（仅包含正在通过的细胞的观测值，并根据变化的分割点标记为近通过 或远凌日

表1. 轮廓衍生特征的定义。对于特征1–8，细胞和细胞核轮廓采用相同的计算方法，因此面积、周长、圆度和纵横比各给出一个公式。公式适用于具有 $n$ 个边界点 $\omega_j (j = 1, ..., n)$的任意轮廓。边界点按顺序排列并构成闭合轮廓，因此 $\omega_1 = \omega_{n+1}$。该边界内包含 $N$ 个像素 $\Omega_i$，每个像素都有坐标向量 $\mathrm{r}_i = (x_i, y_i)$。质心坐标为$(X, Y)$，计算方式为$(1/N)\Sigma_i^N \mathrm{r}_i$。$\mathrm{r}_{\mathrm{cell}} = (X_{\mathrm{cell}}, Y_{\mathrm{cell}})$和$\mathrm{r}_{\mathrm{nuc}} = (X_{\mathrm{nuc}}, Y_{\mathrm{nuc}})$分别是细胞和细胞核的质心。特征9–12由向量 s 和 v 推导得出（见图2a）。向量 s 是细胞和细胞核质心之差，$\mathrm{r}_{\mathrm{cell}} - \mathrm{r}_{\mathrm{nuc}}$。速度向量 v 是细胞核的速度 $[\mathrm{r}_{\mathrm{cell}}(t) - \mathrm{r}_{\mathrm{nuc}}(t - \Delta t)]/\Delta t$。

| # 特征 | 单位 | 公式 |
|---|---|---|
| 1 细胞面积（C. Area） | $\mu\mathrm{m}^2$ | $N \times$（像素面积 $= 0.047 \mu\mathrm{m}^2$） |
| 2 细胞核面积（N. Area） | $\mu\mathrm{m}^2$ | |
| 3 细胞周长（C.Perim.） | $\mu\mathrm{m}$ | $\sum_1^n \lvert\omega_{j+1} - \omega_j\rvert$ |
| 4 细胞核周长（N.Perim.） | $\mu\mathrm{m}$ | |
| 5 细胞圆度（C. Circ.） | – | $4\pi$（面积）/（周长）$^2$ |
| 6 细胞核圆度（N. Circ.） | – | |
| 7 细胞长宽比（C. Asp.） | – | $R_x/R_y$，其中 |
| 8 细胞核长宽比（N. Asp.） | – | $R_x = \sqrt{\sum_{i=1}^N (x_i - X)^2/N}$，$R_y = \sqrt{\sum_{i=1}^N (y_i - Y)^2/N}$ |
| 9 间距（Sep.） | $\mu\mathrm{m}$ | $\lvert\mathbf{s}\rvert = \lvert\mathbf{r}_{\mathrm{cell}} - \mathbf{r}_{\mathrm{nuc}}\rvert$ |
| 10 点积 | μ米²/分钟 **秒 ·速度** | |
| 11 取向一致性（Align.） | – | $\mathbf{s} \cdot \mathbf{v}/(\lvert\mathbf{s}\rvert\lvert\mathbf{v}\rvert)$ |
| 12 前导投影（L. Proj.） | $\mu\mathrm{m}$ | $\mathbf{s} \cdot \mathbf{v}/\lvert\mathbf{v}\rvert$ |

表 2. 分类器所用的特征集、对应特征及特征选择过程中得到的正则化参数 $C$。较小的 $C$ 会产生更简单、正则化程度更高的模型，并容许一定的训练误差；较大的 $C$ 则能更贴合训练数据，但存在过拟合的风险。

| **特征集** | **特征** | $C$ **(逻辑回归/支持向量分类器)** |
|---|---|---|
| 最佳 2 | 细胞面积、细胞圆度 | 10/0.01 |
| 最佳 5 | 细胞长宽比、细胞纵横比变化速率、细胞圆度、细胞核面积、细胞核速度范数 | 0.001/10 |
| 最佳 7 | 细胞面积、细胞圆度、细胞核面积、细胞核圆度、细胞核周长、细胞核速度范数、间距 | 0.01/1 |
| 基础特征 | $Best\ 7$ 不含细胞核速度范数 +取向一致性、细胞长宽比、细胞周长、点积，前导投影、细胞核长宽比 | 0.01/1 |
| 动力学特征：细胞面积速度、细胞核面积速度、细胞速度范数，细胞核速度范数、细胞周长速度、细胞核周长速度，细胞圆度变化速率、细胞核圆度变化速率、分离速度，点速度、细胞纵横比变化速率、细胞核纵横比变化速率，排列速度，领先细胞投影速度 | | |
| All | 基线 +动力学特征 | 0.01/1 |

$t_c$），因此为过渡任务调优的数值并不适用于此处。表3显示，在较宽的范围内，结果对 $C$ 并不敏感：对于径向基函数支持向量分类器，在所有特征集上，当 $C \in [0.1, 10]$ 时，互信息和精确率-召回率曲线下面积仅有中等幅度的变化；而当 $C \leq 0.01$ 时，互信息则降至接近零，也就是说，预测标签几乎不包含有关真实标签的信息。对于逻辑回归，在 $[10^{-3}, 10^2]$范围内，结果对 $C$ 并不敏感。因此，$C = 1$ 的选择对于所有特征集和两种分类器而言，都处于性能稳定的范围内。

为补偿随 $t_c$ 强烈变化的类别不平衡，每个类别的训练观测 $a$ = 近过渡/远过渡 均按照 $w_a = N_{\mathrm{train}}/(2N_{\mathrm{train},\ a})$ 进行加权。特征使用仅根据训练数据计算得到的统计量进行标准化，使其具有零均值和单位方差。将分类器应用于其他数据集时（模型泛化至 5μm 和 15μm 纤连蛋白），使用各测试条件的均值和标准差对该条件下的特征进行标准化。这样可以消除不同条件之间细胞平均形态的差异，同时不使用任何有关测试细胞过渡时间的信息。

为确保分类器始终在训练期间未见过的细胞上接受测试，我们使用了重复 $K$ 折

按细胞分组的交叉验证。每次重复 $r$ =1，……， $R$ 29 个细胞都会被随机打乱，并分配到 $K = 4$ 个折中，每折包含 7–8 个细胞。对于每一折 $k$，训练一个分类器 $f_{r,\,k}$ 使用其余三折中的观测值进行训练，并用于预测第 $k$ 折中的观测值。由此得到一个折外预测

$$\hat{y}_i = f_{r,k(c(i))}(\mathbf{x}_i) \tag{1}$$

对于每个观测值，分类器从未见过与之对应的细胞。我们使用不同的随机折分配进行了 100 次重复实验，并报告各次重复实验的均值和标准差。

对于每次重复实验和每个分割点，将全部 $N$ 观测值的折外预测汇总为一个单独的$2 \times 2$ 混淆矩阵，其元素为 $N_{ab}$，即真实类别为 $a$ 且预测类别为 $b$的观测值数量。给定 $p_{ab} = N_{ab}/N$ 以及边缘分布 $p_a = \sum_b p_{ab}$ 和$q_b = \sum_a p_{ab}$，真实类别与预测类别之间的互信息为

$$I(y;\hat{y}) = \sum_{a,b} p_{ab}\,\log_2 \frac{p_{ab}}{p_a\,q_b}, \tag{2}$$

其中，对于 $p_{ab} = 0$ 设为零的项不计入。

互信息受真实标签熵的上界约束，即 $I(y;\hat{y}) \le H(y)$，其中

$$H(y) = -p\log_2 p - (1-p)\log_2(1-p), \tag{3}$$

其中， $p(t_c)$ 是分割点 $t_c$处{style id='3'>近通过 快照所占的比例。由于 $p$ 随 $t_c$单调增加（从 $t_c = 5$ 分钟时的 0{style id='8'>.06 增至 $t_c = 115$ 分钟时的 0{style id='11'>.97，见 **图 S4(a)**）， $H(y)$ 在 $t_c = 45$ 分钟时达到最大值，此时两个类别处于平衡状态；在扫描范围的两端则较小。因此，互信息随 $t_c$ 变化的最大值，可能部分反映类别平衡，而非分类能力。为消除这种依赖性，我们进一步报告归一化互信息

$$U(t_c) = \frac{I(y;\hat{y})}{H(y)} \in [0,1], \tag{4}$$

因此，对于不包含任何真实类别信息的分类器， $U = 0$；对于完美分类器， $U = 1$ 。对于分布外条件， $H$（$y$）根据相应测试数据的类别平衡计算。由于扫描范围两端的 $H$（$y$）较小， $U$ 在那里对统计波动更加敏感。

我们将分类器的性能与一个空模型进行了比较，在该模型中，特征与通过时间之间的关系被破坏。在每个训练集中，过渡时间 $T_i$ 会在训练观测值之间随机置换，并根据相同的规则 $T_i \le t_c$，依据置换后的时间分配标签。这会在每个 $t_c$ 中保留训练集的类别平衡，同时移除$\mathbf{x}_i$ 与 $T_i$ 之间的任何关联。空分类器基于这些置换后的标签进行训练，并在未作修改的留出折上进行评估，其互信息的计算方式与真实分类器完全相同。在每一折和每次重复中，都会重新计算空模型。

为检验归一化互信息 $U$ 的最大值位置有多稳健，我们对轨迹进行自助抽样。我们从 15μ米数据的29条轨迹中进行有放回重采样，共1000次，并根据每次重采样中抽取的细胞的折外预测结果，计算每个 $t_c$ 的$U$（不重新训练）。我们发现， $U$ 的最大值最常出现在 $t_c = 30$ 分钟：在64%（SVC-RBF）和49%（LR）的重采样中如此，而其他任何 $t_c$ 出现这一情况的比例均不超过13%。不过，其位置在不同重采样之间仍有较大变化（SVC-RBF的95%区间为$10 - 115$ 分钟）。对于SVC-RBF， $U$ 在 $t_c = 30$ 分钟时显著高于其最小值（$t_c = 95$ 分钟； $p = 0.022$，经邦费罗尼校正，针对22项比较）；而对于LR，与其他任何 $t_c$ 的比较均不显著。$t_c = 25$-35分钟与 $t_c = 45$-80分钟之间的差异并不显著（SVC-RBF： $p = 0.16$；LR： $p = 0.38$），因此最大值的界定并不精确。由此可见，过渡前信号最有可能在约30分钟的分割点处最强，并且在 $t_c$的大部分范围内仍高于零模型。

数据和代码可用性用于计算这些特征的数据和 Python 脚本可在 Zenodo（数字对象标识符：10.5281/zenodo.23344526）获取；机器学习分析所用的脚本可在 GitHub 获取：https://github.com/statbiophys/ cell-transit。

**致谢**

本研究得到居里研究所的一项机构资助（2020-2021 年），以及法国国家研究署（ANR）的资助，资助项目编号为 ANR- 22-CE30-0030（SQUEEZACTIV 项目）和 ANR-24- CE95-0005（DEMOQCELL 项目）。CSand MM 感谢 Thierry Dubois 在细胞处理方面给予的建议。

[1] A. T. Salminen、Z. Allahyari、S. Gholizadeh、M. C. McCloskey、R. Ajalik、R. N. Cottle、T. R. Gaborski 和 J. L. McGrath，Frontiers in Medical Technology **2**，600616（2020）。

[2] J. Perea Paizal、S. H.Au 和 C. Bakal，BritishJournal of Cancer **124**，58（2021）。[3] J. Roche，Cancers **10**，52（2018）。

| 特征集 | 分类器 | $C = 0.001$ | $C = 0.01$ | $C = 0.1$ | $C = 1$ | $C = 10$ | $C = 100$ |
|---|---|---|---|---|---|---|---|
| *Mutual information* $[10^{-3}$ *bits]* | | | | | | | |
| 最佳 2 | 径向基函数支持向量分类器 | 0.9 | 5.0 | 26.7 | 25.6 | 27.4 | 24.7 |
| 最佳 2 | LR | 18.2 | 18.4 | 18.4 | 18.4 | 18.4 | 18.4 |
| 最佳 5 | 径向基函数支持向量分类器 | 0.8 | 0.7 | 18.9 | 23.4 | 22.9 | 18.4 |
| 最佳 5 | LR | 15.6 | 16.2 | 16.3 | 16.4 | 16.4 | 16.4 |
| 最佳 7 | 径向基函数支持向量分类器 | 0.8 | 0.7 | 20.0 | 15.4 | 16.9 | 11.3 |
| 最佳 7 | LR | 13.3 | 13.7 | 14.3 | 15.7 | 16.4 | 16.5 |
| base | 径向基函数支持向量分类器 | 0.8 | 0.6 | 11.7 | 12.5 | 16.0 | 13.0 |
| base | LR | 11.4 | 11.3 | 13.1 | 13.0 | 12.9 | 12.8 |
| all | 径向基函数支持向量分类器 | 0.8 | 0.6 | 15.6 | 15.1 | 14.6 | 11.9 |
| all | LR | 13.0 | 15.5 | 18.7 | 20.3 | 21.3 | 21.5 |
| *PR-AUC* | | | | | | | |
| 最佳 2 | 径向基函数支持向量分类器 | 0.581 | 0.588 | 0.622 | 0.650 | 0.653 | 0.633 |
| 最佳 2 | LR | 0.581 | 0.580 | 0.580 | 0.580 | 0.580 | 0.580 |
| 最佳 5 | 径向基函数支持向量分类器 | 0.612 | 0.605 | 0.628 | 0.639 | 0.623 | 0.615 |
| 最佳 5 | LR | 0.632 | 0.631 | 0.629 | 0.629 | 0.629 | 0.629 |
| 最佳 7 | 径向基函数支持向量分类器 | 0.604 | 0.599 | 0.614 | 0.626 | 0.627 | 0.606 |
| 最佳 7 | LR | 0.615 | 0.616 | 0.613 | 0.615 | 0.616 | 0.616 |
| base | 径向基函数支持向量分类器 | 0.608 | 0.600 | 0.610 | 0.614 | 0.607 | 0.592 |
| base | LR | 0.595 | 0.598 | 0.606 | 0.610 | 0.610 | 0.610 |
| all | 径向基函数支持向量分类器 | 0.608 | 0.601 | 0.620 | 0.630 | 0.619 | 0.601 |
| all | LR | 0.611 | 0.616 | 0.626 | 0.630 | 0.630 | 0.630 |

表3. **通过时间分类器对正则化参数的敏感性** $C$。 对 SVC-RBF 和逻辑回归的预测标签与真实标签之间的互信息，以及精确率-召回率曲线下面积（以近通过为正类），在分割点$t_c$ = 20、30、45、60、80、100 分钟以及将细胞随机划分为 4 折的 5 次结果上取平均（对 15 μm 数据进行细胞分组交叉验证）。SVC-RBF 对每个特征集使用中位数启发式方法 $\gamma$。对于 $C \leq 0.01$，SVC-RBF 的互信息降至接近零，表明其预测标签几乎不包含有关真实标签的信息。对于 $C \in [0.1, 10]$，SVC-RBF 的性能基本稳定。逻辑回归对 $C$基本不敏感，但 *all*特征集除外，其性能提升至 $C \approx 1$。在正文中预测通过时间的分类任务中，我们使用 $C = 1$。

[4] V. Mittal, 《Annual Review of Pathology: Mechanisms of Disease》**13**, 395（2018）。[5] G. Charras 和 E. Sahai, Nature reviews Molecular cell biology **15**, 813（2014）。[6] C. D. Paul、P. Mistriotis 和 K. Konstantopoulos, Na- ture reviews cancer **17**, 131（2017）。[7] S. St¨oberl、J. Flommersfeld、M. M. Kreft、M. Benoit、C. P. Broedersz 和 J. O. R¨adler, Science advances **10**, eadm9195（2024）。[8] P. M. Davidson、C. Denais、M. C. Bakshi 和 J. Lammerding, Cellular and molecular bioengineering **7**, 293（2014）。[9] L. A. Lautscham、C. K¨ammerer、J. R. Lange、T. Kolb、C. Mark、A. Schilling、P. L. Strissel、R. Strick、C. Gluth、A. C. Rowat、 *et al.*, Biophysical journal **109**, 900（2015）。[10] Z. Sadjadi、D. Vesperini、A. M. Laurent、L. Barnefske、E. Terriac、F. Lautenschl¨ager 和 H. Rieger, Biophysi- cal Journal **121**, 4615（2022）。[11] A. L. McGregor、C.-R. Hsia 和 J. Lammerding, 《细胞生物学当前观点》**40**, 32（2016）。[12] P. Friedl、K. Wolf 和 J. Lammerding, Current opinion in cell biology **23**, 55（2011）。[13] A. Arjonen、R. Kaukonen 和 J. Ivaska, Cell adhesion & migration **5**, 421（2011）。

[14] J. Guck、F. Lautenschl¨ager、S. Paschke 和 M. Beil, 《Integrative Biology》**2**, 575（2010）。[15] T. L¨ammermann 和 M. Sixt, 《细胞生物学当前观点》**21**, 636（2009）。[16] M. Bergert、S. D. Chandradoss、R. A. Desai 和 E. Paluch, 《美国国家科学院院刊》**109**, 14434（2012）。[17] V. Ruprecht、S. Wieser、A. Callan-Jones、M. Smutny、H. Morita、K. Sako、V. Barone、M. Ritsch-Marte、M. Sixt、R. Voituriez、 *et al.*, 《细胞》**160**, 673（2015）。[18] P. Radler、T. Viehboeck、Z.-H. Luo、N. Masla´c、K. Schmidt、R. Hauschild、M. K. Nobu、S. Bulgheresi、T. E. Stradal、K. Rottner、 *et al.*, 《自然》, **1**（2026）。[19] S. Amiri、Y. Zhang、A. Gerardos、C. Sykes 和 P. Ron-ceray, 《物理评论研究》**6**, 043030（2024）。[20] J. Keys、B. C. Cheung、M. A. Elpers、M. Wu 和 J. Lammerding, 《细胞科学杂志》**137**, jcs260623（2024）。[21] J. Renkawitz、A. Kopf、J. Stopp、I. De Vries、M. K. Driscoll、J. Merrin、R. Hauschild、E. S. Welf、G. Danuser、R. Fiolka、 *et al.*, 《自然》**568**, 546（2019）。[22] V. Venturini、F. Pezzano、F. Catala Castro、H.-M. H¨akkinen、S. Jim´enez-Delgado、M. Colomer-Rosell、M. Marro、Q. Tolosa-Ramon、S. Paz-L´opez、M. A. Valverde、 *et al.*, 《科学》**370**, eaba2644（2020）。

[23] T. Karling 和 H. Weavers，EMBO reports 26，1238（2025）。[24] A. van der Net、R. Boot、I. Van Dijk、J. Conboy、P. Boukany 和 G. Koenderink，Biomicrofluidics 19（2025）。[25] A. L. Godeau、M. Leoni、J. Comelles、T. Guyomar、M. Lieb、H. Delanoe-Ayari、A. Ott、S. Harlepp、P. Sens 和 D. Riveline，bioRxiv，2020（2020）。[26] D. B. Br¨uckner、M. Schmitt、A. Fink、G. Ladurner、J. Flommersfeld、N. Arlt、E. Hannezo、J. O. R¨adler 和 C. P. Broedersz，《物理评论X》12，031041（2022）。[27] I. M. Wortel、I. Niculescu、P. M. Kolijn、N. S. Gov、R. J. de Boer 和 J. Textor，Biophysical journal 120，2609（2021）。[28] A. Moure 和 H. Gomez，Computer Methods in Applied Mechanics and Engineering 320，162（2017）。[29] S. U. Hirway 和 S. H. Weinberg，Computational and Systems Oncology 3，e1044（2023）。[30] E. Bahar 和 H. Yoon，*Healthcare*，9，911（2021）。[31] H. Umar、M. R. Aliyu、A. G. Usman、U. M. Ghali、S. I. Abba 和 D. U. Ozsahin，Scientific Reports 13，22242（2023）。[32] J. C. Kimmel、A. S. Brack 和 W. F. Marshall，IEEE/ACM Transactions on Computational Biology and Bioinformatics 18，562（2019）。[33] A. Mencattini、D. Di Giuseppe、M. Comes、P. Casti、F. Corsi、F. Bertani、L. Ghibelli、L. Businaro、C. Di Na-tale、M. Parrini、*et al.*，Scientific Reports 10，7653（2020）。[34] C. Z. Eddy、H. Raposo、A. Manchanda、R. Wong、F. Li 和 B. Sun，Scientific Reports 11，20434（2021）。[35] Z. Zhang、L. Chen、B. Humphries、R. Brien、M. S. Wicha、K. E. Luker、G. D. Luker、Y.-C. Chen 和 E. Yoon，Integrative Biology 10，758（2018）。[36] L. S. Shapley，“17. n人博弈的一个值”，载于 *Contributions to the Theory of Games (AM-28) Volume II*（普林斯顿大学出版社，1953），第307–318页。[37] S. M. Lundberg 和 S.-I. Lee（Curran Associates, Inc.，2017）。[38] T.-C. Lin、C.-H. Yang、L.-H. Cheng、W.-T. Chang、Y.-R. Lin 和 H.-C. Cheng，Cells 9，27（2019）。[39] K. M. Lugo-Cintron、M. M. Gong、J. M. Ayuso、L. A. Tomko、D. J. Beebe、M. Virumbrales-Munoz 和 S. M. Ponik，Cancers 12，1173（2020）。[40] J. H. Longstreth 和 K. Wang，American Journal of Physiology-Cell Physiology 326，C1212（2024）。[41] A. W. Holle、N. Govindan Kutty Devi、K. Clar、A. Fan、T. Saif、R. Kemkemer 和 J. P. Spatz，Nano letters 19，2280 2019( )。[42] S. R. Gabbireddy、K. W. Vosatka、A. J. Chung 和 J. S. Logue，Scientific Reports 11，17804（2021）。[43] S. Amiri、I. Bos、E. Reyssat 和 C. Sykes，Soft matter 21，893（2025）。[44] C.-R. Hsia、J. McAllister、O. Hasan、J. Judd、S. Lee、R. Agrawal、C.-Y. Chang、P. Soloway 和 J. Lammerd-ing，Iscience 25（2022）。[45] A. T. Tran、E. O. Wisniewski、P. Mistriotis、K. Stoletov、M. Parlani、A. Amitrano、B. Ifemembi、S. J. Lee、K. Bera、Y. Zhang、*et al.*，Nature Materials 24，1476（2025）。[46] V. Todorovski、F. McCluggage、Y. Li、A. Meid、J. P. Spatz、A. W. Holle、A. H. Fox 和 Y. S. Choi，Communications Biology 6，145（2023）。[47] K. Vaiˇziulyt˙e、A.-S. Mac´e、A. Battistella、W. Beng、K. Schauer 和 M. Coppey，Elife 11，e69229（2022）。[48] M. Yamao、H. Naoki、K. Kunida、K. Aoki、M. Matsuda 和 S. Ishii，《Scientific Reports》**5**，17527（2015）。[49] Ultralytics，“Ultralytics YOLOv8”，（2023），访问日期：2024-10-23。[50] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cour-napeau、M. Brucher、M. Perrot 和 E. Duchesnay，《Journal of Machine Learning Research》**12**，2825（2011）。

## 补充信息

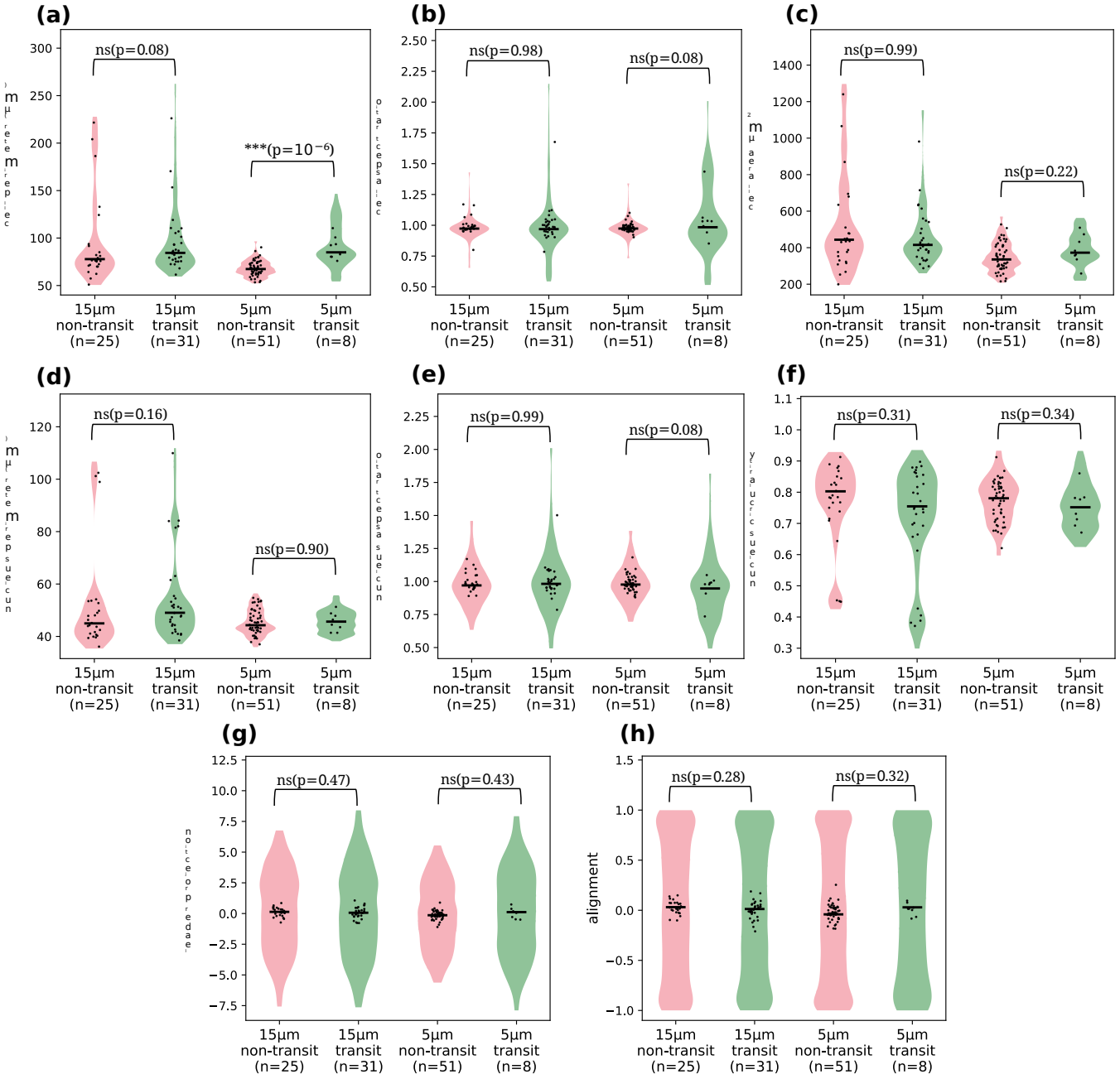


图 S1. 15 和 5µm 狭窄图案中，通过（通过，红色）和未通过（未通过，绿色）细胞每个快照的轮廓衍生特征分布。（**a**） *cell perimeter*、（**b**） *cell aspect-ratio*（**c**） *cell area*（**d**） *nucleus perimeter*、（**e**） *nucleus aspect-ratio*、（**f**） *nucleus circularity*、（**g**） *leader-projection*以及（**h**） *alignment*。红色和绿色小提琴图面积表示所有时间帧中快照的数据，而黑色点表示单条轨迹的时间平均值。统计显著性基于对时间平均值进行曼-惠特尼U检验所得的 $p$ 值，以避免跟踪时间较长的细胞使数据产生偏差。

表 S1. 各测试情境的精确率-召回率曲线下面积：100次划分的均值 [95% CI]，个针对精确率-召回率曲线下面积调优的模型。 15 $\mu$m：对各划分进行校正重采样 $t$ （Nadeau 与 Bengio（2003））； 5 $\mu$m 和纤连蛋白：分层细胞自助法（2000次重复，在类别内对完整细胞进行重采样，使用划分 $b$ mod 100 的模型对重复抽样 $b$ 进行评分）。随机基准：患病率。

| 特征集 | 模型 | 15 $\mu$m | 5 $\mu$m | 成纤维细胞 15 $\mu$m |
|---|---|---|---|---|
| 最佳 2 | 径向基函数支持向量分类器 | 0.653 [0.50, 0.81] | 0.702 [0.36, 0.93] | 0.701 [0.45, 0.90] |
| | LR | 0.652 [0.50, 0.80] | 0.670 [0.31, 0.92] | 0.708 [0.47, 0.90] |
| 最佳 5 | 径向基函数支持向量分类器 | 0.731 [0.56, 0.91] | 0.314 [0.10, 0.67] | 0.530 [0.35, 0.75] |
| | LR | 0.529 [0.34, 0.72] | 0.226 [0.09、0.53] | 0.562 [0.35, 0.75] |
| 最佳 7 | 径向基函数支持向量分类器 | 0.733 [0.59、0.88] | 0.577 [0.21, 0.90] | 0.618 [0.37, 0.83] |
| | LR | 0.638 [0.46, 0.82] | 0.393 [0.15, 0.76] | 0.651 [0.38, 0.86] |
| 基础特征 | 径向基函数支持向量分类器 | 0.673 [0.50, 0.84] | 0.619 [0.24、0.90] | 0.593 [0.37, 0.84] |
| | LR | 0.580 [0.39, 0.76] | 0.624 [0.28、0.88] | 0.606 [0.34、0.83] |
| All | 径向基函数支持向量分类器 | 0.680 [0.50, 0.86] | 0.540 [0.17, 0.87] | 0.627 [0.40, 0.83] |
| | LR | 0.632 [0.46, 0.81] | 0.327 [0.13, 0.67] | 0.639 [0.39, 0.84] |
| 随机基准 | | 0.290 | 0.022 | 0.415 |

表 S2。精确率-召回率曲线下面积（PR-AUC）的配对差异， $\Delta = A - B$ （100次划分的均值） [95% 置信区间]，针对精确率-召回率曲线下面积调优的模型。15 $\mu$m：对各划分进行校正重采样（Nadeau 与 Bengio（2003））； 5 $\mu$m 和纤连蛋白：分层细胞自助法（2000次重复，在类别内对完整细胞进行重采样，每次重复 $b$ 使用划分 $b$ 的模型评分，模 100）。粗体：置信区间不包含 0。*：在同一情境的 25 项比较中进行 Holm 校正的 $p < 0.05$。置信区间未经多重性校正。

| 比较（$A - B$） | 15 $\mu$m | 5 $\mu$m | 成纤维 15 $\mu$细胞 |
|---|---|---|---|
| *SVC-RBF: feature sets* | | | |
| 最佳 2 − 最佳 5 | −0.08 [−0.28, 0.12] | **+0.39 [0.05, 0.62]** | +0.17 [−0.01, 0.33] |
| 最佳 2 − 最佳 7 | −0.08 [−0.23, 0.07] | +0.13 [−0.08, 0.43] | +0.08 [−0.08, 0.29] |
| 最佳 2 − 基础特征 | −0.02 [−0.17, 0.13] | +0.08 [−0.10, 0.38] | +0.11 [−0.06, 0.28] |
| 最佳 2 − 全部 | −0.03 [−0.18, 0.13] | +0.16 [−0.07, 0.52] | +0.07 [−0.08, 0.24] |
| 最佳 5 − 最佳 7 | −0.00 [−0.16, 0.15] | −0.26 [−0.53, 0.06] | −0.09 [−0.25, 0.12] |
| 最佳 5 − 基础特征 | +0.06 [−0.09, 0.21] | −0.30 [−0.54, 0.02] | −0.06 [−0.22, 0.09] |
| 最佳 5 − 全部 | +0.05 [−0.09, 0.19] | −0.23 [−0.50, 0.17] | −0.10 [−0.24, 0.06] |
| 最佳 7 − 基础特征 | +0.06 [−0.01, 0.13] | −0.04 [−0.28, 0.19] | +0.02 [−0.10, 0.14] |
| 最佳 7 − 全部 | +0.05 [−0.04, 0.15] | +0.04 [−0.29, 0.45] | −0.01 [−0.15, 0.12] |
| 基础特征 −全部 | −0.01 [−0.09, 0.07] | +0.08 [−0.16, 0.44] | −0.03 [−0.11, 0.07] |
| *LR: feature sets* | | | |
| 最佳 2 − 最佳 5 | +0.12 [−0.08, 0.33] | **+0.44 [0.11, 0.68]** | +0.15 [−0.03, 0.36] |
| 最佳 2 − 最佳 7 | +0.01 [−0.14, 0.17] | +0.28 [−0.01, 0.51] | +0.06 [−0.11, 0.32] |
| 最佳 2 − 基础特征 | +0.07 [−0.06, 0.21] | +0.05 [−0.08, 0.22] | +0.10 [−0.06, 0.35] |
| 最佳 2 − 全部 | +0.02 [−0.13, 0.17] | **+0.34 [0.05, 0.58]** | +0.07 [−0.09, 0.31] |
| 最佳 5 − 最佳 7 | −0.11 [−0.26, 0.04] | −0.17 [−0.42, 0.02] | −0.09 [−0.26, 0.12] |
| 最佳 5 − 基础特征 | −0.05 [−0.23, 0.13] | **−0。40 [−0.65, −0。07]** | −0.04 [−0.21, 0.16] |
| 最佳 5 − 全部 | −0.10 [−0.26, 0.05] | −0.10 [−0.35, 0.08] | −0.08 [−0.22, 0.09] |
| 最佳 7 − 基础特征 | +0.06 [−0.03, 0.14] | −0.23 [−0.46, 0.03] | +0.05 [−0.09, 0.17] |
| 最佳 7 − 全部 | +0.01 [−0.05, 0.06] | +0.07 [−0.04, 0.21] | +0.01 [−0.12, 0.11] |
| 基础特征 −全部 | −0.05 [−0.13, 0.02] | **+0.30 [0.04, 0.53]** | −0.03 [−0.15, 0.06] |
| *SVC-RBF − LR, per feature set* | | | |
| 最佳 2 | +0.00 [−0.11, 0.11] | +0.03 [−0.04, 0.11] | −0.01 [−0.10, 0.07] |
| 最佳 5 | **+0.20 [0.03, 0.38]** | +0.09 [−0.15, 0.39] | −0.03 [−0.16, 0.13] |
| 最佳 7 | +0.10 [−0.02, 0.21] | +0.18 [−0.19, 0.48] | −0.03 [−0.24, 0.22] |
| 基础特征 | +0.09 [−0.03, 0.22] | −0.01 [−0.32, 0.22] | −0.01 [−0.18, 0.22] |
| All | +0.05 [−0.03, 0.13] | +0.21 [−0.08, 0.47] | −0.01 [−0.17, 0.19] |

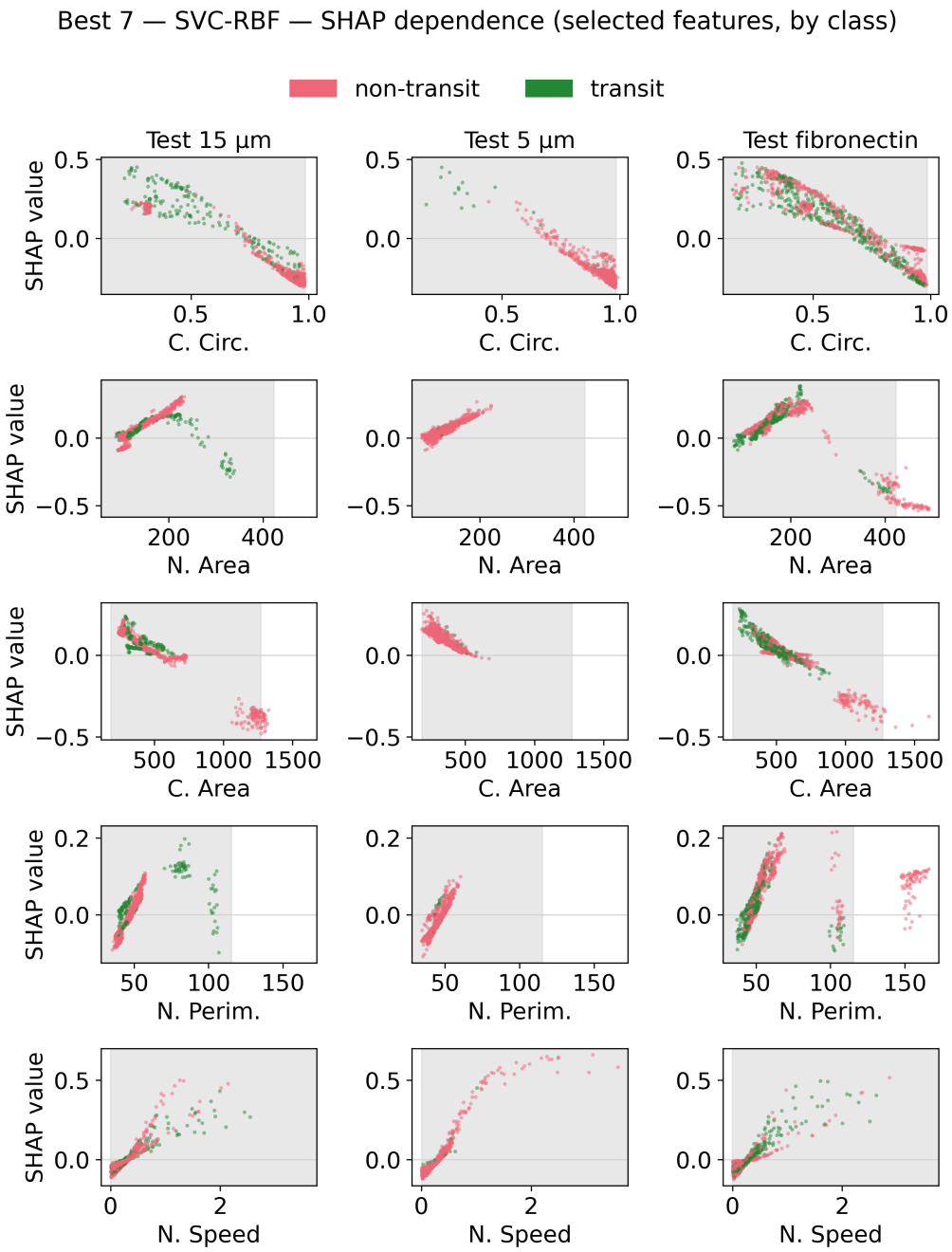


图 S2. *cell circularity*， *cell area*， *nucleus area*， *nucleus perimeter* 和 *nucleus speed*的 Shapley值。阴影区域表示 15µm 狭窄通道训练集中特征值的范围。绿色点表示通过轨迹，红色点表示未通过轨迹。以细胞面积和细胞核面积为例，可以看到，与有或没有纤连蛋白的 15µm 狭窄通道相比， 5µm 狭窄通道中的轨迹取值范围非常有限。

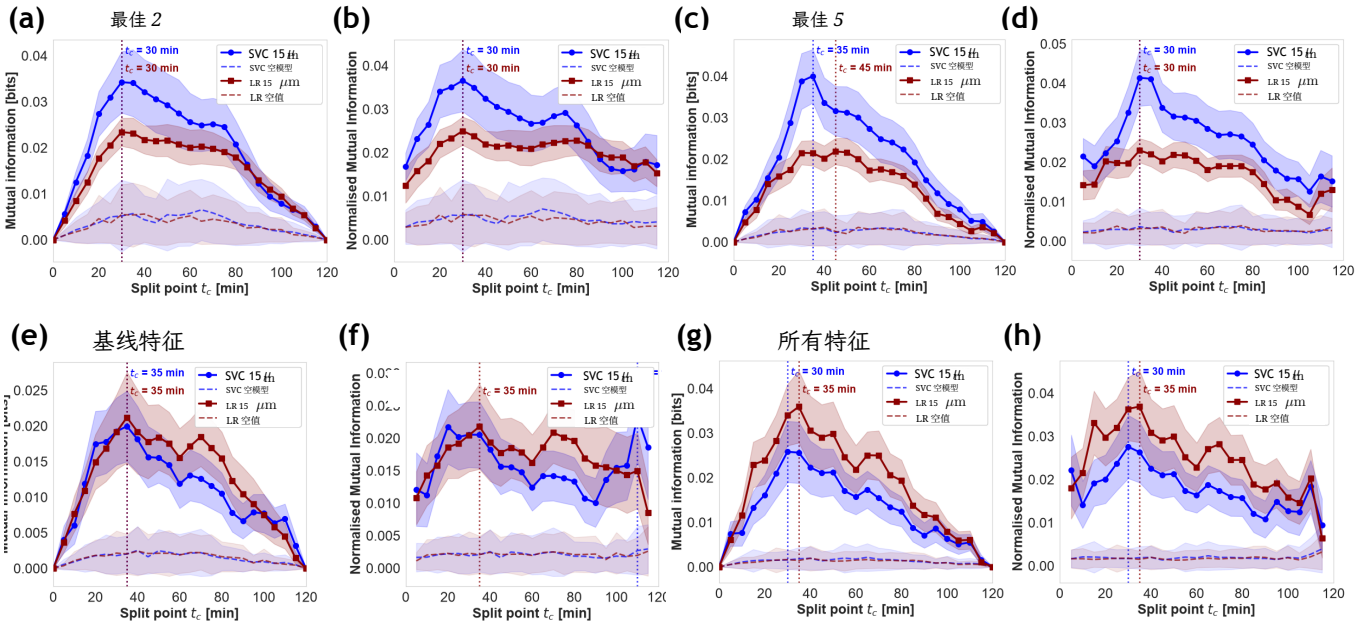


图 S3。**对以下对象的通过时间预测结果：**最佳 ***2***、最佳 ***5***、基线和全部**特征集**。实线表示 100 次重复的均值，阴影区域表示标准差。我们使用带径向基函数核的支持向量分类器（SVC-RBF，蓝色）和逻辑回归（LR，红色）。虚线表示零模型，即在标记前随机打乱训练时间点的通过时间，同时保持类别平衡。分类器在使用**(a-b)** *best 2*、**(c-d)** *best 5*、**(e-f)** *base* 和**(g-h)** *all* 特征集的 15μ 米狭窄通道上进行训练和测试。互信息 $I$ 表示真实类别与预测类别之间关于分割点 $t_c$ 的函数关系，以及归一化互信息。两个分类器的最大值均以点线标出，对于所有特征集，该最大值都保持在约 $t_c = 30 - 35$ 分钟处。在所有分割点上，两个分类器的表现都高于随机基准。

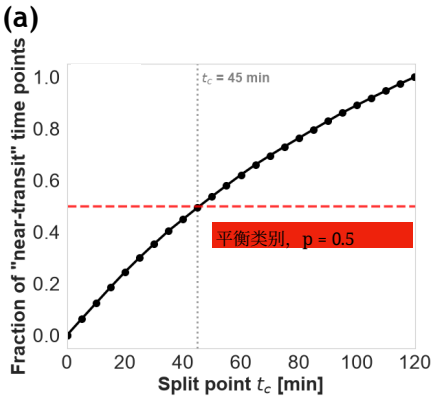


图 S4。**数据集中**近通过**轨迹的比例 (a)**被标记为 *near-transit* 的时间点比例 $p(t_c)$，随分割点 $t_c$ (15μ 米狭窄通道数据) 变化。由于通过时间 $\Delta t \leq t_c$ 的每个时间点都被标记为 *near-transit*，$p(t_c)$ 是通过时间的累积分布。两个类别是平衡的（$p = 0.5$，红色虚线），对应于 $t_c = 45$ 分钟。