多物种定位的拓扑

Abhinav Natarajan[1] Thomas Chaplin[1] Joshua A. Bull[2] Eoghan J. Mulholland-Illingworth[3]
Simon J. Leedham[3,5] Helen M. Byrne[2,3,4] Maria-Jose Jimenez[6,10]
Heather A. Harrington[7,8,9,1 0]

2026年3月4日

摘要

多物种数据中的空间关系能够指示并影响系统的结果与行为，涵盖从癌症中的疾病进展到生态学中的珊瑚礁韧性；因此，量化这些关系是各科学领域的一项重要课题。持续同调（PH）是拓扑数据分析（TDA）中的关键数学与计算工具，能够对数据形状进行多尺度描述。尽管它能有效描述物种的空间组织，例如病理学中的细胞模式，却无法检测不同类型物种之间的形状关系。传统上，持续同调分析单物种数据，这限制了对不同物种间相互作用的空间分析。借助拓扑数据分析和计算几何领域的近期进展，我们提出一种可扩展的方法，用于量化多物种数据中的高阶相互作用。该框架能够区分数据中形状特征或模式的存在情况，这些特征或模式可能：（i）为多种点的种类所共有；（ii）存在于某些物种中，但在其他物种出现时消失；（iii）只有在将多个物种结合起来考察时才可见；以及（iv）由某些物种形成，并在其他物种存在时仍然可见。我们通过两个示例应用展示了这一方法。我们识别出：（1）合成肿瘤微环境模型中的不同行为模式；以及（2）在疾病进展期间，结直肠癌组织样本中变化最显著的种间空间相互作用。

arXiv:2603.03237v1 [math.AT] 3 Mar 2026

科学领域中的许多相互作用本质上具有空间性，涉及多个物种协同作用时彼此间的关系，但对这类多物种空间组织的量化仍然有限。此外，复杂系统的涌现属性往往由高阶相互作用驱动，而这些相互作用无法简化为成对连接。例如，伤口愈合需要成纤维细胞、内皮细胞和免疫细胞协同作用，以促进血管生成和免疫反应；三级淋巴结构的形成则涉及T细胞、B细胞和成纤维细胞的协同作用。在癌症生物学中，肿瘤微环境内不同细胞群体（例如肿瘤细胞、巨噬细胞和内皮细胞）的空间组织及其相互作用，深刻影响着肿瘤进展、转移和治疗反应 [1]。多重成像技术的发展 [2–4] 能够通过功能标志物的表达，生成具有单细胞分辨率的组织样本空间分辨图谱；经过后处理后，得到的是各细胞类型中单个细胞的坐标（x， y）。量化这些图谱中不同细胞类型之间的空间相互作用，有助于确定并描述区分健康组织与病变组织的主要相互作用。

[1]牛津大学数学研究所，拉德克利夫天文台区，安德鲁·怀尔斯大楼，伍德斯托克路，牛津 OX2 6GG，英国。[2]沃尔夫森数学生物学中心，牛津大学数学研究所，牛津 OX2 6GG，英国。[3]人类遗传学中心，罗斯福大道，牛津大学，牛津，英国。[4]路德维希癌症研究所，纳菲尔德医学系，牛津大学，牛津 OX3 7DQ，英国。[5]转化胃肠病学中心，约翰·拉德克利夫医院，牛津大学，以及牛津NIHR生物医学研究中心，牛津，英国。[6]塞维利亚大学。高等计算机工程学院。雷娜·梅赛德斯大道 s/n，41012 塞维利亚，西班牙。[7]马克斯·普朗克分子细胞生物学与遗传学研究所，德累斯顿 01307，德国。[8]德累斯顿系统生物学中心，德累斯顿 01307，德国。[9]数学学院，德累斯顿工业大学，德累斯顿 01062，德国。[10]通讯作者。
Maria-Jose Jimenez：majiro@us.es；Heather A. Harrington：harrington@mpi.cbg.de

多种方法都可用于测量多物种相互作用，但目前没有一种方法能够以实用且易于解释的方式量化高阶拓扑关系。空间统计学方法根据细胞的空间分布及其距离推导统计量，并以此刻画多物种相互作用 [5–10]。网络方法则依据细胞间的连通性信息量化成对相互作用 [11, 12]。超图已用于对基因组学和代谢组学中的多方相互作用进行建模 [13, 14], ，但这类结构捕捉的是组合关系，而非空间拓扑特征，因此不太适用于空间数据。拓扑数据分析（TDA）提供了描述多尺度空间连通性的工具，例如连通分支、环和空洞；其中最重要的是持续同调（PH）。通过结合每个物种的PH，可以像 [15–17] 那样研究多物种空间关系。较新的TDA方法能够直接分析成对相互作用，采用Dowker持久性和多物种见证过滤 [18, 19], ，或混合条形码 [20]。这些拓扑方法只有通过诱导的成对关系，才能应用于三个或更多物种。统计描述符和拓扑描述符也可以整合起来，发挥各自的优势，Bull 等人 [21, 22] 对结直肠癌中多物种成对空间相互作用的分析便展示了这一点。

直接量化多物种数据中的高阶相互作用仍是一项重大挑战。这一挑战既是数学层面的——如何定义和解释高阶相互作用，也是计算层面的，因为数据中可能的相互作用数量会随着物种数量的增加而呈指数增长。彩色拓扑数据分析（Chromatic TDA）是拓扑数据分析领域的一项新进展 [23, 24], ，能够解决其中许多问题；然而，在实际应用和解释这一框架时，理论与计算方面仍存在不足。在此基础上，我们提出了一种新颖的拓扑框架，用于量化多物种数据中的高阶空间拓扑相互作用（总结见图 1）。该框架利用彩色 Delaunay 过滤 [25] 的离散 Morse 理论提高计算效率，并生成可解释的多物种空间关系拓扑特征。这些特征能够检测多种空间关系的存在，例如：(1) 某些物种中存在、但在其他物种出现时被掩盖的拓扑特征（图2a 和图2b）；(2) 只有在同时考虑多个物种时才会出现的特征（图2c）；以及 (3) 多个物种共有的特征（图2d）。特别是，该方法能够检测仅考虑成对相互作用时无法发现的空间特征。

我们在合成数据与实证空间生物数据上展示所提出的框架（第2节和第3节），其中分析由点的种类的单体、成对项和三元组所产生的空间特征。结果表明，该框架与现有的成对分析始终一致，但对多物种空间相互作用的几何给出了更完整、更稳健的描述。

1 拓扑模型

1.1 预备知识

拓扑数据分析的核心思想是将数据集解释为几何对象，并借助代数拓扑工具进行研究。在经典拓扑数据分析范式中，我们可以将一个 d 维点云 $X \subset \mathrm{R}^d$ 与一个 过滤单纯复形 联系起来——它是由尺度参数r索引的单纯复形嵌套序列 $F_r(X)$，可作为底层空间的多尺度近似，并假定 X 是从该空间中采样得到的。在每个尺度 r 下，复形 $F_r(X)$ 的 m 维连通性——例如，当 $m = 0$，分别为 1 和 2 时，对应的连通分支、环和空洞——由一个称为 $F_r(X)$ 的次m同调的向量空间 $H_m(F_r(X))$ 描述。在任意两个尺度值之间，例如 $r \le s$，复形的包含映射 $F_r(X) \hookrightarrow F_s(X)$ 会在相应的同调向量空间 $H_m(F_r(X)) \to H_m(F_s(X))$ 之间诱导一个线性映射。所有尺度值下的 m 次同调向量空间及其间的线性映射，共同构成这些

向量空间，共同构成 X [26–29] 的次m持续同调 $PH_m(X)$。持续同调（PH）描述数据在每个尺度下的拓扑特征，以及这些特征随尺度的演变。PH 可以通过其持久图完整描述（并可视化）；持久图由平面上的一组点构成，其中每个点（b，d）代表一个在尺度 b 出现、并在尺度 d 消亡的拓扑特征。该特征的生命周期为 $d-b$（从视觉上看，即点（b，d）到对角线的垂直距离）；长寿命特征通常被解释为底层空间的重要特征，而短寿命特征往往被视为噪声。在温和的假设下，$PH_m(X)$ 对 X [30, 31] 中的扰动具有稳定性，因此它是一种稳健的特征描述符，适合作为机器学习和统计推断流程的输入（向量化之后）。

对于多物种数据——其中 X 是由属于不同物种的点组成的不交并 $\bigsqcup_{i=0}^{s} X_i$ ——人们已尝试多种方法，修改传统的拓扑数据分析流程，以纳入物种标签提供的额外信息。在 [18], 中，作者使用*Dowker*持久性（定义见 [32], ），并引入多物种见证过滤。在这种设定下，固定两种点的亚种群 X_i 和 X_j，根据 X_i 构建单纯复形，并按照其单纯形与 X_j中各点的相对接近程度进行过滤，随后计算该过滤的持续同调。 [20], 提出的混合条形码将标准持续同调与像持续同调（下文将介绍）相结合，以捕捉当 X_j 中的点存在时，X_i 中的拓扑特征生命周期如何缩短。这两种方法都仅限于分析成对点云之间的相互作用。

1.2 彩色拓扑数据分析

Montesano 等人 [24] 提出了一种名为染色拓扑数据分析的框架，它结合了核、余核和像持久性。该框架的主要思想是，点的包含关系 $X_i \subseteq X_i \cup X_j$在每个尺度上诱导出复形的包含映射 $F_r(X_i) \subseteq F_r(X_i \cup X_j)$，而这些包含映射又在同调空间 ι_r之间诱导出线性映射：$H_m(F_r(X_i)) \to H_m(F_r(X_i \cup X_j))$。这些线性映射在所有尺度上汇集起来，构成持续同调模之间的映射 ι：$PH_m(X_i) \to PH_m(X_i \cup X_j)$。该映射的核——它本身由一系列向量空间 $\{\ker(\iota_r)\}_{r\geq 0}$以及线性映射 $\{\ker(\iota_r) \to \ker(\iota_s)\}_{r\leq s}$组成，贯穿各个尺度——描述了 X_i 中存在、但 $X_i \cup X_j$中不存在的拓扑特征，以及这些特征随尺度的演变。ι 的余核描述了存在于 $X_i \cup X_j$ 中、但单独在 X_i 中不存在的特征。ι的像描述了 X_i 中仍然存在于 $X_i \cup X_j$中的特征。核、余核和像也可以用持久图来描述。将这些图与 PH_m（X_i，定义域）和PH_m（$X_i \cup X_j$，陪域）的持久图结合起来，便可对X_i 与 $X_i \cup X_j$之间的空间关系作出丰富的数学描述；这些图属于 [24]提出的六图组。将 X_i和 X_j 替换为针对不相交颜色集合 I和 J的 $\sqcup_{i\in I} X_i$和 $\sqcup_{j\in J} X_j$，即可将上述方法扩展到两种以上的物种。然而，该方法对所选物种 i和 j并不对称；当 X_i和 X_j 的空间密度差异很大时，这会带来问题。此外，在考虑所有可能的包含关系后，持久图的总数会随着物种数量的增加而急剧上升。

作者通过定义 k 色彩子过滤来解决这些问题。该过滤是 $F_\bullet(X)$ 中所有单纯形的集合，其中每个单纯形所含点的不同颜色最多为 k 种。我们将这一子过滤到完整色彩过滤中的包含称为 k 色彩包含映射。该映射在持续同调上诱导的映射的核图、余核图和像图能够捕捉物种之间许多重要的空间关系。然而，这种方法仍存在一些问题。首先，它无法识别多个点种类中同时出现的特征（见图 1c）；其次，解释与该映射相关联的图中的特征并非易事。我们将在补充信息附录中详细讨论这一点。

1.3 多尺度多物种空间特征（M^2S^2）

我们提出一个框架，以解决现有拓扑方法在分析多物种数据时的各种不足。对于物种子集 $I \subset \{0、\ldots、s\}$，我们用 X_I 表示 X 中标签属于 I 的点的子集。固定一个整数 k，我们考虑由 X_I 张成的复形的不交并，其中 I 遍历颜色的 k 元组合：

$$F_r^{(k)}(X) := \bigsqcup_{\substack{I \subset \{0,\ldots,s\} \\ |I|=k}} F_r(X_I). \tag{1}$$

将 $F_{(k)r}(X)$ 中的每个单形映射到 $F_r(X)$ 中的对应单形，得到映射 $F_{(k)r}(X) \to F_r(X)$，进而得到映射 $PH_m(F^k(X)) \to PH_m(F(X))$，我们将其称为 k- 色粘合（次数为 m）；该映射将由每个 k 元物种组合中的点张成的过滤粘合在一起。我们考察该映射的核、余核和像持久图。这些图可以向量化以得到特征向量，例如使用持久性统计量 [33]或持久性景观 [34]，来描述数据子集之间的空间相互作用。

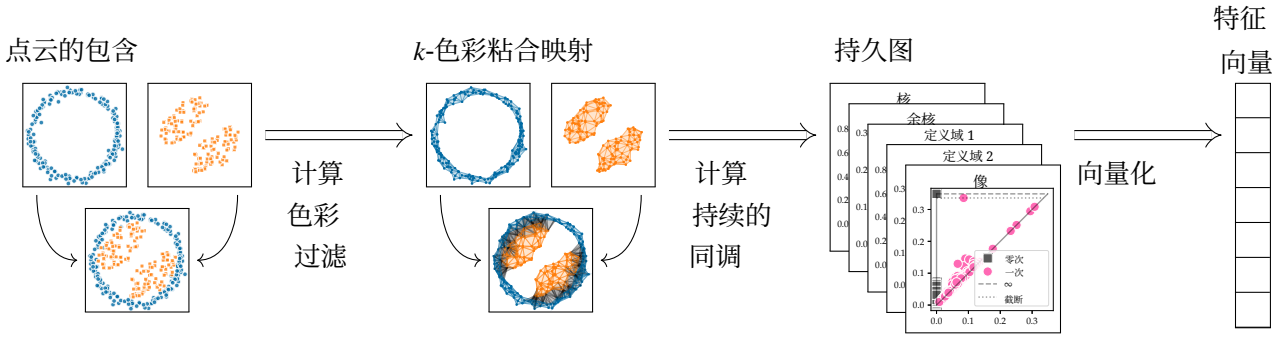
为了解释这些图所编码的空间关系，我们称一个拓扑特征为k色的，如果存在一组由 k 个不同物种标签组成的集合 I，使得该特征存在于 X_I 中。那么，k 色粘合映射在次数 m 下的余核会标识出非 k 色的 m 维特征；也就是说，这些特征不会出现在任何由至多 k 种颜色的点组成的数据子集中。关于 $k = 1$，请参见图2b中的第三个例子；关于 $k = 2$，请参见图2c。k 色映射的像描述了在加入所有其他物种的点后仍然存在的k色特征。请参见图2a中的空心圆或图2b中的第二个例子。k 色映射的核会标识出在加入其他物种的点后消失的 k色特征。一个特征可能会因被另一物种的点填充而消失（图2a中的蓝色圆被橙色点填充），也可能因两个特征合并为一个由不同物种子集共有的共址特征而消失（图2b中的两个共址单色圆合并，图2d中的三个共址双色圆合并为一个三色圆）。更多详细示例请参见图2。

一个重要的技术细节是如何根据数据点构造过滤。在拓扑数据分析中，传统的选择是Čech过滤和Vietoris–Rips过滤，但由于这些过滤中单纯形数量庞大，它们不适用于当前情境。alpha过滤中的单纯形较少，但也不适用于此处，因为点云的包含并不一定会诱导出alpha过滤的包含。色彩*alpha*过滤 [24] 正是为了解决后一缺陷而提出的。我们则采用彩色*Delaunay–Čech*过滤，它在拓扑上等价于色彩alpha过滤，且计算效率更高 [25]。

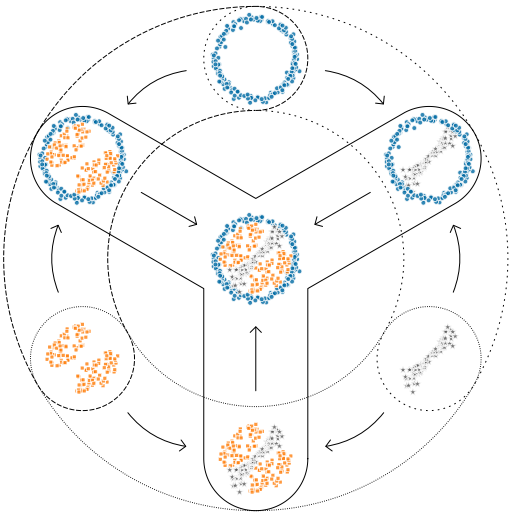
2 数据集描述

2.1 肿瘤微环境模型生成的合成数据

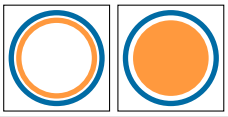
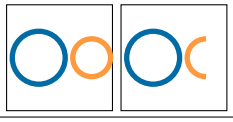
《基于代理的模型（ABM）》在 [36] 中介绍，用于模拟肿瘤微环境的时空动态，其间肿瘤细胞、基质细胞、坏死细胞、血管和巨噬细胞相互作用。巨噬细胞进一步分为抗肿瘤（M_1）和促肿瘤（M_2）两种状态。模拟探究了两个调节巨噬细胞行为的参数变化所产生的影响：χ_c^m，即对CSF-1梯度的趋化敏感性；以及$c_{1/2}$，即巨噬细胞外渗达到半最大值时的CSF-1浓度。不同的



(a) M^2S^2流程。一组彼此之间具有包含关系的点云产生了一个过滤的粘合映射（ k-色彩粘合映射），该映射诱导出一个持续同调映射。该映射可用定义域、核、余核和像的持久图来概括。这里，定义域图是每个点云各自通常的持久图，而其他图则编码点云之间的各种空间关系。随后使用持久性统计量将这些图向量化，得到概括点云之间空间关系的特征向量。

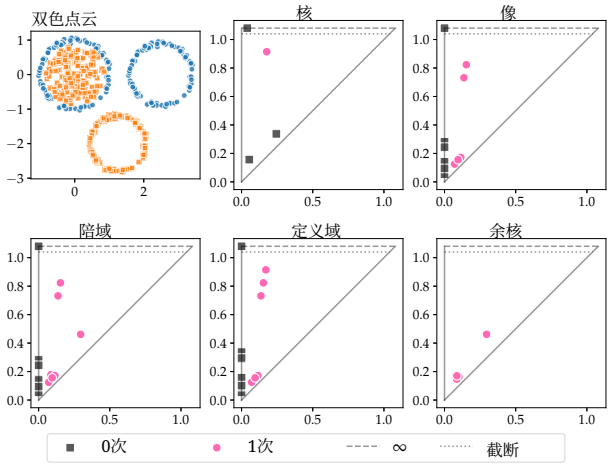


（b）三色示例所用的粘合映射。每个虚线／点线区域中的箭头构成两个物种的 1-色粘合映射。实线围成区域中的箭头构成二色粘合映射。共有三个 1-色粘合映射和一个二色粘合映射，总计四个映射，均用于（a）中所示的流程。

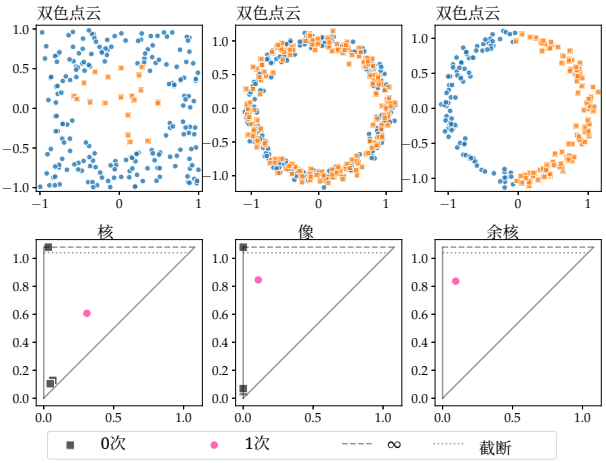
示例	Dowker 持久性 [18]	k-色彩包含 [24]	k-色彩粘合映射（拟议）
	✗	像	像
	✗	像	像
	需要 3 P持续性图	✗	核
	需要 3 P持续性图	✗	核

(c) 我们的方法相较于其他拓扑方法的区分能力。第一列展示了带标签点云的空间模式示例。其余各列分别对应一种用于量化这些模式的拓扑方法，并指出该方法是否能够区分第一列中的示例。第二列指的是使用 Dowker 持久图，但该方法无法直接处理三色空间关系。组合 3 个成对 Dowker 图可以区分最后两行中的示例。对于其余各列，我们明确指出哪些图（定义域、陪域、核、像、余核）能够区分这些构型，适用于 $k = \#$ 种颜色 $- 1$。

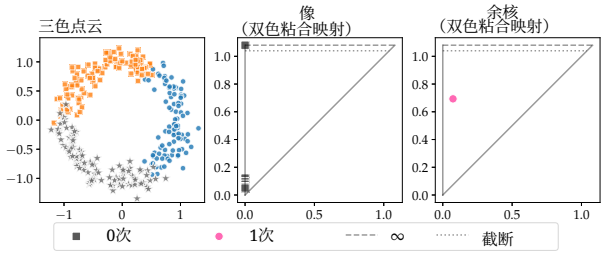
图 1：M^2S^2流程 (a) 汇集了每种点的种类对应的持久图中的信息，以及根据 (b) 所示方案将不同种类组合粘合后产生的核、余核和像图。这构成了一种更易解释且更具判别力的工具 (c)。



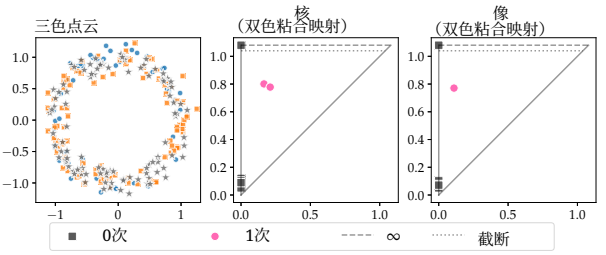
(a) 一个双色点云，以及由 1-染色粘合映射产生的相关核、像、陪域、定义域和余核持久图。1次核中的特征表明存在一个单色环，该环被另一种颜色的点填充。1次像中的两个长寿命特征表明，将两种颜色的点放在一起考虑时，有两个单色环仍然存在。1次定义域中的三个长寿命特征表明总共有三个单色环。余核持久图中缺少长寿命特征，说明每个双色特征都对应于至少一种颜色的点所形成的某个单色特征。



（b）当单色环被不同颜色的点部分或稀疏填充时，1-染色粘合映射的1次核会捕捉到这一现象，如左侧第一个示例所示。与环被完全或密集填充的情况相比，这一特征的生命周期更短；参见图2a。在第二个示例中，加入另一种颜色的点后，每个单色环仍保持为环，并且两个单色环对应于同一个双色环。因此，我们在1次像中看到一个持续时间较长的特征。从左侧数第三个示例中，1次余核里的显著长寿命特征表明，双色点云中存在一个在其任一单色子集中都不存在的环。



（c）对于三色点云，除了考虑由1-色粘合映射产生的图之外，我们还可以考虑由双色粘合映射产生的图。在此示例中，1次余核中的显著特征证明了存在一个在任何双色点子集中都不存在的三色环。1次像中没有特征，表明任何双色点子集中都不存在在三种颜色合并考虑时仍保留的大环。综合这些观察结果可知，点云中不存在大的双色环。



(d) 在此例中，我们再次考虑双色粘合映射。每个双色点子集都包含一个环，且三个双色环彼此重合，并对应于同一个三色环。1次核中的两个特征表明，当同时考虑三种颜色时，三个双色环中的两个与第三个环被视为相同。显然，三色环属于双色环的像，这一点由1次像中的单个长寿命特征所示。

图2：玩具示例展示了由 1 色和 2 色粘合映射得到的核、像、陪域、定义域和余核持久图所捕获的信息。为清晰起见，图中省略了持久度小于 0.05 的特征。

确定了若干参数范围，在这些范围内，基于智能体的模型呈现出三种定性行为之一，分别对应癌症免疫编辑的经典阶段 [37]：清除，其中 M_1 巨噬细胞消灭肿瘤；平衡，其中巨噬细胞控制肿瘤但不将其消灭；以及逃逸，其中 M_2 巨噬细胞促进肿瘤向血管系统迁移。我们为每个参数考虑九个不同取值，从而得到 81 种参数组合；每对参数最多进行 20 次随机模拟，共生成 1485 个点云，均取自模拟的最终时刻（t = 500小时）。随后， [36], 根据相应模拟中不同细胞类型的空间分布，对每组参数所对应的行为结果进行定性分类；示例见图3a。

2.2 结直肠癌标本的实证数据

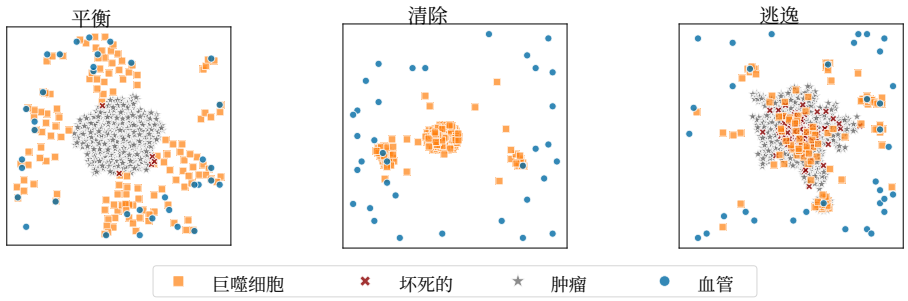
这些数据包含来自牛津直肠癌队列的43份腺瘤内癌标本，代表了从腺瘤向癌转变的结直肠肿瘤。研究采用多重免疫荧光染色和图像分割，根据标志物表达识别每份标本中11种点的坐标 [21]：上皮细胞、中性粒细胞、巨噬细胞、细胞毒性T细胞、辅助性T细胞、调节性T细胞、骨膜蛋白、CD146、CD34、平滑肌肌动蛋白和足糖萼蛋白。每份标本包含多个感兴趣区域（ROI），根据组织病理学评估将其归类为腺瘤（良性）或癌（恶性）。质量控制后，从41份患者标本中获得了12030个对齐良好的ROI，其中癌ROI有7360个，腺瘤ROI有4670个；每位患者的ROI数量在109至700个之间（含两端）。每个ROI均表示为R^2中的带标签点云。

3 结果

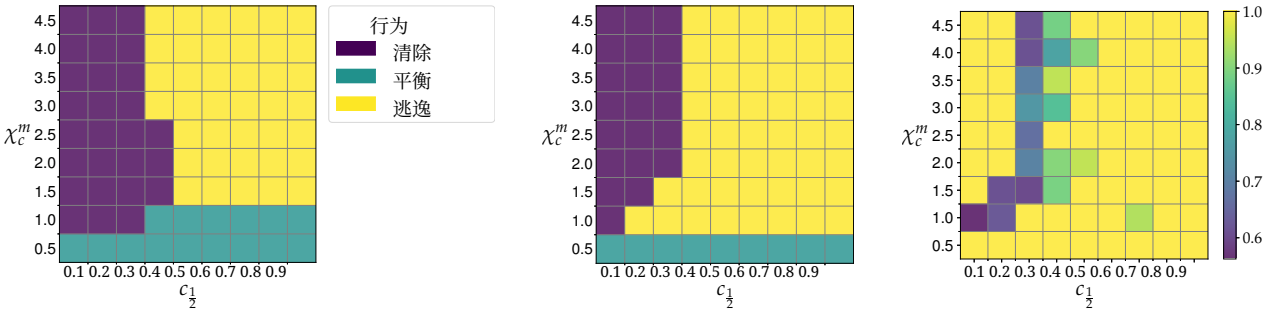
我们展示了多尺度多物种空间特征（M^2S^2）在两个数据集的点位数据上的有效性：（1）肿瘤微环境计算模型的数据 [36] ，其中包含五种细胞；（2）来自结直肠癌标本的实证数据，其中包含十一种物种。在合成数据集中，多物种空间关系能够区分不同的定性模型行为。在结直肠癌数据集中，多物种空间关系能够识别出在腺瘤进展为癌的过程中其空间关系发生改变的物种组合。其中包括仅凭成对关系无法捕获的三方空间关系。

3.1 M ^{2}S^2 区分合成数据集中的定性行为模式 et

我们利用每次模拟中五种细胞类型之间的空间关系（合并M1型和M2型巨噬细胞），来识别导致三种不同明行为结果——清除、平衡和逃逸的模型参数组合（图3a）。更确切地说，我们从每次模拟中获得一个多物种空间特征，并使用无监督机器学习将这些特征分配到三个簇中。根据由该参数组合生成的一组模拟所获得的簇标签中出现频率最高的标签，为每种模型参数组合指定标签。所得的聚类结果（图3c）与主观分配结果（图3b）一致。我们还考察了聚类纯度分数（图3d），以研究模型在行为模式之间过渡时的表现。清除与逃逸之间存在明显边界，但在平衡与逃逸之间看不到边界区域，这表明模型行为的这一变化比从清除到逃逸的过渡更加 abrupt。



(a) 基于智能体的模型中每种行为机制的示例样本。为清晰起见，图中省略了基质细胞；图中及我们的分析均未区分 M1 型和 M2型巨噬细胞。



(b) 对基于智能体的模型参数进行主观分类。(c) 使用 k-均值进行聚类的结果，以及 (d) 聚类的纯度得分。$k = 3$。

图3：合成基于智能体的模型数据集的聚类结果和聚类纯度分数。基于在最终模拟时间（$t = 500$）通过我们的流程获得的空间特征，对基于智能体的模型参数$c_{\frac{1}{2}}$ 和 χ_c^m的所有组合进行聚类。每种参数组合的颜色表示根据该参数组合的大多数模拟所指定的簇标签。纯度分数按每种参数组合中多数类别的模拟所占比例计算。

3.2 M²S² 识别结直肠癌数据集中的重要空间相互作用

我们将该流程应用于从一组包含 11 种点的腺瘤内癌组织样本中获取的感兴趣区域（图4a）。我们的目标是识别在腺瘤进展为癌的过程中空间关系发生改变的物种组合，并评估三方空间关系相较于单纯成对关系的重要性。对于每个感兴趣区域，我们计算 5 种免疫细胞和 5 种基质细胞的多物种空间特征（即排除用于配准的广泛表达的基准上皮标记物；见 [8]），每次考虑最多由三种物种构成的组合。为了量化腺瘤与癌之间空间关系的变化，我们根据多物种空间特征训练分类器，以区分腺瘤和癌感兴趣区域，并考察相关的特征重要性得分。我们发现，由于感兴趣区域空间结构存在患者间异质性，基于所有患者的感兴趣区域共同训练的分类器，其表现不如针对每位患者分别训练的分类器；因此，我们为每位患者分别训练一个分类器。我们使用这些分类器计算所考察的每种物种组合的汇总重要性得分，并对所有患者取均值。该分析进行两次：一次忽略物种三元组，另一次将其纳入。我们强调，我们分析的目标并非对腺瘤和癌感兴趣区域本身进行分类，而是识别在腺瘤进展为癌的过程中发生改变的重要空间相互作用。

分析揭示了显著的骨膜蛋白–巨噬细胞两两空间相互作用，这与文献中的既往发现一致；研究认为，骨膜蛋白可能在招募肿瘤相关

巨噬细胞（TAMs）至肿瘤部位的过程中发挥作用 [38–40]。在分类过程中纳入三阶交互作用不会显著影响分类准确率（图4b），但确实揭示了潜在的重要三阶空间相互作用，并改变了成对和单个特征的重要性得分（图4c）。特别值得注意的是，巨噬细胞–骨膜蛋白–中性粒细胞和巨噬细胞–骨膜蛋白–平滑肌肌动蛋白三阶交互作用的得分较高。

讨论

我们提出了一种拓扑数据分析框架，用于量化多物种空间数据中涉及多个点的种类的高阶空间关系。该框架区分数据中拓扑特征的存在性——例如连通分支、环和空洞——以及这些特征在不同物种子集之间的组合排列。该方法利用染色拓扑数据分析和计算几何领域的最新进展，将核、余核和像持久性与一种称为 *k*-染色粘合映射的新构造相结合。该框架适用于分析涉及三种或更多物种的高阶空间相互作用，而无需像现有文献中通常所做的那样，将其简化为一组成对相互作用。

我们在来自肿瘤微环境基于主体的模型的合成数据以及来自结直肠癌组织样本的实证数据上验证了我们的框架。我们在合成数据集上的结果与 Stolz 等人已有的分析一致 [18]。在结直肠癌数据集中，我们的结果揭示了：(1) 骨膜蛋白–巨噬细胞相互作用的重要性，这与现有文献 [38–40], 一致；以及 (2) 可能重要的涉及骨膜蛋白、巨噬细胞、中性粒细胞和平滑肌肌动蛋白的三阶交互作用。最重要的是，我们的方法可扩展至高阶相互作用，而不限于成对相互作用。我们预计，在多物种空间数据丰富的各种生物学情境中，例如微生物群落 [41], 、组织结构 [42], 和生态系统 [43]的研究中，我们的方法都将发挥作用。

4 材料与方法

4.1 持续同调计算

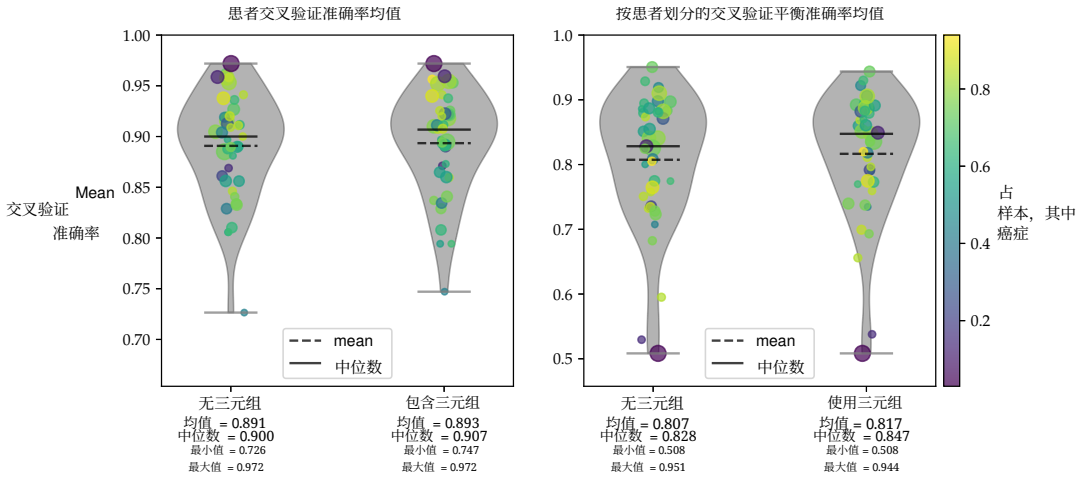
对于所考虑的每一物种组合，我们计算属于这些物种的点的彩色Delaunay–Čech过滤 [25]，并由该过滤计算持续同调。对于每一个体物种，我们计算0次和1次的一参数持久同调。对于成对（分别为三元组）物种，我们考虑第1.3节所述的1-色彩（分别为2-色彩）粘合映射，并计算与该映射相关联的核（0次和1次）、像（0次和1次）以及余核（1次）持久图。我们省略0次余核图，因为对于点云而言它们总是平凡的。我们不计算次数大于1的持久图，因为对于二维点云而言它们总是空的。

4.2 持久图的向量化

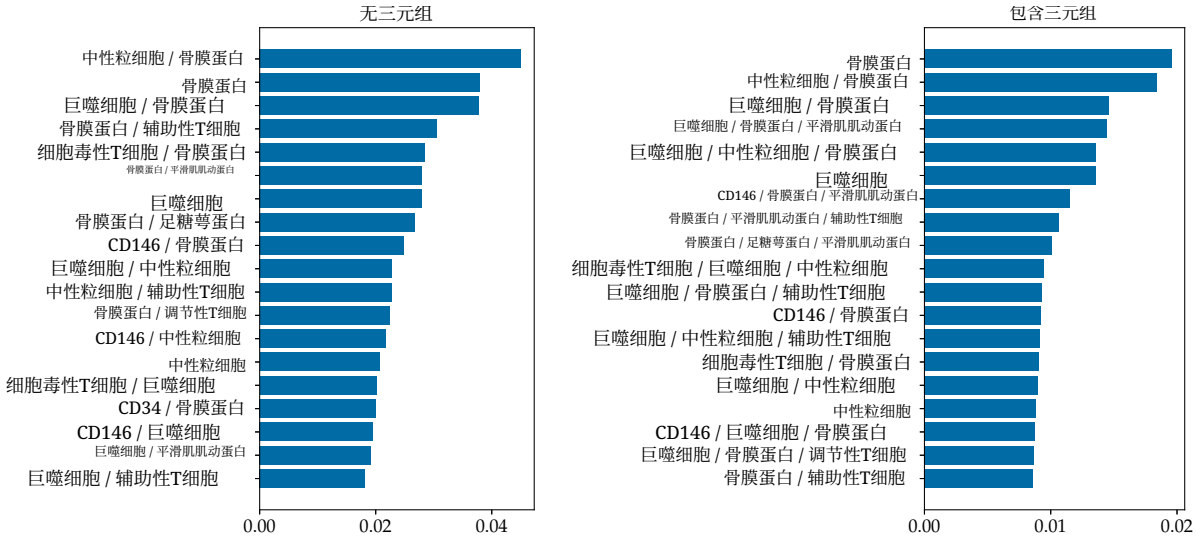
我们使用持久性统计量 [33], 对持久图进行向量化，这些统计量是由每个持久图中特征的出生时间、死亡时间、寿命（死亡时间减去出生时间）和中点时间（出生时间与死亡时间的平均值）构建的统计摘要。对于每个持久图，我们纳入出生时间、死亡时间、寿命的均值、标准差、中位数、极差，以及第10、第25、第75和第90百分位数



(a) 结直肠癌数据集中，腺瘤内癌组织样本（中间）的感兴趣区域示例。



(b) 结直肠癌数据集中每位患者在交叉验证折上的平均分类准确率，比较纳入和未纳入三阶交互作用的情况。每个点代表一位患者，阴影区域为所有患者准确率分布的小提琴图。点的面积表示该患者可用感兴趣区域的数量，颜色表示该患者癌变感兴趣区域的数量。左图显示原始分类准确率，右图显示平衡准确率（真阳性率和真阴性率的平均值），以校正类别不平衡。



(c) 结直肠癌数据集中各物种组的累积特征重要性得分（对所有患者取均值），分别为未纳入三阶交互作用（左）和纳入三阶交互作用（右）的情况。物种组按特征重要性降序排列，每个子图展示排名前二十的组。通过随机森林分类器计算这些特征重要性；这些分类器针对每位患者进行训练，以区分腺瘤和癌的感兴趣区域，所用指标为不纯度平均减少量（MDI）。

Fi图 4：结直肠癌数据集中腺瘤和癌的感兴趣区域分类结果。

以及该图中特征的中点。我们还纳入了特征数量和该图的持久熵 [44, 45] 。因此，我们从持久图中得到一个34维特征向量。对于0次的定义域图和像图，我们不统计出生时间、中点和生命周期，因为连通分支总是在时间零点出生，因此其生命周期和中点可由其死亡时间推得。对于这些图，我们得到一个10维特征向量。表1总结了各物种组合所贡献的特征。

组合	图的数量 图在线观看中文字幕	图 类型	特征向量 维数
个体	2	定义域（次数 0,1）	10 + 34 = 44
Pair	5	核（次数 0, 1） 像（次数 0, 1） 余核（次数 1）	10 + 4* 34 = 146
三元组	5	核（次数 0、1） 像（次数 0、1） 余核（次数 1）	10 + 4* 34 = 146

表1：持久图向量化的维数。

4.3 基于主体的模型模拟

该模型是一个二维、非格点、混合的基于力的主体模型，细胞运动由机械相互作用以及对可扩散物质（氧气、集落刺激因子-1、TGF-β、CXCL12 和表皮生长因子）的趋化反应所驱动 [36]。血管被建模为静止状态，且不与其他细胞发生物理相互作用；其余所有细胞类型均可运动，并根据局部环境信号发生相互作用。巨噬细胞具有连续表型 $\Omega \in [0, 1]$，介于抗肿瘤状态（$\Omega \approx 0$）和促肿瘤状态（$\Omega \approx 1$）之间。

4.4 基于主体的模型模拟的聚类

我们的流程输入包括 5 种个体细胞类型、10 对细胞类型和 10 个细胞类型三元组，因此每次模拟共有 25 种细胞类型组合。将这些组合对应的特征向量拼接起来，为每次模拟生成一个 3140 维特征向量。随后将这些特征向量输入 k-均值聚类算法，该算法使用 $k = 3$，且最多迭代 100000 次。

4.5 实证数据集获取

从牛津直肠癌队列中筛选出43例腺瘤内癌标本，并在获得伦理批准（REC 11/YH/0020）后取用福尔马林固定石蜡包埋组织块。采用 Akoya Biosciences OPAL方案（马尔伯勒，马萨诸塞州）和徕卡 BOND RXm 自动染色仪（Leica Microsystems，德国）进行多重免疫荧光（mIF）染色。每份样本切取并染色两张连续多重免疫荧光切片，以检测上皮标志物（E-钙黏蛋白）、免疫标志物（CD4、CD8、FoxP3、CD68、MPO）和间充质标志物（CD34、CD146、骨膜蛋白、足糖蛋白、 α-SMA）；这些标志物依据文献选取，具有预后相关性，并侧重于免疫-基质相互作用。随后对所得像进行细胞分割（HALO, Indica Labs）处理，并根据标志物表达分配物种。后处理后的像被划分为 1000px $\times$ 1000px (500μm $\times$500μm) 感兴趣区域（ROI），并以两张切片上均存在的上皮E-钙黏蛋白作为

参考标记，使用密度自适应点集配准（DARE）算法的改良版本进行对齐 [46]。更多详情请参见 [21]。

4.6 经验数据集中腺瘤和癌的分类

数据包括 10 种单独物种（不包括上皮）、45 对物种和 120 个物种三元组，因此每个样本共包含 175 种物种组合。将这些组合对应的特征向量拼接起来，为每个感兴趣区域生成一个 24530 维的特征向量。由于物种组合数量众多，相对于每位患者可用的样本数量，这些特征向量的维数非常高。因此，在训练分类器之前，我们使用主成分分析（PCA）降低特征向量的维数。对于每种物种组合 C，令 n_C 表示与 C 中特征相对应的特征空间维数，我们选取 min(5，⌊0。05n_C⌋) 个主成分（针对每位患者分别计算）。所有选定的物种组合总共产生 2535 个主成分。将数据投影到这些主成分上所得的结果用作分类器的输入；该分类器为包含 500 棵梯度提升决策树的随机森林 [47, 48]。为减少过拟合，最大树深度、叶节点最大数量以及每个叶节点所需的最小样本数分别设为 3、6 和 5；每次分裂时抽取 3% 的特征。梯度提升通过收缩进行正则化，学习率设为 0.4，每棵树进行 25 轮提升。分类器采用分层五折交叉验证进行训练和评估，训练-测试划分比例为 70:30。特征重要性使用平均不纯度下降（MDI） [47] 计算。分配给每组物种的净重要性得分，是该组对应的所有特征的 MDI 得分之和，再对队列中的所有患者取平均值。

4.7 处理缺失物种

在合成数据集和经验数据集中，对于每次模拟/每个样本，我们都会丢弃代表数量少于 3 个的物种，以避免计算彩色Delaunay三角剖分时遇到技术困难。对于某次模拟/某个样本中不存在的物种组合，其对应特征在该模拟/样本的拼接特征向量中设为零，从而确保所有模拟/样本的特征向量维数一致。

4.8 代码

计算彩色滤过并绘制持久图时，我们使用 chalc Python 包的 14.0.0 版本（PyPI 项目页面：https://www.pypi.org/project/chalc，GitHub 源代码：https: // github.com/abhinavnatarajan/chalc）。chalc 中的持续同调计算使用 lophat Rust crate 的 0.11.0 版本（项目页面：https://crates.io/crates/lophat，GitHub 源代码：https://github.com/tomchaplin/lophat），该 crate 实现了用于计算持续同调的无锁算法 [49]，并采用清除优化 [50]。核、余核和像图使用 phimaker Python 包的 0.3.0 版本计算（PyPI 项目页面：https://www.pypi. org/project/phimaker，GitHub 源代码：https://github.com/tomchaplin/phimaker），该包采用了 [51]中介绍的算法。随机森林分类和 k-均值聚类使用 Python 包 scikit-learn 的 v1.6 版本完成 [52]（DOI：10.5281/zenodo.14340513）。

用于从数据集中计算特征向量、运行分类器并可视化结果的代码可在 https://github.com/abhinavnatarajan/M2S2_demo 获取。

数据存档

基于主体的模型模拟可在 https://github.com/JABull1066/MacrophageSensitivityABM/ 获取。腺瘤癌变数据集的点云可在 https://github.com/JABull1066/ SHIFT-Score-Carcinoma-in-Adenoma/ 获取。

补充信息附录（SI）

补充信息附录包含有关拓扑数据分析的数学背景，以及用于阐释数学框架的其他图（图 S1–S5）：

1. 图 S1 展示了一个示例，以突出我们的框架与 k-色彩子过滤之间的差异。2. 图 S2 将我们的框架与 Dowker持久性进行比较，并进一步扩展了图 1c 表格前两行中的示例。3. 图 S3 和 S4 扩展了图 1c 表格第三行和第四行中的示例，将通过 k-色彩粘合映射获得的持久图与通过 k-色彩子过滤获得的持久图进行比较。4. 图 S5 扩展了图 2 中的示例，展示了其中部分示例的其余持久图。

致谢

作者感谢 U. Tillmann 就本文提供的有益讨论和意见。A.N. 获得牛津大学克拉伦登奖学金、默顿学院和数学研究所的支持。J.A.B. 获得英国癌症研究中心（CR-UK）编号为 CTRQQR-2021\100002 的资助，该资助通过英国癌症研究中心牛津中心提供。E.J.M.I 获得李·普拉西托医学研究奖学金（牛津大学）的资助。S.J.L. 获得英国癌症研究中心项目资助（DRCNPG-Jun22/100002）。数学与生物学融合获得英国癌症研究中心 CRC-STARS 战略资助（SEBCRCS-2024/100001）的支持。M.J.J. 获得由西班牙大学、研究与创新部/西班牙国家研究署/10.13039/501100011033 和欧洲区域发展基金/欧盟资助的 PID2024-155867NB-I00 项目资助，以及由西班牙大学、研究与创新部/西班牙国家研究署/10.13039/501100011033 资助的 IMUS–María de Maeztu 资助 CEX2024-001517-M（玛丽亚·德·梅斯图卓越单位支持计划）。A.N.、T.C.、H.M.B.、M.J.J. 和 H.A.H. 感谢英国拓扑数据分析中心英国工程与自然科学研究委员会资助 EP/R018472/1 提供的支持。HAH 衷心感谢英国皇家学会资助 RGF/EA/201074、UF150238，以及英国工程与自然科学研究委员会资助 EP/Y028872/1 和 EP/Z531224/1。

利益声明

作者声明不存在利益冲突。

参考文献

[1] Nicole M Anderson与M Celeste Simon。“肿瘤微环境”。载于：《当代生物学》 30.16（2020），R921–R925。[2] Yury Goltsev等人。“利用 CODEX 多重成像对小鼠脾脏结构进行深度分析”。载于：《细胞》 174.4（2018），第968–981页。

[3] Maryam Pourmaleki等人。“将空间分辨多重蛋白质谱分析推向临床肿瘤学”。载于：癌症发现 13.4（2023），第824–828页。[4] Charlotte Giesen等人。“利用质谱流式细胞术对肿瘤组织进行亚细胞分辨率的高度多重成像”。载于：自然方法 11.4（2014），第417–422页。[5] Joshua A Bull等人。“结合多种空间统计学可增强对肿瘤内免疫细胞定位的描述”。载于：科学报告 10.1（2020），第18624页。[6] Yeman B Hagos等人。“CD8+ FOXP3+ 与CD4+CD8+ 细胞的高度滤泡间空间共定位可预测滤泡性淋巴瘤的良好预后”。载于：血液肿瘤学 40.4（2022），第541–553页。8+ Yuzhou Feng等人。“利用SPIAT和spaSim进行空间分析以表征和模拟组织微环境”。载于：自然通讯 14.1（2023），第2697页。3+ Joshua A Bull等人。“用于多重成像数据空间分析的扩展相关函数”。载于：生物成像 4（2024），e2。issn：2633-903X。4+K. Schladitz与A. J. Baddeley。“三阶点过程特征”。载于：斯堪的纳维亚统计学杂志 27.4（2000），第657–671页。issn：1467-9469。doi：10.1111/1467-9469.00214。8+ Martin Kerscher。“宇宙中大尺度结构的统计分析”。载于：统计物理学与空间统计学。Klaus R. Mecke与Dietrich Stoyan编。柏林，海德堡：施普林格，2000，第36–71页。isbn：978-3-540-45043-6。doi：10.1007/3-540-45043-2_3。[7] Praveen Weeratunga等人。“单细胞空间分析揭示新冠肺炎肺中未成熟中性粒细胞与CD8 T细胞的炎症灶”。载于：自然通讯 14.1（2023），第7216页。[8] Praveen Weeratunga等人。“特发性肺纤维化中再生肺泡龛的时空细胞图谱”。载于：自然通讯 16.1（2025），第7150页。[9] Song Feng等人。“用于识别致病性病毒应答关键基因的生物网络超图模型”。载于：BMC生物信息学 22.1（2021），第287页。issn：1471-2105。doi：10.1186/s12859-021-04197-2。[10] Sean M. Colby等人。“引入分子超网络以在多维代谢组学数据中进行发现”。载于：蛋白质组研究杂志 23.11（2024），第4789–4801页。issn：1535-3893。doi：10.1021/acs.jproteome. 3c00634。[11] Oliver Vipond等人。“多参数持续同调景观识别肿瘤中的免疫细胞空间模式”。载于：美国国家科学院院刊 118.41（2021），e2102166118。[12] Bernadette J Stolz等人。“多尺度拓扑刻画动态肿瘤血管网络”。载于：科学进展 8.23（2022），eabm2456。[13] Jingjie Yang等人。“肿瘤-免疫相互作用及其动力学的拓扑分类：J. Yang等人”。载于：数学生物学杂志 91.3（2025），第25页。[14] Bernadette J Stolz等人。“面向多物种数据的关系持续同调及其在肿瘤微环境中的应用”。载于：数学生物学通报 86.11（2024），第128页。[15] Iris HR Yoon等人。“利用拓扑数据分析解析肺腺癌中细胞外基质与细胞空间模式的多样性及序列”。载于：bioRxiv（2024），第2024–01页。[16] Hubert Wagner等人。“混合条形码：量化点云之间的几何-拓扑相互作用”。2024。arXiv：2402.15058 [math.AT]。[21] Joshua A Bull等人。“整合多种统计方法以分析结直肠瘤变中的分期判别性细胞相互作用”。载于：bioRxiv（2024），第2024–06页。[22] Joshua A. Bull等人。“MuSpAn：多尺度空间分析工具箱”。载于：bioRxiv（2024）。doi：10.1101/ 2024.12.06.627195。eprint：https://www.biorxiv.org/content/early/2024/12/08/2024.12.06.627195.full.pdf。[23] Sebastiano Cultrera di Montesano等人。“染色拓扑数据分析”。2024年6月6日。arXiv：2406.04102[cs.CG]。[24] Sebastiano Cultrera di Montesano等人。“色彩阿尔法复形”。载于：数据科学基础 8（2026），第30–62页。doi：10.3934/fods.2025003。[25] Abhinav Natarajan等人。“着色Delaunay三角剖分的莫尔斯理论”。2024。doi：10.48550/arXiv. 2405.19303。arXiv：2405.19303 [math.AT]。

[26] Patrizio Frosini。“通过尺寸函数测量形状。”载于：《智能机器人与计算机视觉 X：算法与技术》。第1607卷。SPIE，1992年，第122–133页。doi：10.1117/12.57059。[27] Vanessa Robins。“从近似中计算同调的探索。”载于：《拓扑论文集》24（1999）。[28] H. Edelsbrunner、D. Letscher和A. Zomorodian。“拓扑持续性与简化。”载于：《第41届计算机科学基础年度研讨会论文集》。2000年，第454–463页。doi：10.1109/SFCS.2000.892133。[29] Afra Zomorodian和Gunnar Carlsson。“计算持续同调。”载于：《离散与计算几何》33.2（2005），第249–274页。[30] David Cohen-Steiner、Herbert Edelsbrunner和John Harer。“持久图的稳定性。”载于：《第二十一届计算几何年度研讨会论文集》。2005年，第263–271页。[3l] David Cohen-Steiner等。“利普希茨函数具有Lp稳定持续性。”载于：《计算数学基础》10.2（2010年4月1日），第127–139页。issn：1615-3383。doi：10.1007/s10208-010-9060-6。[32] Samir Chowdhury和Facundo Mémoli。“函子性Dowker定理与非对称网络的持续同调。”载于：《应用与计算拓扑期刊》2.1（2018），第115–175页。issn：2367-1734。doi：10.1007/s41468-018-0020-6。[33] Dashti Ali等。“拓扑数据分析中的向量化方法综述。”载于：《IEEE模式分析与机器智能汇刊》45.12（2023），第14069–14080页。[34] Peter Bubenik。“使用持续景观进行统计拓扑数据分析。”载于：机器学习研究期刊16.1（2015），第77–102页。issn：1532-4435。[35] Herbert Edelsbrunner。“球并及其对偶形状。”载于：第九届计算几何年度研讨会论文集。1993年，第218–231页。[36] Joshua A Bull和Helen M Byrne。“基于主体的肿瘤-巨噬细胞相互作用模型中的空间与表型异质性量化。”载于：《PLOS计算生物学》19.3（2023），e1010994。[37] Gavin P. Dunn、Lloyd J. Old和Robert D. Schreiber。“癌症免疫编辑的三大要素。”载于：《免疫学年度评论》22.第22卷，2004年（2004），第329–360页。issn：1545-3278。doi：https://doi.org/10.1146/annurev.immunol.22.012703.104803。[38] Wenchao Zhou等。“胶质母细胞瘤干细胞分泌的骨膜蛋白可募集M2肿瘤相关巨噬细胞并促进恶性生长。”载于：《自然细胞生物学》17.2（2015年2月），第170–182页。issn：1476-4679。doi：10.1038/ncb3090。[39] Meng Tang等。“卵巢癌细胞与巨噬细胞通过骨膜蛋白相互交流，促进巨噬细胞募集。”载于：《癌症科学》109.5（2018），第1309–1318页。issn：1349-7006。doi：10.1111/cas.13567。[40] Tian Wei等。“骨膜蛋白缺乏可减少PD-l+ 肿瘤相关巨噬细胞浸润，并增强抗PD-1疗法对结直肠癌的疗效。”载于：《细胞报告》42.2（2023年2月28日）。issn：2211-1247。doi：10.1016/j.celrep. 2023.112090。[41] Karoline Faustand Jeroen Raes。“微生物相互作用：从网络到模型。”载于：《自然综述：微生物学》10.8（2012年7月16日），第538–550页。issn：1740-1534 1740-1526。doi：10.1038/nrmicro2832。[42] Alejo Torres-Cano等。“空间组织化的细胞群落塑造胰腺的功能性组织结构。”载于：《科学进展》11.46（2025），eadx5791。issn：2375-2548。doi：10.1126/sciadv.adx5791。[43] Corina E. Tarnita。“空间生态学中的自组织。”载于：《当代生物学》34.20（2024年10月），R965–R970。issn：0960-9822。doi：10.1016/j.cub.2024.09.032。[44] H. Chintakunta等。“基于熵的持续条形码。”载于：《模式识别》48.2（2015年2月），第391–401页。doi：10.1016/j.patcog.2014.06.023。[45] M. Rucco等。“通过持久熵刻画独特型免疫网络。”载于：《ECCS 2014论文集》。施普林格国际出版，2016年，第117–128页。doi：10.1007/978-3-319-29228-1_11。[46] Felix Järemo Lawin等。“密度自适应点集配准。”载于：《2018年IEEE/CVF计算机视觉与模式识别会议》。2018年6月，第3829–3837页。doi：10.1109/CVPR.2018.00403。[47] Leo Breiman。“随机森林。”载于：《机器学习》45.1（2001），第5–32页。issn：1573-0565。doi：10.1023/A: 1010933404324。[48] Jerome H. Friedman。“贪婪函数逼近：梯度提升机。”载于：《统计年鉴》29.5（2001），第1189–1232页。issn：0090-5364。

[49] Dmitriy Morozov和Arnur Nigmetov。“无锁持续同调的探索。”载于：第*32*届*ACM*算法与体系结构并行性研讨会论文集。SPAA ’20。美国纽约州纽约市：美国计算机协会，2020年，第555–557页。isbn：978-1-4503-6935-0。doi：10.1145/3350755.3400244。[50] Ulrich Bauer、Michael Kerber和Jan Reininghaus。“清除与压缩：分块计算持续同调。”载于：数学与可视化 （2013）。issn：978-3-319-04098-1。doi：10.1007/978-3-319-04099- 8_7。[51] David Cohen-Steiner等。“核、像与余核的持续同调。”载于：*ACM-SIAM*离散算法研讨会论文集。2009年。doi：10.1137/1.9781611973068.110。[52] Fabian Pedregosa等。“Scikit-learn：Python中的机器学习。”载于：机器学习研究期刊 12（null，2011年11月1日），第2825–2830页。issn：1532-4435。

《多物种定位的拓扑》的补充信息

Abhinav Natarajan[1] Thomas Chaplin[1] Joshua A. Bull[2] Eoghan J. Mulholland-Illingworth[3]
Simon J. Leedham[3,5] Helen M. Byrne[2,3,4] Maria-Jose Jimenez[6,10]
Heather A. Harrington[7,8,9,10]

2026年3月4日

1 拓扑数据分析（TDA）背景

本节概述我们在研究中使用的拓扑数据分析方法。关于单纯复形和同调的更多细节，请参阅 [1]。关于拓扑数据分析的综述，请参阅 [2–4]。

1.1 单纯复形

有限点云 $X \subset \mathrm{R}^d$ 的拓扑是离散的，未必能反映数据的几何特征，例如点之间的两两距离，或不同尺度下簇、环和空洞的存在。我们不直接考察数据点本身，而是可以考察数据的 ϵ-邻域（即以数据点为中心、半径为 ϵ 的开球的并集），其中 $\epsilon \geq 0$ 是尺度参数。这个邻域记作 $B_\epsilon(X)$，通常能以更有意义的方式捕捉 X 的几何特征。为了进行具体计算，我们需要为 $B_\epsilon(X)$ 建立一个组合模型。

利用组合结构对数据点之间的关系建模并非新想法；例如，在许多数据分析领域中，常用邻近图和最近邻图来建模数据点之间的成对关系。我们使用单纯复形，它是简单无向图的高阶推广。

定义 1.1. 单纯复形是有限集合构成的集合，其中每个有限集合称为单形（复数形式为单纯形），并且每个单形的非空子集也都是单形。

如果 σ 是一个单纯复形中的单形，且 $\tau \subset \sigma$，我们称 τ 是 sigma 的面，并记作 $\tau \leq \sigma$。单纯复形 K 的一个子集若也是单纯复形，则称为 K 的子复形 。维数

[1]牛津大学数学研究所，拉德克利夫天文台区，Andrew Wiles 大楼，伍德斯托克路，牛津 OX2 6GG，英国。[2]牛津大学数学研究所沃尔夫森数学生物学中心，牛津 OX2 6GG，英国。[3]牛津大学人类遗传学中心，罗斯福大道，牛津，英国。[4]牛津大学纳菲尔德医学系路德维希癌症研究所，牛津 OX3 7DQ，英国。[5]牛津大学约翰·拉德克利夫医院转化胃肠病学中心，以及牛津NIHR生物医学研究中心，牛津，英国。[6]塞维利亚大学。高等计算机工程学院。雷娜·梅赛德斯大道 s/n，41012 塞维利亚，西班牙。[7]马克斯·普朗克分子细胞生物学与遗传学研究所，德累斯顿 01307，德国。[8]德累斯顿系统生物学中心，德累斯顿 01307，德国。[9]德累斯顿工业大学数学学院，德累斯顿 01062，德国。[10]通讯作者。Maria-Jose Jimenez：majiro@us.es；Heather A. Harrington：harrington@mpi.cbg.de

单形 $\sigma \in K$ 的维数定义为 $\dim(\sigma) = \#\sigma - 1$，也就是 σ 的基数减一；如果 $\dim(\sigma) = n$，我们称 σ 是一个 n- 单形 。 K 的维数是 K 中任意单形的最大维数。因此，简单无向图是一个 1 维单纯复形，其中顶点是 0 单形，边是 1 单形；与每条边关联的顶点是该边的面；子图对应于子复形。

单纯复形 K 通过将每个 n-单形 $\sigma \in K$ 与一个 n 维的几何单纯形 $|\sigma|$对应起来，构造出一个拓扑空间 $|K|$，称为它的几何实现。 n 维几何单纯形是 R^d 中 $n+1$ 个仿射无关点的凸包。举例来说，与 0 、1、2 和 3 维单形对应的几何单纯形分别是点、线段、三角形和四面体。要得到 K 的几何实现，就要按照 K 的组合结构，将这些几何单纯形沿着它们的公共面粘合起来。因此， K 是拓扑空间 $|K|$ 的组合模型。每个拓扑空间都与某个单纯复形的几何实现弱同伦等价，因此单纯复形是逼近空间拓扑的有力工具。

虽然单纯复形是对空间拓扑的组合近似，但我们也需要对空间之间的关系进行建模。

定义 1.2. 函数 $f\colon L \to K$ 从一个单纯复形到另一个单纯复形，称为单纯映射，如果 $\tau \le \sigma$ 属于 L，则 $f(\tau) \le f(\sigma)$ 属于 K。

一个重要的例子是 L 是 K的子复形，此时包含映射 $L \hookrightarrow K$ 是一个单纯映射。单纯映射是拓扑空间之间连续映射的组合模型，也就是说，单纯映射 $f\colon L \to K$ 会在 L 和 K的几何实现之间诱导一个连续映射 $|f|\colon |L| \to |K|$。在我们的研究中，我们用单纯映射来模拟由各个物种的点所形成的空间如何彼此粘合。

1.2 Čech复形与Alpha复形

我们再次考虑数据的 ϵ-邻域， $B_\epsilon(X)$。对 $B_\epsilon(X)$进行建模的一个单纯复形是切赫复形 $\check{\mathrm{C}}_\epsilon(X)$，其定义为

$$\check{\mathrm{C}}_\epsilon(X) = \left\{ \sigma \subseteq X : \bigcap_{x \in \sigma} B_\epsilon(x) \neq \emptyset \right\}. \tag{1}$$

也就是说，$\check{\mathrm{C}}_\epsilon(X)$ 中的单纯形是数据的子集；点集 $\sigma \subseteq X$ 是 $\check{\mathrm{C}}_\epsilon(X)$ 中的一个单形，当且仅当以 σ 中各点为中心、半径为 ϵ 的开球有公共交点。以下结果是代数拓扑中的一个经典结果——神经引理——的推论，它表明 $\check{\mathrm{C}}_\epsilon(X)$ 编码了 X 在尺度 ϵ 下的空间结构。

定理 1.3. 切赫复形 $\check{\mathrm{C}}_\epsilon(X)$ 的几何实现与 ϵ-邻域 $B_\epsilon(X)$同伦等价。

另一种能够编码 X 空间结构的单纯复形是阿尔法复形 $\mathrm{A}_\epsilon(X)$ [5]。点集 $\sigma \subset X$ 是 $\mathrm{A}_\epsilon(X)$ 中的一个单形，当且仅当 σ 存在一个半径不超过 ϵ 的外接球 S，且该球所界定的开球不包含 X 中的任何点。阿尔法复形以如下方式编码与切赫复形相同的拓扑信息。

定理 1.4。 对于每个 $\epsilon \ge 0$，$\mathcal{A}_\epsilon(X)$都是$\check{\mathrm{C}}_\epsilon(X)$的子复形，而几何实现的诱导包含映射 $|\mathcal{A}_\epsilon(X)| \hookrightarrow |\check{\mathrm{C}}_\epsilon(X)|$ 是同伦等价。

在 X 中的点满足温和条件且 ϵ 足够大的情况下，阿尔法复形 $\mathrm{A}_\epsilon(X)$ 就是广为人知的德洛内三角剖分 $\mathrm{Del}(X)$，它常用于计算机图形学、地理信息系统以及有限元/有限体积分析。出于计算目的，阿尔法复形通常比切赫复形更受青睐，因为它的单纯形更少，计算效率也更高。

1.3 同调

从单纯复形 K 中提取拓扑信息的一种方法是研究它的同调。对于每个$n \geq 0$，令 $C_n(K)$ 表示 K 中由形式上的 Z/2Z 线性组合构成的向量空间，其中的单形均为 n 维；若 $n < 0$ 或 $n > \dim(K)$，则 $C_n(K) = 0$。n 维单形 $\sigma \in C_n(K)$ 的边界 记为 $\partial\sigma$，它是 $C_{n-1}(K)$ 中由其 $(n-1)$ 维面构成的形式和：

$$\partial\sigma := \sum_{\substack{\tau \subset \sigma \\ \dim(\tau)=n-1}} \tau.$$

通过对 $C_n(K)$ 中所有单形作线性扩张，我们得到边界映射 $\partial_n\colon C_n(K) \to C_{n-1}(K)$。不难验证 $\partial_n \circ \partial_{n+1} = 0$，也就是说，$\mathrm{im}(\partial_{n+1}) \subseteq \ker(\partial_n)$。因此，商向量空间

$$H_n(K) := \ker(\partial_n)/\mathrm{im}(\partial_{n+1})$$

定义良好，并称为的同调 K 在次数 n。非正式地说，$H_n(K)$由 K中的 n维“洞”生成；对于 $n = 0$, 1 和 2，它们分别对应于 K 中的连通分支、环和空洞。

同调的一个关键事实是它具有函子性：单纯映射 $f\colon K \to L$ 会诱导一个线性映射$H_n(f)\colon H_n(K) \to H_n(L)$，并且这些线性映射满足以下性质：

1. 对于恒等映射 $\mathrm{id}_K\colon K \to K$，有 $H_n(\mathrm{id}_K) = \mathrm{id}_{H_n(K)}$。2. 对于任意两个单纯映射 $f\colon K \to L$和 $g\colon L \to M$，有 $H_n(g \circ f) = H_n(g) \circ H_n(f)$。

这对我们的研究很重要，因为同调的函子性使我们能够追踪一个空间中的拓扑特征如何映射到另一个空间中的拓扑特征，并比较不同空间的拓扑特征。

1.4 过滤与持续同调

令 $F_\bullet(X)$ 表示 X 的 Čech 过滤或 alpha过滤。迄今所述流程中的一个问题是尺度参数 ϵ 的选择。它会显著影响 $|F_\epsilon(X)|$的拓扑；若事先不了解数据，就不清楚该如何选择 ϵ。因此，我们不固定 ϵ 的单一取值，而是考虑单纯复形的整个族 $\{F_\epsilon(X)\}_{\epsilon\in[0,\infty]}$，它是对 X 几何的多尺度表示。

定义 1.5（过滤）。设 K 是一个单纯复形。$[0,\infty]$ 索引的过滤是由 K 的子复形构成的族 $\{K_t\}_{t\in[0,\infty]}$，满足：若 $s \leq t$，则 K_s 是 K_t 的子复形，并且

$$K = K_\infty = \bigcup_{t\in[0,\infty]} K_t.$$

如果 K 配备了一个过滤，我们就称它为过滤单纯复形。$K_\bullet$ 的子过滤是由单纯复形构成的族 $\{L_t\}_{t\in[0,\infty]}$，满足对每个 $t \in [0,\infty]$，L_t 都是 K_t 的子复形。

容易验证，$\{\check{\mathrm{C}}_\epsilon(X)\}_{\epsilon\in[0,\infty]}$和$\{\mathrm{A}_\epsilon(X)\}_{\epsilon\in[0,\infty]}$都是$X$的幂集 $\mathrm{P}(X)$上的过滤；它们分别称为*Čech*过滤和*alpha*过滤。

定义 1.6（持续同调）。设 $K_\bullet$ 是单纯复形 K的一个过滤。$K_\bullet$的 n次持续同调是由向量空间$\{H_n(K_t)\}_{t\in[0,\infty]}$以及由包含映射 $K_s \hookrightarrow K_t$诱导的线性映射$\{H_n(K_s) \to H_n(K_t)\}_{s\le t}$组成的集合。我们用 $\mathrm{PH}_n(K_\bullet)$表示 $K_\bullet$的 n 次持续同调。

$\mathrm{PH}_n(F_\bullet(X))$ 同时编码了 X 在每个尺度 ϵ 下的 n 维拓扑特征，以及这些拓扑特征随尺度变化的演变。

1.5 持久模与持久图

持续同调是持久模的一个例子，也就是说，它由配备线性映射 $\{V_t\}_{t\in[0,\infty]}$ 和 $\{v_{s,t}\colon V_s \to V_t\}_{s\le t}$ 的一系列向量空间组成，并满足以下性质。

1. $v_{t,t} = \mathrm{id}_{V_t}$。 2. $v_{r,t} = v_{s,t} \circ v_{r,s}$

当 $r \le s \le t$ 时。

持久模的另一个例子是区间模 $I^{[s,t)}$，定义如下：

$$\left(I^{[s,t)}\right)_r = \begin{cases} \mathbb{Z}/2\mathbb{Z} & \text{if } s \le r < t, \\ 0 & \text{otherwise,} \end{cases}$$

所有非零线性映射均为 Z/2Z 上的恒等映射。持久模理论的标准结果表明，当 $K_\bullet$ 是有限 单纯复形的过滤时，对于任意 $n \ge 0$，其持续同调 $\mathrm{PH}_n(K_\bullet)$ 都能唯一分解为区间模的直和（例如参见 [7] 中的主要结果）。

定理 1.7（持久同调的区间分解）。设 $\{K_t\}_{t\in[0,\infty]}$是有限单纯复形的一个过滤，且 $n \ge 0$ 固定。则存在唯一的区间多重集 $\{[b_i, d_i)_{i\in I}$，其中每个区间都包含在 $[0,\infty]$中，使得持续同调 $\mathrm{PH}_n(K_\bullet)$同构于直和 $\bigoplus_{i\in I} I^{[b_i,d_i)}$。

因此，我们可以将过滤 $K_\bullet$ 的持续同调可视化为平面上的点的多重集 $\{(b_i,\ d_i)\}$。点 $(b_i,\ d_i)$ 对应于 $\mathrm{PH}_n(K_\bullet)$ 的区间分解中的区间 $[b_i, d_i)$，因此表示一个 n维拓扑特征，当且仅当 $t \in [b_i, d_i)$ 时，该特征在 K_t 中存在。

2 k-色彩粘合映射的补充细节

假设 X 是R^d 中的一个有限点集，它被划分为子集 X_0，……，X_s；我们将$\{0$，……，$s\}$中的指标称为标签，并将每个 X_i 称为一个物种。我们采用与主文稿相同的记号；对于标签的一个子集 I，我们令 X_I 表示 X 中标签属于 I的点所构成的子集：

$$X_I := \bigcup_{i\in I} X_i.$$

对于每个这样的子集 $I \subseteq\{0$、……、 $s\}$，我们都有 $\check{\mathrm{C}}_\bullet(X_I)$ 是 $\check{\mathrm{C}}_\bullet(X)$ 的子过滤。将 $\check{\mathrm{C}}_\bullet(X_I)$ 在所有包含 k 个不同标签的标签子集 I 上取不交并，我们便得到 k-色彩粘合映射：

$$q_{k,\bullet} : \bigsqcup_{\substack{I\subseteq\{0,\ldots,s\} \\ \#I=k}} \check{\mathrm{C}}_\bullet(X_I) \to \check{\mathrm{C}}_\bullet(X).$$

k-色彩粘合映射在每个尺度上都会产生单纯映射，并将由每个 k 元组物种张成的复形粘合起来。由同调的函子性，进而得到一族线性映射 $(\epsilon \geq 0)$：

$$H_n(q_{k,\epsilon}) : \bigoplus_{\substack{I \subseteq \{0,\dots,s\} \\ \#I=k}} H_n(\check{\mathrm{C}}_\epsilon(X_I)) \to H_n(\check{\mathrm{C}}_\epsilon(X)).$$

这些线性映射的核、余核和像在 ϵ 的所有取值上汇集成持久模，这是代数中的标准结果所致——例如参见 [8, 引理 9.1]。这些持久模也具有区间分解，并可通过其持久图加以描述和可视化。

2.1 彩色滤过

如正文所述，使用 Čech过滤时会遇到一个计算难题：过滤中的单纯形数量会随着数据点数量呈指数级增长。对于 alpha过滤，这并不是问题；但另一方面，$\mathrm{A}_\bullet(X_I)$ 不一定是 $\mathrm{A}_\bullet(X)$ 的子过滤（例如参见 [9, 图 2]）。这意味着，对于 alpha过滤，k-彩色粘合映射没有良好定义。彩色*Delaunay*三角剖分 [10], 、耦合阿尔法过滤 [11], 、色彩*alpha*过滤 [11, 12], 以及彩色*Delaunay–Čech*过滤 [9]解决了 alpha过滤的这一问题。在我们的工作中，我们使用彩色Delaunay–Čech过滤 $\mathrm{ChrDel}\ \check{\mathrm{C}}_\bullet(X)$。它是 Čech过滤的一个计算上可行的子过滤，并且在每个尺度 $\epsilon \in [0, \infty]$ 上，$|\mathrm{ChrDel}\ \check{\mathrm{C}}_\epsilon(X)|$ 与 $|\check{\mathrm{C}}_\epsilon(X)|$ 同伦等价，因此能够捕捉与 Čech过滤相同的拓扑信息。

2.2 与 k-色彩包含映射的比较

设 $F_\bullet(X)$表示Čech过滤、色彩alpha过滤或彩色Delaunay–Čech过滤中的任意一种。在 [12], 中，作者将 $F_\bullet(X)$的 k-色彩子过滤 ——此处记为 $F_\bullet(X;\ k)$——定义为由 $F_\bullet(X)$中那些至多包含 k个不同物种的点的单纯形所组成。我们将包含$i_{k,\bullet}\colon\ F_\bullet(X;\ k) \hookrightarrow F_\bullet(X)$称为 k-色彩包含映射。 k-色彩粘合映射经由k-色彩包含映射分解，如下面的交换图所示：

$$\begin{array}{ccc} \bigsqcup_{\substack{I \subseteq \{0,\dots,s\} \\ \#I=k}} F_\bullet(X_I) & \xrightarrow{q_{k,\bullet}} & F_\bullet(X) \\ \downarrow & \nearrow_{i_{k,\bullet}} & \\ F_\bullet(X;k) & & \end{array} \tag{2}$$

因此，这两种映射密切相关，并且实际上对于 $k = 1$是重合的。然而一般而言，它们捕捉到关于数据的不同信息，图 1给出了一个例子来说明这些差异。来自k-色彩粘合映射的图更易于解释，并且捕捉到了 k-色彩包含映射所未捕捉到的差异，如图 1、3和4所示。

我们指出，也可以考虑其他类型的包含；例如，可以考虑将 $F_\bullet(X;\ k)$ 包含到 $F_\bullet(X;\ k+1)$ 中。

3 正文示例的补充图

1. 图 1 展示了一个示例，以突出 k-色彩粘合映射与 k-色彩包含映射之间的差异。

2. 图2 将我们的框架与 Dowker持久性进行了比较，并扩展了正文中图1(c)表格前两行的示例。
3. 图 3 和图 4 扩展了正文中图1(c)表格第三行和第四行的示例，比较了通过 k-色彩粘合映射得到的持久图与通过 k-色彩子过滤得到的持久图。 4. 图5 扩展了正文中图2的示例，展示了其中一些示例的其余持久图。

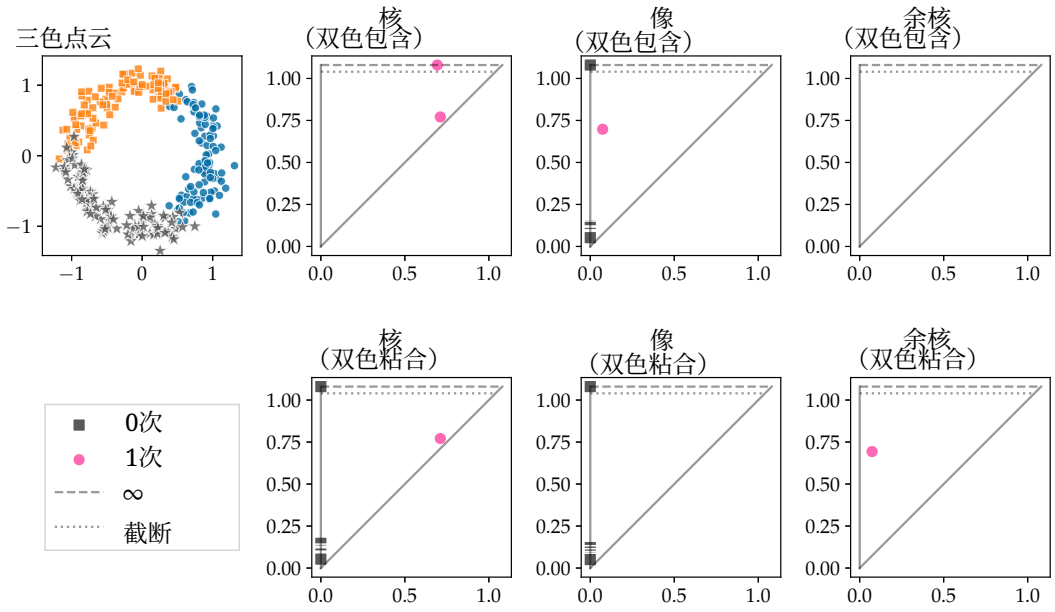


图 1：三色点云中 2-色包含映射和 2 色粘合映射对应的核、像和余核持久图的比较。在尺度 0.4 处，色彩 Delaunay–Čech 过滤的 2-色子过滤包含过滤中的每条边，因为每条边至多包含 2 种不同物种的点。值得注意的是，它只包含一个连通分支，即由所有点共同构成的环。因此，2-色包含映射在 0 次的核或 1 次的余核中都不包含任何信息。另一方面，由每一对物种张成的子过滤的不交并仅包含同一物种点之间的边。因此，2 色粘合映射包含三个连通分支，其中两个通过 2 色粘合映射与第三个分支对应。这两个连通分支在 0 次核中产生两个特征（死亡时间为无穷大）。余核在 1 次包含一个圈，因为这个圈在考虑任意一对物种时都不存在，而在将所有三种物种一起考虑时才会出现。

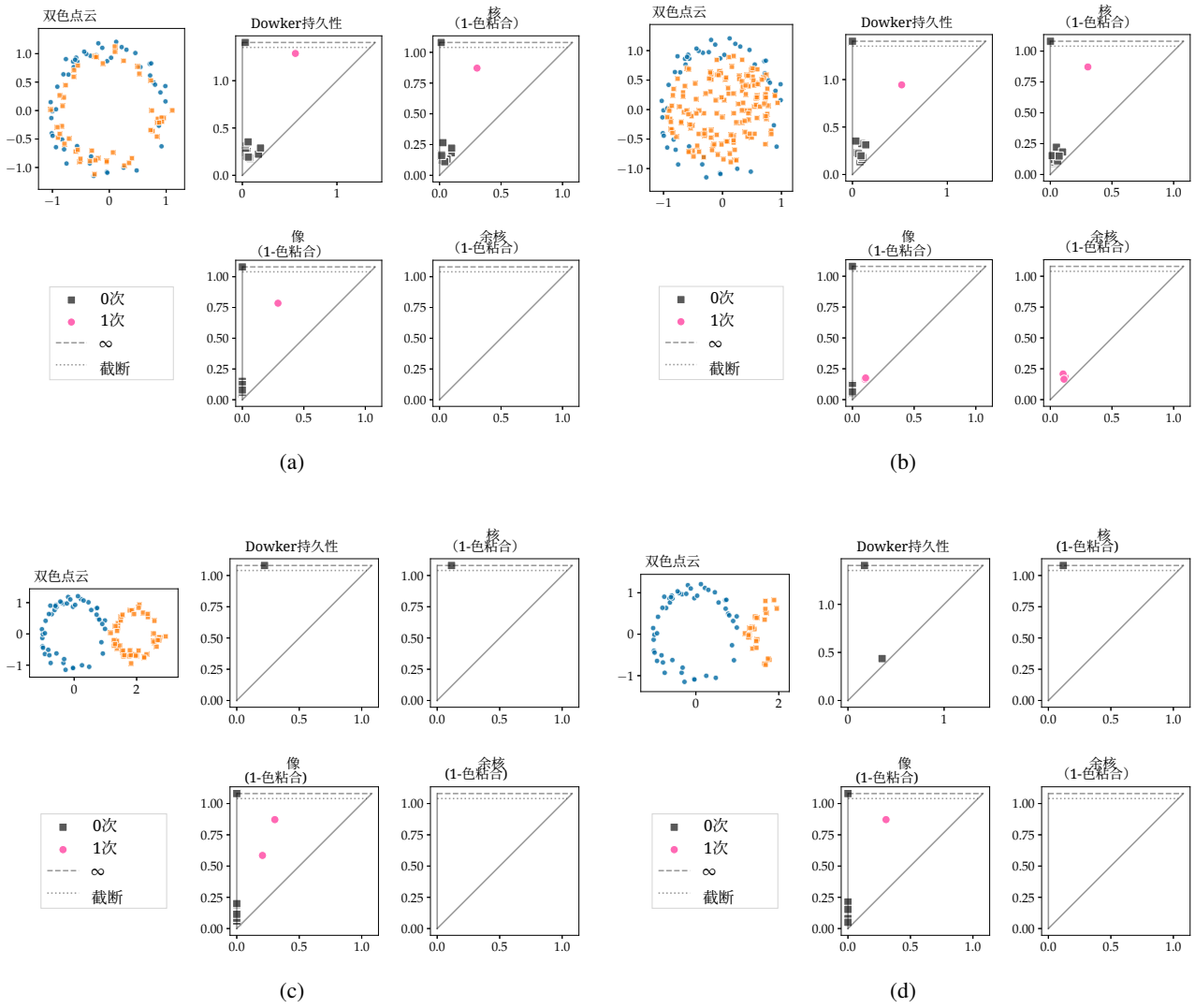
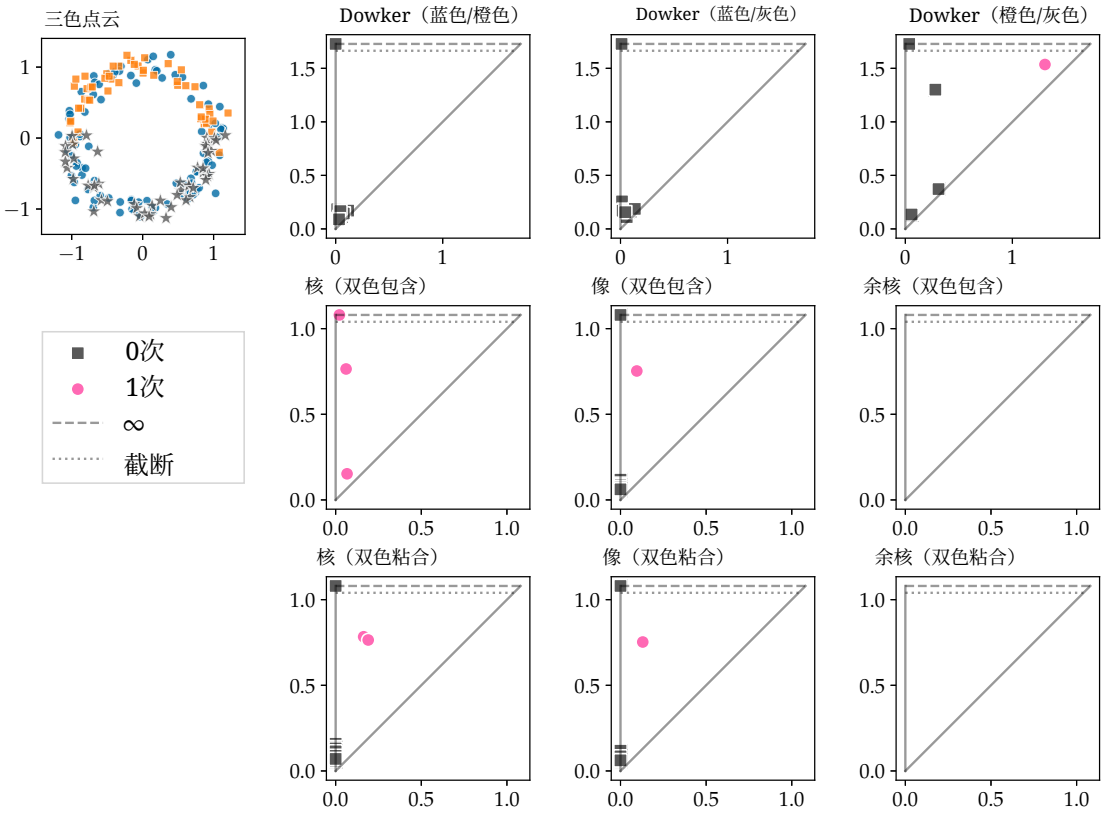
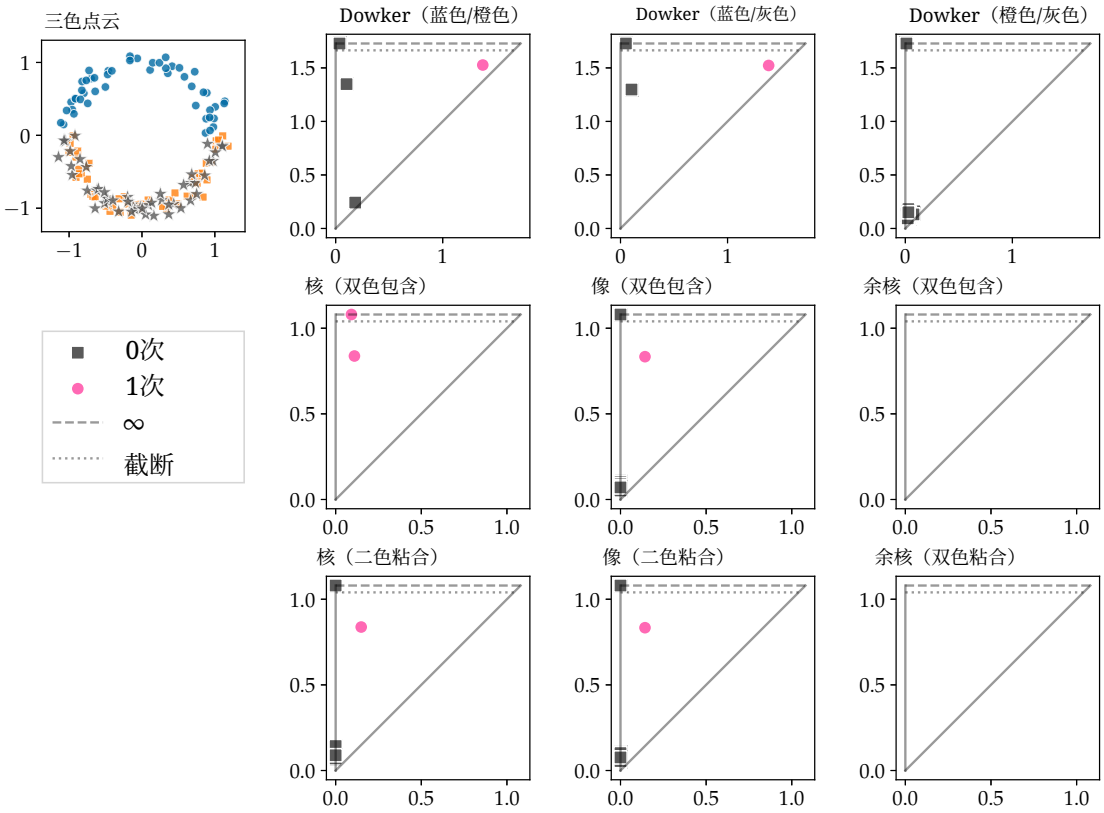


图2：与 Dowker 持久性以及 1-色粘合的核、像和余核对应的图，用于说明主文稿中图1(c)表格的前两行。左上和右上示例（a）和（b）可以通过 1-染色粘合映射的持久图加以区分，但无法通过相应的 Dowker 持久图加以区分。同样，左下和右下示例（c）和（d）可以通过 1-染色粘合映射加以区分，但无法通过相应的 Dowker 持久图加以区分。

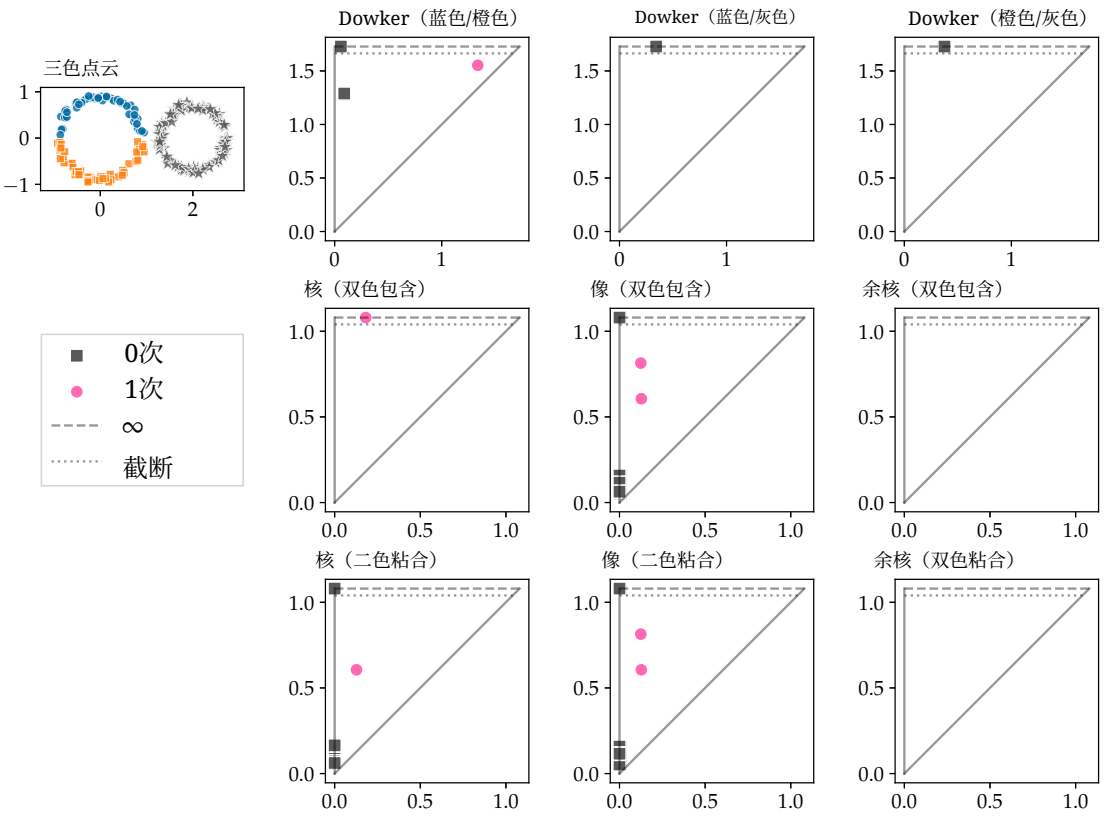


(a) 三种两两颜色组合各自形成一个环，但在任何 Dowker 图中都无法清楚地看到这一点。当同时考虑全部三种颜色时，双色粘合映射会将其中两个环与第三个环识别。这种识别产生了1次核中的两个显著特征（左下），以及像图中圈的存在（中下）。虽然双色包含映射的像图（中间一行，中间）也包含一个显著的1次特征，但核中有一个具有无限寿命的1次特征。该特征对应于过滤中由至多两种颜色的边路径形成、且从未被至多两种颜色的三角形填满的环。因此，对该特征的解释不太明确。也不清楚如何从双色包含映射的图中恢复数据中存在多个重合环这一事实；这些环各自由两种颜色形成。

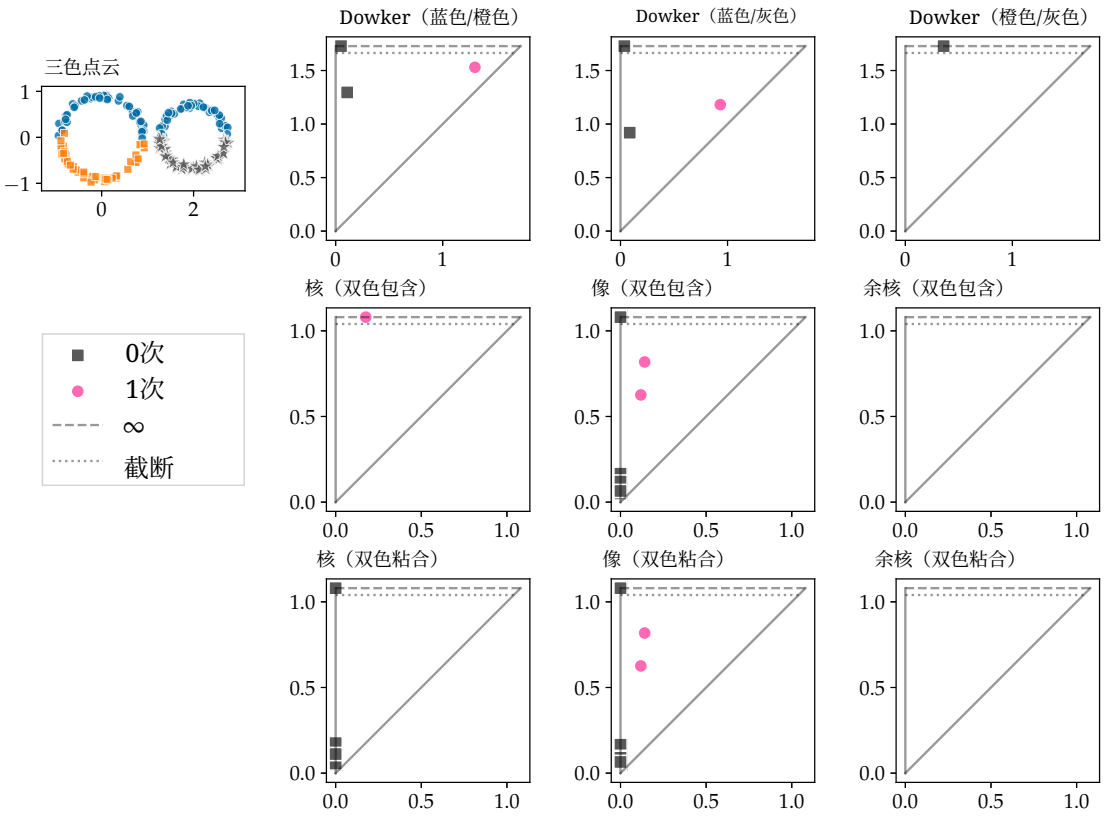


(b) 在此例中，橙色/蓝色和灰色/蓝色（分别对应正方形/圆形和星形/圆形）的物种对各自形成一个环，在 Dowker 图中无法清晰辨认。当将三种颜色放在一起考虑时，这两个环会通过 2-色粘合映射合并。这使得 1 次核（左下）的显著特征只有一个，像图（中下）的显著特征也只有一个。对 2-色包含映射所得图（上排）的解读与前一个例子 (a) 类似。

图3：这些例子展示了主文稿中图 1(c) 表格的第三行。这里展示的两种点云构型，可以通过三种成对组合对应的三个 Dowker 图，或通过 2-色粘合映射的核图加以区分，但无法通过 2-色包含映射所得的图加以区分。对 2-色粘合映射所得图中的信息也更容易解读，因为这些图中的每个特征都直接对应于某种颜色组合中的一个拓扑特征。

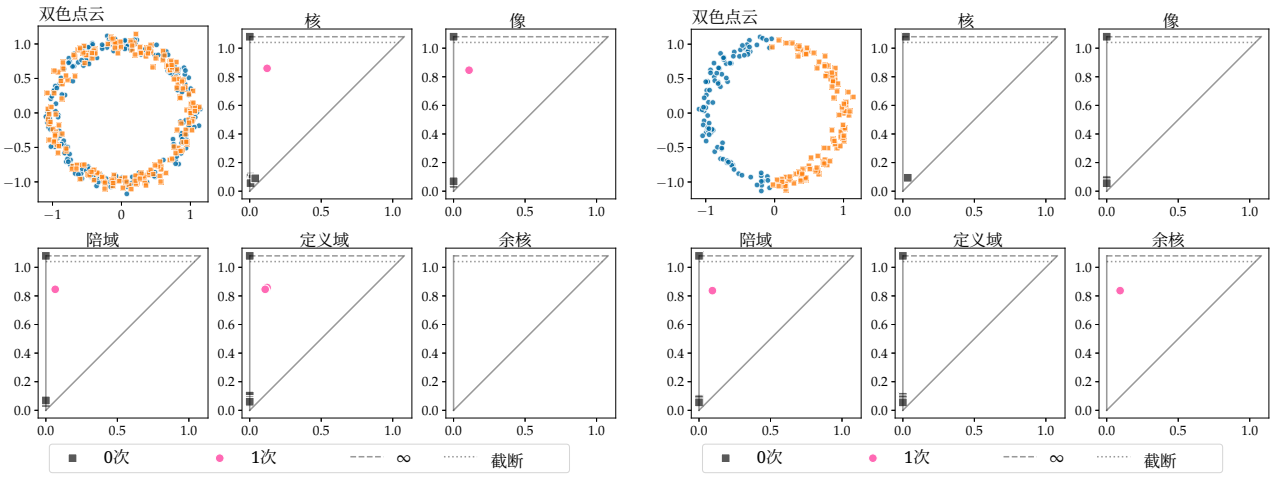


(a) 数据中有两个显著的环，其中一个由橙色/蓝色（正方形/圆形）物种对构成，另一个则由灰色（星形）物种构成。Dowker图都无法很好地捕捉这些信息。灰色环也出现在灰色/蓝色和灰色/橙色组合中；当同时考虑这三种颜色时，双色粘合映射会将这两者对应起来。因此，在1次核中出现了一个显著特征（左下行）。无论是通过双色粘合（下行中间）还是双色包含（上行中间）映射，像图中都能看到灰色环和橙色/蓝色环这两个显著特征。



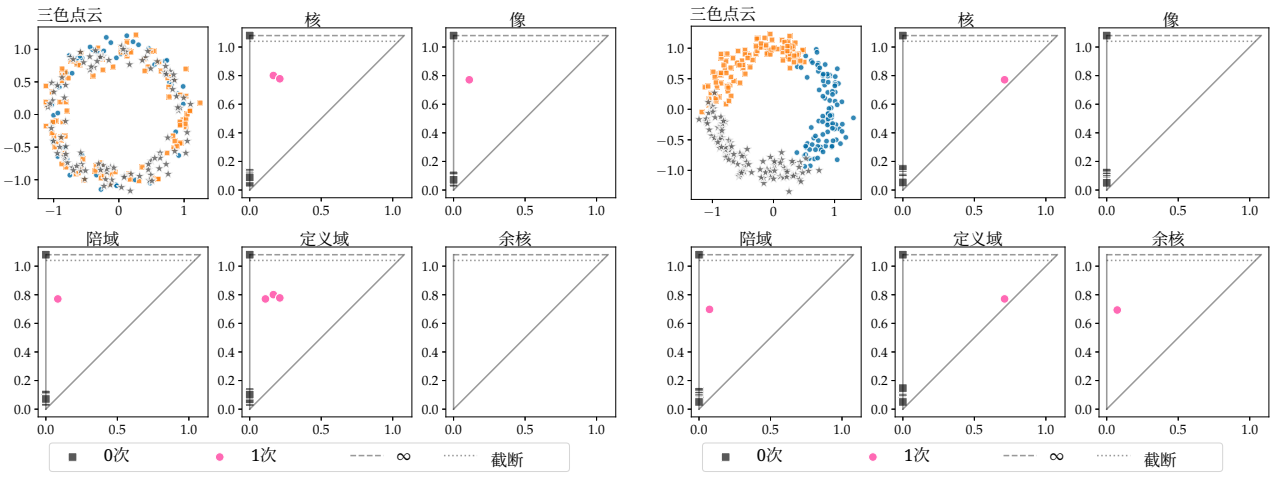
(b) 此例与 (a) 的不同之处在于，灰色环被灰/蓝环取代，也就是说，灰色/橙颜色对中不再有明显的环。因此，1 次的双色粘合映射的核中不再有明显特征（左下）。

图4：这些例子展示了主文稿中图1(c)表格的第四行。这里展示的两个点云构型，既可以通过三组颜色对的三个 Dowker 图加以区分，也可以通过双色粘合映射的核加以区分；但无法通过双色包含映射的图加以区分。与图3一样，双色粘合映射的图所包含的信息更容易解读，因为这些图中的每个特征都直接对应于某种颜色组合中的一个拓扑特征。



(a)

(b)



(c)

(d)

图5：正文图2中示例的附加持久图。示例 (a) 和 (b)（分别为左上与右上）展示双色点云，以及相应1色粘合映射的图。示例 (c) 和 (d)（分别为左下与右下）展示三色点云，以及相应2色粘合映射的图。当多个环在全部颜色集合中彼此被等同起来时，其中一个可见于像图，其余则位于核图中（参见 (a) 中的2种颜色情形以及 (c) 中的三种颜色情形）。1-染色粘合映射的余核图刻画了 (b) 中由两种颜色形成、但并非由一种颜色形成的环，而双色粘合映射的余核刻画了 (d) 中由3种颜色形成、但并非由2种颜色形成的环。

参考文献

[1] Allen Hatcher。《代数拓扑》。剑桥；纽约：剑桥大学出版社，2002年。544页。isbn：978-0-521-79160-1 978-0-521-79540-1。[2] Steve Y. Oudot。《持久性理论：从箭图表示到数据分析》。第209卷。数学调查与专著 SURV。美国数学会，2015年。218页。isbn：978-1-4704-2795-5。[3] Herbert Edelsbrunner 和 Dmitriy Morozov。“持续同调。”载于：《离散与计算几何手册》。第3版。Chapman and Hall/CRC，2017年。isbn：978-1-315-11960-1。[4] Gunnar Carlsson。“持续同调与应用同伦论。”2020年。doi：10.48550/arXiv.2004. 00738。arXiv：2004。00738 [math.AT]。[5] Herbert Edelsbrunner。“球并及其对偶形状。”载于《第九届计算几何年度研讨会论文集》。1993年，第218–231页。[6] Boris Delaunay。“关于空球。”法文。载于《苏联科学院通报。数学科学类》6（1934年），第793–800页。[7] William Crawley-Boevey。“逐点有限维持久模的分解。”载于《代数及其应用杂志》14.5（2015年）。issn：0219-4988。doi：10.1142/S0219498815500668。[8] Serge Lang。《代数》。第211卷。数学研究生教材。纽约，NY：施普林格，2002年。isbn：978-1-4612-6551-1 978-1-4613-0041-0。doi：10.1007/978-1-4613-0041-0。[9] Abhinav Natarajan 等人。“色彩德洛内三角剖分的莫尔斯理论。”2024年。doi：10.48550/arXiv. 2405.19303。arXiv：2405。19303 [math.AT]。[10] Ranita Biswas 等人。“色彩德洛内镶嵌的规模。”2022年。arXiv：2212。03121 [math.CO]。[11] Yohai Reani 和 Omer Bobrowski。“耦合阿尔法复形。”载于《计算几何杂志》14.1（2023年10月10日），第221–256页。issn：1920-180X。doi：10.20382/jocg.v14i1a9。[12] Sebastiano Cultrera di Montesano 等人。“色彩阿尔法复形。”载于《数据科学基础》8（2026年），第30–62页。doi：10.3934/fods.2025003。