Cite this: DOI: 00.0000/xxxxxxxxxx

# 形状几乎就是一切！持久同调特征（PHFs）是信息丰富的输入，有助于高效的分子机器学习。†

Ella M. Gale*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

三维形状对化学很重要，但究竟有多重要？当输入简单且与问题契合时，机器学习的效果最佳。与机器学习中通常使用的数据集相比，化学数据集往往规模很小，因此我们需要充分利用每个数据点。持久同调用于拓扑数据分析，可衡量点云在不同尺度下的拓扑形状属性。本文探究持久同调能够从分子结构中捕捉到哪些信息，并创建持久同调特征（PHFs），以编码分子的形状，同时舍弃原子标签、化合价、电荷、化学键等大部分符号细节。我们在一系列化学数据集上展示了 PHF 的实用性：QM7、亲脂性、Delaney 和 Tox21。PHF 的表现与最佳基准相当。PHF 信息密度很高，且比迄今发现的其他编码方法小得多，这意味着机器学习算法的能效更高。尽管舍弃了大量化学细节，PHF 仍取得成功，这凸显了化学中有多少内容可以简化为拓扑形状。

arXiv:2304.07554v1 [cs.LG] 15 Apr 2023

## 1 引言

什么是化学键？原子和化学键的存在早已被提出假说；人们发现原子具有物理实在性，但化学键并非如此，[1–3] 由此衍生的官能团也并非人们常以为的那种独立存在的对象。[4] 化学键对化学家来说是一种实用的简化方式，而为描述化学键制定的简单规则往往也存在反例。那么，分子究竟是什么？对于某个特定问题，众多可供计算机处理的分子描述中，哪一种最合适？最后，某种特定描述的成功或失败，能告诉我们有关特定问题的化学知识吗？

机器学习最恰当的描述，是发现、操纵并利用高维数据中的模式。分子机器学习（MML）是将机器学习方法应用于化学问题。特征化（即输入机器学习算法的分子描述）的选择至关重要，却并非显而易见。可用的特征化方式有很多，例如基于分子性质的指纹[5–8]、物理化学性质[9,10] 、密度泛函理论[11] 等。化学本质上至少是一个三维问题。分子形状对其功能至关重要，例如药物作用、手性、光学性质、电子结构、反应性、可用的反应路径、受挫现象以及许多其他化学概念，本质上都与形状有关。我这项工作的一个关键动机是：如果向机器学习算法提供正确且最便于使用的信息，也就是让问题尽可能简单，那么算法就能以更高效、更稳定的方式解决问题。对于涉及三维性质的问题，这需要采用三维特征化，或能够编码三维性质。如何将这种形状编码为机器可读的格式，以供机器学习使用，仍是一个尚未解决的问题。

早期的人工智能，尤其是化学领域的人工智能，会使用符号作为输入来构建符号人工智能。最近，人们发现，要映射一个大型化学反应数据集中不到一半的数据，需要 10,000 多条规则，[?] 这远超人类专家手工编码所能胜任的数量。非符号神经网络领域近期取得的巨大成功（例如，参见[12–14]）以及其他机器学习技术，表明了一种不同的方法：对于某些任务，我们可以摒弃符号。值得探讨的是，这种破除传统观念的做法在化学领域能走多远：能否完全不借助以人为中心的符号，进行任何有用的机器学习？如果去掉原子标签、化学键、电荷、化合价及所有其他符号，仅将分子简化为一组空间中的点，会发生什么？这种方法能否让我们学到任何有用的东西?

本文将把一种基于持久同调描述分子的新方法应用于分子机器学习问题。这种新方法以化学家不熟悉的方式编码分子的三维结构和形状，而不直接编码原子或键的属性。这种方法的成功或失败有助于我们加深对化学问题的理解。这种特征化还可以与其他特征化有效结合，以极高的效率取得最先进水平的结果。

由于持久同调领域十分新兴，在化学界鲜为人知，而且我们认为它有望在化学科学中发挥重要作用，因此本文将全面回顾相关文献，并阐释这些技术的应用，帮助化学家将这些方法纳入自己的工具箱。第 1.1 节将概述拓扑学在化学中的既有应用，以便我们区分这种拓扑学

化学学院，布里斯托大学，坎托克斯克洛斯，布里斯托，英国。电子邮件：*ella.gale@bristol.ac.uk*

† 电子补充信息（ESI）可用：[此处应包含任何可用补充信息的详细内容]。参见数字对象标识符：00.0000/00000000。

方法与其他方法的不同之处；第 2 节将介绍持久同调，第 3 节则通过化学实例展示其应用，完整的数学细节见附录。第 1.6.1 节概述所研究的化学问题，第 1.1 节介绍可用的特征化方法并回顾机器学习模型。这些实验的结果见第 5 节。示例代码见附录 7.1。完整的数学细节见附录 8。

### 1.1 分子机器学习的特征化

#### 1.1.1 特征化方法

要将分子输入机器学习算法，首先必须找到一种方法，对与问题相关的分子特征进行编码（这就是特征化的过程）。没有哪一种特征化方法优于其他方法；具体选择取决于问题类型，而所选的特征化方法也会决定训练模型的选择，见表 1。

尽管 SMILES、SMIRKS 和 SELFIES 等基于字符串的编码是为输入计算机而设计的，但它们并不特别适合机器学习；虽然在训练用于语言处理的机器学习模型方面已有进展[15]，，但生成的分子往往显得“奇怪”或“不符合化学规律”。对此的一种解释是，尽管化学式看起来像遵循语法规则的语言，但实际上未必如此。[16,17] 我认为，分子的三维特性是一项至关重要的特征，必须将其作为输入；如果不对三维结构进行某种编码，算法就无法学习键角和扭转角，而这些信息对于预测有效的化学结构是必需的。

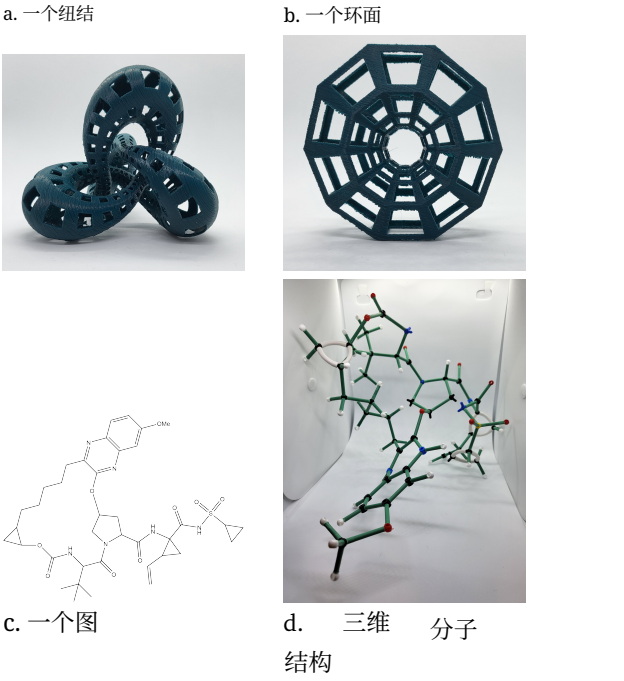



**图1** 拓扑学研究的对象：左上：纽结（纽结拓扑）；右上：环面（同调）；左下：图（图拓扑）；右下：三维分子结构。请注意，分子图是二维表示，只包含分子之间的连接关系，不包含立体化学，因此无法体现分子的三维形状。图中展示的是格卡瑞韦，这是一种用于治疗丙型肝炎的大环药物；图中有意未绘出其立体化学，因为分子图中不包含立体化学信息，因此无法利用图拓扑进行拓扑分析。

### 1.2 化学中的拓扑学

拓扑学是数学的一个分支，主要研究形状。具体而言，它关注那些在连续变形下保持不变的性质（例如拉伸、扭转和变形等操作，而非切割、穿线等操作）。拓扑学中的若干思想已见于化学和生物学文献（尽管化学家使用的“拓扑学”一词，通常比数学家使用时涵盖的范围广得多[18]）。为阐明这一领域的贡献，我们将简要回顾拓扑学思想及其对化学的影响。为区分各个子领域，在提及拓扑学时，我们会注明其所属领域。

#### 1.2.1 多面体拓扑学

古希腊人认识到五种多面体，它们都与化学密切相关，分别是：四面体、立方体、八边形、二十面体和十二面体。它们出现在晶体结构中（四面体孔和八面体孔）、分子几何结构以及分子中（如立方烷、$SF_6$）。关于多面体异构化的拓扑学研究[19] 描述了多面体的动态性质，例如四面体被压扁成扁平的菱形，或扁平的正方形变形为菱形；这些性质可能有助于描述化学反应或有机金属化合物。

**表 1** 分子机器学习中使用的常见和新型特征化方法。NN 指前馈神经网络，RF 指随机森林，KRR 指核岭回归。*此方法基于几何机器学习，目前正在我们团队内开发。

| Type | 维数<br>输入 | 维数<br>特征化 | 数学<br>基础 | 示例 | 示例<br>模型 |
|---|---|---|---|---|---|
| 化学式 | 一维字符串 | 1-D | N/A | SMILES、SMARTS模式、SMIRKS | MTR 或文本神经网络 |
| 分子指纹 | 一维字符串 | 1-D | N/A | 扩展连接指纹、MACCS | 神经网络、随机森林、核岭回归 |
| 物理化学性质 | N/A | 1-D | N/A | RDKit | 多任务回归 |
| 库仑矩阵 | 3-D | 2-D | 图拓扑 | 库仑矩阵 | DTNN |
| 库仑矩阵的特征值 | 3-D | 1-D | 图拓扑 | CM 特征值 | 多任务回归 |
| 图 | 2-D | 2-D | 图拓扑 | 图卷积 | 图卷积的 |
| 图 | 2-D | 2-D | 图拓扑 | 编织卷积 | 编织神经网络 |
| Grid | 3-D | 1-D | 几何 | Grid | MTR |
| 二十面体球面/球面* | 3-D | 3-D | 几何 | 二十面体球面映射 | 球面神经网络 |
| PHF（本研究） | 3-D | 1-D | 持久同调 | PHF | 多任务回归 |

### 1.2.2 纽结拓扑

纽结拓扑是代数拓扑学的一个分支，研究由一圈绳子构成的纽结（见图 1），例如三叶结和凯尔特结（注意：这里指的是数学纽结，与船上或鞋带上用两端绳子打成的结不同），并启发了若干有趣的化学结构。[20] 纽结拓扑的一个分支是链环拓扑，研究一组彼此不相交、但可以相互链接的纽结；这同样启发了化学结构，例如 DNA链环。? 纽结通过投影到二维空间中进行分类和比较（想象三维分子结构模型投射在墙上的影子），并对各种交叉和弯曲进行计数。[21]

尽管蛋白质或 DNA 这类长分子在纽结拓扑上对应的是平凡而乏味的纽结结构——平凡结（一条两端开口的一维线），化学家仍扩展了纽结拓扑这一数学概念，以研究 DNA 和蛋白质结构。例如[22] 将纽结理论扩展至拟纽结，或扩展至具有自由末端的蛋白质；其中的重叠和交叉可像纽结拓扑一样提供有关结构的信息。我认为，电路拓扑是纽结拓扑的一种形式。纳入硫桥等分子内相互作用后便形成电路，其纽结拓扑比平凡结更复杂。Scalvini 等人最近表明，这些电路与蛋白质折叠有关。[23] 尽管化学家将这些方法称为拓扑学方法，但有人指出，它们实际上研究的是长程几何。[18] DNA 超螺旋以便压缩到狭小空间，通常被称为一种拓扑性质[24], ，但它并非真正的数学拓扑性质，因为解开超螺旋属于非破坏性变形。

### 1.2.3 图拓扑。

在化学领域，「拓扑学」一词大多指图拓扑，这是一个不同的子领域，研究分子图（即化学式）之间的连接关系，见图 1c。化学家使用「拓扑不变量」一词时，通常指这些图或子图的不变量*。拓扑不变量是在拉伸、弯曲和形变过程中保持不变的量；它们无法直接测量，但应当是定量的，因此我们只能进行计数，而这些特征必须能够从物体内部加以验证；我们无法「拉回」并从外部观察其结构。本文提出的新型拓扑特征并非来自分子图上的图拓扑。

网络拓扑研究网络的形状及其连接关系中的不变量，并已被用于尝试根据标量场和拓扑连接关系来定义化学键†。[1] 许多分子特征化方法都基于应用于分子图的图论方法，因此必然涉及图拓扑。有人提出，分子图必然是二维描述[25]。然而，将化学式写成二维图后，可以使用邻接矩阵或连接矩阵来构造一个拓扑图不变量，也就是一个定义连接关系且不受原子置换、图标记和距离影响的单一数值。图可以用谱方法进行分析，这已被用于预测生物学性质、[26] 沸点[27] 和无机结构[1] （本文还尝试从拓扑学推导出化学家能够识别的化学键）。

### 1.2.4 代数拓扑及其衍生领域：持久同调与拓扑数据分析

本文所述的新型分子特征化方法，正是从拓扑学这一领域汲取灵感。尽管拓扑学的现代根源可追溯至 1700 年欧拉的工作，且这一领域直到 1900 年代才真正兴起，但从数学发展的角度看，拓扑学是较晚出现的数学分支之一，目前仍由数学家积极推动发展，因此新的研究成果和子领域仍在不断涌现。拓扑数据分析 是应用数学中一个非常新的领域，研究如何将代数拓扑和组合拓扑技术应用于数据分析，尤其是持久同调。持久同调将同调（参见下文）推广到点云（如原子），关注的是能够在多个长度尺度上持续存在的特征。拓扑数据分析利用持久同调来考察高维数据中的模式，见[28–30]，并已被应用于[31] 化学领域，以探寻 Delaney 溶解度数据集[32]中的模式。

* 不变量是指在某种操作下保持不变的量。

†在拓扑学领域中，这些在技术上称为键；但由于它们并不总是与化学键相对应，这种术语容易引起误解。

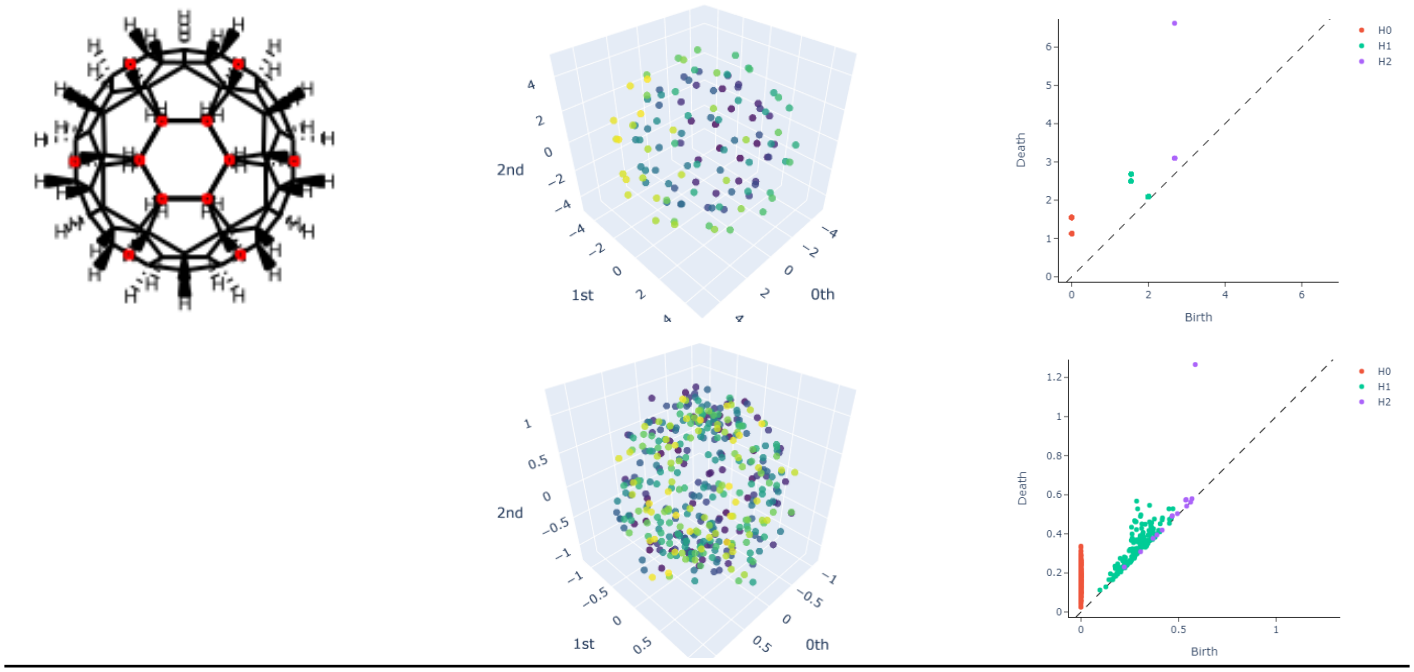



**图 2** 巴基球在同调上与球面相似，也可以看作是均匀噪声球面点分布的简化形式，或者更确切地说，是噪声球面的晶态版本。随着振动加入这一模型，显然，持久图会发生“扭曲”，打破点的对称性和重数，并使图形逐渐趋近于噪声球面。上：左：巴克明斯特富勒烯结构；中：巴克明斯特富勒烯点云；右：持久图。下：左和中：由球面上的均匀噪声点构成的点云；右：噪声球面的持久图。

### 1.3 什么是代数拓扑?

形式化的数学描述见附录；此处将以非形式化、图示化的方式概述相关概念。代数拓扑关注比较物体（维数不限）的形状，而同调是一种比较物体“曲面”（严格来说是流形）形状的方法。直线与曲线之间的差异等几何关系不会保留下来，但各部分之间的连接关系会保留下来；孔洞或空腔的数量则用于对形状进行分类。测量“孔洞”——也就是字面意义上缺少某物的部分——并不容易；而且我们无法退后观察物体，只能依据曲面的性质来找出其整体结构。这可以通过寻找沿曲面的路径，或构建曲面的覆盖来实现。请注意，在拓扑学中，三维物体的形状指的是该物体外部的曲面。分类依据是物体中的孔洞数量，因此球面和环面属于不同的物体类别；这些类别由贝蒂数确定。例如，图 1a 中线框环面的曲面有一个孔洞（不过请注意，若计算三维打印结构中的孔洞数量，所得数值可能会大得多）。如果两个物体具有相同的贝蒂数，则称它们同伦等价。例如，茶杯与甜甜圈同伦等价，因为二者都只有一个孔洞；如果它们由柔软、有弹性的材料制成，就可以在不破坏或撕裂材料的情况下，将其中一个平滑地变形成另一个（另见图 12）。

### 1.4 同调

流形是一种类似于欧几里得空间的空间。同调研究流形之间的映射。我们期望这些空间是连通的，并允许它们像由可拉伸、可弯曲的材料制成那样发生变形（拓扑变形）。在这类操作中保持不变的任何性质都是拓扑不变量。切割、粘贴和粘合等操作并非拓扑不变。‡ 保持不变的量包括任意维数孔洞的数量。0 维孔洞对应于连通，因此连通区域的数量也是拓扑不变量。同调群 $H_k(X)$ 用于度量一个空间中的孔洞和连通区域的数量。贝蒂数与之相关，它们是同调群的秩，也表示在物体被切分成彼此分离的区域之前可以进行的切割次数。例如，开环反应会切开一个分子并改变其同调群，但不会使分子失去连通性。孔洞是某种缺失之物，因此很难想象如何对其进行测量；解决方法是考察曲面上的路径（你可以留在曲面上绕过孔洞行走，却不能穿过它，见附录 8 中的图 12）。

‡ 拓扑学的很大一部分内容，就是通过切割、拼接和粘合来构造形状。

**1.5 持久同调。**

持久同调将同调思想从曲面推广到点云。由于点按定义没有体积，一组点如何形成形状、体积和曲面并不显而易见，尽管这一想法直观上颇具吸引力。解决方法是在每个点上放置半径为 $\eta$ 的球（从而在三维空间中形成体积），并考察其形状。通常并不知道哪个半径最适合表示底层数据点的形状，因此解决方法是使用不同半径的球。这种方法的优点是能够以某种方式描绘同调特征（连通区域和孔洞）的大小；例如，随着球半径增大，孔洞可能会被填满。

同调特征，$H_k$ 与其所在空间的维数相关。连通区域是$H_0$（一维点）。平坦的二维孔洞是$H_1$，$k$ 取 1，这是因为在拓扑学中，二维曲面上的孔洞是一维特征：拓扑学关注的是曲面，而孔洞的表面就是界定其边缘的线段。空腔，例如足球中心的空腔，可视为三维空间中的孔洞，是$H_2$，因为三维孔洞由包围它的二维曲面（足球的外皮）界定。同调特征可以延伸到任意高的维数，但本文只关注三维空间中存在的特征（$\mathbb{R}^3$）。

**1.6 同调特征**

持久性衡量一个特征在多大的长度尺度范围内存在。例如，巴基球中心的大空腔会在很大的半径范围内持续存在。一个连通区域在半径为 0 时出生，并在该连通区域与另一个连通区域合并时死亡（这种方法会自然地描绘出孔洞和空腔，因为当球足够大时，每个对象都会成为一个连通分量）。条形码图绘制了连通区域持续存在的半径范围。持久图绘制了不同同调特征的出生时刻与死亡时刻（见图 2），提供了丰富的信息（打个比方，可以把这个过程视为形状光谱学）。化学键也包含在内，因为与一个原子相连的最短化学键会是半径大于 0 时最先出生的连通特征；随着这些特征合并，化学键和官能团对应的特征之间的距离也会显现，见图 2 中的红点。不过，这项技术无法区分形式化学键与空间上彼此靠近的原子。我认为这是一个优点，因为化学键只是对真实情况的符号化简写，有时反而会遮蔽问题，而非阐明问题。$H_1$ 特征可以用来表示形式化学环的大小，见图 2 中的绿点。最后闭合的特征给出了对象的拓扑类别，因此从图 2 可以看出，巴基球在同调意义上等价于一个球面，其紫色$H_2$ 特征远离对角线。有趣的是，研究多维数据集持久同调的数学家认为（但尚未证明），在较大尺度范围内无法持续存在的特征是噪声，或与数据集的结构无关。[33] 在这些示例中，我们发现，持续时间不长的特征在小半径下与化学上极其重要的特征有关，包括化学键、角度、扭转角和官能团。

基于持久同调的特征编码了一种粗略的形状，类似于填充空间模型§，但它们并未将原子类型纳入形状，因此会“忽略”手性¶，但能保留异构体之间的差异。我也在开发一种基于几何机器学习工具的三维输入方法，该方法能够保留手性（论文即将发表），将这些输入类型结合起来，可能有助于进一步开发特征化方法。

现代数学中的一个关键思想可追溯到黎曼于1854年所作的一次讲座：每个由$n$ 个实数参数化的数学对象，都可以视为$n$维空间中的一个点。图拓扑可被视为‘数据的欧几里得化’[34]，即将数据点移入三维空间以研究其拓扑结构。分子图是分子形式键结构的二维表示。作为化学家，我们常用楔形键或位置来编码分子结构的某些方面，这些化学式或许可以（非正式且非数学意义上地）看作二维多一点；不过需要指出的是，补足这些信息的是人类的感知和化学知识：机器学习算法无法从图中获取这些信息。‖ 图1中的药物绘图省略了其六个立体中心，以展示真实的分子图。随后，对分子图进行拓扑分析，会将这个二维化学式映射到欧几里得空间，并仅保留二维结构及原子之间的连接关系，而不保留距离或角度（这些属于几何学的范畴）。本文提出的技术则是‘离散化欧几里得空间’，具体而言是三维空间；随后依据形状构建图并计算拓扑特征。换言之，以三维结构作为输入，并由于使用持久同调，对距离进行了一定程度的处理，因此也保留了部分几何信息。欧几里得空间与拓扑空间之间的关系具有遗忘性，因此转换过程中会丢失信息，也就“无法”** 直接从拓扑特征还原出分子。

本文将使用持久同调工具研究并编码三维化学结构。通过尽可能全面的文献综述，我认为这是首次将这些工具应用于这些特定问题。将化学结构输入计算机有许多方法，其中一些基于图拓扑。我提出，三维结构对于某些化学学习任务至关重要。本文使用的持久同调编码将对形状性质进行编码，以供机器学习使用。由此得到的持久图为我们提供了有关形状拓扑性质的大量信息。我们将通过对持久图进行向量化和嵌入，将这些信息“提炼”为

§ 旁注：用拓扑学的语言来说，填充空间模型与球棍模型互为对偶，因为球是 0-单形，棍是 1-单形，而填充空间模型由 3-单形（体积）构成。

¶ 缺少原子标签使得难以区分手性对映异构体，但键长差异得以保留，因此手性或许并未完全消失。||这未必完全正确：也有可能在数据足够多的情况下，机器学习算法能够从这些二维表示中学习到有关三维空间的信息。**这就是逆问题。我认为这并非不可能，但需要根据拓扑特征猜测分子结构，并不断完善这些猜测。

**表2** 同调特征（$H_k$）及其非正式定义。化学示例是以该同调特征为主要属性的分子。

| 同调特征 | 描述 | 化学示例 |
|---|---|---|
| $H_0$ | 连通区域，无孔洞 | 任意长链烃 |
| $H_1$ | 一维孔洞：存在于二维空间中，就像纸上打出的一个孔 | 教科书形式苯、血红素（不含铁） |
| $H_2$ | 二维孔洞：存在于三维空间中，就像足球的内部。注意：从拓扑学术语来看，足球是包裹在二维空腔上的二维曲面 | 巴克明斯特富勒烯 |
| $H_k, k > 2$ | 高维空腔。 | N/A |

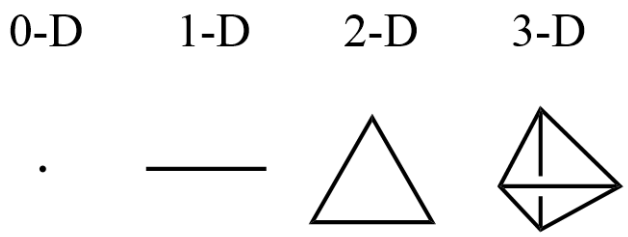



**图 3** 0 维、1 维、2 维和 3 维单形对化学家来说都非常熟悉。

每个分子长度不超过18的向量，从而形成迄今为止最精简的三维化学结构编码。借助这些新颖的化学结构编码，我们可以探究哪些问题主要依赖于三维结构、该方法编码的信息是否足够，以及何时可能需要纳入额外信息。文中还研究了每项任务所需的额外信息，重点关注成本低且易于计算的非拓扑特征。

#### 1.6.1 尝试解决的分子机器学习问题

我们在若干标准的分子机器学习问题上测试了我们的方法，见表3，重点关注药物开发所需的数据，例如溶解度和毒性。

## 2 方法论

### 2.1 以简单分子为例构建持久同调拓扑指标

为了更清楚地阐述本文讨论的概念，我们将简要概述一些小分子的持久图。

### 2.2 点云

首先，我们将原子位置转换为点云，这一过程会去除原子类型、形式键、原子大小、化合价、质量和电荷，只保留形状信息。剩余的唯一化学信息，是机器学习算法能够根据原子间距离推断出的信息。

### 2.3 单形

鉴于我们关注的是曲面，也就是分子经过平滑处理后的底层“形状”，下一步却要用直线和平面对象来量化这种形状，这听起来或许有些奇怪。这是因为拓扑学并不区分曲线与直线，而我们只关注各部分之间的连接关系，因此我们将任务简化，使用称为单形的简单对象来量化这种连接关系。

三维单形对于化学家来说十分熟悉，因为它们就是四面体。这一概念可以扩展到任意空间维数，不过此处我们只关注四种：点（0-D）、线段（1-D）、三角形（2-D）和四面体（3-D），见图 3。请注意，单形的各部分可以是任意长度，不必像图 3 中画的那样等长。多胞体是一种具有“平坦”侧面的几何对象，而单纯形是给定空间中可能存在的最简单多胞体。例如，三角形是有三个边的平坦二维对象；增加边数会形成较不简单的多胞体，而去掉一条边则会变成两条一维线段（也就是说，它不再是多胞体）。理解“平坦性”的最佳方式是：$n$单纯形的切线在$n+1$维空间中彼此平行。也就是说，$x$ 中的一条线段沿其全长在$y$方向的切线都彼此平行；二维空间（$x$，$y$）中的三角形，其整个面上的切线都在 $z$ 方向彼此平行，依此类推（不过请注意，要使这种构造成立，我们必须在二维空间中进行），另见附录 8.4.1。

### 2.4 构建单纯复形

如果以半径为 $\eta$ 的球分别以两个原子（此处将其建模为点）为中心，且球的半径为两点间距离的一半，使球彼此接触或重叠，那么这两个点就连通了。在化学中，这通常对应于一个形式键。计算的下一步是在描述该对象的两个 0-单形（点）之间画一条 1-单形（直线）。对于三个原子，如果球彼此重叠，就在它们之间画一个三角形。单形具有层级结构，因此一个三角形由三个 0-单形（点）、三个 1-单形（线段）和一个 2-单形（三角形）组成。同理，一个四面体由四个 0-单形（点）、六个 1-单形（线段）、四个 2-单形（三角形）和一个 3-单形（四面体）组成[††]。请注意，图 1 中展示的环面实际上是由线单形构成的、建立在环面上的单纯复形。

对于阈值 $\eta$，过程如下：

1. 在任意两个彼此距离小于 $\eta$ 的点之间添加一条线段（1-单形）（即球彼此重叠）；
2. 在任意三个彼此距离小于 $\eta$ 的点之间添加一个三角形（2-单形）；
3. 在任意四个彼此距离小于 $\eta$ 的点之间添加一个四面体（3-单形）。

†† A. 高阶单形可以存在于高维空间中。B. 你可以用点、线段、正方形和立方体创建立方复形，但它们不那么简单，因为包含更多低阶组成部分。

**表 3** 研究的数据集。

| 数据集 | 问题领域 | 问题类型 | 目标 |
|---|---|---|---|
| QM7 | 量子力学 | 回归 | 原子化能 |
| QM8 | 量子力学 | 回归 | 电子光谱与激发态<br>小分子的能量。 |
| Delaney | 物理化学 | 回归 | 水溶解度数据<br>常见有机小分子 |
| FreeSolv | 物理化学 | 回归 | 小分子在水中的实验与计算水合自由能。 |
| 亲脂性 | 药物开发 | 回归 | 辛醇/水分配系数的实验结果（pH 7.4 下的 logD）。 |
| Tox21 | 药物开发 | 分类 | 针对 12 种生物靶标的定性毒性测量 |

| 流形 | 维度 | $\beta_0$ | $\beta_1$ | $\beta_2$ |
|---|---|---|---|---|
| 圆，$S^1$ | 1-D | 1 | 1 | - |
| 球面，$S^2$ | 2-D | 1 | 0 | 1 |
| 环面，$T^2$ | 2-D | 1 | 2 | 1 |

**表 4** 拓扑曲面（流形）及其贝蒂数示例。注意：圆是球面的二维版本，因此圆的边界是球面的一维版本。

结果是一个单纯复形（由单形构成的连通对象）。

我们用定义单形的点来表示单形，因此，连接$a$ 和$b$ 的线段写作 $[a,b]$, ，这实际上是满足“平坦性”约束（另见附录 8.4.2）的$a$ 和$b$ 之间所有点的集合。

### 2.5 Vectoris-Rips复形的构建

在不同的球半径下创建一系列单纯复形，并将它们组合成一个称为 Vectoris-Rips复形的分层复形。

### 2.6 条形码与持久性图

生成元出现时所对应的过滤值称为出生，而其消失时所对应的过滤值称为死亡，并且

$$persistence = death - birth$$

我们可以将生成元的持久性绘制成持久性图（见图2和图7）、条形码图或贝蒂曲线（见图6）。

持久图是一种简洁易懂的方式，可在过滤发生变化（即球的直径增大）时，表示条形码图中的信息（尤其是点很多时）。出生时刻指特征的产生，死亡时刻则指特征消亡（与另一个特征合并）。出生—死亡线上的点表示产生后立即消失的点，而线上许多点对应的是非持久性特征，被视为噪声。[33] 特征越持久，就被认为越重要。在一系列过滤范围内持续存在的点远离这条线；我们通过观察这些点来读取持久图，从而确定存在哪些同调群生成元（即有多少个簇、孔洞和空腔）。

### 2.7 持久性图的向量化

持久图便于人类解读，但不太适合输入机器学习算法。为此，我们使用从持久图中导出的五种特征：

1. 持久熵[35],2. 点的数量,
3. 瓶颈距离，[36]4.
Wasserstein振幅，[36]5. 持
久性图像，[37]

每个特征都是三维的，因此每个输入对应一个长度为 18 的输入向量。关于这些特征如何计算的说明见附录 8。

这种输入格式信息密度很高，非常适合机器学习。MML 的输入存在两个问题：输入向量通常较为稀疏（即包含许多零），而且无论分子的大小和复杂程度如何，每个分子的输入向量都必须相同；通常的做法是以数据集中最大的分子来确定输入大小，这也会增加稀疏性。例如，神经网络很难从稀疏数据集中学习。[38]

### 2.8 计算细节

数据集取自 MoleculeNet[39], ，并使用 DeepChem 中可用的特征化方法进行特征化，包括扩展连接指纹、网格、图卷积、MACCS、RDKit、库仑矩阵以及库仑矩阵的特征值（CM_eig）。使用 RDKit 创建并最小化分子结构。[40] 使用 Giotto-TDA 拓扑数据分析模块[41,42]生成 PHF 特征。TDA 输出特征保存为 HDF5 文件。所有特征均在 DeepChem 中使用适用于各特征的标准基准模型[39]进行训练；例如，PHF 和其他一维特征输入到每层含 1000 个单元的两层前馈神经网络中。图特征化结果输入图卷积神经网络。所有对照实验均使用 DeepChem v2.3 中的配置和模型进行，具体如 MoleculeNet 论文[39]所述。

### 2.9 软件

所有代码均位于 GitHub 项目“化学中的图与拓扑学”中[43] ，该项目基于 Giotta-tda[41,42] 和 RDKit[40]构建。附录 7 中提供了示例代码。此软件包包含用于提取分子结构特征的代码，以及用于提取

| Term | 非正式定义 |
|---|---|
| 拓扑不变量 | 在拓扑变形下保持不变的性质。 |
| 同伦等价 | 拓扑形状相同，其中一个可以连续变形为另一个，也就是说，你可以将 $X$ 映射到 $Y$。 |
| 同胚等价 | 与同伦等价类似，但你可以将$X$ 变形为$Y$，再将$Y$ 变形为$X$。 |
| 连通空间 | 不是两个互不相交的非空开集的并。苯不是单个连通空间，在图4中的单纯复形$X_0$ 和$X_1$ 中，它是单纯形中的一个连通空间复形$X_2$, $X_3$, $X_4$ 和$X_5$。 |
| $H_k$ ($X$) | $k$ 维$X$的同调群，生成围绕维数为$k$的孔洞的路径。 |
| 贝蒂数 $\beta_k$ | $H_k$的秩，即在$X$ 不再连通之前可以切割的次数。 |
| $\beta_0$ | 连通分量的数量。 |
| $\beta_1$ | 一维“平坦”孔洞的数量， |
| $\beta_2$ | 二维孔洞（空腔）的数量。 |

**表 5**术语

表 **??** 中所列数据集特征的脚本。要按照 MoleculeNet 论文[39], 中的方式运行代码，需要使用 DeepChem 2.3[39], ; [44] 中提供了该软件的一个分支版本以及验证脚本。另附有使用 DeepChem 2.6 的更新版本。*

## 3 苯示例

为了阐明第 2 节中的概念，并探究持久同调特征化保留了哪些信息，下面我们将通过一种非常简单的分子——苯，逐步讲解所采用的方法。

### 3.0.1 创建 Vectoris-Rips复形。

首先，我们按照图4a所示，将原子构成点云。然后，我们增大 $\eta$ 的值，以构造一系列复形，并记录同调群（$H_k$）的生成与消亡位置。贝蒂数 $\beta_k$ 表示给定同调类型的特征数量。

我们记录复形$X_n$中的所有单纯形特征。
如果 $\eta$ 小于$r_{CH}/2$，我们就只有点：

$$X_0 = \{C_1, C2, ...C_6, H_1, H_2, ...H_6\}$$

贝蒂数为 $\beta_0 = 12$, $\beta_1 = 0$, $\beta_2 = 0$，也就是说有12个连通分量，复形$X_0$ 由一组点组成。这里，为了便于理解，我们用原子为它们标注；但在该方法中，这些是等价的。

如果 $\eta$ 大于$r_{CH}/2$但小于$r_{CC}/2$，碳氢原子之间会形成1-单形：我们有12个点（0-单形）和6条线段（1-单形）：

$$X_1 = \{X_0, [C_1, H_1], ...[C_6, H_6],$$

此时我们有6个连通分量（簇）（对应于碳氢键中碳点和氢点组成的簇），见图4c，因此 $\beta_0 = 6$, $\beta_1 = 0$, $\beta_2 = 0$，从拓扑角度看，这是6个簇。

如果 $\eta$ 大于$r_{CC}/2$，我们就有 12 个点、12 条线段和 12 个三角形：

$$X_2 = \{X_0, X_1, [C_1, C_2], ...[C_6, C_1],$$
$$[H_1, C_1, C_2], [H_1, C_1, C_6], ...$$
$$[H_6, C_6, C_5], [H_6, C_6, C_1]\},$$

CCH三角形相互重叠形成一个环，见图4。由于这是一个一维曲面，我们现在知道该结构中有一个二维孔洞，贝蒂数为 $\beta_1 = 1$。因此 $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 0$。此时，苯与一个圆同伦等价（$S^1$），见表4。从概念上说，这对化学家而言略显奇怪，因为苯环结构是由相互重叠的HCH三角形构成的，而非由碳碳键组成的环。

增大 $\eta$，我们将 CCC 键角作为 2-单纯形加入：

$$X_3 = \{X_0, X_1, X_2, [C_1, C_2, C_3], ...[C_6, C_1, C_2]\}$$

见图4d。由于拓扑结构本来就是一个圆，因此从拓扑学角度看，这项添加不会带来任何变化。我们有 12 个点、24 条线段和 18 个三角形。

当 $\eta$ 为苯分子直径的一半时，我们会添加跨越环的连接关系以及新的三角形：

$$X_4 = \{X_0, X_1, X_2, X_3, [C_1, C_4], ...[C_3, C_6], [C_1, C_3, C_4], [C_6, C_2, C_3]\}$$

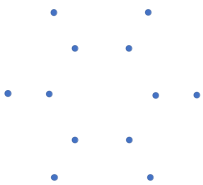
贝蒂数为 $\beta_1 = 1$、$\beta_1 = 0$、$\beta_2 = 1$，环随之消亡。最终，环两侧的氢原子连接起来：

$$X_5 = \{X_0, X_1, X_2, X_3, X_4, [H_1, C_1, C_4, H_4], ...[H_6, C_6, C_3, H_3]\}$$

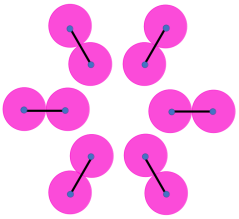
此时所有元素都彼此连通，苯同伦等价于球面，并形成单个簇，见图4e，$\beta_0 = 1$,$\beta_1 = 0$, $\beta_2 = 0$ 由于可以将这种连接想象成跨越顶部，因此在三维空间中连接起来的是孔洞。如果设想氢原子略微向上弯曲并偏离平面，那么显然会在碳原子平面上方形成一个三维空腔。RDKit 版本的结构只有五位小数的精度（事实上，六个 C-H 化学键的数值并不完全相同，但在持久性图中看不出来，因为它们彼此重叠），这可以解释空腔的出现。† 我们希望这个例子也清楚地说明了拓扑学与组合学之间的联系。最终结果是维托里斯-里普斯复形，它是一组由复形构成的集合，如图5所示。

* DeepChem 2.3 和 2.6 之间的底层代码有较大改动。

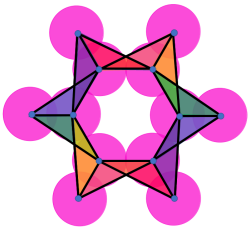
†不过，如果苯完全处于平面内，那么根据氢原子高出平面的距离趋近于零这一推理，会存在一个无穷小的点，使空腔出现又消失。这会出现在出生=死亡线上。



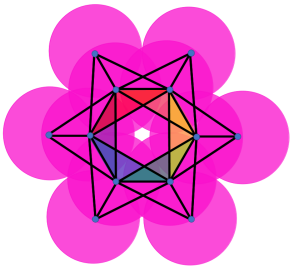
a. $X_0$ 是12个0-单纯形（点），$\beta_0 = 12, \beta_1 = 0, \beta_2 = 0$，类：簇



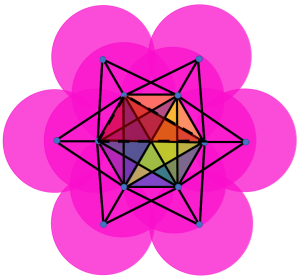
b. $X_1$：添加了 6 个 1-单形，$\beta_0 = 6, \beta_1 = 0, \beta_2 = 0$，类：簇



$X_2$：添加了 6 个 1-单形、12 个 2-单形，$\beta_0 = 1, \beta_1 = 1, \beta_2 = 0$，类：圆



c. $X_3$：添加了 6 个 2-单形，$\beta_0 = 1, \beta_1 = 1, \beta_2 = 0$，类：圆



e. $X_4$，$\beta_0 = 1, \beta_1 = 0, \beta_2 = 0$，类：簇

**图 4** 随着球的大小增加，单纯复形是如何构建的。

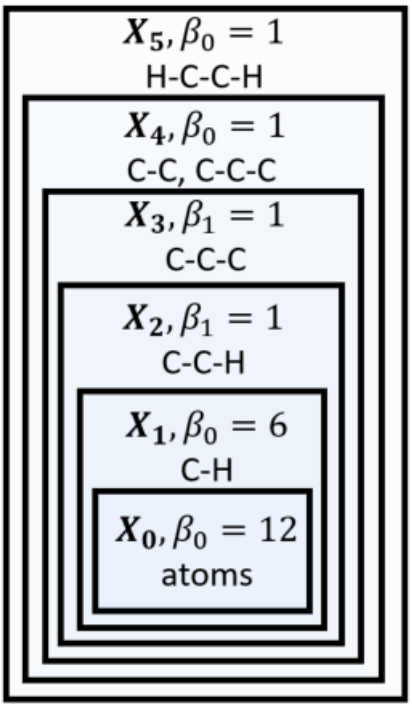



**图 5** 苯的 Vectoris-Rips复形。该复形具有层级性和组合性；$\eta$ 的值越大，所构成的复形就越包含该球大小下的单形，以及在更小半径下找到的单纯复形。这里只显示最大的非零贝蒂数。

从苯的例子来看，我们很容易想把单形与化学领域能够识别并命名的事物联系起来，例如化学键、角度和扭转角；但这种思路有两个问题。首先，角度和扭转角是几何属性，涉及直线和可测量的角度。拓扑学不测量也不保留角度，直线与曲线是等价的，‡ 而且这里没有使用原子标签，因此我们写下 [C, H, H]时，它仅表示空间中三个点之间的连接。其次，并非所有连通部分都是化学家能够识别并标记的事物。例如，船式构象中的环己烷上方的氢原子会通过单形连接起来，但它们之间相隔五个化学键（见图 7）。相邻蛋白质链上的近邻原子也会连接起来，反离子也是如此。如果这种拓扑学方法适用于某个问题，我们可以认为，解决该问题需要这些较长程的属性。

### 3.0.2 创建条形码图

图 5 展示了最终的维托里斯-里普斯复形，它是由一组组单形构成的层级结构，对应于 $\eta$ 的离散取值；然而，拓扑特征存在于连续的长度尺度范围内。我们关注这些同调生成元在不同长度尺度下的持久性，这正是条形码图中相对于连续尺度所绘制的内容。

条形码图记录了名称颇为戏剧性的同调特征的出生与死亡

‡为方便起见，我们使用直线和扁平形状；它们是绘制连接关系的最简单方式，因此称为单形。

在长度尺度上的变化，见图6（在该领域中称为过滤参数，因为我们正在过滤单纯复形）。苯中有12个点（0-单纯形），其中总有一个点特征会延续到无穷远（我们不绘制它），因此图6的条形码图上只有11个点。12个连通区域在0Å处出生，并在图中标为红色。其中六个在C-H键长处死亡，与另外六个合并，形成6个簇。请注意，原子没有编号或区分，因此彼此之间没有差异，条形码图上的排列仅为方便起见。* 这些0-单纯形中有六个在1.08Å处死亡，即C-H键距离；这对应于图4b和单纯复形$X_1$。当过滤参数为1.4Å（C-C键长）时，最后剩余的簇死亡，同时一个新的同调群$H_1$特征出生；也就是说，这些簇合并成一个二维孔洞，其周围有一个一维表面（碳环），这就是单纯复形$X_2$，，见图4c。这个环持续了1.024Å后闭合（单纯复形$X_4$和图4e）。在这个过滤参数下，苯与单个簇无法区分，增大球的尺寸也不会改变该复形的同调。这个单一簇持续到∞Å，但这个最后的点没有绘制在条形码图上，这也解释了为什么图6中只有11个特征。

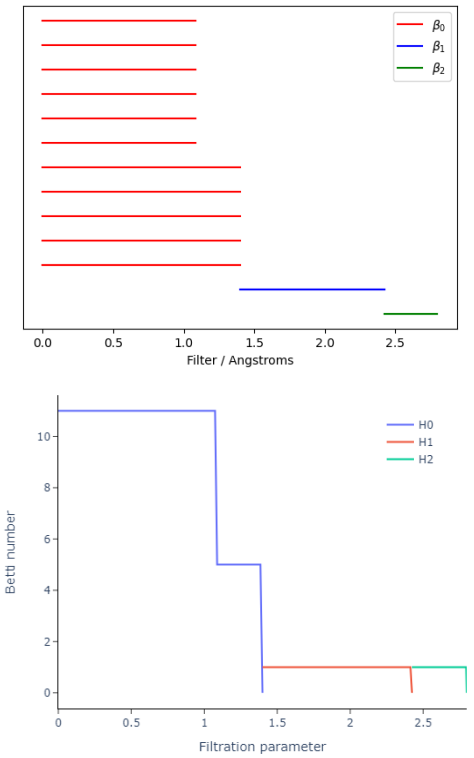



**图6** 苯的条形码图（上）和贝蒂曲线（下）。

### 3.0.3 持久图

每个拓扑特征都由一对数字描述，$\eta_t$其中 $\eta$ 是当前球半径，而$t$是达到的最大半径。随后将其转换为点的重数（$b$，$d$）（见附录8，其中$b$是特征出生时的半径，而$d$是特征消亡时的半径）。持久图尽管被称为图，实际上是点的多重集，如表6中苯的示例所示。每个拓扑特征都关联着三个数字：其出生时的半径、消亡时的半径，以及该特征对应的贝蒂数。还有一个通常从持久图中排除的特征，即从倒数第二个特征消亡时的点开始（表6的最后一行），并在无穷远处消亡的特征——该特征未绘制。† 对于分子而言，由于这一原因，始终会有$N-1$个原始特征（请注意，条形码图中缺少一个氢原子）。

表6中持久图的数据可以绘制在一种也令人困惑地称为持久图的图上，图2和图7展示了此类图的示例。这些持久图将数据绘制为由点组成的多重集（$b$,$d$），也就是说，我们以出生为横轴、以死亡为纵轴，并用颜色表示贝蒂数（第三列）。如果有多个点具有相同的（$b$,$d$），它们就会重叠绘制。在软件中，将鼠标悬停在这些点上即可查看该值的重数；也可以在图中于点旁标注其数量。显然，对称性很高的分子以及处于完美基态结构中的分子，其持久图的重数也会很高。随着分子移动或结构发生轻微变化，这些重合的点会彼此分开，如图7所示。

**表 6** 苯的持久图

| 出生 | 死亡 | $k$ 在$H_k$ |
|---|---|---|
| 0. | 1.08233893 | 0. |
| 0. | 1.08233905 | 0. |
| 0. | 1.08233905 | 0. |
| 0. | 1.08233905 | 0. |
| 0. | 1.08233905 | 0. |
| 0. | 1.08233905 | 0. |
| 0. | 1.39904296 | 0. |
| 0. | 1.39904296 | 0. |
| 0. | 1.39904296 | 0. |
| 0. | 1.39904296 | 0. |
| 0. | 1.39904296 | 0. |
| 1.39904296 | 2.42321348 | 1. |
| 2.42321348 | 2.79808593 | 2. |

苯分子的同调类是什么？持久性在这里很有用，我们选择舍弃那些不具持久性的特征。对此并没有正式的数学规则。通常，我们可以将其近似为持久性最大、排名最高的$H_0$；根据上述数据，这表明苯与一个圆同伦等价，也就是说，它有一个（扁平的）孔洞。连通区域（$H_0$）也会持续很长时间，这告诉我们，这个较大的孔洞结构由

*将这些特征的出生与死亡按其到某一点的距离排序，以便对特征进行空间排序，绘制出相应图形，或许会很有趣；但这并不属于纯拓扑学范畴，本文未对此进行研究。

†这有一种贴切的隐喻性解释，与我们对人类视觉系统的理解相符：在很远的距离上，每种形状都会变成一个点。

| 特征化器 | 特征数量 | MTR大小 |
|---|---|---|
| **PHF** | **36** | **1,039,001** |
| Grid | 2052 | 3,055,001 |
| ECFP | 1024 | 2,027,001 |
| MACCS | 166 | 1,169,001 |
| RDKit | 208 | 1,211,001 |

**表 7** 特征大小和多任务回归器模型大小。

多个簇构成（这在化学上说得通）。在拓扑数据分析中，点越靠近线，就越可能是噪声。在这些化学结构中，这些“噪声”点似乎编码了有关结构的一些有趣信息。例如，在苯中，$H_2$ 可以被视为噪声而忽略，但考虑到苯是三维结构而非完全平面的，这确实说得通；这可能导致分子上方或下方短暂出现一个空腔。

创建这些拓扑特征的过程看起来可能相当复杂，不过需要注意的是，只需四行 Python 代码就能完成（见附录 7.1）。有兴趣的读者可参阅附录 8，了解更深入、更正式的数学细节。

## 4 其他简单分子

通过第 3 节中的苯示例，我们现在已经能够解读持久图。现在显而易见，像己烷这样的链状分子（见图 7）只会形成簇，而持久性图清楚地显示了代表两种键长的两个点。

图 7 展示了一组$C_6H_x$ 分子的持久图，我们可以看到，低持久性特征确实有助于区分不同的结构。环己烷的持久图与苯有两点不同。首先，$H_0$ 特征的重数不同（但在图中不可见）。其次，环己烷打破了环的对称性：椅式构象有两个空腔，即平面上方和下方的三维孔洞；船式构象则从船的顶部多出一个环，并有两个空腔。在化学中，较不持久的特征似乎仍然是不可或缺的信息。

同调只计算拓扑形状，但持久同调中使用的过滤复形也包含某些几何特征，因为单纯复形是由空间中的点及其关系构建而成。因此，环己烷的两种不同形式显然有所不同，参见图 7 中它们的持久图。请注意，图拓扑无法区分这两种结构，因为两种情况下原子之间的图连接方式相同。

持久同调只考察形状，会丢失有关原子类型的信息，因此对映异构体的信息大多也会丢失。例如，S-丙氨酸和 D-丙氨酸的持久图精确到小数点后三位都相同（该精度下的差异很可能是数值计算所致）。

## 5 结果

PHF 特征比其他一维输入特征更少，见表 7。

### 5.1 QM7

PHF 方法适用于 QM 任务（至少适用于回归模型的特征化），这一点很有意思，因为它的规模小得多，而且不包含库仑矩阵能量，也不包含 CM 特征值或 RDKit 的化学特征。针对 CM_eig 输入与 PHF，以及 CM_eig 与 PCA-PHF 的两个独立样本进行的 t 检验，所得 p值均为 $1 \times 10^{-5}$，这表明 TDAF方法与库仑矩阵特征值方法之间不存在统计学上的显著差异（这是因为 CM_eig 的方差较大，而且在 10 次重复运行中平均绝对误差非常高）。

### 5.2 亲脂性

亲脂性数据集的结果见图 9。PHF 的表现与最佳的一维方法扩展连接指纹（ECFP）以及最佳图方法（GC）相当。PHF、ECFP、MACCS 和 GC 的结果之间没有统计学差异。

### 5.3 Delaney（溶解度）数据集

在 Delaney 数据集任务中，PHF 的表现与最佳特征化方法相当。PHF 与扩展连接指纹、MACCS 和 RDKit 特征化方法之间的差异在统计上并不显著。

### 5.4 SAMPL 数据集

SAMPL 数据集规模很小（约 600 个点），在此情况下，扩展连接指纹和 MACCS 方法明显优于 PHF 方法。

## 6 结论

本文提出了一种基于持久同调的新型分子特征化方法，称为 PHF（或称 TDAF，即拓扑数据分析特征化）。这种特征化方法去除了大多数与化学相关的符号，例如原子类型、化学键、化合价和电荷，同时保留了拓扑不变量，例如连通区域的数量、平面孔洞和空腔的数量；此外，由于采用了滤过，还编码了少量几何信息。可以从这些数据中推断出一些化学信息，但其内容主要是拓扑形状不变量。我已证明，这些 PHF 在一系列化学数据集（如亲脂性、溶解度和 QM7）上的表现可达到最先进水平。

考虑到 PHF 编码的化学信息如此之少，它们竟然能发挥作用，实在令人惊叹；这一结果理应深刻改变我们对化学的看法。我不太愿意说三维形状就是一切，但它的重要性确实超出了人们此前的认识。不过，有一点需要说明：神经网络并不会像我们一样“思考”，[45–47] 我们也不应因此被诱导去期待它们（至少目前如此）。因此，这一结果并不是在建议人类化学家抛弃我们用于编码化学信息的所有符号和标签，而只是说，在对统计学习机器（即机器学习算法）进行编程时，我们应愿意舍弃这些符号。

PHF 所需的数据量非常小（向量长度仅为 18），因此比迄今发现的任何其他化学特征化方法都更精简。相应地，它们的信息密度也非常高。需要指出的是，与某些较简单的特征化器（例如

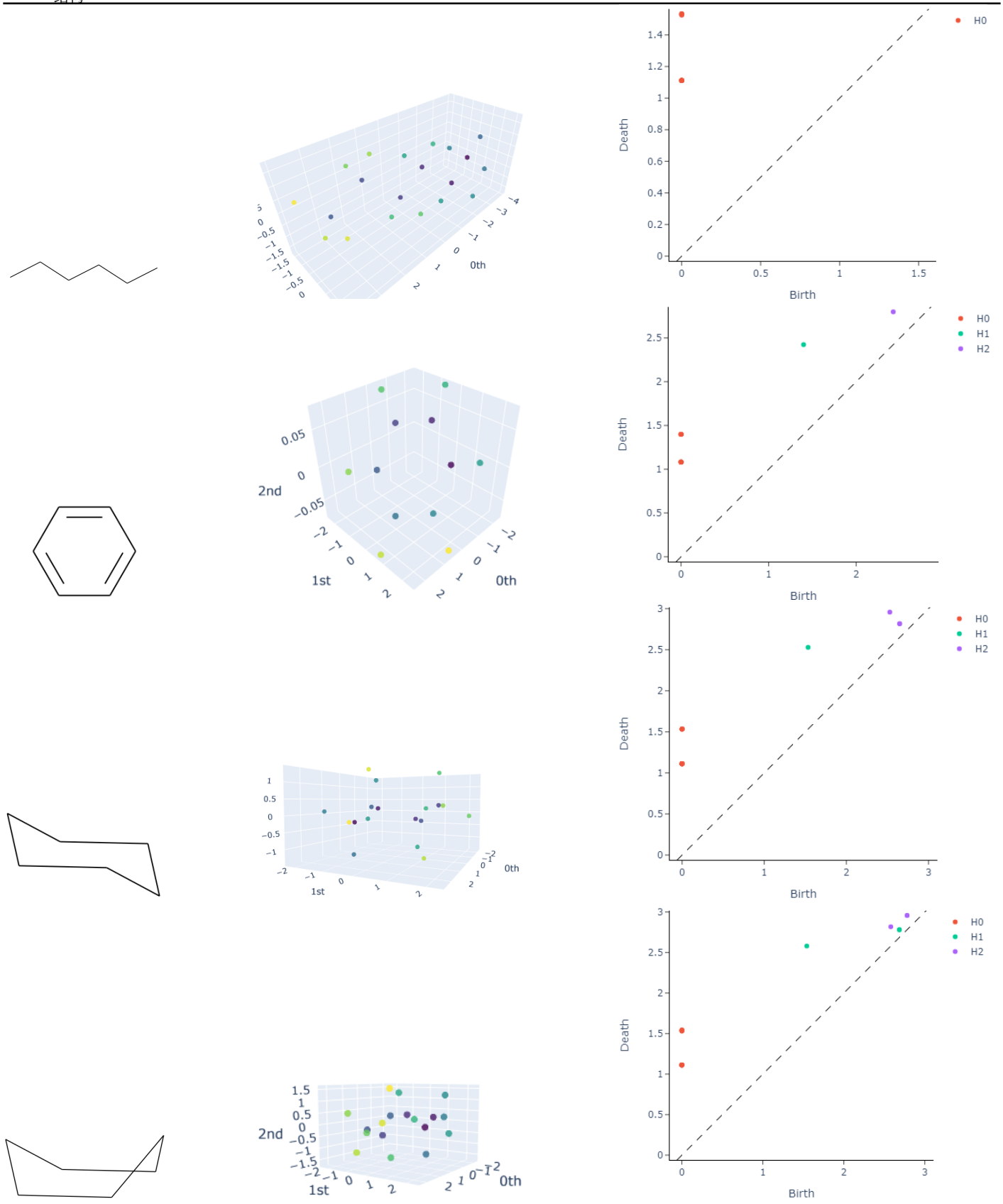



**图 7** 某些 $C_6H_x$ 分子的不同持久图。持久同调能够捕捉环己烷船式异构体与椅式异构体之间的差异。

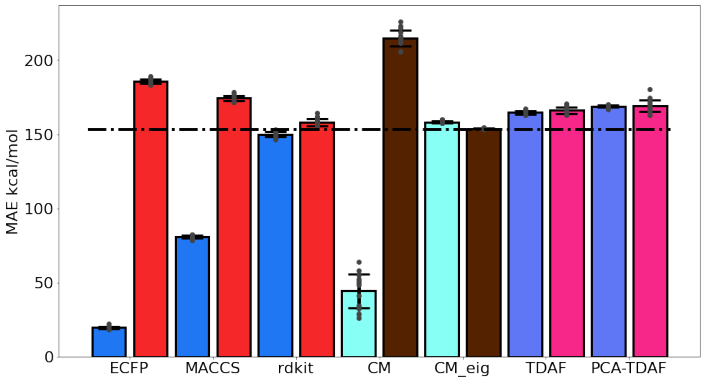



**图 8** QM7 数据集任务（计算原子化能）的单任务回归模型。蓝色、薄荷绿和紫色分别表示基于一维、库仑矩阵和 PHF 特征在验证数据集上的结果。红色、棕色和粉色分别表示基于一维、库仑矩阵和 PHF 特征在测试数据集上的结果。

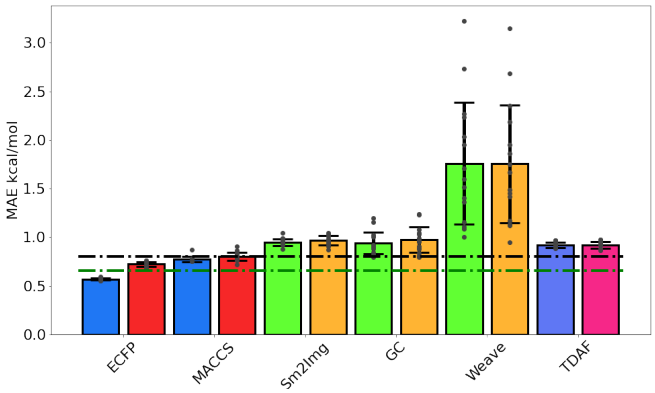



**图 9** 亲脂性数据集上的性能。蓝色、绿色和紫色分别表示基于一维、二维和 PHF 特征在验证数据集上的结果。红色、橙色和粉色分别表示基于一维、二维和 PHF 特征在测试数据集上的结果。

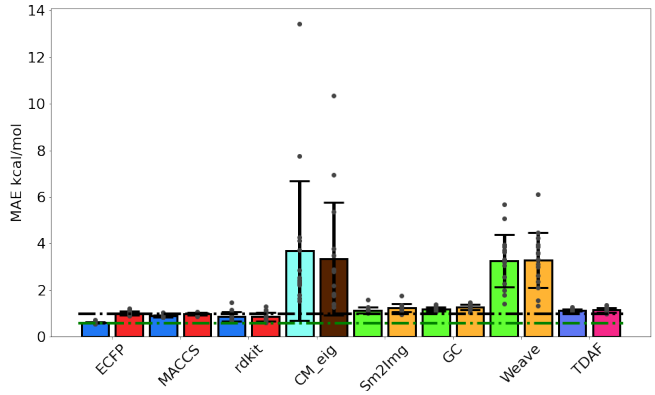



**图 10** Delaney 数据集上的性能。蓝色、薄荷绿、绿色和紫色分别表示基于一维、库仑矩阵、二维和 PHF 特征在验证数据集上的结果。红色、棕色、橙色和粉色分别表示基于一维、库仑矩阵、二维和 PHF 特征在测试数据集上的结果。

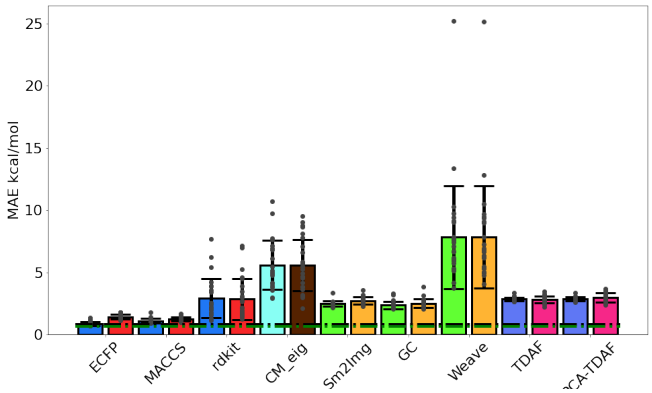



**图 11** SAMPL数据集上的表现。蓝色、薄荷色、绿色和紫色分别表示一维、基于库仑矩阵、二维和 PHF 特征在验证数据集上的结果。红色、棕色、橙色和粉色分别表示一维、基于库仑矩阵、二维和 PHF 特征在测试数据集上的结果。

例如扩展连接指纹）。数据集规模较小，意味着我们可以使用更简单、成本更低的模型（见表 7），能源消耗也更低，对庞大而昂贵的计算机的需求也较少。这还有助于推动机器学习普及，让它不再局限于资金雄厚的跨国公司，而是惠及使用廉价笔记本电脑工作的学生。

PHF 表现如此出色，这也证明，将人类知识和创造力用于重新表述问题、使其更简单（即让机器更容易处理），能够大幅提升机器学习效果；这比单纯向复杂模型投入尽可能多的数据和信息更为有效。使用更小、更简单的模型，意味着我们不再那么依赖复杂的数学方法来防止过拟合，也不再那么需要庞大的数据集（不过，由于化学数据集规模很小，我们仍然需要更多数据）。

## 7 附录：代码示例

用于计算苯的持久图的示例代码。首先导入所需的包（注意，运行 graphs-and-topology-for-chemists.py 即可完成这些操作）。

```
from rdkit import Chem import rdkit.Che
m.AllChem as AllChem from gtda.homology import
VietorisRipsPersistence from gtda.diagrams import
PersistenceEntropy import numpyas np
```

此函数根据 RDKit 的分子结构生成坐标。它是 graphs-and-topology-for-chemists.py 的一部分，这里为便于教学而完整展示。

```
def generate_structure_from_smiles(smiles):

    # 根据 SMILES 生成三维结构

    分子 =Chem.MolFromSmiles(smiles)
    分子 =Chem.AddHs(mol)
```

状态 = AllChem.EmbedMolecule(mol) 状态 = AllChem.UFFOptimizeMolecule(mol)

构象 = mol.GetConformer() 坐标 = conformer.GetPositions() 坐标 = np.array(coordinates)

返回坐标

此函数创建 Vietoris-Rips 复形，计算不同长度尺度下的持久性，并生成持久性图。

def smiles_转为_持久性_图(smiles)：坐标=从_smiles(smiles)生成_结构_ # 将结构转换为点云形式 # 没有原子类型

# 追踪连通分量、环和空腔 同调_维数 = [0, 1, 2]

# 持久性计算 持久性 =Vietoris–Rips 持久性( 度量="欧几里得", 同调_维数=同调_维数, n_任务=6, collapse_边=True, ) 重塑_坐标=坐标[None, :, :]图_基本=持久性.fit_变换(重塑_坐标) 返回坐标, 图_基本

### 7.1 创建拓扑特征

从 SMILES 字符串生成小型数据输入（仅包含持久熵）时，只需输入以下代码块即可。

smiles="c1ccccc1" coords, diagrams_basic=smiles_to_persistence_diagrams(smiles) persistence_entropy = PersistenceEntropy() # 计算拓扑特征矩阵 X_basic = persistence_entropy.fit_transform(diagrams_basic)

用于生成绘图、计算更多特征以及创建完整特征数据集的代码，详见《面向化学家的图与拓扑学》。

## 8 附录：数学细节

若想进一步了解拓扑学，建议有兴趣的读者参阅《Topology: illustrated》，该书通过大量图片对数学内容进行了清晰解释。[34]

### 8.1 代数拓扑详述

空间是一个由数学对象构成的集合，这些对象可以视为点（尽管它们未必是点），并且这些对象之间存在选定的关系——这些关系构成了该空间的规则。例如，原子的位置可以视为空间中的点。

### 8.2 度量空间

度量空间 是由一个满足度量‡“规则”的函数定义的点的空间（例如，$\{x, y, z, a, b\}$），该函数$d$是距离函数，并遵循标准的距离度量规则：

$$d(x,y) = 0 \Leftrightarrow x = y\,, \tag{1}$$

也就是说，如果$x$ 与$y$ 这两个点之间的距离为零，那么这两个点必定相同，反之亦然，

$$d(x,y) = d(y,x)\,, \tag{2}$$

以及三角不等式：

$$d(x,z) \leq d(x,y) + d(y,z)\,. \tag{3}$$

三维空间中的欧几里得距离函数满足上述条件，因此它是描述三维分子结构的自然空间。欧几里得距离函数也称为$l^2$ 范数，如三维空间中的$x = \sqrt{(i^2 + j^2 + k^2)}$所示，其中$x$ 是该空间中的一个“点”，而$i$、$j$ 和$k$ 是沿笛卡尔坐标轴的距离 （抱歉，这里滥用了术语，它们并非单位向量）。利用点积和欧几里得范数，可以构造一个向量空间（字面意义上说，就是“点”为向量的空间） $\mathbb{R}^3$，它是由实数三元组构成的向量空间，其中最熟悉的是笛卡尔坐标$(i, j, k)$，向量从原点$(0, 0, 0)$定义（本文中提到三维空间时，指的就是这种空间）。

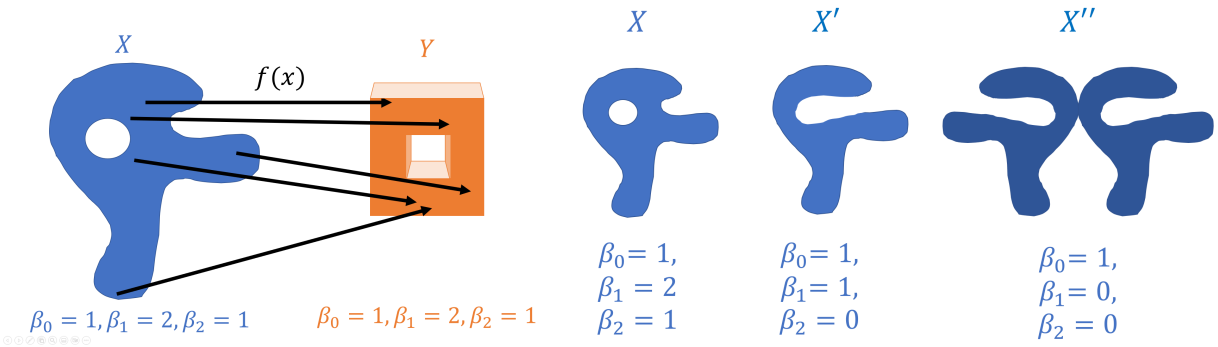
（三维）点云是 $\mathbb{R}^3$ 的一个有限子集，并具有由欧几里得空间诱导的度量。因此，三维点云位于一个有限度量空间中：$(X,d)$（令 $d$ 为欧几里得距离函数，$X$ 为按某种顺序排列的点列表：$X = x_1 < x_2 < x_3 \ldots < x_n$）。由此我们可以构造一个距离矩阵，其中第 $(x_i,x_j)$ 项表示 $x_i$ 与 $x_j$ 之间的距离，即 $d(x_i,x_j)$。

因此，分子中原子的原子位置可以描述为一个点云，其度量为欧几里得距离函数。

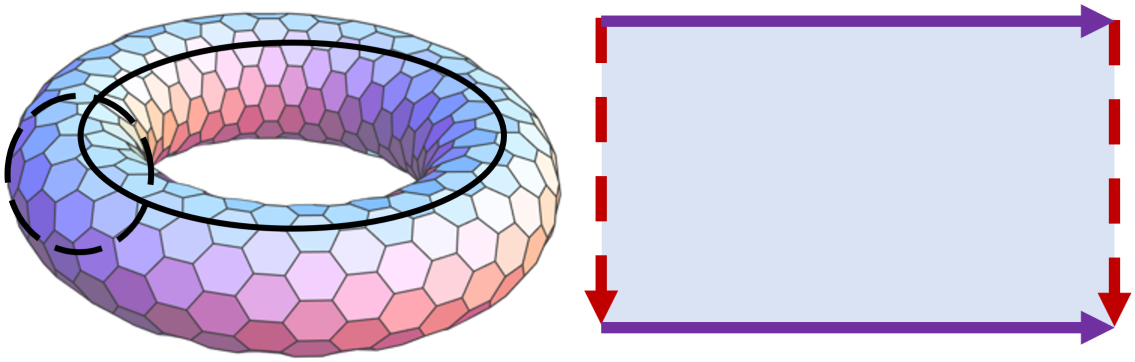
### 8.3 同调

在 $\mathbb{R}^3$ 中，我们最多只能有维数为 2 的同调特征，也就是说，我们的拓扑空间是曲面（维数为 2 的流形）。图 12 展示了一个环面曲面 12c 的示例，它由一张可拉伸的二维薄片 12d 制成。这个环面有两个维数为 2 的生成元：$H_2$ ，它们生成曲线族：其中一组绕中央大孔一圈或多圈（其中一条曲线以黑色画在环面上）；另一组绕环面的内部一圈或多圈（其中一条曲线以黑色虚线画在环面上）。这些圆在拓扑学中称为 1-循环，同一群中的所有 1-循环彼此同伦等价，因为它们可以平滑地变形为

‡ 度量字面上就是一把尺子




a. 从$X$到$Y$的连续映射。b. 切割$X$会改变其类，但不会使其断开。



c. 环面上的生成元 d. 环面映射

**图 12** 同调概念。a. 二维拓扑空间$X$和$Y$彼此同胚等价，并且都与一个环面同胚。函数$y = f(x)$通过变形将$X$映射到$Y$上。蓝色和橙色区域各自连通，但彼此不连通。b. 切割曲面$X$会使其类从环面变为圆柱面/圆，而$X'$仍然连通。需要切割两次才能将流形展平，这也解释了为什么 $\beta_1 = 2$ 在原始$X$中。$X''$与平面同胚。c. 要了解环面的拓扑结构，可以观察环面上的路径。绕着环行进的路径有无穷多条，包括绕行不止一周的路径，但它们同伦等价，因此我们有一个生成元$H_1$，它生成所有这些退化路径。还有第二个$H_1$生成元，它环绕环面内部的孔洞，并生成一组无限的路径。因此，对于环面，$\beta_1 = 2$。环面上绘制的圆是由$H_1$同调群生成的 1-循环。d. 将一张平面弯成圆柱面，并粘合紫色边缘，使箭头相匹配；然后再将圆柱面弯曲，并粘合红色虚线边缘，使箭头相匹配，即可制成环面。

彼此。环面表面上存在所谓的平凡 1-循环，它们没有围住孔洞，并且与一个点同伦等价（因为它们可以变形（即收缩）到彼此重合的一个点。注意，这并非同胚，因为 1-循环一旦变成点，就无法反向还原成 1-循环。

图 12 中的流形 $X$ 存在于三维空间中。存在一个连续映射 $y = f(x)$，可以将 $X$ 变形为 $Y$，而这两个流形都与图 12c 中的环面同伦等价（也同胚等价；图中未画出反向函数 $x = g(y)$）。

如果沿着 $X$ 中的一个生成元切开，就会得到 $X'$，其拓扑类别随之改变（$X$ 与 $X'$ 不同伦等价），因为同伦不变量的数量发生了变化：具体而言，新流形少了一个孔洞，因此 $\beta_1$ 现在为 1。物体中心仍有一个孔洞，而 $X'$现在与圆柱面同伦等价；由于圆柱面可以压扁成圆，它也与圆同伦等价。第二种操作并不同胚，原因与将圆收缩为点时相同。我们可以切开剩下的孔洞，得到流形 $X''$；它现在已被展开，与平面同伦等价。由于需要切两次才能将 $X$压平，$\beta_1 = 2$，这也正是环面上生成元的数量，$H_2$，。

## 8.4 单形

### 8.4.1 平坦性

现在，我们将尝试更严格地定义“平坦性”这一概念。

给定一组$n$ 个点 $\{a_1, a_2, \ldots a_n\}$，这些点的凸组合是由$x$ 表示的任意点

$$x = \sum_{i=0}^{n} \lambda_i a_i \,, \tag{4}$$

其中 $\lambda$ 为实数，且必须满足：

1. $0 \leq \lambda_i \leq \lambda 1;, \forall i$
2. $\sum_i \lambda = 1.$

凸包 是所有凸组合的集合（即所有满足这些约束的点的集合）。

设有两个点 {a、b}，它们是从原点出发的向量，且不平行于任何笛卡尔坐标轴。第一个约束使这些点位于 a 和 b 之间的凸包内，第二个约束则使它们位于平面上。线性组合是在方程 4 中不受 $\lambda_i$ 约束的组合，它们构成由 (0,a,b) 定义的平面，并且该平面可以延伸到二维 $\pm\infty$。仿射包只有第二个约束，它定义了一条位于平面 (0,a,b) 内的线段，该线段经过点 (a,b)，并向两侧延伸至 $+\infty$ 和 $-\infty$（即一维）。凸包是位于平面 (0,a,b) 内、连接 a 和 b 的线段，因此是有限的，也就是在点 a 和 b 之间画出的那条线段。它是这两点之间最短的直线，因此是“平坦的”。三角形的凸包是连接点 (a,b,c) 的三条直线，它们当然位于平面 (a,b,c) 内。凸集中的点还必须几何独立，也就是说，集合中的任意两个点都不相同。

### 8.4.2 凸集与单纯复形

凸集（单形就是其中一例）是路径连通的（也就是说，可以在集合内行走），且没有孔洞或空腔等；相反，它会包围这些孔洞和空腔，这也是拓扑学发现这些特征的一种方式。单纯复形$X$是所发现的所有单形构成的超集。单纯复形是 $\mathbb{R}^3$ 中有限个单形的并集，并满足以下条件：其中任一单形的每个面都属于该复形；两个单形的非空交集是它们各自的一个面。

一个抽象单纯复形是由两个集合组成的有序对（$V$，$X$），其中$X$ 中的元素是 V 的子集，并且满足：对于$V$（向量空间）中的每个$v$ （向量），单元素集合$v$ 也属于$X$；若$x$ 属于$X$，且$y$ 是$x$的子集，则$y$ 也属于$X$。这定义了单纯复形的层次结构，见图 5。

$X$ 中的元素称为单形，单形$x$的维数定义为 $|x| = \#x - 1$，其中 $\#x$ 表示$x$的基数。维数为$n$ 的单形称为$n$-单形。这就是为什么四面体由 4 个三角形构成，却是 3-单形；由 3 条线段构成的三角形是 2-单形；由 2 个点构成的线段是 1-单形。请注意，虽然这里采用单形，但也可以使用其他形状。例如，基于超立方体的集合称为立方复形（点、线段、正方形、立方体、超立方体），但其每个层级包含的单形（1、2、4、8）比单纯复形（1、2、3、4）更多，因此效率较低；胞腔复形是另一种复形，使用点、闭线段、圆盘、球和超球。

过滤复形是由一系列单纯复形组成的集合$\{X_\eta\}_{\eta \in \mathbb{R}^3}$，其中$X^\eta$ 对于每个 $\eta \leq t$ 都是$X_t$ 的子复形。因此，过滤复形由较简单的复形逐步构建而成；取值较高的 $\eta$ 所对应的复形包含取值较低的 $\eta$所对应的复形。

维托里斯-里普斯复形是一种使用距离函数构建的过滤复形。设 $(X,d)$ 为一个有限度量空间（即由分子点云构成的欧几里得距离度量空间）。$X$的 Vietoris-Rips 复形是过滤复形 $VR^\eta(X)$：如果$X$ 的某个子集中的所有两两距离都小于或等于 $\eta$，则该子集包含在这个复形中，作为一个单纯形。Vietoris-Rips 复形的定义如下：

$$VR_{\eta(X)} = \{[v_0, \ldots v_n | \forall i, j, d(v_i, v_j) \leq \eta\} \,. \tag{5}$$

基本上，维托里斯-里普斯复形是一组单纯复形：当单纯形之间的距离小于 $\eta$ 时，将它们连接起来形成这些复形；它还包含距离小于 $\eta$时形成的子单纯复形。图4展示了不同 $\eta$取值对应的单纯复形。

**8.4.2.1 同胚** 源自希腊语“homoios”（相似）和“morpho”（形状），表示两个对象的形状相似，因为它们在拓扑上等价。

若$X$ 和$Y$ 是拓扑空间，且$f: X \to Y$ 是一个将 X 映射到 Y 上的函数，则当$f$:

$$f(x) = f(y) \implies x = y \,,$$

对于每个 $y \in Y$，都存在一个 $x \in X$，使得 $y = f(x)$，也就是说，$x$映射到$y$，$f$是连续的（没有间隙），并且 $f^{-1}$ 是连续的（也就是说，存在

逆映射，而且它也没有间隙）。[34] 这意味着我们可以将$X$变形为$Y$，再变回来。

在这种变形过程中保持不变的性质称为拓扑不变量，这些正是我们在这个领域中计数的性质，例如连通分量、孔洞、空腔等。而不会保持不变的性质则是我们在拓扑学中舍弃的性质，例如距离、角度、几何、手性、体积等。

**8.4.2.2 同伦与同伦等价**如果$f$和$g$是映射，$X$和$Y$是空间，那么两个连续映射$f,\ g\colon X \to Y$在存在连续映射$H\colon X \times [0,1]$的情况下称为同伦，其中$H(x,\ 0) = f(x)$，且$H(x,\ 1) = g(x)$；也就是说，$H$通过光滑变形将$f$映射到$g$（反之亦然），即我们可以将一种形状的弹性材料塑造成另一种形状。

如果存在映射$G\colon Y \to X$，使得对于映射$f\colon X \to Y$，$f \cdot y$与$Y$上的恒等映射同伦，且$g \cdot f$与$f$同伦，那么两个空间$X$和$Y$就是同伦等价的。

对于任意拓扑空间$X$、阿贝尔群$A$和整数$k$，$k > 0$，都可以指定一个群$H_k(X,A)$。若$f$和$g$同伦，则$H_k(f,A) = H_k(g,A)$。若$X$和$Y$同伦等价，则$H_k(X,A)$与$H_k(Y,A)$同构。

**8.4.2.3 贝蒂数** 对于任意域$F$，$H_k(X,F)$都是$F$上的向量空间。其维数（若为有限维）即为贝蒂数$\beta_k(X,F)$（系数取自 F）。第$k$个贝蒂数表示相互独立的$k$维曲面的数量。若两个空间同伦等价，则它们的所有贝蒂数都相等。（例如，甜甜圈表面与茶杯表面同伦等价，因为它们的贝蒂数分别为：$\beta_0 = 1$、$\beta_1 = 2$和$\beta_2 = 1$。

### 8.5 多重集

数学集合是一组元素，其中没有定义元素之间的顺序，例如：$\{1,5,3,2\}$或$\{a,b,c,d\}$，其中$\{\}$通常用于表示一个集合。我们人类可能会认为这里存在数值或字母顺序，但在数学中，集合并没有这样的顺序。集合中的项目称为元素。多重集将这一概念扩展为允许同一项目出现多个副本，其中重数表示该项目在集合中出现的次数。例如：$\{a,a,a,b,b,c\}$是一个多重集，其中元素 a 的重数为 3，元素 b 的重数为 2，元素 c 的重数为 1。例如，一个持久图可能包含多重集：$\{(b_1,d_1),(b_1,d_1),(b_1,d_1),(b_2,d_2),(b_3,d_3)\}$，其中$(b_1,d_1)$的重数为 3，而$(b_2,d_2)$和$(b_3,d_3)$的重数均为 1。

#### 8.5.1 持久图

持久图是一个点的多重集，其中包含所有具有$d_i < +\infty$的点的出生半径和死亡半径，也就是说，我们忽略延伸至无穷远的特征。持久图$D$写作$D = \{(b_i,d_i)\}_{i\in I}$

### 8.6 持久熵

持久图$E(D)$的熵为

$$E(D) = -\sum_{i\in I} p_i \log(p_i)\,,$$

其中$p_i$为

$$p_i \frac{(d_i - b_i)}{\sum_{i\in I}(d_i - b_i)}\,.$$

由于$(d_i - b_i)$是条形码的长度，也就是特征的持久性，因此持久熵衡量的是条形码长度的分布。

## 致谢

E.G. 谨向资助本研究的英国工程与自然科学研究委员会（EPSRC）、参与讨论和校对的 Oliver Matthews、DeepChem 项目团队，以及出版了《Topology Illustrated》的 Peter Saveliev 致谢；这本书帮助我们更容易地理解了所有这些概念。

## Notes and references

1 I. D. Brown，结构化学，2002，**13**，339–355。2 C. A. Coulson，*Journalof the Chemical Society(Resumed)*，1955，2069–2084。3 M. Weisberg，科学哲学，2008，**75**，932–946。4 P. G. Mezey，*Journal of Mathematical Chemistry*，2012，**50**，926–933。5 D. Gluck，化学文献杂志，1965，**5**，43–51。6 D. Rogers 和 M. Hahn，化学信息与建模杂志，2010，**50**，742–754。7 D. Lowis，*Notes*，1998，**1**，。8 J. L. Durant、B. A. Leland、D. R. Henry 和 J. G. Nourse，化学信息与计算机科学杂志，2002，**42**，1273–1280。9 P. M. Murray、F. Bellany、L. Benhamou、D.-K. Bu˘car、A. B. Tabor 和 T. D. Sheppard，有机与生物分子化学，2016，**14**，2373–2384。10 J. D. Moseley 和 P. M. Murray，*Journal of ChemicalTechnol-ogy & Biotechnology*，2014，**89**，623–632。11 D. J. Durand 和 N. Fey，化学评论，2019，**119**，6561–6594。12 A. Krizhevsky、I. Sutskever 和 G. E. Hinton，《神经信息处理系统进展》，2012，第1097–1105页。13 I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio，《神经信息处理系统进展》，2014，第2672–2680页。14 C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna，IEEE计算机视觉与模式识别会议论文集，2016，第2818–2826页。15 P. Schwaller、R. Petraglia、V. Zullo、V. H. Nair、R. A. Haeuselmann、R. Pisoni、C. Bekas、A. Iuliano 和 T. Laino，《*Chemical Science*》，2020，**11**，3316–3325。16 P. Ball，*Chemistry World*，2021。17 M. Gordin，*Chemistry World*，2015。18 M. Francl，自然化学，2009，**1**，334–335。19 R. King，*Journal of Mathematical Chemistry*，1991，**7**，51–68。20 S. D. Fielden、D. A. Leigh 和 S. L. Woltering，*Angewandte*

《*Chemie International Edition*》，2017，**56**，11166–11194。21 D. Sumners，*Journal ofmathematical chemistry*，1987，**1**，1–14。22 C. Adams、J. Devadoss、M. Elha mdadi 和 A. Mashaghi，*Journal of Mathematical Chemistry*，2020，**58**，1711–1736。23 B. Scalvini、V. Sheikhhassani 和 A. Mashaghi，*PhysicalChem-istryChemical Physics*，2021。24 C. R. Calladine 和 H. Drew，《理解*DNA*：分子及其作用机制》，学术出版社，1997。25 D. Rouvray，《*Journal of Molecular Structure: THEOCHEM*》，1995，**336**，101–114。26 I. Fabiˆc-Petraˆc、B. Jerman-Blaˇziˆc 和 V. Batagelj，*Journal ofmathematicalchemistry*，1991，**8**，121–134。27 S. C. Basak、G. J. Niemi 和 G. D. Veith，*Journal of Mathematical Chemistry*，1991，**7**，243–272。28 G. Carlsson，《美国数学会通报》，2009，**46**，255–308。29 L. Wasserman，《统计学及其应用年度评论》，2018，**5**，501–532。30 J. Murugan 和 D. Robertson，*arXiv*预印本 *arXiv:1904.11044*，2019。31 M. Pirashvili、L. Steinberg、F. Belchi Guillamon、M. Niranjan、J. G. Frey 和 J. Brodzki，《化学信息学杂志》，2018，**10**，1–14。32 J. S. Delaney，化学信息与计算机科学杂志，2004，**44**，1000–1005。33 N. Atienza、R. Gonzalez-Diaz 和 M. Rucco，国际软件技术联合会议：应用与基础，2016，第3–12页。34 P. Saveliev，*Topology Illustrated*，2016。

35 M. Rucco、F. Castiglione、E. Merelli 和 M. Pettini，《ECCS 2014 论文集：欧洲复杂系统会议》，2016 年，第 117–128 页。36 M. Kerber、D. Morozov 和 A. Nigmetov，几何有助于比较持久图，2017 年。37 H. Adams、T. Emerson、M. Kirby、R. Neville、C. Peterson、P. Shipman、S. Chepushtanova、E. Hanson、F. Motta 和 L. Ziegelmeier，《机器学习研究杂志》，2017 年，**18**，。38 E. M. Gale、N. Martin 和 J. S. Bowers，*arXiv* 预印本 *arXiv:1806.03934*，2018 年。39 Z. Wu、B. Ramsundar、E. N. Feinberg、J. Gomes、C. Geniesse、A. S. Pappu、K. Leswing 和 V. Pande，《*Chemical Science*》，2018 年，**9**，513–530。40 G. Landrum，*RDKit*：开源化学信息学软件，http://www.rdkit.org/。41 G. Tauzin，*Giotto-TDA*，2020 年，https://github.com/ giotto-ai/giotto-tda。42 G. Tauzin、U. Lupo、L. Tunstall、J. Pérez、M. Caorsi、A. Medina-Mardones等，*arXiv* 预印本 *arXiv:2004.02551*，2020 年。43 E. M. Gale，*GandT*：化学中的图与拓扑，2022 年，https://github.com/ellagale/GandT。44 E. M. Gale，*DeepChem* 验证，2020 年，https://github.com/ellagale/deepchem_2.3_validation。45 E. M. Gale、N. Martin、R. Blything、A. Nguyen 和 J. S. Bowers，《视觉研究》，2020 年，**176**，60–71。46 I. J. Goodfellow、J. Shlens 和 C. Szegedy，*arXiv* 预印本 *arXiv:1412.6572*，2014 年。47 J. B. Marin Dujmovi´c、Gaurav Malhotra，*bioRxiv 2020.02.25.964361*；数字对象标识符：*https://doi.org/10.1101/2020.02.25.964361*，2020 年。