匿名模式分子指纹及其在性质识别中的应用

Xue Liu[1, 3, †]、Qian Cheng [2, 3, †]、Dan Sun[2, 3], Xing Li[2, 3], Wei Wei [2, 1, 3, 4, *]和 Zhiming Zheng[1, 3, 4]

[1]北京航空航天大学人工智能研究院，北京，100191，中国[2]北京航空航天大学数学科学学院，北京，100191，中国[3]教育部数学、信息学与行为语义重点实验室，100191，中国[4]中关村实验室，北京，100094，中国

arXiv:2301.01620v1 [stat.AP] 4 Jan 2023

摘要——分子指纹是重要的化学信息学工具，可根据分子在不同官能团、原子序列及其他拓扑结构方面的特征，将分子映射到向量空间。本文旨在研究一种新型分子指纹 *Anonymous-FP*，它能够充分感知由小、中、大分子尺度连接所形成的内在相互作用。具体而言，我们从每个分子中采样可能存在的原子链，并将其扩展为特定的匿名模式。随后，借助自然语言处理技术 *PV-DBOW* 对分子指纹 *Anonymous-FP* 进行编码。我们研究了*Anonymous-FP* 在分子性质识别中的应用，结果显示其具有信息含量丰富、实验性能优异和结构意义完整等显著优势。在实验验证过程中，原子链的尺度或其匿名方式对*Anonymous-FP*的整体表征能力有显著影响。总体而言，典型尺度 $r = 8$ 可提升其在一系列真实世界分子上的表现；具体而言，在所有 NCI数据集上，准确率均可达到 93% 以上。

索引词——分子指纹、随机游走、匿名模式游走、典型尺度、性质识别

1 引言

发现新分子极大地造福了人类社会，而准确预测未知分子的性质仍是一项开放挑战。在药物化学领域 [1]，药物设计 [2]，《生物信息学》 [3] 等领域中，分子均可表示为图；图中的节点依据彼此间的边相互作用，并整合为一个整体以实现特定性质。在分子平面图中，节点的位置由不同原子占据，边则依据原子之间的化学键形成。拓扑结构的差异会导致化学或物理性质的多样性。例如，图 1 展示了 [4] 4-硝基联苯与5-硝基苊分子之间典型的异构现象；这两种分子均收录于 MUTAG 数据集[5]，且被赋予了不同的标签。二者具有相同的分子式 $C_{12}H_9NO_2$，但由于碳链结构不同，表现出截然不同的性质（见表格）。尤其是，5-硝基苊广泛用于制药工程，是药物合成的重要原材料。

如何推断分子的物理化学性质，或仅根据图拓扑区分不同分子，近年来受到了机器学习等多个领域的广泛关注。而这一切的核心都不可避免地归结为图同构问题（简称图同构问题）[6]，[7]。如果存在一个从该图的顶点到另一图的顶点的双射 f，且该映射能够保持邻接关系，那么这两个图就是同构的。也就是说，对于图 $G_1 = (\mathcal{V}_1, \mathcal{E}_1)$和 $G_2 = (\mathcal{V}_2, \mathcal{E}_2)$，对于所有 v_i, $v_j \in \mathcal{V}_1, i \neq j$,

$$(v_i, v_j) \in \mathcal{E}_1 \Leftrightarrow (f(v_i), f(v_j)) \in \mathcal{E}_2. \tag{1}$$

显然，图同构问题（图同构问题）属于 NP 类，而子图同构已被证明是 NP完全问题；然而，图同构是否属于 NP完全问题仍未可知 [8]。因此，将图同构问题（图同构问题）的理论成果直接应用于实际的分子相似性度量并不可行。近年来，大量关于分子指纹的文献为分子表示、识别和比较提供了有效方法，MACCS [9], PubChem指纹 [10], Morgan指纹 [11] 等都是典型类型。

分子指纹 [12], [13] 尝试通过判断预定义结构键集合中的某些化学片段是否存在，将分子编码为一串比特。这些化学片段是对子结构模式的简化或抽象，需要借助领域知识加以识别。比特串的长度取决于目标分子所包含的化学片段数量，而分子指纹中的每个索引都表示一种化学结构的键。当每个分子的组成部分很可能都包含在结构

†这些作者对本研究的贡献相同，应视为共同第一作者 * 通讯作者：*weiw@buaa.edu.cn*

Property	4-Nitrobiphenyl	5-Nitroacenaphthene
Molecular Formula	$C_{12}H_9NO_2$	$C_{12}H_9NO_2$
Molecular Weight	199.2 g/mol	199.2 g/mol
Melting Point	113-114 °C	101.5-102.5 °C
Boiling Point	340 °C	377.49 °C
Water Solubility	No	Yes
Medical Application	No	Yes

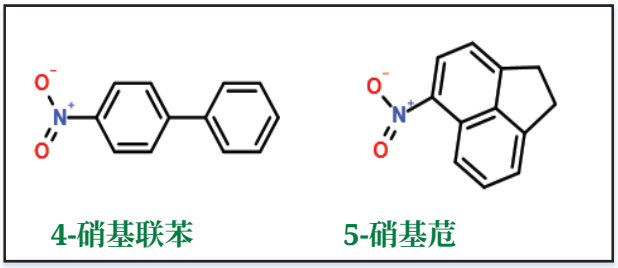


图 1. 异构体分子 4-硝基联苯和 5-硝基苊的结构图。4-硝基联苯是一种芳香族化合物，而 5-硝基苊属于杂芳香族化合物。

键集合中时，分子指纹最为有用；然而，当分子包含键集合之外的子结构时，就会面临挑战。

为了全面理解键集合，最直接且最简单的策略是通过随机游走模型遍历所有可能的（子）结构，然后基于嵌入技术将每个已知或未知结构映射为唯一代码。这样，任何子结构仅由其包含的原子以及每条原子链中原子的连接方式决定，而不受先验知识中其合理化学定义或功能的影响。这里，每个挖掘出的子结构值得关注之处在于其拓扑尺度，它对应于通过随机游走得到的每条原子链的长度r 。特别地，我们将目标分子中表达能力最强的子结构的尺度称为典型尺度 。

本研究。 本文将分子图分解为一系列 r尺度的匿名原子链，并基于这些结构提出一种新的分子指纹方法——匿名模式分子指纹 （简称***Anonymous-FP***），其中 r 表示从起始原子到终止原子的原子链长度。我们的方法分为两个步骤：通过随机游走采样匿名原子链，以及通过嵌入技术对匿名原子链进行编码。

步骤 1：采样匿名原子链。我们从分子中的每个原子出发，采样t 次 r尺度的随机游走，并将其收集成一个集合，该集合表示原子链分解。显然， r尺度的原子链出现的概率会随着距离 r的增加而降低。为避免分布稀疏，我们采用一种名为基于匿名化的随机游走[14] 的特殊编码模式，将每条原子链转换为由索引组成的匿名模式（即匿名原子链）。该序列的每个位置表示链中相应原子首次出现的顺序。

这种方案使 *Anonymous-FP* 适用于缺少全局结构键、甚至某些原子缺少标签的分子。此外，该方案的计算复杂度也较低，因为匿名编码通常具有统计显著性，尤其有助于理解分布稀疏的长 r 尺度原子链。

步骤 2：对匿名原子链进行编码。 在分子具有不同的匿名原子链分解结果时，要评估两个分子之间的相似性，面临着链数量不同这一挑战。受自然语言处理（自然语言处理）文档嵌入技术***PV-DBOW*** [15],的启发，该技术将每个文档编码为向量表示，我们将分子视为文档，将匿名原子链视为其中相互关联的词。随后，我们将每个分子嵌入欧几里得空间，表示为一个向量，并将这种固定大小的向量称为分子指纹，命名为*Anonymous-FP*。我们的方法源于相似性性质原理（***SPP***）[16], [17], ，该原理指出，如果分子具有相似的结构特征、相似的匿名原子链分解结果，并且在嵌入空间中彼此接近，那么它们就表现出相似的物理化学性质。

我们使用一系列真实世界的分子数据集（MUTAG、PTC、PROTEINS、DD 和 NCI数据集）评估 *Anonymous-FP* 在性质识别（如性质分类）方面的效率。我们将*Anonymous-FP* 与核方法（Graphlet核、Weisfeiler-Lehman 核）、嵌入方法（Graph2vec、AWE）和图神经网络（PATCHY-SAN、GraphSAGE）的性能进行了比较。实验结果表明，在分类准确率方面，我们的 *Anonymous-FP* 具有显著优势。我们系统地分析了分子图表示能力与原子链长度 r 及采样数量 t之间的相关性。同时，将达到最高准确率的尺度定为典型尺度。随后，我们提出了一个更有趣的发现：在所有 NCI数据集中，当 r 变为8时，分类准确率均达到最高，且超过93%。

2 相关工作

本节回顾本文所用主要方法的相关工作，包括分子指纹、无权图的基本定义、分子图嵌入，以及*PV-DBOW* 在自然语言处理中的技术。

分子指纹。分子指纹是重要的化学信息学工具，专用于通过向量表示和比较来搜索、描述并验证分子结构特征。各种先驱性指纹大致可分为四类，包括子结构键指纹、拓扑指纹、药效团指纹及其他类型。

子结构键指纹根据一组结构键中子结构的存在情况，将每个分子编码为位字符串，但当键集合中缺少子结构时，其有效性会降低。MACCS [9],PubMed指纹 [10], 和 BCI指纹 [18] 指纹是典型示例。拓扑指纹（例如 Molprint2D [19], ECFP [20], 和 MP-MFP [21]）会查找

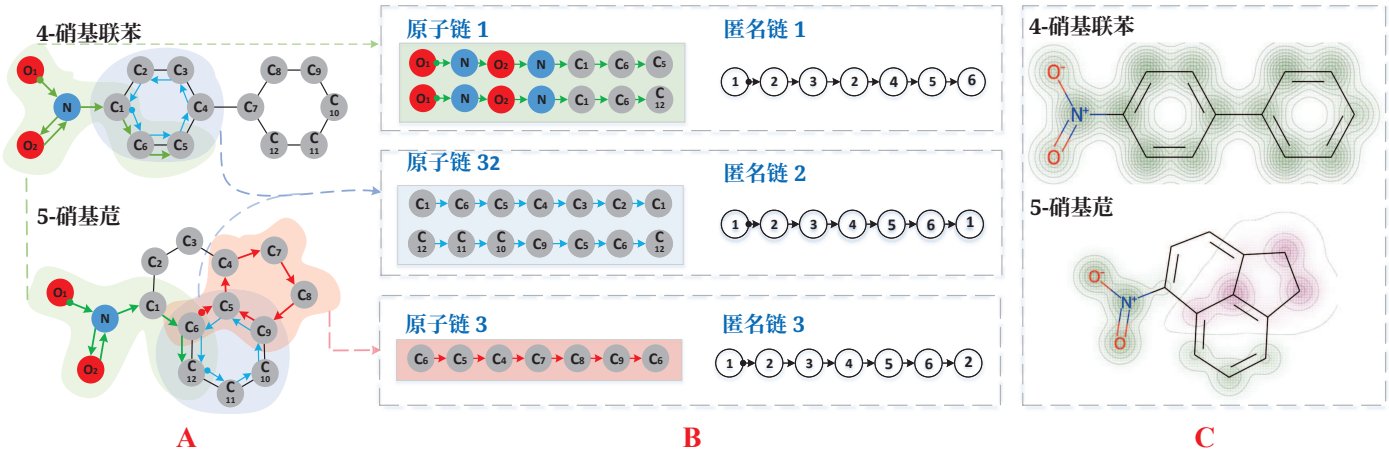


图 2. 原子链 与匿名原子链之间关系的概览。如子图（A）所示，氧原子、氮原子和碳原子分别标为红色、蓝色和灰色。它们分别以不同方式从 1 到 12 排列。在子图（B）中，匿名原子链 1–3 对应于 5 种不同的 6 尺度原子链，这些原子链从这两个分子中随机采样，并分别以浅绿色、浅蓝色和浅红色阴影标示。在这三条匿名原子链中，匿名原子链 1 和匿名原子链 2 为两个分子共有，而匿名原子链 3 则是 5-硝基苊所独有的。这也表明了这两个分子之间的差异，并可在子图（C）中大致体现。

原子链，然后对每条原子链进行哈希处理以创建指纹。药效团指纹会考虑目标特征列表中的分子特征 [22]。其他类型通常使用规范 SMILES [23], 蛋白质-配体相互作用 [24], 以及其他结构相互作用生成指纹。

图表示。令 $G = (\mathcal{V}, \mathcal{E})$ 表示一个无权图，其中 n 个顶点位于集合 $\mathcal{V}$ 中， m 条边位于集合 $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$中，邻接矩阵 $\mathrm{A} \in \mathbb{R}^{n \times n}$ 编码了图中各顶点之间的连接关系，其定义如下：

$$\mathrm{A}_{i,j} = \begin{cases} 1, \ if \ (v_i, v_j) \in \mathcal{E} \\ 0, \ otherwise \end{cases}. \tag{2}$$

顶点 v_i 的度数 d_{v_i} 定义为邻接矩阵 A 中第 i 行元素之和，这恰好等于 v_i的一跳邻居数量：

$$d_{v_i} = \sum_{j=1}^{n} \mathrm{A}_{i,j}. \tag{3}$$

转移矩阵 P 记录了智能体从顶点 v_i移动到其一跳邻域中任意顶点的所有转移概率 $Prob(v_j|v_i)$，且每个元素 $\mathrm{P}_{i,j} = Prob(v_j|v_i)$ 满足

$$\mathrm{P}_{i,j} = \begin{cases} \dfrac{1}{d_{v_i}}, \ if \ (v_i, v_j) \in \mathcal{E} \\ 0, \ otherwise \end{cases}. \tag{4}$$

显然，对于每个顶点 v_i,

$$\sum_{j=1}^{n} \mathrm{P}_{i,j} = 1. \tag{5}$$

分子图嵌入。 分子图嵌入与分子的向量表示有关。它将分子图 G 映射为欧几里得空间中的一个 d维向量，即

$$\varphi : G \to \mathbb{R}^d, \tag{6}$$

其中 φ表示一个嵌入函数。

PV-DBOW。 在自然语言处理（自然语言处理）中，*PV-DBOW* 技术用于对文档级别的情感进行无监督嵌入。具体而言，给定一个包含词集 $\mathcal{V} = \{\omega_1, \ldots, \omega_\nu\}$的文档集合 $\mathcal{D} = \{D_1, \ldots, D_N\}$，对于包含词序列 $\mathcal{V}_i = \{\omega_1, \ldots, \omega_l\} \subseteq \mathcal{V}$的目标文档 $D_i \in \mathcal{D}$，目标是通过最大化以下对数概率，学习文档 D_i 的低维向量表示 $\boldsymbol{D}_i$：

$$\sum_{\omega_j \in \mathcal{V}_i} \log \mathbf{Pr}(\omega_j | D_i). \tag{7}$$

上面的条件概率$\mathbf{Pr}(w_j|D_i)$ 定义为 softmax 函数：

$$\mathbf{Pr}(\omega_j | D_i) = \frac{\exp(\boldsymbol{D}_i \cdot \boldsymbol{\omega}_j)}{\sum_{\omega_m \in \mathcal{V}} \exp(\boldsymbol{D}_i \cdot \boldsymbol{\omega}_m)}, \tag{8}$$

其中， $\boldsymbol{\omega}_j$ 是 ω_j对应的表示向量。

3 方法

本节介绍新提出的分子指纹方法*Anonymous-FP*的详细内容。在该方法中，分子集合 $\mathcal{G} = \{G_1, \ldots, G_N\}$ 中的每个分子都被分解为一组 r尺度的原子链s，然后我们将它们转换为 r尺度的匿名原子链s，并将分子图嵌入高维向量空间。

3.1 原子链与匿名原子链

在一个分子图 $G_i = (\mathcal{V}_i, \mathcal{E}_i)$ 中，邻接矩阵为$\mathbf{A}$，转移矩阵为$\mathbf{P}$， r尺度的原子链记为马尔可夫链 $w = (v_0, v_1, .., v_r)$。该链是通过智能体从根原子 v_0 逐步行走至终点 v_r 这一过程推导得出的。

接着，匿名原子链将每条原子链转换为一个整数序列，用于记录首次出现的位置。更具体地说，对于 r尺度的原子链 w，其匿名原子链a 是由算子 ψ定义的整数序列，

$$a = \psi(w) = [f(v_0), f(v_1), \ldots, f(v_r)], \tag{9}$$

其中，f是位置函数，使得 $f(v_i) = |\{v_0 \ldots v_i\}|$，其中$\hat{}$是满足 $\hat{v_i} = v_i$的最小整数。

假设智能体在 G_i中的每个根顶点处采样t r尺度的原子链，并将其收集为一个集合，作为分子图 G_i的结构分解，记为 $\mathcal{W}_i = \{w_i^1, \ldots, w_i^t\}$。该原子链集合对应于一个匿名原子链集合$\mathcal{A}_i = \{a_i^1, \ldots, a_i^{\tau_i}\}$，显然 $|\mathcal{A}_i| \leq |\mathcal{W}_i|$。

现在，我们将每个分子结构分解$\mathcal{W}_i$ 收集到一个并集中，记为

$$\mathcal{W} = \bigcup_{i=1}^{N} \mathcal{W}_i = \{w_1, \ldots, w_\mu\}. \tag{10}$$

相应地，匿名原子链集合的并集记为

$$\mathcal{A} = \bigcup_{i=1}^{N} \mathcal{A}_i = \{a_1, \ldots, a_\nu\}. \tag{11}$$

将每条原子链转换为其匿名模式的过程如图2所示。模式转换的基本思想源于两个原因。

(1) 增强对未知结构的表征。在各种先驱分子指纹方法中，生成指纹前始终需要充分了解原子、基团或其他子结构。然而，在匿名模式中，进行随机游走的观察者只根据原子在随机游走中的首次出现来记录每个原子，而不考虑其实际原子类别。这有助于按照统一规则，将化学和生物信息学领域中任何已知或较少为人所知的子结构转换为唯一的数值序列。

(2) 降低计算复杂度。对于每个 r 尺度的原子链 $w = (v_0, v_1, \ldots, v_r)$，其出现概率为

$$P(w) = \prod_{i=0}^{r-1} \mathrm{P}_{i,i+1}, \tag{12}$$

其中，运算符号遵循第2节中的定义。因此，在 G_i 中选择匿名模式 $a = \psi(w)$ 的概率等于

$$P(a) = \sum_{w \in W_i, a = \psi(w)} P(w). \tag{13}$$

此外，我们使用图 3 中的统计数据来验证进行匿名模式处理时的简化效果。这里，我们将尺度定义为 r （范围为 6 到 10），并将原子类别总数定义为 $C > 10$。可能的原子链数量和匿名原子链数量分别以表格和直方图的形式报告。随着 r的增加，可能的原子链数量呈指数级增长，其数量为 C^{r+1}，并且大于 r尺度匿名原子链的数量。这

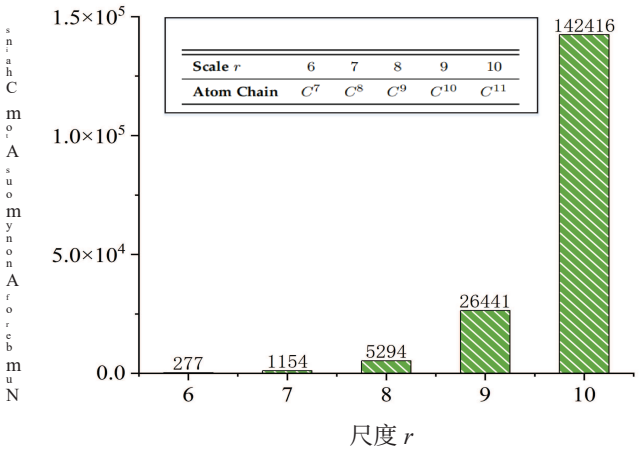


图 3. 目标分子由 C 类原子组成时，可能的原子链和匿名原子链的统计数据（$C > 10$）。

一结果可能归因于这样一个事实：大量分布稀疏的原子链都被压缩为少量匿名原子链位，因此整体计算复杂度显著降低。

3.2 Anonymous-FP

我们提出一种新颖的分子指纹方法，其基础是从分子中挖掘出的匿名原子链。在我们的工作中，采用了一种自然语言处理 技术 *PV-DBOW* 将每个分子编码为定长向量嵌入，并将匿名原子链视为文档（即分子）中包含的词。我们的思路源于相似性性质原理（*SPP*）[16], [17] 即具有相似匿名原子链的分子在嵌入空间中彼此接近。

在数学框架中，我们假设一个 d 维向量 $\boldsymbol{G}_i$ 作为分子 G_i的表示，并假设矩阵$\mathbf{M}$ 作为匿名原子链集合 $\mathcal{A} = \{a_1, \ldots, a_\nu\}$的编码，其中每个行向量$\mathbf{a}_i$ 对应一个匿名原子链 a_i。

总体目标是通过最小化目标函数，将目标分子 G_i 嵌入到d维向量$\mathbf{G}_i$ 中

$$\mathcal{L} = -\sum_{a_j \in \mathcal{A}_i} \log Prob(a_j | G_i). \tag{14}$$

上述条件概率 $Prob(a_j|G_i)$ 定义为 softmax 函数:

$$Prob(a_j | G_i) = \frac{\exp(\boldsymbol{G}_i \cdot \boldsymbol{a}_j)}{\sum_{a_m \in \mathcal{A}} \exp(\boldsymbol{G}_i \cdot \boldsymbol{a}_m)}, \tag{15}$$

其中，$\boldsymbol{G}_i$ 和 $\boldsymbol{a}_m$ 分别是 G_i 和 a_m对应的表示向量。

由于匿名原子链集合 ν 的规模 $\mathcal{A}$往往非常庞大，因此对 softmax 项（15）进行枚举需要消耗大量计算资源。因此，我们采用负采样方法 [25] 来近似计算对数概率（14），即从目标分子 $\tilde{\mathcal{A}}_i^{a_j}$ 中随机抽取一小部分匿名原子链 $\boldsymbol{G}_i$作为负样本，即:

$$\tilde{\mathcal{A}}_i^{a_j} = \mathcal{A}_i / \{a_j\}. \tag{16}$$

在迭代训练中，仅更新目标匿名原子链 a_j 和负样本，而不是集合 $\mathcal{A}$ 中的所有元素。这种策略效率较高，尤其适用于计算复杂度极高的任务。因此，目标函数（14）可改写如下：

$$\begin{aligned}\mathcal{L} &= -\sum_{a_j \in \mathcal{A}_i} \log Prob(a_j|G_i) \\ &= \log\sigma(\boldsymbol{a}_j \cdot \boldsymbol{G}_i) + \sum_{k=1}^{K} \mathbb{E}_{a_k \in \tilde{\mathcal{A}}_i^{a_j}} \log\sigma(-\boldsymbol{a}_k \cdot \boldsymbol{G}_i) \\ &= \log\sigma(\boldsymbol{a}_j \cdot \boldsymbol{G}_i) + \sum_{k=1}^{K} \mathbb{E}_{a_k \in \mathcal{A}_i/\{a_j\}} \log\sigma(-\boldsymbol{a}_k \cdot \boldsymbol{G}_i)\end{aligned} \tag{17}$$

其中 $\sigma(x) = \frac{1}{1+\exp(-x)}$ 表示 Sigmoid 函数，$\boldsymbol{a}_k$ 是负样本的向量嵌入 a_k 从 $\tilde{\mathcal{A}}_i^{a_j}$ 中采样 K 次。我们使用随机梯度下降优化该损失函数（17），并更新 $\boldsymbol{G}_i$ 和 $\boldsymbol{a}_j$。学习过程结束后，我们将这个 d维向量 $\boldsymbol{G}_i$ 称为分子*Anonymous-FP* 的指纹 G_i。

Anonymous-FP 吸收了子结构键指纹、拓扑指纹、药效团指纹及其他指纹类型的优点。原因有两方面。一方面，它重新表述并略微扩展了子结构键指纹和拓扑指纹：所有子结构都通过随机游走模型以原子链的形式采样。随后，原子链被转换为其匿名模式，从而在一定程度上摆脱了对具体子结构键及原子的先验知识的依赖。另一方面，*Anonymous-FP*采用*PV-DBOW* 生成指纹，将目标图视作文档，并将其中包含的匿名原子链视为文档中的词。因此，结构相互作用背后的机制被隐含在向量表示中，这构成了该分子指纹方法最初的出发点。

算法1 和图4 概述了*Anonymous-FP*的框架。初始化时，首先通过正态分布$\mathcal{N}(0, 0.01)$ 随机设定分子嵌入向量 $\boldsymbol{G}_i$ 以及匿名原子链向量$\boldsymbol{a}_j$、 $j = 1, \ldots, \nu$，随后通过梯度下降迭代计算这些嵌入向量，直至收敛。

算法 1*Anonymous-FP*

Input: molecules set $\mathcal{G} = \{G_1, \ldots, G_N\}$; anonymous atom chains set $\mathcal{A} = \bigcup_{i=1}^{N} \mathcal{A}_i = \{a_1, \ldots, a_\nu\}$; scale r; vector dimension d

输出： *Anonymous-FP* $\boldsymbol{G}_i$ 用于分子 G_i

初始化 $\boldsymbol{G}_i = [\epsilon_s]^{1\times d},\ \epsilon_s \sim N(0,\ 0.01)$

初始化 $\boldsymbol{a}_j = [\varepsilon_s]_{1\times d},\ \varepsilon_s \sim N(0, 0.01),\ j = 1,\ \ldots,\ \nu$

遍历每个匿名原子链 a_j 在 $\mathcal{A}_i$ **执行**

  从集合 $\mathcal{A}$中采样负样本集 $\tilde{\mathcal{A}}_i^{a_j}$

  $\mathcal{L} = \log\sigma(\boldsymbol{a}_j \cdot \boldsymbol{G}_i) + \sum_{k=1}^{K} \mathbb{E}_{a_k \in \tilde{\mathcal{A}}_i^{a_j}} \log\sigma(-\boldsymbol{a}_k \cdot \boldsymbol{G}_i)$

  $\boldsymbol{G}_i = \boldsymbol{G}_i - \alpha\frac{\partial\mathcal{L}}{\boldsymbol{G}_i}$

  $\boldsymbol{a}_j = \boldsymbol{a}_j - \alpha\frac{\partial\mathcal{L}}{\boldsymbol{a}_j}$

结束循环

返回 *Anonymous-FP* $\boldsymbol{G}_i$

4 实验

本节中，为了验证所提出方法的效率，我们在分子图数据集上针对分子图分类任务开展了广泛实验。该任务是一种监督模式，其训练数据由输入数据（即每个分子的*Anonymous-FP* ）与期望输出标签（即目标理化性质）构成。结果表明，与若干常用基线方法相比，我们的方法能够取得更优的性能。接下来还将简要讨论*Anonymous-FP* 的表达能力，以及对步行驱动样本的全面探索。

4.1 数据集

Anonymous-FP 在一系列真实世界的分子图数据集上进行了测试：**NCI-1**、**NCI-109**、**PROTEINS**、**DD**、**MUTAG**和**PTC**。每个数据集都属于特定类型的物理化学性质，并以活性和非活性作为标签。相关统计数据见表 1 ，简要描述如下。

表1 基准图数据集的统计信息。各列分别表示数据集名称、正标签图数量、图数量、类别数量以及节点/边的平均数量。

数据集	正标签图	总数	类别	平均节点数	平均边数
NCI-1	2055	4110	2	29.87	32.30
NCI-109	2063	4126	2	29.68	32.13
PROTEINS	556	1112	2	39.06	72.82
DD	589	1178	2	284.32	715.66
MUTAG	94	188	2	17.93	19.79
PTC	172	344	2	14.29	14.69

- **NCI-1**、**NCI-109** [26] 是按照抗癌性质（活性或阴性）划分的化合物数据集。这些数据集由美国国家癌症研究所（NCI）公开发布。
- **PROTEINS** [27] 是一组蛋白质图，其中节点表示二级结构元件，边表示它们在氨基酸序列中或三维空间中的邻近关系。
- **DD** [28] 是一个蛋白质结构数据集，其中节点表示氨基酸，边表示空间上的接近关系；该数据集分为酶和非酶。
- **MUTAG** [5] 是一个由芳香族和杂芳香族硝基化合物构成的数据集，这些化合物根据其是否对细菌具有致突变效应进行标注。
- **PTC** [29] 由化学分子的图表示构成，并根据其对雄性和雌性大鼠的致癌性进行标注。

4.2 基线方法

为充分展示我们模型的卓越性能，我们将其与一系列基线方法进行比较。

- **Graphlet核** [30]: Graphlet核（GK）通过统计共同的 k 节点图元来衡量图相似性

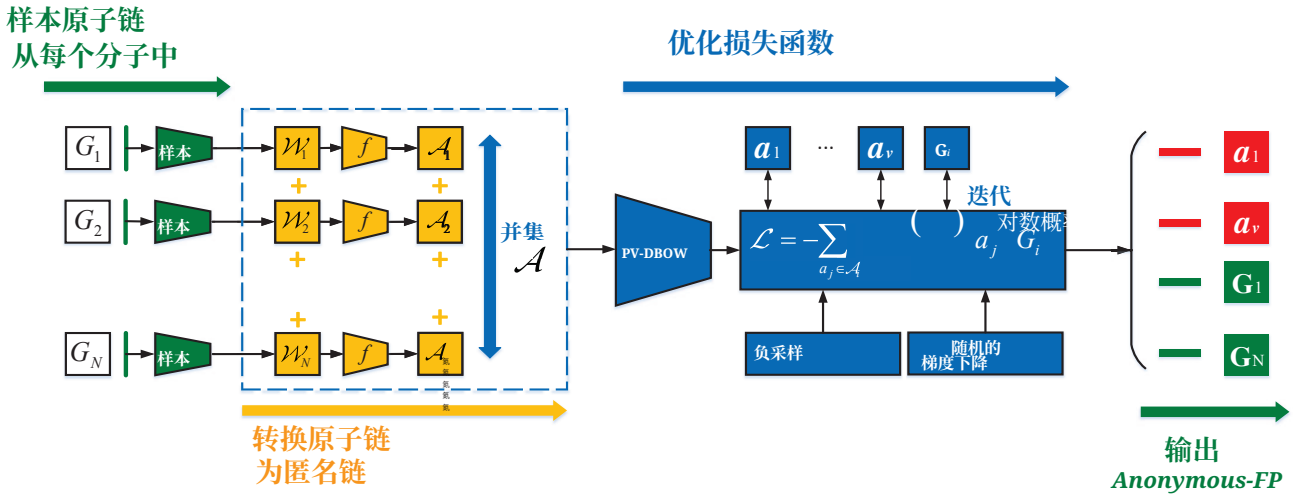


图 4. *Anonymous-FP* 算法：从初始图数据中采样原子链，然后将每条原子链转换为其匿名原子链。通过取所有匿名原子链集合的并集，构建集合 $\mathcal{A}$，随后将其输入嵌入模型 *PV-DBOW*。之后，迭代输出 *Anonymous-FP* 。

图元 ，这确保了计算复杂度受限于多项式时间，杂度受限于多项式时间。

- **Weisfeiler-Lehman 核** [31]：Weisfeiler-Lehman 核（WL）将图数据映射为 Weisfeiler-Lehman 序列，其节点属性表示图拓扑和标签信息。由于运行时间随图的边数和 Weisfeiler-Lehman 图序列的长度呈线性增长，WL 核被广泛用于图的同构测试。

- **Graph2vec** {v4>：Graph2vec 将有根子图视为词，将图视为句子或文档，然后在自然语言处理 中使用 Skip-gram 来获得显式的图嵌入。
- **AWE** [33]：AWE 使用匿名随机游走以无监督方式嵌入整个图，但与我们的方法相比，它采用了不同的嵌入策略。AWE 利用匿名游走的邻域，而我们的工作则关注全局范围内共现的匿名游走。

- **PATCHY-SAN** [34]：类似于卷积神经网络（CNN）[35]，[36]，，PATCHY-SAN（PSCN）提出了一个框架，用于对任意图数据执行卷积操作。
- **GraphSAGE** [37]：GraphSAGE 利用归纳式框架，通过采样和聚合 1 跳及 2 跳邻域特征来计算图嵌入。

4.3 实现与超参数

本文使用 Python 3.6.12、TensorFlow 1.2.0、Scikit-learn 0.24.1、NumPy 1.22.1 和 NetworkX 2.6.3 作为计算环境，所有实验均在配备 2 颗英特尔至强 CPU 和 4 块 NVIDIA GeForce GTX 1080 Ti GPU 的工作站上进行。我们首先将每个数据集随机划分为 10 个等份，选取 9 份用于训练，1 份用于测试效率。为进行公平评估，我们将所有数据集的 *Anonymous-FP* 大小统一设为 128。事实上，*Anonymous-FP* 表示与超参数尺度 r 以及采样数量 t 之间存在紧密的内在关联。为探究这种关联，我们将 *Anonymous-FP* 视为 r 的函数，其范围从短尺度 6 逐步增至 10；同时也将其视为 t 的函数，其取值从 10 增至 160。随后，我们使用径向基函数核 [38] 的支持向量机建立分子图分类器，以测试并验证*Anonymous-FP* 的判别能力，并讨论在超参数 r、 t控制下分类准确率的变化趋势。

4.4 性能评估指标

大多数评估指标都源自以下五个术语：准确率、精确率、召回率、F1 分数和 ROC-AUC。

- **准确率** [39]，[40] 用于计算模型对活性或非活性项目做出正确预测的概率。
- **精确率** [39]，[40] 用于估算模型做出正确活性类别预测的概率。
- **召回率** [39]，[40]，也称为真正率或灵敏度，表示被正确预测为活性化学物质的比例。
- **F1 分数** [39]，[40]，是精确率和召回率的平衡指标，取值范围为 0 到 1。F1 分数越高，表示分类器的性能越好。
- **ROC-AUC** [41]。受试者工作特征（ROC）曲线用于展示预测变量与真实结果的比较情况。通常，ROC 曲线反映了在不同阈值下，灵敏度（真正率）如何随特异度（真负率）的变化而变化。变量的预测能力通常用曲线下面积（AUC）概括，该值可通过对线段下方面积进行积分计算得出。

表 2 我们的方法与基线方法在真实世界分子数据集上的平均分类准确率（均值 ± 标准差 %）。最佳结果以 **粗体**标示。

算法	NCI-1	NCI-109	PROTEINS	DD	MUATG	PTC
Graphlet核 [30]	62.28 ± 0.29	62.60 ± 0.19	71.67 ± 0.55	78.45 ± 0.26	80.63 ± 3.07	57.26 ± 1.41
Weisfeiler-Lehman 核 [31]	80.13 ± 0.50	80.22 ± 0.3	72.92 ± 0.56	77.95 ± 0.70	81.66 ± 2.11	56.97 ± 2.01
Graph2vec [32]	73.22 ± 1.81	74.26 ± 1.47	73.30 ± 2.05	58.64 ± 0.01	83.15 ± 9.25	60.17 ± 6.86
AWE [33]	62.72 ± 1.67	63.21 ± 1.42	70.01 ± 2.52	71.51 ± 4.02	87.87 ± 9.76	59.14 ± 1.83
PATCHY-SAN [34]	78.59 ± 1.89	-	75.89 ± 2.76	77.12 ± 2.41	92.63 ± 4.21	60.00 ± 4.82
GraphSAGE [37]	74.73 ± 1.34	74.17 ± 2.89	74.01 ± 4.27	75.78 ± 3.91	78.75 ± 1.18	-
Anonymous-FP	95.74 ± 0.96	95.95 ± 0.74	72.32 ± 4.77	94.83 ± 1.67	81.58 ± 4.85	61.14 ± 5.29
典型尺度 r	8	8	9	8	10	9
采样数量 t	40	70	40	110	30	130

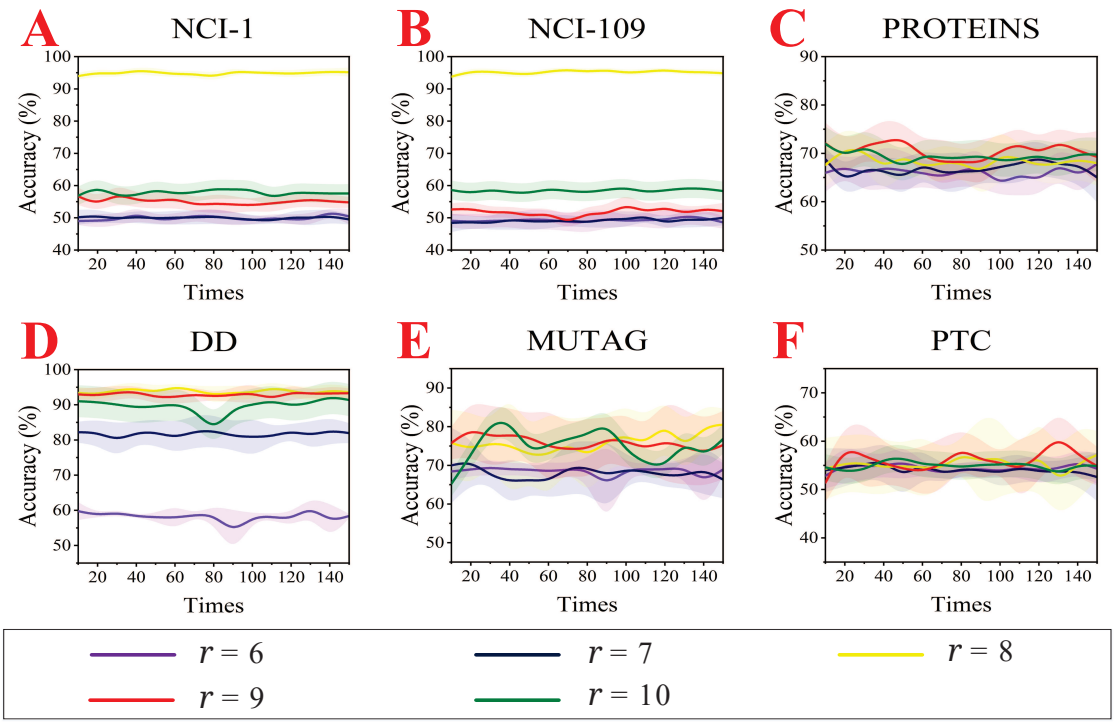


图 5. 使用径向基函数核的支持向量机分类器在 NCI-1（A）、NCI-109（B）、PROTEINS（C）、DD（D）、MUTAG（E）和 PTC（F）上的监督式图分类性能。结果以尺度 r 和采样数量 t为变量绘制。阴影表示每个采样点分类结果的标准差。

4.5 总体结果

4.5.1 准确率

表2 总结了基线方法和*Anonymous-FP*得出的分类结果，同时还列出了*Anonymous-FP* 达到最佳性能时所采用的典型尺度 r 和采样次数 t 。从表中可以看出，*Anonymous-FP* 在 NCI-1、NCI-109、DD 和 PTC 上表现出强大的判别能力，其分类准确率均远高于各数据集上表现最好的基线方法，分别达到95.74%、95.95%、94.83%和61.14%。然而，在 PROTEINS 和 MUTAG 数据集上，*Anonymous-FP*未能优于基线方法，分别比最佳结果低约3.57% 和10.95%。

此外，我们进一步考察了*Anonymous-FP* 分类性能随连续 rs 和 ts 变化的模式，趋势如图 5所示。

我们首先展示分类性能随尺度 $r \in [6, 7, 8, 9, 10]$ 的变化情况，该尺度刻画了分子中从小尺度到大尺度的原子链和匿名原子链。

当 r 取值为 6 或 7 时，对于所有数据集，随着采样数量 t 增加，准确率在整个采样过程中都在初始值附近波动。具体而言，NCI-1 或 NCI-109 上的准确率几乎等于50%，这意味着*Anonymous-FP* 对这两个均衡标注数据集几乎没有区分能力。对于 DD 和 PROTEINS，当尺度为 $r = 7$ 时，增加采样数量 t 仍然无法提高分类准确率。

对于中等规模的 r 值，即 $r = 8$，结果表明 *Anonymous-FP* 与六个数据集上较高的分类性能密切相关。总体来看，对于 NCI-1、NCI-109 和 DD数据集，当规模为 $r = 8$时，分类性能优于最佳准确率。对于 NCI-1 和 NCI-109，我们可以清楚地看到增长趋势

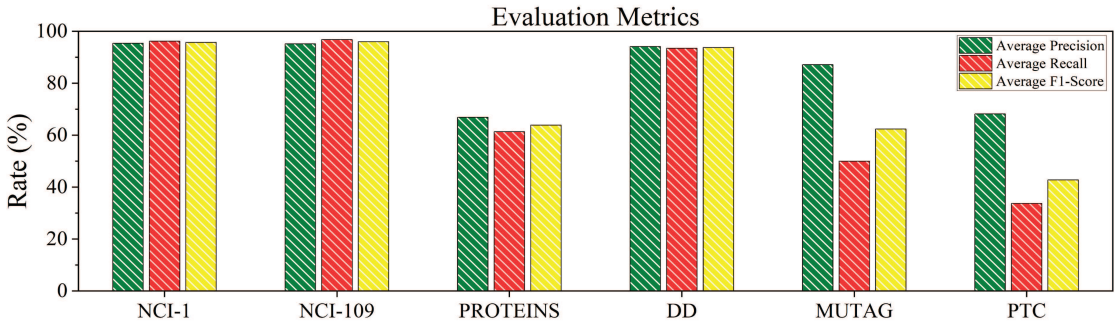


图 6. 评估 在不同数据集上的指标，包括平均精确率、平均召回率和平均 F1 分数，报告 d.

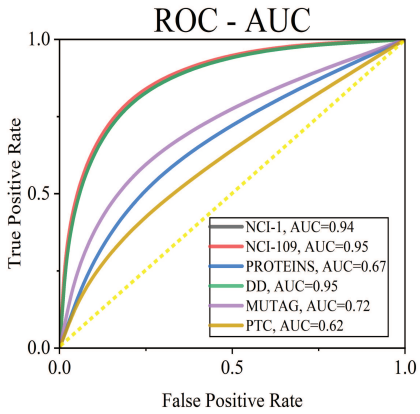


图 7. 不同数据集的 ROC 曲线和 AUC 值。我们展示了不同数据集的标准 ROC 曲线，并用不同颜色区分。这里，横轴为假阳性率，纵轴为真正率。对角虚线表示恒等线。

在规模 r 变为 8 时，增长幅度骤增，几乎达到30%，曲线分别达到95.74% 和95.95%的最大值，并在高位附近波动。对于 DD数据集，当 $r = 8$时，也表现出最佳性能。虽然 MUTAG 上并非如此，其最佳性能是在规模为10时达到的。

对于较大的 r 尺度，即 9 或 10，*Anonymous-FP* 在 NCI-1、NCI-109 上的表现不如尺度r等于 8 时具有竞争力。一个显著的现象是，尺度 r从 9增加或减小后，准确率便会急剧下降，分类准确率不超过 60%。

因此，基于上述分析，我们发现 *Anonymous-FP* 与超参数匿名原子链尺度 r 密切相关。具体而言，基于 8 尺度匿名原子链的 *Anonymous-FP* 具有显著的分子区分能力。然而，这种分类性能可能会突然消失，尤其是在 NCI-1 和 NCI-109 中，当尺度发生相对较小的变化时，甚至是 r 从 8 增加到 9 时也是如此。但仍有必要进一步研究，以验证这一现象是否普遍存在于其他 NCI分子数据集 中。

4.5.2 精确率、召回率、*F1* 分数和 *ROC-AUC*

精确率、召回率、F1 分数和 ROC-AUC 也是评估 *Anonymous-FP* 的重要指标。结果见图 6 和图7。

在图6中，我们比较了*Anonymous-FP* 在 NCI-1、NCI-109、PROTEINS、DD、MUTAG 和 PTC 分子数据集上的精确率、召回率和 F1 分数。在 NCI-1、NCI-109 和 DD 上，*Anonymous-FP* 能够将精确率、召回率和 F1 分数提升至 90 % 以上。这表明，当前方法有望提高预测阳性样本中真实阳性样本所占比例、真实阳性样本中被正确预测为阳性的比例，以及整体平衡表现。在 PROTEINS、MUTAG 和 PTC 上，通常精确率高于召回率和 F1 分数。这意味着选择通过 *Anonymous-FP* 对分子进行建模，有利于提高大量阳性样本的预测精确率。图7 中的 ROC-AUC 用于在一系列阈值中确定最佳模型；曲线下面积（AUC）越大，模型的综合表现越好。对于 NCI-1、NCI-109 和 DD，AUC 不低于 0.94，这与图6 和表2的结果一致。PROTEINS、MUTAG 和 PTC 的 AUC 分别为 0.67、0.72 和 0.62。

4.6 NCI数据集上的补充实验

此外，我们将所提出的方法应用于更多 NCI 数据集，详情汇总于表 3。每个实例代表一组对特定癌症具有活性或非活性作用的分子，每组均按平衡的先验标签分布划分。在这些实验中，我们关注采样数量 t（范围为 10 至 160）和匿名原子链尺度 r（取值为 6、7、8、9 或 10）对分类的影响。表 4 展示了所提出方法取得的最佳结果、典型尺度r以及每个 NCI 数据集的采样时间 t。

在图 8 中，对于每个数据集，都可以清楚地看到，当 $r = 7$ 时，无论 t取何值，我们的方法表现都不尽如人意。这意味着在这种情况下，采样时间 t 对结果影响甚微。一旦尺度 r 变为 8，所有 NCI 数据集的结果便迅速升至 93% 以上，随后随着 t 增加，结果保持在最佳水平附近。值得注意的是，这 6 个数据集对尺度 $r = 9$的反应各不相同：NCI-33、NCI-41 和 NCI-47 受尺度 $r = 9$ 的影响较小，而 NCI-81 、 NCI-83 或

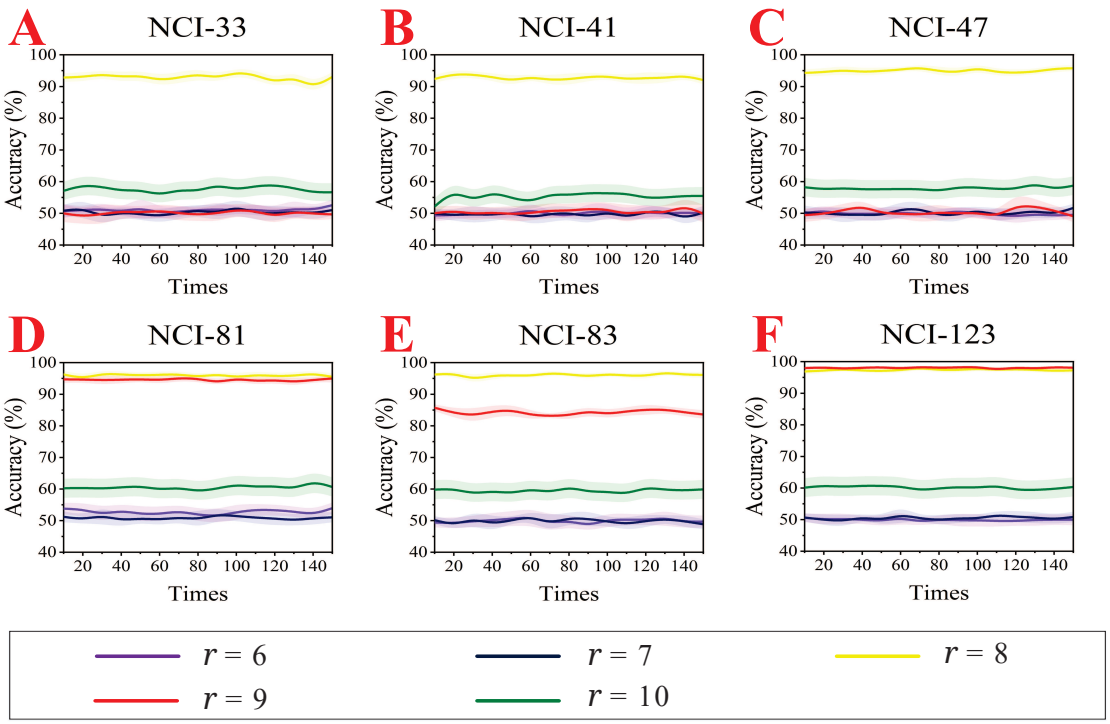


图8. 使用径向基函数核的支持向量机分类器进行NCI-33（A）、NCI-41（B）、NCI-47（C）、NCI-81（D）、NCI-83（E）和NCI-123（F）的监督式图分类性能。

当尺度 $r = 9$ 时，NCI-123 的表现能够提升。此外，图 9 还通过箱线图报告了整体精确率、召回率和 F1 分数，这些指标可作为补充指标。最后，上述分析结果明确支持了我们的推断：由8尺度匿名原子链生成的 *Anonymous-FP* 能够更好地区分分子，而且这一结论普遍适用于大量 NCI数据库。

表3 NCI数据集的统计信息。各列分别为数据集名称、正标签图数量、总图数以及相应数据集的描述。

数据集	正标签图	总图数	肿瘤描述
NCI-33	1467	2934	黑色素瘤
NCI-41	1350	2700	前列腺
NCI-47	1735	3470	中枢神经系统
NCI-81	2081	4162	结肠
NCI-83	1959	3918	乳腺
NCI-123	2715	5430	白血病

表4 各个NCI数据集的分类结果（均值 ± 标准差 %）、典型尺度 r和采样数量 t 。

数据集	平均准确率	典型尺度 r	采样数量 t
NCI-33	95.53 ± 1.17	8	100
NCI-41	93.92 ± 1.30	8	20
NCI-47	96.02 ± 0.92	8	70
NCI-81	96.57 ± 0.85	8	140
NCI-83	96.91 ± 0.57	8	130
NCI-123	97.78 ± 0.63	8	100

5 结论与后续展望

分子通常以顶点—边拓扑图结构的形式存在，该结构由一组原子链构成。原子链的匿名模式反映了分子内各项之间的强关联，并承载着数据的底层语义。本文提出了一种新颖的分子指纹方法 *Anonymous-FP*，利用嵌入式匿名原子链分解来区分分子。该方法的优势在于，它采用*NLP* 技术 *PV-DBOW* 将每个分子编码为一个向量，该向量可用于表征分子整体结构，而无需理解每个原子的化学符号。

值得注意的是，匿名原子链的尺度 r 在分子性质表示中发挥着重要作用。通常，典型尺度 $r = 8$ 能够显著提高一系列数据集上的区分准确率，而且这一有趣现象对于更多NCI数据集也普遍成立。然而，尺度 $r = 8$ 通常能够促进表示的潜在原因目前仍然未知。此外，进一步深入了解适用于更大分子表示的、更有效的指纹设计也颇为重要。

致谢

本研究得到国家自然科学基金（项目编号：62276013、62141605、62050132）、北京市自然科学基金（项目编号：1192012）以及中央高校基本科研业务费的资助。

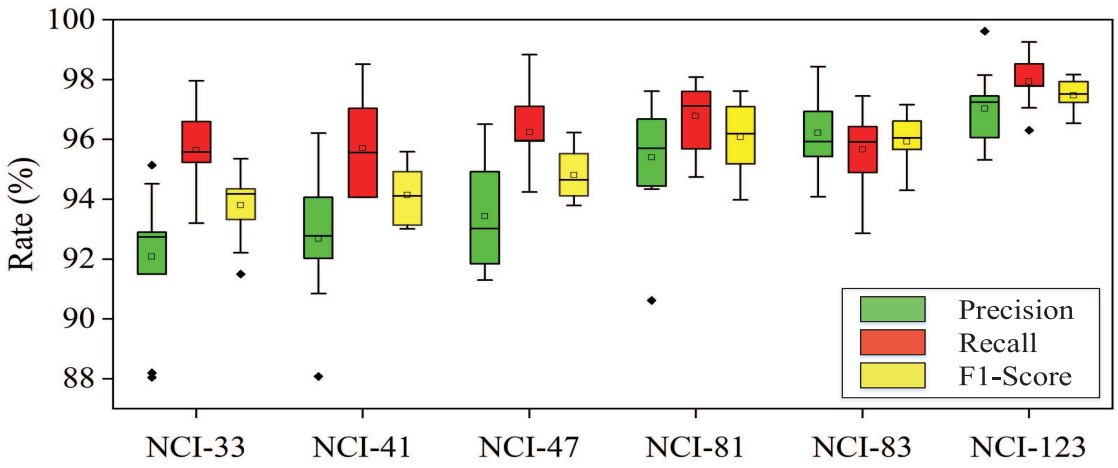


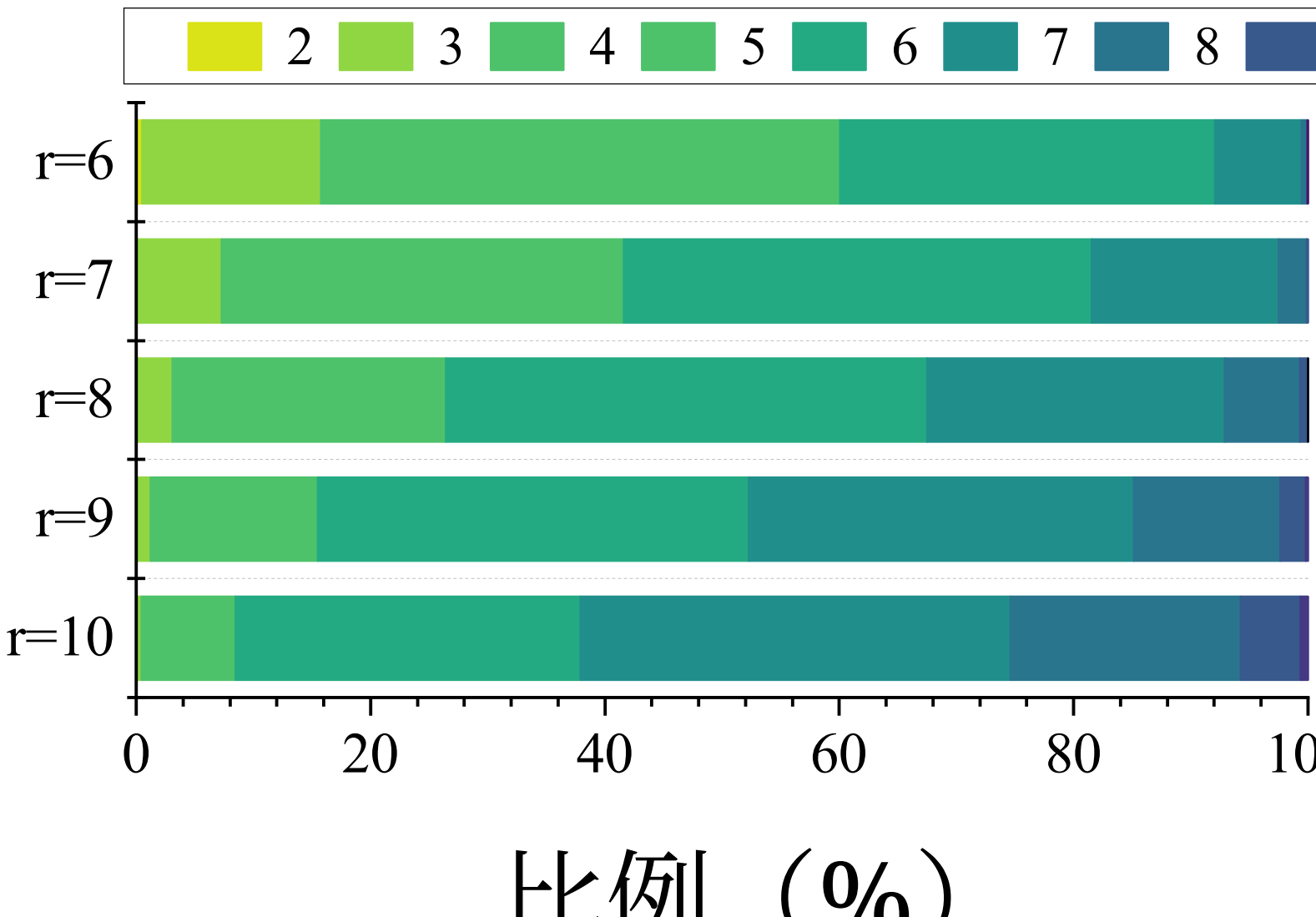
图 9. 报告了 8 种尺度的 *Anonymous-FP* 在不同数据集上的精确率、召回率和 F1 分数箱线图。箱线图包含数据中的极值（最大值和最小值）、下四分位数和上四分位数、中位数以及极端离群值。在每个箱线图中，除以实心节点表示的离群值外，大多数数值都分布在下须线和上须线之间。下四分位数和上四分位数构成箱体，实线表示中位数。

参考文献

[1] A. H. Beckett 和 J. B. Stenlake。《实用药物化学》。实用药物化学，1968年。[2] Martin EM Noble、Jane A Endicott 和 Louise N Johnson。蛋白激酶抑制剂：从结构中探究药物设计。《科学》，303(5665)：1800–1805，2004年。[3] John Boaz Lee、Ryan Rossi 和 Xiangnan Kong。利用结构注意力进行图分类。载于《第24届美国计算机协会SIGKDD知识发现与数据挖掘国际会议论文集》，第1666–1674页，2018年。[4] Naveen Chhabra、Madan L Aseri 和 Deepak Padmanabhan。药物异构及其意义综述。《国际应用与基础医学研究杂志》，3(1)：16，2013年。[5] Asim Kumar Debnath、Rosa L. Lopez De Compadre、Gargi Debnath、Alan J. Shusterman 和 Corwin Hansch。致突变芳香族和杂芳香族硝基化合物的构效关系：与分子轨道能量和疏水性的相关性。《药物化学杂志》，34(2)：786–797，1991年。[6] Martin Grohe 和 Pascal Schweitzer。图同构问题。《ACM通讯》，63(11)：128–134，2020年。[7] László Babai。群、图与算法：图同构问题。载于《2018年里约热内卢国际数学家大会论文集》，第3319–3336页。世界科学出版社，2018年。[8] Juris Hartmanis。《计算机与难解性：NP完全性理论指南》（Michael R. Garey 和 David S. Johnson 著）。《SIAM评论》，24(1)：90，1982年。[9] Robert D Brown 和 Yvonne C Martin。与配体-受体结合相关的二维和三维结构描述符的信息含量。《化学信息与计算机科学杂志》，37(1)：1–9，1997年。[10] Sunghwan Kim、Paul A Thiessen、Evan E Bolton、Jie Chen、Gang Fu、Asta Gindulyte、Lianyi Han、Jane He、Siqian He、Benjamin A Shoemaker 等。PubChem物质和化合物数据库。《核酸研究》，44(D1)：D1202–D1213，2016年。[11] Robert C Glen、Andreas Bender、Catrin H Arnby、Lars Carlsson、Scott Boyer 和 James Smith。圆形指纹：灵活的分子描述符，从物理化学到吸收、分布、代谢和排泄的应用。《IDrugs》，9(3)：199，2006年。[12] Alice Capecchi、Daniel Probst 和 Jean-Louis Reymond。一种统摄一切的分子指纹：药物、生物分子和代谢组。《化学信息学杂志》，12(1)：1–15，2020年。[13] Aleksandrs Leitis、Andreas Tittl、Mingkai Liu、Bang Hyun Lee、Man Bock Gu、Yuri S. Kivshar 和 Hatice Altug。用于宽带分子指纹提取的角度复用全介质超表面。《科学进展》，5(5)：eaaw2871，2019年。

[14] Silvio Micali 和 Zeyuan Allen Zhu。从独立且匿名的实验中重建马尔可夫过程。《离散应用数学》，200:108–122，2016年。[15] Quoc Le 和 Tomas Mikolov。句子和文档的分布式表示。载于《国际机器学习会议》，第 1188–1196 页。PMLR，2014年。[16] Jürgen Bajorath。面向信息学应用的分子相似性概念。载于《生物信息学》，第 231–245 页。Springer，2017年。[17] Miquel Duran-Frigola、Eduardo Pauls、Oriol Guitart-Pla、Martino Bertoni、Víctor Alcalde、David Amat、Teresa Juan-Blanco 和 Patrick Aloy。借助化学检查器，将小分子相似性原理扩展至生物学的所有层面。《Nature Biotechnology》，38(9):1087–1096，2020年。[18] John M Barnard 和 Geoff M Downs。化学片段生成与聚类软件。《化学信息与计算机科学杂志》，37(1):141–142，1997年。[19] Jianxin Duan、Steven L. Dixon、Jeffrey F. Lowrie 和 Woody Sherman。二维指纹的分析与比较：通过八种指纹方法深入了解数据库筛选性能。《分子图形与建模杂志》，29(2):157–170，2010年。[20]Mickaël Lelimousin、Marjolaine Noirclerc-Savoye、Christelle Lazareno-Saez、Bernhard Paetzold、Sophie Le Vot、Richard Chazal、Pauline Macheboeuf、Martin J. Field、Dominique Bourgeois 和 Antoine Royant。ECFP 和 Cerulean 的内在动力学对荧光量子产率的调控。《生物化学》，48 10046，2009 42 10038 19754158 ( ): – 。PMID: 。[21] Ling Xue、Jeffrey W. Godden、Florence L. Stahura 和 Jürgen Bajorath。轮廓缩放提高了包含数值描述符和结构键的分子指纹的相似性搜索性能。《化学信息与计算机科学杂志》，43(4):1218–1225，2003年。PMID: 12870914。[22] Adrià Cereto-Massagué、María José Ojeda、Cristina Valls、Miquel Mulero、Santiago Garcia-Vallvé 和 Gerard Pujadas。虚拟筛选中的分子指纹相似性搜索。《方法》，71:58–63，2015年。虚拟筛选。[23] David Vidal、Michael Thormann 和 Miquel Pons。Lingo：一种高效的全息文本方法，用于计算生物物理性质和分子间相似性。《化学信息与建模杂志》，45(2):386–393，2005年。PMID: 15807504。[24] Istyastono EP Radifar M，Yuniarti N。PyPLIF：基于 Python 的蛋白质-配体相互作用指纹。《生物信息》，9 328，2013 6 325( ): – 。[25] Junyang Chen、Zhiguo Gong、Wei Wang 和 Weiwen Liu。HNS：用于网络表示学习的分层负采样。《信息科学》，542:343–356，2021年。[26] John Boaz Lee、Ryan Rossi 和 Xiangnan Kong。使用结构注意力进行图分类。载于*Proceedings of the 24th ACM*

美国计算机协会知识发现与数据挖掘特别兴趣小组国际会议，第1666–1674 页，2018年。[27] Karsten M. Borgwardt、Ong Cheng Soon、Schönauer Stefan、S. V. N. Vishwanathan、Alex J. Smola 和 Kriegel Hans-Peter。通过图核预测蛋白质功能。《生物信息学》，21(suppl_1):i47–i56，2005年。[28] Paul D. Dobson 和 Andrew J. Doig。无需比对即可区分酶结构与非酶结构。《Journal of Molecular Biology》，330(4):771–783，2003年。[29] C. Helma、R. D. King、S. Kramer 和 A. Srinivasan。2000—2001年预测毒理学挑战。《生物信息学》，17(1):107–108，2001年。[30] Nino Shervashidze、SVN Vishwanathan、Tobias Petri、Kurt Mehlhorn 和 Karsten Borgwardt。用于大型图比较的高效 Graphlet 核。载于《人工智能与统计》，第 488–495 页，2009年。[31] Nino Shervashidze、Pascal Schweitzer、Erik Jan、Van Leeuwen 和 Karsten M. Borgwardt。Weisfeiler-Lehman图核。《机器学习研究杂志》，1(3):1–48，2010年。[32] Annamalai Narayanan、Mahinthan Chandramohan、Rajasekar Venkatesan、Lihui Chen、Yang Liu 和 Shantanu Jaiswal。Graph2vec：学习图的分布式表示。*arXiv*预印本 *arXiv:1707.05005*，2017年。[33] Sergey Ivanov 和 Evgeny Burnaev。匿名游走嵌入。载于《国际机器学习会议》，第 2186–2195 页。PMLR，2018年。[34] Mathias Niepert、Mohamed Ahmed 和 Konstantin Kutzkov。学习用于图的卷积神经网络。载于《国际机器学习会议》，第 2014–2023 页。PMLR，2016年。[35] Thomas N Kipf 和 Max Welling。利用图卷积网络进行半监督分类。载于国际学习表征会议论文集，2017年。[36] Jiuxiang Gu、Zhenhua Wang、Jason Kuen、Lianyang Ma、Amir Shahroudy、Bing Shuai、Ting Liu、Xingxing Wang、Gang Wang、Jianfei Cai 等。卷积神经网络的最新进展。《模式识别》，77:354–377，2018年。[37] William L Hamilton、Rex Ying 和 Jure Leskovec。大型图上的归纳式表示学习。载于第*31*届神经信息处理系统国际会议论文集，第 1025–1035 页，2017年。[38] Murilo VF Menezes、Luiz CB Torres 和 Antonio P Braga。支持向量机二分类中径向基函数核的宽度优化：一种基于密度估计的方法。《模式识别快报》，128:1–7，2019年。[39] Tuomas Kynkäänniemi、Tero Karras、Samuli Laine、Jaakko Lehti- nen 和 Timo Aila。用于评估生成模型的改进精确率与召回率指标。神经信息处理系统进展，32，2019年。[40] Michael Buckland 和 Fredric Gey。召回率与精确率之间的关系。《美国信息科学学会期刊》，45(1):12–19，1994年。[41] John Muschelli。二元预测变量的受试者工作特征曲线和曲线下面积：一种可能具有误导性的指标。《分类学杂志》，37(3):696–708，2020年。


2
3
4
5
6
7
8
r=6
r=7
r=8
r=9
r=10
0
20
40
60
80
10
比例（%）