混合图位置编码下节点定位的信息论极限

Zimo Yan[1], Zheng Xie[1],Chang Liu[1],Yiqin Lv[1],Runfan Duan[1]。

[1]国防科技大学，中国长沙。*{yanzimo20, xiezheng81, liuchang_, lvyiqin98, duanrunfan24 }@nudt.edu.cn*

arXiv:2603.25030v1 [cs.IT] 26 Mar 2026

摘要——**位置编码已成为图学习中的标准组成部分，尤其适用于图 Transformer 及其他需要区分结构相似节点的模型，但其根本性的可识别性仍未得到充分理解。本文研究一种混合位置编码下的节点定位，该编码结合了锚点距离剖面与量化低频谱特征。我们将定位问题表述为一个观测映射问题，其难度由编码所产生的不同代码数量决定，并建立了一个信息论逆定理，指出由锚点数量、谱维数和量化级别共同决定的不可实现机制。实验进一步支持了这一观点：在随机 3-正则图上，实证交叉点与预测的缩放关系高度吻合；而在两个真实世界 DDI 图上，可识别性强烈依赖于图本身：在所测试的编码下，DrugBank 仍存在大量冗余，而 Decagon 派生图在谱信息足够丰富时几乎变为单射。总体而言，这些结果表明，位置编码不应仅被视为一种启发式架构组件，而应被理解为一种依赖于图的结构分辨机制。**

关键词：图位置编码；节点可识别性；谱嵌入；距离编码；图神经网络；药物-药物相互作用图

I. 引言

图神经网络和图 Transformer 在关系数据上取得了优异表现，但其有效性往往取决于能否在底层图中有意义地区分节点位置。因此，位置编码和结构编码已成为现代图学习中的重要课题 [1], [2], [3], [4]。

现有的图位置编码方法大致可分为三类。一类采用谱编码，通常由拉普拉斯特征向量或其不变变体推导而来，用于捕捉图的全局几何结构 [5], [6]。另一类采用基于距离或结构的编码，例如节点到锚点的最短路径距离、随机游走信息或相关结构描述符 [7], [8]。还有一类工作侧重于学习式或混合编码器，在图 Transformer 或其他图模型中组合多种位置信号 [2], [9], [4]。

本研究得到湖南省重大科技创新平台项目 [2024JC1003] 和国防科技大学研究生创新项目[XJQY2024065]的资助。作者谨向所有审稿人致以诚挚谢意，感谢他们认真审阅本文并提出宝贵建议。*通讯作者：郑燮（xiezheng81@nudt.edu.cn）

尽管取得了这些进展，仍有三项局限性。现有文献大多通过下游准确率间接评估位置编码，因此尚未明确编码本身能否唯一地区分节点。谱方法还存在符号、基底和稳定性方面的问题，需要谨慎处理 [5], [6]。更广泛地说，位置编码的影响显然取决于图，但从理论角度看，这种依赖关系仍未得到充分理解[10], [11]。

研究动机：由此引出以下问题：我们能否刻画图位置编码在何种情况下包含足够的信息来定位节点，以及这种可识别性如何共同取决于距离特征、谱特征和量化？

为回答这一问题，我们通过一种混合观测映射研究节点定位，该映射结合了锚点距离剖面与量化低频谱特征。随后，我们针对这一映射展开信息论逆定理分析，并确定一个由锚点数、谱维数和量化级别共同决定的不可能机制。在随机正则图和两个真实世界 DDI 图上的实验进一步表明，同一种编码在不同图结构上的表现可能大相径庭。

概要与主要贡献：本文的主要贡献如下：

1）我们通过一种结合锚点距离特征与量化低频谱特征的混合观测映射研究节点定位，并推导出一个通用的信息论逆定理，确定由锚点数、谱维数和量化级别共同决定的不可能机制。2）对于典范能量嵌入，我们进一步通过逐桶碰撞与平衡细化收紧通用像集大小界，从而得出结构更明确的不可能性判据。3）我们通过在随机 3-正则图和两种真实世界 DDI 图构造上的实验验证理论，展示了合成场景中类似相变的交叉，以及 DrugBank 与 Decagon 派生图之间显著的图依赖性差异。

本文其余部分安排如下：第二节回顾相关工作，第三节介绍问题表述，第四节阐述理论，第五节展示实验与实证分析，第六节总结全文。

II. 文献综述

图位置编码已成为现代图学习中的核心课题，尤其是在图 Transformer 及相关架构中；这些架构需要显式的位置或结构信息来区分结构相似的节点 [12], [13],[14]。现有方法大致可分为三类：谱位置编码、基于距离和结构感知的编码，以及学习式或混合式位置/结构编码器。

A. 谱位置编码

谱位置编码通过拉普拉斯特征向量、特征值或相关谱坐标描述节点位置，因此能够注入纯局部消息传递无法获取的全局几何信息 [9], [15],[16]。代表性例子包括可学习的结构与位置表示 [9], 、等变且稳定的谱编码 [15],以及诸如 SignNet/BasisNet [5] 这样的符号/基不变模型。近期研究进一步探讨了谱编码的稳定性与表达能力 [6], [17],，开发了随机特征传播和基于层的定位编码等替代性谱构造或受谱启发的构造 [18], [19], ，并将讨论扩展至有向图；在有向图中，位置信息更难以一致地定义 [20]。与此同时，这些方法受到特征向量符号和基底不确定性的影响，可能对图扰动反应敏感，而且在大型图上通常依赖特征分解这一非平凡的预处理步骤。

B. 距离与结构感知编码

第二类方法通过相对结构信息来刻画节点位置，例如基于锚点的距离、最短路径距离、电阻距离、随机游走统计量或图重连模式 [21], [22], [23]。距离编码为这一思路提供了一种特别清晰、注重表达能力的表述 [7], ，而 P-GNN 及相关的基于锚点的方法则展示了如何直接利用到采样参考节点的距离来构建位置信息 [21]。Graphormer 和 GRPE 将相对结构偏置直接融入 Transformer 注意力 [2], [22], ；相关研究还表明，也可以通过重连策略注入位置信息，或通过在图 Transformer 中明确比较绝对与相对位置方案来分析位置信息 [8], [23]。更广泛地说，结构增强型图编码器也会利用拓扑导出的相似性信息，即使其机制并未被明确表述为位置编码 [24]。近期研究也开始明确探讨锚点距离与谱视角之间的联系 [25]。与谱坐标相比，这些方法通常更具可解释性，但其有效性可能高度依赖锚点选择、距离定义和预处理等设计决策，而且高度对称的节点仍可能具有相同的结构特征。

C. 学习式与混合位置/结构编码器

第三类方法学习或组合多种位置信号，而非依赖单一的手工编码 [12], [26]。Graph-Bert 和 GraphiT 是较早将结构与位置信息明确编码到基于注意力的架构中的 Transformer 风格图模型 [12],[13]。GPS 在统一架构中整合了消息传递、Transformer 层以及位置／结构输入 [3],；而 GRIT 强调图归纳偏置，无需依赖显式消息传递 [27]。GPSE 直接从图结构中学习可迁移的位置与结构表示 [4], ，近期研究则进一步探讨了位置／结构编码的基础模型风格泛化 [28],、高效位置编码学习 [29],、图神经网络和图 Transformer 的大规模基准测试 [30], ，以及量子位置编码等替代编码系列 [31]。这类方法的灵活性颇具吸引力，但模型与训练复杂度的增加，也使得更难将编码本身的贡献与骨干架构及优化过程的贡献区分开来。

D. 与我们工作的关系

我们的工作与结合全局和相对结构信息的混合方法最为接近，但研究目标不同。现有大多数研究通过下游准确率、可迁移性或泛化能力间接评估位置编码[28], [29], [30]。相比之下，我们关注的是固定编码能否实现节点定位。从这个意义上说，我们的视角与近期关于位置编码在随机图中作用的理论研究 [10], ，以及将锚点距离编码与扩散几何联系起来、或研究基于符号与基不变拉普拉斯编码的节点可辨识性的近期分析 [25], [32]更为相关。不过，我们特别关注确定性混合观测映射下的节点可辨识性，该映射结合了锚点距离与量化低频谱特征。这一视角也适用于药物-药物相互作用（DDI）和多重用药建模等图学习应用 [33], [34], ；在这些应用中，出色的预测性能并不一定意味着底层位置编码在节点层面提供了足够的结构分辨率。

III. 问题表述

设 $G = (V,\ E)$ 是一个具有 $|V| = n$ 的简单无向图。对于 $u,\ v \in V$，令

$$\mathrm{SPD}(u, v) := \mathrm{dist}_G(u, v)$$

表示最短路径距离。给定一个有序锚点集

$$\mathcal{A} = \{a_1, \ldots, a_k\} \subset V,$$

我们将顶点 $v \in V$ 的锚点距离剖面定义为

$$\mathbf{d}_{\mathcal{A}}(v) := \big(\mathrm{SPD}(v, a_1), \ldots, \mathrm{SPD}(v, a_k)\big) \in \mathbb{Z}^k_{\geq 0}.$$

在整个理论中，我们考虑随机 r-正则图

$$G \sim \mathcal{G}_{n,r},$$

其中 $r \geq 3$ 为固定值，且 nr 为偶数。令$\mathbf{L}$ 表示作用于 G上的拉普拉斯型算子，例如组合拉普拉斯 $L = D - A$ 或归一化拉普拉斯矩阵 $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$。由于 G 是 r-正则的，这些算子仅相差一个仿射缩放，因此在下文的形式体系中，这一区别无关紧要。

Let

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

设 $\mathbf{L}$ 的特征值为这些值，并令 $\Phi_m \in \mathbb{R}^{n\times m}$ 为所考虑的低频 m 维谱子空间的一组标准正交基。由于该基并不唯一，我们采用预先指定的基不变截断谱特征

$$\mathcal{S}_m : V \to \mathbb{R}^{p_m},$$

该特征仅取决于底层谱子空间，而与特征基的具体选择无关。典型示例包括逐坐标能量特征，或由行 $\Phi_m(v,\ \cdot)$ 得出的其他正交不变描述符。

对于量化分辨率 $\eta > 0$，令

$$Q_\eta : \mathbb{R}^{p_m} \to (\eta\mathbb{Z})^{p_m}$$

为一个确定性的逐坐标量化器。例如，可以取

$$(Q_\eta(x))_j = \eta \left\lfloor \frac{x_j}{\eta} \right\rfloor, \quad j = 1, \ldots, p_m.$$

Q_η 的确切形式在此并不重要；它的作用是对有限分辨率进行建模，并导出一种稳健的观测不可区分性概念。

我们将距离和谱分量组合成观测映射

$$F_{G,\mathcal{A}}^{(\eta)} : V \to \mathcal{Y}_{k,m}^{(\eta)}(G,\ \mathcal{A}),\ F_{G,\mathcal{A}}^{(\eta)}(v) = \left(\mathbf{d}_{\mathcal{A}}(v)\quad Q_\eta(\mathcal{S}_m(v))\right),$$

其中

$$\mathcal{Y}_{k,m}^{(\eta)}(G,\mathcal{A}) \coloneqq \mathrm{Im}\big(F_{G,\mathcal{A}}^{(\eta)}\big)$$

是由该编码诱导的可观测编码集合。

现在，设 $v_\star$ 是 V 中均匀随机选取的源顶点，且与 G 相互独立。在给定 G 的条件下，锚点集 $\mathcal{A}$ 从 V 中均匀无放回抽取，且与 $v_\star$ 相互独立。因此，可用数据为

$$F_{G,\mathcal{A}}^{(\eta)}(v_\star) = \big(\mathbf{d}_{\mathcal{A}}(v_\star),\, Q_\eta(\mathcal{S}_m(v_\star))\big).$$

定位问题是根据观测结果（G，$\mathcal{A}$）恢复 $v_\star$。

对于固定的 $(G,\mathcal{A})$，重建过程是任意映射

$$s_{G,\mathcal{A}} : \mathcal{Y}_{k,m}^{(\eta)}(G,\mathcal{A}) \to V.$$

其条件错误概率为

$$\mathrm{Err}(s_{G,\mathcal{A}}) \coloneqq \mathbb{P}\Big(s_{G,\mathcal{A}}\big(F_{G,\mathcal{A}}^{(\eta)}(v_\star)\big) \neq v_\star \,\Big|\, G,\mathcal{A}\Big).$$

我们的目标是刻画 (k,m,η) 的各种机制，分析它们如何随n变化；在这些机制下，从信息论角度看，定位是不可能的，也就是说

$$\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}})$$

在 $(G,\mathcal{A})$ 上以高概率保持远离零。

关键的组合对象是观测映射的原像。对于任意

$$y \in \mathcal{Y}_{k,m}^{(\eta)}(G,\mathcal{A}),$$

定义

$$\mathcal{P}_{G,\mathcal{A}}^{(\eta)}(y) \coloneqq \{v \in V :\ F_{G,\mathcal{A}}^{(\eta)}(v) = y\}.$$

当 $y = F_{G,\mathcal{A}}^{(\eta)}(v)$ 对某个 $v \in V$ 成立时，为简洁起见，我们也记作 $\mathcal{P}_{G,\mathcal{A}}^{(\eta)}(v)$。较大的原像意味着观测歧义：如果多个顶点产生相同的编码，那么任何重构映射都必须将该编码映射到单一输出，因此其成功概率不可能超过原像大小的倒数。观测映射的组合结构，尤其是其像集大小及其原像的几何结构，是理解这一问题的关键。这一观点构成了下一节反向分析的基础。

IV. 理论分析

本节分两层发展相关理论。我们首先将节点定位归约为观测映射的像大小计数问题，从而建立一般性反向定理。随后，我们针对典范能量嵌入给出模型特定细化：通过逐桶碰撞与均衡性原理，进一步收紧一般的量化谱码复杂度界。所有证明汇总于附录 A。

A. 观测映射恒等式与一般性反向定理

我们从一般性的计数归约开始。核心思想是，最优定位误差恰好由观测映射的像集大小决定，因此问题归结为界定混合编码能够诱导出多少个不同的编码。锚点距离分量控制距离桶的数量，而谱分量控制每个桶内可能出现的不同量化特征数量。

固定一个图 $G=(V,\ E)$、一个锚点集 $\mathcal{A} = \{a_1, \ldots, a_k\} \subset V$、一个谱截断级别 $m \geq 1$ 以及一个量化分辨率 $\eta > 0$。回顾量化映射

$$F_{G,\mathcal{A}}^{(\eta)}(v) \coloneqq \Big(\mathbf{d}_{\mathcal{A}}(v),\ Q_\eta(\mathcal{S}_m(v))\Big), \quad v \in V, \tag{1}$$

带像

$$\mathcal{Y}_{k,m}^{(\eta)}(G,\mathcal{A}) \coloneqq \mathrm{Im}(F_{G,\mathcal{A}}^{(\eta)}). \tag{2}$$

其中 $\mathbf{d}_{\mathcal{A}}(v) \in \mathbb{Z}_{\geq 0}^k$ 是锚点距离向量

$$\mathbf{d}_{\mathcal{A}}(v) \coloneqq \big(\mathrm{SPD}(v,a_1), \ldots, \mathrm{SPD}(v,a_k)\big), \tag{3}$$

而 $\mathcal{S}_m :\ V \to \mathbb{R}^{p_m}$ 是一个 $O(m)$-不变的谱嵌入，其中 Q_η 表示分箱宽度为η的逐坐标量化。

对于 $y \in \mathcal{Y}_{k,m}^{(\eta)}(G,\mathcal{A})$，定义 y 在 $F_{G,\mathcal{A}}^{(\eta)}$ 下的原像为

$$\mathcal{P}_{G,\mathcal{A}}^{(\eta)}(y) \coloneqq \big(F_{G,\mathcal{A}}^{(\eta)}\big)^{-1}(\{y\}) = \{v \in V :\ F_{G,\mathcal{A}}^{(\eta)}(v) = y\}. \tag{4}$$

当 $y = F^{(\eta)}_{G,\mathcal{A}}(v)$ 时，我们用 $\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v)$ 表示 $\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)$。

下文中，源顶点 $v_\star$ 在 V 上均匀分布，且与 $(G,\mathcal{A})$ 独立。重构映射是任意可测函数。可测截面是一种满足以下条件的特殊重构映射

$$F^{(\eta)}_{G,\mathcal{A}}(s_{G,\mathcal{A}}(y)) = y \tag{5}$$

对每个 $y \in \mathcal{Y}^{(\eta)}_{k\ m}(G,\mathcal{A})$,

$$s_{G,\mathcal{A}}:\ \mathcal{Y}^{(\eta)}_{k,m}(G,\mathcal{A}) \to V. \tag{6}$$

给定 $s_{G,\mathcal{A}}$，其条件错误概率为

$$\mathrm{Err}(s_{G,\mathcal{A}}) \coloneqq \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \neq v_\star \ \Big|\ G,\mathcal{A}\Big). \tag{7}$$

引理 1. 对于任意固定的 $(G,\mathcal{A})$，令

$$Y \coloneqq F^{(\eta)}_{G,\mathcal{A}}(v_\star). \tag{8}$$

那么对于任意可测映射

$$s_{G,\mathcal{A}} : \mathcal{Y}^{(\eta)}_{k,m}(G,\mathcal{A}) \to V, \tag{9}$$

有

$$\begin{aligned}\mathbb{P}\Big(s_{G,\mathcal{A}}(Y) = v_\star \ \Big|\ G,\mathcal{A}\Big) &= \mathbb{E}\left[\frac{\mathbf{1}\Big\{s_{G,\mathcal{A}}(Y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)\Big\}}{\big|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)\big|} \ \Big|\ G,\mathcal{A}\right] \\ &\le \mathbb{E}\left[\frac{1}{\big|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)\big|} \ \Big|\ G,\mathcal{A}\right].\end{aligned} \tag{10}$$

此外，

$$\mathbb{E}\left[\frac{1}{\big|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)\big|} \ \Big|\ G,\mathcal{A}\right] = \frac{\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|}{n}. \tag{11}$$

任何满足以下条件的可测截面都能达到该上界

$$\tilde{s}_{G,\mathcal{A}} : \mathcal{Y}^{(\eta)}_{k,m}(G,\mathcal{A}) \to V \tag{12}$$

满足

$$F^{(\eta)}_{G,\mathcal{A}}\big(\tilde{s}_{G,\mathcal{A}}(y)\big) = y, \quad \forall\, y \in \mathcal{Y}^{(\eta)}_{k,m}(G,\mathcal{A}). \tag{13}$$

因此，

$$\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) = 1 - \frac{\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|}{n}. \tag{14}$$

对于$\mathbf{t} \in \mathbb{Z}^k_{\ge 0}$，定义距离桶

$$B_{\mathbf{t}}(G,\mathcal{A}) \coloneqq \{v \in V :\ \mathbf{d}_{\mathcal{A}}(v) = \mathbf{t}\}. \tag{15}$$

引理 2. 假设$\mathrm{diam}(G) \le C_{\mathrm{diam}}$对数$n$。则

$$\big|\{\mathbf{d}_{\mathcal{A}}(v) : v \in V\}\big| \le (C\log n)^k, \tag{16}$$

尤其是，非空桶的数量 $B_{\mathbf{t}}$（$G,\mathcal{A}$）至多为 （C 对数 n）^{k}。

假设 1. 令 $\mathcal{S}_m(V) \coloneqq \{\mathcal{S}_m(v) :\ v \in V\} \subset \mathbb{R}^{p_m}$. 固定$m \ge 1$ 和 $\eta \in (0,1)$。存在常数 $C_{\mathrm{ent}} \ge 1, c_{\mathrm{ent}} > 0$ 以及一个 $\mathcal{E}_{\mathrm{ent}}$ 事件（取决于 G），满足 $\mathbb{P}(\mathcal{E}_{\mathrm{ent}}) = 1 - o(1)$，使得在 $\mathcal{E}_{\mathrm{ent}}$ 上，

$$\big|Q_\eta(\mathcal{S}_m(V))\big| \le \Big(\frac{C_{\mathrm{ent}}}{\eta}\Big)^{c_{\mathrm{ent}}m}. \tag{17}$$

假设 1 并非旨在作为对谱分量最终且精确的描述。相反，它是一个通用的复杂度假设，在该假设下，逆命题可采用简洁的信息预算形式。随后，命题4和推论1将对典范能量嵌入进行明确处理。

引理 3。 假设 $\mathrm{diam}(G) \le C_{\mathrm{diam}}$ 对数 n，且假设 *1* 在 $\mathcal{E}_{\mathrm{ent}}$ 上成立。那么，在事件 $\{\mathrm{diam}(G) \le C_{\mathrm{diam}}$ 对数 $n\} \cap \mathcal{E}_{\mathrm{ent}}$发生时，

$$\begin{aligned}\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| &\le (C\log n)^k \cdot \big|Q_\eta(\mathcal{S}_m(V))\big| \\ &\le (C\log n)^k \cdot \Big(\frac{C_{\mathrm{ent}}}{\eta}\Big)^{c_{\mathrm{ent}}m},\end{aligned} \tag{18}$$

where $C = C(r) > 0$ depends only on C_{diam}.

令 Φ_m 为 $\mathbf{L}$ 对应前 m 个特征值的标准正交特征基矩阵，并用 $\Phi_m(v,\cdot)$ 表示第 v 行。考虑 $O(m)$ 不变的能量型谱嵌入

$$\mathcal{S}_m(v) \coloneqq \big(\Phi_m(v,1)^2, \ldots, \Phi_m(v,m)^2\big) \in \mathbb{R}^m. \tag{19}$$

命题4。 对于嵌入 $\mathcal{S}_m$（v）$=$（Φ_m（v,j）^{2}）$_{j\in[m]}$以及 $\eta \in (0,\ 1)$，有

$$\big|Q_\eta(\mathcal{S}_m(V))\big| \le \big|Q_\eta([0,1]^m)\big| \le \Big(\frac{2}{\eta}\Big)^m. \tag{20}$$

特别地，当 $C_{\mathrm{ent}} = 2$、 $c_{\mathrm{ent}} = 1$ 和 $\mathcal{E}_{\mathrm{ent}}$ 等于整个概率空间时，假设 *1* 成立。

现在可以陈述一般性反向定理。

定理 5。 固定 $r \ge 3$，并令$G \sim \mathcal{G}_{n,r}$。令 $v_\star \sim \mathrm{Unif}(V)$，并令 $\mathcal{A} = \{a_1,\ldots,a_k\}$ 从 V 中均匀无放回抽样，且与 $v_\star$ 和 G 独立。固定$m \ge 1$ 和 $\eta \in (0,1)$。假设直径界$\mathrm{diam}(G) \le C_{\mathrm{diam}}\log n$ 以概率 $1 - o(1)$ 成立，并且谱嵌入 $\mathcal{S}_m$ 在分辨率 η 下满足假设 *1*。若存在 $\varepsilon_0 \in (0,1)$ 使得

$$k\log\log n + c_{\mathrm{ent}}m\log\Big(\frac{C_{\mathrm{ent}}}{\eta}\Big) \le (1-\varepsilon_0)\log n, \tag{21}$$

则存在 $\varepsilon_1 \in (0,\varepsilon_0)$，使得对 $(G,\mathcal{A})$，以概率 $1-o(1)$有，

$$\inf_{s_{G,\mathcal{A}}} \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \neq v_\star \ \Big|\ G,\mathcal{A}\Big) \ge 1 - n^{-\varepsilon_1}. \tag{22}$$

特别地，对于所有充分大的 n，左侧有任意固定常数 $\delta \in (0,1)$ 作为下界。

定理 5 确定了一个一般性的亚临界不可能区域：当

$$k\log\log n + c_{\mathrm{ent}}m\log\Big(\frac{C_{\mathrm{ent}}}{\eta}\Big) \tag{23}$$

严格低于 $\log n$ 时，编码无法产生足够多的不同观测，因而无法以

趋零误差定位均匀随机节点。这是一个单侧逆命题，而非尖锐阈值定理。

假设 1 应被理解为关于量化谱编码的一般复杂度假设。其作用是以抽象形式隔离出支配逆向结论的信息预算。然而，对于典范能量嵌入，该假设可以得到明确验证：命题 4给出了具体常数 $C_{\mathrm{ent}} = 2$ 和 $c_{\mathrm{ent}} = 1$。因此，定理 5 专门化为以下无需假设的推论。

推论1。固定 $r \geq 3$，并令$G \sim \mathcal{G}_{n,\ r}$。令 $v_\star \sim$ 均匀分布于 (V)，并令 $\mathcal{A} = \{a_1, \ldots, a_k\}$ 从 V 中均匀无放回地抽取，且独立于 $v_\star$ 和 G。固定 $m \geq 1$和 $\eta \in (0,\ 1)$，并取

$$\mathcal{S}_m(v) = \big(\Phi_m(v,1)^2, \ldots, \Phi_m(v,m)^2\big). \tag{24}$$

假设以概率 $1 - o(1)$ 有 $\mathrm{diam}(G) \leq C_{\mathrm{diam}}$对数 n。如果存在 $\varepsilon_0 \in (0,\ 1)$ 使得

$$k \log\log n + m \log\Big(\frac{2}{\eta}\Big) \leq (1 - \varepsilon_0) \log n, \tag{25}$$

那么存在 $\varepsilon_1 \in (0,\ \varepsilon_0)$，使得在 $(G,\ \mathcal{A})$ 上以概率 $1-o(1)$，

$$\inf_{s_{G,\mathcal{A}}} \ \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \neq v_\star \ \Big|\ G, \mathcal{A}\Big) \geq 1 - n^{-\varepsilon_1}. \tag{26}$$

推论1应被视为针对典范能量型谱嵌入的一般性反向定理的具体实例。

B. 通过分桶碰撞与均衡性进行模型特定细化

现在，我们针对典范能量型谱嵌入细化一般计数论证。我们不再通过全局复杂度界来控制每个距离桶中的谱代码数量，而是利用一个结构性更强的原理：如果量化谱特征在每个非单元素距离桶内充分频繁地发生碰撞，且已占用的量化编码分布没有过度失衡，那么该桶内不同观测的数量就必然较少。这给出了一种模型特定细化，并与实验中测量的碰撞和占用统计直接相关。

对于任意有限的 $B \subset V$，定义在 B 中取得的量化谱编码集合为

$$Q_\eta(\mathcal{S}_m(B)) \coloneqq \{Q_\eta(\mathcal{S}_m(v)) :\ v \in B\}. \tag{27}$$

Let

$$M_{\eta,m}(B) \coloneqq \big|Q_\eta(\mathcal{S}_m(B))\big| \tag{28}$$

表示在 B中取得的不同量化谱编码的数量。对于每个

$$z \in Q_\eta(\mathcal{S}_m(B)), \tag{29}$$

将其占用数定义为

$$N_{\eta,m}(B; z) \coloneqq \big|\{v \in B :\ Q_\eta(\mathcal{S}_m(v)) = z\}\big|. \tag{30}$$

对于任意有限的 $B \subset V$，若满足 $|B| \geq 2$，则将尺度 η 下的量化碰撞密度定义为

$$\mathrm{Coll}^{Q}_{\eta,m}(B) \coloneqq \tfrac{1}{|B|(|B|-1)} \sum_{\substack{u,v \in B \\ u \neq v}} \mathbf{1}\Big\{Q_\eta(\mathcal{S}_m(u)) = Q_\eta(\mathcal{S}_m(v))\Big\}. \tag{31}$$

当 $|B| = 1$时，无需使用该量，因为该桶恰好贡献一个观测。

我们还将桶平衡系数定义为

$$\mathrm{Bal}_{\eta,m}(B) \coloneqq \frac{M_{\eta,m}(B)}{|B|} \cdot \max_{z \in Q_\eta(\mathcal{S}_m(B))} N_{\eta,m}(B; z). \tag{32}$$

根据构造，$\mathrm{Bal}_{\eta,\ m}(B) \geq 1$。条件 $\mathrm{Bal}_{\eta,\ m}(B) \leq \beta$ 表示：在 B中，没有任何占用的量化编码所携带的质量超过占用编码平均质量的 β 倍。

对于由 $\mathcal{A}$诱导的距离划分，回顾

$$B_{\mathbf{t}}(G, \mathcal{A}) \coloneqq \{v \in V :\ \mathbf{d}_{\mathcal{A}}(v) = \mathbf{t}\}, \tag{33}$$

并令

$$D(G, \mathcal{A}) \coloneqq \big|\{\mathbf{d}_{\mathcal{A}}(v) : v \in V\}\big| \tag{34}$$

表示可取距离桶的数量。

命题 6。假设$\mathrm{diam}(G) \leq C_{\mathrm{diam}}$ 对数 n。设对于每个可取到的距离向量$\mathbf{t}$，以下条件之一成立：

- $|B_{\mathbf{t}}(G, \mathcal{A})| = 1$; *or*
- $|B_{\mathbf{t}}(G, \mathcal{A})| \geq 2$, *and*

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \geq c_0\, \eta^{c_1 m}, \tag{35}$$

连同

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \leq \beta, \tag{36}$$

对于某些常数 c_0、$c_1 > 0$ 和 $\beta \geq 1$。则

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \leq D(G, \mathcal{A}) \cdot \Big(1 + \frac{\beta}{c_0\, \eta^{c_1 m}}\Big). \tag{37}$$

特别是，

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \leq (C \log n)^k \cdot \Big(1 + \frac{\beta}{c_0\, \eta^{c_1 m}}\Big). \tag{38}$$

命题 6 应被理解为一种第二层计数原理。单元素桶仅凭距离分量便已得到处理，而每个非单元素桶则进一步由两个可测量的量来约束：桶内量化碰撞的下界，以及已占用的量化编码不平衡程度的上界。

推论 2. 假设 $\mathrm{diam}(G) \leq C_{\mathrm{diam}}$ 对数 n 的概率为$1 - o(1)$，并且假设存在一个事件 $\mathcal{E}_{\mathrm{ref}}$，其满足 $\mathbb{P}(\mathcal{E}_{\mathrm{ref}}) = 1 - o(1)$，使得对于 $\mathcal{E}_{\mathrm{ref}}$ 上的每个可取到的距离向量$\mathbf{t}$，以下条件之一成立：

- $|B_{\mathbf{t}}(G, \mathcal{A})| = 1$; *or*
- $|B_{\mathbf{t}}(G, \mathcal{A})| \geq 2$, *and*

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \geq c_0\, \eta^{c_1 m}, \tag{39}$$

以及

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \leq \beta, \tag{40}$$

对于某些常数 c_0、$c_1>0$ 和 $\beta\ge 1$。那么，以概率 $1-o(1)$ 在 $(G、\mathcal{A})$ 上，

$$\left|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\right| \le (C\log n)^k \cdot \Big(1+\frac{\beta}{c_0\,\eta^{c_1 m}}\Big). \tag{41}$$

特别地，由于 $\eta\in(0,1)$，

$$\left|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\right| \le (C\log n)^k \cdot \Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m}. \tag{42}$$

将精细的像集大小界与引理 1 结合，可得以下精细的不可能性推论。

推论 3. *在定理 5 的设定下，进一步假设存在常数 c_0、$c_1>0$ 和 $\beta\ge 1$，以及一个事件 $\mathcal{E}_{\mathrm{ref}}$，满足 $\mathbb{P}(\mathcal{E}_{\mathrm{ref}})=1-o(1)$，且在 $\mathcal{E}_{\mathrm{ref}}$ 上，对于每个可取到的距离向量$\mathbf{t}$，以下条件之一成立：*

- $|B_{\mathbf{t}}(G,\mathcal{A})|=1$*; or*
- $|B_{\mathbf{t}}(G,\mathcal{A})|\ge 2$*, and*

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}(G,\mathcal{A})) \ge c_0\,\eta^{c_1 m}, \tag{43}$$

以及

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}(G,\mathcal{A})) \le \beta. \tag{44}$$

如果存在 $\varepsilon_0\in(0,1)$ 使得

$$k\log\log n + c_1 m\log(1/\eta) \le (1-\varepsilon_0)\log n, \tag{45}$$

那么存在 $\varepsilon_1\in(0,\varepsilon_0)$ 使得在 $(G,\mathcal{A})$ 上，概率为 $1-o(1)$，

$$\inf_{s_{G,\mathcal{A}}} \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \neq v_\star \;\Big|\; G,\mathcal{A}\Big) \ge 1-n^{-\varepsilon_1}. \tag{46}$$

推论 3 是定理 5 的模型特定细化。一旦分桶量化碰撞足够强，且已占用的量化编码在每个非单点距离桶内保持适度均衡，就可以用更精确的分桶计数原理取代通用的量化谱码复杂度界。

V. 实证评估

A. 实验设置

我们在合成图和真实世界图上评估所提出的编码，重点关注编码本身的可识别性，而非端到端预测。研究对象是确定性观测映射

$$F(v)=\big(d_A(v),Q_\eta(S_m(v))\big), \tag{47}$$

其中，$d_A(v)$ 是锚点距离轮廓，$Q_\eta(S_m(v))$ 是相对量化的低频谱编码。除非另有说明，锚点均采用均匀无放回采样；谱分量使用归一化拉普拉斯矩阵的前 m 个非平凡特征向量，且量化方式为相对量化。

对于合成研究，我们使用随机 3-正则图，

$$\begin{aligned} n &\in \{500,1000,2000,4000\},\\ k &\in \{1,2,3,4,6,8\},\\ m &\in \{0,1,2,5\},\\ \eta &\in \{0.9,0.7,0.5,0.3,0.1\}, \end{aligned}$$

对于每种配置，取 20 个独立采样的图的平均值；其中包含 $m=0$，作为仅距离基线。我们按以下方式组织合成结果：

$$\rho_{\mathrm{eng}}=\frac{k\log\log n + m\log(2/\eta)}{\log n},$$

并报告经验锚点阈值

$$k_{\mathrm{emp}}(n,m,\eta)=\min\{k:\mathbb{E}[\mathrm{err}(F)]\le 0.1\}.$$

在真实世界研究中，我们考虑两种药物相互作用图构造：一种是 DrugBank 图，其

$$(N_{\mathrm{DB}},E_{\mathrm{DB}})=(1684,189774),$$

另一种是 Decagon 派生图，其

$$(N_{\mathrm{DEC}},E_{\mathrm{DEC}})=(40,466).$$

在这些固定图上，我们评估

$$k\in\{1,2,3,4,6,8\},\quad m\in\{0,2,5,10\},\quad \eta\in\{0.9,0.5,0.25,0.1\},$$

并对 30 次独立的随机锚点采样结果取平均。

我们的主要指标是归一化像集大小

$$\begin{aligned} \mathrm{succ}(F)&=\frac{|\mathrm{Im}(F)|}{n},\\ \mathrm{err}(F)&=1-\mathrm{succ}(F). \end{aligned} \tag{48}$$

因此，误差越小，表明节点可辨识性越强。完整的数据集构建、基线定义、量化规则和详细的诊断流程见附录 H。

B. 随机正则图上的类相变行为

图 1 最清楚地展示了随机正则图设置下的类相变行为。在面板 (a) 中，锚点预算进入转变区域后，平均定位误差会急剧下降。

面板 (b) 展示了相应的结构效应：以理论指导的预算比 ρ_{eng} 为横轴作图时，平均平均原像大小会迅速收缩至接近 1。这表明，经验误差的下降伴随着观测映射纤维的实际缩小。

面板 (c) 右侧的区块进一步表明，对于不同的谱维数 $m\in\{0,1,2,5\}$，以 ρ_{eng} 为横轴绘制误差图，相比单独使用锚点数，能更一致地呈现不同图规模下的转变。主要交叉点仍集中在 $\rho_{\mathrm{eng}}\approx 1$ 附近，尽管转变的陡峭程度会随 m 而变化。对此应将其解读为与逆命题边界一致的经验标度律，而非锐利阈值的证据。

综合来看，这些图表表明，相较于单独考察原始锚点数量，依据理论的信息预算更能自然地解释这一过渡；误差下降也恰好伴随着观测歧义的组合式坍缩。

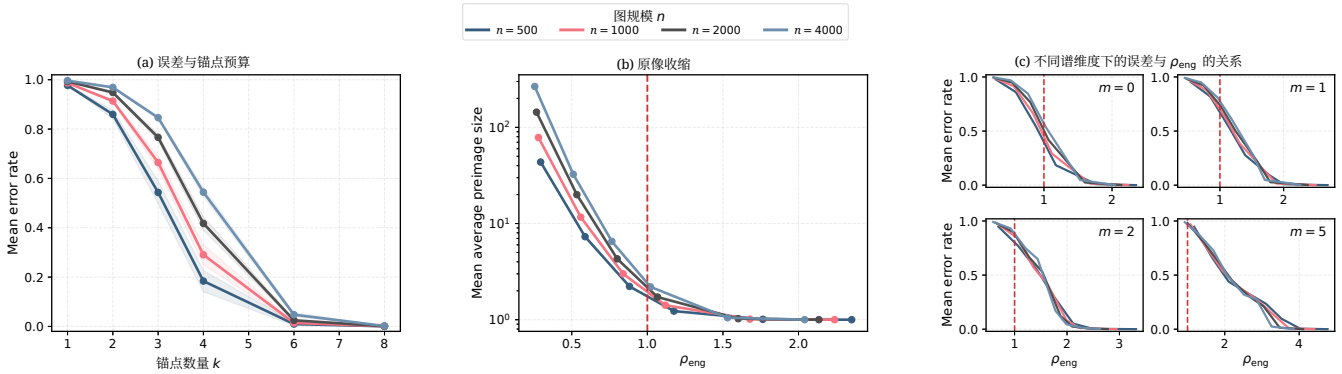


图 1：随机 3-正则图上的类相变行为。(a) 平均定位误差与锚点数量 k 的关系。(b) 平均平均原像大小与预算比 ρ_{eng}的关系，显示了低误差区域中观测映射纤维的坍缩。(c) 不同谱维数 $m \in \{0, 1, 2, 5\}$ 下，误差曲线随 ρ_{eng} 变化的情况。在不同谱维数下，相较于单独使用锚点数，理论指导的预算比能够更一致地刻画这一过渡。

表 I：最精细量化级别$\eta = 0.1$下的阈值诊断。

n	m	k_{emp}	ρ_{emp}	$\lvert$像$(F)\rvert/n$	平均原像	$\lvert Q_\eta(S_m(V))\rvert$
500	0	6	1.764	0.992	1.008	1
	1	3	1.364	0.906	1.105	51.75
	2	2	1.552	0.956	1.046	285.6
	5	1	2.704	1.000	1.000	499.45
1000	0	6	1.679	0.987	1.014	1
	1	4	1.553	0.963	1.038	66
	2	2	1.427	0.915	1.093	444.2
	5	1	2.448	0.999	1.001	997.8
2000	0	6	1.601	0.973	1.027	1
	1	4	1.462	0.937	1.067	76.3
	2	3	1.589	0.967	1.034	692.75
	5	1	2.237	0.999	1.001	1990.1
4000	0	6	1.530	0.948	1.055	1
	1	6	1.892	0.995	1.005	87.65
	2	3	1.488	0.940	1.064	1033.6
	5	1	2.061	0.998	1.002	3962

注：谱码本越大，锚点阈值越小。对于 $m = 5$，码本大小已接近 n，且 $k_{\mathrm{emp}} = 1$ 适用于所有测试过的图规模。

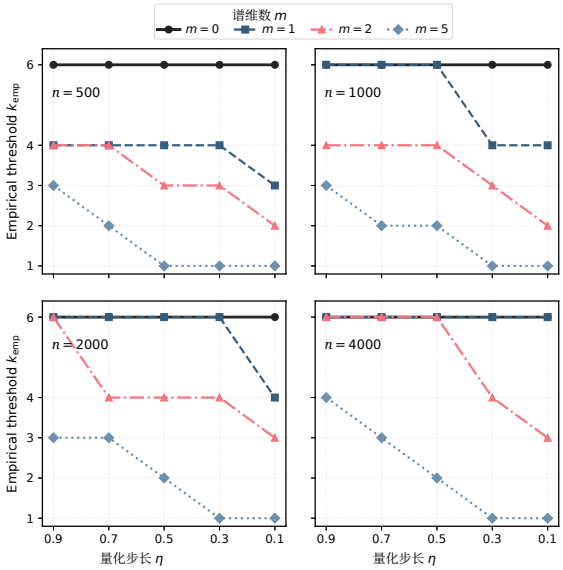


图 2：随机 3-正则图的经验阈值 k_{emp} 与量化步长 η 的关系。每个子图固定 n，每条曲线对应一个谱维数 m。

C. 与理论一致的观测映射诊断

图 2 展示了 k_{emp} 随 n、m 和 η 的变化情况。

最显著的模式是仅距离基线的稳定性：对于所有测试的 n 和 η，当 $m = 0$ 时，$k_{\mathrm{emp}} = 6$。加入谱通道后，阈值降低，而且量化越精细，收益越明显。在 $\eta = 0.1$ 时，对于 $m = 1$，阈值从 6 降至 (3，4，4，6)；对于 $m = 2$，降至 (2，2，3，3)；当 $m = 5$ 时，对于所有测试的图规模，阈值均降至 1。曲线在 η 上并非严格单调，这符合预期，因为 k_{emp} 是一个通过有限次试验估计出的整数阈值。

表 I 展示了在 η =0.1 时相应的阈值统计。达到阈值时，归一化像集大小已经很高（0.9057–0.9999），平均原像大小则接近 1。谱码数量随 m 急剧增加，在 $m = 5$ 时几乎达到 n；这正是锚点阈值降至 1 的机制。

D. 分桶细化诊断

为考察精细化分桶计数界背后的机制，我们分析了带有$n = 2000$的随机 3-正则图中非单例距离桶内的谱细化，并采用三种具有代表性的机制：碰撞密集型（k，m，η）=（2，1，2。0）、过渡（2，2，1。0）和近似单射（2，5，0。3）。这三种机制是辅助诊断设置，旨在涵盖碰撞密集型、过渡和近似单射谱，并不局限于相变实验所用的主网格。每种机制下，我们使用 20 个图实例，并对每个图进行 10 次锚点重采样。由于三种机制使用相同的图实例和相同的锚点重采样，因此各行的距离桶完全相同，而这种比较能够隔离谱

分量在残余桶内的影响。

表 II 显示出清晰的递进关系。在高碰撞区域，桶内成对加权碰撞较大（0.727），而桶内归一化代码数中位数较小（0.185），表明非单元素桶内部存在显著压缩。在近单射区域，加权碰撞可以忽略不计（4.45×10^{-4}），桶内归一化代码数中位数基本为一，说明谱分量几乎完全解析了剩余的桶。过渡区域处于这两个极端之间。

平衡统计量进一步清晰地揭示了这一图景。过渡区域的上尾平衡系数最大，这表明碰撞本身并不能决定桶内像集大小：占用数越不均匀，对应的已占用谱代码数量越多。这与精化命题一致，其中碰撞和平衡性都会影响分桶计数界。只关注较大的非单例桶时，排序关系依然稳定；当截断值从$|B_{\mathfrak{t}}| \geq 3$ 提高到 $|B_{\mathfrak{t}}| \geq 10$时，桶内归一化代码数的差异则更加显著。

E. 两个真实世界 *DDI* 图上的可识别性

本实验旨在作为域内案例研究，而非宣称结果普遍适用于各类图族。我们比较了设置部分介绍的两个真实世界 DDI 图上的可识别性。

这两个图的可识别性截然不同。在 DrugBank 上，对于所有测试设置，观测映射仍存在大量碰撞。即使在 $(k,\ m,\ \eta)$ =(8，10，0.1) 时，我们仍有

$$\mathrm{err}(F) = 0.9023, \quad \frac{|\mathrm{Im}(F)|}{n} = 0.0977,$$

顶点平均原像大小为 79.33，单例原像顶点比例为 0.0384。

对于此处考察的小型 Decagon 派生图，仅靠距离信息并不足够，但加入谱信息后，在测试设置下，该编码几乎是单射。在 (8，0，0.1) 时，误差为 0.4250，归一化像集大小为 0.5750。在 (8，10，0.1) 时，这些数值变为

$$\mathrm{err}(F) = 0.0158, \quad \frac{|\mathrm{Im}(F)|}{n} = 0.9842,$$

顶点平均原像大小为 1.03，单例原像顶点比例为 0.9683。

总体趋势与这一对比一致。在 Decagon 派生图上，随着 m 从 0 增至 10，平均误差从 0.704 降至 0.192；随着 η 从 0.9 降至 0.1，平均误差从 0.579 降至 0.293；随着 k 从 1 增至 8，平均误差则从 0.585 降至 0.264。在 DrugBank 上，相应变化要小得多。

表 IV 展示了桶级别上相同的模式。在 DrugBank 上，单例距离桶很少见，而非单例桶仍存在大量碰撞。在基于 Decagon 构建的图上，近单射区域与更多单例距离桶及弱得多的桶内碰撞相关。在这两个药物相互作用图中，可识别性不仅取决于 $(k,\ m,\ \eta)$，还取决于底层图结构所诱导的观测映射几何。

F. 消融研究

我们在随机 3-正则图上进行消融实验，以评估距离分量和谱分量的作用、(k, m, η) 的交互作用，以及对锚点选择的敏感性。除非另有说明，消融实验使用

$$\begin{aligned} n &\in \{500, 1000, 2000\}, \\ k &\in \{1, 2, 3, 4, 6, 8\}, \\ m &\in \{0, 2, 5, 10\}, \\ \eta &\in \{0.9, 0.5, 0.25, 0.1\}, \end{aligned}$$

在 20 次独立图试验中取平均。

a) 表示与锚点。：表 V 表明，位置信息不可或缺。缺少位置信息时，误差接近 1。纯谱特征的效果仍然较弱，平均误差为 0.9328。纯距离特征提供的信息丰富得多，将平均误差降至 0.2912。完整观测映射进一步将平均误差降至 0.0789，平均原像大小为 1.0908，单例观测比率为 0.9310。在随机正则图上，锚点距离提供了主要的判别信号，而量化谱项主要用于细化同一距离桶内的节点。

更复杂的锚点策略相较于均匀随机采样只能带来有限的提升。随机锚点的平均误差为 0.0941，最远点启发式的平均误差为 0.0832，基于度的启发式的平均误差为 0.0755。随机锚点在重采样下也很稳定：对于 $n = 1000$，图内误差的平均标准差为 0.0150；对于 $n = 2000$，则为 0.0199，观测到的最大值为 0.0290。

b) 超参数交互：图 3 显示，k、m 和 η 的影响彼此耦合。固定 $k = 4$ 和 $n = 2000$ 时，平均误差从 $m = 0$ 处的 0.4156 降至 (m, η) =(2, 0.1) 处的 0.1308、(5, 0.1) 处的 0.0393，以及 (10, 0.1) 处的 9.5×10^{-4}。固定 $m = 5$ 和 $n = 2000$ 时，平均误差从 (k, η) =(1, 0.1) 处的 0.5501 降至 (4, 0.1) 处的 0.0314、(6, 0.1) 处的 5.75×10^{-4}，以及 (8, 0.1) 处的 7.5×10^{-5}。在固定 k 时，较小的 η 可进一步提升效果。这些结果与理论图景相符：锚点距离决定粗粒度划分，而谱信息则提升距离桶内的分辨率。

关于量化校准的更多结果及其他诊断细节见附录 H。

VI. 讨论与结论

A. 讨论

本文研究了在混合图位置编码下的节点定位，该编码结合了锚点距离剖面与量化低频谱特征。本文的核心理论结论是，定位能力受信息预算支配，而信息预算由上下文大小 k、谱维数 m 和量化分辨率 η 共同决定。我们的逆定理通过观测映射的像集大小对此进行了形式化

表 II：在 $n = 2000$条件下，随机 3-正则图上的逐桶细化诊断。数值为 20 次图试验和每次试验 10 次锚点重采样的平均值。最后四列报告了在阈值 $|B_t| \geq 3$ 和$|B_t| \geq 10$下的大桶诊断。

机制	总体误差	单例比例	加权碰撞	中位数 $M(B_t)/\|B_t\|$	$q_{0.9}$ 均衡性	大规模碰撞 (≥ 3)	大规模碰撞 (≥ 10)	大规模中位数 $M/\|B\|$ ($\geq 3, \geq 10$)
碰撞密集型	0.894	0.202	**0.727**	**0.185**	3.23	0.727	0.727	0.154, 0.093
过渡	0.703	0.202	0.253	0.465	**6.97**	0.253	0.253	0.413, 0.311
近似单射	**0.012**	0.202	4.45×10^{-4}	0.9999	1.89	4.45×10^{-4}	4.46×10^{-4}	0.9999, 0.9999

表 III：两个真实世界 DDI 图的结构统计数据。

统计量	DrugBank DDI	Decagon 派生的药物相互作用
n	1684	40
$\|E\|$	189774	466
平均度	225.385	23.300
密度	0.1339	0.5974
直径	5	3
平均最短路径长度	2.086	1.414
平均聚类系数	0.550	0.874
传递性	0.496	0.801
度方差	38041.436	106.610
度数基尼系数	0.478	0.246

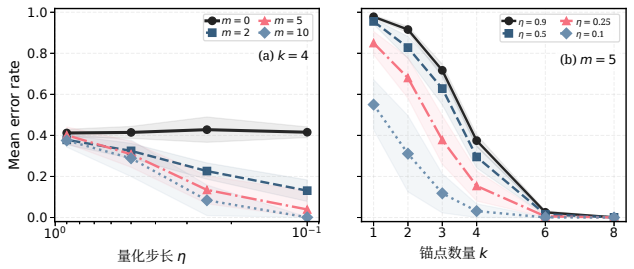


图3：在 $n = 2000$ 下随机 3-正则图上的超参数交互。上：固定 $k = 4$ 时，平均误差随 η 的变化。下：固定 $m = 5$ 时，平均误差随 k 的变化。

映射 F，并表明，当诱导码空间相对于 n 过小时，该编码在信息论意义上不足以在误差趋零的情况下定位一个均匀随机节点。

实验从两个互补的角度支持了这一图景。在随机正则图上，经验观察到的类似相变的交叉点可由理论指导比率 $\rho_{\text{eng}}(n,\ k,\ m,\ \eta)$很好地刻画，这与逆定理所揭示的缩放规律一致。对于本文考察的两种药物相互作用图，可识别性显然取决于图：在所有测试设置下，DrugBank 中仍存在大量碰撞；而较小的 Decagon 派生图一旦纳入足够丰富且量化足够精细的谱信息，便几乎变为单射。总体而言，所提出的编码可视为一种依赖于图的结构分辨机制，其有效性同时取决于（$k,\ m,\ \eta$）以及底层图的冗余度。

B. 局限性

需要指出若干局限性。首先，当前理论主要是单向的：它在亚临界区间证明了不可能性，但尚未提供相应的可达性保证或精确的转变边界。因此，经验观察到的交叉点应被理解为与逆命题边界相一致的类相变标度律，而非精确阈值的证明。

第二，当前分析针对随机正则图以及一种特定的正交不变谱摘要展开。因此，目前的研究范围尚未涵盖更广泛的图模型、更加异质的结构状态或学习得到的位置表示。

第三，真实世界验证的范围有意保持狭窄。我们研究了两种具有代表性的药物相互作用图构造，以支持域内图依赖性的论断，但并未声称这一结论适用于差异显著的各类图族。

第四，正文实验采用经过校准的相对量化规则，而核心理论则以固定的绝对分辨率量化器为基础。尽管附录表明，在经过适当校准的绝对量化下，定性趋势依然成立，但理论与实验之间仍有必要实现更紧密的匹配。

最后，实验关注的是编码本身的可识别性，而非完整的端到端下游预测器。因此，应将当前结果解读为位置编码的结构性限制与诊断，而不是对学习型图模型中预测性能的完整描述。

C. 未来方向

一个自然的后续步骤是推导相匹配的反向界与可达性结果，从而更精确地刻画过渡，并使其适用于较弱假设。此外，将这一框架从随机正则图拓展到更广泛的图族也很有价值，包括加权图、异质图和动态图。

另一个重要方向是加强理论中使用的抽象绝对分辨率量化器与实践中采用的校准量化规则之间的匹配。在方法层面，以自适应或可学习的变体取代当前的手工编码也值得探索，例如可学习锚点选择、多尺度谱摘要和任务感知量化规则。

在实证方面，一项重要的拓展是将可识别性与下游图学习性能更直接地联系起来，使当前分析不仅可作为结构性判据，也能为图神经网络和图 Transformer 中位置编码的设计提供实用原则。

表 IV：两个真实世界 DDI 图上的代表性可辨识性结果。这里 $|\mathrm{Im}(F)|/n$ 表示归一化像集大小，其余列报告原像和基于距离分桶的诊断指标。

图	k	m	η	误差	$\lvert\mathrm{Im}(F)\rvert/n$	顶点平均原像	单元素原像占比	单元素距离桶占比	非单元素碰撞中位数	非单元素平衡中位数
DrugBank DDI	1	0	0.5	0.9971	0.0029	959.65	0.0007	0.0007	1.0000	1.0000
DrugBank DDI	1	10	0.1	0.9918	0.0082	949.07	0.0037	0.0007	0.9637	3.4002
DrugBank DDI	4	0	0.1	0.9773	0.0227	264.77	0.0057	0.0057	1.0000	1.0000
DrugBank DDI	4	10	0.1	0.9711	0.0289	288.81	0.0104	0.0064	1.0000	1.0000
DrugBank DDI	8	0	0.1	0.9093	0.0907	92.46	0.0340	0.0340	1.0000	1.0000
DrugBank DDI	8	10	0.1	0.9023	0.0977	79.33	0.0384	0.0346	1.0000	1.0000
Decagon 派生的药物相互作用	1	0	0.5	0.9175	0.0825	24.14	0.0283	0.0283	1.0000	1.0000
Decagon 派生的药物相互作用	1	10	0.1	0.0233	0.9767	1.05	0.9533	0.0275	0.0015	1.3520
Decagon 派生的药物相互作用	4	0	0.1	0.6958	0.3042	9.47	0.1792	0.1792	1.0000	1.0000
Decagon 派生的药物相互作用	4	10	0.1	0.0200	0.9800	1.04	0.9600	0.1792	0.0000	1.0000
Decagon 派生的药物相互作用	8	0	0.1	0.4250	0.5750	3.83	0.4367	0.4367	1.0000	1.0000
Decagon 派生的药物相互作用	8	10	0.1	0.0158	0.9842	1.03	0.9683	0.4458	0.0000	1.0000

表 V：随机 3-正则图上的表示消融与锚点策略。

设置	描述	错误前像	单一		碰撞
表示消融					
NoPE	无位置编码	0.999	1166.7	0.000	1.000
谱	仅谱	0.933	25.00	0.261	0.198
距离	锚点距离仅	0.291	1.439	0.743	0.0009
Full	距离 +谱	**0.079**	**1.091**	**0.931**	**0.0002**
锚点策略（完整特征）					
随机	均匀锚点	0.094	1.112	0.920	0.0002
最远	最远点启发式	0.083	1.097	0.926	0.0002
度	基于度的启发式	**0.076**	**1.087**	**0.934**	**0.0002**

声明

a) 数据和材料的可用性：本研究开发的 MetaMolGen 模型源代码已在 GitHub 上公开，地址为 https://github.com/yzz980314/Graph_节点_可识别性_通过_观测_映射。本研究使用的数据集均已公开。DrugBank 数据库可通过以下地址访问： https://go.drugbank.com/。Decagon 多药联用数据集可从斯坦福 SNAP 项目页面获取，地址为 https://snap.stanford.edu/decagon/。多药联用副作用网络的相关公开数据条目也可通过以下地址获取：

h

t

t

p

s://snap.stanford.edu/biodata/datasets/10017/10017-ChChSe-Decagon.html。用于构建实验图并复现所报告结果的代码和预处理脚本将在论文录用后公开。

b) 竞争利益：作者声明不存在任何竞争利益。

c) 资金支持：本研究得到湖南省重大科技创新平台项目 [2024JC1003] 和国防科技大学研究生创新项目 [XJQY 2024065]的资助。

d) 作者贡献：Z.Y. 和 Z.X. 提出了研究方法、开发了软件、开展了实验并撰写了初稿。C.L. 负责项目监督、筹措资金，并审阅和编辑了稿件。

e) 致谢：作者谨向所有审稿人致以诚挚的谢意，感谢他们认真审阅稿件并提出富有见地的建议。

参考文献

[1] V. P. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson, “图神经网络基准测试”，机器学报

机器学习研究，第 24 卷，第 43 期，第 1–48 页，2023 年。 [在线]。获取地址：http://jmlr.org/papers/v24/22-0567.html [2] C. Ying、T. Cai、S. Luo、S. Zheng、G. Ke、D. He、Y. Shen 和 T. Liu，《Transformer 在图表示学习中的表现真的很差吗？》，载于神经信息处理系统进展，2021 年，第 28 877–28 888 页。[3] L. Rampášek、M. Galkin、V. P. Dwivedi、A. T. Luu、G. Wolf 和 D. Beaini，《通用、强大且可扩展的图 Transformer 配方》，载于神经信息处理系统进展，2022 年。[4] S. Cantürk、R. Liu、O. Lapointe-Gagné、V. Létourneau、G. Wolf、D. Beaini 和 L. Rampášek，《图位置与结构编码器》，载于第 41 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 235 卷。PMLR，2024 年，第 5533–5566 页。 [在线]。获取地址：https://proceedings.mlr.press/v235/canturk24a.html [5] D. Lim、J. D. Robinson、L. Zhao、T. E. Smidt、S. Sra、H. Maron 和 S. Jegelka，《用于谱图表示学习的符号与基不变网络》，载于国际学习表征会议，2023 年。[6] Y. Huang、W. Lu、J. Robinson、Y. Yang、M. Zhang、S. Jegelka 和 P. Li，《图的高表达能力位置编码的稳定性》，载于国际学习表征会议，2024 年。[7] P. Li、Y. Wang、H. Wang 和 J. Leskovec，《距离编码：设计可证明更强大的图表示学习神经网络》，载于神经信息处理系统进展，2020 年。[8] R. B. Gabrielsson、M. Yurochkin 和 J. Solomon，《利用位置编码对图神经网络进行重连》，机器学习研究汇刊，2023 年。[9] V. P. Dwivedi、A. T. Luu、T. Laurent、Y. Bengio 和 X. Bresson，《具有可学习结构与位置表示的图神经网络》，载于国际学习表征会议，2022 年。[10] N. Keriven 和 S. Vaiter，《图神经网络能在随机图上计算哪些函数？位置编码的作用》，载于神经信息处理系统进展，2023 年。[11] H. Li、M. Wang、T. Ma、S. Liu、Z. Zhang 和 P. Chen，《什么因素提升了图 Transformer 的泛化能力？对自注意力与位置编码的理论探究》，载于第 41 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 235 卷。PMLR，2024 年，第 28 784–28 829 页。[在线]。获取地址：https://proceedings.mlr.press/v235/li24bo.html [12] J. Zhang、H. Zhang、C. Xia 和 L. Sun，《Graph-bert：学习图表示只需注意力》，arXiv 预印本 arXiv:2001.05140，2020 年。[13] G. Mialon、D. Chen、M. Selosse 和 J. Mairal，《GraphiT：在 Transformer 中编码图结构》，arXiv 预印本 arXiv:2106.05667，2021 年。[14] C. Yuan、K. Zhao、E. E. Kuruoglu、L. Wang、T. Xu、W. Huang、D. Zhao、H. Cheng 和 Y. Rong，《图 Transformer 综述：架构、理论与应用》，CoRR，第 abs/2502.16533 卷，2025 年。[15] H. Wang、H. Yin、M. Zhang 和 P. Li，《用于构建更强大图神经网络的等变且稳定的位置编码》，载于国际学习表征会议，2022 年。[16] D. Kreuzer、D. Beaini、W. L. Hamilton、V. Létourneau 和 P. Tossou，《利用谱注意力重新思考图 Transformer》，载于神经信息处理系统进展，第 34 卷，2021 年，第 21 618–21 629 页。 [在线]。获取地址：https://proceedings.neurips.cc/paper_files/paper/2021/hash/b4fd1d2cb085390fbbadae65e07876a7-Abstract.html [17] B. Zhang、L. Zhao 和 H. Maron，《谱不变图神经网络的表达能力》，载于第 41 届国际机器学习会议论文集，系列：机器学习研究会议论文集

第 235 卷。PMLR，2024 年，第 60 496–60 526 页。[在线]。获取地址：https://proceedings.mlr.press/v235/zhang24ck.html [18] M. Eliasof、F. Frasca、B. Bevilacqua、E. Treister、G. Chechik 和 H. Maron，《通过随机特征传播实现图位置编码》，载于第 40 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 202 卷。PMLR，2023 年，第 9202–9223 页。[在线]。获取地址：https://proceedings.mlr.press/v202/eliasof23a.html [19] Y. He、C. Bodnar 和 P. Liò，《用于图神经网络的基于层的位置编码》，载于第 2 届 NeurIPS神经表示中的对称性与几何研讨会论文集，系列：机器学习研究会议论文集，第 228 卷。PMLR，2024 年，第 1–18 页。[在线]。获取地址：https://proceedings.mlr.press/v228/he24a.html [20] Y. Huang、H. P. Wang 和 P. Li，《有向图的良好位置编码是什么？》，载于国际学习表征会议，2025 年。[在线]。获取地址：https://openreview.net/forum?id=s4Wm71LFK4[21]J. You、R. Ying 和 J. Leskovec，《位置感知图神经网络》，载于第 36 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 97 卷。PMLR，2019 年，第 7134–7143 页。[在线]。获取地址：https://proceedings.mlr.press/v97/you19b.html [22] W. Park、W. Chang、D. Lee、J. Kim 和 S. won Hwang，《GRPE：图 Transformer 的相对位置编码》，arXiv 预印本 arXiv:2201.12787，2022 年。[23] M. Black、Z. Wan、G. Mishne、A. Nayyeri 和 Y. Wang，《通过位置编码比较图 Transformer》，载于第 41 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 235 卷。PMLR，2024 年，第 4103–4139 页。[在线]。获取地址：https://proceedings.mlr.press/v235/black24b.html [24] J. Zhang、M. Li、Y. Xu、H. He、Q. Li 和 T. Wang，《Strucgcn：用于图嵌入的结构增强图卷积网络》，信息融合，第 117 卷，文章编号 102893，2025 年。[25] Z. Yan、Z. Xie、R. Duan、C. Liu 和 W. Du，《通过基于锚点的扩散几何近似打通距离与谱位置编码》，arXiv 预印本 arXiv:2601.04517，2026 年。[26] J. Yeom、T. Kim、R. Chang 和 K. Song，《用于图 Transformer 的结构与位置集成编码》，模式识别快报，第 183 卷，第 104–110 页，2024 年。[27] L. Ma、C. Lin、D. Lim、A. Romero-Soriano、P. K. Dokania、M. Coates、P. H. S. Torr 和 S.-N. Lim，《无需消息传递的 Transformer 图归纳偏置》，载于第 40 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 202 卷。PMLR，2023 年，第 23 321–23 337 页。[在线]。获取地址：https://proceedings.mlr.press/v202/ma23c.html [28] B. J. Franks、M. Eliasof、S. Cantürk、G. Wolf、C.-B. Schönlieb、S. Fellenz 和 M. Kloft，《迈向图基础模型：位置与结构编码泛化能力研究》，机器学习研究汇刊，2025 年。[在线]。获取地址：https://openreview.net/forum?id=mSoDRZXsqj [29] C. I. Kanatsoulis、E. Choi、S. Jegelka、J. Leskovec 和 A. Ribeiro，《利用图神经网络学习高效位置编码》，载于国际学习表征会议，2025 年。[在线]。获取地址：https://openreview.net/forum?id=AWg2tkbydO [30] F. Grötschla、J. Xie 和 R. Wattenhofer，《图神经网络与图 Transformer 位置编码基准测试》，CoRR，第 abs/2411 12732 卷，2024 年。。[31]S. Thabet、M. Djellabi、I. O. Sokolov、S. Kasture、L. Henry 和 L. Henriet，《用于图神经网络的量子位置编码》，载于第 41 届国际机器学习会议论文集，系列：机器学习研究会议论文集，第 235 卷。PMLR，2024 年，第 47 965–47 996 页。[在线]。获取地址：https://proceedings.mlr.press/v235/thabet24a.html [32] Z. Yan、Z. Xie、C. Liu 和 Y. Wang，《通过拉普拉斯谱编码解决图神经过程中的节点可辨识性问题》，arXiv 预印本 arXiv:2511.19037，2025 年。[33] M. Zitnik、M. Agrawal 和 J. Leskovec，《利用图卷积网络建模多重用药副作用》，生物信息学，第 34 卷，第 13 期，第 i457–i466 页，2018 年。[34] Q. Liu、E. Yao、C. Liu、X. Zhou、Y. Li 和 M. Xu，《M2gcn：用于多重用药副作用建模的多模态图卷积网络》，应用智能，第 53 卷，第 6 期，第 6814–6825 页，2023 年。

目录

I 引言 1

II 文献综述 2

II-A 谱位置编码 . . . . . 2 II-B 距离与结构感知编码 2 II-C 学习式与混合位置/结构编码器 . . . . . 2 II-D 与我们工作的关系.2

III 问题表述 2

IV 理论分析 3

IV-A 观测映射恒等式与一般性反向定理.3 IV-B 通过逐桶碰撞与均衡性进行模型特定细化 5

V 实证评估 6

V-A 实验设置.6 V-B 随机正则图上的类相变行为.6 V-C 观测映射的理论一致性诊断.7 V-D 逐桶细化诊断 . . . 7 V-E 两个真实世界 DDI 图上的可识别性.8 V-F 消融研究.8

VI 讨论与结论 8

VI-A 讨论.8 VI-B 局限性.9 VI-C 未来方向.9

参考文献 10

附录 13

A 引理 1 的证明.13 B 引理 2 的证明.14 C 引理 3 的证明.14 D 命题4 的证明.14 E 定理 5 的证明.15 F 命题 6 的证明.15 G 推论 2 的证明.17 H 推论 3 的证明.18 I 数据集.18 J 基线.19 K 编码配置.19 L 评估指标.19 M 消融实验方案.19 N 观测映射的完整诊断表.19 O 量化校准.20 P 逐桶细化诊断. . . . . 20 Q 补充稳健性细节. 20

附录

本附录给出了第四节中理论结果的完整证明。为方便起见，我们按照正文中的顺序重述每项结果，然后给出其证明。

A. 引理 1 的证明

引理 (引理 1). 对于任意固定的 $(G, \mathcal{A})$，令

$$Y := F^{(\eta)}_{G,\mathcal{A}}(v_\star). \tag{49}$$

那么对于任意可测映射

$$s_{G,\mathcal{A}} : \mathcal{Y}^{(\eta)}_{k,m}(G, \mathcal{A}) \to V, \tag{50}$$

有

$$\begin{aligned}\mathbb{P}\Big(s_{G,\mathcal{A}}(Y) = v_\star \;\Big|\; G, \mathcal{A}\Big) &= \mathbb{E}\left[\frac{\mathbf{1}\Big\{s_{G,\mathcal{A}}(Y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)\Big\}}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)|} \;\middle|\; G, \mathcal{A}\right] \\ &\le \mathbb{E}\left[\frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)|} \;\middle|\; G, \mathcal{A}\right].\end{aligned} \tag{51}$$

此外，

$$\mathbb{E}\left[\frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)|} \;\middle|\; G, \mathcal{A}\right] = \frac{\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|}{n}. \tag{52}$$

该上界由任意满足以下条件的可测截面达到

$$\tilde{s}_{G,\mathcal{A}} : \mathcal{Y}^{(\eta)}_{k,m}(G, \mathcal{A}) \to V \tag{53}$$

满足

$$F^{(\eta)}_{G,\mathcal{A}}\big(\tilde{s}_{G,\mathcal{A}}(y)\big) = y, \quad \forall\, y \in \mathcal{Y}^{(\eta)}_{k,m}(G, \mathcal{A}). \tag{54}$$

因此，

$$\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) = 1 - \frac{\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|}{n}. \tag{55}$$

证明。令

$$Y := F^{(\eta)}_{G,\mathcal{A}}(v_\star). \tag{56}$$

由于 $v_\star$ 在 V 上均匀分布，对于每个

$$y \in \mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}}), \tag{57}$$

给定 $Y = y$ 时， $v_\star$ 的条件分布在原像上均匀分布

$$\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y) = \{v \in V :\ F^{(\eta)}_{G,\mathcal{A}}(v) = y\}. \tag{58}$$

事实上，对于任意 $v \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)$，

$$\begin{aligned}\mathbb{P}(v_\star = v \mid Y = y, G, \mathcal{A}) &= \frac{\mathbb{P}(v_\star = v \mid G, \mathcal{A})}{\mathbb{P}(Y = y \mid G, \mathcal{A})} \\ &= \frac{1/n}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)|/n} = \frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)|}.\end{aligned} \tag{59}$$

现在固定任意可测映射 $s_{G,\mathcal{A}}$。在给定 $Y = y$ 的条件下，事件

$$\{s_{G,\mathcal{A}}(Y) = v_\star\} \tag{60}$$

只有在以下条件成立时才会发生

$$s_{G,\mathcal{A}}(y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y). \tag{61}$$

因此

$$\mathbb{P}\big(s_{G,\mathcal{A}}(Y) = v_\star \mid Y = y, G, \mathcal{A}\big) = \begin{cases} \dfrac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)|}, & s_{G,\mathcal{A}}(y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y), \\ 0, & s_{G,\mathcal{A}}(y) \notin \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y). \end{cases} \tag{62}$$

等价地，

$$\mathbb{P}\big(s_{G,\mathcal{A}}(Y) = v_\star \mid Y = y, G, \mathcal{A}\big) = \frac{\mathbf{1}\Big\{s_{G,\mathcal{A}}(y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)\Big\}}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)|}. \tag{63}$$

对 Y 在给定 $(G, \mathcal{A})$ 的条件下取期望，得到

$$\mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) = v_\star \;\Big|\; G, \mathcal{A}\Big) = \mathbb{E}\left[\frac{\mathbf{1}\Big\{s_{G,\mathcal{A}}(Y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)\Big\}}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)|} \;\middle|\; G, \mathcal{A}\right]. \tag{64}$$

由于示性函数至多为 1，我们得到

$$\mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) = v_\star \;\Big|\; G, \mathcal{A}\Big) \le \mathbb{E}\left[\frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)|} \;\middle|\; G, \mathcal{A}\right]. \tag{65}$$

因为

$$Y = F^{(\eta)}_{G,\mathcal{A}}(v_\star), \tag{66}$$

我们还有

$$|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(Y)| = |\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)|. \tag{67}$$

Thus

$$\mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) = v_\star \;\Big|\; G, \mathcal{A}\Big) \le \mathbb{E}\left[\frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)|} \;\middle|\; G, \mathcal{A}\right]. \tag{68}$$

接下来，令

$$\mathcal{Y} := \mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}}), \quad N_y := |\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y)|, \quad y \in \mathcal{Y}. \tag{69}$$

由于 $v_\star$ 在 V上均匀分布，因此有

$$\mathbb{P}(Y = y \mid G, \mathcal{A}) = \frac{N_y}{n}. \tag{70}$$

因此

$$\begin{aligned}\mathbb{E}\left[\frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)|} \;\middle|\; G, \mathcal{A}\right] &= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid G, \mathcal{A}) \cdot \frac{1}{N_y} \\ &= \sum_{y \in \mathcal{Y}} \frac{N_y}{n} \cdot \frac{1}{N_y} \\ &= \sum_{y \in \mathcal{Y}} \frac{1}{n} = \frac{|\mathcal{Y}|}{n} = \frac{|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})|}{n}.\end{aligned} \tag{71}$$

现在令 $\tilde{s}_{G,\mathcal{A}}$ 为任意可测截面，即

$$F^{(\eta)}_{G,\mathcal{A}}\big(\tilde{s}_{G,\mathcal{A}}(y)\big) = y, \quad \forall\, y \in \mathcal{Y}. \tag{72}$$

Then

$$\tilde{s}_{G,\mathcal{A}}(y) \in \mathcal{P}^{(\eta)}_{G,\mathcal{A}}(y), \quad \forall\, y \in \mathcal{Y}, \tag{73}$$

因此，上一个公式中的示性函数恒等于 1。于是

$$\mathbb{P}\Big(\tilde{s}_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) = v_\star \;\Big|\; G,\mathcal{A}\Big) = \mathbb{E}\left[\frac{1}{|\mathcal{P}^{(\eta)}_{G,\mathcal{A}}(v_\star)|}\;\middle|\; G,\mathcal{A}\right] = \frac{|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})|}{n}. \tag{74}$$

因此，最大的条件成功概率等于

$$\frac{|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})|}{n}, \tag{75}$$

因此

$$\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) = 1 - \frac{|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})|}{n}. \tag{76}$$

证明完毕。 □

B. 引理 *2* 的证明

引理 (引理 2)。 假设 $\mathrm{diam}(G) \le C_{\mathrm{diam}}$ 对数 n。则

$$\big|\{\mathbf{d}_{\mathcal{A}}(v) : v \in V\}\big| \le (C\log n)^k, \tag{77}$$

特别地，非空桶的数量 $B_{\mathbf{t}}(G,\mathcal{A})$至多为 $(C$ 对数 $n)^k$。

证明。假设

$$\mathrm{diam}(G) \le C_{\mathrm{diam}} \log n. \tag{78}$$

任取顶点 $v \in V$。对于每个锚点 $a_i \in \mathcal{A}$，最短路径距离满足

$$0 \le \mathrm{SPD}(v,a_i) \le \mathrm{diam}(G) \le C_{\mathrm{diam}} \log n. \tag{79}$$

The因此，锚点距离向量的每个坐标 向量

$$\mathbf{d}_{\mathcal{A}}(v) = \big(\mathrm{SPD}(v,a_1),\ldots,\mathrm{SPD}(v,a_k)\big) \tag{80}$$

取值于有限集合

$$\{0,1,\ldots,\lfloor C_{\mathrm{diam}} \log n\rfloor\}. \tag{81}$$

因此，可能的 k 维距离向量总数至多为

$$\big(\lfloor C_{\mathrm{diam}} \log n\rfloor + 1\big)^k. \tag{82}$$

将加性常数吸收到 C中，我们得 到

$$\big|\{\mathbf{d}_{\mathcal{A}}(v) : v \in V\}\big| \le (C\log n)^k. \tag{83}$$

由于每个非空桶 $B_{\mathbf{t}}(G,\ \mathcal{A})$ 都对应一个已达到的向量 $\mathbf{t}$，因此非空桶的数量也受同一数量的限制。证明至此完成。 □

C. 引理 *3* 的证明

引理 (引理 3)。 假设 $\mathrm{diam}(G) \le C_{\mathrm{diam}}$ 对数 n，并且假设 *1* 在 $\mathcal{E}_{\mathrm{ent}}$ 上成立。那么，在事件 $\{\mathrm{diam}(G) \le C_{\mathrm{diam}}$ 对数 $n\} \cap \mathcal{E}_{\mathrm{ent}}$ 下，

$$\begin{aligned}\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| &\le (C\log n)^k \cdot \big|Q_\eta(\mathcal{S}_m(V))\big| \\ &\le (C\log n)^k \cdot \Big(\frac{C_{\mathrm{ent}}}{\eta}\Big)^{c_{\mathrm{ent}} m},\end{aligned} \tag{84}$$

其中 $C = C(r) > 0$ 仅取决于 C_{diam}。

该证明结合了距离桶计数与量化谱码界。

证明。 在该事件上固定 $(G,\mathcal{A})$

$$\{\mathrm{diam}(G) \le C_{\mathrm{diam}} \log n\} \cap \mathcal{E}_{\mathrm{ent}}. \tag{85}$$

根据引理 2，可取距离向量的数量至多为 $(C\log n)^k$。

固定一个这样的距离向量$\mathbf{t}$ ，并考虑对应的桶

$$B_{\mathbf{t}}(G,\mathcal{A}) = \{v \in V :\ \mathbf{d}_{\mathcal{A}}(v) = \mathbf{t}\}. \tag{86}$$

在这个桶内，观测映射中唯一可能变化的部分是量化谱项

$$Q_\eta(\mathcal{S}_m(v)). \tag{87}$$

因此，在 $B_{\mathbf{t}}(G,\mathcal{A})$ 内可取到的不同观测值数量至多为

$$\big|Q_\eta(\mathcal{S}_m(V))\big|. \tag{88}$$

对所有可取距离桶求和，得到

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \le (C\log n)^k \cdot \big|Q_\eta(\mathcal{S}_m(V))\big|. \tag{89}$$

最后，由假设 1 可得

$$\big|Q_\eta(\mathcal{S}_m(V))\big| \le \Big(\frac{C_{\mathrm{ent}}}{\eta}\Big)^{c_{\mathrm{ent}} m}, \tag{90}$$

因此

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \le (C\log n)^k \cdot \Big(\frac{C_{\mathrm{ent}}}{\eta}\Big)^{c_{\mathrm{ent}} m}. \tag{91}$$

证明至此完成。 □

D. 命题*4*的证明

命题 (命题4)。 对于嵌入 $\mathcal{S}_m(v) \ =(\Phi_m(v,j)^2)_{j\in[m]}$ 和 $\eta \in (0,1)$，有

$$\big|Q_\eta(\mathcal{S}_m(V))\big| \le \big|Q_\eta([0,1]^m)\big| \le \Big(\frac{2}{\eta}\Big)^m. \tag{92}$$

特别地，假设 *1* 成立，其中 $C_{\mathrm{ent}} = 2$、 $c_{\mathrm{ent}} = 1$ 和 $\mathcal{E}_{\mathrm{ent}}$ 均等于整个概率空间。

证明。 对于能量嵌入，

$$\mathcal{S}_m(v) = \big(\Phi_m(v,1)^2,\ldots,\Phi_m(v,m)^2\big) \in \mathbb{R}^m. \tag{93}$$

由于 Φ_m 是标准正交特征基矩阵，因此每个元素都满足

$$|\Phi_m(v,j)| \le 1, \tag{94}$$

因此，每个坐标的平方都位于 $[0,1]$ 中。于是

$$\mathcal{S}_m(V) \subset [0,1]^m. \tag{95}$$

Thus

$$\big|Q_\eta(\mathcal{S}_m(V))\big| \le \big|Q_\eta([0,1]^m)\big|. \tag{96}$$

在分箱宽度为 η 的逐坐标量化下，每个坐标区间 $[0,1]$ 至多与

$$\left\lfloor \frac{1}{\eta} \right\rfloor + 1 \tag{97}$$

个量化区间相交。由于 $\eta \in (0,\ 1)$，因此有

$$\left\lfloor \frac{1}{\eta} \right\rfloor + 1 \le \frac{2}{\eta}. \tag{98}$$

因此

$$\big|Q_\eta([0,1]^m)\big| \le \left(\left\lfloor \frac{1}{\eta} \right\rfloor + 1 \right)^m \le \left(\frac{2}{\eta} \right)^m. \tag{99}$$

由此得证。 □

E. 定理 *5* 的证明

定理 (定理 5)。 固定 $r \ge 3$，并令 $G \sim \mathcal{G}_{n,\ r}$。令 $v_\star \sim \mathrm{Unif}(V)$，并令 $\mathcal{A} = \{a_1, \ldots, a_k\}$ 从 V 中均匀无放回抽样，且独立于 $v_\star$ 和 G。固定 $m \ge 1$，并令 $\eta \in (0,\ 1)$。假设直径界 $\mathrm{diam}(G) \le C_{\mathrm{diam}} \log n$ 以概率 $1 - o(1)$ 成立，并且谱嵌入 $\mathcal{S}_m$ 在分辨率 η 下满足假设 *1*。若存在 $\varepsilon_0 \in (0,\ 1)$ 使得

$$k \log\log n + c_{\mathrm{ent}} m \log\Big(\frac{C_{\mathrm{ent}}}{\eta} \Big) \le (1 - \varepsilon_0) \log n, \tag{100}$$

那么存在 $\varepsilon_1 \in (0, \varepsilon_0)$，使得在 $(G, \mathcal{A})$ 上，以概率 $1 - o(1)$，

$$\inf_{s_{G,\mathcal{A}}} \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \ne v_\star \ \Big|\ G, \mathcal{A} \Big) \ge 1 - n^{-\varepsilon_1}. \tag{101}$$

特别地，对于所有充分大的 n，左侧都以下界为任意固定常数 $\delta \in (0,1)$。

证明。 固定陈述中的 ε_0，并考虑事件

$$\mathcal{E} := \{\mathrm{diam}(G) \le C_{\mathrm{diam}} \log n\} \cap \mathcal{E}_{\mathrm{ent}}. \tag{102}$$

根据假设，

$$\mathbb{P}(\mathcal{E}) = 1 - o(1). \tag{103}$$

在事件 $\mathcal{E}$ 上，引理 3 给出

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \le (C \log n)^k \cdot \Big(\frac{C_{\mathrm{ent}}}{\eta} \Big)^{c_{\mathrm{ent}} m}. \tag{104}$$

将此界代入引理 1，可得

$$\begin{aligned} \inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) &= 1 - \frac{\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|}{n} \\ &\ge 1 - \frac{(C \log n)^k \cdot \big(\frac{C_{\mathrm{ent}}}{\eta} \big)^{c_{\mathrm{ent}} m}}{n}. \end{aligned} \tag{105}$$

现在估计分子。取对数，

$$\log\Big((C\log n)^k \cdot \Big(\tfrac{C_{\mathrm{ent}}}{\eta} \Big)^{c_{\mathrm{ent}} m} \Big) = k \log(C \log n) + c_{\mathrm{ent}} m \log\Big(\tfrac{C_{\mathrm{ent}}}{\eta} \Big). \tag{106}$$

由于

$$k \log\log n \le (1 - \varepsilon_0) \log n, \tag{107}$$

we have

$$k \le \frac{(1 - \varepsilon_0) \log n}{\log\log n}. \tag{108}$$

因此

$$k \log C \le \frac{(1 - \varepsilon_0) \log C}{\log\log n} \log n = o(\log n). \tag{109}$$

因此，存在 $\varepsilon_1 \in (0, \varepsilon_0)$，使得对所有充分大的 n，

$$k \log(C \log n) + c_{\mathrm{ent}} m \log\Big(\frac{C_{\mathrm{ent}}}{\eta} \Big) \le (1 - \varepsilon_1) \log n. \tag{110}$$

因此，该假设意味着

$$(C \log n)^k \cdot \Big(\frac{C_{\mathrm{ent}}}{\eta} \Big)^{c_{\mathrm{ent}} m} \le n^{1 - \varepsilon_1} \tag{111}$$

对所有充分大的 n均成立。因此

$$\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) \ge 1 - \frac{n^{1-\varepsilon}}{n} = 1 - n^{-\varepsilon_1}. \tag{112}$$

根据 $\mathrm{Err}(s_{G,\mathcal{A}})$ 的定义，这恰好是

$$\inf_{s_{G,\mathcal{A}}} \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \ne v_\star \ \Big|\ G, \mathcal{A} \Big) \ge 1 - n^{-\varepsilon_1}. \tag{113}$$

由于 $\mathbb{P}(\mathcal{E}) = 1 - o(1)$，该界以概率$1 - o(1)$ 在 $(G, \mathcal{A})$ 上成立。证明至此完成。 □

F. 命题 *6* 的证明

命题 (命题 6（重述）)。 假设 $\mathrm{diam}(G) \le C_{\mathrm{diam}}$ 对数n。设对于每个可取到的距离向量 $\mathbf{t}$，以下条件之一成立：

- $|B_{\mathbf{t}}(G, \mathcal{A})| = 1$; *or*
- $|B_{\mathbf{t}}(G, \mathcal{A})| \ge 2$, *and*

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \ge c_0\, \eta^{c_1 m}, \tag{114}$$

并且

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \le \beta, \tag{115}$$

其中 c_0、$c_1 > 0$ 和 $\beta \ge 1$ 为常数。则

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \le D(G, \mathcal{A}) \cdot \Big(1 + \frac{\beta}{c_0\, \eta^{c_1 m}} \Big). \tag{116}$$

特别地，

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big| \le (C \log n)^k \cdot \Big(1 + \frac{\beta}{c_0\, \eta^{c_1 m}} \Big). \tag{117}$$

证明。 在命题的假设下，固定 $(G, \mathcal{A})$。令

$$\mathcal{T}(G, \mathcal{A}) := \{\mathbf{d}_{\mathcal{A}}(v) : v \in V\} \tag{118}$$

表示可取距离向量的集合，因此

$$D(G, \mathcal{A}) = |\mathcal{T}(G, \mathcal{A})|. \tag{119}$$

对于每个 $\mathbf{t} \in \mathcal{T}(G,\mathcal{A})$，记

$$B_{\mathbf{t}} \coloneqq B_{\mathbf{t}}(G,\mathcal{A}) = \{v \in V :\ \mathbf{d}_{\mathcal{A}}(v) = \mathbf{t}\}, \tag{120}$$

并令

$$b_{\mathbf{t}} \coloneqq |B_{\mathbf{t}}|. \tag{121}$$

另令

$$Q_\eta(\mathcal{S}_m(B_{\mathbf{t}})) = \{Q_\eta(\mathcal{S}_m(v)) :\ v \in B_{\mathbf{t}}\} \tag{122}$$

为桶$B_{\mathbf{t}}$内取得的量化谱编码集合，并定义

$$M_{\mathbf{t}} \coloneqq M_{\eta,m}(B_{\mathbf{t}}) = \big|Q_\eta(\mathcal{S}_m(B_{\mathbf{t}}))\big|. \tag{123}$$

对每个

$$z \in Q_\eta(\mathcal{S}_m(B_{\mathbf{t}})), \tag{124}$$

定义占用数

$$N_{\mathbf{t}}(z) \coloneqq N_{\eta,m}(B_{\mathbf{t}}; z) = \big|\{v \in B_{\mathbf{t}} :\ Q_\eta(\mathcal{S}_m(v)) = z\}\big|. \tag{125}$$

那么根据构造，

$$\sum_{z \in Q_\eta(\mathcal{S}_m(B_{\mathbf{t}}))} N_{\mathbf{t}}(z) = b_{\mathbf{t}}. \tag{126}$$

现在我们分别对每个桶的 $M_{\mathbf{t}}$ 进行界定。

情形 1： $b_{\mathbf{t}} = 1$。

在此情形下，桶中恰好包含一个顶点，因此恰好贡献一个观测值。更准确地说，

$$M_{\mathbf{t}} = 1. \tag{127}$$

情形 2： $b_{\mathbf{t}} \geq 2$。

在此情形下，命题的假设给出

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}) \geq c_0\, \eta^{c_1 m} \tag{128}$$

and

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}) \leq \beta. \tag{129}$$

我们先改写平衡条件。根据定义，

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}) = \frac{M_{\mathbf{t}}}{b_{\mathbf{t}}} \cdot \max_{z \in Q_\eta(\mathcal{S}_m(B_{\mathbf{t}}))} N_{\mathbf{t}}(z). \tag{130}$$

因此

$$\frac{M_{\mathbf{t}}}{b_{\mathbf{t}}} \cdot \max_z N_{\mathbf{t}}(z) \leq \beta, \tag{131}$$

这意味着

$$\max_z N_{\mathbf{t}}(z) \leq \beta\, \frac{b_{\mathbf{t}}}{M_{\mathbf{t}}}. \tag{132}$$

接下来，我们来给出二次占用和的界：

$$\sum_z N_{\mathbf{t}}(z)^2 \leq \Big(\max_z N_{\mathbf{t}}(z)\Big) \sum_z N_{\mathbf{t}}(z). \tag{133}$$

利用前面的不等式和恒等式 $\sum_z N_{\mathbf{t}}(z) = b_{\mathbf{t}}$，我们得到

$$\sum_z N_{\mathbf{t}}(z)^2 \leq \beta\, \frac{b_{\mathbf{t}}}{M_{\mathbf{t}}} \cdot b_{\mathbf{t}} = \beta\, \frac{b_{\mathbf{t}}^2}{M_{\mathbf{t}}}. \tag{134}$$

另一方面，

$$\begin{aligned} \sum_z N_{\mathbf{t}}(z)\big(N_{\mathbf{t}}(z)-1\big) &= \sum_z N_{\mathbf{t}}(z)^2 - \sum_z N_{\mathbf{t}}(z) \\ &= \sum_z N_{\mathbf{t}}(z)^2 - b_{\mathbf{t}}. \end{aligned} \tag{135}$$

因此

$$\sum_z N_{\mathbf{t}}(z)\big(N_{\mathbf{t}}(z)-1\big) \leq \beta\, \frac{b_{\mathbf{t}}^2}{M_{\mathbf{t}}} - b_{\mathbf{t}}. \tag{136}$$

现在我们使用碰撞下界。根据 $\mathrm{Coll}^{Q}_{\eta,\,m}(B_{\mathbf{t}})$ 的定义，

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}) = \tfrac{1}{b_{\mathbf{t}}(b_{\mathbf{t}}-1)} \sum_{\substack{u,v \in B_{\mathbf{t}} \\ u \neq v}} \mathbf{1}\Big\{Q_\eta(\mathcal{S}_m(u)) = Q_\eta(\mathcal{S}_m(v))\Big\}. \tag{137}$$

按照共同的量化编码对有序对 $(u,\ v)$ 进行分组，可得

$$\sum_{\substack{u,v \in B_{\mathbf{t}} \\ u \neq v}} \mathbf{1}\Big\{Q_\eta(\mathcal{S}_m(u)) = Q_\eta(\mathcal{S}_m(v))\Big\} = \sum_z N_{\mathbf{t}}(z)\big(N_{\mathbf{t}}(z)-1\big). \tag{138}$$

因此

$$\sum_z N_{\mathbf{t}}(z)\big(N_{\mathbf{t}}(z)-1\big) = b_{\mathbf{t}}(b_{\mathbf{t}}-1)\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}). \tag{139}$$

利用 $\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}})$ 的下界，我们得到

$$\sum_z N_{\mathbf{t}}(z)\big(N_{\mathbf{t}}(z)-1\big) \geq b_{\mathbf{t}}(b_{\mathbf{t}}-1) c_0\, \eta^{c_1 m}. \tag{140}$$

合并同一数量的上界和下界，得到

$$b_{\mathbf{t}}(b_{\mathbf{t}}-1) c_0\, \eta^{c_1 m} \leq \beta\, \frac{b_{\mathbf{t}}^2}{M_{\mathbf{t}}} - b_{\mathbf{t}}. \tag{141}$$

在两边加上 $b_{\mathbf{t}}$ ：

$$b_{\mathbf{t}} + b_{\mathbf{t}}(b_{\mathbf{t}}-1) c_0\, \eta^{c_1 m} \leq \beta\, \frac{b_{\mathbf{t}}^2}{M_{\mathbf{t}}}. \tag{142}$$

在左侧提取 $b_{\mathbf{t}}$ ：

$$b_{\mathbf{t}}\Big(1 + (b_{\mathbf{t}}-1) c_0\, \eta^{c_1 m}\Big) \leq \beta\, \frac{b_{\mathbf{t}}^2}{M_{\mathbf{t}}}. \tag{143}$$

由于 $b_{\mathbf{t}} \geq 1$，我们可以将两边除以 $b_{\mathbf{t}}$，得到

$$1 + (b_{\mathbf{t}}-1) c_0\, \eta^{c_1 m} \leq \beta\, \frac{b_{\mathbf{t}}}{M_{\mathbf{t}}}. \tag{144}$$

等价地，

$$M_{\mathbf{t}} \leq \beta\, \frac{b_{\mathbf{t}}}{1 + (b_{\mathbf{t}}-1) c_0\, \eta^{c_1 m}}. \tag{145}$$

现在我们简化这个界。根据定义，$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}) \leq 1$，并且由于

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}) \geq c_0\, \eta^{c_1 m}, \tag{146}$$

由此可得

$$c_0\, \eta^{c_1 m} \leq 1. \tag{147}$$

因此

$$1 - c_0\, \eta^{c_1 m} \geq 0. \tag{148}$$

由此可得

$$1+(b_{\mathbf{t}}-1)c_0\,\eta^{c_1 m}=b_{\mathbf{t}}c_0\,\eta^{c_1 m}+\big(1-c_0\,\eta^{c_1 m}\big)\geq b_{\mathbf{t}}c_0\,\eta^{c_1 m}. \tag{149}$$

将此下界代入分母可得

$$M_{\mathbf{t}}\leq\beta\,\frac{b_{\mathbf{t}}}{b_{\mathbf{t}}c_0\,\eta^{c_1 m}}=\frac{\beta}{c_0\,\eta^{c_1 m}}. \tag{150}$$

因此，对于每个可取距离向量 $\mathbf{t}$，我们已经证明

$$M_{\mathbf{t}}\leq 1\quad\text{if } b_{\mathbf{t}}=1, \tag{151}$$

and

$$M_{\mathbf{t}}\leq\frac{\beta}{c_0\,\eta^{c_1 m}}\quad\text{if } b_{\mathbf{t}}\geq 2. \tag{152}$$

现在对所有可取距离向量求和。 Let

$$\mathcal{T}_1\coloneqq\{\mathbf{t}\in\mathcal{T}(G,\mathcal{A}):\ b_{\mathbf{t}}=1\} \tag{153}$$

and

$$\mathcal{T}_{\geq 2}\coloneqq\{\mathbf{t}\in\mathcal{T}(G,\mathcal{A}):\ b_{\mathbf{t}}\geq 2\}. \tag{154}$$

Then

$$\mathcal{T}(G,\mathcal{A})=\mathcal{T}_1\cup\mathcal{T}_{\geq 2},\quad\mathcal{T}_1\cap\mathcal{T}_{\geq 2}=\varnothing, \tag{155}$$

and

$$|\mathcal{T}_1|+|\mathcal{T}_{\geq 2}|=D(G,\mathcal{A}). \tag{156}$$

像集大小等于所有可取距离桶中不同量化谱编码数量之和：

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|=\sum_{\mathbf{t}\in\mathcal{T}(G,\mathcal{A})}M_{\mathbf{t}}. \tag{157}$$

因此

$$\begin{aligned}\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|&=\sum_{\mathbf{t}\in\mathcal{T}_1}M_{\mathbf{t}}+\sum_{\mathbf{t}\in\mathcal{T}_{\geq 2}}M_{\mathbf{t}}\\&\leq\sum_{\mathbf{t}\in\mathcal{T}_1}1+\sum_{\mathbf{t}\in\mathcal{T}_{\geq 2}}\frac{\beta}{c_0\,\eta^{c_1 m}}\\&=|\mathcal{T}_1|+|\mathcal{T}_{\geq 2}|\frac{\beta}{c_0\,\eta^{c_1 m}}\\&\leq D(G,\mathcal{A})\cdot\Big(1+\frac{\beta}{c_0\,\eta^{c_1 m}}\Big).\end{aligned} \tag{158}$$

这就证明了第一个界。

最后，由引理 2 可得

$$D(G,\mathcal{A})\leq(C\log n)^k, \tag{159}$$

因此

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|\leq(C\log n)^k\cdot\Big(1+\frac{\beta}{c_0\,\eta^{c_1 m}}\Big). \tag{160}$$

证明至此完成。 □

G. 推论 2 的证明

推论 (推论 2（重述）**)**。假设 $\mathrm{diam}(G)\quad\leq C_{\mathrm{diam}}$ 对数 n 的概率为 $1-o(1)$，并假设存在一个事件 $\mathcal{E}_{\mathrm{ref}}$，满足 $\mathbb{P}(\mathcal{E}_{\mathrm{ref}})=1-o(1)$，使得在 $\mathcal{E}_{\mathrm{ref}}$ 上，对于每个可取到的距离向量$\mathbf{t}$，以下条件之一成立：

- $|B_{\mathbf{t}}(G,\mathcal{A})|=1$*; or*
- $|B_{\mathbf{t}}(G,\mathcal{A})|\geq 2$*, and*

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}(G,\mathcal{A}))\geq c_0\,\eta^{c_1 m}, \tag{161}$$

结合

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}(G,\mathcal{A}))\leq\beta, \tag{162}$$

对于某些常数 c_0、 $c_1>0$ 和 $\beta\geq 1$。那么，以概率$1-o(1)$在 $(G$、 $\mathcal{A})$ 上，

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|\leq(C\log n)^k\cdot\Big(1+\frac{\beta}{c_0\,\eta^{c_1 m}}\Big). \tag{163}$$

特别地，由于 $\eta\in(0,\ 1)$，

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|\leq(C\log n)^k\cdot\Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m}. \tag{164}$$

证明。 考虑事件

$$\mathcal{E}\coloneqq\{\mathrm{diam}(G)\leq C_{\mathrm{diam}}\log n\}\cap\mathcal{E}_{\mathrm{ref}}. \tag{165}$$

根据假设，

$$\mathbb{P}(\mathcal{E})=1-o(1). \tag{166}$$

在事件 $\mathcal{E}$ 上固定 $(G,\mathcal{A})$。此时直径界

$$\mathrm{diam}(G)\leq C_{\mathrm{diam}}\log n \tag{167}$$

成立，并且对于每个可取到的距离向量 $\mathbf{t}$，命题 6 中出现的分桶假设均得到满足。因此，命题 6 适用，并给出

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|\leq D(G,\mathcal{A})\cdot\Big(1+\frac{\beta}{c_0\,\eta^{c_1 m}}\Big). \tag{168}$$

现在应用引理 2，可得

$$D(G,\mathcal{A})\leq(C\log n)^k. \tag{169}$$

因此

$$\big|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\big|\leq(C\log n)^k\cdot\Big(1+\frac{\beta}{c_0\,\eta^{c_1 m}}\Big). \tag{170}$$

这证明了第一个结论。

对于第二个结论，注意到 $\eta\in(0,\ 1)$ 意味着

$$\eta^{-c_1 m}\geq 1. \tag{171}$$

因此

$$1\leq\eta^{-c_1 m} \tag{172}$$

并且

$$\frac{\beta}{c_0\,\eta^{c_1 m}}=\frac{\beta}{c_0}\eta^{-c_1 m}. \tag{173}$$

将这两个界相加可得

$$1+\frac{\beta}{c_0\,\eta^{c_1 m}}\leq\Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m}. \tag{174}$$

代入前一个估计式可得

$$\left|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\right| \le (C\log n)^k \cdot \Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m}. \tag{175}$$

由于 $\mathbb{P}(\mathcal{E}) = 1 - o(1)$，这两个界都以概率$1 - o(1)$ 在 $(G, \mathcal{A})$ 上成立。证明至此完成。 □

H. 推论 *3* 的证明

推论 (推论 3（重述）)。 在定理 *5* 的设定下，进一步假设存在常数 c_0、$c_1 > 0$ 和 $\beta \ge 1$，以及一个满足 $\mathbb{P}(\mathcal{E}_{\mathrm{ref}}) = 1 - o(1)$ 的事件 $\mathcal{E}_{\mathrm{ref}}$，使得在 $\mathcal{E}_{\mathrm{ref}}$ 上，对于每个可取到的距离向量 $\mathbf{t}$，以下条件之一成立：

- $|B_{\mathbf{t}}(G, \mathcal{A})| = 1$*; or*
- $|B_{\mathbf{t}}(G, \mathcal{A})| \ge 2$*, and*

$$\mathrm{Coll}^{Q}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \ge c_0\, \eta^{c_1 m}, \tag{176}$$

连同

$$\mathrm{Bal}_{\eta,m}(B_{\mathbf{t}}(G, \mathcal{A})) \le \beta. \tag{177}$$

若存在 $\varepsilon_0 \in (0,1)$ 使得

$$k\log\log n + c_1 m \log(1/\eta) \le (1-\varepsilon_0)\log n, \tag{178}$$

则存在 $\varepsilon_1 \in (0, \varepsilon_0)$ 使得以概率 $1-o(1)$在 $(G, \mathcal{A})$ 上，

$$\inf_{s_{G,\mathcal{A}}} \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \ne v_\star \;\Big|\; G, \mathcal{A}\Big) \ge 1 - n^{-\varepsilon_1}. \tag{179}$$

证明。 考虑事件

$$\mathcal{E} \coloneqq \{\mathrm{diam}(G) \le C_{\mathrm{diam}} \log n\} \cap \mathcal{E}_{\mathrm{ref}}. \tag{180}$$

根据定理 5 和本推论的假设，

$$\mathbb{P}(\mathcal{E}) = 1 - o(1). \tag{181}$$

在事件 $\mathcal{E}$ 上固定 $(G, \mathcal{A})$。于是可应用推论 2，得到

$$\left|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\right| \le (C\log n)^k \cdot \Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m}. \tag{182}$$

将此估计代入引理 1，得到

$$\begin{aligned}\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) &= 1 - \frac{\left|\mathrm{Im}(F^{(\eta)}_{G,\mathcal{A}})\right|}{n} \\ &\ge 1 - \frac{(C\log n)^k \cdot \big(1+\frac{\beta}{c_0}\big)\eta^{-c_1 m}}{n}.\end{aligned} \tag{183}$$

现在估计分子。取对数可得

$$\begin{aligned}&\log\left((C\log n)^k \cdot \Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m}\right) \\ =&k\log(C\log n) + \log\Big(1+\frac{\beta}{c_0}\Big) + c_1 m\log(1/\eta).\end{aligned} \tag{184}$$

利用

$$\log(C\log n) = \log\log n + \log C, \tag{185}$$

我们可以将其改写为

$$k\log\log n + c_1 m\log(1/\eta) + k\log C + \log\Big(1+\frac{\beta}{c_0}\Big). \tag{186}$$

根据假设，

$$k\log\log n + c_1 m\log(1/\eta) \le (1-\varepsilon_0)\log n. \tag{187}$$

特别地，

$$k\log\log n \le (1-\varepsilon_0)\log n, \tag{188}$$

因此

$$k \le \frac{(1-\varepsilon_0)\log n}{\log\log n}. \tag{189}$$

两边乘以常数 $\log C$，得到

$$k\log C \le \frac{(1-\varepsilon_0)\log C}{\log\log n}\log n = o(\log n). \tag{190}$$

此外，

$$\log\Big(1+\frac{\beta}{c_0}\Big) = O(1) = o(\log n). \tag{191}$$

因此，这两个额外项

$$k\log C + \log\Big(1+\frac{\beta}{c_0}\Big) \tag{192}$$

与 $\log n$ 相比可以忽略。因此，存在$\varepsilon_1 \in (0, \varepsilon_0)$，使得对于所有充分大的 n，

$$k\log(C\log n) + \log\Big(1+\frac{\beta}{c_0}\Big) + c_1 m\log(1/\eta) \le (1-\varepsilon_1)\log n. \tag{193}$$

等价地，

$$(C\log n)^k \cdot \Big(1+\frac{\beta}{c_0}\Big)\eta^{-c_1 m} \le n^{1-\varepsilon_1} \tag{194}$$

对于所有充分大的 n。

将其代入先前关于误差的下界，得到

$$\inf_{s_{G,\mathcal{A}}} \mathrm{Err}(s_{G,\mathcal{A}}) \ge 1 - \frac{n^{1-\varepsilon_1}}{n} = 1 - n^{-\varepsilon_1}. \tag{195}$$

根据 $\mathrm{Err}(s_{G,\mathcal{A}})$ 的定义，这恰好是

$$\inf_{s_{G,\mathcal{A}}} \mathbb{P}\Big(s_{G,\mathcal{A}}(F^{(\eta)}_{G,\mathcal{A}}(v_\star)) \ne v_\star \;\Big|\; G, \mathcal{A}\Big) \ge 1 - n^{-\varepsilon_1}. \tag{196}$$

由于 $\mathbb{P}(\mathcal{E}) = 1 - o(1)$，因此该结论以概率$1 - o(1)$ 关于 $(G, \mathcal{A})$ 成立。证明完毕。 □

I. 数据集

我们在合成图和真实世界图上评估所提出的编码。

在合成数据研究中，我们使用随机 3-正则图来考察逆定理定性预测的类相变行为。情况 $m = 0$ 被作为仅距离基线处理，不纳入形式化的谱论述。参数网格为

$$\begin{aligned}n &\in \{500, 1000, 2000, 4000\},\\ k &\in \{1, 2, 3, 4, 6, 8\},\\ m &\in \{0, 1, 2, 5\},\\ \eta &\in \{0.9, 0.7, 0.5, 0.3, 0.1\},\end{aligned}$$

每种配置均进行 20 次相互独立的图试验。

在真实世界研究中，我们考虑两种药物相互作用图构造。DrugBank 图由经过处理的 DrugBank 训练/测试 CSV 文件中的阳性相互作用对构建而成，随后仅保留其最大连通分量，最终得到

$$N_{\mathrm{DB}} = 1684, \quad E_{\mathrm{DB}} = 189774.$$

Decagon 派生图采用相同方式根据处理后的 Decagon 划分构建，得到

$$N_{\mathrm{DEC}} = 40, \quad E_{\mathrm{DEC}} = 466.$$

对于真实图实验，我们评估

$$\begin{aligned} k &\in \{1, 2, 3, 4, 6, 8\}, \\ m &\in \{0, 2, 5, 10\}, \\ \eta &\in \{0.9, 0.5, 0.25, 0.1\}, \end{aligned}$$

所有报告的指标均取 30 次独立随机锚点采样的平均值。

J. 基线

由于我们的目标是评估编码本身的可识别性，而非学习得到的预测器，因此基线是在特征层面定义的：

$$\begin{aligned} \text{NoPE}: &\quad F(v) = 0, \\ \text{Distance}: &\quad F(v) = d_A(v), \\ \text{Spectral}: &\quad F(v) = Q_\eta(S_m(v)), \\ \text{Full}: &\quad F(v) = \big(d_A(v), Q_\eta(S_m(v))\big). \end{aligned} \tag{197}$$

对于完整特征，我们还比较了三种锚点选择策略：无放回随机抽样、基于度的选择，以及最远点贪心策略。

K. 编码配置

这些实验不涉及端到端学习。相反，评估对象是确定性观测映射

$$F(v) = \big(d_A(v), Q_\eta(S_m(v))\big). \tag{198}$$

对于每个锚点集 $A = \{a_1, \ldots, a_k\}$，距离分量为

$$d_A(v) = \big(\mathrm{SPD}(v, a_1), \ldots, \mathrm{SPD}(v, a_k)\big). \tag{199}$$

谱分量由归一化拉普拉斯矩阵构造

$$\mathcal{L} = I - D^{-1/2} A D^{-1/2}, \tag{200}$$

使用前 m 个非平凡特征向量：

$$S_m(v) = \big(\Phi_m(v, 1)^2, \ldots, \Phi_m(v, m)^2\big). \tag{201}$$

在所有实验中，我们使用相对量化。给定谱矩阵 S_m，我们定义

$$\Delta_\eta = \eta \cdot \max_{i,j} |(S_m)_{ij}|, \tag{202}$$

并对每个坐标进行量化，方法是

$$\big(Q_\eta(x)\big)_j = \Delta_\eta \cdot \mathrm{round}\left(\frac{x_j}{\Delta_\eta}\right), \quad j \in [m]. \tag{203}$$

因此，η 越小，对应的相对量化越精细。除非另有说明，所有正文实验均采用式 (202)–(203) 中的相对量化规则。绝对量化则在本附录中单独讨论，作为校准和稳健性检查。

对于合成相变研究，结果按以下方式组织

$$\rho_{\mathrm{eng}} = \frac{k \log\log n + m \log\big(C_{\mathrm{ent}}/\eta\big)}{\log n}, \quad C_{\mathrm{ent}} = 2. \tag{204}$$

L. 评估指标

对于固定的图和参数配置，我们测量观测映射的像集大小，并定义

$$\begin{aligned} \mathrm{succ}(F) &= \frac{|\mathrm{Im}(F)|}{n}, \\ \mathrm{err}(F) &= 1 - \mathrm{succ}(F). \end{aligned} \tag{205}$$

误差越小，获得唯一结构编码的顶点就越多。

对于合成实验，我们还定义锚点经验临界数量

$$k_{\mathrm{emp}}(n, m, \eta) = \min\{k : \mathbb{E}[\mathrm{err}(F)] \le 0.1\}. \tag{206}$$

对于合成实验，每种配置的期望值通过对独立采样的 20 个图实例取平均来近似。对于真实图实验，图保持固定，重复试验对应于 30 次独立的随机锚点采样。

M. 消融实验方案

为了解每个组件的作用，我们在随机 3-正则图上进行消融实验，

$$\begin{aligned} n &\in \{500, 1000, 2000\}, \\ k &\in \{1, 2, 3, 4, 6, 8\}, \\ m &\in \{0, 2, 5, 10\}, \\ \eta &\in \{0.9, 0.5, 0.25, 0.1\}, \end{aligned}$$

使用 20 次独立图试验。

消融研究包括三个部分：比较四种特征变体、在完整特征下对（k、m、η）进行敏感性分析，以及比较三种锚点选择策略。

有关数据集、基线和超参数配置的详细信息，请参阅 https://github.com/yzz980314/Graph_Node_Identifiability_via_Observation_Maps。

N. 观测映射的完整诊断表

对于每种配置（n，m，η），我们将 k_{emp} 定义为使平均错误率不超过 0.1 的最小锚点数。我们还记录相应的理论指导比率 $\sim\rho_{\mathrm{emp}}$ 和谱码复杂度 $|Q_\eta(S_m(V))|$，其中

$$\begin{aligned} F(v) &= \big(d_A(v),\, Q_\eta(\widetilde{S}_m(v))\big), \\ \widetilde{S}_m(v) &= n\big(\phi_1(v)^2, \ldots, \phi_m(v)^2\big). \end{aligned}$$

表 VI：完整经验阈值表 k_{emp}。

n	m	$\eta=0.9$	$\eta=0.7$	$\eta=0.5$	$\eta=0.3$	$\eta=0.1$
500	0	6	6	6	6	6
	1	4	4	4	4	3
	2	4	4	3	3	2
	5	3	2	1	1	1
1000	0	6	6	6	6	6
	1	6	6	6	4	4
	2	4	4	4	3	2
	5	3	2	2	1	1
2000	0	6	6	6	6	6
	1	6	6	6	6	4
	2	6	4	4	4	3
	5	3	3	2	1	1
4000	0	6	6	6	6	6
	1	6	6	6	6	6
	2	6	6	6	4	3
	5	4	3	2	1	1

表 VII：理论指导的阈值比 ρ_{emp}。

n	m	$\eta=0.9$	$\eta=0.7$	$\eta=0.5$	$\eta=0.3$	$\eta=0.1$
500	0	1.764	1.764	1.764	1.764	1.764
	1	1.304	1.345	1.399	1.481	1.364
	2	1.433	1.514	1.328	1.492	1.552
	5	1.524	1.433	1.409	1.820	2.704
1000	0	1.679	1.679	1.679	1.679	1.679
	1	1.794	1.831	1.879	1.394	1.553
	2	1.350	1.423	1.520	1.389	1.427
	5	1.417	1.319	1.563	1.653	2.448
2000	0	1.601	1.601	1.601	1.601	1.601
	1	1.706	1.739	1.783	1.851	1.462
	2	1.811	1.344	1.432	1.567	1.589
	5	1.326	1.491	1.446	1.515	2.237
4000	0	1.530	1.530	1.530	1.530	1.530
	1	1.627	1.657	1.698	1.759	1.892
	2	1.723	1.784	1.865	1.478	1.488
	5	1.502	1.398	1.346	1.399	2.061

主要的阈值模式很简单：(i) 仅使用距离的基线表现稳定，对于所有测试过的 n 和 η，$k_{\mathrm{emp}}=6$ 均成立；(ii) 增大 m 并细化 η 通常会降低 k_{emp}；(iii) k_{emp} 的降低伴随着 $\sim$|$Q_\eta(S_m(V))$|的急剧增长。

O. 量化校准。

正文始终采用相对量化。在本附录中，我们还将相对量化与绝对量化进行比较，作为校准和稳健性检查。在测试范围内，相对量化始终近乎单射，平均误差至多为 6.7×10^{-5}，谱码比至少为 0.9933。然而，对于绝对量化，性能高度依赖于步长的校准。随着 η 从 5×10^{-3} 降至 2×10^{-4}，平均误差从 0.1686 单调下降至 10^{-4}，而占用谱码比则从 0.0593 上升至 0.9473。对于 $\eta\le10^{-4}$，绝对量化也基本变得

~表 VIII：谱码复杂度 $|Q_\eta(S_m(V))|$ 在阈值处。

n	m	$\eta=0.9$	$\eta=0.7$	$\eta=0.5$	$\eta=0.3$	$\eta=0.1$
500	0	1.0	1.0	1.0	1.0	1.0
	1	9.1	11.7	15.7	23.8	51.8
	2	31.2	46.6	68.7	117.7	285.6
	5	213.2	290.6	395.6	469.8	499.5
1000	0	1.0	1.0	1.0	1.0	1.0
	1	10.5	14.4	17.5	27.1	66.0
	2	46.2	64.5	97.2	169.3	444.2
	5	345.2	491.6	693.9	900.9	997.8
2000	0	1.0	1.0	1.0	1.0	1.0
	1	12.4	16.0	22.0	32.5	76.3
	2	59.6	78.0	128.6	240.5	692.8
	5	561.4	820.3	1196.5	1677.7	1990.1
4000	0	1.0	1.0	1.0	1.0	1.0
	1	14.4	21.1	25.1	37.1	87.7
	2	74.2	109.8	163.6	315.2	1033.6
	5	898.6	1345.2	2091.1	3123.6	3962.0

单射。这表明，定性的可识别性趋势并非仅仅依赖于相对量化的实现方式，而是在经过适当校准的绝对量化方案下同样成立。

P. 逐桶细化诊断。

图4报告了不同配置下的逐桶细化统计。对于（n，k，m，η）=（2000，4，5，0。25）这样的中间区间，误差仍为0。1521，单例节点比例仅为0。3803，桶内平均碰撞水平仍为0。2603。相比之下，在（2000，6，10，0。1）这样的近单射区域中，误差降至 10^{-4}，单例节点比例升至0。9422，桶内平均碰撞水平变得可以忽略不计（0。0039）。因此，逐桶统计在中间区间最具信息量；而一旦距离划分已经几乎完全实现区分，这些统计就基本失去实际意义。

Q. 补充稳健性细节。

为完整起见，在重复进行随机锚点重采样时，$n=1000$ 的平均图内误差范围为 0.0848，$n=2000$为 0.1189；观测到的最大范围分别为 0.1870 和 0.2120。与图间变化相比，这些波动幅度较小，不会改变总体趋势。

表 IX：绝对量化步长的校准。相对量化在整个测试范围内都保持近似单射；表中报告了绝对量化趋势的详细情况。

η	平均误差	平均原像	谱码比
5×10^{-3}	0.1686	1.2464	0.0593
2×10^{-3}	0.0528	1.0604	0.2489
10^{-3}	0.0142	1.0149	0.4908
5×10^{-4}	0.0018	1.0018	0.7354
2×10^{-4}	0.0001	1.0001	0.9473
10^{-4}	0.000025	1.000025	0.9933
5×10^{-5}	0.0000	1.0000	0.9994
2×10^{-5}	0.0000	1.0000	1.0000

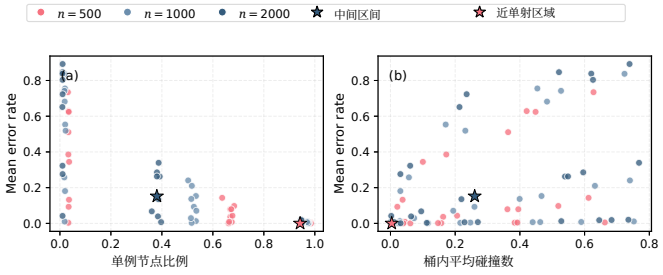


图 4：随机 3-正则图上的分桶诊断。左：平均误差与单例节点比例的关系。右：平均误差与桶内平均碰撞数的关系。高亮点展示了一个中间区间和一个近单射区域。