基于整数规划的分子设计与通过超平面划分数据集

Jianshen Zhu[1],Naveed Ahmed Azam[2, *], Kazuya Haraguchi[1], Liang Zhao[3], Hiroshi Nagamochi[1]和 Tatsuya Akutsu[4]

[1] 京都大学应用数学与物理学系，京都 606-8501，日本[2] 奎艾德阿扎姆大学数学系，伊斯兰堡 45320，巴基斯坦[3] 京都大学人类生存综合研究生院（止止馆），京都 606-8306，日本[4] 京都大学化学研究所生物信息学中心，宇治 611-0011，日本

* 通讯作者

arXiv:2305.00801v1 [cs.CE] 27 Apr 2023

摘要

近期，有人提出了一种新颖的框架，用于设计具有所需化学性质的化合物分子结构。该框架通过求解混合整数线性规划（MILP），推断出所需的化学图；该线性规划模拟了由双层模型定义的、作用于化学图的特征函数的计算过程，以及由机器学习方法构建的预测函数。为了提高该框架中预测函数的学习性能，我们设计了一种方法：利用化学空间中的一个超平面，将给定数据集 $\mathcal{C}$ 划分为两个子集 $\mathcal{C}^{(i)}$和 $i = 1, 2$，使得第一个（分别为第二个）子集中的大多数化合物的观测值低于（分别高于）阈值 θ。我们通过组合分别独立基于 $\mathcal{C}^{(i)}$ 构建的预测函数 ψ_i和 $i = 1, 2$，为数据集 $\mathcal{C}$ 构建预测函数 ψ。计算实验结果表明，对于若干难以构建优良预测函数的化学性质，所提出的方法提高了学习性能。

关键词：机器学习、整数规划、化学信息学、材料信息学、定量构效关系/定量构性关系、分子设计。

1 引言

背景 在生物信息学和机器学习的诸多应用领域中，药物设计正受到越来越多的关注 [1, 2]。因此，针对化学结构的计算分析已开展了大量研究。这类研究主要涉及两个主题：预测给定化学结构的化学性质，以及设计具有所需化学性质的化学结构。这些主题也在化学信息学领域得到了广泛研究，其中前者称为定量构效关系（QSAR） [3] ，后者称为逆定量构效关系（逆QSAR） [4, 5, 6]。

对于预测任务，统计方法和机器学习方法已得到广泛应用 [3]。在大多数此类研究中，包含两个阶段：学习阶段和预测阶段。在学习阶段，根据训练数据推导出预测函数；训练数据由若干对

化学结构及其活性（或性质）组成，其中每个化学结构都表示为一个无向图（称为化学图），然后转换为由实数组成的向量，称为特征或描述符。在预测阶段，只需将上述推导出的预测函数应用于从给定化学图获得的特征向量。在传统的定量构效关系研究中，通常使用基于回归的方法推导预测函数 [3],；而近年来，包括基于人工神经网络（ANN）的方法在内的基于机器学习的方法 [7, 8], 已得到广泛应用。需要注意的是，使用图卷积网络（GCN）时 [9], ，可以直接处理化学图，因此不一定需要将其转换为特征向量。

对于设计任务，也会使用预测函数，通常与上述情况一样从现有数据中推导得到。随后，通过预测函数 [4, 5, 6] ，根据给定的化学活性推断化学结构，其中可能会施加额外约束，以限制可能的结构。在传统方法中，这一推断任务分为两个阶段：（i）使用预测函数的逆函数，从给定的化学活性推导特征向量；（ii）根据给定的特征向量重建化学结构，这两个阶段往往会交织在一起。对于（i），通常会采用一些优化方法和/或采样方法。对于（ii），通常会采用一些枚举方法。然而，两者都是逆问题，计算起来都很困难。例如，已知可能的化学图数量庞大 [10] ，而从给定的特征向量推断化学图通常是 NP难的 [11]。因此，大多数现有方法在（i）和（ii）中都采用启发式方法，因而无法保证得到最优解或精确解。

最近，人们提出了使用人工神经网络解决设计任务的不同方法。人工神经网络的一个吸引人之处在于可以使用生成模型，其中包括自编码器和生成对抗网络。此外，如前所述，通过使用 GCN，可以直接处理化学结构 [9]。结合这些技术，或许可以在无需求解逆问题的情况下设计出新颖的化学结构 [12]。事实上，近年来人们利用各种人工神经网络模型开展了大量研究，其中包括变分自编码器 [13], 语法变分自编码器 [14], 生成对抗网络 [15], 循环神经网络 [16, 17], 和可逆流模型 [18, 19]。然而，这些方法属于启发式方法（尽管基于某些统计模型），因此无法保证解的最优性或精确性。

框架 一种用于推断化学图的新框架已经开发出来 [20, 21, 22, 23]，其基础是这样一种思路：将机器学习方法构建的预测函数的计算过程表述为混合整数线性规划（MILP）。该框架的独特之处在于，一旦预测函数被表述出来，就可以通过应用混合整数线性规划求解器精确求解逆问题。该框架包含两个主要阶段：第一阶段构建针对某种化学性质的预测函数 η ，第二阶段则基于该函数 η，推断具有该性质目标值的化学图。对于某种化学性质 π，设 $\mathcal{C}_\pi$ 为一个化学图数据集，其中每个化学图 $\mathbb{C} \in \mathcal{C}_\pi$ 的性质 π 的观测值 $a(\mathbb{C})$均已知。在第一阶段，我们针对正整数 K引入特征函数 $f\colon\ \mathcal{G} \to \mathbb{R}^K$ ，其中化学图的描述符以一种称为双层模型的特殊方式，基于局部图结构定义。随后，我们构建一个预测函数

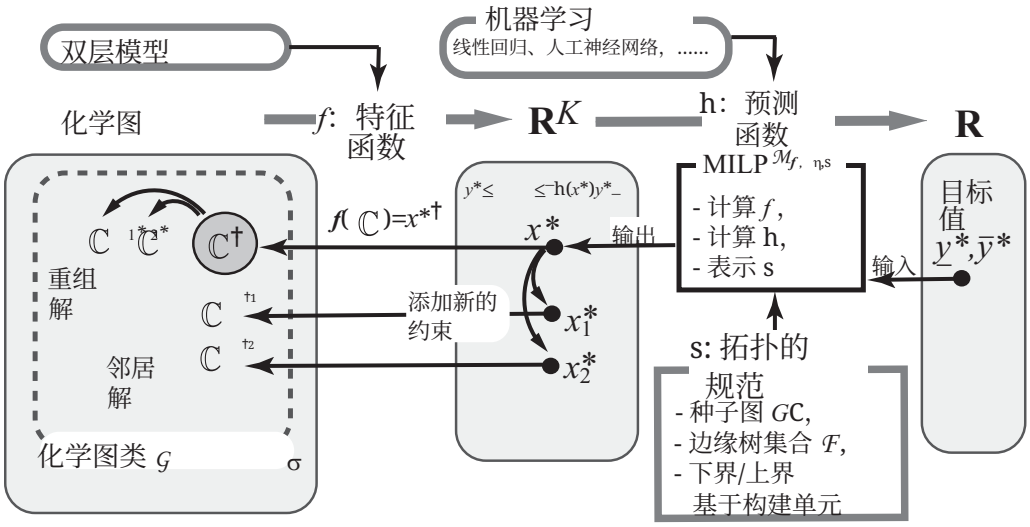


图 1：推断所需化学图的示意图 $\mathbb{C} \in \mathcal{G}_\sigma$ 借助 $\underline{y^*} \le \eta(f(\mathbb{C})) \le \overline{y^*}$。

η，采用线性回归、决策树和人工神经网络等机器学习方法，使每个 $\mathbb{C} \in \mathcal{C}_\pi$ 对应的特征向量 $x = f(\mathbb{C})$的输出 $y = \eta(x) \in \mathbb{R}$ 充当真实值 $a(\mathbb{C})$的预测值。在推断所需化学图的第二阶段，我们不仅指定性质 π 的目标化学值，还指定待推断化学图的抽象结构。后者由一组基于双层模型的规则描述，称为拓扑规范 σ；并用 $\mathcal{G}_\sigma$ 表示满足 σ中规则的所有化学图的集合。用户选择拓扑规范 σ，并选择两个实数 $\underline{y^*}$ 和$\overline{y^*}$ 作为目标化学值的区间。第二阶段的任务是推断化学图 $\mathbb{C}^* \in \mathcal{G}_\sigma$，使得$\underline{y^*} \le \eta(f(\mathbb{C}^*)) \le \overline{y^*}$ （示意见图1）。为此，我们构建一个混合整数线性规划$\mathcal{M}_{f,\eta,\sigma}$，以表示：(i) 特征函数 f中从化学图 $\mathbb{C}$ 计算 $x := f(\mathbb{C})$的过程；(ii) 预测函数 η中从向量 $x \in \mathbb{R}^K$ 计算 $y := \eta(x)$的过程；以及 (iii) 关于 $\mathbb{C} \in \mathcal{G}_\sigma$的约束。给定一个区间及其中的$\underline{y^*}$, $\overline{y^*} \in \mathbb{R}$，我们求解混合整数线性规划$\mathcal{M}_{f,\eta,\sigma}$，以找到一个特征向量 $x^* \in \mathbb{R}^K$ 和一个化学图 $\mathbb{C}^\dagger \in \mathcal{G}_\sigma$，使得 $f(\mathbb{C}^\dagger) = x^*$且 $\underline{y^*} \le \eta(x^*) \le \overline{y^*}$ （若混合整数线性规划实例不可行，则表明 $\mathcal{G}_\sigma$ 中不包含符合要求的化学图）。在第二阶段中，接下来我们基于解 $\mathbb{C}^\dagger$生成其他符合要求的化学图。为此，我们设计了以下两种方法。

第一种方法无需求解任何新的混合整数线性规划即可构造 $\mathbb{C}^\dagger$ 的异构体。在此方法中，我们首先将化学图 $\mathbb{C}^\dagger$ 分解为一组化学无环图 $T_1^\dagger, T_2^\dagger, \ldots, T_q^\dagger$，然后利用 Azam 等人 [24]提出的动态规划算法，为每棵树 $T_i^\dagger$ 构造一组异构体 $\mathcal{T}_i$，使得 $f(T_i^*) = f(T_i^\dagger)$。最后，我们为每个$i = 1$ （2, . . . , q）选取一个异构体 $T_i^* \in \mathcal{T}_i$，并将它们组装成 $\mathbb{C}^\dagger$的一个异构体 $\mathbb{C}^* \in \mathcal{G}_\sigma$，使得 $f(\mathbb{C}^*) = x^* = f(\mathbb{C}^\dagger)$。第一种方法会生成这样的异构体 $\mathbb{C}_1^*, \mathbb{C}_2^*, \ldots$，我们称之为 $\mathbb{C}^\dagger$的重组解。

第二种方法通过求解 MILP $\mathcal{M}_{f,\ \eta,\ \sigma}$ 并添加一组新的线性约束 [23]，从而构造新的解。我们首先准备任意的 p_{dim} 线性函数 $\theta_j : \mathbb{R}^K \to \mathbb{R}, j = 1, 2, \ldots, p_{\text{dim}}$，并考虑 $\mathbb{C}^\dagger$的一个邻域，该邻域由满足线性约束 $(k - 0.5)\delta \le |\theta_j(f(\mathbb{C}^*)) - \theta_j(f(\mathbb{C}^\dagger))| \le (k + 0.5)\delta, j = 1, 2, \ldots, p_{\text{dim}}$的化学图集合 $\mathbb{C}^*$ 定义，其中

一个较小的实数 $\delta > 0$ 和一个整数 $k \geq 1$。通过系统地改变整数 k，我们可以搜索 MILP $\mathcal{M}_{f,\ \eta,\ \sigma}$ 在约束 Θ 下的新解 $\mathbb{C}_1^\dagger, \mathbb{C}_2^\dagger, \ldots \in \mathcal{G}_\sigma$，使得特征向量 $x^* = f(\mathbb{C}^\dagger)$ $x_1^* = f(\mathbb{C}_1^\dagger)$ $x_2^* = f(\mathbb{C}_2^\dagger) \ldots$ 都略有不同。我们称这些化学图 $\mathbb{C}_1^\dagger, \mathbb{C}_2^\dagger, \ldots$ 为 $\mathbb{C}^\dagger$ 的邻域解。

该框架能够推断出含有 50 个非氢原子的化合物，其主要原因在于：在双层模型中，化学图的描述符是在局部图结构上定义的，因此，用于表示化学图所必需的混合整数线性规划可以构造为相当紧凑的形式，并可由标准求解器高效求解。

贡献 该框架中的描述符主要由基于双层模型的局部图结构频率组成，在该模型中，化学图 $\mathbb{C}$ 被视为内部结构和外部结构的组合（详情请参见第 3 节）。为了推导出用于推断化学图的紧凑混合整数线性规划公式，采用当前的描述符定义非常重要。然而，对于某些化学性质，使用特征函数 f 构建的预测函数的性能仍然较低。为了使用相同的双层模型提高预测函数的学习性能，我们提出了一种方法：在特征空间中使用超平面将给定数据集划分为两个子集，先分别为每个子集构建预测函数，然后通过组合这些预测函数获得针对原始数据集的预测函数 η（详情请参见第 5 节）。基于 Zhu 等人提出的相同混合整数线性规划 $\mathcal{M}_{f,\ \eta,\ \sigma}$ 公式，我们实现了该框架，以处理新提出的预测类型。根据我们对包括气味阈值在内的一些化学性质所进行的计算实验结果 [30], ，我们观察到，所提出的数据集划分与预测函数组合方法提高了这些化学性质预测函数的性能。需要指出的是，自支持向量机器发展以来，使用超平面进行预测问题的研究已经广泛开展 [25]。然而，现有方法只能应用于预测问题。我们的研究新颖且独特之处在于，开发了一种针对化学图的高效混合整数线性规划公式，从而使逆问题（即设计问题）变得易于处理。

本文组织如下。第2节介绍图和化合物建模的一些概念。第3节回顾双层模型以及该模型定义的描述符。第4节回顾通过线性回归构建的预测函数。第5节提出一种通过超平面划分数据集的方法，以及用于求解此类超平面的线性规划 formulation。第6节报告针对22种化学性质开展的计算实验结果，例如自燃温度、可燃极限和气味阈值。第7节作出一些总结性评论。技术细节见附录：附录A列出特征函数中的所有描述符；附录 B全面介绍一种拓扑规范；附录 C详细说明计算实验中使用的测试实例。

2 预备知识

本节介绍有关图、化合物建模以及我们所选描述符的一些概念和术语。

令 $\mathbb{R}$、$\mathbb{R}_+$、$\mathbb{Z}$ 和 $\mathbb{Z}_+$ 分别表示实数集、非负实数集、整数集和非负整数集。对于两个整数 a 和 b，令 $[a,b]$ 表示满足 $a \le i \le b$ 的整数 i 所构成的集合。对于向量 $x \in \mathbb{R}^p$，其第 j 个分量记为 $x(j)$。

图 给定一个图 G，令 $V(G)$ 和 $E(G)$ 分别表示顶点集和边集。对于图 G 的子集 $V' \subseteq V(G)$（分别为 $E' \subseteq E(G)$），令 $G - V'$（分别为 $G - E'$）表示从 G 中删除顶点 V'（分别删除边 E'）后得到的图；删除 V' 中顶点关联的所有边后，得到 $G - V'$。端点为 u 和 v 的路径称为 u,v-路径。

我们将有根图定义为带有一个指定顶点的图，该顶点称为根。对于可能带有根的图 G，叶顶点定义为度为 1 的非根顶点。称与叶顶点 v 相邻的边 uv 为叶边，并分别用 $V_{\text{leaf}}(G)$ 和 $E_{\text{leaf}}(G)$ 表示 G 中叶顶点和叶边的集合。对于图或有根图 G，我们定义图 $G_i, i \in \mathbb{Z}_+$，它们由 G 删除叶顶点集 i 次后得到，使得

$$G_0 := G; \quad G_{i+1} := G_i - V_{\text{leaf}}(G_i),$$

我们称顶点 v 为树顶点，若对于某个 $i \ge 0$，有 $v \in V_{\text{leaf}}(G_i)$。定义每个树顶点 $v \in V_{\text{leaf}}(G_i)$ 的高度 $\mathsf{ht}(v)$ 为 i；对于与树顶点相邻的每个非树顶点 v，定义其高度 $\mathsf{ht}(v)$ 为 $\mathsf{ht}(u)+1$，其中取与 v 相邻的树顶点 u 的 $\mathsf{ht}(u)$ 最大值；对于不与任何树顶点相邻的非树顶点，则不定义其高度。我们称满足 $\mathsf{ht}(v) = k$ 的顶点 v 为叶 k-分支。有根树 T 的高度 $\mathsf{ht}(T)$ 定义为顶点 $v \in V(T)$ 的 $\mathsf{ht}(v)$ 的最大值。

2.1 化合物建模

我们回顾 Zhu 等人提出的一种化合物建模方法 [21]。

为表示一种化合物，我们引入一组化学元素，例如 H（氢）、C（碳）、O（氧）、N（氮）等。为区分具有多种化合价的化学元素 a，例如 S（硫），我们将化合价为 i 的化学元素 a 记为 $\mathtt{a}_{(i)}$；对于只有一种化合价的化学元素，则不使用此类后缀（i）。设 Λ 为一组化学元素 $\mathtt{a}_{(i)}$。例如，$\Lambda = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{P}, \mathtt{S}_{(2)}, \mathtt{S}_{(4)}, \mathtt{S}_{(6)}\}$。设 $\mathsf{val}\colon \Lambda \to [1,6]$ 为一个化合价函数。例如，$\mathsf{val}(\mathtt{H}) = 1$，$\mathsf{val}(\mathtt{C}) = 4$，$\mathsf{val}(\mathtt{O}) = 2$，$\mathsf{val}(\mathtt{P}) = 5$，$\mathsf{val}(\mathtt{S}_{(2)}) = 2$，$\mathsf{val}(\mathtt{S}_{(4)}) = 4$，且 $\mathsf{val}(\mathtt{S}_{(6)}) = 6$。对于每种化学元素 $\mathtt{a} \in \Lambda$，令 $\mathsf{mass}(\mathtt{a})$ 表示 a 的质量。

化合物用化学图表示，化学图定义为一个元组 $\mathbb{C} = (H, \alpha, \beta)$，由简单连通无向图 H 及函数 $\alpha\colon V(H) \to \Lambda$ 和 $\beta\colon E(H) \to [1,3]$ 构成。化合物中的原子集和化学键集分别由顶点集 $V(H)$ 和边集 $E(H)$ 表示。分配给顶点 $v \in V(H)$ 的化学元素由$\alpha(v)$ 表示，两个相邻顶点 $u, v \in V(H)$ 之间的键多重性由边 $e = uv \in E(H)$ 的 $\beta(e)$ 表示。若两个元组 (H_i, α_i, β_i) $i = 1, 2$ 存在同构 ϕ，即存在双射 $\phi\colon V(H_1) \to V(H_2)$，满足 $uv \in E(H_1)$ $\alpha_1(u) = \mathtt{a}, \alpha_1(v) = \mathtt{b}, \beta_1(uv) = m \leftrightarrow \phi(u)\phi(v) \in E(H_2)$ $\alpha_2(\phi(u)) = \mathtt{a}, \alpha_2(\phi(v)) = \mathtt{b}, \beta_2(\phi(u)\phi(v)) = m$，则称它们同构。当 H_i

以顶点 $r_i, i = 1, 2$ 为根时，若化学图 (H_i, α_i, β_i) $i = 1, 2$ 存在同构 ϕ，且满足 $\phi(r_1) = r_2$，则称它们根同构（r-同构）。

为方便记号，对于化学图$\mathbb{C} = (H, \alpha, \beta)$，我们使用函数 $\beta_{\mathbb{C}}$：$V(H) \to [0, 12]$，其中 $\beta_{\mathbb{C}}(u)$ 表示与顶点u相 incident 的边的键重数之和；即，

$$\beta_{\mathbb{C}}(u) \triangleq \sum_{uv \in E(H)} \beta(uv) \text{ for each vertex } u \in V(H).$$

对于每个顶点 $u \in V(H)$，定义电子度 $\mathrm{eledeg}_{\mathbb{C}}(u)$ 为

$$\mathrm{eledeg}_{\mathbb{C}}(u) \triangleq \beta_{\mathbb{C}}(u) - \mathrm{val}(\alpha(u)).$$

对于每个顶点 $u \in V(H)$，令 $\deg_{\mathbb{C}}(v)$ 表示在 $\mathbb{C}$中与 u 相邻的顶点数。

对于化学图 $\mathbb{C} = (H, \alpha, \beta)$，令 $V_{\mathtt{a}}(\mathbb{C})$，$\mathtt{a} \in \Lambda$ 表示满足 $\alpha(v) = \mathtt{a}$ 属于 $\mathbb{C}$ 的顶点 $v \in V(H)$ 的集合，并将氢抑制化学图 $\langle \mathbb{C} \rangle$ 定义为从 H 中移除所有顶点 $v \in V_{\mathtt{H}}(\mathbb{C})$后得到的图。

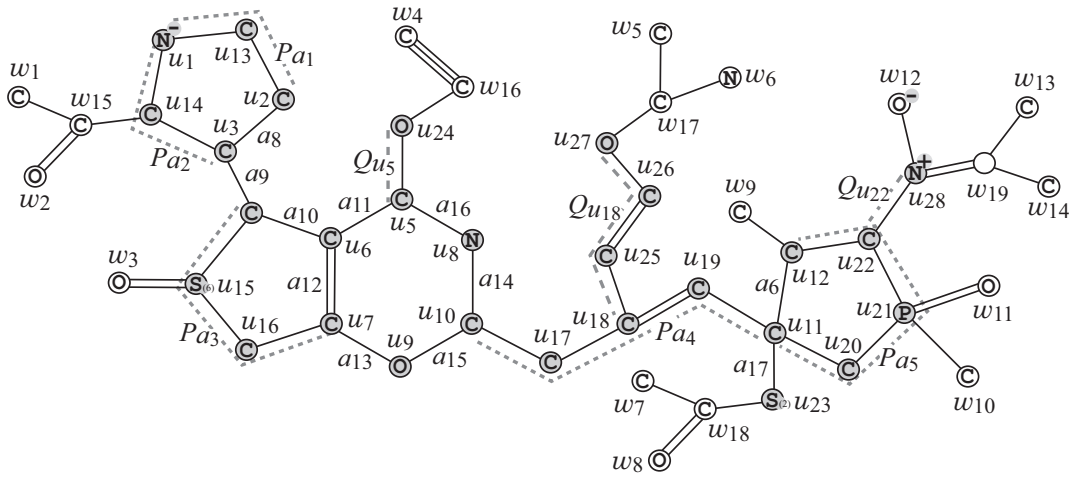


图2：从化学图 $\mathbb{C}$ 中移除所有氢原子后得到的氢抑制化学图 $\langle \mathbb{C} \rangle$ 的示意图，其中对于 $\rho = 2$，$V_{\text{外部}}(\mathbb{C}) = \{w_i \mid i \in [1, 19]\}$ 且$V_{\text{内部}}(\mathbb{C}) = \{u_i \mid i \in [1, 28]\}$。

3 双层模型

本节回顾 Shi 等人提出的双层模型 [20]。

设 $\mathbb{C} = (H, \alpha, \beta)$ 为一个化学图，$\rho \geq 1$ 为一个整数，我们称之为分支参数。

$\mathbb{C}$ 的双层模型是将氢抑制化学图 $\langle \mathbb{C} \rangle$ 按以下方式划分为“内部”和“外部”。如果图 $\mathbb{C}$ 中的顶点 $v \in V$（$\langle \mathbb{C} \rangle$）满足 $\mathrm{ht}(v) < \rho$（分别地，边$e \in E$（$\langle \mathbb{C} \rangle$）与外部顶点相 incident），则称其为外部顶点（分别地，外部边），并将外部顶点集和外部边集分别记为 V^{ex}（$\mathbb{C}$）和 E^{ex}（$\mathbb{C}$），

分别记 $V^{\text{int}}(\mathbb{C}) = V(\langle\mathbb{C}\rangle) \setminus V_{\text{ex}}(\mathbb{C})$ 和 $E^{\text{int}}(\mathbb{C}) = E(\langle\mathbb{C}\rangle) \setminus E^{\text{ex}}(\mathbb{C})$。我们称 $V^{\text{int}}(\mathbb{C})$ 中的顶点（分别地，$E^{\text{int}}(\mathbb{C})$ 中的边）为内部顶点（分别地，内部边）。外部边集 $E^{\text{ex}}(\mathbb{C})$ 由若干连通图构成，每个图均视为一棵有根树 T，其根为 ht(v) 最大的顶点 $v \in V(T)$。用 $\mathcal{T}_{\text{ex}}(\langle\mathbb{C}\rangle)$ 表示 $\langle\mathbb{C}\rangle$中这些化学根树的集合。$\mathbb{C}$ 的内部 $\mathbb{C}^{\text{int}}$ 定义为 $\langle\mathbb{C}\rangle$的子图（$V_{\text{int}}(\mathbb{C})$ $E^{\text{int}}(\mathbb{C})$）。

图2展示了一个氢抑制化学图 $\langle\mathbb{C}\rangle$的例子。对于枝参数 $\rho = 2$，图2中的化学图 $\langle\mathbb{C}\rangle$ 的内部通过移除 1 $\rho = 2$ 次顶点集合得到；也就是说，先移除 $\langle\mathbb{C}\rangle$ 中度为1的顶点集合 $V_1 = \{w_1, w_2, \ldots, w_{14}\}$，再移除 $\langle\mathbb{C}\rangle - V_1$中度为1的顶点集合 $V_2 = \{w_{15}, w_{16}, \ldots, w_{19}\}$，其中被移除的顶点成为 $\langle\mathbb{C}\rangle$的外部顶点。

对于每个内部顶点 $u \in V^{\text{int}}(\mathbb{C})$，令 $T_u \in \mathcal{T}_{\text{外部}}(\langle\mathbb{C}\rangle)$ 表示以 u为根的化学树（其中 T_u 可能由顶点 u组成），并将 $\mathbb{C}[u]$ 定义为 T_u 中原本与 $\mathbb{C}$里的 T_u 相连的氢原子放回后得到的化学根树。令 $\mathcal{T}(\mathbb{C})$ 表示 $\mathbb{C}[u]$、$u \in V_{\text{内部}}(\mathbb{C})$ 中的 ρ枝树集合。图3展示了图2中示例 $\mathbb{C}$ 的2-边缘树集合 $\mathcal{T}(\mathbb{C}) = \{\mathbb{C}[u_i] \mid i \in [1, 28]\}$。

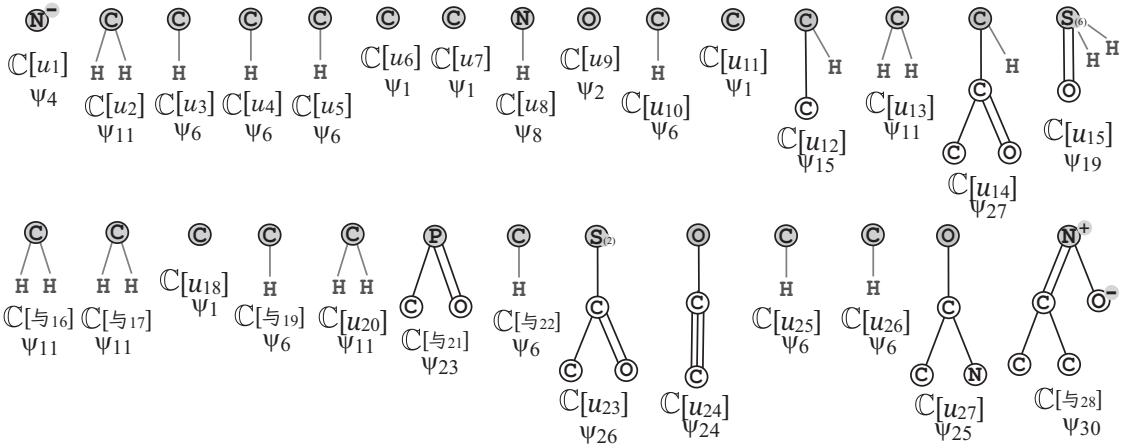


图3：图2中示例的 2-边缘树集合 $\mathcal{T}(\mathbb{C})$ $\mathbb{C}[u_i], i \in [1, 28]$，其中每棵树的根均以灰色圆圈表示，图中省略了连接到非根顶点的氢原子。

特征函数 满足 $\alpha(u) = \mathtt{a}$、$\deg^{\langle\mathbb{C}\rangle}(u) = d$、$\alpha(v) = \mathtt{b}$、$\deg^{\langle\mathbb{C}\rangle}(v) = d'$ 且 $\beta(e) = m$ 的内部边 $e = uv \in E^{\text{int}}(\mathbb{C})$的特征由元组$(\mathtt{a}d, \mathtt{b}d', m)$表示，称为边 e的边构型，其中我们称元组$(\mathtt{a}, \mathtt{b}, m)$为边 e的邻接构型。

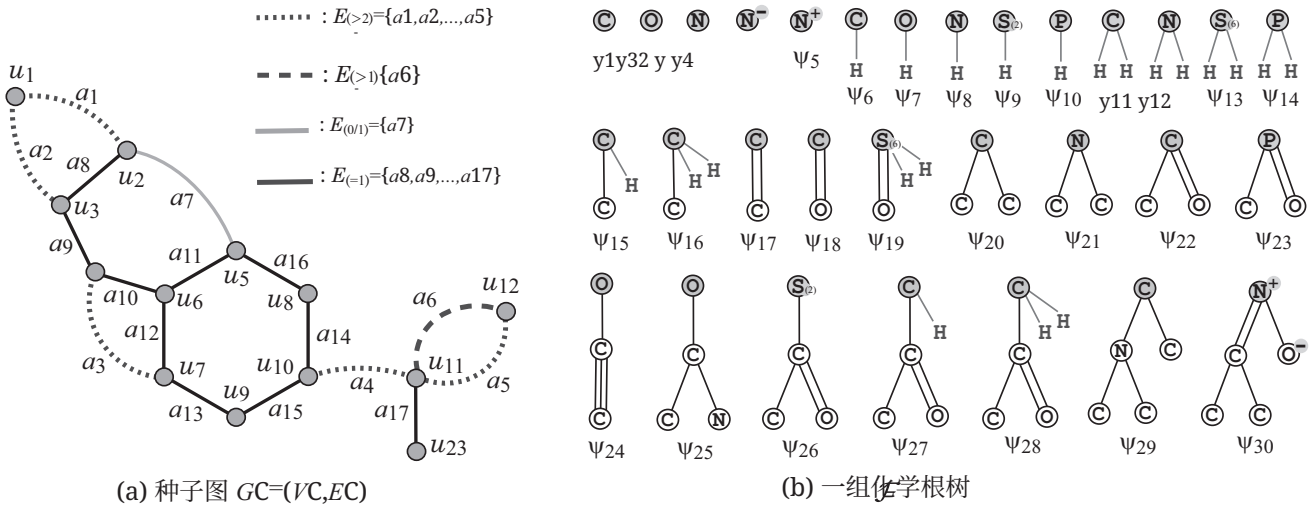
在采用双层模型的框架中，特征向量 f 主要由内部边的边构型频数，以及所有化学根树集合 $\mathbb{C}[u]$ 中各内部顶点 u上的化学根树频数构成。所有这些描述符见附录A $x(1)$ $x(2)$ . . ., $x(K_1)$，它们称为线性描述符。我们将针对某一性质 π在数据集上构建的描述符集合记为 $D_\pi^{(1)}$:= $\{x(k) \mid k \in [1, K_1]\}$。Zhu

等人 [28][1] 引入了一个二次项 $x(i)x(j)$（或 $x(i)(1-x(j))$），$1 \le i \le j \le K_1$ 作为一种新的描述符，其中假设每个 $x(i)$都已归一化到 0 到 1 之间。该项 $x(i)x(j)$,$1 \le i \le j \le K_1$（或 $x(i)(1-x(j))$ $i,j \in [1,K_1]$）称为二次描述符，记作 $D_\pi^{(2)}:=\{x(i)x(j) \mid 1 \le i \le j \le K_1\} \cup \{x(i)(1-x(j)) \mid i,j \in [1,K_1]\}$ 即二次描述符集合。

为了构造预测函数，我们使用并集 $D_\pi^{(1)} \cup D_\pi^{(2)}$。在构造预测函数时，这组描述符通常过多，因此我们将其缩减为较小的一组描述符，以构造特征函数 f： $\mathbb{R}^K \to \mathbb{R}$，其中 K 是所得描述符的数量。我们将 $\mathbb{R}^K$ 称为特征空间。为缩减描述符，我们采用 Zhu 等人 [28]提出的方法。

拓扑规范 一项拓扑规范 σ 被描述为以下规则的集合:

(i) 一个种子图 G_{C}，作为目标化学图 $\mathbb{C}$的抽象形式；(ii) 一组 $\mathcal{F}$ 化学根树，作为候选项，用于构成以 $\mathbb{C}$ 中每个外部顶点u为根的树 $\mathbb{C}[u]$；以及 (iii) 目标化学图中连通分量数量的下界和上界，例如化学元素、双键/三键以及 $\mathbb{C}$ 中内部顶点的数量。



(a) 种子图 GC=(VC,EC)　　(b) 一组化学根树

图4: （a）种子图 G_{C} 的示意图，其中 V_{C} 中的顶点以灰色圆表示，$E_{(\ge 2)}$ 中的边以点线表示，$E_{(\ge 1)}$ 中的边以虚线表示，$E_{(0/1)}$ 中的边以灰色粗线表示，$E_{(=1)}$ 中的边以黑色实线表示；（b）包含30棵化学根树 ψ_i、$i \in [1,30]$ 的集合 $\mathcal{F}=\{\psi_1,\psi_2,\ldots,\psi_{30}\} \subseteq \mathcal{F}(\mathcal{C}_\pi)$，其中每棵树的根均以灰色圆表示。图中省略了连接到非根顶点的氢原子。

图 4(a) 和 (b) 分别展示了一个种子图 G_{C} 和一组化学根树 $\mathcal{F}$ 的示例。给定一个种子图 G_{C}，目标化学图 $\mathbb{C}$ 的内部由 G_{C} 构造，即将一些边 $a=uv$ 替换为端点 u 和 v 之间的路径 P_a，并

[1]文章的完整版本可在 https://arxiv.org/abs/2209.13527 获取

通过向某些顶点 v连接新的路径 Q_v 来实现。例如，图2中的化学图 $\mathbb{C}$ 由图4(a)中的种子图 G_{C} 按如下方式构造。

- 首先，将G_{C} 中的五条边 $a_1 = u_1u_2$、$a_2 = u_1u_3$、$a_3 = u_4u_7$、$a_4 = u_{10}u_{11}$和 $a_5 = u_{11}u_{12}$分别替换为新路径 $P_{a_1} = (u_1、u_{13}、u_2)$、$P_{a_2} = (u_1、u_{14}、u_3)$、$P_{a_3} = (u_4、u_{15}、u_{16}、u_7)$、$P_{a_4} = (u_{10}、u_{17}、u_{18}、u_{19}、u_{11})$ 和 $P_{a_5} = (u_{11}、u_{20}、u_{21}、u_{22}、u_{12})$，从而得到 $\langle\mathbb{C}\rangle$的一个子图 G_1。

- 接下来，将三条新路径 $Q_{u_5} = (u_5、u_{24})$、$Q_{u_{18}} = (u_{18}、u_{25}、u_{26}、u_{27})$和$Q_{u_{22}} = (u_{22}、u_{28})$连接到该图 G_1 上，从而得到图2中 $\langle\mathbb{C}\rangle$ 的内部。
- 最后，将从集合 $\mathcal{F}$ 中选出的 28 棵树附加到内部，并为内部指定化学元素和键重数，从而得到图2中的化学图 $\mathbb{C}$ 。在图3中，$\psi_1 \in \mathcal{F}$ 被选作 $\mathbb{C}[u_i]$、$i \in \{6, 7, 11\}$。类似地，ψ_2 对应 $\mathbb{C}[u_9]$，ψ_4 对应 $\mathbb{C}[u_1]$，ψ_6 对应 $\mathbb{C}[u_i]$，$i \in \{3, 4, 5, 10, 19, 22, 25, 26\}$，$\psi_8$对应 $\mathbb{C}[u_8]$，ψ_{11}对应 $\mathbb{C}[u_i]$，$i \in \{2, 13, 16, 17, 20\}$，$\psi_{15}$对应 $\mathbb{C}[u_{12}]$，ψ_{19} 对应 $\mathbb{C}[u_{15}]$，ψ_{23} 对应 $\mathbb{C}[u_{21}]$，ψ_{24} 对应 $\mathbb{C}[u_{24}]$，ψ_{25} 对应 $\mathbb{C}[u_{27}]$，ψ_{26} 对应 $\mathbb{C}[u_{23}]$，ψ_{27} 对应 $\mathbb{C}[u_{14}]$ ，以及ψ_{30} 对应 $\mathbb{C}[u_{28}]$。

有关拓扑规范的完整说明，请参见附录 B。

4 预测函数

设 $\mathcal{C}$ 为一组化学图 $\mathbb{C}$ ，其观测值为 $a(\mathbb{C}) \in \mathbb{R}$。设 D 为一组描述符，其中 $K = |D|$ ； f 为一个特征函数，它将化学图 $\mathbb{C}$ 映射为向量$f(\mathbb{C}) \in \mathbb{R}^K$，其中 $x(d)$表示描述符 $d \in D$的值。为简化记号，对于带索引的图 $\mathbb{C}_i \in \mathcal{C}$，我们记 $a_i = a(\mathbb{C}_i)$ 和 $x_i = f(\mathbb{C}_i)$。

4.1 评估

对于预测函数 $\eta : \mathbb{R}^K \to \mathbb{R}$，定义一个误差函数

$$\mathrm{Err}(\eta; \mathcal{C}) \triangleq \sum_{\mathbb{C}_i \in \mathcal{C}} (a_i - \eta(f(\mathbb{C}_i)))^2 = \sum_{\mathbb{C}_i \in \mathcal{C}} (a_i - \eta(x_i))^2,$$

并定义决定系数 $\mathrm{R}^2(\eta, \mathcal{C})$ 为

$$\mathrm{R}^2(\eta, \mathcal{C}) \triangleq 1 - \frac{\mathrm{Err}(\eta; \mathcal{C})}{\sum_{\mathbb{C}_i \in \mathcal{C}} (a_i - \widetilde{a})^2} \text{ for } \widetilde{a} = \frac{1}{|\mathcal{C}|} \sum_{\mathbb{C} \in \mathcal{C}} a(\mathbb{C}).$$

我们通过五折交叉验证来评估一种基于一组 D 描述符构建预测函数的方法，具体如下。一次 r 五折交叉验证运行按以下步骤执行：将一个数据集 $\mathcal{C}$ 随机划分为五个子集 $\mathcal{C}^{(k)}$、$k \in [1, 5]$ ，使得$|\mathcal{C}^{(i)}|$ 与 $|\mathcal{C}^{(j)}|$ 之差至多为 1。对于每个 $k \in [1, 5]$，令 $\mathcal{C}_{\text{train}} := \mathcal{C} \setminus \mathcal{C}^{(k)}$、$\mathcal{C}_{\text{test}} := \mathcal{C}^{(k)}$ ，并执行该方法，在一个训练集 $\mathcal{C}_{\text{train}}$ 上构建预测函数 $\eta^{(k)} : \mathrm{R}^{|D|} \to \mathbb{R}$ ，然后计算$g_r^{(k)} := \mathrm{R}^2(\eta^{(k)}, \mathcal{C}_{\text{test}})$。令 $\mathrm{R}^2_{\mathrm{CV}}(\mathcal{C}, D, p)$ 表示 $\{g_{r_i}^{(k)} \mid k \in [1, 5], i \in [1, p]\}$ 次 p 五折交叉验证运行r_1、r_2、...、r_p结果的中位数。

4.2 线性回归

对于一组 D 描述符，超平面定义为一对 (w,b)，其中包含一个向量 $w \in \mathbb{R}^K$ 和一个实数 $b \in \mathbb{R}$。给定超平面 (w,b)，通过设定以下内容来定义预测函数$\eta_{w,\ b}$：$\mathbb{R}^K \to \mathbb{R}$

$$\eta_{w,b}(x) \triangleq w \cdot x + b = \sum_{d \in D} w(d)x(d) + b.$$

多维线性回归（MLR）给定一个数据集 $\mathcal{C}$ 和一组描述符 D，多维线性回归 MLR($\mathcal{C}$, D) 返回一个超平面 (w,b)，其带有 $w \in \mathbb{R}^K$，并使 Err($\eta_{w,\ b}$; $\mathcal{C}$) 最小化。然而，这样的超平面 (w,b) 可能包含过多非零实数 $w(d)$。为避免这一情况，有人提出在误差函数中加入额外惩罚项 τ 进行最小化。其中，套索函数 [26] 定义为

$$\frac{1}{2|\mathcal{C}|}\mathrm{Err}(\eta_{w,b};\mathcal{C}) + \lambda\tau, \quad \tau = \sum_{d \in D} |w(d)| + |b|,$$

其中 $\lambda \in \mathbb{R}_+$ 是一个给定的非负数。

调整线性回归（ALR）我们介绍一种近期的学习方法，称为调整线性回归。该方法通过线性回归实现，其效果等同于没有隐藏层的人工神经网络，其中每个输入节点都可以具有非线性激活函数（有关该思想的详情，见 [27]）。令 $\mathcal{C} = \{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_m\}$, $A = \{a_i = f(\mathbb{C}_i) \mid i \in [1,m]\}$ 和$X = \{x_i = f(\mathbb{C}_i) \in \mathbb{R}^K \mid i \in [1,m]\}$。令 D^+（分别为 D^-）表示描述符 $d \in D$ 的集合，其中相关系数 $\sigma(X[d], A)$ 在 $X[d] = \{x_i(d) \mid i \in [1,m]\}$ 与 A 之间为非负（分别为负）。我们首先求解以下线性规划，其中包含常数 $\lambda \geq 0$、实变量b和非负实变量 $c_q(0)$ $q \in [0,2]$, $w_q(d)$ $q \in [0,2], d \in D$。

线性规划

$$\begin{aligned}\text{Minimize: } \frac{1}{2m}\sum_{i\in[1,m]} \Big| & c_0(0)a_i + c_1(0)a_i^2 + c_2(0)(1-(a_i-1)^2) \\ & - \sum_{d\in D^+}[w_0(d)x_i(d) + w_1(d)x_i(d)^2 + w_2(d)(1-(x_i(d)-1)^2)] \\ & + \sum_{d\in D^-}[w_0(d)x_i(d) + w_1(d)x_i(d)^2 + w_2(d)(1-(x_i(d)-1)^2)] - b\Big| + \lambda\tau\end{aligned}$$

满足约束条件

$$\tau = \sum_{d\in D} w_0(d) + |b|, \quad c_0(0) + c_1(0) + c_2(0) = 1. \tag{1}$$

该最小化问题的最优解可通过求解含有 $O(m+|D|)$ 个变量和约束的线性规划得到。根据最优解，我们接下来计算如下超平面（w^*, b^*），以获得线性预测函数 $\eta_{w^*,\ b^*}$。令 $c_q^*(0)$ $q \in [0,2]$、$w_q^*(d)$ $q \in [0,2]$、$d \in D$ 和 b^* 分别表示最优解中变量 $c_q(0)$ $q \in [0,2]$、$w_q(d)$ $q \in [0,2]$、$d \in D$ 和 b 的取值。令 $D^\dagger$ 表示满足 $w_0^*(d) > 0$ 的描述符 $d \in D$ 集合。接着，对于满足 $w_0^*(d) = 0$ 的 $d \in D$，我们设定 $w^*(d) := 0$

$w^*(d) := w_0^*(d)/(w_0^*(d) + w_1^*(d) + w_2^*(d))$，适用于 $d \in D^+ \cap D^\dagger$；$w^*(d) := -w_0^*(d)/(w_0^*(d) + w_1^*(d) + w_2^*(d))$，适用于 $d \in D^- \cap D^\dagger$；以及$w^* := (w_0^*(1)\ w_0^*(2) \ldots, w_0^*(|D|)) \in \mathbb{R}^K$。

描述符约简与线性回归（RLR）最后，我们回顾 Zhu 等人最近提出的一种学习方法 [28]，该方法旨在提升双层模型的学习性能。给定一组描述符 $x(1)\ x(2) \ldots, x(K)$，该方法首先在原有线性描述符集合中，为每两个描述符添加一个二次描述符 $x(i)x(j)$（或 $x(i)(1 - x(j))$）。这会大幅增加描述符的数量，从而增加学习所需的运行时间，或导致对数据集过拟合。接下来，该方法将线性描述符和二次描述符集合约简为一个更小的集合，以生成性能更高的预测函数（约简过程的详细信息见 [28]）。最后，该方法使用多元线性回归（MLR），基于选出的描述符构建预测函数。本文将这种基于约简和线性回归的方法称为 RLR。

5 通过超平面划分数据集

本节提出一种方法，利用特征空间中的超平面将给定数据集划分为两个子集，使得第一个（分别为第二个）子集中的大多数化合物 $\mathbb{C}$ 的观测值 a（$\mathbb{C}$）小于（分别为大于）阈值 θ。

对于性质 π，设 $\mathcal{C} = \{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_n\}$ 为一组化学图。假设所有条目和观测值均已归一化，其中 $\min\{a_i \mid \mathbb{C}_i \in \mathcal{C}\} = 0$ 和 $\max\{a_i \mid \mathbb{C}_i \in \mathcal{C}\} = 1$，且对于每个 $d \in D$，$\min\{x_i(d) \mid \mathbb{C}_i \in \mathcal{C}\} = 0$ 和 $\max\{x_i(d) \mid \mathbb{C}_i \in \mathcal{C}\} = 1$。

对于满足 $0 < \theta < 1$ 的阈值 θ，我们找到一个超平面（w，b），满足 $w \in \mathbb{R}^K$ 和 $b \in \mathbb{R}$，将集合 $\mathcal{C}$ 划分为子集

$$\mathcal{C}^{(1)} := \{\mathbb{C}_i \in \mathcal{C} \mid wx_i - b \leq 0\} \text{ and } \mathcal{C}^{(2)} := \{\mathbb{C}_i \in \mathcal{C} \mid wx_i - b > 0\}$$

使得 $\mathcal{C}^{(1)}$（分别为 $\mathcal{C}^{(2)}$）包含尽可能多的具有 $a_i \leq \theta$（分别为 $a_i > \theta$）的化合物 $\mathbb{C}_i \in \mathcal{C}$。然后，我们将每个子集 $\mathcal{C}^{(j)}$、$j = 1$、2 视为新的数据集，并构建预测函数ψ_j，再通过组合函数 ψ_1 和 ψ_2，得到针对原始集合 $\mathcal{C}$ 的预测函数。

用于寻找超平面的线性规划公式要寻找一个超平面来划分给定的数据集，我们按照 Freed and Glover [29] 提出的、用于分离分类数据的思路，构建一个线性规划。定义集合 $\mathcal{C}_{\leq\theta} := \{\mathbb{C}_i \in \mathcal{C} \mid a_i \leq \theta\}$ 和 $\mathcal{C}_{>\theta} := \{\mathbb{C}_i \in \mathcal{C} \mid a_i > \theta\}$，并选择具有 $a_s = 0$ 的化合物 $\mathbb{C}_s \in \mathcal{C}_{\leq\theta}$ 以及具有 $a_t = 1$ 的化合物 $\mathbb{C}_t \in \mathcal{C}_{>\theta}$，使得 $\mathbb{C}_s \in \mathcal{C}^{(1)}$,$\mathbb{C}_t \in \mathcal{C}^{(2)}$。

线性规划的表述如下，其中超平面 (w, b) 及其 $w \in \mathbb{R}^K$ 和 $b \in \mathbb{R}$作为该线性规划的一个最优解而得到：

LP(θ)
常数：$a_i \in \mathbb{R}$, $x_i \in \mathbb{R}^K$, $\forall \mathbb{C}_i \in \mathcal{C}$；索引 s 和 t，满足 $a_s = 0$ 和 $a_t = 1$；$\theta \in \mathbb{R}$；

变量： $w \in \mathbb{R}^K, b \in \mathbb{R}$，非负变量 $\delta_i \geq 0, \forall \mathbb{C}_i \in \mathcal{C}$；约束：

$$wx_s - b \leq 0,$$
$$wx_t - b \geq 0,$$
$$\delta_i \geq wx_i - b + (a_i - \theta)^2, \quad \forall \mathbb{C}_i \in \mathcal{C}_{\leq\theta},$$
$$\delta_i \geq -(wx_i - b) + (a_i - \theta)^2, \quad \forall \mathbb{C}_i \in \mathcal{C}_{>\theta},$$

目标函数：

$$\text{minimize} \quad \sum_{\mathbb{C}_i \in \mathcal{C}} \delta_i.$$

上述线性规划包含 $O(|D| + |\mathcal{C}|)$ 个变量和 $O(|\mathcal{C}|)$ 个约束。我们求解该线性规划以获得一个最优解 (w, b)，并计算 $\mathcal{C}^{(1)} = \{\mathbb{C}_i \in \mathcal{C} \mid wx_i - b \leq 0\}$和 $\mathcal{C}^{(2)} = \{\mathbb{C}_i \in \mathcal{C} \mid wx_i - b > 0\}$。记

$$a^{(j)}_{\min} := \min\{a_i \mid \mathbb{C}_i \in \mathcal{C}^{(j)}\} \text{ and } a^{(j)}_{\max} := \max\{a_i \mid \mathbb{C}_i \in \mathcal{C}^{(j)}\}, j = 1, 2,$$

其中 $a^{(1)}{}_{\text{最小}} = 0$ 和 $a^{(2)}_{\max} = 1$。当 $\mathcal{C}^{(1)} = \mathcal{C}_{\leq\theta}$ 和 $\mathcal{C}^{(2)} = \mathcal{C}_{>\theta}$ 成立时，范围 $[a^{(1)}{}_{\text{最小}} = 0, a^{(1)}_{\max}]$和 $[a^{(2)}{}_{\text{最小}} a^{(2)}_{\max} = 1]$ 互不重叠（即 $a^{(1)}_{\max} \leq \theta < a^{(2)}{}_{\text{最小}}$）。否则，$a^{(2)}{}_{\text{最小}} \leq a^{(1)}_{\max}$ 成立；即使在这种情况下，如果 $a^{(2)}{}_{\text{最小}}$ 和 $a^{(1)}_{\max}$ 非常接近，两个子集 $\mathcal{C}^{(1)}$ 和 $\mathcal{C}^{(2)}$ 仍然良好分离。我们从候选项集合中选择一个阈值 θ，使得在满足 $|\mathcal{C}^{(1)}|$ 和 $|\mathcal{C}^{(2)}|$ 各自的大小接近原始大小 $|\mathcal{C}|$的一半这一条件下，$a^{(1)}_{\max} - a^{(2)}{}_{\text{最小}}$ 达到最小值。

框架第一阶段的实现 在框架的第一阶段，我们为性质 π 和特征函数 f 的描述符集 D，针对数据集 $\mathcal{C}$ 构造预测函数 ψ： $\mathcal{C} \to \mathbb{R}^K$，具体如下。对于选定的阈值 θ，我们求得一个超平面 (w, b)，$w \in \mathbb{R}^K, b \in \mathbb{R}$，作为上述线性规划 LP($\theta$) 的最优解，并据此将 $\mathcal{C}$ 划分为子集 $\mathcal{C}^{(j)}, j = 1, 2$。对于每个 $j = 1, 2$，选择一组描述符 $\widetilde{D}_j$，并为描述符集为 $\widetilde{D}_j$ 的数据集 $\mathcal{C}^{(j)}$ 构造预测函数 ψ_j： $\mathbb{R}^{|\widetilde{D}_j|} \to \mathbb{R}$。在计算实验中，我们使用 LLR、人工神经网络、ALR 和 RLR 为 $\mathcal{C}^{(j)}, j = 1, 2$ 构造预测函数，并从中选择学习性能最佳的一个作为 ψ_j（当使用 LLR、人工神经网络或 ALR 时，$\widetilde{D}_j$ 是线性描述符集 D 的一个 $\widetilde{}$ 子集；当使用 RLR 构造 ψ_j 时，D_j 由 D 的一些线性和二次描述符组成）。给定一个特征向量 $x \in \mathbb{R}^K$，若 $wx - b \leq 0$，则使用预测函数 ψ_1；否则使用预测函数 ψ_2。因此，预测函数 ψ 定义为

$$\psi(x) := \begin{cases} \psi_1(x) & \text{if } wx - b \leq 0, \\ \psi_2(x) & \text{otherwise,} \end{cases}$$

其中，超平面 (w, b) 是预测函数 ψ 的一部分。

框架第二阶段的实现 在框架的第二阶段，我们需要一个混合整数线性规划公式来模拟预测函数 ψ 的计算过程。对于使用 LLR、人工神经网络、ALR 或 RLR 构建的预测函数，此类公式已有相关研究 [21, 27, 28]。对于第一阶段中性质 π 对应的数据集 $\mathcal{C}$，令 (w, b) 表示将 $\mathcal{C}$ 划分为子集 $\mathcal{C}^{(j)}, j = 1, 2$ 的超平面，并令 ψ_j 表示为 $\mathcal{C}^{(j)}$ 构建的预测函数。假设对于特征函数 f、拓扑规范 σ 以及每个预测函数 $\psi_j, j = 1, 2$，我们都有一个混合整数线性规划公式 $\mathcal{M}_{f,\ \eta_j,\ \sigma}$，用于在第二阶段推断满足 σ 的化学图。

令 $\underline{y}^*$ 和 $\overline{y}^*$ 分别为性质 π 的目标值下限和上限。回想一下，化合物 $\mathbb{C} \in \mathcal{C}$ 的观测值 $a(\mathbb{C})$ 会被归一化为介于 0 和 1 之间的值 $\nu(a(\mathbb{C}))$。令 $\nu(\underline{y}^*)$ 和 $\nu(\overline{y}^*)$ 分别表示 $\underline{y}^*$ 和 $\overline{y}^*$ 的归一化值，其中假设 $\nu(\overline{y}^*) \leq \max\{a^{(1)}_{\max}, \theta\}$ 或 $\min\{a^{(2)} \min\ \theta\} \leq \nu(\underline{y}^*)$（否则，我们考虑两个目标实例 $[\nu(\underline{y}^*)$、$\max\{a^{(1)}_{\max}, \theta\}]$ 和 $[\min\{a^{(2)} \min\ \theta\}, \nu(\overline{y}^*)])$。对于前一种情况（分别地，后一种情况），我们求解混合整数线性规划 $\mathcal{M}_{f,\eta_1,\sigma}$，并附加约束 $wx_i - b \leq 0$（分别地，混合整数线性规划 $\mathcal{M}_{f,\eta_2,\sigma}$，并附加约束 $wx_i - b \geq 0$），以推断所需的化学图 $\mathbb{C}^\dagger$。

6 结果

采用新的数据集划分方法，并构建混合整数线性规划（MILP）来处理双层模型中的二次描述符后，我们实现了用于化学图推断的框架，并开展了实验以评估计算效率。我们在一台配备以下处理器和内存的电脑上执行了实验：处理器：Core i7-9700（3.0GHz；最高4.7 GHz），内存：16 GB DDR4 RAM。为了通过多元线性回归（MLR）或人工神经网络（ANN）构建预测函数，我们使用了 `scikit-learn` 版本1.0.2、Python 3.8.12、用于MLR的MLPRegressor，以及用于ANN的ReLU激活函数。

6.1 框架第一阶段的结果

化学性质 我们针对以下22种化学性质实现了第一阶段 s
单体的：HSDB通过PubChem [30]提供的自燃温度（AT）、生物半衰期（BHL）、临界压力（CP）、临界温度（CT）、解离常数（DC）、爆炸下限（FLML）、爆炸上限（FLMU）、闭杯闪点（FP）、熔点（MP）、树木折射率（RFIDT）、气味阈值下限（ODRL）、气味阈值上限（ODRU）和蒸气压（VP）；ESOL [31]提供的溶解度（SL）；A. Dashti等 [32]提供的有机化合物自燃温度（ATO）；S. Yuan等 [33]提供的有机化合物爆炸上限（FLMUO）；S. Kondo等 [34]提供的气体爆炸下限（FLMLG）和气体爆炸上限（FLMUG）；以及最高占据分子轨道能量（HOMO），

最低未占分子轨道能量（Lumo）、 Homo 与 Lumo 之间
的能量差（Gap）以及
MoleculeNet [35],提供的电偶极矩（mu），其中从 Homo到 mu 的所有这些数据均
基于同一个数据集 QM9。

数据集 QM9 包含超过 130,000 种化合物。在我们的实验中，我们使用从该数据集中随机选取的 1,000 种化合物。对于这些性质，我们没有将原始数据集中的任何聚合物作为离群值排除。

设置数据集 对于每种性质 π，我们首先选取一组 Λ 化学元素，然后收集一个关于化学图的数据集 $\mathcal{C}_\pi$，这些化学图由化学元素集合 Λ 构成。为构建数据集 $\mathcal{C}_\pi$，我们剔除了不满足以下任一条件的化合物：图是连通的；碳原子数至少为四；每个原子的非氢邻居数最多为 4。

我们将分支参数 ρ 设为 2，引入由化学模型中未抑制氢的双层图定义的线性描述符，并使用线性描述符和二次描述符集合 $D_\pi^{(1)}$和 $D_\pi^{(2)}$（详情见附录A）。

对于性质 $\pi \in \{\text{BHL, FlmL, FlmLG, FlmU, FlmUO, FlmUG, OdrL, OdrU, Vp}\}$，我们采用观测值 $a(\mathbb{C})$ $\mathbb{C} \in \mathcal{D}_\pi$ 的对数测量值；在实验中，我们将 $\log(a(\mathbb{C}) + c)$ ($c = 10^{-8}$ 若 $\pi =$Vp，否则为 $c = 0$) 用作观测值。

我们对每个线性描述符的取值范围以及观测值 $a(\mathbb{C})$ $\mathbb{C} \in \mathcal{C}_\pi$ 的取值范围进行归一化。

我们进行了一项实验，对比以下五种构建预测函数的方法。

(i) LLR：在由线性描述符构成的集合 $D_\pi^{(1)}$上使用 Lasso 线性回归（实现细节见 [21] ）；(ii) 人工神经网络：在由线性描述符构成的集合 $D_\pi^{(1)}$上使用人工神经网络（实现细节见 [21] ）；(iii) ALR：在由线性描述符构成的集合 $D_\pi^{(1)}$上使用调整线性回归（实现细节见 [27] ）；(iv) RLR：Zhu 等人提出的学习方法 [28]，该方法从由线性描述符和二次描述符构成的集合 $D_\pi^{(1)} \cup D_\pi^{(2)}$中选择一组描述符，然后对所选描述符应用 MLR（多维线性回归）来构建预测函数；(v) HPS：我们的方法，通过计算一个超平面，将给定数据集分成两个子集，并分别为每个子集构建预测函数。通过初步实验，我们预先确定阈值 θ 和超平面 (w, b) $w \in \mathbb{R}^{|D_\pi^{(1)}|}$ 及 $b \in \mathbb{R}$，该超平面位于线性描述符的特征空间中。在交叉验证中，我们将训练数据集 $\widehat{\mathcal{C}}_\pi \subseteq \mathcal{C}_\pi$ 分成两个子集 $\widehat{\mathcal{C}}_\pi^{(j)}, j = 1, 2$，并通过应用上述四种方法 (i)–(iv) 之一，为每个子集 $\widehat{\mathcal{C}}_\pi^{(j)}$ 构建预测函数 ψ_j。

表1显示了我们为每种化学性质准备的数据集的规模和范围，以构建预测函数，其中符号表示如下：- π：实验中使用的化学性质名称。

表1：设置数据集的结果。

π	Λ	$\lvert\mathcal{C}_\pi\rvert$	$\underline{n},\ \overline{n}$	$\underline{a},\ \overline{a}$	$\lvert\Gamma\rvert$	$\lvert\mathcal{F}\rvert$	K_1
At	λ_1	400	4, 85	64.0, 715.0	23	160	216
At	λ_3	448	4, 85	64.0, 715.0	28	181	254
AtO	λ_5	443	2, 32	170.0, 680.0	16	205	262
BHL	λ_1	300	5, 36	-1.522, 2.865	20	70	117
BHL	λ_3	514	5, 36	-1.522, 2.865	26	101	164
Cp	λ_1	125	4, 63	4.7×10^{-6}, 5.52	8	75	107
Cp	λ_4	131	4, 63	4.7×10^{-6}, 5.52	8	79	115
Ct	λ_1	125	4, 63	56.1, 3607.5	8	76	108
Ct	λ_4	132	4, 63	56.1, 3607.5	8	81	117
Dc	λ_1	141	5, 44	0.5, 17.11	20	62	109
Dc	λ_3	161	5, 44	0.5, 17.11	25	69	128
FlmL	λ_6	254	4, 67	-0.585, 0.875	19	126	177
FlmLG	λ_8	233	1, 13	-0.221, 1.158	10	152	199
FlmU	λ_6	219	4, 67	0.107, 1.681	19	119	170
FlmUO	λ_7	78	2, 10	0.732, 1.903	7	61	99
FlmUG	λ_7	233	1, 13	0.462, 2.0	10	152	199
Fp	λ_1	368	4, 67	-82.99, 300.0	20	131	181
Fp	λ_3	424	4, 67	-82.99, 300.0	25	161	228
Mp	λ_1	467	4, 122	-185.33, 300.0	23	142	195
Mp	λ_3	577	4, 122	-185.33, 300.0	32	176	253
OdrL	λ_1	64	4, 13	0.0002, 725.0	13	49	88
OdrL	λ_3	83	4, 22	0.0002, 725.0	16	60	107
OdrU	λ_1	64	4, 13	0.024, 6000.0	13	49	88
OdrU	λ_3	83	4, 22	0.015, 6000.0	16	60	107
RfIdT	λ_1	166	4, 26	1.3326, 1.613	14	98	139
Sl	λ_1	673	4, 55	-9.332, 1.11	27	154	216
Sl	λ_3	915	4, 55	-11.6, 1.11	42	207	299
Vp	λ_1	392	4, 55	-8.0, 3.416	22	133	185
Vp	λ_3	482	4, 55	-8.0, 3.416	30	165	236
Homo	λ_2	977	6, 9	-0.3335, -0.1583	59	190	296
Lumo	λ_2	977	6, 9	-0.1144, 0.1026	59	190	296
Gap	λ_2	977	6, 9	0.1324, 0.4117	59	190	296
mu	λ_2	977	6, 9	0.04, 6.8966	59	190	296

- Λ: 数据集 $\mathcal{C}_\pi$ 中使用的一组选定元素； Λ 是以下八个集合之一：$\Lambda_1 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}\}$；
$\Lambda_2 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{F}\}$； $\Lambda_3 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{S}_{(2)}, \mathtt{S}_{(6)}\}$； $\Lambda_4 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{Pb}\}$；
$\Lambda_5 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{S}_{(2)}, \mathtt{S}_{(4)}, \mathtt{S}_{(6)}, \mathtt{Cl}, \mathtt{Br}, \mathtt{F}\}$； $\Lambda_6 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{P}_{(2)}, \mathtt{P}_{(5)}\}$； $\Lambda_7 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{Br}\}$；
$\Lambda_8 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{Br}, \mathtt{F}\}$，其中对于化学元素 $\mathtt{a}$ 和整数 $i \geq 1$，$\mathtt{a}_{(i)}$ 表示

化合价为 i 的化学元素 a 。 - $|\mathcal{C}_\pi|$: 性质 π 在元素集 Λ 上的数据集 $\mathcal{C}_\pi$ 的大小。 - $\underline{n}$, $\overline{n}$: 化合物 $\mathbb{C}$ 在 $\mathcal{C}_\pi$ 中所含非氢原子数 $n(\mathbb{C})$ 的最小值和最大值。 - $\underline{a}$, $\overline{a}$: 对化合物 $\mathbb{C}$ 在 $\mathcal{C}_\pi$ 中的 π ，在将其归一化到 0 到 1 之间之前， $a(\mathbb{C})$（或原始观测值的对数）的最小值和最大值。 - $|\Gamma|$: 化合物在 $\mathcal{C}_\pi$ 中的内部边的不同边构型数。 - $|\mathcal{F}|$: 化合物 $\mathcal{C}_\pi$ 中所有 2-边缘树组成的集合中的非同构化学根树数。 - K_1: 线性描述符集合 $D_\pi^{(1)}$ 的大小 $|D_\pi^{(1)}|$ ，其中满足 $|D_\pi^{(2)}| = (3K_1^2 + K_1)/2$ 。

构建预测函数 对于每种化学性质 π，我们通过四种方法 (i)–(iv) 中的一种构建预测函数。

对于方法 (i)–(iv)，由于 Zhu 等人 [21, 27, 28] ，我们采用了相同的实现方式，具体细节从略。

表2展示了构建预测函数的结果，其中符号含义如下：- π、Λ：一个具有化学性质 π 以及从数据集 $\mathcal{C}_\pi$中选取的一组化学元素 Λ 的实例。- LLR：对于通过方法 (i) 构建的预测函数，在十次五折交叉验证中测试 R^2 的中位数。- 人工神经网络：对于通过方法 (ii) 构建的预测函数，在十次五折交叉验证中测试 R^2 的中位数。- ALR：对于通过方法 (iii) 构建的预测函数，在十次五折交叉验证中测试 R^2 的中位数。- RLR：对于通过方法 (iv) 构建的预测函数，在十次五折交叉验证中测试 R^2 的中位数。- HPS：HPS 获得的预测函数在十次五折交叉验证中测试 R^2 的中位数，其中预测函数通过组合针对第一子集和第二子集构建的预测函数获得。- LLR、ANN、ALR、RLR 或 HPS 的得分后标有“*”，表示在性质 π方面五种方法中的最佳性能；- θ: HPS 中用于拆分数据集 $\mathcal{C}_\pi$的阈值。- $a^{(1)}_{\max}$、$a^{(2)}{}_{\min}$：HPS 中某个化合物 $\mathbb{C} \in \mathcal{C}^{(1)}$ 的最大观测值 $a(\mathbb{C})$和某个化合物 $\mathbb{C} \in \mathcal{C}^{(2)}$ 的最小观测值 $a(\mathbb{C})$：- $|\mathcal{C}^{(1)}|$、$|\mathcal{C}^{(2)}|$：HPS 中第一子集 $\mathcal{C}^{(1)}$ 和第二子集 $\mathcal{C}^{(2)}$ 的大小。- $\mathcal{C}^{(1)}$-R^2：HPS 中用于针对第一子集构建预测函数的方法名称（(i)–(iv) 之一），以及该预测函数在十次五折交叉验证中第一子集上的测试 R^2 中位数。- $\mathcal{C}^{(2)}$-R^2: HPS 中用于针对第二子集构建预测函数的方法名称（(i)–(iv) 之一），以及该预测函数在十次五折交叉验证中第二子集上的测试 R^2 中位数。

在 HPS 中为子集 $\mathcal{C}^{(1)}$ 和 $\mathcal{C}^{(2)}$ 构建两个预测函数，耗时最多约 13 秒。执行 RLR 时，最耗时的部分是选择线性和二次描述符子集

，耗时为 1162 至 44356 秒。

表2：构建单体预测函数的结果。

π, Λ	LLR	ANN	ALR	RLR	HPS	θ	$a_{\max}^{(1)}$ $a_{\min}^{(2)}$	$\lvert\mathcal{C}^{(1)}\rvert$, $\lvert\mathcal{C}^{(2)}\rvert$	$\mathcal{C}^{(1)}$-R^2	$\mathcal{C}^{(2)}$-R^2
At, λ_1	0.363	0.465	0.449	0.505	*0.808	0.40	0.505, 0.293	147, 253	RLR 0.220	RLR 0.521
At, λ_3	0.391	0.476	0.442	0.501	*0.765	0.35	0.555, 0.178	121, 327	人工神经网络 0.126	RLR 0.532
AtO, λ_5	0.710	0.715	0.716	0.829	*0.894	0.45	0.447, 0.451	223, 220	人工神经网络 0.747	RLR 0.610
BHL, λ_1	0.580	0.648	0.679	0.759	*0.818	0.45	0.598, 0.259	124, 176	RLR 0.650	RLR 0.560
BHL, λ_3	0.688	0.751	0.706	*0.834	0.831	0.50	0.753, 0.183	240, 274	RLR 0.620	RLR 0.569
Cp, λ_1	0.429	0.592	0.805	0.677	*0.887	0.55	0.549, 0.562	59, 66	RLR 0.759	RLR 0.632
Cp, λ_4	0.559	0.768	0.546	0.841	*0.850	0.60	0.598, 0.601	68, 63	RLR 0.661	RLR 0.564
Ct, λ_1	0.069	0.290	0.903	*0.937	0.863	0.15	0.150, 0.151	60, 65	RLR 0.713	LLR 0.740
Ct, λ_4	0.037	0.236	*0.941	0.860	0.920	0.15	0.150, 0.150	61, 71	RLR 0.791	RLR 0.883
Dc, λ_1	0.559	0.662	0.529	0.908	*0.939	0.45	0.440, 0.456	68, 73	RLR 0.787	RLR 0.810
Dc, λ_3	0.574	0.628	0.534	0.829	*0.919	0.40	0.386, 0.400	78, 83	RLR 0.551	RLR 0.760
FlmL, λ_6	0.412	0.524	0.707	0.604	*0.833	0.50	0.499, 0.502	106, 113	RLR 0.920	RLR 0.363
FlmLG, λ_8	0.850	0.786	0.824	0.925	*0.935	0.25	0.243, 0.254	38, 40	RLR 0.838	RLR 0.833
FlmU, λ_6	0.146	0.295	0.311	0.538	*0.703	0.45	0.445, 0.451	167, 66	RLR 0.477	RLR 0.344
FlmUO, λ_7	0.442	0.573	0.221	0.642	*0.851	0.45	0.429, 0.455	107, 147	RLR 0.945	RLR 0.559
FlmUG, λ_7	0.556	0.649	0.443	0.655	*0.837	0.35	0.346, 0.363	126, 107	RLR 0.679	RLR 0.292
Fp, λ_1	0.607	0.742	0.653	0.899	*0.920	0.40	0.399, 0.402	178, 190	RLR 0.684	RLR 0.880
Fp, λ_3	0.593	0.634	0.626	0.846	*0.880	0.45	0.449, 0.452	261, 163	RLR 0.697	RLR 0.645
Mp, λ_1	0.815	0.840	0.850	0.873	*0.923	0.50	0.493, 0.504	272, 195	RLR 0.607	RLR 0.778
Mp, λ_3	0.786	0.843	0.803	0.898	*0.920	0.55	0.547, 0.542	329, 248	人工神经网络 0.659	RLR 0.747
OdrL, λ_1	-0.034	-0.402	-0.098	0.449	*0.901	0.50	0.498, 0.503	29, 35	RLR 0.695	RLR 0.836
OdrL, λ_3	-0.055	-0.172	0.041	0.532	*0.827	0.50	0.498, 0.503	38, 45	RLR 0.543	RLR 0.657
OdrU, λ_1	0.008	0.170	0.214	0.641	*0.931	0.55	0.541, 0.550	32, 32	RLR 0.755	RLR 0.800
OdrU, λ_3	0.164	0.266	0.392	0.654	*0.955	0.50	0.496, 0.500	40, 43	RLR 0.936	RLR 0.749
RfIdT, λ_1	0.679	0.770	0.677	*0.876	*0.876	0.30	0.294, 0.303	89, 77	RLR 0.873	RLR 0.594
Sl, λ_1	0.771	0.831	0.788	0.894	*0.923	0.45	0.449, 0.439	109, 564	RLR 0.781	RLR 0.847
Sl, λ_3	0.807	0.867	0.813	0.897	*0.916	0.50	0.548, 0.500	128, 787	RLR 0.859	RLR 0.832
Vp, λ_1	0.871	0.937	0.893	0.969	*0.986	0.45	0.449, 0.453	155, 237	RLR 0.930	RLR 0.949
Vp, λ_3	0.830	0.922	0.871	0.959	*0.977	0.45	0.449, 0.453	210, 272	RLR 0.845	RLR 0.935
Gap, λ_2	0.784	0.763	0.744	0.876	*0.907	0.25	0.298, 0.162	131, 846	RLR 0.524	RLR 0.863
Homo, λ_2	0.704	0.608	0.657	0.804	*0.847	0.65	0.792, 0.557	849, 128	RLR 0.733	RLR 0.768
Lumo, λ_2	0.841	0.843	0.819	0.920	*0.948	0.70	0.700, 0.701	700, 277	RLR 0.874	RLR 0.855
mu, λ_2	0.366	0.442	0.399	0.645	*0.708	0.20	0.506, 0.021	169, 808	RLR 0.708	RLR 0.593

在表2中，在大多数情况下，HPS在五种方法中取得了测试集R^2 中位数的最佳得分。尤其是，对于性质 At、 FlmL、FlmU、 FlmUO、 FlmUG、 OdrL和 OdrU，相较于方法(i)–(iv)的提升显著。

根据表2中的数值 $a_{\max}^{(1)}, a^{(2)}$ 最小值，我们可以看到原始集合 $\mathcal{C}_\pi$ 被一个超平面分成两个子集 $\mathcal{C}^{(j)}$, $j = 1$，每个性质 π对应 2 个元素。对于具有 λ_1 的性质 At，有 $a_{\max}^{(1)} = 0.505 > a^{(2)}$ 最小值 $= 0.293$, 这意味着没有超平面能够将 $\mathcal{C}_{\leq\theta} = \{\mathbb{C}_i \in \mathcal{C}_\pi \mid a_i \leq \theta\}$ 和$\mathcal{C}_{>\theta} = \{\mathbb{C}_i \in \mathcal{C}_\pi \mid a_i > \theta\}$ 按照 $\theta = 0.40$ 分开。另一方面，对于性质 AtO，有

$a^{(1)}_{\max} = 0.447 < \theta = 0.45 < a^{(2)}$ 最小值$= 0.451$，这意味着 $\mathcal{C}_\pi$ 被一个超平面分成$\mathcal{C}^{(1)} = \mathcal{C}_{\leq\theta}$ 和 $\mathcal{C}^{(2)} = \mathcal{C}_{>\theta}$。

对于某些性质，例如 At with λ_1，HPS 的 R^2 中位数明显大于 $\mathcal{C}^{(1)}$-R^2 和 $\mathcal{C}^{(2)}$-R^2 的 R^2 中位数。对于性质 $\pi =$ At with λ_1，HPS 的 R^2 中位数为 0.808，而 $\mathcal{C}^{(1)}$（分别为 $\mathcal{C}^{(2)}$）的中位数为 0.220（分别为 0.521）。出现这种情况的原因可能是，$\mathcal{C}^{(1)}$（分别为 $\mathcal{C}^{(2)}$）的观测值范围 $[a^{(1)}{}_{\min} = 0, a^{(1)}_{\max} = 0.505]$（分别为 $[a^{(2)}{}_{\min} = 0.293, a^{(2)}_{\max} = 1]$）会再次归一化到 $[0, 1]$，并在其上构建预测函数 ψ_1（分别为 ψ_2），然后评估其学习性能 $\mathcal{C}^{(1)}$-R^2（分别为 $\mathcal{C}^{(2)}$-R^2）。然而，在评估 HPS 的 R^2 中位数时，预测函数 $\psi_j, j = 1, 2$ 所产生的误差，是在相对于 $\mathcal{C}_\pi$ 原始观测值范围更宽的范围上衡量的，因此误差值较小。

6.2 框架第二阶段的结果

为了执行第二阶段，我们使用了基于 Zhu 等人 [21] 准备的种子图的一组七个实例 I_{a}、I^i_{b}、$i \in [1, 4]$、I_{c} 和 I_{d}。实例 I_{a} 和 I_{c} 具有受限种子图，实例 I^i_{b} 具有抽象种子图，实例 I_{c} 和 I_{d} 具有受限的枝树集合。我们在此展示它们的种子图 G_{C}（有关 I_{a} 的详细信息，请参见附录 B；有关 I^i_{b}、$i \in [1, 4]$、I_{c} 和 I_{d} 的详细信息，请参见附录 C）。

I_{a} 的种子图 G_{C} 由图4(a)中的图给出。I^1_{b} 的种子图 G^1_{C}（分别对应G^i_{C}、$i = 2$、I^i_{b}的3、4以及$i = 2$的3、4）如图5所示。

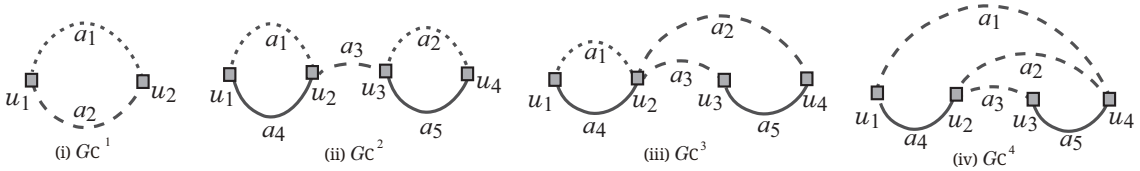


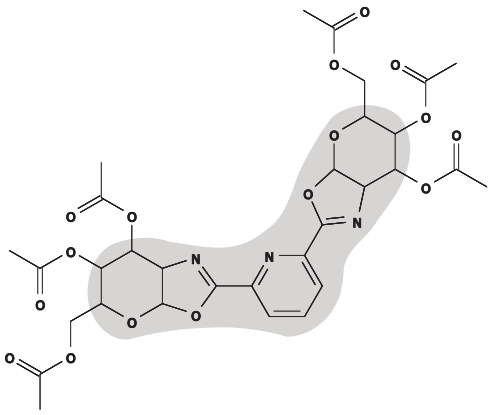
图5：(i) I^1_{b} 和 I_{d}的种子图 G^1_{C}；(ii) I^2_{b}的种子图 G^2_{C}；(iii) I^3_{b}的种子图 G^3_{C}；(iv) I^4_{b}的种子图 G^4_{C}。

引入实例 I_{c} 以推断化学图 $\mathbb{C}^\dagger$，使得：- $\mathbb{C}^\dagger$ 的核心部分与图6(a)中的化学图 $\mathbb{C}_A$（CID 24822711）的核心部分相同（其中，图6(a)中阴影区域表示 I_{c} 的种子图 G_{C}）。- $\mathbb{C}^\dagger$ 非核心部分中每种边构型的频数与图6(b)中的化学图$\mathbb{C}_B$（CID 59170444）相同。

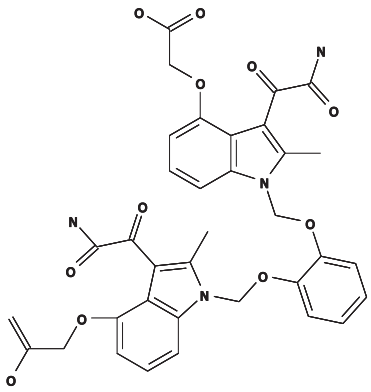
引入实例 I_{d} 以推断化学图 $\mathbb{C}^\dagger$，使得：- $\mathbb{C}^\dagger$ 是单环的（其中，I_{d} 的种子图由图5(i)中的 G^1_{C} 给出）；并且- $\mathbb{C}^\dagger$ 中边构型的频率向量是通过合并图6(c)和(d)中化学图 $\mathbb{C}_A$（CID 10076784）和 $\mathbb{C}_B$（CID 44340250）的边构型频率向量而得到的。

求解逆问题的混合整数线性规划我们执行了求解混合整数线性规划的阶段，以推断具有两种性质 $\pi \in \{$At, FLML$\}$ 的化学图。

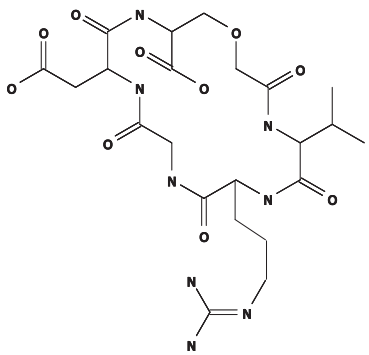
对于混合整数线性规划公式 $\mathcal{M}_{f,\ \eta,\ \sigma}$，我们对每个 At 使用预测函数 η，其中 Λ_3和 FLML 由方法(v)构造，并且 HPS 在表2中取得了中位数测试 R^2。



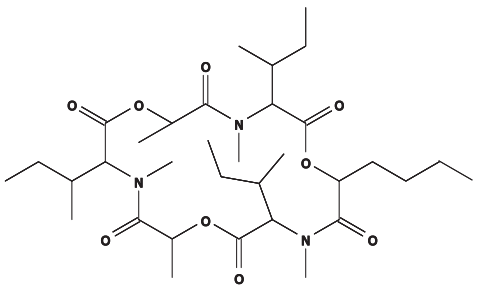
(a) $\mathcal{C}^{\mathcal{A}}$: CID 24822711



(b) $\mathcal{C}^{\mathcal{B}}$: CID 59170444



(c) $\mathcal{C}^{\mathcal{A}}$: CID 10076784



(d) $\mathcal{C}^{\mathcal{B}}$: CID 44340250

图6：实例 I_{c} 和 I_{d}的化合物示意图：(a) $\mathbb{C}_A$：CID 24822711；(b) $\mathbb{C}_B$：CID 59170444；(c) $\mathbb{C}_A$：CID 10076784；(d) $\mathbb{C}_B$：CID 44340250，其中省略了氢原子。

为使用该公式求解混合整数线性规划，我们使用了 CPLEX 12.10 版本。表3和表4展示了这一阶段针对这两个性质的实验计算结果，其中我们定义如下：

- n_{LB}：待推断化学图中非氢原子数量的下界 $\mathbb{C}$；- $\underline{y}^*$，$\overline{y}^*$：待推断化学图的值 $a(\mathbb{C})$的下界和上界$\underline{y}^*, \overline{y}^* \in \mathbb{R}$；对于 AT，我们使用归一化前原始值 $a(\mathbb{C})$的范围。对于 FLML，我们使用对数尺度 $\log a(\mathbb{C})$作为目标值范围。- #v（分别为 #c）：混合整数线性规划中的变量（分别为约束）数量；- 求解时间：求解混合整数线性规划所需的时间（秒）；- n：通过求解混合整数线性规划推断出的化学图 $\mathbb{C}^\dagger$ 中非氢原子的数量 $n(\mathbb{C}^\dagger)$；- $\mathrm{n}^{\mathrm{int}}$：推断出的化学图 $\mathbb{C}^\dagger$中内部顶点的数量 $\mathrm{n}^{\mathrm{int}}(\mathbb{C}^\dagger)$；以及 - η：推断出的化学图 $\mathbb{C}^\dagger$的预测性质值 $\eta(f(\mathbb{C}^\dagger))$。图7(a)展示了由 I_{c} 推断出的化学图 $\mathbb{C}^\dagger$，其 AT的$\underline{y}^*, \overline{y}^*$为 $=(360, 370)$

表3：使用 Λ_3为 AT推断化学图 $\mathbb{C}^\dagger$ 并生成重组解的结果。

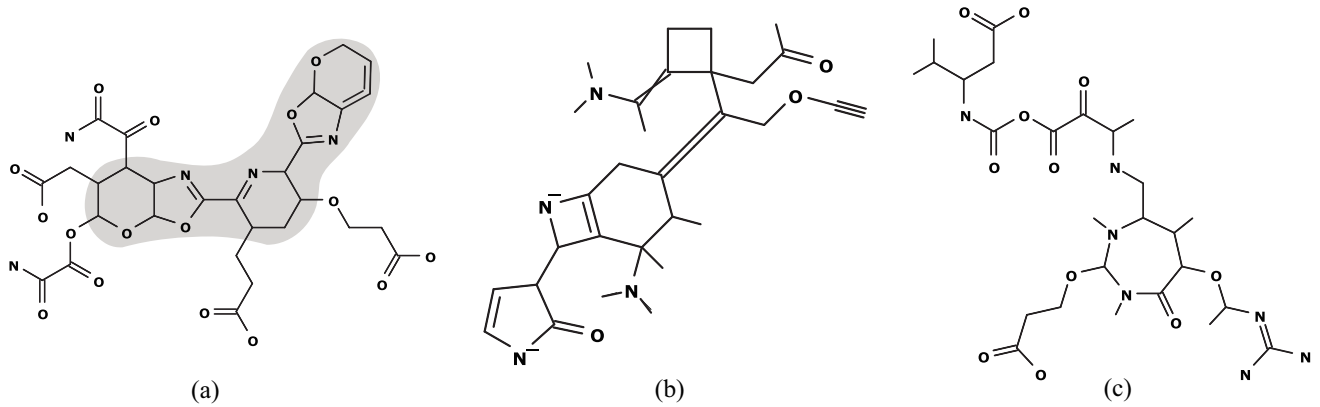
即	n_{LB}	$\underline{y}^*$, $\overline{y}^*$	#v	#c	求解时间	n	$\mathrm{n}^{\mathrm{int}}$	η	动态规划时间	$\mathbb{C}$-LB	#$\mathbb{C}$
I_{a}	30	240, 250	9846	9250	3.61	44	25	241.98	0.111	12	12
I_{b}^1	35	240, 250	10428	6904	3.07	37	15	247.71	0.0481	8	8
I_{b}^2	45	240, 250	13113	10019	12.8	50	25	241.51	0.155	432	100
I_{b}^3	45	190, 200	12909	10022	10.3	50	25	197.34	0.267	208	100
I_{b}^4	45	300, 310	12705	10025	12.1	48	29	300.18	0.192	432	100
I_{c}	50	360, 370	7875	8734	1.29	50	33	361.62	0.0163	1	1
I_{d}	40	230, 240	5479	6775	2.88	43	23	237.72	0.184	10496	100

表4：推断化学图 $\mathbb{C}^\dagger$ 并为FLML 生成重组解的结果，使用 $\Lambda_{6\circ}$

实例。	n_{LB}	$\underline{y}^*$, $\overline{y}^*$	#v	#c	求解时间	n	$\mathrm{n}^{\mathrm{int}}$	η	动态规划时间	$\mathbb{C}$-LB	#$\mathbb{C}$
I_{a}	30	-0.55, -0.5	11794	12922	4.17	38	23	-0.551	0.0692	1	1
I_{b}^1	35	-0.15, -0.1	11107	10423	2.82	35	9	-0.140	0.0163	2	2
I_{b}^2	45	-0.5, -0.45	13426	13871	28.6	50	28	-0.494	12.8	207594	100
I_{b}^3	45	-0.5, -0.45	12974	13535	20.2	45	25	-0.483	17.2	2215023	100
I_{b}^4	45	-0.45, -0.4	12746	13534	12.5	46	27	-0.445	0.0889	5040	100
I_{c}	50	-0.55, -0.5	9386	11000	0.878	50	33	-0.509	0.0162	1	1
I_{d}	40	0.2, 0.25	6184	7821	9.18	44	23	0.204	0.16	21600	100

见表 3。

图7(b)（分别为图7(c)）展示了表4中根据 I_{a} 推断得到的化学图 $\mathbb{C}^\dagger$，其参数为（$\underline{y}^*$,$\overline{y}^*$）$=$（$-0.55, -0.5$）（分别为 I_{d}，参数为（$\underline{y}^*$,$\overline{y}^*$）$=$(0.2, 0.25)）。



(a) (b) (c)

图 7：

从表3和表4中可以看出，一般而言，变量和约束数量较多的实例比规模较小的实例需要更长的运行时间。使用我们的混合整数线性规划公式，本实验中的所有实例均在几秒到约30秒内求解。

生成重组解 令 $\mathbb{C}^\dagger$ 为通过求解逆问题的混合整数线性规划 $\mathcal{M}_{f,\,\eta,\,\sigma}$ 得到的化学图。此处，我们执行一个阶段，生成 $\mathbb{C}^\dagger$ 的重组解 $\mathbb{C}^* \in \mathcal{G}_\sigma$，使得 $f(\mathbb{C}^*) = x^* = f(\mathbb{C}^\dagger)$。

当化学异构体总数超过 100 时，我们执行一种算法，生成至多 100 个 $\mathbb{C}^\dagger$ 的化学异构体。为此，我们使用动态规划算法 [21]。该算法首先将 $\mathbb{C}^\dagger$ 分解为一组无环化学图，接着用另一个无环化学图 T' 替换每个无环化学图 T，该替代图具有与 T 相同的特征向量，最后将所得无环化学图组装成 $\mathbb{C}^\dagger$ 的一个化学异构体 $\mathbb{C}^*$。该算法无需生成全部化学异构体，即可计算化学异构体总数 $\mathbb{C}^\dagger$ 的下界。

表3和表4展示了本阶段针对两种性质 $\pi \in \{\text{AT}, \text{FLML}\}$ 的实验计算结果，其中下列符号表示：
- 动态规划时间：运行动态规划算法以计算所有化学异构体 $\mathbb{C}^*$ of $\mathbb{C}^\dagger$ 数量下界并生成所有（或最多100个）化学异构体 $\mathbb{C}^*$ 所需的时间（秒）；- $\mathbb{C}$-LB：所有化学异构体 $\mathbb{C}^*$ of $\mathbb{C}^\dagger$ 数量的下界；以及 - #$\mathbb{C}$：本阶段生成的所有（或最多100个）化学异构体 $\mathbb{C}^*$ of $\mathbb{C}^\dagger$ 的数量。

从表3和表4中，我们可以观察到本阶段的运行时间以及生成的重组解数量。

对于所测试的性质，I_{b}^2、I_{b}^3 和 I_{d} 中的化学图 $\mathbb{C}^\dagger$ 存在大量化学异构体 $\mathbb{C}^*$，其中化学异构体数量的下界 $\mathbb{C}$-LB 无需生成所有异构体即可得出。计算下界并生成最多100个目标化学图所需的时间不超过18秒。对于某些化学图 $\mathbb{C}^\dagger$，我们的算法找到的化学异构体数量较少。这是因为，在 $\mathbb{C}^\dagger$ 的分解中，有些无环化学图除了原始图之外，不存在其他可选的无环化学图。

生成邻域解 令 $\mathbb{C}^\dagger$ 为通过求解用于逆问题的混合整数线性规划 $\mathcal{M}_{f,\,\eta,\,\sigma}$ 得到的化学图。我们对 $\mathbb{C}^\dagger$ 执行了生成邻域解的阶段。

我们为逆问题选择一个混合整数线性规划，其预测函数为 η，使得在生成重组解的阶段，该混合整数线性规划的一个解 $\mathbb{C}^\dagger$ 仅对应两个异构体 $\mathbb{C}^*$；即，性质 AT 对应的实例 I_{c}（其中 Λ_3），以及性质 FLML 对应的实例 I_{a}、I_{b}^4 和 I_{c}（其中 Λ_6）。

在本实验中，我们在混合整数线性规划 $\mathcal{M}_{f,\,\eta,\,\sigma}$ 中额外加入一组 Θ 约束，包含两个关于线性描述符和二次描述符的线性约束，具体如下。对于这两个约束，我们使用 RLR 为表2中满足 Λ_3 的性质 $\pi \in \{\text{MP}, \text{SL}\}$ 构建的预测函数 η_π。

我们将 η_{MP} 和 η_{SL} 分别视为从 $\mathbb{R}^{|D_\pi^{\text{union}}|}$ 到 $\mathbb{R}$ 的函数，其中 $\pi \in \{\text{AT}, \text{FLML}\}$。我们设置 $p_{\text{dim}} := 2$，并令 Θ 由两个线性约束 $\theta_1 := \eta_{\text{MP}}$ 和 $\theta_2 := \eta_{\text{SL}}$组成。我们选择$\delta \in \{0.01, 0.05, 0.1, 0.15\}$，其定义了一个二维网格空间，其中 $\mathbb{C}^\dagger$映射到原点（有关邻居的详细信息，请参见 [23] ）。我们在网格搜索空间中选择原点$\mathbb{C}^\dagger$ 的 48 个邻居组成集合 N_0。对于每个实例，我们按邻居与原点之间距离递增的顺序检查 N_0 中邻居的可行性，

对于每个可行邻居 $z \in N_0$，输出增强型混合整数线性规划实例的一个可行解 $\mathbb{C}_z^\dagger$。我们将检查邻居可行性的时间限制设为 300 秒；如果相应的混合整数线性规划未能在时间限制内求解，我们就跳过该邻居。此外，如果我们发现一个不可行邻居 $z' \in N_0$，且 z' 比 z 更接近原点，那么即使不测试 z 的可行性，我们也会忽略任何邻居 $z \in N_0$。

表5展示了三个实例的实验计算结果，其中符号含义如下：

- (实例，π)：拓扑规范 I 和性质 π；- n：受测实例中的非氢原子数；- δ：网格搜索空间中一个子区域的大小；- #sol：从邻域集 N_0获得的新化学图数量；- #infs：搜索过程中发现不可行的 N_0 中邻居数量；- #ign：搜索过程中被忽略的 N_0 中邻居数量；- #TO：搜索过程中可行性检查耗时超过300秒时间限制的 N_0 中邻居数量。

表5：生成 $\mathbb{C}^\dagger$ 的邻域解的结果。

(实例，π)	n	δ	#sol	#infs	#ign	#TO
(I_{c},AT)	50	0.01	9	0	0	39
(I_{a},FLML)	30	0.05	41	0	0	7
(I_{b}^4,FLML)	45	0.15	38	0	0	10
(I_{c},FLML)	40	0.05	15	0	0	33

在许多求解器（如 CPLEX）中，混合整数线性规划采用基于分支定界法的算法求解，对于规模相同的实例，有时会耗费极长的运行时间。我们引入时间限制，以限定在 N_0 中测试邻居可行性的运行时间，从而跳过此类实例。从表5可以看出，对于四个实例中的每一个，都成功生成了一定数量的$\mathbb{C}^\dagger$ 邻域解。

7 结论 remarks

在化学图推断框架中，预测函数的描述符主要定义为双层模型中局部图结构出现的频率，以这种方式定义描述符对于在框架第二阶段推导出紧凑的混合整数线性规划公式至关重要。为提升基于相同描述符定义的预测函数的性能，本文提出了一种方法：通过特征空间中的一个超平面将给定数据集划分为两个子集，使第一和第二子集分别主要由观测值低于和高于某一阈值的化合物构成。通过独立构建第一和第二子集的预测函数，并将二者结合，得到一个预测函数；其中，超平面用于确定对给定特征向量采用哪一个预测函数。

实验结果表明，所提出的方法提升了可燃极限和气味阈值等化学性质的学习性能；对于包含约50 个非氢原子的化学图推断实例，第二阶段的混合整数线性规划也可以求解。未来的工作将进一步拓展这种数据集划分方法：当数据集规模足够大时，反复将给定数据集划分为观测值范围更窄的较小子集。

参考文献

[1] Lo, Y-C., Rensi, S. E., Torng, W., Altman, R. B.：化学信息学和药物发现中的机器学习。Drug Discovery Today 23，1538–1546（2018年）[2] Tetko, I. V., Engkvist, O.：从大数据到人工智能：化学信息学面临新挑战。J. Cheminformatics 12，74（2020年）[3] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., 等。定量构效关系建模：你从哪里来？将走向何方？J. Med. Chem. 57，4977–5010（2014年）[4] Miyao, T., Kaneko, H., Funatsu, K.：用于化学结构生成的逆向定量构效关系分析（从 y 到 x）。J. Chem. Inf. Model. 56，286–299（2016年）[5] Ikebata, H., Hongo, K., Isomura, T., Maezono, R., Yoshida, R.：利用化学语言模型进行贝叶斯分子设计。J. Comput. Aided Mol. Des. 31，379–391（2017年）[6] Rupakheti, C., Virshup, A., Yang, W., Beratan, D. N.：在小分子空间中发现多样化最优分子的策略。J. Chem. Inf. Model. 55，529–537（2015年）[7] Ghasemi, F., Mehridehnavi, A., P´erez-Garrido, A., P´erez-S´anchez, H.：用于定量构效关系研究的神经网络和深度学习算法：优点与缺点。Drug Discovery Today 23，1784–1790（2018年）[8] Kim, J., Park, S., Min, D., Kim, W.：利用深度学习开展近期药物发现的综合综述。Int. J. Molecular Science 22(18), 9983（2022年）[9] Kipf, T. N., Welling, M.：利用图卷积网络进行半监督分类。arXiv:1609.02907（2016年）[10] Bohacek, R. S., McMartin, C., Guida, W. C.：基于结构的药物设计的艺术与实践：分子建模视角。Med. Res. Rev. 16，3–50（1996年）[11] Akutsu, T., Fukagawa, D., Jansson, J., Sadakane, K.：根据路径频率推断图。Discrete Appl. Math. 160, 10-11，1416–1428（2012年）[12] Xiong, J., Xiong, Z., Chen, K., Jiang, H., Zheng, M.：用于自动化从头药物设计的图神经网络。Drug Discovery Today 26，1382–1393（2022年）

[13]G´omez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hern´andez-Lobato, J. M., S´anchez- Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., Aspuru- Guzik, A.：利用数据驱动的分子连续表示进行自动化化学设计。ACS Cent. Sci. 4, 268–276（2018年）[14] Kusner, M. J., Paige, B., Hern´andez-Lobato, J. M.：语法变分自编码器。第34届国际机器学习会议论文集第70卷，1945–1954（2017年）[15] De Cao, N., Kipf, T.：MolGAN：用于小分子图的隐式生成模型。arXiv:1805.11973（2018年）[16] Segler, M. H. S., Kogej, T., Tyrchan, C., Waller, M. P.：利用循环神经网络为药物发现生成聚焦分子库。ACS Cent. Sci. 4，120–131（2017年）[17] Yang, X., Zhang, J., Yoshizoe, K., Terayama, K., Tsuda, K.：ChemTS：用于从头分子生成的高效 Python 库。STAM 18，972–976（2017年）[18] Madhawa, K., Ishiguro, K., Nakago, K., Abe, M.：GraphNVP：用于生成分子图的可逆流模型。arXiv:1905.11600（2019年）[19] Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., Tang, J.：GraphAF：用于分子图生成的基于流的自回归模型。arXiv:2001.09382（2020年）[20] Shi, Y., Zhu, J., Azam, N. A., Haraguchi, K., Zhao, L., Nagamochi, H., Akutsu, T.：一种基于双层模型和整数规划的逆定量构效关系方法。Int. J. Molecular Sciences 22，2847（2021年）[21] Zhu, J., Azam, N. A., Haraguchi, K., Zhao, L., Nagamochi, H., Akutsu, T.：一种基于线性回归和整数规划的分子设计方法。第12届国际生物科学、生物化学与生物信息学会议，日本东京，1月7日至10日，#TJ0002，21–28（2022年）[22] Ido, R., Cao, S., Zhu, J., Azam, N. A., Haraguchi, K., Zhao, L., Nagamochi, H., Akutsu, T.：一种基于线性回归和整数规划的聚合物推断方法。第20届亚太生物信息学会议（APBC2022），2022年4月26日至28日[23] Azam, N. A., Zhu, J., Haraguchi, K., Zhao, L., Nagamochi, H., Akutsu, T.：基于人工神经网络、整数规划和网格邻居搜索的分子设计。BIBM 2021：360–363（2021年）[24] Azam, N. A., Zhu, J., Sun, Y., Shi, Y., Shurbevski, A., Zhao, L., Nagamochi, H., Akutsu, T.：一种基于人工神经网络和整数规划、用于推断具有有界分支高度的无环化合物的新方法。Algorithms for Molecular Biology 16，18（2021年）

[25] Cortes, C., Vapnik, V.：支持向量网络. 机器学习 20, 273-297 (1995)

[26] Tibshirani, R.：通过套索进行回归收缩与变量选择.《英国皇家统计学会杂志B辑》58 ,
267–288（1996年）[27] Zhu, J., Haraguchi, K., Nagamochi, H., Akutsu, T.：调整线性回归及其
s
在逆定量构效关系中的应用。第13届生物信息学模型国际会议 ,
方法与算法，2月9日至11日。#14，144–151（2022年）[28] Zhu, J., Azam, N. A., Cao, S., Ido, R., Haraguchi, K., Zhao, L., Nagamochi, H., Akutsu, T.
:
基于整数规划和二次描述符的分子设计，采用双层 d
模型。第21届生物信息学国际会议（InCoB2022），11月21日至2日（2022年）[29] 3
Freed, N., Glover, F.：用于判别问题的简单而强大的目标规划模型
-
lems. Europian J. Operations Research 7, 44–60 (1981)[30] HSDB 注释（来自 PubChem）： |https://pubchem.ncbi.nlm.nih.gov/—[31] MoleculeNet 上的 ESOL： | http://moleculenet.ai/datasets-1—[32] Dashti, A., Jokar, M., Amirkhani, F., Mohammadi, A. H.,：定量结构性质
y
用于估算有机化合物自燃温度的关系模型。J .
Molecular Liquids Feb 15; 300: 111797 (2020)[33] Yuan, S., Jiao, Z., Quddus, N., Kwon, J. S., Mashuga, C. V.：开发定量
e
结构—性质关系模型，以利用机器学习预测上限可燃性限值。 e
学习。Industrial & Engineering Chemistry Research 58(8):3531-7 (2019)[34] Kondo S., Urano, Y., Tokuhashi, K., Takahashi, A., Tanaka, K.：可燃性预测
y
通过 F 数分析气体。J. Hazardous Materials，3 月 30 日；82(2): 113-28 (2001)[35] MoleculeNet 上的 QM9： |http://moleculenet.ai—

附录

A 描述符的完整描述

对于化学图 $\mathbb{C}=(H,\alpha,\beta)$ 中的两个函数 α 和 β，我们定义如下函数 ac：$V(E)\to(\Lambda\setminus\{\mathtt{H}\})\times(\Lambda\setminus\{\mathtt{H}\})\times[1,3]$、cs：$V(E)\to(\Lambda\setminus\{\mathtt{H}\})\times[1,6]$ 以及 ec：$V(E)\to((\Lambda\setminus\{\mathtt{H}\})\times[1,6])\times((\Lambda\setminus\{\mathtt{H}\})\times[1,6])\times[1,3]$。

为了表示 $\mathbb{C}$ 外部的一个特征，$\mathcal{T}(\mathbb{C})$ 中的一个化学根树称为 $\mathbb{C}$ 的边缘构型。

我们还表示 $\mathbb{C}$ 外部的叶边。对于一条叶边 $uv\in E(\langle\mathbb{C}\rangle)$，若 $\deg^{\langle\mathbb{C}\rangle}(u)=1$，则定义 e 的邻接构型为有序元组 $(\alpha(u)\ \alpha(v)\ \beta(uv))$。定义

$$\Gamma_{\mathrm{ac}}^{\mathrm{lf}}\triangleq\{(\mathtt{a},\mathtt{b},m)\mid\mathtt{a},\mathtt{b}\in\Lambda,m\in[1,\min\{\mathrm{val}(\mathtt{a}),\mathrm{val}(\mathtt{b})\}]\}$$

为叶边所有可能邻接配置的集合。

为了表示化学图 $\mathbb{C}=(H,\alpha,\beta)$ 中内部顶点 $v\in V^{\mathrm{int}}(\mathbb{C})$ 的一种特征，使得 $\alpha(v)=\mathtt{a}$ 且 $\deg^{\langle\mathbb{C}\rangle}(v)=d$（即与 v 相邻的非氢原子数为 d），我们使用一对 $(\mathtt{a},d)\in(\Lambda\setminus\{\mathtt{H}\})\times[1,4]$，称其为顶点$v$的化学符号 $\mathsf{cs}(v)$。我们将 $(\mathtt{a},d)$ 视为一个单独的符号 $\mathtt{a}d$，并定义 Λ_{dg} 为所有化学符号$\mu=\mathtt{a}d\in(\Lambda\setminus\{\mathtt{H}\})\times[1,4]$的集合。

我们定义一种如下描述内部边的方法。设化学图$\mathbb{C}=(H,\alpha,\beta)$ 中的内部边 $e=uv\in E^{\mathrm{int}}(\mathbb{C})$ 为 $e=uv\in E^{\mathrm{int}}(\mathbb{C})$，满足 $\alpha(u)=\mathtt{a}$、$\alpha(v)=\mathtt{b}$ 且 $\beta(e)=m$。为了描述这条边 e，我们使用一个三元组 $(\mathtt{a},\mathtt{b},m)\in(\Lambda\setminus\{\mathtt{H}\})\times(\Lambda\setminus\{\mathtt{H}\})\times[1,3]$，称其为该边的邻接构型 $\mathsf{ac}(e)$。我们在 Λ 的元素上引入全序 $<$，以便在记号上区分 $(\mathtt{a},\mathtt{b},m)$ 和 $(\mathtt{b},\mathtt{a},m)$ $(\mathtt{a}\neq\mathtt{b})$。对于三元组$\nu=(\mathtt{a},\mathtt{b},m)$，令$\overline{\nu}$表示三元组 $(\mathtt{b},\mathtt{a},m)$。

设 $e=uv\in E^{\mathrm{int}}(\mathbb{C})$ 是化学图 $\mathbb{C}=(H,\alpha,\beta)$ 中满足 $\mathsf{cs}(u)=\mu$、$\mathsf{cs}(v)=\mu'$ and$\beta(e)=m$ 的内部边 $e=uv\in E^{\mathrm{int}}(\mathbb{C})$。为表征这条边 e，我们使用一个元组 $(\mu,\mu',m)\in\Lambda_{\mathrm{dg}}\times\Lambda_{\mathrm{dg}}\times[1,3]$，并将其称为边 e 的边构型 $\mathsf{ec}(e)$。我们在 Λ_{dg} 的元素上引入全序$<$，以便在记号上区分 (μ,μ',m) 和 (μ',μ,m) $(\mu\neq\mu')$。对于元组 $\gamma=(\mu,\mu',m)$，令 $\overline{\gamma}$ 表示元组 (μ',μ,m)。

设 π 是一种化学性质，我们将基于化学图 $\mathbb{C}$ 的特征向量 $f(\mathbb{C})$，构造一个从该向量到预测值 $y\in\mathbb{R}$ 的预测函数 η，以预测 $\mathbb{C}$ 的化学性质。

我们首先选择一组 Λ 化学元素，然后收集一个数据集 $\mathcal{C}_\pi$，其中包含化合物 C，其化学元素均属于 Λ；在这里，我们将 $\mathcal{C}_\pi$ 视为一组化学图 $\mathbb{C}$，用于表示 $\mathcal{C}_\pi$中的化合物 C。为了定义化学图$\mathbb{C}\in\mathcal{C}_\pi$的内部/外部，我们接下来选择一个分支参数 ρ，其中我们建议使用 $\rho=2$。

令 $\Lambda^{\mathrm{int}}(\mathcal{C}_\pi)\subseteq\Lambda$（分别地，$\Lambda^{\mathrm{ex}}(\mathcal{C}_\pi)\subseteq\Lambda$）表示所有化学图$\mathbb{C}\in\mathcal{C}_\pi$中，$\mathbb{C}$ 的内部顶点集合$V^{\mathrm{int}}(\mathbb{C})$（分别地，外部顶点集合 $V^{\mathrm{ex}}(\mathbb{C})$）所用化学元素的集合；令 $\Gamma^{\mathrm{int}}(\mathcal{C}_\pi)$表示所有化学图 $\mathbb{C}\in\mathcal{C}_\pi$中，$\mathbb{C}$ 的内部边集合 $E^{\mathrm{int}}(\mathbb{C})$所用边构型的集合。令 $\mathcal{F}(\mathcal{C}_\pi)$表示所有化学图 $\mathbb{C}\in\mathcal{C}_\pi$中，与 $\mathcal{T}(\mathbb{C})$内某个化学根树 ψ r-同构的化学根树的集合；其中，化学根树 $\psi\in\mathcal{F}(\mathcal{C}_\pi)$可能只由一个化学元素 $\mathtt{a}\in\Lambda\setminus\{\mathtt{H}\}$构成。

我们将有限元素集 A 的整数编码定义为一个双射 $\sigma: A \to [1, |A|]$，其中以 $[A]$ 表示整数集 $[1, |A|]$。为集合$\Lambda^{\text{int}}(\mathcal{C}_\pi)$、$\Lambda^{\text{ex}}(\mathcal{C}_\pi)$、$\Gamma^{\text{int}}(\mathcal{C}_\pi)$和 $\mathcal{F}(\mathcal{C}_\pi)$分别引入整数编码。令 $[\mathtt{a}]^{\text{int}}$（分别地，$[\mathtt{a}]^{\text{ex}}$）表示元素$\mathtt{a} \in \Lambda^{\text{int}}(\mathcal{C}_\pi)$（分别地，$\mathtt{a} \in \Lambda^{\text{ex}}(\mathcal{C}_\pi)$）的编码整数；令 $[\gamma]$ 表示 $\Gamma^{\text{int}}(\mathcal{C}_\pi)$中的元素 γ 所对应的编码整数，并令 $[\psi]$表示 $\mathcal{F}(\mathcal{C}_\pi)$中的一个元素 ψ。

PubChem 中非氢原子数不超过 100 的化合物有超过 99% 在氢抑制图 $\langle\mathbb{C}\rangle$ [24] 中的度至多为 4。我们假设，本文研究的化学图 $\mathbb{C}$ 在氢抑制图 $\langle\mathbb{C}\rangle$ 中满足 $\deg^{\langle\mathbb{C}\rangle}(v) \leq 4$。

I在我们的模型中，对于每个 $\mathtt{a} \in$，我们使用整数质量$^*(\mathtt{a}) = \lfloor 10 \cdot \text{mass}(\mathtt{a}) \rfloor \quad \Lambda.$

对于化学性质 π，我们将化学图 $\mathbb{C} = (H, \alpha, \beta) \in \mathcal{C}_\pi$ 的描述符集合 $D_\pi^{(1)}$ 定义为以下非负整数 $\text{dcp}_i(\mathbb{C})$，$i \in [1, K_1]$，其中 $K_1 = 14 + |\Lambda^{\text{int}}(\mathcal{C}_\pi)| + |\Lambda^{\text{ex}}(\mathcal{C}_\pi)| + |\Gamma^{\text{int}}(\mathcal{C}_\pi)| + |\mathcal{F}(\mathcal{C}_\pi)| + |\Gamma_{\text{ac}}^{\text{lf}}|$。

1. $\text{dcp}_1(\mathbb{C})$：$\mathbb{C}$中非氢原子的数量为 $|V(H)| - |V_{\mathtt{H}}|$。

2. $\text{dcp}_2(\mathbb{C})$：$\mathbb{C}$ 的秩（即使图变为无环所需删除的最少边数）。

3. $\text{dcp}_3(\mathbb{C})$：$\mathbb{C}$中内部顶点的数量为 $|V^{\text{int}}(\mathbb{C})|$。

4. $\text{dcp}_4(\mathbb{C})$：所有原子的~~质~~量* 的平均值 $\text{ms}(\mathbb{C})$，即 $\text{ms}(\mathbb{C})$ $\frac{1}{|V(H)|}\sum_{v \in V(H)}$ ~~质~~量$^*(\alpha(v))$。

5. $\text{dcp}_i(\mathbb{C})$，$i = 4 + d, d \in [1, 4]$：氢抑制化学图 $\langle\mathbb{C}\rangle$中度为 $\deg^{\langle\mathbb{C}\rangle}(v) = d$ 的非氢顶点 $v \in V(H) \setminus V_{\mathtt{H}}$ 的数量 $\text{dg}_d^{\mathtt{H}}(\mathbb{C})$。

6. $\text{dcp}_i(\mathbb{C})$，$i = 8 + d, d \in [1, 4]$：$\mathbb{C}$的内部 $\mathbb{C}^{\text{int}} = (V_{\text{int}}(\mathbb{C}), E^{\text{int}}(\mathbb{C}))$中内部度为 $\deg_{\mathbb{C}^{\text{int}}}(v) = d$ 的内部顶点数量 $\text{dg}_d^{\text{int}}(\mathbb{C})$。

7. $\text{dcp}_i(\mathbb{C})$，$i = 12 + m, m \in [2, 3]$：$\mathbb{C}$中键多重度为m 的内部边数量 $\text{bd}_m^{\text{int}}(\mathbb{C})$；即 $\text{bd}_m^{\text{int}}(\mathbb{C}) |\{e \in E^{\text{int}}(\mathbb{C}) \mid \beta(e) = m\}|$。

8. $\text{dcp}_i(\mathbb{C})$，$i = 14 + [\mathtt{a}]^{\text{int}}, \mathtt{a} \in \Lambda^{\text{int}}(\mathcal{C}_\pi)$：化学元素 $\mathtt{a}$ 在集合 $V^{\text{int}}(\mathbb{C})$ 中的频数 $\text{na}_{\mathtt{a}}^{\text{int}}(\mathbb{C}) = |V_{\mathtt{a}}(\mathbb{C}) \cap V_{\text{内部}}(\mathbb{C})|$；该集合由 $\mathbb{C}$ 中的内部顶点构成。

9. $\text{dcp}_i(\mathbb{C})$，$i = 14 + |\Lambda^{\text{int}}(\mathcal{C}_\pi)| + [\mathtt{a}]^{\text{ex}}, \mathtt{a} \in \Lambda^{\text{ex}}(\mathcal{C}_\pi)$：化学元素 $\mathtt{a}$ 在集合 $V^{\text{ex}}(\mathbb{C})$ 中的频数 $\text{na}_{\mathtt{a}}^{\text{ex}}(\mathbb{C}) = |V_{\mathtt{a}}(\mathbb{C}) \cap V^{\text{ex}}(\mathbb{C})|$；该集合由 $\mathbb{C}$ 中的外部顶点构成。

10. $\text{dcp}_i(\mathbb{C})$，$i = 14 + |\Lambda^{\text{int}}(\mathcal{C}_\pi)| + |\Lambda^{\text{ex}}(\mathcal{C}_\pi)| + [\gamma], \gamma \in \Gamma^{\text{int}}(\mathcal{C}_\pi)$：$\mathbb{C}$ 中内部边集合 $E^{\text{int}}(\mathbb{C})$ 内的边构型 γ 的频数 $\text{ec}_\gamma(\mathbb{C})$。

11. $\text{dcp}_i(\mathbb{C})$，$i = 14 + |\Lambda^{\text{int}}(\mathcal{C}_\pi)| + |\Lambda^{\text{ex}}(\mathcal{C}_\pi)| + |\Gamma^{\text{int}}(\mathcal{C}_\pi)| + [\psi], \psi \in \mathcal{F}(\mathcal{C}_\pi)$：在 $\mathbb{C}$中 ρ-悬挂树集合里，边缘构型 ψ 的频数 $\text{fc}_\psi(\{\text{v13})$。

12. $\text{dcp}_i(\mathbb{C})$，$i = 14 + |\Lambda^{\text{int}}(\mathcal{C}_\pi)| + |\Lambda^{\text{ex}}(\mathcal{C}_\pi)| + |\Gamma^{\text{int}}(\mathcal{C}_\pi)| + |\mathcal{F}(\mathcal{C}_\pi)| + [\nu], \nu \in \Gamma_{\text{ac}}^{\text{lf}}$：$\langle\mathbb{C}\rangle$中叶边集合里邻接构型 ν 的频数 $\text{ac}_\nu^{\text{lf}}(\mathbb{C})$。

本文还使用了一种生成二次描述符的方法。为此，我们首先通过将最小值和最大值分别缩放至 0 和 1，把每个描述符 dcp$_i(\mathbb{C})$ $i \in [1, K_1]$ 归一化为介于 0 和 1 之间的值 $x(i)$。然后构造集合 $D_\pi^{(2)} := \{x(i)x(j) \mid 1 \le i \le j \le K_1\} \cup \{x(i)(1-x(j)) \mid i, j \in [1, K_1]\}$，其中包含 $(3K_1^2 + K_1)/2$ 二次描述符。随后，我们按照 Zhu 等人提出的程序 [28]，将并集 $D_\pi^{(1)} \cup D_\pi^{(2)}$缩减为一个子集，以构造预测函数。

B 指定目标化学图

给定预测函数 η 和目标值 $y^* \in \mathbb{R}$，若化学图 $\mathbb{C}^*$ 满足$\eta(x^*) = y^*$，其中特征向量为 $x^* = f(\mathbb{C}^*)$，则称其为目标化学图。本节介绍一组规则，用于在框架第二阶段以灵活的方式指定目标化学图的拓扑子结构。

我们首先说明如何将化学图 $\mathbb{C} = (H, \alpha, \beta)$化简为一种抽象形式，并以此为基础定义指定规则。为说明这一化简过程，我们使用化学图 $\mathbb{C} = (H, \alpha, \beta)$，其 $\langle\mathbb{C}\rangle$ 如图2所示。

R1 移除所有 ρ-枝树：通过移除每棵 ρ-枝树 $\mathbb{C}[u] \in \mathcal{T}(\mathbb{C})$ $u \in V_{\text{int}}(\mathbb{C})$的非根顶点，可得到 $\mathbb{C}$ 的内部 $H^{\text{int}} = (V_{\text{int}}(\mathbb{C})\ E^{\text{int}}(\mathbb{C}))$。图 8 展示了图2中带有 $\rho = 2$ 的化学图 $\mathbb{C}$ 的内部 H^{int}。

R2 移除一些叶路径：如果顶点 v 是 H^{int} 的叶顶点，且 H^{int} 中 Q 的每个内部顶点的度均为 2，并且我们将 Q 视为以顶点 u 为根，那么我们称 H^{int} 中从 u 到 v 的路径 Q 为叶路径。若连通子图 S 可由内部 H^{int} 中的子图 $\mathbb{C}$ 移除 $V(Q_u)$ 中的顶点 $H \setminus \{u\}, u \in X$ 得到，其中 X 是内部顶点的子集，$\{Q_u \mid u \in X\}$ 是叶 u, v-路径 Q_u 的集合，且任意两条路径 Q_u 和 $Q_{u'}$ 不共享顶点，则称 S 为循环基。图 9(a) 展示了图 8 中一组叶路径 $\{Q_{u_5} = (u_5, u_{24})$ $Q_{u_{18}} = (u_{18}, u_{25}, u_{26}, u_{27})$ $Q_{u_{22}} = (u_{22}, u_{28})\}$ 对应的内部 H^{int} 的循环基 $S = H^{\text{int}} - \bigcup_{u \in X}(V(Q_u) \setminus \{u\})$。

R3 某些纯路径的收缩：如果 S 中一条路径的每个内部顶点的度均为 2，则称该路径为纯路径。在 S 中选取一组 $\mathcal{P}$ 纯路径，使得任意两条路径除了端点外不共享顶点。如果图 S' 是通过将 S 中每条纯 u、v-路径替换为一条边 $a = uv$ 而从图 S 得到的，则称图 S' 为图 S （相对于 $\mathcal{P}$）的收缩，其中 S' 可能在同一对相邻顶点之间包含重边。图 9(b) 展示了一个收缩图 S'，它由化学图 S 收缩每条 uv-路径 $P_a \in \mathcal{P}$ 为一条新边 $a = uv$ 而得到，其中 $a_1 = u_1u_2$、$a_2 = u_1u_3$、$a_3 = u_4u_7$、$a_4 = u_{10}u_{11}$ 和 $a_5 = u_{11}u_{12}$ 以及 $\mathcal{P} = \{P_{a_1} = (u_1$、$u_{13}$、$u_2)$ $P_{a_2} = (u_1$、u_{14}、$u_3)$ $P_{a_3} = (u_4$、u_{15}、u_{16}、$u_7)$ $P_{a_4} = (u_{10}$、u_{17}、u_{18}、u_{19}、$u_{11})$ $P_{a_5} = (u_{11}$、u_{20}、u_{21}、u_{22}、$u_{12})\}$ 是图 9(a) 中的纯路径。

我们将定义一组规则，使得通过反向应用过程 R3 至 R1，可以从一个图（下一节中称为种子图）得到一个化学图。我们指定

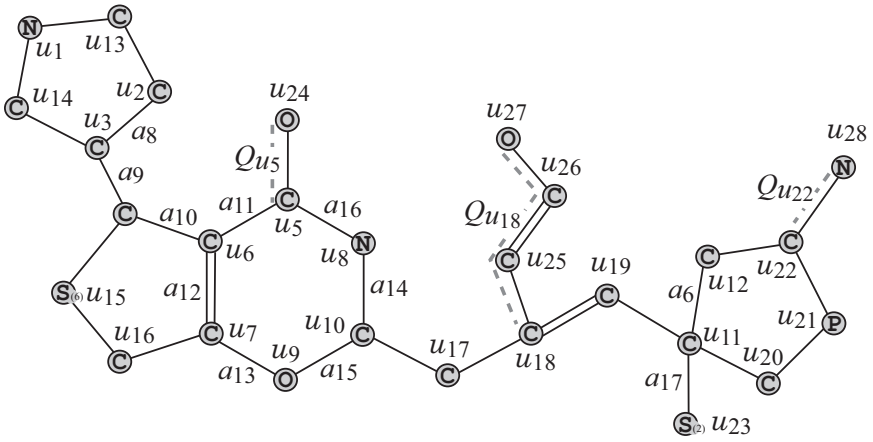
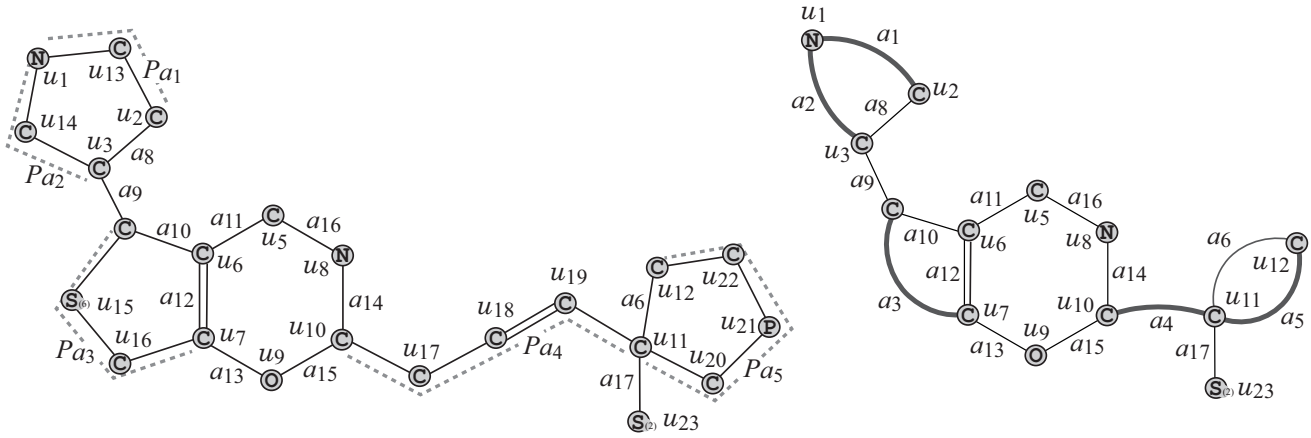


图 8：图2中化学图 $\mathbb{C}$ 的内部 H^{int}，其中含有 $\langle\mathbb{C}\rangle$，对应于 $\rho = 2$。



(a) {style id='1'>S</style} 的循环基 {style id='3'>H</style}int

(b) {style id='1'>S'</style} 的收缩 {style id='3'>S</style}

图9：(a) 图 8 中内部 H^{int} 的循环基 $S = H^{\text{int}} - \bigcup_{u\in\{u_5,u_{18},u_{22}\}}(V(Q_u) \setminus \{u\})$；(b) (a) 中纯路径集 $\mathcal{P} = \{P_{a_1}, P_{a_2}, \ldots, P_{a_5}\}$ 的 S 的收缩 S'，其中收缩纯路径得到的新边以粗线表示。

目标化学图的拓扑子结构，并用元组（$G_{\text{C}}, \sigma_{\text{int}}, \sigma_{\text{ce}}$）表示，称为目标规格，其定义遵循以下规则。

种子图

种子图 $G_{\text{C}} = (V_{\text{C}}, E_{\text{C}})$ 定义为一种图（可能包含重边），其边集 E_{C} 由四个集合 $E_{(\geq 2)}$、$E_{(\geq 1)}$、$E_{(0/1)}$ 和 $E_{(=1)}$组成，其中每个集合都可以为空。种子图在 R3 中起到最抽象形式 S' 的作用。图4(a)展示了一个种子图 G_{C}的示例，其中 $V_{\text{C}} = \{u_1, u_2, \ldots, u_{12}, u_{23}\}$、$E_{(\geq 2)} = \{a_1, a_2, \ldots, a_5\}$、$E_{(\geq 1)} = \{a_6\}$、$E_{(0/1)} = \{a_7\}$ 和 $E_{(=1)} = \{a_8, a_9, \ldots, a_{16}\}$。

.....的细分 S G_{C} 是根据以下规则从种子图 G_{C} 构建的图

:

- 每条边 $e = uv \in E_{(\geq 2)}$ 都替换为一条长度至少为 2 的 u, v-路径 P_e；

- 每条边 $e = uv \in E_{(\geq 1)}$ 都替换为一条长度至少为 1 的 u, v-路径 P_e（等价地，直接使用 e，或将其替换为一条长度至少为 2 的 u, v-路径 P_e）；- 每条边 $e \in E_{(0/1)}$ 要么使用，要么舍弃，其中 $E_{(0/1)}$ 必须选作 $E(G_{\mathrm{C}})$ 的非分离边子集，否则无法保证最终化学图 $\mathbb{C}$ 的连通性；- 每条边 $e \in E_{(=1)}$ 始终直接使用。

在从种子图构造目标化学图时，我们允许将边的消除作为一条可选规则应用于 $E_{(0/1)}$，尽管过程 R3 中并未包含此类操作。细分 S 在 R2 中起到循环基的作用。目标化学图 $\mathbb{C} = (H, \alpha, \beta)$ 将包含 S，作为 $\mathbb{C}$ 的内部 H^{int} 的子图。

内部规范

作为目标化学图 $\mathbb{C}$ 的内部 H^{int} 的图 H^* 将按如下方式构造。首先，通过将每条边 $e = uu' \in E_{(\geq 2)} \cup E_{(\geq 1)}$ 替换为一条纯的 u, u'-路径 P_e，构造种子图 G_{C} 的细分 S。接下来，通过在每个顶点 $v \in V_{\mathrm{C}}$ 或每条纯的 u, u'-路径P_e 的某个内部顶点 $v \in V(P_e) \setminus \{u, u'\}$ 处附加一条叶路径 Q_v，构造 S 的超图 H^*，其中边 $e = uu' \in E_{(\geq 2)} \cup E_{(\geq 1)}$可能满足 $Q_v = (v)$ $E(Q_v) = \emptyset$（即，不向 v附加任何新边）。我们引入以下规则，用于规定 H^*的大小、纯路径 P_e的长度 $|E(P_e)|$、叶路径 Q_v的长度 $|E(Q_v)|$、叶路径的数量Q_v 以及每条内部边的键多重性；我们将所规定常数的集合称为内部规范 σ_{int}:

- 目标化学图 $\mathbb{C}$ 的内部顶点数量下界 $\mathsf{n}_{\mathrm{LB}}^{\mathrm{int}}$、上界 $\mathsf{n}_{\mathrm{UB}}^{\mathrm{int}} \in \mathbb{Z}_+$。

- 对于每条边 $e = uu' \in E_{(\geq 2)} \cup E_{(\geq 1)}$,

  纯 u, u'-路径P_e 的长度 $|E(P_e)|$ 的下界 $\ell_{\mathrm{LB}}(e)$ 和上界 $\ell_{\mathrm{UB}}(e)$。（为方便记号，设 $\ell_{\mathrm{LB}}(e) := 0$, $\ell_{\mathrm{UB}}(e) := 1$, $e \in E_{(0/1)}$ 和 $\ell_{\mathrm{LB}}(e) := 1$,$\ell_{\mathrm{UB}}(e) := 1$, $e \in E_{(=1)}$。）

  附着在纯 u, u'-路径 P_e 的内部顶点 v 上的叶路径 Q_v 数量的下界 $\mathsf{bl}_{\mathrm{LB}}(e)$ 和上界 $\mathsf{bl}_{\mathrm{UB}}(e)$。

  附着在纯 u, u'-路径 P_e 的内部顶点 $v \in V(P_e) \setminus \{u, u'\}$ 上的叶路径 Q_v 最大长度 $|E(Q_v)|$ 的下界 $\mathsf{ch}_{\mathrm{LB}}(e)$ 和上界 $\mathsf{ch}_{\mathrm{UB}}(e)$。

- 对于每个顶点 $v \in V_{\mathrm{C}}$,

  附加到 v 的叶路径 Q_v 数量的下界 $\mathsf{bl}_{\mathrm{LB}}(v)$ 和上界 $\mathsf{bl}_{\mathrm{UB}}(v)$，其中 $0 \leq \mathsf{bl}_{\mathrm{LB}}(v) \leq \mathsf{bl}_{\mathrm{UB}}(v) \leq 1$。

一个叶路径 Q_v（连接至 v）的长度 $|E(Q_v)|$ 的下界 $\mathsf{ch}_{\mathrm{LB}}(v)$ 和上界 $\mathsf{ch}_{\mathrm{UB}}(v)$。

- 对于每条边 $e = uu' \in E_{\mathrm{C}}$，边数的下界 $\mathsf{bd}_{m,\ \mathrm{LB}}(e)$ 和上界 $\mathsf{bd}_{m,\ \mathrm{UB}}(e)$，其中键多重性为 $m \in [2,3]$ in u, u'-路径 P_e 的边数；这里我们将 P_e,$e \in E_{(0/1)} \cup E_{(=1)}$ 视为单条边 e。

我们将满足内部规范 σ_{int} 的图 H^* 称为 G_{C} 的 σ_{int}-扩展，其中每条边的键多重性均已确定。

表6展示了一个　图4中的种子图 G_{C} 的内部规范 σ_{int} 示例　.

表6：内部规范 σ_{int} 的示例1。

$\mathrm{n}_{\mathrm{LB}}^{\mathrm{int}} = 20$	$\mathrm{n}_{\mathrm{UB}}^{\mathrm{int}} = 28$

	a_1	a_2	a_3	a_4	a_5	a_6
$\ell_{\mathrm{LB}}(a_i)$	2	2	2	3	2	1
$\ell_{\mathrm{UB}}(a_i)$	3	4	3	5	4	4
$\mathsf{bl}_{\mathrm{LB}}(a_i)$	0	0	0	1	1	0
$\mathsf{bl}_{\mathrm{UB}}(a_i)$	1	1	0	2	1	0
$\mathsf{ch}_{\mathrm{LB}}(a_i)$	0	1	0	4	3	0
$\mathsf{ch}_{\mathrm{UB}}(a_i)$	3	3	1	6	5	2

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{23}
$\mathsf{bl}_{\mathrm{LB}}(u_i)$	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathsf{bl}_{\mathrm{UB}}(u_i)$	1	1	1	1	1	0	0	0	0	0	0	0	0
$\mathsf{ch}_{\mathrm{LB}}(u_i)$	0	0	0	0	1	0	0	0	0	0	0	0	0
$\mathsf{ch}_{\mathrm{UB}}(u_i)$	1	0	0	0	3	0	1	1	0	1	2	4	1

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}
$\mathsf{bd}_{2,\ \mathrm{LB}}(a_i)$	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
$\mathsf{bd}_{2,\mathrm{UB}}(a_i)$	1	1	0	2	2	0	0	0	0	0	0	1	0	0	0	0	0
$\mathsf{bd}_{3,\mathrm{LB}}(a_i)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathsf{bd}_{3,\ \mathrm{UB}}(a_i)$	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

图10展示了图4中的种子图 G_{C} 在表6中的内部规范 σ_{int} 下的一种 σ_{int}-扩展 H^* 示例。

化学规范

设 H^* 是一个图，作为目标化学图 $\mathbb{C}$ 的内部 H^{int}，其中 H^* 中每条边的键多重性均已确定。最后，我们介绍一组规则，用于通过选择化学元素 $\mathtt{a} \in \Lambda$，并为每个内部顶点 $v \in V^{\mathrm{int}}$ 指定一个 ρ-枝树 ψ，从 H^* 构造目标化学图 $\mathbb{C}$。我们介绍以下规则，用于规定 $\mathbb{C}$ 的大小、允许用作 ρ-悬挂树的一组化学根树，以及化学元素、化学符号和边构型的频数上下界。我们将这组预先设定的常数称为化学规格 σ_{ce}:

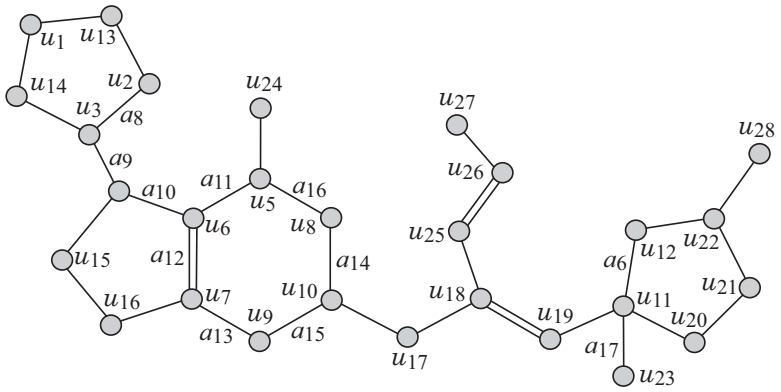


图10：图4中的种子图 G_{C} 在内部规范 σ_{int} （见表6）下得到的图 H^* 的示意图。

- 顶点数量的上下界 $n_{\mathrm{LB}}, n^* \in \mathbb{Z}_+$，其中 $\mathbf{n}_{\mathrm{LB}}^{\mathrm{int}} \leq n_{\mathrm{LB}} \leq n^*$。

- 化学根树 ψ 中满足 $\mathsf{ht}(\langle\psi\rangle) \leq \rho$ 的子集 $\mathcal{F}(v) \subseteq \mathcal{F}(D_\pi)$ $v \in V_{\mathrm{C}}$ 和 $\mathcal{F}_E \subseteq \mathcal{F}(D_\pi)$，其中要求：以 $v \in V_{\mathrm{C}}$（分别以不属于 V_{C} 的内部顶点 v）为根的每个 ρ-枝树 $\mathbb{C}[v]$ 均属于 $\mathcal{F}(v)$（分别属于 $\mathcal{F}_E$）。令 $\mathcal{F}^*:=\mathcal{F}_E \cup \bigcup_{v\in V_{\mathrm{C}}} \mathcal{F}(v)$ 和 Λ^{ex} 表示 $\mathcal{F}^*$ 中所有化学根树里分配给非根顶点的化学元素集合。

- 子集 $\Lambda^{\mathrm{int}} \subseteq \Lambda^{\mathrm{int}}(D_\pi)$，其中要求：分配给 $\mathbb{C}$ 中内部顶点 v 的每个化学元素 $\alpha(v)$ 均属于 Λ^{int}。令 $\Lambda := \Lambda^{\mathrm{int}} \cup \Lambda^{\mathrm{ex}}$ 和 $\mathsf{na}_{\mathsf{a}}(\mathbb{C})$（分别为 $\mathsf{na}_{\mathsf{a}}^{\mathrm{int}}(\mathbb{C})$ 和 $\mathsf{na}_{\mathsf{a}}^{\mathrm{ex}}(\mathbb{C})$）表示满足 $\alpha(v) = \mathsf{a}$ 于 $\mathbb{C}$ 的顶点数（分别为内部顶点数和外部顶点数）。

- 一组化学符号 $\Lambda_{\mathrm{dg}}^{\mathrm{int}} \subseteq \Lambda \times [1,4]$ 和一组边构型 $\Gamma^{\mathrm{int}} \subseteq \Gamma^{\mathrm{int}}(D_\pi)$（$\mu$、$\mu'$、$m$），其中 $\mu \leq \mu'$；我们要求，$\mathbb{C}$ 中内部边 e 的边构型 $\mathsf{ec}(e)$ 属于 Γ^{int}。我们不区分（μ、μ'、m）和（μ'、μ、m）。

- 定义 $\Gamma_{\mathrm{ac}}^{\mathrm{int}}$ 为满足以下条件的邻接构型集合：$\Gamma_{\mathrm{ac}}^{\mathrm{int}} := \{(\mathsf{a}, \mathsf{b}, m) \mid (\mathsf{a}d, \mathsf{b}d', m) \in \Gamma^{\mathrm{int}}\}$。令 $\mathsf{ac}_\nu^{\mathrm{int}}(\mathbb{C})$ $\nu \in \Gamma_{\mathrm{ac}}^{\mathrm{int}}$ 表示满足 $\mathsf{ac}(e) = \nu$ in $\mathbb{C}$ 的内部边 e 的数量。

- 子集 $\Lambda^*(v) \subseteq \{\mathsf{a} \in \Lambda^{\mathrm{int}} \mid \mathsf{val}(\mathsf{a}) \geq 2\}$, $v \in V_{\mathrm{C}}$，我们要求，分配给种子图中顶点 $v \in V_{\mathrm{C}}$ 的每个化学元素 $\alpha(v)$ 都属于 $\Lambda^*(v)$。

- 关于内部顶点数量的下界函数和上界函数 $\mathsf{na}_{\mathrm{LB}}, \mathsf{na}_{\mathrm{UB}}: \Lambda \to [1, n^*]$ 和 $\mathsf{na}_{\mathrm{LB}}^{\mathrm{int}}, \mathsf{na}_{\mathrm{UB}}^{\mathrm{int}}:$ $\Lambda^{\mathrm{int}} \to [1, n^*]$，其中 v 满足 $\alpha(v) = \mathsf{a}$，且属于 $\mathbb{C}$。

- 下界函数和上界函数 $\mathsf{ns}_{\mathrm{LB}}^{\mathrm{int}}, \mathsf{ns}_{\mathrm{UB}}^{\mathrm{int}}$：关于 $\mathbb{C}$ 中满足 $\mathsf{cs}(v) = \mu$ 的内部顶点数量v 的 $\Lambda_{\mathrm{dg}}^{\mathrm{int}} \to [1, n^*]$。

- 关于内部边数量的下界函数和上界函数 $\mathsf{ac}_{\mathrm{LB}}^{\mathrm{int}}, \mathsf{ac}_{\mathrm{UB}}^{\mathrm{int}}$：$\Gamma_{\mathrm{ac}}^{\mathrm{int}} \to \mathbb{Z}_+$，满足 $\mathsf{ac}(e) = \nu$ 属于 $\mathbb{C}$ 的 e。

- 下界函数和上界函数 $\mathsf{ec}_{\mathrm{LB}}^{\mathrm{int}}$,$\mathsf{ec}_{\mathrm{UB}}^{\mathrm{int}}$： $\Gamma^{\mathrm{int}} \to \mathbb{Z}_+$ 内部边的数量 e，使得 $\mathsf{ec}(e) = \gamma$ 在 $\mathbb{C}$ 中。

- 关于内部顶点数量的下界函数和上界函数 $\mathsf{fc}_{\mathrm{LB}}$, $\mathsf{fc}_{\mathrm{UB}}$： $\mathcal{F}^* \to [0, n^*]$，使得 $\mathbb{C}[v]$ 在 $\mathbb{C}$ 中与 $\psi \in \mathcal{F}^*$ r-同构。

- 关于具有邻接构型 ν 的 ac_{C} 中叶边数量的下界函数和上界函数 $\mathsf{ac}_{\mathrm{LB}}^{\mathrm{lf}}$, $\mathsf{ac}_{\mathrm{UB}}^{\mathrm{lf}}$： $\Gamma_{\mathrm{ac}}^{\mathrm{lf}} \to [0, n^*]$ uv。

我们将满足化学规格 σ_{ce} 的化学图 $\mathbb{C}$ 称为 G_{C} 的 $(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$-扩展，并用 $\mathcal{G}(G_{\mathrm{C}}, \sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$ 表示所有 G_{C} 的 $(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$-扩展的集合。

表7展示了一个　图4中种子图 G_{C} 的化学规范 σ_{ce} 示例 .

图2展示了一个示例 $\mathbb{C}$，说明如何根据表7中的化学规范 σ_{ce}，由图10中的 σ_{int}-扩展 H^* 得到 G_{C} 的$(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$-扩展。

C　用于化学图推断的测试实例

我们准备了以下实例(a)-(d)，用于开展框架第二阶段的实验。

在化学图推断的第二阶段，我们使用两个性质 $\pi \in \{\textsc{At}, \textsc{FLmL}\}$，并将化学元素集合 $\Lambda(\pi)$ 定义如下： $\Lambda(\textsc{At}) = \Lambda_3 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{S}_{(2)}, \mathtt{S}_{(6)}\}$ 和$\Lambda(\textsc{FLmL}) = \Lambda_6 = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{Cl}, \mathtt{P}_{(2)}, \mathtt{P}_{(5)}\}$。

(a) $I_{\mathrm{a}} = (G_{\mathrm{C}}, \sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$：B节中为说明目标规范而介绍的实例。对于每个性质 π，我们将表7中的 $\Lambda = \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{S}_{(2)}, \mathtt{S}_{(6)}, \mathtt{P}_{(5)}\}$ 替换为 $\Lambda(\pi) \cap \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}, \mathtt{S}_{(2)}, \mathtt{S}_{(6)}, \mathtt{P}_{(5)}\}$，并从 σ_{ce} 中删除所有无法由替换后的元素集构造的化学符号、边构型和边缘配置（即包含化学元素 $\mathtt{P}_{(5)}$的项目）。

(b) $I_{\mathrm{b}}^i = (G_{\mathrm{C}}^i, \sigma_{\mathrm{int}}^i, \sigma_{\mathrm{ce}}^i)$, $i = 1, 2, 3, 4$：一个秩至多为 2 的化学图推断实例。在这四个实例 I_{b}^i, $i = 1, 2, 3, 4$ 中，$(\sigma_{\mathrm{int}}, \sigma_{\mathrm{ce}})$ 中的以下规格是共同的。

对于给定性质 $\pi \in \{\textsc{At}, \textsc{FLmL}\}$，设 $\Lambda := \Lambda(\pi)$，设 $\Lambda_{\mathrm{dg}}^{\mathrm{int}}$ 为数据集 $\mathcal{C}_\pi$ 中出现的 $\Lambda \times [1, 4]$ 里的所有可能符号的集合，并设 Γ^{int} 为数据集 $\mathcal{C}_\pi$ 中出现的所有边构型的集合。设 $\Lambda^*(v) := \Lambda$, $v \in V_{\mathrm{C}}$。

下界 ℓ_{LB}、$\mathsf{bl}_{\mathrm{LB}}$、$\mathsf{ch}_{\mathrm{LB}}$、$\mathsf{bd}_{2,\mathrm{LB}}$、$\mathsf{bd}_{3,\mathrm{LB}}$、$\mathsf{na}_{\mathrm{LB}}$、$\mathsf{na}_{\mathrm{LB}}^{\mathrm{int}}$、$\mathsf{ns}_{\mathrm{LB}}^{\mathrm{int}}$、$\mathsf{ac}_{\mathrm{LB}}^{\mathrm{int}}$、$\mathsf{ec}_{\mathrm{LB}}^{\mathrm{int}}$ 和 $\mathsf{ac}_{\mathrm{LB}}^{\mathrm{lf}}$ 均设为 0。

设置上界 $\mathsf{na}_{\mathrm{UB}}(\mathtt{a}) := n^*$, $\mathsf{na} \in \{\mathtt{H}, \mathtt{C}\}$, $\mathsf{na}_{\mathrm{UB}}(\mathtt{a}) := 5$, $\mathsf{na} \in \{\mathtt{O}, \mathtt{N}\}$, $\mathsf{na}_{\mathrm{UB}}(\mathtt{a}) :=2$, $\mathsf{na} \in \Lambda \setminus \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}\}$。其他上界 ℓ_{UB}, $\mathsf{bl}_{\mathrm{UB}}$, $\mathsf{ch}_{\mathrm{UB}}$, $\mathsf{bd}_{2,\mathrm{UB}}$, $\mathsf{bd}_{3,\mathrm{UB}}$, $\mathsf{na}_{\mathrm{UB}}^{\mathrm{int}}$, $\mathsf{ns}_{\mathrm{UB}}^{\mathrm{int}}$, $\mathsf{ac}_{\mathrm{UB}}^{\mathrm{int}}$, $\mathsf{ec}_{\mathrm{UB}}^{\mathrm{int}}$ 和 $\mathsf{ac}_{\mathrm{UB}}^{\mathrm{lf}}$ 均设置为 $n(G^*)$ 的上界 n^*。

我们将 n_{LB} 指定为参数，并设置 $n^* := n_{\mathrm{LB}} + 10$, $\mathsf{n}_{\mathrm{LB}}^{\mathrm{int}} := \lfloor (1/4) n_{\mathrm{LB}} \rfloor$ 和 $\mathsf{n}_{\mathrm{UB}}^{\mathrm{int}} := \lfloor (3/4) n_{\mathrm{LB}} \rfloor$。

表7：化学规范 σ_{ce}的示例2。

$n_{\mathrm{LB}} = 30$, $n^* = 50$。
分支参数： $\rho = 2$

集合 $\mathcal{F}(v)$ $v \in V_{\mathrm{C}}$和 $\mathcal{F}_E$中的每一个都设为 图4(b)中高度为 $\mathrm{ht}(\langle\psi\rangle) \leq \rho = 2$ 的化学根树 $\mathcal{F}$ ψ 的集合。

$\Lambda = \{\mathtt{H}, \mathtt{C}, \mathtt{N}, \mathtt{O}, \mathtt{S}_{(2)}, \mathtt{S}_{(6)}, \mathtt{P} = \mathtt{P}_{(5)}\}$	$\Lambda^{\mathrm{int}}_{\mathrm{dg}} = \{\mathtt{C2}, \mathtt{C3}, \mathtt{C4}, \mathtt{N2}, \mathtt{N3}, \mathtt{O2}, \mathtt{S}_{(2)}2, \mathtt{S}_{(6)}3, \mathtt{P4}\}$

$\Gamma^{\mathrm{int}}_{\mathrm{ac}}$	$\nu_1 = (\mathtt{C}, \mathtt{C}, 1)$ $\nu_2 = (\mathtt{C}, \mathtt{C}, 2)$ $\nu_3 = (\mathtt{C}, \mathtt{N}, 1)$ $\nu_4 = (\mathtt{C}, \mathtt{O}, 1)$ $\nu_5 = (\mathtt{C}, \mathtt{S}_{(2)}, 1)$ $\nu_6 = (\mathtt{C}, \mathtt{S}_{(6)}, 1)$ $\nu_7 = (\mathtt{C}, \mathtt{P}, 1)$
Γ^{int}	$\gamma_1 = (\mathtt{C2}, \mathtt{C2}, 1)$ $\gamma_2 = (\mathtt{C2}, \mathtt{C3}, 1)$ $\gamma_3 = (\mathtt{C2}, \mathtt{C3}, 2)$ $\gamma_4 = (\mathtt{C2}, \mathtt{C4}, 1)$ $\gamma_5 = (\mathtt{C3}, \mathtt{C3}, 1)$ $\gamma_6 = (\mathtt{C3}, \mathtt{C3}, 2)$, $\gamma_7 = (\mathtt{C3}, \mathtt{C4}, 1)$ $\gamma_8 = (\mathtt{C2}, \mathtt{N2}, 1)$ $\gamma_9 = (\mathtt{C3}, \mathtt{N2}, 1)$ $\gamma_{10} = (\mathtt{C3}, \mathtt{O2}, 1)$ $\gamma_{11} = (\mathtt{C2}, \mathtt{C2}, 2)$ $\gamma_{12} = (\mathtt{C2}, \mathtt{O2}, 1)$, $\gamma_{13} = (\mathtt{C3}, \mathtt{N3}, 1)$ $\gamma_{14} = (\mathtt{C4}, \mathtt{S}_{(2)}2, 2)$ $\gamma_{15} = (\mathtt{C2}, \mathtt{S}_{(6)}3, 1)$ $\gamma_{16} = (\mathtt{C3}, \mathtt{S}_{(6)}3, 1)$ $\gamma_{17} = (\mathtt{C2}, \mathtt{P4}, 2)$, $\gamma_{18} = (\mathtt{C3}, \mathtt{P4}, 1)$

$\Lambda^*(u_1) = \Lambda^*(u_8) = \{\mathtt{C}, \mathtt{N}\}$, $\Lambda^*(u_9) = \{\mathtt{C}, \mathtt{O}\}$, $\Lambda^*(u) = \{\mathtt{C}\}$, $u \in V_{\mathrm{C}} \setminus \{u_1, u_8, u_9\}$

	H	C	N	O	$\mathtt{S}_{(2)}$	$\mathtt{S}_{(6)}$	P
$\mathrm{na}_{\mathrm{LB}}(\mathtt{a})$	40	27	1	1	0	0	0
$\mathrm{na}_{\mathrm{UB}}(\mathtt{a})$	65	37	4	8	1	1	1

	C	N	O	$\mathtt{S}_{(2)}$	$\mathtt{S}_{(6)}$	P
$\mathrm{na}^{\mathrm{int}}_{\mathrm{LB}}(\mathtt{a})$	9	1	0	0	0	0
$\mathrm{na}^{\mathrm{int}}_{\mathrm{UB}}(\mathtt{a})$	23	4	5	1	1	1

	C2	C3	C4	N2	N3	O2	$\mathtt{S}_{(2)}2$	$\mathtt{S}_{(6)}3$	P4
$\mathrm{ns}^{\mathrm{int}}_{\mathrm{LB}}(\mu)$	3	5	0	0	0	0	0	0	0
$\mathrm{ns}^{\mathrm{int}}_{\mathrm{UB}}(\mu)$	8	15	2	2	3	5	1	1	1

	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	ν_7
$\mathrm{ac}^{\mathrm{int}}_{\mathrm{LB}}(\nu)$	0	0	0	0	0	0	0
$\mathrm{ac}^{\mathrm{int}}_{\mathrm{UB}}(\nu)$	30	10	10	10	1	1	1

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{16}	γ_{17}	γ_{18}
$\mathrm{ec}^{\mathrm{int}}_{\mathrm{LB}}(\gamma)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathrm{ec}^{\mathrm{int}}_{\mathrm{UB}}(\gamma)$	4	15	4	4	10	5	4	4	6	4	4	4	2	2	2	2	2	2

	$\psi \in \{\psi_i \mid i = 1, 6, 11\}$	$\psi \in \mathcal{F}^* \setminus \{\psi_i \mid i = 1, 6, 11\}$
$\mathrm{fc}_{\mathrm{LB}}(\psi)$	1	0
$\mathrm{fc}_{\mathrm{UB}}(\psi)$	10	3

	$\nu \in \{(\mathtt{C}, \mathtt{C}, 1), (\mathtt{C}, \mathtt{C}, 2)\}$	$\nu \in \Gamma^{\mathrm{lf}}_{\mathrm{ac}} \setminus \{(\mathtt{C}, \mathtt{C}, 1), (\mathtt{C}, \mathtt{C}, 2)\}$
$\mathrm{ac}^{\mathrm{lf}}_{\mathrm{LB}}(\nu)$	0	0
$\mathrm{ac}^{\mathrm{lf}}_{\mathrm{UB}}(\nu)$	10	8

对于每个性质 π，令 $\mathcal{F}(\mathcal{C}_\pi)$ 表示 $\mathcal{C}_\pi$ 中化合物里的 2-边缘树集合，并选取一个子集 $\mathcal{F}^i_\pi \subseteq \mathcal{F}(\mathcal{C}_\pi)$，满足 $|\mathcal{F}^i_\pi| = 45 - 5i$, $i \in [1, 5]$。对于每个实例 I^i_{b}，设置$\mathcal{F}_E := \mathcal{F}(v) := \mathcal{F}^i_\pi$, $v \in V_{\mathrm{C}}$以及 $\mathrm{fc}_{\mathrm{LB}}(\psi) := 0$, $\mathrm{fc}_{\mathrm{UB}}(\psi) := 10$, $\psi \in \mathcal{F}^i_\pi$。

实例 I^1_{b} 由图5(i)中的秩-1种子图 G^1_{C} 给出，实例 I^i_{b}、 $i = 2$、3、4由图5(ii)-(iv)中的秩-2种子图 G^i_{C}、 $i = 2$、3、4给出。

(i) 例如 I^1_{b}，选择图5(i)中的单环图 $G^1_{\mathrm{C}} = (V_{\mathrm{C}}, E_{\mathrm{C}} = E_{(\geq 2)} \cup E_{(\geq 1)})$ 作为种子图，其中 $V_{\mathrm{C}} = \{u_1, u_2\}$, $E_{(\geq 2)} = \{a_1\}$ and $E_{(\geq 1)} = \{a_2\}$。我们将线性约束 $\ell(a_1) \leq \ell(a_2)$ and $5 \leq \ell(a_1) + \ell(a_2) \leq 15$ 纳入侧约束。

(ii) 例如 I_{b}^{2}，选择图5(ii)中的图 $G_{\mathrm{C}}^{2}=(V_{\mathrm{C}},E_{\mathrm{C}}=E_{(\geq 2)}\cup E_{(\geq 1)}\cup E_{(=1)})$ 作为种子图，其中 $V_{\mathrm{C}}=\{u_1,u_2,u_3,u_4\}$, $E_{(\geq 2)}=\{a_1,a_2\}$, $E_{(\geq 1)}=\{a_3\}$ and $E_{(=1)}=\{a_4,a_5\}$。我们加入线性约束 $\ell(a_1)\leq\ell(a_2)$ and $\ell(a_1)+\ell(a_2)+\ell(a_3)\leq 15$。(iii) 例如 I_{b}^{3}，选择图5(iii)中的图 $G_{\mathrm{C}}^{3}=(V_{\mathrm{C}},E_{\mathrm{C}}=E_{(\geq 2)}\cup E_{(\geq 1)}\cup E_{(=1)})$ 作为种子图，其中 $V_{\mathrm{C}}=\{u_1,u_2,u_3,u_4\}$, $E_{(\geq 2)}=\{a_1\}$, $E_{(\geq 1)}=\{a_2,a_3\}$ and$E_{(=1)}=\{a_4,a_5\}$。我们加入线性约束 $\ell(a_1)\leq\ell(a_2)+\ell(a_3)$, $\ell(a_2)\leq\ell(a_3)$ and$\ell(a_1)+\ell(a_2)+\ell(a_3)\leq 15$。(iv) 例如 I_{b}^{4}，选择图5(iv)中的图 $G_{\mathrm{C}}^{4}=(V_{\mathrm{C}},E_{\mathrm{C}}=E_{(\geq 2)}\cup E_{(\geq 1)}\cup E_{(=1)})$ 作为种子图，其中 $V_{\mathrm{C}}=\{u_1,u_2,u_3,u_4\}$, $E_{(\geq 1)}=\{a_1,a_2,a_3\}$ and $E_{(=1)}=\{a_4,a_5\}$。我们加入线性约束 $\ell(a_2)\leq\ell(a_1)+1$, $\ell(a_2)\leq\ell(a_3)+1$, $\ell(a_1)\leq\ell(a_3)$ and$\ell(a_1)+\ell(a_2)+\ell(a_3)\leq 15$。

我们定义 (c) 和 (d) 中的实例，以寻找具有由给定的两个化学环图 $G_A=(H_A=(V_A,E_A)$ $\alpha_A,\beta_A)$ 和 $G_B=(H_B=(V_B,E_B)$ $\alpha_B,\beta_B)$ 构成的中间结构的化学图。令 Λ_A^{int} 和 $\Lambda_{\mathrm{dg},\ A}^{\mathrm{int}}$ 分别表示 G_A 中内部顶点的化学元素集和化学符号集，Γ_A^{int}表示G_A中内部边的边构型集，$\mathcal{F}_A$ 表示 G_A 中的 2-边缘树集。类似地，在 G_B中定义集合 Λ_B^{int}、$\Lambda_{\mathrm{dg},\ B}^{\mathrm{int}}$、$\Gamma_B^{\mathrm{int}}$和$\mathcal{F}_B$。

(c) $I_{\mathrm{c}}=(G_{\mathrm{C}},\sigma_{\mathrm{int}},\sigma_{\mathrm{ce}})$：旨在推断化学图 $G^{\dagger}$ 的一个实例，使得 $G^{\dagger}$ 的核心等于 G_A 的核心，并且$G^{\dagger}$ 的非核心中每种边构型的频数等于 G_B中的频数。我们使用图6(a)和(b)中的化合物 CID 24822711 和 CID 59170444，分别作为 G_A 和 G_B。将种子图 $G_{\mathrm{C}}=(V_{\mathrm{C}},E_{\mathrm{C}}=E_{(=1)})$ 设为 G_A的核心。设 $\Lambda:=\{\mathtt{H},\mathtt{C},\mathtt{N},\mathtt{O}\}$，并将 $\Lambda\times[1,4]$中所有可能的化学符号组成的集合设为$\Lambda_{\mathrm{dg}}^{\mathrm{int}}$。设 $\Gamma^{\mathrm{int}}:=\Gamma_A^{\mathrm{int}}\cup\Gamma_B^{\mathrm{int}}$，且 $\Lambda^{*}(v):=\{\alpha_A(v)\}$, $v\in V_{\mathrm{C}}$。设 $\mathsf{n}_{\mathrm{LB}}^{\mathrm{int}}:=\min\{\mathsf{n}^{\mathrm{int}}(G_A),\mathsf{n}^{\mathrm{int}}(G_B)\}$, $\mathsf{n}_{\mathrm{UB}}^{\mathrm{int}}:=\max\{\mathsf{n}^{\mathrm{int}}(G_A),\mathsf{n}^{\mathrm{int}}(G_B)\}$,$n_{\mathrm{LB}}:=\min\{n(G_A)\ n(G_B)\}-10=40$，以及 $n^{*}:=\max\{n(G_A)\ n(G_B)\}+5$。将下界 ℓ_{LB}、$\mathsf{bl}_{\mathrm{LB}}$、$\mathsf{ch}_{\mathrm{LB}}$、$\mathsf{bd2}_{,\mathrm{LB}}$、$\mathsf{bd3}_{,\mathrm{LB}}$、$\mathsf{na}_{\mathrm{LB}}$、$\mathsf{na}_{\mathrm{LB}}^{\mathrm{int}}$、$\mathsf{ns}_{\mathrm{LB}}^{\mathrm{int}}$、$\mathsf{ac}_{\mathrm{LB}}^{\mathrm{int}}$ 和 $\mathsf{ac}_{\mathrm{LB}}^{\mathrm{lf}}$ 设为 0。将上界 $\mathsf{na}_{\mathrm{UB}}(\mathtt{a}):=n^{*}$、$\mathsf{na}\in\{\mathtt{H},\mathtt{C}\}$、$\mathsf{na}_{\mathrm{UB}}(\mathtt{a}):=5$、$\mathsf{na}\in\{\mathtt{O},\mathtt{N}\}$、$\mathsf{na}_{\mathrm{UB}}(\mathtt{a}):=2$、$\mathsf{na}\in\Lambda\setminus\{\mathtt{H},\mathtt{C},\mathtt{O},\mathtt{N}\}$ 设定，并将其他上界 ℓ_{UB}、$\mathsf{bl}_{\mathrm{UB}}$、$\mathsf{ch}_{\mathrm{UB}}$、$\mathsf{bd2}_{,\mathrm{UB}}$、$\mathsf{bd3}_{,\mathrm{UB}}$、$\mathsf{na}_{\mathrm{UB}}^{\mathrm{int}}$、$\mathsf{ns}_{\mathrm{UB}}^{\mathrm{int}}$、$\mathsf{ac}_{\mathrm{UB}}^{\mathrm{int}}$ 和 $\mathsf{ac}_{\mathrm{UB}}^{\mathrm{lf}}$ 设为 n^{*}。令 $\mathsf{ec}_{\mathrm{LB}}^{\mathrm{int}}(\gamma)$ 表示 G_A 中满足 $\gamma\in\Gamma^{\mathrm{int}}$ 的核心边数，$\mathsf{ec}_{\mathrm{UB}}^{\mathrm{int}}(\gamma)$ 表示 G_A 中满足边构型 γ 的内部边数。令 $\mathcal{F}_B^{(p)}$、$p\in[1,2]$ 表示 G_B中与 p-悬挂树 r-同构的化学根树集合；设 $\mathcal{F}_E:=\mathcal{F}(v):=\mathcal{F}(1)_{\ B}\cup\mathcal{F}(2)_{\ B}$, $v\in V_{\mathrm{C}}$，以及 $\mathsf{fc}_{\mathrm{LB}}(\psi):=0$, $\mathsf{fc}_{\mathrm{UB}}(\psi):=10$, $\psi\in\mathcal{F}_B^{(1)}\cup\mathcal{F}(2)_{\ B}$。

(d) $I_{\mathrm{d}}=(G_{\mathrm{C}}^{1},\sigma_{\mathrm{int}},\sigma_{\mathrm{ce}})$：一个旨在推断化学单环图 $G^{\dagger}$ 的实例，使得 $G^{\dagger}$ 中边构型的频率向量是由合并 G_A 和 G_B的频率向量得到的向量。我们将图6(c)和(d)中的化学单环化合物 CID 10076784 和 CID 44340250 分别用于 G_A 和 G_B。将种子图设为图5(i)中的单环种子图 $G_{\mathrm{C}}^{1}=(V_{\mathrm{C}},E_{\mathrm{C}}=E_{(\geq 2)}\cup E_{(\geq 1)})$，其包含 $V_{\mathrm{C}}=\{u_1,u_2\}$、$E_{(\geq 2)}=\{a_1\}$ 和 $E_{(\geq 1)}=\{a_2\}$。

设 $\Lambda := \{\mathtt{H}, \mathtt{C}, \mathtt{N}, \mathtt{O}\}$, $\Lambda^{\text{int}}_{\text{dg}} := \Lambda^{\text{int}}_{\text{dg},A} \cup \Lambda^{\text{int}}_{\text{dg},B}$ 和 $\Gamma^{\text{int}} := \Gamma^{\text{int}}_A \cup \Gamma^{\text{int}}_B$。设 $\mathsf{n}^{\text{int}}_{\text{LB}} := \min\{\mathsf{n}^{\text{int}}(G_A), \mathsf{n}^{\text{int}}(G_B)\}$, $\mathsf{n}^{\text{int}}_{\text{UB}} := \max\{\mathsf{n}^{\text{int}}(G_A), \mathsf{n}^{\text{int}}(G_B)\}$, $n_{\text{LB}} := \min\{n(G_A)\ n(G_B)\} = 40$ 和 $n^* := \max\{n(G_A)\ n(G_B)\}$。将下界 ℓ_{LB}, bl_{LB}, ch_{LB}, $\mathsf{bd2}_{\text{LB}}$, $\mathsf{bd3}_{\text{LB}}$, na_{LB}, $\mathsf{na}^{\text{int}}_{\text{LB}}$, $\mathsf{ns}^{\text{int}}_{\text{LB}}$, $\mathsf{ac}^{\text{int}}_{\text{LB}}$ 和 $\mathsf{ac}^{\text{lf}}_{\text{LB}}$ 设为 0。将上界 $\mathsf{na}_{\text{UB}}(\mathtt{a}) := n^*$, $\mathsf{na} \in \{\mathtt{H}, \mathtt{C}\}$, $\mathsf{na}_{\text{UB}}(\mathtt{a}) := 5$, $\mathsf{na} \in \{\mathtt{O}, \mathtt{N}\}$, $\mathsf{na}_{\text{UB}}(\mathtt{a}) := 2$, $\mathsf{na} \in \Lambda \setminus \{\mathtt{H}, \mathtt{C}, \mathtt{O}, \mathtt{N}\}$ 设为相应值，并将其他上界 ℓ_{UB}, bl_{UB}, ch_{UB}, $\mathsf{bd2}_{\text{UB}}$, $\mathsf{bd3}_{\text{UB}}$, $\mathsf{na}^{\text{int}}_{\text{UB}}$, $\mathsf{ns}^{\text{int}}_{\text{UB}}$, $\mathsf{ac}^{\text{int}}_{\text{UB}}$ 和 $\mathsf{ac}^{\text{lf}}_{\text{UB}}$ 设为 n^*。对于每种边构型 $\gamma \in \Gamma^{\text{int}}$，令 $x^*_A(\gamma^{\text{int}})$（分别为 $x^*_B(\gamma^{\text{int}})$）表示具有 γ 且位于 G_A（分别为 G_B）中的内部边数量，$\gamma \in \Gamma^{\text{int}}$ 并设 $x^*_{\min}(\gamma) := \min\{x^*_A(\gamma)\ x^*_B(\gamma)\}$, $x^*_{\max}(\gamma) := \max\{x^*_A(\gamma)\ x^*_B(\gamma)\}$, $\mathsf{ec}^{\text{int}}_{\text{LB}}(\gamma) := \lfloor (3/4)x^*_{\min}(\gamma) + (1/4)x^*_{\max}(\gamma) \rfloor$ 和 $\mathsf{ec}^{\text{int}}_{\text{UB}}(\gamma) := \lceil (1/4)x^*_{\min}(\gamma) + (3/4)x^*_{\max}(\gamma) \rceil$。设 $\mathcal{F}_E := \mathcal{F}(v) := \mathcal{F}_A \cup \mathcal{F}_B$, $v \in V_{\text{C}}$ 和 $\mathsf{fc}_{\text{LB}}(\psi) := 0$, $\mathsf{fc}_{\text{UB}}(\psi) := 10$, $\psi \in \mathcal{F}_A \cup \mathcal{F}_B$。我们将线性约束 $\ell(a_1) \leq \ell(a_2)$ and $5 \leq \ell(a_1) + \ell(a_2) \leq 15$ 纳入侧约束。