CONFHIT：基于保形方法的生成式设计及无需预言机的保证

Siddhartha Laghuvarapu[1], **Ying Jin**[2*], **Jimeng Sun**[1*1] 伊利诺伊大学厄巴纳-香槟分校 Siebel计算与数据科学学院，美国伊利诺伊州[2] 宾夕法尼亚大学沃顿商学院统计与数据科学系，美国宾夕法尼亚州 {sl160,jimeng}@illinois.edu, yjinstat@wharton.upenn.edu

arXiv:2603.07371v1 [cs.LG] 7 Mar 2026

摘要

深度生成模型在科学发现中的成功，不仅需要具备生成新颖候选项的能力，还需要可靠地保证这些候选项确实满足预期属性。近期的保形预测方法为实现此类保证提供了途径，但将其应用于药物发现中的生成建模，受到预算限制、缺乏预言机访问以及分布偏移的制约。为应对这些挑战，我们提出 CONFHIT，这是一个在上述条件下提供有效性保证的模型无关框架。CONFHIT 将两个核心问题形式化：（i）认证：能否保证一个生成批次中至少包含一个命中样本，并达到用户指定的置信水平；（ii）设计：能否在不削弱这一保证的前提下，将生成结果精炼为一个紧凑的候选集。CONFHIT 利用历史样本与生成样本之间的加权可交换性，免去对实验预言机的需求；构造多样本密度比加权保形 p 值，以量化命中样本的统计置信度；并提出嵌套检验程序，在维持统计保证的同时，对由多个生成样本构成的候选集进行认证和精炼。在具有代表性的生成式分子设计任务和广泛的方法中，CONFHIT 始终能在多个置信水平下提供有效的覆盖保证，同时保持认证集紧凑，从而为生成建模确立了一种有原则且可靠的框架。

1 引言

深度生成建模展现出探索高维空间的卓越能力，推动了多个领域的进步，例如文本生成（Radford 等人，2018年）、图像合成（Ho 等人，2022）、蛋白质工程（Madani 等人，2020）和分子发现（Gómez-Bombarelli 等人，2018年）。然而，在药物发现等关键领域，成功所需的不仅仅是生成：尽管已有强大的生成模型问世，旨在加速早期发现（Madani 等人，2021年；Hoogeboom 等人，2022；Yim 等人，2023），其实际效用取决于生成的候选项是否确实满足关键的生化性质。由于这些性质只能通过成本高昂的湿实验或体内实验加以验证，因此至关重要的是预先评估并确保生成样本（命中样本）的可行性，从而引出一个核心问题：

> 给定一个生成模型，如何构建生成样本的批次，使其能够以较高的统计置信度保证至少包含一个有效命中？

保形预测为黑盒预测模型提供了一种模型无关框架，用于建立统计保证（Vovk 等人，2005），近期还被扩展为可校准任意生成模型，使其生成的样本集合以高概率包含至少一个高质量实例（Quach 等人，2023；Ulmer 等人，2024年；Shahrokhi 等人，2025年）。尽管此类保证可能极具价值，但在药物发现等资源受限问题中直接应用现有方法，仍受到若干挑战的限制。（i）认证：在生成预算有限且不对生成模型作任何假设的情况下，并非总能生成有效命中。因此，必须明确说明何时能够提供保证、何时不能。从技术层面看，（ii）缺少预言机访问：现有方法依赖预言机对已有输入对应的新生成样本进行评估（例如，将其与金标准输出进行比较）。在药物发现中，这意味着需要合成生成样本并通过实验验证；在同样资源受限的场景下，这难以实现（Kladny 等人，2024年）。（iii）分布偏移：生成的候选项可能服从不同于校准数据的分布，从而违反可交换性假设。

*这些作者共同指导了这项工作。

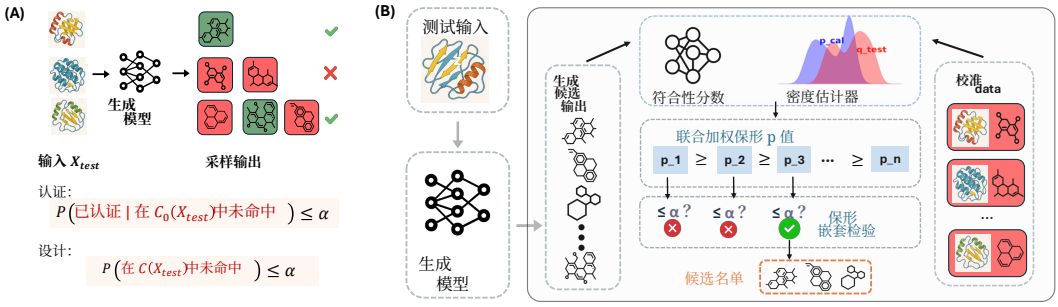


图 1：（a）问题设置：给定一个输入，以至少 $1-\alpha$ 的概率认证并生成一个候选集，使其中至少包含一个“命中”（绿色）。（b）CONFHIT 工作流程。给定一系列嵌套的候选批次，我们估计带标签数据与生成样本之间的密度比，计算每个批次的保形 p 值，以量化该批次包含命中的置信度，并返回 p 值低于 α 的最小批次。

我们提出 CONFHIT，这是一种模型无关框架，可为资源受限场景中的生成建模提供可靠性保证。为应对预算限制，我们将生成问题拓展为两个相互关联的基本问题：

1. 认证。给定一个输入 X_{test}、一批 $\mathcal{C}_0(X_{\text{test}})$ 生成样本以及一个未知的属性指示函数 $A(\cdot) \in \{0,1\}$，我们能否在给定置信水平 $1-\alpha \in (0,1)$ 下保证该批次 $\mathcal{C}(X_{\text{test}})$ 至少包含一个命中？2. 设计。在置信度较高的情况下，我们能否设计一个紧凑的候选集 $\mathcal{C}(X_{\text{test}})$，同时保证其以至少 $1-\alpha$ 的概率包含一个有效命中？

这里 $A(\cdot) \in \{0,1\}$ 是一个预言机，仅当样本满足所需属性时才返回 1。在分子设计中，它可以指示某个分子是否 x 优于种子分子 X_{test}，具体体现在活性方面；这些问题可确保对 $\mathcal{C}(X_{\text{test}})$ 进行实验验证1a）不太可能是徒劳之举。我们指出，在资源充足的情况下，可以结合认证与细化：不断增加生成预算，直至实现认证，随后将集合精简为紧凑的认证子集，从而通过现有方法获得相同的有效性保证。

为消除预言机访问并应对分布偏移，CONFHIT 利用历史（带标签）样本与生成样本之间的可交换性结构——而非早期方法中针对新旧模型输入所采用的生成样本间可交换性——并估计它们的密度比，从而导出加权可交换性（Tibshirani 等人，2019年；Jin 与 Candès，2023a）。针对认证问题，对于任意可行性评分函数，我们构造一个加权置换p值，用以量化反对全局零假设的证据；该零假设认为生成批次中的样本均不可行。针对设计问题，CONFHIT 随后考察候选批次嵌套序列的认证置信度，并在给定置信水平 $\alpha \in (0,1)$ 下返回能够通过认证的最小批次。我们证明，无论评分函数和生成模型如何，该程序返回不含任何可行样本的批次的概率均低于 $\alpha \in (0,1)$。

我们在两项代表性生成式设计任务上展示了 CONFHIT 的稳健保证：**(i)约束分子优化**，旨在寻找满足目标性质、同时与给定分子骨架保持相似的新分子；**(ii) 基于结构的药物发现**，旨在生成对给定蛋白质具有活性的分子。我们在下文总结我们的贡献：

- 我们形式化了在资源受限场景下具有共形有效性保证的生成建模任务：给定输入上下文（如先导分子或蛋白质口袋），认证并生成一组候选项，使其在预先指定的置信水平 $1-\alpha$ 下至少包含一个命中。
- 我们针对认证问题引入了一类密度比加权的多样本检验保形 p 值，并证明其在分布偏移下，对给定批次生成样本中至少存在一个命中进行认证的有效性。
- 我们提出了一个通用的嵌套检验框架，以实现设计问题的有效性保证。具体而言，我们证明：给定我们的 p 值序列，一旦 p 值降至 α 以下即停止，即可实现有限样本误差控制。
- 我们针对该方法的两个关键要素——评分建模和密度比估计——提出了实用策略，并展示了这些策略在两项标准分子设计任务（约束分子优化和基于结构的药物发现）中的稳健表现。

相关工作。保形预测（CP）（Vovk 等人，2005）通过利用可交换校准数据构建预测集，为任意黑盒模型提供无分布假设的不确定性量化（Papadopoulos 等人，2002）。近期进展将保形预测扩展到了生成任务（Angelopoulos 等人，2021年；Quach 等人，2023；Shahrokhi 等人，2025年；Kladny 等人，2024年）。然而，这些方法需要预言机访问，对于药物发现这类资源受限场景而言并不实用。

在科学研究和药物发现任务中，保形预测迄今主要用于性质预测中回归和分类问题的预测推断（Sun et al., 2017；Svensson et al., 2018；Cortés-Ciriano & Bender, 2018；2019；Zhang et al., 2021），以及通过重要性权重估计将其扩展到协变量偏移情形（Fannjiang et al., 2022；Laghuvarapu 等人, 2023；Prinster et al., 2023；Fannjiang & Park, 2025）。除构造预测集外，保形选择还利用 p 值对化合物进行优先级排序，并控制选择性误差（Bai et al., 2024；Jin & Candès, 2023b；Bai & Jin, 2024）。由于我们研究的是一个不同的生成式设计问题，因此我们的形式化方法、推理方案和方法论与以往工作截然不同。

与此同时，关于科学设计任务中生成式建模的文献日益增多，包括目标导向的小分子优化（Gómez-Bombarelli 等人，2018年；Jin 等人，2020）、基于结构的配体设计（Corso 等人，2022年；Guan 等人，2023）、大规模蛋白质生成器（Madani 等人，2020；Ingraham 等人，2019年；Strokach& Kim，2022年）以及材料发现（Xie 等人，2022年；Zeni 等人，2023）。我们的贡献与该文献正交，因为 CONFHIT 是模型无关的，并为从这些生成模型构建满足性质要求的样本提供保证。

综上所述，与现有工作相比，CONFHIT 是首个能够（i）消除对预言机访问的需求、（ii）校正历史样本与生成样本之间的协变量偏移，以及（iii）在生成式设计中同时为认证和紧凑设计提供有限样本保证的框架。

2 问题表述

为便于说明，我们对第 1 节中的符号略作调整。

数据与分布。我们假设可以获取一组来自未知分布 P 的独立同分布标注（校准）样本 $\mathcal{D}_{\text{calib}} = \{(X_i, Y_i)\}_{i=1}^{n}$，例如来自既往筛选活动且已知性质的分子。这里，$X_i \in \mathcal{X}$ 是特征，例如某个分子的化学结构；$Y_i \in \{0,1\}$ 是二元标签，例如由预言机判定的 $Y_i = A(X_i)$。生成模型 $\mathcal{M}$ 从另一个分布 Q 生成独立同分布样本 $\{X_{n+j}\}_{j\geq 1}$，这些样本的标签未知 $\{Y_{n+j}\}_{j\geq 1}$（若由预言机判定，其性质为 $Y_{n+j} = A(X_{n+j})$）。尽管原则上可以计算 Q 的密度，但这样做通常代价高昂。我们改为假设这两个分布之间存在协变量偏移：对于某个函数 $w\colon \mathcal{X} \to \mathbb{R}^+$，$\mathrm{d}Q/\mathrm{d}P(x,y) = w(x)$。这意味着 X 包含了预测 Y 所需的全部信息；在药物发现领域，这一假设尤为合理且常见，因为生物学性质由结构决定（尽管其决定机制非常复杂）。

保证。现在，我们正式阐述以下两项目标的保证：(i) 认证一组生成候选样本中存在命中；(ii) 构造一组至少包含一个命中的样本。对于目标 (i)，给定预先指定的置信水平 $\alpha \in (0,1)$ 和一组预算为 $N \in \mathbb{N}^+$ 的生成样本 $\mathcal{C}_0^{\text{new}} := \{X_{n+j}\}_{j=1}^{N}$，我们旨在提出一个检验 $\psi(\mathcal{D}_{\text{calib}}, \mathcal{C}_0^{\text{new}}) \in \{0,1\}$，使得

$$\text{(Certification)} \qquad \mathbb{P}\big(\psi(\mathcal{D}_{\text{calib}}, \mathcal{C}_0^{\text{new}}) = 1 \,\big|\, A(X_{n+j}) = 0,\ \forall 1 \leq j \leq N\big) \leq \alpha, \tag{1}$$

也就是说，错误地认证低质量候选集的概率上界为 α。对于目标 (ii)，我们希望找到一个（随机）停止点 $\hat{N} \leq N$，使得

$$\text{(Design)} \qquad \mathbb{P}\big(A(X_{n+j}) = 0,\ \forall 1 \leq j \leq \hat{N}\big) \leq \alpha, \tag{2}$$

输出样本列表 $\hat{\mathcal{C}}^{\text{new}} := (X_{n+1}, \ldots, X_{n+\hat{N}})$，其中包含一个命中，同时控制空声明的误差。我们将 $\hat{N} = 0$ 定义为：当我们对在生成预算内声明任何正样本的置信度不足时（认证失败），可适用此定义。[1]

3 方法

我们提出共形嵌套检验框架，以解决认证问题（第 3.1 节）和设计问题（第 3.2 节），并证明该框架具有模型无关的有限样本保证。

[1] 这与生成模型不确定性量化领域的现有研究一致（Quach 等人，2023），这些研究要求提供“零假设”选项（即该方法无法确信能够生成优质样本）。

3.1 认证：联合加权保形 p 值

我们首先解决认证问题：给定一组生成样本，如何量化其中包含命中的置信度？我们的关键策略是借鉴保形推断的思想构造一个 p 值，以检验 $\{X_{n+j}\}_{j=1}^{N}$ 中是否存在任何命中。请回顾，带标签校准数据与新的测试样本之间存在协变量偏移$w(x) = \mathrm{d}Q/\,\mathrm{d}P(x,y)$。为了明确起见，我们暂且假设密度比 $w(\cdot)$已知，并将在注释 3.3 中讨论估计密度比的情形。

我们的 p 值利用非活跃校准数据 $\{X_i:\ i \in \mathcal{I}_0\}$ 其中 $\mathcal{I}_0 := \{i \in [n]:\ Y_i = 0\}$。我们用 $\boldsymbol{X} = (X_1, \ldots, X_{n_0}, X_{n+1}, \ldots, X_{n+N})$ 表示（非活跃）校准协变量和测试协变量构成的向量。按照拆分保形预测（Vovk 等人，2005），令 $V:\ \mathcal{X}^{n_0+N} \to \mathbb{R}$ 为任意函数，其训练过程独立于 $\{(X_i, Y_i)\}_{i\geq 1}$，后者随后被视为固定值。我们将 V称为一致性评分函数，其选择将在本小节末尾附近讨论。一般性不失，假设符合性分数越大，表示对 $\{Y_{n+j}\}_{j=1}^{N}$ 中存在正值的置信度越高。

为了处理多个测试点，同时避免引入过多计算开销，我们采用随机化方法（非随机化版本见附录 A.1）。令 Π_N 为 $\{1, \ldots, n_0, n+1, \ldots, n+N\}$ 的所有置换的集合，并将恒等映射记为 $\pi^{(0)}$，即对于任意 i，$\pi^{(0)}(i) = i$。形式上，固定任意 $B \in \mathbb{N}^+$，我们抽取 $\pi^{(1)}, \ldots, \pi^{(B)}$ 独立同分布$\sim \mathrm{Unif}(\Pi_N)$，即置换空间上的均匀分布。定义

$$p_N^{\mathrm{rand}} = \frac{\sum_{b=0}^{B} \bar{w}(\pi^{(b)}; \boldsymbol{X})\, \mathbb{1}\{V(\pi_0; \boldsymbol{X}) \leq V(\pi^{(b)}; \boldsymbol{X})\}}{\sum_{b=0}^{B} \bar{w}(\pi^{(b)}; \boldsymbol{X})}. \tag{3}$$

这里，对于任意置换 $\pi \in \Pi_N$，我们将置换数据 $V(\pi;\ \boldsymbol{X}) = V(X_{\pi(1)}, \ldots, X_{\pi(n_0)}\ X_{\pi(n+1)}, \ldots, X_{\pi(n+k)})$, 的符合性分数定义为，并且 $\bar{w}(\pi;\ \boldsymbol{X}) = \prod_{j=1}^{k} w(X_{\pi(n+j)})$,是置换 π 后的联合似然比，其中 $w(x) = \mathrm{d}Q/\,\mathrm{d}P(x,y)$ 是密度比。

我们构造的 p 值依赖于对分布偏移下非活跃校准数据 $\{X_i$与 $i \in \mathcal{I}_0\}$ 测试样本 $\{X_{n+j}\}_{j=1}^{N}$ 之间可交换性结构的分析，受篇幅所限，详见附录 A.2 。该方法通过扩展的加权可交换性结构，推广了文献中的保形 p 值（Tibshirani 等人，2019年），适用于多个测试样本（Hu 和 Lei，2024年；Jin 与 Candès，2023a）；这些联系将在附录 A.3中讨论。

下面的定理确立了随机化 p 值（3）在有限样本下的有效性。值得注意的是，无论抽取的置换数量是多少，该结论均成立。证明见附录 B.4。

定理 3.1。 在协变量偏移假设下，对于任意固定的 $t \in [0,1]$，都有

$$\mathbb{P}\big(\, p_N^{\mathrm{rand}} \leq t \,\big|\, \max_{1\leq j\leq N} Y_{n+j} = 0\big) \leq t.$$

因此，认证检验函数 $\psi(\mathcal{D}_{calib},\ \{X_{n+j}\}_{j=1}^{N}) = \mathbb{1}\{p_N^{\mathrm{rand}} \leq \alpha\}$ 达到 (1).

继承了保形推断的无模型特性后，只要评分函数 V 的训练过程独立于校准数据和测试数据，p_k 和 p_k^{rand} 的有效性就不受评分函数的影响。

备注 3.2 (异常值检测)。 我们的 p 值也可以视为对以下情况进行检验：是否存在至少 一个异常值位于 $\{X_{n+j}\}_{j=1}^{m}$，其中 $\{X_i\}_{i\in\mathcal{I}_0}$ 被视为校准内点。关于协变量偏移下异常值检测的相关讨论，请参阅附录*Bates 等人，2023；Jin 与 Candès，2023a)* 中关于两者联系的讨论。*A.4*。

备注 3.3(估计密度比)。 在许多实际场景中，密度比 $w(\cdot)$ 未知或难以计算。在这种情况下，我们通过 $\hat{w}(\cdot)$ 估计密度比，并在构造 p 值时使用 $\hat{w}(\cdot)$ 代替 $w(\cdot)$。为简洁起见，我们不写出此类插件式 p 值的精确公式。例如，我们可以使用全部带标签数据中的一个随机子集以及独立生成的测试样本来训练密度比。鉴于ConfHit 的成功取决于密度比估计的准确性，我们在第 3.3 节中详细讨论了使用估计权重时 *ConfHit* 的鲁棒性，并介绍了若干实用的诊断方法 *3.3*。

符合性分数函数。 由于当 p_N 较小时我们会认证候选集，因此当 $X_{n+1:n+N}$ 包含正值时，V必须增大，以表示较高的置信度。为此，我们假设可以使用一个预训练模型 $\hat{\mu}$: $\mathcal{X} \to \mathbb{R}$，该模型根据 X_{n+j}预测未知属性 Y_{n+j}，并将用于得分 V。我们提出几种自然的选择：(i) 最大池化： $V(x_{1:n_0},\ x_{n+1:n+N}) =$

$\max_{1\le j\le N}\hat{}\mu(x_{n+j})$，(ii) 预测值之和：$V(x_{1:n_0}, x_{n+1:n+N}) = \sum_{j=1}^{N}\hat{}\mu(x_{n+j})$，(iii) 排名和：$V(x_{1:n_0}, x_{n+1:n+N}) = \sum_{j=1}^{N} R_{n+j}$，其中 R_{n+j} 是所有得分中的排名，(iv) 似然比：$V(x_{1:n_0}, x_{n+1:n+N}) = \sum_{j=1}^{N}\log(\hat{\mu}(x_{n+j})\,1-\hat{\mu}(x_{n+j}))$。我们在附录 A.5 中解释这些得分。我们注意到，CONFHIT 与模型无关：实践者可以选择任何合适的模型来构建得分V。得分/模型的质量会影响检验效能，但不会影响 CONFHIT 的错误控制。最后，当带标签数据有限时，实践者可能希望避免样本拆分；我们在附录 A.6 中提供了这样一种变体，其中带标签数据同时用于训练 $\hat{}\mu$ 和构造 p 值。

3.2 设计：保形嵌套检验

在获得给定候选集的有效 p 值和认证检验后，我们着手解决设计问题：如何提出一个紧凑的生成样本集，使其以高概率至少包含一个命中？为此，我们将（2）与构造较小候选子集的问题联系起来，使这些子集能够通过我们的方法进行认证，并确立其统计保证。

具体而言，对于每个 $1 \le k \le N$，我们考虑原假设

$$H_k\colon Y_{n+j} = 0,\ \forall\, 1 \le j \le k. \tag{4}$$

也就是说，H_k 认为前 k 个生成实例均不满足所需属性。从假设检验的角度看，拒绝 H_k 因此表明，我们有足够的置信度，可以判定$\{X_{n+j}\}_{j=1}^{k}$中存在命中。第3.1 节中的认证策略可直接用于获得 p 值 p_k，以认证由 $\{X_{n+j}\}_{j=1}^{k}$ 索引且满足 k 的每个子集 $\mathbb{P}$（$p_k \le t \,|\, H_k$ 为真） $\le t$，其中 $t \in [0,1]$。

$\hat{}$ 我们对设计问题的解决方案看起来很简单：确定一个索引 N，使得 $H_{\hat{N}}$ 可以被拒绝（即可以有把握地断定 $\{X_{n+j}\}_{j=1}^{k}$ 包含一次命中）。具体而言，假设我们可以为每个固定的 H_k构造一个递减的 p 值序列 $p_k \in [0,1]$。那么，我们设定

$$\hat{N} := \inf\left\{k\colon p_k \le \alpha\right\}. \tag{5}$$

也就是说，我们取第一个通过显著性水平 $\alpha \in (0,1)$的 p 值。定理 3.4 证实了这种嵌套检验策略的有效性，其证明见附录 B.1。

定理 3.4。假设 *p* 值 $p_k \in [0,1]$ 满足：*(i)* 单调性：$p_1 \ge p_2 \ge \cdots p_N$。*(ii)* 有效性：对于任意固定的 $k \ge 1$ 和 $t \in [0,1]$，若 $p_k \le t \,|\, H_k$ 为真，则有 $\mathbb{P}(p_k \le t \,|\, H_k$ 为真 $) \le t$。那么，我们有

$$\mathbb{P}(\max_{1\le j\le \hat{N}} Y_{n+j} = 0) \le \alpha$$

$\hat{}$对应的生成阈值 N 按 (5) 中的方式计算。这里 H_k 定义见 (4)。

我们注意到，任何满足定理 $\{\tilde{p}_k\}$ 3.4 中条件 (ii) 的 p 值，例如在认证问题中构造的 p 值，都可以转化为 $p_k = \max_{k\le j\le N}\tilde{p}_j$，使其满足定理 3.4中的两个条件。或许令人意外的是，尽管我们同时考察多个候选批次，但事实证明，使用一列单调的 各自有效的 p 值就足以实现我们的目标，无需对多重性进行调整。我们指出，我们的理论保证的关键在于假设具有嵌套性质：如果 H_k 为真，那么所有“更早的”假设$\{H_\ell\}_{\ell\le k}$ 也必定为真。因此，我们将该方法称为“保形嵌套检验”。

ConfHit：汇总全流程。至此，我们已完成 CONFHIT 的所有组成部分。我们在算法 1中总结了整个流程，其中包括认证（针对每个嵌套子集）和设计。请注意，输入 $\hat{}w(\cdot)$ 表示一个估计的密度比函数；当密度比已知时，该函数与 $w(\cdot)$ 相同。如果连最大的批次都无法认证，CONFHIT 就会标记“置信度不足”，以明确传达生成任务的难度。

3.3 密度比估计的鲁棒性与诊断

密度比在实践中通常需要估计，其质量自然会影响ConfHit的性能。历史测定数据与新生成的分子之间的差异是药物发现中的一个持续问题，几乎需要在该领域的所有保形预测方法中加以解决（Krstajic，2021年；Laghuvarapu 等人，2023；Jin 与 Candès，2023a；Fannjiang& Park，2025年）。估计密度比是统计学中的一个标准问题，被用于机器学习、统计学和因果推断等许多领域；除我们实验中的核密度估计之外，已有丰富的工具箱，例如核方法或基于分类的方法（Horvitz &

算法 1 CONFHIT

输入：校准数据 $\{X_i\}_{i=1}^{n_0}$，其中 $Y_i = 0$，测试数据 $\{X_{n+j}\}_{j=1}^{N}$，符合性分数 V： $\mathcal{X}^{n_0+k} \to \mathbb{R}$，（估计的）权重函数$\hat{w}$： $\mathcal{X} \to \mathbb{R}^+$，置信水平 $\alpha \in (0, 1)$，预算 $N \in \mathbb{N}^+$，蒙特卡洛样本量$B \in \mathbb{N}^+$1：对 $k = 1$，.....， N执行 2：设 $\Pi = \{$为 $\{1, \ldots, n_0, n+1, \ldots, n+k\}\}$的所有排列。// 序贯共形p值 3：抽取 $\pi^{(b)}$ 独立同分布$\sim$ Unif(Π)，针对 $b = 1$，.....， B。4：按（3）计算p 值 p_k，其中 $w(x)$：$= \hat{w}(x)$。5：结束循环 6：设 $p_k = \max_{k' \geq k} p_{k'}$，对 $k = N - 1$，.....，1迭代进行。// 单调化p值 7：计算$\hat{N} = \text{argmax}\{k \in [N]$： $p_k \leq \alpha\}$，若 $p_N > \alpha$则使用$\hat{N} = 0$。输出：置信

shortli st $\mathcal{C} := \{X_{n+j}\}_{j=1}^{\hat{N}}$ 或“信心不足 ugh” w$_{\text{hen } \hat{N}} = 0$.

Thompson，1952；Shimodaira，2000；Sugiyama 等，2007；Nguyen 等，2010；Sugiyama 等，2012；Shapiro，2003）。下文将详细阐述（i）对 CONFHIT 估计误差鲁棒性的理论分析，以及（ii）三种用于诊断估计质量的方法。

定理 3.5 表明，使用估计密度比时，误差膨胀取决于激进(分别是 保守)的权重对较小（分别是较大）p 值的影响；证明见附录 B.5。

定理 3.5（对估计误差的鲁棒性）。设 $\hat{p}_N$ 为使用任意估计权重 $\hat{w}(\cdot)$ 得到的 p 值 *(3)*。定义估计误差的正部/负部 $\hat{\delta}^{\pm}(\pi; \boldsymbol{X}) = [\hat{\bar{w}}(\pi; \boldsymbol{X}) - \bar{w}(\pi; \boldsymbol{X})]_{\pm}$，其中 $\hat{\bar{w}}(\pi; \boldsymbol{X}) = \prod_{j=1}^{N} \hat{w}(X_{\pi(n+j)})$。那么，对于任意固定的 $t \in (0, 1)$，有

$$\mathbb{P}\left(\hat{p}_N^{rand} \leq t \,\middle|\, \max_{1 \leq j \leq N} Y_{n+j} = 0\right) \leq t + \mathbb{E}\left[\frac{\hat{t} \cdot \hat{\Delta}^+ + (1-\hat{t}) \cdot \hat{\Delta}^-}{\sum_{b=0}^{B} \bar{w}(\pi^{(b)}; \boldsymbol{X})} \,\middle|\, \max_{1 \leq j \leq N} Y_{n+j} = 0\right],$$

其中，令 $\hat{p}_N(v) := \frac{\sum_{b=0}^{B} \hat{\bar{w}}(\pi^{(b)}; \boldsymbol{X}) \mathbb{1}\{v \leq V(\pi^{(b)}; \boldsymbol{X})\}}{\sum_{b=0}^{B} \bar{w}(\pi^{(b)}; \boldsymbol{X})}$，我们定义 $\hat{v} = \textit{argmax}\{V(\pi^{(b)}; \boldsymbol{X})$： $\pi \in \Pi$， $\hat{p}_N(V(\pi^{(b)}; \boldsymbol{X})) \leq t\}$ 为得分截断值，相应的 p 值为 $\hat{t} := \hat{p}_N(\hat{v})$，并且 $\hat{\Delta}^+ = \sum_{b=0}^{B} \hat{\delta}^+(\pi^{(b)}; \boldsymbol{X}) \mathbb{1}\{V(\pi^{(b)}; \boldsymbol{X}) < \hat{v}\}$， $\hat{\Delta}^- = \sum_{b=0}^{B} \hat{\delta}^-(\pi^{(b)}; \boldsymbol{X}) \mathbb{1}\{V(\pi^{(b)}; \boldsymbol{X}) \geq \hat{v}\}$

虽然定理 3.5揭示了保证如何随（未知的）估计误差而退化，但在实践中进行稳健性诊断往往很有用。下述三种方法将在“稳健性检查”第 4.5 节的实验中展示。

(1) 平衡性检查。根据因果推断中的平衡思想（Ben-Michael 等，2021年），通过正确的密度比进行重新加权，可以使校准数据中任意协变量函数的经验均值接近测试数据中的对应均值（在零假设成立且没有命中样本的情况下）。一种自然的诊断方法是检查不平衡程度 $|\frac{1}{n_0}\sum_{i=1}^{n_0} \hat{w}(X_i) f(X_i) - \frac{1}{k}\sum_{j=1}^{k} f(X_{n+j})|$ 对于某些 f： $\mathcal{X} \to \mathbb{R}$。

(2) 验证分布偏移。第二种方法是通过合成分布偏移来考察估计的可靠性。可以在带标签数据中人为制造分布偏移，例如进行骨架划分；将其中一折作为校准数据，另一折作为测试数据后，就可以利用已知标签来考察 p 值的均匀性以及密度比估计算法的质量。

(3) 敏感性分析。最后，可以进行敏感性分析，以探查 CONFHIT 对估计权重扰动的鲁棒性。与因果推断类似（Rosenbaum，2005），我们可以假设正确权重与估计权重之间的距离在一定范围内，并在假定范围内改变权重，同时考察最坏情况下的 p 值和检验结果。

尽管生成模型本身并不天然具备内置的可靠性概念，CONFHIT 为其增添了适用于实际部署的新能力；它与模型无关，只依赖一个标量量值——密度比。密度比可以得到验证和压力测试，从而使不确定性可测，并且比单纯信任原始模型更加可靠（第 4.5节“其他基线”）。

4 实验

4.1 任务、模型和数据集

我们在两个具有代表性的条件分子生成任务上评估了我们的框架，并通过采用涵盖变分自编码器（VAE）的五种生成模型，突出展示了我们方法的模型无关特性、

自回归变换器、扩散模型和贝叶斯流网络。为了评估，我们使用计算预言机来判断生成样本的属性，因为湿实验评估不可行（Jin 等人，2020）；运行 CONFHIT 时不使用该预言机。

附录 1：约束分子优化（CMO）。给定一个种子分子 X，它未能满足目标属性，目标是生成一个分子 X，使其与 X 的 Tanimoto 距离较小，同时满足该属性。此任务反映了先导化合物优化流程，其中实验测定提供真实标注。按照使用计算预言机的标准做法，我们在实验中使用属性*DRD2* 受体结合（*DRD2*），若成功（$Y = 1$），则$\mathrm{DRD2}(x) > 0.5$。附录中报告了药物相似性定量估计（QED）提升的更多全面结果E.2。

为展示 CONFHIT 与通用生成模型的兼容性，我们采用了两种最先进的 CMO 模型：1.*H*GRAPH2GRAPH，一种分层变分自编码器，可编辑骨架级基序，同时保持有效性（Jin 等人，2020）；2.*SELF*-EDIT，一种自回归 Transformer，可生成 SELFIES 字符串以优化属性，同时保持与种子分子的相似性（Jiao 等人，2023）。

我们从 ChEMBL 数据（Yang 等人，2019年）中 Chemprop 训练划分的一个子集（Gaulton 等人，2012）用作带标签校准数据 $\{X_i, X_i, Y_i\}_{i=1}^n$（种子输入、从既往筛选活动中抽取的样本以及预言机属性），遵循Jin 等人（2020）的方法。测试划分中的非活性分子用作测试输入 $\{\bar{X}^{(\ell)}\}_{\ell=1}^L$。对于每个 $\bar{X}^{(\ell)}$，模型会生成 $\{X_{n+j}^{(\ell)}\}_{j=1}^m$。

附录 2：基于结构的药物发现（SBDD）。此处的输入 X 是一个三维蛋白质结合口袋，任务是生成能够与其结合的配体 X，因此标签 Y 取决于$(X,\ X)$。按照标准做法，仅为评估之用，我们使用 AutoDock Vina（Trott 和 Olson，2010）作为计算预言机，并将得分低于 -7.5 千卡/摩尔$^{-1}$的配体标记为命中；该阈值涵盖了 CrossDock（Francoeur 等人，2020a）中约75% 的已知活性样本。

在本任务中，我们将 CONFHIT 应用于三个最先进的 SBDD 模型：*TargetDiff*(Guan 等人，2023），一种以口袋网格为条件的 SE(3) 等变扩散模型；*DecompDiff*(Guan 等人，2024年），使用分解先验分离结构成分；*MolCRAFT* (Qu 等人，2024年），一种完全在连续参数空间中运行的贝叶斯场网络。

校准数据为来自 CrossDocked（$\{(X_i,\ X_i)\}_{i=1}^n$ 的蛋白质-配体对（Francoeur 等人，2020b）。遵循 Guan 等人（2023）中的相同划分，我们将测试划分中的蛋白质作为输入$\{\bar{X}^{(\ell)}\}_{\ell=1}^L$，模型针对每个输入生成候选配体 $\{X_{n+j}\}_{j=1}^N$ 给定每个输入 $\bar{X}^{(\ell)}$。

4.2 实现细节

一致性评分函数。我们使用性质预测模型 $\hat{\mu}(\cdot)$定义 V：

- **CMO**。对于这两种属性（QED 和 DRD2），我们使用 Chemprop 库（Yang 等人，2019年）和 Jin 等人(2020年$\mu\cdot$，以分子图为输入训练一个二元分类器。
- **SBDD**。我们在 PDBBind 晶体结构数据（Wang 等人，2005）上训练等变图神经网络$\mu\cdot$(Satorras 等人，2021年），以预测结合亲和力，使用的是 TARGETDIFF (Guan 等人，2023）代码库。

对于这两项任务，μ 均使用独立于校准集和测试集的数据进行训练。附录 C中给出了μ 和生成模型的训练细节、超参数及计算资源。为简洁起见，我们主要展示采用最大池化评分 V $x_{1:n_0},\ \ x_{n+1:n+k} = \max^1\!\le j\le k\,\mu\ (x_{n+j})$ 的结果。与其他评分函数的比较见图4。

密度比估计。我们利用核密度估计（KDE）来构造 p 值中的权重。根据附录 A.2 中引理 A.2 的讨论，所需权重 $w(\cdot)$ 等于生成样本的边缘密度比，与命中状态无关。此外，由于归一化，代入输入/生成对联合分布的密度（记为 $\tilde{q}(\bar{x}, x)$ 和 $\tilde{p}(\bar{x}, x)$）也会得到相同的 p 值。因此，我们直接估计 $\tilde{p}(\cdot)$ 和 $\tilde{q}(\cdot)$，以确保有足够的数据用于估计。遵循 CODRUG（Laghuvarapu 等人，2023）的方法，我们从属性模型的倒数第二层提取潜在特征 $g(\bar{x}, x)$，并在校准样本和测试样本上拟合高斯核密度估计 $\hat{p}(\bar{x}, x)$ 和 $\hat{q}(g(\bar{x}, x))$，形成权重 $\tilde{w}(\bar{x}, x) = \hat{q}(g(\bar{x}, x))/\hat{p}(g(\bar{x}, x))$。

4.3 结果：认证

本节展示认证实验的结果，实验目标（1）是在固定大小 N 的批次中，以 α的显著性水平认证是否存在命中样本。我们的评估重点包括：（i）错误控制；（ii）检验效能，即成功认证命中的频率。我们展示了 CMO 和 SBDD 任务

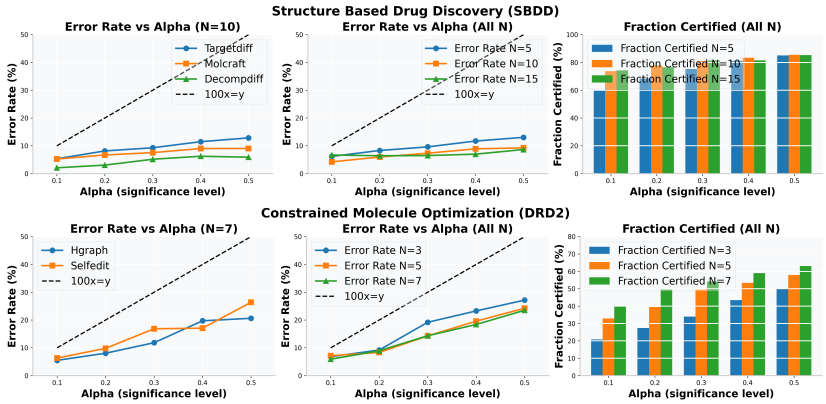


图 2：**认证结果**。左：不同模型和误差水平下，固定 N 时 α 在 SBDD（上）和 CMO（下）中的实际错误率。中：改变预算 N时的平均错误率。右：在不同误差水平 α 和预算 N 下的检验效能，即获得认证的活性分子所占比例。虚线表示理想的 $y = x$ 误差界。我们的方法在各种情境下始终实现有效覆盖率。结果为 5 次随机运行的平均值；误差条以及其他 N 取值的结果见附录 E.2

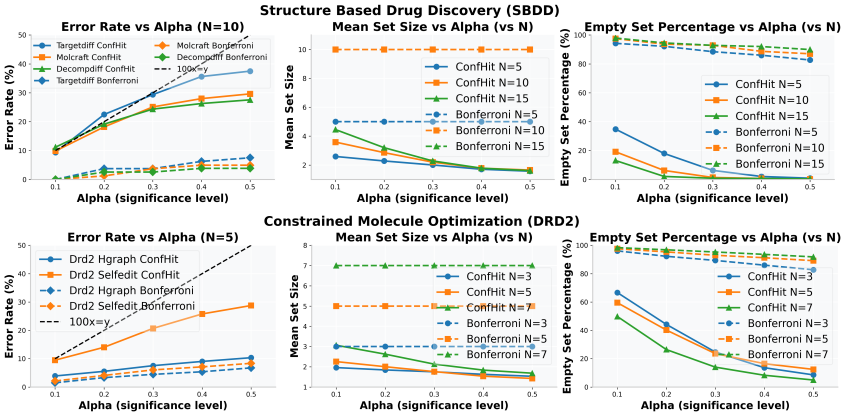


图 3：**设计结果**。不同方法在固定 N 下的错误率（左）；不同 N 取值下各方法的平均集合大小（中）；以及不同 N 取值和目标水平 α下的空集百分比（右）。上排展示 SBDD 的结果，下排展示 CMO 的结果。错误率图中的黑色虚线表示理想的 $y = x$ 界。CONFHIT 在实现严格错误控制的同时，生成的集合显著更小。结果为 5 次随机运行的平均值；更多结果见附录 E.2。

在不同生成模型上的结果，如图2所示。左图报告经验错误率，即认证不包含任何命中样本的集合的频率。在所有设置下，我们都观察到严格的错误控制。中图展示了不同预算 N 下的错误率（为便于可视化，按方法取平均），可以看到错误率始终低于目标误差水平。最后，右图绘制了在不同 α下，包含活性物质的集合被认证的比例；CONFHIT 能以令人满意的检验效能检测出真命中。正如预期，该比例会随着显著性水平和集合大小 N 的增加而上升。

4.4 结果：设计

本节将介绍公式（2）中设计问题的实验：我们对生成批次进行剪枝，以获得紧凑子集，同时保持严格的错误控制。如图 3所总结，我们的嵌套过程在保持统计保证的同时细化候选集。

由于我们在独特约束下处理一个新颖问题，因此不存在可直接比较的基线（保形生成建模中的现有方法（Quach 等人，2023；Kladny 等人，2024年）由于无法获得预言机而不具备可比性）。相反，我们考虑两个基线，以展示 CONFHIT 的优势：**（i）Bonferroni校正**。给定 N 个测试样本，我们将单测试样本p值（将每个候选样本作为（3）中的测试集）设为 α/N。该方法为任何测试样本提供更强的无错误认证保证，但可能较为保守。**（ii）仅认证**。第 4.3 节中的认证流程，该流程作为不进行剪枝的基线。

与 Bonferroni校正相比，我们的方法在 CMO 和 SBDD任务上均显著优于前者。在严格的水平下（$\alpha = 0.1$），Bonferroni校正对近 100% 的 SBDD 输入产生空集，而 ConfHit将空集频率降至 $N = 15$[2]。此外，在 Bonferroni校正能够进行发现的少数情况下，该方法通常会耗尽全部预算；相比之下，CONFHIT 始终将候选剪枝至 2–5 个分子（完整预算的 30–50%）。由于其保守性，Bonferroni校正的错误率较低或接近于零，而 CONFHIT 则实现了自适应且紧致的覆盖率。

与仅认证方法相比，其结果见图2，嵌套过程成功地将有希望的初始大集合剪枝为较小的集合。例如，在 SBDD 的 $\alpha = 0.1$ 条件下，$N = 15$嵌套过程将平均集合大小缩减至约 4 个分子，同时不损害错误控制。在 CMO 中也有类似改进：嵌套剪枝通常能将认证集合大小减半，同时使实际错误率与名义错误率更加吻合。

总而言之，CONFHIT 能够在严格控制错误的同时生成紧凑的认证集。它基本避免了空洞结果，并产生规模更小、更具可操作性的候选名单，同时提高了后续实验验证的覆盖率。这使我们的框架在科学发现工作流中特别有价值，因为实验预算要求兼顾统计置信度与实际可行性。

4.5 额外研究

检验统计量的选择。虽然我们报告的结果采用最大池化符合性分数，图 4使用五种符合性分数选项 $V(\cdot)$——最小值、均值、似然比（*LR*）、秩和以及最大值——在预算为 $N = 5$ 的 TARGETDIFF SBDD任务上评估 CONFHIT；其他任务和预算下的结果呈现相似模式，详见附录 E.2。总体而言，CONFHIT 始终使实际错误率接近目标 α，证实了其模型无关有效性。不过，分数的选择确实会影响检验效能：最大值最常获得认证（例如，在 $\alpha = 0.1$ 时，73 % 的随机种子获得认证），而秩和最为保守。

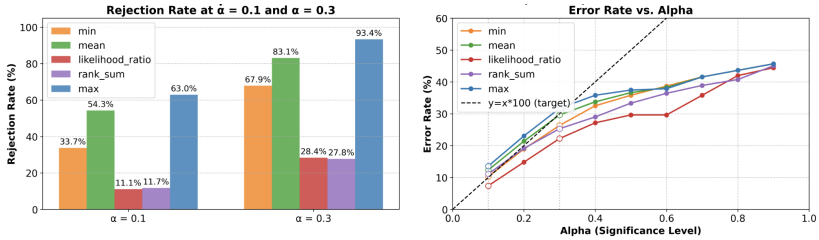


图 3：TargetDiff（SBDD，N=5）上分数统计量的比较。左：拒绝（检验效能），在 $\alpha = 0.1$，0：误差与 α。所有结果均保持有效；最大 统计量显示出最高的检验效能。

性质预测模型的选择。正如我们在第3.1节末尾所讨论的，性质预测模型的选择只影响检验效能。我们通过向预测添加非信息性扰动，对较弱的性质预测模型开展了消融研究。ConfHit 即便较弱的预测器降低了检验效能，仍能保持稳健的错误率控制；参见附录 F。

分布偏移调整。关于 ConfHit 在认证和设计两方面的保证均依赖于有效的 p 值，而这进一步取决于分布偏移调整。作为消融研究，我们将 ConfHit 与未进行分布偏移调整的版本进行比较，即取 $w(\cdot) \equiv 1$。我们观察到，运行 ConfHit 时若不进行分布偏移调整，可能会违反覆盖保证，尤其是在严格的目标错误率下，例如使用 TargetDiff 模型执行 SBDD任务，以及 SelfEdit 模型执行 CMO 任务。这些差异在图5a 中予以突出显示。此外，我们绘制了在不同数据集上用负样本计算的 p 值分布，并观察到它们近似均匀；参见附录 E.3。

稳健性检查。我们在第 3.3 节展示稳健性诊断，并在附录 H.1 中进行敏感性分析，在附录 H.2 中使用现实分子骨架划分进行合成验证，并在附录 H.3 中基于关键特征进行平衡性检查。首先，通过幂律扰动权重，我们展示了如何检验 CONFHIT 输出的稳健性，并通过额外评估说明错误控制仅轻微退化。同时，CONFHIT 在合成骨架划分下实现了严格的错误控制，且估计密度比改善了关键特征的平衡性。

[2]返回空集意味着在生成预算 N 下置信度不足；如果不对生成能力做出强假设，这种情况就不可避免。

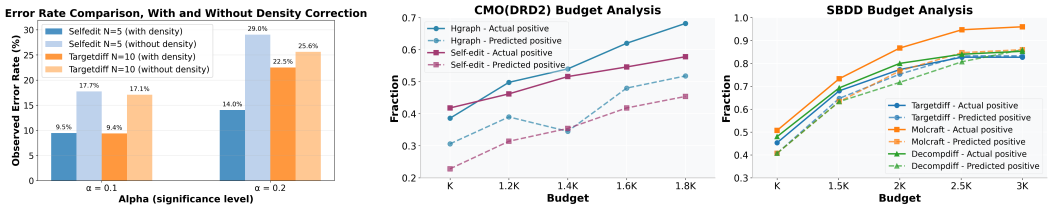


图 5：(a) **分布偏移调整**（左）。未进行密度校正时运行 CONFHIT 所产生的覆盖率违例。(b) **预算分析**（中和右）。随着生成预算增加，至少有一次命中的输入所占比例。实线：实际命中样本所占比例；虚线：预测的命中样本所占比例。

其他基线。 我们还评估了两种基线。(i) 一种 **启发式基线**，它使用估计概率来校准生成批次大小（附录 G.1）。当估计质量适度下降时，它会违反错误控制；相比之下，ConfHit 与模型无关，并能在相同的退化情况下保持有效性（附录 F），这表明有原则的不确定性量化能够提供稳健保证。(ii) 一种 **预言机基线**，它（不切实际地）为生成样本标注（Quach 等人，2023），但不进行协变量偏移调整（附录G.2）。即使能够访问预言机，忽略分布偏移也会导致误差较高且空集频率较高（可能是因为它依赖输入之间的可交换性，输入，这会降低样本量）。

4.6 总体预算下的运行

最后，我们展示 ConfHit 在多个任务之间分配固定预算时的实用价值。在药物发现项目中，科学家通常会使用多个生成输入，但只能通过实验验证有限数量的样本。在这种情境下，ConfHit 对有效命中的置信度，可作为预算分配的实用启发式依据——决定每项任务合成多少个候选项。给定 K 个输入（任务）和总预算 B，我们为每个输入设定最大规模 N；随后，ConfHit 会为每个输入选出包含命中样本的最小子集，其置信水平为 α。然后，我们逐步调整 α 的值，直至预算用尽。最后，我们通过 $(1-\alpha)$ 减去空集所占比例（即 ConfHit 在水平为 α 时无法认证命中样本），估算输出集中含有命中样本的任务比例。具体步骤见附录 D。

在图5b中，我们比较了不同预算水平（以输入数量 K 的倍数衡量）下的预测阳性与实际阳性。在CMO和SBDD中，活性分子的实际比例始终超过我们的估计，体现了CONFHIT的可靠性与实用价值。作为对认证与设计结果的补充，这一预算分配视角证实了CONFHIT在与更复杂（尽管属于启发式）的应用相结合时具有稳健的实证性能。当然，我们承认这只是一种缺乏正式理论保证的启发式方法，CONFHIT在此类任务中的应用值得未来审慎研究。

5 结论

在本工作中，我们提出了CONFHIT，这是一个为条件生成模型提供有限样本保证的模型无关框架。通过用密度比对校准数据进行重新加权，CONFHIT校正了从历史化合物到模型生成的候选项的分布偏移，从而无需外部实验预言机——而这是生成建模中所有现有保形预测方法所要求的假设。基于这些调整后的权重，我们推导出嵌套保形 p 值，用以认证采样到至少一个可行分子的概率。在受约束分子优化和基于结构的设计这两项标准基准上，以及在广泛的方法与预算范围内，CONFHIT在保持紧凑认证集的同时，始终提供有效的覆盖保证。我们的结果表明，CONFHIT是一个兼具原则性与实用性的认证与设计框架，能够在严格的资源约束下实现可靠的生成建模。

局限性覆盖保证依赖于密度比估计；当校准集较小，或特征提取器与目标域的对齐效果不佳时，估计可能存在较大噪声。我们的实验聚焦于小分子；要将这一方法扩展到具有更大、更复杂结构的生成空间的蛋白质或其他大分子，还需要进一步研究。目前的验证依赖于计算机模拟预言机；未经湿实验验证，其可迁移性仍有待检验。在更广泛的化学和实验环境中证明其鲁棒性，是未来研究的重要方向。

6 伦理声明

由于该框架与模型无关，且面向条件分子生成，我们预计它将有益于化学、生物学和材料科学领域的多种发现流程。这项工作建立在分子建模的长期实践基础之上，因此我们预计不会造成负面的社会影响。不过，我们也承认其存在潜在的双重用途，例如设计有害化合物；要确保负责任地使用，必须加强监督并实施访问控制。

大语言模型使用 本文仅使用大语言模型来润色文本并协助处理与代码相关的任务。

7 可复现性声明

所有实验环境均在第 4.1 节中进行了说明，实现细节的扩展讨论见附录 C。计算资源和运行时间的信息见第 C.4 节。本文实验所用的代码、数据集和配置文件均包含在
h
ttps://github.com/siddharthal/CONFHIT-Conformal-Generative-Design-with-Oracle-Free-Guarantees中，并在附录C.5中进一步详细说明。

参考文献

Anastasios N. Angelopoulos、Stephen Bates、Emmanuel J. Candès、Michael I. Jordan 和 Lihua Lei。先学习后检验：校准预测算法以实现风险控制。*arXiv* 预印本 *arXiv:2110.01052*，2021年。

Tian Bai 和 Ying Jin。优化共形选择：通过符合性分数优化实现强有力的选择性推断。*arXiv* 预印本 *arXiv:2411.17983*，2024年。Tianyu Bai、Puchen Tang、Yifan Xu、Vladimir Svetnik、Abbass Khalili、Xusheng Yu 和 Andrew Yang。用于药物发现中高效、准确化合物筛选的保形选择。*arXiv* 预印本 *arXiv:2402.12345*，2024年。Stephen Bates、Emmanuel Candès、Lihua Lei、Yaniv Romano 和 Matteo Sesia。使用保形 p 值检验离群值。《统计年鉴》，51(1):149–178，2023。Eli Ben-Michael、Avi Feller、David A Hirshberg 和 José R Zubizarreta。因果推断中的平衡之道。*arXiv* 预印本 *arXiv:2110.14831*，2021年。Gabriele Corso、Hannes Stärk、Bowen Jing、Regina Barzilay 和 Tommi Jaakkola。DiffDock：用于分子对接的扩散步骤、扭转与转动。*arXiv* 预印本 *arXiv:2210.01776*，2022年。Isidro Cortés-Ciriano 和 Andreas Bender。深度置信度：一种计算高效的框架，用于计算深度神经网络的可靠预测误差。《化学信息与建模杂志》，59(3):1269–1281，2018年。Isidro Cortés-Ciriano 和 Andreas Bender。保形预测在计算药物发现中的概念与应用。*arXiv* 预印本 *arXiv:1908.03569*，2019年。Clara Fannjiang 和 Ji Won Park。机器学习引导设计中的可靠算法选择。*arXiv* 预印本 *arXiv:2503.20767*，2025年。

Clara Fannjiang、Stephen Bates、Anastasios N. Angelopoulos、Jennifer Listgarten 和 Michael I. Jordan。面向生物分子设计的反馈协变量偏移下保形预测。美国国家科学院院刊，119(43):e2204569119，2022年。

Ronald Aylmer Fisher 和 Ronald A Fisher。实验设计。施普林格，1971年。

Patrick Francoeur、Ruben J. Gomez-Bombarelli、Jianfeng Pei 和 Rohit Singh。用于基于结构的药物设计的三维卷积神经网络与交叉对接数据集。《化学信息与建模杂志》，60(9):4200–4215，2020a。doi: 10.1021/acs.jcim.0c00411。

Paul G Francoeur、Tomohide Masuda、Jocelyn Sunseri、Andrew Jia、Richard B Iovanisci、Ian Snyder 和 David R Koes。用于基于结构的药物设计的三维卷积神经网络与交叉对接数据集。《化学信息与建模杂志》，60(9):4200–4215，2020b。

Anna Gaulton、Louisa J Bellis、A Patricia Bento、Jon Chambers、Mark Davies、Anne Hersey、Yvonne Light、Shaun McGlinchey、David Michalovich、Bissan Al-Lazikani 等。ChEMBL：用于药物发现的大规模生物活性数据库。《核酸研究》，40(D1):D1100–D1107，2012。

Jiaqi Guan、Wesley Wei Qian、Xingang Peng、Yufeng Su、Jian Peng 和 Jianzhu Ma。用于靶标感知分子生成和亲和力预测的三维等变扩散。载于 国际学习表征会议（*ICLR*），2023年。网址https://openreview.net/forum?id=kJqXEPXMsE0。

Jiaqi Guan、Xiangxin Zhou、Yuwei Yang、Yu Bao、Jian Peng、Jianzhu Ma、Qiang Liu、Liang Wang 和 Quanquan Gu。DECOMPDIFF：采用分解先验的扩散模型，用于基于结构的药物设计。*arXiv* 预印本 *arXiv:2403.07902*，2024年。

Rafael Gómez-Bombarelli、Jennifer N. Wei、David Duvenaud等。使用分子的数据驱动的连续表征进行自动化学设计。《*ACS*中央科学》，4(2):268–276，2018年。

Jonathan Ho、Chitwan Saharia、WilliamChan、DavidJ Fleet、MohammadNorouzi和TimSalimans。用于高保真图像生成的级联扩散模型。《机器学习研究期刊》，23(47):1–33，2022年。

Emiel Hoogeboom、Victor Garcia Satorras等。用于分子generati的等变扩散三维in。载于国际机器学习会议，2022年。

Daniel G Horvitz和Donovan J Thompson。有限总体无放回抽样的一种推广。《美国统计学会期刊》，47(260):663–685，1952年。

Xiaoyu Hu 和 Jing Lei。利用保形预测和加权秩和进行双样本条件分布检验。《美国统计学会期刊》，119(546):1136–1154，2024年。

John Ingraham、Vikas K. Garg、Regina Barzilay 和 Tommi Jaakkola。用于基于图的蛋白质设计的生成模型。载于神经信息处理系统进展，2019年。

Miao Jiao、Yiqing Wang、Xiufeng Chen 和 Tien Huynh。SELF-EDIT：使用 SELFIES 编辑变换器进行结构约束分子优化。《应用智能》，53(18):21455–21471，2023年。doi: 10.1007/s10489-023-04915-8。

Wengong Jin、Regina Barzilay 和 Tommi Jaakkola。使用结构基元分层生成分子图。载于 第*37*届国际机器学习会议论文集（*ICML*），第 4839–4848 页，2020。URL https://proceedings.mlr.press/v119/jin20a.html。

Ying Jin 和 Emmanuel J Candès。通过加权，在协变量偏移下进行无模型选择性推断保形 p 值。*arXiv* 预印本 *arXiv:2307.09291*，2023a。

Ying Jin 和 Emmanuel J Candès。利用保形 p 值进行基于预测的选择。《机器学习研究期刊》，24(244):1–41，2023b。

Klaus-Rudolf Kladny、Bernhard Schölkopf 和 Michael Muehlebach。通过序贯贪心筛选提升样本效率的保形生成建模。*arXiv* 预印本 *arXiv:2410.01660*，2024年。

Damjan Krstajic。对应用于二元分类场景的保形预测方法进行批判性评估。《化学信息与建模杂志》，61(10):4823–4826，2021年。doi: 10. 1021/acs.jcim.1c00549。URLhttps://doi.org/10.1021/acs.jcim.1c00549。PMID: 34550693。

Siddhartha Laghuvarapu、Zhen Lin 和 Jimeng Sun。CODRUG：协变量偏移下基于密度估计的保形药物性质预测。《神经信息处理系统进展》，36:37728–37747，2023。Ali Madani、Bryan McCann、Nikhil Naik、Nitish Shirish Keskar、Namrata Anand、Raphael R. Eguchi、Po-Ssu Huang 和 Richard Socher。Progen：用于蛋白质生成的语言建模。arXiv 预印本 arXiv:2004.03497，2020。Ali Madani、Bryan McCann、Nikhil Naik 等。Progen：用于蛋白质生成的语言建模。arXiv 预印本 arXiv:2004.03497，2021。XuanLong Nguyen、Martin J Wainwright 和 Michael I Jordan。通过凸风险最小化估计散度泛函和似然比。《IEEE 信息论汇刊》，56(11):5847–5861，2010。Harris Papadopoulos、Kostas Proedrou、Vassilis Vovk 和 Alex Gammerman。用于回归的归纳式置信机。载于《第13届欧洲机器学习会议（ECML）论文集》，第345–356页，2002。Drew Prinster、Suchi Saria 和 Anqi Liu。Jaws-x：解决标准协变量偏移和反馈协变量偏移下保形预测的效率瓶颈。载于《第40届国际机器学习会议（ICML）论文集》，2023。Yanru Qu、Keyue Qiu、Yuxuan Song、Jingjing Gong、Jiawei Han、Mingyue Zheng、Hao Zhou 和 Wei-Ying Ma。MOLCRAFT：连续参数空间中的基于结构的药物设计。arXiv 预印本 arXiv:2404.12141，2024。Victor Quach、Adam Fisch、Tal Schuster、Adam Yala、Jae Ho Sohn、Tommi S Jaakkola 和 Regina Barzilay。保形语言建模。arXiv 预印本 arXiv:2306.10193，2023。Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。通过生成式预训练改进语言理解。技术报告，OpenAI，2018。URL https://cdn.openai.com/research-covers/language-unsupervised/ language_ understanding_paper.pdf。Paul R Rosenbaum。观察性研究中的敏感性分析。《行为科学统计学百科全书》，4:1809–1814，2005。Vıctor Garcia Satorras、Emiel Hoogeboom 和 Max Welling。E (n) 等变图神经网络。载于《国际机器学习会议》，第9323–9332页。PMLR，2021。Hooman Shahrokhi、Devjeet Raj Roy、Yan Yan、Venera Arnaoudova 和 Janaradhan Rao Doppa。通过化简为保形回归，获得深度生成模型的保形预测集。arXiv 预印本 arXiv:2503.10512，2025。Alexander Shapiro。蒙特卡罗抽样方法。《运筹学与管理科学手册》，10:353–425，2003。Hidetoshi Shimodaira。通过对对数似然函数加权，改进协变量偏移下的预测推断。《统计规划与推断杂志》，90(2):227–244，2000。Alexey Strokach 和 Philip M. Kim。用于蛋白质设计的深度生成建模。《结构生物学当前观点》，72:226–236，2022。doi: 10.1016/j.sbi.2021.11.008。Masashi Sugiyama、Shinichi Nakajima、Hisashi Kashima、Paul Buenau 和 Motoaki Kawanabe。利用模型选择进行直接重要性估计及其在协变量偏移适应中的应用。《神经信息处理系统进展》，20，2007。Masashi Sugiyama、Taiji Suzuki 和 Takafumi Kanamori。机器学习中的密度比估计。剑桥大学出版社，2012。Jiangming Sun、Lars Carlsson、Ernst Ahlberg、Ulf Norinder、Ola Engkvist 和 Hongming Chen。应用蒙德里安交叉保形预测估计大型不平衡生物活性数据集的预测置信度。《化学信息与建模杂志》，57(7):1591–1598，2017。

Fredrik Svensson、Natalia Aniceto、Ulf Norinder、Isidro Cortés-Ciriano、Ola Spjuth、Lars Carlsson 和 Andreas Bender。用于定量构效关系模型的保形回归——量化预测不确定性。《化学信息与建模杂志》，58(5)：1132–1140，2018年。

Ryan J Tibshirani、Rina Foygel Barber、Emmanuel Candes 和 Aaditya Ramdas。协变量偏移下的保形预测。神经信息处理系统进展，32，2019年。

Oleg Trott 和 Arthur J Olson。AutoDock Vina：通过新的评分函数、高效优化和多线程提高分子对接的速度与准确性。计算化学杂志，31(2)：455–461，2010年。

Dennis Ulmer、Chrysoula Zerva 和 André F. T. Martins。利用最近邻进行非可交换保形语言生成。*arXiv* 预印本 *arXiv:2402.00707*，2024年。网址：https://arxiv.org/abs/2402.00707。

Vladimir Vovk、Alexander Gammerman 和 Glenn Shafer。《随机世界中的算法学习》，第 29 卷。施普林格，2005。

Renxiao Wang、Xueliang Fang、Yipin Lu、Chao-Yie Yang 和 Shaomeng Wang。PDBbind数据库：方法与更新。《药物化学杂志》，48(12):4111–4119，2005。

Tian Xie、Xiang Fu、Octavian-Eugen Ganea、Regina Barzilay 和 Tommi Jaakkola。用于周期性材料生成的晶体扩散变分自编码器。发表于《国际学习表征会议》，2022年。

Kevin Yang、Kyle Swanson、Wenhao Jin、Connor Coley、Philipp Eiden 等。分析用于性质预测的学习型分子表征。《化学信息与建模杂志》，59(8):3370–3388，2019年。doi: 10.1021/acs.jcim.9b00237。

Jason Yim、Andrew Campbell、Andrew Y. K. Foong、Michael Gastegger、José Jiménez-Luna、Sarah Lewis、Victor Garcia Satorras、Bastiaan S. Veeling、Regina Barzilay 和 Tommi Jaakkola。利用 SE(3)流匹配实现快速蛋白质主链生成。*arXiv* 预印本 *arXiv:2310.05297*，2023。网址：https://arxiv.org/abs/2310.05297。

Claudio Zeni、Robert Pinsler、Daniel Zügner、Andrew Fowler、Matthew Horton 等。Mattergen：用于无机材料设计的生成模型。*arXiv* 预印本 *arXiv:2312.03687*，2023。

Jin Zhang、Ulf Norinder 和 Fredrik Svensson。基于深度学习的毒性保形预测。《化学信息与建模杂志》，61(6)：2648–2657，2021年。

A 补充讨论

A.1 非随机保形p值

令 Π 为 $\{1,\dots,n_0,n+1,\dots,n+k\}$的所有置换的集合。对于任意置换$\pi \in \Pi$，我们将对观测特征进行置换 π 后得到的得分记为

$$V(\pi;\boldsymbol{X}) = V(X_{\pi(1)},\dots,X_{\pi(n_0)},X_{\pi(n+1)},\dots,X_{\pi(n+k)}).$$

W我们另外将恒等映射记为 π_0，即对任意 i，$\pi_0(i) = i$。最后，我们定义

$$p_k = \frac{\sum_{\pi\in\Pi} \bar{w}(\pi;\boldsymbol{X})\,\mathbb{1}\{V(\pi_0;\boldsymbol{X}) \le V(\pi;\boldsymbol{X})\}}{\sum_{\pi\in\Pi} \bar{w}(\pi;\boldsymbol{X})}. \tag{6}$$

此处，$\bar{w}(\pi;\boldsymbol{X}) = \prod_{j=1}^{k} w(X_{\pi(n+j)})$, 为置换 π之后的联合似然比，其中 $w(x)$为密度比。简而言之，p_k 考虑非活性校准样本与测试样本的所有排列，并计算由观测数据得到的得分$V(\pi_0;\boldsymbol{X})$在所有置换数据得分中的加权秩。

确定性p值的有效性由如下定理确立，其证明见附录B.3。

定理 A.1. 在协变量偏移假设下，对于任意固定的 $t \in [0,1]$，均有

$$\mathbb{P}\big(\, p_k \le t \,\big|\, max_{1\le j\le k} Y_{n+j} = 0\big) \le t.$$

A.2 联合加权可交换性

在本部分中，我们刻画零假设情形下非活跃校准数据 $\{X_i\}_{i\in\mathcal{I}_0}$ 和测试样本 $\{X_{n+j}\}_{j=1}^k$ 的可交换性结构

$$H_k\colon \max_{1\le j\le k} Y_{n+j}=0.$$

为一般起见，我们使用索引 k 表示 $\{X_{n+j}\}_{j=1}^k$ 可能是原始提议集的一个子集。以下引理刻画了给定标签时这些数据的分布。其证明只需应用贝叶斯法则，完整证明见附录 B.2 。

引理 *A.2.* 给定 $\{Z_i\}_{i=1}^n\cup\{Z_{n+j}\}_{j\ge1}$ 以及 $max_{1\le j\le k}Y_{n+j}=0$，特征$\{X_i\}_{i\in\mathcal{I}_0}$ 独立同分布地来自 $P_0:=P_{X\mid Y=0}$，而 $\{X_{n+j}\}_{j=1}^k$ 独立同分布地来自 $Q_0:=Q_{X\mid Y=0}$。此外，$dQ_0/dP_0(x)=w(x)$，其中 $w(x)=dQ\,dP(x,y)=dQ_X dP_X(x)$ 是密度比。

引理 A.2 指出，在给定所有标签的条件下，在“零假设”事件 H_k 成立时，“边际”密度比 $w(x)$ 足以校准 $\mathcal{I}_0$中的非活跃校准数据与新生成数据之间的分布偏移。下面我们将讨论其对密度比估计的影响。

备注 A.3。 当 $w(\cdot)$ 未知时，引理 *A.2* 证明了可以使用带标签数据的随机子集和一组独立的生成样本（无论它们是否活跃）来估计 $w(\cdot)$。另一种方法是直接估计 $\{X_i\}_{i\in\mathcal{I}_0}$ 与 $\{X_{n+j}\}_{j=1}^k$之间的密度比。然而，当 k 较小时，这种方法会导致相当有限的样本量。

引理 A.2 使我们能够刻画校准点与测试点之间的可交换性结构，将加权可交换性概念（Tibshirani 等人，2019年）扩展到多个测试点。这是条件排列式结果的一个推论，该结果扩展了保形预测中的基础性结果（Vovk 等人，2005；Tibshirani 等人，2019年）。

令 $[\boldsymbol{X}]=[X_1,\dots,X_{n_0},X_{n+1},\dots,X_k]$ 为特征组成的无序集合，并固定任意值$[\boldsymbol{x}]=[x_1,\dots,x_{n_0},x_{n_0+1},\dots,x_{n_0+k}]$。那么，给定 $[\boldsymbol{X}]=[\boldsymbol{x}]$后，剩余的唯一随机性在于 $\boldsymbol{x}$ 中的哪个值对应于每个 X_i。命题 A.4 直接由引理 A.2推出，它将 Tibshirani 等人 (2019年) 中的加权可交换性概念扩展到了多个测试点。

命题 A.4。 令 Π 为 $[n_0+k]$的所有置换的集合。那么，对于任意固定的值$x_1,\dots,x_{n_0},x_{n_0+1},\dots,x_{n_0+k}$，以及任意置换 $\pi\colon\ [n_0+k]\to[n_0+k]$，

$$\mathbb{P}^{\text{cond}}\Big(X_1=x_{\pi(1)},\dots,X_{n_0+k}=x_{\pi(n_0+k)}\,\Big|\,[\boldsymbol{X}]=[\boldsymbol{x}]\Big)=\frac{\bar{w}(\pi;x_{1:n_0+k})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';x_{1:n_0+k})},$$

其中我们定义 $\bar{w}(\pi;\ x_{1:n_0+k}):=\prod_{j=1}^k w(x_{\pi(n_0+j)})$，而 $\mathbb{P}^{\text{cond}}(\cdot)$ 表示在对 $\{Z_i\}_{i=1}^n\cup\{Z_{n+j}\}_{j\ge1}$ 以及 最大值$_{1\le j\le k}Y_{n+j}=0$进行条件化后的分布。

命题 A.4 描述了在给定其实现值的 $(X_1,\ \dots\dots,\ X_{n_0+k})$ 无序集合的条件下，$\boldsymbol{x}$ 取其任意置换值的概率。这种条件排列式结果是许多保形推断结果的基础，从而将我们的 p 值与现有的保形 p 值联系起来。具体而言，当 $w(\cdot)\equiv1$ 且 $m=1$时，它退化为标准的可交换性条件，并且 p 值（6）与经典保形 p 值一致。一般地，当 $m=1$时，它退化为 Tibshirani 等人（Tibshirani 等人，2019年）研究的加权可交换性，并且我们的 p 值（6）与加权保形 p 值（Tibshirani 等人，2019年；Jin 与 Candès，2023a）一致。我们将在下一小节讨论这些联系。

A.3 ONETESTPOINT 的保形 p 值

我们的 p 值建立在保形预测思想之上（Vovk 等人，2005）。为提供更多背景，我们简要讨论广泛使用的单个测试点的保形 p 值，这是推导保形预测集的关键概念。给定任意函数 $s\colon\ \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ （通常称为“非一致性分数”），其训练过程独立于校准数据 $\{(X_i,\ Y_i)\}_{i=1}^n$ 以及一个新的测试点 X_{n+1}，保形 p 值定义为

$$p(y)=\frac{1+\sum_{i=1}^n\mathbb{1}\{s(X_i,Y_i)\le s(X_{n+1},y)\}}{n+1}. \tag{7}$$

当 $\{(X_i, Y_i)\}_{i=1}^{n+1}$ 在 $i \in [n+1]$ 上可交换（例如独立同分布）时，可以证明 $p(Y_{n+1})$ 被 Unif$[0,1]$ 随机占优。该性质是为未知的Y_{n+1}构造预测集的关键。近来，基于可交换性的保形 p 值被用于多重检验场景，以检测离群值（Bates 等人，2023）并识别具有理想标签值的测试实例（Jin 与 Candès，2023b）。针对协变量偏移设定的扩展也分别在个体p值（Tibshirani 等人，2019年）和多重检验（Jin 与 Candès，2023a）中得到研究。

这里简要讨论备注 3.2。注意，当 $m = 1$，$n := n_0$ 以及 $w(x) \equiv 1$ 在 (7) 中，令

$$V(x_1, \ldots, x_{n_0}, x_{n+1}) = s(x_{n+1}, 0),$$

其中我们用 $y = 0$ 作为占位符，我们的 p 值 (6) 化简为 $p(0)$ 在 (7) 中。

A.4 协变量偏移下的异常值检测关联

更明确的关联在于我们的 p 值与异常值检测 p 值之间的关系。形式上，对于一个与校准“内点” X_{n+1} 独立的测试点 $\{X_i\}_{i=1}^{n} \overset{\text{独立同分布}}{\sim} P$，Bates 等人 (2023) 提议使用保形 p 值（为与当前记号保持一致，符号取反）

$$p = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\{s(X_i) \le s(X_{n+1})\}}{n+1}$$

以检验原假设 H_0：$X_{n+1} \sim P$，其中 s：$\mathcal{X} \to \mathbb{R}$ 是一个得分，用于表示在 P 下，值x 相较于正常值是异常值的可能性。在原假设 H_0 下，可以证明 p 值 p 被 Unif$[0,1]$ 随机变量随机支配。

Jin 与 Candès（2023a）将这一问题扩展到协变量偏移下的异常值检测，其中H_0： $X_{n+1} \sim Q$ 对于某个分布 Q 满足 $\mathrm{d}Q/\mathrm{d}P(x) = w(x)$，给定密度比 $w(\cdot)$。相应的加权保形 p 值为

$$p_w = \frac{w(X_{n+1}) + \sum_{i=1}^{n} w(X_i)\,\mathbb{1}\{s(X_i) \le s(X_{n+1})\}}{\sum_{i=1}^{n+1} w(X_i)}. \tag{8}$$

在这种情况下，我们的 p 值（6）通过令8而化为（

$$V(x_1, \ldots, x_{n_0}, x_{n+1}) = s(x_{n+1}).$$

也就是说，在我们的设定中，评分函数可以视为新测试点的“异常值”得分。

同样地，为保持一般性，我们使用索引 k（而不是 N）来表示 $\{X_{n+j}\}_{j=1}^{k}$ 可能是原始提议集的一个子集。最后，我们注意到，根据引理 A.2，我们的零假设 H_k：$\max_{1 \le j \le k} Y_{n+j} = 0$ 意味着 $X_{n+j} \overset{\text{独立同分布}}{\sim} Q_0$，其中 $\mathrm{d}Q_0/\mathrm{d}P_0(x) = w(x)$，且 $\{X_i\}_{i \in \mathcal{I}_0} \overset{\text{独立同分布}}{\sim} P_0$。因此，我们的 p 值 p_k 可以视为检验全局原假设

$$\tilde{H}_k : \{X_{n+j}\}_{j=1}^{k} \overset{\text{i.i.d.}}{\sim} Q_0$$

使用校准“内点” $\{X_i\}_{i \in \mathcal{I}_0} \overset{\text{独立同分布}}{\sim} P_0$。因此，$p_k$ 可被视为将 p_w （8）扩展到多个测试点的同时推断。

A.5 评分函数的选择

受到经典置换检验的启发，我们提出几种自然的选择：

- 最大池化。 由于我们的目标是判断任意一个测试样本是否足够有潜力，一个自然的选择是 $V(x_{1:\,n_0}, x_{n+1:\,n+k}) = \max_{1 \le j \le k} \hat{\mu}(x_{n+j})$，即预测测试分数的最大值。
- 预测值之和。受 Fisher 在因果推断中的随机化检验（Fisher &Fisher，1971年）启发，我们可以比较测试数据与校准数据之间预测值的平均值。在置换意义下，这等价于设定 $V(x_{1:\,n_0}, x_{n+1:\,n+k}) = \sum_{j=1}^{k} \hat{\mu}(x_{n+j})$。
- 秩和统计量。受 Wilcoxon 秩和检验的启发，我们比较测试点的秩和。对置换而言，我们设定 $V(x_{1:n_0}, x_{n+1:n+k}) = \sum_{j=1}^{k} R_{n+j}$，其中 R_{n+j} 是 $\hat{\mu}(X_{n+j})$ 在 $\{\hat{\mu}(X_i)\}_{i=1}^{n_0} \cup \{\hat{\mu}(X_{n+\ell})\}_{\ell=1}^{k}$ 中的秩（按升序排列）。

• 似然比统计量。受似然比检验的启发，我们注意到 $\hat{\mu}(x)$ 可视为估计量 $\mathbb{P}(Y=1 \mid X=x)$，因此 $\hat{\mu}(x)/(1-\hat{\mu}(X_{n+j}))$ 可作为在给定 X_{n+j} 的情况下检验 $Y_{n+j}=0$ 的似然比估计量。据此，我们考虑测试点的联合似然比，并设定 $V(x_{1:n_0}, x_{n+1:n+k}) = \sum_{j=1}^{k} \log(\hat{\mu}(x_{n+j})/(1-\hat{\mu}(x_{n+j}))$.

A.6 CONFHIT 的非拆分版本

正文中，我们讨论了当符合性分数基于预训练模型时的 ConfHit $\hat{\mu}$：基于 $\mathcal{X} \to \mathbb{R}$，这在拆分保形预测中是标准做法（Vovk 等人，2005）。这使我们的框架能够自然地兼容文献中的任何预训练预测模型（例如我们实验中使用的模型）。然而，实践者可能希望在模型训练和 p 值校准中都使用所有带标签数据。在本部分中，我们介绍 CONFHIT 的一种扩展：带标签的非活跃数据同时用于训练和校准，但仍能维持 I 类错误控制。值得注意的是，这一特性是我们的问题所独有的（因为我们以零假设事件上的标签为条件）。

形式上，令 $\mathcal{A}$ 为任意通用训练算法，它基于带标签的输入数据 $(Z_1, \ldots, Z_N)$ 生成预测模型 $\mathcal{A}(Z_1, \ldots, Z_N)$，其中 $Z_i = (X_i, Y_i)$。我们假设 $\mathcal{A}$ 的训练过程对其输入数据具有置换不变性，这一点对于许多常用算法都成立。

我们首先考虑认证问题。回顾一下，带标签的非活跃数据为 $\{(X_i, 0)\}_{i=1}^{n_0}$，带标签的活跃数据（不失一般性）为 $\{(X_i, 1)\}_{i=n_0+1}^{n}$，测试数据为 $\{X_{n+j}\}_{j=1}^{k}$. 我们考虑插补数据 $Z_{n+j} = (X_{n+j}, 0)$（针对 $1 \le j \le k$），并

$$\tilde{\mathcal{D}} = (Z_1, \ldots, Z_n, \tilde{Z}_{n+1}, \ldots, \tilde{Z}_{n+k}).$$

然后，我们令 $\hat{\mu}(\cdot) = \mathcal{A}(\tilde{\mathcal{D}})$，即使用插补数据训练得到的模型。最后，我们按照第 3.1 节中的相同方式，定义符合性分数 V 以及 p 值。随后，可以按照第 3.2 节中的相同方式，合并每个 $k \in \mathbb{N}^+$ 对应的这些 p 值，以解决设计问题。

不难看出，我们的 p 值在给定 $\max_{1 \le j \le k}\{Y_{n+j}\} = 0$ 的条件下仍然有效。由于训练得到的模型 $\hat{\mu}$ 对 $\mathcal{D}$ 中数据的置换保持不变，因此，对 $\{1, \ldots, n_0, n+1, \ldots, n+k\}$ 进行任意置换时，训练得到的模型都保持不变；符合性分数之间的加权可交换性也可通过完全相同的论证得出。

B 技术证明

B.1 定理的证明 3.4

定理 3.4 的证明。我们定义 $K_0 =$ 最大值$\{k \le N:$ 最大值$_{1 \le j \le k} Y_{n+j} = 0\}$，即所有更早的分子均为非正数时最后一个实例。我们还令 $K_0 = 0$ 当 $Z_{n+1} = 1$ 时，使得 K_0 有定义。注意，K_0 关于 $\{Y_{n+j}\}_{j \ge 1}$ 是可测的。给定 $\{Y_{n+j}\}_{j \ge 1}$，$\hat{K}_0$ 因此变为确定的，根据 N 的定义，我们有

$$\begin{aligned}\mathbb{P}\Big(\max_{1 \le j \le \hat{N}} Y_{n+j} = 0 \,\Big|\, \{Y_{n+j}\}_{j \ge 1}\Big) &= \mathbb{P}\Big(\hat{N} > 0,\ \max_{j \le \hat{N}} Z_{n+j} = 0 \,\Big|\, \{Y_{n+j}\}_{j \ge 1}\Big) \\ &= \mathbb{P}\Big(0 < \hat{N} \le K_0 \,\Big|\, \{Z_{n+j}\}_{j \ge 1}\Big)\, \mathbb{1}\{K_0 > 0\}. \qquad (*)\end{aligned}$$

由于 p 值是单调的，我们知道 $N \le K_0$ 等价于 $p_{K_0} \le \alpha$，因此

$$\begin{aligned}(*) &= \mathbb{P}\Big(p_{K_0} \le \alpha \,\Big|\, \{Z_{n+j}\}_{j \ge 1}\Big)\, \mathbb{1}\{K_0 > 0\} \\ &= \mathbb{P}\Big(p_{K_0} \le \alpha \,\Big|\, \{Z_{n+j}\}_{j \ge 1},\ \max_{j \le K_0} Z_{n+j} = 0\Big)\, \mathbb{1}\{K_0 > 0\} \le \alpha.\end{aligned}$$

这里第二行用到了 $\max_{j \le K_0} Z_{n+j}$ 关于 $\{Z_{n+j}\}_{j \ge 1}$ 可测这一事实，以及有效性条件 (ii)。最后，利用塔式性质，我们得到所需结果。 □

B.2 引理的证明 A.2

引理的证明 A.2。在给定所有校准标签和测试标签以及 最大值$_{1 \le j \le k} Y_{n+j} = 0$ 的条件下，显然，$\mathcal{I}_0$ 中的特征独立同分布于 $P_0 := P_{X \mid Y=0}$，而特征 $\{X_{n+j}\}_{j=1}^{k}$ 则是

独立同分布于 $Q_0 := Q_{X\,|\,Y=0}$。根据贝叶斯法则，这两个分布之间的关系为

$$\frac{\mathrm{d}Q_0}{\mathrm{d}P_0}(x) = \frac{Q(Y=0\,|\,X=x)}{P(Y=0\,|\,X=x)}\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x) = w(x),$$

其中，P_X、Q_X 分别是 X 在 P 与 Q下的分布。此处，我们利用了如下事实：由于协变量偏移假设，$Q(Y=0\,|\,X=x) = P(Y=0\,|\,X=x)$。 □

B.3 证明定理 A.1

定理 *A.1* 的证明。在整个过程中，我们在 $\{Z_i\}_{i\geq 1}$ 与 $\max_{j\leq k} Z_{n+j} = 0$的条件下进行讨论，并令 $\mathbb{P}^{\text{cond}}$ 表示在这些信息上条件化后的分布。对于任意固定向量 $\boldsymbol{x} \in \mathcal{X}^{n_0+k}$，在 $[\boldsymbol{X}] = [\boldsymbol{x}]$的条件下，我们令 $p_k(\pi;\boldsymbol{x})$为当数据实现为$(X_1,\ldots,X_{n_0},X_{n+1},\ldots,X_{n+k}) = (x_{\pi(1)},\ldots,x_{\pi(n_0)}\ x_{\pi(n+1)},\ldots,x_{\pi(n+k)})$时所得到的 p 值。为简单起见，我们假设$V(\pi;\boldsymbol{X})$几乎必然没有结，但所有论证在一般情况下仍然成立。

在引理A.2的条件下，命题A.4 意味着

$$\begin{aligned}&\mathbb{P}^{\text{cond}}(p_k = p_k(\pi;\boldsymbol{x})\,|\,[\boldsymbol{X}] = [\boldsymbol{x}])\\ &= \mathbb{P}^{\text{cond}}(\boldsymbol{X} = (x_{\pi(1)},\ldots,x_{\pi(n+k)})\,|\,[\boldsymbol{X}] = [\boldsymbol{x}]) = \frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})}.\end{aligned}$$

因此，注意到根据定义

$$p_k(\pi;\boldsymbol{x}) = \frac{\sum_{\tilde{\pi}\in\Pi}\bar{w}(\tilde{\pi}\circ\pi;\boldsymbol{x})\,\mathbb{1}\{V(\tilde{\pi}\circ\pi;\boldsymbol{x})\geq V(\pi;\boldsymbol{x})\}}{\sum_{\tilde{\pi}\in\Pi}\bar{w}(\tilde{\pi};\boldsymbol{x})} = \frac{\sum_{\tilde{\pi}\in\Pi}\bar{w}(\tilde{\pi};\boldsymbol{x})\,\mathbb{1}\{V(\tilde{\pi};\boldsymbol{x})\geq V(\pi;\boldsymbol{x})\}}{\sum_{\tilde{\pi}\in\Pi}\bar{w}(\tilde{\pi};\boldsymbol{x})},$$

并令 π_t 为使 $p_k(\pi_t;\boldsymbol{x}) \leq t$成立且 $V(\pi_t;\boldsymbol{x})$ 最大的排列，我们有

$$\begin{aligned}\mathbb{P}^{\text{cond}}(p_k\leq t\,|\,[\boldsymbol{X}] = [\boldsymbol{x}]) &= \sum_{\pi\in\Pi}\mathbb{1}\{p_k(\pi;\boldsymbol{x})\leq t\}\cdot\mathbb{P}^{\text{cond}}\big(p_k = p_k(\pi;\boldsymbol{x})\,\big|\,[\boldsymbol{X}] = [\boldsymbol{x}]\big)\\ &= \sum_{\pi\in\Pi}\mathbb{1}\{p_k(\pi;\boldsymbol{x})\leq t\}\cdot\frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})}\\ &= \sum_{\pi\in\Pi}\mathbb{1}\{V(\pi;\boldsymbol{x})\geq V(\pi_t;\boldsymbol{x})\}\cdot\frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})} = p_k(\pi_t;\boldsymbol{x})\leq t.\end{aligned}$$

这里，第一个等号使用了上文所述的条件分布；第三个等号使用了 $p_k(\pi;\boldsymbol{x})$ 关于 $V(\pi;\boldsymbol{x})$ 的单调性以及 π_t 的定义；最后一个等号则根据定义成立。随后，对 $[\boldsymbol{X}]$ 进行边缘化（但仍以 $\{Z_i\}_{i=1}^{n+k}$ 和 $\max_{1\leq j\leq k} Y_{n+j} = 0$ 为条件），即可得到所需结果。 □

B.4 定理3.1 的证明

定理*3.1*的证明。与往常一样，全文中我们均以 $\{Y_i\}_{i\geq 1}$ 和$\max_{1\leq j\leq k} Y_{n+j} = 0$为条件，并用 $\mathbb{P}^{\text{cond}}$ 表示给定这些信息后的分布。我们以事件 $[\boldsymbol{x}] = [\boldsymbol{x}]$ 为条件，其中对任意固定向量 $\boldsymbol{x} = (x_1,\ \ldots\ldots,\ x_{n_0},\ x_{n+1},\ \ldots\ldots,\ x_{n+k})$。与往常一样，标签$\{Z_i\}_{i=1}^{n+k}$ 也作为条件。随机性由以下因素产生：（i）$[\boldsymbol{x}]$中哪些值对应于每个数据点；（ii）随机抽取的排列。

令 $\hat{}\pi$ 为 $\{1,\ldots,n_0,n+1,\ldots,n+k\}$ 的唯一置换，使得 $\boldsymbol{X} = (x_{\hat{\pi}(1)},\ \ldots,\ x_{\hat{\pi}(n+k)})$。那么，对于任意置换 $\pi\in\Pi$，我们有

$$V(\pi;\boldsymbol{X}) = V(\pi\circ\pi^{(0)};\boldsymbol{X}) = V(\pi\circ\hat{\pi};\boldsymbol{x}),$$

对于 $\bar{w}(\pi;\ \boldsymbol{X}) = \bar{w}(\pi\circ\hat{}\pi;\ \boldsymbol{x})$，同理可得。根据命题 A.4，我们知道 $\mathbb{P}(\hat{}\pi = \pi\,|\,[\boldsymbol{X}] = [\boldsymbol{x}])$ $\propto \bar{w}(\pi;\boldsymbol{x})$，其中回顾一下 $\bar{w}(\pi;\ \boldsymbol{x}) = \prod_{j=1}^{k} w(x_{\pi(n+j)})$ 的定义。此外，我们有

$$p_k^{\text{rand}} = \frac{\bar{w}(\hat{\pi};\boldsymbol{x}) + \sum_{b=1}^{B}\bar{w}(\hat{\pi}\circ\pi^{(b)};\boldsymbol{x})\,\mathbb{1}\{V(\hat{\pi};\boldsymbol{x})\leq V(\hat{\pi}\circ\pi^{(b)};\boldsymbol{x})\}}{w(\hat{\pi};\boldsymbol{x}) + \sum_{b=1}^{B} w(\hat{\pi}\circ\pi^{(b)};\boldsymbol{x})},$$

其中 $\boldsymbol{x}$ 是固定向量，随机性来自$\hat{}\pi$ 和 $\{\pi^{(b)}\}_{b=1}^{B}$。

为简化记号，对于 $b = 1, \ldots, B$，我们记 $\hat{}\pi^{(0)} = \hat{}\pi$ 和 $\hat{}\pi^{(b)} = \hat{}\pi \circ \pi^{(b)}$。现在我们研究 ($\hat{}\pi^{(0)}, \hat{}\pi^{(1)}, \ldots, \hat{}\pi^{(B)}$) 的联合分布。对于任意固定的排列 $\pi_0, \ldots, \pi_B \in \Pi$，我们注意到

$$\begin{aligned}
&\mathbb{P}^{\text{cond}}\Big(\hat{\pi}^{(0)} = \pi_0, \hat{\pi}^{(1)} = \pi_1, \ldots, \hat{\pi}^{(B)} = \pi_B \,\Big|\, [\boldsymbol{X}] = [\boldsymbol{x}]\Big) \\
&= \mathbb{P}^{\text{cond}}\Big(\hat{\pi} = \pi_0, \pi^{(1)} = \pi_0^{-1} \circ \pi_1, \ldots, \hat{\pi}^{(B)} = \pi_0^{-1} \circ \pi_B \,\Big|\, [\boldsymbol{X}] = [\boldsymbol{x}]\Big) \\
&= \frac{1}{|\Pi|^B} \cdot \mathbb{P}^{\text{cond}}(\hat{\pi} = \pi_0 \,|\, [\boldsymbol{X}] = [\boldsymbol{x}]) \propto \bar{w}(\pi_0; \boldsymbol{x}),
\end{aligned}$$

其中 π_0^{-1}是 π_0的逆置换。上式中，最后一个等式利用了以下事实： $\pi^{(1)}, \ldots, \pi^{(B)}$是从 Π 中独立同分布地均匀采样的，并且相互独立。因此，当我们进一步以置换 $[\hat{}\pi^{(0)}, \hat{}\pi^{(1)}, \ldots, \hat{}\pi^{(B)}] = [\pi_0, \ldots, \pi_B]$ 所实现值的无序集为条件时，对于 $\pi_0, \ldots, \pi_B$的任意（固定）取值，我们知道

$$\begin{aligned}
&\mathbb{P}^{\text{cond}}\Big((\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}) = (\pi_{\sigma(0)}, \ldots, \pi_{\sigma(B)}) \,\Big|\, [\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}] = [\pi_0, \ldots, \pi_B],\ [\boldsymbol{X}] = [\boldsymbol{x}]\Big) \\
&= \frac{\bar{w}(\pi_{\sigma(0)}; \boldsymbol{x})}{\sum_{\sigma \in \tilde{\Pi}} \bar{w}(\pi_{\sigma(0)}; \boldsymbol{x})},
\end{aligned}$$

˜其中 Π 是 $\{0, 1, \ldots, B\}$的所有置换的集合。这也意味着，对于每个$b \in \{0, 1, \ldots, B\}$,

$$\mathbb{P}^{\text{cond}}\Big(\hat{\pi}^{(0)} = \pi_b \,\Big|\, [\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}] = [\pi_0, \ldots, \pi_B]\Big) = \frac{\bar{w}(\pi_b; \boldsymbol{x})}{\sum_{b'=0}^{B} \bar{w}(\pi_{b'}; \boldsymbol{x})}. \tag{9}$$

另一方面，对于任意 $\sigma \in \tilde{\Pi}$，当($\hat{}\pi^{(0)}, \hat{}\pi^{(1)}, \ldots, \hat{}\pi^{(B)}$) =($\pi_{\sigma(0)}, \ldots, \pi_{\sigma(B)}$)时，我们有

$$\begin{aligned}
p_k^{\text{cond}} &= \frac{\bar{w}(\pi_{\sigma(0)}; \boldsymbol{x}) + \sum_{b=1}^{B} \bar{w}(\pi_{\sigma(b)}; \boldsymbol{x})\, \mathbb{1}\{V(\pi_{\sigma(0)}; \boldsymbol{x}) \le V(\pi_{\sigma(b)}; \boldsymbol{x})\}}{\bar{w}(\pi_{\sigma(0)}; \boldsymbol{x}) + \sum_{b=1}^{B} \bar{w}(\pi_{\sigma(b)}; \boldsymbol{x})} \\
&= \frac{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})\, \mathbb{1}\{V(\pi_{\sigma(0)}; \boldsymbol{x}) \le V(\pi_b; \boldsymbol{x})\}}{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})}.
\end{aligned} \tag{10}$$

也就是说， p_k^{cond}的值只取决于数据置换 $\hat{}\pi^{(0)}$ 在（π_0、.....、 π_B）中取哪个值。因此，

$$\begin{aligned}
&\mathbb{P}^{\text{cond}}\Big(p_k^{\text{rand}} \le t \,\Big|\, [\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}] = [\pi_0, \ldots, \pi_B],\ [\boldsymbol{X}] = [\boldsymbol{x}]\Big) \\
&= \sum_{\sigma \in \tilde{\Pi}} \mathbb{P}\Big((\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}) = (\pi_{\sigma(0)}, \ldots, \pi_{\sigma(B)}) \,\Big|\, [\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}] = [\pi_0, \ldots, \pi_B]\Big) \\
&\qquad \times \mathbb{1}\left\{\frac{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})\, \mathbb{1}\{V(\pi_{\sigma(0)}; \boldsymbol{x}) \le V(\pi_b; \boldsymbol{x})\}}{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})} \le t\right\} \\
&= \sum_{b^*=0}^{B} \mathbb{P}\Big(\hat{\pi}^{(0)} = \pi_{b^*} \,\Big|\, [\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(B)}] = [\pi_0, \ldots, \pi_B]\Big) \\
&\qquad \times \mathbb{1}\left\{\frac{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})\, \mathbb{1}\{V(\pi_{b^*}; \boldsymbol{x}) \le V(\pi_b; \boldsymbol{x})\}}{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})} \le t\right\} \\
&= \sum_{b^*=0}^{B} \frac{\bar{w}(\pi_{b^*}; \boldsymbol{x})}{\sum_{b'=0}^{B} \bar{w}(\pi_{b'}; \boldsymbol{x})}\, \mathbb{1}\left\{\frac{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})\, \mathbb{1}\{V(\pi_{b^*}; \boldsymbol{x}) \le V(\pi_b; \boldsymbol{x})\}}{\sum_{b=0}^{B} \bar{w}(\pi_b; \boldsymbol{x})} \le t\right\} \le t.
\end{aligned}$$

这里，第一个等号由（10）得出，第三个等号由（9）得出，最后一个不等式则由定义得出。最后，对 $[\boldsymbol{X}]$ 和 $[\hat{}\pi^{(0)}, \ldots, \hat{}\pi^{(B)}]$ 进行边缘化，即可得到所需结果。 □

B.5 定理的证明 定理 3.5

本部分证明使用估计密度比的联合加权保形 p 值的鲁棒性。为保持一般性，我们使用一般索引 k 而非总预算 N，以表明 $\{X_{n+j}\}_{j=1}^{k}$ 可以是原始样本的一个子集（如第 3.2 节所示）。为便于理解，我们给出确定性p值的详细证明（附录 A.1）；正文中随机化p值的结果则可用相同思路自然推出。

定理 3.5 的证明。令 $\mathbb{P}^{\text{cond}}(\cdot)$ 表示在给定所有标签 $\{Y_i\}_{i\geq 1}$ 以及事件 $\max_{1\leq j\leq k} Y_{n+j} = 0$ 条件下的分布。按照定理 A.1 证明中的论证，对于任意固定值 $\boldsymbol{x}$，令 $p_k(\pi;\boldsymbol{x})$ 表示当数据实现为 $(X_1,\ldots,X_{n_0},X_{n+1},\ldots,X_{n+k}) = (x_{\pi(1)},\ldots,x_{\pi(n_0)}\ x_{\pi(n+1)},\ldots,x_{\pi(n+k)})$ 时，使用正确权重 $w(\cdot)$ 获得的 p 值；类似地，我们定义 $\hat{p}_k(\pi;\boldsymbol{x})$。于是有 $\hat{p}_k(V(\pi;\ \boldsymbol{x}))\ = \hat{p}_k(\pi;\ \boldsymbol{x})$，其中 $\hat{p}_k(v)$ 的定义见定理 3.5，并且根据定义，我们有

$$\sum_{\pi\in\Pi} \mathbb{1}\{\hat{p}_k(\pi;\boldsymbol{x}) \leq \hat{t}\} \cdot \frac{\hat{\bar{w}}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})} = \hat{p}_k(\hat{v}) \leq \hat{t} \leq t.$$

此外，根据 $\hat{t}$ 的定义，对于任意 $\pi\in\Pi$，$\hat{p}_k(\pi;\boldsymbol{x})\ \leq \hat{t}$ 当且仅当 $\hat{p}_k(\pi;\ \boldsymbol{x})\ \leq t$。因此，我们有

$$\begin{aligned}
&\mathbb{P}^{\text{cond}}\big(\hat{p}_k \leq t \,\big|\, [\boldsymbol{X}=\boldsymbol{x}]\big) = \sum_{\pi\in\Pi} \mathbb{1}\{\hat{p}_k(\pi;\boldsymbol{x}) \leq t\} \cdot \frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})} \\
&= \sum_{\pi\in\Pi} \mathbb{1}\{\hat{p}_k(\pi;\boldsymbol{x}) \leq \hat{t}\} \cdot \frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})} \\
&\leq \hat{t} + \sum_{\pi\in\Pi} \mathbb{1}\{\hat{p}_k(\pi;\boldsymbol{x}) \leq \hat{t}\}\left(\frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})} - \frac{\hat{\bar{w}}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})}\right) \\
&\leq \hat{t} + \sum_{\pi\in\Pi} \mathbb{1}\{V(\pi;\boldsymbol{x}) \geq \hat{v}\}\left(\frac{\bar{w}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})} - \frac{\hat{\bar{w}}(\pi;\boldsymbol{x})}{\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})}\right) \\
&\leq t + \frac{\sum_{\pi\in\Pi}\bar{w}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})} - \frac{\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}}{\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})} =: t + \hat{\Delta}.
\end{aligned}$$

重新排列各项后，我们得到

$$\begin{aligned}
\hat{\Delta} &= \frac{\big(\sum_{\pi\in\Pi}\bar{w}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big)\big(\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)}{\big(\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})\big)\big(\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})\big)} \\
&\quad - \frac{\big(\sum_{\pi\in\Pi}\bar{w}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)\big(\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big)}{\big(\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})\big)\big(\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})\big)} \\
&\leq \frac{\big(\sum_{\pi\in\Pi}[\hat{\bar{w}}(\pi;\boldsymbol{x})+\hat{\delta}^{-}(\pi;\boldsymbol{x})]\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big)\big(\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)}{\big(\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})\big)\big(\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})\big)} \\
&\quad - \frac{\big(\sum_{\pi\in\Pi}[\hat{\bar{w}}(\pi;\boldsymbol{x})-\hat{\delta}^{+}(\pi;\boldsymbol{x})]\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)\big(\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big)}{\big(\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})\big)\big(\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})\big)} \\
&\leq \frac{\big(\sum_{\pi\in\Pi}\hat{\delta}^{-}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big)\big(\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)}{\big(\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})\big)\big(\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})\big)} \\
&\quad + \frac{\big(\sum_{\pi\in\Pi}\hat{\delta}^{+}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)\big(\sum_{\pi\in\Pi}\hat{\bar{w}}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big)}{\big(\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})\big)\big(\sum_{\pi'\in\Pi}\hat{\bar{w}}(\pi';\boldsymbol{x})\big)},
\end{aligned}$$

其中我们使用了权重的非负性以及 $\hat{\delta}^{\pm}(\pi;\ \boldsymbol{x})$ 的定义。此外，注意到 $\hat{t} = \sum_{\pi\in\Pi}\hat{\bar{w}}(\pi\ \boldsymbol{X}\ \mathbb{1}\{v\leq V\ \pi\ \boldsymbol{X}\ \};)\hat{}(;)\hat{}(;)$，因此我们有 $\sum_{\pi\in\Pi}\bar{w}\ \pi\ \boldsymbol{X}\ \}$

$$\hat{\Delta} \leq \frac{(1-\hat{t})\cdot\big(\sum_{\pi\in\Pi}\hat{\delta}^{-}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})\geq\hat{v}\}\big) + \hat{t}\cdot\big(\sum_{\pi\in\Pi}\hat{\delta}^{+}(\pi;\boldsymbol{x})\,\mathbb{1}\{V(\pi;\boldsymbol{x})<\hat{v}\}\big)}{\sum_{\pi'\in\Pi}\bar{w}(\pi';\boldsymbol{x})},$$

由此完成了对 p_k 相关结果的证明。

最后，所有论证依然成立（参照定理 3.1 的证明中的一些论证），适用于 $\hat{}p_k$，只需将 Π 替换为 $\{\pi^{(b)}\}_{b=0}^{B}$，并额外以以下元素的无序集合为条件：$\{\hat{}\pi,\ \pi^{(1)},\ \ldots,\ \pi^{(B)}\}$，其中 $\hat{}\pi$ 是满足以下条件的置换：$\boldsymbol{X} = (x_{\pi(1)},\ \ldots,\ x_{\pi(n+k)})$。为简洁起见，此处省略详细内容。□

C 实验设置详情

C.1 受约束分子优化

训练生成器时，我们使用了 HGraph2Graph(Jin 等人，2020) 代码库中公开提供的训练代码。我们保留了默认的训练集/测试集划分，并随机从训练集中留出 20% 用于校准。测试集包含 800 个样本，也取自同一代码库。

对于 Self-Edit (Jiao 等人，2023) 基线，我们使用了相同的数据划分，并从其官方代码库获取了训练代码。两个模型均采用默认配置进行训练，未作修改。

在所有情况下，我们针对每个输入最多采样 50 个候选分子，直到获得足够数量的符合条件的候选项。我们将不符合条件的 候选项定义为无效（例如，化学上无效）、重复，或相对于训练数据属于分布外（OOD）的候选项。我们仅保留至少能生成 $N \in \{5, 10\}$ 个有效候选项的测试样本。

QED 的性质预言机来自 RDKit，DRD2 的性质预言机则来自 Therapeutics Data Commons 库。

密度估计与评分函数。我们使用 Chemprop（Yang 等人，2019年）库，针对 QED 和 DRD2 任务训练了一个分子性质预测器。模型采用一个 3 层消息传递神经网络，最终隐藏层维度为 4，用于特征提取；训练时使用默认的 Chemprop 训练器设置，共训练 5 个 epoch。我们使用了`MulticomponentMessagePassing` 模块，未作任何修改。

在进行密度估计时，我们采用 CODRUG 方法，提取倒数第二层特征，并使用 scipy 库拟合带宽为 1 的高斯核密度估计量（KDE）。为了过滤分布外样本，我们计算校准集上的第 95 百分位密度，并剔除测试时生成的、密度低于该阈值的分子。我们继续抽取额外样本，直到达到目标数量 N 个有效候选项。

进行序贯 p 值计算时，我们为每个测试实例使用 2000 个蒙特卡罗样本。

C.2 超参数选择及设计依据

密度比模型与特征选择。 对于密度估计，我们采用 CODRUG 流程，该流程已被证明在协变量偏移下进行小分子密度估计时具有稳健性。我们使用性质预测器的倒数第二层表示，得到一个 4 维特征向量；该向量表达能力强，同时维度足够低，能够实现稳定的核密度估计。对于带宽选择，我们采用与 CODRUG 相同的策略：对带宽 $\{0.1, 1, 10\}$ 进行 5 折交叉验证，计算每个留出划分上的核密度估计似然，并选择平均似然最高的带宽。这样可以避免事后调参，并确保核密度估计尽可能贴合校准分布。

预测模型选择。 保形有效性不受预测模型复杂度的影响，但使用更优的预测器和区分能力更强的评分函数时，该方法的检验效能会提高。为避免不必要的调参，我们选择已有报告显示性能出色且实现公开可用的模型。对于 CMO 任务（QED/DRD2），我们训练标准的 Chemprop 消息传递神经网络。对于 SBDD 任务，我们使用 TARGETDIFF 实现中的默认等变图神经网络预测器。这些选择简单、稳定，足以生成有信息量的保形评分。

检验统计量的选择。 由于该流程作用于生成分子的集合，得分统计量的选择会影响检验效能，但不会影响有效性。我们采用该集合中的**最大**得分，这在此情境下是一种自然的启发式方法：如果至少存在一个优质候选项，最大得分就能立即捕捉到这一证据；而均值或最小值等统计量会被较弱的候选项稀释信号。根据经验，我们发现，与其他方法相比，最大值统计量带来更高的检验效能和更低的拒绝率。

排列次数（B）。 我们设置为 $B = 2000$ 次蒙特卡洛置换。我们发现，这一数量足以确保性能稳定，同时计算成本适中。使用 ONFHIT 流程处理 $B = 2000$ 仅需 **4 分钟即可完成 800 个样本**（见图6）。我们还评估了 $B \in \{500, 1000, 2000, 5000\}$，观察到这些取值下的经验覆盖率方差可忽略不计，这证实了蒙特卡洛近似能够快速收敛。运行时间随 B线性增长，而 2000 则在统计稳定性与效率之间取得了有效平衡。

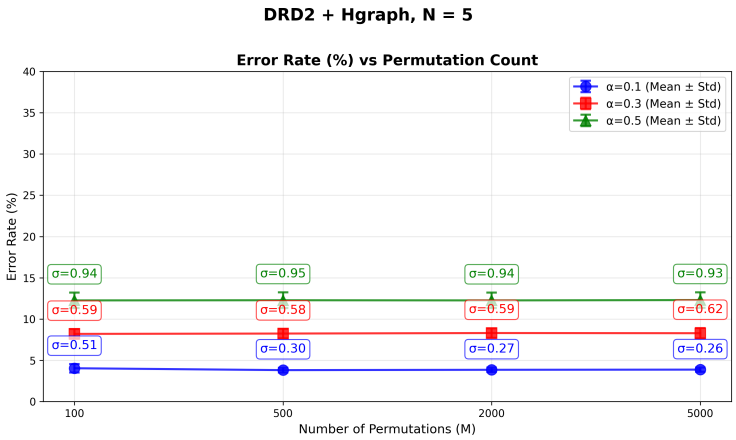


图 6：DRD2 使用 HGraph2Graph 时，错误率随蒙特卡洛置换次数 B 的变化情况（N = 5）。经验错误率保持稳定，方差极小。

C.3 基于结构的药物发现

对于此任务，我们使用了一个预训练的生成模型检查点，该检查点取自官方 ARGET-DIFF (Guan 等人, 2023) 代码库，并使用了与之对应的测试集。

为了估算得分并提取特征，我们使用其提供的代码库训练了一个基于等变图神经网络的模型。我们修改了最后一个隐藏层，使其输出一个 8 维嵌入。对于验证数据，我们使用了 CrossDocked2020（Francoeur 等人，2020a）数据集中的蛋白质-配体复合物，并排除了所有训练集和测试集条目，以避免数据泄漏。

密度估计按上述方法进行，使用带宽为 1 的高斯核密度估计。

C.4 计算资源

所有实验均在单台 NVIDIA A100 GPU 上进行。各组件的大致运行时间如下：

- 性质预测器训练（QED/DRD2）：20 分钟
- HGRAPH2GRAPH 生成器训练：10 小时
- SELF-EDIT 训练：48 小时
- 用于 SBDD 的密度估计器训练：24 小时
- 使用 TARGETDIFF 进行推理：30 分钟
- CONFHIT 流程每测试批次：30 分钟（超过 800 个样本；≈2.25 秒/样本）– 候选生成（CMO 任务）：14 分钟（≈1.05 秒/样本）

– 预测器特征提取：6 分钟（≈0.45 秒/样本）– 核密度估计计算：3 分钟（≈0.23 秒/样本）– 运行保形预测流程：4 分钟（≈0.30 秒/样本）

C.5 代码和数据可用性

我们随代码一并提供以下内容：

- 模型检查点。
- 用于 SBDD 校准的提取的特征和 CrossDocked 标识符。
- 校准和测试集划分。
- 用于在不同数据集和模型上复现结果的配置文件。
- 安装说明。

代码公开发布于 https://github.com/siddharthal/CONFHIT-Conformal-Generative-Design-with-Oracle-Free-Guarantees，并根据 MIT 许可证分发。

D 在总体预算下工作

在第4.6节中，我们讨论了 ConfHit 在固定总预算下构建预测集时提供实用启发式方法的应用。 B 详细算法见算法 2。

算法 2 在总体预算下工作 输入：校准数据 $\{X_i\}_{i=1}^{n_0}$，包含 $Y_i = 0$；测试批次 $\{X^{(t)}\}_{t=1}^{T}$，每批大小为 N；符合性分数 V；权重函数 $\hat{}w$；置信水平网格 $\mathcal{A} \subset (0,1)$；总预算 $B \in \mathbb{N}^{+}$。 1：初始化 $\hat{}$ $P_{\max} \leftarrow 0, \mathcal{C}^{\star} \leftarrow \emptyset, \alpha^{\star} \leftarrow 0$。 2：对于每个 $\alpha \in \mathcal{A}$，执行 3：对于每个测试批次 $t = 1, \ldots, T$，以水平 α 运行算法 1，得到修剪后的预测集 $\mathcal{C}_t^{(\alpha)}$。 4：计算总成本 $C = \sum_{t=1}^{T} |\mathcal{C}_t^{(\alpha)}|$。 5：记录空集数量 $E = |\{t : \mathcal{C}_t^{(\alpha)} = \emptyset\}|$。 6：如果 $C > B$，则 7：按大小降序排列 $\mathcal{C}_t^{(\alpha)}$。 8：迭代删除集合，直到总成本 $C \le B$。 9：令 D 为已删除集合的数量。 10：否则 11：设置 $D \leftarrow 0$。 12：结束 if 13：将正例数估计为 $\hat{P}(\alpha) = (1-\alpha) - E - D$。 $\hat{}\hat{}$ 14：如果 $P(\alpha) > P_{\max}$，则 15：更新 $\hat{}$ $P_{\max} \leftarrow \hat{P}(\alpha)$, $\mathcal{C}^{\star} \leftarrow \{\mathcal{C}_t^{(\alpha)}\}, \alpha^{\star} \leftarrow \alpha$。 16：结束 if 17：结束 for 输出：最佳正例数估计值 $\hat{}$ $P_{\max}$ 及其对应集合 $\mathcal{C}^{\star}$，显著性水平为 $\alpha^{\star}$。

E 附加实验

E.1 详细结果

本节为第4.3 节和4.4节中的图表提供误差线。这些表格列出了第4.3 节和4.4 节中所示的错误率、认证比例、空集比例和平均集合大小，以及在不同方法下进行 5 次随机运行所计算的标准差。此外，我们报告了受约束分子优化（CMO）任务的指标，使用药物相似性的定性估计（QED）属性；当$Y = 1$时视为成功，其中$\mathrm{QED}(x) > 0.9$，遵循Jin 等人（2020）的方法。

E.1.1 认证结果

表1：DRD2 HGRAPH 认证结果_

N	α	错误率（%）	认证比例（%）
3	0.1	6.5 ± 1.1	22.6 ± 2.4
3	0.2	9.6 ± 1.4	33.2 ± 4.1
3	0.3	18.5 ± 1.9	39.8 ± 4.7
3	0.4	26.8 ± 2.1	47.6 ± 3.5
3	0.5	26.0 ± 2.0	57.5 ± 6.8
5	0.1	7.3 ± 2.0	43.0 ± 4.3
5	0.2	9.3 ± 1.2	53.0 ± 4.6
5	0.3	12.2 ± 3.2	68.8 ± 6.2
5	0.4	21.8 ± 2.6	65.1 ± 4.6
5	0.5	21.6 ± 2.2	70.2 ± 6.1
7	0.1	5.5 ± 0.6	51.5 ± 5.1
7	0.2	8.1 ± 1.2	60.6 ± 2.5
7	0.3	11.8 ± 3.1	66.2 ± 5.2
7	0.4	19.7 ± 3.2	70.4 ± 3.6
7	0.5	20.6 ± 2.2	74.5 ± 6.2

表 2：DRD2 SELFEDIT 认证结果_

N	α	错误率 (%)	认证比例 (%)
3	0.1	7.1 ± 0.4	19.1 ± 2.1
3	0.2	8.7 ± 1.7	21.5 ± 3.1
3	0.3	19.8 ± 1.2	28.2 ± 3.2
3	0.4	19.8 ± 1.6	39.3 ± 4.1
3	0.5	28.3 ± 3.3	42.5 ± 5.2
5	0.1	7.0 ± 1.0	22.8 ± 3.1
5	0.2	7.4 ± 1.3	25.9 ± 2.5
5	0.3	16.4 ± 2.2	30.2 ± 4.1
5	0.4	17.4 ± 2.6	41.7 ± 4.6
5	0.5	26.9 ± 3.2	45.7 ± 5.1
7	0.1	6.3 ± 0.9	29.0 ± 2.2
7	0.2	9.8 ± 1.6	38.3 ± 3.1
7	0.3	16.8 ± 2.1	42.3 ± 4.7
7	0.4	17.1 ± 2.2	47.6 ± 4.9
7	0.5	26.4 ± 4.2	51.6 ± 5.6

表 3：QED_HGRAPH 认证结果

N	α	错误率（%）	认证比例（%）
3	0.1	7.8 ± 1.2	49.3 ± 2.0
3	0.2	13.4 ± 1.6	47.8 ± 1.8
3	0.3	16.3 ± 2.1	54.0 ± 1.0
3	0.4	19.4 ± 2.7	56.8 ± 1.0
3	0.5	21.8 ± 3.2	60.8 ± 3.4
7	0.1	7.9 ± 0.5	52.2 ± 3.2
7	0.2	14.6 ± 1.6	56.5 ± 2.0
7	0.3	17.9 ± 2.1	56.7 ± 2.2
7	0.4	20.3 ± 1.2	65.5 ± 3.1
7	0.5	22.5 ± 3.4	72.0 ± 2.7

表 4：QED_SELFEDIT 认证结果

N	α	错误率（%）	认证比例（%）
3	0.1	7.6 ± 1.2	42.6 ± 6.1
3	0.2	10.4 ± 0.6	54.6 ± 3.9
3	0.3	14.2 ± 1.3	52.2 ± 2.7
3	0.4	15.7 ± 2.1	57.0 ± 3.6
3	0.5	17.2 ± 3.7	62.7 ± 4.1
7	0.1	11.9 ± 1.4	52.2 ± 3.5
7	0.2	21.6 ± 2.1	62.3 ± 4.0
7	0.3	29.7 ± 2.5	61.2 ± 2.9
7	0.4	30.8 ± 1.2	77.9 ± 4.3
7	0.5	32.9 ± 3.2	78.4 ± 1.6

表 5：TARGETDIFF 认证结果

N	α	误差率（%）	认证比例（%）
5	0.1	8.4 ± 2.6	42.9 ± 2.5
5	0.2	10.8 ± 3.0	52.1 ± 2.9
5	0.3	11.4 ± 2.3	57.7 ± 3.8
5	0.4	14.9 ± 4.4	64.3 ± 3.6
5	0.5	17.1 ± 3.0	68.5 ± 3.4
10	0.1	5.3 ± 2.5	57.8 ± 4.3
10	0.2	8.2 ± 3.3	63.9 ± 3.6
10	0.3	9.3 ± 2.0	68.8 ± 2.6
10	0.4	11.5 ± 2.7	72.4 ± 4.0
10	0.5	12.8 ± 3.9	76.6 ± 3.7
15	0.1	10.1 ± 1.1	69.5 ± 3.3
15	0.2	7.6 ± 1.2	70.0 ± 4.6
15	0.3	8.7 ± 1.5	73.1 ± 5.1
15	0.4	9.8 ± 2.0	74.5 ± 2.5
15	0.5	12.0 ± 1.8	79.7 ± 2.3

表 6：MOLCRAFT 认证结果

N	α	误差率 (%)	认证比例 (%)
5	0.1	7.0 ± 1.2	40.1 ± 2.7
5	0.2	11.0 ± 1.8	47.9 ± 3.5
5	0.3	12.7 ± 1.0	55.5 ± 4.9
5	0.4	15.1 ± 1.2	63.1 ± 4.9
5	0.5	16.6 ± 1.2	68.1 ± 3.8
10	0.1	5.2 ± 1.6	58.3 ± 5.8
10	0.2	6.7 ± 1.1	64.2 ± 5.8
10	0.3	7.5 ± 1.4	67.6 ± 4.9
10	0.4	9.0 ± 1.6	71.1 ± 7.1
10	0.5	9.0 ± 2.9	73.5 ± 7.5
15	0.1	7.3 ± 1.0	55.4 ± 4.1
15	0.2	9.1 ± 2.0	59.8 ± 5.8
15	0.3	9.1 ± 3.6	68.7 ± 5.9
15	0.4	9.5 ± 3.0	69.1 ± 7.2
15	0.5	11.5 ± 4.9	76.2 ± 7.8

表 7：DECOMPDIFF 认证结果

N	α	误差率 (%)	认证比例（%）
5	0.1	3.2 ± 2.2	56.6 ± 1.4
5	0.2	3.1 ± 1.9	63.5 ± 1.2
5	0.3	4.7 ± 2.0	69.0 ± 2.4
5	0.4	5.2 ± 2.4	73.4 ± 0.8
5	0.5	5.4 ± 2.6	77.0 ± 1.2
10	0.1	2.1 ± 1.6	74.4 ± 4.0
10	0.2	3.1 ± 2.5	76.4 ± 3.5
10	0.3	5.2 ± 3.1	79.6 ± 3.4
10	0.4	6.2 ± 2.8	82.7 ± 3.6
10	0.5	5.9 ± 3.0	82.7 ± 2.1
15	0.1	2.6 ± 3.2	83.2 ± 5.7
15	0.2	2.5 ± 3.2	83.9 ± 5.0
15	0.3	1.6 ± 2.7	85.8 ± 5.3
15	0.4	1.6 ± 2.8	85.1 ± 6.2
15	0.5	2.5 ± 3.1	87.1 ± 5.2

E.1.2 设计结果

DRD2 HGRAPH_

表 8：DRD2_HGRAPH 设计结果

N	α	错误率（%）	平均集合大小	空集（%）
3	0.1	4.2 ± 0.3	1.9 ± 0.0	68.8 ± 1.3
3	0.2	6.6 ± 0.5	1.8 ± 0.0	46.1 ± 1.3
3	0.3	9.1 ± 0.6	1.8 ± 0.0	25.3 ± 1.8
3	0.4	11.0 ± 0.8	1.6 ± 0.0	15.0 ± 1.3
3	0.5	13.0 ± 0.8	1.5 ± 0.0	9.9 ± 0.8
5	0.1	5.4 ± 1.6	2.4 ± 0.7	47.9 ± 2.6
5	0.2	7.9 ± 2.5	2.1 ± 0.7	28.7 ± 2.8
5	0.3	10.4 ± 3.3	1.8 ± 0.6	15.8 ± 2.1
5	0.4	12.4 ± 4.0	1.6 ± 0.5	11.3 ± 1.9
5	0.5	13.9 ± 4.4	1.4 ± 0.5	8.8 ± 1.5
7	0.1	3.6 ± 0.8	3.1 ± 0.1	53.0 ± 1.2
7	0.2	6.3 ± 0.7	2.6 ± 0.1	30.0 ± 1.4
7	0.3	9.0 ± 1.0	2.1 ± 0.0	15.8 ± 0.7
7	0.4	11.2 ± 1.1	1.8 ± 0.0	9.2 ± 0.5
7	0.5	12.9 ± 1.2	1.7 ± 0.0	5.9 ± 0.6

DRD2 SELFEDIT_

表 9：DRD2_SELFEDIT 设计结果

N	α	错误率（%）	平均集合大小	空集（%）
3	0.1	9.8 ± 1.3	2.0 ± 0.0	64.5 ± 0.5
3	0.2	14.6 ± 2.2	1.9 ± 0.1	42.3 ± 0.2
3	0.3	19.6 ± 3.2	1.7 ± 0.0	24.0 ± 0.9
3	0.4	22.5 ± 3.2	1.6 ± 0.0	12.3 ± 0.1
3	0.5	24.5 ± 4.0	1.5 ± 0.0	7.3 ± 0.6
5	0.1	9.5 ± 1.2	2.6 ± 0.1	53.0 ± 1.5
5	0.2	14.0 ± 2.4	2.3 ± 0.1	33.3 ± 2.6
5	0.3	20.7 ± 2.1	2.0 ± 0.0	14.9 ± 0.6
5	0.4	25.8 ± 2.6	1.7 ± 0.1	8.1 ± 1.1
5	0.5	28.8 ± 3.1	1.6 ± 0.0	4.2 ± 0.4
7	0.1	11.8 ± 1.3	3.1 ± 0.1	46.6 ± 0.8
7	0.2	19.5 ± 3.3	2.6 ± 0.0	22.9 ± 3.5
7	0.3	24.8 ± 6.3	2.1 ± 0.1	12.2 ± 1.9
7	0.4	28.6 ± 7.1	1.8 ± 0.1	7.4 ± 1.7
7	0.5	31.6 ± 6.3	1.7 ± 0.0	4.0 ± 0.6

QED_HGRAPH

表 10：QED_HGRAPH 设计结果

N	α	错误率 (%)	平均集合大小	空集 (%)
3	0.1	10.1 ± 0.7	1.2 ± 0.0	70.7 ± 1.9
3	0.2	17.0 ± 0.6	1.3 ± 0.0	45.4 ± 1.6
3	0.3	21.2 ± 1.4	1.3 ± 0.1	26.8 ± 0.9
3	0.4	25.2 ± 1.5	1.3 ± 0.0	13.2 ± 1.2
3	0.5	28.2 ± 1.7	1.2 ± 0.0	6.9 ± 0.6
5	0.1	12.2 ± 0.7	2.6 ± 0.0	27.2 ± 0.2
5	0.2	18.8 ± 1.1	2.3 ± 0.1	7.1 ± 0.2
5	0.3	24.4 ± 1.2	1.8 ± 0.0	0.0 ± 0.0
5	0.4	27.7 ± 1.3	1.6 ± 0.1	0.0 ± 0.0
5	0.5	29.9 ± 1.0	1.5 ± 0.1	0.0 ± 0.0
7	0.1	10.4 ± 1.2	2.5 ± 0.1	56.6 ± 0.8
7	0.2	17.5 ± 1.2	2.0 ± 0.0	31.7 ± 1.8
7	0.3	22.0 ± 1.4	1.7 ± 0.1	18.9 ± 1.6
7	0.4	26.1 ± 1.5	1.4 ± 0.0	9.4 ± 1.5
7	0.5	29.1 ± 1.6	1.3 ± 0.0	4.9 ± 1.1

QED_SELFEDIT

表 11：QED_SELFEDIT 设计结果

N	α	错误率 (%)	平均集合大小	空集 (%)
3	0.1	10.1 ± 0.4	1.2 ± 0.0	64.8 ± 1.8
3	0.2	14.4 ± 0.9	1.3 ± 0.1	45.2 ± 1.5
3	0.3	18.8 ± 1.4	1.3 ± 0.1	27.1 ± 1.8
3	0.4	21.5 ± 1.9	1.2 ± 0.1	13.3 ± 1.0
3	0.5	23.4 ± 1.9	1.2 ± 0.0	7.0 ± 1.0
5	0.1	11.9 ± 0.8	2.7 ± 0.1	26.0 ± 2.4
5	0.2	20.1 ± 1.2	2.3 ± 0.1	7.0 ± 0.7
5	0.3	25.8 ± 1.4	1.9 ± 0.1	0.0 ± 0.0
5	0.4	28.4 ± 1.5	1.6 ± 0.1	0.0 ± 0.0
5	0.5	29.9 ± 2.1	1.5 ± 0.2	0.0 ± 0.0
7	0.1	18.2 ± 2.7	2.1 ± 0.2	56.7 ± 1.5
7	0.2	26.9 ± 3.8	1.9 ± 0.2	34.7 ± 2.9
7	0.3	34.5 ± 3.0	1.6 ± 0.1	18.3 ± 2.2
7	0.4	39.2 ± 3.0	1.4 ± 0.1	10.2 ± 2.5
7	0.5	42.1 ± 2.5	1.3 ± 0.0	5.7 ± 1.6

TARGETDIFF

表 12：TARGETDIFF 设计结果

N	α	错误率（%）	平均集合大小	空集（%）
5	0.1	10.0 ± 1.5	2.5 ± 0.1	37.9 ± 2.6
5	0.2	17.4 ± 1.5	2.2 ± 0.1	21.4 ± 3.1
5	0.3	27.2 ± 3.5	2.0 ± 0.1	8.6 ± 0.0
5	0.4	34.2 ± 2.9	1.7 ± 0.1	2.1 ± 1.4
5	0.5	37.0 ± 1.2	1.6 ± 0.0	0.4 ± 0.7
10	0.1	9.4 ± 2.7	3.7 ± 0.1	21.9 ± 8.0
10	0.2	22.5 ± 3.5	2.9 ± 0.0	6.2 ± 3.5
10	0.3	29.4 ± 0.9	2.3 ± 0.0	1.2 ± 0.0
10	0.4	35.6 ± 0.9	1.8 ± 0.0	0.0 ± 0.0
10	0.5	37.5 ± 3.5	1.6 ± 0.1	0.0 ± 0.0
15	0.1	11.2 ± 1.8	4.7 ± 0.1	15.0 ± 5.3
15	0.2	21.9 ± 2.7	3.5 ± 0.0	3.8 ± 1.8
15	0.3	31.9 ± 0.9	2.4 ± 0.0	1.2 ± 1.8
15	0.4	36.9 ± 2.7	1.8 ± 0.0	0.6 ± 0.9
15	0.5	38.1 ± 2.7	1.6 ± 0.1	0.0 ± 0.0

MOLCRAFT

表 13：MOLCRAFT 设计结果

N	α	错误率（%）	平均集合大小	空集（%）
5	0.1	9.8 ± 1.2	2.6 ± 0.1	40.7 ± 1.9
5	0.2	18.7 ± 1.4	2.3 ± 0.2	20.7 ± 5.6
5	0.3	26.0 ± 3.1	2.0 ± 0.1	7.7 ± 0.7
5	0.4	29.3 ± 5.3	1.7 ± 0.0	2.4 ± 0.0
5	0.5	32.9 ± 4.2	1.6 ± 0.0	1.2 ± 1.2
10	0.1	9.9 ± 1.3	3.6 ± 0.3	22.2 ± 3.3
10	0.2	18.2 ± 2.0	3.0 ± 0.2	9.9 ± 2.1
10	0.3	25.1 ± 0.7	2.2 ± 0.1	2.9 ± 1.4
10	0.4	28.0 ± 1.4	1.8 ± 0.1	2.1 ± 0.7
10	0.5	29.6 ± 2.5	1.7 ± 0.1	0.4 ± 0.7
15	0.1	11.1 ± 0.0	4.3 ± 0.0	16.0 ± 0.0
15	0.2	19.8 ± 0.0	3.2 ± 0.0	1.2 ± 0.0
15	0.3	23.5 ± 0.0	2.2 ± 0.0	1.2 ± 0.0
15	0.4	27.2 ± 0.0	1.8 ± 0.0	1.2 ± 0.0
15	0.5	28.4 ± 0.0	1.6 ± 0.0	1.2 ± 0.0

DECOMPDIFF

表 14：DECOMPDIFF 设计结果

N	α	错误率（%）	平均集合大小	空集（%）
5	0.1	9.6 ± 2.6	2.6 ± 0.3	25.0 ± 2.2
5	0.2	16.7 ± 4.7	2.3 ± 0.2	11.7 ± 0.7
5	0.3	24.2 ± 0.7	2.0 ± 0.0	2.5 ± 2.2
5	0.4	26.7 ± 0.7	1.7 ± 0.1	1.2 ± 0.0
5	0.5	28.8 ± 1.2	1.6 ± 0.1	0.8 ± 0.7
10	0.1	11.2 ± 0.5	3.4 ± 0.1	13.5 ± 0.9
10	0.2	19.2 ± 0.1	2.7 ± 0.0	2.6 ± 0.0
10	0.3	24.4 ± 3.6	2.1 ± 0.1	0.0 ± 0.0
10	0.4	26.3 ± 0.9	1.7 ± 0.1	0.0 ± 0.0
10	0.5	27.6 ± 0.9	1.6 ± 0.1	0.0 ± 0.0
15	0.1	9.6 ± 0.9	4.3 ± 0.4	8.3 ± 2.7
15	0.2	19.2 ± 3.6	2.9 ± 0.1	1.3 ± 1.8
15	0.3	23.7 ± 4.5	2.3 ± 0.1	0.0 ± 0.0
15	0.4	26.9 ± 3.6	1.8 ± 0.0	0.0 ± 0.0
15	0.5	27.6 ± 2.7	1.7 ± 0.1	0.0 ± 0.0

E.2 比较不同的检验统计量

在本节中，我们补充了第4.5 节比较不同检验统计量的额外结果。虽然检验效能会因所选检验统计量而变化，但 CONFHIT 在各种统计量上均取得了有效覆盖率。

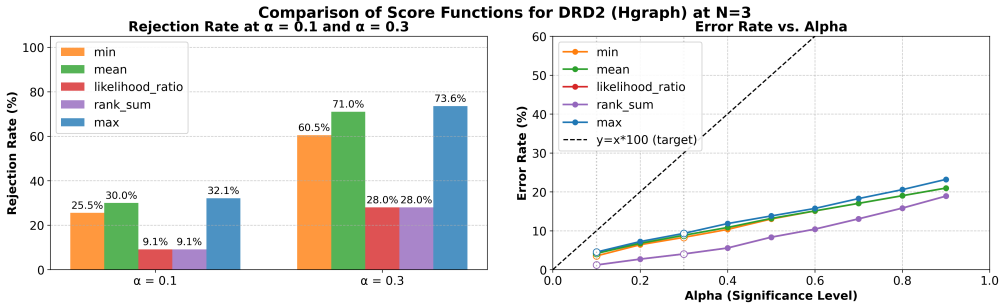


图 7：DRD2（HGRAPH2GRAPH，N=3）上分数统计量的比较。左：$\alpha = 0.1$,0.3 处的拒绝（检验效能）。右：误差与 α 的关系

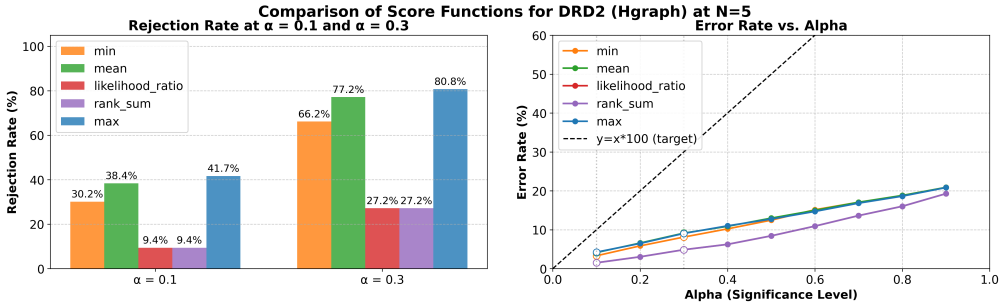


图 8：DRD2（HGRAPH2GRAPH，N=5）上分数统计量的比较。左：$\alpha = 0.1$,0.3 处的拒绝（检验效能）。右：误差与 α 的关系

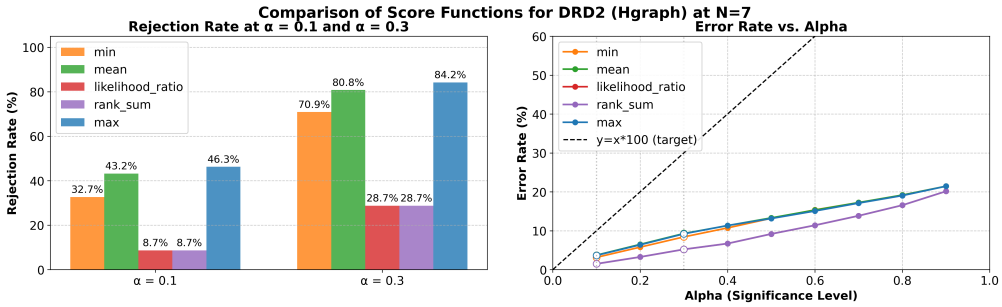


图 9：DRD2（HGRAPH2GRAPH，N=7）上分数统计量的比较。左：$\alpha = 0.1$,0.3 处的拒绝（检验效能）。右：误差与 α 的关系

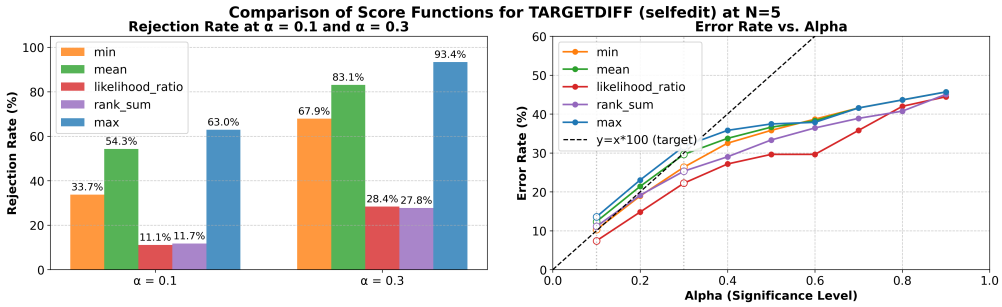


图 10：SBDD（TARGETDIFF，N=5）上分数统计量的比较。左：$\alpha = 0.1$，0.3 时的拒绝（检验效能）。右：误差与 α 的关系

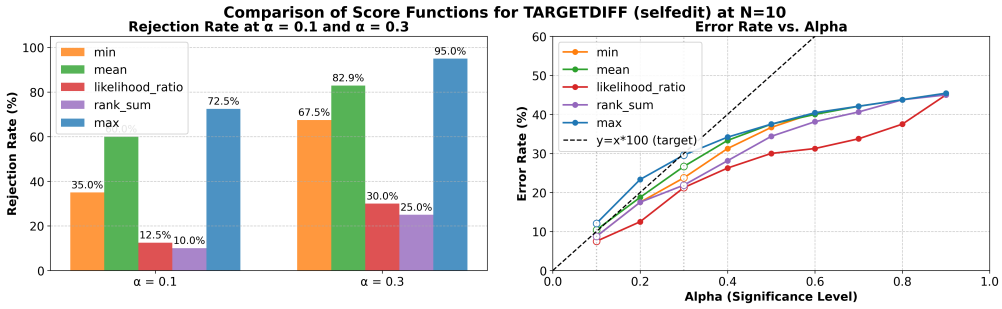


图 11：SBDD（TARGETDIFF，N=10）上分数统计量的比较。左：
0.1 0.3 α $\alpha =$ ，时的拒绝（检验效能）。右：误差与 的关系

E.3 P 值

在本节中，我们展示了第4.5节所述保形 p 值图在各项任务和数据集上的有效性。每个面板中的分布都接近均匀分布（虚线；另见报告的与均匀分布之间的 KL 散度），证实了我们的 p 值近似有效（尽管略显保守）。

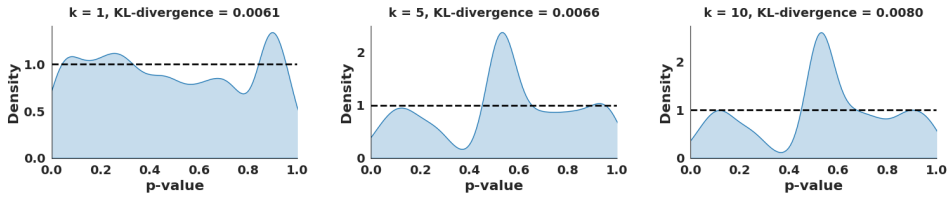


图 12：使用 SELFEDIT 时 QED 的保形嵌套 p 值密度，集合大小为 $k = 3$，5、7。

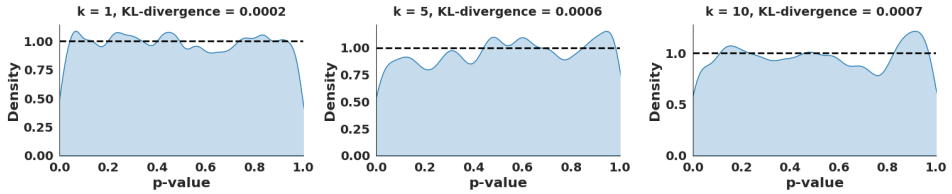


图 13：使用 HGRAPH 时 DRD2 的保形嵌套 p 值密度，集合大小为 $k = 3$、5、7。

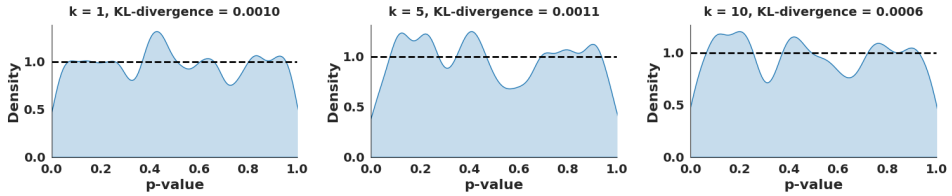


图 14：使用 TARGETDIFF 时 SBDD 的保形嵌套 p 值密度，集合大小为 $k = 1$、5、10。

F 预测器质量的影响

本节探讨预测器的准确性如何影响 CONFHIT 的性能。我们考虑三种情形：

1. 普通预测器：校准分类器输出 $p = \Pr(\text{active} \mid x)$。2. 噪声预测器：通过添加均匀高斯噪声得到，结果为

$$\tilde{p} = \frac{p + \mathcal{N}(0,1)}{2}.$$

3. 逆向预测器：通过将 p 替换为 $1 - p$得到。

图15 展示了 CONFHIT 在这些扰动设置下的表现。即使预测器受到干扰或被反转，该方法仍能实现有效的错误控制，经验错误率始终低于目标线 $y = 100\alpha$。这表明 CONFHIT 的统计有效性并不依赖于预测器的准确性。

预测器质量仅影响检验效能。随着预测器的信息量降低，保形集的效率下降，这会导致空集出现频率更高，或根据任务不同而产生更大的集合。尽管存在这种预期的性能下降，该方法仍始终保持其覆盖保证。

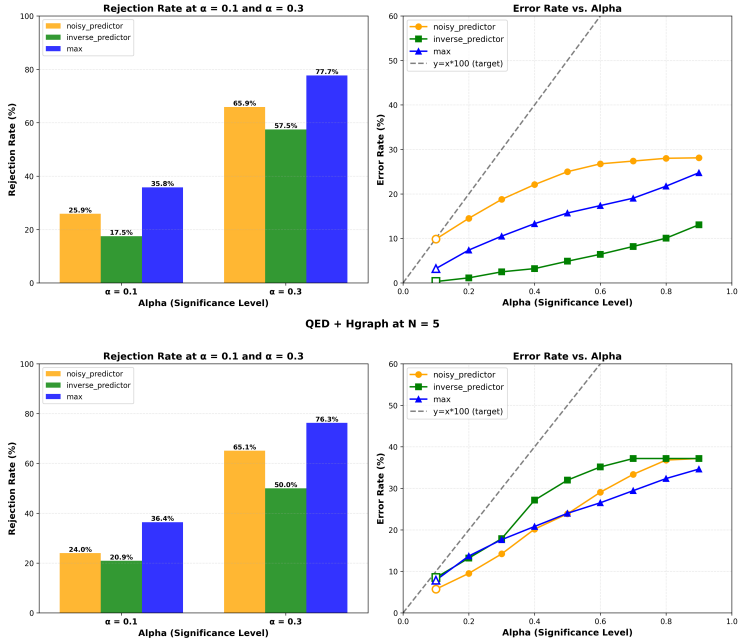


图 15：预测器质量对 CONFHIT 在 CMO 任务上的错误率的影响。在使用 Hgraph 模型时，无论是 DRD2（上图）还是 QED （下图），噪声预测器和反向预测器都会降低检验效能，但在 α 的所有取值下都能维持有效的错误控制。

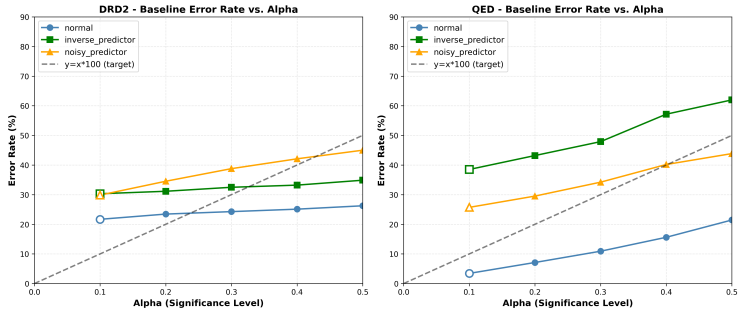


图 16：使用预测概率的启发式基线结果（改变预测器质量），展示了使用 Hgraph 模型、 $N = 5$ 时 QED 的拒绝率和错误率。随着预测器质量下降，性能迅速下滑，并严重违反错误控制要求。

G 其他基线

G.1 启发式基线

我们评估一种简单的基于置信度的启发式方法。给定成功的预测概率$\hat{p}$，该启发式方法选择满足$(1 - \hat{p})^n \leq \alpha$的最小 n。这提供了一个自然的基线，它仅依赖于分类器分数，而无需任何不确定性校准。

图 16 和图表明，这种启发式方法无法可靠地控制错误率。预测器准确时，错误率有时会贴合目标线；但在预测器存在噪声或方向相反时，误差会急剧增加，且往往超过预期水平。DRD2 和 QED 均可见这种性能退化：随着预测器质量下降，错误曲线不断上升，拒绝率也显著波动，说明其表现不稳定。

相比之下，CONFHIT 在所有情况下都保持在目标线以下。预测器质量只影响检验效能，而启发式基线一旦预测器变得不可靠，就会失去错误控制。

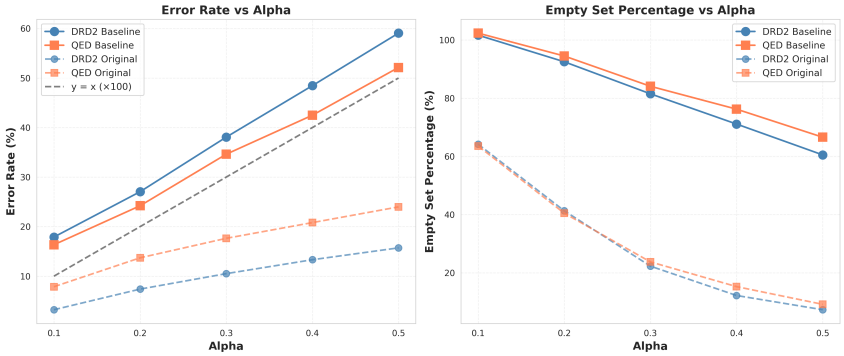


图 17：用于分子优化的保形语言模型基线。使用 Hgraph 模型时，QED 和 DRD2 在 N = 5 下的错误率（左）和空集百分比（右）随 α 变化。尽管该方法依赖于现实场景中无法获得的预言机，但它未能维持覆盖率，且表现不如 CONFHIT。

经验曲线表明，启发式方法的表现会从尚可转为糟糕，而 CONFHIT 在所有预测器设置下都能保持稳定的保证。

G.2 与基于预言机的方法比较

我们改造了一个基于预言机的保形语言模型（Quach 等人，2023）基线，如算法 3所示，将预测分数之和作为候选集的总置信度，而不进行任何密度校正或低质量样本过滤。如图 17所示，该基线无法维持有效保证：经验错误率始终高于目标水平，并且在所有 α中，空集百分比都明显偏离预期行为。尽管依赖理想化的预言机，该方法的检验效能较低，不及 CONFHIT；而 CONFHIT 凭借其经过校准的评分规则和密度感知校正，同时实现了更高的检测功效和有效覆盖率。最后，我们强调，该基线无法与 CONFHIT 直接比较，因为它假设可以访问完美的预言机预测器——这在实际分子设计工作流中是不现实的要求。

算法 3 约束分子优化的保形采样 输入：输入分子 x；生成器 $p_\theta(\cdot \mid x)$；评分函数 $\mathcal{F}$；校准阈值 λ；采样预算 $k_{\max}$输出：候选集 C_λ，经认证至少包含一个改进分子 1: $C_\lambda \leftarrow \varnothing$ **{初始化输出集} 2: 对** $k = 1$至 $k_{\max}$ **执行 3: 采样** $y_k \sim p_\theta(\cdot \mid x)$ **{生成候选项} 4:** $C_\lambda \leftarrow C_\lambda \cup \{y_k\}$ **{加入集合} 5: 如果** $\mathcal{F}(x, C_\lambda) \geq \lambda$，**则 6: 跳出 {设置认证标记} 7: 结束如果 8: 结束循环 9: 返回** C_λ

H 密度比估计的鲁棒性与诊断

H.1 敏感性分析

我们通过对似然比进行幂变换，研究似然比设定错误所带来的影响 w^γ，并考察一系列指数 γ。图 18 展示了此类变换如何影响负样本的 p 值的 KL 散度，以及由此产生的错误率和拒绝（空集）率。在实践中，这些结果可用于判断分析是否足够稳健。在此，我们

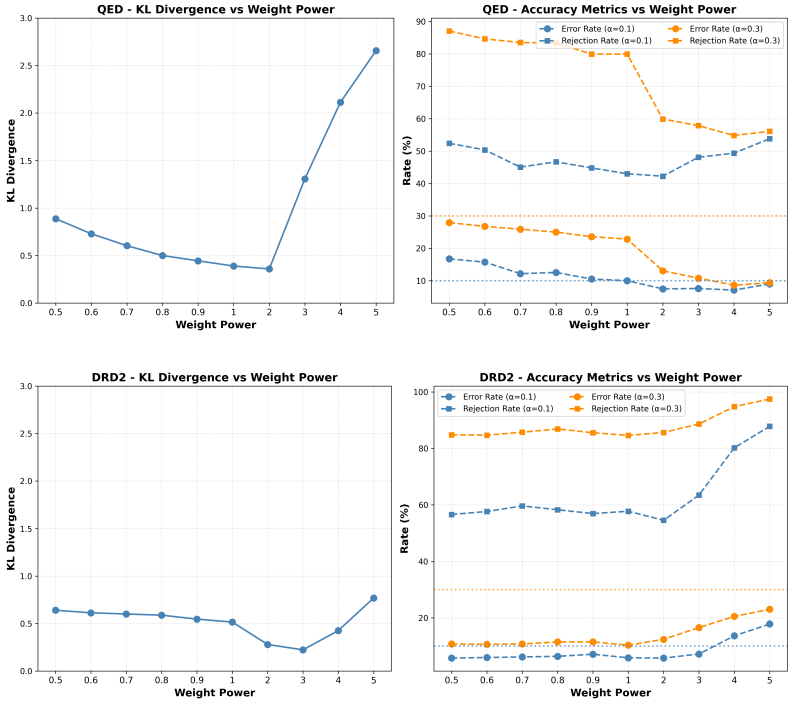


图18：使用 HGraph2Graph 模型时，似然比权重的幂变换 w^γ 对 QED 和 DRD2 的影响；在不同的 α 下 $N = 5$。在所有变换下，CONFHIT 均能在拒绝率发生显著变化时，保持错误控制的平稳退化。

还评估了此类扰动下的错误控制表现，以进一步展示 CONFHIT 的稳健性能。

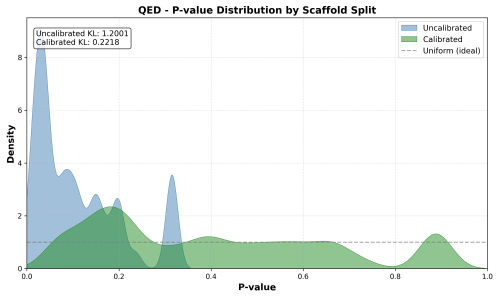
此处的 KL 散度衡量经验 p 值分布与均匀分布（0,1）之间的偏差。数值越小，表示变换后的似然比仍能产生接近均匀分布的 p 值；数值越大，则表明偏离理想均匀形态的程度越高。在这两个数据集中，我们发现，在中度错误设定下，KL 散度仅受到中等程度的影响。只有在放大权重（$\gamma > 3$）时，QED 案例中的 KL 散度才会急剧上升，显示出明显的均匀性偏离。

对于 $\alpha = 0.1$，错误率始终接近目标水平；对于 $\alpha = 0.3$，则低于目标水平。受到影响的只有拒绝率：KL 散度较小时，拒绝率较低；p 值偏离均匀分布越明显，拒绝率越高。这表明，轻微的错误设定主要会改变方法的保守程度，而不会影响其有效性。

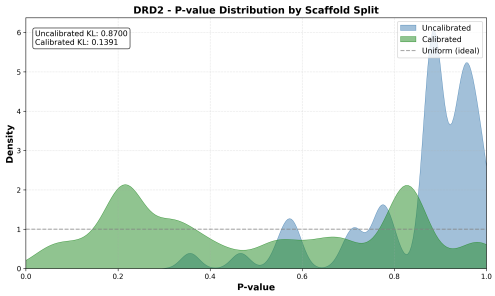
DRD2 也呈现出类似的模式。权重扁平化会降低 KL 散度，而较大的 γ 值会使其增大。错误控制始终保持稳定，而在 KL 值较高的区域，拒绝率会上升；该区域对应着 p 值分布更严重的失真。总体而言，这些结果表明，ConfHit 对似然比的中度错误设定具有鲁棒性。

H.2 在校准集上进行骨架划分

为评估校准程序的鲁棒性，我们对校准负样本进行分子骨架划分。我们选取也出现在测试集中的前 30 个分子骨架，由此产生非平凡的分布偏移，反映了骨架多样性发挥重要作用的真实药物发现情境。此设置提供了一项诊断性测试：如果校准程序有效，那么在这一偏移划分上得到的 p 值应仍接近均匀分布。



(a) QED + Hgraph：校准显著降低了 p 值分布偏离均匀性的程度。



(b) DRD2 + Hgraph：校准后的 p 值与均匀分布高度吻合。

图 19：QED（左）和 DRD2（右）在骨架划分校准负样本上的 p 值分布。两种情况下，校准后得到的 p 值都与均匀分布高度吻合，表明密度校正在基于骨架的分布偏移下仍然可靠。

图 19a和 19b 展示了应用密度比校正前后的 p 值分布。在 QED 和 DRD2 中，未经校准的 p 值都明显偏离理想的均匀分布，表现为较大的 KL 散度。校准后，p 值明显更接近均匀分布，KL 散度显著降低，且整个区间内的密度看起来更加均匀。

校准后的 p 值与均匀分布基准高度吻合，这表明校准过程成功纠正了与骨架变化相关的分布差异。这说明，即使测试集的化学空间不同于训练分布，该方法在测试集上仍可能保持良好的校准效果。在实际应用中，若能部分了解测试集的化学空间，便可进行简单的合理性检查：如果相应分子骨架上的 p 值看起来接近均匀分布，就可以确信该方法在实际测试化合物上也会表现可靠。

H.3 平衡性检查

最后，为了诊断密度校正的影响，我们按照第 3.3 节所述，比较重新加权前后校准集与测试集的特征均值。图20 和21展示了每个特征维度的均值和标准差，其中校准后的测试特征按似然比加权 $p_{\text{cal}}(x)/p_{\text{test}}(x)$。

对于 QED（图20），校准前的测试特征与校准分布存在显著差异，尤其是在第一维。应用密度校正后，加权后的测试特征在所有维度上都与校准均值更加吻合。特征向量均值之间的余弦距离从 0.0623 降至 0.0105，表明分布匹配效果显著改善。

DRD2 也呈现出类似趋势（图21）：重新加权后，多个维度上的较大差异显著减小。余弦距离从0.0491 降至 0.0075，证实校正后的测试分布与校准分布更加接近。

这些平衡性检查提供了一种实用的诊断方法：给定一个无标签测试集，可以评估密度校正是否使其特征分布更接近校准集，这表明校准程序很可能仍然有效。

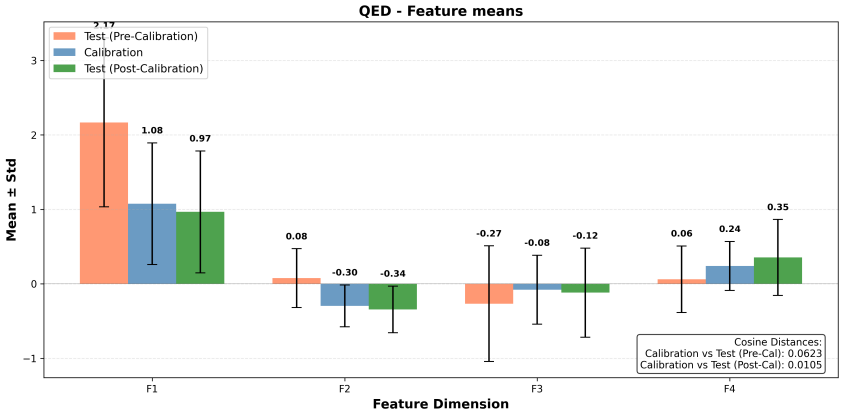


图 20： $N = 5$ 处使用 Hgraph 模型得到的 QED 特征均值。密度校正使测试特征更接近校准分布。

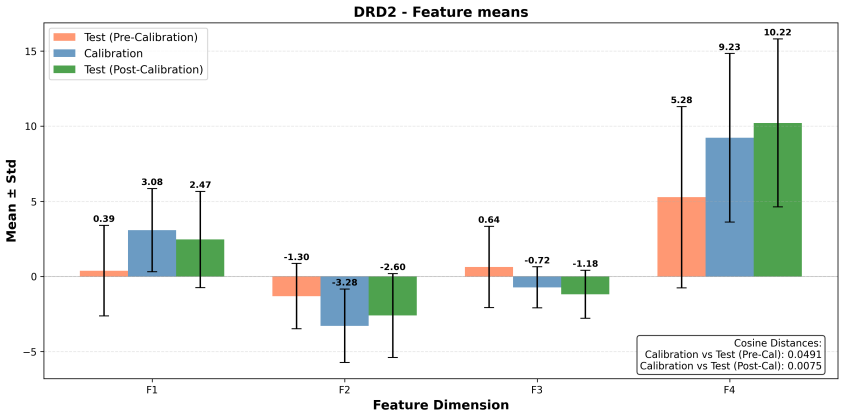


图 21： $N = 5$ 处使用 Hgraph 模型得到的 DRD2 特征均值。密度校正缩小了校准分布与测试分布之间的差异。