一种半监督　　用于活性悬崖估计的分子学习框架　　n

Fang Wu斯坦福大学
fa
ngwu97@stanford.edu

arXiv:2601.04507v1 [cs.CE] 8 Jan 2026

摘要

机器学习（ML）能够快速、准确地进行分子性质预测，这对药物发现和材料设计具有重要意义。其成功的核心在于相似性原则，即假设相似分子具有相近的性质。然而，活性悬崖对这一原则构成挑战，其存在会导致现有机器学习算法的性能急剧下降，尤其是基于图的方法。为在低数据场景下克服这一障碍，我们提出了一种新颖的半监督学习（SSL）方法，称为 SemiMol。该方法利用大量未标注数据上的预测结果作为后续训练的伪信号。具体而言，我们引入一个额外的指导模型，用于评估代理标签的准确性和可信度，因为现有的伪标签方法需要概率输出以体现模型的置信度，因此无法应用于回归任务。此外，我们设计了一种自适应课程学习算法，以可控的速度逐步引导目标模型学习困难样本。在 30 个活性悬崖数据集上开展的大量实验表明，SemiMol 显著增强了基于图的机器学习架构，且性能优于最先进的预训练方法和半监督学习基线。

1 引言

相似结构相似性质原理（SSP） [Bender 和 Glen, 2004] 一直是药物化学中的核心前提，即相似分子往往表现出相似的性质，这些性质可以是物理性质（例如沸点），也可以是生物性质（例如活性）。长期经验为这一基本论断提供了依据，例如，苯乙胺类很可能在中枢神经系统中具有活性，而 β-内酰胺则常具有抗菌活性 [Martin 等， 2002]。计算化学家也利用这一前提分析化合物库的分子多样性，并从中筛选化合物用于高通量筛选（HTS） [Johnson 等， 1989]。

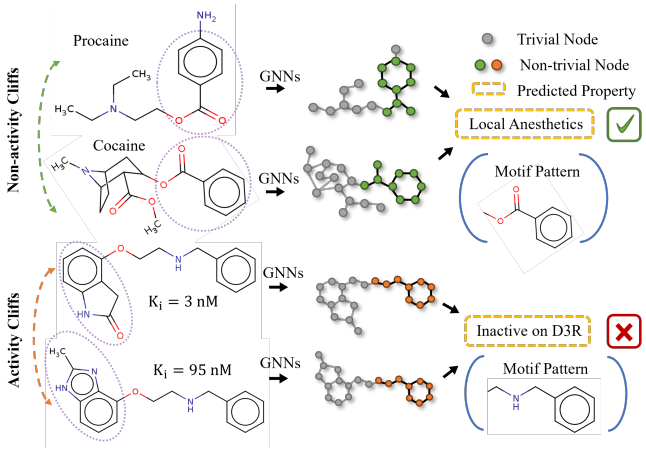

图1：对于非活性悬崖，图神经网络成功发现了 COC(=O)c1ccccc1 的基序，并将具有这种模式的分子归为局部麻醉药。对于活性悬崖，图神经网络受到 COCCNCc1ccccc1 这一共享基序模式的误导，错误地预测相关分子没有活性。

化学相似性（或分子相似性）这一概念，为机器学习（ML）模型在发现和设计具有预期功能的新药方面的广泛应用提供了理论依据。值得注意的是，机器学习算法通常通过识别具有统计学重要性的模式，并在高维特征空间中为邻近样本赋予相近的数值来发挥作用 [Mitchell 和 Mitchell， 1997]。它们学习训练集中反复出现的基序，将携带相同基序的分子聚类，并对其作出相似的预测 [Wu 等，2023b]。因此，相似结构相似性质原理为机器学习奠定了坚实的科学基础，并与其内在运行逻辑高度契合。

尽管 ML 在预测分子性质方面取得了巨大成功，但这一原则存在一个特殊例外，它为理解潜在的构效（或结构-性质）关系提供了重要见解 [Stumpfe *et al.*， 2020]。这一例外由活性悬崖构成 [Maggiora，2006], 即一对结构相似但生物活性存在显著差异的分子。这些离群值可能对机器学习模型产生不利影响，误导模型严重错误地预测某些分子的活性，即使模型总体上具有较高的预测能力

（见图1）。尽管已有大量研究致力于定义活性悬崖并探究其影响 [van Tilborg *et al.*， 2022], 但如何缓解这一阻碍 ML 发展的难题仍有待探索。

尤其是，数据的数量和质量是深度学习（DL）算法能否成功的关键因素之一，但深度学习方法在活性悬崖估计上的评估始终处于低数据场景。同时，分子建模领域的图研究社区一直试图复现自然语言处理（NLP）中自监督预训练取得的成功，并声称通过在大规模数据集上进行预训练带来了显著提升 [Rong 等人，2020；Zhou 等人， 2022]。因此，我们不禁要问：“自监督学习能为活性悬崖估计带来多大益处？”对此，我们观察到，在许多情况下，预训练带来的收益微乎其微。我们的第一个重要发现是，自监督图预训练并非总是比非预训练方法具有统计学上的显著优势，其平均提升比例仅为 6.35%。其次，尽管近期有若干研究在预训练阶段利用了分子的三维空间信息 [Fang等人， 2022], 融入三维几何并未在识别分子活性悬崖方面展现出显著优势。第三，骨干图神经网络的设计对活性悬崖估计准确率的影响，大于预训练机制的选择。

这一现象促使我们探索一种更高效的方案，以利用丰富的未标记分子，并缓解活性悬崖识别中的数据稀缺问题。为此，我们提出了一种新颖且有效的半监督学习（SSL）机制，称为 SemiMol。该机制利用未标注数据上的预测结果作为后续训练的监督信号。然而，由于带标签样本和无标签样本通常来自不同的数据分布，且深度学习模型往往难以检测这种差异，因此伪标签可能并不可靠。此外，现有的伪标签方法需要概率输出以体现模型的置信度，因而无法应用于回归任务 [Rizve 等人，2021；Wu 等人，2023a]。为克服这一障碍，我们引入了一个额外的指导模型，用于评估代理标签的准确性和可信度。随后，这些置信度分数会指导目标分子模型以渐进的方式有选择地关注不同的数据点。在 SemiMol 中，我们不采用固定阈值，也不通过百分位数分数手动调整阈值 [Cascante-Bonilla 等人， 2021], 而是提出了一种用于课程学习的自适应准则, 以确定应将哪些伪标记样本纳入训练集。借助这些精心设计、可控的学习节奏，目标分子模型能够可靠地利用无标签数据，同时避免过度依赖带标签数据，也不会受到部分伪标签严重不准确的影响。

我们在 MoleculeACE [vanTilborg等人，2022] 上进行了大量实验和可视化，以验证该方法的有效性。结果表明，我们的 SemiMol 显著提升了基于图的深度学习模型在低数据场景下的表征能力，并且优于最先进的预训练图神经网络、基于描述符的机器学习模型、基于序列的语言模型，以及在全部 30 个活性悬崖数据集上表现出色的半监督学习方法。这些结果有力地表明，SemiMol 能够准确捕捉充分的化学和生物信息，从而识别生物活性中的活性悬崖。我们的工作融入了生物信息学图学习领域更广泛的发展趋势，旨在系统地探索低数据场景下分子性质预测的最佳解决方案。

2 方法

2.1 预备知识与背景

任务定义。假设存在一组可获取的分子数据 $\mathcal{D}^{\star} = \{(x_i^{\star}, y_i^{\star})\}_{i=1}^{N}$，其中包含带标签数据点 $x_i^{\star}$。这些数据可以采用不同格式，例如一维序列、二维图和三维结构； $y_i^{\star}$ 可以是离散属性（例如毒性和药物反应），也可以是连续属性（例如水溶性和自由能）。为与活性悬崖估计问题的设定保持一致，这里仅考虑连续属性。不过，我们的方法很容易扩展到二元或多标签情形。目标分子模型 $f: \mathcal{X} \rightarrow \mathcal{Y}$ 旨在捕捉分子及其属性之间的映射关系。它可以采用任意类型的架构，例如 Transformer、图神经网络或几何神经网络。我们还有另一组未见数据 $\mathcal{D}^{test}$ 用于评估模型的性能。通常， $\mathcal{D}^{\star}$ 会被划分为训练集和验证集，分别记作 $\mathcal{D}^{train} = \left\{\left(x_i^{train}, y_i^{train}\right)\right\}_{i=1}^{N_1}$ 和 $\mathcal{D}^{val} = \left\{\left(x_i^{val}, y_i^{val}\right)\right\}_{i=1}^{N_2}$。此外，还有一组无标签数据点 $\mathcal{D}^{*} = \{x_i^{*}\}_{i=1}^{M}$，其中无标签数据 M 的数量远大于带标签数据 N 的数量（例如，大 100 倍或 10000 倍）。

半监督学习。 半监督学习（SSL）在解决对大量标注数据集的需求方面日益受到关注。现有的半监督算法大致可分为三类：一致性正则化、代理标签方法和生成模型。一致性正则化基于一个简单的概念：神经网络内部的随机性（例如使用丢弃）或数据增强变换，不应改变模型对同一输入所作的预测，并且会施加辅助损失。这一研究方向包括 π模型 [Laine 和 Aila、 2016], 时间集成 [Laine 和 Aila、 2016], 以及平均教师 [Tarvainen 和 Valpola。 2017]代理标签方法将无标签数据上的代理标签视为目标，分为两类：自训练 [Yarowsky， 1995], 即由模型本身生成代理标签；以及多视图学习 [Zhao等， 2017], 即由在数据不同视图上训练的模型生成代理标签。生成模型依靠变分自编码器（VAE） [Ehsan Abbasnejad 等和生成对抗网络（GAN） [Odena， 2016] 更准确地捕捉联合分布 $p(y|x)$。

2.2 SemiMol 的研究动机

基于置信度的代理标注。在大多数情况下，已标记数据 $\mathcal{D}^{\star}$ 和未标记数据 $\mathcal{D}^{*}$ 遵循显著不同的数据分布，即，$\mathrm{P}(x_i^{\star}, y_i^{\star}) \neq \mathrm{P}(x_i^{*}, y_i^{*})$。然而——

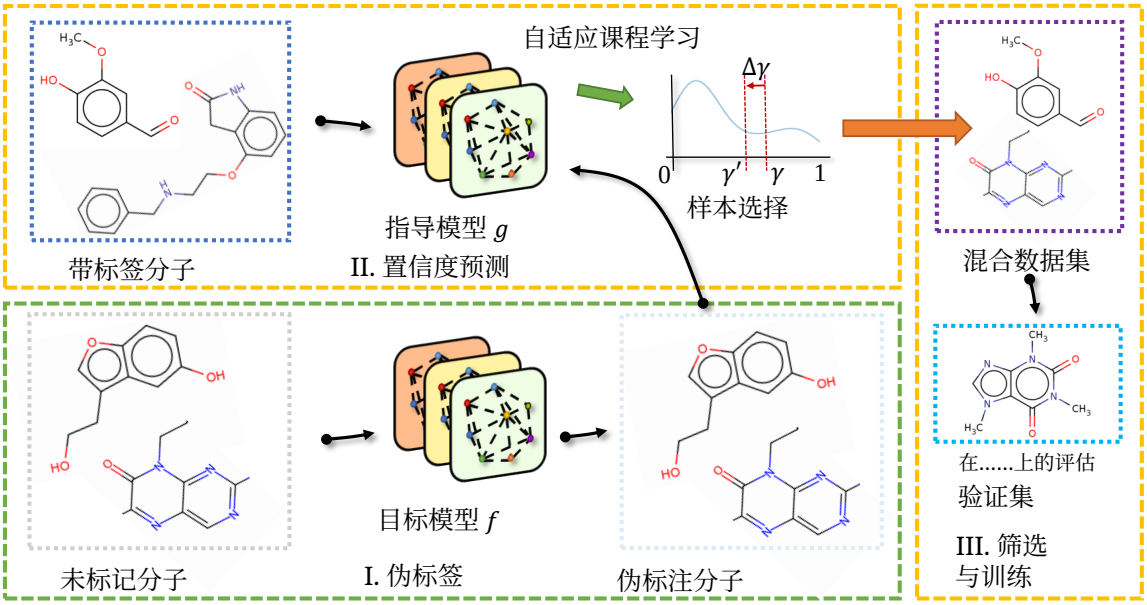

图 2：SemiMol 示意图。目标模型首先为未标注分子数据生成预测结果。随后，指导模型分析这些代理标签的置信度。之后，采用自适应课程学习方案，逐步引导目标模型处理困难样本。通过反复迭代这些过程，目标模型能够最大限度地从大量未标注分子中提取任务特定信息。

另一方面，许多现有的半监督算法 [Yarowsky，1995；Laine 和 Aila，2016；Sohn 等，2020；Xie等，2020] 假设 $\mathcal{D}^*$ 和 $\mathcal{D}^*$之间不存在分布偏移。它们很可能只是将已标记数据 $\mathcal{D}^*$中的一致信息强化到未标记样本 $\mathcal{D}^*$，而不是从 $\mathcal{D}^*$中挖掘辅助信息，代理标注 [Lee 等人，2013]也不例外。另一方面，尽管代理标注具有通用性且与模态无关，但与近期的半监督算法 [Rizve 等，2021]相比，其性能相对较差。这是因为在训练过程中，一些伪标签$\hat{}y_i^*$可能严重错误，这是由于经典深度学习模型 [Hendrycks等，2021]的泛化能力较差。如果我们直接使用先前训练的模型预测出的伪标签进行后续训练，前面训练轮数中积累的一致性偏差信息可能会提高对错误预测的置信度，最终导致错误不断累积的恶性循环 [Arazo 等，2020]。如果已标记数据 $\mathcal{D}^*$ 因不可避免的实验误差而包含噪声，情况甚至会更糟。因此，有必要了解伪标注的质量和可靠性，并智能地选择其中的一个子集，以减少训练中存在的噪声 [Rizve 等，2021]。

回归任务的置信度。 尽管置信度对于伪标签选择至关重要，基于置信度的选择也能降低伪标签错误率，但神经网络校准不佳，使得这一方案仍不够完善。具体而言，校准不佳的网络也可能对错误预测给出较高的置信度分数（即，$\hat{}y_i^* \to 0$or $\hat{}y_i^* \to 1$）[Rizve等，2021]。更重要的是，现有的大多数方法，例如不确定性感知伪标签选择（UPS）——（UPS）[Rizve 等，2021] 都依赖目标模型输出$\hat{}y_i$ 作为置信度指标。这些方法倾向于通过 $y_i^* = \mathbb{1}[\hat{}y_i^* \geq \gamma_1]$ 或 $y_i^* = \mathbb{1}[\hat{}y_i^* \leq \gamma_2]$生成硬标签，其中 γ_1 和 γ_2 是预先设定的置信度阈值。然而，如果 $\mathcal{Y}$ 属于连续标签空间，这种选择机制便不再适用，因为网络不再输出类别概率。换言之，对于回归任务，$\hat{}y_i^*$不包含任何置信度信息。值得注意的是，大量生化问题都属于回归问题，包括分子性质预测 [Wu 等，2018]，3D 结构预测 [Jumper等，以及结合亲和力预测 [Wang 等，2005]。这给基于概率输出的代理标注算法带来了不可回避但至关重要的挑战 [Rizve 等，2021]。

基于这两个动机，我们不再依赖模型 $\hat{}y_i^*$ 的输出判断代理标签的可靠性，而是为目标分子模型 f配备一个额外的指导模型 g。它充当评判器，预测标签可观测性，即判断标签是真还是假。引入 g 后，置信度预测与性质预测得以解耦，并能大幅减少伪标签生成过程引入的噪声。

2.3 SemiMol 框架

指导模型引导的半监督学习。 为克服过度依赖带来的障碍并减少代理标签中的噪声，我们提出了一种指导模型引导的半监督框架，称为 SemiMol（见图 2），以智能地筛选出噪声较少的部分伪标签。SemiMol 由两个组件组成：一个目标分子模型 f 和一个指导模型 g。前者 f 负责预测属性 $\hat{}y$。相比之下，后者 g 则负责衡量

监督信号的可靠性 p，可将其解释为伪标签的置信概率。我们将 SemiMol 的完整流程划分为两个阶段。首先，我们保留伪标签 $\{\hat{y}_i^*\}_{i=1}^M$。生成代理标签的方法有多种，例如通过邻域图进行标签传播 [Iscen 等,2019]。这里我们采用 [Lee 等人的方法， 2013] 要求分子模型 f 直接为未标记数据集中的样本 $\mathcal{D}^*$生成标注。接下来，我们构建一个同时包含已标记样本和伪标记样本的新数据集，表示为 $\mathcal{D}' = \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*)\}_{i=1}^M$，并基于这个新数据集训练目标分子模型 f 和指导模型 g。我们反复执行这两项操作，直到 f在验证集 $\mathcal{D}^{val}$上达到最佳性能。

具体而言，指导模型 $g: (\mathcal{X} \times \mathcal{Y} \times \mathcal{H}_f) \to \mathcal{P}$ 预测给定标签 y_i' 属于真实标签集 $\{y_i^\star\}_{i=1}^N$ 还是伪标签集 $\{y_i^*\}_{i=1}^M$ 的置信度度量 p_i $(0 \le p \le 1)$。它处理三项内容：带标签的数据样本 $(x_i', y_i') \in \mathcal{D}'$ 以及一个附加损失项 $\mathcal{H}_f(f(x_i')\ y_i')$，其中 $\mathcal{H}_f$对于回归任务通常选用均方根误差（RMSE）损失或平均绝对误差（MAE）损失，对于分类问题则选用交叉熵（CE）损失。在这里，我们将 $\mathcal{H}_f(\cdot)$视为 g 输入的一部分，为主要的分子性质预测任务提供更多信息。最后，指导模型 g 通过二元交叉熵损失（BCE）进行监督，具体如下

$$\begin{aligned}\mathcal{L}_g\left(\mathcal{D}', \{\hat{y}_i'\}_{i=1}^{N+M}\right) &= \sum_{(x_i', y_i') \in \mathcal{D}'} \mathrm{BCE}(p_i, c_i) \\ &= \sum_{(x_i', y_i') \in \mathcal{D}'} \mathrm{BCE}\Big(g\big(x_i', y_i', \mathcal{H}_f\left(\hat{y}_i', y_i'\right)\big), c_i\Big),\end{aligned} \tag{1}$$

其中， $c_i \in [0, 1]$ 是一个整数，表示可观测性掩码。它指示 y_i 是否为伪标记（$c_i = 0$）样本。

基于课程学习的伪标签生成。与此同时，目标分子模型 f 从指导模型 g 接收判别信息$\{p_i\}_{i=1}^{N+M}$，并利用这些信息高效地选择伪标记样本，以反向传播其梯度。值得注意的是，用于确定每轮训练中纳入哪些未标记样本 $\mathcal{D}^*$ 的标准，对伪标签生成的成功至关重要。以往研究探索了各种不确定性指标。例如，一些方法选择置信度最高的实例 [Zhu， 2005]，另一些方法则检索特征空间中最近的样本 [Shi 等，2018]。随后，一些研究采用固定阈值、手动调整阈值或使用百分位数分数等技术 [Cascante-Bonilla 等， 2021]。在此，我们提出一种自适应课程学习算法，使伪标记样本能够进入或退出新的训练集。

具体而言，我们首先初始化百分比阈值 γ，并使用它生成硬标签，构成混合集合 $\mathcal{D}'' = \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*) | p_i^* \ge \gamma\}$，以训练目标模型 f。

随着训练的进行，我们会根据模型在验证集上的表现自适应地调整置信度截断值 γ。如果评估认为模型的能力更强

算法 1 SemiMol算法

1: 输入：目标模型 f**、指导模型** g**、已标记数据** $\mathcal{D}^\star$**、无标签数据** $\mathcal{D}^*$**、伪标签更新频率** k**、阈值百分比** γ**、步进阈值百分比** $\Delta\gamma$**2: 初始化并预训练目标模型** f_0 **和指导模型** g_0**3: 对于训练轮数** $n = 0$**，1，2，...，执行 4: 如果** n **模** $k == 0$**，则 5:** $\hat{y}_i^* \leftarrow f(x_i^*)$ $\forall x_i^* \in \mathcal{D}^*$ ▷ **每隔** k **个训练轮数迭代分配伪标签 6: 结束如果**7 $\mathcal{D}' \leftarrow \mathcal{D}^\star \cup \{ x_i^* \ y_i^* \}_{i=1}^M$**: (，) :** 8 $y_i' \leftarrow f(x_i')$ $\forall x_i' \in \mathcal{D}'$**9:** $p_i \leftarrow g(x_i', y_i', \mathcal{H}_f(.))$ $\forall (x_i', y_i') \in \mathcal{D}'$ ▷ **预测置信度分数 10:** $\mathcal{L}_g\left(\mathcal{D}', \{\hat{y}_i'\}_{i=1}^{N+M}\right) \leftarrow$ **公式 1 11:** $\mathcal{D}'' \leftarrow \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*) | p_i^* \ge \gamma\}$ ▷ **构建训练集 12:** $\mathcal{L}_f(\mathcal{D}'') \leftarrow$**公式 2 13: 根据** $\mathcal{L}_g(.)$ **和**$\mathcal{L}_f(.)$ **更新** f **和** g **的参数 14: 在验证集上评估** f **并获得性能**s**15: 如果** $n > 0$ **和** $s < s'$**，则 16:** $\gamma \leftarrow \gamma - \Delta\gamma$ ▷ **自动降低阈值 17: 结束如果 18:** $s' \leftarrow s$ ▷ **记录上一训练轮数的验证指标 19: 结束循环**

（例如， 评估指标均方根误差变小），我们提出将阈值降低为 $\gamma = \gamma - \Delta\gamma$，其中 $\Delta\gamma$ 为步进百分比。这会得到一个更大的混合集合 $\mathcal{D}''$，目标模型的损失 f 可写为:

$$\mathcal{L}_f(\mathcal{D}'') = \sum_{(x_i'', y_i'') \in \mathcal{D}''} \mathcal{H}_f(f(x_i''), y_i''), \tag{2}$$

我们的自适应课程学习策略能够抑制概念漂移或确认偏差，因为它可以防止训练早期阶段由训练不足的网络预测的错误标注不断累积 [Cascante- Bonilla 等， 2021]。我们的 SemiMol 伪代码如算法 1 所示。

实现 SemiMol 的建议。在执行 SemiMol 之前，通常先通过在标注数据集 $\mathcal{D}^\star$ 上进行常规监督学习，获得训练良好的分子靶标模型 f_0，然后根据该模型 f_0生成的伪标签初始化指导模型 g_0。经验证，这种方法能够提高训练的稳定性和鲁棒性。此外，我们选择每隔k 个训练轮数分配一次伪标签，其中 k 的合理设置对 SemiMol 的成功至关重要。如果伪标签更新过于频繁，训练过程往往会不稳定。

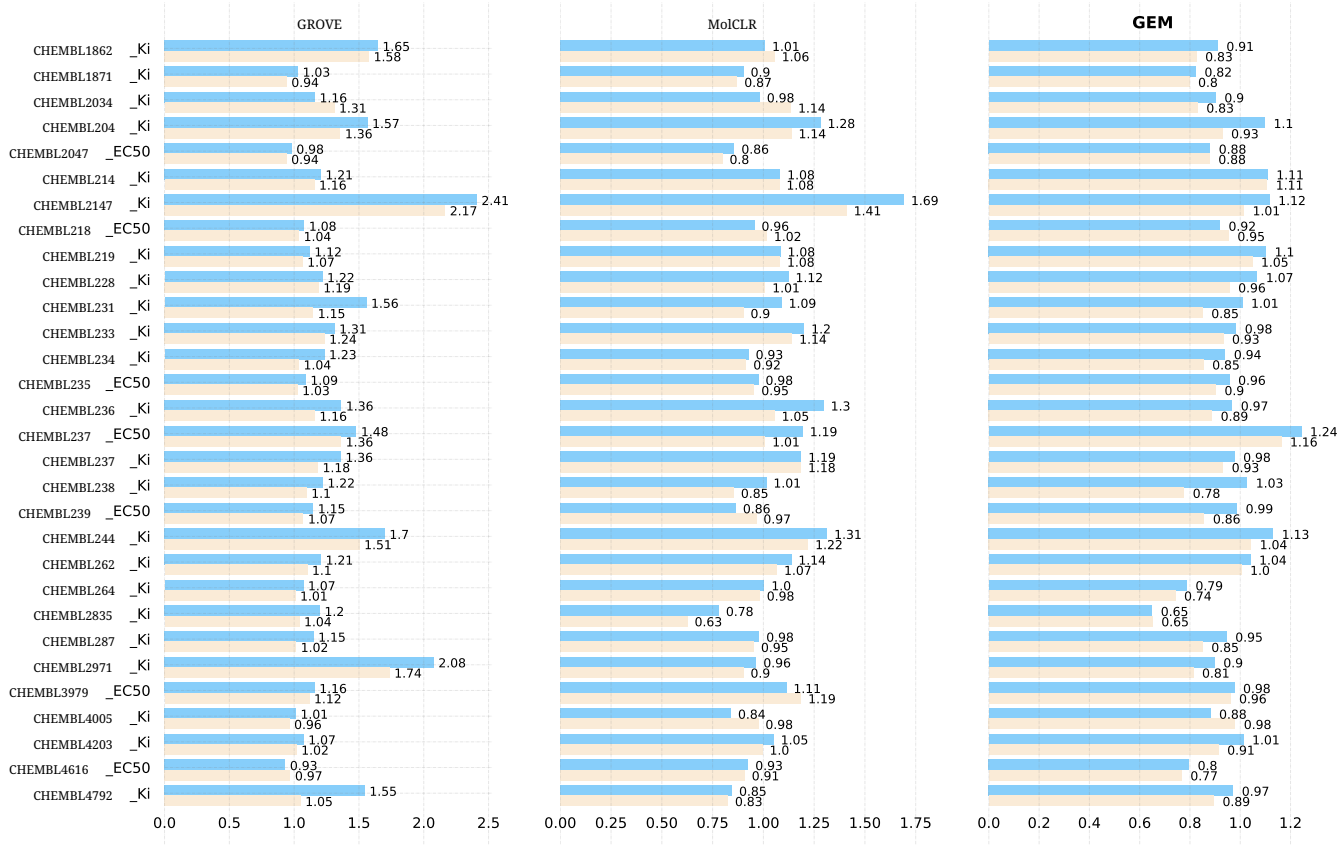

图 3：不同 图神经网络在 30 个活性悬崖数据集上进行预训练与否的对比 ．浅蓝色对应于未经过预训练时的性能，而古董白色代表经过预训练后的性能。

3 实验

3.1 数据与实验设置

本节中的所有评估均在 MoleculeACE（活性悬崖估计）数据集上进行。MoleculeACE 是一个开放访问的基准测试平台，可通过 GitHub https://github.com/molML/MoleculeACE 获取。该平台包含针对 30 个大分子靶标的 35,000 多个分子，每个靶标对应一个数据集。尤其值得注意的是，MoleculeCAE 的 30 个数据集中有 12 个的训练集包含不超过 1K 个分子，属于典型的小数据场景。此外，我们采用图同构网络 (GIN) [Xu 等人， 2018] 以及图多集 Transformer (GMT) [Baek 等人， 2021] 来聚合节点特征，并将其作为骨干网络。更多实验细节和数据集统计信息详见附录。

3.2 图神经网络预训练能否缓解活性悬崖问题?

数据的数量和质量始终被视为深度学习算法成功的最关键因素之一。然而，在低数据场景下评估深度学习方法对活性悬崖的表现仍然存在困难。值得注意的是，过去几年见证了自监督学习在多个领域的盛行 [Devlin 等，2018；Brown 等，2020；Dosovitskiy 等， 2020], 分子建模领域的图研究社区一直在尝试复现其在自然语言处理领域取得的成功。各种研究试图挖掘大规模未标记分子中的信息 [Rong等，2020；Fang 等，2022；Wang 等，2022；Zhou等， 2022], 并声称取得了显著的性能提升 [Rong 等，2020；You 等，2020；Zhu 等，2021；Zhou 等， 2022]。然而，这些预训练框架仅在不含活性悬崖分子的数据集上进行了验证，其是否同样能够有效缓解活性悬崖估计中的数据稀缺问题，仍然未知。

在此，我们研究了几种最先进的预训练图神经网络，包括 GROVE [Rong 等、 2020], MolCLR [Wang 等、2022], 以及 GEM [Fang等，并考察它们在活性悬崖数据集上的效果。具体而言，GROVE 采用 Transformer 式架构 [Vaswani 等，显著增强了分子表征学习方案的表征能力和应用范围。MolCLR 在节点和图两个层级对分子图进行数据增强，并采用对比学习策略，使图神经网络能够推广到更广阔的化学空间。GEM 提出了一种基于几何的图神经网络，并采用专门的几何层级自监督学习技术来捕捉分子几何知识。图 3 展示了

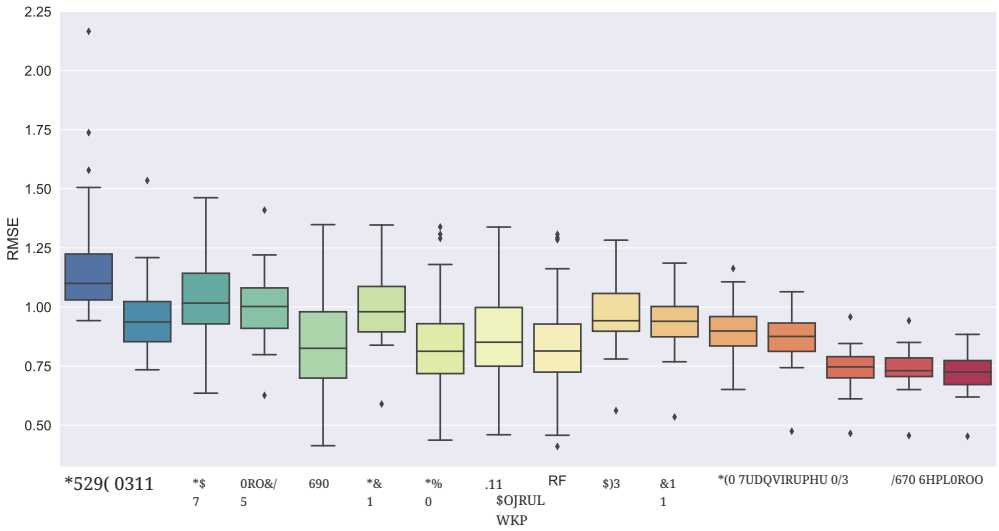

图 4：各机器学习模型在 30 个活性悬崖数据集上的整体均方根误差表现 ts.

实证结果。可以发现，在所选的三个预训练框架中，预训练通常都有助于解决活性悬崖估计问题。在全部 30 个数据集上，GROVE、MolCLR 和 GEM 的平均表现分别提升了 8.57%、3.92% 和 6.54%。然而，这些提升并非总是显著，有时甚至微乎其微或呈负向。例如，GROVE 在两个数据集上经过预训练后表现更差，MolCLR 在六个数据集上经过预训练后竞争力下降，而 GEM 在三个数据集上受到预训练的负面影响。

这一现象与近期一项研究 [Sun，2022] 中的结论一致：在许多设置下，自监督图预训练的统计表现并不总是优于非预训练方法。此外，局部和全局分子三维结构可能并非解决我们问题的关键，因为 GEM 已将它们纳入自监督学习，但收益甚微。最后，我们比较了不同图神经网络在自监督学习前后的表现。GROVE 预训练前后的平均均方根误差分别为 1.3107 和 1.1871。MolCLR 预训练前后的平均均方根误差分别为 1.0556 和 1.0071。GEM 预训练前后的平均均方根误差分别为 0.9699 和 0.9035。图神经网络的架构设计是提升我们所研究的活性悬崖问题表现的一个更基础的因素。未经预训练的 GEM 轻松超越了经过预训练的 MolCLR 和 GROVE。

为什么图神经网络预训练没有益处？ 由于若干潜在原因，我们未能在 MoleculeNet 等传统分子数据集中复现预训练带来的收益。首先，大多数无监督预训练的目标标签与我们的活性悬崖估计任务并未恰当对齐，甚至相互矛盾。例如，GROVE [Rong 等人， 2020] 要求图神经网络判断一个分子是否含有某种基序。然而，这 85 种预定义基序可能会误导对活性悬崖分子性质的估计。其次，正如 [Sun 所指出的， 2022], 当缺乏高质量的人工设计特征时，预训练通常才会发挥作用。在我们的实验中，我们为图神经网络注入了丰富的特征，包括额外的节点特征（如氢受体匹配、酸性匹配）以及键特征（如环信息）。最后，一些预训练方法（例如掩码节点标签预测）可能过于简单，无法为我们的活性悬崖估计任务迁移足够的知识。

3.3 SemiMol 能否改善活性悬崖估计？

整体表现。 我们在 MoleculeACE 上将 SemiMol 与多种机器学习和深度学习算法进行比较，并分别在图 4 和附录中展示活性悬崖分子及所有分子的结果。可以看出，在 30 个活性悬崖数据集上，SemiMol 的均方根误差最低，优于其他基线方法。更重要的是，我们发现 SemiMol 带来了平均 26.53% 的提升，显著高于最先进的预训练方法所带来的收益。这一现象表明，由半监督学习赋能的图神经网络可能成为比现有算法更有影响力的工具，能够克服活性悬崖这一障碍，并加速新药发现进程。

在所有方法中，非预训练图神经网络（包括 MPNN [Gilmer等、 2017], GAT [Veli˘ckovi´c等、 2017],GCN [Kipf 和 Welling，以及 2016], AFP [Baek 等）的表现通常不如其他方法。与此同时，预训练图神经网络的性能差异很大。例如，GEM 的表现远胜 GROVE。然而，正如第 3.2 节所分析的，预训练的益处有限。这种差异更多与骨干图神经网络的能力有关，而非预训练策略。此

数据集	CHEMBL1862 抑制常数	CHEMBL1871抑制常数	CHEMBL2034 抑制常数	CHEMBL204 抑制常数	CHEMBL2047 半数有效浓度
GIN	1.019	0.854	0.951	1.179	0.895
+ π-模型	0.984	0.793	0.902	1.060	0.880
+ 半监督生成对抗网络	0.955	0.781	0.834	0.927	0.854
+ SemiMol	**0.714**	**0.671**	**0.703**	**0.801**	**0.629**

表 1：不同半监督学习算法在 MoleculeACE 的 5 个数据集上的性能。

数据集	CHEMBL214 抑制常数	CHEMBL2147 抑制常数	CHEMBL218 半数有效浓度	CHEMBL219 抑制常数	CHEMBL228 抑制常数
SemiMol					
+ 固定阈值	0.788	0.704	0.781	0.914	0.821
+ 百分位数分数	0.739	0.683	0.744	0.850	0.779
+ 自适应步调	**0.726**	**0.672**	**0.723**	**0.835**	**0.758**

表 2：在 MoleculeACE 的 5 个数据集上与不同课程学习方法的比较。

外，基于序列的模型，例如长短期记忆网络 [Hochreiter 和 Schmidhuber，以及 Transformer [Vaswani 等，凭借出色的性能而尤为突出。实证结果表明，如果微调阶段没有提供额外的无标签数据，与基于图和基于描述符的模型相比，它们是解决活性悬崖问题的更佳选择。

与其他半监督方法的比较。多种半监督学习算法已被提出，以针对特定任务的方式利用丰富的未标注资源。总体而言，它们可以分为三类：一致性正则化、代理标签方法和生成模型。我们从每一类中选取一种基线方法，并研究其在分子性质预测方面的改进效果。值得注意的是，随机节点丢弃和边扰动在生化意义上并不合理，而数据增强可能会改变分子图的语义 [Sun 等， 2021]。因此，我们采用 π-模型 [Laine 和 Aila， 2016] 进行比较。该模型对网络而非输入图应用不同的随机变换（即丢弃）。此外，我们选择 Semi-GAN [Odena， 2016], , 该方法引入判别器来判断输入是否带有标签。如前所述，UPS [Rizve等， 2021] 利用预测不确定性来指导伪标签选择过程，因此不适用于我们的场景。我们从 MoleculeACE 中选取五个数据集进行比较，并采用 GIN 作为骨干网络架构。如表 1 所示，SemiMol 在总体均方根误差方面优于所有其他半监督方法。此外，SemiMol 克服了 UPS 的缺点，并可用于回归任务。上述所有证据表明，与现有半监督学习机制相比，SemiMol 是一种具有更强泛化能力和更广泛应用范围的先进代理标注算法。

消融研究。 我们还进行了额外实验，以说明自适应课程学习机制的有效性，并在表 2 中报告了结果。可以发现，与固定阈值和基于百分位分数的课程学习范式相比，我们的自适应调度在整体均方根误差方面表现更优 [Cascante- Bonilla 等， 2021]。

分类任务的性能。值得注意的是，MoleculeACE 中的数据集全是回归任务，而我们提出的 SemiMol 相较于其他代理标注算法的一项主要优势是——SemiMol 能够处理连续性质空间。尽管如此，SemiMol 也可用于分类任务。为更好地展示其优越性，我们采用分类数据集 CYP3A4 [Rao 等人， 2022], 并评估其性能。CYP3A4 包含通过实验测得的细胞色素 P450 3A4 抑制作用中的性质悬崖， [其中包括 3,626 种活性抑制剂/底物和 5,496 种非活性化合物。我们在表 3 中报告了不同方法的相应性能，其中最佳结果以粗体标出，次佳结果加下划线。可以看出，SemiMol 的表现优于所有以消息传递神经网络为骨干网络的半监督基线方法。

模型	CYP3A4 ROC-AUC	CYP3A4 Cliff ROC-AUC
GCN	0.7659	0.6801
GAT	0.7734	0.6828
MPNN	0.7811	0.6867
π-模型	0.7842	0.6903
半监督生成对抗网络	0.8015	0.7122
UPS	<u>0.8438</u>	<u>0.7546</u>
SemiMol	**0.8568**	**0.7703**

表 3：CYP3A4 分类数据集上的性能。

4 结论

过去十年间，机器学习技术在药物发现中的应用日益广泛，但活性悬崖（AC）这一课题却鲜少受到科学界关注。本文旨在找出在存在活性悬崖的情况下提升模型预测能力的最佳实践，并提出一种名为 SemiMol 的有效半监督学习（SSL）方法。该方法引入指导模型来评估代理标签的置信度，并利用自适应课程训练范式，更高效地引导目标模型处理困难样本。我们证明，即使在低数据场景下，基于图的深度学习模型在药物发现中的表现也不逊于更简单的机器学习算法。我们期待更高效、更适用的半监督学习方法能够推动活性悬崖识别领域的发展。

参考文献

[Arazo等， 2020] Eric Arazo、Diego Ortego、Paul Albert、Noel E O’Connor 和 Kevin McGuinness。《深度半监督学习中的伪标签与确认偏差》。载于*2020* 年国际联合神经网络大会（*IJCNN*），第 1–8 页。IEEE，2020 年。

[Baek等， 2021] Jinheon Baek、Minki Kang 和 Sung Ju Hwang。利用图多重集合池化实现图表示的准确学习。*arXiv* 预印本 *arXiv:2102.11533*，2021 年。

[Bender 和 Glen， 2004] Andreas Bender 和 Robert C Glen。分子相似性：分子信息学中的关键技术。《有机与生物分子化学》，2(22)：3204–3218，2004 年。[Brown 等， 2020] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习器。神经信息处理系统进展，33：1877–1901，2020 年。[Cascante-Bonilla等， 2021] Paola Cascante-Bonilla、Fuwen Tan、Yanjun Qi 和 Vicente Ordonez。课程式标注：重新审视半监督学习中的伪标签。在*AAAI*人工智能会议论文集中，第 35 卷，第 6912–6920 页，2021 年。[Devlin等， 2018] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018 年。[Dosovitskiy 等， 2020] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一幅图像胜过 16×16 个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*，2020 年。[Ehsan Abbasnejad 等， 2017] M Ehsan Abbasnejad、Anthony Dick 和 Anton van den Hengel。用于半监督学习的无限变分自编码器。在*IEEE*计算机视觉与模式识别会议论文集中，第 5888–5897 页，2017年。[Fang等， 2022] Xiaomin Fang、Lihang Liu、Jieqiong Lei、Donglong He、Shanzhuo Zhang、Jingbo Zhou、Fan Wang、Hua Wu 和 Haifeng Wang。增强几何的分子表示学习用于性质预测。《自然-机器智能》，4(2)：127–134，2022 年。[Fey 和 Lenssen， 2019] Matthias Fey 和 Jan Eric Lenssen。利用 PyTorch Geometric 进行快速图表示学习。*arXiv* 预印本 *arXiv:1903.02428*，2019 年。[Gaulton等， 2012] 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 年。

[Gilmer 等人， 2017] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals 和 George E Dahl。《用于量子化学的神经消息传递》。载于机器学习国际会议论文集，第 1263–1272 页。PMLR，2017年。[Hendrycks 等人， 2021] Dan Hendrycks、Steven Basart、Norman Mu、Saurav Kadavath、Frank Wang、Evan Dorundo、Rahul Desai、Tyler Zhu、Samyak Parajuli、Mike Guo 等人。《鲁棒性的诸多面貌：对分布外泛化的批判性分析》。载于 IEEE/CVF国际计算机视觉会议论文集，第 8340–8349 页，2021年。[Hochreiter and Schmidhuber， 1997] Sepp Hochreiter 和 J¨urgen Schmidhuber。《长短期记忆》。Neural Computation，9(8)：1735–1780，1997年。[Iscen 等人， 2019] Ahmet Iscen、Giorgos Tolias、Yannis Avrithis 和 Ondrej Chum。《用于深度半监督学习的标签传播》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 5070–5079 页，2019年。[Johnson 等人， 1989] M Johnson、M Lajiness 和 G Maggiora。《分子相似性：设计药物筛选程序的基础》。Progress in Clinical and Biological Research，291：167–171，1989年。[Jumper 等人， 2021] John Jumper、Richard Evans、Alexan- der Pritzel、Tim Green、Michael Figurnov、Olaf Ron- neberger、Kathryn Tunyasuvunakool、Russ Bates、Au- ˇgustin Z´ıdek、Anna Potapenko 等人。《利用 AlphaFold 进行高精度蛋白质结构预测》。Nature，596(7873)：583–589，2021年。[Kimber 等人， 2021] Talia B Kimber、Maxime Gagnebin 和 Andrea Volkamer。《MaxSMI：通过 SMILES增强和深度学习结合置信度估计，最大限度提高分子性质预测性能》。Artificial Intelligence in the Life Sciences，1：100014，2021年。[Kipf 和 Welling， 2016] Thomas N Kipf 和 Max Welling。《基于图卷积网络的半监督分类》。arXiv 预印本 arXiv:1609.02907，2016年。[Laine 和 Aila， 2016] Samuli Laine 和 Timo Aila。《用于半监督学习的时间集成》。arXiv 预印本 arXiv:1610.02242，2016年。[Landrum 等人， 2013] Greg Landrum 等人。《RDKit：用于化学信息学、计算化学和预测建模的软件套件》。Greg Landrum，8，2013。[Lee 等人， 2013] Dong-Hyun Lee 等人。《伪标签：一种简单高效的深度神经网络半监督学习方法》。载于表征学习挑战研讨会，ICML，第 3 卷，第 896 页，2013。[Ma 等人， 2019] Yanjun Ma、Dianhai Yu、Tian Wu 和 Haifeng Wang。《PaddlePaddle：源自工业实践的开源深度学习平台》。Frontiers of Data and Domputing，1(1)：105–115，2019年。[Maggiora， 2006] Gerald M Maggiora。《论离群值与活性悬崖：定量构效关系为何常令人失望》，2006年。

[Martin 等人， 2002] Yvonne CMartin、James L Kofron 和 Linda M Traphagen。《结构相似的分子是否具有相似的生物活性？》《药物化学杂志》，45(19)：4350–4358，2002年。

[Mitchell 和 Mitchell， 1997] Tom M Mitchell 和 Tom M Mitchell。机器学习，第 1 卷。麦格劳-希尔，纽约，1997年。

[Odena， 2016] Augustus Odena。利用生成对抗网络进行半监督学习。arXiv 预印本 arXiv:1606.01583，2016。[Pedregosa 等人， 2011] Fabian Pedregosa，Ga¨el Varoquaux，Alexandre Gramfort，Vincent Michel，Bertrand Thirion，Olivier Grisel，Mathieu Blondel，Peter Prettenhofer，Ron Weiss，Vincent Dubourg 等。Scikit-learn：Python 中的机器学习。《机器学习杂志》，12：2825–2830，2011。[Ramsundar 等人， 2019] Bharath Ramsundar，Peter East- man，Patrick Walters，Vijay Pande，Karl Leswing 和 Zhenqin Wu。《生命科学深度学习》。O'Reilly Media，2019。https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837。[Rao 等人， 2022] Jiahua Rao，Shuangjia Zheng，Yutong Lu 和 Yuedong Yang。可解释图神经网络在分子性质预测中的定量评估。Patterns，3(12)：100628，2022。[Rizve 等人， 2021] Mamshad Nayeem Rizve，Kevin Duarte，Yogesh S Rawat 和 Mubarak Shah。为伪标签辩护：一种面向不确定性的半监督学习伪标签选择框架。arXiv 预印本 arXiv:2101.06329，2021。[Rong 等人， 2020] Yu Rong，Yatao Bian，Tingyang Xu，Weiyang Xie，Ying Wei，Wenbing Huang 和 Junzhou Huang。在大规模分子数据上进行自监督图变换器学习。《神经信息处理系统进展》，33：12559–12571，2020。[Shi 等人， 2018] Weiwei Shi，Yihong Gong，Chris Ding，Zhiheng MaXiaoyu Tao 和 Nanning Zheng。利用极小极大特征进行传导式半监督深度学习。载于《欧洲计算机视觉会议》（ECCV）论文集，第 299–315 页，2018。[Shortle， 2009] David Shortle。一条序列加一个突变等于两种折叠。《美国国家科学院院刊》，106(50)：21011–21012，2009。[So hn 等人， 2020] Kihyuk Sohn，David Berthelot，Nicholas Carlini，Zizhao Zhang，Han Zhang，Colin A Raffel，Ekin Dogus Cubuk，Alexey Kurakin 和 Chun-Liang Li。FixMatch：通过一致性和置信度简化半监督学习。《神经信息处理系统进展》，33：596–608，2020。[Sterling 和 Irwin， 2015] Teague Sterling 和 John J I rwin。ZINC 15——面向所有人的配体发现。《化学信息与建模杂志》，55(11)：2324–2337，2015。[Stumpfe 等人， 2020] Dagmar Stumpfe，Huabin Hu 和 J¨urgen Bajorath。探索活性悬崖的进展。《计算机辅助分子设计杂志》，34：929–942，2020。

[Sun 等， 2021] Mengying Sun、Jing Xing、Huijun Wang、Bin Chen 和 Jiayu Zhou。Mocl：通过从分子图中进行知识感知对比学习，获得数据驱动的分子指纹。载于第 *27* 届 *ACM SIGKDD*知识发现与数据挖掘大会论文集，第 3585–3594 页，2021年。

[Sun， 2022] RuoxiSun。图神经网络预训练是否有助于分子表征？*arXiv* 预印本 *arXiv:2207.06010*，2022年。

[Tarvainen 和 Valpola， 2017] Antti Tarvainen 和 Harri Valpola。均值教师是更好的榜样：加权平均的一致性目标提升了半监督深度学习的效果。神经信息处理系统进展，第 30 卷，2017年。

[van Tilborg等, 2022] Derek van Tilborg、Alisa Alenicheva 和 Francesca Grisoni. 揭示分子机器学习在活性悬崖上的局限性。《化学信息与建模杂志》，62(23):5938–5951, 2022.

[Vaswani 等, 2017] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin. 注意力就是你所需要的一切。神经信息处理系统进展, 30, 2017年.

[Veith等, 2009] Henrike Veith、Noel Southall、Ruili Huang、Tim James、Darren Fayne、Natalia Artemenko、Min Shen、James Inglese、Christopher P Austin、David G Lloyd 等. 对化学库中细胞色素 P450 同工酶选择性进行全面表征。《自然-生物技术》,27(11):1050–1055, 2009.

[Veli ˇckovi´c 等， 2017] Petar Veli ˇckovi´c、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。图注意力网络。*arXiv* 预印本 *arXiv:1710.10903*，2017 年。[Wang 等， 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年。[Wang 等， 2022] Yuyang Wang、Jianren Wang、Zhonglin Cao 和 Amir Barati Farimani。通过图神经网络进行分子表征的分子对比学习。《自然·机器智能》，4(3):279–287，2022年。[Wu等， 2018] Zhenqin Wu、Bharath Ramsundar、Evan N Feinberg、Joseph Gomes、Caleb Geniesse、Aneesh S Pappu、Karl Leswing 和 Vijay Pande。MoleculeNet：分子机器学习基准测试。《化学科学》，9(2):513–530，2018年。[Wu等，2023a] Fang Wu、Huiling Qin、Wenhao Gao、Siyuan Li、Connor W Coley、Stan Z Li、Xianyuan Zhan 和 Jinbo Xu。InstructBio：面向生化问题的大规模半监督学习范式。*arXiv* 预印本 *arXiv:2304.03906*，2023年。

[Wu 等，2023b] Fang Wu、Dragomir Radev 和 Stan Z Li。Molformer：用于三维异质分子图的基于基序的 Transformer。*Rn*，1:1，2023年。[Xie等， 2020] Qizhe Xie、Minh-Thang Luong、Eduard Hovy 和 Quoc V Le。通过带噪学生模型进行自训练，提升 ImageNet 分类效果。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10687–10698页，2020年。[Xiong等， 2019] Zhaoping Xiong、Dingyan Wang、Xi- aohong Liu、Feisheng Zhong、Xiaozhe Wan、Xutong Li、Zhaojun Li、Xiaomin Luo、Kaixian Chen、Hualiang Jiang 等。利用图注意力机制推动药物发现分子表征的边界。《药物化学杂志》，63(16):8749–8760，2019年。[Xu 等， 2018] Keyulu Xu、Weihua Hu、J ure Leskovec 和 Stefanie Jegelka。图神经网络的能力有多强？*arXiv* 预印本 *arXiv:1810.00826*，2018年。[Yarowsky， 1995] David Yarowsky。无监督词义消歧可媲美监督方法。载于第*33*届计算语言学协会年会，第189–196页，1995年。[You等， 2020] Yuning You、Tianlong Chen、Zhangyang Wang 和 Yang Shen。自监督何时有助于图卷积网络？载于国际机器学习会议，第10871–10880页。PMLR，2020年。[Zhao等， 2017] Jing Zhao、Xijiong Xie、Xin Xu 和 Shiliang Sun。多视图学习综述：最新进展与新挑战。信息融合，38:43–54，2017年。[Zhou等， 2022] Gengmo Zhou、Zhifeng Gao、Qiankun Ding、Hang Zheng、Hongteng Xu、Zhewei Wei、Linfeng Zhang 和 Guolin Ke。Uni-Mol：通用三维分子表征学习框架。国际学习表征会议，2022年。[Zhu等， 2021] Yanqiao Zhu、Yichen Xu、Feng Yu、Qiang Liu、Shu Wu 和 Liang Wang。通过自适应增强进行图对比学习。载于*2021*年万维网会议论文集，第2069–2080页，2021年。[Zhu， 2005] Xiaojin Jerry Zhu。半监督学习文献综述。2005年。

实验设置

所有实验均在两块 A100 图形处理器上运行。我们使用 PyTorch Geometric 软件包（PyG） [Fey 和 Lenssen， 2019] v.2.2.0 构建所有图神经网络。使用 Adam 优化器传播梯度。我们尝试使用 ReduceLROnPlateau 调度器调整学习率，衰减系数为 0.6，耐心值为 10，但发现恒定学习率的效果也很好。运行 SemiMol 实验时需要搜索的超参数列于表 4。

表 4：SemiMol 的超参数设置。

超参数搜索空间	符号	值
训练设置		
学习率	-	[1e-4, 5e-5, 1e-6, 5e-6]
批量大小	-	[32, 64, 128]
训练轮数	-	[100, 300]
丢弃率	-	[0.2, 0.4]
SemiMol 设置		
损失权重	λ	[0.1、1.0、10.0、100.0]
初始置信度阈值	γ	[0.8、0.9、0.95]
置信度步进百分比	$\Delta\gamma$	[0.05、0.1、0.2]
伪标签更新频率	k	[5, 10, 15, 20]
神经网络设置		
消息传递神经网络层数	-	[2, 3, 4, 5, 6]
节点表示的隐藏维度	-	[32, 64, 128, 256, 512]
边表示的隐藏维度	-	[64, 128, 256]
GMT 中的头数	-	[4, 8 ,12]
GMT 中的隐藏维度	-	[64, 128, 256, 512]
全连接层的数量	-	[1, 2, 3]

基线模型

这里我们详细介绍 ACE 问题中的一些关键基线模型。

基于描述符的模型

KNN、SVM、GBM、随机森林和多层感知机。每种算法都结合了四种类型的分子描述符，即由人工设计、旨在捕捉预定化学信息的数值特征。我们探索了复杂程度各异的分子描述符：(1) 扩展连接指纹（ECFP），以二进制数组的形式编码以原子为中心的径向子结构；(2) 分子访问系统（MACCS）键，以二进制数组编码预定义子结构是否存在；(3) 加权整体不变分子（WHIM）描述符，用于表征分子三维尺寸、形状、对称性和原子分布；以及 (4) 与类药性相关的十一种理化性质。

基于图的模型

GCN、GAT、MPNN 和 AFP 均使用 PyTorch Geometric 软件包实现。在 GCN、GAT 和 MPNN 中，全局池化通过图多重集 Transformer 实现。

基于 SMILES 的模型

所有 SMILES 字符串均编码为独热向量。长度超过 200 个字符的 SMILES 字符串会被截断（平均占 0.71%）。使用 RDKit［Landrum 等人， 2013]生成非规范 SMILES 字符串。**长短期记忆网络** [Hochreiter and Schmidhuber， 1997] 是一种循环神经网络（RNN），能够通过追踪长程依赖，从字符串序列中学习。在我们的研究中，长短期记忆网络模型先通过下一个字符预测，对合并所有训练集并去除重复项后得到的 SMILES（36K 个分子）进行预训练，然后再应用迁移学习进行生物活性预测。

Transformer [Vaswani 等人， 2017] 以类似图的方式一次性处理整个序列，并利用位置嵌入捕捉位置信息。Transformer 采用自注意力机制，学习序列中哪些部分与特定任务更相关。ChemBERTa 架构在 1000 万个化合物上进行了预训练，并结合迁移学习用于生物活性预测。**一维卷积神经网络** 使用一个卷积核大小为 1 的一维卷积层，后接一个全连接层。

Data

无标签数据

我们使用 ZINC15 [Sterling 和 Irwin， 2015] 数据库收集无标签分子数据，该数据可从 DeepChem [Ramsundar 等人， 2019]下载。ZINC15 支持四种不同的数据规模：25 万、100 万、1000 万和 2.7 亿。为简便起见，我们使用 25 万规模的数据作为无标签数据集。随后，使用 RDKit [Landrum 等人， 2013] 将这些无标签 SMILES 转换为二维图。我们希望未来的工作能利用 ChemBL [Gaulton 等人， 2012], Chembridge 和 Chemdiv 等其他资源来丰富无标签语料库。

活性断崖数据

MoleculeACE 是一种用于评估机器学习模型对活性断崖化合物预测性能的工具。它收集并整理了针对 30 个大分子靶标的生物活性数据。整理后的数据集共包含 48.7K 个分子，其中 35.6K 个为独特分子；它涵盖多个靶标家族和不同训练场景，因而能够模拟典型的药物发现数据集。对于每个大分子靶标，通过考察成对分子的结构相似性以及效力差异来识别活性悬崖。采用三种不同方法量化任意两种分子间的分子相似性，分别是子结构相似性、骨架相似性以及 SMILES 字符串的相似性。

由于 MoleculeACE 是一个相对较新的基准测试，此前没有研究全面计算其统计数据，因此我们在表 5 中计算了关键统计数据，以便读者全面了解其分子分布。MoleculeACE 中的所有数据集均采用 [van Tilborg 等人， 2022]建议的划分方式。数据统计信息及这些数据集的简要介绍见表 5。有关数据整理和活性悬崖检测的更多详情，请参阅 [van Tilborg 等人， 2022] 。

表 5：MoleculeACE 中 30 个回归数据集的描述，包括分子数量以及真实属性的最大值、最小值、均值和标准差。

	化合物数量	Min	Max	Mean	Std
CHEMBL4203 抑制常数	731	-4.90	0.50	-2.50	1.02
CHEMBL2034 抑制常数	750	-4.27	1.00	-1.20	1.09
CHEMBL233 抑制常数	3,142	-4.80	2.00	-1.68	1.29
CHEMBL4616 半数有效浓度	682	-4.15	1.00	-1.20	0.93
CHEMBL287 抑制常数	1,328	-4.67	1.44	-1.52	1.09
CHEMBL218 半数有效浓度	1,031	-4.99	1.52	-2.25	1.05
CHEMBL264 抑制常数	2,862	-4.93	1.60	-1.25	1.10
CHEMBL219 抑制常数	1,859	-4.95	1.74	-1.77	1.07
CHEMBL2835 抑制常数	615	-3.66	1.00	-0.33	0.95
CHEMBL2147 抑制常数	1,456	-5.00	2.00	-1.14	1.96
CHEMBL231 抑制常数	973	-5.00	1.31	-2.07	1.26
CHEMBL3979 半数有效浓度	1,125	-4.79	1.22	-2.27	1.15
CHEMBL237 半数有效浓度	955	-4.70	1.96	-1.56	1.39
CHEMBL244 抑制常数	3,097	-5.00	2.00	-1.86	1.64
CHEMBL4792 抑制常数	1,471	-4.25	1.15	-2.07	1.11
CHEMBL1871 抑制常数	659	-4.75	0.60	-1.95	1.03
CHEMBL237 抑制常数	2,602	-4.91	1.85	-1.66	1.33
CHEMBL262 抑制常数	856	-5.00	1.05	-2.49	1.06
CHEMBL2047 半数有效浓度	631	-4.96	0.52	-2.77	0.98
CHEMBL239 半数有效浓度	1,721	-4.95	1.59	-2.65	1.09
CHEMBL2971 抑制常数	976	-5.00	1.22	-1.14	1.45
CHEMBL204 抑制常数	2,754	-6.39	2.00	-2.10	1.51
CHEMBL214 抑制常数	3,317	-4.80	1.85	-1.66	1.14
CHEMBL1862 抑制常数	794	-5.00	1.73	-1.41	1.51
CHEMBL234 抑制常数	3,657	-4.86	1.70	-1.54	1.18
CHEMBL238 抑制常数	1,052	-4.97	0.35	-2.34	1.09
CHEMBL235 半数有效浓度	2,349	-5.00	1.74	-2.57	1.09
CHEMBL4005 抑制常数	960	-4.35	1.52	-1.31	1.07
CHEMBL236 抑制常数	2,598	-5.00	2.00	-2.01	1.35
CHEMBL228 抑制常数	1,704	-4.86	1.89	-1.67	1.21

在 CYP3A4 中，通过 RDKit 软件包中的 MMPA46 [Landrum 等人， 2013]，将活性化合物与非活性化合物进行了比较。这产生了 106 对分子对及其对应的子结构，涉及 46 个活性化合物和 51 个非活性化合物。

在图神经网络模型训练期间，所有涉及已发现分子对的化合物都被放入测试集。CYP3A4 可在 Mol-XAI 的 GitHub 上获取 [Rao 等人， 2022] https://github.com/biomed-AI/MolRep，数据存储于 https://drive.google.com/drive/folders/1Lbb5KBat4Au 1yMHKyYnAxf6xVFNWtkW。

为进行全面比较，我们选取随机森林（RF）、梯度提升（GBM）、支持向量回归（SVM）、K近邻（KNN）、消息传递神经网络（MPNN） [Gilmer等人、 2017], 图卷积网络（GCN） [Kipf 和 Welling、 2016], 图注意力网络（GAT） [Veli˘ckovi´c 等人、 2017], 注意力指纹（AFP） [Xiong 等人、 2019], 长短期记忆网络（LSTM） [Hochreiter and Schmidhuber、 1997], 1 D 卷积神经网络（CNN） [Kimber 等人以及 Transformer [Vaswani 等人作为基线模型。我们按照 MoleculeACE [van Tilborg等人 2022]中的建议，使用高斯过程进行贝叶斯优化，以优化超参数。这些架构的具体细节如下：

- KNN、SVM、GBM 和随机森林回归模型使用 sklearn [Pedregosa 等人， 2011] 实现。对于 KNN，我们从候选值 $k = [3, 5, 11, 21]$ 中优化邻居数量 k。对于 SVM，我们从候选值 $\gamma = [1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}$, 或 $1 \times 10^{-1}]$和 $C = [1, 10, 100, 1000, 10,000]$ 中优化核系数 γ 和正则化参数 C。对于 GBM，我们从候选值 $n_{\mathrm{b}} = [100, 200, 400]$ 和 $m_{\mathrm{d}} = [5, 6, 7]$ 中优化提升阶段数量 n_{b} 和最大模型深度 m_{d}。对于随机森林，我们从候选列表 $t = [100, 250, 500, 1000]$ 中选择决策树数量 t。
- 对于基于序列的方法，SMILES 字符串被编码为独热向量。长度超过 200 个字符的 SMILES 字符串会被截断。LSTM 模型包含 4 层，其隐藏单元数分别为 1024、1024、256 和 256。Transformer 模型及其对应的 SMILES 分词基于 ChemBERTa [Baek等人， 2021]。
- 对于基于图的基线方法，我们主要搜索 $[32, 64, 128, 256, 512]$, 卷积层的数量、 $[1, 2, 3, 4, 5]$, 隐藏多重集 Transformer节点以及 $[64, 128, 256, 512]$ 隐藏预测器特征 $[128, 256, 512]$。对于 AFP，我们调节时间步 $[1, 2, 3, 4, 5]$。

我们还选取了三种最先进的预训练图神经网络，以检验预训练对活性悬崖的有效性。GROVE [Rong等人， 2020] 从其官方 GitHub 下载，网址为 https://github.com/tencent-ailab/grover。具体而言，我们使用 GROVE 的基础预训练模型。MolCLR 从其官方 GitHub 下载，网址为 https://github.com/yuyangw/MolCLR；GEM 从其官方 GitHub 下载，网址为 https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/pretrained compound/ ChemRL/GEM。GROVE 和 MolCLR 均基于 PyG 实现，而 GEM 则基于其自行构建的 PaddlePaddle 平台实现。 [Ma 等人， 2019]。

蛋白质中的活性悬崖

在正文中，我们主要讨论机器学习算法在非活性悬崖小分子与活性悬崖小分子上的表现差异。然而，这种现象也可能出现在大分子中，例如蛋白质。例如，一些先前的研究 [Shortle, 2009] 发现，单个氨基酸替换就可能彻底改变蛋白质的折叠方式。由这一次突变引起的结构变化，不能被简单归结为“何谓不同折叠方式”这一语义问题。具体而言，一种构象由三螺旋束组成，而另一种构象则具有一个四股 β折叠片层和一条单独的 α螺旋。尽管序列中的其他片段相同，这一个氨基酸仍可能造成巨大的结构差异。这种“序列相似、性质（即结构）不同”的情况，可能给蛋白质结构预测等任务带来巨大挑战；它可以看作大分子中活性悬崖的另一种形式。我们期待未来能有更多研究，探索机器学习算法在此类大分子建模中的表征行为，并考察它们是否会遭遇与表征活性悬崖小分子相同的问题。