arXiv:2611.32371v1 [q-bio.BM] 24 Nov 2026

基于描述的药物-药物相互作用基准揭示了依赖数据划分的分子表示排名

Manel Gil-Sorribes[1] **Isaac Filella-Merce**[1] **Alexis Molina**[1,2]

[1]Nostrum Biodiscovery[2]
AITHYRA

manel.gil@nostrumbiodiscovery.com
alexis.molina@nostrumbiodiscovery.com

摘要

药物-药物相互作用（DDI）预测在临床上至关重要，也对分子表示学习提出了独特挑战。与单分子性质预测不同，DDI 预测要求模型将两种分子表示组合起来，得出关系结果。DrugBank 提供的是由编者撰写的药物-药物相互作用描述，而非现成的分类标签，因此必须先从文本中构建分类预测目标。我们通过将处理后的描述映射到六种文本衍生类别，构建了一个药物-药物相互作用配对数据集，并设定了四种留出方案，以限制训练集与测试集之间的化学重叠程度，范围从随机配对到骨架不相交评估。随后，我们使用该基准，在五个随机种子和多个分类器头下，对比分子指纹、预训练序列表示和图衍生特征。在高重叠的随机划分中，分子指纹表现最佳；而在药物留出和骨架留出方案中，图衍生表示表现最佳，且随着分子重叠受到的限制增加，性能差距也随之扩大。改变分类器头会影响绝对性能，但不会改变不同表示的相对排名。这些结果表明，在固定的描述衍生标签空间下，表示排名会随测试集新颖性的变化而改变。

1 引言

药物-药物相互作用（DDI）预测不同于许多分子性质任务，因为预测目标属于一对分子，而非单个分子。下游预测器必须结合两个分子表示，以推断出一种无法由任一单分子表示直接编码的相互作用。DrugBank 大规模提供了由整理人员撰写的自然语言相互作用描述 [1]。然而，将这些描述转换为分类预测目标本身就是一种数据集构建选择。

评估引入了第二个选择。现有 DDI 研究采用配对和药物留出方案[2], , 以及知识图谱形式 [3]。留出配对并不一定会留出其中的分子内容：测试配对中仍可能包含训练集中出现的药物、骨架或密切相关的

结构。骨架感知评估解决了这一问题 [4, 5]。
Morgan 指纹编码固定的局部环境 [6], MoLFormer 和 ChemBERTa 学习分子字符串表示 [7, 8]，而基于图的表示则通过图神经网络和消息传递来编码分子连接性 [9, 10]。先前的药物-药物相互作用实验已表明，当仍存在大量化学重叠时，指纹可能更具优势 [11]。

提交至第 40 届神经信息处理系统大会（NeurIPS 2026）。研讨会：分子科学智能体系统。请勿传播。

因此，我们要探究：当标签保持不变而测试群体的分子新颖性发生变化时，表示排名是否仍然保持不变。

为评估不同分子表示在不同留出方案下对药物-药物相互作用（DDI）的影响，我们构建了一个包含 1,392,791 个无序 DrugBank 配对和六类文本衍生的相互作用类别的数据集。我们用“数据集”指代这些已标注配对，用“基准”指代该数据集以及固定的划分清单和评估协议。该基准包含随机配对、单药未见、双药未见和骨架不相交情形，从而能够在限制与训练数据重叠的情况下，评估同一表征网格。

我们的贡献总结如下：

- 一个包含 1,392,791 个无序 DrugBank 配对和六类可追溯的描述衍生类别的药物-药物相互作用（DDI）数据集，
- 一个包含四种评估方案的基准，明确控制配对、药物、骨架和分子相似性重叠，且
- 在五个模型随机种子下重复进行实验，结果表明，分子表示的排名会随测试集新颖性而变化，而分类器头复杂度对结论的影响较小。

2 数据集构建与基准设计

2.1 DDI 数据集与文本衍生标签

从 DrugBank 药物-药物相互作用记录出发，将有向条目规范化为每个无序 DrugBank ID 配对对应一行，得到涵盖 4,437 种药物的 1,392,791 个已标注配对；附录B和表8提供了具有代表性的描述示例。随后，从每条描述中移除主体药物和受影响药物的名称，以减少直接身份线索。使用以 PubMedBERT 为骨干网络的生物医学 SentenceTransformer 对清理后的描述进行嵌入 [12, 13]。表1总结了已标注数据集中的分子结构覆盖范围。所报告的划分清单基于完整的已标注配对表构建，在特征构建过程中处理不可用的分子特征。有关数据集构建的更多细节见附录A。

表1：数据集与分子结构覆盖范围。完整的已标注数据集提供了用于构建所报告划分清单的配对总体。结构解析子集包含两个端点均具有有效且可由 RDKit 解析的 SMILES 的配对。

表或子集	Rows	配对	药物	Role
完整标注的药物-药物相互作用数据集	1,392,791	1,392,791	4,437	报告划分清单的基础
两种药物均具有有效SMILES的配对	1,158,783	1,158,783	3,468	结构完全解析的子集

为每个药物对生成分类标签，我们对相应的药物-药物相互作用描述嵌入进行了聚类。首先，我们抽取了 50,000 条描述，用于评估不同的聚类配置。在测试过的凝聚式设置中，六类别方案取得了最高的综合分离度与平衡度得分。随后使用 MiniBatchKMeans 对完整数据集进行可扩展的分配 [14]。表2总结了与所得类别相关的反复出现的描述信号。完整的聚类诊断结果和代表性措辞模式见附录B。

由于这些类别仍可追溯到 DrugBank 中反复出现的措辞，因此可以对照源描述检查其内容。文本编码器用于生成基准标签，且不作为药物-药物相互作用预测的分子表征进行评估（图1）。

我们通过将这一固定标签空间与四种预定义的训练集、验证集和测试集划分相结合，构建了药物-药物相互作用基准。因此，在不同评估方案中，标签保持不变，而训练集与测试集之间共享的分子信息的数量和类型则受到控制。这种区分是该基准的核心：不同划分之间的差异可以基于同一预测目标进行解读，而非基于独立构建的药物-药物相互作用数据集。

表2：由六种文本衍生类别表示的 DrugBank 描述中反复出现的信号。这些类别概括的是描述性语言，而非专家裁定机制。

类别	反复出现的描述信号
0	暴露、血清浓度、蛋白质结合、代谢物
1	暴露、疗效、不良反应、生物利用度
2	排泄、血清水平、疗效变化
3	中枢神经系统效应、血压、凝血、心脏效应
4	出血、QTc、高血压、电解质和安全性效应
5	代谢和吸收

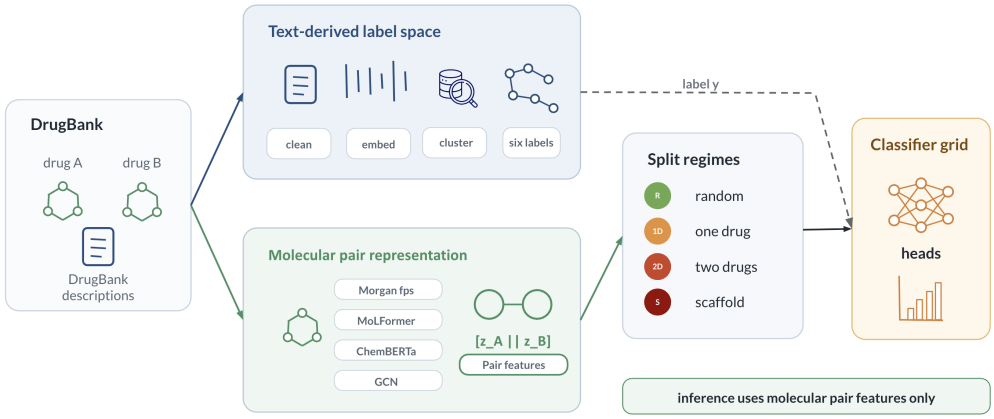

图1：基准构建与评估。DrugBank描述用于获取六种文本衍生标签，而分子表示则是在配对级、药物级和骨架级留出条件下评估的推理输入。

2.2 基 准机制与表征评估

评估

四种评估方案在保持标签空间和训练协议固定的同时，改变了留出群体。随机配对留出用于评估未见组合，同时保留较高的单个药物重叠。单药未见评估要求测试端点中恰有一个在训练集中未出现，而双药未见评估则将两个端点都从训练集中移除。骨架不相交评估还要求训练集与测试集之间的 DrugBank ID、规范 SMILES 和 Bemis-Murcko框架重叠为零[5]。ECFP4 相似度 [6] 使用 RDKit [15], 计算，并将验证集/测试集到训练集的最大 Tanimoto相似度限制为 0.70。表 3 报告了由此得到的划分规模。所有方案的训练集与测试集无序药物对重叠均为零。骨架不相交评估还确保训练集与测试集之间的药物身份重叠为零；在结构已解析化合物中，规范 SMILES 和 Bemis-Murcko骨架重叠也为零，且最大 ECFP4 Tanimoto 相似度为 0.70。

表 3：不同基准方案下的训练集、验证集和测试集规模。药物留出和骨架留出设置仅保留满足相应划分约束的配对。

划分	训练集	Val.	Test	总计
随机	974,954	139,279	278,558	1,392,791
一种未见药物	675,377	191,699	397,605	1,264,681
双药未见	691,248	13,179	54,956	759,383
骨架不相交	806,341	10,970	36,621	853,932

所有骨架不相交的验证集和测试集配对均已完成结构解析。在 806,341 个训练对中，有 613,853 个配对的两个端点均具有有效结构；对于不可用的分子特征，则采用附录 C 中所述的零向量回退进行处理。

该基准评估了四种分子输入形式，它们在构建方式和预训练信号方面均有所不同。Morgan 指纹是固定的、由规则定义的环状描述符 [6]。MoLFormer 和 ChemBERTa 使用通过掩码标记

目标学习得到的自监督分子字符串表示 [7, 8]。GCN 特征使用在 PCBA-1328 [16, 17] 上训练的 Tensor-DTI 药物编码器。成对特征按照规范的 DrugBank ID 顺序进行拼接。为测试对下游分类器容量的敏感性，四种表征均使用相同的预定义分类器头网格和五个模型随机种子进行评估。架构和优化细节见附录 C。

3 基准揭示了什么

随着分子新颖性增加，基准揭示了明显的表征排名逆转：在高重叠的随机划分中，Morgan 指纹表现最佳；而在一种未见药物、双药未见和骨架不相交情形下，Tensor-DTI 图衍生特征领先。

表4：不同划分方案下的表征性能。每个单元格报告了在预定义的分类器头网格内，五个随机种子下观测到的最佳宏平均 F1 ± 标准差。粗体表示每种方案中的领先表征，下划线表示第二佳表征。

表征	随机	一种未见药物	双药未见	骨架不相交
Morgan 对	**0.899 ± 0.000**	0.631 ± 0.002	0.412 ± 0.003	0.287 ± 0.005
MoLFormer 配对	0.876 ± 0.001	0.632 ± 0.001	0.420 ± 0.005	0.308 ± 0.005
ChemBERTa对	0.866 ± 0.001	0.623 ± 0.001	0.398 ± 0.005	0.299 ± 0.005
GCN 对	0.856 ± 0.001	**0.635 ± 0.001**	**0.432 ± 0.004**	**0.324 ± 0.005**

表4量化了这一逆转。在单药未见情形下，图卷积网络相对于 Morgan 的优势为 0.004 个宏平均 F1；在双药未见情形下增至 0.020，在骨架不相交情形下增至 0.037。相对于最强的非图卷积网络替代方法，其优势在这些设置下也逐步扩大，分别从 0.003 增至 0.012 和 0.016。在骨架互斥测试集中，图卷积网络在已填充的相似度区间内也均优于 Morgan（图2）。因此，如果基准仅限于随机划分，Morgan 指纹将被排在首位，从而掩盖分子新颖性条件下所观察到的逆转。由于不同情形下的预测目标、表征网格和训练协议均保持不变，结果表明，表征排名高度依赖于训练集与测试集之间分子重叠的程度。更多配置排名、配对比较和诊断指标见附录E。

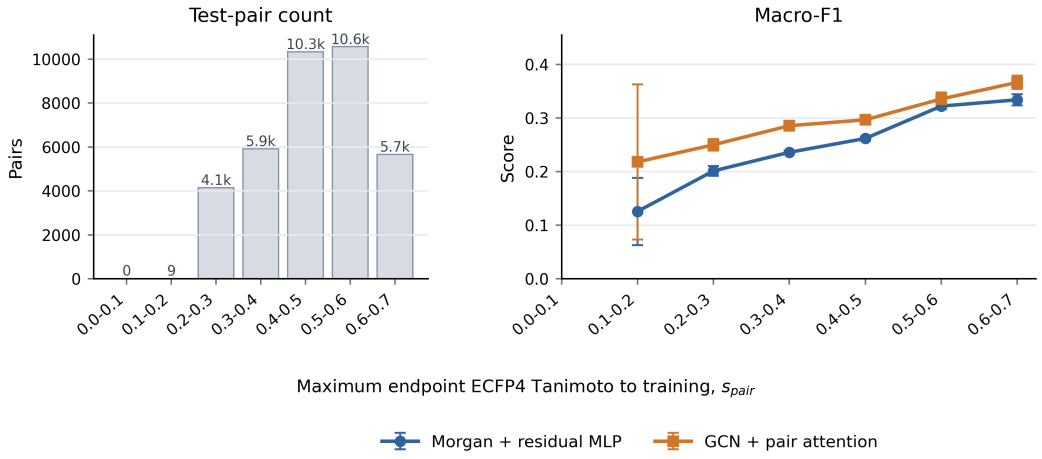

图2：骨架不相交性能随其与训练化学空间的相似度变化的情况。左：按端点与训练集之间的ECFP4 Tanimoto相似度最大值统计的测试对数量。右：Morgan 和图卷积网络的五个随机种子的宏平均 F1。误差线表示均值 ±标准差。

分类器头部变化小于分子表示和划分效应，并且改变注意力头不会改变任何设置下的获胜分子表示。

4 局限性

这六个类别源自DrugBank描述中反复出现的语言，而非经专家裁定的机制，因此可能继承来源数据中的标注或措辞偏差。这些分子表示在架构和预训练数据方面也存在差异，因此观察到的图派生优势不能归因于其中任何单一因素。

5 结论

药物-药物相互作用预测中的表征排名会随评估机制显著变化。我们还提出了一种基于描述构建的药物-药物相互作用基准，将固定的六类别标签空间与配对级、药物级和骨架级留出相结合，并明确控制分子相似性。在五个随机种子下，高重叠随机抽样中 Morgan 指纹表现最佳，而在单药未见、双药未见和骨架不相交设置中，GCN 特征表现最佳。在所评估的网格中，分类器注意力头的变化会影响同一表征下的绝对性能，但不会改变这种依赖划分的排名。

本研究表明，表征的选择取决于训练集与测试集之间允许重复出现哪些信息。随机药物对划分主要评估未见组合，而药物身份和局部化学模式在训练集中仍有充分体现。药物留出和骨架不相交设置则分别检验向未见化合物和化学邻域迁移的互补能力。没有哪一种单一机制适用于所有应用，但评估划分应反映预期的部署场景。

这些结果进一步凸显了这样做的必要性：在明确控制重叠程度的条件下，将学习得到的分子表示与简单的指纹基线进行比较。仅报告随机划分下的得分，可能会偏向于利用重复出现的化学信息的表示，同时掩盖其在分子新颖性条件下呈现出的不同排名。我们的基准提供了一种明确这些条件的方法；更广泛而言，药物-药物相互作用基准应在报告性能的同时，报告训练集与测试集之间药物、骨架及分子相似性重叠的程度。缺少这一背景信息时，表示排名可能在很大程度上反映评估设计，而非分子特征所捕获的可迁移信息。

参考文献

[1] Craig Knox、Mike Wilson、Christen M. Klinger、Mark Franklin、Eponine Oler、Alex Wilson、Allison Pon、Jordan Cox、Na Eun Chin、Seth A. Strawbridge 等。DrugBank 6.0：2024年 DrugBank 知识库。《核酸研究》，52(D1)：D1265–D1275，2024年。doi：10.1093/nar/gkad976。[2] Yi Zhong、Gaozheng Li、Ji Yang、Houbing Zheng、Yongqiang Yu、Jiheng Zhang、Heng Luo、Biao Wang 和 Z uquan Weng。通过局部-全局自注意力学习基于基序的图，用于药物-药物相互作用预测。《自然·机器智能》，6：1094–1105，2024年。doi：10.1038/ s42256-024-00888-6。[3] Yue Yu、Kexin Huang、Chao Zhang、Lucas M. Glass、Jimeng Sun 和 Cao Xiao。SumGNN：通过高效知识图谱摘要进行多类型药物相互作用预测。《生物信息学》，37(18)：2988–2995，2021年。doi：10.1093/bioinformatics/btab207。[4] 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。《化学科学》，9(2)：513–530，2018年。doi：10.1039/C7SC02664A。[5] Guy W. Bemis 和 Mark A. Murcko。已知药物的性质。1. 分子骨架。《药物化学杂志》，39(15)：2887–2893，1996年。doi：10.1021/jm9602928。[6] David Rogers 和 Mathew Hahn。扩展连接指纹。《化学信息与建模杂志》，50(5)：742–754，2010年。doi：10.1021/ci100050t。[7] Jerret Ross、Brian Belgodere、Vijil Chenthamarakshan、Inkit Padhi、Youssef Mroueh 和 Payel Das。大规模化学语言表征捕获分子结构与性质。《自然·机器智能》，4：1256–1264，2022年。doi：10.1038/s42256-022-00580-7。[8] Seyone Chithrananda、Gabriel Grand 和 Bharath Ramsundar。ChemBERTa：用于分子性质预测的大规模自监督预训练。*arXiv* 预印本 *arXiv:2010.09885*，2020年。doi：10.48550/arXiv.2010.09885。[9] Thomas N. Kipf 和 Max Welling。使用图卷积网络进行半监督分类。国际学习表征会议，2017。URL https:// openreview.net/forum?id= SJU4ayYgl。[10] 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。载于第*34*届国际机器学习会议论文集，第70卷，第1263–1272页。PMLR，2017年。[11] Manel Gil-Sorribes 和 Alexis Molina。使用分子指纹解决药物-药物相互作用预测中的模型过度复杂问题。《化学信息学杂志》，18(1)：38，2026。doi：10.1186/s13321-025-01128-8。[12] Yu Gu、Robert Tinn、Hao Cheng、Michael Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。面向生物医学自然语言处理的领域特定语言模型预训练。《*ACM*医疗保健计算汇刊》，3(1)：1–23，2021年。doi：10.1145/3458754。[13] Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。载于*2019*年自然语言处理实证方法会议暨第*9*届国际自然语言处理联合会议论文集，第3982–3992页，中国香港，2019年。计算语言学协会。doi：10.18653/v1/D19-1410。[14] D. Sculley。网络规模的k均值聚类。载于第*19*届国际万维网会议论文集，WWW '10，第1177–1178页，美国纽约州纽约，2010年。计算机协会。doi：10.1145/1772690.1772862。[15] Greg Landrum、Paolo Tosco、Brian Kelley 等。RDKit：发行版 2024_09_5。Zenodo 软件发行版，2025年。

[16] Manel Gil-Sorribes、Júlia Vilalta-Mor、Isaac Filella-Mercè、Robert Soliva、Álvaro Ciudad、Víctor Guallar 和 Alexis Molina。Tensor-DTI：通过对比嵌入学习增强生物分子相互作用预测。*arXiv* 预印本 *arXiv:2601.05792*，2026 年。doi：10.48550/arXiv.2601.05792。[17] Sunghwan Kim、Jie Chen、Tiejun Cheng、Asta Gindulyte、Jia He、Siqian He、Qingliang Li、Benjamin A. Shoemaker、Paul A. Thiessen、Bo Yu、Leonid Zaslavsky、Jian Zhang 和 Evan E. Bolton。PubChem 2023 年更新。《核酸研究》，51(D1)：D1373–D1380，2023 年。doi：10.1093/nar/gkac956。[18] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q. Weinberger。现代神经网络的校准。载于：第 *34* 届国际机器学习会议论文集，第 70 卷，第 1321–1330 页。PMLR，2017 年。

A 基于数据集构建的更多细节

DrugBank 相互作用记录被规范化为每个无序 DrugBank ID 对对应一行。在进行文本嵌入之前，从每条相互作用描述中移除主体药物名称和受影响的药物名称，以减少直接身份线索。清理后的语料库包含 417 条规范化描述字符串，其中 380 条出现在评估数据集中。

在全部已标注配对中，有 1,158,783 对的两端均具有可由 RDKit 解析的有效 SMILES，对应 3,468 种药物。不过，所报告的划分清单是根据完整的已标注配对表构建的。包含未收录于分子特征库中的药物的配对仍会保留，并在特征构建过程中将无法获取的端点表征替换为零向量。DrugBank 源数据仍受适用许可条款约束。基准测试脚本、汇总文件和论文资源将另行发布。

B 文本聚类与标签解读

使用 SentenceTransformer 检查点 pritamdeka/S-PubMedBert-MS-MARCO[1],对清理后的描述进行嵌入；该检查点以 PubMedBERT 为底层模型，并采用 Sentence-BERT 框架 [12, 13]。在对完整表格进行可扩展分配之前，先基于 50,000 个描述嵌入的样本评估候选聚类数量。综合选择得分为

$$\text{selection score} = \text{silhouette} - \text{Davies-Bouldin} + \text{cluster size balance},$$

其中，簇大小平衡度为最小簇大小除以最大簇大小。Calinski-Harabasz 指数单独报告，作为一项额外的诊断指标，不纳入综合得分。

表5比较了抽样的候选粒度。采用 $k = 6$ 的凝聚层次聚类具有所列最强的综合得分。随后使用MiniBatchKMeans [14] 在选定的六类别粒度上对完整语料库进行分配，这是一种实用的基准选择，而非数学上唯一的最优解。

表5：基于50,000个SentenceTransformer描述嵌入样本的聚类粒度诊断。综合选择得分为轮廓系数减去Davies-Bouldin指数再加上簇大小平衡度。Calinski-Harabasz指数予以报告，但不计入该得分。采用 $k = 6$的凝聚层次聚类具有所列最强得分，并促使采用六类别标签粒度进行全语料库分配。

方法	k	Sil.	Davies-Bouldin 指数	Calinski-Harabasz 指数	平衡	得分
凝聚式	6	0.565	1.205	19,135	0.719	0.079
凝聚式	8	0.644	1.045	20,807	0.253	-0.148
凝聚式	50	0.943	1.095	34,605	0.003	-0.148
凝聚式	30	0.892	1.112	24,878	0.020	-0.199
凝聚式	5	0.506	1.184	18,390	0.386	-0.292

Ben基准的流行度按频率加权，因为某些清理后的模板会反复出现多次 s. Fig因此，图3和表6同时报告按配对加权的出现率和唯一模板数量。

聚类0、1、2和5主要由少数反复出现的模板构成，而聚类3和4的模板多样性更高。同时报告按配对加权的计数和唯一模板数量，可以清楚地呈现这一差异。

表7和表8提供了描述层面的可解释性辅助工具。它们总结了基准目标所使用的反复出现的措辞模式，但并非外部生物学验证。

这些类别可以追溯到源描述中反复出现的措辞，但不应将其解读为经过独立判定的机制性或临床标签。

[1]https: //huggingface.co/pritamdeka/S-PubMedBert-MS- MARCO

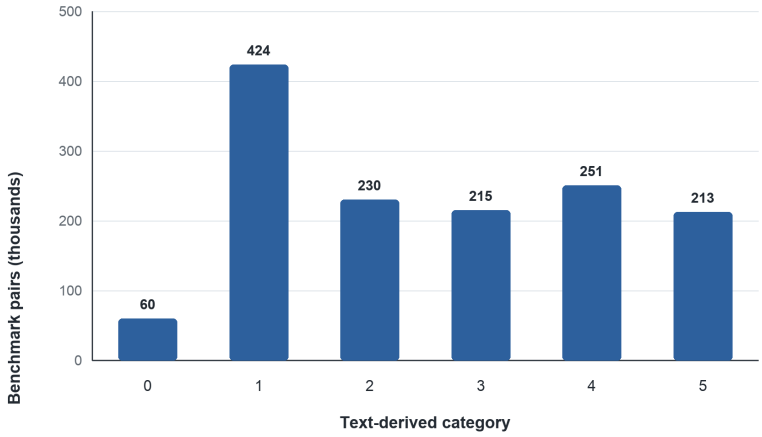

图3：六种文本派生的相互作用描述类别的簇大小分布 .

表6：用于模型评估的基准表的簇统计。配对百分比按基准表行的频率加权，而模板计数则将规范化描述视为唯一字符串。

聚类	配对	配对百分比	清理后的模板	前五模板行百分比
0	60,169	4.3	23	99.8
1	423,895	30.4	37	99.8
2	230,493	16.5	17	99.9
3	215,150	15.4	104	69.3
4	250,567	18.0	182	54.7
5	212,517	15.3	18	100.0

表7：六种文本派生的相互作用描述类别的文字摘要。这些摘要将每个基准标签与反复出现的 DrugBank 措辞模式联系起来，并定性呈现标签的可解释性。

聚类	描述级摘要	代表性信号
0	药代动力学暴露变化、代谢抑制或诱导、蛋白质结合与分布、治疗效果调节	血清、浓度、蛋白质、结合、活性、代谢物，代谢物
1	药代动力学暴露变化、药效学毒性累加性、治疗作用调节	效应、不良、风险、严重程度、治疗、疗效、生物利用度能力
2	药代动力学暴露变化、治疗作用调节tion	水平、排泄、血清、疗效、降低
3	治疗作用调节、心脏电生理、中枢神经系统抑制、血压和血管张力、凝血与出血、血糖调节	抑制剂，活性，神经，中枢，CNS，抗高血压药，降压药，致心律失常
4	药效学毒性叠加，心脏电生理学血压和血管张力，凝血与出血，肾脏与电解质安全性，血糖调节	严重程度，风险，高血压，出血，延长，QTc，高血血钾水平，出血
5	药代动力学暴露变化、代谢抑制或诱导、治疗作用调节	代谢、吸收

表8：文本衍生标签的代表性措辞模式。这些类似模板的模式提取自经过清理的高频描述，其中省略了药物名称，并展示了反复出现的相互作用描述用语。

聚类	代表性描述模式
0	合用时血清浓度可能升高，蛋白质结合及活性代谢物相关表述
1	合用时治疗疗效可能降低，合用时治疗疗效可能提高
2	排泄率可能降低，导致血清水平升高，血清水平及排泄相关表述
3	抗凝活性可能增强，中枢神经系统抑制活性可能增强
4	出血风险或严重程度可能增加，高血压风险或严重程度可能增加
5	联合使用时，代谢可能会增加 ，联合使用时，代谢可能会减少

C 分子表示与模型头

本节分别讨论分子表示选择、分子对构建、分类器头容量以及通用训练协议。Morgan 指纹使用半径 2 和 2048 位，MoLFormer 使用 ibm/MoLFormer-XL-both-10pct，ChemBERTa 使用 DeepChem/ChemBERTa-77M-MLM，GCN 特征使用在 PCBA-1328 [16, 17] 上训练的 Tensor-DTI 药物编码器。所有分子对特征均按照规范化的 DrugBank ID 顺序拼接端点向量。

表 9 给出了模型头和优化设置。辅助回归头仅将描述嵌入用作训练时的目标。相互作用文本不是推理时的模型输入。其作用是将分子对表示正则化，使其趋向于文本衍生的标签空间。

表 9：最终流式 Torch 实验中使用的模型头和训练超参数。数值取自用于基准网格的训练脚本。

组件	设置
线性头	从输入特征到六个类别的单个线性层
多层感知机头	512 个单元的隐藏层、ReLU、0.2 的 dropout、256 个单元的隐藏层、ReLU、0.2 的 dropout、六分类输出
残差多层感知机	512 维投影后接两个残差块，每个块使用线性-ReLU-丢弃 0.2-线性，并进行层归一化
门控多层感知机	512 维值分支与 512 维 Sigmoid 门控相乘，随后依次接丢弃 0.2、256 单元隐藏层、丢弃 0.2 和输出层
特征注意力多层感知机	512 维值分支与学习得到的 Sigmoid 相乘特征注意力向量，随后输入与门控多层感知机相同的 256 单元注意力头
配对注意力头	两个药物向量投影至 256 维，使用四头自注意力机制在两个标记上应用 0.1 的 dropout，经层归一化的 512 维展平标记输出输入带有 0.2 dropout 的 512 单元注意力头
辅助回归头	512单元和256单元编码器，dropout为0.2，交叉熵分类器以及使用余弦嵌入损失训练的768维回归头，$\lambda_{\text{cls}} = 1.0$，$\lambda_{\text{reg}} = 1.0$
优化	AdamW，学习率 10^{-3}，权重衰减 10^{-4}，训练30轮，批大小1024，评估批大小 8192
数据处理	流式特征构建，并将训练集特征标准化应用于留出批次，最终 Torch 运行使用分配给每个划分的所有行，扩展比较还包括八种浅层基线配置。对于表征特征库中不存在的药物，将其指定为零向量。
随机种子	数据构建随机种子 42，模型随机种子 11，22, 33, 44, 以及 Torch 的 55 grid

共享表示和优化控制有助于将性能差异归因于结果表中记录的表征、划分方案和头部选择。数据构建使用随机种子 42，模型结果则汇总了随机种子 11、22、33、44 和 55 的结果。端点向量采用规范的 DrugBank ID 顺序，特征标准化仅在训练分区上进行拟合。

D 划分构建与验证

这四种方案分别考察互补的泛化形式。随机评估测试未见过的药物对组合，同时保留较高的单个药物重叠度。单药未见和双药未见评估分别要求一个或两个留出端点。骨架不相交评估则限制药物身份、规范 SMILES、Bemis-Murcko 骨架和分子相似性重叠。因此，这些方案应被视为互补的泛化测试，而非单一且严格单调递增的难度尺度。

划分规模见表 3。表 10 列出了相应的重叠情况和分配检查。

所有 47,591 个骨架不相交的验证集和测试集配对，两个端点均具有有效的分子结构。在 806,341 个训练对中，613,853 个包含两个有效结构；另有 192,488 个至少包含一个不可用或无效结构，并通过附录 C 所述的零向量特征回退予以保留。

单药未见划分的测试集规模最大，因为许多配对符合其单端点留出规则。骨架不相交划分的验证集和测试集规模较小，是由全局一致的分子划分及跨划分排除所致，而非任意子采样。表 10记录了重叠和分配检查结果。

表 10：划分套件的划分验证检查。所有设置下，训练集与测试集的配对重叠均为零。骨架不相交划分中，训练集与测试集的 DrugBank ID、规范 SMILES 和 Bemis-Murcko 骨架重叠也均为零。验证集/测试集与训练集之间的最大 ECFP4 Tanimoto 相似度为 0.70，没有任何值超过阈值。

划分	配对重叠	训练集与测试集的药物重叠	验证检查
随机	0	4,169/4,207	行级打乱，大多数测试药物在训练集中已见过
一种未见药物	0	2,995/3,874	0 行违反单个未见药物规则
双药未见	0	0/862	0 行违反双未见药物规则
骨架不相交	0	0/656	0 个规范 SMILES/骨架重叠，最大 ECFP4 Tanimoto 为 0.70

每种设置的训练集与测试集无序药物对重叠均为零。骨架不相交划分还具有零训练集与测试集 DrugBank ID 重叠，并且在结构已解析化合物中，规范 SMILES 和 Bemis-Murcko 骨架重叠均为零。分子相似性约束已在表 10 中得到验证。这些控制措施从不同于双药未见评估所单独隔离的终点新颖性的维度，检验了化学邻域泛化。

E 扩展结果

以下小节保留了支撑正文主要论文获胜结论的配置级证据。它们区分了表征效应与分类器注意力头变化，并报告补充性的基于概率的指标，同时不改变将宏平均 F1 作为主要排名指标的做法。

E.1 表现最佳的配置

表 11 报告了每个评估数据划分中排名前三的五随机种子 Torch 配置，并针对每种设置分别选择表征和注意力头。该表支持正文中的获胜结论。

表 11：每个评估数据划分中排名前三的五随机种子 Torch 配置。Morgan 对特征在随机评估中表现最佳，而 GCN 对特征在全部三个药物留出和骨架留出设置中表现最佳。对于相同表征，骨架不相交设置也是唯一一个成对注意力优于残差多层感知机的领先设置。

划分	排名	配置	Runs	准确率	宏平均 F1
随机	1	Morgan 对 +辅助正则化多层感知机	5	0.905 ± 0.000	**0.899 ± 0.000**
	2	Morgan 对 + 多层感知机	5	0.902 ± 0.001	0.897 ± 0.001
	3	Morgan 对 + 残差多层感知机	5	0.902 ± 0.000	0.896 ± 0.001
一种未见药物	1	GCN 对 + 残差多层感知机	5	0.672 ± 0.001	**0.635 ± 0.001**
	2	MoLFormer 对 +残差多层感知机	5	0.669 ± 0.001	0.632 ± 0.001
	3	Morgan 对 + 残差多层感知机	5	0.669 ± 0.003	0.631 ± 0.002
Twodrugs 未见过 1		GCN 对 +残差多层感知机	5	0.496 ± 0.005	**0.432 ± 0.004**
	2	GCN 对 +多层感知机	5	0.507 ± 0.004	0.432 ± 0.007
	3	GCN 对 +辅助正则化多层感知机	5	0.496 ± 0.002	0.426 ± 0.002
骨架不相交	1	GCN 对 +成对注意力	5	0.343 ± 0.006	**0.324 ± 0.005**
	2	GCN 对 +多层感知机	5	0.345 ± 0.004	0.319 ± 0.009
	3	GCN 对 +辅助正则化多层感知机	5	0.336 ± 0.004	0.319 ± 0.004

Morgan 占据随机设置的前列，而图卷积网络在单药未见、双药未见和骨架不相交评估中表现领先。成对注意力取得了观测到的最高骨架不相交图卷积网络结果。

E.2 表征与注意力头的配对比较

表12中的差异是使用相同的五个模型随机种子逐个计算的，因此这些区间描述的是配对配置差值，而不是独立运行比较。

表12：关键基准比较的五个随机种子宏平均F1配对差值。差值按随机种子逐一计算，为第一种配置减去第二种配置。区间为基于随机种子 11、22、33、44 和 55 的 95% 配对 t 区间。

比较	对比	划分	差值
随机顶部配置	Morgan 辅助回归与 Morgan 多层感知机	随机	0.002 [0.001，0.004]
单药表征边际	图卷积网络残差 vs MoLFormer 残差	一种未见药物	0.004 [0.002，0.005]
双药物表征边距	图卷积网络残差 vs MoLFormer 残差	双药未见	0.012 [0.004，0.020]
骨架表示差值	图卷积网络成对注意力对比 MoLFormer 残差	骨架不相交	0.015 [0.009, 0.022]
骨架头部差值	GCN 对成对注意力与 GCN 残差	骨架不相交	0.009 [-0.002，0.020]

对于单药未见评估，GCN-MoLFormer 的差值虽小但为正；对于双药未见评估，差值则更大。在骨架不相交评估中，所列的 GCN 配置优于所列的 MoLFormer 配置，因为其成对区间不包含零。相比之下，成对注意力与残差 GCN 之间的区间包含零，因此无法确定注意力头是否更优。

E.3 基于概率的诊断

除了宏平均 F1，表 13 还报告了每种情形下领先配置的平衡准确率、MCC、宏平均受试者工作特征曲线下面积、宏平均精确率-召回率曲线下面积和置信度 ECE [18]。这些指标评估类别平衡判别和置信度校准的互补方面。ECE 并非完整的认知不确定性分析。

表 13：每个评估数据划分中领先五随机种子配置的基于概率的诊断指标。AUROC 和 AUPRC 是宏平均的一对多多分类指标，由归档预测概率重新计算得出。ECE 是 15 分箱置信度校准误差。

划分	配置	平衡准确率	MCC	AUROC	精确率-召回率曲线下面积 / 期望校准误差
随机	Morgan 对 + 辅助回归多层感知器	0.888 ± 0.001	0.875 ± 0.000	0.991 ± 0.000	0.964 ± 0.000 / 0.019 ± 0.001
一种未见药物	GCN 对 + 残差多层感知机	0.624 ± 0.003	0.563 ± 0.001	0.887 ± 0.001	0.670 ± 0.002 / 0.171 ± 0.003
双药未见	GCN 对 + 残差多层感知机	0.419 ± 0.003	0.330 ± 0.007	0.758 ± 0.001	0.415 ± 0.006 / 0.318 ± 0.003
骨架不相交	GCN 对 +成对注意力	0.325 ± 0.004	0.192 ± 0.007	0.689 ± 0.005	0.310 ± 0.005 / 0.322 ± 0.005

在更严格的泛化设置下，区分能力和校准均有所下降。随机划分结果在未见药物和骨架不相交评估下无法保持，而宏平均 F1 仍是比较不平衡六分类任务的主要指标。

E.4 泛化设置比较

表 14 的每一行都比较了在药物对未见和骨架不相交评估下相同的训练设置、表征和注意力头。由于评估的药物对群体不同，该比较属于描述性比较，而不是针对相同样本进行的配对检验。

在全部 36 组匹配比较中，骨架不相交的宏平均 F1 均较低。这支持将其视为一种独立的化学泛化压力测试，而不是另一种药物留出划分。

表 14：双药未见与骨架不相交评估之间的匹配比较。每行比较不同划分中相同的训练设置、分子表示和模型头。正值 Δ表示骨架不相交的宏平均 F1 更高。负值 Δ 表示双药未见的宏平均 F1 更高。

匹配组	设置	骨架互斥表现更优	平均值 Δ	中位数 Δ
所有匹配的附录设置	36	0	-0.085	-0.093
Torch 设置	28	0	-0.093	-0.098
Morgan 对设置	9	0	-0.100	-0.122
图卷积网络对设置	9	0	-0.082	-0.098
Torch Morgan 对设置	7	0	-0.110	-0.124
Torch 图卷积网络对设置	7	0	-0.091	-0.106

E.5 分类器头摘要

表15对不同表征取平均，因此它提供的是分类头稳健性的补充视角，而非对特定表示的比较的替代。

表15：完整评估配置集中的分类器头摘要。各项数值为每种划分下四种成对表示的五个随机种子均值的平均值，因此旨在作为注意力头层面的辅助证据，而非取代特定表示的比较。

划分	Rank	Head	平均准确率	平均宏 F1
随机	1	残差多层感知机	0.880	0.874
	2	特征注意力	0.869	0.861
	3	门控多层感知机	0.868	0.860
	4	辅助正则化多层感知机	0.863	0.854
一种未见药物	1	残差多层感知机	0.668	0.630
	2	门控多层感知机	0.663	0.619
	3	特征注意力	0.660	0.616
	4	辅助正则化多层感知机	0.654	0.613
双药未见	1	残差多层感知机	0.478	0.416
	2	门控多层感知机	0.475	0.401
	3	辅助正则化多层感知机	0.465	0.398
	4	MLP	0.470	0.398
骨架不相交	1	残差多层感知机	0.321	0.302
	2	MLP	0.314	0.293
	3	特征注意力	0.315	0.293
	4	辅助正则化多层感知机	0.313	0.293

残差多层感知机是平均表现最稳定强劲的注意力头。这并不与成对注意力产生单个表现最佳的骨架不相交图卷积网络配置相矛盾，因为注意力头层面的平均表现不应取代特定表示的比较。