arXiv:2601.15763v1 [cs.CE] 22 Jan 2026

NMRGym：用于核磁共振分子结构解析的综合基准

Zheng Fang
zfang723@connect.hkust-gz.edu.cn
香港科技大学
与技术（广州）
广州，中国

Chen Yang
chenyangmiles@gmail.com
香港科技大学
（广州）
广州，中国

Haitao Yu
hyu382@connect.hkust-gz.edu.cn
香港科技大学
与技术（广州）
中国广州

Haoming Luo
wangnan8008@gmail.com
同济大学
中国上海

Haitao He
hehaitao@mail.nwpu.edu.cn
西北工业
大学
山西，中国

Jiaqing Xie
xiejiaqing@pjlab.org.cn
上海人工智能
实验室
中国上海

Zhuo Yang
yangzhuo@pjlab.org.cn
上海人工智能
实验室
中国上海

Yuqiang Li
liyuqiang@pjlab.org.cn
上海人工智能
实验室
中国上海

Jun Xia
junxia@hkust-gz.edu.cn
香港科技大学
与技术（广州）
广州，中国

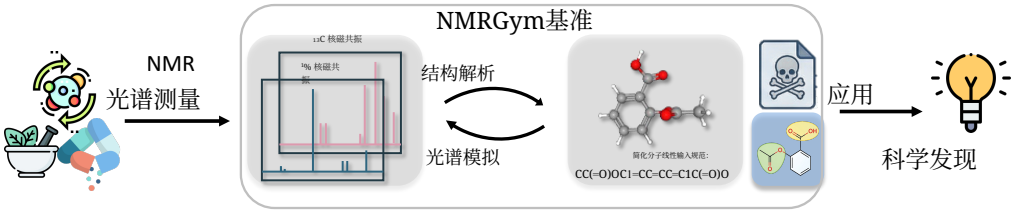


图1：NMRGym框架。一个连接实验核磁共振波谱测定与分子结构的综合数据集与基准，旨在促进结构解析和光谱模拟，从而加速科学发现。

摘要

核磁共振（NMR）波谱学是小分子结构解析的基石。尽管深度学习已展现出在自动化结构解析和光谱模拟方面的巨大潜力，但当前进展严重受制于对合成数据集的依赖；当应用于现实世界的实验光谱时，这会引入显著的领域偏移。此外，缺乏标准化的评估协议和严格的数据划分策略，常常导致不公平的比较和数据泄漏。为应对这些挑战，我们推出了**NMRGym**，这是迄今为止规模最大、最全面的标准化数据集与基准，源自高质量实验核磁共振数据。该数据集包含**269,999**个独特分子配对高保真 ^{1}氢和 ^{13}碳光谱后，NMRGym 弥合了合成近似与现实世界多样性之间的关键鸿沟。我们实施严格的质量控制流程，并统一数据格式，以确保公平比较。为严格防止数据泄漏，我们采用基于分子骨架的划分。此外，我们提供细粒度的峰-原子级标注，以支持未来的应用。借助这一资源，我们建立了一个全面的评估套件，涵盖多种下游任务，包括结构解析、基于核磁共振的官能团预测、基于核磁共振的毒性预测和光谱模拟，并对具有代表性的最先进的方法进行基准评测。最后，我们发布了一个带有自动化排行榜的开源排行榜，以促进社区协作并规范未来研究。该数据集、基准和排行榜均公开发布于 https://AIMS-Lab-HKUSTGZ.github.io/NMRGym/。

允许为个人或课堂使用而免费制作本作品的全部或部分数字副本或纸质副本，但前提是副本不得为牟利或商业利益而制作或分发，且副本必须包含本声明以及首页上的完整引文。本作品中由作者以外的其他人员拥有版权的组件必须予以尊重。经适当署名后进行摘要摘编是允许的。除上述情况外，如需复制、再版、发布到服务器或向列表重新分发，必须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 向我们申请许可。*KDD '26*，韩国济州© 2026 版权所有归所有者/作者所有。出版权授权予 ACM。ACM ISBN 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXX.XXXXXXX

CCS 概念

• **应用计算** → **生命与医学科学**；**化学**。

关键词

核磁共振、光谱学、结构解析、光谱模拟

ACM 参考格式：
Zheng Fang、Chen Yang、Haitao Yu、Haoming Luo、Haitao He、Jiaqing Xie、Zhuo Yang、Yuqiang Li 和 Jun Xia。2026 年。NMRGym：核磁共振分子结构解析综合基准。载于《第 *32* 届 *ACM SIGKDD* 知识发现与数据挖掘大会（*KDD '26*）论文集》。ACM，纽约，纽约州，美国，13 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

核磁共振（NMR）波谱学长期以来一直被视为小分子结构解析的基石，为连接性、立体化学和官能团环境提供了详细信息，而这些信息构成了现代有机分析的基础 [10, 33]。一维（1D）NMR 技术，尤其是质子（^{1}H）和碳（^{13}C）光谱学，仍然是分析化学中用于分子**结构解析**的最广泛工具[21]。传统的核磁共振结构解析流程将经验规则、基于物理学的预测和数据库匹配有机结合起来：化学家首先解读已确立的化学位移趋势和耦合模式，进而组装局部分子片段 [10, 33]。为提高这一过程的可靠性，密度泛函理论（DFT）等基于物理学的方法 [16] 经常用于**光谱模拟**。这些模拟发挥着两个关键作用：利用理论数据扩充参考数据库，并通过将预测光谱与实验观测进行严格比较来验证候选结构[31]。然而，尽管密度泛函理论具有较高的精确性，其高昂的计算成本限制了它在大规模数据库扩展中的常规应用 [9, 44]。因此，NMRShiftDB [37] 等数据库驱动工具在很大程度上依赖于高质量实验核磁共振数据的有限供应，而这类数据的获取仍然成本高昂且耗时 [14]。这些局限性共同凸显了对更具可扩展性、数据驱动方法的迫切需求。

受深度学习在分子科学领域变革性影响的启发， [22], 近期核磁共振分析研究聚焦于两项关键挑战：结构解析和 光谱模拟。在结构解析方面，方法已从传统的数据库驱动搜索 [19] 迅速发展到生成式深度学习范式。这些方法包括基于 Transformer 的序列翻译模型 [18, 38, 48]，用于解读光谱模式；以及新兴的基于扩散的方法 [46, 49] ，能够生成分子图或三维构象。与此同时，正向光谱模拟也从依赖固定向量描述符 [5] 发展到复杂的二维 [20, 23, 26] 和三维 [20, 47, 50] 几何表示，从而能够更准确地模拟化学环境。

尽管取得了进展，该领域仍缺乏基于高质量实验数据构建的标准化基准。这一缺陷给未来发展带来了三个关键障碍：**1. 领域偏移**。 目前大多数方法主要在合成数据集 [2, 15, 18] 上进行验证。因此，在此类理想化数据上训练的模型应用于真实世界的实验光谱 [46] 时，往往会遭遇严重的性能下降。**2. 评估协议不一致**。 光谱元数据存在异质性——从详细的峰属性到稀疏的化学位移列表不等——迫使模型依赖多种不同的输入格式。这种缺乏标准化的情况不仅限制了模型的广泛泛化能力，也使得不同方法之间无法进行公平的定量比较。**3. 数据泄漏风险**。 由于缺乏严格的划分协议，结构相似的分子常常会同时出现在训练集和测试集[45]中。因此，有必要采用骨架划分，以避免指标虚高，并准确评估真实的泛化能力。尽管 NMRShiftDB [37] 等实验数据集已经存在，但其用途受到数据规模的严重限制，无法满足现代深度学习方法对数据量的巨大需求。

为解决这些问题，我们推出了**NMRGym**，这是迄今为止最全面、标准化程度最高的实验核磁共振数据集和基准。通过精心设计的质量控制（QC）流程统一不同的光谱格式，我们整理出**269,999**组高保真实验分子-光谱对，为在复杂的真实实验场景中严格评估模型提供了强大的资源。最后，我们采用基于分子骨架的划分，严格防止数据泄漏，并利用这一严谨框架，对涵盖四项下游任务的代表性最先进方法进行系统基准测试，为未来研究建立可靠的基线。

O我们的主要贡献总结如下 s:

- **大规模实验数据**。 我们发布了目前规模最大的标准化实验核磁共振数据集，包含 269,999 个分子。该资源提供高保真的真实世界光谱，而非合成近似数据，有效缩小了领域差距，满足了现代深度学习模型对数据的大量需求。

- **严谨且公平的基准**。 我们建立了一套全面的评估套件，涵盖四项下游任务（结构解析、基于核磁共振的官能团预测、基于核磁共振的毒性预测和光谱模拟）。尤为重要的是，我们引入了基于分子骨架的划分，以评估分布外（OOD）鲁棒性，并提供细粒度的峰-原子级标注，助力开发具有物理可解释性的模型。

- **交互式评估排行榜**。 我们开发了一个开源排行榜，配备自动化排行榜和可视化工具。该排行榜简化了模型提交和评估流程，促进社区协作，并支持对最先进方法进行直接、透明的比较。

2 相关工作

2.1 核磁共振结构解析

结构解析的首要目标是将实验核磁共振谱逆向映射为分子结构。早期的数据驱动方法将结构解析重新定义为序列到序列的翻译任务。这些模型的一个关键区别在于其光谱编码策略。例如，CLAMS [38] 将核磁共振谱处理为二维图像，并利用卷积神经网络提取视觉特征。相比之下，NMRFormer [2] 采用 1D-CNN 架构，自动学习光谱分词。为实现更明确的表示，NMR2Struct [18] 和 NMRMind [48] 等模型通过编码特定化学位移范围及其对应的信号强度，手动对光谱进行分词，使 Transformer 能够关注离散的光谱输入。

表1：常用核磁共振数据集概览。"Exp."表示光谱是否来自现实世界实验。"Assign."表示是否提供峰级归属。

数据集	# 分子数	划分	Exp.	分配。
密度泛函理论模拟数据集				
QM9-NMR [15]	~ 131k	随机。	✗	✗
SpectraBase-Mnova [18]	~ 143k	随机	✗	✗
Multispec [1]	~ 790k	N.A.	✗	✗
Pistachio-Mnova [2]	~ 1.03M	N.A.	✗	✗
Mind-Gaussian [48]	~ 2.25M	随机	✗	✗
机器学习模拟数据集				
NN-NMR [29]	16 k	随机	✗	✗
ShiftML [8]	~ 257k	随机	✗	✗
Pubchem-NMRNet [47]	~ 106M	N.A.	✗	✓
实验数据集				
NMRShiftDB [37]	43,580	随机	✓	✓
NMRBank [39]	149,135	N.A.	✓	✗
NMRGym（我们的）	**269,999**	骨架	✓	✓

为解决自回归解码中常见的化学无效性问题，研究者提出了 NMR-Solver [19] 等基于搜索的框架。这些方法将神经表示与严格的化学约束先验相结合，并利用 NMRNet [47], 等大规模模拟数据库进行基于相似度的检索和结构优化。近期，受生成式建模领域突破的启发，[17, 30, 32, 40], 该领域已转向基于扩散的范式。DiffNMR [49]采用扩散框架，逐步去噪并生成分子二维图。与此同时，ChefNMR [46] 直接生成三维原子构象，从而捕捉分子结构固有的几何对称性。

2.2 核磁共振光谱模拟

光谱模拟，又称正向问题，旨在直接根据分子结构预测核磁共振谱，从而避免量子力学计算的高昂计算成本。这些模拟器的效能很大程度上取决于其分子表示方式。早期研究采用基于向量的表示 [5, 11, 28], ，将定长描述符映射到化学位移或耦合常数。为更好地反映有机分子的拓扑连接性，研究人员转向基于二维图的表示 [20, 23, 26]。这些模型利用图神经网络或消息传递神经网络（MPNN）来捕捉每个原子的局部化学环境。近期，最先进的方法已转向三维分子表示 [12, 47, 50], ，其中纳入了显式的原子坐标。通过利用三维感知图神经网络，这些模型能够考虑长程空间相互作用，这对于实现密度泛函理论级精度的化学位移预测至关重要。

2.3 核磁共振数据与基准

尽管核磁共振机器学习在架构方面取得了进展，但可用数据集的匮乏、异质性以及缺乏标准化，仍然严重制约着这一领域。如表 1 所示，当前研究通常依赖两类数据来源：合成数据集和实验数据集。合成数据集通过密度泛函理论或基于机器学习的模拟器生成，例如 QM9-NMR [15],Multispec [1], 和 Pubchem-NMRNet [47] ，包含超过 1 亿条记录，为训练深度神经网络提供了必要的数据规模。然而，这些模拟谱往往无法捕捉真实采集过程中的细微特征，例如溶剂效应、基线噪声和复杂的耦合模式。因此，基于这些合成先验训练的模型常常会遭遇显著的领域偏移，导致其在实验样本上应用时泛化能力较差 [46]。相比之下，NMRShiftDB [37] 等实验数据集最能真实反映化学现实，但规模往往有限。

3 方法

3.1 预备知识

我们将核磁共振谱 S 正式定义为由 N个峰组成的无序集合，记作 $S = \{s_1, s_2, \ldots\ldots, s_N\}$，其中每个 $s_i \in \mathbb{R}$ 都表示化学位移值（单位为百万分之一）。详细的化学定义见附录A。由于核磁共振峰的物理测量结果不受置换影响，我们严格将 S 视为集合，而非序列。相应的分子结构以简化分子线性输入规范 [41] 序列 $Y = \{y_1, \ldots\ldots, y_L\}$ 表示。最后，图2展示了所提出的数据集及基准构建流程的概览。

3.2 数据收集与预处理

数据采集。 为构建全面且多样化的基准，我们汇集了来自广泛公共化学数据库的实验核磁共振谱。我们的主要数据来源包括 CH-NMR-NP [4], CASCADE [13], SDBS [36], PubChem [24], NMRShiftDB [37],HMDB [42], NP-MRD [43], 和 NMRMind [48]。数据采集过程采用混合方式，结合了公共 API、数据库批量转储以及受文献 [48] 中方法启发编写的文献挖掘脚本。这种多来源策略确保数据集覆盖广阔的化学空间，范围从简单的有机分子到复杂的天然产物。

质量控制（QC）。我们使用 RDKit[27]为所有分子生成规范化的简化分子线性输入规范（SMILES）[41]，并将其作为唯一标识符，以检测和删除不同来源中的重复条目。此前研究表明，协同使用碳谱和质子谱能显著提升模型性能 [3, 48],，因此我们严格筛选数据集，仅保留同时具有配对 ^{1}H 和 ^{13}C 核磁共振谱的样本。为确保数据集与有机化学研究相关，我们根据元素组成筛选分子，仅保留含有以下重原子的分子：$\{C, H, O, N, F, P, S, Cl, Br, I\}$。此外，不同于通常会限制序列长度的标准小分子数据集，我们特意保留了序列较长的复杂天然产物（具体来自 NP-MRD 数据集）。纳入这些样本对于评估该基准在重原子结构和大环化合物上的泛化能力至关重要。

数据格式标准化。 由于报告标准不一致，要统一异构来源的核磁共振数据格式面临重大挑战。为确保数据严格统一，我们处理了两项主要的不一致问题。首先，关于 J 耦合常数，我们排除了这一属性，这不仅是因为它在公开数据集中较为稀缺，也因为当前的模型架构通常并非为处理这种特定模态而设计。

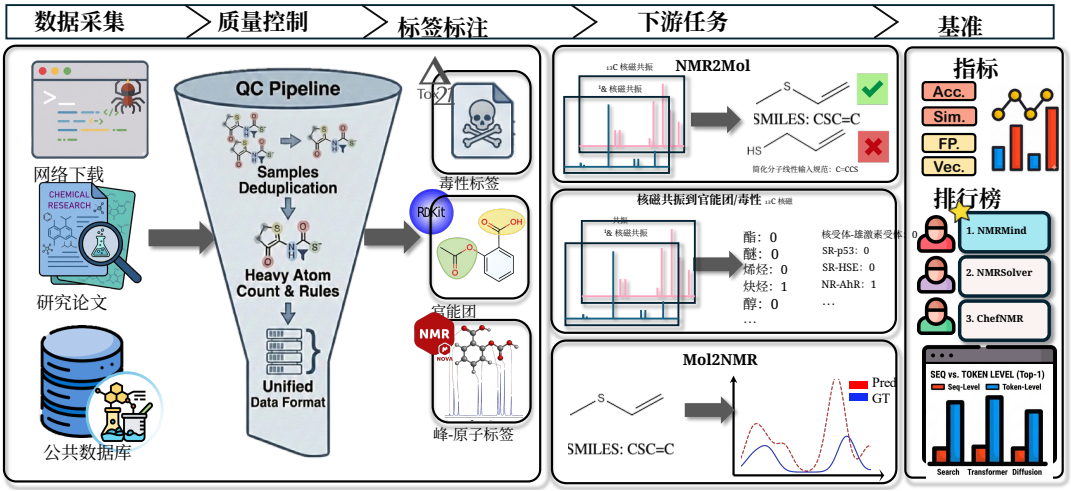


图2：NMRGym 数据整理工作流概览。该流程整合了从异构来源进行的数据采集、用于标准化的严格质量控制，以及全面的标签标注（包括毒性、官能团和峰-原子归属），以支持下游生成和预测任务。

其次，我们处理了峰强度问题。在 ^{1}H 核磁共振中，通常通过对信号积分来推断质子化学计量。我们没有强迫模型学习如何将不一致的原始强度值复杂地映射为离散原子数，而是采用了基于多重性的编码策略。具体而言，化学位移出现的频次隐式地表示了强度；例如，若某个信号对应两个质子（2H），就会在输入序列中将其化学位移值重复两次进行标记。这种方法将质子数约束明确嵌入统一的基于序列的格式中，有效减轻了模型从可变强度输入中解码化学计量信息的负担。

标签标注。 除谱图-SMILES 配对外，NMRGym 还提供多种标注，以支持各种下游任务。*1.* 官能团。 遵循 NMR-Former [2], 中确立的方案，我们使用 RDKit [27] 通过 SMARTS 子结构匹配来识别官能团。附录 C 详细列出了所追踪官能团的完整定义。
2. 毒性。为了展示该数据集在实际安全性评估中的实用价值，我们纳入了来自 Tox21 [35] 的毒性标签。鉴于原始 Tox21 数据集规模有限，我们发布了一个有针对性的子集，专门用于评估实际应用中的分子毒性预测。12 项毒性测定的详细定义见附录 D。

3. 峰-原子归属。为了促进可解释性和结构验证——这是解析工作流中的关键环节——我们使用 Mnova [7] 软件生成了峰-原子级归属。通过提供这些详细标注，我们旨在支持未来围绕可解释验证开展的研究。

骨架划分。 为了严格评估模型对未见化学结构的泛化能力，我们刻意避免采用会导致训练集和测试集之间存在结构冗余的随机划分。相反，我们采用了受近期基准测试协议启发的骨架划分策略 [6]。我们使用了 DeepChem [34] 提供的 Bemis-Murcko 骨架分解算法，根据分子的核心结构框架对分子进行聚类。共享同一骨架的分子被严格分配到同一子集中。该策略确保测试集中的分子具有与训练期间见过的分子截然不同的分子骨架，从而严格评估分布外泛化能力。详细分析见第 4.2 节。

3.3 任务定义与评估

任务 1：结构解析（NMR2Mol）。 该任务旨在根据输入光谱 S，重构以简化分子线性输入规范序列 $Y = \{y_1, \ldots, y_L\}$ 表示的分子结构。模型的总体目标是最大化条件概率 $P(Y|\mathrm{S})$。此外，对于支持显式化学约束的方法，我们将分子式 F 作为辅助条件纳入其中。在这种情况下，目标调整为最大化联合条件概率 $P(Y|\mathrm{S}, \mathrm{F})$。

目标。

指标。 我们使用三个指标评估生成质量。首先，Top-K完全匹配准确率用于衡量真实 SMILES Y是否出现在排名前-K的候选项中 $\{\hat{}_{Y^{(1)}}, \ldots, \hat{}_{Y^{(K)}}\}$:

$$\mathrm{Acc@}K = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(Y_i \in \{\hat{Y}_i^{(1)}, \ldots, \hat{Y}_i^{(K)}\}). \tag{1}$$

其次，词元级准确率用于计算每个位置上正确预测的词元占真实值序列中对应词元的比例。第三，我们使用三种不同的分子指纹[27]评估结构相似度：Morgan、拓扑扭转和原子对。所有指纹均使用 RDKit 库生成，表示为 2048 位向量。设$\mathbf{v}, \hat{}\mathbf{v} \in \{0,1\}^D$ 分别表示真实分子和预测分子的二进制指纹向量。我们报告 Tanimoto 相似度，其定义为活跃位的交集与并集之比：

$$\mathrm{Sim}_{\mathrm{Tanimoto}}(\mathbf{v}, \hat{\mathbf{v}}) = \frac{\mathbf{v} \cdot \hat{\mathbf{v}}}{\|\mathbf{v}\|^2 + \|\hat{\mathbf{v}}\|^2 - \mathbf{v} \cdot \hat{\mathbf{v}}}. \tag{2}$$

表 2：各方法在序列级和词元级结构解析任务中的性能比较（NMR2Mol）。分子式约束 **表示将分子式用作结构约束。**

	前1名（% ↑）		前5名（% ↑）		前10名（% ↑）	
模型	序列级	词元级	序列级	词元级	序列级	词元级
基于搜索的方法						
NMR-Solver（仅搜索）[19]	6.27 ± 0.00	21.75 ± 0.00	13.96 ± 0.00	34.78 ± 0.00	17.37 ± 0.00	40.13 ± 0.00
+公式条件	**17.92 ± 0.00**	27.64 ± 0.00	**33.97 ± 0.00**	41.06 ± 0.00	**36.48 ± 0.00**	42.79 ± 0.00
基于 *Transformer* 的方法						
CLAMS[38]	0.00 ± 0.00	7.90 ± 0.07	0.00 ± 0.00	12.78 ± 0.13	0.00 ± 0.00	14.70 ± 0.14
NMRFormer[2]	1.75 ± 0.04	22.49 ± 0.02	2.81 ± 0.05	29.58 ± 0.02	3.30 ± 0.03	32.32 ± 0.01
NMR2Struct[18]	0.24 ± 0.08	21.97 ± 0.37	1.05 ± 0.19	33.65 ± 0.55	1.85 ± 0.27	38.67 ± 0.63
NMRMind[48]	11.75 ± 0.17	36。96 ± 0.01	23.22 ± 0.05	54.71 ± 0.00	27.00 ± 0.15	59.08 ± 0.01
+公式条件	15.49 ± 0.08	**40.82 ± 0.06**	29。82 ± 0.09	**60。00 ± 0.00**	34.03 ± 0.09	**64.40 ± 0.11**
基于扩散的方法						
DiffNMR[49]	0.00 ± 0.00	17.37 ± 0.02	0.00 ± 0.00	25.37 ± 0.01	0.00 ± 0.00	28.49 ± 0.02
ChefNMR-S[46]	0.02 ± 0.00	3.18 ± 0.01	0.04 ± 0.01	8.93 ± 0.00	0.05 ± 0.01	11.4 ± 0.03
ChefNMR-S（微调）[46]	1.69 ± 0.02	11.87 ± 0.03	3.87 ± 0.03	23.85 ± 0.07	5.04 ± 0.04	28.18 ± 0.09
ChefNMR-L[46]	0.02 ± 0.00	3.70 ± 0.02	0.07 ± 0.01	10.16 ± 0.00	0.08 ± 0.01	13.29 ± 0.02
ChefNMR-L（微调）[46]	1.93 ± 0.02	13.59 ± 0.14	4.36 ± 0.08	25.92 ± 0.01	5.66 ± 0.01	30。07 ± 0.02

我们还计算余弦相似度，以衡量高维指纹向量之间的取向对齐：

$$\mathrm{Sim}_{\mathrm{Cosine}}(\mathbf{v},\hat{\mathbf{v}}) = \frac{\mathbf{v}\cdot\hat{\mathbf{v}}}{\|\mathbf{v}\|_2\|\hat{\mathbf{v}}\|_2}. \tag{3}$$

对于基于 Transformer 的模型，我们采用束搜索，通过迭代扩展概率最高的部分词元来近似最优序列。对于基于扩散的模型，我们采用 Top-K采样策略，直接通过反向扩散过程生成一批 K 个候选结构。最后，基于搜索的方法根据与输入 S 的相似度指标 [19] 检索分子。

T任务 2：属性预测（NMR2FG 和 NMR2T ox).

目标。我们将官能团预测和毒性预测表述为多标签二分类任务。给定光谱 S，目标是预测一个二元标签向量 $\mathbf{y}\in\{0,1\}^C$，涵盖 C属性。指标。按照标准基准，我们报告宏观/微观 F1、召回率和准确率，以评估各类别及整体表现，并考虑标签不平衡问题。我们还报告子集准确率，这是最严格的指标，要求给定样本的所有 C 标签均被正确预测：

$$\mathrm{Acc}_{subset} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(\mathbf{y}_i = \hat{\mathbf{y}}_i). \tag{4}$$

任务 3：光谱模拟（Mol2NMR）。

目标。此任务聚焦于从分子结构正向预测光谱峰。与以往受限于特定输入的方法不同，我们将输入泛化为分子表示 M（例如，二维分子图或三维构象）。目标是预测化学位移集合 $\hat{S} = \{\hat{s}_1, \ldots\ldots, \hat{s}_{\hat{N}}\}$，使其尽可能接近真实光谱 S。

指标。为评估光谱保真度，我们采用双指标方案：使用向量相似度进行全局检索，使用集合相似度进行细粒度评分，方法如文献所述于 [19]。对于向量相似度，我们使用带宽为 σ_{vec} 0.3 百万分之一（^{1}H）和 2.0 百万分之一（^{13}C）的高斯核，将离散峰平滑为连续向量。然后计算余弦相似度，以衡量光谱包络的重叠程度：

$$\mathrm{Sim}_{\mathrm{vec}}(\mathbf{v},\hat{\mathbf{v}}) = \frac{\mathbf{v}\cdot\hat{\mathbf{v}}}{\|\mathbf{v}\|_2\|\hat{\mathbf{v}}\|_2}. \tag{5}$$

对于集合相似度，我们通过将比较表述为二分图匹配问题，重点关注峰级精确率。使用匈牙利算法 [25], ，我们找到使预测值与真实值的峰之间位移最小的最优置换 π。相似度得分按容差 σ_{set} 进行归一化（ ^{1}H 为 1.0 百万分之一， ^{13}C 为 10.0 百万分之一）：

$$\mathrm{Sim}_{\mathrm{set}}(\mathcal{S},\hat{\mathcal{S}}) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\frac{|s_i - \hat{s}_{\pi(i)}|}{\sigma_{\mathrm{set}}}\right). \tag{6}$$

4 实验

4.1 基准

NMR2SMILES。为确保比较严谨且公平，我们按照各基线方法的官方实现重新进行了复现。表 2 和图 3 展示了基于搜索、基于 Transformer 和基于扩散的范式之间的比较分析。在基于 Transformer 的方法中，NMRMind 为生成模型确立了最先进的水平，在分子式约束下实现了 15.49% 的 Top-1 序列准确率。更重要的是，它展现出卓越的结构理解能力，在指纹相似度指标上显著优于所有基线方法（例如，Top-1 Morgan 相似度达到 65.51%，而最佳的基于搜索的基线方法为 43.31%）。我们将这种稳健性归因于两个因素：在约 5000 万个合成数据点上进行的大规模预训练 [48]，以及引入二维相关谱（例如，异核单量子相干谱、同核相关谱），这使模型能够捕捉比仅使用一维谱的方法更丰富的结构连接性。相比之下，早期的深度

表 3：使用 Tanimoto 相似度评估 Morgan、扭转和原子对指纹的 Top-K 相似度。（NMR2Mol）

	Morgan 相似度（% ↑）			扭转相似度（% ↑）			原子对相似度（% ↑）		
	前1名	前5名	前10名	前1名	前5名	前10名	前1名	前5名	前10名
基于搜索的方法									
NMR-Solver（仅搜索）[19]	33.28 ± 0.00	42.88 ± 0.00	46.84 ± 0.00	34.63 ± 0.00	45.29 ± 0.00	49.50 ± 0.00	45.65 ± 0.00	55.48 ± 0.00	59.17 ± 0.00
+公式条件	43.31 ± 0.00	45.70 ± 0.00	46.11 ± 0.00	44.00 ± 0.00	46.18 ± 0.00	46.54 ± 0.00	46.00 ± 0.00	47.71 ± 0.00	48.01 ± 0.00
基于 *Transformer* 的方法									
CLAMS[38]	0.72 ± 0.12	2.24 ± 0.39	3.49 ± 0.73	0.22 ± 0.07	0.75 ± 0.17	1.20 ± 0.26	0.58 ± 0.11	2.10 ± 0.41	3.51 ± 0.80
NMRFormer[2]	27.87 ± 0.01	40.94 ± 0.06	44.85 ± 0.04	29.11 ± 0.01	43.52 ± 0.07	47.92 ± 0.05	33.25 ± 0.05	49.44 ± 0.04	53.98 ± 0.04
NMR2Struct[18]	35.18 ± 0.71	43.43 ± 0.93	47.47 ± 0.91	32.79 ± 0.83	42.64 ± 0.98	47.28 ± 0.97	39.54 ± 0.63	48.26 ± 0.74	52.16 ± 0.71
NMRMind[48]	61.14 ± 0.12	67.83 ± 0.10	70。23 ± 0.10	63.47 ± 0.09	70.57 ± 0.09	73.06 ± 0.07	69.56 ± 0.07	75.65 ± 0.06	77.67 ± 0.06
+公式条件	**65.51 ± 0.03**	**72.80 ± 0.10**	**75.07 ± 0.08**	**67.94 ± 0.03**	**75.63 ± 0.06**	**77.99 ± 0.07**	**74.86 ± 0.01**	**81.15 ± 0.04**	**82.92 ± 0.03**
基于扩散的方法									
DiffNMR[49]	9.05 ± 0.02	13.33 ± 0.01	14。67 ± 0.02	6.44 ± 0.01	12.22 ± 0.02	14.40 ± 0.01	20.48 ± 0.02	28.95 ± 0.01	31.02 ± 0.02
ChefNMR-S[46]	2.65 ± 0.02	7.04 ± 0.02	9.10 ± 0.02	2.33 ± 0.02	6.85 ± 0.03	9.26 ± 0.02	6.86 ± 0.04	18.15 ± 0.05	23.26 ± 0.10
ChefNMR-S（微调）[46]	17.17 ± 0.07	33。81 ± 0.12	39.42 ± 0.15	18.72 ± 0.08	37.36 ± 0.17	43.79 ± 0.18	26.18 ± 0.09	49.68 ± 0.17	56.40 ± 0.19
ChefNMR-L[46]	3.14 ± 0.02	8.18 ± 0.03	10.45 ± 0.02	2.82 ± 0.02	7.99 ± 0.03	10.68 ± 0.02	7.98 ± 0.01	20.59 ± 0.03	26.05 ± 0.00
ChefNMR-L（微调）[46]	20.29 ± 0.12	37.62 ± 0.00	42.91 ± 0.04	22.15 ± 0.04	41.57 ± 0.03	47.55 ± 0.04	30.68 ± 0.19	53.97 ± 0.10	59.72 ± 0.07

表 4：多标签官能团预测模型（NMR2Func）对比。

模型	宏观准确率（%）	微观准确率（%）	宏观召回率（%）	微平均召回率（%）	宏平均 F1（%）	微平均 F1（%）	准确率（%）
经典分类方法							
CNN	91.95 ± 1.49	91.95 ± 1.49	51.45 ± 0.00	82.67 ± 0.00	38.14 ± 7.16	77.42 ± 1.72	25.37 ± 2.10
MLP	92.11 ± 1.58	92.11 ± 1.58	50。63 ± 0.00	81.88 ± 0.00	38.81 ± 5.19	77.63 ± 2.00	28.18 ± 0.71
XGBoost	**94.85 ± 0.00**	**94.85 ± 0.00**	45.02 ± 0.00	81.00 ± 0.00	**55.47 ± 0.00**	**85.92 ± 0.00**	**45.68 ± 0.00**
随机森林	93.11 ± 0.01	93.11 ± 0.01	41.97 ± 0.03	77.72 ± 0.05	48.64 ± 0.00	81.39 ± 0.03	36.14 ± 0.16
最先进方法							
CLAMS[38]	91.20 ± 2.27	91.20 ± 2.27	**51.89 ± 0.00**	80.52 ± 0.00	39.37 ± 3.39	75.79 ± 3.68	21.48 ± 6.77
NMR2Struct[18]	92.23 ± 0.09	92.23 ± 0.09	58.62 ± 1.12	**83.24 ± 0.49**	55.44 ± 0.03	81.69 ± 0.05	36.25 ± 0.18

表5：毒性预测模型（NMR2Tox）对比。

模型	宏观准确率（%）	微观准确率（%）	宏观召回率（%）	微平均召回率（%）	宏平均 F1（%）	微平均 F1（%）	准确率（%）
MLP	75.88 ± 6.31	75.88 ± 6.31	00.00 ± 0.00	00.00 ± 0.00	18.76 ± 0.93	21.55 ± 1.88	28.45 ± 9.80
CNN	84.22 ± 0.24	84.22 ± 0.24	00.00 ± 0.00	00.00 ± 0.00	**20.58 ± 1.31**	**26.42 ± 0.70**	32.08 ± 0.86
XGBoost	**91.91 ± 0.05**	**91.91 ± 0.05**	**4.76 ± 0.31**	**4.41 ± 0.00**	7.68 ± 0.74	7.94 ± 0.04	50.77 ± 0.52
随机森林	91.77 ± 0.08	91.77 ± 0.08	2.01 ± 0.00	2.22 ± 0.00	3.22 ± 0.13	4.07 ± 0.04	**51.19 ± 0.52**

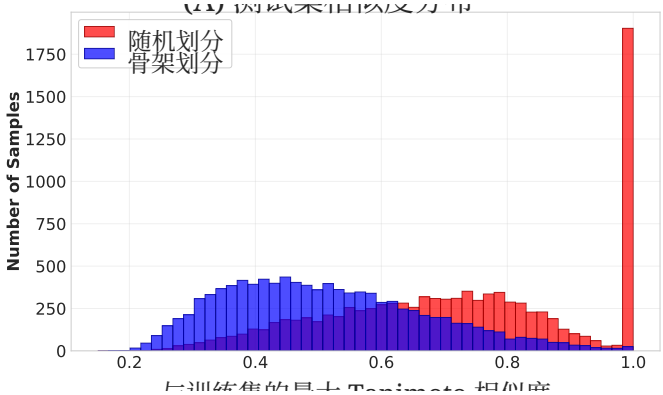
学习基线方法存在明显的架构局限。CLAMS 将核磁共振谱视为密集的二维图像，无法从稀疏峰中提取有效特征（序列准确率为 0.00%）；而 NMR2Struct 则因解码过程中缺乏结构约束，导致语法无效输出的比例很高。

对于基于搜索的基线模型 NMR-Solver，我们在主要比较中明确报告其搜索模式的性能。对于完整测试集评估，我们未采用迭代优化模块，因为其计算成本较高——平均每个样本需要 8 分钟——处理整个测试数据集将耗时一个多月。因此，我们在一个子集上进行了完整的优化评估，结果详见附录 E.2。即使在仅搜索模式下，NMR-Solver 也表现出很强的竞争力（Top-1序列准确率为 17.92%），这得益于其使用了包含超过 1.06 亿种化合物的大型数据库 [19]。

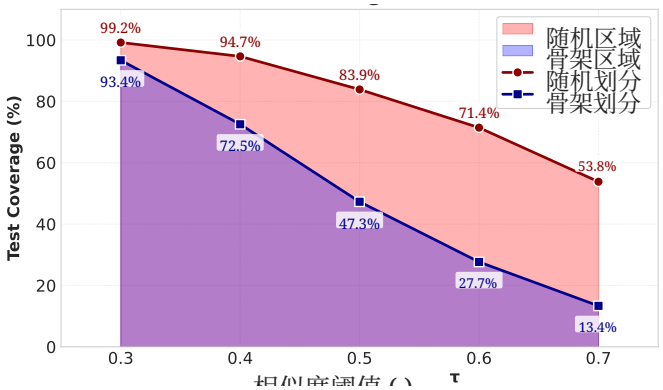
对于基于扩散的方法，DiffNMR 无法生成完全匹配的结果（准确率为 0.00%），但能捕捉到一些结构基序。

我们在零样本和微调设置下对 ChefNMR 进行了评估。鉴于其据称在合成基准测试中达到最先进水平的表现 [18], ，我们首先测试了零样本设置。然而，与作者讨论的局限性一致 [46], ，合成数据与真实世界数据之间显著的领域差距严重限制了其能力，导致序列级准确率接近于零（0.02%）。对 ChefNMR 进行微调仅带来微幅提升（ChefNMR-L 达到 1.93%），但其表现仍显著落后于基于 Transformer 和基于搜索的范式。我们认为，这是由于扩散模型收敛效率低下，通常需要过多的 ~10, 000 个训练轮次才能达到稳定状态 [46]。

NMR2Func。鉴于专门针对从核磁共振谱中识别官能团的深度学习方法十分有限，我们采用标准神经网络架构（卷积神经网络、多层感知机）和经典机器学习算法（XGBoost、随机森林），建立了一个综合基准。附录 B 中提供了详细的超参数和架构配置。



(a) 相似度分布



(b) 覆盖率@τ 曲线

图3：随机（红色）划分与骨架（蓝色）划分之间的数据泄漏比较。（a）测试集与训练集之间最大Tanimoto相似度的分布。骨架划分显示出明显向较低相似度偏移的趋势。（b）Coverage@τ曲线，衡量测试分子中在训练集中具有结构邻居（相似度为 $\geq \tau$）的分子比例。

如表 4 所示，XGBoost 表现稳健，子集准确率（45.68%）和 Micro-F1 得分均为最高。然而，所有模型的 Micro 指标与宏观指标之间都存在显著差异，这凸显了该任务固有的类别不平衡问题，反映出现实世界化学数据的长尾分布特征。尽管面临这一挑战，NMR2Struct 仍取得了颇具竞争力的 Macro-F1 得分（55.44%），表明基于 Transformer 的骨干网络能够有效捕捉不同官能团的光谱特征。

NMR2Tox。我们使用同一组通用基线方法评估了毒性预测（参见附录 B）。表5中的结果凸显了这一任务的极端困难，而以下两个现实约束进一步加剧了这一困难：（1）严重的类别不平衡；（2）数据集资源匮乏的低资源特性（该数据集是主语料库的一个子集）。尽管 XGBoost 和随机森林等集成方法取得了较高的名义准确率（> 91%），但这一指标具有误导性；它们几乎为零的召回率表明，这些方法无法识别少数类毒性样本，这是不平衡少样本场景中的常见问题。相比之下，卷积神经网络基线达到了

表6：光谱模拟模型的性能比较。* 表示闭源商业软件。

方法	覆盖率	集合相似度	向量相似度
^{1}氢核磁共振预测			
Mnova*[7]	99.91%	**0.8790**	0.8443
级联（DFTNN）[13]	97.96%	0.6899	0.8692
DetaNet[50]	85.59%	0.6871	0.8632
NMRNet[47]	**100.0%**	0.7046	**0.8785**
^{13}C 核磁共振预测			
Mnova*[7]	99.41%	0.7268	**0.8948**
级联（DFTNN）[13]	97.96%	0.4474	0.3303
级联（ExpNN-DFT）[13]	97.96%	0.1466	0.1232
级联（ExpNN-FF）[13]	97.96%	0.0330	0.0597
DetaNet[50]	85.59%	0.7784	0.6417
NMRNet[47]	**100.0%**	**0.9080**	0.8492

* 闭源/商业方法

最先进的 Macro-F1（20.58%）和 Micro-F1（26.42%），显著优于基于树的方法。这表明，在数据稀缺且类别不平衡的情况下，深度学习特征提取器在检测细微毒性信号方面具有更好的泛化能力。

Mol2NMR。表6对正向光谱模拟方法与 Mnova 进行了比较分析。Mnova 是商业领域的行业标准，采用由机器学习、HOSE码和基于增量的方法等算法组成的混合集成[7]。所有数据驱动模型均以零样本方式在实验基准上进行评估。

在开源方法中，NMRNet展现出更强的鲁棒性，化学空间覆盖率达到100%，而DetaNet的覆盖率有限，仅为85.59%。我们认为，NMRNet的成功归功于其整合了三维构象信息，并进行了大规模自监督几何预训练，使其能够捕捉对于高保真光谱合成至关重要的细微立体化学效应（在 ^{13}C集相似度方面优于Mnova：0.9080 vs. 0.7268）。相比之下，我们的架构分析揭示了DetaNet的明显局限。尽管采用了先进的E(3)等变张量架构，DetaNet仍依赖固定词表的原子嵌入，受预先设定的最大原子序数限制（例如，仅使用*C,H,O, N, F* 原子类型进行训练）。观察到的覆盖不完整，主要归因于这一设计选择以及天然产物固有的结构复杂性；这些分子复杂的重原子骨架和较高的立体化学密度，常导致DetaNet在推理过程中出现计算收敛失败。

最后，我们强调表现最佳的开源模型 NMRNet 与商业标准 Mnova 之间一个关键的计算差异。Mnova 的混合算法主要受 CPU 限制。因此，在分析复杂天然产物时，计算延迟会显著增加，因为 HOSE码 搜索和基于规则的增量所带来的组合复杂度难以应对复杂的骨架。相比之下，NMRNet 利用针对 GPU 加速优化的深度学习架构，从而实现大规模并行推理，使得

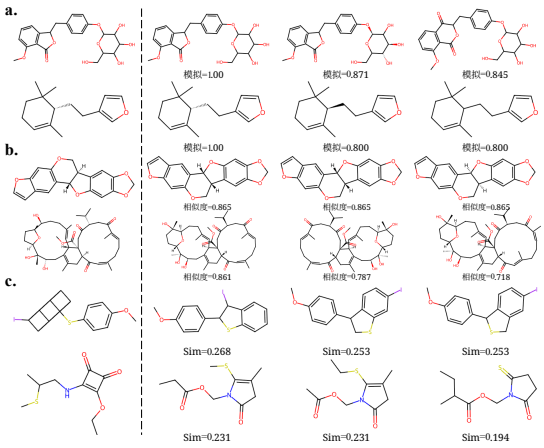


图 4：结构解析结果的定性可视化。我们将结果分为三种代表性情形：(a) 准确解析。**(b)** 高相似度偏差。**(c)**低相似度失败。**注：“Sim”表示使用 Morgan 指纹计算的 Tanimoto 相似度。**

即使对于结构复杂、通常会成为传统基于 CPU 的工作流瓶颈的分子，也能快速进行光谱预测，同时保持较高的计算效率。

4.2 主要结果

数据泄漏分析。为严格评估数据泄漏程度，我们对随机划分和骨架划分进行了定量比较。由于在完整数据集上计算高维成对指纹的计算成本过高，我们从测试集中随机抽取了 1,000 个样本作为代表性子集进行分析。我们计算了每个样本与训练集之间的最大 Tanimoto 相似度（MTS），并特别指出，MTS 为 1.0 的实例通常对应于立体异构体（例如对映异构体）：它们具有相同的二维图连接关系，但三维空间排列不同，而非完全重复。如图 3 所示，随机划分（红色）表现出较高的结构冗余，平均相似度为 0.72，且覆盖率广泛（在 $\tau = 0.5$处为 83.9%）。相比之下，骨架划分（蓝色）导致分布显著向低相似度偏移（$\Delta \approx 0.21$），覆盖率也大幅下降（在 $\tau = 0.5$处为 47.3%），这表明我们的基准有效地减少了结构重叠。

4.3 可视化

为了深入了解模型除整体指标之外的能力，我们在图 4 中展示了表现最佳的深度学习模型 NMRMind [48]，的代表性结构解析结果。生成的分子通过 Morgan 指纹与真实值之间的 Tanimoto 相似度进行评估。**准确解析（面板 a）**。如图 4(a) 所示，NMR- Mind 展现出重建复杂分子骨架的卓越能力。在这些成功案例中，模型能够有效地将光谱峰与正确的原子类型和连接性相匹配，实现完美匹配（相似度 = 1.0）。

高相似度偏差（面板 b）。图 4(b) 展示了 NMRMind 生成的分子，这些分子具有较高的指纹相似度（相似度 > 0.8），但其精确结构并不正确。在这些实例中，模型成功识别出主要的分子骨架（例如稠合环系），但可能会错置特定的官能团取代。这表明，即使 NMRMind 偶尔难以进行细粒度的原子定位，它仍已有效学习到从光谱到子结构的语义映射。

低相似度失败案例 在图 4(c) 中，我们观察到了一些以低相似度分数为特征的失败案例。这些错误出现在具有高度柔性的 大环化合物或处于异常电子环境中的分子中，此时光谱信号较为模糊。这凸显了区分具有相似光谱的结构所面临的挑战。

5 讨论

我们的综合基准揭示了当前人工智能方法在核磁共振光谱分析中的能力与显著局限。尽管基于 Transformer 的模型已经取得了新的最先进结果，但要实现实用化自动化，仍有若干关键挑战需要解决，以弥合这一差距。**模拟与现实差距**。 我们的实验始终呈现出一个趋势：合成训练集与现实世界基准之间的领域差距会导致性能下降。然而，这并不否定合成数据的价值；恰恰相反，鉴于获取成对实验数据集的成本高昂，高保真模拟仍然是扩展该领域基础模型的基石。因此，该领域必须转向先进的适应范式，尤其是迁移学习 和 测试时自适应（*TTA*），从而实现稳健的泛化能力，有效弥合计算机模拟与实验现实之间的差距。

峰-原子可解释性。当前大多数最先进的方法都作为端到端黑箱运行，将光谱输入直接映射为 SMILES 字符串。尽管这一范式有效，但它绕过了基本的化学推理过程：峰指认 （即将特定光谱峰映射到特定原子）。缺乏明确的峰-原子对齐会导致可解释性较差。未来的架构必须超越简单的序列生成，纳入明确的指认机制。

二维数据的稀缺性。我们的结果表明，整合二维光谱约束可能会显著提升解析能力，尤其是对于复杂异构体，因为一维光谱中存在严重的峰重叠。然而，一个主要瓶颈在于，与一维数据集相比，公开可用的二维核磁共振数据十分稀缺。尽管一维光谱随处可见，但它们通常缺乏明确的连接信息，例如异核单量子相干谱（HSQC）提供的直接碳-氢（C-H）相关性，或相关谱（COSY）揭示的质子-质子（H-H）耦合；而这些信息对于无歧义的结构解析至关重要。

6 结论

在本研究中，我们介绍了**NMRGym**，这是规模最大、首个公开发布的标准化基准，源自高质量实验核磁共振数据，旨在弥合合成光谱与实验光谱之间的领域差距，规范未来的研究，并加速*AI4Spectrum*的部署。

参考文献

[1] Marvin Alberts、Oliver Schilter、Federico Zipoli、Nina Hartrampf 和 Teodoro Laino。2024 年。揭示分子结构：用于化学的多模态光谱数据集。《神经信息处理系统进展》37（2024），125780–125808。[2] Marvin Alberts、Federico Zipoli 和 Alain Vaucher。2023年。学习核磁共振的语言：使用Transformer模型从核磁共振谱进行结构解析。载于《用于加速材料设计的人工智能——NeurIPS 2023研讨会》。[3] Marvin Alberts、Federico Zipoli 和 Alain Vaucher。2023年。学习核磁共振的语言：使用Transformer模型从核磁共振谱进行结构解析。载于《用于加速材料设计的人工智能——NeurIPS 2023研讨会》。[4] Katsuo Asakura。2015年。天然产物核磁共振光谱数据库“CH-NMR-NP”。《日本合成有机化学杂志》73，12（2015），1247–1252。doi: 10 5059 73 1247 。/yukigoseikyokaishi。. Y. Binev、[5] M. M. B. Marques 和 J. Aires-de Sousa。2007年。使用针对化学位移训练的关联神经网络预测氢-1核磁共振耦合常数。《化学信息与建模杂志》47，6（2007），2089–2097。[6] Roman Bushuiev、Anton Bushuiev、Niek de Jonge、Adamo Young、Fleming Kretschmer、Raman Samusevich、Janne Heirman、Fei Wang、Luke Zhang、Kai Dührkop 等。2024年。MassSpecGym：用于分子发现与鉴定的基准。《神经信息处理系统进展》37（2024），110010–110027。[7] Tim Claridge。2009年。Mnova软件评述：核磁共振数据处理、分析与预测软件。[8] Manuel Cordova、Edgar A Engel、Artur Stefaniuk、Federico Paruzzo、Albert Hofstetter、Michele Ceriotti 和 Lyndon Emsley。2022年。面向化学和结构多样分子固体的化学位移机器学习模型。《物理化学杂志C辑》126，39（2022），16710–16720。[9] Robert Ditchfield。1974年。抗磁性的自洽微扰理论：I. 用于核磁共振化学位移的规范不变LCAO方法。《分子物理学》27，4（1974），789–807。[10] Leslie D Field、Sev Sternhell 和 John R Kalman。2012年。《从光谱解析有机结构》。约翰·威立父子出版公司。[11] W. Gerrard、L. A. Bratholm、M. J. Packer、A. J. Mulholland、D. R. Glowacki 和 C. P. Butts。2020年。IMPRESSION——使用机器学习以接近量子化学精度预测三维化学结构的核磁共振参数。《化学科学》11（2020），508–515。[12] Y. Guan 等。2021年。使用三维图神经网络以密度泛函理论精度实时预测氢-1和碳-13化学位移。《化学科学》12，36（2021），12012–12026。[13] Yanfei Guan、S V Shree Sowndarya、Liliana C Gallegos、Peter C St. John 和 Robert S Paton。2021年。使用三维图神经网络以密度泛函理论精度实时预测氢-1和碳-13化学位移。《化学科学》12，36（2021），12012–12026。doi:10.1039/D1SC03343C CASCADE网络服务器：http://nova.chem.colostate.edu/cascade/。[14] Kehan Guo、Yili Shen、Gisela Abigail Gonzalez-Montiel、Yue Huang、Yujun Zhou、Mihir Surve、Zhichun Guo、Prayel Das、Nitesh V Chawla、Olaf Wiest 等。2025年。光谱学中的人工智能：推动化学从预测走向生成及更远。《arXiv》预印本 arXiv:2502.09897（2025）。[15] Amit Gupta、Sabyasachi Chakraborty 和 Raghunathan Ramakrishnan。2021年。加速跨化学空间的碳-13核磁共振屏蔽预测：分子原子核机器学习的基准，以及包含134000个分子的新数据。《机器学习：科学与技术》2，3（2021），035010。[16] Pierre Hohenberg 和 Walter Kohn。1964年。非均匀电子气。《物理评论》136，3B（1964），B864。[17] Emiel Hoogeboom、Vıctor Garcia Satorras、Clément Vignac 和 Max Welling。2022年。用于三维分子生成的等变扩散。载于《国际机器学习会议》。PMLR，8867–8887。[18] Frank Hu、Michael S Chen、Grant M Rotskoff、Matthew W Kanan 和 Thomas E Markland。2024年。使用多任务机器学习从常规一维核磁共振谱进行准确高效的结构解析。《ACS中央科学》10，11（2024），2162–2170。[19] Yongqi Jin、Jun-Jie Wang、Fanjie Xu、Xiaohong Ji、Zhifeng Gao、Linfeng Zhang、Guolin Ke、Rong Zhu 等。2025年。NMR-Solver：通过大规模光谱匹配和物理引导的片段优化实现自动化结构解析。《arXiv》预印本 arXiv:2509.00640（2025）。[20] E. Jonas 和 S. Kuhn。2019年。带量化不确定性的核磁共振光谱性质快速预测。《化学信息学杂志》11（2019），1–7。[21] Eric Jonas、Stefan Kuhn 和 Nils Schlörer。2022年。核磁共振化学位移预测：综述。《化学中的磁共振》60，11（2022），1021–1031。[22] John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov、Olaf Ronneberger、Kathryn Tunyasuvunakool、Russ Bates、Augustin Žídek、Anna Potapenko 等。2021年。高度准确的蛋白质结构预测与AlphaFold。《自然》596，7873（2021），583–589。[23] S. Kang 等。2020年。不使用原子级标注的核磁共振化学位移预测建模。《化学信息与建模杂志》60，8（2020），3765–3769。[24] Sunghwan Kim、Jie Chen、Tiejun Cheng、Asta Gindulyte、Jia He、Siqian He、Qingliang Li、Benjamin A Shoemaker、Paul A Thiessen、Bo Yu 等。2025年。PubChem 2025年更新。《核酸研究》53，D1（2025年），D1516–D1525。[25] Harold W Kuhn。1955年。用于指派问题的匈牙利算法。《海军研究后勤季刊》2，1-2（1955年），83–97。[26] Y. Kwon 等。2020年。用于核磁共振化学位移预测的神经消息传递。《化学信息与建模杂志》60，4（2020年），2024–2030。[27] Greg Landrum。2013年。RDKit 文档。Release 1，1-79（2013年），4。[28] M. Lin 等。2022年。一种从顺磁性电池材料的动态核磁共振位移中揭示离子传输机制的机器学习方案。《化学科学》13（2022年），7863–7872。[29] Min Lin、Jingfang Xiong、Mintao Su、Feng Wang、Xiangsi Liu、Yifan Hou、Riqiang Fu、Yong Yang 和 Jun Cheng。2022年。一种从顺磁性电池材料的动态核磁共振位移中揭示离子传输机制的机器学习方案。《化学科学》13，26（2022年），7863–7872。[30] Zhiyuan Liu、Yanchen Luo、Han Huang、Enzhi Zhang、Sihang Li、Junfeng Fang、Yaorui Shi、Xiang Wang、Kenji Kawaguchi 和 Tat-Seng Chua。2025年。NEXT-MOL：三维扩散与一维语言建模结合，实现三维分子生成。arXiv 预印本 arXiv:2502.12638（2025年）。[31] Michael W Lodewyk、Matthew R Siebert 和 Dean J Tantillo。2012年。1H 和碳-13化学位移的计算预测：天然产物、机理研究和有机合成化学的实用工具。《化学评论》112，3（2012年），1839–1862。[32] Alex Morehead 和 Ji anlin Cheng。2024年。用于三维分子生成与优化的几何完备扩散。《化学通讯》7，1（2024年），150。[33] Yong-Cheng Ning。2011年。《有机光谱解析》。约翰·威立父子出版公司。[34] Bharath Ramsundar。2018年。使用 DeepChem 进行分子机器学习。博士学位论文。斯坦福大学。[35] Ann M Richard、Ruili Huang、Suramya Waidyanatha、Paul Shinn、Bradley J Collins、Inthirany Thillainadarajah、Christopher M Grulke、Antony J Williams、Ryan R Lougee、Richard S Judson 等。2020年。Tox21 10K化合物库：推动毒理学发展的协作化学。《毒理学化学研究》34，2（2020年），189–216。[36] Takeshi Saito 和 Shinichi Kinugasa。2011年。有机化合物光谱数据库的开发与发布：大型数据库持续服务与成功的关键。《综合学》 4, 1 2011 35 44 10 5571 4 35 ( )，–。[37] 2003 doi:。/syntheng.。Christoph Steinbeck、Stefan Krause 和 Stefan Kuhn。。NMRShiftDB：构建一个采用开源组件的免费化学信息系统。《化学信息与计算机科学杂志》43，6（2003年），1733–1739。[38] Xiaofeng Tan。2025年。一种基于 Transformer 的生成式化学语言人工智能模型，用于有机化合物结构解析。化学信息学杂志 17，1 2025 103( )，。[39] Qinggong Wang、Wei Zhang、Mingan Chen、Xutong Li、Zhaoping Xiong、Jiacheng Xiong、Zunyun Fu 和 Mingyue Zheng。2025年。NMRExtractor：利用大语言模型从开源科学出版物构建实验核磁共振数据库。《化学科学》16，25（2025年），11548–11558。[40] Yuyang Wang、Ahmed A Elhag、Navdeep Jaitly、Joshua M Susskind 和 Miguel Angel Bautista。2023年。吞下苦药：简化且可扩展的构象生成。arXiv 预印本 arXiv:2311.17932（2023年）。[41] David Weininger。1988年。SMILES：一种化学语言和信息系统。1. 方法论与编码规则简介。《化学信息与计算机科学杂志》28，1（1988年），31–36。[42] David S Wishart、AnChi Guo、Eponine Oler、Fei Wang、Afia Anjum、Harrison Peters、Raynard Dizon、Zinat Sayeeda、Siyang Tian、Brian L Lee 等。2022年。人类代谢组数据库5.0：2022年人类代谢组数据库。《核酸研究》50，D1（2022年），D622–D631。[43] David S Wishart、Zinat Sayeeda、Zachary Budinski、AnChi Guo、Brian L Lee、Mark Berjanskii、Manoj Rout、Harrison Peters、Raynard Dizon、Robert Mah 等。2022年。NP-MRD：天然产物磁共振数据库。《核酸研究》50，D1（2022年），D665–D677。[44] Krzysztof Wolinski、James F Hinton 和 Peter Pulay。1990年。用于核磁共振化学位移计算的规范无关原子轨道法的高效实现。《美国化学会志》112，23（1990年），8251–8260。[45] Zhenqin Wu、Bharath Ramsundar、Evan N Feinberg、Joseph Gomes、Caleb Geniesse、Aneesh S Pappu、Karl Leswing 和 Vijay Pande。2018年。MoleculeNet：分子机器学习基准。《化学科学》9，2（2018年），513–530。[46] Ziyu Xiong、Yichi Zhang、Foyez Alauddin、Chu Xin Cheng、Joon Soo An、Mohammad R Seyedsayamdost 和 Ellen D Zhong。2025年。用于从核磁共振谱解析小分子结构的原子扩散模型。arXiv 预印本 arXiv:2512.03127（2025年）。[47] Fanjie Xu、Wentao Guo、Feng Wang、Lin Yao、Hongshuai Wang、Fujie Tang、Zhifeng Gao、Linfeng Zhang、Weinan E、Zhong-Qun Tian 等。2025年。迈向用于深度学习核磁共振化学位移预测的统一基准与框架。《自然·计算科学》（2025年），1–9。[48] Xi Xue、Hanyu Sun、Jingying Sun、Luc Patiny、Xiangying Liu、Kai Chen、Jingjie Yan、Liangning Li、Xue Liu、Shu Xu 等。2025年。NMRMind：一种实现从多维核磁共振到结构解析的 Transformer 模型。

《分析化学》97，41（2025年），22603–22614。[49] 杨青松、吴冰岚、刘旭伟、陈博、李伟、龙根、陈鑫和肖明军。2025年。DiffNMR：用于核磁共振谱解析的扩散模型。*arXiv* 预印本 *arXiv:2507.08854*（2025年）。[50] 邹子涵、张雨瑾、梁立军、魏明志、冷建才、姜军、罗毅和胡伟。2023年。一种用于预测特定有机分子光谱的深度学习模型。《自然·计算科学》3，11（2023年），957–964。

化学预备知识

形式上，我们将光谱建模为化学位移域 δ 上的连续函数 $x(\delta)$：$\mathbb{R} \to \mathbb{R}$。忽略自旋 - 自旋相互作用时，信号是 N 个相互独立的共振峰的叠加：

$$x(\delta) = \sum_{n=1}^{N} I_n \cdot \mathcal{L}(\delta; \mu_n, \lambda_n) + \xi(\delta), \tag{7}$$

其中，I_n 和 μ_n 分别表示第 n 个原子核的强度和化学位移，$\xi(\delta)$ 表示加性高斯噪声。谱线形状 L 通常服从半峰宽为 λ_n 的洛伦兹分布：

$$\mathcal{L}(\delta; \mu_n, \lambda_n) = \frac{1}{\pi} \frac{\lambda_n}{(\delta - \mu_n)^2 + \lambda_n^2}. \tag{8}$$

在实验条件下，标量耦合（J-耦合）会在相邻原子核之间引入自旋-自旋相互作用，使共振信号分裂为多重峰。该模型可推广为对 K_n 个子峰的嵌套求和：

$$x(\delta) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} I_{n,k} \cdot \mathcal{L}(\delta; \mu_{n,k}, \lambda_n) + \xi(\delta), \tag{9}$$

其中，相对位置 $\mu_{n,\,k}$ 和强度 $I_{n,\,k}$ 由耦合常数（J值）和分子拓扑决定。

至关重要的是，这些光谱参数的可获取性因数据来源而有显著差异。高保真数据集通常提供全面的标注，包括精确的化学位移（μ）、峰强度（I）和耦合常数（J）。

B 实现细节

我们评估了两类基线方法：经典机器学习算法和深度神经网络。为确保结果可复现，我们使用固定的随机种子（例如 42、123、456）进行确定性训练。

B.1 数据预处理

所有核磁共振谱均通过高斯展宽，从离散的化学位移转换为连续的光谱表示。我们将 ^{1}H 核磁共振谱（0 – 12 百万分之一）和 ^{13}C 核磁共振谱（0 – 220 百万分之一）各离散为 1,800 个区间。我们应用高斯核（σ = 0.05百万分之一）对信号进行平滑处理：

$$y(x) = \sum_{i} \exp\left(-\frac{(x - s_i)^2}{2\sigma^2}\right), \tag{10}$$

其中 s_i 表示各个峰的位置。随后将所得光谱进行最小-最大归一化至 $[0, 1]$。对于基于向量的模型（多层感知机、XGBoost、随机森林），将两种光谱拼接成一个 3,600 维向量。对于基于二维数据的模型（卷积神经网络），则将光谱重塑为 60×60 双通道图像。

B.2 经典机器学习基线方法

XGBoost。我们使用梯度提升框架，为每种官能团训练了一个由相互独立的二元分类器组成的集成模型。主要超参数包括 200 轮提升、最大树深度为 8，以及 0.1 的学习率。为缓解类别不平衡，我们排除了极其罕见的类别（< 10 个样本），并根据验证损失采用早停。

随机森林。我们采用多输出分类策略，其中每个随机森林由 200 棵决策树组成，最大深度为 30。为应对标签分布偏斜，我们采用类别平衡加权（class_weight='balanced'），自动根据类别频率的倒数调整权重。

B.3 深度学习基线

多层感知机。多层感知机通过三个全连接层（$3600 \to 1024 \to 512 \to 22$）处理展平后的 3,600 维光谱向量。正则化方法包括 ReLU 激活，并在前两层之后应用 Dropout（$p = 0.2$）。

卷积神经网络）。卷积神经网络架构由四个卷积块和一个三层全连接头组成。每个卷积块包含一个 3×3 2D 卷积、批归一化、ReLU 激活和池化（最大池化或自适应平均池化）。通道深度依次为 $1 \to 64 \to 128 \to 256 \to 512$。分类头（$512 \to 1024 \to 512 \to 22$）采用 Dropout（$p = 0.3$，0。2）以防止过拟合。

B.4 训练与优化

损失函数。为应对严重的类别不平衡（正类比例为< 1% ~ 30%），我们将所有神经网络模型的标准交叉熵替换为**FocalLoss** 。损失函数定义如下：

$$L_{FL} = -\sum_{i} [\alpha(1-p_i)^{\gamma} y_i \log(p_i) + (1-\alpha) p_i^{\gamma} (1-y_i) \log(1-p_i)]. \tag{11}$$

我们设置聚焦参数 $\gamma = 5.0$ 和 $\alpha = 0.25$，以降低简单负样本的权重，并强调困难的少数类样本。

优化。深度模型使用 Adam（学习率 $= 3\times10^{-6}$，$\gamma = 0.975$）进行优化，并采用阶梯衰减调度器（每个 epoch 衰减一次）。训练最多进行 50 个训练轮次，并使用早停（耐心值为 = 6 个训练轮次）；批量大小为 256。

偏置初始化。最终分类层的偏置采用经验对数先验概率 ($b_i = \log(p_i/(1-p_i))$) 进行初始化，使初始预测与边缘类别分布相一致。

B.5 推理策略

我们没有使用固定阈值（例如 0.5），而是通过最大化验证集上的 F1 分数，计算**每个类别的最优阈值**。对于基准率各异的多标签任务，这种自适应阈值策略至关重要。

C 官能团定义

官能团的定义见表 8。

D 毒性标签定义

NMRGym 毒性子集包含 12 个源自 Tox21 数据挑战赛的二元标签。这些标签表示定量高通量筛选（qHTS）检测的实验结果，分为核受体（NR）信号通路和应激反应（SR）通路。官能团的定义见表 9。

表 8：NMRGym 中包含的 20 种官能团标签的定义。

	定义
醇	[OX2H]
羧酸	[CX4;!$(C([OX2H])[O,S,#7,#15])][CX3](=O)[OX2H1]
酯	[#6][CX3](=O)[OX2H0][#6]
醚	[OD2]([#6])[#6]
醛	[CX3H1](=O)[#6]
酮	[#6][CX3](=O)[#6]
烯烃	[CX3]=[CX3]
炔烃	[$([CX2]#C)]
苯	c1ccccc1
伯胺	[NX3;H2;!$(NC=[#6]);!$(NC#[#6])][#6]
仲胺	[NH1,nH1]
叔胺	[NH0,nH0]
酰胺	[NX3][CX3](=[OX1])[#6]
氰基	[NX1]#[CX2]
氟	[#6][F]
溴	[#6][Br]
磺酰胺	[#16X4]([NX3])(=[OX1])(=[OX1])[#6]
砜	[#16X4](=[OX1])(=[OX1])([#6])[#6]
硫化物	[#16X2H0]
磷酸	[$(P(=[OX1])([$([OX2H]),$([OX1-]),$([OX2]P)])([$([OX2H]),$([OX1-]),$([OX2]P)])[$([OX2H]),$([OX1-]),$([OX2]P)]),$([P+]([OX1-])([$([OX2H]),$([OX1-]),$([OX2]P)])([$([OX2H]),$([OX1-]),$([OX2]P)])[$([OX2H]),$([OX1-]),$([OX2]P)])]

表7：基线模型的超参数设置。

方法	超参数	值
通用	高斯 σ	0.05 百万分之一
	光谱分箱（^{1}H / ^{13}C）	1800 / 1800
	随机种子	42, 1337, 2024
卷积神经网络 / 多层感知机	焦点损失（γ, α）	5.0, 0.25
	学习率	3×10^{-6}
	批量大小	256
	优化器	Adam
	Dropout	$0.2 - 0.3$
XGBoost	n 个估计器_	200
	最大_深度	8
	学习_率	0.1
随机森林	n_估计器	200
	最大_深度	30
	类别_权重	平衡的

E 更多结构解析结果

E.1 使用余弦指标评估指纹相似度

E.2 NMR-Solver（搜索 +优化）在子集上的表现

鉴于 NMRSolver 的完整流程涉及计算开销巨大的三维构象生成和组合优化——处理整个测试集需要一个多月的时间——我们整理了一个包含 6,329 个样本的子集，用于与最先进的基于 Transformer 和基于扩散的方法进行基准比较。

如图5所示，NMRSolver（包含优化模块的完整流程）表现出显著优于基线方法的性能，尤其是在结构解析中最严格的指标——序列级准确率方面。最值得注意的是，NMRSolver的Top-1序列准确率达到22.88%，大幅领先NMRMind（15.66%）逾7个百分点（相对提升 ~46%）。有趣的是，尽管NMRMind在词元级准确率方面表现不俗（例如，前5名：60.0%，而NMR-Solver为57.7%），但它难以将这些预测正确的子结构组合成准确的分子结构。

这一差异凸显了我们提出的优化模块所发挥的关键作用。通过优化结构候选，优化过程能够有效纠正纯序列生成模型（如 NMRMind）无法解决的拓扑错误。此外，在结构相似度方面，NMRSolver 在前10名和前5名排名中持续检索到化学保真度更高的候选（Morgan相似度：0.78 和 0.82，而 NMRMind 分别为 0.73 和 0.76），进一步验证了优化增强型流程的可行性和稳健性。

表 9：NMRGym 子集中包含的 12 个毒性标签的定义。

标签 ID	全名	生物学靶点 / 机制
核受体（*NR*）面板		
核受体-雄激素受体	雄激素受体	雄激素受体的激动作用。
NR-AR-LBD	雄激素受体配体结合域	AR 配体结合域的拮抗作用（荧光素酶）。
NR-AhR	芳烃受体	芳烃受体信号传导的激活。
NR-芳香化酶	芳香化酶	芳香化酶（CYP19A1）的抑制。
NR-ER	雌激素受体	雌激素受体的激动作用 α（ERα）。
NR-ER-LBD	雌激素受体配体结合域	ER配体结合域的拮抗作用。
NR-PPAR-γ	PPARγ	过氧化物酶体增殖物激活受体 γ的激动作用。
应激反应（*SR*）面板		
SR-ARE	抗氧化反应元件	激活 Nrf2抗氧化通路（氧化应激）。
SR-ATAD5	ATAD5	诱导遗传毒性（DNA损伤反应）。
SR-HSE	热休克反应	热休克因子反应元件的激活。
SR-MMP	线粒体膜电位	线粒体膜电位的破坏。
SR-p53	p53	p53 DNA损伤应答通路的激活。

表10：使用余弦相似度对前K名 Morgan、扭转和原子对指纹进行相似度评估。

	Morgan 相似度（% ↑）			扭转相似度（% ↑）			原子对相似度（% ↑）		
	前1名	前5名	前10名	前1名	前5名	前10名	前1名	前5名	前10名
基于搜索的方法									
NMR-Solver（仅搜索）[19]	43.94 ± 0.00	53.90 ± 0.00	57.75 ± 0.00	44.92 ± 0.00	56.27 ± 0.00	60.39 ± 0.00	60。38 ± 0.00	68.82 ± 0.00	71.77 ± 0.00
+公式条件	45。68 ± 0.00	47.41 ± 0.00	47.69 ± 0.00	46.08 ± 0.00	47.68 ± 0.00	47.92 ± 0.00	48.20 ± 0.00	49.30 ± 0.00	49.48 ± 0.00
基于 *Transformer* 的方法									
CLAMS[38]	1.57 ± 0.28	4.80 ± 0.88	7.39 ± 1.60	0.59 ± 0.19	1.92 ± 0.46	3.07 ± 0.67	1.79 ± 0.33	5.87 ± 1.19	9.46 ± 2.15
NMRFormer[2]	36.03 ± 0.05	52.45 ± 0.06	57.02 ± 0.04	37.14 ± 0.04	54.87 ± 0.07	59.88 ± 0.06	42.74 ± 0.08	61.89 ± 0.05	66.79 ± 0.05
NMR2Struct[18]	48.93 ± 0.66	57。20 ± 0.79	61.01 ± 0.75	48.14 ± 0.80	58.43 ± 0.83	62.83 ± 0.78	62.17 ± 0.54	69。69 ± 0。54	72.63 ± 0.48
NMRMind[48]	70。81 ± 0.12	76.77 ± 0.08	78.77 ± 0.08	71.68 ± 0.10	78。27 ± 0.11	80.46 ± 0.09	79.42 ± 0.04	84.20 ± 0.15	85.68 ± 0.09
+公式条件	**74.09 ± 0.03**	**80.53 ± 0.06**	**82.39 ± 0.07**	**75.86 ± 0.04**	**82.70 ± 0.05**	**84.64 ± 0.06**	**82.84 ± 0.01**	**87.96 ± 0.04**	**89.23 ± 0.03**
基于扩散的方法									
DiffNMR[49]	16.14 ± 0.02	23.09 ± 0.01	25.12 ± 0.02	11.58 ± 0.01	21.08 ± 0.02	24.41 ± 0.01	32.09 ± 0.02	43.30 ± 0.01	45.76 ± 0.02
ChefNMR-S[46]	4.70 ± 0.04	12.35 ± 0.03	15.83 ± 0.04	4.07 ± 0.04	11.74 ± 0.06	15.68 ± 0.05	10.41 ± 0.07	26.90 ± 0.08	33.97 ± 0.14
ChefNMR-S（微调）[46]	23.39 ± 0.09	44.72 ± 0.13	51.13 ± 0.18	24.67 ± 0.09	47.83 ± 0.18	55。00 ± 0。20	32.80 ± 0.11	60.50 ± 0.16	67。61 ± 0.21
ChefNMR-L[46]	5.55 ± 0.03	14.24 ± 0.04	18.04 ± 0.02	4.89 ± 0.02	13.62 ± 0.04	17.98 ± 0.02	12.05 ± 0.01	30.29 ± 0.01	37.74 ± 0.01
ChefNMR-L（微调）[46]	27。52 ± 0.20	49.08 ± 0.06	54.80 ± 0.04	29.08 ± 0.11	52.54 ± 0.07	58.89 ± 0.05	38.29 ± 0.29	65.13 ± 0.16	70.86 ± 0.05

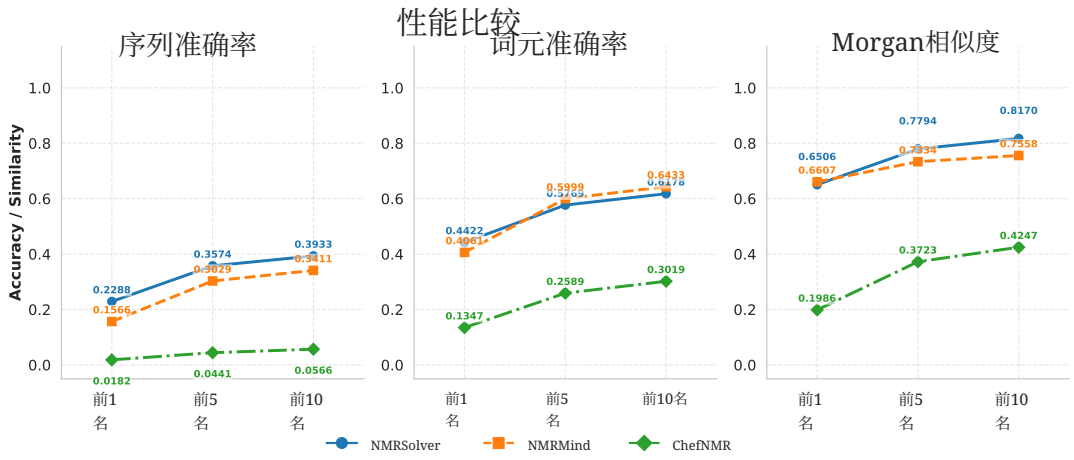


图5：thehard 子集上的性能比较。评估前1名、前5名和前10名排名中的序列准确率、词元准确率和Morgan-田中本体相似度。将NMRSolver（采用包含优化模块的完整流程）与NMRMind及基线模型ChefNMR进行比较。结果表明，优化策略能够有效改进结构预测。