arXiv:2304.03361v1 [physics.chem-ph] 6 Apr 2023

基于少量数据的核磁共振化学位移预测

预印本

Herman Rull爱沙尼亚塔尔图大学计算机科学系
herman.rull@ut.ee

Markus Fischer德国莱比锡大学医学物理与生物物理研究所

Stefan Kuhn爱沙尼亚塔尔图大学计算机科学系

2023年4月10日

摘要

使用机器学习方法预测核磁共振中的化学位移，通常会利用尽可能多的可用数据，以获得最佳结果。在某些情况下，例如对于异核，并没有如此大量的数据可用。我们展示了一种新颖的机器学习模型，该模型能够利用相对较少的数据取得良好结果。我们通过预测特定溶剂中小分子的^{19}F和^{13}C核磁共振化学位移来对此进行展示。

1 引言

核磁共振（NMR）中的化学位移预测是化学信息学领域一个由来已久的问题。 [Dailey 和 Shoolery（1955） Dailey、Shoolery] 的研究或许是该领域最早的发表成果。这里，我们将预测定义为使用现有数据的方法，而不是从头算起的计算。随着时间的推移，人们开发出了各种用于此类预测的方法。特别是，机器学习方法已被应用于这一领域，从早期的方法（如小型神经网络 [Kvasnicka 等（1992）Kvasnicka、Sklenak 和 Pospichal], ）一直发展到最新的卷积神经网络和图神经网络。关于该领域的概览，请参阅近期的综述[Jonas 等（2022年）Jonas、Kuhn 和 Schl¨orer] 。

对于监督学习方法，例如前文提到的神经网络，需要带注释的数据集，而所使用的数据点数量是预测结果质量的一个重要因素。对文献的综述表明，所使用的数据集通常规模较大，由数万个分子组成。表 1概述了近期出版物中所使用的结构数量。需要注意的是，有时会进行初步筛选，例如， [Tsai等人(2022年)Tsai、Amichetti、Zanardi、Grimson、Daranas 和 Sarotti] 中的17,000个结构，就是从 170,000 个分子中依据结构多样性进行筛选的结果（使用了 RDKit 中的 Morgan 指纹和 MaxMin 算法）。

文献参考	[Jonas 和 Kuhn（2019）Jonas 和 Kuhn]	[Unzueta 等人(2021年)Unzueta、Greenwell 和 Beran]	^{13}C [Kwon
数量 使用的结构	32,538	57,456	

表1：关于化学位移预测的论文示例及使用的结构数量 d.

遗憾的是，许多研究并未考虑训练数据量对预测质量的影响。一个例外是 [Kuhn 等人（2022年）Kuhn、Borges、Venturini 和 Sansotera] , , 该研究表明，一些机器学习方法只有在训练样本超过 5000 个时才具有足够的预测能力。然而，在许多实际应用中，可用的实验数据量相当有限。例如，杂核的核磁共振化学位

特征	描述 [单位] (类型)
原子序数	独热编码的 [数据集中的所有原子](array[bool])
原子半径	来自 Mendeleev 库的斯莱特数据 [pm](整数)
中子	中子数 [-](整数)
电负性	鲍林标度 [-](浮点数)
电子亲和能	来自 Mendeleev 库的值 [eV](浮点数)

表 2：2023模型中使用的原子特征。

特征	描述 [单位] (类型)
键长	原子中心之间的距离 [$\mathring{A}$](float)
键类型	独热编码[单键、双键、三键、芳香键]（数组[布尔值]）

表3：2023年模型中使用的键特征。

移、特定类别的化合物、使用特定溶剂测得的核磁共振谱，或其他特定实验条件。对于此类数据量较少的情况（此处定义为结构少于 5000 个），现有模型要么无法提供良好的解决方案，要么尚未在如此小的数据集上经过测试。

本文提出了一种图神经网络，即使数据量较少也能实现良好的预测结果；在数据量较大时，其性能也能与最先进的模型相媲美。我们的研究仅针对溶液中的小型有机分子。固态核磁共振化学位移、生物大分子或无机化合物的预测需要采用不同且更具针对性的模型。因此，本文不涉及这些情况。

2 材料和方法

预测模型可应用于许多不同的核磁共振参数，例如耦合常数、弛豫时间或峰形。在此，我们希望重点关注核磁共振化学位移的预测。现代（溶液）核磁共振实验是极佳的数据来源，因为它们能够提供噪声很小的各向同性化学位移信息[Zangger(2015)]。在核磁共振波谱学中，化学位移等同于一个原子的共振频率，其数值由某个原子的（化学）环境决定。如何表示这一环境是一项艰巨的任务。单个原子可以用一个向量表示，其中存储了其属性。然而，这种表示方式无法适用于整个分子。因此，我们不采用类似张量的表示，而是将分子构建为图。在图中，原子可以描述为节点，而连接原子的化学键可以描述为边。利用图神经网络进行几何学习[Battaglia et al.(2018)Battaglia, Hamrick, Bapst, Sanchez-Gonzalez,Zambaldi, Malinowski, Tacchetti, Raposo, Santoro,Faulkner,Gulc

Gori 等人（2005）Gori、Monfardini 和 Scarselli] 提供了一种适用于在此类特定结构上运行机器学习算法的工具。在图中，信息沿边传递，因此来自相连原子的信息最为重要。这与真实分子相符：相邻原子之间的键数较少，对核磁共振化学位移值的影响最大。

在这项工作中，我们基于[Fischer 等人（2022）Fischer、Schwarze、Ristic 和 Scheidt] , 开发了一个模型，用于学习分子中的原子性质。本文将其称为“2023模型”。该模型使用消息传递图网络[Duvenaud 等人（2015）Duvenaud、Maclaurin、Aguilera-Iparraguirre、G´omez-Bombarelli、Hirzel、Aspuru-Guzik 和 Adams，Gilmer 等人（2017）Gilmer、Schoenholz、Riley、Vinyals 和 Dahl], , 通过图中的边传递信息，从而逐步在节点中构建局部信息。遵循[Fischer 等人（2022）Fischer、Schwarze、Ristic 和 Scheidt] , 的方法，我们使用一种带有额外边聚合函数的消息传递图网络模块，如图1所示。由于我们使用该网络来预测小分子的节点级特征，因此忽略了[Fischer 等人（2022）Fischer、Schwarze、Ristic 和 Scheidt]所述图网络的第二阶段。我们使用表2和表3中列出的一组特征来描述原子和化学键。

之所以选择这一特定的特征选择方法，是因为初步实验测量了每个特征对最终预测结果的影响。随后，将表现最佳的特征进行组合，直到预测质量不再提高为止。

为了优化预测准确度，我们仔细为2023模型选择了最佳超参数。最重要的超参数包括消息传递步数、学习率和权重衰减。

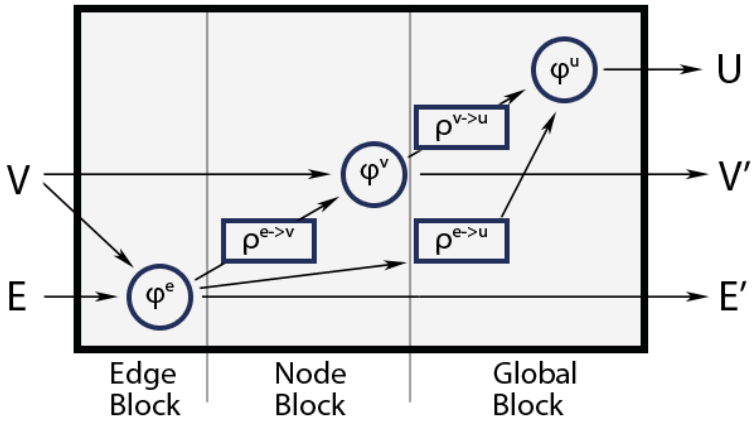


图1：2023模型的消息传递图网络中的信息流。除了节点和边聚合函数之外，还有一个额外的边聚合函数输入到全局更新函数中（引自 [Fischer等人（2022年）Fischer、Schwarze、Ristic和Scheidt]）。

我们使用4折交叉验证，在 ^{19}F 数据的训练集上优化超参数。表现最佳的模型使用了6个消息传递步数、 10^{-3}的学习率和0。01的权重衰减。

所有编程均使用 Python 完成，并以 RDKit [rdk(2023（访问于2023年2月24日）)] 版本 2022.9.5 作为主要库。此外，使用 mendeleev [men(2014–)] 版本 0.12.1 计算部分原子性质。本文的补充信息中包含一本 Jupyter笔记本，其中提供了代码和说明。

作为比较，我们使用另外两种预测方法。其中一种是分层排序的球形环境（HOSE）代码 [Bremser(1978)]，，这是一种长期以来用于将原子及其环境描述为字符串的方法。借助这些字符串，可以查找其他相似原子的化学位移，并将其用于预测。从机器学习的角度来看，这可以称为最近邻搜索。所使用的 HOSE 代码实现是化学开发工具包（CDK）[Willighagen 等人(2017)Willighagen、Mayfield、Alvarsson、Berg、Carlsson、Jeliazkova、Kuhn、Pluskal、 ´o、Spjuth、Torrance、Evelo、Guh 的 HOSE 代码实现移植版。

并可在 [hos(2023年 (访问日期：2023年2月24日))] 获取。这会生成标准 HOSE编码，而不是 [Kuhn和 Johnson(2019)Kuhn，和 Johnson] 的立体增强型 HOSE编码。

我们将 [Jonas and Kuhn(2019)Jonas, and Kuhn] 提出的模型作为一种现代机器学习模型，并在本文中称其为“2019模型”。该模型使用卷积图神经网络，将某个原子的特征向量与其邻居的特征向量结合起来进行预测。对这些方法的评估通常采用75:25的训练集与测试集划分。由于数据集规模较小，我们决定不单独设置验证集。

所有数据均取自开放的核磁共振数据库 nmrshiftdb2 [Kuhn 等人（2012）Kuhn、Schl¨orer、Kolshorn 和 Stoll]。该数据库包含化学位移值列表，以及针对多种不同原子核的各种一维和二维核磁共振实验的原始数据。正如第 3 节各小节中所述，我们在此关注特定的子集。需要注意的是，我们所使用的数据集由随机选取的结构组成。通过确保结构多样性或化学空间中的均匀分布，或许可以利用小型数据集优化训练过程。我们没有采用这种做法，而是假设数据呈随机分布。具体而言，我们纳入了 nmrshitdb2 中的所有实验数据（前提是它们符合第 3 节所使用的子集）。这与其他研究有所不同，例如 [Jonas 和 Kuhn（2019）Jonas，以及 Kuhn]，后者将选择范围限制为仅含常见元素的分子。这也解释了使用 2019模型和来自 nmrshiftdb2 的数据时结果略有不同，此外数据库也会随时间发生变化。

为了比较模型在各种条件下的性能，我们报告三个数值：平均绝对误差（平均绝对误差）、均方根误差（均方根误差）以及误差的标准差 σ 。标准差根据模型的所有预测结果计算，用于衡量误差相对于均值的变化程度。

		100	250	500	1000	2500	5000	10000	25000	44370
2019 模型	平均绝对误差（百万分率）	70.31	64.84	61.64	57.77	31.81	3.65	2.40	2.11	1.82
	均方根误差（百万分率）	83.30	79.86	76.36	70.71	36.27	5.41	3.35	4.09	3.13
	σ (百万分率)	52.29	52.51	52.69	48.15	28.05	6.64	5.08	5.13	4.57
2023 模型	平均绝对误差（ppm）	24.8	21.45	22.77	17.11	15.0	11.18	9.63	8.21	7.65
	均方根误差（ppm）	46.65	40.82	44.66	45.27	40.27	32.82	29.01	25.63	24.53
	σ (ppm)	45.29	40.44	44.41	45.13	40.11	32.77	28.95	25.58	24.48
HOSE code	平均绝对误差（ppm）	20.81	18.99	17.68	18.85	16.14	15.2	14.02	12.14	10.98
	均方根误差（ppm）	35.29	33.44	30.94	30.32	30.03	29.19	27.95	25.84	24.60
	σ (百万分率)	34.72	33.26	30.74	30.29	30.01	29.18	27.95	25.83	24.5
	缺失预测结果	17.6	19.6	26.3	33.9	41.4	51.0	57.9	76.2	85.2

表 4：使用递增数量的光谱预测 ^{13}C 位移的结果

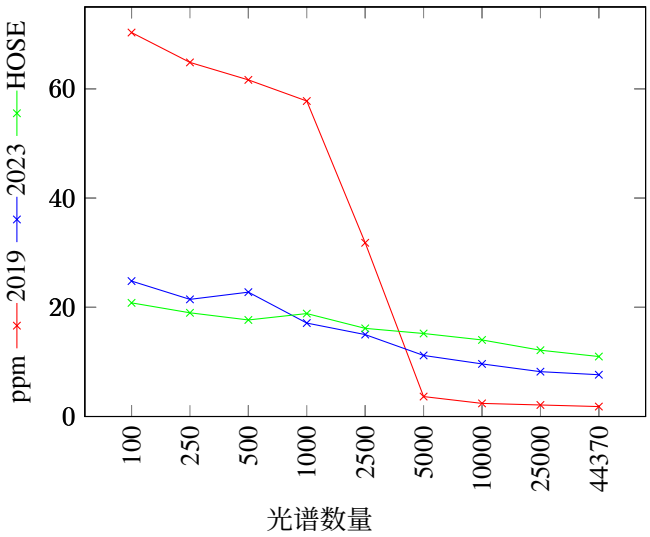


图 2：使用不断增加的样本数量进行 a ^{13}C 核磁共振化学位移预测的平均绝对误差。

3 结果

3.1 整体表现

首先，我们希望将新的2023模型与HOSE编码和2019模型进行比较。为此，我们分析了不同模型在使用数量不断增加的分子进行训练时的预测性能。结果如图2和表4所示。可以清楚地看到，在使用最多2500个数据点（结构）进行训练时，新模型的表现优于2019模型，而从5000个数据点开始，2019模型的表现更好。2019模型在该范围内的显著改进已在[Kuhn等人（2022）Kuhn、Borges、Venturini和Sansotera], 中被观察到，而对于这种自发改进的解释至今仍未得出。

HOSE编码具有良好的预测能力，这一点此前已在其他已发表的研究中得到证实。然而，如表4所示，在某些情况下无法基于HOSE编码进行预测。如果训练集中不存在相似度足够高（即至少有一个球层）的示例，就可能出现这种情况。表中还显示，2019模型结果的标准差显著低于HOSE编码，说明该模型比基于HOSE编码的预测更加稳定。

3.2 杂原子核

^{13}C 和 ^{1}H 是核磁共振波谱学中最常用的原子核，主要因为具有磁响应性的同位素天然丰度较高，并且存在于有机化合物中。其他原子核也用于某些应用，但

		100	250	500	957
2019模型	平均绝对误差（百万分率）	79.43	72.68	69.65	57.82
	均方根误差（百万分率）	82.54	84.04	73.23	61.86
	σ （百万分率）	47.32	93.13	43.18	41.61
2023模型	平均绝对误差（百万分率）	22.25	15.94	13.32	9.77
	均方根误差（百万分率）	45.19	38.46	34.02	27.95
	σ (百万分率)	43.57	37.56	33.44	27.73
HOSE代码	平均绝对误差（百万分率）	12.21	10.53	7.87	7.38
	均方根误差（ppm）	25.97	29.68	20.16	23.33
	σ (百万分率)	25.70	29.45	20.13	23.32
	缺失预测	1.93	2.88	7.38	4.75

表 5：使用逐渐增加的谱图数量预测 ^{19}F 位移的结果

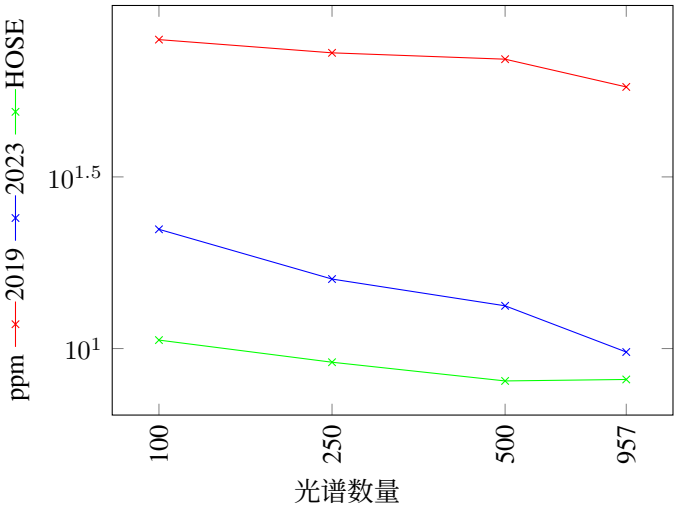


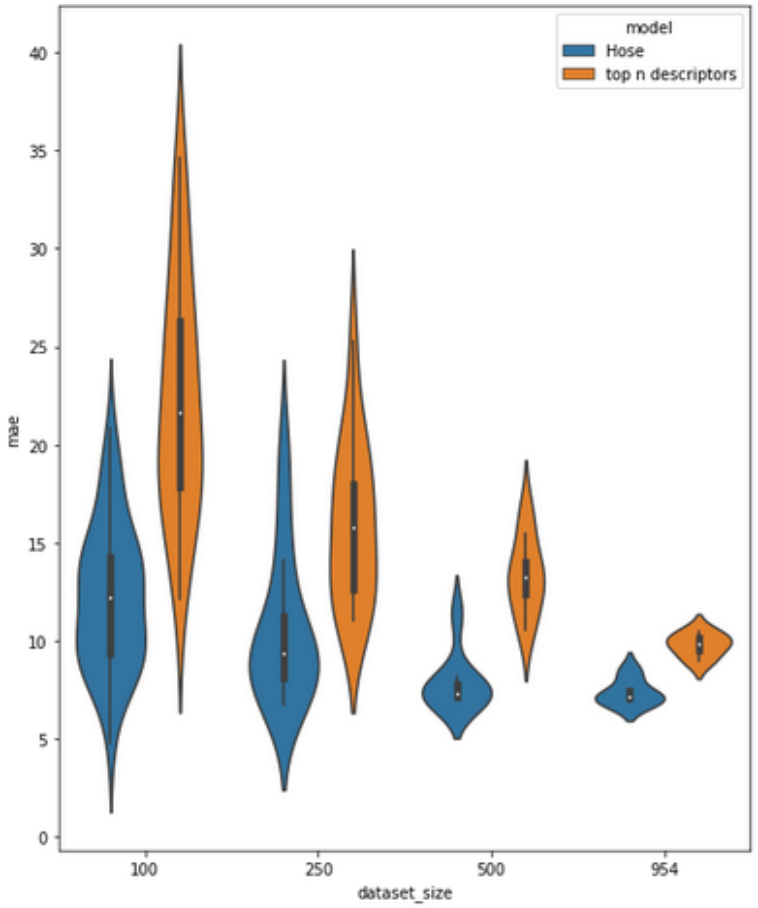
图 3： 使用递增数量的样本，在对数刻度上的一个 ^{19}F 核磁共振化学位移预测的平均绝对误差 le.

可获得的数据量要小得多。因此，它们是我们模型的良好测试案例，在此我们以 ^{19}F光谱为例。在nmrshiftdb2中，目前有957个具有实测 ^{19}F 光谱的结构。我们不考虑nmrshiftdb2中通过从头计算得到的光谱，并且在极少数记录了多条光谱的情况下，每种化合物只使用一条光谱。

我们使用与 ^{13}C 第3.1节中相同的机器学习模型和HOSE编码。通过针对某个原子核专门优化模型，或许可以改进预测，但这不在本工作的范围内。一般而言，^{19}F 的行为应与 ^{13}C 和 ^{1}H相似，但例如对于金属而言可能并非如此。

表5和图3展示了基于 957 ^{19}F nmrshiftdb2中的光谱进行预测的结果，使用100、250、500和957（全部）张光谱训练模型。HOSE编码和2019模型的结果与 [Kuhn等人（2022年）Kuhn、Borges、Venturini和Sansotera] 中的结果相似（差异源于论文中使用了较旧版本的nmrshiftdb2），在这些少量数据的情况下，HOSE编码明显优于2019模型。我们预计，随着数据量增加，该模型的性能将得到提升，与 ^{13}C类似；但是，对于这一核素，可用数据量有限。图3中以蓝色显示的新模型，在使用100张光谱训练时优于2019模型，并且在使用957张光谱训练时性能显著提升，几乎达到了HOSE编码模型的水平（请注意，图3使用对数刻度）。使用957张光谱时，2023年模型的标准差超过了HOSE编码的标准差，因此在此处能够提供更稳定的预测。图4展示了2019模型和HOSE编码的分布，显示出随着数据量增加而取得的显著改进。

基于HOSE编码的预测结果相当准确，但正如之前所讨论的，其固有缺点是可能完全无法给出预测结果。机器学习模型即使面对相似度较低的分子，也始终会预测化学位移，因为它们能够进行泛化。因此，“缺失预测”这一类别仅用于HOSE编码。



Figu图 4：HOSE 代码与 GNN 预测的准确度及其分布比较 n.

3.3 溶剂

所使用的溶剂是影响特定化合物化学位移值的主要因素之一，因为它会影响分子的化学环境、形成氢键的可能性、所研究分子电荷状态的变化等。出于预测目的，通常的做法是忽略溶剂（例如 [Jonas 和 Kuhn(2019年)Jonas，以及 Kuhn] 或 [Kwon 等人(2020年)Kwon、Lee、Choi、Kang 和 Kang]）。更准确的预测需要使用溶剂信息。这其中的一个问题是，针对特定溶剂的光谱数量相对较少，即使对于 ^{13}C 和 ^{1}H 光谱也是如此。例如，nmrshiftdb2目前有 2324 ^{13}C份在氘代氯仿-D1中的核磁共振光谱、456份在六氘代二甲基亚砜中的光谱，以及351份在氘代甲醇中的光谱（这些是数据库中最常见的溶剂）。

我们使用这些数据为每种溶剂分别训练模型，并将结果与使用所有 ^{13}C 光谱所得到的数值进行比较。结果如表6所示。需要注意的是，这些模型与之前使用所有溶剂和 ^{19}F 原子核进行预测时所使用的模型相同，并未针对特定溶剂的预测进行优化。我们仍然可以作出以下观察：

- 特定溶剂训练相比整体模型能产生更好的结果。例如，对于氘代氯仿-D1，2019模型和2023模型在2324条光谱数据上达到的平均绝对误差分别为24.06和4.55百万分率，而使用所有溶剂时，在2500条光谱数据上，平均绝对误差分别为31.81和14.60。
- 总体趋势与我们之前看到的类似：2019模型的预测质量起初误差较高，在超过1000个光谱后显著改善。在较小的数据集上，2023模型的表现优于2019模型，这是由于其在使用最多2500个光谱进行训练时能够快速改进。HOSE编码总体表现良好，但没有太大改善。
- 对于 Chloroform-D1，2023模型使用1000条光谱时的误差低于5百万分率。这优于使用全部数据的2023模型。2019模型需要接近5000条光谱，误差才能低于5百万分率。

是必需的。这意味着，鉴于现有数据，我们的新模型优于2019模型。对于六氘代二甲基亚砜和氘代甲醇，我们的新模型仅使用456个或分别351个谱图进行训练，就能取得相当不错的结果。这些结果远优于使用相同数量谱图训练的2019模型。

			100	250	500	1000	2324
氘代氯仿（CDCl3）	2019模型	平均绝对误差（百万分率）	64.69	58.24	58.62	50.91	24.06
		均方根误差（百万分率）	80.32	74.71	74.63	66.18	34.38
		σ (百万分率)	53.17	53.02	52.01	48.61	29.47
	2023模型	平均绝对误差（ppm）	18.89	7.47	5.23	4.32	4.12
		均方根误差（ppm）	34.76	15.69	12.42	10.62	10.53
		σ (百万分率)	31.99	15.02	12.17	10.47	10.48
	HOSE code	平均绝对误差（百万分率）	5.35	4.81	4.32	4.03	3.19
		均方根误差（百万分率）	8.79	8.53	8.03	7.76	6.99
		σ (百万分率)	8.77	8.51	8.02	7.76	6.99
		缺失预测结果	11.92	14.23	17.30	19.42	30.25
六氘代二甲基亚砜（六氘代二甲基亚砜, C2D6SO）	2019模型	平均绝对误差（ppm）	91.67	93.09	85.84 (456)	n/a	n/a
		均方根误差（ppm）	100.41	100.92	93.76 (456)	n/a	n/a
		σ (百万分率)	46.28	44.75	45.01 (456)	n/a	n/a
	2023模型	平均绝对误差（百万分率）	24.03	5.82	5.75(456)	n/a	n/a
		均方根误差（百万分率）	37.35	10.50	8.85（456）	n/a	n/a
		σ (百万分率)	31.26	9.92	7.61(456)	n/a	n/a
	HOSE code	平均绝对误差（ppm）	4.96	4.27	3.65（456）	n/a	n/a
		均方根误差（ppm）	7.61	6.88	6.21(456)	n/a	n/a
		σ (百万分率)	7.60	6.87	6.21(456)	n/a	n/a
		缺失预测结果	9.95	12.0	10.0(456)	n/a	n/a
氘代甲醇（CD_3OD）	2019模型	平均绝对误差（百万分率）	78.60	71.92	69.41 (351)	n/a	n/a
		均方根误差（ppm）	89.66	84.17	82.03 (351)	n/a	n/a
		σ (百万分率)	49.35	49.95	49.75（351）	n/a	n/a
	2023模型	平均绝对误差（百万分率）	18.65	7.10	5.53 (351)	n/a	n/a
		均方根误差（百万分率）	32.69	15.31	9.94 (351)	n/a	n/a
		σ (百万分率)	29.10	14.53	9.11 (351)	n/a	n/a
	HOSE code	平均绝对误差（百万分率）	4.67	4.24	3.64 (351)	n/a	n/a
		均方根误差（百万分率）	8.15	8.05	7.37 (351)	n/a	n/a
		σ (百万分率)	8.13	8.04	7.37（351）	n/a	n/a
		缺失预测结果	8.42	6.75	12.5 (351)	n/a	n/a

表 6：使用逐渐增加的谱图数量对 ^{13}C 位移进行预测的结果。不适用表示可用数据不足，括号中的数字表示所用化合物的数量，与顶部标题不同。

图5显示了使用全部数据和仅使用CDCl3训练的两个模型所达到的平均绝对误差。显然，2023模型的表现优于2019模型。此外，对于两个模型而言，CDCl3预测结果不仅优于使用全部数据得到的预测结果，而且2023模型的改进幅度高于2019模型。

为进一步验证我们的结果，我们仅预测了CDCl3光谱的化学位移，但使用了全部数据进行训练，并采用2019模型。为此，我们将CDCl3光谱分成四个等份，并使用所有非CDCl3数据以及其中三份数据训练模型。随后将第四份数据用作测试集，预测其中的所有化学位移。这些测试集的平均误差为：平均绝对误差 2.04、均方根误差2.65，以及 δ 2.60。这证实了，在数据量充足的情况下，2019模型仅使用CDCl3数据也能够取得良好结果，同时也表明2023模型在CDCl3数据上的良好结果并非由这些数据造成。

为验证化合物的分布是否不受溶剂影响，我们在图6中绘制了化学空间图。可以看出，三种溶剂中的分布均与整个数据库的分布相似。需要注意的是，所有方法都会受到任何潜在失真的同等影响。

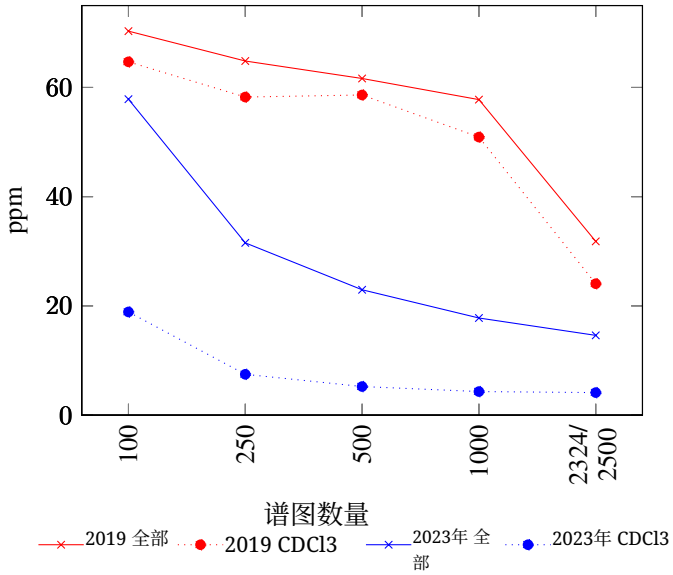


图 5：使用不断增加的样本数量进行 a ^{13}C 核磁共振化学位移预测的平均绝对误差。

4 讨论

本文设计了一种基于消息传递图神经网络的新模型，用于预测化学位移。与现有模型不同，该新网络旨在处理少量训练数据。

使用 ^{19}F 数据测试这一新模型表明，可以将误差率显著降低到标准深度学习模型数值以下。具体而言，当使用来自 nmrshiftdb2的全部 ^{19}F 数据进行训练时，我们的模型达到9.95百万分率的平均绝对误差，而标准深度学习模型的平均绝对误差仅达到57.82百万分率。以类似方式，还可以改进使用特定溶剂时对 ^{13}C 化学位移的预测结果。使用新模型时，所有谱图的平均绝对误差为4.5百万分率，而标准模型仅达到24百万分率。这清楚地表明，我们的模型在较小的数据集上表现显著更好。

分析表4和表5后，我们可以得出结论：使用约500个结构训练新模型即可取得良好效果。这一结果是经验性的，可能取决于结构的多样性等因素，但由于该结论对 ^{19}F 和 ^{13}C 均成立，因此可将其视为一个大致的阈值。

我们的模型仅针对预测 ^{19}F 化学位移数据进行了优化，其他测试案例中则直接使用该模型。这意味着，对于特定的原子核或溶剂等情况，专门定制的模型或许能取得更好的效果。提升性能的一种方法是针对特定数据集调整特征选择。采用单一模型的方法适用于有机化合物的溶液核磁共振研究，也可以针对某一类化合物专门训练模型。另一方面，在某些领域，这种方法可能行不通，例如预测无机化合物或固体时。不过，由于数据可用性是研究中经常面临的问题，这种总体方法仍可能具有实用价值。

在本研究中，我们并未针对不同数据量测试所有目前可用的模型。其中一些模型发表于2019模型之后，并声称使用大型数据集可以取得略好的结果。因此，它们在小型数据集上的表现也可能更好；不过，我们预计它们与2019模型之间的差异微乎其微，因为这些模型并非专门针对较小的数据集构建和测试的。

5 结论

我们介绍了一种新颖的机器学习模型，即使在样本数量较少的情况下，也能够取得合理的核磁共振化学位移预测结果。在 ^{13}C 核磁共振谱上，使用1000个数据点进行训练时，该模型的平均绝对误差小于5百万分率。我们对 ^{19}F 化学位移和特定于溶剂的 ^{13}C 化学位移进行了测试。在所有情况下，我们的模型都优于现有的最先进机器学习模型。通过针对特定数据集优化模型的各个方面，或许还能进一步改进，但在这项工作中，我们希望重点关注即使是

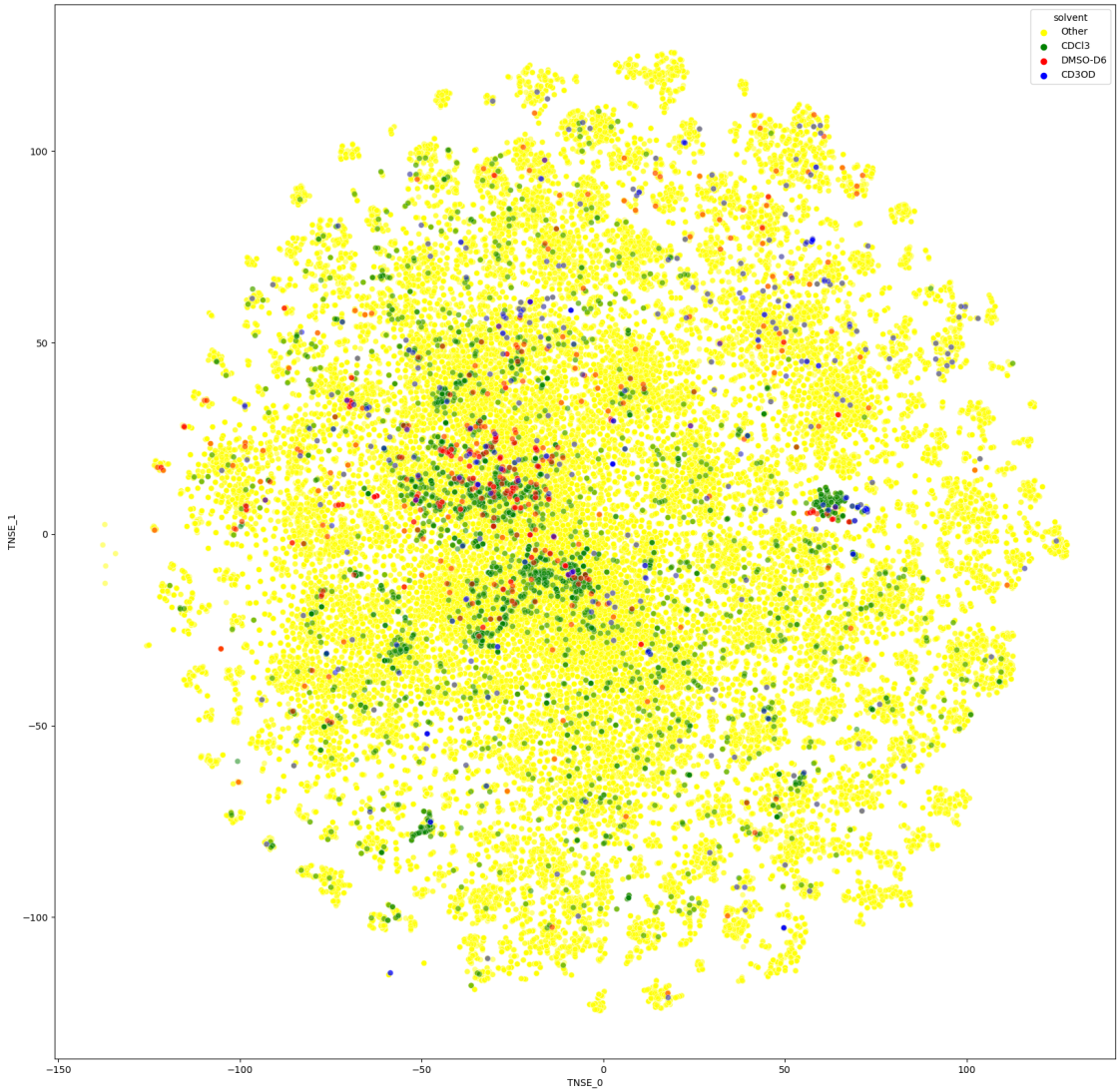


图6：nmrshiftdb2中按溶剂区分的化合物在化学空间中的分布图。计算使用扩展连接性（ECFP）指纹来计算描述符，并使用t分布随机邻域嵌入（t-SNE）进行降维。图中绘制了两个主要成分。使用来自[Simon（2023年（访问于2023年3月8日））]的代码。

通用化模型所达到的效果。这种方法或许可以扩展到其他领域，例如无机化合物，但因此可能需要进一步调整。

致谢

作者感谢参与 UT“机器学习”模块、共同研究这一问题的所有参与者：Karl Kristjan Kaup、Artur Kurvits、Ellen Leib、Dmytro Pashchenko、Kyrylo Riazantsev 和Daniel W¨unsch；同时感谢HolgerA. Scheidt（莱比锡大学）对稿件的帮助。S.K 感谢德蒙福特大学为计算设施提供的资助（VC2020年新员工 L SL 2020年）。

补充信息

项目的代码及说明以 Jupyter笔记本形式提供于
https://colab.research.google.com/drive/1yKTRjpWzR8T199eCokuJfd9Y5o2oNtPp。

参考文献

[Dailey 和 Shoolery（1955年）Dailey，和 Shoolery] Dailey，B. P.；Shoolery，J.N.《美国化学会志》1955年，77，3977–3981。[Kvasnicka 等人（1992年）Kvasnicka、Sklenak 和 Pospichal] Kvasnicka，V.；Sklenak，S.；Pospichal，J.《化学信息与计算机科学杂志》1992年，32，742–747。[Jonas 等人（2022年）Jonas、Kuhn和Schl¨orer] Jonas，E.；Kuhn，S.；Schl¨orer，N.《化学磁共振：MRC》2022年，60，1021—1031。[Tsai等人（2022年）Tsai、Amichetti、Zanardi、Grimson、Daranas 和 Sarotti] Tsai，Y.-H.；Amichetti，M.；Za-nardi，M. M.；Grimson，R.；Daranas，A. H.；Sarotti，A.M.《有机快报》2022年，24，7487–7491，PMID：35508069。[Jonas 和 Kuhn（2019年）Jonas，和Kuhn] Jonas，E.；Kuhn，S.《化学信息学杂志》2019年，11，50。[Unzueta 等人（2021年）Unzueta、Greenwell 和 Beran] Unzueta，P. A.；Greenwell，C.S.；Beran，G. J.O.《化学理论与计算杂志》2021年，17，826–840。[Kwon 等人（2020年）Kwon、Lee、Choi、Kang和 Kang] Kwon，Y.；Lee，D.；Choi，Y.-S.；Kang，M.；Kang，S.《化学信息与建模杂志》2020年，60，2024–2030，PMID：32250618。[Gerrard 等人（2020年）Gerrard、Bratholm、Packer、Mulholland、Glowacki和Butts] Gerrard，W.；Bratholm，L.A.；Packer，M. J.；Mulholland，A.J.；Glowacki，D. R.；Butts，C. P.《化学科学》2020年，11，508–515。[ModgraphConsultantsLtd（2023年（访问日期：2023年2月24日））] Modgraph Consultants Ltd，NMR Predict Desktop。https://www.modgraph.co.uk/product_nmr_desktop.htm，2023年（访问日期：2023年2月24日）。[Guan 等人（2021年）Guan、Shree Sowndarya、Gallegos、St. John 和 Paton] Guan，Y.；ShreeSowndarya，S. V.；Galle-gos，L.C.；St.John，P. C.；Paton，R. S.《化学科学》2021年，12，12012–12026。[Kuhn 等人（2022年）Kuhn、Borges、Venturini和 Sansotera] Kuhn，S.；Borges，R.M.；Venturini，F.；Sansotera，M.《数据集规模与机器学习——开放核磁共振数据库作为案例研究》。2022年 IEEE 第46届年度计算机、软件与应用会议（COMPSAC）。2022年；第1632–1636 页。[Zangger（2015年）] Zangger，K.《核磁共振波谱学进展》2015年，86-87，1–20。[Battaglia 等人（2018年）Battaglia、Hamrick、Bapst、Sanchez-Gonzalez、Zambaldi、Malinowski、Tacchetti、Raposo、Santor

o, Faulkner,Gulc

Battaglia, P. W. 等。关系归纳偏置、深度学习与图网络。2018年；https://arxiv.org/abs/1806.01261.[Gori 等（2005年）Gori、Monfardini 和 Scarselli] Gori, M.;Monfardini, G.; Scarselli, F. 图域学习的新模型。2005年IEEE 国际神经网络联合会议论文集，2005 年。2005年；第 729–734 页，第2 卷。[Fischer 等（2022年）Fischer、Schwarze、Ristic 和 Scheidt] Fischer, M.; Schwarze, B.; Ristic, N.;Scheidt, H. A.计算生物学与化学 **2022年**，*100*，107750。[Duvenaud 等（2015年）Duvenaud、Maclaurin、Aguilera-Iparraguirre、G´omez-Bombarelli、Hirzel、Aspuru-Guzik 和 A

Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; G´omez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. 用于学习分子指纹的图卷积网络。2015；https://arxiv.org/abs/1509.09292。

[{"id": 0," output": "水坝"}]

[Gilmer 等人（2017年）Gilmer、Schoenholz、Riley、Vinyals 和 Dahl] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. 用于量子化学的神经消息传递。2017年；https://arxiv.org/abs/ 1704.01212.[rdk（2023年（访问于2023年2月24日））] RDKit：开源化学信息学工具。2023年（访问于2023年2月24日）；https://www.rdkit.org.[men（2014年至今）] mendeleev——用于查询化学元素、离子和同位素性质的 Python 资源，版本 0.12.1。https://github.com/lmmentel/mendeleev，2014年至今。[Bremser（1978年）] Bremser, W.分析化学学报**1978年**，*103*，355–365。[Willighagen 等人（2017年）Willighagen、Mayfield、Alvarsson、Berg、Carlsson、Jeliazkova、Kuhn、Pluskal、 ´o、Spjuth、Torrance、Evelo、Guh

Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.;´o, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C.*J Cheminform* **2017**，*9*，33。[hos（2023年（访问日期：2023年2月24日））] 使用Python生成分子的HOSE代码。2023年（访问日期：2023年2月24日）；https://github.com/Ratsemaat/HOSE_代码_生成器。[Kuhn 和 Johnson（2019）Kuhn，以及 Johnson] Kuhn, S.; Johnson, S. R. *ACS Omega* **2019**，*4*，7323–7329。[Kuhn 等（2012）Kuhn、Schl¨orer、Kolshorn，以及 Stoll] Kuhn, S.; Schl¨orer, N. E.; Kolshorn, H.; Stoll, R. *Journal ofCheminformatics***2012**，*4*，P52。[Simon（2023年（访问日期：2023年3月8日））] Simon, E. 使用 UMAP 映射化学空间。2023年（访问日期：2023年3月8日）；https://gist.github.com/ElanaPearl/444b3331f61485bbe8862db27cb2b968。