Training and Application of MolBert, a Molecular Structure Understanding Enhancement Model

Article · May 2024

CITATIONS
0

READS
198

5 authors, including:

Xiaoyu Yang
Chinese Academy of Sciences
37 PUBLICATIONS **713** CITATIONS
SEE PROFILE

Qing Peng
King Fahd University of Petroleum and Minerals
320 PUBLICATIONS **6,295** CITATIONS
SEE PROFILE

计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,xxxx,xx(x):x−x [doi: 10.15888/j.cnki.csa.00xxxx]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041.

分子结构理解增强模型 MolBert 的构建及应用验证①

徐早辉 [1,2]，杨小渝 [1,2,*]，王亚鑫 [3]， 谭伟民 [3]，马新杰 [4]，彭庆 [5]

1 中国科学院计算机网络信息中心
2 中国科学院大学
3 中海油常州研究院
4 北京迈高材云科技有限公司
5 中国科学院力学研究所
*通讯作者: 杨小渝, E-mail: kxy@cnic.cn

摘　要: 近年来，使用数据驱动的方法预测分子的性质引发了广泛关注。现阶段分子性质预测研究中传统机器学习算法如随机森林（RF）在回归任务中取得了部分成功。 但因为极度缺乏标注数据导致模型的泛化能力差，容易过拟合，实际预测效果不理想。针对上述问题，受到近年来自然语言处理领域的基于 Transformers 的大规模预训练模型的启发，我们基于 4 亿条 SMILES 非标注序列化分子结构数据，训练出了大规模分子结构理解增强模型 MolBert。经实验验证，仅采用 50 条训练数据的情况下，基于 MolBert 构建的分子性质预测模型的预测效果已和传统 10 万条训练数据构建的模型预测效果相当。作为一个大模型，MolBert 解决了“人工智能+化学”的一些传统瓶颈问题，包括样本数据量小导致预测模型泛化能力差等问题，对人工智能在化学中的应用，将起到极大的促进作用。

关键词: 人工智能+化学；大模型；Transformers; BERT；自监督训练; SMILES；分子性质预测

Training and Application of MolBert, a Molecular Structure Understanding Enhancement Model

Zaohui Xu[1,2], Xiaoyu Yang[1,2,*],Yaxin Wang[3], Weimin Tan[3], Xinjie Ma[4], Pengqing[5]

1 Computer Network Information Center, Chinese Academy of Sciences
2 University of Chinese Academy of Sciences
3 CNOOC Changzhou Paint & Coatings Industry Research Institute Co., Ltd.
4 Beijing MaiGao MatCloud Technolgiy Co. Ltd.
5 Institute of Mechanics, Chinese Academy of Sciences
*Corresponding author: Xiaoyu Yang, E-mail: kxy@cnic.cn

Abstract: In recent years, the data-driven methods for predicting the properties of molecules have attracted extensive attention. Currently, molecular property prediction using traditional machine learning algorithms such as random forest (RF) have achieved partial success in regression tasks. However, the lack of labeled data leads to poor generalization ability of the model, easy overfitting, and unsatisfactory actual prediction results. To address the above problems, inspired by the Transformers-based large-scale pre-training models in natural language processing, we have introduced and trained MolBert, a large-scale molecular structure understanding enhancement model, which is created by 1 billion pieces of unlabeled serialized molecular structure data from SMILES. MolBert is experimentally verified. MolBert can be used to predict the molecular properties of a molecule with only 50 pieces of training data, whose precision is equivalent to that of the traditional model constructed with 100,000 pieces of training data. As a large model, MolBert has overcomed some traditional bottlenecks of AI+Chemistry including poor generalization ability of prediction model due to small sample data, thus it might play a great role in promoting the application of AI in chemistry.
Key words: transformers; BERT; pre-train; fine-tune; SMILES; Prediction of molecular properties

近年来，物质分子的性质预测的研究受到了广泛关注，发展迅速。随着计算机技术的发展和许多新的预测方法和工具的不断涌现，人工智能和机器学习技术在分子性质预测中也得到了广泛应用[1-4]。相比其他学科领域，分子化学领域的整体可探索空间极大， 但是想要预测分子的性质往往需要大量的分子结构和性

① 基金项目:国家自然科学基金“生成式 AI 驱动的高分子智能设计方法与技术研究 ”(编号 62376258)
收稿时间:xxxx-xx-xx;收到修改稿时间:xxxx-xx-xx

质数据构建预测模型，而从一般的实验和理论计算中获得大批量标注数据是比较困难的，需要花费大量的精力或计算资源，因此分子化学领域的数据具有标注数据规模小、多样性高、实验批量获取困难的特征，进而导致了相关领域的开源数据集十分缺乏[1,5]。

在分子性质预测模型方面，随机森林（Random Forest，RF）在回归任务中取得了部分成功，但因为极度缺乏标注数据，同时由于数据较少导致模型的泛化能力差，容易过拟合，实际预测效果不理想[6]。受到近年来自然语言处理领域的基于 Transformers 的大规模预训练模型的启发，我们可以利用大量的非标注序列化分子信息训练出大规模分子结构理解增强模型[7-9]。Transformers 类型的模型利用大量的非标注文本信息对模型进行预训练，使其学习到文本的序列信息，之后下游的相关任务如文本分类、文本情感分析等可以直接对该模型在相关数据集上进行微调（fine-tune）即可，并且可以使用少量标注数据训练可以得到十分惊人的效果。借鉴此种思想，我们从 MatCloud 等平台或数据库获取了 SMILES (Simplified Molecular Input Line Entry System)分子序列信息数据[10,11]，用以训练大规模分子结构理解增强模型，我们采用了 Transformers 模型架构，构建分子结构理解增强模型 MolBert 并在小样本的分子结构标注数据上开展了应用。使用 MolBert 分子结构理解增强模型模型在 50 条训练数据的情况下构建的预测模型，其预测效果已和传统 10 万条训练数据构建的模型预测效果类似。

本论文将探讨大规模预训练模型在化学小样本分子性质预测领域的应用和研究，帮助解决分子性质预测面临的小样本回题，不仅拓宽了有监督训练数据来源，又降低了标注数据的需求量，进一步丰富了分子性质预测的方法和手段。

1 国内外研究现状分析

1.1 材料化学领域标注数据少的小样本问题

小样本问题指的是拥有的标注数据量很少的情况，这个问题通常发生在新兴的材料化学领域，或者是某些特殊材料的研究中，因为这些领域可能没有大量的标注数据可供使用。另外，对于某些特殊的材料，比如某些稀有材料，可能也很难获得大量的标注数据。

小样本问题可能会导致机器学习模型出现过拟合现象，也就是模型在训练数据上表现得很好，但是在真实数据上表现不佳。这是因为模型没有足够的数据来学习，所以可能会学习到训练数据中的噪声或者特征，而不能很好地泛化到真实数据。

另一方面，分子、聚合物的结构数据较多，如 MatCloud 高通量多尺度材料智算平台和一些开源分子序列信息数据库中有大量蕴含分子结构的 SMILES 分子序列数据，可以用以训练大规模分子结构理解增强模型，并使该模型仅需要少量标注数据即可用于下游的分子性质预测。

1.2 预训练分子结构理解增强模型

预训练指的是在没有真实数据的情况下，使用大量的非标注数据来预先训练模型，使模型能够学习到一些通用的特征。预训练模型能够提高模型的泛化能力，也能够减少下游任务的训练时间。目前主要是采用自然语言处理（NLP，Natural Language Processing）的 Transformers 技术进行大规模的模型预训练[9-12]。

1.2.1 Transformers 简介

Transformer 是 Ashish Vaswani 和 Niki Parmar 等人在 2017 年发表的论文《Attention is All You Need》中介绍的一种深度学习模型[7]。它主要用于自然语言处理任务，如语言翻译和文本总结，但也用于各种其他任务，如图像分类。该模型是一个 Seq2Seq（Input a sequence, output a sequence）模型[12]，其核心是自注意力机制，它允许模型在处理文本时聚焦在文本中的特定部分，而不是像传统的卷积神经网络（Convolutional Neural Networks，CNN）或循环神经网络（Recurrent Netural Network，RNN）那样在文本中扫描整个序列。在传统的神经网络中是按顺序处理输入，因此很难处理长输入序列，但 Transformer 中的注意力机制允许模型直接并行处理输入，使其更高效、更有效地处理长序列。Transformer 模型还使用自注意力机制，允许模型直接在输入序列的不同部分之间建立连接。这与依赖经常性连接的传统模型，如 LSTM (Long Short-Term Memory)，形成鲜明对比，LSTM 可能会使远程依赖项建模变得困难。

Transformer 模型由编码器 Encoder 和 解码器 Decoder 两个部分组成，编码器 Encoder 由 6 个编码 block 组成，同样解码器 Decoder 是 6 个解码 block 组成。与所有的生成模型相同的是，编码器的输出会作为解码器的输入。

Transformer 模型的核心思想是自注意力机制（self-attention），它能够有效地捕捉输入序列中不同

位置之间的关系，无论距离有多远。这种注意力机制允许模型根据输入的其他位置信息来赋予每个位置更多或更少的重要性。

Transformer 模型由编码器（Encoder）和解码器（Decoder）两个主要部分组成，通常用于序列到序列的任务，如机器翻译、文本生成等。编码器负责将输入序列编码成一系列特征表示，而解码器则利用这些特征表示生成目标序列。模型中的编码器和解码器均由多层堆叠的自注意力层和前馈神经网络层组成。
在自然语言处理领域，Transformer 模型的一些重要变体包括 BERT、GPT 和 T5 等，它们在各自的任务中都取得了显著的成就。这个架构的成功引发了许多后续模型的发展，成为了深度学习中重要的基础模型之一。

1.2.2 大规模预训练模型

大规模预训练模型是利用 Transformers 的模块进行组合拼接后的模型，它是在大型文本数据集上进行训练。近年来，预训练模型在自然语言处理（NLP）中变得越来越流行，因为它们可以利用大量的非标注数据进行自监督的预训练，而后在少量的训练数据和计算资源的情况下调整适配各种任务。

自监督学习（Self-supervised Learning）是指直接从大规模的无监督数据中挖掘自身监督信息来进行监督学习和训练的一种机器学习方法（可以看成是无监督学习的一种特殊情况），自监督学习需要标签，不过这个标签不来自于人工标注，而是来自于数据本身，如随机掩码（Mask- Language Model）和下句预测（Next-Sentence-Predict）都是常用的自监督训练方式，使得大规模模型能够从海量的非标注文本数据中学习到序列以及结构的知识[10]。

大规模预训练模型中最知名的预训练模型是 BERT[8]，该模型由 12 层 Encoder block 组成，是 Google 于 2018 年开发。 BERT 是一种基于变换器的模型，它使用注意机制来处理前向和后向方向的输入序列。这使得 BERT 能够捕捉文本中单词之间的上下文关系，这对于诸如命名实体识别和文本分类之类的任务非常有用。自 BERT 发布以来，它已成为许多 NLP 领域最先进模型的基础，并已被微调用于各种任务，包括问答，机器翻译、情感分析和信息抽取等领域。

另一个流行的预训练模型是 GPT-3（Generative Pre-trained Transformer 3），该模型由 96 层 Decoder block 组成，是由 OpenAI 在 2020 年开发的[13]。GPT-3 使用序列到序列架构生成类似人类的文本，它在大量网页数据集上进行了训练，并且能够在广泛的领域上生成连贯和连贯的文本，包括翻译，摘要和问答等任务。

除了 BERT 和 GPT-3 之外，还有许多其他预训练模型在近年来得到开发，包括 RoBERTa（A Robustly Optimized BERT Pretraining Approach）[14]，XLNet（eXtreme Transformer Language Model）[15] 和 ALBERT（A Lite BERT）[16]。这些模型在多种 NLP 任务上取得了良好的成绩，并已被研究人员和从业人员广泛使用。

预训练模型的一个关键用途是它们可以在最少的训练数据的情况下为特定任务进行微调。例如，只需几千个标记的例子，BERT 就可以被微调用于情感分析任务，而从头开始训练的模型则需要数万甚至数十万个标记的例子才能达到可比性能。这使得预训练模型特别适合标记训练数据稀缺或难以获得的任务，这也是论文使用大规模预训练模型来实现分子结构理解增强模型的理论基础之一。预训练模型还具有将知识从一个任务转移到另一个任务的优势。例如，训练用于执行自然语言生成的模型可以被微调用于机器翻译，训练用于情感分析的模型可以被微调用于文本分类。

尽管预训练模型具有许多优点，但它们也有一些局限性。一个局限性是它们训练的是大型数据集，使用起来可能资源密集。另一个局限性是预训练模型可能难以应对需要领域特定知识或与模型训练的任务显著不同的任务。例如，在通用网页文本上训练的预训练模型可能在法律文件分类任务上表现不佳，因为它可能没有暴露在这种特定领域的知识中。因此需要使用相关领域的数据对模型进行预训练。

1.3 分子结构理解增强模型的国内外现状

目前在国际上，主要的分子结构预训练模型包括 ChemBerta[17]、MoLFormer[18]等。这些模型旨在通过预训练捕捉分子结构的内部关系。在国内，我们开发的 MolBert 模型也在此领域中占有一席之地。2022 年，我们就提出了利用 MatCloud 平台沉淀的数千万分子结构数据和自研算法进行分子结构数据的合成，参照 BERT 架构用于预训练分子结构模型，以增强下游构

建“结构-性质”预测模型时对分子结构知识的理解（因此我们称其为结构理解增强模型）。自 2022 年下半年开始，我们针对数亿小分子进行预训练和模型的迭代优化。到 2023 年初，我们已经取得显著成果，并启动了相关学生研究工作的开题。

ChemBerta 的研究工作始于 2020 年，而 MolFormer 的工作也在 2022 年前后开始。从模型预测性能来看，MolBert 和 MoLFormer 的表现相近。由于 MolFormer 的预训练数据量较大，其表现略胜一筹；但是两者的表现均优于 ChemBerta（见 4.2 详述）。在数据来源方面，ChemBerta 和 MoLFormer 主要基于现有的结构数据，而 MolBert 不仅使用现有数据，还将引入了自主合成的数据，并且目前正在向高分子领域拓展。从数据量来看， MolBert 的预训练用了 4 亿数据，而 MoLFormer 的预训练用了 11 亿数据， 但 MoLFormer 只是略好于 MolBert(如基于 200 条数据的溶解度预测， MolBert 的 R2 为 0.816， MoLFormer 的 R2 为 0.825，详见见 4.2 详述)。

ChemBerta 是一个基于 transformer 编码器的模型，它通过掩盖语言建模的方式进行预训练从而捕捉分子之间的关系和特征，该模型基于 RoBERTa，相较于 BERT 模型没有使用分子间反应的信息，使用 7 千万的分子 SMILES 序列对模型进行预训练[12]。但由于其收集的分子 SMILES 序列并不够丰富，tokenizer 分词器模型没有针对分子结构如官能团进行特殊设计，并且预训练任务没有对化学式结构和反应方式相贴合。因此其实际效果并非十分理想。它在疏水性的预测任务上，仅取得了 R2 为 0.7 左右的效果。

MoLFormer 是 IBM 基于 Transformer 架构的分子预训练模型，它收集和整理了 11 亿个未标记分子序列上进行了预训练，数据规模较本论文的 4 亿分子数更大，目的是学习如何表征分子。研究结果表明，MoLFormer 能够通过捕捉分子内部原子之间的空间关系，也证明了预训练化学语言模型能够通过 SMILES 字符串预测多种分子性质（包括量子化学性质）。 与 MolBert 相比，其训练数据 2 倍多于 MolBert, 在 200 个样本的训练集和 20 个样本的测试集的实验设置下，MoLFormer 表现略好于 MolBert。

2　技术路线、模型构建以及训练算法

2.1 技术路线

本论文将基于 Transformers 大规模预训练模型，构建分子结构理解增强模型，使得即便不超过 200 条数据的小样本，在采用分子结构理解增强模型后，其预测精度可以达到或超过 10 万量级样本数据的预测精度。

分子属性预测模型的训练可以分为三个步骤：（1）收集大量的分子结构序列数据；（2）基于 Transformers 模型对大量序列化的分子数据进行预训练，得到分子结构理解增强模型；（3）基于初始得到的分子结构理解增强模型，针对有限的分子标注数据对模型进行微调构建分子属性预测模型，并进一步优化分子结构理解增强模型。

我们从 MatCloud 等开源分子结构数据库中，收集了 4 亿个化合物作为大规模序列化分子数据集，用于基于 Transformer 模型进行自监督预训练的大规模序列化分子数据集。预训练前需要重新设计 tokenizer 策略，使其适配分子的序列。同时重新设计模型的结构，减小模型的隐藏层宽度并且增加模型的深度，使其与分子序列相匹配。最后在有标注的数据集上验证最终模型预测的准确率，得到预测效果的数据，并且与基于 10 万量级标注数据训练所得到的传统 RF 随机森林回归模型进行预测精度的性能对比，并不断优化大模型。

为达到上述研究目标，本论文的技术路线如下：

1) 分子结构关键信息提取模型的研发

由于分子序列的特点与传统文本截然不同，因此直接延用传统基于自然语言文本设计的 tokenizer 分词器在分子序列上的应用并不是十分理想，所以需要研究适用于分子结构 SMILES 序列的 tokenizer 分词器，对分子的序列进行 token 字符化，这样才能够使得模型可以读取到分子的序列信息。

2) 预训练算法的设计

研究将设计自监督预训练阶段的算法，针对化学序列的特点进行创新式的设计，让训练方法与化学式结构和反应方式相贴合。

3) 分子结构理解增强模型的构建

研究基于 transformers 的大规模预训练模型的结构，将设计由一层的分子信息嵌入层和多层的分子信息提取层共同组成的分子结构理解增强模型。

4) 下游应用

由分子结构理解增强模型添加微调小模型以及

Fine-tune 微调得到分子性质预测模型。Fine-tune 的关键即是结合预训练大模型设计出贴合下游任务的新模型，再以新模型作为后续任务的基石，完成相关如回归、分类等任务。

5) 分子结构理解增强模型的使用评判和完善。

论文将采集 10 万量级的标注样本并训练相关预测模型，用于评判和完善分子结构理解增强模型。不断优化和完善分子结构理解增强模型，使得即便是不超过 200 的标注数据，在采用分子结构理解增强模型后，其预测精度可达到 10 万量级的标注样本的预测精度。

2.2 预训练模型与性质预测模型的构建

图 1 为分子结构理解增强模型 MolBert 的完整结构，它的设计以及训练流程主要包括以下内容：

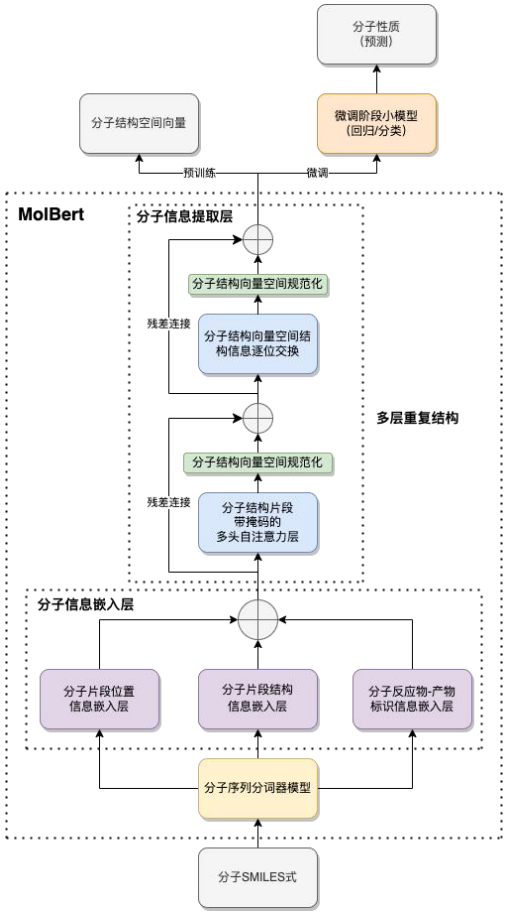

图 1 分子结构理解增强模型 MolBert 结构图

（1）收集和整理的几亿条分子结构 SMILES 数据；

（2）训练用于分子结构关键信息提取的 tokenizer 分词器模型；

（3）用于训练基于 transformers 架构所构建分子结构理解增强模型；

（4）最后使用少量的标注数据训练适用于下游预测任务的微调小模型如多层感知机（MLP，Multilayer Perceptron），即是完整的分子性质预测模型。

因此，总的来说预训练分子结构理解增强模型由一层的分子序列分词器模型、一层的分子信息嵌入层和多层的分子信息提取层共同组成。性质预测模型为基于预训练模型重新添加下游小模型所构建，并经过微调有监督训练得到。

2.2.1 分子序列分词器模型设计

以 SMILES 式表达的分子，需要经过分子 SMILES 式 tokenizer 分词器模型对 SMILES 式进行切分，得到便于模型处理的分子结构片段（token）。这里的难点在于分词器的切分规则，因为一个 SMILES 有多种切分方式。如乙酸乙酯的 SMILES 表达为 O=C(OCC)C，可以切分为[O=C(O,C,C,),C]（1 个序列，5 个片段，片段以逗号区分），也可以切分为[O=C(O,C,C),C]（1 个分子序列，4 个片段）。

与传统 subwords 算法如 BPE、WordPiece 不同，传统方法是依据前后单词、字母出现的频率进行切分[19,20]。而化学元素种类有限，因此使用单个元素作为 token 所代表的信息又太少。因此需要重新设计一个 tokenizer 算法，如依据官能团、有意义的基团作为最基本的组成部分，再以此进行切分，最后依据大量分子 SMILES 序列化信息，选择互信息最大的（也就是两个分子片段关联性最强）分子片段进行组合，扩充分子序列分词器词表，最终达到算法的预期规模。

同时设计将按影响化合物性质的关键因素，进行切分。有机化学中，决定化合物特殊性质的原子或原子团叫官能团。官能团就是原子之间不同结构组合的集团粒子。常见的官能团有羟基、羧基、醛基、酮基、酯基、卤原子、氨基、双键、叁键等。此外，分子中重复单元、手性分子、支链、环状链段、或者其他链段，对性能也有较大的影响。

为此，我们将针对上述几点影响因素，设计了专门的 SMILES 切分算法，对分子 SMILES 式进行切分，并且定义了在训练过程中具有特殊功能的字符，用于训练。<cls>字符会添加到分子序列的开头，用于聚集分子序列信息，下游预测任务以及分类任务均会采用该字符下的隐藏层参数作为输出，再经过重新构建的

微调小模型得到预测的数据结果。<pad>字符用于填充分子序列，由于分子序列长度并不统一，分词器 tokenizer 将会统一将分子序列扩充到指定长度以便于后续的处理。<sep>字符用于分割分子序列，如在预训练阶段的分子反应预测任务时将会有多条分子序列作为一段输入，分子序列拼接处将会使用<sep>字符作为分割。<unk>字符将作为在输入阶段不在 tokenizer 词表内的片段的替代，用于解决词表不足的情况。<mask>字符为掩蔽字符，在预训练阶段的掩码预测任务中，会将被掩蔽的 token 替换为<mask>字符，用于后续的预训练任务。特殊字符列表如表 1。

表 1 特殊字符

特殊字符	编号	含义
<cls>	0	聚集分子序列信息
<pad>	1	对齐分子序列长度
<sep>	2	分割分子序列或为终点
<unk>	3	未知分子序列片段
<mask>	4	遮蔽分子序列片段

如在对乙酸乙酯为例子，分子序列分词器切分为[<cls>,O=C(O,C,C,),C,<sep>]，最大长度为 10 的情况下需要进行长度对齐，即需要加入三个<pad>字符，变为[<cls>,O=C(O,C,C,),C,<sep>,<pad>, <pad>,<pad>]。当有片段不属于词表时会将对应的片段替换为<unk>，在训练阶段对不同片段进行遮蔽时会将对应的片段替换为<mask>。

2.2.2 分子信息嵌入层

分子信息嵌入层的作用是将上一步分子 SMILES 式 tokenizer 分词器模型所得到的分子结构 token 片段序列进行向量化，赋予每个分子结构 token 片段各种信息。分子信息嵌入层分为分子片段结构信息嵌入层、分子片段位置信息嵌入层以及分子反应物-产物标识信息嵌入层三种。

分子片段结构信息嵌入层负责将切分后生成的分子结构 token 片段转化为算法模型可识别的向量空间，参数规模为(batch_num, seq_len, hidden_state)。以乙酸乙酯为例子切分为[<cls>,O=C(O,C,C,),C,<sep>]，向量空间为(1, 7, hidden_state)，1 代表 batch size，即一次输入给模型的数据量，多条分子 SMILES 式同时进行训练或性质预测时可以为更大的数字，7 代表分子 SMILES 式由上一步得到的 token 片段序列的 token 个数，并且一般会由特殊字符<pad>扩充到最大长度如 128，hidden_state 代表各个 token 的向量宽度，也是以下自注意力层、信息嵌入层等的参数宽度。该层赋予了分子结构 token 片段最基本的标识信息，即赋予了区分不同片段的性质的能力，特别是官能团片段。

分子片段位置信息嵌入层用以提供区分该片段在 SMILES 式的位置信息的能力，即分子的性质是与分子结构片段的位置相关。以乙酸乙酯为例，最简单的位置信息可以为[1,2,3,4,5,6,7]。

分子反应物-产物标识信息嵌入层，用于区分预训练阶段的反应物与产物的信息。如乙酸与乙醇发生酯化反应 CC(O)=O + CCO -> O=C(OCC)C 经过分子 SMILES 式 tokenizer 分词器模型切分为[<cls>,c, C(O)=O,<sep>, C,C,O, <sep>, O=C(O,C,C,),C]，则标识信息向量为[1,1,1,1,1,1,1,1,0,0,0,0,0]。

分子片段结构信息嵌入层、分子片段位置信息嵌入层、分子反应物-产物标识信息嵌入层共同组成了分子信息嵌入层，将分子 SMILES 式 tokenizer 分词器模型切分得到的分子结构 token 片段赋予各种信息，最终得到模型训练输入的分子结构向量空间。

2.2.3 分子信息提取层

分子信息提取层是分子结构理解增强模型最关键的结构，其作用是提取分子结构 token 片段的性质信息，并且可以得到分子结构 token 片段之间的关联信息。

分子结构片段多头自注意力层将会对分子信息嵌入层给到的分子信息进行重要性的计算，并且在隐藏层纬度将分子信息进行切分即“多头”，使不同的注意力头可以计算到不同向量空间的值。该层主要用于计算不同分子片段之间的注意力参数，即计算出哪些片段是更为关键的信息，同时多头结构可以让模型注意到不同维度的信息，再汇总综合判断不同分子片段的重要性。

分子结构向量空间规范化用于处理分子结构片段向量。分子结构片段向量空间经规范化处理后，得到均值为 0，方差为 1 的输出，目的是让模型的参数处于较为合理的数值区间，使得模型始终是处于一个较快的收敛区间内加快模型的收敛，让模型训练速度和效果更好。

残差连接（Skip-Connect）可以降低模型复杂度以减少过拟合，并且可以防止梯度消失，使得模型可以训练的更深效果更好[21]。残差连接的函数表达式是 g(x)=f(x)+x，x 为主模型输入前的向量空间，f(x)即是

经过主模型计算后的结果，而残差连接的做法是将其相加然后作为整体输出。

分子结构向量空间结构信息逐位交换作用是添加了非线性映射，以增强模型的表达能力，更加深入的提取特征信息。在分子结构片段多头自注意力层中，主要是进行矩阵乘法，即都是线性变换，而线性变换的学习能力不如非线性变换的学习能力强。它的输入是一个词向量，经过一系列线性变换和激活函数处理之后，输出另一个词向量。主要是交换模型参数间的信息，考虑注意力机制可能对复杂过程的拟合程度不够，通过增加两层网络来增强模型的能力。

分子信息提取层可以进一步叠加层数，上一层的输出可以作为下一层的输入，参数总量将会达到千万级别，增加模型可以学习到的分子结构信息。最终，输出一个包含分子结构信息的 3 维向量空间，参数规模为(batch_num, seq_len, hidden_state)。

2.2.4 微调模型设计

为了贴合下游任务，需要在上述原有的分子结构理解增强模型的基础上，设计出 Fine-tune 微调的小模型。

考虑到多层感知机（MLP，Multilayer Perceptron）具有很强的隐式交叉能力，将所有保留将稠密特征和离散特征的 Embedding 一起输入到 MLP，以隐式的方式学习其非线性表达。

可以使用分子结构理解增强模型+前馈神经网络得出预测数值作为回归任务的新模型如图 2 结构。

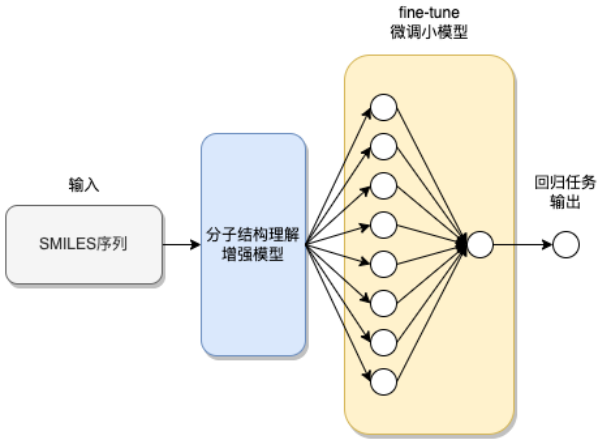

图 2 Fine-tune 微调回归模型示意图

2.3 训练算法

2.3.1 分子结构理解增强模型预训练算法

分子结构理解增强模型是一种预训练模型，该模型需要大量的分子序列数据，以及合适的预训练算法使得模型能够学习到普适的分子结构知识，从而迁移用于下游的只需要少量标注信息的分子性质预测任务的训练。预训练算法分为掩码-预测训练和反应物-产物预测训练两种。

首先是掩码-预测训练，对于海量 SMILES 化学式序列信息进行 tokenizer 之后的部分 token 进行遮掩和预测，使模型能够学习到化学物质普遍的 SMILES 序列化的结构信息。

其次是反应物-产物预测训练，反应数据取自化学反应数据库。反应预测可以看作是一个 NMT（Neural Machine Translation，神经机器翻译）任务，其中反应物是一种语言，产物是另一种语言。多个反应物或产物可以使用特殊字符[SEP]进行分割，这个训练任务更多是学习 SMILES 序列化的性质信息。

2.3.2 分子性质预测模型 Fine-tune 算法

Fine-tune 即为微调，是深度学习领域的专业名词，需要在原有大模型的基础上，贴合下游任务，设计出符合需求的新模型，再在新模型的基础之上使用少量的样本对整个新模型进行训练。在此过程中原本预训练大模型的参数也会参与训练，由于此次参与有监督训练的训练集远小于预训练时的训练集，因此被称为 Fine-tune 微调。

分子结构理解增强模型+前馈神经网络得出预测数值作为回归任务使用均方差 MSELoss，作为模型最终的损失函数，记录预测值与真实值的均方差，以及作为后续反向传播微调更新模型的依据。

分子结构理解增强模型+前馈神经网络+softmax 函数得出不同类型的概率作为分类任务使用交叉熵 Cross Entropy loss，作为模型最终的损失函数，记录预测值与真实值的均方差，以及作为后续反向传播微调更新模型的依据。

3 训练环境以及细节设置和部署

3.1 训练环境

本论文代码运行的外部环境为中国科技云平台（China Science & Technology Cloud），代码运行在 python 3.8、pytorch 1.10、cuda 10.1 电脑环境下，预训练阶段采用大约 4 亿条分子序列数据，使用 4 张英伟达 P100 训练 5 个训练周期（一次完整的在训练集上的训练为一个训练周期）共用时 4-5 天，Fine-tune 阶段

使用 1 张英伟达 P100 训练 5 个 epoch 共用时 4-5 小时。

3.2 MolBert 模型有关参数以及文件

模型的参数如下，分词器模型 tokenizer 词表大小为 10000 个 token，模型的序列长度为 512，分子结构片段多头子注意力层宽度为 768，多头数目为 6，分子信息逐位交换层宽度为 3072，整体模型为 6 层结构，完整模型的参数量约为 5000 万。

Tokenizer 分词器模型保存为 tokenizer.json 和 vocab.txt 两个文件，预训练分子结构理解增强模型保存为 pth 文件，Fine-tune 微调后的分子性质预测模型保存为 pth 文件，具体文件列表以及作用可见下表 2.

表 2 模型文件列表

模型	文件	作用
Tokenizer	tokenizer.json	根据编码生成 one-hot 向量
	vocab.txt	词典文件，仅为单个字符
MolBert	model.pth	预训练模型
Finetune r	model_r.pth	微调后的回归模型
Finetune c	model_c.pth	微调后的分类模型

3.3 模型部署

模型可以作为本地部署和在线部署，本地部署即在当前使用的电脑上进行模型的运算以及推理任务，而远程部署即将模型算法安置在服务器上，用户使用浏览器远程访问使用即可。

本地部署：用户在自己的电脑上使用 PyTorch 的 torch.load 函数加载经过微调训练的模型文件（.pth），以及使用 transformers 库的 tokenizer_load 函数加载分词器模型（包括 tokenizer.json 和 vocab.txt）。然后，使用分词器处理 SMILES 分子序列，将其转换为模型可理解的数值输入。最后，输入到经过微调的预训练模型中，得到分子的目标结构预测，可以是具体的数值或分类概率。

在线部署（Model as a Service）：首先在服务器上完成多种分子性质预测模型的本地部署。用户通过浏览器访问，选择想要预测的分子数据类型（如溶解度、疏水参数、极性表面积等），然后输入分子的 SMILES 式。提交后，系统使用分词器处理 SMILES 式，再由分子性质预测模型进行快速预测（约 1-2 秒）。最终，预测结果会展示在用户界面上。

两种部署方式都利用了分词器和微调后的预训练模型，但在线部署提供了更便捷的远程访问和使用方式。

4 实验方式以及实验结果分析

4.1 实验方式

基于分子结构理解增强模型 MolBert 的分子性质预测模型、ChemBerta 与传统随机森林回归模型 RF 在溶解度（mol/L）、疏水参数（XlogP）、极性表面积（PSA）三项分子性质预测问题上的对比。

分子结构理解增强模型采用 200 条数据作为训练集进行 5 个 epoch 的有监督训练并且采用 50 条作为验证集对模型训练结果进行校验，RF 采用 10 万条数据进行有监督训练。分子结构理解增强模型直接使用 SMILES 分子结构序列作为输入，而 RF 提前使用开源分子性质提取工具 RDKit 提取化学分子中的特征信息如电荷、极性表面积、亲脂性等等。

训练完成后统一采用 200 条的测试集验证模型的分子性质预测能力，预测定量分析算法选择均方误差(Mean Square Error, MSE)以及可决系数 (coefficient of determination, R2)。

4.2 实验结果分析

MSE 均方误差是指参数估计值与参数真值之差平方的期望值，即可以理解为分子性质预测值与实际值的差距，因此越小越好。如图 3 可见分子结构理解增强模型 MolBert 在 200 条小样本有监督训练下误差明显小于 ChemBerta 以及传统模型 RF 在 10 万量级样本下训练的预测效果。由于 MolFormer 模型采用了 11 亿的数据进行预训练（2 倍于 MolBert 预训练的数据量），其预测效果略好于 MolBert 的预测效果。

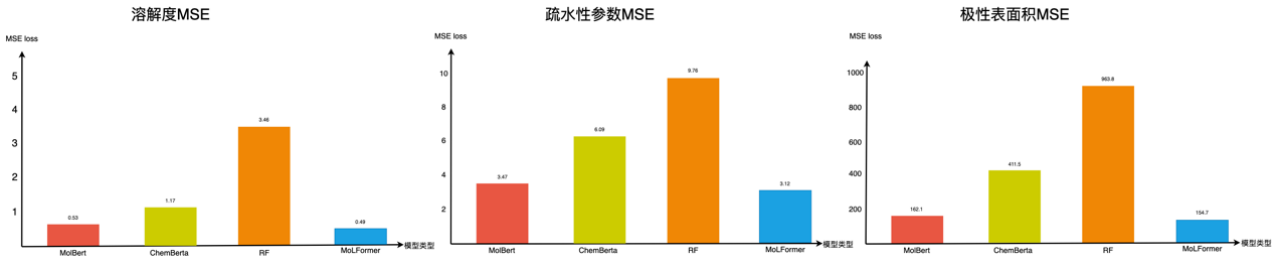

图 3 各模型在三种属性测试集的 MSE 数据对比

可决系数，亦称测定系数、决定系数、可决指数，表示一个随机变量与多个随机变量关系的数字特征，用来反映回归模式说明因变量变化可靠程度的一个统计指标。因此该数值越大则回归模型的拟合程度越高。如图 4 所示，分子结构理解增强模型 MolBert 的拟合程度明显高于 ChemBerta 以及传统 RF 模型，同样证明了在小样本下 MolBert 性能的优越性，但同样的效果不及最新的 MolFormer 模型

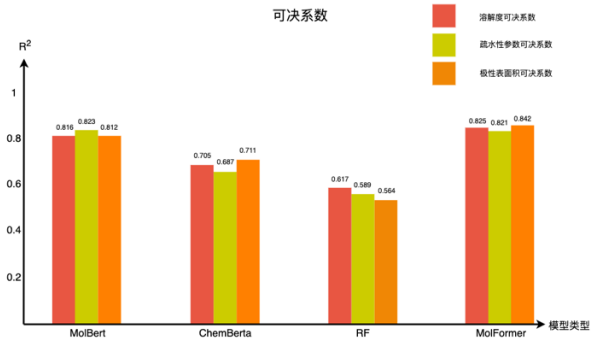

图 4 各模型在三种属性测试集的可决系数上的对比

本论文选择 100 条未在训练集样本以及验证集样本中的分子疏水参数(XlogP)真实值与分子结构理解增强模型以及传统 RF 模型的预测值，得到如图 5 的对比数据。

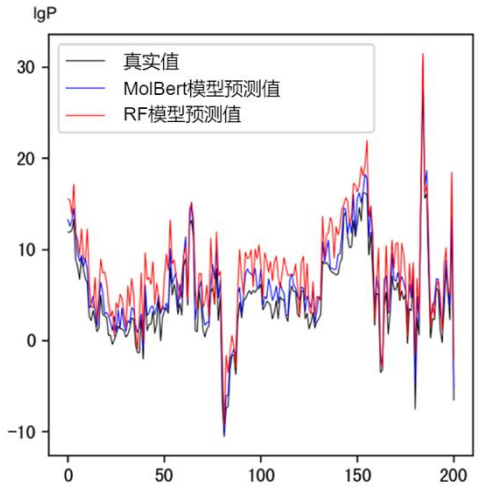

图 5 200 条疏水参数真实值与模型预测值对比数据

由数据以及折线图可以看出，分子结构理解增强模型预测分子性质的误差明显小于传统 RF 模型，并且分子结构理解增强模型所预测的分子性质变化趋势也更加贴合真实数据。由此可以得出基本结论，基于分子结构理解增强模型可以使得在较小样本的微调下，其分子性质的预测能力将超过传统 RF 模型在 10 万条数据下的预测能力。

当采用分子结构理解增强模型 MolBert 训练分子性质预测模型时，进一步探究训练集大小对预测性能的影响。分别采用 50 条、200 条、500 条、1000 条以及 2000 条训练集做有监督微调训练，统一采用 200 条数据作为评测集，各个模型在三种数据集上的可决系数的变化趋势如图 6。

可见分子结构理解增强模型 MolBert 在 50 条训练数据的情况下其预测效果已和传统模型 10 万条训练数据得到的效果类似，而进一步增加训练数据所达到的瓶颈在 86%左右。

而 ChemBerta 模型在各个训练集数据量情况下分子性质预测任务的可决系数的变化趋势，其预测效果明显弱于分子结构理解增强模型在相同数据量下的预测能力。

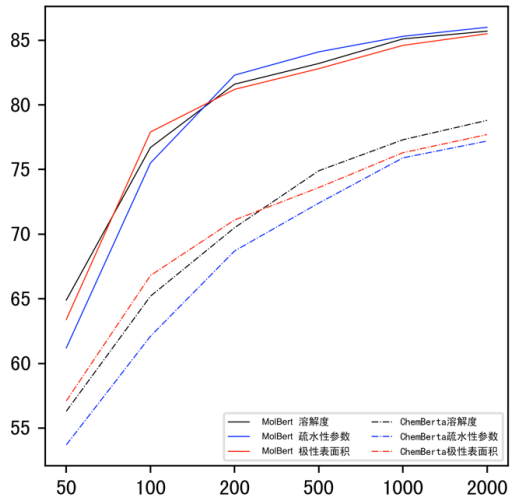

图 6 可决系数的变化趋势

5 总结

本论文基于自然语言处理中的 Transformers 大规模预训练模型的自监督学习机制，构建了分子结构理解模型 MolBert。该模型能够从大量的非标注的分子序列中学习到的分子结构知识，并且通过迁移学习下游任务可以通过微调方式，迁移用到回归任务以及分类任务模型的构建中，从而可大幅减少下游有监督训练任务的数据量。在 200 条小样本数据的情况下，模型的预测效果优于 10 万条样本训练下的传统随机森林 RF 模型，证明了分子结构理解增强模型“知识迁移”的学习效果。

本论文得到的预训练分子结构理解模型可以方便的应用于各种性质领域的预测任务，得益于该模型的可迁移性，只需要在特定领域的小样本下对模型进行微调即可实现较好的预测效果，可以大大减少分子性质预测所需要的数据量以及算力成本。因此该项研究在新兴前沿领域如分子制药、材料化学等等方面都推动作用，对于“人工智能+化学”的跨学科交叉研究，有着积极的意义。

由于目前所掌握的开源分子序列数据仍然十分有限，对于大规模预训练模型来说仍然可以继续增大训练数据以及模型规模来进一步提高模型的效果，引入更大量的开源数据，并针对垂类领域如聚合物分子，合成大量数据，提高在专业领域下的准确性。在未来我们将会进一步探索预训练模型规模的上限，并且调整模型结构以及相关分子信息提取算法以加深在化学领域的适应性，以达到更加出色的效果。

参考文献

1 Y Q. Yao, Z. Shen, Y. Wang and D. Dou, "Property-Aware Relation Networks for Few-Shot Molecular Property Prediction," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2024.3368090, 2024

2 K.Yang, K. Swanson, W.Jin, et al. Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model. 2019 Aug 26;59(8):3370-3388. doi: 10.1021/acs.jcim.9b00237.

3.A. Radford, K. Narasimhan, T. Salimans et al. Improving Language Understanding by Generative Pretraning,https://www.cs.ubc.ca/~amuham01/LING530/papers/redford2018improving.pdf, 2020

4. C.Bilodeau C. W. Jin, T. Jaakkola et al. Generative models for molecular discovery: Recent advances and challenges. WIREs Comput Mol Sci. 2022; e1608

5 Y. Rong, Y. Bian, T. Xu, “Self-Supervised Graph Transformer on Large-Scale Molecular Data”, 34th Conference on Neural Information Processing Systems (NeurIPS 2020)

6 T. Julie and S.Wager. “GENERALIZED RANDOM FORESTS By Susan Athey.” (2018).

7 Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30

8 J.Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

9 W.Liu, P. Zhou, Z.Zhao, Z.Wang, et al. K-BERT: Enabling Language Representation with Knowledge Graph. Proceedings of the AAAI Conference on Artificial Intelligence, 34(03), 2901-2908,2020 https://doi.org/10.1609/aaai.v34i03.5681

10. X. Yang, Z. Wang, X. Zhao et al., 《MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources》, Computational Materials Science, Vol146, Page 319–333, April 2018, doi: 10.1016/j.commatsci.2018.01.039.

11 杨小渝,林海青,王娟,等.支撑材料基因工程的高通量材料集成计算平台[J].计算物理,2017, 34(6):8.DOI:CNKI:SUN:JSWL.0.2017-06-009.

12. Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney and Daniel S. Weld. “S2ORC: The Semantic Scholar Open Research Corpus.” Annual Meeting of the Association for Computational Linguistics , 2020

13 Tom B. Brown, Benjamin Mann, Nick Ryder, Language Models are Few-Shot Learners,NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing SystemsDecember 2020Article No.: 159Pages 1877–1901

14 L. Zhuang, L. Wayne, S. Y, Z. Jun, “A Robustly Optimized BERT Pre-training Approach with Post-training”, Proceedings of the 20th Chinese National Conference on Computational Linguistics, Huhhot, China: Chinese Information Processing Society of China, Aug.2021, pp

1218–1227.

15 Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdino, Q. V. Le, 《XLNet: Generalized Autoregressive Pretraining for Language Understanding》, Advances in Neural Information Processing Systems, Curran Associates, Inc., 2019.

16 Z. Lan, M. Chen, S. Goodman et al. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.” ArXiv abs/1909.11942 (2019): n. pag.

17 S. Chithrananda, G. Grand, B. Ramsundar, @ChemBerta: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction”. arXiv, Oct, 2020 doi: 10.48550/arXiv.2010.09885.

18. J. Ross, B. Belgodere, V. Chenthamarakshan et al. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 4, 1256–1264 (2022). https://doi.org/10.1038/s42256-022-00580-7

19 R. Sennrich, B. Haddow, and A. Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

20 M. Schuster,K. Nakajima, “Japanese and Korean voice search”, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2012, pp 5149–5152. doi: 10.1109/ICASSP.2012.6289079.

21 K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.