BioT5+：利用 IUPAC 集成和多任务调整实现通用生物理解

Qizhi Pei[1] , Lijun Wu[2*] , Kaiyuan Gao[3] , Xiaozhuan Liang[4] , Yin Fang[4] , Jinhua Zhu[5] , Shufang Xie[1] , Tao Qin[2] , Rui Yan *[1,6]

[1]中国人民大学高岭学院[2] 微软研究院 [3] 华中科技大学[4] 浙江大学 中国科学技术大学 [56] 下一代智能搜索工程技术研究中心

和建议，教育部

{qizhipei,shufangxie,ruiyan}@ruc.edu.cn
apeterswu@gmail.com
im_kai@hust.edu.cn {liangxiaozhuan,fangyin}@zju.edu.cn
teslazhu@mail.ustc.edu.cn taoqin@microsoft.com

摘要

最近，计算生物学的研究趋势越来越多地集中在整合文本和生物实体建模上，尤其是在分子和蛋白质方面。然而，像 BioT5 这样的前人所做的努力面临着在不同任务中进行泛化的难题，而且缺乏对分子结构的细致理解，特别是对其文本代表（如 IUPAC）的理解。本文介绍了 BioT5+，它是 BioT5 框架的扩展，专为加强生物研究和药物发现而量身定制。BioT5+ 融合了几项新功能：整合 IUPAC 名称以促进分子理解、纳入来自 bioRxiv 和 PubChem 等来源的大量生物文本和分子数据、多任务结构调整以实现跨任务的通用性，以及数字标记化技术以提高数字数据的处理能力。这些改进使 BioT5+ 能够弥合分子表征与其文本描述之间的差距，提供对生物实体更全面的理解，并在很大程度上改进了生物文本和生物序列的基础推理。该模型经过大量实验的预训练和微调，包括 *3 类问题（分类、回归、生成）*、*15 种任务和 21 个基准数据集*，在大多数情况下*都*表现出卓越的性能和最先进的结果。BioT5+ 能够捕捉生物数据中错综复杂的关系，从而为生物信息学和计算生物学做出了重大贡献。我们的代码可在 https://github.com/QizhiPei/BioT5.

* 通讯作者：Lijun Wu (apeterswu@gmail. com) and Rui Yan (ruiyan@ruc.edu.cn)

1 导言

分子和蛋白质是药物发现中的两个关键生物实体，是生物活动的基础（Dara 等，2022 年；AI4Science 和 Quantum，2023 年）。分子可以用 SMILES（Weininger，1988 年；Weininger 等人，1989 年）或 SELFIES（Krenn 等人，2020 年）序列表示，蛋白质可以用 FASTA（Lipman 和 Pearson，1985 年；Pearson 和 Lipman，1988 年）序列描述。随着语言模型（LMs）的发展，越来越多的工作侧重于通过对分子和蛋白质的生物序列建模来理解它们（Chithrananda 等人，2020 年；Rives 等人，2021 年；Lin 等人，2022 年）。

值得注意的是，生物文献（Canese and Weis, 2013; White, 2020）中有大量关于分子和蛋白质的信息。当这些文献中提到一个生物实体时，其上下文主要围绕着对该实体某些特征的描述。因此，越来越多的工作致力于文本和生物实体的联合建模（Pei 等人，2024 年），如 Galactica（Tay-lor 等人，2022 年）、MolXPT（Liu 等人，2023 年 c）、BioT5（Pei 等人，2023 年）和 BioMedGPT（Luo 等人，2023 年 c），它们都是基于文本、分子和蛋白质序列训练的科学模型。尽管取得了这些成就，但仍有大量改进的机会：（1）先前的工作忽视了分子文本名称建模的重要性，如国际纯粹与应用化学联合会（IUPAC）提供了标准和系统的命名方法，以确保整个科学界的统一性和清晰性。与 SMILES 和

IUPAC 和 SELFIES 与自然语言更为相似，这一点从其在科学文献中的广泛应用就可见一斑（Klinger 等，2008 年）。(2) 以前的模型主要是专业模型，需要为每个下游任务训练一个单独的模型，因此缺乏通用性，增加了训练和开发成本（刘等人，2023c；裴等人，2023）。(3) 以往基于 T5（Raffel 等人，2020 年）和 GPT（Brown 等人，2020 年）架构的模型大多只关注分类任务，因为它们没有对数值数据进行专门的标记化，这导致它们对回归任务的适应性不理想。

为了应对上述挑战，我们在本文中介绍了 BioT5+，它是 BioT5 框架（Pei 等人，2023 年）的高级迭代版本，旨在通过丰富的数据集成、多任务能力和解决回归任务的能力来增强生物研究和药物发现。简而言之，BioT5+ 包含以下重大改进：

(1) *增强对分子的理解*：通过将 IUPAC 名称整合到 BioT5+ 框架中，该模型可以更深入地理解分子结构。这种整合使 BioT5+ 能够解释通常出现在科学文献中的化学名称，缩小了正式分子表征（如 SELFIES）与文本描述之间的差距。因此，这增强了对分子的理解，有助于对分子特性和活性进行更准确的预测和分析。

(2) *扩展的生物文本和分子数据*：与 BioT5 相比，BioT5+ 包含来自 bioRxiv（Sever 等人，2019 年）和 PubMed（Canese 和 Weis，2013 年；White，2020 年）等来源的大量生物文本数据，以及来自 PubChem（Kim 等人，2019 年）的高质量分子数据。这种扩展不仅扩大了模型的知识基础，还丰富了对生物实体的文字理解。

(3) *多任务指令调整*：BioT5+ 针对下游任务采用多任务指令调整策略，而不是针对每个任务单独进行专门的模型训练。通过利用统一的多任务训练框架，BioT5+ 可以无缝整合来自不同任务的知识，增强其在不同生物和化学领域的预测能力和泛化能力。

(4) *高级数字标记化*：过度

鉴于数值表示法的局限性，BioT5+ 从 Llama（Touvron 等人，2023a）模型中汲取灵感，整合了先进的基于字符的数值标记化策略。这种技术可以更细致、更一致地表示数值。通过我们设计的预训练和多任务指令调整，BioT5+ 的有效性在 *3 类问题（分类、生成和回归）*、*15 种不同任务和 21 个基准数据集上得到了*验证，包括分子属性预测、逆合成、分子描述生成、药物与靶标相互作用等。BioT5+ 显示了极具竞争力的结果，在大多数任务中都达到了最先进的性能。这种强大的性能凸显了 BioT5+ 在捕捉和分析生物数据固有的复杂关系和属性方面的更强能力，标志着 BioT5+ 在生物数据分析领域的重大突破。在计算生物学领域向前迈进了一步。

2 相关工作

2.1 生物跨模态模型

最近，LLM 的进步促使人们更加关注分子、原蛋白和文本的联合建模，旨在通过文本增强对生物实体的理解。

分子-文本。MolT5（Edwards等人，2022年）使用T5（Raffel等人，2020年）屏蔽跨度预测目标，对一般文本和分子SMILES进行联合训练。MoMu（Su 等人，2022 年）采用了分子图和相关文本的对比学习，而 MolFM（Luo 等人，2023 年 b）则进一步将知识图嵌入到分子表示中。MolXPT （Liu 等人，2023c）使用 GPT（Brown 等人，2020 年）框架对分子 SMILES 和包装文本进行联合训练。MolCA（Liu等人，2023d）通过跨模态投影仪和单模态适配器整合二维分子图感知，增强了LM。GIT-Mol（Liu 等人，2023a）是一种多模态 LLM，可协同图形、图像、SMILES 和分子标题。Text+Chem T5（Christofidel- lis 等人，2023 年）是一种多领域、多任务语言模型，能够同时处理分子和自然语言。

蛋白质-文本有几项著名的研究集中于蛋白质和文本的联合建模。ProteinDT （Liu等人，2023b）提出了一个文本引导的蛋白质设计框架。BioTranslator （Xu 等人，2023b）是一个跨模态翻译系统，可注释

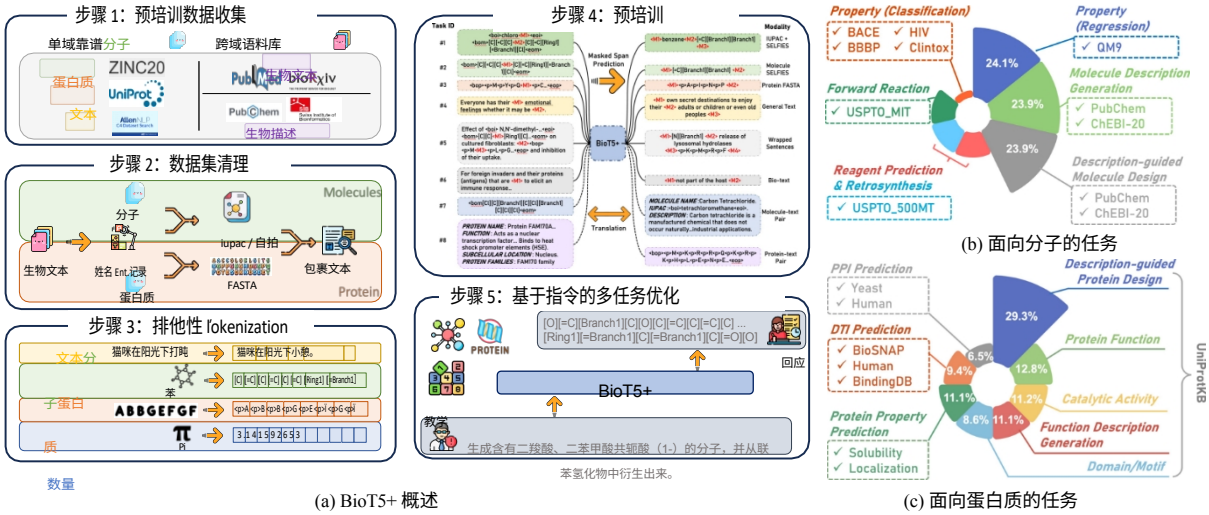


图 1：(a)：BioT5+ 框架概览。(b) (c)：BioT5+ 下游任务的组成，分为两类：(b) 面向分子的任务和 (c) 面向蛋白质的任务。任务名称及其指导数据集和各自所占百分比均注释在所附饼状图的每个部分附近。

使用文本描述生成各种生物实例。Prot2Text（Abdine 等人，2023 年）在编码器-解码器框架中结合了 GNN 和 LLM，以自由文本方式生成蛋白质功能。

除上述模型外，还有一些模型是在更多样化的模态中训练出来的：DeepEIK （Luo 等人，2023a）是一种多模态模型，它整合了来自药物、蛋白质和文本等多模态输入的特征。BioT5（Pei等人，2023年）是一个基于T5（Raffel等人，2020年）的模型，它对文本、分子SELFIES和蛋白质FASTA序列进行联合训练，有效地缩小了文本数据和生物数据之间的差距。

尽管这些模型取得了成功，但它们对单一任务训练的依赖限制了它们的通用性，并阻碍了计算生物学中更具通用性和适应性的方法的发展。

2.2 生物任务的指令调整

指令调整是一种应用于预训练 LLM 的流行技术，即使用专门的指令数据集对 LLM 进行训练，从而使 LLM 具备理解特定任务指令的能力。最近，人们对探索各种生物任务的指令调整越来越感兴趣。其中值得注意的是 Mol-Instructions 的开发（Fang 等人，2023 年），这是一个专门为生物领域设计的综合指令数据集，其中包括面向分子的指令、面向蛋白质的指令和生物分子文本指令。InstructMol（Cao 等人，2023 年）是一种多模态 LLM，利用指令调整来对齐分子图、分子 SELFIES 和自然语言。与这些方法不同，我们的 BioT5+ 专门针对生物领域进行了预训练。在双

逻辑指令使 BioT5+ 不仅能理解生物实体，还能在各种生物任务中进行泛化。

3 BioT5+ 框架

本节介绍 BioT5+ 框架，图 1 是其概览。图 3 介绍了预训练所涉及的任务。

3.1 IUPAC 整合的直觉

国际理论化学和应用化学联合会（IUPAC）命名系统提供了一套标准化的化合物命名规则，可精确描述分子结构及其组成部分（官能团、链和环），是化学命名法的基石。通常，国际理论化学和应用化学联合会（IUPAC）的名称是由单个分子的组成部 分名称构成的，反映了分子的结构。其中包括表示各种化学基团和结构特征的前缀、后缀和后缀，对分子进行全面描述。例如，阿司匹林的 IUPAC 名称是 "2-乙酰氧基苯甲酸"。这里的 "2-乙酰氧基 "指的是苯环第二个碳上的乙酰氧基，而 "苯甲酸 "则表示苯环上带有一个羧酸基团。IUPAC 名称与自然语言的相似性，加上它们在科学文献中的广泛使用，使它们成为模型预训练的理想对象。通过在包含 IUPAC 名称的文献中对模型进行预训练，BioT5+ 可以建立对分子与其化学性质的各种文字描述之间关系的细致理解。

3.2 训练前语料库

作为 BioT5 的扩展，BioT5+ 的大部分预训练语料与 BioT5 相同、

因此，我们将简要提及常见要素，而主要关注 BioT5+ 中引入的新功能。

预培训语料库由 4 个类别组成：(1) *单模态数据*，包括来自 PubChem 的带有 IUPAC 名称的分子 SELF- IES（Kim 等人，2019 年）、来自 ZINC20 的分子 SELFIES（Irwin 等人，2020 年）、来自 Uniref50 的蛋白质 FASTA（Suzek 等人，2007 年）以及来自 "Colossal Clean Crawled Corpus"（C4）（Raffel 等人，2020 年）的一般文本。对于来自 PubChem 的分子，我们将 IUPAC 名称和 SELFIES 连接起来进行预训练，如图 3 所示。(2) *封装文本*，其中分子或基因/蛋白质名称后缀有相应的序列表示。我们使用 BERN2（Sung 等人，2022 年）--一种基于神经的双逻辑领域命名实体识别（NER）系统--来检测 PubMed（White，2020 年）和 bioRxiv（Sever 等人，2019 年）摘要中出现的分子和蛋白质并对其进行分类。对于分子名称，我们首先将其标准化为 IUPAC 名称，然后附加相应的 SELFIES。对于基因/蛋白质名称，我们将直接附加其 FASTA 序列。为了生成高质量的包装文本，我们还分析了 BERN2（Sung 等人，2022 年）预测的置信度得分分布。只有那些置信度分数较高的实体才会被保留下来，以确保附加序列数据的准确性和有效性。(3) *生物文本*，包括 PubMed（White，2020 年）中心全文文章，以及 PubMed（White，2020 年）摘要和 bioRxiv（Sever 等人，2019 年）摘要中未产生 (2) 中可识别命名实体的*生物*文本。(4) *分子-描述对*和*蛋白质-描述对*。分子-文本数据来自 PubChem（Kim 等人，2019 年），我们还在文本描述中添加了 IUPAC 名称。所有存在于下游 Mol- Instructions 数据集（Fang 等人，2023 年）和 ChEBI- 20（Edwards 等人，2022 年）中的分子和蛋白质都被排除在外，以防止数据泄露。蛋白质文本数据与 BioT5（Pei 等人，2023 年）相同。

显著区别。BioT5+ 与 BioT5 的主要区别如下：(1) BioT5+ 在分子预训练数据中整合了 IUPAC，包括与 SELFIES、包装文本和分子-文本翻译数据相结合的 IUPAC 名称。更多详情见附录 D 部分。

这些高质量数据包括来自 PubChem 的 IUPAC 名称和 SELFIES，以及来自 bioRxiv 和 PubMed Central 的综合性文章。

3.3 令牌化

BioT5（Pei 等人，2023 年）已经证明了采用独立标记化和嵌入技术的优势。BioT5+ 继承了这一优势，专门针对生物实体应用专门的标记化和嵌入技术。这种方法明确区分了生物语义空间和文本语义空间。对于分子 SELFIES，每个具有化学意义的原子组（由于其括号格式（如 [C]）而与文本词汇自然区分开来）都使用 SELFIES 定义的固有标记集作为单独标记进行标记化。对于蛋白质 FASTA 序列，为了确保明确的模态区分，每个氨基酸都被标记为带有前缀 <p> 的单独标记，以区别于标准的大写英文字母。

同时，数值数据的标记化也值得专门考虑和设计。直接应用 T5（Raffel 等人，2020 年）从自然语言中提取的句子片段（Kudo 和 Richardson，2018 年）进行数字标记化可能会导致不一致（Liu 和 Low，2023 年）。例如，数字 "1024 "可能被标记为 "10 "和 "24"，而 "2048 "则可能被分割为 "2"、"0 "和 "48"。这种不规则的分割给模型带来了挑战，因为它无法将嵌入式映射到数字上，尤其是当数字代表的位数不同时。相比之下，Llama（Touvron 等人，2023a,b）和 ChatGLM（等人，2023 年）等模型采用了基于字符的数字标记化方法，将每个数字标记为一个单独的标记。这种方法已被证明能在各种算术任务中取得优异的结果（Liu 和 Low，2023 年；Nogueira 等人，2021 年）。因此，在 BioT5+ 中，我们也采用了这种基于字符的数字标记化方法，而无需修改原始词典。第 4.3 节显示了这种方法与原始 T5（Raffel 等人，2020 年）和 BioT5（Pei 等人，2023 年）相比在数字标记化方面的功效，为其在处理数字数据方面的卓越性能提供了经验证据。

3.4 模型和培训

模型结构。BioT5+ 采用了与 BioT5 相同的结构（Pei 等人，2023 年），该结构包括

遵循 T5-v1.1-base[1] 配置，带有词汇大小 35、076 和 252M 参数。

预培训。基于预培训语料库 de-
在第 3.2 节中介绍了 BioT5+ 的预训练。

(1) 针对*特定模式的 T5 目标*：这一类涉及将 T5 目标（屏蔽跨度预测）单独应用于每种模式，包括带有 IUPAC 名称的分子 SELFIES（任务 #1）、分子 SELFIES（任务 #2）、蛋白质 FASTA 序列（任务 #3）和一般文本内容（任务 #4）。

(2) *封装文本的 T5 目标*：将 T5 目标应用于从科学语料库中提取的 "包装 "文本（任务 #5）。(3) *生物文本的 T5 目标*：将 T5 目标应用于生物领域的文本（任务 6）。(4) *双向翻译任务*：这涉及分子 SELFIES 文本对（任务 #7）和蛋白质 FASTA 文本对（任务 #8）之间的双向翻译。通过这些策略性的结构化预训练任务，BioT5+ 能够很好地学习文本信息中所代表的生物实体的复杂关系和特征。

基于多任务指令的微调。在综合预训练阶段之后，BioT5+ 将进行基于多任务指令的微调。与 BioT5 不同的是，BioT5 的每个下游任务都有一个专门的微调模型，而我们则按照 Fang 等人，2023 年和 Cao 等人，2023 年的方法对下游任务进行分类，并进行多任务指令微调，这不仅节省了重复微调的成本，还简化了评估多个任务时的模型部署。图 1 展示了基准任务和数据集的相关分组和信息，并按领域（如面向分子或蛋白质的任务）进行了简单划分。这种方法有双重目的：首先，它弥补了预训练和微调阶段之间的差距，确保了学习能力的平稳过渡和整合。其次，它激活并利用了 BioT5+ 在各种任务中的通用能力，展示了它在处理各种生物问题时的多功能性和适应性。

表 1：MoleculeNet 上分子性质预测任务（分类）的性能（AUROC）比较（吴等人，2018 年）基准（**最佳**、次佳）。 *指 LoRA（Hu 等人，2022 年）调整。

4 实验和结果

如图 1 所示，BioT5+ 在 21 个成熟的下游领域进行了广泛评估。

[1]https://huggingface.co/docs/transformers/model_doc/t5v1.1

方法 # 分子	BACE ↑ 1513	BBBP ↑ 2039	HIV ↑ 41127	Clintox ↑ 1478
单一任务专家模式				
图表	75.4	69.7	78.5	76.0
GraphMVP-C	81.2	72.4	77.0	77.5
MGSSL	79.7	70.5	79.5	80.7
MolCLR	89.0	73.8	80.6	93.2
GEM	85.6	72.4	80.6	90.1
Uni-Mol	85.7	72.9	80.8	91.9
KV-PLM	71.9	66.9	68.8	84.3
墨木	76.7	70.5	75.9	79.9
MolFM	83.9	72.9	78.8	79.7
MolXPT	88.4	**80.0**	78.1	95.3
BioT5	**89.4**	77.7	**81.0**	**95.4**
基于 LLM 的通用模型				
银河-6.7B	58.4	53.5	72.2	78.4
银河-30B	72.7	59.6	75.9	82.2
Galactica-120B	61.7	66.1	74.5	82.6
Vicuna-v1.5-13B-16k (4 发)	49.2	52.7	50.5	-
Vicuna-v1.3-7B*	68.3	60.1	58.1	-
拉马-2-7B-聊天*	74.8	65.6	62.3	-
InstructMol-G-6.9B	85.9	64.0	74.0	-
InstructMol-GS-6.9B	82.3	70.0	68.9	-
BioT5+	**86.2**	**76.5**	**76.3**	**92.3**

基准数据集可分为 7 个面向分子的任务和 8 个面向蛋白质的任务，包含 3 类问题：分类、再ression 和生成。根据 Fang 等人的研究（2023 年），我们将下游任务分为不同的组别，以同样的方式进行多任务指令调整，下游数据集和基线的详情见附录 G 部分。

4.1 面向分子的任务

面向分子的任务涵盖不同的主题。由于我们在预训练中加入了分子的 IUPAC 名称，因此我们在一些面向分子的任务中也使用了 IUPAC 名称，如分子属性预测和分子描述生成。更多详情见以下章节和附录。

4.1.1 分子特性预测

分子性质预测是生物信息学中的一项重要任务，其重点是确定给定分子表现出的特定性质。继 Cao 等人，2023 年之后，我们探索了 BioT5+ 在 MoleculeNet（Wu 等人，2018 年）基准数据集上的能力。在分类任务中，我们重点关注 4 个基准数据集：BACE、BBBP、HIV 和 Clintox。每个样本都包含一条指令，详细说明要预测的性质和分子的 IUPAC 名称，要求模型生成简单的 "是 "或 "否 "预测。在回归任务方面，我们重点研究了 QM9 数据集中的 3 个回归基准。

表 2：化学反应相关任务的性能比较（**最佳**、次佳）。∗ 表示 LoRA 调整。

模型	精确↑	BLEU↑	列文思坦↓	RDK FTS↑	MACCS FTS↑	摩根 FTS↑	有效性↑
试剂预测							
拉马-7B	0.000	0.003	28.040	0.037	0.001	0.001	0.001
银河-6.7B	0.000	0.141	30.760	0.036	0.127	0.051	0.995
文本+化学 T5-223M	0.000	0.225	49.323	0.039	0.186	0.052	0.313
谟尔说明-7B	0.044	0.224	23.167	0.237	0.364	0.213	1.000
Llama-7B* (LoRA)	0.000	0.283	53.510	0.136	0.294	0.106	1.000
InstructMol-G-6.9B	0.070	**0.890**	24.732	0.469	**0.691**	0.426	1.000
InstructMol-GS-6.9B	0.129	0.610	19.664	0.444	0.539	0.400	1.000
BioT5+	**0.257**	0.695	**12.901**	**0.539**	0.621	**0.512**	1.000
前向反应预测							
拉马-7B	0.000	0.020	42.002	0.001	0.002	0.001	0.039
银河-6.7B	0.000	0.468	35.021	0.156	0.257	0.097	0.946
文本+化学 T5-223M	0.239	0.782	20.413	0.705	0.789	0.652	0.762
谟尔说明-7B	0.045	0.654	27.262	0.313	0.509	0.262	1.000
Llama-7B* (LoRA)	0.012	0.804	29.947	0.499	0.649	0.407	1.000
InstructMol-G-6.9B	0.153	0.906	20.155	0.519	0.717	0.457	1.000
InstructMol-GS-6.9B	0.536	0.967	10.851	0.776	0.878	0.741	1.000
BioT5+	**0.864**	**0.993**	**3.403**	**0.949**	**0.975**	**0.935**	1.000
回溯合成							
拉马-7B	0.000	0.036	46.844	0.018	0.029	0.017	0.010
银河-6.7B	0.000	0.452	34.940	0.167	0.274	0.134	0.986
文本+化学 T5-223M	0.141	0.765	24.043	0.685	0.765	0.585	0.698
谟尔说明-7B	0.009	0.705	31.227	0.283	0.487	0.230	1.000
Llama-7B* (LoRA)	0.000	0.283	53.510	0.136	0.294	0.106	1.000
InstructMol-G-6.9B	0.114	0.586	21.271	0.422	0.523	0.285	1.000
InstructMol-GS-6.9B	0.407	0.941	13.967	0.753	0.852	0.714	1.000
BioT5+	**0.642**	**0.969**	**6.710**	**0.897**	**0.930**	**0.866**	1.000

表 3：MoleculeNet（Wu 等人，2018 年）基准 QM9 数据集上分子性质预测任务（回归）的性能（MAE）比较（**最佳**、次佳）。$\Delta\epsilon$ 表示 HOMO-LUMO 间隙。

方法	HOMO ↓	LUMO ↓	$\Delta\epsilon$ ↓	平均值 ↓
基于 LLM 的通用模型				
Llama2-7B（5 发 ICL）	0.7367	0.86410	.5152 0	.7510
维库纳-13B（5 发 ICL）	0.7135	3.6807	1.5407	1.9783
谟尔说明-7B	0.0210	0.0210	0.0203	0.0210
InstructMol-G-6.9B	0.0060	0.0070	0.0082	0.0070
InstructMol-GS-6.9B	0.0048	0.0050	0.0061	0.0050
BioT5+	**0.0022**	**0.0024**	**0.0028**	**0.0025**

带有 IUPAC 名称的 SELFIES 分子，包括 HUMO、LUMO 和 HUMO-LUMO 间隙。

结果分类和回归任务的结果分别如表 1 和表 3 所示。BioT5+ 的表现优于其他通用模型基线。值得注意的是，在分类任务方面，BioT5+ 超越了 Galactica（泰勒等人，2022 年）等模型，后者是在庞大的科学文献语料库中经过大量训练的。同样，InstructMol（Cao 等人，2023 年）尽管包含二维图信息和 LLMs，但 BioT5+ 在分类和回归任务上的表现都优于它。这种性能的提高可归因于 BioT5+ 在预训练中整合了 IUPAC 名称、包装文本、生物文本和分子文本对。在这些二维语料库中存在分子特性描述，可使模型获得对分子特性的全面理解。然而，与单任务专家模型相比，BioT5+ 显示出了一些差距。这种差异是可以理解的，部分原因是单任务模型本身易于调整，部分原因是一些基线包含了额外的分子信息，如二维和三维结构。

4.1.2 化学反应相关任务

在计算化学领域，与化学反应相关的任务至关重要，因为它们可以加快研发进程。按照 Cao 等人（2023 年）的观点，我们重点关注 3 项此类任务：试剂预测、正向反应预测和合成。

结果主要结果见表 2，完整结果见表 10。虽然 LLM 在预培训期间接触了一些分子数据，但他们在化学反应相关任务的直接零点测试中表现极差。Mol-Instructions（Fang等人，2023年）在Llama（Touvron等人，2023年a）的基础上对分子导向任务进行了多任务指令调整。InstructMol（Cao等人，2023年）引入了分子图编码器，为Vicuna（Chiang等人，2023 年）编码二维分子图信息。我们的 BioT5+ 采用了与 Mol-Instructions （Fang 等人，2023 年）相同的训练设置，在化学反应相关任务的几乎所有指标上都表现优异。这一结果证明了分子数据和文本数据联合预训练的有效性。

4.1.3 分子描述生成

分子描述生成的目的是为给定分子生成详细翔实的描述。为了与 BioT5+ 的预训练保持一致，这里的输入也包括带有 IUPAC 的分子 SELFIES。分子特性预测通常侧重于特定属性，而分子描述生成则不同，它需要解释和传达分子的全面描述。这种描述不仅包括其分子组成和特性，还包括其潜在的应用和作用、

表 4：ChEBI-20（Edwards 等人，2022 年）数据集上分子描述生成任务的性能比较。

模型	BLEU-2↑	BLEU-4↑	ROUGE-1↑	红宝石-2↑	红磨坊	METEOR↑
单一任务专家模式						
变压器	0.061	0.027	0.204	0.087	0.186	0.114
T5 基座	0.511	0.423	0.607	0.451	0.550	0.539
MoT5-基础	0.540	0.457	0.634	0.485	0.568	0.569
莫姆（MolT5-base）	0.549	0.462	-	-	-	0.576
MolFM（MolT5-碱基）	0.585	0.498	0.653	0.508	0.594	0.607
MolXPT	0.594	0.505	0.660	0.511	0.597	0.626
GIT-Mol-graph	0.290	0.210	0.540	0.445	0.512	0.491
GIT-Mol-SMILES	0.264	0.176	0.477	0.374	0.451	0.430
GIT-Mol-（图形+SMILES）	0.352	0.263	0.575	0.485	0.560	0.430
文本+化学 T5	0.625	0.542	0.682	0.543	0.622	0.648
BioT5	0.635	0.556	0.692	0.559	0.633	0.656
MolCA	0.639	0.555	0.697	0.558	0.636	0.669
基于检索的 LLM						
GPT-3.5-涡轮增压（10 发 MolReGPT）	0.565	0.482	0.623	0.450	0.543	0.585
GPT-4-0314（10 射 MolReGPT）	0.607	0.525	0.634	0.476	0.562	0.610
基于 LLM 的通用模型						
GPT-3.5-涡轮增压（零喷射）	0.103	0.050	0.261	0.088	0.204	0.161
BioMedGPT-10B	0.234	0.141	0.386	0.206	0.332	0.308
谟尔说明-7B	0.249	0.171	0.331	0.203	0.289	0.271
InstructMol-G-6.9B	0.466	0.365	0.547	0.365	0.479	0.491
InstructMol-GS-6.9B	0.475	0.371	0.566	0.394	0.502	0.509
BioT5+	**0.666**	**0.591**	**0.710**	**0.584**	**0.650**	**0.681**

表 5：ChEBI-20（Edwards 等人，2022 年）数据集上描述引导的分子设计任务的性能比较。基本事实 Text2Mol（Edwards 等人，2021 年）得分为 0.609。

模型	*BLEU*↑	精确↑	*列*文思坦↓	MACCS *FTS*↑	RDK *FTS*↑	摩根 *FTS*↑	*FCD*↓	*Text2Mol*↑	有效性↑
单一任务专家模式									
变压器	0.499	0.000	57.660	0.480	0.320	0.217	11.32	0.277	0.906
T5 基座	0.762	0.069	24.950	0.731	0.605	0.545	2.48	0.499	0.660
MolT5-碱基	0.769	0.081	24.458	0.721	0.588	0.529	2.18	0.496	0.772
墨木基地	0.815	0.183	20.520	0.847	0.737	0.678	-	0.580	0.863
MolFM-base	0.822	0.210	19.445	0.854	0.758	0.697	-	0.583	0.892
GIT-Mol	0.756	0.051	26.315	0.738	0.582	0.519	-	-	0.928
MolXPT	-	0.215	-	0.859	0.757	0.667	0.45	0.578	0.983
BioT5	0.867	0.413	15.097	0.886	0.801	0.734	0.43	0.576	**1.000**
基于检索的 LLM									
Llama2-7B（2 射 MolReGPT）	0.693	0.022	36.77	0.808	0.717	0.609	4.90	0.149	0.761
GPT-3.5-涡轮增压（10 发 MolReGPT）	0.790	0.139	24.91	0.847	0.708	0.624	0.57	0.571	0.887
GPT-4-0314（10 射 MolReGPT）	0.857	0.280	17.14	0.903	0.805	0.739	0.41	**0.593**	0.899
基于 LLM 的通用模型									
Llama2-7B (0-shot)	0.104	0.000	84.18	0.243	0.119	0.089	42.01	0.148	0.631
GPT-3.5-涡轮增压（0-射程）	0.489	0.019	52.13	0.705	0.462	0.367	2.05	0.479	0.802
BioT5+	**0.872**	**0.522**	**12.776**	**0.907**	**0.835**	**0.779**	**0.353**	0.579	**1.000**

表 6：PEER 基准的性能（准确度）比较（**最佳**、次佳）。* 表示线性探测。

	模型溶解度定位		酵母	人类
单任务 Spe	*专业模型*			
DDE	59.77 ± 1.21	77.43 ± 0.42	55.83 ± 3.13	62.77 ± 2.30
莫兰	57.73 ± 1.33	55.63 ± 0.85	53.00 ± 0.50	54.67 ± 4.43
LSTM	70.18 ± 0.63	88.11 ± 0.14	53.62 ± 2.72	63.75 ± 5.12
变压器	70.12 ± 0.31	75.74 ± 0.74	54.12 ± 1.27	59.58 ± 2.09
美国有线电视新闻网	64.43 ± 0.25	82.67 ± 0.32	55.07 ± 0.02	62.60 ± 1.67
ResNet	67.33 ± 1.46	78.99 ± 4.41	48.91 ± 1.78	68.61 ± 3.78
ProtBert	68.15 ± 0.92	91.32 ± 0.89	63.72 ± 2.80	77.32 ± 1.10
ProtBert*	59.17 ± 0.21	81.54 ± 0.09	53.87 ± 0.38	83.61 ± 1.34
ESM-1B	70.23 ± 0.75	**92.40 ± 0.35**	57.00 ± 6.38	78.17 ± 2.91
ESM-1B*	67.02 ± 0.40	91.61 ± 0.10	**66.07 ± 0.58**	**88.06 ± 0.24**
BioT5	**74.65 ± 0.49**	91.69 ± 0.05	64.89 ± 0.43	86.22 ± 0.53
多任务通用模式				
美国有线电视新闻网	70.63 ± 0.34	82.67 ± 0.72	54.50 ± 1.61	69.03 ± 2.68
变压器	70.03 ± 0.42	76.27 ± 0.57	54.00 ± 1.17	67.33 ± 2.68
ESM-1B	70.46 ± 0.16	**92.50 ± 0.26**	64.76 ± 1.42	83.00 ± 0.88
BioT5+	**74.37 ± 0.19**	90.41 ± 0.07	**66.16 ± 0.43**	**85.09 ± 0.40**

由 SELFIES 代表和 IUPAC 名称整合得出。我们使用的评价指标与 Fang 等人的研究相同，2023 年。

结果如表 4 所示，我们的 BioT5+ 优于所有单任务专家、基于检索的 LLM 和多任务通用基线。BioT5+ 之所以能取得如此优异的成绩，得益于其在预训练过程中的全面学习。该模型有效地吸收了分子的多维度和丰富的文本描述。

4.1.4 描述引导的分子设计

描述引导的分子设计本质上是分子描述生成的逆任务，它要求根据提供的文本描述生成分子。在 BioT5+ 设置中，我们在分子的文本描述中不包含 IUPAC 名称，以防止模型学习从 IUPAC 名称到其 SELFIES 表示的简单映射，从而确保模型不会忽略文本中提供的其他描述元素。

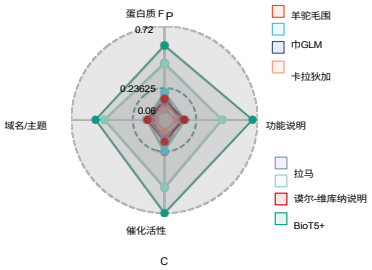
结果表 5 列出了由描述引导的分子设计任务的结果。我们的 BioT5+ 超越了所有比较基线。这一成绩凸显了 BioT5+ 预训练的功效，因为模型已经对分子知识有了深刻的理解。

4.2 面向蛋白质的任务

4.2.1 蛋白质描述生成

蛋白质描述生成任务包括从给定的蛋白质序列中获取相关文本信息。根据 Fang 等人的研究，我们主要关注 4 项相关的生成任务：蛋白质功能生成、催化活性生成和蛋白质描述生成、

表 7：在单一任务设置下生成分子描述任务的 IUPAC 和附加数据的消减情况。B-2 代表 BLEU-2，R-1 代表 ROUGE-1。



C

表 8：3 个 DTI 数据集的性能（AUROC）比较（**最佳**、次佳）。

方法	BIOSNAP	人类	结合数据库
单一任务	*列表型号*		
SVM	0.862±0.007	0.940±0.006	0.939±0.001
射频	0.860±0.005	0.952±0.011	0.942±0.011
DeepConv-DTI	0.886±0.006	0.980±0.002	0.945±0.002
GraphDTA	0.887±0.008	0.981±0.001	0.951±0.002
MolTrans	0.895±0.004	0.980±0.002	0.952±0.002
毒品禁令	0.903±0.005	0.982±0.002	0.960±0.001
BioT5	0.937±0.001	**0.989±0.001**	0.963±0.001
多任务 Gener	*模型*		
BioT5+	**0.939±0.001**	**0 . 987±0**.001	0 . **964±0.001**

表 9：在 QM9 数据集上消减默认 T5 标记符号生成器和基于字符的标记符号生成器（BioT5+）。

方法	HOMO ↓	LUMO ↓	$\Delta\epsilon$ ↓	平均值 ↓
T5 默认标记符	0.0024	0.0026	0.0032	0.0027
BioT5+	0.0022	0.0024	0.0028	0.0025

域/主题词生成和功能描述生成。

结果如图 2 所示，在 4 项任务中，BioT5+ 超越了所有比较基线。这一结果凸显了 BioT5+ 将复杂的蛋白质序列解释为平均文本信息的高级能力，表明 BioT5+ 通过预训练获得了对蛋白质结构和功能的全面理解。

4.2.2 蛋白质特性预测

蛋白质性质预测任务包括根据蛋白质的氨基酸序列预测蛋白质的特定性质，如溶解性、结构或功能。继 BioT5（Pei 等人，2023 年）之后，我们重点研究 PEER（Xu 等人，2022 年）基准中的 2 项蛋白质特性预测任务，该基准专门针对蛋白质序列理解而设计：

(1) 溶解性预测：输入蛋白是否可溶。(2) 定位预测：输入蛋白质是 "膜结合型 "还是 "可溶性 "。这两项任务都是二元分类任务，模型需要生成 "是 "或 "否 "的预测结果。这些结果将在下面的第 4.2.3 节中进行总结。

4.2.3 蛋白质相关相互作用预测

在药物发现过程中，生物实体之间的相互作用预测非常重要。

模型	B-2↑	B-4↑	R-1↑	R-2↑	R-L↑	METEOR↑
BioT5+（单一任务）	0.671	0.597	0.715	0.590	0.655	0.687
BioT5+（单一任务）wo IUPAC	0.661	0.584	0.706	0.578	0.647	0.677
BioT5+（单一任务）其他数据	0.666	0.591	0.711	0.586	0.651	0.681
BioT5+	0.676	0.606	0.721	0.599	0.660	0.691

图 2：蛋白质描述基因迭代任务的性能（ROUGE-L）比较。

蛋白质-蛋白质相互作用（PPI）和药物-靶标相互作用（DTI）就是两个重要的例子。这两项任务对于了解生物过程和确定潜在治疗靶点至关重要。为此，我们效仿 Pei 等人，于 2023 年纳入了 PEER（Xu 等人，2022 年）基准的 2 个 PPI 数据集（包括酵母和人类），以及 3 个 DTI 数据集（包括 BioSNAP（Zitnik 等人，2018 年）、人类（Liu 等人，2015 年；Chen 等人，2020 年）和 BindingDB（Liu 等人，2007 年））。

结果如表 6 所示，在 PEER 基准测试中，我们的 BioT5+ 表现出了卓越的性能，在 4 项任务中的 3 项任务中超过了其他多任务模型，取得了与单任务专家模型相当的结果。值得注意的是，在酵母 PPI 预测任务中，BioT5+ 的表现超过了所有基线模型。考虑到基线模型 ESM-1b（Rives 等人，2021 年）在大量蛋白质序列上进行了专门的预训练，并拥有比 BioT5+ 多一倍以上的参数，这一点就显得尤为重要。此外，如表 8 所示，BioT5+ 在 DTI 任务中也表现出了卓越的性能（完整结果见表 11），在 BioSNAP 和 BindingDB 数据集上始终优于其他方法。值得注意的是，许多基线方法涉及分子和蛋白质编码器的专门设计。这些结果充分证明了 BioT5+ 对生物文本、分子和蛋白质进行联合预训练的有效性。这种全面的理解体现在 BioT5+ 能够准确预测蛋白质的性质、相互作用和药物靶标间的作用，使其成为计算生物学领域的一个宝贵工具。

4.2.4 描述引导的蛋白质设计

对于描述引导的蛋白质设计，模型需要根据特定的设计要求（如蛋白质结构和功能）生成蛋白质氨基酸序列。由于这项任务缺乏成熟的基准，我们在附录表 20 中列出了一些测试用例及其对应的序列相似性得分，以提供直接的参考。

我们的模型与现有的模型，如 Galactica（Taylor 等人，2022 年）和 Mol- Instructions（Fang 等人，2023 年）之间的比较。

4.3 消融研究

在本节中，我们将进行消融研究，以了解我们的设计在 BioT5+ 中的有效性。具体来说，我们重点研究了以下三种情况：**（1）不包含分子的 IUPAC 名称。**如表 7 所示，去除 IUPAC 名称会导致分子描述生成任务的性能明显下降。这一下降凸显了 IUPAC 名称在分子理解相关任务中的重要作用。**(2) 在预训练中不添加 PubMed Cen- tral 和 bioRxiv 数据**。表 7 和表 13 中的结果表明，这两个数据集在增强分子理解方面起着至关重要的作用。省略这两个数据集会导致分子描述生成和描述引导的分子设计任务的性能略有下降，但降幅明显。**(3) 对数字使用 T5 去错标记符，而不是基于字符的标记符。**表 9 中的结果表明，在回归任务中，基于字符的数字标记化方法比默认的 T5 标记化器更有效。我们还在附录 F 部分进行了一项消融研究，以进一步对比单任务和多任务调整策略。

5 结论

作为 BioT5 框架的高级迭代版本，BioT5+ 标志着在组合生物学和药物发现领域取得了重大进展。通过整合 IUPAC 名称、扩展生物文本和分子数据源、采用多任务指令调整以及结合先进的数字标记化技术，BioT5+ 成功地弥合了分子表征与其文本描述之间的差距。BioT5+ 增强了对分子结构的理解，并能处理复杂的生物数据，这一点已在广泛的任务中得到证实，BioT5+ 在大多数任务中都取得了一流的性能。这一成功彰显了 BioT5+ 作为理解和分析生物实体的多功能强大工具的潜力。

6 局限性

尽管 BioT5+ 取得了重大进展，但仍存在一些局限性，需要进一步改进。

在未来的工作中需要解决这些问题。首先，该模型在通用于各种生物逻辑任务方面面临挑战，这是一个不同于普通 NLP 设置的问题。每种生物任务都具有复杂性和独特性，因此很难开发出放之四海而皆准的解决方案，这就凸显了在这一领域需要更专业的方法。其次，BioT5+ 目前的规模有限，无法理解和整合来自图像等其他模式的信息，限制了其在多模式生物数据分析中的适用性。BioT5+ 不具备通用聊天机器人的功能，也无法回答生物学特定范围之外的一般领域问题。这一限制凸显了开发更大、更多功能模式的必要性，这些模式能够处理更广泛的数据类型，回答生物和通用领域中更广泛的问题。

7 伦理方面的考虑

虽然 BioT5+ 是一项重大进步，但它的功能，尤其是根据文字描述生成分子和预测化学反应产物的功能，引发了重要的伦理问题。其中一个值得关注的问题是，这项技术可能会被滥用于生成有害或危险分子，从而对公共安全和环境健康构成威胁。此外，BioT5+ 预测和生成新分子的能力也可能导致知识产权和专利方面的问题。利用人工智能驱动的方法可以轻松设计和合成新化合物，这可能会扰乱传统的研究实践，并引发有关这些发现的所有权问题。

8 致谢

本研究得到了国家自然科学基金（NSFC）（批准号：62122089）、北京市杰出青年科学基金项目（批准号：BJWZYJH012019100020098）和国家自然科学基金面上项目（批准号：62122089）的资助。BJJWZYJH012019100020098 和
中国人民大学智慧社会治理平台、中国人民大学 "双一流 "建设重大需求与规划交叉学科平台、中央高校基本科研业务费、中国人民大学科研基金。裴启智获得中国人民大学 "2023杰出创新人才培养资助计划 "支持。