BioT5:用化学知识和自然语言联想丰富生物学中的跨模态整合

裴启智 1，5，魏章 2，朱金华 2，吴 2，高开元 3，吴立军 4 *,英策夏 4 *,茵燕 1，6*

1 中国人民大学高岭人工智能学院
2 中国科学技术大学
3 华中科技大学 4 微软研究院
5 教育部下一代智能搜索与推荐工程研究中心
6 北京大数据管理与分析方法重点实验室

{qizhipei,ruiyan}@ruc.edu.cn
{weizhang_cs,teslazhu,wu_2018}@mail.ustc.edu.cni
m_kai@hust.edu.cn{lijuwu,yinxia}@microsoft.
com

摘要

生物学研究的最新进展利用分子、蛋白质和自然语言的整合来促进药物发现。然而，目前的模型表现出一些局限性，如无效分子微笑的产生，上下文信息的利用不足，以及结构化和非结构化知识的平等对待。为了解决这些问题

问题，我们提出 BioT5，这是一个全面的预训练框架，丰富了化学知识和自然语言联想在生物学中的跨模态集成。BioT5 利用自拍进行 100%可靠的分子表示，并从非结构化生物文献中生物实体的周围环境中提取知识。此外，BioT5 区分结构化和非结构化知识，从而更有效地利用信息。经过微调后，BioT5 在广泛的任务中表现出优异的性能，展示了其捕捉生物实体的潜在关系和属性的强大能力。我们的准则可从以下网址获得 https://github.com/QizhiPei/BioT5。

1 介绍

分子和蛋白质是药物发现中的两个基本生物实体(Dara et al.,2022).近一个世纪以来，小分子药物一直是制药工业的基石，因为它们具有独特的优势，例如口服可用性、不同的作用模式等(AstraZeneca,2023).蛋白质是生命科学的基础，其功能是作为药物靶标或疾病途径中的关键元素。如图所示 1，两者都有

通讯作者：吴立军(lijuwu@microsoft.com)、英策夏(@microsoft.com)和

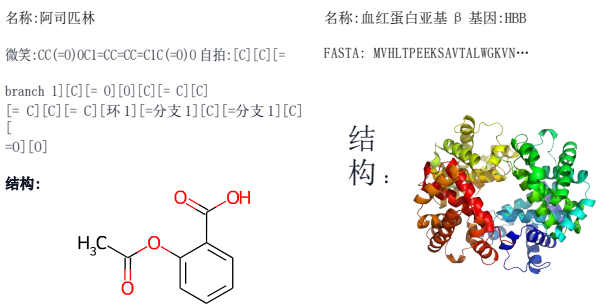

图 1:分子和蛋白质的表示。分子可以用它的名字、生物序列(微笑和自拍)和 2D 图结构来表示。蛋白质可以通过其名称、相应的基因名称、生物序列(FASTA)和 3D 结构来表示。

分子和蛋白质可以用序列来表示。分子可以用 SMILES 序列来描述(Weininger,1988；Weininger et al.,1989)，它是通过深度优先搜索遍历分子图并应用特定的分支规则得到的。蛋白质可以由 FASTA 序列(Lipman and Pearson,1985；Pearson and Lipman,1988)，它概括了蛋白质中的氨基酸。分子和蛋白质的顺序格式便于变压器模型的应用(Vaswani et al.,2017)和预训练技巧(Liu et al.,2019；Radford et al.,2019)从自然语言处理(NLP)到生物医学领域。切伯塔(Chithranandaet al.,2020)和 ESM(Rives et al.,2021；Lin et al.,2022)分别对分子微笑和蛋白质 FASTA 应用掩蔽语言建模，而 MolGPT(Bagal et al.,2022)和 ProtGPT2(Ferruzet al.,2022)利用 GPT 式的模型来生成分子和蛋白质。

科学文献(Beltagy et al.,2019；Caneseand Weis,2013)和生物数据库(Kim

et al.,2023; Boutet et al.,2007)充当分子和蛋白质的知识库。这些资源详细描述了各种生物实体之间的特性、实验结果和相互作用，这些不能仅从分子或蛋白质序列中明确推断出来。因此，最近的趋势包括将文本与分子和蛋白质一起建模，允许文本描述增强分子和蛋白质的表示。MolT5(Edwards et al.,2022)采用T5(Raffel et al.,2020)分子微笑和生物医学文献框架。MolXPT(Liuet al.,2023b)和卡拉狄加(Taylor et al.,2022)是在文本和生物实体上训练的GPT模型，例如微笑和FASTA序列。深海EIK(Luo et al.,2023)使用注意力(Vaswaniet al.,2017)机制。尽管他们取得了成功，但仍然有很大的改进空间:(I)以前的工作经常依赖于微笑来代表分子。然而，解决产生无效微笑的问题仍然是一个需要克服的挑战(Edwards et al.,2022; Li et al.,2023). (ii)围绕分子或蛋白质名称的上下文信息可以为理解生物实体的相互作用和性质提供有价值的见解。开发有效的方法来利用这些信息值得进一步关注。(iii)现有研究倾向于同等对待结构化数据(例如，来自数据库的分子-文本对)和非结构化数据(例如，文献中的文本序列)。然而，可以更有效地利用结构化数据来进一步提高整体性能。

为了应对上述挑战，在本文中，我们介绍了BioT5，这是一个全面的预训练框架，包括文本、分子和蛋白质。BioT5利用自拍(Krenn et al.,2020)来表示小分子，因为它相对于SMILES的优点是SELFIES提供了更健壮和容错的分子表示，消除了SMILES经常遇到的非法结构的问题。BioT5预训练主要有两个步骤:

(1) *数据收集和处理:我们收集文本、分子和蛋白质数据，以及包含分子-文本平行数据和蛋白质-文本平行数据的现有数据库。对于来自生物领域的文本数据(PubMed)，我们使用命名实体识别和实体链接来提取分子和蛋白质提及，用相应的自拍或FASTA替换它们*序列。跟随Liu et al. (2023b)，我们称之为“包装”文本。文本标记、FASTA序列和自拍是独立标记的(请参见一节)3.2 了解更多详情)。

(2) *模型训练:BioT5利用一个共享的编码器和一个共享的解码器来处理各种模态。标准T5采用“恢复屏蔽跨度”目标，其中每个屏蔽跨度及其相应部分共享相同的sent inel令牌。为简单起见，我们将上述训练目标函数称为“T5目标”。有三种类型的预训练任务:(I)独立地将标准T5目标应用于分子自拍、蛋白质FASTA和普通文本，确保模型在每种模态中都具有能力。(ii)将T5目标应用于来自生物领域的包装文本，其中所有文本、FASTA和SELFIES标记可以被屏蔽和恢复。(iii)对于结构化的分子文本数据，我们引入了翻译目标。具体来说，BioT5被训练成将分子自拍翻译成相应的描述，反之亦然。同样，翻译目标也适用于蛋白质文本数据。*

在预训练后，我们在15个任务上对获得的BioT5进行微调，这些任务包括分子和蛋白质性质预测、药物-靶标相互作用预测、蛋白质-蛋白质相互作用预测、分子标题和基于文本的分子生成。BioT5在10个任务上实现了最先进的性能，并在5个任务上展示了与特定领域大型模型相当的结果，展示了我们提出的方法的优越能力。BioT5模型为整合化学知识和自然语言联想建立了一个有前途的途径，以增加当前对生物系统的理解。

2 相关著作

在这一节中，我们简要回顾了生物学中跨模态模型以及分子和蛋白质表示的相关工作。

2.1 生物学中的跨模态模型

生物学领域的语言模型已经获得了相当大的关注。其中，比奥伯特(Lee et al.,2020)和BioGPT(Luoet al.,2022)，在有效理解科学文本方面特别成功。最近，关注联合建模的跨模态模型

带有生物序列的文本已经出现。他们可以分为以下三类。

交叉文本-分子模态 MolT5(Ed-wards et al.,2022)是 T5(Raffel et al.,2020)为基础的模型，该模型是在分子微笑和一般文本语料库上联合训练的。宿墨(Suet al.,2022)使用对比学习对分子图和相关文本数据进行训练。MolXPT(Liu et al.,2023b)是 GPT(Radfordet al.,2018)-基于分子微笑、生物医学文本和包装文本的预训练模型。与 BioT5 不同，这些模型都使用微笑来表示分子，这导致了生成分子时的有效性问题。

跨文本-蛋白质模态蛋白质 DT(Liuet al.,2023a)是一个多模态框架，使用语义相关的文本进行蛋白质设计。生物翻译(Xu et al.,2023a)是一个跨模态翻译系统，专门设计用于基于用户编写的文本来注释生物学实例，如基因表达载体、蛋白质网络和蛋白质序列。

跨越三个或三个以上的生物模态 Galac-tica(Taylor et al.,2022)是基于 GPT 的通用大型语言模型，在各种科学领域进行训练，包括科学论文语料库、知识库(例如，公共化学(Kim et al.,2023)分子，UniProt(uni,2023)蛋白质)、代码和其他来源。迪佩克(Luo et al.,2023)融合来自多模态输入(药物、蛋白质和文本)的特征。然后注意(Vaswani et al.,2017)机制进行文本信息去噪和异构特征融合。

我们的工作在几个方面不同于以前的研究:(1)我们主要集中在两个生物学模态——分子和蛋白质——以文本作为知识库和桥梁来丰富分子和蛋白质领域的基础关系和性质；(2)我们使用多任务预训练以更全面的方式对这三种模态之间的联系进行建模。

(3)我们用自拍代替微笑来表示分子，这样更加健壮，解决了分子生成任务中的有效性问题。

2.2 分子和蛋白质的表示

分子表示分子的表示和建模长期以来一直是生物信息学中的一个挑战。有许多方法来表示分子：名称、指纹(Rogersand Hahn,2010a)，微笑(Weininger,1988；Weininger et al.,1989)，英制(Heller et al.,2013)，DeepSMILES(O'Boyle and Dalke,2018)，SELF-IES(Krenn et al.,2020)、2D 分子图等。SMILES(简化的分子输入行输入系统)是最常用的方法，它是分子结构的一种简洁的文本表示。它使用一系列字符对原子、键和其他分子特征进行编码。然而，SMILES 有几个缺点(Krennet al.,2022)，例如缺乏句法和语义鲁棒性，这显著影响了由深度学习模型生成的分子的有效性(Edwards et al.,2022).为了解决这个问题，引入了 SELFIES(自引用嵌入串)作为 100%健壮的分子串表示(Krenn et al.,2020).自拍字母表中的每种符号排列都会产生一种化学上有效的分子结构，确保每张自拍对应一种有效的分子。与第节介绍的现有作品不同 2.1 使用微笑来表示分子，我们在 BioT5 中使用带有单独编码的自拍来实现下游分子生成任务的 100%有效性。

蛋白质表达蛋白质也可以用多种方式表达，如通过其名称、相应的基因名称、FASTA 格式或 3D 几何结构。FASTA 格式是编码蛋白质序列的常见选择，它使用单字母代码来代表 20 种不同的氨基酸。在 BioT5 中，我们还采用 FASTA 格式来表示蛋白质。

不像 Edwards et al.（2022)和 Taylor et al.（2022)共享生物序列表征和自然语言表征之间的字典，BioT5 使用单独的字典和生物特定表征来明确区分生物模态。我们将在第 3 节对此进行进一步的分析 3.2。

3 BioT5

BioT5 预培训的概述如图所示 2。我们结合来自不同模态的数据来执行多任务预训练。

3.1 训练前语料库

如图所示 2BioT5 的预训练语料分为三类:(1)单模态数据，包括分子自拍、蛋白质 FASTA 和普通文本。对于小分子，我们使用锌 20(Irwin et al.,2020)数据集并将微笑转换为自拍。蛋白质

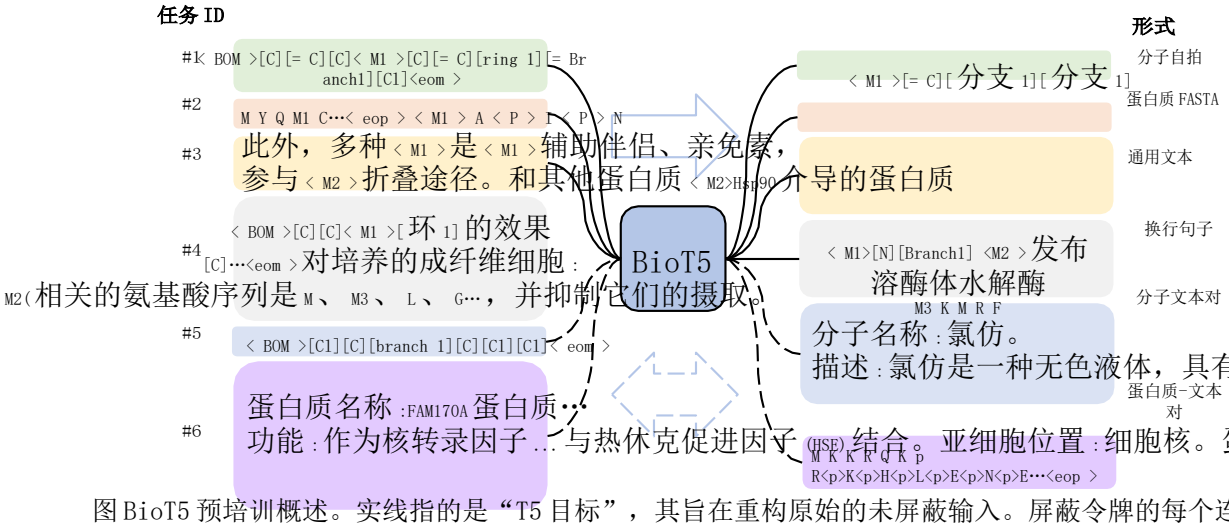

图 BioT5 预培训概述。实线指的是“T5 目标”，其旨在重构原始的未屏蔽输入。屏蔽令牌的每个连续区间被替换为标记令牌，描述为
< M1 >，< M2 >，和< M3 >。我们将此目标应用于分子自拍(任务#1)、蛋白质 FASTA(任务#2)、普通文本(任务#3)和包装文本(任务#4)。虚线表示生物序列之间的双向翻译
和结构化文本描述(任务#5 和#6)。

FASTA ， 我 们 从 Uniref50(Suzek et al.,2007)数据集，过滤掉超过指定长度的蛋白质，导致对于一般文本的 27M 蛋白质的集合，我们使用“庞大的干净爬行语料库”(C4)数据集(Raffel et al.,2020).(2)包装文本，其中分子名称被替换为它们相应的自拍照，基因名称被加上相关的蛋白质 FASTA。我们使用 3300 万 篇 PubMed 文 章 (Canese and Weis,2013)和应用 BERN2(Sung et al.,2022)进行命名实体识别。没有被生物序列替换或附加的科学句子被保留下来，作为一般文本的补充。详细的过程如图所示 4 并在附录中讨论 B。(3)分子描述对和蛋白质描述对。对于分子文 本 数 据 ， 我 们 从 PubChem(Kim et al.,2019)，排除下游 ChEBI- 20 数据集中存在的所有分子(Edwards et al.,2022)以避免潜在的数据泄漏。对于蛋白质-文本数据，我们从 Swiss- Prot(Boutet et al.,2007)，其中包含各种蛋白质属性的高质量注释。详情见附录 E.1。

3.2 分离标记化和嵌入

在大多数以前的工作中，分子和蛋白质的表示没有用

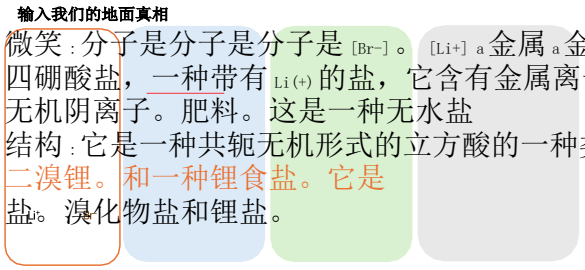

图 3:标记化的例子。MolT5 将“Br”(溴原子)处理为“B”(硼原子)和“r”，导致不正确的描述，包括四硼酸盐(与“B”相关)。BioT5 保留化学上有意义的基团“[Br-1]”作为完整的标记，从而产生正确的输出。

对 细 节 的 充 分 关 注 。 MolT5(Edwards et al.,2022)使用与原始 T5 相同的字典，因为它从原始 T5 检查点开始预训练。原始 T5 字典是使用句子片断(Kudoand Richardson,2018).然而，直接使用这个分子微笑词典是次优的，因为一些化学上有意义的标记，如官能团或完整的原子，将被不准确地识别。例如，在图 1 所示的分子中 3，在 SMILES 中用“Br”表示的溴原子被 MolT5 标记为“B”(硼原子)和“r”。因此，MolT5 错误地将该分子描述为二溴甲苯(与“Br”相关)和四硼酸盐(与“B”相关)。Galac-的基于字符的标记化

蒂卡(Taylor et al.,2022)也有同样的问题。

除了令牌化方法之外，针对不同模态的共享令牌嵌入(Ed-wards et al.,2022；Taylor et al.,2022)也是有疑问的。在多语言任务中，共享嵌入允许模型准确地表示借词和同源词的含义，这些词和同源词保留了它们在不同语言中的原始含义。然而，分子、蛋白质和文本代表了完全不同的语言。这三种不同情态中的同一个标记具有不同的语义。例如，记号“C”在自然语言中表示字符C，在分子中表示碳原子，在蛋白质中表示半胱氨酸(20种氨基酸之一)。研究由Beltagy et al.（2019)和Gu et al. （2021)进一步强调特定领域词汇的重要性。

为了解决上面提到的问题，我们对分子、蛋白质和文本使用不同的词汇。在BioT5中，分子由SELFIES字符串表示，其中每个化学上有意义的原子团都包含在括号中，并标记为SELFIES标记。比如[C][=C][Br] [C]，[=C]，[Br]。对于蛋白质，为了区分文本中大写字母的氨基酸，我们为每个氨基酸引入一个特殊的前缀< p>。举个例子，
对于文本，我们
使用与原始T5相同的字典。通过这一点，我们明确地区分了不同模态的语义空间，这保持了每个独特模态的内在完整性，并防止模型将跨模态的含义合并。

3.3 模型和培训

模型架构BioT5采用与T5模型相同的架构(Raffel et al.,2020).我们遵循T5-v1.1-base中使用的配置[1]。
BioT5的词汇表大小为35，073，不同于默认配置，因为我们为分子自拍和

蛋白质氨基酸。BioT5模型总共包括252M个参数。
预训练在预训练阶段，模型以多任务方式在六个任务上被训练，这六个任务可以被分类为三种类型:(1)独立地将T5目标应用于每个单一模态，包括分子自拍(任务#1)、蛋白质FASTA(任务#2)和一般文本(任务#3)。(2)将T5目标应用于科学语料库的包装文本(任务#4)。(3)分子自身-文本对(任务#5)和蛋白质FASTA-文本对(任务#6)的双向翻译。通过这些预训练任务从文本信息中有效地学习生物实体的潜在联系和属性，BioT5允许对生物领域的整体理解，从而有助于在各种生物任务中增强预测和生成能力。
微调BioT5可以在涉及分子、蛋白质和文本的各种下游任务上进行微调。为了统一不同的下游任务并减少预训练和微调之间的差距(Brown et al.,2020)阶段，我们采用基于提示的微调(Gao et al.,2021)方法，该方法便于将各种任务格式转换成序列生成格式。

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

4 实验和结果

我们在15个成熟的下游任务上评估BioT5，这些任务可以分为三种类型:单实例预测、多实例预测和跨模态生成。我们在附录中包括了关于微调数据集、基线和提示的细节F。

对于第节中介绍的下游二元分类任务4.1和4.2评估度量(如AUROC和AUPRC)的计算需要预测标签的软概率。当我们使用基于提示的微调方法时，输出要么是正标签的Yes，要么是负标签的No。要获得适当的标签分布，请遵循Liu et al. （2023b)，我们首先提取Yes和No标记的概率(分别表示为ppos和pneg)并归一化它们。阳性标记的结果概率是ppos，阴性标记的结果概率是pneg。

$p_{pos}+p_{neg}$ $p_{pos}+p_{neg}$

4.1 单实例预测

4.1.1 分子性质预测

分子性质预测旨在确定给定的分子是否表现出特定的性质。分子网(Wu et al.,2018)是一个广泛用于分子性质预测的基准，包含涵盖许多分子方面的不同数据集，如量子力学、物理化学、生物物理学等。符合Liuet al.（2023b)，我们在六个二元分类任务上进行实验，包括BBBP、Tox21、ClinTox、HIV、BACE和SIDER。跟随(Fang

Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER	Avg
#Molecules	2039	7831	1478	41127	1513	1427	-
#Tasks	1	12	2	1	1	27	-
G-	70.3 1.6	75.2 0.3	59.9 8.2	75.9 0.9	79.2 0.3	58.4 0.6	69.8
G-Motif	70.0 0.1	74.3 0.1	81.2 3.0	62.5 0.9	82.6 0.7	64.8 0.6	70.9
GROVERbase	72.4 1.6	75.9 0.5	79.1 2.8	77.0 1.2	81.2 0.9	63.9 1.2	72.6
GROVERlarge	72.2 2.1	75.0 0.2	91.2 3.5	78.1 0.5	82.4 0.9	58.9 1.4	72.3
GraphMVP							74.9
MGSSL							74.8
MolCLR							76.3
GEM							79.0
MoMu	70.5						
MolXPT	-	78.1		74.0	-	61.5	-
BioT5	77.7 0.6	77.9 0.2	95.4 0.5	78.2 0.1	82.1 0.3	63.7 0.4	79.0

表 MoleculeNet 上的性能比较(最好，第二好)。评估指标是 AUROC。基准结果主要来自 MolXPT(Liu et al.,2023b).

型号#参数。溶解度定位 DDE 205.3K 59.77 1.21					
77.43 0.42					
莫兰	123.4K	57.73	1.33	55.63	0.85
LSTM	26.7 米	70.18	0.63	88.11	0.14
变压器	21.3 米	70.12	0.31	75.74	0.74
CNN	5.4M 米	64.43	0.25	82.67	0.32
	11.0 米	67.33	1.46	78.99	4.41
普罗伯特	419.9 米	68.15	0.92	91.32	0.89
普罗伯特*	419.9 米	59.17	0.21	81.54	0.09
ESM-1b	652.4 米	70.23	0.75	92.40	0.35
ESM-1b *	652.4 米	67.02	0.40	91.61	0.10
BioT5	252.1M	74.65	0.49	91.69	0.05

表 2:不同方法在溶解度和定位预测任务上的性能比较(最好，第二好)。评估标准是准确性。*表示仅调整预测头。基线结果来自同级(Xu et al.,2022).

et al.,2022)，我们采用脚手架分裂，相比随机分裂更有挑战性。

基线我们将 BioT5 与两种类型的基线进行比较:(1)使用分子图作为输入的预训练图神经网络(GNN)，其是 G 上下文(Rong et al.,2020)，G 基序(Ronget al.,2020)，GROVERbase(Rong et al.,2020)，GROVERlarge(Rong et al.,2020)，GraphMVP(Liuet al.,2022)，MGSSL(Zhang et al.,2021)MolCLR(Wang et al.,2022)和宝石(Fang et al.,2022);(2)预先训练的语言模型基线，它们是 KV-PLM(Zeng et al.,2022)，Galactica(Taylor et al.,2022)，嫫母(Su et al.,2022)和 MolXPT(Liu et al.,2023b).

结果结果列于表中 1 所有统计数据来自三次随机运行。从

从这些结果中，我们可以看到 BioT5 在 MoleculeNet 的大多数下游任务上都超过了基线。与根据 2D/3D 分子数据预先训练的 GNN 基线相比，BioT5 表现出优异的性能，强调了文本中知识的有效性。此外，BioT5 输出形成其他语言模型基线，这可能归因于科学上下文文本或现有生物数据库条目中分子特性描述的存在。

4.1.2 蛋白质性质预测

蛋白质性质预测至关重要，因为它提供了对蛋白质行为和功能的关键见解。我们在对等基准上集中于两个蛋白质性质预测任务(Xu et al.,2022):蛋白质溶解度预测，旨在预测给定蛋白质是否可溶，以及蛋白质定位预测，即将蛋白质分类为“膜结合”或“可溶”。基线我们将 BioT5 与同行基准中提供的三种类型的基线进行比较:(1)特征工程师，包括两个蛋白质序列特征描述符:二肽偏离预期均值(DDE)(Saravanan and Gautham,2015)和莫兰相关性(Moran)(Feng and Zhang,2000);(2)蛋白质序列编码器，包括 LSTM(Hochreiter and Schmidhuber,1997)、变压器(Vaswani et al.,2017)，CNN(O’Shea andNash,2015)和 ResNet(He et al.,2016);(3)预训练的蛋白质语言模型，其使用蛋白质 FASTA 序列的广泛集合进行预训练，包括 ProtBert(Elnaggaret al.,2021)和 ESM-1b(Rives et al.,2021).两者

	生物快照			人类		BindingDB		
方法	奥罗克	AUPRC	准确(性)	AUROC AUPRC		奥罗克	AUPRC	准确(性)
SVM	0.862	0.864	0.777	0.940 0.006 0.920 0.009		0.939	0.928	0.825
无线电频率(radio frequency)	0.007	0.004	0.011	0.952 0.011 0.953 0.010		0.001	0.002	0.004
DeepConv-DTI	0.886 0.006 0.890 0.006 0.805 0.009 0.980 0.005 0.804			0.981 0.002 0.945 0.002 0.925 0.005 0.882 0.007			0.921 0.016	0.880 0.012
GraphDTA	0.887 0.008	0.890 0.007	0.800 0.007	0.981 0.001 0.982 0.002		0.951 0.002	0.934 0.002	0.888 0.005
MolTrans	0.903 0.005 0.902 0.004	0.834 0.008	0.982 0.002	0.980 0.003 0.960 0.001		0.948 0.002		
DrugBAN	0.895	0.897	0.825	0.980 0.002 0.978 0.003		0.952 0.002	0.936	0.887 0.006
	0.004	0.005	0.010				0.001	0.904 0.004

表 3:在 BindingDB、Human 和 BioSNAP 数据集上的性能比较。(最好，第二好)。基线结果来自 DrugBAN(Bai et al.,2023).

型号#参数。酵母人类				
DDE	205.3K	55.83 3.13	62.77	2.30
莫兰	123.4K	53.00 0.50	54.67	4.43
LSTM	26.7 米	53.62 2.72	63.75	5.12
变压器	21.3 米	54.12 1.27	59.58	2.09
CNN	5.4M	55.07 0.02	62.60	1.67
	11.0 米	48.91 1.78	68.61	3.78
普罗伯特	419.9 米	63.72 2.80	77.32	1.10
普罗伯特*	419.9 米	53.87 0.38	83.61	1.34
ESM-1b	652.4 米	57.00 6.38	78.17	2.91
ESM-1b *	652.4 米	66.07 0.58	88.06	0.24
BioT5	252.1M	64.89 0.43	86.22	0.53

表 4:在酵母和人类数据集上的性能比较(最好，第二好)。评估标准是准确性。*表示仅调整预测头。基线结果来自同级(Xu et al.,2022).

对 ProtBert 和 ESM-1b 进行了两种设置的研究(I)冻结蛋白质语言模型参数，只训练预测头；㈡微调所有模型参数。

结果结果显示在表中 2，所有统计数据来自三次随机运行。在蛋白质溶解度预测任务中，BioT5 的表现优于同行中的所有基线(Xu et al.,2022)基准。在蛋白质定位预测任务中，BioT5 是所有方法中第二好的。值得注意的是，ProtBert 和 ESM-1b 都是在一个大的蛋白质序列语料库上进行预训练的，这个语料库与我们的相当，甚至更大。此外，这些模型比 BioT5 大两到三倍。这些证明了 BioT5 通过整合文本信息在蛋白质性质预测中增强预测能力的潜力。

4.2 多实例预测

4.2.1 药物-靶标相互作用预测

药物-靶标相互作用(DTI)预测在药物发现中起着至关重要的作用，因为它旨在预测给定的药物(分子)和靶标(蛋白质)可以相互作用。我们看到-

选择三个广泛使用的具有二元分类设置的 DTI 数据集，它们是 BioSNAP(Zitniket al.,2018)，BindingDB(Liu et al.,2007)和人类(Liu et al.,2015；Chen et al.,2020).

我们将 BioT5 与两种类型的基线进行比较:(1)传统的机器学习方法，包括 SVM(Cortes and Vapnik,1995)和随机森林(RF)(Ho,1995);(2)深度学习方法包括 DeepConv-DTI(Leeet al.,2019)，GraphDTA(Nguyen et al.,2021)，MolTrans(Huang et al.,2021)和 DrugBAN(Baiet al.,2023)，其中药物和目标特征首先由设计良好药物编码器和蛋白质编码器提取，然后融合用于预测。

结果 BioSNAP，Human 和 BindingDB 数据集的结果如表所示 3。所有统计数据都是从五次随机运行中获得的。在 BioSNAP 和 BindingDB 数据集上，BioT5 在各种性能指标上持续优于其他方法，包括 AUROC、AUPRC 和准确性。对于人类数据集，尽管基于深度学习的模型通常表现出很强的性能，但 BioT5 模型表现出了相对于基线模型的轻微优势。值得注意的是，与大多数基于深度学习的基线相比，我们的 BioT5 不依赖于为分子或蛋白质量身定制的特定设计。对 BioT5 优异性能的一种可能解释是，SELFIES 和 FASTA 表征有效地捕捉了分子和蛋白质的复杂结构和功能，并且它们之间的相互作用信息可以在上下文科学文献或数据库中相应的文本条目中得到很好的描述。

4.2.2 蛋白质相互作用预测

蛋白质-蛋白质相互作用(PPI)预测在理解蛋白质功能和结构中起着至关重要的作用，因为它旨在确定潜力

模型	#Params。	BLEU-2	布鲁-4	胭脂-1	胭脂-2	胭脂-L	流星	Text2Mol
RNN	56 米	0.251	0.176	0.450	0.278	0.394	0.363	0.426
变压器	76 米	0.061	0.027	0.204	0.087	0.186	0.114	0.057
t5-小型	77 米	0.501	0.415	0.602	0.446	0.545	0.532	0.526
T5-base	248 米	0.511	0.423	0.607	0.451	0.550	0.539	0.523
t5-大	7.83 亿	0.558	0.467	0.630	0.478	0.569	0.586	0.563
molt 5-小	77 米	0.519	0.436	0.620	0.469	0.563	0.551	0.540
MolT5 碱基	248 米	0.540	0.457	0.634	0.485	0.578	0.569	0.547
molt 5-大	7.83 亿	0.594	0.508	0.654	0.510	0.594	0.614	0.582
GPT-3.5 涡轮增压(零排量)	> 175B	0.103	0.050	0.261	0.088	0.204	0.161	0.352
GPT-3.5 涡轮增压(10 发 MolReGPT)	> 175B	0.565	0.482	0.623	0.450	0.543	0.585	0.560
MolXPT	350 米	0.594	0.505	0.660	0.511	0.597	0.626	0.594
BioT5	252 米	**0.635**	**0.556**	**0.692**	**0.559**	**0.633**	**0.656**	**0.603**

表 5:分子标题任务的性能比较(最好，第二好)。胭脂分数是 F1 值。地面真实分子与相应文本描述之间的 Text2Mol 得分为 0.609。基线结果来自 MolT5(Edwards et al.,2022)，MolXPT(Liu et al.,2023b)，以及 MolReGPT(Li et al.,2023).

模型	#Params	蓝色	精确 ↑	莱文斯坦 ↓	MACCS FTS	RDK·FTS	摩根 FTS	FCD ↓	Text2Mol ↑	有效性 ↑
RNN	。56 米	0.652	0.005	38.09	0.591	0.400	0.362	4.55	0.409	0.542
变压器	76 米	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	0.906
t5-小型	77 米	0.741	0.064	27.703	0.704	0.578	0.525	2.89	0.479	0.608
T5-base	248 米	0.762	0.069	24.950	0.731	0.605	0.545	2.48	0.499	0.660
t5-大	7.83 亿	0.854	0.279	16.721	0.823	0.731	0.670	1.22	0.552	0.902
molt 5-小	77 米	0.755	0.079	25.988	0.703	0.568	0.517	2.49	0.482	0.721
MolT5 碱基	248 米	0.769	0.081	24.458	0.721	0.588	0.529	2.18	0.496	0.772
molt 5-大	7.83 亿	0.854	0.311	16.071	0.834	0.746	0.684	1.20	0.554	0.905
GPT-3.5 涡轮增压(零排量)	> 175B	0.489	0.019	52.13	0.705	0.462	0.367	2.05	0.479	0.802
GPT-3.5 涡轮增压(10 发 MolReGPT)	> 175B	0.790	0.139	24.91	0.847	0.708	0.624	0.57	0.571	0.887
MolXPT	350 米	-	0.215	-	0.859	0.757	0.667	0.45	**0.578**	0.983
BioT5	252 米	**0.867**	**0.413**	**15.097**	**0.886**	**0.801**	**0.734**	**0.43**	0.576	**1.000**

表 6:基于文本的分子生成任务的性能比较(最好，第二好)。跟随 Edwardset al.(2022)、BLEU、Exact、Levenshtein 和 Validity 在所有生成的分子上计算，而其他度量仅在语法上有效的分子上计算。地面实况的 Text2Mol 分数是 0.609。基线结果来自 MolT5(Edwards et al.,2022)，MolXPT(Liu et al.,2023b)，以及 MolReGPT(Li et al.,2023).

蛋白质对之间的相互作用。关注对等方(Xu et al.,2022)基准，我们在两个 PPI 数据集上执行微调:酵母(Guo et al.,2008)和人类(Pan et al.,2010).

基线用于比较的基线与第节中的相同 4.1.2。

结果结果如表所示 4。所有统计数据都是三次随机运行的结果。在两个 PPI 数据集上，BioT5 显示出比几乎所有基线模型都优越的性能。值得注意的是，BioT5 的性能优于 ProtBert 和 ESM- 1b(对所有参数进行了微调)。这一结果强烈强调了在 BioT5 的预训练过程中整合文本信息的关键作用，这有效地建立了蛋白质之间的深刻联系。尽管我们的模型更小，但它能够利用科学文本中嵌入的非结构化信息和生物数据库中的结构化信息，将蛋白质的全面知识封装在它们不同的上下文中。

4.3 跨模态生成

在本节中，我们评估 BioT5 在跨通道生成任务中的性能。具体来说，我们在分子标题和基于文本的分子生成任务上对 BioT5 进行了微调。这两项任务是由 MolT5(Edwards et al.,2022)并且都使用 ChEBI-20 数据集(Edwardset al.,2021).附录中介绍了评价指标和一些有趣的案例 D 和 G。

4.3.1 分子标题

对于给定的分子，分子克隆任务的目标是提供给定分子的描述。当我们使用自拍序列来表示分子时，这个任务可以被公式化为一个奇异的序列到序列的翻译任务。

基准基准包括:RNN(Medskerand Jain,2001)，变压器(Vaswani et al.,2017)，T5(Raffel et al.,2020)，MolT5(Edwardset al.,2022)，GPT-3.5 涡轮增压[2]零距离射击和

[2]https://openai.com/blog/openai-api

10 发 MolReGPT(Li et al.,2023)设置，而 MolXPT(Liu et al.,2023b).

结果结果如表所示 5。BioT5 仅具有与 MolT5-base 几乎相同数量的参数，但在所有指标上优于所有基线模型，包括那些具有更多参数的模型。Text2Mol 分数为 0.603，这非常接近于地面真实分子和相应描述之间的 Text2Mol 分数 0.609。我们可以将这种优异的性能归功于 BioT5 预训练中诱导的非结构化上下文知识和结构化数据库知识，这有助于模型学习文本和分子之间的复杂关系。

4.3.2 基于文本的分子生成

这是一个与分子标题相反的任务。给定预期分子的自然语言描述，目标是生成符合描述的分子。

基线比较的基线与第节中的基线相同 4.3.1。

结果结果列于表中 6。BioT5 仅使用与 MolT5-base 相似的参数，但在几乎所有的矩阵中都表现出优异的性能。值得注意的是，BioT5 的精确匹配分数超过 molt 5-Large 32.8%，同时保持 1.0 的有效性。这表明 BioT5 不仅生成了与给定文本描述相对应的更多相关分子，而且确保了所生成分子的 100%有效性。BioT5 整体性能的提高可归因于结合了背景知识和数据库知识，以及利用自拍进行分子表征。

5 结论和未来工作

在本文中，我们提出了 BioT5，这是一个全面的预训练框架，能够通过利用具有 100%鲁棒分子表示的结构化和非结构化数据源来捕获生物实体的潜在关系和属性。我们的方法有效地丰富了化学知识和自然语言联想在生物学中的跨模态整合，证明了在各种任务中的显著改进。

对于未来的工作，我们的目标是进一步丰富我们的模型，纳入更多的生物数据类型，如基因组学或转录组学数据，以创建一个更全面的生物预训练框架。此外，我们计划评估 BioT5 预测的可解释性，旨在为研究中的生物系统提供更多的见解。因此，我们预计我们的工作将在计算生物学领域的人工智能模型应用中引发进一步的创新，最终导致对生物系统的更深入理解，并促进更有效的药物发现。

6 限制

BioT5 的一个限制是对每个下游任务进行全参数微调。这样做是因为我们没有观察到使用指令调整(Wei et al.,2022)方法。另一个原因是，使用指令组合来自不同任务的数据会导致数据泄漏。例如，已经注意到 BindingDB 的训练集与 BioSNAP 和 Human 的测试集之间的重叠。此外，我们仅证明了 BioT5 在文本、分子和蛋白质模式中的能力。存在许多其他生物模态，例如 DNA/RNA 序列和细胞，并且在单个模态内或跨多个模态存在许多其他任务。此外，BioT5 主要关注生物实体的序列格式，但其他格式，如 2D 或 3D 结构，也具有重要意义。我们把对这些的进一步探索留给未来的工作。

7 风险

尽管 BioT5 在研究和药物应用方面有潜在的好处，但应防止误用的风险。BioT5 可能无法产生治疗特定疾病的有效分子，并可能产生具有不良副作用的化合物。此外，BioT5 可能被用来制造危险的分子。

8 承认

本工作得到了国家重点研究发展项目(编号:2020YFB1406702)、国家自然科学基金项目(编号:62122089)、北京市杰出青年科学基金项目(编号:BJJWZYJH012019100020098)和 Intelligent 社会治理平台，“双一流”重大创新规划跨学科平台，中国人民大学，中央高校基础研究基金，中国人民大学研究基金。

参考

2023.uni prot:2023 年通用蛋白质知识库。核酸研究，51(D1):D523 - D531。

何塞·胡安·阿尔马格罗·阿尔门特罗斯、卡斯帕·卡埃·索恩-德比、索伦·卡埃·索恩德比、亨里克·尼尔森和奥勒·温瑟。2017.Deeploc:蛋白质预测使用深度学习的亚细胞定位。生物信息。，33(21):3387 - 3395.

阿斯利康。2023.小分子的美好未来:瞄准不可欺骗的人。

Viraj Bagal、Rishal Aggarwal、P. K. Vinod 和 U. Deva Priyakumar。2022.Molgpt:分子生成使用变换器-解码器模型。化学博士。Inf。模型。，62(9):2064 - 2076.

白佩珍，菲利普，，奈杰尔格林，比诺约翰和卢。2021.用于药物-靶标相互作用预测的低偏倚评估的分层聚类分裂。2021 年 IEEE 生物信息学和生物医学国际会议(BIBM)，第 641-644 页。IEEE。

白佩珍、菲利普·米利科维奇、比诺·约翰和·卢。2023.具有域适应性的可解释双线性注意力网络改善了药物靶标预测。自然机器智能，5(2):126 - 136。

萨坦杰夫·班纳吉和阿龙·拉维。2005.METEOR:一种改进的机器翻译自动评测方法与人类判断的相关性。《机器翻译和/或总结的内在和外在评估方法研讨会论文集》@ACL 2005，美国密歇根州安阿伯，2005 年 6 月 29 日，第 65 - 72 页。计算机语言学协会。

伊兹·贝尔塔吉、凯尔·罗和阿曼·科汉。2019.Scib-ert:科技文本的预训练语言模型。《2019 年自然语言处理 Empiri- cal 方法会议和第九届自然语言处理国际联合会议论文集》，EMNLP-IJCNLP 2019，中国香港，2019 年 11 月 3-7 日，第 3613-3618 页。计算语言学协会。

Emmanuel Boutet、Damien Lieberherr、Michael Tog- nolli、Michel Schneider 和 Amos Bairoch。2007.uniprotkb/Swiss-prot:uniprot 知识库的人工注释部分。植物生物信息学:方法和协议，89-112 页。

罗德尼·布里斯特、丹索·阿科-阿杰、益铭·鲍和奥尔加·布林科娃。2015.Ncbi 病毒基因组资源。核酸研究，43(D1):D571 - D577。

Tom Brown，Benjamin Mann，Nick Ryder，Melanie Subbiah，Jared D Kaplan，Prafulla Dhariwal，Arvind Neelakantan，Pranav Shyam，Girish Sastry，Amanda Askell，等。语言模型是一次性学习者。神经信息处理系统进展，33:1877 - 1901。

达科·布蒂娜。1999.无监督数据库聚类基于 daylight 的指纹和 tanimoto 相似-ity:一种快速、自动化的小型集群方式大型数据集。化学博士。Inf。计算机。Sci。，39(4):747 - 750.

凯茜.卡内斯和莎拉.维斯。2013.Pubmed:圣经数据库。《NCBI 手册》，第 2 卷第 1 期。

董，青，梁一增。2013.属性:生成各种模式的工具周的诗。生物信息。，29(7):960 - 962.

陈力帆，，谭，，王，钟飞生，，杨天标，，陈开先，姜华良，。2020.Transformercpi:通过基于序列的深度学习，利用自我注意机制和标签反转实验，改善化合物-蛋白质相互作用预测。生物信息学，36(16):4406 - 4414。

Seyone Chithrananda、Gabriel Grand 和 Bharath Ramsundar。2020.Chemberta:用于分子性质预测的大规模自我监督预处理。arXiv 预印本 arXiv:2010.09885。

科琳娜·科尔特斯和弗拉基米尔·瓦普尼克。1995.支持向量网络。机器学习，20(3):273 - 297。

Suresh Dara、Swetha Dhamercherla、Surender Singh Jadav、CH Madhu Babu 和 Mohamed Jawed Ahsan。2022.药物发现中的机器学习:综述。人工智能评论，55(3):1947 - 1999。

约瑟夫·L·杜兰特，伯顿·A·利兰，道格拉斯·R·亨利，詹姆斯·G·诺斯。2002.用于药物发现的 mdl 关键字的再优化。化学信息与计算机科学杂志，42(6):1273 - 1280。

卡尔·爱德华兹，段曼丽，凯文·罗斯，加勒特·洪克，赵京贤和亨吉。2022.Trans-分子与自然语言的关系。《2022 年自然语言处理经验方法会议论文集》，EMNLP 2022，阿拉伯联合酋长国阿布扎比，2022 年 12 月 7-11 日，第 375-413 页。计算机语言学协会。

卡尔爱德华兹，翟，和恒基。2021.Text2mol:用 nat-进行跨模态分子检索乌拉尔语言查询。《2021 年自然语言处理经验方法会议论文集》，EMNLP 2021，虚拟事件/蓬塔卡纳，多米尼加共和国，2021 年 11 月 7-11 日，第 595-607 页。计算语言学协会。

Ahmed Elnaggar、Michael Heinzinger、Christian Dal- lago、Ghalia Rehawi、王禹、Llion Jones、Tom Gibbs、Tamas Feher、Christoph 安格雷尔、Martin Steinegger 等。通过自我监督学习来理解生活语言。IEEE 模式分析和机器智能汇刊，44(10):7112 - 7127。

方、、刘、、何东龙、张善卓、、吴华、。2022.用于性质预测的几何增强的分子表示学习。自然机器智能，4(2):127–134。

智与。2000.基于氨基酸疏水指数的膜蛋白类型预测。蛋白质化学杂志，19:269–275。

诺埃利亚·费鲁兹、斯特芬·施密特和比尔特·哈克。2022.Protgpt2 是用于蛋白质设计的深度无监督语言模型。自然通讯，13(1):4348。

高天宇、亚当·费舍尔和齐丹·陈。2021.使预先训练的语言模型更好 learners。《计算语言学协会第 59 届年会暨第 11 届国际自然语言处理联合会议论文集》，ACL/IJCNLP 2021，(第 1 卷:长篇论文)，虚拟事件，2021 年 8 月 1-6 日，第 3816-3830 页。计算语言学协会。

、Robert Tinn、、、Naoto Usuyama、、Tristan Naumann、高剑锋和 Hoifung Poon。2021.面向生物医学自然语言处理的特定领域语言模型预处理。ACM 医疗保健计算汇刊，3(1):1–23。

郭，，俞，温志宁，。2008.利用支持向量机结合自协方差预测蛋白质序列中的蛋白质相互作用。核酸研究，36(9):3025–3030。

迦娜·黑斯廷斯、加雷斯·欧文、阿德里亚诺·德克尔、马斯·恩尼斯、纳姆拉塔·卡莱、文卡特什·穆图克里希南、史蒂夫·特纳、尼尔·斯温斯顿、佩德罗·门德斯和克里斯托弗·斯坦贝克。2016.2016 年的 Chebi:改进的服务和不断扩大的代谢物收集。核酸研究，44(D1):d 1214–d 1219。

何、、、任、。2016.用于图像识别的深度残差学习。IEEE 计算机视觉和模式识别会议论文集，第 770-778 页。

斯蒂芬·海勒、艾伦·麦克诺特、斯蒂芬·斯坦、德米特里·切霍夫斯科伊和伊戈尔·普莱特涅夫。2013.全球化学结构标识符标准。化学信息学杂志，5(1):1–9。

天金豪。1995.随机决策森林。《第三届文件分析与识别国际会议论文集》，第 1 卷，第 278-282 页。

Sepp Hochreiter 和 Jürgen Schmidhuber。1997.长短期记忆。神经计算，9(8):1735–1780。

黄克新，，卢卡斯玻璃，和孙。2021.MolTrans:用于药物-靶标相互作用预测的分子相互作用转换器。生物信息学，37:830–836。

John J Irwin、Khanh G Tang、Jennifer Young、Chin- zorig Dandarchuluun、Benjamin R Wong、Munkhzul Khurelbaatar、Yurii S Moroz、John Mayfield 和 Roger A Sayle。2020.zinc 20——用于配体发现的免费超大规模化学数据库。化学信息与建模杂志，60(12):6065–6073。

Sameer Khurana、Reda Rawi、Khalid Kunji、Gwo- 于闯、Halima Bensmail 和 Raghvendra Mall。2018.Deepsol:深度学习框架基于序列的蛋白质溶解度预测。生物信息。，34(15):2605–2613.

金成焕、、、阿斯塔·金杜-莱特、、何、、李、本杰明·A·舒梅克、保罗·A·泰森、等。公共化学 2023 更新。核酸研究，51(D1):d 1373–d 1380。

金成焕、保罗·泰森、程铁军、张建、阿斯塔·金杜莱和埃文·波顿。2019.Pug-view:编程访问集成在 pubchem 中的化学注释。化学信息学杂志，11(1):1–11。

托马斯·n·基普夫和马克斯·韦林。2017.Semi-基于图卷积的监督分类 networks。在 2017 年 4 月 24 日至 26 日在法国 ICLR 土伦举行的第五届学习表征国际会议上，会议记录。OpenReview.net。

Mario Krenn，Qianxiang Ai，Senja Barthel，Carson，Angelo Frei，Nathan C Frey，Pascal Friederich，Théophile Gaudin，Alberto Alexander Gayle，Kevin Maik Jablonka 等。自拍和分子弦表示的未来。图案，3(10):100588。

Mario Krenn、Florian HSE、AkshatKumar Nigam、Pas- cal Friederich 和 Alan Aspuru-Guzik。2020.自参照嵌入字符串(自拍):一个 100% ro- bust 分子字符串表示。机器学习:科学与技术，1(4):045024。

工藤多久和约翰·理查森。2018.例句:一个简单且独立于语言的子词 tok-神经文本处理的 enizer 和 detokenizer。《2018 自然语言处理经验方法会议论文集》，EMNLP 2018:系统演示，比利时布鲁塞尔，2018 年 10 月 31 日-11 月 4 日，第 66–71 页。计算语言学协会。

格雷格·兰德勒姆。2021.Rdkit:开源 cheminfor-matics 软件。GitHub 发布。

英戈·李，钟洙·金和何正南。2019.DeepConv-DTI:通过对蛋白质序列进行卷积的深度学习来预测药物-靶标相互作用。PLoS 计算生物学，15。

李真旭，尹元真，金成东，金东妍，金善奎，陈浩素和姜在宇。2020.Biobert:用于生物医学文本挖掘的预训练生物医学语言表示模型。生物信息学，36(4):1234－1240。

、刘运清、范、、、、肖。2023.使用大型语言模型进行分子标题翻译的分子发现:chatgpt 的观点。arXiv 预印本 arXiv:2306.06615。

林金耀。2004.Rouge:一个自动评估摘要的包。在 74-81 页的《文本摘要分支》。

、Halil Akin、Roshan Rao、Brian Hie、Zhu、、Allan dos Santos Costa、Maryam Fazel-Zarandi、Tom Sercu、Sal 等。进化尺度上的蛋白质序列的语言模型使得精确的结构预测成为可能。BioRxiv。

戴维·J·李普曼和威廉·R·皮尔逊。1985.快速灵敏的蛋白质相似性搜索。科学，227(4693):1435－1441。

卡洛琳·科普斯科姆。2000.医学主题词。医学图书馆协会公报，88(3):265。

、孙建江、、关、、周水耕。2015.通过构建高度可信的阴性样本来改进化合物-蛋白质相互作用预测。生物信息学，31(12):i221－i229。

刘、王、刘维扬、琼·拉森比、郭宏宇和。2022.Pre-用 3d ge-训练分子图表示ometry。在 2022 年 4 月 25 日至 29 日于 ICLR 2022 举行的第十届国际学习代表大会虚拟活动中。OpenReview.net。

刘、、朱、陆家瑞、、聂伟力、安东尼·吉特、肖、、郭宏宇和阿尼玛·阿南德·库玛。2023a。文本引导的蛋白质设计框架。arXiv 预印本 arXiv:2302.04611。

刘体清，林玉梅，，罗伯特·N·约里森，迈克尔·K·吉尔森。2007.Bindingdb:一个实验确定的蛋白质-配体结合亲和力的网络数据库。核酸研究，35(增刊_ 1):D198－D201。

刘、米勒·奥特、纳曼·戈亚尔、·杜、曼-达尔·乔希、·陈、奥梅尔·利维、、卢克·塞特勒莫耶和韦塞林·斯托扬诺夫。2019.Roberta:一种稳健优化的 bert 预训练方法 arXiv 预印本 arXiv:1907.11692

刘泽群、、夏颖策、、舒、、、。2023b。Molxpt:用文本包装分子生成性预训练。第 61 届计算机协会年会论文集

语言学(第 2 卷:短文)，ACL 2023，加拿大多伦多，2023 年 7 月 9-14 日，第 1606-1616 页。计算语言学协会。

伊利亚·洛希洛夫和弗兰克·哈特。2019.Decoupled 权重衰减正则化。在 2019 年 5 月 6 日至 9 日在美国路易斯安那州新奥尔良举行的 2019 年 ICLR 第七届国际学习代表大会上。OpenRe-view.net。

罗、、孙、、夏颖策、、、潘海峰和。2022.Biogpt:用于生物医学文本生成和挖掘的预训练生成转换器。生物信息学简报，23(6)。

罗、、洪、、、吴玉帅、聂在勤。2023.用显性和隐性知识进行人工智能药物发现。arXiv 预印本 arXiv:2305.01523。

拉里·R·梅德斯克和 LC·贾恩。2001.递归神经网络。设计与应用，5:64－67。

弗雷德里克·P·米勒、艾格尼丝·F·范多姆和约翰·麦克布雷斯特。2009.Levenshtein 距离:信息论，计算机科学，string(计算机科学)，string metric，damerau? 莱文斯坦距离，拼写检查，汉明距离。

皮奥特·纳罗特。2023.nanoT5。

Thin Nguyen、Hang Le、T. Quinn、特里·阮明、Thuc Duy Le 和 Svetha Venkatesh。2021.用图形神经网络预测药物与靶标的结合亲和力。生物信息学，37(8):1140－1147。

诺埃尔·奥博伊尔和安德鲁·达尔克。2018.deepsmiles:smiles 的改编版，用于化学结构的机器学习。

凯龙·奥谢和瑞安·纳什。2015.卷积神经网络导论。arXiv 预印本 arXiv:1511.08458。

潘小永、张亚男和沈红彬。2010.基于潜在主题特征从氨基酸序列大规模预测人类蛋白质相互作用。蛋白质组研究杂志，9(10):4992－5001。

Kishore Papineni、Salim Roukos、Todd Ward 和魏-朱婧。2002.Bleu:一种自动评估的方法机器翻译的现状。计算机语言学协会第 40 届年会论文集，2002 年 7 月 6-12 日，美国宾夕法尼亚州费城，第 311-318 页。ACL。

威廉·皮尔逊和大卫·李普曼。1988.改进的生物序列比较工具。美国国家科学院学报，85(8):2444－2448。

Suraj Peri、J .丹尼尔·纳瓦罗、Ramars Amanchy、Troels Z Kristiansen、Chandra Kiran Jonnalagadda、梵持 Surendranath、Vidya Niranjan、Babylakshmi Muthusamy、TKB·甘地、Mads Gronborg 等人，2003 年。开发人类蛋白质参考数据库作为研究人类系统生物学的初始平台。基因组研究，13(10):2363–2371。

Kristina Preuer、Philipp Renz、Thomas Unterthiner、Sepp Hochreiter 和 Günter Klambauer。2018.弗雷歇化学网距离:一种生成性度量药物发现中的分子模型。化学博士。Inf。模型。，58(9):1736–1741.

亚历克·拉德福德、卡蒂克·纳拉辛汉、蒂姆·萨利曼斯、伊利亚·苏茨基弗等人，2018 年。通过生成性预训练提高语言理解能力。

亚历克拉德福德，杰弗里吴，Rewon 儿童，大卫栾，达里奥阿莫代伊，伊利亚苏茨基弗，等人 2019 年。语言模型是无人监督的多任务学习者。OpenAI 博客，1(8):9。

科林·拉弗尔、诺姆·沙泽尔、、凯瑟琳·李、·纳朗、迈克尔·马泰纳、周燕琪、和彼得·刘。2020.用统一的文本-文本转换器探索迁移学习的局限性。机器学习研究杂志，21(1):5485–5551。

Alexander Rives，Joshua Meier，Tom Sercu，Goyal，，Jason Liu，Demi Guo，Myle Ott，C Lawrence Zitnick，Jerry Ma，等。生物结构和功能来自于对 2．5 亿个蛋白质序列的无监督学习。美国国家科学院院刊，118(15):e2016239118。

斯蒂芬·罗伯逊和雨果·萨拉戈萨。2009.The 概率相关性框架:BM25 和 be-yond。找到了。趋势信息。Retr。，3(4):333–389.

大卫·罗杰斯和马修·哈恩。2010 年 a。扩展连接指纹。化学信息与建模杂志，50(5):742–754。

大卫·罗杰斯和马修·哈恩。2010 年 b。Extended-连接指纹。化学博士。Inf。模型。，50(5):742–754.

俞蓉，卞亚涛，，徐，谢维扬，，，黄。2020.大规模分子数据的自监督图形转换器。神经信息处理系统进展，33:12559–12571。

Vijayakumar Saravanan 和 Namasivayam Gautham。2015.利用计算生物学进行精确的线性 b 细胞表位预测:一种新的基于氨基酸组成的特征描述符。组学：综合生物学杂志，19(10):648–658。

纳丁·施耐德、罗杰·a·塞尔和格雷戈里·a·兰·德拉姆。2015.整理好你的原子--一个开放的--一种新的健壮分子的源代码实现 lar 规范化算法。化学博士。Inf。模型。，55(10):2111–2120.

马丁·施泰因格和约翰内斯·索丁。2018.在线性时间内聚类巨大的蛋白质序列集。自然通讯，9(1):2542。

提格.斯特林和约翰. j .欧文。2015.ZINC 15 -每个人的配体发现。化学博士。Inf。模型。，55(11):2324–2337.

苏冰、杜大钊、赵阳、、江、饶安义、、、嵇。2022.将分子图与自然语言联系起来的分子多模态基础模型。arXiv 预印本 arXiv:2209.05481。

吴镇星、郑敏彬、崔永和、金东妍、李金旭和姜在宇。2022.Bern2:高级神经生物医学命名实体识别和规范化工具。生物信息学，38(20):4837–4839。

Suzek、黄宏展、Peter McGarvey、Raja Mazumder 和 Cathy H Wu。2007.Uniref:全面且非冗余的 uniprot 参考集群。生物信息学，23(10):1282–1288。

罗斯·泰勒、马尔钦·卡尔达斯、吉列姆·库库鲁勒、托马斯·西沙龙、安东尼·哈特肖恩、埃尔维斯·萨拉维亚、安德鲁·波尔顿、维克多·克尔凯兹和罗伯特·斯托伊尼克。2022.卡拉狄加:科学的大型语言模型。arXiv 预印本 arXiv:2211.09085。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan Gomez、ukasz Kaiser 和 Illia Polosukhin。2017.你需要的只是关注。神经信息处理系统进展，30。

、王、和阿米尔·巴拉蒂·法里马尼。2022.基于图形神经网络的表征的分子对比学习。自然机器智能，4(3):279–287。

贾森·魏、马腾·博斯马、赵永健、顾开文、亚当斯·禹卫、布赖恩·莱斯特、、戴安祖和郭伟民。2022.Finetuned 语言模型是零射击学习者。第十届国际学习代表大会，ICLR 2022，虚拟活动，2022 年 4 月 25-29 日。OpenReview.net。

大卫·魏宁格。1988.化学语言和信息系统。1.方法论和编码规则介绍。化学信息与计算机科学杂志，28(1):31–36。

戴维·魏宁格、阿瑟·魏宁格和约瑟夫·魏宁格。1989.微笑。2.生成唯一微笑符号的算法。化学信息与计算机科学杂志，29(2):97–101。

大卫·S·维沙特、扬尼克·D·弗南格、安·C·郭、埃尔维斯·J·罗、安娜·马尔库、杰森·R·格兰特、坦维尔·萨杰德、、卡琳·李、齐纳特·萨耶达等 2018。药物银行 5.0:2018 年药物银行数据库的重大更新。核酸研究，46(D1):d 1074–d 1082。

秦镇·吴、巴拉思·拉姆松达、埃文·N·范伯格、约瑟夫·戈梅斯、迦勒·格尼塞、安尼施·S·帕普、卡尔·莱斯温和维贾伊·潘德。2018.分子机器学习的基准。化学科学，9(2):513–530。

许汉文，艾迪·沃里克，潘海峰，拉斯·奥尔特曼和王胜。2023a。使用生物翻译器进行零剂量生物医学分类的多语言翻译。自然通讯，14(1):738。

徐明浩、袁新玉、桑提亚哥·米雷特和汤集安。2023b。蛋白质序列和生物医学文本的多模态学习。arXiv 预印本 arXiv:2301.12040。

、张作柏、陆家瑞、、张、、刘润成和。2022.Peer:一个全面的多任务蛋白质序列理解基准。神经信息处理系统进展，35:35156–35173。

曾珍妮，，，孙茂松。2022.一个连接分子结构和生物医学文本的深度学习系统，其理解能力堪比人类专业人员。自然通讯，13(1):862。

、刘琦、、和李志刚。2021.基于模体的图形自监督学习在分子性质预测中的应用。神经信息处理系统进展，34:15870–15882。

Marinka Zitnik、Rok Sosic 和 Jure Leskovec。2018.生物快照数据集:斯坦福生物医学网络数据集。注意:*http://snap。斯坦福。edu/生物数据被引用，5(1)*。

A 再现性

我们BioT5的代码可从以下网址获得https://github.com/QizhiPei/BioT5。

B NER和实体链接过程

我们遵循KV-PLM(Zeng et al.,2022)和MolXPT(Liu et al.,2023b)对科学文本中出现的生物实体名称进行命名实体识别(NER)和实体链接。更具体地说，我们首先利用BERN2(Sunget al.,2022)，一种生物医学领域的高级神经命名实体识别(NER)工具，用于识别分子或蛋白质修饰的所有实例。随后，我们将它们映射到可公开访问的知识库中的对应实体。对于分子，我们使用ChEBI(Hastingset al.,2016)和网格(Lipscomb,2000)数据库，对于蛋白质，我们使用NCBI基因(Brister et al.,2015)数据库。然后我们可以得到匹配实体对应的分子自拍照和蛋白质FASTA。如图所示4，用于

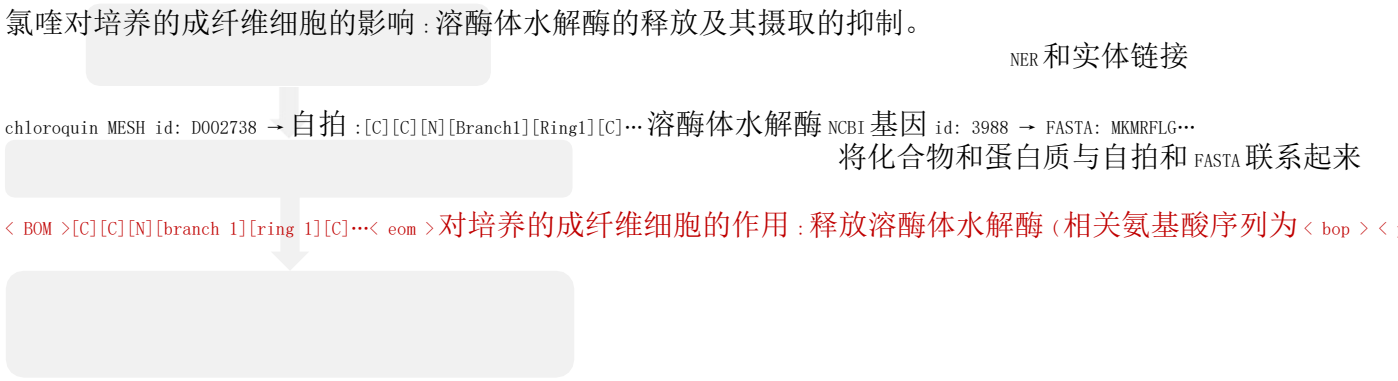

图4:包装文本匹配和映射过程。

分子，我们直接把所有检测到的名字都换成它的自拍串；对于蛋白质，由于长度限制，如果一个句子由多个蛋白质实体组成，我们只随机选择一个将蛋白质FASTA附加到名称上。添加蛋白质FASTA而不是替换的动机是基因被转录和翻译产生蛋白质。因此，与直接代表分子的分子名称不同，基因名称与蛋白质FASTA之间的关系是间接的。注意，替换或附加不会出现在每个句子中。只有那些检测到生物实体的人才会进行上述过程。

C 字典和自拍转换

对于分子相关的数据集，当只提供微笑时，我们利用自拍[3]把微笑转换成自拍的包。

D 分子文本生成度量

我们跟随Edwards et al.（2022)对分子标题和基于文本的分子生成任务使用相同的评估指标。为了确保公平的比较，在计算这些指标之前，我们将分子SEIFLES转换为SMILES。

D.1 分子标题度量

在分子标题任务中，像BLEU(Papineni et al.,2002)，胭脂(Lin,2004)，还有流星(Banerjee and Lavie,2005)用于评估生成的描述与基本事实描述的接近程度。我们还采用Text2Mol度量，它是由Ed-wards et al.（2021)并使用预先训练的模型来测量描述和基本事实分子之间的相似性。更高的相似性意味着给定的文本描述与分子更相关，并且也计算基础事实描述和分子之间的Text2Mol分数用于比较。

D.2 基于文本的分子生成度量

由于分子可以用生物序列结构来表示，NLP度量如BLEU(Pa-pineni et al.,2002)并且生成的和真实的微笑之间的精确匹配分数被直接应用于评估。此外，我们还报告了分子特异性指标的性能：三分子指纹(FTS)相似性得分-MACCS(Durant et al.,2002)，RDK(Schneideret al.,2015)，而摩根(Rogers and Hahn,2010a)；莱文斯坦距离(Miller et al.,2009)；FCD分数(Preuer et al.,2018)，其根据基于预先训练的“化学网”的生物信息来测量分子相似性；有效性，这是可由RDKit处理的有效微笑的百分比(Landrum,2021).Text2Mol度量也用于测量分子微笑和基本事实描述之间的相似性。

[3]https://github.com/aspuru-guzik-group/selfies

E 培训前详细信息

E.1 特殊代币

在 BioT5 的预训练中，我们对从 PubChem(Kim et al.,2023) 和 Swiss-Prot(Boutet et al.,2007)分开。我们使用特殊的标记来格式化这些数据库条目的文本描述，这些标记用作嵌入科学上下文和结构的锚。对于分子，我们用分子名称和描述来表示其名称和描述，包括性质、功能等。对于蛋白质，类似于 Xu et al. （2023b)，我们使用蛋白质名称、功能、亚细胞定位
阳离子和蛋白质家族来表示它的名称、功能、在细胞中的位置和拓扑结构以及它所属的家族。通过顺序连接这些字段，省略任何缺失的字段，创建完整的文本描述。通过特殊的标记，我们可以有效地编码与每个生物实体相关的复杂信息。

E.2 超参数

我们使用代码库 nanoT5(Nawrot,2023)进行 BioT5 预培训。我们在八个 NVIDIA 80GB A100 GPUs 上对 BioT5 进行了 350K 步的预训练。每个 GPU 的批处理大小为 96，其中一个批处理包含六种类型的数据。分子-文本和蛋白质-文本对的“翻译”方向是以 0.5 的概率为每个样品随机选择的。我们使用 AdamW(Loshchilov and Hutter,2019)使用均方根(RMS)缩放优化器进行优化。学习率调度程序是余弦退火，基本学习率设置为 1e 2，最小学习率设置为 1e 5。热身步数 10000，辍学率 0.0。预训练的最大输入长度为 512。与绝对位置编码不同，T5(Raffel et al.,2020)使用相对位置编码。这使得模型对于不同长度的输入是灵活的，这有助于下游的微调。

F 微调细节

在本节中，我们提供了有关下游任务的详细信息，包括数据集、比较基线和提示。有关下游任务的一些统计数据显示在表中 7 显示提示时，⟨SELFIES⟩指的是分子自身，⟨FASTA⟩指的是蛋白质 FASTA。

F.1 单实例预测

F.1.1 分子性质预测

所有数据集分别使用 8 ： 1 ： 1 的比率进行分割，用于训练、验证和测试。我们使用支架分裂法，根据贝米斯-穆尔科支架表示法对分子进行分类。

数据集

(1) BBBP(血脑屏障通透性)旨在帮助屏障通透性的建模和预测。它包括使用二元标记分类的化合物，表明它们是否可以穿透血脑屏障。

(2) tox 21(“21 世纪毒理学”)倡议建立了一个公众可访问的数据库，该数据库量化了各种化合物的毒性水平。该数据集包括大约 8000 种化合物的定性毒性评估(二元标签)，针对 12 种不同的生物途径，如核受体和应激反应机制。

(3) ClinTox 数据集对比了 FDA 批准的药物和由于毒性问题而在临床试验中失败的药物。该数据集结合了 1，491 种具有确定化学结构的药物化合物的两个分类目标:(I)在临床试验中是否存在毒性；(ii)FDA 批准或未批准。

(4) HIV 数据集评估了超过 40，000 种化合物对 HIV 复制的抑制潜力。筛选结果分为三类:确认无活性(CI)、确认活性(CA)和确认中度活性(CM)。随后，后两个标签被合并，将任务转化为非活动(CI)和活动(CA 和 CM)类别之间的二元分类。

(5) BACE 数据集提供了针对人类 β 分泌酶 1 (BACE-1)的抑制剂集合的定量 IC50 值和定性二元标记。

(6) SIDER(副作用资源)是一个全面的数据库，包括上市药物及其相应的药物不良反应(ADR)。根据 MedDRA 分类，SIDER 中的药物副作用分为 27 个系统器官类别。该数据集编码了 1，427 种已批准药物的数据。

基线

(1) 格罗弗(Rong et al.,2020)包含 Mes-

任务/数据集	任务类型	#火车	#验证	#测试
分子性质预测				
BBBP	分子分类	1,631	204	204
Tox21	分子分类	6,264	783	784
克林托克斯	分子分类	1,181	148	148
艾滋病病毒	分子分类	32,901	4,113	4,113
BasicAutomaticCheckoutEquipment 基本自动检验装置	分子分类	1,210	151	152
帮派成员	分子分类	1,141	143	143
蛋白质性质预测				
溶解度预测	蛋白质分类	62,478	1,999	1,999
本地化预测	蛋白质分类	5,184	1,749	1,749
药物-靶标相互作用预测				
生物快照	分子-蛋白质分类	19,224	2,747	5,493
人类	分子-蛋白质分类	4,197	600	1,200
BindingDB	分子-蛋白质分类	34,439	4,920	9,840
蛋白质相互作用预测				
酵母	蛋白质对分类	4,945	394	394
人类	蛋白质对分类	35,669	237	237
分子标题和基于文本的分子生成				
切比-20	分子文本翻译	26,407	3,301	3,300

表 7:下游任务描述，包括任务或数据集名称、类型和每个分割的大小。

sage 在变压器式架构内传递网络，并在没有任何监督的情况下在大规模分子数据集上进行预训练。G-Contextual 和 G-Motif 是 GROVER 的两个变体，分别在上下文属性预测任务和 Motif 预测任务上进行预训练。

(2) GraphMVP(Liu et al.,2022)通过利用分子 2D 拓扑结构和 3D 几何视图之间的对应性和一致性来采用自我监督学习。

(3) MGSSL(Zhang et al.,2021)结合了一个新颖的图形神经网络自监督基序生成框架。

(4)MolCLR(Wang et al.,2022)是一个自我监督的学习框架，它利用了大量未标记的独特分子(大约 1000 万)

(5)宝石(Fang et al.,2022)具有特别设计的基于几何的图形神经网络结构和几个专用的几何级自监督学习策略，以有效地捕获分子几何知识。

(6)KV-PLM(Zeng et al.,2022)是为分子表征学习设计的基于 BERT 的模型，在预训练过程中，分子微笑被附加在其名称之后。这

分子名称和微笑序列的组合允许模型捕捉文本和结构信息，从而增强其在各种下游任务中的性能。

(7) 卡拉狄加(Taylor et al.,2022)是一个基于 GPT 的大型语言模型，它是在各种语料库上预先训练的，如论文、代码、微笑、蛋白质序列等。

(8) 嫫母(Su et al.,2022)通过对比学习使用分子图及其语义相关的文本数据进行预训练。

(9) MolXPT(Liu et al.,2023b)是一个统一的基于 GPT 的语言模型，用于文本和在“包装”文本上预先训练的分子，其中分子名称被替换为相应的微笑。提示

对于上面提到的六个 MoleculeNet 数据集，提示仅在任务定义上有所不同。因此，我们将只为第一个数据集提供指令和输出，其余的数据集将遵循相同的格式。

(1) BBBP

任务定义:定义:BBBP 数据集的分子性质预测任务(二元分类任务)。血脑屏障穿透(BBBP)数据集是为模型设计的

屏障渗透率的测定和预测。如果给定的分子能够穿透血脑屏障，请用“是”表示。否则，请回答“否”。

说明:现在完成下面的例子-输入:分子:⟨bom⟩⟨SELFIES⟩⟨eom⟩输出:。输出:抑制剂为是，反之为否。

(2) Tox21

任务定义:定义:Tox21 数据集的分子性质预测任务(二元分类任务)。Tox21 数据集包含 8k 化合物对 12 种不同靶标的定性毒性测量，包括核受体和应激反应途径。如果给定的分子能够激活/改变/影响目标，请用“是”表示。否则，请回答“否”。其中 target 代表每个子任务的相应受体、结构域、元素、基因、潜能或途径。

(3) 克林托克斯

任务定义:定义:ClinTox 数据集的分子性质预测任务(二元分类任务)。ClinTox 数据集比较了 FDA 批准的药物和因毒性原因未通过临床试验的药物。如果给定的分子是子任务，请用“是”表示。否则，请回答“否”。在哪里⟨Subtask⟩要么是有毒的，要么是 FDA 批准的。

(4) 艾滋病病毒

任务定义:定义:HIV 数据集的分子特性预测任务(二元分类任务)。HIV 数据集由药物治疗计划(DTP)艾滋病抗病毒筛选引入，该筛选测试了超过 40，000 种化合物抑制 HIV 复制的能力。如果给定的分子能够抑制艾滋病毒复制，请用“是”表示。否则，请回答“否”。

(5) BasicAutomaticCheckoutEquipment 基本自动检验装置

任务定义:定义:BACE 数据集的分子性质预测任务(二元分类任务)。BACE 数据集提供了一组人 β-分泌酶 1 (BACE-1)抑制剂的定性(二元标记)结合结果。如果给定的分子能抑制 BACE-1，请用“是”表示。否则，请回答“否”。

(6) 帮派成员

任务定义:定义:SIDER 数据集的分子性质预测任务(二元分类任务)。副作用资源(SIDER)是上市药物和不良药物反应(ADR)的数据集。如果给定的分子会引起⟨side effect⟩的副作用，请通过

“是的”。否则，请回答“否”。其中副作用是指每个子任务的相应副作用。

F.12 蛋白质性质预测数据集

(1) 溶解度预测是预测蛋白质是否可溶。我们遵循与 DeepSol 相同的分割方法(Khurana et al.,2018).

(2) 定位预测旨在预测蛋白质是“膜结合的”还是“可溶的”，这是亚细胞定位预测任务的简单版本。我们遵循与 DeepLoc 相同的分割方法(Armenteros et al.,2017).

基线

(1) 特征工程师。二肽与预期平均值的偏差(Saravanan and Gau-tham,2015)特征描述符，由 400 个维度组成，基于蛋白质序列中的二肽频率。莫兰特征描述符(莫兰相关性)(Feng and Zhang,2000)，有 240 个维度，描述了蛋白质序列中氨基酸属性的分布。

(2) 蛋白质序列编码器，包括 LSTM(Hochreiter and Schmidhuber,1997)、变压器(Vaswani et al.,2017)，CNN(O’Sheaand Nash,2015)和 ResNet(He et al.,2016).最后一层中的氨基酸特征被聚集用于最终预测。

(3) 预先训练的蛋白质语言模型。波特(Elnaggar et al.,2021)和 ESM-1b(Riveset al.,2021)都使用掩蔽语言建模(MLM)目标在大规模蛋白质序列数据集上进行了预训练。具体来说，ProtBert 是根据从 BFD 数据库(Steineg-ger and Söding,2018)，而 ESM-1b 是在来自 UniRef50(Suzek et al.,2007).

提示

(1) 溶解度预测

任务定义:溶解度数据集的蛋白质溶解度预测任务(二元分类任务)。如果给定的蛋白质是可溶的，请用“是”表示。否则，请回答“否”。

指令现在完成下面的例子-输入:蛋白质:⟨bom⟩⟨FASTA⟩⟨eom⟩输出:。

输出:对于可溶性蛋白质是或否。

(2) 本地化预测

任务定义:蛋白质亚细胞定位任务(一个二元分类任务)。如果给定的蛋白质是膜结合的，请用“是”表示。否则(蛋白质是可溶的)，通过“否”回答。
指令现在完成下面的例子-输入:蛋白质:⟨bom⟩⟨FASTA⟩⟨eom⟩输出:。
输出:膜结合蛋白为是，可溶性蛋白为否。

F.2 多实例预测

F21 药物-靶标相互作用预测数据集

(1) 生物快照(Zitnik et al.,2018)是从DrugBank数据库(Wishart et al.,2018)并由创建者Huang et al. (2021)和Zitnik et al.(2018).它由4510种药物和2181种蛋白质组成。这个数据集是平衡的，既包含经过验证的阳性相互作用，也包含从看不见的配对中随机选择的相同数量的阴性样本。
(2) BindingDB(Liu et al.,2007)是一个可访问的在线数据库，包含实验验证的结合亲和力。它的主要焦点是类似药物的小分子和蛋白质之间的相互作用。我们跟随Bai et al.(2023)来使用BindingDB数据集的修改版本，它以前是由Bai et al. (2021)具有减小的偏差。
(3) 人类(Liu et al.,2015；Chen et al.,2020)是通过包含高度可信的阴性样本来构建的。跟随Bai et al. (2023)，我们还使用人类数据集的平衡版本，它包含相同数量的阳性和阴性样本。
基线
我们比较了BioT5与以下六个模型在DTI任务上的性能。
(1) 支持向量机(Cortes and Vap-nik,1995)(SVM)在连接的指纹ECFP4(Rogers and Hahn,2010b)(扩展连接指纹，最多四个键)和PSC(Cao et al.,2013)(伪氨基酸组成)特征。
(2) 随机森林(Ho,1995)(RF)在连接的指纹ECFP4和PSC特征上。
(3) DeepConv-DTI(Lee et al.,2019)使用完全连接的神经网络对ECFP4药物指纹进行编码，使用卷积神经网络(CNN)以及全局最大池层从蛋白质序列中提取特征。然后将药物和蛋白质特征连接起来输入一个完全连接的神经网络进行最终预测。
(4) GraphDTA(Nguyen et al.,2021)使用图形神经网络(GNNs)对药物分子图形进行编码，使用CNN对蛋白质序列进行编码。药物和蛋白质表示的导出向量被连接用于相互作用预测。
(5) 莫尔特兰(Huang et al.,2021)使用变压器架构来编码药物和蛋白质。然后使用基于CNN的交互模块来捕捉他们的交互。
(6) 德鲁巴(Bai et al.,2023)使用图形卷积网络(GCN)(Kipf and Welling,2017)和1D CNN来编码药物和蛋白质序列。然后采用双线性注意网络来学习药物和蛋白质之间成对的局部相互作用。产生的联合表示由完全连接的神经网络解码。提示
任务定义:定义:数据集数据集的药物靶标相互作用预测任务(二元分类任务)。如果给定的分子和蛋白质可以相互作用，请用“是”表示。否则，请回答“否”。其中数据集是上述三个DTI数据集之一。
说明:现在完成以下示例-输入:分子:bom SELFIES eom蛋白质:bom FASTA eom输出:。
输出:阳性标签为是，否则为否。

F22 蛋白质相互作用预测数据集

(1) 酵母(Guo et al.,2008)涉及到确定两种酵母蛋白是否相互作用。负对来自不同的亚细胞位置。跟随(Xu et al.,2022)，数据集根据蛋白质序列相似性被分割并去除冗余，这允许跨不同蛋白质序列的泛化的评估。
(2) 人类(Pan et al.,2010)包括确定两种人类蛋白质是否相互作用。它包含来源于人类蛋白质参考数据库(HPRD)的阳性蛋白质对(Periet al.,2003)和来自不同亚细胞位置的负对。数据集分割方案类似于酵母PPI预测的方案，训练/验证/测试的比例为8 : 1 : 1。基线比较的基线与秒-中的蛋白质性质预测任务相同

象征式互动 F.1.2。

提示

任务定义:数据集数据集的蛋白质相互作用预测任务(二元分类任务)。如果给定的两种酵母蛋白(蛋白 A 和蛋白 B)可以相互作用，请用“是”表示。否则，请回答“否”。{Dataset}要么是酵母要么是人类。说明:现在完成下面的例子-输入:蛋白质_A: ⟨bom⟩⟨FASTA⟩⟨eom⟩蛋白质_B: ⟨bom⟩⟨FASTA⟩⟨eom⟩输出:。
输出:阳性标签为是，否则为否。

F.3 跨模态生成

F.3.1 分子标题数据集

我们使用 Text2mol 创建的 ChEBI-20 数据集(Edwards et al.,2021)，它由 33，010 个分子-文本对组成，20 表示每个文本描述超过 20 个单词。数据集被分成 8 : 1 : 1 用于训练、验证和测试。

基线

(1) RNN(Medsker and Jain,2001)在 ChEBI-20 数据集上从头开始训练。

(2) 变压器(Vaswani et al.,2017)包含 6 个编码器和解码器层，在 ChEBI-20 数据集上从头开始训练。

(3) T5(Raffel et al.,2020)直接在来自公共检查点的 ChEBI-20 数据集上进行微调[4]有三种不同的型号:小型、基本型和大型。注意，在最初的 T5 预训练中没有引入分子域知识。

(4) MolT5(Edwards et al.,2022)在来自锌-15 数据集(Ster-ling and Irwin,2015)和来自 C4 数据集的一般文本(Raffel et al.,2020)以便 MolT5 具有关于这两个域的先验知识。它还有三种不同的尺寸:小号、中号和大号。然后在 ChEBI-20 数据集上进一步微调。

(5) GPT-3.5 涡轮增压(Li et al.,2023)由直接调用 OpenAI API 使用，无需进一步微调。输入包括以下五个部分 Li et al. (2023):角色标识、任务描述、示例、输出指令和用户输入提示。这些样本是由摩根指纹公司(Butina,1999)分子标题任务的相似性和到 BM25(Robertson and Zaragoza,2009)用于基于文本分子生成任务。

(6) MolXPT(Liu et al.,2023b)是由 PubChem(Kim et al.,2023)，生物医学文本来自 PubMed(Canese andWeis,2013)，以及用相应的微笑替换分子名称的“包装”文本。提示与基础事实输出为是或否的分类任务不同，分子标题任务的输出是文本序列。

任务定义:定义:给你一张分子自拍。你的工作是用英语生成适合分子自拍的分子描述。
指令:现在完成以下示例-输入:<bom>{SELFIES}<eom>输出:。
输出:{Text Description}

F.3.2 基于文本的分子生成

这是分子标题的反向任务。输入是所需分子的文本描述，输出是相应的分子本身。数据集和比较基线与第节中的分子标题相同 F.3.1 所以这里只提供提示。

提示

任务定义:定义:给你一个英文的分子描述。你的工作是生成符合描述的分子自拍照。
指令:现在完成以下示例-输入:{Text Description}输出:。
输出:<bom>{SELFIES}<eom>

G 个案研究

在本节中，我们展示了分子标题和基于文本的分子生成任务中不同模型的几个示例输出。数字 5 显示了分子标题任务的案例。在实施例(1)中，BioT5 的描述最符合基本事实，成功地定位了取代基和“吡啶和芳基硫醇的成员”的位置。在例(2)中，MolT5 错误地描述了分子中含有硼，而 BioT5 的描述更准确。在示例(3)中，当 MolT5 生成重复输出时，BioT5 和 T5 生成语义连贯的输出，并且 BioT5 的输出与基本事实更好地匹配。对于示例(4)中的复杂分子，BioT5 的输出更加全面和准确。值得注意的是，只有 BioT5 描述了这种分子

[4] https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md#t511

作为SARS冠状病毒主要抑制剂的蛋白酶原，可能来源于我们对蛋白质知识的整合。数字6显示了基于文本的分子生成任务的案例。从这些案例中，我们有几个发现:(i) BioT5更有可能产生与地面真相完全匹配的分子。(ii)通过使用自拍，BioT5不会产生无效分子，特别是对于实施例(3)和(4)中所示的复杂和更长的分子。(iii)有些分子实际上是短蛋白质。实施例(3)显示了一种33元多肽分子，其由33个氨基酸残基按顺序连接而成。因此，蛋白质和分子之间的界限可能并不总是清晰的，利用两者的信息可以带来互惠互利。

输入 T5 MolT5 Ours 地面真相

1 该分子是被硫烷基取代的硫醇	位置 4。它是一种代谢物。它是硫醇，也是苯的一员。它来源于硫醇的氢化物。	该分子是由硫氰酸与苯的缩甲醛反应生成的单硫代氨基甲酸酯。它是硫代氨基甲酸和单硫代氨基甲酸酯的一员。	该分子是在 2 位被硫烷基取代的吡啶。它具有腐蚀抑制剂和过敏原的作用。它是吡啶和芳基硫醇的一员。	该分子是在 C-2 位被硫烷基取代的吡啶。它作为荧光猝灭剂和过敏原发挥作用。它是吡啶和芳基硫醇的一员。
Li^+ Br^-	该分子是金属卤化物、金属阳离子和锂分子实体。它有渗透物和阻燃剂的作用。	该分子是金属四硼酸盐、金属离子和一价无机阴离子。它是二溴锂的共轭酸。	该分子是具有 Li(+)抗衡离子的金属溴化物盐。它有肥料的作用。它是无机溴化物盐和锂盐。	这种分子是锂盐，其中的抗衡离子是溴化物。无水盐形成类似于食盐的立方体晶体。它是溴化物盐和锂盐。
四	该分子是氨的季铵盐，其中 2-和 3-位的氢被甲基取代。它是一种心脏毒剂。它是季铵盐和季铵盐。它含有四甲基铵。 该分子是由 L-天冬氨酸、L-苯丙氨酸和两个 L-脯氨酸单元通过肽键连接而成的四肽。它是一种代谢物。它来源于一种 L-天冬氨酸、一种 L-苯丙氨酸和一种 L-脯氨酸。	该分子是碘-碘衍生物 该分子是一种五肽，由 L-可待因酰、L-苯丙氨酸和 L-(2-萘基)乙酰胺残基依次连接而成。它是一种代谢物。它是一种多肽，是萘和五肽的一员。	该分子是季铵盐，即乙基三甲基铵的一碘化物盐。它是季铵盐、有机碘化物盐和季铵盐。它含有乙基三甲铵。 该分子是由 1H-吲哚-2-羧酸的羧基与(2S，3R)-环己基-L-丙氨酸的伯氨基的形式缩合产生的仲甲酰胺。它是 SARS 冠状病毒主要蛋白酶的抑制剂，并抑制细胞培养中的新型冠状病毒复制(EC50 = 0.72 muM)。它扮演着欧共体的角色 3.4.22.69(SARS 冠状病毒主要蛋白酶)抑制剂和抗冠状病毒剂。它是一种仲甲酰胺、吡咯烷-2-酮、寡肽、吲哚甲酰胺和 L-丙氨酸衍生物的成员。	该分子是季铵盐，其基本单元包括乙基三甲基铵阳离子和碘阴离子。它是季铵盐和碘化物盐。 该分子是由 1H-吲哚-2-羧酸的羧基与 3-环己基-N-(2S)-1-氧代-3-[(3S)-2-氧代吡咯烷-3-基]丙-2-基-L-丙氨酸酰胺的伯氨基的形式缩合产生的仲甲酰胺。它是 SARS 冠状病毒主要蛋白酶的抑制剂，并抑制细胞培养中的新型冠状病毒复制(EC50 = 0.53 muM)。作为 3.4.22.69(SARS 冠状病毒主蛋白酶)抑制剂和抗冠状病毒剂发挥作用。它是一种吲哚甲酰胺、吡咯烷-2-酮的成员、一种醛、一种次级甲酰胺和一种寡肽。

图 5:分子标题案例。

输入 T5 MolT5 Ours 地面真相

1 该分子是 3’，5’-环嘌呤核苷酸，即 3’，5’-环 AMP 在腺嘌呤环的 8 位带有一个额外的溴取代基。一种环腺苷酸依赖性蛋白激酶的激活剂，但对环腺苷酸磷酸二酯酶的降解有抗性。作为蛋白激酶激动剂和抗抑郁剂发挥作用。它是一种 3’，5’-环嘌呤核苷酸、一种有机溴化合物和一种腺苷核糖核苷酸。它来源于一个 3 ’，5 ’-环 AMP。				
2 该分子是 5-甲基嘧啶-2，4-氨基嘧啶二胺，其中一个 甲基的氢被 2-环丙基- 7，8-二甲氧基-2H-色烯-5-基取代。它是一种氨基嘧啶，是色烯和环丙烷的成员。	无效的			
3 该分子是由组氨酸、甘氨酸， Asp、Gly、Ser、Phe、Ser、Asp、Glu、Met、Asn、Thr、Ile、Leu、Asp、Asn、Leu、Ala、Ala、Arg、Asp、Phe、Ile、Asn、Trp、Leu、Ile、Gln、Thr、Lys、Ile、Thr 和 Asp 残基依次连接。一种用于治疗短肠综合征的胰高血糖素样肽-2 受体激动剂。作为胰高血糖素样肽-2 受体激动剂、代谢物、抗氧化剂和保护剂发挥作用。		无效的		
4 该分子是 Cy5 染料和有机高氯酸盐。它有一个 作为荧光染料的作用。它包含一个 diIC18(5)(1+)。	无效的	无效的		

图 6:基于文本的分子生成案例。