# BioT5+: Towards Generalized Biological Understanding with IUPAC Integration and Multi-task Tuning

**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]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Microsoft Research [3]Huazhong University of Science and Technology
[4]Zhejiang University [5]University of Science and Technology of China
[6]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education
{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

## Abstract

Recent research trends in computational biology have increasingly focused on integrating text and bio-entity modeling, especially in the context of molecules and proteins. However, previous efforts like BioT5 faced challenges in generalizing across diverse tasks and lacked a nuanced understanding of molecular structures, particularly in their textual representations (e.g., IUPAC). This paper introduces BioT5+, an extension of the BioT5 framework, tailored to enhance biological research and drug discovery. BioT5+ incorporates several novel features: integration of IUPAC names for molecular understanding, inclusion of extensive bio-text and molecule data from sources like bioRxiv and PubChem, the multi-task instruction tuning for generality across tasks, and a numerical tokenization technique for improved processing of numerical data. These enhancements allow BioT5+ to bridge the gap between molecular representations and their textual descriptions, providing a more holistic understanding of biological entities, and largely improving the grounded reasoning of bio-text and bio-sequences. The model is pre-trained and fine-tuned with a large number of experiments, including *3 types of problems (classification, regression, generation), 15 kinds of tasks, and 21 total benchmark datasets*, demonstrating the remarkable performance and state-of-the-art results in most cases. BioT5+ stands out for its ability to capture intricate relationships in biological data, thereby contributing significantly to bioinformatics and computational biology. Our code is available at https://github.com/QizhiPei/BioT5.

* Corresponding authors: Lijun Wu (apeterswu@gmail.com) and Rui Yan (ruiyan@ruc.edu.cn)

## 1 Introduction

Molecules and proteins are two crucial bio-entities in drug discovery, forming the foundation of biological activities (Dara et al., 2022; AI4Science and Quantum, 2023). A molecule can be represented by its SMILES (Weininger, 1988; Weininger et al., 1989) or SELFIES (Krenn et al., 2020) sequence, and a protein can be described by a FASTA (Lipman and Pearson, 1985; Pearson and Lipman, 1988) sequence. With the advancement of Language Models (LMs), an increasing body of work focuses on understanding the molecules and proteins by modeling their biosequences (Chithrananda et al., 2020; Rives et al., 2021; Lin et al., 2022).

Notably, biological literature (Canese and Weis, 2013; White, 2020) is full of extensive information on molecules and proteins. When a biological entity is mentioned in such literature, its context is predominantly centered around a description of some characteristics of the entity. Consequently, there has been a growing body of work dedicated to the joint modeling of text and biological entities (Pei et al., 2024), such as Galactica (Taylor et al., 2022), MolXPT (Liu et al., 2023c), BioT5 (Pei et al., 2023) and BioMedGPT (Luo et al., 2023c), which are all scientific models trained on text, molecule and protein sequences. Despite their achievements, substantial opportunities for enhancement still remain: (1) Prior works neglect the importance of modeling the textual name of molecules, such as International Union of Pure and Applied Chemistry (IUPAC), which provides a standard and systematic naming method for ensuring uniformity and clarity across the scientific community. Different from SMILES and



*Findings of the Association for Computational Linguistics: ACL 2024*, pages 1216–1240
August 11-16, 2024 ©2024 Association for Computational Linguistics

SELFIES, IUPAC bears a closer resemblance to natural language that is evident in its widespread adoption within scientific literature (Klinger et al., 2008). (2) Previous models were predominantly specialist models, necessitating the training of a separate model for each downstream task, thereby lacking in generality and increasing the training and developing cost (Liu et al., 2023c; Pei et al., 2023). (3) Most of the previous models based on T5 (Raffel et al., 2020) and GPT (Brown et al., 2020) architectures only focus on the classification tasks since they do not implement specialized tokenization for numerical data, which results in their suboptimal adaptation to regression tasks.

To address the above challenges, in this paper, we introduce BioT5+, an advanced iteration of the BioT5 framework (Pei et al., 2023), designed to augment biological research and drug discovery with enriched data integration, multi-task capabilities, and the ability to solve regression tasks. Shortly speaking, BioT5+ incorporates following significant enhancements:

(1) *Enhanced Molecule Understanding*: By integrating IUPAC name into BioT5+ framework, the model can achieve a deeper comprehension of molecular structures. This integration allows BioT5+ to interpret chemical names as they commonly appear in scientific literature, bridging the gap between formal molecular representations (such as SELFIES) and their textual descriptions. Consequently, this enhances the understanding of molecules and facilitates more accurate predictions and analyses of molecular properties and activities.

(2) *Expanded Bio-text and Molecule Data*: Compared to BioT5, BioT5+ includes an extensive corpus of bio-text data from sources like bioRxiv (Sever et al., 2019) and PubMed (Canese and Weis, 2013; White, 2020), alongside high-quality molecular data from PubChem (Kim et al., 2019). This expansion not only broadens the knowledge base of the model but also enriches the contextual understanding of biological entities.

(3) *Multi-task Instruction Tuning*: BioT5+ employs multi-task instruction tuning strategy for downstream tasks rather than the separate specialized model training for each task. By leveraging a unified and multi-task training framework, BioT5+ can seamlessly integrate knowledge from diverse tasks, enhancing its predictive power and generalization capabilities across different biological and chemical domains.

(4) *Advanced Numerical Tokenization*: To overcome the limitations of the numerical representations, BioT5+ integrates an advanced character-based numerical tokenization strategy, drawing inspiration from the Llama (Touvron et al., 2023a) model. This technique allows for a more nuanced and consistent representation of numerical values.

With our designed pre-training and multi-task instruction tuning, the effectiveness of BioT5+ is verified on *3 types of problems (classification, generation, and regression), 15 different tasks, and 21 benchmark datasets*, including molecule property prediction, retrosynthesis, molecule description generation, drug-target interaction, and so on. BioT5+ has shown highly competitive results, achieving state-of-the-art performance in most of the tasks. This robust performance underscores the enhanced capability of BioT5+ to capture and analyze the intricate relationships and properties inherent in biological data, marking a significant step forward in computational biology.

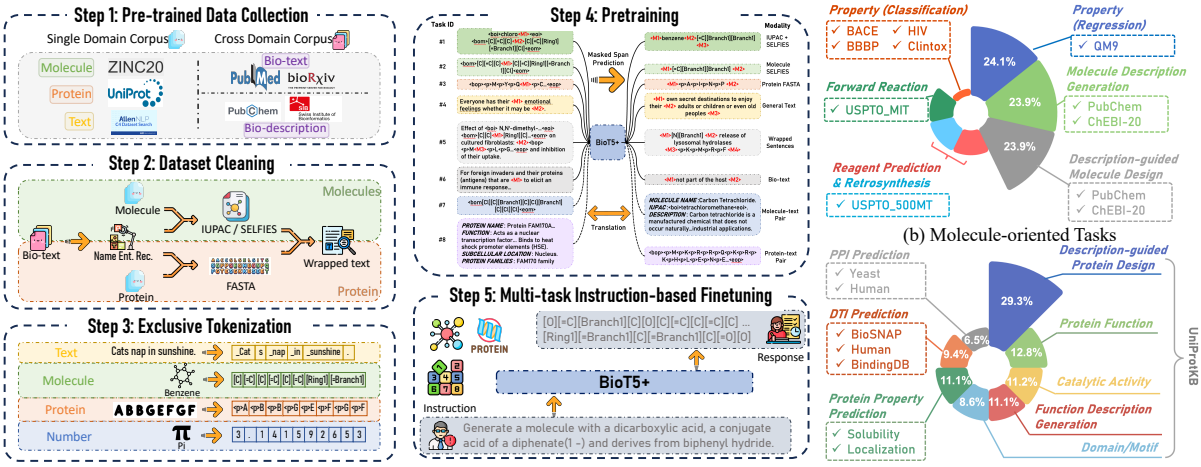
## 2 Related Work

### 2.1 Biological Cross-modal Models

Recent advancements in LLMs have led to an increased focus on jointly modeling molecules, proteins, and text, aiming to enhance the understanding of bio-entities through text.

**Molecule-Text.** MolT5 (Edwards et al., 2022) is jointly trained on general text and molecule SMILES using T5 (Raffel et al., 2020) masked span prediction objective. MoMu (Su et al., 2022) employs contrastive learning on molecular graphs and related text, and MolFM (Luo et al., 2023b) further incorporates knowledge graph embedding for molecule representation. MolXPT (Liu et al., 2023c) is jointly trained on molecule SMILES and wrapped text using GPT (Brown et al., 2020) framework. MolCA (Liu et al., 2023d) enhances LMs by integrating 2D molecular graph perception through a cross-modal projector and uni-modal adapter. GIT-Mol (Liu et al., 2023a) is a multi-modal LLM that synergizes graphs, images, SMILES, and molecule captions. Text+Chem T5 (Christofidellis et al., 2023) is a multi-domain, multi-task language model capable of concurrently processing molecules and natural language.

**Protein-Text.** Several notable works focus on jointly modeling proteins and text. ProteinDT (Liu et al., 2023b) presents a text-guided protein design framework. BioTranslator (Xu et al., 2023b) is a cross-modal translation system, which can annotate




(a) Overview of BioT5+ (b) Molecule-oriented Tasks (c) Protein-oriented Tasks

Figure 1: (a): The overview of BioT5+ framework. (b) (c): the composition of BioT5+ downstream tasks, which is divided into two categories: (b) molecule-oriented tasks and (c) protein-oriented tasks. The names of the tasks, along with their instruction datasets and respective percentages, are annotated near each segment of the accompanying pie charts.

various biological instances using textual descriptions. Prot2Text (Abdine et al., 2023) combines GNN and LLM in an encoder-decoder framework to generate protein functions in a free-text style.

In addition to the models mentioned above, there are other models trained in a more diverse range of modalities: DeepEIK (Luo et al., 2023a) is a multi-modal model which integrates features from multi-modal inputs including drugs, proteins, and text. BioT5 (Pei et al., 2023) is a T5-based (Raffel et al., 2020) model that undergoes joint training on text, molecule SELFIES, and protein FASTA sequences, effectively bridging the gap between textual and biological data.

Despite the successes of these models, their focus on single-task training limits their versatility and hinders the development of a more generalized and adaptable approach in computational biology.

### 2.2 Instruction Tuning for Biological Tasks

Instruction tuning is a popular technique applied to pre-trained LLMs where they are trained with specialized instruction datasets, thus equipping LLMs with the ability to understand task-specific instructions. Recently, there has been a growing interest in exploring instruction tuning for various biological tasks. Notable among these efforts is the development of Mol-Instructions (Fang et al., 2023), a comprehensive instruction dataset specifically designed for the biological domain, which includes molecule-oriented instructions, protein-oriented instructions, and biomolecular text instructions. InstructMol (Cao et al., 2023), a multi-modal LLM, employs instruction tuning to align molecular graphs, molecule SELFIES, and natural language. Different from these approaches, our BioT5+ is specifically pre-trained for the biological domain. Further instruction tuning within biological instructions enables BioT5+ to not only understand bio-entities but also generalize across various biological tasks.

## 3 BioT5+ Framework

This section introduces BioT5+ framework and an overview is shown in Figure 1. The tasks involved in pre-training are presented in Figure 3.

### 3.1 Intuition for IUPAC Integration

IUPAC naming system provides a standardized set of rules for naming chemical compounds, which allows for the precise description of molecular structures and their components (functional groups, chains, and rings), making it a cornerstone in chemical nomenclature. Typically, an IUPAC name is constructed from the names of individual molecules' constituent parts, reflecting their structure. This includes prefixes, infixes, and suffixes that indicate various chemical groups and structural features, providing a comprehensive description of the molecule. For instance, the IUPAC name for Aspirin is "2-acetyloxybenzoic acid". Here, "2-acetyloxy" refers to an acetoxy group attached to the second carbon of a benzene ring, and "benzoic acid" indicates a benzene ring with a carboxylic acid group. The resemblance of IUPAC names to natural language, coupled with their prevalent use in scientific literature, makes them an ideal candidate for model pre-training. By pre-training the model on literature that includes IUPAC names, BioT5+ can establish a nuanced understanding of the relationship between molecules and various textual descriptions of their chemical properties.

### 3.2 Pre-training Corpus

As an extension of BioT5, the majority of the pre-training corpus for BioT5+ is identical to BioT5,

hence we will briefly mention the common elements while focusing primarily on the novel aspects introduced in BioT5+.

The pre-training corpus consists of 4 classes: (1) *Single-modal data*, including molecule SELFIES with IUPAC name from PubChem (Kim et al., 2019), molecule SELFIES from ZINC20 (Irwin et al., 2020), protein FASTA from Uniref50 (Suzek et al., 2007), and general text from "Colossal Clean Crawled Corpus" (C4) (Raffel et al., 2020). For the molecule from PubChem, we concatenate the IUPAC name and SELFIES for pre-training as shown in Figure 3. (2) *Wrapped text*, where the molecule or gene/protein names are suffixed with corresponding sequence representation. We employ BERN2 (Sung et al., 2022), a neural-based Named Entity Recognition (NER) system in the biological domain, to detect and classify occurrences of molecules and proteins within the abstracts of PubMed (White, 2020) and bioRxiv (Sever et al., 2019). For the molecule name, we first standardize the name to its IUPAC name and then append the corresponding SELFIES. For the gene/protein name, we will directly append its FASTA sequence. For the generation of high-quality wrapped text, we also analyze the confidence score distribution predicted by BERN2 (Sung et al., 2022). Only those entities with higher confidence scores were retained to ensure the accuracy and relevance of the appended sequence data. Further detailed descriptions of this process are provided in Appendix Section C. (3) *Bio-text*, including PubMed (White, 2020) Central full text articles, and bio-texts from PubMed (White, 2020) abstracts and bioRxiv (Sever et al., 2019) abstracts that do not yield identifiable named entities in (2). (4) *Molecule-description pairs* and *protein-description pairs*. The molecule-text data is collected from PubChem (Kim et al., 2019) and we also add the IUPAC name to the text description. All molecules and proteins that exist in the downstream Mol-Instructions dataset (Fang et al., 2023) and ChEBI-20 (Edwards et al., 2022) are excluded to prevent data leakage. The protein-text data is the same as BioT5 (Pei et al., 2023).

**Remarkable Difference.** The primary distinctions between BioT5+ and BioT5 are as follows: (1) BioT5+ integrates IUPAC in the molecular pre-training data, encompassing IUPAC names combined with SELFIES, wrapped text, and molecule-text translation data. More details are in Appendix Section D. (2) BioT5+ incorporates a broader spectrum of high-quality data, including IUPAC names and SELFIES from PubChem, as well as comprehensive articles from bioRxiv and PubMed Central.

### 3.3 Tokenization

BioT5 (Pei et al., 2023) has already demonstrated the advantages of employing separate tokenization and embedding techniques. BioT5+ inherits this advantage to apply specialized tokenization and vocabulary specifically for bio-entities. This method explicitly differentiates between the biological semantic space and the textual semantic space. For molecule SELFIES, each chemically meaningful atom group, naturally distinguished from textual vocabulary due to its bracketed format like [C], is tokenized as an individual token using the inherent token set defined by SELFIES. For protein FASTA sequences, to ensure a clear modal distinction, each amino acid is tokenized into a separate token with the prefix <p>, differentiating them from standard upper-case English letters.

Concurrently, the tokenization of numerical data is worth dedicated consideration and design. The direct application of T5 (Raffel et al., 2020) dictionary derived from nature language using SentencePiece (Kudo and Richardson, 2018) for numerical tokenization can lead to inconsistencies (Liu and Low, 2023). For instance, the number 1024 might be tokenized into "10" and "24", while "2048" could be split into "2", "0", and "48". This irregular segmentation poses a challenge for the model in consistently mapping embeddings to numbers, especially when the number of digits they represent varies. In contrast, models like Llama (Touvron et al., 2023a,b) and ChatGLM (et.al., 2023) adopt a character-based approach to numerical tokenization, where each digit is tokenized as an individual token. This method has been demonstrated to yield superior results in various arithmetic tasks (Liu and Low, 2023; Nogueira et al., 2021). Accordingly, in BioT5+ we also implement this character-based approach for numerical tokenization without modifying the original dictionary. The efficacy of this method over the original T5 (Raffel et al., 2020) and BioT5 (Pei et al., 2023) to numerical tokenization are shown in Section 4.3, providing empirical evidence of its superior performance in handling numerical data.

### 3.4 Model and Training

**Model architecture.** BioT5+ adopts the same architecture as the BioT5 (Pei et al., 2023), which

follows the T5-v1.1-base[1] configuration with vocab size 35,076 and 252M parameters.

**Pre-training.** Based on the pre-training corpus described in Section 3.2, the pre-training for BioT5+ is conducted in a multi-task way with eight tasks that fall into 4 categories: (1) *Modality-Specific T5 Objectives*: This category involves the application of the T5 objective (masked span prediction) to each modality in isolation, including molecule SELFIES with IUPAC name (Task #1), molecule SELFIES (Task #2), protein FASTA sequences (Task #3), and general textual content (Task #4). (2) *T5 Objectives on Wrapped Text*: Applying the T5 objective to "wrapped" text extracted from scientific corpora (Task #5). (3) *T5 Objectives on Biotext*: Applying the T5 objective to text in biological domain (Task #6). (4) *Bidirectional Translation Tasks*: This involves the bidirectional translation between molecule SELFIES-text pairs (Task #7) and protein FASTA-text pairs (Task #8). Through these strategically structured pre-training tasks, BioT5+ is adept at learning the intricate relationships and characteristics of bio-entities as represented in textual information.

**Multi-task Instruction-based Fine-tuning.** After the comprehensive pre-training phase, BioT5+ undergoes multi-task instruction-based fine-tuning. Unlike BioT5 where each downstream task has a specialized fine-tuned model, we follow Fang et al., 2023 and Cao et al., 2023 to categorize downstream tasks and conduct multi-task instruction tuning, which not only saves the repeated tuning cost but also eases the model deployment for the evaluation of multiple tasks. The relevant groupings and information about benchmark tasks and datasets are illustrated in Figure 1, which is simply split by the domains, e.g., molecule-oriented or protein-oriented tasks. This methodology serves a dual purpose: firstly, it bridges the gap between the pre-training and fine-tuning phases, ensuring a smoother transition and integration of learned capabilities. Secondly, it activates and harnesses the general capabilities of BioT5+ across various tasks, demonstrating its versatility and adaptability in handling diverse biological problems.

## 4 Experiments and Results

As shown in Figure 1, BioT5+ is extensively evaluated across 21 well-established downstream benchmark datasets, which can be classified into 7 molecule-oriented tasks and 8 protein-oriented tasks with 3 types of problems: classification, regression, and generation. Following Fang et al., 2023, we categorize downstream tasks into different groups for multi-task instruction tuning in the same way, and details about the downstream datasets and baselines are in Appendix Section G.

Table 1: Performance (AUROC) comparison on molecule property prediction tasks (classification) on MoleculeNet (Wu et al., 2018) benchmark (**Best**, Second Best). * means LoRA (Hu et al., 2022) tuning.

| METHOD<br># MOLECULES | BACE ↑<br>1513 | BBBP ↑<br>2039 | HIV ↑<br>41127 | Clintox ↑<br>1478 |
|---|---|---|---|---|
| *Single-task Specialist Models* | | | | |
| GraphCL | 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 |
| MoMu | 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** |
| *LLM-based Generalist Models* | | | | |
| Galactica-6.7B | 58.4 | 53.5 | 72.2 | 78.4 |
| Galactica-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-shot) | 49.2 | 52.7 | 50.5 | - |
| Vicuna-v1.3-7B* | 68.3 | 60.1 | 58.1 | - |
| Llama-2-7B-chat* | 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** |

### 4.1 Molecule-oriented Tasks

The molecule-oriented tasks cover different topics. As we incorporate IUPAC name for molecule in the pre-training, we also use IUPAC in some molecule-oriented tasks, such as molecule property prediction and molecule description generation. More details are in the following sections and Appendix.

#### 4.1.1 Molecule Property Prediction

Molecule property prediction is a crucial task in bioinformatics, focusing on the determination of specific properties exhibited by a given molecule. Following Cao et al., 2023, we explore the ability of BioT5+ on MoleculeNet (Wu et al., 2018) benchmark. For classification tasks, we focus on 4 benchmark datasets: BACE, BBBP, HIV, and Clintox. Each sample includes an instruction detailing the property to be predicted and the molecule SELFIES with IUPAC name, with models required to generate a simple "yes" or "no" prediction. For regression tasks, we focus on 3 regression benchmarks from QM9 dataset, which aims to predict quantum mechanical properties of molecules, based on the

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

Table 2: Performance comparison on chemical reaction-related tasks (**Best**, Second Best). $*$ means LoRA tuning.

| MODEL | EXACT↑ | BLEU↑ | LEVENSHTEIN↓ | RDK FTS↑ | MACCS FTS↑ | MORGAN FTS↑ | VALIDITY↑ |
|---|---|---|---|---|---|---|---|
| *Reagent Prediction* | | | | | | | |
| Llama-7B | 0.000 | 0.003 | 28.040 | 0.037 | 0.001 | 0.001 | 0.001 |
| Galactica-6.7B | 0.000 | 0.141 | 30.760 | 0.036 | 0.127 | 0.051 | 0.995 |
| Text+Chem T5-223M | 0.000 | 0.225 | 49.323 | 0.039 | 0.186 | 0.052 | 0.313 |
| Mol-Instructions-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 |
| *Forward Reaction Prediction* | | | | | | | |
| Llama-7B | 0.000 | 0.020 | 42.002 | 0.001 | 0.002 | 0.001 | 0.039 |
| Galactica-6.7B | 0.000 | 0.468 | 35.021 | 0.156 | 0.257 | 0.097 | 0.946 |
| Text+Chem T5-223M | 0.239 | 0.782 | 20.413 | 0.705 | 0.789 | 0.652 | 0.762 |
| Mol-Instructions-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 |
| *Retrosynthesis* | | | | | | | |
| Llama-7B | 0.000 | 0.036 | 46.844 | 0.018 | 0.029 | 0.017 | 0.010 |
| Galactica-6.7B | 0.000 | 0.452 | 34.940 | 0.167 | 0.274 | 0.134 | 0.986 |
| Text+Chem T5-223M | 0.141 | 0.765 | 24.043 | 0.685 | 0.765 | 0.585 | 0.698 |
| Mol-Instructions-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 |

Table 3: Performance (MAE) comparison on molecule property prediction tasks (regression) on QM9 dataset from MoleculeNet (Wu et al., 2018) benchmark (**Best**, Second Best). $\Delta\epsilon$ means HOMO-LUMO gap.

| METHOD | HOMO ↓ | LUMO ↓ | $\Delta\epsilon$ ↓ | AVG ↓ |
|---|---|---|---|---|
| *LLM-based Generalist Models* | | | | |
| Llama2-7B (5-shot ICL) | 0.7367 | 0.8641 | 0.5152 | 0.7510 |
| Vicuna-13B (5-shot ICL) | 0.7135 | 3.6807 | 1.5407 | 1.9783 |
| Mol-Instructions-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** |

molecule SELFIES with IUPAC name, including HUMO, LUMO, and the HUMO-LUMO gap.

**Results.** The results for classification and regression tasks are shown in Table 1 and Table 3 respectively. BioT5+ demonstrates superior performance over other generalist model baselines. Notably, for classification tasks, BioT5+ surpasses models like Galactica (Taylor et al., 2022), which is extensively trained on a vast corpus of scientific literature. Similarly, InstructMol (Cao et al., 2023), despite its inclusion of 2D graph information and LLMs, BioT5+ outperforms on both classification and regression tasks. This enhanced performance can be attributed to the integration of IUPAC names, wrapped text, bio-text, and molecule-text pairs in BioT5+ pre-training. The presence of molecule property descriptions in the context of these diverse corpora allows the model to acquire a comprehensive understanding of molecular properties. However, when compared to single-task specialist models, BioT5+ showed some gaps. This discrepancy is understandable and can be attributed partly to the ease of tuning inherent in single-task models and partly to the fact that some baselines incorporated additional molecular information, such as 2D and 3D structures.

#### 4.1.2 Chemical Reaction-related Tasks

In computational chemistry, tasks related to chemical reactions are of vital importance as they can speed up development processes. Following Cao et al., 2023, we focus on 3 such tasks: reagent prediction, forward reaction prediction, and retrosynthesis.

**Results.** The main results are presented in Table 2 and full results are in Table 10. While LLMs have been exposed to some molecular data during pre-training, their direct zero-shot testing on chemical reaction-related tasks demonstrated extremely poor performance. Mol-Instructions (Fang et al., 2023) conducts multi-task instruction tuning based on Llama (Touvron et al., 2023a) with molecule-oriented tasks. InstructMol (Cao et al., 2023) introduces a molecule graph encoder to encode 2D molecular graph information for Vicuna (Chiang et al., 2023). Our BioT5+ follows the same training setting with Mol-Instructions (Fang et al., 2023) and shows superior performance across almost all metrics on chemical reaction-related tasks. This outcome demonstrates the effectiveness of joint pre-training on both molecular and textual data.

#### 4.1.3 Molecule Description Generation

The objective of molecule description generation is to generate a detailed and informative description for a given molecule. To be consistent with BioT5+ pre-training, the input here also consists of molecule SELFIES with IUPAC. Unlike molecule property prediction, which often focuses on specific attributes, molecule description generation involves interpreting and conveying a comprehensive narrative of the molecule. This narrative encompasses not only its molecular composition and properties but also its potential applications and roles,

Table 4: Performance comparison on molecule description generation task on ChEBI-20 (Edwards et al., 2022) dataset.

| MODEL | BLEU-2↑ | BLEU-4↑ | ROUGE-1↑ | ROUGE-2↑ | ROUGE-L↑ | METEOR↑ |
|---|---|---|---|---|---|---|
| *Single-task Specialist Models* | | | | | | |
| Transformer | 0.061 | 0.027 | 0.204 | 0.087 | 0.186 | 0.114 |
| T5-base | 0.511 | 0.423 | 0.607 | 0.451 | 0.550 | 0.539 |
| MoT5-base | 0.540 | 0.457 | 0.634 | 0.485 | 0.568 | 0.569 |
| MoMu (MolT5-base) | 0.549 | 0.462 | - | - | - | 0.576 |
| MolFM (MolT5-base) | 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-(graph+SMILES) | 0.352 | 0.263 | 0.575 | 0.485 | 0.560 | 0.430 |
| Text+Chem 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 |
| *Retrieval Based LLMs* | | | | | | |
| GPT-3.5-turbo (10-shot MolReGPT) | 0.565 | 0.482 | 0.623 | 0.450 | 0.543 | 0.585 |
| GPT-4-0314 (10-shot MolReGPT) | 0.607 | 0.525 | 0.634 | 0.476 | 0.562 | 0.610 |
| *LLM-based Generalist Models* | | | | | | |
| GPT-3.5-turbo (zero-shot) | 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 |
| Mol-Instructions-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** |

Table 5: Performance comparison on description-guided molecule design task on ChEBI-20 (Edwards et al., 2022) dataset. The ground truth Text2Mol (Edwards et al., 2021) score is 0.609.

| MODEL | BLEU↑ | EXACT↑ | LEVENSHTEIN↓ | MACCS FTS↑ | RDK FTS↑ | MORGAN FTS↑ | FCD↓ | TEXT2MOL↑ | VALIDITY↑ |
|---|---|---|---|---|---|---|---|---|---|
| *Single-task Specialist Models* | | | | | | | | | |
| Transformer | 0.499 | 0.000 | 57.660 | 0.480 | 0.320 | 0.217 | 11.32 | 0.277 | 0.906 |
| T5-base | 0.762 | 0.069 | 24.950 | 0.731 | 0.605 | 0.545 | 2.48 | 0.499 | 0.660 |
| MolT5-base | 0.769 | 0.081 | 24.458 | 0.721 | 0.588 | 0.529 | 2.18 | 0.496 | 0.772 |
| MoMu-base | 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** |
| *Retrieval-based LLMs* | | | | | | | | | |
| Llama2-7B (2-shot MolReGPT) | 0.693 | 0.022 | 36.77 | 0.808 | 0.717 | 0.609 | 4.90 | 0.149 | 0.761 |
| GPT-3.5-turbo (10-shot MolReGPT) | 0.790 | 0.139 | 24.91 | 0.847 | 0.708 | 0.624 | 0.57 | 0.571 | 0.887 |
| GPT-4-0314 (10-shot MolReGPT) | 0.857 | 0.280 | 17.14 | 0.903 | 0.805 | 0.739 | 0.41 | **0.593** | 0.899 |
| *LLM-based Generalist Models* | | | | | | | | | |
| 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-turbo (0-shot) | 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** |

Table 6: Performance (accuracy) comparison on PEER benchmark (**Best**, Second Best). * means linear probing.

| MODEL | SOLUBILITY | LOCALIZATION | YEAST | HUMAN |
|---|---|---|---|---|
| *Single-task Specialist Models* | | | | |
| DDE | 59.77 ± 1.21 | 77.43 ± 0.42 | 55.83 ± 3.13 | 62.77 ± 2.30 |
| Moran | 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 |
| Transformer | 70.12 ± 0.31 | 75.74 ± 0.74 | 54.12 ± 1.27 | 59.58 ± 2.09 |
| CNN | 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 |
| *Multi-task Generalist Models* | | | | |
| CNN | 70.63 ± 0.34 | 82.67 ± 0.72 | 54.50 ± 1.61 | 69.03 ± 2.68 |
| Transformer | 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** |

as derived from the integration of SELFIES representations and IUPAC names. We use the same evaluation metrics as Fang et al., 2023.

**Results.** As presented in Table 4, our BioT5+ outperforms all compared single-task specialist, retrieval-based LLMs, and multi-task generalist baselines. This superior performance can be attributed to the comprehensive learning during the pre-training of BioT5+. The model has effectively assimilated a multi-dimensional and rich textual description of molecules.

#### 4.1.4 Description-guided Molecule Design

Description-guided molecule design is essentially the inverse task of molecule description generation, which requires generating a molecule based on a provided textual description. In BioT5+ setting, we do not include IUPAC name in the textual description of the molecule to prevent the model from learning a simplistic mapping from IUPAC name to its SELFIES representation, thereby ensuring the model does not overlook the other descriptive elements provided in the text.

**Results.** Table 5 presents the results for description-guided molecule design task. Our BioT5+ surpasses all the compared baselines. This achievement underscores the efficacy of BioT5+ pre-training, where the model has acquired a profound understanding of molecular knowledge.

### 4.2 Protein-oriented Tasks

#### 4.2.1 Protein Description Generation

The task of protein description generation involves deriving relevant textual information from a given protein sequence. Following Fang et al., 2023, we mainly focus on 4 related generation tasks: protein function generation, catalytic activity generation,

Table 7: Ablation of IUPAC and additional data on molecule description generation task with single task setting. B-2 stands for BLEU-2 and R-1 is ROUGE-1.

| MODEL | B-2↑ | B-4↑ | R-1↑ | R-2↑ | R-L↑ | METEOR↑ |
|---|---|---|---|---|---|---|
| BioT5+(single task) | 0.671 | 0.597 | 0.715 | 0.590 | 0.655 | 0.687 |
| BioT5+(single task) wo IUPAC | 0.661 | 0.584 | 0.706 | 0.578 | 0.647 | 0.677 |
| BioT5+(single task) wo additional data | 0.666 | 0.591 | 0.711 | 0.586 | 0.651 | 0.681 |
| BioT5+ | 0.666 | 0.591 | 0.710 | 0.584 | 0.650 | 0.681 |

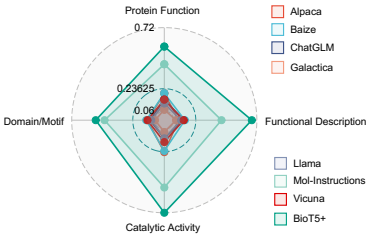



Figure 2: Performance (ROUGE-L) comparison on protein description generation tasks.

Table 8: Performance (AUROC) comparison on the 3 DTI datasets(**Best**, Second Best).

| METHOD | BIOSNAP | HUMAN | BINDINGDB |
|---|---|---|---|
| *Single-task Specialist Models* | | | |
| SVM | 0.862±0.007 | 0.940±0.006 | 0.939±0.001 |
| RF | 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 |
| DrugBAN | 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 |
| *Multi-task Generalist Models* | | | |
| BioT5+ | **0.939±0.001** | 0.987±0.001 | **0.964±0.001** |

Table 9: Ablation of default T5 tokenizer and character-based tokenizer (BioT5+) on QM9 dataset.

| METHOD | HOMO ↓ | LUMO ↓ | Δε ↓ | AVG ↓ |
|---|---|---|---|---|
| T5 default tokenizer | 0.0024 | 0.0026 | 0.0032 | 0.0027 |
| BioT5+ | 0.0022 | 0.0024 | 0.0028 | 0.0025 |

domain/motif generation, and functional description generation.

**Results.** As shown in the Figure 2, BioT5+ surpasses all compared baselines across 4 tasks. This result highlights the advanced ability of BioT5+ to interpret complex protein sequences into meaningful textual information, indicating that BioT5+ gains a comprehensive understanding of protein structures and functions through pre-training.

#### 4.2.2 Protein Property Prediction

Protein property prediction task involves predicting specific properties of proteins, such as solubility, structure, or function, based on their amino acid sequences. Following BioT5 (Pei et al., 2023), we focus on 2 protein property prediction tasks from PEER (Xu et al., 2022) benchmark, which is specifically designed for protein sequence understanding: (1) Solubility prediction: whether the input protein is soluble or not. (2) Localization prediction: either the input protein is "membrane-bound" or "soluble". Both of these tasks are binary classification tasks and the model needs to generate a "yes" or "no" prediction. The results are summarized together with the ones in the following Section 4.2.3.

#### 4.2.3 Protein-related Interaction Prediction

In drug discovery, the prediction of interactions between bio-entities is very important, from which Protein-Protein Interaction (PPI) and Drug-Target Interactions (DTI) are two key examples. These two tasks are essential for understanding biological processes and identifying potential therapeutic targets. To facilitate this, we follow Pei et al., 2023 to incorporate 2 PPI dataset from PEER (Xu et al., 2022) benchmark including Yeast and Human, and 3 DTI datasets including BioSNAP (Zitnik et al., 2018), Human (Liu et al., 2015; Chen et al., 2020), and BindingDB (Liu et al., 2007).

**Results.** As shown in Table 6, in the PEER benchmark, our BioT5+ demonstrates exceptional performance, surpassing other multi-task models in 3 out of the 4 tasks and achieving results comparable to single-task specialist models. Notably, in the Yeast PPI prediction task, BioT5+ exceeded the performance of all baseline models. This is particularly significant considering that the baseline ESM-1b (Rives et al., 2021) was specifically pre-trained on a vast array of protein sequences and possesses more than double the number of parameters compared to BioT5+. Furthermore, BioT5+ also showed superior performance in DTI tasks as in Table 8 (full results in Table 11), consistently outperforming other methods on the BioSNAP and BindingDB datasets. It is noteworthy that many baseline methods involved specialized designs for molecule and protein encoders. These results underscore the effectiveness of the joint pre-training of BioT5+ on bio-text, molecules, and proteins. This comprehensive understanding is evident in the ability of BioT5+ to accurately predict protein properties, interactions, and drug-target interactions, making it a valuable tool in the field of computational biology.

#### 4.2.4 Description-guided Protein Design

For description-guided protein design, the model needs to generate protein amino acid sequences based on specific design requirements, such as protein structures and functions. Due to the absence of a well-established benchmark for this task, we present in Appendix Table 20 a selection of test cases along with their corresponding sequence similarity scores to provide a direct

comparison between our model and existing models like Galactica (Taylor et al., 2022) and Mol-Instructions (Fang et al., 2023).

### 4.3 Ablation Studies

In this section, we conduct ablation studies to investigate the effectiveness of our design in BioT5+. Specifically, we focus on the following 3 scenarios: **(1) Do not incorporate IUPAC name of the molecule.** As shown in Table 7, removing the IUPAC name results in a noticeable performance drop in the molecule description generation task. This decline highlights the significant role that IUPAC name plays in tasks related to molecular understanding. **(2) Do not add PubMed Central and bioRxiv data in pre-training.** Results in Table 7 and Table 13 indicate that these two datasets play a crucial role in enhancing molecular understanding. The omission of them leads to a slight but noticeable decrease in performance on molecule description generation and description-guided molecule design tasks. **(3) Use T5 default tokenizer for numbers instead of character-based tokenizer.** The results in Table 9 demonstrate that the character-based approach for tokenizing numbers is more effective than the default T5 tokenizer on regression tasks. We also conduct an ablation study to further contrast single-task and multi-task tuning strategies in Appendix Section F.

## 5 Conclusions

BioT5+, as an advanced iteration of the BioT5 framework, represents a significant stride in computational biology and drug discovery. By integrating IUPAC names, expanding bio-text and molecular data sources, employing multi-task instruction tuning, and incorporating an advanced numerical tokenization technique, BioT5+ has successfully bridged the gap between molecular representations and their textual descriptions. The enhanced understanding of molecular structures and its ability to process complex biological data have been demonstrated across a wide range of tasks, with BioT5+ achieving state-of-the-art performance in most of them. This success highlights the potential of BioT5+ as a versatile and powerful tool in understanding and analyzing biological entities.

## 6 Limitations

Despite the significant advancements achieved by BioT5+, there remain certain limitations that need to be addressed in future work. Firstly, the model faces challenges in generalizing across various biological tasks, a problem that is distinct from common NLP settings. The intricate and unique nature of each biological task makes it difficult to develop a one-size-fits-all solution, highlighting the need for more specialized approaches within this domain. Secondly, the current scale of BioT5+ is somewhat limited and cannot comprehend and integrate information from other modalities, such as images, restricting its applicability in multi-modal biological data analysis. BioT5+ is not equipped to function as a universal chatbot or to answer queries spanning general domain questions outside the specific scope of biology. This constraint highlights the need for developing larger, more versatile models capable of handling a wider range of data types and answering a broader array of questions in both biological and general domains.

## 7 Ethical Considerations

While BioT5+ presents a significant advancement, its capabilities, particularly in generating molecules based on textual descriptions and predicting chemical reaction products, raise important ethical considerations. One of the concerns is the potential misuse of this technology to generate harmful or dangerous molecules, which could pose risks to public safety and environmental health. Moreover, the ability of BioT5+ to predict and generate novel molecules may also lead to issues surrounding intellectual property rights and patenting. The ease with which new compounds can be designed and synthesized using AI-driven methods could potentially disrupt traditional research practices and raise questions about the ownership of these discoveries.

## 8 Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC Grant No. 62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China. Qizhi Pei is supported by the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China.

## References


Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2023. Prot2text: Multimodal protein's function generation with GNNs and transformers. In *NeurIPS 2023 AI for Science Workshop*.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. 2023. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *CoRR*, abs/2311.16208.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. 2020. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, Ch Madhu Babu, and Mohamed Jawed Ahsan. 2022. Machine learning in drug discovery: A review. *Artif. Intell. Rev.*, 55(3):1947–1999.

Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. 2002. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 375–413. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 595–607. Association for Computational Linguistics.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.

Aohan Zeng et.al. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *CoRR*, abs/2306.08018.

Zhi-Ping Feng and Chun-Ting Zhang. 2000. Prediction of membrane protein types based on the hydrophobic index of amino acids. *Journal of protein chemistry*, 19:269–275.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kexin Huang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. MolTrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37:830 – 836.

John J Irwin, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. 2020. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, 60(12):6065–6073.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Roman Klinger, Corinna Kolárik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. 2008. Detection of IUPAC and iupac-like chemical names. In *Proceedings 16th International Conference on Intelligent Systems for Molecular Biology (ISMB), Toronto, Canada, July 19-23, 2008*, pages 268–276.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Greg Landrum. 2021. Rdkit: Open-source cheminformatics software. GitHub release.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, 15.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*.

David J Lipman and William R Pearson. 1985. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441.

Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. 2015. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229.

Pengfei Liu, Yiming Ren, and Zhixiang Ren. 2023a. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *CoRR*, abs/2308.06911.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training molecular graph representation with 3d geometry. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. 2023b. A text-guided protein design framework. *CoRR*, abs/2302.04611.

Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms GPT-4 on arithmetic tasks. *CoRR*, abs/2305.14201.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. 2007. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023c. Molxpt: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023,*

*Toronto, Canada, July 9-14, 2023*, pages 1606–1616. Association for Computational Linguistics.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023d. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yizhen Luo, Kui Huang, Massimo Hong, Kai Yang, Jiahuan Zhang, Yushuai Wu, and Zaiqin Nie. 2023a. Empowering ai drug discovery with explicit and implicit knowledge. *arXiv preprint arXiv:2305.01523*.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023b. Molfm: A multimodal molecular foundation model. *CoRR*, abs/2307.09484.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023c. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *CoRR*, abs/2308.09442.

Frederic P Miller, Agnes F Vandome, and John McBrewster. 2009. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance.

Thin Nguyen, Hang Le, T. Quinn, Tri Minh Nguyen, Thuc Duy Le, and Svetha Venkatesh. 2021. GraphDTA: Predicting drug-target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147.

Rodrigo Frassetto Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of the transformers with simple arithmetic tasks. *CoRR*, abs/2102.13019.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

William R Pearson and David J Lipman. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, and Rui Yan. 2024. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore. Association for Computational Linguistics.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.*, 58(9):1736–1741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2021. STOUT: SMILES to IUPAC names using neural machine translation. *J. Cheminformatics*, 13(1):34.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

Vijayakumar Saravanan and Namasivayam Gautham. 2015. Harnessing computational biology for exact linear b-cell epitope prediction: a novel amino acid composition-based feature descriptor. *Omics: a journal of integrative biology*, 19(10):648–658.

Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. Get your atoms in order - an open-source implementation of a novel and robust molecular canonicalization algorithm. *J. Chem. Inf. Model.*, 55(10):2111–2120.

Richard Sever, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, and John R Inglis. 2019. biorxiv: the preprint server for biology. *BioRxiv*, page 833400.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. 2007. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

David Weininger, Arthur Weininger, and Joseph L Weininger. 1989. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Canwen Xu, Daya Guo, Nan Duan, and Julian J. McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *CoRR*, abs/2304.01196.

Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B Altman, and Sheng Wang. 2023b. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14(1):738.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. 2022. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173.

Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. 2008. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. In *Proceedings 16th International Conference on Intelligent Systems for Molecular Biology (ISMB), Toronto, Canada, July 19-23, 2008*, pages 232–240.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. 2021. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *The*

*Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Marinka Zitnik, Rok Sosic, and Jure Leskovec. 2018. Biosnap datasets: Stanford biomedical network dataset collection. *Note: http://snap. stanford. edu/biodata Cited by*, 5(1).

## A Additional Results

For the molecule description generation task and description-guided molecule design task, we also compare the performance of BioT5+ with baseline methods on Mol-Instruction (Fang et al., 2023) test sets. The results are shown in Table 15 and Table 16. BioT5+ demonstrate superior performance in almost all metrics compared to baseline methods, which underscore BioT5+'s advanced capabilities in understanding complex molecular data.

## B Hyper-parameters

**Pre-training.** The pre-training process spans 300K steps and is executed on eight NVIDIA 80GB A100 GPUs with batch size 96 per GPU. To balance the data from different tasks during pre-training, we adopt a batch-level balancing strategy, where each batch evenly includes data from all eight different tasks, ensuring a more balanced and comprehensive pre-training process. For small datasets, such as molecule-text pairs and protein-text pairs, we employ a round-robin strategy to repeat their usage multiple times, compensating for their limited size. The dropout rate is maintained at 0.0 and the maximum input length during pre-training is set at 512. Optimization is performed using the AdamW (Loshchilov and Hutter, 2019) optimizer with Root Mean Square scaling. A cosine annealing learning rate scheduler is employed, with the base rate set at $1e-2$ and the minimum rate at $1e-5$.

**Multi-task Fine-tuning.** For multi-task fine-tuning, the dropout rate is searched in [0.0, 0.05, 0.1], and the learning rate is searched in [5e-5, 1e-4, 2e-4, 5e-4]. The total number of steps is 100K and warmup steps is 6% of total steps. The batch size is set to 768 for molecule-oriented tasks and 96 for protein-oriented tasks. The best hyper-parameters for molecule-oriented and protein-oriented task are shown in Table 12.

## C NER and Entity Linking

In general, our approach adheres to the same Named Entity Recognition (NER) and Entity Linking process as BioT5 (Pei et al., 2023) for bio-entity name mentions in biological text using BERN2 (Sung et al., 2022). However, we have implemented some modifications to the process: (1) Upon analyzing the confidence scores of the identified bio-entities, we observed a long-tailed distribution, with the majority of bio-entity confidence scores exceeding 0.9. Based on this empirical finding, we set a threshold of 0.9, retaining only those NER results that surpass this score. (2) With the introduction of IUPAC names into our workflow, we now assess whether a recognized molecular name is an IUPAC name. If it is, we exclusively append the SELFIES representation; if not, we append both the IUPAC name and the SELFIES. This dual approach ensures a more comprehensive and accurate representation of molecular entities in our analysis.

## D IUPAC Incorporation

In our downstream tasks of molecule property prediction and molecule description generation, it is effective to enrich molecules with their IUPAC name. One reason is that IUPAC names are more commonly found in bio-text. By explicitly incorporating IUPAC names in the molecular context during pre-training, the model more readily learns relevant molecular knowledge and establishes connections between the molecule and its contextual information. Additionally, IUPAC names inherently contain structural information about the molecule, such as functional groups and other structural components. This information allows the model to better understand the molecular structure, thus predicting molecule properties with higher accuracy and generating more accurate and detailed descriptions.

### D.1 Mapping Process

Initially, we normalize the SMILES sequences provided in the dataset and map them to their respective PubChem (Kim et al., 2019) CIDs. Subsequently, these CIDs are used to retrieve the corresponding IUPAC names. However, for some molecules, their SMILES sequences do not correspond to a PubChem CID. In such cases, we employ STOUT (Rajan et al., 2021), a highly accurate SMILES to IUPAC name translator utilizing transformers, to convert these SMILES sequences

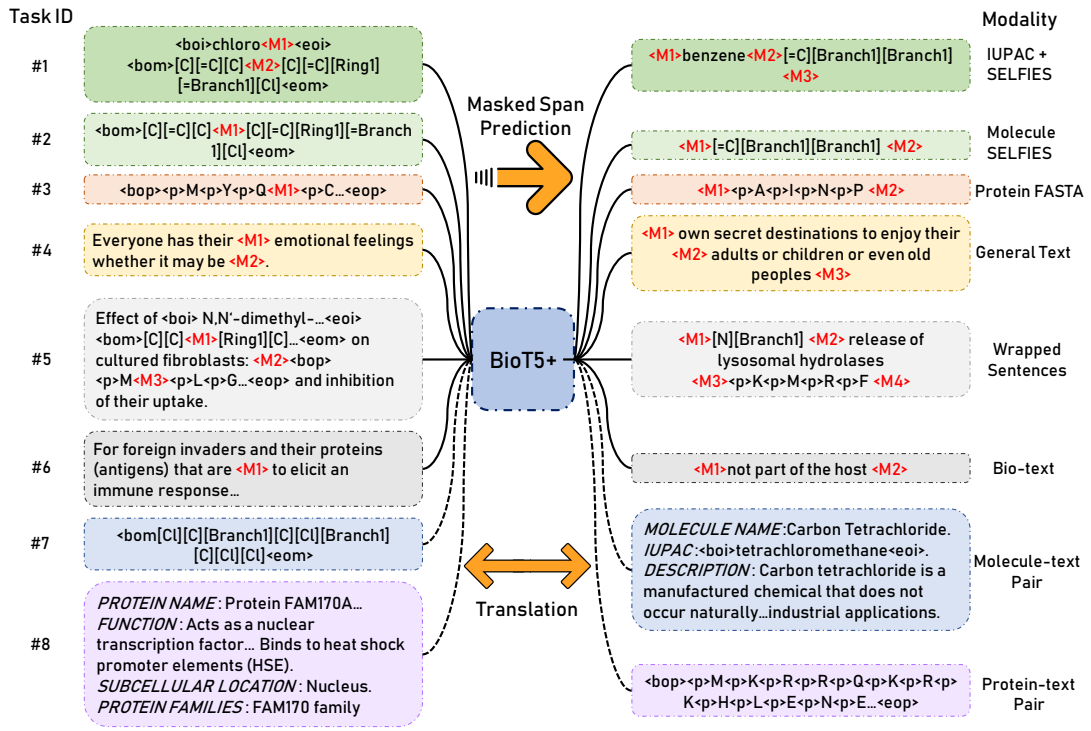



Figure 3: Overview of BioT5+ pre-training. The solid line refers to the masked span prediction task proposed by T5 (Raffel et al., 2020). Each consecutive span of masked tokens is substituted with a sentinel token, represented as <M1>, <M2>, and <M3>. We apply this pre-training task to molecule IUPAC + SELFIES (task #1), molecule SELFIES (task #2), protein FASTA (task #3), general text (task #4), wrapped text (task #5), and bio-text (task #6). The dashed line symbolizes the bidirectional translation between structured text description and biological sequences. (task #7 and #8).

into IUPAC names. This multi-step process ensures that each molecule is accurately equipped with its IUPAC name, facilitating more effective prediction and generation tasks.

## E Comparison with BioT5

Regarding the size of pre-training data and pre-processing techniques, BioT5+ introduces several enhancements over BioT5 (Pei et al., 2023):

- Addition of 28.8M PubChem molecular data, including molecule SELFIES along with their IUPAC names.
- Addition of 28.8M full articles from PubMed.
- Addition of 2.3M bioRxiv abstracts as wrapped biotext.
- In the molecule-text bidirectional translation task, IUPAC representations are included in the text description.
- For wrapped biotext pre-training, detected molecule names in BioT5 are directly replaced with corresponding SELFIES. In BioT5+, we first determine whether a molecule name is an IUPAC name; if so, it is appended with the corresponding SELFIES. If not, both the IUPAC name and SELFIES are appended. Besides, we further ensure data quality by only processing detected bioentities with a confidence score (from BERN2 (Sung et al., 2022)) above 0.9.
- The numbers that appear in the pre-training corpus are tokenized character by character in BioT5+'s tokenizer.

## F Additional Ablation Study

To contrast single-task and multi-task tuning strategies, we further fine-tune BioT5+ with a single-task setting on three MoleculeNet (Wu et al., 2018) classification tasks including BACE, BBBP, Clintox, and three QM9 (Fang et al., 2023) regression tasks including HOMO, LUMO, and HOMO-LUMO gap. The consolidated results are shown in Table 14 (we also summarize the results in Table 7 and Table 13 for generation tasks on ChEBI-20 (Edwards et al., 2022) dataset here), covering 3 types of

Table 10: Performance comparison on chemical reaction-related tasks (**Best**, Second Best). $*$ means LoRA tuning.

| MODEL | EXACT↑ | BLEU↑ | LEVENSHTEIN↓ | RDK FTS↑ | MACCS FTS↑ | MORGAN FTS↑ | VALIDITY↑ |
|---|---|---|---|---|---|---|---|
| *Reagent Prediction* | | | | | | | |
| Alpaca-7B | 0.000 | 0.026 | 29.037 | 0.029 | 0.016 | 0.001 | 0.186 |
| Baize-7B | 0.000 | 0.051 | 30.628 | 0.022 | 0.018 | 0.004 | 0.099 |
| ChatGLM-6B | 0.000 | 0.019 | 29.169 | 0.017 | 0.006 | 0.002 | 0.074 |
| Llama-7B | 0.000 | 0.003 | 28.040 | 0.037 | 0.001 | 0.001 | 0.001 |
| Vicuna-7B | 0.000 | 0.010 | 27.948 | 0.038 | 0.002 | 0.001 | 0.007 |
| Galactica-6.7B | 0.000 | 0.141 | 30.760 | 0.036 | 0.127 | 0.051 | 0.995 |
| Text+Chem T5-223M | 0.000 | 0.225 | 49.323 | 0.039 | 0.186 | 0.052 | 0.313 |
| Mol-Instructions-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 |
| *Forward Reaction Prediction* | | | | | | | |
| Alpaca-7B | 0.000 | 0.065 | 41.989 | 0.004 | 0.024 | 0.008 | 0.138 |
| Baize-7B | 0.000 | 0.044 | 41.500 | 0.004 | 0.025 | 0.009 | 0.097 |
| ChatGLM-6B | 0.000 | 0.183 | 40.008 | 0.050 | 0.100 | 0.044 | 0.108 |
| Llama-7B | 0.000 | 0.020 | 42.002 | 0.001 | 0.002 | 0.001 | 0.039 |
| Vicuna-7B | 0.000 | 0.057 | 41.690 | 0.007 | 0.016 | 0.006 | 0.059 |
| Galactica-6.7B | 0.000 | 0.468 | 35.021 | 0.156 | 0.257 | 0.097 | 0.946 |
| Text+Chem T5-223M | 0.239 | 0.782 | 20.413 | 0.705 | 0.789 | 0.652 | 0.762 |
| Mol-Instructions-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 |
| *Retrosynthesis* | | | | | | | |
| Alpaca-7B | 0.000 | 0.063 | 46.915 | 0.005 | 0.023 | 0.007 | 0.160 |
| Baize-7B | 0.000 | 0.095 | 44.714 | 0.025 | 0.050 | 0.023 | 0.112 |
| ChatGLM-6B | 0.000 | 0.117 | 48.365 | 0.056 | 0.075 | 0.043 | 0.046 |
| Llama-7B | 0.000 | 0.036 | 46.844 | 0.018 | 0.029 | 0.017 | 0.010 |
| Vicuna-7B | 0.000 | 0.057 | 46.877 | 0.025 | 0.030 | 0.021 | 0.017 |
| Galactica-6.7B | 0.000 | 0.452 | 34.940 | 0.167 | 0.274 | 0.134 | 0.986 |
| Text+Chem T5-223M | 0.141 | 0.765 | 24.043 | 0.685 | 0.765 | 0.585 | 0.698 |
| Mol-Instructions-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 |

Table 11: Performance comparison on the BindingDB, Human and BioSNAP datasets(**Best**, Second Best).

| | BIOSNAP | | | HUMAN | | BINDINGDB | | |
|---|---|---|---|---|---|---|---|---|
| METHOD | AUROC | AUPRC | ACCURACY | AUROC | AUPRC | AUROC | AUPRC | ACCURACY |
| *Single-task Specialist Models* | | | | | | | | |
| SVM | 0.862±0.007 | 0.864±0.004 | 0.777±0.011 | 0.940±0.006 | 0.920±0.009 | 0.939±0.001 | 0.928±0.002 | 0.825±0.004 |
| RF | 0.860±0.005 | 0.886±0.005 | 0.804±0.005 | 0.952±0.011 | 0.953±0.010 | 0.942±0.011 | 0.921±0.016 | 0.880±0.012 |
| DeepConv-DTI | 0.886±0.006 | 0.890±0.006 | 0.805±0.009 | 0.980±0.002 | 0.981±0.002 | 0.945±0.002 | 0.925±0.005 | 0.882±0.007 |
| 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 | 0.895±0.004 | 0.897±0.005 | 0.825±0.010 | 0.980±0.002 | 0.978±0.003 | 0.952±0.002 | 0.936±0.001 | 0.887±0.006 |
| DrugBAN | 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 | 0.904±0.004 |
| BioT5 | 0.937±0.001 | 0.937±0.004 | 0.874±0.001 | **0.989±0.001** | **0.985±0.002** | 0.963±0.001 | **0.952±0.001** | **0.907±0.003** |
| *Multi-task Generalist Models* | | | | | | | | |
| BioT5+ | **0.939±0.001** | **0.942±0.002** | **0.875±0.001** | 0.987±0.001 | **0.985±0.002** | **0.964±0.001** | **0.952±0.001** | 0.906±0.003 |

tasks: MoleculeNet for classification task, ChEBI-20 for generation task, and QM9 for regression task. Our findings indicate that: (1) Single-task fine-tuning yields different performance in different tasks. For the tasks such as BACE, BBBP, Clintox, and ChEBI-20, single-task tuned BioT5+ performs closely or slightly better than the multi-task-tuned version. (2) Multi-task fine-tuning can still be advantageous for tasks with inherent correlations, such as the prediction of HOMO, LUMO, and the HOMO-LUMO gap. This evidence points to the potential for cross-task generalization, even if it is not uniformly applicable across all tasks.

## G Fine-tuning Details

We adopt multi-task instruction tuning on molecule-oriented tasks and protein-oriented tasks. To facilitate a fair comparison with earlier studies, given the wider variety of categories in our fine-tuning dataset compared to the Mol-Instructions (Fang et al., 2023) dataset, we perform multi-task instruction tuning for both molecule-oriented and protein-oriented tasks, using both the Mol-Instruction dataset and an alternative dataset excluding Mol-

Table 12: Best hyper-parameters for multi-task instruction fine-tuning.

| Hyper-parameter | Molecule | | Protein | |
|---|---|---|---|---|
| | Mol-Instructions | Others | Mol-Instructions | Others |
| Dropout Rate | 0.05 | 0.1 | 0.05 | 0.05 |
| LR | 2e-4 | 5e-4 | 1e-4 | 1e-4 |
| Batch Size | 768 | 768 | 96 | 96 |
| Steps | 100,000 | 100,000 | 100,000 | 100,000 |

Table 13: Ablation of additional data on the description-guided molecule design task.

| Model | BLEU↑ | Exact↑ | Levenshtein↓ | MACCS FTS↑ | RDK FTS↑ | Morgan FTS↑ | FCD↓ | Text2Mol↑ | Validity↑ |
|---|---|---|---|---|---|---|---|---|---|
| BioT5+(single task) | 0.877 | 0.535 | 12.777 | 0.909 | 0.842 | 0.784 | 0.350 | 0.580 | 1.000 |
| BioT5+(single task) wo additional data | 0.875 | 0.516 | 12.840 | 0.904 | 0.833 | 0.777 | 0.358 | 0.579 | 1.000 |
| BioT5+ | 0.872 | 0.522 | 12.776 | 0.907 | 0.835 | 0.779 | 0.353 | 0.579 | 1.000 |

Instructions for each sub-domain. Taking into account the varying sizes of the datasets involved, we report the results across differing epoch spans to accommodate these discrepancies. All results are derived from 3 random runs.

## G.1 Molecule-oriented Tasks

### G.1.1 Molecule Property Prediction

**Classification** We focus on the following four datasets with scaffold splits setting:
(1) BACE dataset provides both qualitative binary labels and quantitative IC50 measurements for various inhibitors aimed at human beta-secretase 1 (BACE-1).
(2) BBBP (Blood-Brain Barrier Penetration) dataset, designed to assist in predicting and modeling permeability of the blood-brain barrier, consists of compounds classified by binary labels that denote their ability to penetrate the barrier.
(3) HIV dataset includes more than 40,000 compounds evaluated for their ability to inhibit HIV replication. They were initially categorized into Confirmed Inactive (CI), Confirmed Active (CA), and Confirmed Moderately Active (CM). Later, CA and CM categories were merged, simplifying the classification into a binary system of inactive (CI) versus active (CA and CM).
(4) Clintox dataset differentiates between FDA-approved drugs and those that failed clinical trials due to toxicity. It features two distinct classification tasks with known chemical structures: (i) determining whether they exhibited toxicity in clinical trials, and (ii) assessing their FDA approval status.

We compare our BioT5+ with (1) single-task specialist models including GraphCL (You et al., 2020), GraphMVP-C (Liu et al., 2022), MGSSL (Zhang et al., 2021), MolCLR (Wang et al., 2022), GEM (Fang et al., 2022), Uni-Mol (Zhou et al., 2023), KV-PLM (Zeng et al., 2022), MoMu (Su et al., 2022), MolFM (Luo et al., 2023b), MolXPT (Liu et al., 2023c), and BioT5 (Pei et al., 2023); (2) LLM-based generalist models including Galactica (Taylor et al., 2022), Vicuna (Chiang et al., 2023), Llama2 (Touvron et al., 2023b), and InstructMol (Cao et al., 2023). The baseline results are mainly derived from original papers, MolFM (Luo et al., 2023b), InstructMol (Cao et al., 2023). The evaluation metric is AUROC, and we follow the same AUROC calculation method with MolXPT (Liu et al., 2023c) and BioT5 (Pei et al., 2023) based on the logits of "yes" and "no" predictions.

**Regression** For the regression task, we focus on three tasks from the QM9 dataset, which encompasses over 134,000 stable organic molecules with no more than nine heavy atoms, characterized by their geometric, energetic, electronic, and thermodynamic properties. Following Mol-Instructions (Fang et al., 2023), our focus is on three specific subtasks within QM9: (1) "HOMO" for the highest occupied molecular orbital energy. (2) "LUMO" for the lowest unoccupied molecular orbital energy. (3) "GAP" denoting the energy difference between HOMO and LUMO. All of them are measured in Hartree units. We use the processed QM9 dataset in instruction format and corresponding splits from Mol-Instruction (Fang et al., 2023).

We compare our BioT5+ with LLM-based generalist models, including Llama2 (Touvron et al., 2023b), Vicuna (Chiang et al., 2023), Mol-Instructions (Fang et al., 2023), and InstructMol (Cao et al., 2023). The baseline results come from InstructMol (Cao et al., 2023). The evaluation

Table 14: Performance comparison of single-task and multi-task versions of BioT5+ across various datasets and tasks. *Mol2text* means molecule description generation, and *text2mol* means description-guided molecule design.

| DATASET/TASK | TASK TYPE | SINGLE-TASK | MULTI-TASK | METRIC |
|---|---|---|---|---|
| BACE | classification | 87.8 | 86.2 | AUROC ↑ |
| BBBP | classification | 78.5 | 76.5 | AUROC ↑ |
| Clintox | classification | 97.0 | 92.3 | AUROC ↑ |
| ChEBI-20-*mol2text* | generation | 0.687 | 0.681 | METEOR ↑ |
| ChEBI-20-*text2mol* | generation | 0.877/0.535 | 0.872/0.522 | BLEU/Exact Match ↑ |
| HOMO | regression | 0.0029 | 0.0022 | MAE ↓ |
| LUMO | regression | 0.0029 | 0.0024 | MAE ↓ |
| HOMO-LUMO gap | regression | 0.0040 | 0.0028 | MAE ↓ |

Table 15: Performance comparison on molecule description generation task on Mol-Instrutions (Fang et al., 2023) dataset (**Best**, Second Best).

| MODEL | BLEU-2↑ | BLEU-4↓ | ROUGE-1↑ | ROUGE-2↑ | ROUGE-L↑ | METEOR↑ |
|---|---|---|---|---|---|---|
| *Molecule Description Generation* | | | | | | |
| Alpaca-7B | 0.068 | 0.014 | 0.178 | 0.041 | 0.136 | 0.107 |
| Baize-7B | 0.064 | 0.015 | 0.189 | 0.053 | 0.148 | 0.106 |
| ChatGLM-6B | 0.055 | 0.011 | 0.163 | 0.036 | 0.121 | 0.105 |
| Llama-7B | 0.059 | 0.014 | 0.164 | 0.066 | 0.148 | 0.184 |
| Vicuna-7B | 0.052 | 0.011 | 0.151 | 0.055 | 0.130 | 0.168 |
| Galactica-6.7B | 0.024 | 0.008 | 0.074 | 0.015 | 0.063 | 0.065 |
| Mol-Instructions-7B | 0.217 | 0.143 | 0.337 | 0.196 | 0.291 | 0.254 |
| Text+Chem T5-223M | 0.062 | 0.036 | 0.126 | 0.075 | 0.119 | 0.139 |
| MolT5-248M | 0.002 | 0.001 | 0.036 | 0.001 | 0.034 | 0.033 |
| **BioT5+** | **0.549** | **0.497** | **0.758** | **0.701** | **0.747** | **0.715** |

metric is MAE.

### G.1.2 Chemical Reaction-related Tasks

In computational chemistry, tasks centered around chemical reactions are crucial, as they significantly enhance research and development efficiency and support the advancement of eco-friendly chemistry methods. These tasks typically require understanding the interplay among reactants, reagents, and products, usually represented in the format of "reactant > reagent > product". In line with Fang et al., 2023 and Cao et al., 2023, we concentrates on three specific tasks: (1) Reagent Prediction: This critical task entails the identification of key components such as catalysts, solvents, or auxiliary substances necessary for executing a chemical reaction. The input for this task comprises the reactants and the intended product, challenging the model to deduce the required reagents for the reaction process. (2) Forward Reaction Prediction: This task focuses on forecasting the potential outcomes of a chemical reaction. Provided with the reactants and reagents, the objective is to accurately predict the products that would result from the chemical reaction, thereby aiding in the planning and optimization of chemical synthesis. (3) Retrosynthesis: An essential task in synthetic chemistry, retrosynthesis involves working backwards from a target product to identify plausible reactant combinations for its synthesis. The input is a specific product compound, and the challenge lies in determining the most efficient and feasible reactants needed to produce it. All of these chemical reaction-related data in instruction format and corresponding splits are sourced from Mol-Instruction (Fang et al., 2023) dataset.

We compare our BioT5+ with LLM-based generalist models, including Alpaca (Taori et al., 2023), Baize (Xu et al., 2023a), ChatGLM (et.al., 2023), Llama (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Galactica (Taylor et al., 2022), Text+Chem T5 (Christofidellis et al., 2023), Mol-Instructions (Fang et al., 2023), and InstructMol (Cao et al., 2023). The baseline results are sourced from InstructMol (Cao et al., 2023). The evaluation metrics are exact match, BLEU (Papineni et al., 2002), Levenshtein distance (Miller et al., 2009), three molecular fingerprints (FTS) similarity scores including MACCS (Durant et al., 2002), RDK (Schneider et al., 2015), and Morgan (Rogers and Hahn, 2010), and validity score (whether the SMILES can be successfully processed by RDKit (Landrum, 2021)).

Table 16: Performance comparison on description-guided molecule design task on Mol-Instructions (Fang et al., 2023) dataset (**Best**, Second Best).

| MODEL | EXACT↑ | BLEU↑ | LEVENSHTEIN↓ | RDK FTS↑ | MACCS FTS↑ | MORGAN FTS↑ | VALIDITY↑ |
|---|---|---|---|---|---|---|---|
| *Description-guided Molecule Design* | | | | | | | |
| Alpaca-7B | 0.000 | 0.004 | 51.088 | 0.006 | 0.029 | 0.000 | 0.002 |
| Baize-7B | 0.000 | 0.006 | 53.796 | 0.000 | 0.000 | 0.000 | 0.002 |
| ChatGLM-6B | 0.000 | 0.004 | 53.157 | 0.005 | 0.000 | 0.000 | 0.005 |
| Llama-7B | 0.000 | 0.003 | 59.864 | 0.005 | 0.000 | 0.000 | 0.003 |
| Vicuna-7B | 0.000 | 0.006 | 60.356 | 0.006 | 0.001 | 0.000 | 0.001 |
| Galactica-6.7B | 0.000 | 0.192 | 44.152 | 0.135 | 0.248 | 0.088 | 0.992 |
| Mol-Instructions-7B | 0.002 | 0.345 | 41.367 | 0.231 | 0.412 | 0.147 | 1.000 |
| Text+Chem T5-223M | 0.097 | 0.508 | 41.819 | 0.352 | 0.474 | 0.353 | 0.721 |
| MolT5-248M | **0.112** | 0.546 | 38.276 | 0.400 | 0.538 | 0.295 | 0.773 |
| **BioT5+** | 0.079 | **0.795** | **30.728** | **0.567** | **0.687** | **0.410** | 1.000 |

### G.1.3 Molecule Description Generation

Molecule description generation task requires the model to generate a comprehensive description for a given molecule, encompassing its properties, functions, and potential applications. Unlike the straightforward prediction of molecular properties, this task poses a significantly greater challenge. It demands an in-depth and holistic understanding of the molecule from the model, necessitating not only the recognition of its structural and chemical characteristics but also an integration of this knowledge into a coherent, detailed narrative. There are two well-established benchmark datasets for this task: ChEBI-20 (Edwards et al., 2022) and Mol-Instructions (Fang et al., 2023) with recommended splits, and the corresponding results are presented in Table 4 and Table 15.

We compare our BioT5+ with (1) single-task specialist models, including Transformer (Vaswani et al., 2017), T5 (Raffel et al., 2020), MolT5 (Edwards et al., 2022), MoMu (Su et al., 2022), MolFM (Luo et al., 2023b), MolXPT (Liu et al., 2023c), GIT-Mol (Liu et al., 2023a), Text+Chem T5 (Christofidellis et al., 2023), BioT5 (Pei et al., 2023), and MolCA (Liu et al., 2023d); (2) retrieval-based LLM, including GPT-3.5-turbo (Li et al., 2024) and GPT-4 (OpenAI, 2023). (3) LLM-based generalist models, including GPT-3.5-turbo (Li et al., 2024), BioMedGPT (Luo et al., 2023c), Mol-Instructions (Fang et al., 2023), and InstructMol (Cao et al., 2023). The baseline results are mainly derived from MolXPT (Liu et al., 2023c), BioT5 (Pei et al., 2023), MolCA (Liu et al., 2023d), and InstructMol (Cao et al., 2023). The evaluation metrics are common NLP metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

### G.1.4 Description-guided Molecule Design

Description-guided molecule design is the inverse task of molecule description generation. This task centers around designing molecules based on detailed textual descriptions. Here, the model is presented with a comprehensive description encompassing various aspects of a desired molecule, such as its intended functions, properties, and potential applications. The challenge lies in accurately interpreting this textual information and translating it into a specific molecular structure. This task is considerably complex as it requires the model to have a deep understanding of the relationship between molecular characteristics and their corresponding textual descriptors. This task is crucial for advancing drug discovery and material design, where precise molecular configurations are often derived from elaborate functional requirements. We use the same two benchmark datasets: ChEBI-20 (Edwards et al., 2022) and Mol-Instructions (Fang et al., 2023) as the molecule description generation task, and the corresponding results are shown in Table 5 and Table 16.

The compared baselines are a subset of that listed in Section G.1.3 except for Llama2 (Touvron et al., 2023b), with results mainly sourced from BioT5 (Pei et al., 2023) and MolReGPT (Li et al., 2024). The evaluation metrics are BLEU (Papineni et al., 2002), exact match, Levenshtein distance (Miller et al., 2009), three molecular fingerprints (FTS) similarity scores including MACCS (Durant et al., 2002), RDK (Schneider et al., 2015), and Morgan (Rogers and Hahn, 2010), FCD score (Preuer et al., 2018), Text2Mol score (Edwards et al., 2021), and validity score.

### G.2 Protein-oriented Tasks

#### G.2.1 Protein Description Generation

In computational biology, the task of protein description generation is of paramount importance, as it entails extracting insightful textual information from protein sequences. Aligning with Fang et al., 2023, our focus is on four intricate generation tasks, each taking a protein sequence as its input: (1) Protein Function Generation: This task aims to produce outputs that consist of Gene Ontology (GO) terms, providing a multifaceted description of the protein's functions. These GO terms cover three key domains: cellular component, biological process, and molecular function, offering a holistic view of the protein's role and interactions within a cellular context. (2) Catalytic Activity Generation: Here, the focus is on delineating the specific catalytic activities of the protein, moving beyond merely identifying its Enzyme Commission (EC) number. The output targets a detailed characterization of the chemical reactions facilitated by the protein, capturing its dynamic role in metabolic and biochemical pathways. (3) Domain/Motif Generation: This task involves pinpointing and describing domains or motifs within the protein sequence. Domains and motifs are essential elements, recognized as compact, folded three-dimensional structures that play pivotal roles in the protein's function and stability. The identification of these features is crucial for understanding protein folding, function, and interactions. (4) Functional Description Generation: The goal here is to generate a comprehensive and detailed textual description that encapsulates a protein's function, its subcellular localization, and its involvement in various biological processes. This output seeks to provide an extensive narrative, encompassing the diverse functionalities, roles, and significance of the protein within a biological system. These tasks collectively aim to deepen our understanding of proteins, facilitating advancements in fields such as drug discovery, molecular biology, and bioinformatics. All of these data in instruction format and corresponding splits are sourced from Mol-Instruction (Fang et al., 2023) dataset.

We compare our BioT5+ with LLM-based generalist models, including Alpaca (Taori et al., 2023), Baize (Xu et al., 2023a), ChatGLM (et.al., 2023), Galactica (Taylor et al., 2022), Llama (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), and Mol-Instructions (Fang et al., 2023). The baseline results come from Mol-Instructions (Fang et al., 2023). The evaluation metric is ROUGE-L (Lin, 2004).

### G.3 Description-guided Protein Design

Similar to description-guided molecule design, in the description-guided protein design task, the primary objective is to generate amino acid sequences of proteins that meet specific user-defined design requirements. This task necessitates a profound and comprehensive understanding of protein structures and functions from the model. It involves the intricate process of translating complex textual descriptions, which may include functional targets, structural characteristics, and desired biological activities, into precise amino acid sequences. As there is no well-established benchmark for this task, we show some cases for this task in Section 20.

### G.4 Protein Property Prediction

The protein property prediction task plays a pivotal role in computational biology, focusing on the prediction of specific protein attributes such as solubility, structural characteristics, or functional roles, based on their amino acid sequences or structural information. Following Pei et al., 2023, we centered on two key protein property prediction tasks within the PEER (Xu et al., 2022) benchmark. (1) Solubility Prediction: This task involves determining the solubility status of a given protein. It seeks to predict if a protein, when introduced into a solvent, will dissolve or remain insoluble. This property is crucial as it influences the protein's functionality and its interaction with other biomolecules. (2) Localization Prediction: The second task focuses on identifying the cellular localization of proteins, distinguishing whether a given protein is "membrane-bound" or "soluble". Membrane-bound proteins are those that are associated with or integrated into the cell membrane, playing key roles in various cellular processes such as signal transduction and transport. In contrast, soluble proteins are those that are not associated with the membrane and are typically involved in various intracellular activities. We use the same data and splitting methods as BioT5 (Pei et al., 2023).

We compare our BioT5+ with (1) single-task specialist models, including DDE (Dipeptide Deviation from Expected Mean) (Saravanan and Gautham, 2015), Moran feature descriptor (Moran correlation) (Feng and Zhang, 2000), LSTM (Hochreiter and Schmidhuber, 1997), Trans-

formers (Vaswani et al., 2017), CNN (O'Shea and Nash, 2015), ResNet (He et al., 2016), ProtBert (Elnaggar et al., 2021), ESM-1b (Rives et al., 2021), and BioT5 (Pei et al., 2023); (2) multi-task generalist models, including CNN (O'Shea and Nash, 2015), Transformers (Vaswani et al., 2017), and ESM-1b (Rives et al., 2021). Note that here the multi-task generalist results are derived from PEER (Xu et al., 2022), where contact prediction, fold classification, and secondary structure prediction are combined with the original task. We report the best results obtained from training each of these tasks in conjunction with the primary task. The baseline results are derived from PEER (Xu et al., 2022). The evaluation metric is Accuracy.

### G.5 Protein-Protein Interaction

The Protein-Protein Interaction (PPI) task is an essential component in the field of computational biology, focusing on the prediction of interactions between two proteins based on their amino acid sequences. In this task, the input consists of the amino acid sequences of two distinct proteins, and the output is a binary classification: "yes" if the proteins are predicted to interact, and "no" if they are not. This task holds substantial biological significance as protein-protein interactions are fundamental to most biological processes, including signal transduction, cellular metabolism, and immune responses. Following Pei et al., 2023, we use Yeast and Human PPI datasets with corresponding splits from the PEER (Xu et al., 2022) benchmark, which include proteins related to yeast and humans respectively.

The compared baselines are the same with that in Section G.4, with results derived from PEER (Xu et al., 2022). The evaluation metric is Accuracy.

### G.6 Drug-Target Interaction

The Drug-Target Interaction (DTI) task focuses on the prediction of interactions between a drug molecule and a protein target. This task involves inputting the molecular structure of a drug, encoded as a SELFIES representation, alongside the amino acid sequence of a target protein. The output is a binary decision: "yes" indicates a predicted interaction between the drug and the protein, and "no" suggests no interaction. Understanding and predicting these interactions is crucial for drug discovery and development, offering insights into the mechanism of action of drugs, identifying potential off-target effects, and aiding in the design of novel therapeutics with improved efficacy and reduced side effects. We use the same data and corresponding splits with BioT5 (Pei et al., 2023).

We compare our BioT5+ with single-task specialist models, including Support Vector Machine (Cortes and Vapnik, 1995) (SVM), Random Forest (Ho, 1995) (RF), DeepConv-DTI (Lee et al., 2019), GraphDTA (Nguyen et al., 2021), MolTrans (Huang et al., 2021), DrugBAN (Bai et al., 2023), and BioT5 (Pei et al., 2023). The baseline results are sourced from DrugBAN (Bai et al., 2023) and BioT5 (Pei et al., 2023). The evaluation metric is AUROC, AUPRC, and Accuracy.

## H Case Studies

In this section we show some cases for our fine-tuned BioT5+ model. We only showcase key parts of inputs, intentionally omitting the instruction context due to the varying requirements for instruction context across different baseline methods.

### H.1 Molecule-oriented Cases

We show some selected cases for molecule-oriented tasks, including molecule description generation in Table 17, description guided molecule design in Table 18, and chemical reaction-related tasks in Table 19.

### H.2 Description-guided Protein Design

We also show some cases of description-guided protein design task in Table 20. We compute the normalized Smith-Waterman (SW) alignment score (Yamanishi et al., 2008) between the output and ground truth protein FASTA sequence. This method involves comparing segments of variable lengths within the protein sequences and is optimized to maximize the similarity metric, effectively evaluating the correspondence between the model output and the ground truth. In the model output of amino acid sequences, occurrences of more than two consecutive amino acids at the terminus were manually truncated.

Table 17: **Molecule Description Generation Cases**. Given a molecule, the model needs to generate the corresponding text description. We compare BioT5+ with Galactica (Taylor et al., 2022) (6.7B), Mol-Instructions (Fang et al., 2023), and Text+Chem T5 (Christofidellis et al., 2023) (base-augm version). We adjust the instructions and molecule representations (SMILES or SELFIES) for different models according to their papers.

| | |
|---|---|
| Input | [N][C][=Branch1][C][=O][N][O][C][C@H1] [Branch1][C][NH3+1][C][=Branch1][C][=O][O-1] |
| Target | The molecule is an L-alpha-amino acid zwitterion arising from transfer of a proton from the carboxy to the amino group of O-ureido-L-serine. It is a tautomer of an O-ureido-L-serine. |
| Galactica | The molecule is a zwitterion, so it has a positive charge on the nitrogen atom and a negative charge on the oxygen atom. The positive charge is on the nitrogen atom and the negative charge is on the oxygen atom. |
| Mol-Instructions | The molecule is an amino acid zwitterion obtained by transfer of a proton from the carboxy to the amino group of N-methyl-L-aspartic acid. It is an amino acid zwitterion and a N-methyl-L-aspartate. It is functionally related to a N-methyl-L-aspartic acid. It is a tautomer of a N-methyl-L-aspartic acid. It is a tautomer of a N-methyl-L-aspartic acid zw |
| Text+Chem T5 | The molecule is an N-acyl-D-alpha-amino acid zwitterion obtained by transfer of a proton from the carboxy to the amino group of N-carbamoyl-D-ornithine; major species at pH 7.3. It is a tautomer of a N-carbamoyl-D-ornithine. |
| BioT5+ | The molecule is an amino acid zwitterion resulting from the transfer of a proton from the carboxy to the amino group of O-ureido-L-serine. It is a tautomer of an O-ureido-L-serine. |
| Input | [O][=C][Branch1][C][O][C@H1][Branch1][C][O][C][=Branch1] [C][=O][C][O][P][=Branch1][C][=O][Branch1][C][O][O] |
| Target | The molecule is the ketoaldonic acid phosphate formed formally from D-erythronic acid by oxidation of the 3 -hydroxy group to an oxo group and phosphorylation at the 1 -hydroxy group. It is a conjugate acid of a (R)-2 -hydroxy-3 -oxo-4 -(phosphonatooxy)butanoate(3 -). It is an enantiomer of a (S)-2 -hydroxy-3 -oxo-4 -(phosphonooxy)butanoic acid. |
| Galactica | The molecule is a carboxylic acid. |
| Mol-Instructions | The molecule is a metabolite found in Escherichia coli. It has a role as a human metabolite and a Saccharomyces cerevisiae metabolite. It is a conjugate acid of a 3-phosphoglycerate(2-). It is an enantiomer of a 3-phosphoglycerate(1-). It is a tautomer of a 3-phosphoglyceric acid zwitterion. It is a tautomer of a 3-phosphoglyceric acid zwitterion. |
| Text+Chem T5 | The molecule is the D-enantiomer of 2-dehydro-3-deoxy-D-glucuronic acid. It has a role as an Escherichia coli metabolite and a mouse metabolite. It is a conjugate acid of a (R)-2-dehydro-3-deoxy-D-glucuronate(2-). It is an enantiomer of a (S)-2-dehydro-3-deoxy-L-glucuronic acid. |
| BioT5+ | The molecule is the ketoaldonic acid phosphate formed formally from L-erythronic acid by oxidation of the 3 -hydroxy group to an oxo group and phosphorylation at the 1 -hydroxy group. It derives from a L-erythronic acid. It is a conjugate acid of a (R)-2 -hydroxy-3 -oxo-4 -(phosphonatooxy)butanoate(3 -). It is an enantiomer of a (S)-2 -hydroxy-3 -oxo-4 -(phosphonooxy)butanoic acid. |
| Input | [NH3+1][C@@H1][Branch1][#C][C][C][=C][NH1][C][=C][C][=C][C] [=C][Ring1][=Branch2][Ring1][=Branch1][C][=Branch1][C][=O][O] |
| Target | The molecule is the L-enantiomer of tryptophanium. It has a role as an animal metabolite and a plant metabolite. It is a conjugate acid of a L-tryptophan. It is an enantiomer of a D-tryptophanium. |
| Galactica | The molecule is [NH3+1][C@@H1](CC1=C[NH1]C2=CC=CC=C12)C(=O)O. |
| Mol-Instructions | The molecule is an amino acid zwitterion. It has a role as a human metabolite. It is functionally related to an alanine. It is a tautomer of a L-alanine. It contains a L-alanine. It is a tautomer of a D-alanine zwitterion. It is an enantiomer of a D-alanine zwitterion. It is a tautomer of a D-alanine zwitterion. It is an enantiomer of a L-alanine zwitterion. |
| Text+Chem T5 (Ours) | The molecule is an ammonium ion resulting from the protonation of the amino group of (S)-nefopam. It is a conjugate acid of a (S)-nefopam. It is an enantiomer of a (R)-nefopam(1+). |
| BioT5+ | The molecule is the L-enantiomer of tryptophanium. It has a role as an Escherichia coli metabolite and a Saccharomyces cerevisiae metabolite. It is a conjugate acid of a L-tryptophan. It is an enantiomer of a D-tryptophanium. |

Table 18: **Description-guided Molecule Design**. Given a text description, the model needs to generate the molecule that fits the description. We compare BioT5+ with Galactica (Taylor et al., 2022) (6.7B), Mol-Instructions (Fang et al., 2023), and Text+Chem T5 (Christofidellis et al., 2023) (base-augm version). We adjust the instructions for different models according to their papers.

| | |
|---|---|
| Input | The molecule is any secondary alcohol that is one of the eight possible diastereoisomers of 5 -methyl-2 -(propan-2 -yl)cyclohexan-1 -ol. It has a role as a volatile oil component. It is a p-menthane monoterpenoid and a secondary alcohol. |
| Target SMILES | CC1CCC(C(C)C)C(O)C1 |
| Target SELFIES | [C][C][C][C][C][Branch1][=Branch1][C][Branch1]<br>[C][C][C][C][Branch1][C][O][C][Ring1][#Branch2] |
| Galactica | CC1=CC=C(C)C(C)=C1NC(=O)C1=CC=C(Cl)C=C1 |
| Mol-Instructions | [C][C@@H1][C][C][C@@H1][Branch1] |
| Text+Chem T5 | CC(C)C1CCC(CC1)O |
| BioT5+ | [C][C][C][C][C][Branch1][=Branch1][C][Branch1]<br>[C][C][C][C][Branch1][C][O][C][Ring1][#Branch2] |
| Input | The molecule is a phenylsulfate oxoanion that is the conjugate base of 2 -aminophenyl hydrogen sulfate, obtained by deprotonation of the sulfo group; major species at pH 7.3. It is a conjugate base of a 2 -aminophenyl hydrogen sulfate. |
| Target SMILES | NC1=CC=CC=C1OS(=O)(=O)[O-1] |
| Target SELFIES | [N][C][=C][C][=C][C][=C][Ring1][=Branch1]<br>[O][S][=Branch1][C][=O][=Branch1][C][=O][O-1] |
| Galactica | C[C@H]1C[C@@H]2[C@H]1[C@@H]1C[C@H]3[C@@H]4C[C@H](F)C5=CC<br>(=O)C=C[C@]5(C)[C@@]4(F)[C@@H](O)C[C@]3(C)[C@]2(C(=O)CO)O1 |
| Mol-Instructions | [C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1]<br>[N][S][=Branch1][C][=O][=Branch1][C][=O][O-1] |
| Text+Chem T5 | C1=CC=C(C(=C1)N)OS(=O)(=O)[O-] |
| BioT5+ | [N][C][=C][C][=C][C][=C][Ring1][=Branch1]<br>[O][S][=Branch1][C][=O][=Branch1][C][=O][O-1] |
| Input | The molecule is a dicarboxylic acid. It is a conjugate acid of a diphenate(1 -). It derives from a hydride of a biphenyl. |
| Target SMILES | O=C(O)C1=CC=CC=C1C2=CC=CC=C2C(=O)O |
| Target SELFIES | [O][=C][Branch1][C][O][C][=C][C][=C][C][=C][Ring1][=Branch1]<br>[C][=C][C][=C][C][=C][Ring1][=Branch1][C][=Branch1][C][=O][O] |
| Galactica | C[C@@H]1[C@@H]2C[C@H]1[C@@H]1O[C@@H]12 |
| Mol-Instructions | [C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1][C][=Branch1]<br>[C][=O][O][C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1]<br>[C][=Branch1][C][=O][O][C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1][C] |
| Text+Chem T5 | C1=CC=C(C(=C1)C2=CC=CC=C2)C(=O)O |
| BioT5+ | [O][=C][Branch1][C][O][C][=C][C][=C][C][=C][Ring1][=Branch1]<br>[C][=C][C][=C][C][=C][Ring1][=Branch1] |

Table 19: **Chemical Reaction-related Tasks**. We compare BioT5+ with Galactica (Taylor et al., 2022) (6.7B), Mol-Instructions (Fang et al., 2023), and Text+Chem T5 (Christofidellis et al., 2023) (base-augm version) on reagent prediction, forward reaction prediction, and retrosynthesis tasks. We adjust the instructions for different models according to their papers.

| | |
|---|---|
| *Reagent Prediction* | |
| Input | [O][=N+1][Branch1][C][O-1][O-1].[O][=C][NH1][C][=Branch1][C][=O][C][=C][C][=C][C][=C][Ring1][=Branch1][NH1][Ring1][O]»[O][=C][NH1][C][=Branch1][C][=O][C][=C][C][Branch1][=Branch1][N+1][=Branch1][C][=O][O-1][=C][C][=C][Ring1][=Branch2][NH1][Ring1][=C] |
| Target SMILES | O=S(=O)(O)O.[K+] |
| Target SELFIES | [O][=S][=Branch1][C][=O][Branch1][C][O][O].[K+1] |
| Galactica | O=C1[NH1]C(=O)C2=CC=CC=C2[NH1]1 |
| Mol-Instructions | [C][C][=Branch1][C][=O][O].[C][C][=Branch1][C][=O][O].[C][C][=Branch1][C][=O][O].[O][=C][Branch1][C][O-1][O-1].[O][=N+1][Branch1][C][O-1][O-1].[O][=N+1][Branch1][C][O-1][O-1].[O][=N+1][Branch1][C][O-1] |
| BioT5+ | [O][=S][=Branch1][C][=O][Branch1][C][O][O].[K+1] |
| *Forward Reaction Prediction* | |
| Input | [C][C][=N][O][C][Branch1][=C][C][=C][C][=C][Branch1][C][Cl][C][=C][Ring1][#Branch1][C][O][=N][Ring1][=C].[Cl][C][Cl].[O][=Mn][=O] |
| Target SMILES | Cc1noc(-c2ccc(Cl)cc2C=O)n1 |
| Target SELFIES | [C][C][=N][O][C][Branch1][=C][C][=C][C][=C][Branch1][C][Cl][C][=C][Ring1][#Branch1][C][=O][=N][Ring1][=C] |
| Galactica | CC1=CC=C(C2=CC(C3=CC=C(Cl)C=C3Cl)N3NN=NC3=N2)C=C1 |
| Mol-Instructions | [C][C][=N][O][C][Branch1][=C][C][=C][C][=C][Branch1][C][Cl][C][=C][Ring1] |
| Text+Chem T5 | CC1=NOC(C2=CC=C(Cl)C=C2C=O)=N1 |
| BioT5+ | [C][C][=N][O][C][Branch1][=C][C][=C][C][=C][Branch1][C][Cl][C][=C][Ring1][#Branch1][C][=O][=N][Ring1][=C] |
| *Retrosynthesis* | |
| Input | [C][C][C][C][C][=C][C][=C][Branch1][P][C][=C][C][=C][Branch1][#Branch1][C][=Branch1][C][=O][O][C][C][=C][Ring1][#Branch2][C][=C][Ring1][S].[C][C][C][O][C][Ring1][Branch1].[O].[Na+1].[OH1-1] |
| Target SMILES | BrCc1ccccc1.Oc1ccc(I)nc1Cl |
| Target SELFIES | [Br][C][C][=C][C][=C][C][=C][Ring1][=Branch1].[O][C][=C][C][=C][Branch1][C][I][N][=C][Ring1][#Branch1][Cl] |
| Galactica | CC1=CC=C(C2=CC=C(C(=O)O)C=C2)C=C1S(=O)(=O)[O-] |
| Mol-Instructions | [C][C][C][C][C][=C][C][=C][Branch1][=Branch1][C][=C][Ring1][=Branch1][C][=C][Ring1][Branch2][C][=C][C][=C][Branch1][=Branch1][C][=Branch1][C][=O][O][C][=C][Ring1][=Branch2].[O][=C][Branch1][C][Cl][C][=Branch1][C][=O] |
| Text+Chem T5 | CCCCC1=CC=C(C2=CC=C(C(=O)O)C=C2)C=C1 |
| BioT5+ | [Br][C][C][=C][C][=C][C][=C][Ring1][=Branch1].[O][C][=C][C][=C][Branch1][C][I][N][=C][Ring1][#Branch1][Cl] |

Table 20: **Description-guided Protein Design**. Given a text description, the model needs to generate the protein that fits the description. We compare BioT5+ with Galactica (Taylor et al., 2022) (6.7B) and Mol-Instructions (Fang et al., 2023). We adjust the instructions for different models according to their papers.

| | |
|---|---|
| Input | 1. The designed protein should have at least Helical transmembrane domains to ensure proper localization.<br>2. A protein that can perform RNA binding, cysteine-type endopeptidase activity, serine-type peptidase activity, ribonucleoside triphosphate phosphatase activity, structural molecule activity in viral envelope, lipid droplet, virion membrane, membrane, host cell endoplasmic reticulum membrane to support virion attachment to host cell, fusion of virus membrane with host endosome membrane, clathrin-dependent endocytosis of virus by host cell.<br>3. The protein should be designed to catalyze the following reaction a ribonucleoside 5 '- triphosphate + H2 O = a ribonucleoside 5 '- diphosphate + H(+) + phosphate.<br>4. The designed protein have Helicase ATP-binding, Peptidase C18 domains with specific sequence motifs that are crucial for its activity.<br>5. Target a Basic and acidic residues compositional bias in the protein's composition for improved properties. |
| Target FASTA | MSTNPKPQRKTKRNTNRRPQDVKFPGGGQIVGGVYLLPRRGPRLGVRATRKTSERSQP<br>RGRRQPIPKARRTEGRTWAQPGYPWPLYGNEGMGWAGWLLSPRGSRPSWGPTDPRRRS<br>RNLGKVIDTLTCGFADLMGYIPLVGAPLGGAARALAHGVRVLEDGVNYATGNLPGCSFS<br>IFLLALLSCLTIPASAIEVRNVSGVYHVTNDCSNTSIVYEAADMIMRTPGCVPCVRENNSS<br>RCWVALTPTLAARNSSIP |
| Galactica | MRR...<br>*Normalized SW score*: 0.109 |
| Mol-Instructions | MSTNPKPQRKTKRNTNRRPMDVKFPGGGQIVGGVYLLPRRGPRLGVRAARKTRRSLSD<br>RESWAQLAGWRSADPTGRPEPLDWAPQRTLRGLHGAIFEPNPWPPATERALERRGHRG<br>RRSVTAGWSAREAGGTPEETPSDLSEEDDEGEDRGGTISSSPGSASPDDDAESLDTLLEP<br>AELRRRRRNLGRNGPYSGHPRYRCGCRNVDCDRPCFRCEWPCPRTVRDPCLYCRFSCTS<br>LTEEGDSGNLSNDSANSRNSVSLTAVTSAAEQQQTTNNPNLPVPIISDTILQPQMQQITS...<br>*Normalized SW score*: 0.428 |
| BioT5+ | MSTNPKPQRKTKRNTNRRPQDVKFPGGGQIVGGVYLLPRRGPRLGVRATRKTSERSQP<br>RGRRQPIPKARRPEGRTWAQPGYPWPLYGNEGCGWAGWLLSPRGSRPSWGPTDPRRRS<br>RNLGKVIDTLTCGFADLMGYIPLVGAPLGGAARALAHGVRVLEDGVNYATGNLPGCSFS<br>IFLLALLSCLTVPASAYEVRNASGVYHVTNDCSNASIVYEAADMIMHTPGCVPCVRENN<br>SSRCWVALTPTLAARNSSIP<br>*Normalized SW score*: 0.972 |
| Input | 1. The protein's functional design necessitates the inclusion of a PxxP; stabilizes the interaction of NEF/MHC-I with host AP1 M1 ; necessary for MHC-I internalization motif.<br>2. For general function, the protein need meet that Inhibits the Fas and TNFR- mediated death signals by blocking MAP3 K5 /ASK1 ; Plays a role in optimizing the host cell environment for viral replication without causing cell death by apoptosis<br>3. A protein that localizes to extracellular region, host cell plasma membrane and can perform SH3 domain binding with high specificity and efficiency. |
| Target FASTA | MGGKWSKSSIVGWPAVRERIRQTEPRTEPAAEGVGAASQDLDKYGALTSSNTAHNNADC<br>AWLQAQEEEEGVGFPVKPQVPLRPMTYKAAFDLSFFLKEKGGLDGLIWSRKRQEILDLW<br>VYHTQGFFPDWQNYTPGPGVRLPLTFGWCYKLVPVDPSEVEEANKGENNCLLHPMSQY<br>GMEDEEREVLKWQFDSSLARRHLAREKHPEYYNDC |
| Galactica | MAA<br>*Normalized SW score*: 0.119 |
| Mol-Instructions | MAGHSQFKRLLTREPKVRLHGNLARSLAEQIRDWRAAEDPLYWRADPGAVEAFLRDAGE<br>PAEVVPAAAGDPTAATPEALGRGGDDDRPRSSSGGPSPRSGSGPSGAEREGGLTPDELPPPE<br>SPTSVAEEPVEEEEKEEETENSTEELSDLEEEEDSEDSGTTSLTDTTVSKKKRKSRKDKAKR<br>YKQLISKIDTEEYIKRILDCPICLDCPPDIIHPDIKDLFSDFVDLSKGVSLCDCKTDNGFTCEV<br>CGMMLIGQYSYMDPYRFIEFACSNILNRTFNSHLVPFNLMEICNDITNEIPQNFINRNTFFC...<br>*Normalized SW score*: 0.328 |
| BioT5+ | MGGKWSKSSIVGWPAVRERIRRTEPAAEGVGAASQDLDKHGALTSSNTAHNNADCAWLQ<br>AQEEEEVGFPVRPQVPLRPMTYKGAFDLSFFLKEKGGLEGLIYSKKRQEILDLWVYHTQG<br>FFPDWQNYTPGPGVRYPLTFGWCFKLVPVDPREVEEANEGENNCLLHPMSQHGMEDEDR<br>EVLKWKFDSSLARRHLAREKHPEFYKDC<br>*Normalized SW score*: 0.916 |