BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations
Qizhi Pei$^{1,5}$, Wei Zhang$^2$, Jinhua Zhu$^2$, Kehan Wu$^2$, Kaiyuan Gao$^3$, Lijun Wu$^4$*, Yingce Xia$^4$*, Rui Yan$^{1,6}$*

$^1$Gaoling School of Artificial Intelligence, Renmin University of China
$^2$University of Science and Technology of China
$^3$Huazhong University of Science and Technology
$^4$Microsoft Research
$^5$Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education
$^6$Beijing Key Laboratory of Big Data Management and Analysis Methods

E-mail: {qizhipei,ruiyan}@ruc.edu.cn
{weizhang_cs,teslazhu,wu_2018}@mail.ustc.edu.cn
im_kai@hust.edu.cn {lijuwu,yinxia}@microsoft.com

Abstract
Recent advancements in biological research leverage the integration of molecules, proteins, and natural language to enhance drug discovery. However, current models exhibit several limitations, such as the generation of invalid molecular SMILES, underutilization of contextual information, and equal treatment of structured and unstructured knowledge. To address these issues, we propose BioT5, a comprehensive pre-training framework that enriches cross-modal integration in biology with chemical knowledge and natural language associations. BioT5 utilizes SELFIES for 100% robust molecular representations and extracts knowledge from the surrounding context of bio-entities in unstructured biological literature. Furthermore, BioT5 distinguishes between structured and unstructured knowledge, leading to more effective utilization of information. After fine-tuning, BioT5 shows superior performance across a wide range of tasks, demonstrating its strong capability of capturing underlying relations and properties of bio-entities. Our code is available at https://github.com/QizhiPei/BioT5.

1 Introduction
Molecules and proteins are two essential bio-entities in drug discovery (Dara et al., 2022). Small molecule drugs have been the cornerstone of the pharmaceutical industry for nearly a century, owing to their unique advantages such as oral availability, diverse modes of action, etc (AstraZeneca, 2023). Proteins serve as the foundation of life science, functioning as drug targets or crucial elements in disease pathways. As illustrated in Figure 1, both molecules and proteins can be represented using sequences. A molecule can be depicted by its name, bio-sequence (SMILES and SELFIES), and 2D graph structure. Protein can be represented by its name, corresponding gene name, bio-sequence (FASTA), and 3D structure.

![Figure 1: Representations of molecule and protein.](https://github.com/QizhiPei/BioT5)
et al., 2023; Boutet et al., 2007) serve as knowledge repositories of molecules and proteins. These resources detail properties, experimental results, and interactions between various bio-entities, which cannot be explicitly inferred from molecular or protein sequences alone. Consequently, a recent trend involves jointly modeling text along with molecules and proteins, allowing the textual descriptions to enhance molecular and protein representations. MolT5 (Edwards et al., 2022) adopts the T5 (Raffel et al., 2020) framework to molecular SMILES and biomedical literature. MolXPT (Liu et al., 2023b) and Galactica (Taylor et al., 2022) are GPT models trained on text and bio-entities, such as SMILES and FASTA sequences. DeepEIK (Luo et al., 2023) fuses the encoded features from multi-modal inputs using attention (Vaswani et al., 2017) mechanism. Despite their success, there is still significant room for improvement: (i) Prior work often relies on SMILES to represent molecules. However, addressing the issue of generating invalid SMILES remains a challenge to overcome (Edwards et al., 2022; Li et al., 2023). (ii) The contextual information surrounding molecular or protein names could offer valuable insights for understanding the interactions and properties of bio-entities. Developing effective methods to leverage this information merits further attention. (iii) Existing research tends to treat structured data (e.g., molecule-text pairs from databases) and unstructured data (e.g., text sequences in literature) equally. However, structured data could be utilized more effectively to further enhance overall performance.

To address the above challenges, in this paper, we introduce BioT5, a comprehensive pre-training framework encompassing text, molecules, and proteins. BioT5 leverages SELFIES (Krenn et al., 2020) to represent small molecules since its advantage over SMILES is that SELFIES offers a more robust and error-tolerant molecular representation, eliminating issues of illegitimate structures often encountered with SMILES. There are mainly two steps for BioT5 pre-training:

(1) **Data collection & processing**: We gather text, molecule, and protein data, as well as existing databases containing molecule-text parallel data and protein-text parallel data. For the text data (PubMed) from the biological domain, we employ named entity recognition and entity linking to extract molecular and protein mentions, replacing them with the corresponding SELFIES or FASTA sequences. Following Liu et al. (2023b), we refer to such data as “wrapped” text. Text tokens, FASTA sequences, and SELFIES are tokenized independently (see Section 3.2 for more details).

(2) **Model training**: BioT5 utilizes a shared encoder and a shared decoder to process various modalities. The standard T5 employs the “recover masked spans” objective, wherein each masked span and its corresponding part share the same sentinel token. We refer to the aforementioned training objective function as the “T5 objective” for simplicity. There are three types of pre-training tasks: (i) Applying the standard T5 objective to molecule SELFIES, protein FASTA, and general text independently, ensuring that the model possesses capabilities in each modality. (ii) Applying the T5 objective to wrapped text from the biological domain, where all text, FASTA, and SELFIES tokens can be masked and recovered. (iii) For the structured molecule-text data, we introduce a translation objective. Specifically, BioT5 is trained to translate molecule SELFIES to the corresponding description and vice versa. Likewise, the translation objective is applied to protein-text data.

After pre-training, we fine-tune the obtained BioT5 on 15 tasks covering molecule and protein property prediction, drug-target interaction prediction, protein-protein interaction prediction, molecule captioning, and text-based molecule generation. BioT5 achieves state-of-the-art performances on 10 tasks and exhibits results comparable to domain-specific large models on 5 tasks, demonstrating the superior ability of our proposed method. BioT5 model establishes a promising avenue for the integration of chemical knowledge and natural language associations to augment the current understanding of biological systems.

2 Related Work

In this section, we briefly review related work about cross-modal models in biology and representations of molecule and protein.

2.1 Cross-modal Models in Biology

Language models in the biology field have gained considerable attention. Among these, BioBERT (Lee et al., 2020) and BioGPT (Luo et al., 2022), which are pre-trained on scientific corpora, have been particularly successful in effectively understanding scientific texts. More recently, cross-modal models focusing on jointly modeling
text with bio-sequences have emerged. They can be categorized into the following three groups.

**Cross Text-molecule Modalities**

MoT5 (Edwards et al., 2022) is a T5 (Raffel et al., 2020)-based model, which is jointly trained on molecule SMILES and general text corpus. MoSu (Su et al., 2022) is trained on molecular graphs and related textual data using contrastive learning. MoIXPT (Liu et al., 2023b) is a GPT (Radford et al., 2018)-based model pre-trained on molecule SMILES, biomedical text, and wrapped text. Different from BioT5, these models all use SMILES to represent molecules, which leads to validity issues when generating molecules.

**Cross Text-protein Modalities**

ProteinDT (Liu et al., 2023a) is a multi-modal framework that uses semantically-related text for protein design. BioTranslator (Xu et al., 2023a) is a cross-modal translation system specifically designed for annotating biological instances, such as gene expression vectors, protein networks, and protein sequences, based on user-written text.

**Cross Three or More Biology Modalities**

Galactica (Taylor et al., 2022) is a general GPT-based large language model trained on various scientific domains, including scientific paper corpus, knowledge bases (e.g., PubChem (Kim et al., 2023) molecules, UniProt (uni, 2023) protein), codes, and other sources. DeepEIK (Luo et al., 2023) fuses the feature from multi-modal inputs (drugs, proteins, and text). Then attention (Vaswani et al., 2017) mechanism is adopted to do textual information denoising and heterogeneous features integration.

Our work differs from previous studies in several ways: (1) we primarily focus on two biological modalities—molecule, protein—with text serving as a knowledge base and bridge to enrich the underlying relations and properties in the molecule and protein domains; (2) we use multi-task pre-training to model the connections between these three modalities in a more comprehensive manner. (3) we use SELFIES instead of SMILES to represent molecules, which is more robust and resolves the validity issue in molecule generation tasks.

### 2.2 Representations of Molecule and Protein

**Molecule Representation**

The representation and modeling of molecules have long been a challenge in bioinformatics. There are many methods to represent a molecule: name, fingerprint (Rogers and Hahn, 2010a), SMILES (Weininger, 1988; Weininger et al., 1989), InChI (Heller et al., 2013), DeepSMILES (O’Boyle and Dalke, 2018), SELFIES (Krenn et al., 2020), 2D molecular graph, etc. SMILES (Simplified Molecular-Input Line-Entry System), a compact and textual representation of the molecular structure, is the most common method. It employs a sequence of characters to encode atoms, bonds, and other molecular features. However, SMILES has several drawbacks (Krenn et al., 2020), such as the lack of syntactic and semantic robustness, which significantly affects the validity of molecules generated by deep learning models (Edwards et al., 2022). To address this issue, SELFIES (Self-referencing Embedded Strings) is introduced as a 100% robust molecular string representation (Krenn et al., 2020). Every permutation of symbols within the SELFIES alphabet invariably generates a chemically valid molecular structure, ensuring that each SELFIES corresponds to a valid molecule. Unlike existing works introduced in Section 2.1 that use SMILES for molecule representation, we employ SELFIES with separate encoding in BioT5 to achieve 100% validity in downstream molecule generation tasks.

**Protein Representation**

Protein can also be represented in various ways, such as by its name, corresponding gene name, FASTA format, or 3D geometric structure. The FASTA format is a common choice for encoding protein sequences, which uses single-letter codes to represent the 20 different amino acids. In BioT5, we also employ FASTA format for protein representation.

Unlike Edwards et al. (2022) and Taylor et al. (2022) that share the dictionary between bio-sequence tokens and nature language tokens, BioT5 uses a separate dictionary and biology-specific tokenization to explicitly distinguish biological modalities. We give further analysis of this in Section 3.2.

### 3 BioT5

The overview of the BioT5 pre-training is illustrated in Figure 2. We combine data from different modalities to perform multi-task pre-training.

#### 3.1 Pre-training Corpus

As shown in Figure 2, the pre-training corpus of BioT5 is categorized into three classes: (1) **Single-modal data**, including molecule SELFIES, protein FASTA, and general text. For small molecules, we use the ZINC20 (Irwin et al., 2020) dataset and convert SMILES to SELFIES. For protein
FASTA, we randomly sample proteins from the Uniref50 (Suzek et al., 2007) dataset, filtering out proteins exceeding a specified length, resulting in a collection of 27M proteins. For general text, we use the “Colossal Clean Crawled Corpus” (C4) dataset (Raffel et al., 2020). (2) Wrapped text, where molecule names are replaced with their corresponding SELFIES and gene names are appended with related protein FASTA. We use 33M PubMed articles (Canese and Weis, 2013) and apply BERN2 (Sung et al., 2022) for named entity recognition. The scientific sentences which are not replaced or appended by bio-sequences are remained as a supplement to general text. The detailed process is depicted in Figure 4 and discussed in Appendix B. (3) Molecule-description pairs and protein-description pairs. For molecule-text data, we collect 339K molecule SELFIES along with their corresponding names and descriptions from PubChem (Kim et al., 2019), excluding all molecules present in the downstream ChEBI-20 dataset (Edwards et al., 2022) to avoid potential data leakage. For protein-text data, we obtain 569K protein FASTA-description pairs from SwissProt (Boutet et al., 2007), which contains high-quality annotations of various protein properties. Details are left in Appendix E.1.

### 3.2 Separate Tokenization and Embedding

In most previous works, the representation of molecules and proteins has not been modeled with sufficient attention to detail. MolT5 (Edwards et al., 2022) employs the same dictionary as the original T5, as it starts pre-training from the original T5 checkpoint. The original T5 dictionary is derived from nature language using SentencePiece (Kudo and Richardson, 2018). However, directly utilizing this dictionary for molecule SMILES is suboptimal, as some chemically meaningful tokens, such as functional groups or complete atoms, will be tokenized inaccurately. For example, in the molecule depicted in Figure 3, the bromine atom, symbolized as “[Br-]” as a complete token, thereby producing the correct output.

Figure 2: Overview of BioT5 pre-training. The solid line refers to the “T5 objective”, which aims to reconstruct the original unmasked input. Each consecutive span of masked tokens is replaced with a sentinel token, depicted as \(<M1>-[\text{\textsuperscript{\texttt{ID}}}]>[\text{\texttt{eom}}]\). The character-based tokenization of Galactosalt in MolT5 release of ID folding pathway.

Figure 3: Case for tokenization. MolT5 processes “Br” (bromine atom) as “B” (boron atom) and “r”, resulting in incorrect descriptions including tetraborate (related to “B”). BioT5 retains the chemically meaningful group “[Br-]” as a complete token, thereby producing the correct output.
tica (Taylor et al., 2022) suffers the same issue. In addition to the tokenization method, sharing token embeddings for different modalities (Edwards et al., 2022; Taylor et al., 2022) is also questionable. In multilingual tasks, shared embeddings allow models to accurately represent the meanings of borrowed words and cognates, which retain their original meanings across languages. However, molecules, proteins, and text represent entirely distinct languages. The same token within these three different modalities carries different semantic meanings. For example, the token “C” signifies character C in nature language, the carbon atom in molecules, and cysteine (one of the 20 amino acids) in proteins. Studies by Beltagy et al. (2019) and Gu et al. (2021) further emphasize the significance of domain-specific vocabulary.

To address the issues mentioned above, we employ separate vocabularies for molecule, protein, and text. In BioT5, molecule is represented by SELFIES string, where each chemical meaningful atom group is enclosed within brackets and tokenized as a SELFIES token. For example, \([\text{C}]=[\text{C}][\text{Br}]→[\text{C}],[=\text{C}],[\text{Br}]\). For protein, to differentiate amino acids with capital letters in text, we introduce a special prefix <p> for each amino acid. For example, <p>M<p>K<p>R→<p>M, <p>K, <p>R. For text, we use the same dictionary as the original T5. Through this, we explicitly distinguish the semantic space of different modalities, which maintains the inherent integrity of each unique modality and prevents the model from conflating meanings across modalities.

3.3 Model and Training

Model architecture BioT5 employs the same architecture as T5 models (Raffel et al., 2020). We follow the configuration used in T5-v1.1-base\(^1\). The vocabulary size of BioT5 is 35,073, differing from the default configuration as we incorporate separate vocabulary for molecule SELFIES and protein amino acids. In total, the BioT5 model comprises 252M parameters.

Pre-training During the pre-training phase, the model is trained in a multi-task way on six tasks that can be classified into three types: (1) Applying T5 objective to wrapped text from scientific corpus (task #4). (2) Bidirectional translation for the molecule SELFIES-text pairs (task #5) and protein FASTA-text pairs (task #6). By effectively learning the underlying connections and properties of bio-entities from textual information through these pre-training tasks, BioT5 allows for a holistic understanding of the biological domain, thereby facilitating enhanced prediction and generation abilities in various biological tasks.

4 Experiments and Results

We evaluate BioT5 on 15 well-established downstream tasks, which can be categorized into three types: single-instance prediction, multi-instance prediction, and cross-modal generation. We include details regarding fine-tuning datasets, baselines, and prompts in Appendix F.

For the downstream binary classification tasks presented in Section 4.1 and 4.2, the calculation of evaluation metrics such as AUROC and AUPRC necessitates the soft probability of the predicted label. As we use the prompt-based fine-tuning method, the output is either Yes for the positive label or No for the negative label. To obtain an appropriate label distribution, following Liu et al. (2023b), we first extract the probabilities of Yes and No tokens (denoted as \(p_{\text{pos}}\) and \(p_{\text{neg}}\) respectively) and normalize them. The resulting probability for positive label is \(\frac{p_{\text{pos}}}{p_{\text{pos}}+p_{\text{neg}}}\) and negative label is \(\frac{p_{\text{neg}}}{p_{\text{pos}}+p_{\text{neg}}}\).

4.1 Single-instance Prediction

4.1.1 Molecule Property Prediction

Molecule property prediction aims to determine whether a given molecule exhibits specific properties. MoleculeNet (Wu et al., 2018) is a widely used benchmark for molecule property prediction, encompassing diverse datasets that cover numerous molecular aspects, such as quantum mechanics, physical chemistry, biophysics, etc. In line with Liu et al. (2023b), we conduct experiments on six binary classification tasks, including BBBP, Tox21, ClinTox, HIV, BACE, and SIDER. Following (Fang...
Table 1: Performance comparison on MoleculeNet (Best, Second Best). The evaluation metric is AUROC. The baseline results are mainly sourced from MolXPT (Liu et al., 2023b).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Molecules</th>
<th>#Tasks</th>
<th>BBBP</th>
<th>Tox21</th>
<th>ClinTox</th>
<th>HIV</th>
<th>BACE</th>
<th>SIDER</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2039</td>
<td>1</td>
<td>7831</td>
<td>1478</td>
<td>41127</td>
<td>1513</td>
<td>1427</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison of different methods on solubility and localization prediction tasks (Best, Second Best). The evaluation metric is accuracy. * represents only tuning the prediction head. The baseline results are sourced from PEER (Xu et al., 2022).

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params.</th>
<th>Solubility</th>
<th>Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDE</td>
<td>205.3K</td>
<td>59.77 ± 1.21</td>
<td>77.43 ± 0.42</td>
</tr>
<tr>
<td>Moran</td>
<td>123.4K</td>
<td>57.73 ± 1.55</td>
<td>55.63 ± 0.85</td>
</tr>
<tr>
<td>LSTM</td>
<td>26.7M</td>
<td>70.18 ± 0.63</td>
<td>88.11 ± 0.14</td>
</tr>
<tr>
<td>Transformer</td>
<td>21.3M</td>
<td>70.12 ± 0.31</td>
<td>75.74 ± 0.74</td>
</tr>
<tr>
<td>CNN</td>
<td>5.4M</td>
<td>64.43 ± 0.25</td>
<td>82.67 ± 0.32</td>
</tr>
<tr>
<td>ResNet</td>
<td>11.0M</td>
<td>67.33 ± 1.46</td>
<td>78.99 ± 4.41</td>
</tr>
<tr>
<td>ProtBert</td>
<td>419.9M</td>
<td>68.15 ± 0.92</td>
<td>91.32 ± 0.89</td>
</tr>
<tr>
<td>ProtBert*</td>
<td>419.9M</td>
<td>59.17 ± 0.21</td>
<td>81.54 ± 0.09</td>
</tr>
<tr>
<td>ESM-1b</td>
<td>652.4M</td>
<td>70.23 ± 0.75</td>
<td>92.40 ± 0.35</td>
</tr>
<tr>
<td>ESM-1b*</td>
<td>652.4M</td>
<td>67.02 ± 0.40</td>
<td>91.61 ± 0.10</td>
</tr>
<tr>
<td>BioT5</td>
<td>252.1M</td>
<td>74.65 ± 0.49</td>
<td>91.69 ± 0.05</td>
</tr>
</tbody>
</table>

4.1.2 Protein Property Prediction

Protein property prediction is crucial as it provides critical insights into the behavior and functions of proteins. We concentrate on two protein property prediction tasks on PEER benchmark (Xu et al., 2022): protein solubility prediction, which aims to predict whether the given protein is soluble, and protein localization prediction, which is to classify proteins as either “membrane-bound” or “soluble”.

Baselines We compare BioT5 with three types of baselines provided in PEER benchmark: (1) feature engineers, including two protein sequence feature descriptors: Dipeptide Deviation from Expected Mean (DDE) (Saravanan and Gautham, 2015) and Moran correlation (Moran) (Feng and Zhang, 2000); (2) protein sequence encoders, including LSTM (Hochreiter and Schmidhuber, 1997), Transformers (Vaswani et al., 2017), CNN (O’Shea and Nash, 2015) and ResNet (He et al., 2016); (3) pre-trained protein language models, which are pre-trained using extensive collections of protein FASTA sequences, including ProtBert (Elnaggar et al., 2021) and ESM-1b (Rives et al., 2021).
Drug-target interaction (DTI) prediction plays a crucial role in drug discovery, as it aims to predict whether a given drug (molecule) and target (protein) can interact with each other. We select three widely-used DTI datasets with a binary classification setting, which are BioSNAP (Zitnik et al., 2018), BindingDB (Liu et al., 2007) and Human (Liu et al., 2015; Chen et al., 2020).

**Baselines** We compare BioT5 with two types of baselines: (1) traditional machine learning methods including SVM (Cortes and Vapnik, 1995) and Random Forest (RF) (Ho, 1995); (2) deep learning methods including DeepConv-DTI (Lee et al., 2019), GraphDTA (Nguyen et al., 2021), MolTrans (Huang et al., 2021) and DrugBAN (Bai et al., 2023), in which drug and target feature are firstly extracted by well-design drug encoder and protein encoder then fused for prediction.

**Results** The results on BioSNAP, Human, and BindingDB datasets are presented in Table 3. All statistics are obtained from five random runs. On BioSNAP and BindingDB datasets, BioT5 consistently outperforms other methods in various performance metrics, including AUROC, AUPRC, and accuracy. For the Human dataset, although deep learning-based models generally exhibit strong performance, the BioT5 model demonstrates a slight advantage over the baseline models. It is worth noting that, in contrast to most deep learning-based baselines, our BioT5 does not rely on a specific design tailored for molecules or proteins. A possible explanation for the superior performance of BioT5 is that the SELFIES and FASTA representations effectively capture the intricate structure and function of molecules and proteins, and the interaction information between them may be well-described in the contextual scientific literature or corresponding text entries in databases.

### 4.2 Multi-instance Prediction

#### 4.2.1 Drug-target Interaction Prediction

Drug-target interaction (DTI) prediction plays a vital role in understanding protein functions and structures, as it aims to determine the potential interactions between drug and target proteins.

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params.</th>
<th>Yeast</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDE</td>
<td>205.3K</td>
<td>55.83 ± 3.13</td>
<td>62.77 ± 2.30</td>
</tr>
<tr>
<td>Moran</td>
<td>123.4K</td>
<td>53.00 ± 0.50</td>
<td>54.67 ± 4.43</td>
</tr>
<tr>
<td>LSTM</td>
<td>26.7M</td>
<td>53.62 ± 2.72</td>
<td>63.75 ± 5.12</td>
</tr>
<tr>
<td>Transformer</td>
<td>21.3M</td>
<td>54.12 ± 1.27</td>
<td>59.58 ± 2.09</td>
</tr>
<tr>
<td>CNN</td>
<td>5.4M</td>
<td>55.07 ± 0.02</td>
<td>62.60 ± 1.67</td>
</tr>
<tr>
<td>ResNet</td>
<td>11.0M</td>
<td>48.91 ± 1.78</td>
<td>68.61 ± 3.78</td>
</tr>
<tr>
<td>ProtBert</td>
<td>419.9M</td>
<td>63.72 ± 2.80</td>
<td>77.32 ± 1.10</td>
</tr>
<tr>
<td>ProtBert*</td>
<td>419.9M</td>
<td>53.87 ± 0.38</td>
<td>83.61 ± 1.34</td>
</tr>
<tr>
<td>ESM-1b</td>
<td>652.4M</td>
<td>57.00 ± 6.38</td>
<td>78.17 ± 2.91</td>
</tr>
<tr>
<td>ESM-1b*</td>
<td>652.4M</td>
<td>66.07 ± 0.58</td>
<td>88.06 ± 0.24</td>
</tr>
<tr>
<td>BioT5</td>
<td>252.1M</td>
<td>64.89 ± 0.43</td>
<td>86.22 ± 0.53</td>
</tr>
</tbody>
</table>

Table 4: Performance comparison on Yeast and Human datasets (Best, Second Best). The baseline results derive from DrugBAN (Bai et al., 2023).
Table 5: Performance comparison on molecule captioning task (Best, Second Best). Rouge scores are F1 values. The Text2Mol score between ground truth molecule and corresponding text description is 0.609. The baseline results derive from MolT5 (Edwards et al., 2022), MolXPT (Liu et al., 2023b), and MolReGPT (Li et al., 2023).

Table 6: Performance comparison on text-based molecule generation task (Best, Second Best). Following Edwards et al. (2022), BLEU, Exact, Levenshtein, and Validity are computed on all generated molecules while other metrics are computed only on syntactically valid molecules. The Text2Mol score for ground truth is 0.609. The baseline results derive from MolT5 (Edwards et al., 2022), MolXPT (Liu et al., 2023b), and MolReGPT (Li et al., 2023).

interactions between pairs of proteins. Following PEER (Xu et al., 2022) benchmark, we perform fine-tuning on two PPI datasets: Yeast (Guo et al., 2008) and Human (Pan et al., 2010).

Baselines The baselines for comparison are the same as that in Section 4.1.2.

Results The results are shown in Table 4. All statistics are over three random runs. On two PPI datasets, BioT5 shows superior performance compared to almost all baseline models. Remarkably, BioT5 outperforms both ProtBert and ESM-1b (with full parameters fine-tuned). This result strongly highlights the crucial role of incorporating textual information during the pre-training of BioT5, which effectively establishes profound connections between proteins. Our model, despite being smaller, is able to harness the unstructured information embedded in scientific text and structured information from biological databases, encapsulating the comprehensive knowledge of proteins in their varying contexts.

4.3 Cross-modal Generation

In this section, we evaluate the performance of BioT5 on the cross-modal generation task. Specifically, we fine-tune BioT5 on molecule captioning and text-based molecule generation tasks. These two tasks are proposed by MolT5 (Edwards et al., 2022) and both use the ChEBI-20 dataset (Edwards et al., 2021). The evaluation metrics and some interesting cases are introduced in Appendix D and G.

4.3.1 Molecule Captioning

For the given molecule, the goal of molecule captioning task is to provide a description of the given...
molecule. As we use SELFIES sequences to represent molecules, this task can be formulated as an exotic sequence-to-sequence translation task.

**Baselines** The baselines include: RNN (Medsker and Jain, 2001), Transformer (Vaswani et al., 2017), T5 (Raffel et al., 2020), MolT5 (Edwards et al., 2022), GPT-3.5-turbo\(^2\) with zero-shot and 10-shot MolReGPT (Li et al., 2023) settings, and MolXPT (Liu et al., 2023b).

**Results** The results are shown in Table 5. BioT5 only has nearly the same number of parameters as MolT5-base, but outperforms all baseline models in all metrics, including those that have more parameters. The Text2Mol score is 0.603, which is very close to the Text2Mol score of 0.609 between the ground truth molecule and the corresponding description. We can attribute this superior performance to the unstructured contextual knowledge and structured database knowledge induced in BioT5 pre-training, which helps the model learn the intricate relationship between text and molecules.

### 4.3.2 Text-Based Molecule Generation

This is a reverse task of molecule captioning. Given the nature language description of the intended molecule, the goal is to generate the molecule that fits the description.

**Baselines** The compared baselines are the same as baselines in Section 4.3.1.

**Results** The results are presented in Table 6. BioT5 only uses parameters similar to MolT5-base yet delivers superior performance across nearly all metrics. Notably, the exact match score of BioT5 surpasses the MolT5-Large by 32.8% while maintaining a validity of 1.0. This indicates that BioT5 not only generates more relevant molecules corresponding to the given text descriptions, but also ensures a 100% validity for the generated molecules. The overall enhanced performance of BioT5 can be attributed to the incorporation of both contextual and database knowledge, as well as the utilization of SELFIES for molecular representation.

### 5 Conclusions and Future Work

In this paper, we propose BioT5, a comprehensive pre-training framework capable of capturing the underlying relations and properties of bio-entities by leveraging both structured and unstructured data sources with 100% robust molecular representation. Our method effectively enriches cross-modal integration in biology with chemical knowledge and natural language associations, demonstrating notable improvements in various tasks.

For future work, we aim to further enrich our model by incorporating additional biological data types, such as genomics or transcriptomics data, to create a more holistic biological pre-training framework. Additionally, we plan to evaluate the interpretability of BioT5 predictions, aiming to provide more insights into the biological systems under study. Thus, we foresee our work sparking further innovation in the use of AI models in the field of computational biology, ultimately leading to a deeper understanding of biological systems and facilitating more efficient drug discovery.

### 6 Limitations

One limitation of BioT5 is conducting full-parameter fine-tuning on each downstream task. This is done because we do not observe generalization ability among different downstream tasks using instruction-tuning (Wei et al., 2022) method. Another reason is that combining data from different tasks using instructions results in data leakage. For example, have noticed overlaps between the training set of BindingDB and the test sets of BioSNAP and Human. Additionally, we only demonstrate the ability of BioT5 in text, molecule, and protein modalities. Numerous other biological modalities, such as DNA/RNA sequences and cells, exist, and there are many other tasks within a single modality or across multiple modalities. Moreover, BioT5 primarily focuses on the sequence format of bio-entities, yet other formats, such as 2D or 3D structures, also hold significant importance. We leave further exploration of these to future work.

### 7 Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2020YFB1406702), 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.

\(^2\)https://openai.com/blog/openai-api
References


A Reproducibility

The codes for our BioT5 are available at https://github.com/QizhiPei/BioT5.

B NER and Entity Linking Process

We follow KV-PLM (Zeng et al., 2022) and MolXPT (Liu et al., 2023b) to conduct Named Entity Recognition (NER) and Entity Linking for the bio-entity names appearing in the scientific text. More specifically, we firstly utilize BERN2 (Sung et al., 2022), an advanced neural Named Entity Recognition (NER) tool in biomedical fields, to identify all instances of molecule or protein mentions. Subsequently, we map them to corresponding entities within publicly accessible knowledge bases. For molecule, we use ChEBI (Hastings et al., 2016) and MeSH (Lipscomb, 2000) database, and for protein we use NCBI Gene (Brister et al., 2015) database. Then we can get the corresponding molecule SELFIES and protein FASTA for the matched entities. As shown in Figure 4, for molecule, we directly replace all the detected names with its SELFIES string; for protein, due to the length limitation, if a sentence consists of more than one protein entity, we only randomly choose one to append the protein FASTA to the name. The motivation for appending protein FASTA instead of replacing is that the genes are transcribed and translated to generate proteins. Therefore, unlike the molecule names directly representing the molecule, the relation between gene names and protein FASTA is indirect. Note that the replacement or appendage will not happen in every sentence. Only those with detected bio-entities will be done the above process.

C Dictionary and SELFIES Conversion

For molecule-related datasets, when only SMILES is provided, we utilize selfies\(^3\) package to convert SMILES into SELFIES.

D Molecule-Text Generation Metrics

We follow Edwards et al. (2022) to use the same evaluation metrics for molecule captioning and text-based molecule generation tasks. To ensure a fair comparison, we convert the molecule SEIFLES to SMILES before calculating these metrics.

D.1 Molecule Captioning Metrics

In the molecule caption task, NLP metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are utilized to evaluate the closeness of the generated description to the ground truth description. We also adopt Text2Mol metric, which is proposed by Edwards et al. (2021) and employ pre-trained models to measure the similarity between the description and ground truth molecule. Higher similarity means that the given text description is more relevant to the molecule, and the Text2Mol score between the ground truth description and molecule is also computed for comparison.

D.2 Text-based Molecule Generation Metrics

Since molecules can be represented in bio-sequence structure, NLP metrics like BLEU (Papineni et al., 2002) and Exact Match scores between generated and ground truth SMILES are directly applied for evaluation. Additionally, we also report performance on molecule-specific metrics: three molecule fingerprints (FTS) similarity scores-MACCS (Durant et al., 2002), RDK (Schneider et al., 2015), and Morgan (Rogers and Hahn, 2010a); Levenshtein distance (Miller et al., 2009); FCD score (Preuer et al., 2018), which measures molecule similarities according to biological information based on pre-trained “ChemNet”; validity, which is the percentage of the valid SMILES that can be processed by RDKit (Landrum, 2021). The Text2Mol metric is also used to measure the similarity between the molecule SMILES and ground truth description.

\(^3\)https://github.com/aspuru-guzik-group/selfies

Figure 4: Wrapped text matching and mapping process.
E Pre-training Details

E.1 Special Tokens
In the pre-training of BioT5, we conduct translation tasks on molecule-text pairs and protein-text pairs extracted from PubChem (Kim et al., 2023) and Swiss-Prot (Boutet et al., 2007) separately. We format the text description from these database entries using special tokens, which serve as anchors for embedding scientific context and structure. For molecule, we use `MOLECULE NAME` and `DESCRIPTION` to represent its name and description including properties, functions, etc. For protein, similar to Xu et al. (2023b), we use `PROTEIN NAME`, `FUNCTION`, `SUBCELLULAR LOCATION`, and `PROTEIN FAMILIES` to represent its name, functions, location and topology in the cell, and families it belongs to. A complete text description is created by concatenating these fields sequentially, omitting any missing fields. Through special tokens, we can effectively encode the intricate information associated with each bio-entity.

E.2 Hyper-parameters
We use the codebase nanoT5 (Nawrot, 2023) for BioT5 pre-training. We pre-train BioT5 for 350K steps on eight NVIDIA 80GB A100 GPUs. The batch size is 96 per GPU, in which a batch includes six types of data. The “translation” directions for molecule-text and protein-text pair are randomly selected for each sample with a probability of 0.5. We use AdamW (Loshchilov and Hutter, 2019) with Root Mean Square (RMS) scaling Optimizer for optimization. The learning rate scheduler is cosine annealing with the base learning rate set to $1 \times 10^{-2}$ and the minimum learning rate set to $1 \times 10^{-5}$. The number of warm-up steps is 10,000 and the dropout rate is 0.0. The maximum input length for pre-training is 512. Unlike absolute position encodings, T5 (Raffel et al., 2020) use relative position encodings. This makes the model flexible to inputs of different lengths, which is helpful for downstream fine-tuning.

F Fine-tuning Details
In this section, we provide details about downstream tasks, including datasets, compared baselines, and prompts. Some statistics about downstream tasks are shown in Table 7. When displaying prompts, ⟨SELFIES⟩ refers to the molecule SELFIES, and ⟨FASTA⟩ refers to the protein FASTA.

F.1 Single-instance Prediction

F.1.1 Molecule Property Prediction
All the datasets are split using an 8 : 1 : 1 ratio for train, validation, and test, respectively. We use the scaffold splitting method, in which molecules are categorized according to the Bemis-Murcko scaffold representation.

Datasets
(1) The BBBP (Blood-Brain Barrier Penetration) is curated with the intention of aiding the modeling and forecasting of barrier permeability. It comprises compounds that are categorized using binary labels, indicating whether they can penetrate the blood-brain barrier.

(2) The Tox21 (“Toxicology in the 21st Century”) initiative established a publicly accessible database that quantifies the toxicity levels of various compounds. The dataset encompasses qualitative toxicity assessments (binary labels) for approximately 8,000 compounds, targeting 12 distinct biological pathways such as nuclear receptors and stress response mechanisms.

(3) The ClinTox dataset contrasts FDA-approved drugs with those that have been unsuccessful in clinical trials owing to toxicity issues. This dataset incorporates two classification objectives for 1,491 drug compounds with established chemical structures: (i) Presence or absence of toxicity in clinical trials; (ii) approved or unapproved by FDA.

(4) The HIV dataset assesses the inhibitory potential of over 40,000 compounds on HIV replication. The screening outcomes were classified into three categories: Confirmed Inactive (CI), Confirmed Active (CA), and Confirmed Moderately Active (CM). Subsequently, the latter two labels were combined, transforming the task into a binary classification between inactive (CI) and active (CA and CM) categories.

(5) The BACE dataset presents quantitative IC50 values and qualitative binary labels for a collection of inhibitors targeting human beta-secretase 1 (BACE-1).

(6) The SIDER (Side Effect Resource) is a comprehensive database that consists of marketed drugs and their corresponding adverse drug reactions (ADR). The drug side effects in SIDER are organized into 27 system organ classes, adhering to the MedDRA classifications. This dataset encompasses data for 1,427 approved drugs.

Baselines
(1) GROVER (Rong et al., 2020) incorporates Mes-
Table 7: Downstream task descriptions, including task or dataset name, type, and the size of each split.

<table>
<thead>
<tr>
<th>Task/Dataset</th>
<th>Task Type</th>
<th>#Train</th>
<th>#Validation</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Molecule Property Prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBBP</td>
<td>Molecule-wise Classification</td>
<td>1,631</td>
<td>204</td>
<td>204</td>
</tr>
<tr>
<td>Tox21</td>
<td>Molecule-wise Classification</td>
<td>6,264</td>
<td>783</td>
<td>784</td>
</tr>
<tr>
<td>ClinTox</td>
<td>Molecule-wise Classification</td>
<td>1,181</td>
<td>148</td>
<td>148</td>
</tr>
<tr>
<td>HIV</td>
<td>Molecule-wise Classification</td>
<td>32,901</td>
<td>4,113</td>
<td>4,113</td>
</tr>
<tr>
<td>BACE</td>
<td>Molecule-wise Classification</td>
<td>1,210</td>
<td>151</td>
<td>152</td>
</tr>
<tr>
<td>SIDER</td>
<td>Molecule-wise Classification</td>
<td>1,141</td>
<td>143</td>
<td>143</td>
</tr>
<tr>
<td><strong>Protein Property Prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solubility prediction</td>
<td>Protein-wise Classification</td>
<td>62,478</td>
<td>1,999</td>
<td>1,999</td>
</tr>
<tr>
<td>Localization prediction</td>
<td>Protein-wise Classification</td>
<td>5,184</td>
<td>1,749</td>
<td>1,749</td>
</tr>
<tr>
<td><strong>Drug-target Interaction Prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioSNAP</td>
<td>Molecule-protein Classification</td>
<td>19,224</td>
<td>2,747</td>
<td>5,493</td>
</tr>
<tr>
<td>Human</td>
<td>Molecule-protein Classification</td>
<td>4,197</td>
<td>600</td>
<td>1,200</td>
</tr>
<tr>
<td>BindingDB</td>
<td>Molecule-protein Classification</td>
<td>50,149</td>
<td>5,604</td>
<td>5,505</td>
</tr>
<tr>
<td><strong>Protein-protein Interaction Prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>Protein-pair Classification</td>
<td>4,945</td>
<td>394</td>
<td>394</td>
</tr>
<tr>
<td>Human</td>
<td>Protein-pair Classification</td>
<td>35,669</td>
<td>237</td>
<td>237</td>
</tr>
<tr>
<td><strong>Molecule Captioning and Text-based Molecule Generation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChEBI-20</td>
<td>Molecule-text Translation</td>
<td>26,407</td>
<td>3,301</td>
<td>3,300</td>
</tr>
</tbody>
</table>

Prompts

For the six MoleculeNet datasets mentioned above, the prompts only differ in the Task Definition. Therefore, we will only provide the Instruction and Output for the first dataset, and the remaining datasets will follow the same format.

(1) **BBBP**

**Task Definition:** Definition: Molecule property prediction task (a binary classification task) for the BBBP dataset. The blood-brain barrier penetration (BBBP) dataset is designed for the model-
ing and prediction of barrier permeability. If the given molecule can penetrate the blood-brain barrier, indicate via "Yes". Otherwise, response via "No".

**Instruction**: Now complete the following example -
**Input**: Molecule: ⟨bom⟩⟨SELFIES⟩⟨eom⟩
**Output**: Yes for inhibitor and No instead.

(2) Tox21

**Task Definition**: Definition: Molecule property prediction task (a binary classification task) for the Tox21 dataset. The Tox21 dataset contains qualitative toxicity measurements for 8k compounds on 12 different targets, including nuclear receptors and stress response pathways. If the given molecule can activate/change/affect (target), indicate via "Yes". Otherwise, response via "No".

(3) ClinTox

**Task Definition**: Definition: Molecule property prediction task (a binary classification task) for the ClinTox dataset. The ClinTox dataset compares drugs approved by the FDA and drugs that have failed clinical trials for toxicity reasons. If the given molecule is ⟨Subtask⟩, indicate via "Yes". Otherwise, response via "No".

(4) HIV

**Task Definition**: Definition: Molecule property prediction task (a binary classification task) for the HIV dataset. The HIV dataset was introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, which tested the ability to inhibit HIV replication for over 40,000 compounds. If the given molecule can inhibit HIV replication, indicate via "Yes". Otherwise, response via "No".

(5) BACE

**Task Definition**: Definition: Molecule property prediction task (a binary classification task) for the BACE dataset. The BACE dataset provides qualitative (binary label) binding results for a set of inhibitors of human beta-secretase 1 (BACE-1). If the given molecule can inhibit BACE-1, indicate via "Yes". Otherwise, response via "No".

(6) SIDER

**Task Definition**: Definition: Molecule property prediction task (a binary classification task) for the SIDER dataset. The Side Effect Resource (SIDER) is a dataset of marketed drugs and adverse drug reactions (ADR). If the given molecule can cause the side effect of ⟨side effect⟩, indicate via "Yes". Otherwise, response via "No".

**F.1.2 Protein Property Prediction**

**Datasets**

(1) Solubility prediction is to predict whether a protein is soluble or not. We follow the same splitting method with DeepSol (Khurana et al., 2018).

(2) Localization prediction aims predict whether a protein is “membrane-bound” or “soluble”, which is a simple version of subcellular localization prediction task. We follow the same splitting method with DeepLoc (Armenteros et al., 2017).

**Baselines**

(1) Feature engineers. The DDE (Dipeptide Deviation from Expected Mean) (Saravanan and Gautam, 2015) feature descriptor, consisting of 400 dimensions, is based on the dipeptide frequency within a protein sequence. The Moran feature descriptor (Moran correlation) (Feng and Zhang, 2000), with 240 dimensions, characterizes the distribution of amino acid properties within a protein sequence.

(2) Protein sequence encoders, including LSTM (Hochreiter and Schmidhuber, 1997), Transformers (Vaswani et al., 2017), CNN (O’Shea, 2007), ProtBert (Elnaggar et al., 2021) and ESM-1b (Rives et al., 2021) are both pre-trained on a massive dataset of protein sequences using the masked language modeling (MLM) objective. Specifically, ProtBert is pre-trained on 2.1 billion protein sequences obtained from the BFD database (Steinegger and Söding, 2018), while ESM-1b is pre-trained on a smaller dataset of 24 million protein sequences sourced from UniRef50 (Suzek et al., 2007).

**Prompts**

(1) Solubility prediction

**Task Definition**: Protein solubility prediction task (a binary classification task) for the solubility dataset. If the given protein is soluble, indicate via "Yes". Otherwise, response via "No".

**Instruction**: Now complete the following example -
**Input**: Protein: ⟨bom⟩⟨FASTA⟩⟨eom⟩
**Output**: Yes for soluble protein or No instead.
Task Definition: Protein subcellular localization task (a binary classification task). If the given protein is membrane-bound, indicate via "Yes". Otherwise (the protein is soluble), response via "No".

Instruction: Now complete the following example -

Input: Protein: ⟨bom⟩⟨FASTA⟩⟨eom⟩ Output:

Yes for membrane-bound protein or No for soluble protein.

F.2 Multi-instance Prediction

F.2.1 Drug-target Interaction Prediction

Datasets

1. BioSNAP (Zitnik et al., 2018) is derived from the DrugBank database (Wishart et al., 2018) and was created by Huang et al. (2021) and Zitnik et al. (2018). It consists of 4,510 drugs and 2,181 proteins. This dataset is balanced, containing both validated positive interactions and an equal number of randomly selected negative samples from unseen pairs.

2. BindingDB (Liu et al., 2007) is an accessible online database that contains experimentally validated binding affinities. Its main focus is on the interactions between small drug-like molecules and proteins. We follow Bai et al. (2023) to use a modified version of the BindingDB dataset, which was previously constructed by Bai et al. (2021) with reduced bias.

3. Human (Liu et al., 2015; Chen et al., 2020) is constructed with the inclusion of highly credible negative samples. Following Bai et al. (2023), we also use a balanced version of the Human dataset, which contains an equal number of positive and negative samples.

Baselines

We compare the performance of BioT5 with the following six models on DTI task.

1. Support Vector Machine (Cortes and Vapnik, 1995) (SVM) on the concatenated fingerprint ECFP4 (Rogers and Hahn, 2010b) (extended connectivity fingerprint, up to four bonds) and PSC (Cao et al., 2013) (pseudo-amino acid composition) features.

2. Random Forest (Ho, 1995) (RF) on the concatenated fingerprint ECFP4 and PSC features.

3. DeepConv-DTI (Lee et al., 2019) uses a fully connected neural network to encode the ECFP4 drug fingerprint and a Convolutional Neural Network (CNN) along with a global max-pooling layer to extract features from protein sequences. Then the drug and protein features are concatenated and fed into a fully connected neural network for final prediction.

4. GraphDTA (Nguyen et al., 2021) uses graph neural networks (GNNs) for the encoding of drug molecular graphs, and a CNN is used for the encoding of protein sequences. The derived vectors of the drug and protein representation are concatenated for interaction prediction.

5. MolTrans (Huang et al., 2021) uses transformer architecture to encode drug and protein. Then a CNN-based interaction module is used to capture their interactions.

6. DrugBAN (Bai et al., 2023) use Graph Convolution Network (GCN) (Kipf and Welling, 2017) and 1D CNN to encode drug and protein sequences. Then a bilinear attention network are adopted to learn pairwise local interactions between drug and protein. The resulting joint representation is decoded by a fully connected neural network.

Prompts

Task Definition: Definition: Drug target interaction prediction task (a binary classification task) for the ⟨Dataset⟩ dataset. If the given molecule and protein can interact with each other, indicate via "Yes". Otherwise, response via "No". Where ⟨Dataset⟩ is one of the three DTI datasets mentioned above.

Instruction: Now complete the following example -

Input: Molecule: ⟨bom⟩⟨SELFIES⟩⟨eom⟩ Protein: ⟨bom⟩⟨FASTA⟩⟨eom⟩ Output:

Yes for positive label or No instead.

F.2.2 Protein-protein Interaction Prediction

Datasets

1. Yeast (Guo et al., 2008) involves determining whether two yeast proteins interact or not. The negative pairs are derived from distinct subcellular locations. Following (Xu et al., 2022), the dataset is split and removed redundancy according to protein sequences similarity, which allows for the evaluation of generalization across dissimilar protein sequences.

2. Human (Pan et al., 2010) involves determining whether two human proteins interact or not. It comprises positive protein pairs sourced from the Human Protein Reference Database (HPRD) (Peri et al., 2003) and negative pairs derived from different subcellular locations. The dataset splitting scheme is similar to that of Yeast PPI prediction with an 8 : 1 : 1 ratio for train/validation/test.

Baselines The compared baselines are the same as the protein property prediction task in Sec-
Prompts
Task Definition: Protein protein interaction prediction task (a binary classification task) for the \(\langle\text{Dataset}\rangle\) dataset. If the given two yeast proteins (Protein\_A and Protein\_B) can interact with each other, indicate via "Yes". Otherwise, response via "No". where \(\langle\text{Dataset}\rangle\) is either yeast or human.
Instruction: Now complete the following example -
Input: Protein\_A: \(\langle\text{bom}\rangle\langle\text{FASTA}\rangle\langle\text{eom}\rangle\)
Protein\_B: \(\langle\text{bom}\rangle\langle\text{FASTA}\rangle\langle\text{eom}\rangle\)
Output: .
Output: Yes for positive label or No instead.

F.3 Cross-modal Generation
F.3.1 Molecule Captioning
Datasets
We use ChEBI-20 dataset created by Text2mol (Edwards et al., 2021), which consists of 33,010 molecule-text pairs and 20 means each text description has more than 20 words. The dataset is split into 8 : 1 : 1 for train, validation, and test.
Baselines
(1) RNN (Medsker and Jain, 2001) with 4-layer bidirectional encoder is trained from scratch on ChEBI-20 dataset.
(2) Transformer (Vaswani et al., 2017) containing 6 encoder and decoder layers is trained from scratch on ChEBI-20 dataset.
(3) T5 (Raffel et al., 2020) is directly fine-tuned on ChEBI-20 dataset from public checkpoints \(^4\) with three different model sizes: small, base and large. Note that no molecule domain knowledge is introduced in the original T5 pre-training.
(4) MolT5 (Edwards et al., 2022) is jointly trained on molecule SMILES from ZINC-15 dataset (Sterling and Irwin, 2015) and general text from C4 dataset (Raffel et al., 2020) so that MolT5 has prior knowledge about these two domains. It also contains three different sizes: small, base and large. Then they are further fine-tuned on ChEBI-20 dataset.
(5) GPT-3.5-turbo (Li et al., 2023) is used by directly call OpenAI API without further fine-tuning. The input includes five parts as Li et al. (2023): role identification, task description, examples, output instruction, and user input prompt. The examples are retrieved by Morgan Fingerprint (Butina, 1999) similarity for molecule captioning task and

by BM25 (Robertson and Zaragoza, 2009) for text-based molecule generation task.
(6) MolXPT (Liu et al., 2023b) is jointly trained on molecule SMILES from PubChem (Kim et al., 2023), biomedical text from PubMed (Canese and Weis, 2013), and “wrapped” text in which molecule names are replaced with corresponding SMILES.
Prompts
Different from the classification task in which the ground truth output is either Yes or No, the output for molecule captioning task is text sequence.
Task Definition: Definition: You are given a molecule SELFIES. Your job is to generate the molecule description in English that fits the molecule SELFIES.
Instruction: Now complete the following example -
Input: \(<\text{bom}\rangle\langle\text{SELFIES}\rangle\langle\text{eom}\rangle\)
Output: .
Output: (Text Description)

F.3.2 Text-based molecule generation
This is the reverse task of molecule captioning. The input is the text description of the desired molecule and the output is the corresponding molecule SELFIES. The datasets and compared baselines are the same with molecule captioning in Section F.3.1 so will only provide the prompts here.
Prompts
Task Definition: Definition: You are given a molecule description in English. Your job is to generate the molecule SELFIES that fits the description.
Instruction: Now complete the following example -
Input: \(<\text{bom}\rangle\langle\text{Text Description}\rangle\langle\text{eom}\rangle\)
Output: .
Output: <\text{bom}\rangle\langle\text{SELFIES}\rangle\langle\text{eom}\rangle}

G Case Study
In this section, we show several example outputs from different models in molecule captioning and text-based molecule generation tasks. Figure 5 shows the cases for the molecule captioning task. In example (1), the description of BioT5 matches the ground truth best, successfully localizing the position of the substituent group and “member of pyridines and an aryl thiol”. In example (2), MolT5 mistakenly describes that the molecule contained boron, while BioT5’s description is more accurate. In example (3), while MolT5 generates repetitive output, BioT5 and T5 generate semantically coherent output, and BioT5’s output matches better with ground truth. For a complex molecule in example (4), the output of BioT5 is more holistic and accurate. Notably, only BioT5 describes this molecule.
as an inhibitor of SARS coronavirus main proteinase, which may come from our integration with protein knowledge. Figure 6 shows the cases for the text-based molecule generation task. From the cases, we have several findings: (i) BioT5 is more likely to generate molecules that exactly match the ground truth. (ii) By using SELFIES, BioT5 will not generate invalid molecules, especially for complex and longer molecules shown in examples (3) and (4). (iii) Some molecules are actually short proteins. Example (3) shows a molecule that is a 33-membered polypeptide, which consists of 33 amino acid residues joined in sequence. Therefore, the boundary between proteins and molecules may not always be distinct, and leveraging information from both can provide reciprocal benefits.
The molecule is a quaternary ammonium salt that is ammonia in which the hydrogens at position 2 are substituted by a sulfanyl group. It has a role as a fluorochrome. It is a member of thiocarbamic acids and an oligopeptide.

Figure 6: Text-based molecule generation cases.