Mol2Lang-VLM: Vision- and Text-Guided Generative Pre-trained Language Models for Advancing Molecule Captioning through Multimodal Fusion

Duong Thanh Tran[†], **Nhat Truong Pham**[†], **Nguyen Doan Hieu Nguyen**, **Balachandran Manavalan**[*]

Department of Integrative Biotechnology, Sungkyunkwan University, Republic of Korea
{duongtt, truongpham96, ndhieunguyen, bala2022}@skku.edu
[†]Equal contribution
[*]**Correspondence:** bala2022@skku.edu

Abstract

This paper introduces Mol2Lang-VLM, an enhanced method for refining generative pre-trained language models for molecule captioning using multimodal features to achieve more accurate caption generation. Our approach leverages the encoder and decoder blocks of the Transformer-based architecture by introducing third sub-layers into both. Specifically, we insert sub-layers in the encoder to fuse features from SELFIES strings and molecular images, while the decoder fuses features from SMILES strings and their corresponding descriptions. Moreover, cross multi-head attention is employed instead of common multi-head attention to enable the decoder to attend to the encoder's output, thereby integrating the encoded contextual information for better and more accurate caption generation. Performance evaluation on the CheBI-20 and L+M-24 benchmark datasets demonstrates Mol2Lang-VLM's superiority, achieving higher accuracy and quality in caption generation compared to existing methods. Our code and pre-processed data are available at https://github.com/nhattruongpham/mol-lang-bridge/tree/mol2lang/.

1 Introduction

In the field of cheminformatics, molecule captioning plays a crucial role in helping researchers by automatically generating captions for molecular structures. The accuracy and quality of these captions are vital as they directly impact the understanding of chemical information and scientific discoveries. Traditional techniques primarily rely on unimodal data, often focusing only on textual representations like SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988) strings or SELFIES (Self-referencing Embedded Strings) (Krenn et al., 2020) strings. Although these methods have shown satisfactory results, their dependence on a single modality limits the richness and accuracy of the generated captions.

The rise of multimodal data, which uses information from different sources, presents an opportunity for significant advancements in molecule captioning. Multimodal approaches integrate various forms of molecule, enabling a more comprehensive understanding of molecular characteristics. However, effectively utilizing multimodal data in generative models is challenging and requires advanced techniques to integrate and improve the models effectively.

In this paper, we introduce an enhanced methodology, named Mol2Lang-VLM, to improve generative models in molecule captioning by utilizing multimodal features. Our approach integrates SELFIES strings and high-level features from molecular images in the encoder, while incorporating SMILES features and corresponding descriptions in the decoder. This multimodal integration allows the model to have a deeper understanding of chemical structures within the generative model, which is further refined during the decoder stage.

2 Related Work

2.1 Unimodal Language Models

MolT5 (Edwards et al., 2022) involves translating molecular structures into natural language using a text-to-text transfer transformer (T5) (Raffel et al., 2020) model. This model leverages the robust linguistic capabilities of T5 to understand and generate descriptions of molecular structures accurately. BioT5 (Pei et al., 2023) extends the capabilities of T5 to integrate chemical knowledge and natural language associations into biological contexts. BioT5 employs SELFIES for representing small molecules, as it offers considerable advantages over SMILES. Specifically, SELFIES ensures a more reliable and error-resistant molecular representation, thereby avoiding the problem of invalid structures that frequently occur with SMILES. This

model improves the cross-modal understanding between biological texts and chemical data. While MolT5 and BioT5 are encoder-decoder language models, MolXPT (Liu et al., 2023) utilizes a generative pre-trained Transformer (GPT) (Radford and Narasimhan, 2018) which is a decoder-only language model by introducing a generative pre-training approach by wrapping molecular structures within descriptive texts. MolXPT leverages both text and SMILES sequences for molecular modeling. It wraps SMILES sequences with text, allowing them to influence each other. Specifically, it detects molecule names in text sequences and replaces them with corresponding SMILES representations. ChemBERTa (Chithrananda et al., 2020) is an encoder-only language model that utilizes the RoBERTa (Liu et al., 2019) model that focuses on molecular representation learning and property prediction.

SwinOCSR (Xu et al., 2022) uses Swin Transformer (Liu et al., 2021) architecture for end-to-end optical chemical structure recognition of molecular images. This model can effectively recognize and describe chemical structures from images, providing a significant improvement in the accuracy of vision-language tasks in cheminformatics.

While the aforementioned models have made significant contributions to molecule captioning, their reliance on unimodal data restricts their potential for advancements. By not harnessing the power of multimodal data, these models encounter limitations in terms of information richness, completeness, contextual understanding, generalization, and interpretability.

2.2 Multimodal Language Models

GIT-Mol (Liu et al., 2024) introduces a multimodal large language model that integrates graph, image, and text data to enhance molecular science applications. This model leverages the strengths of different data modalities to provide comprehensive and accurate molecular descriptions. Besides that, MoMu (Su et al., 2022) associates molecular graphs with natural language, providing a sophisticated multimodal foundation model. This model enhances the interpretability and accuracy of molecular captions by integrating graph representations of molecules with their textual descriptions.

While the use of multimodal data is feasible, the aforementioned models face certain issues. These models necessitate significant computational resources and large dataset training. Additionally, scaling up multimodal models can pose challenges.

3 Methodology

3.1 Generative Language Model

We use the T5 (Raffel et al., 2020) architecture as our generative language model. The process begins with the tokenization of the SELFIES string, resulting in token embeddings $X_t \in \mathbb{R}^{L_{\text{enc}} \times d_t}$. Here, L_{enc} represents the length of the encoder input, while d_t denotes the dimensionality of the feature vectors. The encoder comprises a sequence of N encoding layers, with each layer consisting of a Multi-head Self-Attention (MSA) (Vaswani et al., 2017) mechanism (Eq. 1) and a Feed-Forward Network (FFN) (Eq. 2). Following each sub-layer, there is a residual connection that precedes layer normalization (LN). Unlike the original Transformer (Vaswani et al., 2017), T5 incorporates relative position embeddings (Shaw et al., 2018), which are added to the respective logits during the computation of attention weights and shared across all layers in the model.

$$Z_l^{\text{enc}'} = \text{LN}(\text{MSA}(Z_{l-1}^{\text{enc}}) + Z_{l-1}^{\text{enc}}) \quad (1)$$

$$Z_l^{\text{enc}} = \text{LN}(\text{FFN}(Z_l^{\text{enc}'}) + Z_l^{\text{enc}'}) \quad (2)$$

In parallel with the encoder, the decoder similarly consists of N layers. It exhibits three distinctive aspects compared to the encoder: First, target sequences, which are molecular captions, are tokenized into embeddings $Y_t \in \mathbb{R}^{L_{\text{dec}} \times d_t}$, where L_{dec} is the length of the target sequence. They are shifted right by one token to ensure that the ground-truth token is used as the input to predict the next token. Second, a Masked Multi-head Self-Attention (MMSA) (Vaswani et al., 2017) is utilized to ensure auto-regressive generation, maintaining a strict left-to-right processing order (Eq. 3). Third, a Cross Multi-head Attention (CMA) (Vaswani et al., 2017) layer is employed, which enables the decoder to attend to the encoder's output, thereby integrating the encoded contextual information (Eq. 4). Analogous to the encoder, the decoder includes a FFN in each layer (Eq. 5).

$$Z_l^{\text{dec}'} = \text{LN}(\text{MMSA}(Z_{l-1}^{\text{dec}}) + Z_{l-1}^{\text{dec}}) \quad (3)$$

$$Z_l^{\text{dec}''} = \text{LN}(\text{CMA}(Z_l^{\text{dec}'}, Z_N^{\text{enc}}, Z_N^{\text{enc}}) + Z_l^{\text{dec}'}) \quad (4)$$

$$Z_l^{\text{dec}} = \text{LN}(\text{FFN}(Z_l^{\text{dec}''}) + Z_l^{\text{dec}''}) \quad (5)$$

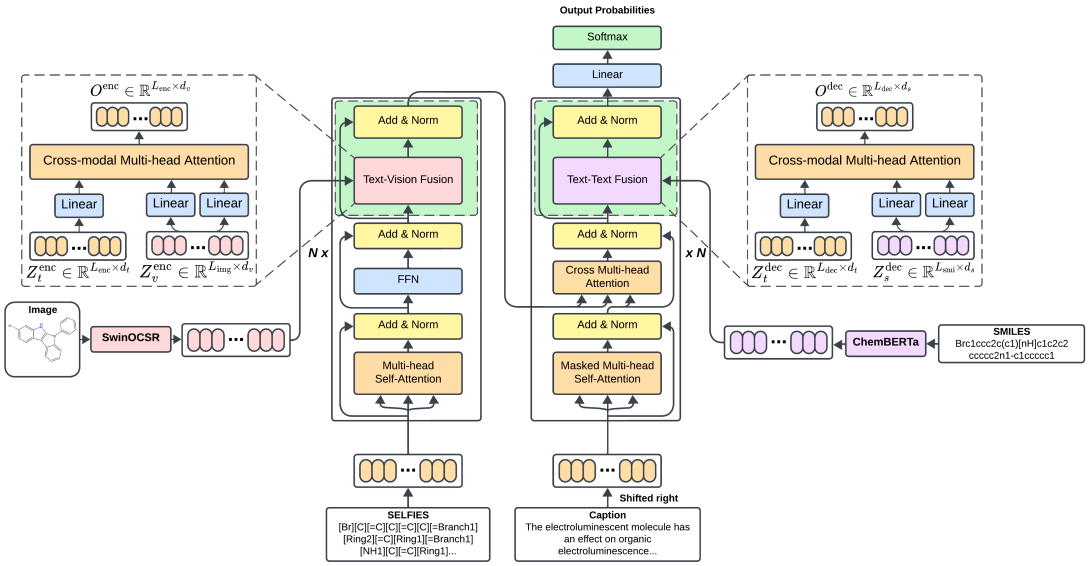

Figure 1: Overview of Mol2Lang-VLM's architecture. The green areas represent the two inserted sub-layers used to fuse the features. The T5 architecture uses relative position embeddings, which are integrated into the Multi-head Attention mechanism and, therefore, are not shown in the figure. Additionally, Cross Multi-head Attention is employed in the decoder instead of Multi-head Attention.

3.2 Vision- and Text-Guided Fusion

Inspired by VG-GPLMs (Yu et al., 2021), we integrate a third sub-layer into both the encoder and decoder of the language model. In the encoder, we insert text-vision fusion at the end of the encoder to learn the cross-modality between the SELFIES string and the molecular image. In the decoder, we replace the last FFN with text-text fusion to capture the relationship between the corresponding caption and the SMILES string. In both fusion processes, we utilize the CMA mechanism to learn the correlation between the two sets of features. The overview of the architecture is exhibited in Figure 1.

In the fusion process of the encoder, the embeddings of SELFIES, denoted as Z_t^{enc}, are linearly projected to the query Q^{enc} (Eq. 6), while the embeddings of the image, denoted as Z_v^{enc}, are linearly projected to the key K^{enc} (Eq. 7) and the value V^{enc} (Eq. 8). These projections are performed before feeding them to the CMA mechanism, which generates the output O^{enc} (Eq. 9).

$$Q^{\text{enc}} = Z_t^{\text{enc}} W_q^{\text{enc}} \tag{6}$$

$$K^{\text{enc}} = Z_v^{\text{enc}} W_k^{\text{enc}} \tag{7}$$

$$V^{\text{enc}} = Z_v^{\text{enc}} W_v^{\text{enc}} \tag{8}$$

$$O^{\text{enc}} = \text{CMA}(Q^{\text{enc}}, K^{\text{enc}}, V^{\text{enc}}) \tag{9}$$

The fusion process of the decoder occurs after the CMA between the decoder's embeddings and the output embeddings of the encoder, resulting in Z_t^{dec}. In this fusion process, Z_t^{dec} are linearly projected to the query Q^{dec} (Eq. 10), while the embeddings of SMILES, denoted as Z_s^{dec}, are also projected to the key K^{dec} (Eq. 11) and the value V^{dec} (Eq. 12). Subsequently, CMA is applied to generate O^{dec} (Eq. 13). The features of SMILES help enhance the overall effectiveness of the features, enabling more effective generation of the desired output.

$$Q^{\text{dec}} = Z_t^{\text{dec}} W_q^{\text{dec}} \tag{10}$$

$$K^{\text{dec}} = Z_s^{\text{dec}} W_k^{\text{dec}} \tag{11}$$

$$V^{\text{dec}} = Z_s^{\text{dec}} W_v^{\text{dec}} \tag{12}$$

$$O^{\text{dec}} = \text{CMA}(Q^{\text{dec}}, K^{\text{dec}}, V^{\text{dec}}) \tag{13}$$

At each fusion, the output is concatenated with the initial embeddings to produce $Z_t^{\text{enc}'}$ and $Z_t^{\text{dec}'}$ (Eq. 14 and 15).

$$Z_t^{\text{enc}'} = (Z_t^{\text{enc}} \oplus O^{\text{enc}}) W_c^{\text{enc}} \tag{14}$$

$$Z_t^{\text{dec}'} = (Z_t^{\text{dec}} \oplus O^{\text{dec}}) W_c^{\text{dec}} \tag{15}$$

Finally, forget gates, denoted as F^{enc} and F^{dec}, are applied to filter out noisy and redundant information introduced during the interactions (Eq. 16 and 17), then point-wise multiplication is applied on O^{enc} and O^{dec} to produce $O^{\text{enc}'}$ and $O^{\text{dec}'}$ (Eq. 18).

$$F^{\text{enc}} = \sigma((Z_t^{\text{enc}} \oplus O^{\text{enc}}) W_f^{\text{enc}}) \tag{16}$$

$$F^{\text{dec}} = \sigma((Z_t^{\text{dec}} \oplus O^{\text{dec}})W_f^{\text{dec}}) \quad (17)$$

$$O^{\text{enc}'} = F^{\text{enc}} \otimes O^{\text{enc}}, O^{\text{dec}'} = F^{\text{dec}} \otimes O^{\text{dec}} \quad (18)$$

4 Implementation Details

4.1 Architectures

We employ BioT5 (Pei et al., 2023) as our generative language model, which uses the *T5-base* version. The model consists of 252 million parameters and has a configuration that includes an embedding dimensionality of 768. It is composed of 12 layers in both the encoder and decoder. The input tokens and output tokens are limited to a maximum length of 512.

To extract visual features from molecular images, we utilize the encoder of SwinOCSR (Xu et al., 2022) which employs the Swin Transformer (Liu et al., 2021) architecture, uses *Swin-L* version. The encoder has a total of 194 million parameters. By inputting images with the size of 224×224, the encoder generates feature embeddings with a length of 49 and a hidden dimensionality of 1536.

To extract features from SMILES representations, we use ChemBERTa (Chithrananda et al., 2020), which is built upon the *RoBERTa-base* architecture with a total of 44 million parameters. The input tokens for ChemBERTa are also limited to a length of 512.

To compute the cross-modality attention in text-vision fusion of the encoder, as well as text-text fusion of the decoder, all features are linearly projected to a gated dimensionality of 256. The text-vision fusion is then integrated at the last two layers of the encoder (the 11th and 12th layers). Concurrently, text-text fusion is incorporated into the initial two layers of the decoder (the 1st and 2nd layers).

4.2 Datasets

L+M-24: The L+M-24 dataset, first introduced from *Language + Molecules Workshop @ ACL 2024* (Edwards et al., 2024), is designed to highlight three key benefits of natural language in molecule design: compositionality, functionality, and abstraction. It contains over 160, 560 molecule-description pairs, which are divided into 80%/20% for train/validation splits.

CheBI-20: The CheBI-20 dataset is widely used in molecular description tasks. It was first introduced in the Text2Mol (Edwards et al., 2021). This dataset contains 33, 010 molecule-description pairs, which are split into 80%/10%/10% for train/validation/test sets.

Since the aforementioned datasets currently lack SELFIES strings and molecular images, we employ *selfies* [1] and *RDKit* [2] package to generate this additional data. We use the prompting template of the molecule captioning task from BioT5 (Pei et al., 2023) to fine-tune the model.

4.3 Configurations

Training: During the training process, we utilize a batch size of 64. To optimize the model, we employ the AdamW (Loshchilov and Hutter, 2019) optimizer. The learning rate scheduler follows a cosine annealing strategy, with a base learning rate of $3e-5$. The warming-up steps for the learning rate scheduler are set to 1 epoch to gradually adjust the learning rate.

Inference: To ensure a fair comparison when evaluating the model, we employ greedy decoding for generating molecular captions by setting the number of beam search to 1, with the decoder starting token as <*pad*>, and the end of sentence token as <*/s*>. Furthermore, post-processing is also applied to skip all special tokens.

5 Results and Discussion

Table 1 presents the performance comparison of Mol2Lang-VLM with all baseline models, such as MolT5-Small, MolT5-Base, MolT5-Large, and BioT5 on the L+M-24 dataset. We used several performance evaluation metrics to evaluate these models, including BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. Notably, Mol2Lang-VLM outperforms all three baseline models in almost all metrics, with BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR values of 77.7, 56.3, 78.6, 59.1, 56.5, and 74.1, respectively. Although Mol2Lang-VLM achieves a lower METEOR of 0.2 compared to MolT5-Large, its number of parameters is approximately 1.5 times lower than MolT5-Large, indicating that the model can learn more efficiently. Compared to BioT5, Mol2Lang-VLM achieves better performance in terms of BLEU-2, BLEU-4, ROUGE-1, and METEOR, with slightly lower scores in ROUGE-2 and ROUGE-L, demonstrating that it generally outperforms BioT5.

We also evaluate our proposed method, along

[1] https://github.com/aspuru-guzik-group/selfies
[2] https://github.com/rdkit/rdkit

Model	#Params	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolT5-Small	77M	70.9	51.2	74.5	55.8	54.4	70.1
MolT5-Base	248M	73.8	53.5	75.0	55.9	53.9	71.8
MolT5-Large	783M	76.9	55.6	77.7	58.0	55.7	**74.3**
BioT5	252M	74.6	54.1	78.5	**59.3**	**56.9**	72.7
Ours	496M	**77.7**	**56.3**	**78.6**	59.1	56.5	74.1

Table 1: Molecule captioning results on the validation split of L+M-24 dataset (**Best**, Second Best). The baseline results are derived from *Language + Molecules Workshop @ ACL 2024* (Edwards et al., 2024). The Text2Mol metric is excluded from the table because Text2Mol is trained on a different distribution of data compared to the L+M-24 dataset.

Model	#Params	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
MolT5-Small	77M	51.9	43.6	62.0	46.9	56.3	55.1	54.0
MolT5-Base	248M	54.0	45.7	63.4	48.5	57.8	56.9	54.7
MolT5-Large	783M	59.4	50.8	65.4	51.0	59.4	61.4	58.2
BioT5	252M	**63.5**	**55.6**	**69.2**	**55.9**	**63.3**	**65.6**	**60.3**
Ours	496M	61.2	52.7	67.4	53.2	61.4	63.3	59.8

Table 2: Molecule captioning results on the test split of CheBI-20 dataset (**Best**, Second Best). The baseline results are derived from MolT5 (Edwards et al., 2022) and BioT5 (Pei et al., 2023).

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Our model w/o forget gate	75.7	54.7	**78.7**	59.0	**56.7**	73.0
Our model w/ forget gate	**77.7**	**56.3**	78.6	**59.1**	56.5	**74.1**

Table 3: Molecule captioning results on the validation split of the L+M-24 dataset to compare between the model with and without a forget gate (**Best**, Second Best).

with all three baseline models and BioT5, on the CheBI-20 dataset. Table 2 displays the performance comparison in terms of BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, and Text2Mol metrics. Interestingly, Mol2Lang-VLM achieves the second-best performance in all metrics, while BioT5 excels on this dataset. This might be acceptable because, in some cases, the fused information may not provide significant additional context or may even introduce noise, making it challenging for the model to effectively utilize the fused embeddings.

Moreover, we conduct an ablation analysis to evaluate Mol2Lang-VLM with and without employing the forget gate. Table 3 compares the performance of these two strategies. Mol2Lang-VLM with the forget gate outperforms the version without it across most metrics, including BLEU-2, BLEU-4, ROUGE-2, and METEOR. The presence of the forget gate mechanism contributes to enhanced caption quality in terms of accuracy and relevance, showcasing the effectiveness of incorporating this mechanism in the model architecture for improved captioning outcomes.

6 Conclusion

This paper introduced Mol2Lang-VLM, a vision- and text-guided generative pre-trained language model designed to enhance molecule captioning performance through multimodal fusion. Our proposed approach achieved comparative results in terms of BLEU, ROUGE, METEOR, and Text2Mol metrics, demonstrating its effectiveness in generating accurate and meaningful captions for molecular structures. The findings highlight the potential of Mol2Lang-VLM in advancing molecule captioning tasks. Future research can explore alternative fusion methods, fine-tuning strategies, and the generalization of the model to other tasks. Additionally, integrating Mol2Lang-VLM with downstream applications and enhancing interpretability can further enhance its practical utility in the field of cheminformatics.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (RS-2024-00344752). This research was supported by the Department of Integrative Biotechnology, Sungkyunkwan Univer-

sity (SKKU) and the BK21 FOUR Project. This work was supported by the Korea Bio Data Station (K-BDS) with computing resources including technical support.

References

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

Carl Edwards, Tuan 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*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+M-24: Building a dataset for Language + Molecules @ ACL 2024. *arXiv preprint arXiv:2403.00791*.

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*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, page 108073.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023. 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)*, pages 1606–1616, Toronto, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

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.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

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. *Preprint*, arXiv:2209.05481.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Zhanpeng Xu, Jianhua Li, Zhaopeng Yang, Shiliang Li, and Honglin Li. 2022. Swinocsr: end-to-end optical chemical structure recognition using a swin transformer. *Journal of Cheminformatics*, 14(1):41.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.