MTSwitch: A Web-based System for Translation between Molecules and Texts

Nijia Han[1], **Zimu Wang**[1], **Yuqi Wang**[1], **Haiyang Zhang**[1], **Daiyun Huang**[2], **Wei Wang**[1,†]
[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
[2]XJTLU Wisdom Lake Academy of Pharmacy, Xi'an Jiaotong-Liverpool University
{Nijia.Han23,Zimu.Wang19,Yuqi.Wang17}@student.xjtlu.edu.cn
{Haiyang.Zhang,Daiyun.Huang,Wei.Wang03}@xjtlu.edu.cn

Abstract

We introduce MTSwitch, a web-based system for the bidirectional translation between molecules and texts, leveraging various large language models (LLMs). It supports two crucial tasks, including molecule captioning (explaining the properties of a molecule) and molecule generation (designing a molecule based on specific properties). To the best of our knowledge, MTSwitch is currently the first accessible system that allows users to translate between molecular representations and descriptive text contents. The system and a screencast can be found in https://github.com/hanninaa/MTSwitch.

1 Introduction

The advent of large language models (LLMs) has significantly transformed the landscape of natural language processing (NLP) (Peng et al., 2023; OpenAI et al., 2024). Their superior language understanding, generation capabilities, and versatility have propelled their utility beyond conversations, now extending into various domains, including bio-medicine (Wang et al., 2023) and molecular chemistry (Liao et al., 2024). This emerging trend, marked by many novel methods being proposed, highlights a promising new research direction (Zeng et al., 2022; Edwards et al., 2022).

Despite the progress made, using the existing models requires familiarisation with invoking the advanced models, which undoubtedly increases the burden on researchers who are not specialists in this field. Consequently, designing an intuitive and user-friendly system for translating molecules and texts is imperative. In this paper, we design the first web-based system, named MTSwitch (Molecule-Text Switch), to translate between texts and molecules. As shown in Figure 1, MTSwitch can translate molecules in SMILES (Simplified Molecular Input Line Entry System) format along with their captions, based on the translation task, the selected model, and the user's input. In this system, we incorporate four models: the trained MolT5 and Meditron from Edwards et al. (2024), Llama 3 (Touvron et al., 2023), and Mistral (Jiang et al., 2023) trained with a subset of L+M-24 (Edwards et al., 2024). Our system empowers users with direct web access, facilitating seamless integration into education and molecular design.

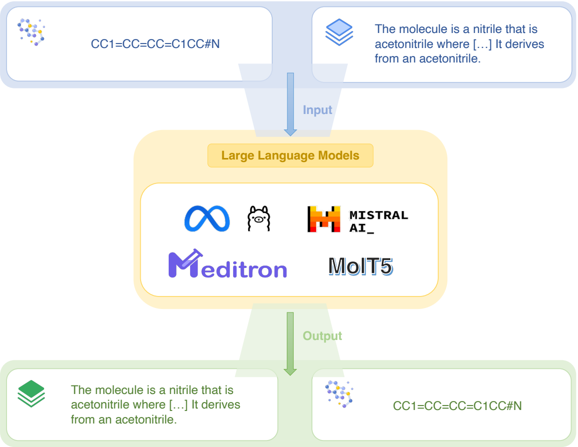
CC1=CC=CC=C1CC#N
The molecule is a nitrile that is acetonitrile where [...] It derives from an acetonitrile.
Input
Large Language Models
MISTRAL AI_
Meditron
MolT5
Output
The molecule is a nitrile that is acetonitrile where [...] It derives from an acetonitrile.
CC1=CC=CC=C1CC#N

Figure 1: System overview of the MTSwitch system.

2 System Overview

MTSwitch offers a user-friendly platform for translation between molecules in SMILES format and natural language, whose web interface is structured as two areas for input and output. As shown in Figure 1, users are prompted to first select an appropriate model (e.g. MolT5 and Meditron) and define their intended task when they access the platform, and then enter their input (either a SMILES notation or a natural language text) in the text box on the left-hand side. Upon submission, the system processes the user input, selects and executes the corresponding model for either molecule captioning or molecule generation, and exhibits the model output in the text box on the right-hand side.

[†] Corresponding author.

Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations, pages 4–6
September 23–27, 2024. ©2024 Association for Computational Linguistics

Model	ROGUE-1↑	ROGUE-2↑	ROGUE-L↑	BLEU-2↑	BLEU-4↑	Meteor↑
Llama 3	66.9	50.1	50.3	61.2	44.1	60.6
Mistral	74.5	55.5	53.6	70.9	51.0	69.0
Meditron	**78.8**	**58.3**	**56.5**	**78.1**	**56.7**	**74.7**
MolT5	75.9	55.4	53.7	75.7	54.2	71.5

Table 1: Experimental results of the molecule captioning task in MTSwitch system with different LLMs, in which the best result for each metric is highlighted in **bold**.

Model	BLEU↑	Levenshtein↓	Validity↑	Uniqueness↑	MACCS↑	RDK↑	Morgan↑
Llama 3	63.6	56.8	92.2	81.2	64.9	57.1	42.1
Mistral	68.5	49.5	91.4	82.6	69.4	62.5	46.1
Meditron	68.4	45.8	**98.8**	97.9	**75.3**	**67.1**	47.8
MolT5	**71.5**	**42.1**	94.5	**99.5**	73.1	65.9	**48.5**

Table 2: Experimental results of the molecule generation task in MTSwitch system with different LLMs, in which the best result for each metric is highlighted in **bold**.

We provide comprehensive model support for the translation between molecules and texts. First, we employ two LLMs pre-trained on the L+M-24 dataset (Edwards et al., 2022), which is a large-scale dataset designed for translating molecules and texts: Meditron and MolT5, which are obtained from the Hugging Face repository[1]. We further fine-tune two state-of-the-art LLMs, Llama 3 (Touvron et al., 2023) and Mistral (Jiang et al., 2023), using a subset of the L+M-24 dataset. During our fine-tuning process, we employ the same instruction and hyperparameters intended for the fine-tuned Meditron model.

This system bridges the gap between complex molecule structures and their textual descriptions, thereby facilitating the understanding and communication within chemistry-related disciplines. By providing an intuitive interface for such translations, MTSwitch aims to reduce the cognitive efforts required to interpret molecular structures or craft SMILES notation from descriptive texts.

3 Evaluation

To evaluate the performance of MTSwitch, we adopted the evaluation metrics introduced by Edwards et al. (2022), which had taken both semantic similarity in natural languages and Fingerprint Tanimoto Similarity (FTS) for molecules into account. ROUGE, BLEU, and METEOR are metrics for evaluating generation quality by measuring n-gram overlap, precision with brevity penalties, and considering synonymy and stemming, respectively. MACCS, RDK, and Morgan are types of molecular fingerprints that represent molecular structures as binary vectors, with MACCS using predefined structural keys, RDK encoding specific molecular features, and Morgan (ECFP) generating circular fingerprints by expanding atom environments. We sampled a subset from the validation set of the L+M-24 dataset and assessed the performance of all models across the two tasks.

Tables 1 and 2 highlight the model performance in molecular captioning and generation, respectively. Meditron performs exceptionally well in most captioning metrics, particularly ROUGE-2 and BLEU-2. While Llama 3 performed well in ROUGE-2 and ROUGE-L, it underperformed in BLEU-4 and Meteor, indicating its potential coherence and lexical diversity issues in specific text generation tasks. For molecular generation, MolT5 and Meditron demonstrated superior performance across key indicators, highlighting their potential for advanced molecular design and synthesis prediction applications. The performances of these two models in our system closely aligned with their originally reported results (Edwards et al., 2024).

4 Conclusion

We introduced MTSwitch, a novel and user-friendly web-based system for bidirectional translation between molecules and texts. Leveraging LLMs, our system demonstrated superior performance, offering users an efficient platform for translating between molecular representations and textual descriptions seamlessly. In the future, we plan to incorporate more models and tasks into the system to make it more comprehensive and effective.

[1] https://huggingface.co/language-plus-molecules

References

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of EMNLP*, pages 375–413.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. *L+M-24*: Building a dataset for language + molecules @ acl 2024. *Preprint*, arXiv:2403.00791.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Chang Liao, Yemin Yu, Yu Mei, and Ying Wei. 2024. From words to molecules: A survey of large language models in chemistry. *Preprint*, arXiv:2402.01439.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yuqi Wang, Zeqiang Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. 2023. Zero-shot medical information retrieval via knowledge graph embedding. In *Proceedings of IOTBDH*, pages 29–40.

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).