Appendix for RTMol: Rethinking Molecule-text Alignment in a Round-trip View

A Related Work

A.1 Molecule-text Modeling

The concept of molecule-text modeling is first brought up by Text2Mol [3], which introduces the ChEBI-20 dataset with pairs of molecule SMILES representations and their textual captions. MolT5 [2] further introduces molecule-caption translation tasks to evaluate the quality of molecule-test alignment. Recent advancements in molecule-text modeling are largely driven by Large Language Models (LLMs) as they demonstrate great potential in language-related applications within scientific domains [6]. LLMs such as ChemDFM [22], ChemLLM [19], and BatGPT-Chem [18] are pre-trained on chemical literature and SMILES data. These models are typically adapted for downstream tasks, such as molecule captioning and text-to-molecule generation, through supervised fine-tuning (SFT) on paired datasets. In-context learning techniques are also leveraged to help LLMs better understand molecules [8, 9]. To overcome the limitations of purely text-based representations, many approaches incorporate richer structural information. MoMu [17], MoleculeSTM [12], and MolCA [14] combines 1D SMILES and 2D graph representations for molecule-to-text generation. 3D-MoLM [10] and 3D-MolT5 [16] leverage the 3D conformation of molecules to help LLM better understand molecules. ChemDFM-X [21] further combines five modalities in the field of chemistry to enhance the ability of molecule understanding. Separately, ether0 [15] leverages reinforcement learning methods to perform chemical reasoning tasks using verifiable rewards. However, the application of reinforcement learning to molecule-text alignment remains less explored. Our work aims to fill this gap by proposing a brand-new round-trip framework to systematically solve this problem.

A.2 Round-trip Strategy

Round-trip process, also known as cycle consistency, is a fundamental concept with broad applications across various scientific and engineering domains. Historically, this principle is employed by human translators for back-translation to verify and improve translation quality [1]. It has since been adapted into machine learning in dual learning for machine translation [7] and to enable end-to-end self-supervised learning [4]. The applicability of cycle consistency also extends into computer vision, where it serves as a powerful constraint in tasks such as semantic alignment [23], unsupervised depth estimation [5], and unpaired image-to-image translation [24]. This principle is also leveraged in chemistry, where a "retrosynthesis-reaction prediction" loop is utilized to evaluate molecular synthesizability and verify retrosynthesis pathways [13, 20]. Beyond these empirical applications, the effectiveness of the round-trip process has been theoretically grounded in the area of density estimation [11].

B Proof on Round-trip Optimization

Assumptions. We make two mild regularity assumptions.

1. **Lipschitz continuity.** Let $\ell_\phi(x \mid y) = \log q_\phi(x \mid y)$. There exists a constant $L > 0$ such that

$$\big|\ell_\phi(x \mid y) - \ell_\phi(x' \mid y)\big| \;\le\; L\, d(x, x') \quad \forall\, x, x', y, \tag{1}$$

   i.e. $\ell_\phi(\cdot \mid y)$ is L–Lipschitz w.r.t. the distance $d(\cdot,\cdot)$ defined in the main text.

2. **Random reconstruction.** For a given y, let $x' \sim q_\phi(x' \mid y)$ be a reconstructed molecule sampled from the Generator.

Inequality derivation. For any x', the Lipschitz condition in Eq. 1 implies

$$\ell_\phi(x \mid y) \;\;=\;\; \log q_\phi(x \mid y) \;\;\ge\;\; \ell_\phi(x' \mid y) - L\, d(x, x'). \tag{2}$$

Taking the expectation over the variable $x' \sim q_\phi(x' \mid y)$ on both sides of Eq. 2 yields

$$\log q_\phi(x \mid y) \;\;\ge\;\; \mathbb{E}_{q_\phi(x'|y)}\big[-L\, d(x, x') + \ell_\phi(x' \mid y)\big]. \tag{3}$$

Setting $\alpha := L$ and the constant $C := \mathbb{E}_{q_\phi(x'|y)}\big[\ell_\phi(x' \mid y)\big]$ (which depends only on y) immediately gives the desired result

$$\log q_\phi(x \mid y) \;\;\ge\;\; \mathbb{E}_{q_\phi(x'|y)}\big[-\alpha\, d(x, x') + C\big]. \tag{4}$$

Interpretation. The inequality states that, under Lipschitz continuity, the log-likelihood assigned by the Generator to any molecule x cannot decay faster than linearly with its molecular distance from *any* sample x' drawn from the same Generator.

C Reinforcement Learning Using Group Relative Policy Optimization

In this work, we use Group Relative Policy Optimization (GRPO) as our optimization strategy.

Given a question x from the dataset, we sample G completions $y_1, \ldots, y_G \sim \pi(\cdot|x)$. Each is assigned a reward $r_1, \ldots, r_G$ and a corresponding advantage:

$$A_i = \frac{r_i - \mathrm{mean}\{r_1, \ldots, r_G\}}{\mathrm{std}\{r_1, \ldots, r_G\}}. \tag{5}$$

Given a single problem x and a group of completions $\{y_i\}$, the per-group objective is:

$$J(\theta, x, y_1, \ldots, y_G) = \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left\{ \text{clip}\left(\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}, A_i, \epsilon \right) - \beta \hat{D}_{\text{KL}}[\pi_{\text{ref}}||\pi_r; x, y_i, z_t] \right\}, \tag{6}$$

where π_θ is the policy being optimized, $\pi_{\theta_{\text{old}}}$ is the policy from which we sampled rollouts, and π_{ref} is a reference policy. clip is the standard PPO clip function:

$$\text{clip}(r, A, \epsilon) = \min\{r \cdot A, \max\{\min\{r, 1+\epsilon\}, 1-\epsilon\} \cdot A\}. \tag{7}$$

D Training Process

The pseudo-code for the training process of RTMol is shown in Algorithm 1.

Algorithm 1 Unified Molecule-Text Round-trip Training

Input: Molecule-text pair dataset $\mathcal{D} = \{(x_i, y_i)\}$; Large language model $\mathcal{L}_\Theta$ with parameters Θ; $\text{Prompt}_{\text{captioner}}$, $\text{Prompt}_{\text{generator}}$; batch size B; step number k; rollout number n

Output: Updated model parameters Θ^*

```
while not converged do
    $\mathcal{G}_\Theta \leftarrow \mathcal{L}_\Theta$ with $\text{Prompt}_{\text{generator}}$
    $\mathcal{C}_\Theta \leftarrow \mathcal{L}_\Theta$ with $\text{Prompt}_{\text{captioner}}$
    for $t = 1$ to $k$ do
        Sample batch $\{(x_j, y_j)\}_{j=1}^{B}$ from $\mathcal{D}$
        for $j = 1$ to $B$ do
            for $s = 1$ to $n$ do
                $x_j'^{(s)} \sim \mathcal{G}_\Theta(y_j)$    // $x_j'^{(s)} \sim q_\phi(x'|y_j)$
                $r_j^{(s)} \leftarrow S(x_j, x_j'^{(s)})$
            end for
        end for
        $\Theta \leftarrow$ Update $\Theta$ using GRPO with rewards $\{r_j^{(s)}\}$
    end for
    $\Theta_0 \leftarrow \Theta$
    for $t = 1$ to $k$ do
        Sample batch $\{x_j\}_{j=1}^{B}$ from $\mathcal{D}$
        for $j = 1$ to $B$ do
            $y_j' \sim \mathcal{C}_\Theta(x_j)$    // $y_j' \sim p_\theta(y|x_j)$
            for $s = 1$ to $n$ do
                $x_j'^{(s)} \sim \mathcal{G}_{\Theta_0}(y_j')$
                $r_j^{(s)} \leftarrow S(x_j, x_j'^{(s)})$
            end for
        end for
        $\Theta \leftarrow$ Update $\Theta$ using GRPO with rewards $\{r_j^{(s)}\}$
        $\Theta_0 \leftarrow \Theta$
    end for
end while
$\Theta^* \leftarrow \Theta$
return $\Theta^*$
```

E Implementation Details

Hyper-parameters used in RTMol during training are shown in Table 1. We implement our framework on top of the VERL library[1] for GRPO training. The experiments are conducted on 8 NVIDIA A800 GPUs, requiring approximately 3, 14, and 50 hours to train ChemT5-0.2B, ChemDFM-8B, and Qwen-3-32B, respectively.

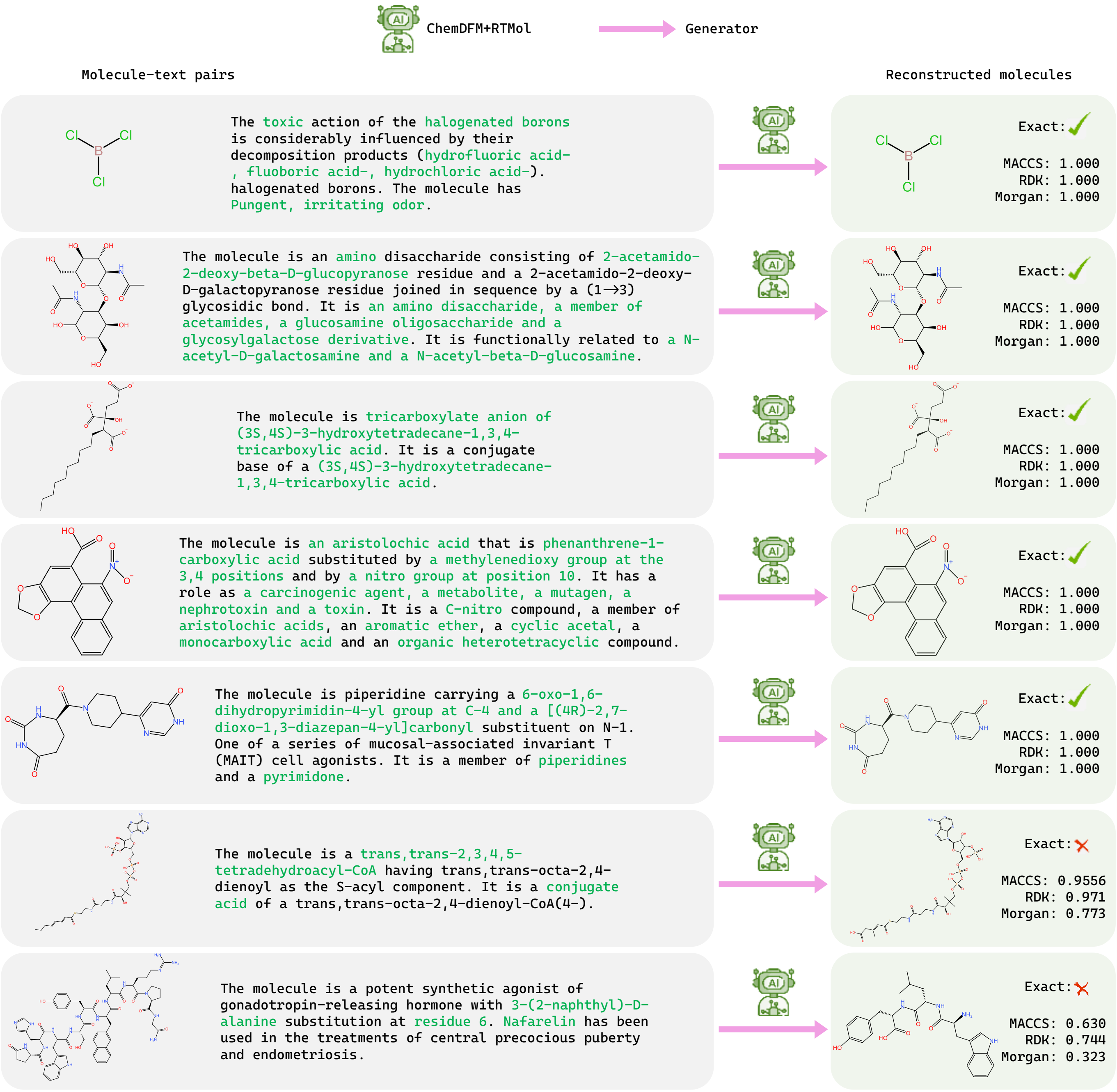
Parameters	Value	Parameters	Value
Train batch size	128	KL loss coefficient	1e-3
PPO mini batch size	64	Learning rate	1e-6
Rollout number	32	Max steps	100

Table 1: Parameters for GRPO training.

F Examples of Generated Molecule Descriptions

Figure 1 demonstrates pairs that ChemDFM+RTMol achieves high round-trip scores.

[1] https://github.com/volcengine/verl



The toxic action of the halogenated borons is considerably influenced by their decomposition products (hydrofluoric acid-, fluoboric acid-, hydrochloric acid-). halogenated borons. The molecule has Pungent, irritating odor.

The molecule is an amino disaccharide consisting of 2-acetamido-2-deoxy-beta-D-glucopyranose residue and a 2-acetamido-2-deoxy-D-galactopyranose residue joined in sequence by a (1→3) glycosidic bond. It is an amino disaccharide, a member of acetamides, a glucosamine oligosaccharide and a glycosylgalactose derivative. It is functionally related to a N-acetyl-D-galactosamine and a N-acetyl-beta-D-glucosamine.

The molecule is tricarboxylate anion of (3S,4S)-3-hydroxytetradecane-1,3,4-tricarboxylic acid. It is a conjugate base of a (3S,4S)-3-hydroxytetradecane-1,3,4-tricarboxylic acid.

The molecule is an aristolochic acid that is phenanthrene-1-carboxylic acid substituted by a methylenedioxy group at the 3,4 positions and by a nitro group at position 10. It has a role as a carcinogenic agent, a metabolite, a mutagen, a nephrotoxin and a toxin. It is a C-nitro compound, a member of aristolochic acids, an aromatic ether, a cyclic acetal, a monocarboxylic acid and an organic heterotetracyclic compound.

The molecule is piperidine carrying a 6-oxo-1,6-dihydropyrimidin-4-yl group at C-4 and a [(4R)-2,7-dioxo-1,3-diazepan-4-yl]carbonyl substituent on N-1. One of a series of mucosal-associated invariant T (MAIT) cell agonists. It is a member of piperidines and a pyrimidone.

The molecule is a trans,trans-2,3,4,5-tetradehydroacyl-CoA having trans,trans-octa-2,4-dienoyl as the S-acyl component. It is a conjugate acid of a trans,trans-octa-2,4-dienoyl-CoA(4-).

The molecule is a potent synthetic agonist of gonadotropin-releasing hormone with 3-(2-naphthyl)-D-alanine substitution at residue 6. Nafarelin has been used used in the treatments of central precocious puberty and endometriosis.

Figure 1: Examples of filtered high-quality molecule-text pairs from the L+M-F and Mol-Instruct-F datasets.

References

[1] Richard W Brislin. Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3):185–216, 1970.

[2] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 375–413. Association for Computational Linguistics (ACL), 2022.

[3] Carl Edwards, ChengXiang Zhai, and Heng Ji. 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, 2021.

[4] Hongchao Fang and Pengtao Xie. An end-to-end contrastive self-supervised learning framework for language understanding. *Transactions of the Association for Computational Linguistics*, 10:1324–1340, 2022.

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[6] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

[7] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in neural information processing systems*, 29, 2016.

[8] Jiatong Li, Wei Liu, Zhihao Ding, Wenqi Fan, Yuqiang Li, and Qing Li. Large language models are in-context molecule learners. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[9] Jiatong Li, Yunqing Liu, Wei Liu, Jingdi Le, Di Zhang, Wenqi Fan, Dongzhan Zhou, Yuqiang Li, and Qing Li. Mol-reflect: Towards in-context fine-grained alignments between molecules and texts. *arXiv preprint arXiv:2411.14721*, 2024.

[10] Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Qiao Liu, Jiaze Xu, Rui Jiang, and Wing Hung Wong. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15):e2101344118, 2021.

[12] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

[13] Songtao Liu, Dandan Zhang, Zhengkai Tu, Hanjun Dai, and Peng Liu. Evaluating molecule synthesizability via retrosynthetic planning and reaction prediction. *arXiv preprint arXiv:2411.08306*, 2024.

[14] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638, 2023.

[15] Siddharth M Narayanan, James D Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G Rodriques, and Andrew D White. Training a scientific reasoning model for chemistry. *arXiv preprint arXiv:2506.17238*, 2025.

[16] Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, and Rui Yan. 3d-molt5: Leveraging discrete structural information for molecule-text modeling. In *International Conference on Learning Representations*, 2024.

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

[18] Yifei Yang, Runhan Shi, Zuchao Li, Shu Jiang, Bao-Liang Lu, Yang Yang, and Hai Zhao. Batgpt-chem: A foundation large model for retrosynthesis prediction. *arXiv preprint arXiv:2408.10285*, 2024.

[19] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

[20] Situo Zhang, Hanqi Li, Lu Chen, Zihan Zhao, Xuanze Lin, Zichen Zhu, Bo Chen, Xin Chen, and Kai Yu. Reasoning-driven retrosynthesis prediction with large language models via reinforcement learning. *arXiv preprint arXiv:2507.17448*, 2025.

[21] Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Yansi Li, Zhongyang Dai, et al. Chemdfm-x: towards large multimodal model for chemistry. *Science China Information Sciences*, 67(12):220109, 2024.

[22] Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Chemdfm: A large language foundation model for chemistry. *arXiv preprint arXiv:2401.14818*, 2024.

[23] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 117–126, 2016.

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.