RUMC: Reward-Guided Monte Carlo Sampling with Uncertainty Awareness for De Novo Molecular Generation

Long Xu[1*], Yongcai Chen[1], Fengshuo Liu[1]
[1]Guangxi Key Lab of Human-Machine Interaction and Intelligent Decision, Nanning Normal University, Nanning, China
[*]Corresponding Author: Long Xu Email: xulong0826@outlook.com

***Abstract*—The design of novel small molecules is crucial for advancing drug discovery, materials science, and chemical biology. Recent advances in deep generative modeling have driven notable progress in molecular generation and property optimization. However, methods such as Monte Carlo sampling still face substantial challenges in thoroughly exploring the vast chemical space and further improving the diversity, validity, and uniqueness of generated molecules. To address these limitations, we introduce RUMC—a reward-guided, uncertainty-aware Monte Carlo sampling framework for de novo molecular generation. RUMC is built upon a transformer-based generative adversarial network and incorporates advanced reinforcement learning strategies. Specifically, RUMC employs a reward-guided sampling mechanism with a deduplicated experience replay buffer to prioritize unique, high-reward molecules and applies robust reward normalization for stable training. Additionally, RUMC utilizes an uncertainty-aware exploration strategy based on Monte Carlo Dropout and the Upper Confidence Bound (UCB) criterion to balance exploration and exploitation during molecular generation. Extensive experiments on the QM9 and ZINC datasets demonstrate that RUMC significantly improves the validity, novelty, and diversity of generated molecules compared to existing methods, while ensuring excellent drug-like properties. Code and data are available at https://github.com/xulong0826/RUMC.**

***Index Terms*—de novo molecular generation; reward-guided; reward buffer; uncertainty sampling; Monte Carlo Dropout**

I. Introduction

The discovery of novel molecules with targeted chemical properties is fundamental to progress in drug design, materials science, and chemical biology [1], [2]. Molecular design traditionally focuses on modifying or optimizing known compounds to achieve desired properties. In contrast, de novo molecular generation, which constructs molecules from scratch rather than relying on existing libraries, enables the exploration of vast chemical spaces and the identification of previously unknown compounds with desirable properties. This approach accelerates the development of new drugs, advanced materials, and functional chemicals by expanding the boundaries of chemical innovation. However, the immense size and complexity of chemical space present significant challenges for efficiently identifying candidates with specific properties, especially in the absence of prior knowledge or heuristic constraints. As the demand for more efficient and innovative molecular generation methods increases, it is crucial to develop approaches that effectively balance novelty, diversity, and property optimization.

Current deep learning-based generative models have greatly advanced molecular discovery [12], [13]. These models are mainly divided into string-based and graph-based approaches. String-based models [10] use SMILES sequences and natural language processing techniques, offering concise representations and efficient training, but often struggle with invalid outputs and limited structural diversity. Graph-based models [3] represent molecules as graphs, capturing richer chemical information and improving property prediction, but require more complex architectures and higher computational cost. Both approaches still have considerable room for improvement, particularly in terms of diversity, sample validity, property optimization, and efficient exploration of chemical space.

To address these limitations, some studies have incorporated Monte Carlo (MC) search into molecular generation frameworks for property optimization [10], [11]. However, in recent studies, MC sampling is generally used to complete molecular sequences and assist in subsequent property prediction, rather than directly optimizing molecular properties. Simple MC sampling typically provides limited improvement in chemical space exploration and does not effectively enhance property optimization. When combined with reinforcement learning (RL), MC sampling can offer intermediate rewards for incomplete molecular sequences by sampling multiple completions, thereby providing more informative feedback during training. Specifically, most existing implementations only partially address critical issues such as mode collapse, limited diversity, and inefficient exploration. These methods often fail to effectively balance exploration and exploitation, normalize reward distributions, and manage experience replay buffers, resulting in limited improvements in molecular generation quality and diversity.

To address these challenges, we propose RUMC, a

reward-guided, uncertainty-aware Monte Carlo sampling framework for de novo molecular generation. RUMC integrates the reward function throughout the generation and sampling process, combining deduplicated experience replay, prioritized high-reward sampling, uncertainty-aware MC Dropout, and reward normalization. Collectively, these innovations address the core challenges of quality, diversity, and efficiency in de novo molecular generation. The main contributions of this work are:

- **Reward-guided sampling and deduplicated replay:** RUMC applies reward signals to select and store molecules. The replay buffer retains only unique, high-reward molecules, and reward normalization stabilizes training. This module directly improves the quality and diversity of generated molecules.
- **Uncertainty-aware MC Dropout and UCB selection:** RUMC employs MC Dropout to estimate the reward mean and variance for each molecule. The Upper Confidence Bound (UCB) score guides the selection of molecules, balancing exploration and exploitation to discover both novel and high-reward candidates.
- **Superior Performance:** Empirical results demonstrate that RUMC outperforms existing methods on benchmark datasets, achieving higher validity, novelty, diversity, and property optimization.

II. Materials and Methods

A. Problem Formulation

The task of de novo molecular generation is to construct a generative model G that transforms a latent variable z, sampled from a prior distribution $p(z)$, into a molecule m within the chemical space $\mathcal{M}$. Here, molecules are encoded as SMILES strings. The aim is to optimize G so that it generates molecules maximizing a reward function $R(m)$, which quantifies target chemical properties such as validity, novelty, and drug-likeness, while ensuring diversity. RUMC addresses this by embedding reward feedback and uncertainty quantification throughout the sampling and training process of a generative adversarial network. The overall framework is illustrated in Figure 1.

B. Adversarial Generation Network

Following previous work [10], RUMC employs a GAN with a transformer-based architecture, comprising a generator and a discriminator.

Generator: In contrast to traditional transformers that use an encoder with a multi-head attention layer to extract features from an input sequence and decode with a decoder, RUMC generates SMILES strings from scratch/noise using a masked transformer encoder for sequence generation. It autoregressively constructs a SMILES string token by token, starting from a beginning-of-sequence ($\langle bos \rangle$) token. At each step, the self-attention mechanism considers the previously generated sequence to produce a probability distribution over the vocabulary for the next token. The causal attention mask prevents the model from attending to future tokens, preserving the sequential nature of generation.

Discriminator: The discriminator, also a transformer encoder, is a binary classifier trained to distinguish real molecules from the training set and synthetic molecules from the generator. Unlike the generator, it processes an entire SMILES string at once, using self-attention layers to capture global chemical context and structural patterns. The output is a single probability score, indicating the model's confidence that the input molecule is authentic.

C. Reward-Guided Sampling and Replay

To guide the generation process toward high-quality molecules, RUMC employs a reward-guided mechanism centered on a deduplicated experience replay buffer (`RewardBuffer`). This buffer serves as a dynamic memory of successful discoveries.

Reward Normalization: For each generated molecule m_i with raw reward r_i, the score is normalized to prevent drastic distributional shifts during training. The normalized reward $\tilde{r}_i$ is calculated as:

$$\tilde{r}_i = \frac{r_i - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}} \tag{1}$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ are the running mean and standard deviation of rewards for all molecules in the buffer. This ensures comparability of rewards across training stages.

Deduplicated High-Reward Buffer: The `RewardBuffer` is implemented as a fixed-size buffer with capacity K. Each time a new molecule m_i with normalized reward $\tilde{r}_i$ is generated, it is added to the buffer only if it is unique (i.e., not already present). If the buffer exceeds its capacity K, the molecule with the lowest reward is removed to maintain buffer quality. This mechanism ensures that the buffer contains a diverse set of high-reward molecules and prevents repeated sampling of suboptimal structures.

Buffer Update and Maintenance: The buffer is updated dynamically during training. Specifically, when a new candidate molecule m_i is generated, if m_i is not already present in the buffer, $(m_i, \tilde{r}_i)$ is added. If the buffer size exceeds the predefined capacity K, the molecule with the lowest normalized reward $\tilde{r}_i$ is removed to maintain buffer quality. If m_i already exists in the buffer, its reward is updated only if the new $\tilde{r}_i$ is higher than the previous value.

Probabilistic Sampling: At each sampling step, with probability p_{buffer}, RUMC samples a molecule directly from the buffer instead of generating a new molecule via MC sampling. The probability of selecting molecule m_i

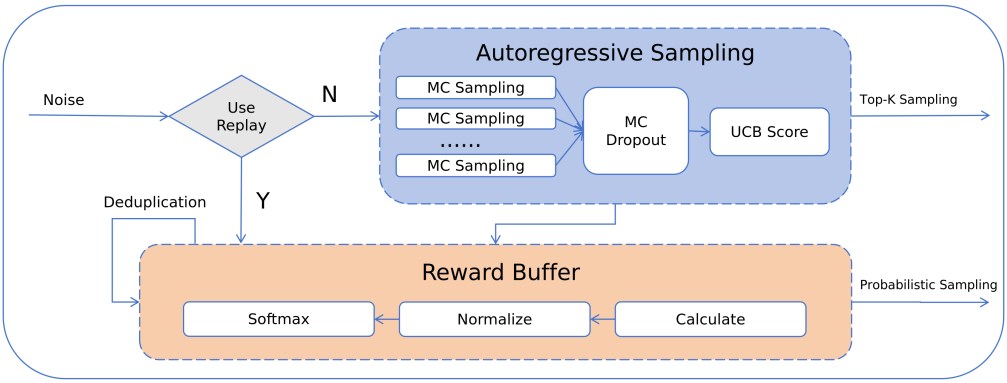

Fig. 1. The overall architecture of the RUMC framework. The generator produces molecules, which are evaluated by a reward function and a discriminator. The reward buffer stores unique, high-reward molecules and normalizes their scores. During sampling, RUMC either performs autoregressive generation guided by MC Dropout and UCB, or directly samples from the high-reward buffer. The feedback loop continuously updates the generator.

from the buffer is proportional to its normalized reward, computed via a softmax function:

$$P(m_i) = \frac{\exp(\tilde{r}_i)}{\sum_{m_j \in \mathcal{B}} \exp(\tilde{r}_j)} \tag{2}$$

where $\mathcal{B}$ denotes the current buffer contents. With probability $1 - p_{\text{buffer}}$, a new molecule is generated using MC sampling. This design allows the model to balance exploitation of high-reward historical samples and exploration of new candidates, accelerating convergence toward desirable regions of chemical space.

D. Uncertainty-Guided Monte Carlo Sampling

To balance exploration of novel chemical structures with exploitation of known high-reward regions, RUMC incorporates an uncertainty-guided sampling strategy. During autoregressive generation, MC Dropout provides a Bayesian approximation of model uncertainty.

By performing N stochastic forward passes with dropout enabled for the same input, a distribution of potential molecules and corresponding reward estimates $\{r_i^{(1)}, r_i^{(2)}, ..., r_i^{(N)}\}$ is obtained without normalization. The Upper Confidence Bound (UCB) score is:

$$\text{UCB}_i = \mathbb{E}[r_i] + \beta \cdot \sqrt{\text{Var}[r_i]} \tag{3}$$

where β is a hyperparameter controlling exploration. High variance indicates model uncertainty, suggesting regions of chemical space with limited knowledge. Prioritizing molecules with high UCB scores encourages selection of candidates with high expected rewards and exploration of uncertain regions, potentially yielding novel and potent structures.

Alternatively, after calculating the UCB scores for all candidate molecules, RUMC can further perform top-k sampling to the UCB scores, ensuring that the final output prioritizes molecules with both high expected rewards and high uncertainty.

E. Training and Loss Function

Following previous work [10], RUMC is trained end-to-end using adversarial loss and policy gradient reinforcement learning, enabling the generator to learn from both discriminator feedback and external property rewards.

The discriminator uses binary cross-entropy loss:

$$\mathcal{L}_D = -\mathbb{E}_{m \sim p_{\text{data}}}[\log D(m)] - \mathbb{E}_{m \sim G}[\log(1 - D(m))] \tag{4}$$

The generator maximizes expected total reward $J(\theta) = \mathbb{E}_{m \sim G_\theta}[R_{\text{total}}(m)]$, where $R_{\text{total}}(m)$ combines discriminator feedback and external property scores. Generator parameters θ are updated via the policy gradient theorem:

$$\nabla_\theta J(\theta) = \mathbb{E}_{m \sim G_\theta}[R_{\text{total}}(m) \nabla_\theta \log G_\theta(m)] \tag{5}$$

This procedure enables robust, property-guided molecular generation by leveraging adversarial and reinforcement learning.

III. Experimental Settings

A. Datasets and Parameter Setting

Dataset. To comprehensively assess RUMC, we employed two established benchmark datasets: QM9 [4] and ZINC [5]. For each dataset, we randomly selected subsets of 5,000 and 10,000 SMILES strings for training. All molecules included up to nine heavy atoms (C, O, N, F), ensuring uniform molecular size and facilitating reliable evaluation of generative performance and property optimization. This selection reflects typical small, drug-like fragments. Additional results for ZINC are presented in the Appendix.

Parameter Setting: For all experiments, the number of Monte Carlo Dropout samples N was set to 16. The experience replay buffer size K was set to 10,000 for QM9 and 30,000 for ZINC. The probability of prioritized sampling from the buffer p_{buffer} was set to 0.2. The exploration coefficient β in the UCB score was set to

TABLE I
PERFORMANCE EVALUATION ON PROPERTY OPTIMIZATIONS COMPARED TO THE CLOSEST RELATED WORK ON QM9 DATASET. HIGHER VALUES ARE BETTER FOR ALL METRICS.

Property	Algorithm	Validity	Uniqueness	Novelty	QED	SA	logP	Diversity
Baseline	LSTM	74.4%	98.3%	96.9%	0.48	0.23	0.29	0.93
	TransEn	89.5%	93.9%	82.7%	0.48	0.24	0.30	0.92
Drug-likeness	Naïve RL	97.0%	59.0%	**100.0%**	0.57	**0.54**	0.47	0.89
	ORGAN	88.1%	65.7%	97.9%	0.55	0.45	0.41	0.86
	TenGAN	97.8%	70.7%	98.0%	0.57	0.50	0.40	**0.90**
	RUMC	**98.2%**	**91.2%**	**100.0%**	**0.62**	0.43	**0.48**	**0.90**
Synthesizability	Naïve RL	97.2%	18.6%	98.2%	0.48	0.75	0.33	0.88
	ORGAN	96.5%	**32.7%**	**99.4%**	0.49	0.71	0.41	0.87
	TenGAN	96.7%	24.2%	97.5%	**0.52**	0.71	**0.50**	**0.90**
	RUMC	**98.5%**	13.4%	97.9%	**0.52**	**0.77**	0.48	**0.90**
Solubility	Naïve RL	92.8%	**63.2%**	**100.0%**	0.44	**0.59**	0.80	0.86
	ORGAN	93.3%	41.3%	99.5%	**0.51**	0.56	0.54	0.89
	TenGAN	97.5%	61.6%	97.6%	**0.51**	0.47	0.55	**0.91**
	RUMC	**98.7%**	49.1%	**100.0%**	**0.51**	0.50	**0.87**	0.90

TABLE II
COMPARISON WITH GRAPH-, VAE-, FLOW-, AND GAN-BASED ALGORITHMS ON THE QM9 DATASET. TOTAL REPRESENTS THE PRODUCT OF VALIDITY, UNIQUENESS, AND NOVELTY.

Algorithm	QED	Validity	Uniqueness	Novelty	Total
JTVAE	0.46	**100.0%**	55.7%	97.1%	54.1%
GraphAF	0.47	37.0%	91.7%	99.6%	33.8%
CharVAE	0.50	17.2%	**99.9%**	94.9%	16.3%
GramVAE	0.48	38.0%	98.8%	93.7%	35.2%
TransVAE	0.52	17.2%	25.2%	97.2%	42.1%
GraphNVP	0.58	83.0%	99.2%	–	–
MoFlow	0.44	95.0%	93.7%	89.0%	79.2%
MolGAN	0.59	99.3%	2.3%	99.7%	2.3%
TenGAN	0.60	97.8%	82.6%	99.8%	80.6%
RUMC	**0.62**	98.2%	91.2%	**99.9%**	**89.5%**

1.0. All experiments were conducted on a workstation equipped with an NVIDIA RTX 5070 Ti GPU and 32 GB RAM.

To evaluate model performance and enable comparison with previous studies, we adopted several widely used metrics [1], [2], [10]: validity (the percentage of chemically valid molecules among all generated samples, checked using RDKit), uniqueness (the proportion of non-repeated molecules among valid samples), novelty (the percentage of valid molecules not present in the training set among all unique molecules), and diversity (the average Tanimoto distance between Morgan fingerprints of any two molecules, radius 4, 2048 bits), drug-likeness (quantified by the QED score), synthesizability (assessed by the synthetic accessibility, SA, score), and solubility (evaluated by the log octanol-water partition coefficient, logP). We also report Total performance, calculated as the product of validity, uniqueness, and novelty percentages.

B. Method Comparison

To comprehensively evaluate the effectiveness of our proposed models, we conducted two sets of comparative experiments. The first set focuses on property optimization methods. As shown in Table I, we compare RUMC with Naïve RL, ORGAN, and TenGAN on the QM9 dataset. These baseline algorithms represent the most relevant prior works for training RNNs to optimize molecular properties. For reference, the LSTM and transformer encoder (TransEn) baselines were trained using maximum likelihood estimation (MLE) without property optimization and their training was stopped immediately after pretraining. The results, discussed in Section III-C, indicate that our proposed models outperform these baselines on most evaluation metrics.

The second set benchmarks RUMC against general molecular generation methods. Specifically, we compared our approach with a range of graph-based algorithms, including JTVAE [7], GraphAF [8], GraphNVP, MoFlow [9], and MolGAN, as well as other SMILES-based models such as CharVAE, GramVAE [6], TransVAE, and TenGAN [10]. This comparison is important because graph-based models generally benefit from richer structural representations, often outperforming SMILES string-based approaches. Meanwhile, discrete GANs on SMILES strings are prone to training instability, which can adversely affect their generative performance.

C. Results and Analysis

Table I summarizes the property optimization results on the QM9 dataset. For all three tasks—drug-likeness, synthesizability, and solubility—RUMC consistently achieves superior performance. In the drug-likeness task, RUMC attains the highest validity among RL and GAN-based methods (98.2%) and leads in QED score (0.62). Notably, RUMC maintains high uniqueness (91.2%) and perfect novelty (100.0%), indicating strong diversity and novelty.

For synthesizability, RUMC achieves the highest SA score (0.77), outperforming Naïve RL and other baselines, while maintaining competitive QED and diversity. The lower uniqueness (13.4%) in this task is expected, as it reflects the limited diversity of easily synthesizable molecules. In the solubility optimization task, RUMC obtains the highest logP score (0.87), again with excellent validity and novelty.

Table II further compares RUMC with various molecular generation architectures. Despite the inherent advantages of graph-based methods, RUMC achieves the highest QED score (0.62) and the best overall balance of validity, uniqueness (91.2%), and novelty (99.9%). Consequently, RUMC also surpasses the next-best model, TenGAN, in the total score (89.5% vs. 80.6%), demonstrating that its reward-guided sampling and uncertainty-aware exploration effectively enhance both diversity and property optimization.

TABLE III
STEPWISE MODULE ADDITION AND PARAMETER SENSITIVITY ANALYSIS OF RUMC ON QM9. BASE REFERS TO THE TRANSFORMER-GAN WITHOUT SPECIALIZED SAMPLING MECHANISMS.

Configuration	QED	Validity	Uniqueness	Novelty	Total
Base	0.61	97.9%	74.9%	99.9%	73.3%
+Buffer	0.61	**98.8%**	77.2%	99.9%	76.2%
+Norm	**0.62**	98.3%	87.1%	99.9%	85.5%
+MC Dropout	**0.62**	98.2%	**91.2%**	**99.9%**	**87.5%**
$p_{\text{buffer}} = 0.1$	**0.62**	**98.9%**	70.0%	99.9%	69.2%
$p_{\text{buffer}} = 0.2$	**0.62**	98.2%	**91.2%**	99.9%	**89.5%**
$p_{\text{buffer}} = 0.3$	**0.62**	97.8%	89.5%	**100.0%**	88.1%

D. Ablation Studies

To validate the contribution of each component, we conducted ablation studies, with the results presented in Table III. The base transformer-GAN achieves reasonable QED (0.61) and validity (97.9%), but its uniqueness is limited (74.9%). Adding the replay buffer (+Buffer) improves validity and slightly increases uniqueness. Subsequently, incorporating reward normalization (+Norm) further boosts uniqueness to 87.1% while maintaining high validity. Finally, the full model with MC Dropout (+MC Dropout) raises uniqueness to 91.2% and the total score to 87.5%, confirming the positive impact of each module.

We also performed a parameter analysis for the probability of prioritized sampling from the buffer, p_{buffer}. The results show that $p_{\text{buffer}} = 0.2$ yields the best balance of validity (98.2%), uniqueness (91.2%), and total score (89.5%). A lower p_{buffer} favors validity but reduces uniqueness, while a higher p_{buffer} increases novelty at the expense of uniqueness. These results highlight the importance of tuning the exploitation-exploration trade-off for optimal performance.

IV. CONCLUSION

In this work, we introduce RUMC, a reward-guided, uncertainty-aware Monte Carlo sampling framework for de novo molecular generation. By integrating reward-guided sampling, deduplicated replay, and uncertainty-aware exploration with robust reward normalization, RUMC addresses the challenges of limited diversity, suboptimal sample quality, and inefficient exploration in molecular design. Our experiments on the QM9 dataset demonstrate that RUMC consistently outperforms reinforcement learning and GAN-based baselines, achieving superior validity, novelty, diversity, and property optimization. Comparisons with state-of-the-art graph-based and generative models further confirm the advantages of RUMC in generating diverse, property-optimized molecules. These findings underscore the potential of combining reward-guided and uncertainty-aware strategies to advance deep generative models in computational chemistry and drug discovery. However, the current framework has not been extensively validated on larger or more complex chemical spaces, which remains an avenue for future work.

REFERENCES

[1] F. Grisoni, M. Moret, R. Lingwood, et al, "Bidirectional molecule generation with recurrent neural networks," Journal of Chemical Information and Modeling, 60(3): 1175-1183, 2020.

[2] C. Li, Y. Yamanishi, "SpotGAN: A reverse-transformer GAN generates scaffold-constrained molecules with property optimization," Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 323-338, Springer, 2023.

[3] X. Xia, J. Hu, Y. Wang, et al, "Graph-based generative models for de Novo drug design," Drug Discovery Today: Technologies, 32: 45-53, 2019.

[4] R. Ramakrishnan, P. O. Dral, M. Rupp, et al, "Quantum chemistry structures and properties of 134 kilo molecules," Scientific Data, 1: 1-7, 2014.

[5] J. J. Irwin, T. Sterling, M. M. Mysinger, et al, "ZINC: a free tool to discover chemistry for biology," Journal of Chemical Information and Modeling, 52(7): 1757-1768, 2012.

[6] M. Simonovsky, N. Komodakis, "GraphVAE: Towards generation of small graphs using variational autoencoders," International Conference on Artificial Neural Networks, 412-422, Springer, 2018.

[7] W. Jin, R. Barzilay, T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," International Conference on Machine Learning, 2323-2332, PMLR, 2018.

[8] C. Shi, M. Xu, Z. Zhu, et al, "GraphAF: a flow-based autoregressive model for molecular graph generation," arXiv:2001.09382, 2020.

[9] C. Zang, F. Wang, "MoFlow: an invertible flow model for generating molecular graphs," Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 617-626, 2020.

[10] C. Li, Y. Yamanishi, "TenGAN: Pure transformer encoders make an efficient discrete GAN for de novo molecular generation," International Conference on Artificial Intelligence and Statistics, 361-369, PMLR, 2024.

[11] Y. Liu, Y. Zhu, J. Wang, et al, "A Multi-Objective Molecular Generation Method Based on Pareto Algorithm and Monte Carlo Tree Search," Advanced Science, 2410640, 2025.

[12] Y. Cheng, Y. Gong, Y. Liu, et al, "Molecular design in drug discovery: a comprehensive review of deep generative models," Briefings in Bioinformatics, 22(6): bbab344, 2021.

[13] X. Zeng, F. Wang, Y. Luo, et al, "Deep generative molecular design reshapes drug discovery," Cell Reports Medicine, 3(12), 2022.