JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

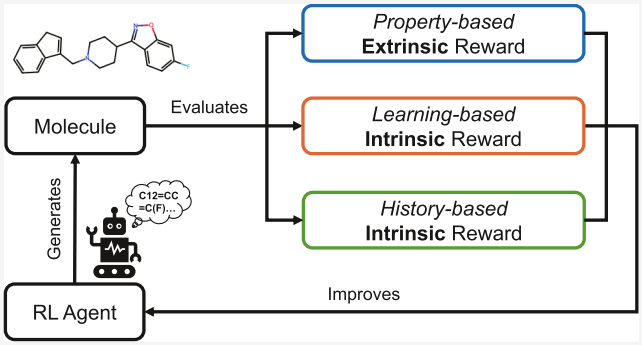
Mol-AIR: Molecular Reinforcement Learning with Adaptive Intrinsic Rewards for Goal-Directed Molecular Generation

Jinyeong Park, Jaegyoon Ahn, Jonghwan Choi,* and Jibum Kim*

Cite This: J. Chem. Inf. Model. 2025, 65, 2283−2296 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Optimizing techniques for discovering molecular structures with desired properties is crucial in artificial intelligence (AI)-based drug discovery. Combining deep generative models with reinforcement learning has emerged as an effective strategy for generating molecules with specific properties. Despite its potential, this approach is ineffective in exploring the vast chemical space and optimizing particular chemical properties. To overcome these limitations, we present Mol-AIR, a reinforcement learning-based framework using adaptive intrinsic rewards for effective goal-directed molecular generation. Mol-AIR leverages the strengths of both history-based and learning-based intrinsic rewards by exploiting random distillation network and counting-based strategies. In benchmark tests, Mol-AIR demonstrates improved performance over existing approaches in generating molecules having the desired properties, including penalized LogP, QED, and celecoxib similarity, without any prior knowledge. We believe that Mol-AIR represents a significant advancement in drug discovery, offering a more efficient path to discovering novel therapeutics.



INTRODUCTION

The development of optimization techniques to efficiently discover molecular structures with target properties is a critical challenge in artificial intelligence (AI)-based drug discovery research. In traditional drug discovery, high-throughput screening (HTS) techniques are employed as a practical method to investigate large numbers of synthetic compounds. They aim to identify hit molecules with desired pharmacological properties, such as activity, toxicity, stability, and binding affinity.[1] However, the HTS approach has limitations in reducing the time and costs of hit discovery.[2] Additionally, the number of drug-like compounds is estimated to be in the range of 10^{33} to 10^{60},[3] making it difficult and time-consuming to discover desired hit and lead molecules against target diseases.

Deep generative models have significantly advanced, and these advancements have been applied to the efficient and effective exploration of molecular structures in drug discovery.[4,5] While HTS involves selecting hits from a known vast chemical library, an approach using deep generative models generates hits directly by creating molecular structures with target properties, a process called goal-directed molecular generation.[6,7] Goal-directed molecular generation faces two key challenges: (1) representing and generating molecular structures using a deep generative model and (2) directing a deep generative model to discover molecules possessing desired chemical properties. To address these challenges, many studies have used the simplified molecular-input line-entry system (SMILES), self-referencing embedded strings (SELFIES),[8,9] and graph-based representation methods[10] in training deep molecular generative models. Researchers have exploited various deep generative models, including recurrent neural networks (RNNs),[11,12] transformers,[13] and graph neural networks,[10] to efficiently handle those string-based or graph-based molecular data.[14,15] Furthermore, Bayesian optimization[10,16] and reinforcement learning (RL) techniques[11,17,18] have been exploited for deep molecular generative models to create molecules with desired chemical properties.

Many studies have demonstrated the effectiveness of molecular structure generation strategies using RL for optimizing various molecular structure properties.[11,12,17−19] The typical RL configuration in AI-based drug design consists of two components: an agent that generates molecular structures and an environment that evaluate the generated molecules. If a generated molecular structure possesses desired target properties, such as a high quantitative estimate of drug-likeness (QED), the agent receives a high reward as feedback from the environment. These property-constrained RL methods are effective in fine-tuning a pretrained molecular generative

Received: September 12, 2024
Revised: February 11, 2025
Accepted: February 12, 2025
Published: February 24, 2025

model so that it can generate a large number of hit molecules.[11,12,17−19] However, owing to the vast size of the chemical space, it is challenging for an agent to perform efficient exploration, resulting in failure to find an optimal policy or state-value function for the desired generation of molecular structures.[15,20,21]

To enhance the exploration capability of property-constrained RL, curiosity strategies have been recently proposed.[21,22] In the curiosity-based RL scheme, there are two types of rewards: extrinsic and intrinsic rewards. An extrinsic reward, aimed at improving target chemical properties, is calculated based on evaluated property scores, whereas an intrinsic reward is used to enhance exploration ability and find new molecular structures without relying on target properties. Intrinsic rewards encourage the learning of diverse molecular structures and the discovery of hits with higher target properties. Thiede et al.[21] demonstrated the effectiveness of using intrinsic rewards to optimize chemical properties, such as penalized LogP (*p*LogP) and QED. However, existing intrinsic reward methods are not effective in tasks involving the generation of compounds structurally similar to specific drugs (e.g., celecoxib). Furthermore, as intrinsic reward is calculated by predefined algorithms, it is necessary to heuristically adjust those calculation algorithms to apply it for various chemical property optimizations.

In this study, we propose a new RL-based framework using a novel intrinsic reward method that can improve the exploration ability of RL for various chemical properties. To the best of our knowledge, the proposed framework is the first molecular optimization framework utilizing a combination of two types of intrinsic rewards based on a random distillation network (RND)[23] and counting-based strategies. Compared to existing intrinsic reward functions, our method demonstrated improved performance in goal-directed molecular generation without any prior knowledge for six chemical properties, including *p*LogP, QED, and drug similarity. The proposed framework significantly outperformed the existing approach in tasks related to identifying hit molecules with structural similarities. Furthermore, we investigated the effectiveness of the proposed framework through an ablation study, hyperparameter analysis, and a proof-of-concept (POC) experiment.

PRELIMINARIES

Molecular Structure Representation. The selection of molecular structure representation methods is crucial in the field of AI-driven drug discovery.[24]

The fundamental elements of representing chemical structures are the atom identities and the molecular connectivity, which are clearly described by both string-based and graph-based representation methods. In graph-based representations, atoms are denoted as nodes and their connections as edges. This method excellently captures the overall structure of compounds, yet it faces the challenge of lacking a standardized convention for configuring node and edge features. Moreover, the choice of a molecular graph representation often hinges on the specific graph traversal algorithm used, rendering the selection of representation a task-specific consideration.[24]

SMILES and SELFIES are widely recognized as standard molecular representation methods in string-based molecular generation tasks.[8,9] Both SMILES and SELFIES represent atomic information through alphabet characters and bond information using symbols such as "-", "=", and "#". The key difference between these two string-representation methods is in their handling of branching and ring structures. While SMILES uses parentheses to denote branches and numeric notations for ring closures, SELFIES utilizes specialized symbols like [Branch] and [Ring] for these purposes.

Although SELFIES was originally introduced to address the syntactical limitations of SMILES and has been widely adopted in recent studies, recent research suggests that SELFIES does not always outperform SMILES-based methods in terms of optimization ability or sample efficiency.[25] While SELFIES tokens may seem flexible, they still have syntax constraints that can restrict exploration if not properly managed. Therefore, the choice between SMILES and SELFIES depends on the specific task and implementation details. In our work, we utilize SELFIES rather than SMILES due to its straightforward rule-based handling and error-correction features,[8,26] allowing us to focus on developing methods to improve exploration capability.

Reinforcement Learning. In RL for goal-directed molecular generation, a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$ is defined as follows: $\mathcal{S}$ is a set of states representing SELFIES strings, $\mathcal{A}$ is a set of actions representing SELFIES characters, $p(s_{t+1}|s_t, a_t)$ is a state transition probability distribution for a next SELFIES character, $r(s_t, a_t)$ denotes the reward function which provides a scalar reward r_t when action a_t is executed in state s_t, and $\gamma \in (0, 1]$ is a scalar discount factor.

RL aims to maximize the expected sum of discounted rewards, which is formalized as the objective function $J(\pi) = \mathbb{E}[\sum_{k=t}^{T} \gamma^{k-t} r_k]$, where $\pi(a_t|s_t)$ represents the policy, which is the probability of taking action a_t given state s_t. To optimize the RL objective, value-based methods learn a state or action-value function, such as $Q(s, a)$, and derive a policy from it. In contrast, policy-based methods directly learn a policy π_θ, parametrized by a vector θ, using techniques like the policy gradient.

Proximal Policy Optimization (PPO),[27] a notable policy gradient algorithm, enhances training stability by implementing constraints on policy modifications. PPO seeks to facilitate stable learning by maintaining updates within a designated trust region, consequently defining the surrogate objective function J^{CLIP}, which is optimized to iteratively update the policy π_θ:

$$J^{\mathrm{CLIP}}(\pi_\theta) = \mathbb{E}[\min(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}\widehat{A}_t,\ g(\epsilon, \widehat{A}_t))] \tag{1}$$

where θ_{old} is the vector of policy parameters before the update, $\hat{A}_t$ is an estimated advantage value at time t, and g is a clipping function defined by

$$g(\epsilon, A) = \begin{cases} (1+\epsilon)A & A \geq 0 \\ (1-\epsilon)A & A < 0 \end{cases} \tag{2}$$

where ϵ is the clipping parameter of PPO and A is an advantage.

RELATED WORKS

Reinforcement Learning for Molecular Generation. Numerous studies have tackled the RL problem by defining an action space, a state, a policy, and an environment. The action space composed of symbol sets that represent molecular structures, a state space made up of symbol substrings, a policy for predicting the next appropriate symbol (action) to append to the current substring (state) up to a certain length, and an environment that evaluates the completed string, providing rewards based on its properties.[11,12,17,19] Policies employ deep neural network models, such as RNNs, to deal with string-based

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

molecular structures. Rewards are allocated proportionally to the molecular structure's target chemical properties or pharmacological efficacy metrics. Policy networks are updated using various policy gradient algorithms, including REINFORCE[11,12,17] and PPO.[21] The generation of a molecular structure with the desired properties yields a high reward, incentivizing the policy network to generate more molecular structures associated with higher rewards.

Intrinsic Rewards for Molecular Generation. The exploration−exploitation trade-off has been a challenge in RL. The agent is tasked with finding an optimal balance between leveraging its accumulated experiences to seek the best policy (exploitation) and probing various episodes to uncover a potentially superior policy (exploration). To mitigate this challenge, simple methods such as epsilon-greedy policy and entropy regularization have been commonly used. However, these methods struggle to promote exploration when the state or action spaces are too large or in environments with sparse rewards.[28] Recently, intrinsic reward methods, which generate rewards internally by modeling human curiosity, have been introduced and considered to be one of the most effective techniques for promoting exploration.[21] Many studies have shown that these methods can handle difficult environments featuring sparse extrinsic rewards, such as video games.[23,29−34]

Exploring diverse molecular structures is a crucial task in the design of chemical compounds, but the immense search space often hinders the effectiveness of RL models. Employing intrinsic reward methods has proven to be a powerful approach to overcome this challenge. For instance, Thiede et al.[21] introduced three distinct types of intrinsic reward methods: count-based, memory-based, and prediction-based intrinsic reward methods. They demonstrated how intrinsic rewards can help the agent to explore diverse molecular structures efficiently in the vast chemical space. Similarly, Blaschke et al.[35] proposed a memory-based approach that employs diversity filters to promote exploration by comparing molecular scaffolds. This method groups compounds with the same scaffold into buckets and penalizes overfilled buckets, thereby encouraging structural novelty.

Count-Based Intrinsic Reward. Count-based intrinsic reward methods are computed by counting how often the agent visits each state. These methods are used in tabular settings[30] as well as more complex models, including context-tree switching density models,[31] pseudocounting,[32] and locality-sensitive hashing (LSH) techniques.[33] They have the advantage of being simple and easy to implement. However, they may be infeasible for problems with vast sizes, such as those involving chemical space.[36]

Thiede et al.[21] implemented the method that computes intrinsic reward values based on the frequency with which molecular structures are observed during training. To achieve this, they use Morgan's fingerprints[37] to convert molecular structures into numerical vectors and LSH for efficient tracking of the occurrence frequencies of these structures. The intrinsic reward for encountering a molecular structure m is calculated as follows:

$$r_{\text{count}}(m) = \frac{1}{\sqrt{\Gamma(\text{LSH}(\text{MF}(m)))} + \epsilon} \tag{3}$$

where MF represents the Morgan fingerprint function, Γ represents a function that records the occurrence frequency of hash values derived from MF, and ϵ is a small positive constant introduced to prevent division by zero. This formulation guarantees that the intrinsic reward for a specific molecule decreases as the molecule is encountered more often, thereby promoting the exploration of new and less frequently observed molecular structures.

Memory-Based Intrinsic Reward. Memory-based intrinsic reward methods involve maintaining a record of previously encountered states in memory. These methods promote exploration; they encourage the discovery of novel states through assessing the novelty of the current state in comparison to stored memories. The higher the difference from stored memories, the higher the reward.[29] However, memory-based methods require the current state to be compared with previous states, making them resource-intensive and requiring extensive memory storage.

In Thiede et al.,[21] a memory data structure was implemented to track molecular structures generated throughout the training process. This method penalizes the agent for generating frequently observed molecular structures by assigning negative intrinsic rewards based on their similarity to stored structures. The memory is organized as a fixed-size First-In-First-Out buffer to maintain computational feasibility. The degree of similarity between the current molecular structure and those in memory is quantified using the Tanimoto similarity coefficient, and the intrinsic reward for a molecule is calculated as follows:

$$r_{\text{memory}}(m) = -\max_{q \in Q}\{\text{TS}(\text{MF}(m), \text{MF}(q))\} \tag{4}$$

where Q denotes the fixed-sized memory and TS represents the Tanimoto similarity function. This method discourages the agent from repeatedly rediscovering similar molecular structures by reducing rewards for such behavior.

Blaschke et al.[35] proposed another memory-based approach using diversity filters to enhance exploration. In this method, molecular scaffolds and their associated compounds are organized into buckets, and compounds with scores above a threshold for the multiparameter objective score are added. Buckets have limited capacity, and once full, additional compounds sharing the same scaffold are penalized with a score of zero. This strategy prevents the agent from generating redundant compounds and encourages it to explore alternative scaffolds.

Prediction-Based Intrinsic Reward. Prediction-based intrinsic reward methods use a neural network model to learn previously visited states and assign high rewards for unvisited states. These methods often incorporate a forward dynamics model to predict subsequent states from the current state and the current action.[23,34] As the predictive model learns, it exhibits lower prediction errors on the states it has already visited compared to the states it has not. Prediction-based methods enable efficient exploration by defining intrinsic rewards based on the prediction errors.

Thiede et al.[21] implemented a method that calculates intrinsic rewards using prediction errors with respect to molecular properties instead of next-state prediction. This strategy is formalized as follows:

$$r_{\text{prediction}}(m) = \left\| \hat{\varphi}(m) - \varphi(m) \right\|_l \tag{5}$$

where φ represents a property oracle, $\hat{\varphi}$ denotes a neural network that approximates this oracle's predictions, and $\|\cdot\|_l$ signifies either the L1 or L2 norm. This method aims to directly align the exploration process with the discovery of molecular structures exhibiting desirable properties by emphasizing the

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

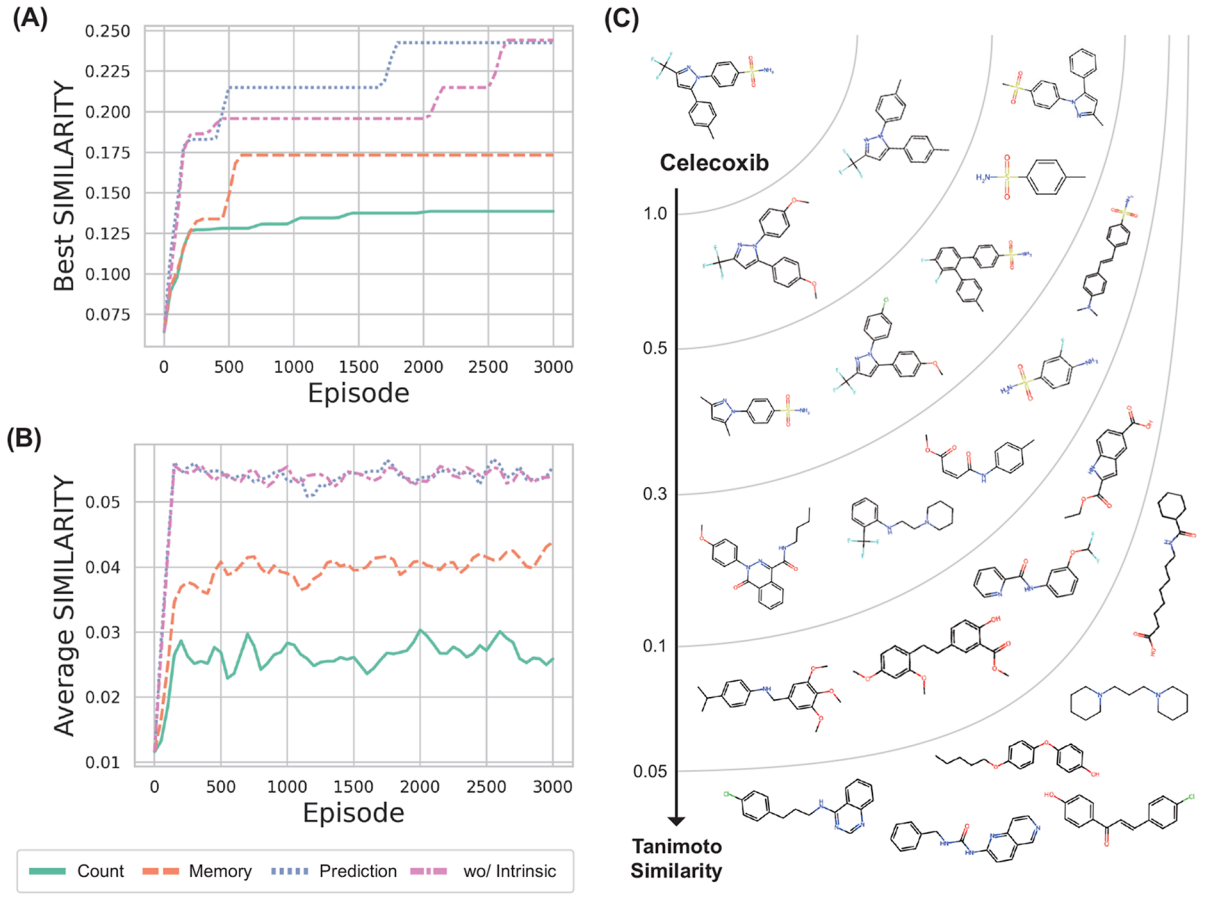


Figure 1. Comparison of traditional intrinsic reward methods for generating celecoxib-like molecular structures. (A) best similarity scores, (B) average similarity scores across training episodes, and (C) examples of celecoxib-like molecular structures at various Tanimoto similarity scores. The similarity scores are calculated using the Tanimoto similarity metric based on molecular fingerprints.

discrepancy between predicted and actual molecular characteristics.

These methods encourage the agent to explore molecular structures that are less predictable, effectively enhancing exploration in the chemical space. However, their methods have certain limitations, which will be presented in the Limitations of Traditional Approaches section. To address these limitations, we combine the advantages of count-based and prediction-based intrinsic methods for efficient exploration in a vast chemical space.

LIMITATIONS OF TRADITIONAL APPROACHES

Thiede et al.[21] proposed three new types of intrinsic reward for enhancing goal-directed molecular generation: count-based, memory-based, and prediction-based rewards. Through evaluation with benchmarks, including *p*LogP, QED, and celecoxib-similarity, their efficacy was examined. Although these benchmark tasks are relatively simple, they provide a standard framework for comparing the performance of our method and are widely used as assessment tools for evaluating AI-based molecular design.[6,38−40] However, the proposed methods still face challenges in discovering new and unknown molecules. This is because solutions occasionally becoming trapped in local optima, hindering efficient exploration within vast chemical spaces.

The count-based and memory-based methods, which are considered history-based approaches, use crafted storage mechanisms to store the information on previously visited states to calculate intrinsic rewards. However, they require meticulous attention to the design of state history management. Moreover, these methods often result in an imbalance between exploration and exploitation, and ultimately lead to excessive exploration by prioritizing the exploration of unvisited states.

The prediction-based method, a learning-based approach, employs a deep neural network to extract meaningful features from past states. It leverages the neural network's ability to learn and remember states without the need for manually designed storage solutions, while promoting more efficient exploration. However, this approach cannot effectively encourage exploration as RL training progresses, particularly when the agent needs to escape out of local optima. Specifically, when the neural network exhibits unexpectedly high generalization performance, it may predict low errors even for unvisited states and hinder the agent from exploring the space further.

History-Based Approach. The history-based approach leverages a predefined storage mechanism to store information on previously visited states. This method prevents revisiting stored states and similar states by utilizing stored state information, thereby encouraging exploration of novel states that are not recorded in the history. History-based approaches employ mechanisms such as hashing functions, scaffold comparisons, Tanimoto similarity, and other techniques to store state information and calculate intrinsic rewards.[21,35] This approach is effective in tasks where continuous exploration of novel molecular structures is crucial. However, it is not suitable

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

for tasks that require finding molecules with structural similarities to a specific molecule.

The main limitation of the existing history-based approaches becomes evident when they are used to discover molecular structures that are similar to specific drugs.[21] Figure 1 shows the structures similar to the drug celecoxib that are discovered across four scenarios: using three types of intrinsic rewards and not using any intrinsic reward.

Figure 1A presents the Tanimoto similarity scores of the molecular structures most closely matched to celecoxib that were identified through a previously proposed RL-based framework by Thiede et al.[21] Figure 1B depicts the average similarity across the discovered molecular structures. The history-based approach, including the count-based and memory-based methods, failed to consistently generate celecoxib-like structures with high Tanimoto similarity. This is primarily because these methods force the agent to find molecular structures that are significantly different from celecoxib, even after identifying similar ones. Therefore, it leads to a lower average similarity compared to the case without intrinsic rewards.

Learning-Based Approach. Learning-based approaches, such as the prediction-based intrinsic reward method, exploit machine learning models to effectively learn previously visited states and calculate intrinsic rewards.[41] Different from history-based approaches, where the state information storage mechanism is predefined, learning-based approaches can dynamically learn and improve this mechanism as the exploration of the agent continues. This feature makes them a versatile tool for goal-directed molecular generation with various target chemical properties. However, the learning-based approach also has limitations when it is solely used. While its high generalizability allows for efficient exploration of an agent, it can also lead to misclassifications where unvisited states are mistakenly identified as visited. This can make it challenging to encourage effective exploration as RL training progresses and can result in the agent getting stuck in local optima.

The limitations of the learning-based approach can be observed when it is solely used for *p*Log*P* optimization. Figure 2 shows the results of molecular optimization for identifying molecules with high *p*Log*P* across four scenarios: using three distinct types of intrinsic rewards and without using any intrinsic reward. Figure 2A shows the highest *p*Log*P* values achieved by various intrinsic reward methods during training.

Two history-based approaches utilizing count-based and memory-based intrinsic rewards, respectively, significantly outperform the case without any intrinsic rewards, achieving consistently higher *p*Log*P* scores. However, the learning-based approach utilizing prediction-based intrinsic reward failed to encourage further exploration and discovery of high *p*Log*P* molecules. Figure 2B shows the average *p*Log*P* scores during a training process. The fluctuations of the two history-based approaches indicate that they successfully explored diverse molecule structures and finally discovered high *p*Log*P* molecules. On the contrary, the plot for the prediction-based method shows minimal fluctuations, indicating that the agent struggles to explore new molecular structures. This is because it has difficulty in performing further exploration when it tries to escape from local optima.

Relying solely on either history-based or learning-based approaches for drug design results in inefficient exploration of vast chemical spaces. This observation motivates the development of a hybrid approach where the two approaches are synergistically combined to calculate adaptive intrinsic rewards. Our goal is to overcome the current limitations of both history-based and learning-based approaches and finally create a more robust framework that efficiently explores the vast chemical space and optimizes molecular structures.

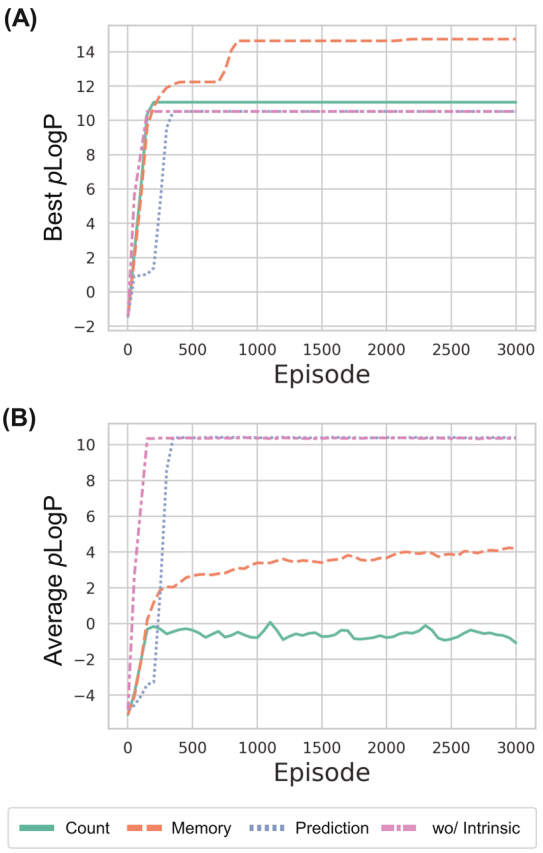


Figure 2. Comparison of traditional intrinsic reward functions in *p*Log*P* optimization. (A) Best *p*Log*P* scores and (B) average *p*Log*P* scores over training episodes.

METHODS

In this study, we introduce Mol-AIR, a molecular optimization framework with adaptive intrinsic rewards that performs efficient exploration for effective goal-directed molecular generation. Mol-AIR integrates the strengths of both history-based and learning-based intrinsic approaches for efficient exploration. This approach synergizes the history-based intrinsic reward (HIR) with the learning-based intrinsic reward (LIR) to achieve efficient exploration of the molecular structure state space to identify target properties.

HIR facilitates the exploration of the chemical space by counting the number of visits to each state. This module encourages the discovery of new molecular structures by prioritizing less visited states. Concurrently, LIR adjusts the balance between exploration and exploitation through the implementation of an RND method.[23] We introduce the RND for efficient exploration in the vast chemical space, as it is known to be effective for encouraging exploration in sparse reward environments in video games using intrinsic rewards. RND computes intrinsic rewards using the difference of outputs between two neural networks, which enables more efficient exploration of the RL agent. By combining HIR and LIR, Mol-AIR provides a powerful framework for navigating the vast and

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

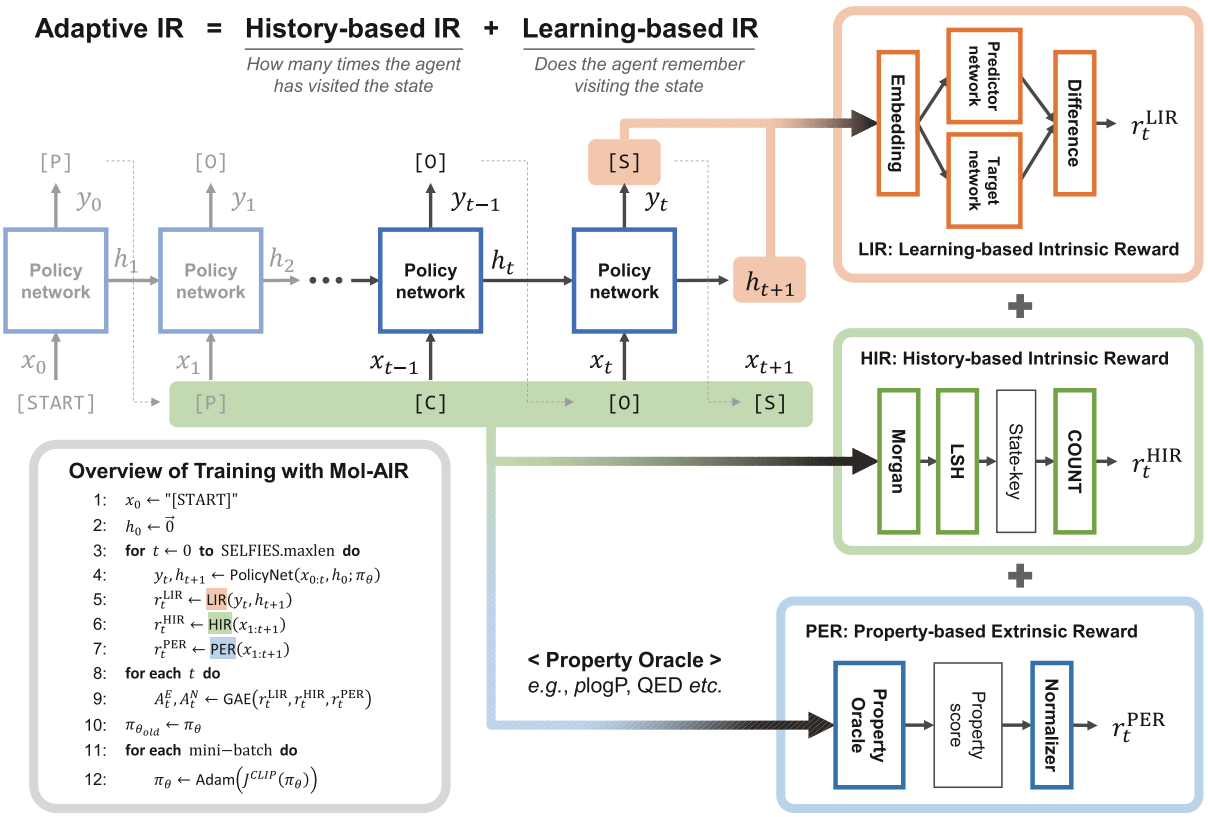


Figure 3. Overview of Mol-AIR.

complex landscape of molecular structures by providing intrinsic rewards. This enables the efficient identification of molecules with desired chemical properties.

Figure 3 illustrates the process of calculating two intrinsic rewards and one extrinsic reward in the training of RL-based models with Mol-AIR. The descriptions of the RL environment are provided in the Supporting Information. At each time step, the agent selects an action, a SELFIES character, using a policy network. After the SELFIES character is added to the SELFIES string constructed so far, extrinsic reward based on target properties, count-based intrinsic reward and RND-based intrinsic reward are calculated. The three rewards are then used to update the policy network through the PPO algorithm, so that the agent not only exploits the current best knowledge of the molecular structure but also explores the chemical space to find better molecular structures. The specific details are given in the subsections that follow.

HIR with Exponential Decay over Visit Counts. In this study, we propose a new count-based intrinsic reward method to efficiently record information about previously visited states. The proposed method tracks the number of times each molecule is visited and also encourages exploration of molecule structures that have not been seen before or have been encountered rarely by assigning high intrinsic rewards to those states. We utilize Morgan fingerprints[37] to capture molecular information and the LSH algorithm to facilitate efficient record-keeping of previously visited structures. Specifically, at time step t, the value of HIR, denoted as r_t^{HIR}, is calculated as follows:

$$r_t^{\mathrm{HIR}} := \exp[-\min\{\tau, \Gamma(\mathrm{LSH}(\mathrm{MF}(x_{1:t+1})))\}] \quad (6)$$

where $x_{1:t+1}$ represents a molecular structure consisting of $t+1$ of SELFIES characters and τ is an upper bound of the frequency of occurrence. In this study, τ was set to 10.

LIR with Random Network Distillation. In this study, we employ an RND-based approach to leverage the advantages of adaptability and flexibility inherent in learning-based approaches. This method involves learning the information on visited states using two neural networks (Supplementary Figure S1) and calculating the intrinsic reward based on the difference in predictions between these two networks. The RND-based method randomly initializes two neural networks with similar architectures and trains only one network to make its outputs identical to those of the other network when given state information as an input. The network being trained is referred to as the predictor network, while the network that retains its initial state is called the target network. Specifically, the predictor network is trained to minimize the error between the outputs of the predictor and target networks. As the training progresses, errors given frequently visited states decrease. The intrinsic reward based on RND at time t, r_t^{LIR}, is also similarly defined by the error form and is as follows:

$$r_t^{\mathrm{LIR}} := \frac{1}{2\sigma^{\mathrm{R}}}\left\|\mathcal{F}(y_t, h_{t+1}) - \mathcal{F}_\phi(y_t, h_{t+1})\right\|^2 \quad (7)$$

where $\mathcal{F}$ represents the target network, $\mathcal{F}_\phi$ represents the predictor network, y_t is the selected SELFIES character at time step t, h_{t+1} is the next hidden state of the policy network, and σ^{R} is the running standard deviation of the RND-based intrinsic returns, which is used to reduce the scale difference of intrinsic rewards over time. In this study, we designed a RNN-based policy network and defined combination of y_t and h_{t+1} as the state value of RL. The intrinsic reward has the error form so that

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

it encourages the agent to more explore unvisited states whose error is greater than frequently visited ones.

PER with Objective Property Oracle. In molecular structure generation tasks, the goal of reinforcement learning is to find molecular structures with superior targeted chemical properties. Once the policy selects an action to append a SELFIES character to the SELFIES string constructed so far, the environment evaluates the chemical properties of the current molecular structure and produces an extrinsic reward based on the evaluation score to update the policy, thereby achieving the goal. Specifically, given an oracle $\Lambda^{(p)}$ that can evaluate the target chemical property p, the value of PER at time step t, r_t^{PER} is calculated as follows:

$$r_t^{\mathrm{PER}} := \Lambda^{(p)}(x_{1:t+1}) - \Lambda^{(p)}(x_{1:t}) \tag{8}$$

As SELFIES substrings can always be converted to structurally valid SMILES strings,[8] in this study, we calculated the extrinsic reward at every time step and reflected it in the policy update.

Two Types of Advantage Estimation. To effectively learn two intrinsic rewards and one extrinsic reward, this study employs an actor-critic structure (Supplementary Figure S2). Following the approach proposed by Burda et al.,[23] we use two critic networks to estimate the values of episodic and nonepisodic advantages at each time step. The episodic advantage encourages exploration and exploitation progress within episodes, while the nonepisodic advantage encourages exploration throughout the entire learning process.

The episodic critic network learns both the property-based extrinsic reward and the count-based intrinsic reward and calculates the episodic state-value $V_\omega^{\mathrm{E}}(x_t, h_t)$ at time step t. The episodic advantage A_t^{E} at time step t is then calculated using the generalized advantage estimation (GAE)[42] as follows:

$$A_t^{\mathrm{E}} = \sum_{l=t}^{L} (\gamma\lambda)^{l-t}(r_l^{\mathrm{PER}} + \alpha\beta r_l^{\mathrm{HIR}} + \gamma V_{l+1}^{\mathrm{E}} - V_l^{\mathrm{E}}) \tag{9}$$

where L is the maximum length of SELFIES, γ is a discount factor, λ is a bias-variance trade-off parameter, α is a weight of intrinsic rewards, and β is a parameter for balancing between HIR and LIR.

The nonepisodic critic network learns the RND-based intrinsic reward and calculates the nonepisodic state-value $V_\psi^{\mathrm{N}}(x_t, h_t)$ at time step t. The nonepisodic advantage A_t^{N} at time step t is calculated using the GAE as follows:

$$A_t^{\mathrm{N}} = \sum_{l=t}^{\infty} (\gamma\lambda)^{l-t}(\alpha r_l^{\mathrm{LIR}} + \gamma V_{l+1}^{\mathrm{N}} - V_l^{\mathrm{N}}) \tag{10}$$

Policy Gradient with AIR and PER. The actor network π_θ, corresponding to the policy, is trained using the sum of two advantage values $A_t^{\mathrm{E}} + A_t^{\mathrm{N}}$ and the PPO-Clip algorithm. Specifically, for a given current SELFIES character x_t, the previous hidden state h_t, and the next SELFIES character y_t, an objective function $J^{\mathrm{CLIP}}(\pi_\theta)$ is calculated, and the actor network is updated using the policy gradient algorithm. Algorithm 1 illustrates the training algorithm of the proposed model and some model parameters are described in Supplementary Table S1.

RESULTS

Implementation Details. The Mol-AIR methodology was implemented in Python 3.7, taking advantage of a suite of open-source tools to facilitate the development and evaluation of the molecular generative model. The key libraries and frameworks used in this implementation included PyTorch 1.11.0 for deep learning models, CUDA 11.3 to harness GPU acceleration, RDKit 2022.9.5 and SELFIES 0.2.4 for cheminformatics support and robust molecular string representation, respectively, and PyTDC 0.4.0 for assessing the chemical properties of generated molecules. The computational experiments were conducted on an Ubuntu 20.04.6 LTS system equipped with 251 GiB of memory and a NVIDIA RTX A6000 GPUs.

Algorithm 1 Training algorithm with Mol-AIR

Require: A family of sets of initial weights for actor, two critic and RND networks $\{\theta, \omega, \psi, \phi\}$, the number of episodes N, the number of environments M, and the maximum length of SELFIES L
Ensure: A trained policy π_θ
for $n = 1, \cdots, N$ **do**
  Initialize a replay buffer $\mathcal{B}$
  for $m = 1, \cdots, M$ **do**
    $x_0 \leftarrow$ "[START]" ▷ Set an initial token
    $h_0 \leftarrow \vec{0}$ ▷ Set an initial hidden state
    for $t = 0, \cdots, L-1$ **do**
      $p_t, h_{t+1} \leftarrow \pi_\theta(x_t, h_t)$
      $y_t \sim p_t(y)$ ▷ Sample the next character
      compute the value of HIR r_t^{HIR} ▷ Eq. (6)
      compute the value of LIR r_t^{LIR} ▷ Eq. (7)
      compute the value of PER r_t^{PER} ▷ Eq. (8)
      if $y_t =$ "[END]" **then**
        break
      end if
    end for
    for each time step t **do**
      compute episodic advantage A_t^E ▷ Eq. (9)
      compute non-episodic advantage A_t^N
    end for
    store all of experiences into the buffer $\mathcal{B}$
  end for
  $\pi_{\theta_{old}} \leftarrow \pi_\theta$ ▷ Keep the old policy for Eq. (1)
  for each mini-batch sampled from $\mathcal{B}$ **do**
    optimize θ wrt PPO loss using Adam
    optimize ω and ψ wrt critic loss using Adam
    optimize ϕ wrt RND loss using Adam
  end for
end for

Benchmark Tasks. To demonstrate the effectiveness of the proposed intrinsic reward method, we set six target properties as benchmarks and searched for the optimal molecular structure for each property. Among the six target properties used in this benchmark test, three (*p*LogP, QED, similarity to the molecule celecoxib) were adopted from previous studies,[6,21] and the remaining three (GSK3B, JNK3, GSK3B+JNK3) are widely used molecular properties in goal-directed molecular generation studies.[43,44] The descriptions of their properties are as follows:

- *p*LogP: The penalized logarithm of the octanol−water partition coefficient simultaneously assesses both the hydrophobicity of a molecule and its chemical feasibility.[16] The range of the *p*LogP score lies between $(-\infty, \infty)$, and the goal of *p*LogP optimization in this study is to find a molecular structure with a high *p*LogP score. Based on the previous study,[21] we normalized *p*LogP scores by multiplying them by 0.1.
- QED: The QED measures the likelihood that a molecule can be used as a drug, considering eight physicochemical properties.[45] The range of the QED score is on the half-open interval $[0, 1)$, and the goal of the QED optimization task is to find a molecule with a high QED score.
- Similarity: This task was designed to evaluate the rediscovery performance of de novo molecular design methods in the Guacamol benchmark.[6] The goal of this similarity task is to find a molecule that is similar to

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

Table 1. Results of Molecular Discovery with the Best Property Scores

intrinsic	best property score					
type	*p*LogP	QED	similarity	GSK3B	JNK3	GSK3B+JNK3
wo/Intrinsic	10.523 ± 0.000	0.898 ± 0.014	0.223 ± 0.019	0.433 ± 0.051	0.240 ± 0.000	0.282 ± 0.003
Count	10.689 ± 0.390	0.869 ± 0.001	0.136 ± 0.003	0.371 ± 0.031	0.233 ± 0.023	0.240 ± 0.009
Memory	13.951 ± 0.891	0.921 ± 0.014	0.174 ± 0.002	0.433 ± 0.031	0.230 ± 0.010	0.263 ± 0.019
Prediction	10.523 ± 0.000	0.897 ± 0.023	0.221 ± 0.019	0.462 ± 0.032	0.223 ± 0.025	0.288 ± 0.022
Mol-AIR	**15.682 ± 0.179**	**0.948 ± 0.001**	**0.308 ± 0.024**	**0.707 ± 0.005**	**0.493 ± 0.070**	**0.401 ± 0.040**

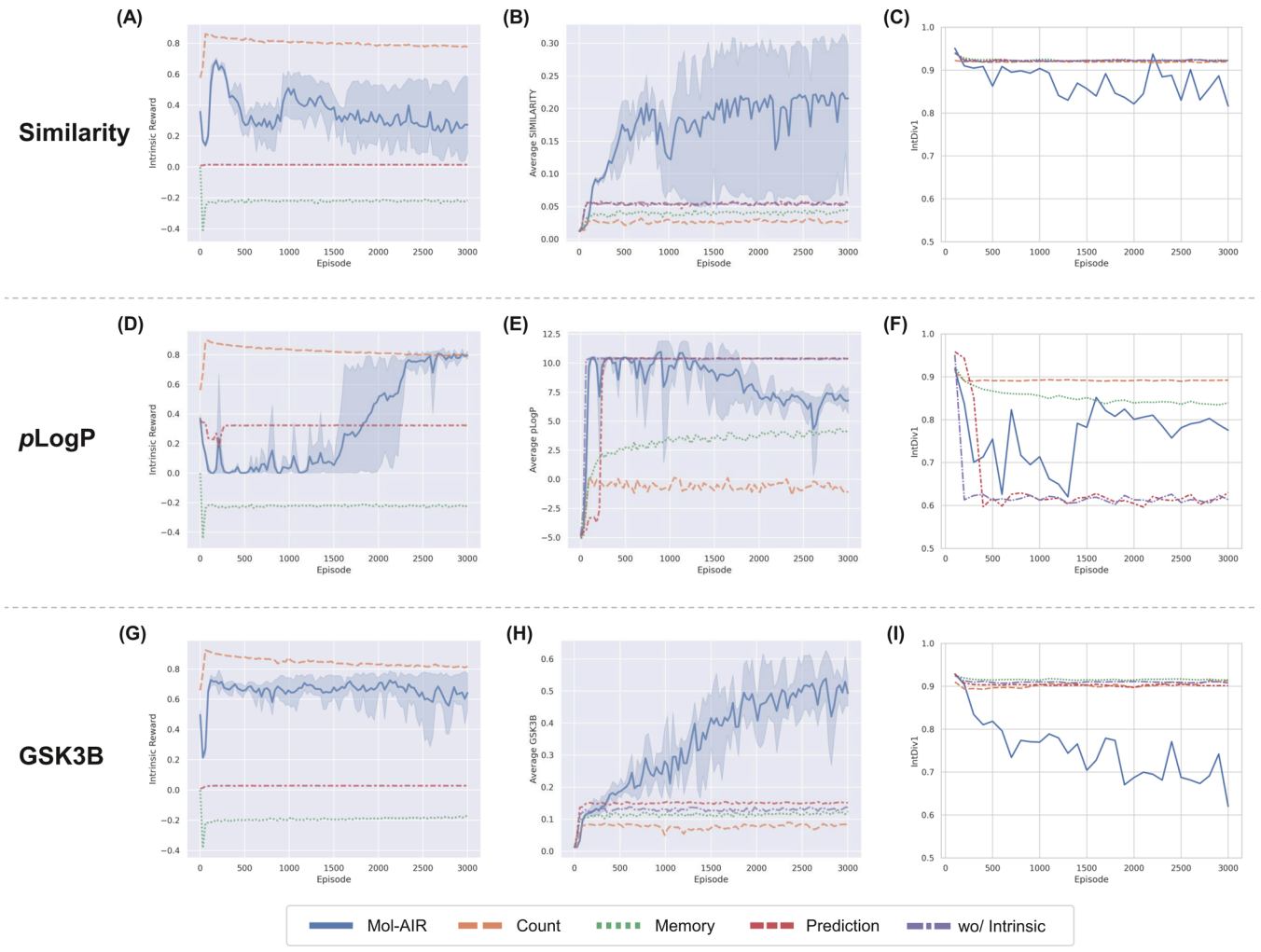


Figure 4. Performance comparison across intrinsic reward methods. (A), (B), and (C) show the average intrinsic rewards, average property scores, and IntDiv1 scores, respectively, over training episodes for experiments with similarity-based extrinsic rewards. (D), (E), and (F) present the corresponding results for *p*LogP-based experiments. (G), (H), and (I) display the results for GSK3B-based experiments.

celecoxib in terms of Tanimoto similarity of Morgan fingerprints. The range of Tanimoto similarity is in the closed interval [0, 1].

- GSK3B: This property measures the inhibitory ability of a molecule against glycogen synthase kinase-3 beta (GSK3B), a potential therapeutic target for Alzheimer's disease (AD) due to its association with AD pathophysiology.[46] The GSK3B score is on the closed interval [0, 1], and the goal of this task is to find a molecule with a high GSK3B score.
- JNK3: This property measures the inhibitory ability of a molecule against c-Jun N-terminal kinase 3 (JNK3), considered a drug target for AD treatment because its overexpression has been found to induce cognitive deficiency.[47] The JNK3 score is on the closed interval [0, 1], and the goal of this task is to find a molecule with a high JNK3 score.
- GSK3B+JNK3: A previous study suggested that a molecule capable of inhibiting both GSK3B and JNK3 simultaneously could be a potential drug candidate for AD treatment.[43] To find molecular candidates for GSK3B and JNK3 dual inhibitors, we set the arithmetic mean of GSK3B and JNK3 scores as an objective score and aim to find a molecule with a high mean score.

For the calculation of *p*LogP, QED, and similarity scores, we used scripts provided by existing research,[21] and for GSK3B and JNK3 scores, we utilized the oracle provided by the Therapeutics Data Commons (TDC) library.[48]

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

Molecular Discovery with the Best Property Scores. We compared our framework with the baseline methods that utilize three intrinsic rewards.[21] In the performance measurement experiments for the baseline techniques, the model parameter values for *p*LogP, QED, and similarity were set according to the values reported in Thiede et al.,[21] and for GSK3B and JNK3, optimal model parameters were determined using a grid search approach (Supplementary Table S2). All experimental results were evaluated based on 3000 iterations of training.

Table 1 presents data indicating that the proposed method, Mol-AIR, outperforms the baseline approaches in finding the best scoring molecules for all six tasks. To ensure the reliability of the results, we conducted three independent runs of each experiment. The mean and standard deviation of the outcomes are listed in Table 1. In particular, in the search for molecular structures similar to celecoxib, where existing methods failed to do so, the proposed framework was able to create molecular structures with a similarity that exceeded 30%. Furthermore, the existing baseline methods failed to reach 0.948, the theoretical optimum of QED, but the proposed Mol-AIR succeeded in reaching the theoretical optimum through efficient exploration.

Table 1 shows that the proposed method, Mol-AIR, is superior to the baseline approaches, outperforming them in finding the best scoring molecules for all six tasks. To ensure the reliability of the results, we conducted three independent runs of each experiment and included the mean and standard deviation of the outcomes in Table 1. In particular, in the search for molecular structures similar to celecoxib, where existing methods failed, the proposed framework was able to generate molecular structures with a similarity exceeding 30%. Furthermore, the existing baseline methods failed to reach 0.948, the theoretical optimum of QED, but the proposed Mol-AIR succeeded in achieving the theoretical optimum through efficient exploration.

To analyze the characteristics of the proposed method, we generated 64 molecular structures during each episode and investigated their average intrinsic rewards, target property values, and diversity using the IntDiv1 metric (Figure 4). IntDiv1 is a standard diversity metric implemented in benchmark frameworks such as MolScore[49] and GuacaMol.[6] This analysis provides insights into the effectiveness of Mol-AIR, particularly in balancing exploration and exploitation.

Figure 4A,B demonstrate that Mol-AIR successfully identified molecular structures similar to the target molecule, celecoxib, through its adaptive intrinsic reward function. By combining the strengths of history-based and learning-based approaches, this function decreases intrinsic rewards after identifying structures similar to celecoxib (episode >1000) to prevent unnecessary exploration of overly diverse molecules. This adaptive mechanism effectively balances exploration and exploitation, facilitating the discovery of desirable molecular structures guided by extrinsic rewards.

For *p*LogP, Mol-AIR provided smaller intrinsic rewards during the exploitation phase (episode <1500) and larger rewards during the exploration phase (episode >1500) (Figure 4D,E). For GSK3B, Mol-AIR showed significant fluctuations in intrinsic rewards, foreshadowing its ability to adaptively adjust exploration and exploitation to discover optimal molecular structures (Figure 4G). This balancing led to continuous improvement in target property scores, with Mol-AIR achieving a higher score than baseline methods (Figure 4H). Results for additional target properties are available in Supplementary Figure S3.

The IntDiv1 results, shown in Figure 4C,F,I, highlight Mol-AIR's adaptive balancing capability compared to baseline methods. Baseline methods generally exhibit relatively small fluctuations in diversity, with most maintaining consistently high IntDiv1 values, while some remain at consistently low values. These results suggest two distinct behaviors: in cases where IntDiv1 values are consistently high, excessive exploration occurs, leading to high diversity but insufficient exploitation, as observed in Figure 4C,I, as well as in the count-based and memory-based methods in Figure 4F. Conversely, consistently low IntDiv1 values, such as those seen in the prediction-based and wo/intrinsic methods in Figure 4F, indicate a lack of sufficient exploration, hindering the discovery of new chemical spaces and resulting in remaining stuck at local optima.

In contrast, Mol-AIR demonstrates significant fluctuations in diversity over training episodes, reflecting its ability to adaptively balance exploration and exploitation. By adjusting diversity levels according to the learning phase, Mol-AIR effectively avoids the pitfalls of excessive or insufficient exploration. This enables Mol-AIR to discover high-quality molecular structures and achieve better performance than baseline methods.

Moreover, the proposed method successfully discovered a sulfur−phosphorus-nitrogen chain with a higher *p*LogP value than the sulfur chain found in previous research (Figure 5).

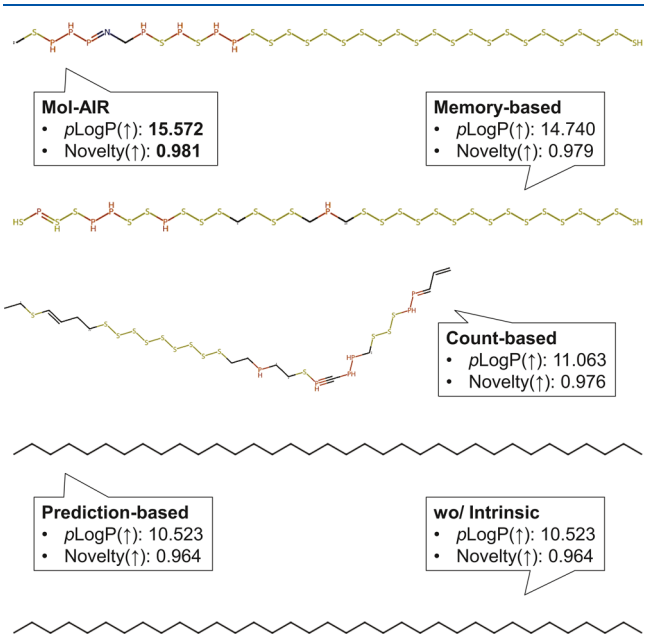


Figure 5. Comparison of the best molecules in *p*LogP optimization task. Novelty represents the degree of structural difference from existing molecules in ChEMBL. The novelty score is defined as one minus the average of Tanimoto similarity scores with all compounds in the ChEMBL.

As shown in Figure 5 and Supplementary Figure S4A, the proposed method initially discovered carbon chains similar to those discovered with prediction-based intrinsic rewards and no intrinsic rewards, found a sulfur chain after training for 750 episodes, and discovered a sulfur−phosphorus-nitrogen chain that was superior to the sulfur−phosphorus chain found with a memory-based intrinsic reward after 2000 episodes of training. Figure 4D also demonstrates that the proposed method

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

Table 2. Results of Ablation Study with the Best Property Scores

intrinsic	best property score					
type	*p*LogP	QED	similarity	GSK3B	JNK3	GSK3B+JNK3
wo/Intrinsic	10.523 ± 0.000	0.890 ± 0.014	0.223 ± 0.019	0.433 ± 0.051	0.240 ± 0.000	0.282 ± 0.003
HIR only	12.569 ± 1.983	0.932 ± 0.013	0.235 ± 0.045	0.647 ± 0.024	0.423 ± 0.048	0.389 ± 0.022
LIR only	10.523 ± 0.000	0.925 ± 0.011	0.284 ± 0.015	0.663 ± 0.009	0.230 ± 0.016	0.397 ± 0.033
Mol-AIR	**15.682 ± 0.179**	**0.948 ± 0.001**	**0.308 ± 0.024**	**0.707 ± 0.005**	**0.493 ± 0.070**	**0.401 ± 0.040**

significantly enhances its exploration capability during episodes 1500 to 2500, achieving higher scores compared to existing approaches. The results for the other properties are shown in Supplementary Figures S4−S9.

Ablation Study. We performed ablation experiments in six benchmark tests to demonstrate the benefits of combining HIR and LIR. To ensure the reliability of the results, we conducted three independent runs of each experiment. The mean and standard deviation of the outcomes are included in Table 2. Table 2 presents the results of the generation of molecular structures with the highest scores for each benchmark test. It demonstrates that the Mol-AIR method, which uses both HIR and LIR, outperforms others in all benchmarks. The use of either HIR or LIR alone decreased performance in all cases, with HIR proving to be more effective than LIR.

To investigate why AIR outperforms the singular application of HIR and LIR, we analyzed intrinsic reward patterns across three benchmark tests: QED, GSK3B, and GSK3B+JNK3. The analysis revealed distinct patterns for AIR, LIR, and HIR (Figure 6). LIR failed to enhance exploration owing to low intrinsic rewards at the start of training but showed an increase in reward magnitude as the neural network progressed and remembered visited states. However, as the prediction accuracy of the neural network model improved for unvisited states through learning and generalization, the trend of increasing intrinsic rewards reversed, leading to a decrease. This pattern indicated that LIR, except at the beginning, was ineffective in inducing long-time exploration.

On the contrary, HIR provided a consistent level of intrinsic reward from the start, as it does not require a learning process. However, its constant encouragement for exploration made exploitation difficult as training progressed. The AIR pattern proposed in this study adequately combines the HIR and LIR patterns. Specifically, similar to LIR, it produces high intrinsic rewards at the beginning of RL to induce strong exploration and then provides steady intrinsic rewards like HIR, assisting RL by allowing attempts at new structures toward the end. Results for other target properties also show that HIR and AIR work together to facilitate appropriate exploration (Supplementary Figure S10). These results confirm that combining HIR and LIR, as in AIR, is more effective in encouraging exploration than using either alone.

Hyperparameter Analysis. For Mol-AIR to achieve optimal exploration, it is crucial to adjust the weight of intrinsic rewards α and the HIR-LIR balancing parameter β in the eqs 9 and 10. As the pattern of extrinsic rewards varies depending on the target property, α requires a heuristic setting. However, β, which controls the exploratory dominance between HIR and LIR, was investigated for the optimal ratio between various target properties with values of β in the set $\{1, 0.1, 0.01, 0.001\}$. Supplementary Table S3 presents the results of comparative experiments on β for six target properties, and Table 3 shows the ranking results. It was observed that a value of β of 0.01 statistically outperforms across multiple target properties. As the scale of HIR is approximately 100 times larger than that of LIR, setting β to 0.01 seems to best balance HIR and LIR, preventing the dominance of excessively large intrinsic rewards and leading to a harmonious exploration between HIR and LIR.

Proof-of-Concept Experiment: Discovery of DRD2 Inhibitors. In the previous section, we demonstrated the exploration capabilities of our methodology through the *p*LogP optimization experiment (Figure 5). While such a benchmark effectively evaluates computational performance, chemically nonsensical structures may be generated if the sole objective is to maximize *p*LogP. To address this limitation and demonstrate that our proposed method can generate synthesizable and chemically meaningful molecules, we conducted a POC experiment focused on discovering dopamine receptor D2 (DRD2) inhibitors.

The DRD2 is a crucial target for therapeutic intervention in treating neuropsychiatric disorders, including schizophrenia, bipolar disorder, and Parkinson's disease.[50] Identifying new inhibitor candidates with improved efficacy and reduced side effects is essential for advancing therapeutic strategies, highlighting the importance of discovering effective DRD2 inhibitors.

The primary objective of this POC experiment was to identify compounds with high inhibitory activity against DRD2. In addition, we included two secondary objectives to ensure the practical viability of the generated molecules: high QED and low synthetic difficulty. Synthetic difficulty was measured using the synthetic accessibility (SA) score.[51] These objectives were incorporated into the Mol-AIR framework to enable the agent to efficiently explore the chemical space and identify valid, potentially synthesizable drug candidates. The oracle function for this POC experiment is defined below and applied to eq 8:

$$\Lambda^{(p)}(x) := \frac{1}{3}(\mathrm{DRD2}(x) + \mathrm{QED}(x) + (1 - \mathrm{SA}(x)/10)) \tag{11}$$

where all scoring functions for DRD2, QED, and SA used in this POC were obtained from Therapeutics Data Commons.[52]

Figure 7 illustrates the results of this POC study, showing that our method successfully generated compounds structurally similar to risperidone, a well-known DRD2 inhibitor. As shown in Figure 7A, we began by pretraining a generator with the ChEMBL v29 data set,[53] applying RDKit's SMILES randomization function for data augmentation.[54] This pretrained model was then fine-tuned using Mol-AIR to bias the generation toward compounds with higher DRD2 inhibitory activity, higher QED, and lower SA scores. After fine-tuning, we generated 10K unique compounds structurally similar to risperidone by conditioning on a fragment of its structure (Supplementary Figure S11). Finally, we analyzed these compounds using AutoDock Vina[55] and ChimeraX.[56,57]

Figure 7B displays a radar chart summarizing various metrics for the 10K generated compounds. Using SELFIES for molecular representation, we achieved a 100% validity rate

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

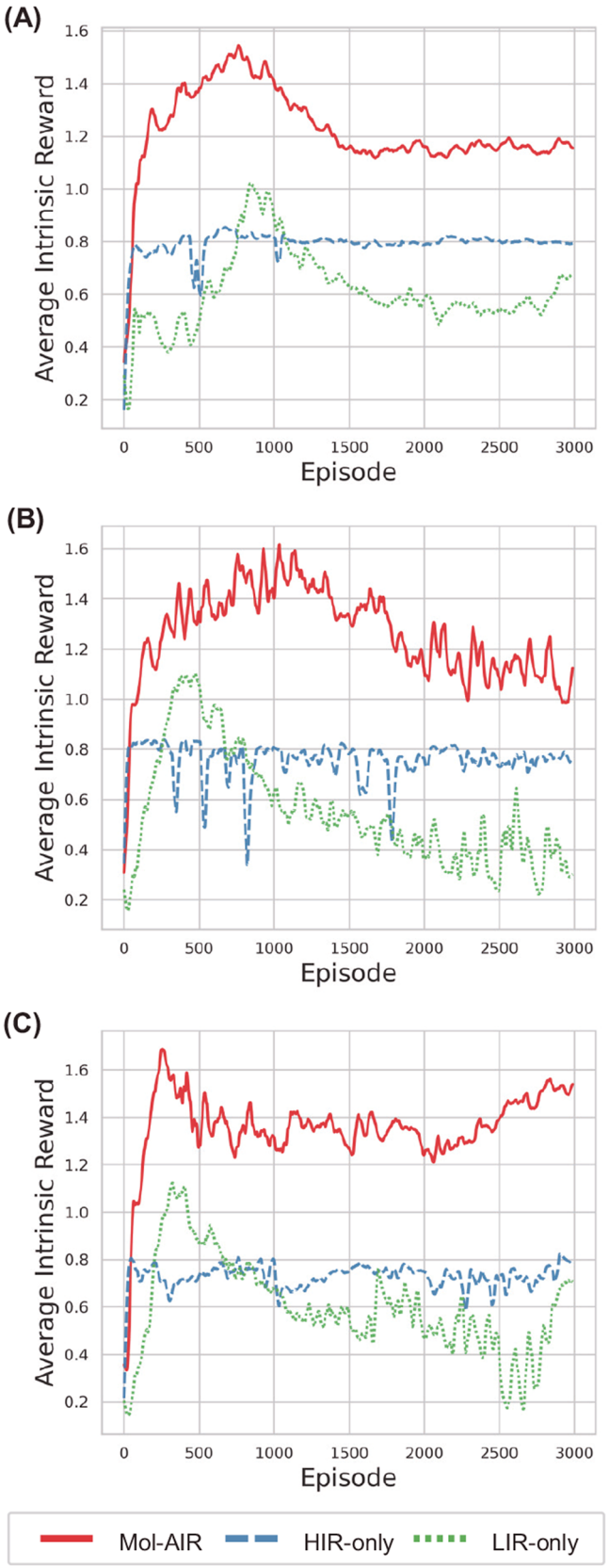


Figure 6. Comparison of intrinsic reward patterns during training among the ablation cases. (A) QED, (B) GSK3B, and (C) GSK3B +JNK3.

and observed an internal diversity score of approximately 0.5, indicating moderate structural variation. Scaffold uniqueness and diversity were not high, as the goal of the POC was to discover chemical structures resembling risperidone and its scaffold. Figure 7C presents a Venn diagram showing how many of these generated molecules exceeded risperidone's DRD2 inhibitory activity score (>0.983), had lower synthetic accessibility (<2.736), and displayed higher QED (>0.658). This analysis identified 84 compounds that met all three criteria (Supplementary Table S4).

To determine whether our findings could be potential hit compounds, we assessed their binding affinity against DRD2. We performed molecular docking simulations using the Protein Data Bank (PDB) entry *6CM4* for DRD2,[58,59] as shown in Figure 7D. The top ten compounds from this set exhibited significantly lower binding energies than risperidone, suggesting they may serve as strong candidates for DRD2 inhibition. Figure 7E highlights the 2D structures of risperidone and representative hits with low binding energies. Finally, Figure 7F visualizes the binding poses in ChimeraX. These results demonstrate that the generated molecules adopt conformations similar to risperidone within the DRD2 binding pocket, supporting Mol-AIR's potential utility for targeted drug discovery aimed at specific scaffolds like risperidone.

Through this POC experiment, we confirm that our methodology is not only capable of optimizing simple properties (e.g., *p*LogP) but also addresses more complex, multiobjective challenges in drug design.

Table 3. Results of Hyperparameter Analysis with Ranking Scores

β	1.0	0.1	0.01	0.001
*p*log*P*	3	2	**1**	4
QED	4	2	**1**	3
Similarity	4	3	**1**	2
GSK3B	4	3	2	**1**
JNK3	**1**	3	2	4
GSK3B+JNK3	4	3	2	**1**
Average Ranking (↓)	3.3	2.7	**1.5**	2.5

CONCLUSIONS

In this study, we propose a new RL-based framework based on a novel intrinsic reward that performs efficient exploration for goal-directed molecular generation. The proposed approach is a hybrid approach for effective exploration, utilizing an adaptive intrinsic reward function that combines the strengths of history-based and learning-based approaches. Our results reveal that this approach is effective for efficient exploration in the chemical space, as it successfully discovers molecules that are better than compounds discovered by existing intrinsic reward methods. An ablation study revealed that the proposed method's two components, LIR and HIR, synergistically contribute to its success. LIR facilitates strong early phase exploration, while HIR ensures sustained exploration later on, ultimately guiding the RL agent toward optimal molecular structures. Mol-AIR successfully performs efficient exploration, balancing exploration and exploitation for goal-directed molecular generation. As a result, we discovered new desired molecular structures for various target chemical properties.

To demonstrate the adaptability of our approach for more complex oracles, we conducted a POC experiment focused on discovering DRD2 inhibitors under the simultaneous constraints of DRD2 inhibitory activity, synthetic accessibility, and QED. Our method successfully generated compounds structur-

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

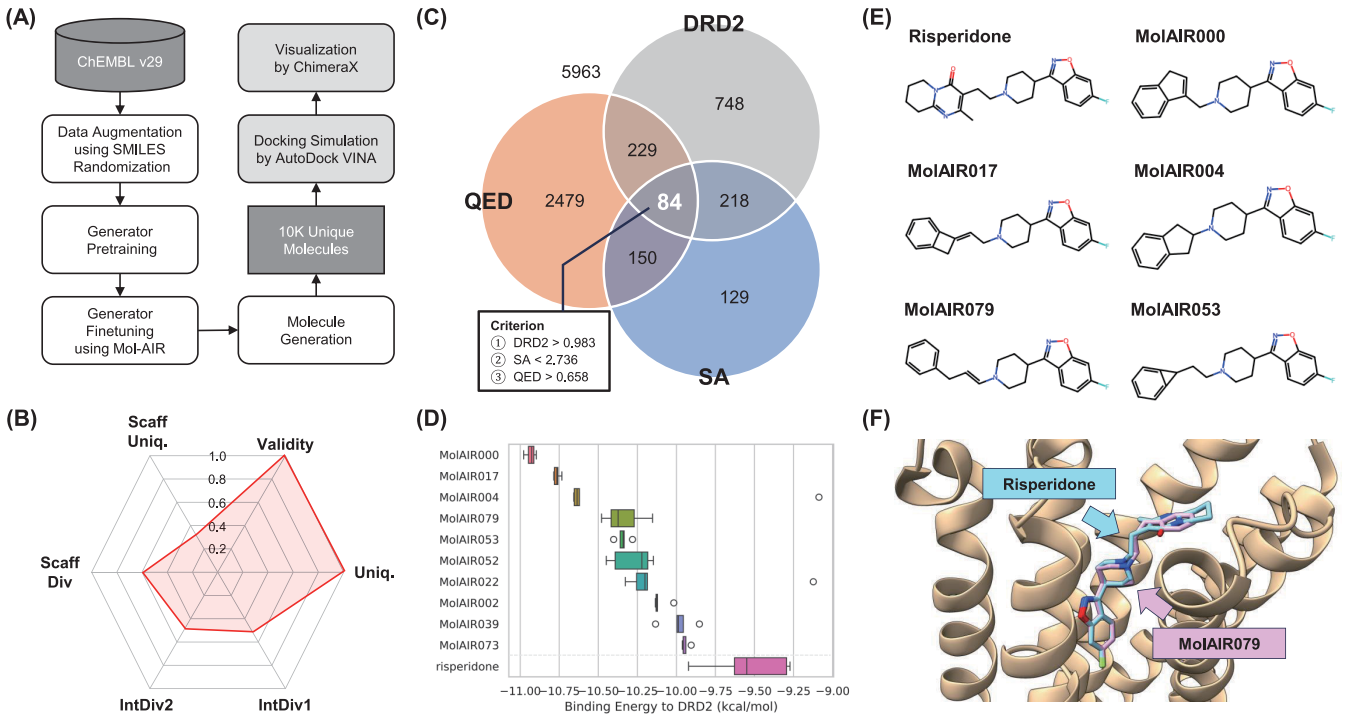


Figure 7. Results of DRD2 inhibitors discovery using Mol-AIR. (A) Overall workflow of the POC experiment for generating risperidone-like compounds. (B) Radar chart summarizing the results of generating the 10K compounds. (C) Venn diagram illustrating the number of generated molecules that exceed risperidone's DRD2 inhibitory activity score, have lower synthetic accessibility, and exhibit higher QED. (D) Top ten binding energy results from docking simulations. (E) 2D structures of risperidone and representative low-binding-energy hits. (F) Docking pose visualization of risperidone and a hit compound discovered by Mol-AIR.

ally similar to risperidone, a known DRD2 inhibitor, showcasing the framework's potential to handle multiple extrinsic objectives. However, we acknowledge that in more complex environments, the likelihood of false positives may increase. Moving forward, we plan to develop techniques to automatically regulate the exploration induced by intrinsic rewards, ensuring a balance between thorough exploration and reliability in diverse real-world applications.

Generating molecular structures similar to existing drugs is a crucial part of lead optimization in drug development. In future work, we aim to conduct a more comprehensive POC study, potentially involving synthesis and biological testing, to further validate our computational findings. By refining the intrinsic rewards and reinforcement learning techniques, we aspire to develop models capable of generating structurally similar molecules with high fidelity, ensuring that our approach remains practical and effective in real-world drug discovery pipelines. Such expanded research will help confirm Mol-AIR's capabilities under more complex conditions and experimental settings, paving the way for broader applications in drug discovery and beyond.

■ ASSOCIATED CONTENT

Data Availability Statement

Our data and source code are available at https://github.com/DevSlem/Mol-AIR.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c01669.

Model Parameters of Mol-AIR (Table S1), Hyperparameter Analysis for Baselines (Table S2), Hyperparameter Analysis for the proposed method (Table S3), Overview of the RND Architecture for Mol-AIR. The numbers between blocks represent the vector lengths (Figure S1), Overview of the Actor-Critic Network Architecture for Mol-AIR. The numbers between blocks represent the vector lengths (Figure S2), The Average Intrinsic reward and Average objective property score of the molecules generated per batch over the training episodes (Figure S3), The objective property value of the best generated molecules in the training run over episodes (Figure S4), The molecules with the highest QED value discovered by each method (Figure S5), The molecules with the highest SIMILARITY value discovered by each method (Figure S6), The molecules with the highest GSK3B value discovered by each method (Figure S7), The molecules with the highest JNK3 value discovered by each method (Figure S8), The molecules with the highest GSK3B+JNK3 value discovered by each method (Figure S9), Results of ablation study of Mol-AIR (Figure S10), and Risperidone fragment used for conditioning in the POC experiment (Figure S11) (PDF)

The list of 84 compounds identified by Mol-AIR for DRD2 inhibition (Table S4) (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Jonghwan Choi − *Division of Software, Hallym University, Chuncheon-si, Kangwon-do 24252, Republic of Korea;* orcid.org/0000-0002-8429-4135; Email: jonghwanc@hallym.ac.kr

Jibum Kim − *Department of Computer Science and Engineering and Center for Brain-Machine Interface, Incheon National*

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296

University, Incheon 22012, Republic of Korea;
Email: jibumkim@inu.ac.kr

Authors

Jinyeong Park − *Department of Computer Science and Engineering, Incheon National University, Incheon 22012, Republic of Korea*

Jaegyoon Ahn − *Department of Computer Science and Engineering, Incheon National University, Incheon 22012, Republic of Korea*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.4c01669

Author Contributions

J.C. and J.K. are co-corresponding authors. J.P.: Data curation, Investigation, Methodology, Software, Validation, Writing−original draft. J.A.: Conceptualization, Methodology. J.C.: Formal analysis, Funding acquisition, Investigation, Visualization, Writing−review and editing. J.K.: Funding acquisition, Resources, Project administration, Supervision, Writing - review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) (IITP-2025-RS-2023-00260175, 30%) grant funded by the Korea government (Ministry of Science and ICT) and the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) (IITP-2025-RS-2024-00437024, 30%) grant funded by the Korea government (Ministry of Science and ICT). This work was also supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) (RS-2024-00345226, 40%).

■ REFERENCES

(1) Pereira, D.; Williams, J. Origin and evolution of high throughput screening. *British journal of pharmacology* **2007**, *152*, 53−61.
(2) Kumar, N.; Acharya, V. Machine intelligence-driven framework for optimized hit selection in virtual screening. *J. Cheminform.* **2022**, *14*, 48.
(3) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* **2013**, *27*, 675−679.
(4) Chen, W.; Liu, X.; Zhang, S.; Chen, S. Artificial intelligence for drug discovery: Resources, methods, and applications. *Mol. Ther. Nucleic Acids* **2023**, *31*, 691−702.
(5) Anstine, D. M.; Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **2023**, *145*, 8736−8750.
(6) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **2019**, *59*, 1096−1108.
(7) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1608.
(8) Krenn, M.; Ai, Q.; Barthel, S.; Carson, N.; Frei, A.; Frey, N. C.; Friederich, P.; Gaudin, T.; Gayle, A. A.; Jablonka, K. M.; et al. SELFIES and the future of molecular string representations. *Patterns* **2022**, *3*, No. 100588.
(9) Jones, G. M.; Story, B.; Maroulas, V.; Vogiatzis, K. D. *Molecular Representations for Machine Learning*; American Chemical Society, 2023.
(10) Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *International conference on machine learning*; PMLR: Stockholm, Sweden, 2018; 2323−2332.
(11) Choi, J.; Seo, S.; Choi, S.; Piao, S.; Park, C.; Ryu, S. J.; Kim, B. J.; Park, S. ReBADD-SE: Multi-objective molecular optimization using SELFIES fragment and off-policy self-critical sequence training. *Computers in Biology and Medicine* **2023**, *157*, No. 106721.
(12) Kim, H.; Ko, S.; Kim, B. J.; Ryu, S. J.; Ahn, J. Predicting chemical structure using reinforcement learning with a stack-augmented conditional variational autoencoder. *J. Cheminform.* **2022**, *14*, 83.
(13) Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **2022**, *62*, 2064−2076.
(14) Cheng, Y.; Gong, Y.; Liu, Y.; Song, B.; Zou, Q. Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief. Bioinform.* **2021**, *22*, No. bbab344.
(15) Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Brief. Bioinform.* **2022**, *23*, No. bbab430.
(16) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268−276.
(17) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48.
(18) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **2019**, *37*, 1038−1040.
(19) Pereira, T.; Abbasi, M.; Ribeiro, B.; Arrais, J. P. Diversity oriented deep reinforcement learning for targeted molecule generation. *J. Cheminform.* **2021**, *13*, 21.
(20) Yang, S.; Hwang, D.; Lee, S.; Ryu, S.; Hwang, S. J. Hit and lead discovery with explorative rl and fragment-based molecule generation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 7924−7936.
(21) Thiede, L. A.; Krenn, M.; Nigam, A.; Aspuru-Guzik, A. Curiosity in exploring chemical spaces: intrinsic rewards for molecular reinforcement learning. *Machine Learning: Science and Technology* **2022**, *3*, No. 035008.
(22) Chadi, M.-A.; Mousannif, H.; Aamouche, A.Curiosity as a Self-Supervised Method to Improve Exploration in De novo Drug Design. *International Conference on Information Technology Research and Innovation (ICITRI)*; IEEE: Jakarta, Indonesia, 2023, *2023*, 151−156.
(23) Burda, Y.; Edwards, H.; Storkey, A.; Klimov, O. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* 2018
(24) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **2020**, *12*, 56.
(25) Gao, W.; Fu, T.; Sun, J.; Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21342−21357.
(26) Cheng, A. H.; Cai, A.; Miret, S.; Malkomes, G.; Phielipp, M.; Aspuru-Guzik, A. Group SELFIES: a robust fragment-based molecular string representation. *Digit. Discov.* **2023**, *2*, 748−758.
(27) Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* 2017
(28) Mokaya, M.; Imrie, F.; van Hoorn, W. P.; Kalisz, A.; Bradley, A. R.; Deane, C. M. Testing the limits of SMILES-based de novo molecular generation with curriculum and deep reinforcement learning. *Nature Machine Intelligence* **2023**, *5*, 386−394.
(29) Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A., et al.

Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038* 2020
(30) Strehl, A. L.; Littman, M. L. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences* **2008**, *74*, 1309−1331.
(31) Ostrovski, G.; Bellemare, M. G.; Oord, A.; Munos, R.Count-based exploration with neural density models. *International conference on machine learning*. 2017, 2721−2730.
(32) Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*; Curran Associates, Inc., **2016**, *29*.
(33) Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Xi Chen, O.; Duan, Y.; Schulman, J.; DeTurck, F.; Abbeel, P. #Exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, **2017**, *30*.
(34) Pathak, D.; Agrawal, P.; Efros, A. A.; Darrell, T. Curiosity-driven exploration by self-supervised prediction. *International conference on machine learning*; PMLR: Sydney, Australia, 2017; 2778−2787.
(35) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for de novo drug design. *J. Chem. Inf. Model.* **2020**, *60*, 5918−5922.
(36) Jo, D.; Kim, S.; Nam, D.; Kwon, T.; Rho, S.; Kim, J.; Lee, D. Leco: Learnable episodic count for task-specific intrinsic reward. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 30432−30445.
(37) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.
(38) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, No. eaap7885.
(39) You, J.; Liu, B.; Ying, Z.; Pande, V.; Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*; Curran Associates, Inc., **2018**, *31*.
(40) Sha, C.; Zhu, F. Goal-directed molecule generation with fine-tuning by policy gradient. *Expert Systems with Applications* **2024**, *246*, No. 123127.
(41) Stadie, B. C.; Levine, S.; Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814* 2015
(42) Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438* 2015
(43) Li, Y.; Zhang, L.; Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *J. Cheminform.* **2018**, *10*, 33.
(44) Fromer, J. C.; Coley, C. W. Computer-aided multi-objective optimization in small molecule discovery. *Patterns* **2023**, *4*, No. 100678.
(45) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chem.* **2012**, *4*, 90−98.
(46) Lauretti, E.; Dincer, O.; Praticò, D. Glycogen synthase kinase-3 signaling in Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2020**, *1867*, No. 118664.
(47) Solas, M.; Vela, S.; Smerdou, C.; Martisova, E.; Martínez-Valbuena, I.; Luquin, M.-R.; Ramírez, M. J. JNK Activation in Alzheimer's Disease Is Driven by Amyloid β and Is Associated with Tau Pathology. *ACS Chem. Neurosci.* **2023**, *14*, 1524−1534.
(48) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.* **2022**, *18*, 1033−1036.
(49) Thomas, M.; O'Boyle, N. M.; Bender, A.; De Graaf, C. MolScore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. *J. Cheminform.* **2024**, *16*, 64.
(50) Zhang, Y.; Yu, J.-G.; Wen, W. Recent Advances in Representative Small-Molecule DRD2 Inhibitors: Synthetic Routes and Clinical Applications. *Eur. J. Med. Chem.* **2024**, *277*, No. 116731.
(51) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.
(52) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* 2021
(53) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research* **2024**, *52*, D1180−D1192.
(54) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **2019**, *11*, 71.
(55) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891−3898.
(56) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein science* **2021**, *30*, 70−82.
(57) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **2023**, *32*, No. e4792.
(58) Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols* **2017**, *1607*, 627−641.
(59) Wang, S.; Che, T.; Levit, A.; Shoichet, B. K.; Wacker, D.; Roth, B. L. Structure of the D2 dopamine receptor bound to the atypical antipsychotic drug risperidone. *Nature* **2018**, *555*, 269−273.

https://doi.org/10.1021/acs.jcim.4c01669
J. Chem. Inf. Model. 2025, 65, 2283−2296