Diverse Mini-Batch Selection in Reinforcement Learning for Efficient Chemical Exploration in de novo Drug Design

Diverse Mini-Batch Selection in Reinforcement Learning for Efficient Chemical Exploration in *de novo* Drug Design

Hampus Gummesson Svensson
Chalmers University of Technology and University of Gothenburg
Molecular AI, Discovery Sciences, R&D, AstraZeneca
Gothenburg, Sweden
hampusgs@gmail.com

Ola Engkvist
Chalmers University of Technology and University of Gothenburg
Molecular AI, Discovery Sciences, R&D, AstraZeneca
Gothenburg, Sweden
ola.engkvist@astrazeneca.com

Jon Paul Janet
Molecular AI, Discovery Sciences, R&D, AstraZeneca
Gothenburg, Sweden
jonpaul.janet@astrazeneca.com

Christian Tyrchan
Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca
Gothenburg, Sweden
christian.tyrchan@astrazeneca.com

Morteza Haghir Chehreghani
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
morteza.chehreghani@chalmers.se

ABSTRACT

In many real-world applications, evaluating the quality of instances is costly and time-consuming, e.g., human feedback and physics simulations, in contrast to proposing new instances. In particular, this is even more critical in reinforcement learning, since it relies on interactions with the environment (i.e., new instances) that must be evaluated to provide a reward signal for learning. At the same time, performing sufficient exploration is crucial in reinforcement learning to find high-rewarding solutions, meaning that the agent should observe and learn from a diverse set of experiences to find different solutions. Thus, we argue that learning from a diverse mini-batch of experiences can have a large impact on the exploration and help mitigate mode collapse. In this paper, we introduce mini-batch diversification for on-policy reinforcement learning and study this framework in the context of a real-world problem, namely, drug discovery. We extensively evaluate how our proposed framework can enhance the effectiveness of chemical exploration in *de novo* drug design, where finding diverse and high-quality solutions is crucial. Our experiments demonstrate that our proposed diverse mini-batch selection framework can substantially enhance the diversity of solutions while maintaining high-quality solutions. In drug discovery, such an outcome can potentially lead to fulfilling unmet medical needs faster.

KEYWORDS

Reinforcement Learning; *de novo* Drug Design; Exploration

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LHWU6232

ACM Reference Format:
Hampus Gummesson Svensson, Ola Engkvist, Jon Paul Janet, Christian Tyrchan, and Morteza Haghir Chehreghani. 2026. Diverse Mini-Batch Selection in Reinforcement Learning for Efficient Chemical Exploration in *de novo* Drug Design. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/LHWU6232

1 INTRODUCTION

In recent years, utilizing reinforcement learning (RL) for fine-tuning of pre-trained generative models has shown great success in various applications [16, 63], including *de novo* drug design [4, 36]. *De novo* drug design is a computational problem that aims to identify novel molecular structures with specific properties without any starting template [33], where generative models have shown great success [39, 57]. When fine-tuning a generative model, the goal is often to align the model's outputs with respect to human preferences or experiments. However, many practical applications require frequent assessment of data and experiences, e.g., via human expert evaluation, computer simulations, field testing, and laboratory experimentation. These assessment methods are often resource-intensive, demanding significant time and financial investment. In *de novo* drug design, resource-intensive computational methods are used to assess the fit of molecules into the binding site of a target protein to predict the strength of each protein-ligand interaction [38]. Consequently, the volume of data that can undergo thorough evaluation is often constrained by budgetary limitations.

In this paper, we tackle this problem in reinforcement learning, where the training instances are provided solely from the agent's interaction with the environment. In particular, we study this problem in the context of *de novo* drug design, where RL techniques are commonly used to fine-tune a pre-trained generative model to produce molecules with desired properties [41, 42]. In general, many successful RL algorithms, e.g., [32, 49], run many copies of the environment in parallel to synchronously or asynchronously learn from

numerous interactions. For synchronous on-policy algorithms, the experiences are accumulated to compute an average loss to update the agent's policy. This is also true for *de novo* drug design [36], where for each policy update, a batch of molecules is first generated in parallel. However, in many real-world applications, including *de novo* drug design, it is impractical to assess all interactions with the environment, where each assessment of an interaction provides a reward signal for the agent. Instead, it is preferable to evaluate a smaller, representative set and learn from it.

At the same time, to avoid mode collapse, exploration mechanisms play a vital role in agent performance, especially in tasks with delayed/sparse reward or for a reward landscape with a vast number of local optima to explore. In *de novo* drug design, a reward can only be obtained when the full molecular structure has been generated. Moreover, diversity among generated molecules is essential since a diverse molecular library increases the likelihood of identifying candidates with unique and favorable pharmacological profiles, thereby enhancing the overall efficiency and success rate of drug development pipelines. In drug design, the reward function is often complex and has many high-rewarding modes that should be found and subsequently exploited to obtain a diverse set of solutions. Thus, chemical exploration and diversification are of integral importance in drug design. In real-world deployment of this *de novo* drug design, it is also often costly and time-consuming to evaluate an instance (i.e., a state-action episode) to obtain a reward. This creates a *reward bottleneck* which limits the policy updates, leading to the need for efficient exploration.

One popular approach to enhance exploration in RL is the addition of an exploration bonus to the reward function, commonly denoted as *intrinsic reward* [5, 11, 51, 55]. Another common approach is maximum entropy RL, where the agent tries to maximize both the reward and entropy simultaneously, i.e., succeeding at a task while still acting as randomly as possible [22, 35]. Our work provides a consistent perspective where, while improving exploration by achieving diverse behaviors, it is important to make sure that the interactions with the environment are of high quality (i.e., receive high rewards). This becomes especially critical when the agent must account for safety considerations, exhibits sensitivity to noise, or operates in environments where numerous trajectories are infeasible. For example, in *de novo* drug design, a molecular representation may not correspond to a chemically viable compound, and minor modifications can readily compromise its validity. In this work, we accomplish this by considering mini-batch diversification in reinforcement learning, where a large number of interactions (obtained from running copies of the environment in parallel) are summarized in a smaller, diverse set of interactions used for updating the policy. This provides an effective way to impose additional exploration in the learning process, while overcoming the reward bottleneck by learning from a smaller set.

In this paper, we argue that providing a diverse mini-batch of interactions makes the agent's exploration more effective and increases the diversity of the forthcoming interactions, especially in *de novo* drug design. Thus, there are two key benefits for such mini-batch diversification: (1) computational aspects to address the reward bottleneck; (2) enhance exploration by diverse behaviors. Therefore, we introduce a framework for diverse mini-batch selection in reinforcement learning. To the best of our knowledge, this is the first effort to study the effects of diverse mini-batch selection in on-policy reinforcement learning to overcome the reward bottleneck and promote exploration. We study the use of determinantal point processes (DPP) [27], the MaxMin algorithm [3] and k-medoids clustering [44] for this task. DPPs provide an effective framework to sample a diverse set based on specified similarity information, while the MaxMin algorithm and k-medoids clustering seek to choose a subset to maximize the coverage of a larger set.

Previous work has proposed a mini-batch diversification scheme based on DPPs for stochastic gradient descent and shown its effectiveness [24, 64], but does not study such a scheme in the context of reinforcement learning. Existing work has used DPPs in diverse sampling for batch Bayesian Optimization [34]. In this paper, we focus on mini-batch diversification for improving exploration and reducing reward computations (i.e., addressing the reward bottleneck) in reinforcement learning. In reinforcement learning, DPPs have previously been used for unsupervised option discovery [13], diverse recommendations for RL-based user preferences [30], and multi-agent RL [37, 52, 62]. All of these are different from our setting and can not be applied to our setting. The work of Zhao et al. [65] employs a DPP to model within-trajectory diversity and prioritize replay in an off-policy setting. In contrast, our work focuses on on-policy learning and how mini-batch diversification enhances exploration through diverse behaviors. Moreover, we consider diversity across trajectories and use it to fine-tune a generative model and also compare different methods for selecting diverse trajectories. Additionally, we investigate the impact of mini-batch diversification on the real-world problem of drug design. The MaxMin algorithm is a popular method used in drug discovery to pick a diverse set [15, 54], but has not been investigated in combination with reinforcement learning. Furthermore, k-medoids clustering is a widely known clustering technique for finding a good partition in non-Euclidean data and has only been used for cluster-based RL [19], which is different from our setting.

To the best of our knowledge, our paper provides the first combinations of these methods with on-policy reinforcement learning to effectively fine-tune a generative model for *de novo* drug design (or any other application). Thereby, the contribution of this paper is twofold:

- We propose a mini-batch diversification framework for on-policy RL to enhance exploration and, at the same time, to address the reward bottleneck issue.
- We extensively investigate the proposed framework on the *de novo* drug design application, and demonstrate its effectiveness via extensive experiments.

Due to the characteristics of the *de novo* drug design problem, it is a suitable problem to employ diverse mini-batch selection and study its effectiveness. We believe that this framework can also help to overcome the reward bottleneck and enhance exploration in other real-world applications of reinforcement learning, especially for fine-tuning a pre-trained generative model in other domains. Exploration is a key challenge in RL, and domain-specific information can easily be incorporated into the proposed framework.

2 RL-BASED *DE NOVO* DRUG DESIGN

The aim of *de novo* drug design is to design novel drug molecules given a set of predefined constraints, but without any known initial structure [33]. A popular approach for *de novo* drug design is to use chemical language models to generate string-based representations of molecules [2, 50]. To steer the chemical language model to promising areas of the chemical space, reinforcement learning can be leveraged [36]. This paper focuses on promoting diversity in RL-based fine-tuning of a chemical language model via mini-batch diversification. An action a in this RL problem corresponds to adding one token to the string representation of the molecule, where $\mathcal{A}$ is the set of possible tokens that can be added, including a start token a^{start} and a stop token a^{stop}. One popular string-based representation of chemical entities is Simplified Molecular Input Line Entry System [60], abbreviated SMILES, where an action correpsonds adding the next token in the SMILES string. The reward function assesses the quality of the molecule represented by the string, and the molecule can only obtain a reward when the full string representation has been generated, i.e., a stop token has been added. This *de novo* drug design problem can be modeled as a Markov decision process (MDP), e.g., see [20] for more details. Our work builds upon the success of the SMILES-based REINVENT [8, 31, 36, 50], since existing evaluations [17, 56] have concluded good performance of REINVENT compared to both other RL-based and non-RL-based approaches for *de novo* drug design. In particular, we focus on improving its chemical exploration and avoiding mode collapse.

The drug-like chemical space is estimated to consist of 10^{33} synthesisable molecules [43]. To explore this space and improve the diversity of the generated molecules, several studies aim to improve the chemical exploration carried out by the RL agent. Without the use of any exploration technique, the policy easily collapses to generating only a few modes of the reward function, which leads to low diversity. To improve the diversity in RL-based *de novo* drug design, Blaschke et al. [9] therefore introduces a count-based method that reduces the reward for similar molecules based on their structure. The work of Park et al. [40] and Wang and Zhu [59] employs memory and learning-based intrinsic motivation to improve the reward of the generated molecules. Moreover, previous work shows that incorporating both structure- and learning-based information into the reward function can improve the overall diversity of *de novo* drug design [21]. Our work takes on a fundamentally alternative perspective to enhance diversity. Rather than just encouraging diverse and explorative behavior via the reward signal, our work studies the effect of maximizing the diversity of the molecules that we evaluate and learn from.

3 DIVERSE MINI-BATCH SELECTION FOR RL

We propose a framework to enhance exploration in reinforcement learning while reducing the number of interactions evaluated. We seek to generate more diverse solutions through reinforcement learning-based fine-tuning of a pre-trained generative model. In this paper, we focus on fine-tuning a chemical language model. We assume delayed rewards and that acquiring a sequence of states and actions is inexpensive compared to the evaluation, which is often true for real-world problems such as *de novo* drug design. Given a large set of interactions, we seek to select a smaller, representative set to use for updating the parameters of our policy. We hypothesize that this affects the agent's exploration of the solution space, which is of vast importance in RL-based *de novo* drug design, while overcoming the reward bottleneck by considering a fixed budget of evaluations. The intuition is that learning from diverse experiences helps the agent to explore more effectively.

Algorithm 1: Diverse Mini-Batch Selection

input: $G, B, k, \theta_0, T, p_0$
$\mathcal{M} \leftarrow \emptyset$
$\theta \leftarrow \theta_0$ ▷ Initial policy parameters
for $g = 1, \ldots, G$ **do**
  for $b = 1, \ldots, B$ **do** ▷ Generate in parallel
    $s_0 \sim p_0(\cdot)$ ▷ Sample first state
    for $t = 0, 1, \ldots, T-1$ **do**
      $a_t \sim \pi_\theta(s_t)$
      Observe next state $s_{t+1} \sim P(\cdot|s_t, a_t)$
    end for
    $\tau_b := s_0, a_0, \ldots, a_{T-1}, s_T$ ▷ Trajectory
  end for
  $\mathcal{B} \leftarrow \{\tau_1, \ldots, \tau_B\}$
  Compute kernel matrix L over $\mathcal{B}$
  Select k representative trajectories from $\mathcal{B}$
  $\forall \tau \in Y$, observe return $r(\tau)$ ▷ Evaluation
  $\mathcal{M} \leftarrow \mathcal{M} \cup (\cup_{\tau \in Y} \{\tau, r(\tau)\})$
  Update θ using RL algorithm
end for
output: $\theta, \mathcal{M}$

Therefore, we suggest enforcing diversity among the selected interactions to improve the efficiency of the exploration. For this purpose, we propose a diverse mini-batch selection framework for reinforcement learning, which is illustrated in Algorithm 1. Here, we focus on trajectories of actions, i.e., an interaction corresponds to a trajectory, which we use as a more general notion of an episode. However, the framework can easily be extended beyond trajectories/episodes. In the *de novo* drug design problem that we consider, an episode corresponds to a fully generated molecule, since each action in the episode corresponds to adding a token in the SMILES representation. Also, we consider policy-based RL, where we directly learn a policy π_θ with policy parameters θ.

Over G training/generative steps, using the agent's current policy π_θ, a batch $\mathcal{B}$ of B trajectories is sampled in parallel over copies of the same environment. Each trajectory has a maximum horizon of T steps, where the true length of each trajectory can depend on some stopping criteria or when the terminal state is reached. If B is chosen such that $B \gg k$ and the agent's policy is stochastic, this set will contain primarily unique items. We let the RL agent in each copy of the environment focus on maximizing the expected return of each trajectory with respect to the reward function, i.e., maximizing the return generated by the agent's policy $\mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$, where τ is a state-action trajectory $S_0, A_0, S_1 \ldots, S_{T-1}, A_{T-1}, S_T$, the return of following τ is denoted by $R(\tau)$ and $\mathbb{E}_{\pi_\theta}[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π_θ. This generates $\mathcal{B}$ under the belief that the agent tries to maximize

each return, without explicitly considering the diversity among individual trajectories. This can be particularly important when the agent has to consider safety concerns, is sensitive to noise, or when many trajectories are not viable. For instance, in *de novo* drug design, a SMILES string is not necessarily chemically feasible, and small changes can easily break its validity. Therefore, it is important that the agent primarily focuses on generating chemically valid SMILES strings of high quality. Moreover, the proposed method can be combined with other exploration techniques, e.g., intrinsic motivation [11, 55], to provide additional domain-specific exploration. Given a large batch of trajectories $\mathcal{B}$, to stay within the given budget of evaluations (per generative step), the next step is to obtain a smaller, diverse mini-batch Y that summarizes $\mathcal{B}$. We study the use of determinantal point processes (DPPs) [27], the MaxMin algorithm [3] and k-medoids clustering [44] for this task. After a set Y of k trajectories has been obtained, each trajectory in Y is evaluated to obtain the corresponding returns and/or state-action rewards. Using the returns and rewards, the policy parameters are updated by employing an arbitrary RL algorithm. The discussed framework is agnostic to the RL algorithm used to update the policy parameters, and yields both the policy parameters θ and a diverse set $\mathcal{M}$ of trajectory-return pairs $\{\tau, R(\tau)\}$.

3.1 Determinantal Point Processes (DPPs)

We propose and study the use of determinantal point processes (DPPs) [27] to sample a diverse mini-batch for RL updates. DPPs provide an effective framework to sample a diverse set based on specified similarity information. To the best of our knowledge, our work is a novel combination of DPP and on-policy reinforcement learning to effectively fine-tune a generative model.

A point process $\mathcal{P}$ is a probability measure over finite subsets of a set $\mathcal{B}$. We consider the discrete case of $\mathcal{B} = \{1, 2, \ldots, B\}$, where B is the number of unique trajectories. In this case, a point process is a probability measure on the power set $2^{\mathcal{B}}$, i.e., the set of all subsets of $\mathcal{B}$. DPPs are a family of point processes characterized by the *repulsion* of items such that similar items are less likely to co-occur in the same sample. Given a kernel, providing a similarity measure between pairs of items, the DPP places a high probability on subsets that are diverse with respect to the kernel. We consider a class of DPPs named L-ensembles [10], which is defined via a real, symmetric matrix L over the entire (finite) domain of $\mathcal{B}$. This matrix is often denoted as the *kernel matrix*. The probability of subset $Y \subseteq \mathcal{B}$ is given by

$$\mathcal{P}_L(Y) \propto \det(L_Y), \tag{1}$$

where $L_Y = [L_{ij}]_{i,j \in Y}$ denotes the restriction of L to the entries indexed by items of Y. Thus, the probability of sampling the set $Y \subseteq \mathcal{B}$ is proportional to the determinant of L_Y restricted to Y. The normalization constant is available in closed form since $\sum_{Y \subseteq \mathcal{B}} \det(L_Y) = \det(L + I)$, where I is the $N \times N$ identity matrix.

Given the larger set $\mathcal{B}$, we want the smaller set Y to contain a pre-defined number of items from $\mathcal{B}$. Thus, we are interested in sampling a subset Y with a fixed cardinality $|Y| = k$ to sample a mini-batch with a fixed size. k-DPPs [26] concern DPPs conditioned on the cardinality of the random subset. Formally, the probability of a k-DPP to sample a subset $Y \subseteq \mathcal{B}$ is given by

$$\mathcal{P}_L^k(Y) = \frac{\det(L_Y)}{\sum_{Y' \subseteq \mathcal{B}: |Y'|=k} \det(L_{Y'})}, \tag{2}$$

where $|Y| = k$. The k-DPP's inherent ability to promote diversity makes it an excellent choice for selecting diverse and representative mini-batches in reinforcement learning. In this way, k-DPP provides a smaller and more diverse set of items from a larger set of items.

How the k-DPP will summarize the larger set is determined by the kernel matrix L. Constructing the kernel matrix entails using domain knowledge, but other information can also be used. Let $q_i \in \mathbb{R}^+$ be a quality term and $\phi_i \in \mathbb{R}^D$, $||\phi_i|| = 1$, a vector of normalized diversity features of the i-th item in $\mathcal{B}$, e.g., the i-th generated SMILES string. Following the work of Kulesza and Taskar [27], the entries of the kernel matrix can then be expressed

$$L_{ij} = q_i \phi_i^T \phi_j q_j, \tag{3}$$

where q_i is a quality term measuring the intrinsic "goodness" of the i-th item, and $\phi_i^T \phi_j \in [-1, 1]$ is a signed measure of similarity between i-th and j-th item. Therefore, utilizing k-DPPs allows for a flexible sampling procedure that behaves differently depending on the information incorporated in the kernel matrix L. It does not directly optimize the determinant of L, but instead includes randomness to encourage additional exploration. In the *de novo* drug design problem studied in this paper, we only consider the similarity between items and do not explore the effects of quality terms. The reason for this is that we focus on pure diversification, and we assume that the items generated by the policy have similar quality. Our preliminary studies on *de novo* drug design did not find any performance gain in incorporating a quality term provided by an oracle. However, we believe that it can be beneficial to include a quality term, but different terms need to be investigated to find a suitable one.

3.2 Maximum Coverage

As an alternative to selecting a representative set by sampling via k-DPPs, we also study mini-batch diversification by maximizing the coverage of the larger set for a fixed cardinality. While k-DPPs provide a sampling procedure to summarize a larger set given a kernel matrix, maximum coverage aims to directly cover as large a part of the space as possible. In this way, we seek to pick the most diverse items subject to the cardinality constraint of k. For a given set $\mathcal{B}$ of B candidate items, let $f(Y)$ be a function that measures the "coverage" of any given set Y of items. The goal is to choose a set Y of k items such that $f(Y)$ is maximized. Here we consider a fixed size of k, but a possible extension could be to choose the smallest set Y such that a sufficient coverage of $\mathcal{B}$ is obtained. Formally, we define this problem by

$$\max_{Y \in [\mathcal{B}]^k} f(Y), \tag{4}$$

where $[\mathcal{B}]^k \triangleq \{X \in 2^{\mathcal{B}} : |X| = k\}$ is the set of all subsets with cardinality k. In this work, we consider coverage functions $f(Y)$ based on dissimilarities between trajectories.

We investigate two algorithms to find an approximate maximum coverage of the large set: (1) the MaxMin algorithm [3], implemented by RDKit [28]; (2) k-medoids clustering [44], using the FasterPAM algorithm [47, 48] implemented by Schubert and

Lenssen [46]. The MaxMin algorithm first picks a starting item, creating a picked set. Then the algorithm iteratively, from the items in the candidate pool, finds the item that has the maximum dissimilarity to molecules in the picked set and adds this item to the picked set. The MaxMin algorithm is widely used in drug discovery to pick a diverse set [15, 54].

k-medoids clustering [44] is a popular technique to cluster non-Euclidean data using arbitrary dissimilarities or input domains. The k-medoids problem aims to split B items into k clusters, where the number of clusters is assumed to be specified beforehand. The medoid of a cluster is defined as the item in the cluster with the minimum average of dissimilarity to all the other items in the cluster, i.e., the item that is most centrally located within the cluster. Unlike several other clustering algorithms, e.g., k-means [1], the medoid is an actual item in the cluster. Thus, the objective is to find a cluster assignment $C_1, \ldots, C_k$ that minimizes

$$\sum_{i=1}^{k} \sum_{x_c \in C_i} d(x_c, m_i), \tag{5}$$

where m_i is the medoid of cluster C_i and d is an arbitrary dissimilarity function. The medoid m of cluster C is defined by $\text{medoid}(C) := \text{argmin}_{x \in C} \sum_{x_c \in C} d(x_c, x)$. While the MaxMin algorithm sequentially adds items to the picked set in a greedy manner, k-medoids simultaneously seeks to optimize all medoids to find the best picks. Finding the global optimum of the k-medoid problem is NP-hard [25]. Instead, the Partitioning Around Medoids (PAM) algorithm [44], which is the standard algorithm for k-medoids clustering, improves the clustering towards a local optimum. In this paper, we use the FasterPAM algorithm [47, 48], which achieves a speedup in runtime compared to the original PAM algorithm, to select k items given by the medoids found by the algorithm.

4 EXPERIMENTAL EVALUATION

We extensively evaluate our framework on *de novo* drug design. We run experiments on three reward functions based on well-established molecule binary bioactivity label optimization tasks: the Dopamine Receptor D2 (DRD2), c-Jun N-terminal Kinases-3 (JNK3), and Glycogen Synthase Kinase 3 Beta (GSK3β) predictive activity models [29, 36] provided by Therapeutics Data Commons [58]. The final (extrinsic) reward also includes the quantitative estimation of drug-likeness (QED) [7], molecular weight, number of hydrogen bond donors, and structural constraints. For full details on the reward functions, see the supplementary material.[1]

To update the policy and generate SMILES, we use the REINVENT framework [31] with its pre-trained policy on the ChEMBL database [18] to generate drug-like bioactive molecules. Previous benchmarks on *de novo* drug design have, for this framework, concluded among the best performances [17, 56], while it is also used in real-world applications [42]. The action space $\mathcal{A}$ consists of 34 tokens, including start and stop tokens.

To measure diversity within a set of molecules, several metrics have been proposed. Hu et al. [23] divides these metrics into two main categories: reference-based and distance-based. A reference-based metric compares a molecular set with a reference set to find the intersection. Distance-based metrics use pairwise distances within the molecular set to measure diversity. We measure the reference-based diversity of the trajectories in $\mathcal{M}$ by the number of molecular scaffolds, also known as Bemis-Murcko scaffolds [6]. Moreover, we evaluate the distance-based diversity by the number of diverse actives[2] metric by Renz et al. [45], which is based on #Circles metric proposed by Xie et al. [61]. When evaluating diversity, we consider only active molecules, defined as those with both QED and predicted activity greater than 0.5.

We compare the use of mini-batch diversification in combination with different techniques to modify the original reward for *de novo* drug design: (1) no modification of the reward, i.e., the agent observes the original (extrinsic) reward; (2) using the popular identical molecular scaffold (IMS) penalty [9], which sets the reward to 0 when M molecules with the same molecular scaffold have been generated; (3) using the TanhRND technique [21], which shows promising empirical results in terms of diversity. No modification of the reward is included as a baseline to investigate if mini-batch diversification can act as an alternative approach to avoid mode collapse by modifying the original reward. We hereafter denote the original reward without any modification as the *extrinsic reward*. For mini-batch diversification with a mini-batch of $k = 64$ SMILES, we first generate $B = 640$ SMILES via multinomial sampling and then select a diverse mini-batch of k SMILES using one of the methods discussed above. Without mini-batch diversification, we directly generate $k = 64$ SMILES via multinomial sampling, which is the standard procedure of the REINVENT framework. We denote these approaches without mini-batch diversification as *diversification-free*.

4.1 Construction of Kernel Matrix

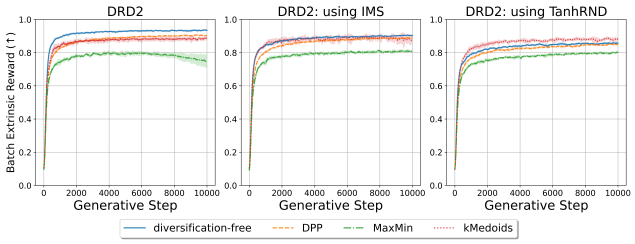
All of the investigated methods for mini-batch diversification (i.e., DPP, the MaxMin algorithm and k-medoids clustering) rely on a kernel matrix L to encode the similarity between different molecules. We construct this kernel matrix based on two other kernel matrices L_T and L_D, which we denote as "base" kernel matrices. The first base kernel matrix L_T is constructed by the Tanimoto similarity between the corresponding 2048-bit Morgan fingerprints (with radius 2 using RDKit [28]) of the generated SMILES. To incorporate more scaffold-based information, we construct the base kernel matrix L_D by computing the Dice coefficients [14, 53] between the scaffolds' atom pair fingerprints [12]. Given these base kernels, we aggregate these base kernel matrices to define the kernel matrix L, which is used for selecting k molecules, by $L = L_T + L_D$. In the supplementary material, we present a study of different combinations of base matrices to define L and argue that the kernel matrix defined here provides the best balance across the diversity metrics.

4.2 Evaluation of Quality

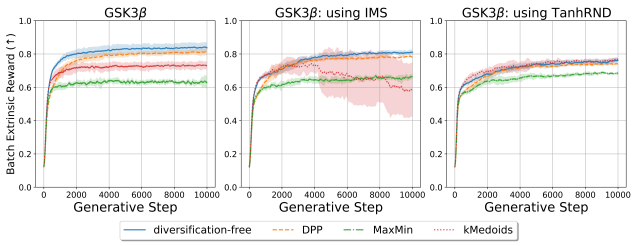
We first assess the quality (i.e., reward) of the generated molecules to evaluate if our proposed framework can maintain high quality while enhancing diversity. Therefore, we study the extrinsic reward of each configuration. The extrinsic reward is the original reward provided for each molecule that we want to maximize, but not the reward observed by the agent when using IMS or TanhRND.

[1] A version including supplementary material is available at https://doi.org/10.48550/arXiv.2506.21158

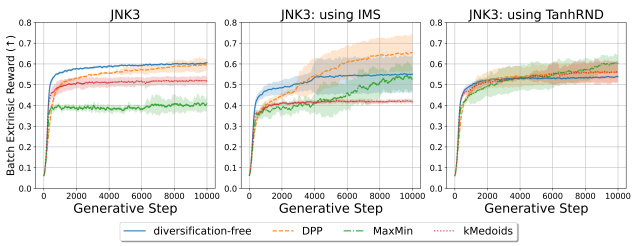
[2] Diverse actives is termed diverse hits in previous work by Renz et al. [45].



(a) DRD2



(b) GSK3β



(c) JNK3

Figure 1: Average extrinsic reward across mini-batch.

Figure 1 displays the average extrinsic rewards for each mini-batch Y of SMILES evaluated in each generative step. The average across 10 independent runs per generative step is plotted over 10 000 generative steps, where the shaded area shows the corresponding standard deviation across the independent runs. For clarity of presentation, we show the moving averages with a window size of 101 (i.e., the current step and upto 50 steps on each side). Each plot of Figures 1a to 1c compares the use of diverse mini-batch selection using k-DPP, the MaxMin algorithm and k-medoids clustering in combination with different techniques of modifying the extrinsic reward for *de novo* drug design. The left plots compare extrinsic rewards with and without mini-batch diversification when the extrinsic reward is not modified. The middle plots compare the extrinsic rewards when using the identical molecular scaffold (IMS) penalty [9] and the right plots display the comparisons when utilizing the TanhRND technique [21].

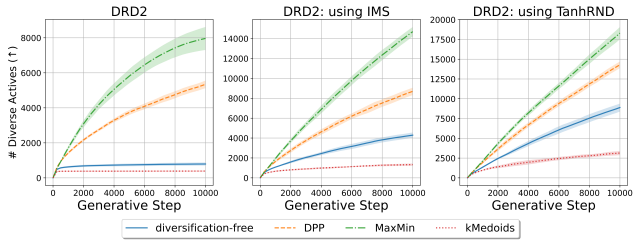
For the DPP and diversification-free methods on the DRD2- and GSK3β-based reward functions (see Figures 1a and 1b), we observe similar trends in terms of extrinsic reward, especially when using IMS or TanhRND. Moreover, on the DRD2 reward, these experiments achieve a reward of 0.8 or higher, while rewards close to 0.8 are achieved on the GSK3β function. The diversification-free experiments converge faster, but the DPP experiments often converge to a similar average reward. Faster convergence tends to indicate that less exploration is performed, which is demonstrated in Figures 2 and 3 below in terms of diversity of the generated molecules. k-medoids shows similar results on DRD2, but achieves more unstable and lower quality on GSK3β. For the MaxMin experiments on the DRD2 and GSK3β problems, we observe that extrinsic rewards are lower than for both the DPP and diversification-free experiments. This is possibly because more exploration is enforced, due to a more diverse mini-batch, at the cost of less exploitation. For the experiments on the JNK3-based reward function (see Figure 1c), we observe similar trends as for DRD2 and GSK3β when not modifying the extrinsic reward (see left plot in Figure 1c). On the other hand, when using the IMS or TanhRND technique to modify the extrinsic reward, all methods display similar extrinsic reward, but different convergence rates. Only k-medoids utilizing IMS performs differently, displaying an early convergence to a reward of around 0.4, which is lower than the other methods. This is likely due to insufficient exploration induced by this configuration. In general, the extrinsic rewards are significantly lower on JNK3, indicating that the JNK3-based reward function is more challenging to optimize. One possible explanation is that there are fewer active molecules for JNK3 in the ChEMBL database. When we evaluate molecules from ChEMBL on the DRD2, GSK3β, and JNK3 oracles, we observe that 2.4%, 1.8%, and 0.3% of the molecules, respectively, have an oracle score above 0.5 (we refer to the supplementary material for more details). Thus, there are fewer good solutions for JNK3. Since we use a model pre-trained on ChEMBL data, which limits the generation to molecules similar to those found in this data, the initial model is less likely to find sufficient solutions for JNK3.

4.3 Evaluation of Distance-Based Diversity

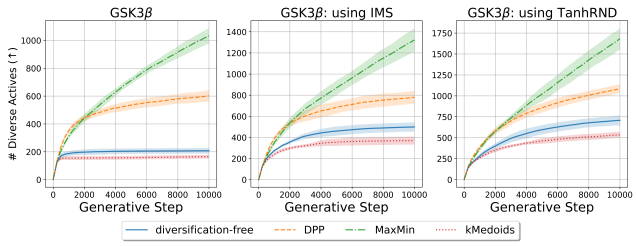
To evaluate the distance-based diversity among the generated molecules, we calculate the number of diverse actives. We use Tanimoto dissimilarity to measure the distance between 2048-bit Morgan fingerprints (with radius 2 and computed by RDKit [28]) and the distance threshold $D = 0.7$ proposed by Renz et al. [45]. Figure 2 shows the total number of diverse actives for every 250th generative step in the *de novo* drug design task for the DRD2-, GSK3β- and JNK3-based reward functions. The lines and shaded area display the mean and standard deviation, respectively, across 10 independent reruns for each configuration. Each plot of Figures 2a to 2c compares the use of DPP in combination with different techniques of modifying the extrinsic reward for *de novo* drug design.

4.3.1 DRD2. Figure 2a displays the cumulative number of diverse actives per generative step on the DRD2-based reward function. We observe that utilizing mini-batch diversification significantly improves the total number of diverse actives found over 10000 generative steps compared to the diversification-free experiments. We observe a significant gain after just a few hundred generative steps. In particular, MaxMin consistently yields the best results in terms of diverse actives, compared to the other methods, and displays a consistent increase in the number of diverse actives found.

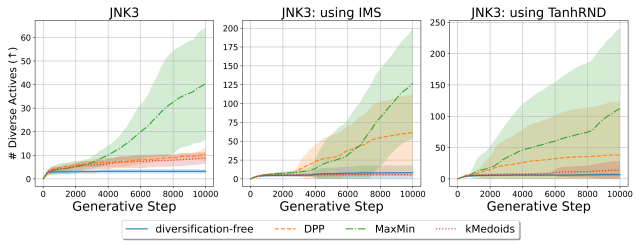
Interestingly, when not using IMS or TanhRND to modify the extrinsic reward (see left plot in Figure 2a), DPP and MaxMin display a considerable increase in distance-based diversity after a few hundred generative steps compared to the diversification-free method,



(a) DRD2



(b) GSK3β



(c) JNK3

Figure 2: Number of diverse activities after g generative steps.

where diversity quickly stagnates. Without mini-batch diversification and any extrinsic reward modification, it is expected that the diversity should stagnate since it has previously been observed that the agent can easily get stuck in a local optimum and will then generate similar molecules [9]. Using mini-batch diversification via DPP or MaxMin overcomes this issue even without modifying the extrinsic reward, which is the standard method for tackling this issue. In addition, we observe that mini-batch diversification in combination with a modification of the extrinsic reward (see the middle and right plot in Figure 2a) yields the largest number of diverse actives, especially when utilizing TanhRND. However, using k-medoids for mini-batch diversification generates fewer diverse activities than the diversification-free methods.

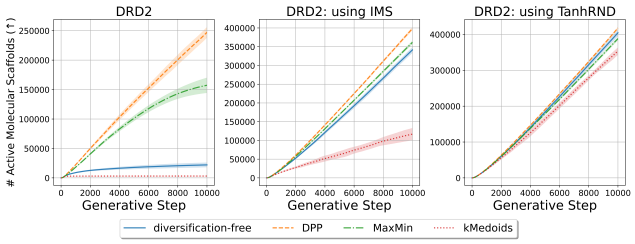
4.3.2 GSK3β. Figure 2b displays the cumulative number of diverse actives per generative step on the GSK3β-based reward function. We observe that utilizing mini-batch diversification via DPP or MaxMin generates significantly more diverse active after a few hundred generative steps. We see this behaviour no matter if we modify the extrinsic reward or not, meaning that mini-batch diversification can successfully be used as an exploration technique to overcome mode collapse and lead to diverse behaviors. Moreover, we notice that, after at most 4000 generative steps, MaxMin yields substantially more diverse actives than the other methods. Also, we note that, similar to the experiments on the DRD2-based reward functions, using k-medoids yields a substantially lower number of diverse actives than the other methods, including diversification-free methods.

4.3.3 JNK3. Figure 2c shows the cumulative number of diverse actives per generative step on the JNK3-based reward function. Firstly, we observe a high standard deviation among all experiments, compared to the other reward functions. This is likely since the JNK3 oracle is more difficult to optimize than the other oracles, and therefore does not have a large margin to the activity threshold of 0.5 for diverse actives. Similar trends in terms of diversity have been observed by previous work [21]. Most approaches using mini-batch diversification keep improving over a large number of generative steps, while the diversification-free experiments generally show a substantially lower number of average diverse actives. For no extrinsic reward modification (see left plot in Figure 2c), MaxMin generates the highest average number of diverse actives, while DPP has lower variance but yields fewer diverse actives. When using the IMS or TanhRND strategy to modify the reward (see middle and right plot in Figure 2c), MaxMin also yields the highest average number of diverse actives, but the runs overlap with DPP since both have high variance. For the experiments using TanhRND (see right plot in Figure 2c), all MaxMin configurations display a larger increase in the average number of diverse actives over time. On this reward function, k-medoids can generate more diverse actives than the diversification-free method when not modifying the (extrinsic) reward, while these two methods display similar performance when modifying the reward.

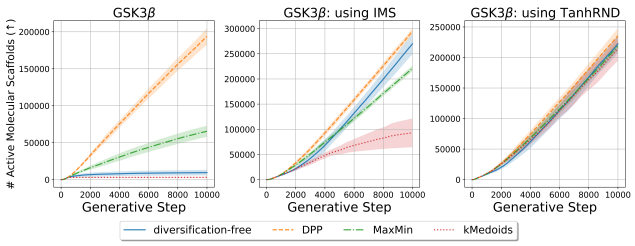
4.4 Evaluation of Reference-Based Diversity

To assess reference-based diversity, we use the number of molecular scaffolds, computed with RDKit [28]. Figure 3 shows the cumulative number of unique active molecular scaffolds per generative step for the DRD2-, GSK3β- and JNK3-based reward functions. The lines and shaded area display the mean and standard deviation, respectively, across 10 independent reruns for each configuration.

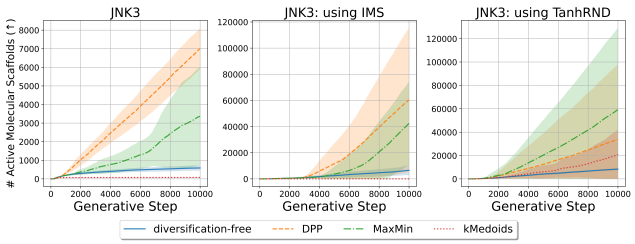
4.4.1 DRD2. Figure 3a displays the cumulative number of active molecular scaffolds, per generative step, evaluated on the DRD2-based reward function. When not modifying the extrinsic reward (see left plot in Figure 3a), using mini-batch diversification via DPP or MaxMin leads to substantially more scaffolds, compared to the diversification-free method, after less than 750 generative steps. In particular, our experiments demonstrate that DPP generates most scaffolds on average. When utilizing the identical molecular scaffold (IMS) penalty [9] for modifying the extrinsic reward (see middle plot in Figure 3a), we observe that DPP generates more molecular scaffolds compared to the other methods. For the TanhRND technique (see right plot in Figure 3a), the diversification-free, MaxMin and DPP methods show similar diversity in terms of molecular scaffolds and perform on par with the best methods when using IMS (see middle plot in Figure 3a). In terms of molecular scaffolds, it is clear that the scaffold-based similarity that mini-batch diversification provides can be important, especially in combination with no or less effective exploration techniques. However, across all experiments, it is clear that k-medoids generates the least amount of



(a) DRD2



(b) GSK3β



(c) JNK3

Figure 3: Number of active molecular scaffolds.

scaffolds, and it is therefore important to choose an appropriate method for mini-batch diversification.

4.4.2 GSK3β. Figure 3b displays the cumulative number of molecular scaffolds for the evaluation on the GSK3β-based reward function. Without any modification of the extrinsic reward (see left plot in Figure 3b), we observe that mini-batch diversification via DPP or MaxMin yields significantly more scaffolds compared to the diversification-free method. The DPP effectively generates more molecular scaffolds, while MaxMin is less effective. For reward modification (see the middle and right plots in Figure 3b), we observe that using mini-batch diversification via DPP generates more scaffolds on average and has lower variance. However, the difference in effectiveness of using DPP is reduced in terms of diverse actives, but DPP can still consistently improve diversity. For MaxMin, which consistently generates the largest number of diverse actives (see Figure 2b), we observe a lower number of scaffolds. Thus, when using the MaxMin algorithm to impose mini-batch diversity, we see that high distance-based diversity does not directly result in high reference-based diversity, and vice versa. When using mini-batch diversification via k-medoids, it generates significantly fewer scaffolds, except when using TanhRND, where it performs on par with the other methods.

4.4.3 JNK3. Figure 3c displays the scaffold diversity for the evaluation on the JNK3-based reward function. When not modifying the extrinsic reward (see left plot in Figure 3c), all DPP-based methods are more effective after around 2000 generative steps. For DPP, we observe the largest average number of molecular scaffolds and notice a more consistent exploration, since the rate of diverse solutions is higher. The MaxMin algorithm does not display the same consistent improvement in the number of scaffolds. When modifying the extrinsic reward (see middle and right plots in Figure 3c), both DPP and MaxMin obtain a higher average number of scaffolds, but they also display a high variability and are therefore not always more effective. This is likely because the agent is not able to effectively optimize the reward (see Figure 1c). In general, as depicted in Figure 1c, the JNK3-based reward is more difficult to optimize for the RL agent. Thus, we notice that the robustness of our proposed mini-batch diversification depends on how well the agent can optimize the given task. This is expected since the mini-batch selection depends on the given larger set $\mathcal{B}$ and, therefore, has limited capabilities to enhance the diversity if the RL agent itself cannot find sufficient solutions.

5 CONCLUSIONS

In this work, we present an easily applicable framework for enhancing mini-batch diversity in on-policy reinforcement learning algorithms. The framework seeks to address the problem of efficient exploration when obtaining a reward is costly. In this paper, we apply our framework to *de novo* drug design; however, it is problem-agnostic. We believe the proposed framework can also be beneficial in other RL applications where efficient exploration and diverse behaviors are crucial. To solve the problem of mini-batch diversification in RL, we study the use of determinantal point processes (DPPs) [27], the MaxMin algorithm [3] and k-medoids clustering [44] for the diversification process. In this way, we seek to summarize a larger set of molecules by selecting a smaller mini-batch of diverse molecules to evaluate, requiring fewer evaluations. We argue that this enhances exploration by focusing on promising, more diverse molecules while keeping rewards high. We demonstrate that our proposed framework for mini-batch diversification can substantially improve the diversity of *de novo* drug design, especially when combined with a domain-specific modification of the extrinsic reward, such as TanhRND [21]. We notice that if the agent alone provides sufficient solutions, our framework can substantially enhance the diversity of the generated solutions. In particular, we observe that DPP-based mini-batch diversification enhances both distance- and reference-based diversity, while the MaxMin algorithm primarily improves distance-based diversity. Therefore, we propose using DPP for diversification, as it allows a more adaptable kernel matrix and a natural way to introduce randomness.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The experimental evaluation was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. We thank Graham Kemp for his valuable input on the manuscript.

REFERENCES

[1] David Arthur and Sergei Vassilvitskii. 2007. k-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, USA, 1027–1035.

[2] Josep Arús-Pous, Thomas Blaschke, Silas Ulander, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. 2019. Exploring the GDB-13 chemical space using deep generative models. *Journal of cheminformatics* 11 (2019), 1–14.

[3] Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. 2002. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure-Activity Relationships* 21, 6 (2002), 598–604.

[4] Sara Romeo Atance, Juan Viguera Diez, Ola Engkvist, Simon Olsson, and Rocío Mercado. 2022. De novo drug design using reinforcement learning with graph-based deep generative models. *Journal of chemical information and modeling* 62, 20 (2022), 4863–4872.

[5] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. 2020. Never Give Up: Learning Directed Exploration Strategies. In *International Conference on Learning Representations*. International Conference on Learning Representations (ICLR), USA. https://openreview.net/forum?id=Sye57xStvB

[6] Guy W Bemis and Mark A Murcko. 1996. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* 39, 15 (1996), 2887–2893.

[7] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. 2012. Quantifying the chemical beauty of drugs. *Nature chemistry* 4, 2 (2012), 90–98.

[8] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. 2020. REINVENT 2.0: an AI tool for de novo drug design. *Journal of chemical information and modeling* 60, 12 (2020), 5918–5922.

[9] Thomas Blaschke, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. 2020. Memory-assisted reinforcement learning for diverse molecular de novo design. *Journal of cheminformatics* 12, 1 (2020), 68.

[10] Alexei Borodin and Eric M Rains. 2005. Eynard–Mehta theorem, Schur process, and their Pfaffian analogs. *Journal of statistical physics* 121 (2005), 291–317.

[11] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by random network distillation. In *International Conference on Learning Representations*. International Conference on Learning Representations (ICLR), USA. https://openreview.net/forum?id=H1lJJnR5Ym

[12] Raymond E Carhart, Dennis H Smith, and RENGACHARI Venkataraghavan. 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* 25, 2 (1985), 64–73.

[13] Jiayu Chen, Vaneet Aggarwal, and Tian Lan. 2023. A unified algorithm framework for unsupervised discovery of skills based on determinantal point process. *Advances in Neural Information Processing Systems* 36 (2023), 67925–67947.

[14] Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 3 (1945), 297–302.

[15] Gabriel HS Dreiman, Magda Bictash, Paul V Fish, Lewis Griffin, and Fredrik Svensson. 2021. Changing the HTS paradigm: AI-driven iterative screening for hit finding. *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 26, 2 (2021), 257–262.

[16] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 79858–79885.

[17] Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. 2022. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., USA, 21342–21357.

[18] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. 2017. The ChEMBL database in 2017. *Nucleic acids research* 45, D1 (2017), D945–D954.

[19] Eoin Martino Grua and Mark Hoogendoorn. 2018. Exploring clustering techniques for effective reinforcement learning based personalization for health and wellbeing. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, USA, 813–820.

[20] Hampus Gummesson Svensson, Christian Tyrchan, Ola Engkvist, and Morteza Haghir Chehreghani. 2024. Utilizing reinforcement learning for de novo drug design. *Machine Learning* 113, 7 (2024), 4811–4843.

[21] Hampus Gummesson Svensson, Christian Tyrchan, Ola Engkvist, and Morteza Haghir Chehreghani. 2025. Diversity-Aware Reinforcement Learning for de novo Drug Design. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, James Kwok (Ed.). International Joint Conferences on Artificial Intelligence Organization, USA, 9194–9204. https://doi.org/10.24963/ijcai.2025/1022 AI4Tech: AI Enabling Technologies.

[22] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*. PMLR, USA, 1352–1361.

[23] Xiuyuan Hu, Guoqing Liu, Quanming Yao, Yang Zhao, and Hao Zhang. 2024. Hamiltonian diversity: effectively measuring molecular diversity by shortest Hamiltonian circuits. *Journal of Cheminformatics* 16, 1 (2024), 94.

[24] Wanming Huang, Richard Yi Da Xu, and Ian Oppermann. 2019. Efficient diversified mini-batch selection using variable high-layer features. In *Asian Conference on Machine Learning*. PMLR, USA, 300–315.

[25] Oded Kariv and S Louis Hakimi. 1979. An algorithmic approach to network location problems. I: The p-centers. *SIAM journal on applied mathematics* 37, 3 (1979), 513–538.

[26] Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Omnipress, Madison, WI, USA, 1193–1200.

[27] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning* 5, 2–3 (2012), 123–286.

[28] Greg Landrum. 2006. RDKit: Open-source cheminformatics. http://www.rdkit.org.

[29] Yibo Li, Liangren Zhang, and Zhenming Liu. 2018. Multi-objective de novo drug design with conditional graph generative model. *Journal of cheminformatics* 10 (2018), 1–24.

[30] Yong Liu, Zhiqi Shen, Yinan Zhang, and Lizhen Cui. 2022. Diversity-promoting deep reinforcement learning for interactive recommendation. In *5th international conference on crowd science and engineering*. Association for Computing Machinery, New York, NY, USA, 132–139.

[31] Hannes H Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. 2024. Reinvent 4: Modern AI–driven generative molecule design. *Journal of Cheminformatics* 16, 1 (2024), 20.

[32] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, USA, 1928–1937.

[33] Varnavas D Mouchlis, Antreas Afantitis, Angela Serra, Michele Fratello, Anastasios G Papadiamantis, Vassilis Aidinis, Iseult Lynch, Dario Greco, and Georgia Melagraki. 2021. Advances in de novo drug design: from conventional to machine learning methods. *International journal of molecular sciences* 22, 4 (2021), 1676.

[34] Elvis Nava, Mojmir Mutny, and Andreas Krause. 2022. Diversified sampling for batched Bayesian optimization with determinantal point processes. In *International Conference on Artificial Intelligence and Statistics*. PMLR, USA, 7031–7054.

[35] Brendan O'Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. 2017. Combining policy gradient and Q-learning. In *5th International Conference on Learning Representations (ICLR 2017)*. International Conference on Learning Representations (ICLR), USA. https://openreview.net/forum?id=B1kJ6H9ex

[36] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. 2017. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* 9 (2017), 1–14.

[37] Takayuki Osogami and Rudy Raymond. 2019. Determinantal reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. Association for the Advancement of Artificial Intelligence, USA, 4659–4666.

[38] Joseph M Paggi, Ayush Pandit, and Ron O Dror. 2024. The art and science of molecular docking. *Annual review of biochemistry* 93, 1 (2024), 389–410.

[39] Chao Pang, Jianbo Qiao, Xiangxiang Zeng, Quan Zou, and Leyi Wei. 2023. Deep generative models in de novo drug molecule generation. *Journal of Chemical Information and Modeling* 64, 7 (2023), 2174–2194.

[40] Jinyeong Park, Jaegyoon Ahn, Jonghwan Choi, and Jibum Kim. 2024. Mol-AIR: Molecular Reinforcement Learning with Adaptive Intrinsic Rewards for Goal-directed Molecular Generation. arXiv:2403.20109 [cs.LG] https://arxiv.org/abs/2403.20109

[41] Atanas Patronov, Kostas Papadopoulos, and Ola Engkvist. 2021. Has artificial intelligence impacted drug discovery? In *Artificial Intelligence in Drug Design*. Springer, USA, 153–176.

[42] Will R. Pitt, Jonathan Bentley, Christophe Boldron, Lionel Colliandre, Carmen Esposito, Elizabeth H. Frush, Jola Kopec, Stéphanie Labouille, Jerome Meneyrol, David A. Pardoe, Ferruccio Palazzesi, Alfonso Pozzan, Jacob M. Remington, René Rex, Michelle Southey, Sachin Vishwakarma, and Paul Walker. 2025. Real-World Applications and Experiences of AI/ML Deployment for Drug Discovery. *Journal of Medicinal Chemistry* 68, 2 (2025), 851–859. https://doi.org/10.1021/acs.jmedchem.4c03044

[43] Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. 2013. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* 27 (2013), 675–679.

[44] LKPJ Rdusseeun and P Kaufman. 1987. Clustering by means of medoids. In *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland, 31 August–4 September 1987*, Vol. 31.

[45] Philipp Renz, Sohvi Luukkonen, and Guünter Klambauer. 2024. Diverse Hits in De Novo Molecule Design: Diversity-Based Comparison of Goal-Directed

Generators. *Journal of Chemical Information and Modeling* 64, 15 (2024), 5756–5761.

[46] Erich Schubert and Lars Lenssen. 2022. Fast k-medoids Clustering in Rust and Python. *Journal of Open Source Software* 7, 75 (2022), 4183.

[47] Erich Schubert and Peter J Rousseeuw. 2019. Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In *International conference on similarity search and applications*. Springer, USA, 171–187.

[48] Erich Schubert and Peter J Rousseeuw. 2021. Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems* 101 (2021), 101804.

[49] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[50] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. 2018. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* 4, 1 (2018), 120–131.

[51] Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2021. State entropy maximization with random encoders for efficient exploration. In *38th International Conference on Machine Learning*. PMLR, USA, 9443–9454.

[52] Hassam Sheikh, Kizza Frisbee, and Mariano Phielipp. 2022. DNS: Determinantal point process based neural network sampler for ensemble reinforcement learning. In *International Conference on Machine Learning*. PMLR, USA, 19731–19746.

[53] Thorvald Sorensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske skrifter* 5 (1948), 1–34.

[54] Youhai Tan, Lingxue Dai, Weifeng Huang, Yinfeng Guo, Shuangjia Zheng, Jinping Lei, Hongming Chen, and Yuedong Yang. 2022. DRlinker: deep reinforcement learning for optimization in fragment linking design. *Journal of Chemical Information and Modeling* 62, 23 (2022), 5907–5917.

[55] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. 2017. #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., USA, 2753–2762. https://proceedings.neurips.cc/paper_files/paper/2017/file/3a20f62a0af1aa152670bab3c602feed-Paper.pdf

[56] Morgan Thomas, Noel M. O'Boyle, Andreas Bender, and Chris De Graaf. 2022. Re-evaluating sample efficiency in de novo molecule generation. arXiv:2212.01385 [cs.CE] https://arxiv.org/abs/2212.01385

[57] Xiaochu Tong, Xiaohong Liu, Xiaoqin Tan, Xutong Li, Jiaxin Jiang, Zhaoping Xiong, Tingyang Xu, Hualiang Jiang, Nan Qiao, and Mingyue Zheng. 2021. Generative models for de novo drug design. *Journal of Medicinal Chemistry* 64, 19 (2021), 14011–14027.

[58] Alejandro Velez-Arce, Kexin Huang, Michelle M Li, xiang lin, Wenhao Gao, Bradley Pentelute, Tianfan Fu, Manolis Kellis, and Marinka Zitnik. 2024. Signals in the Cells: Multimodal and Contextualized Machine Learning Foundations for Therapeutics. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*.

[59] Jing Wang and Fei Zhu. 2024. ExSelfRL: An exploration-inspired self-supervised reinforcement learning approach to molecular generation. *Expert Systems with Applications* 260 (2024), 125410.

[60] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28, 1 (1988), 31–36.

[61] Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. 2023. How Much Space Has Been Explored? Measuring the Chemical Space Covered by Databases and Machine-Generated Molecules. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. International Conference on Learning Representations (ICLR), USA.

[62] Yaodong Yang, Ying Wen, Jun Wang, Liheng Chen, Kun Shao, David Mguni, and Weinan Zhang. 2020. Multi-agent determinantal q-learning. In *International Conference on Machine Learning*. PMLR, USA, 10757–10766.

[63] Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 110935–110971.

[64] C Zhang, H Kjellström, and S Mandt. 2017. Determinantal point processes for mini-batch diversification. In *Uncertainty in Artificial Intelligence-Proceedings of the 33rd Conference, UAI 2017*. AUAI Press, USA.

[65] Kaiyan Zhao, Yiming Wang, Yuyang Chen, Yan Li, Leong Hou U, and Xiaoguang Niu. 2025. Efficient Diversity-based Experience Replay for Deep Reinforcement Learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, James Kwok (Ed.). International Joint Conferences on Artificial Intelligence Organization, USA, 7083–7091. https://doi.org/10.24963/ijcai.2025/788