How Much of the Chemical Space Has Been Explored? Selecting the Right Exploration Measure for Drug Discovery

Anonymous Authors[1]

Abstract

Forming a molecular candidate set that contains a wide range of potentially effective compounds is crucial to the success of drug discovery. While many aim to optimize particular chemical properties, there is limited literature on how to properly measure and encourage the exploration of the chemical space when generating drug candidates. This problem is challenging due to the lack of formal criteria to select good exploration measures. We propose a novel framework to systematically evaluate exploration measures for drug candidate generation. The procedure is built upon three formal analyses: an axiomatic analysis that validates the potential measures analytically, an empirical analysis that compares the correlations of the measures to a proxy gold standard, and a practical analysis that benchmarks the effectiveness of the measures in an optimization procedure of molecular generation. We are able to evaluate a wide range of potential exploration measures under this framework and make recommendations on existing and novel exploration measures that are suitable for the task of drug discovery.

1. Introduction

Drug discovery aims to find molecules that can effectively bind to certain targets, which is often compared to finding needles in a haystack. Indeed, it is estimated that the small organic chemical space has more than 10^{60} molecules (Kirkpatrick & Ellis, 2004; Ruddigkeit et al., 2012). In such a tremendous space, only a small number of molecules can satisfy the bioactivity requirement for a specific target. In contrast, the space that has been explored by the scientists is very limited: the largest authoritative molecular database contains only around 10^8 compounds (Kim et al., 2020).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Machine learning based approaches have demonstrated great potential in helping scientists go beyond the known molecules and efficiently navigate through the huge chemical space, usually through *de novo* molecular generation (Elton et al., 2019; Schwalbe-Koda & Gómez-Bombarelli, 2020; Bian & Xie, 2021; Deng et al., 2022). Such molecular generation methods learn to generate candidate drug designs by optimizing various molecular property scores, like the binding affinity scores that are relevant to the drug efficacy. In practice, these scores can be computationally obtained using biological activity prediction models (Olivecrona et al., 2017; Li et al., 2018), which is the key to obtain massive labeled training data for machine learning.

However, there is a huge gap between the *in silico* property scores of a drug design and its *in vivo* efficacy, as such computationally obtained scores provide limited information about the complex chemical interactions in the real world. Expensive wet-lab experiments are still required to verify the true effectiveness of each candidate molecule. To increase the chance of "hits" in these experiments, it is crucial to generate a variety of qualified candidates that explore (or cover) a broad area of the chemical space (Huggins et al., 2011; Wawer et al., 2014; Ashenden, 2018), rather than a concentrated cluster of molecules with high property scores.

A fundamental question is how to properly measure to what extent a set of candidate molecules explores the chemical space. A reliable way to estimate the variety of a set of molecules is to test their varieties of inhibitory capability, such as half maximal inhibitory concentration (IC50) across a range of biological targets, through wet-lab experiments (Yung-Chi & Prusoff, 1973). These experimental data (such as IC50) provide important reference information for drug design but they are expensive to annotate. An easy-to-calculate exploration measure that faithfully reflects the variety of functionalities is extremely valuable. While a number of studies have evaluated more or less related properties of generated compounds such as *internal diversity* (Brown et al., 2019; Polykovskiy et al., 2020), the validity of these measures is rarely justified. In fact, finding a good measure of exploration itself is challenging. Compared to molecular property scores, the ground truth about the coverage of biological functionalities is harder to obtain. Moreover, the

chemical space is complex and combinatorial, which makes the design of a good measure even more difficult.

In this paper, we address this challenge with a systematic study on measuring to what extent a set of molecular candidates explores the underlining chemical space. We propose a novel framework to select proper exploration measures for molecular generation. The framework is built upon three formal criteria, based on which we are able to investigate and evaluate a series of existing and new measures: (C1) The first criterion is based on an axiomatic analysis with two intuitive axioms that a good exploration measure should satisfy. Surprisingly, multiple heuristic measures that are commonly used in literature, such as *internal diversity*, fail to satisfy one or both of the axioms. (C2) The second criterion compares the exploration measures with a proxy of gold standard: the number of unique biological functionalities covered by the set of molecules. We find that a novel exploration measure not only satisfies both axioms but also better correlates with the gold standard. (C3) The third criterion further evaluates the selected measures in an optimization setting, *i.e.*, jointly optimizing the molecular property scores and each of the exploration measures through a state-of-the-art molecular generation model. We find that by adding to the objective function novelty terms derived from proper exploration measures, the molecular generation models can discover high-quality molecules that span a larger chemical space, in comparison to optimizing molecular properties alone or with alternative exploration measures.

2. Related Work

2.1. Molecular Generation

Most research in drug discovery lays their focus on optimizing molecular properties of generated compounds. Traditionally, rule-based genetic algorithms (GA) and fragment-based combinatorial methods are widely applied to find drug candidates (Brown et al., 2004; Devi et al., 2015; Jensen, 2019). Machine learning, especially deep learning models are proposed to optimize molecular property scores at better efficiency. One such example is the variational autoencoder (VAE), which uses Bayesian inference to find optimal molecules in the latent space (Gómez-Bombarelli et al., 2018; Liu et al., 2018; Jin et al., 2018). Another branch of methods employs reinforcement learning to generate compounds in the explicit chemical space (Olivecrona et al., 2017; You et al., 2018; De Cao & Kipf, 2018; Popova et al., 2019; Shi et al., 2020). Recently, Markov chain Monte Carlo (MCMC) methods also present strong performance in optimizing multiple desired properties simultaneously (Xie et al., 2021). Targeting molecular properties usually leads these algorithms into narrow regions of similar compounds and hinders exploration into the broader chemical space.

2.2. Measuring Exploration

Beyond property optimization, some studies evaluate the molecular generation models with measures that are more or less related to the degree of exploration in the chemical space. Commonly used measures include internal diversity, external diversity, and the number or percentage of unique molecules (You et al., 2018; De Cao & Kipf, 2018; Elton et al., 2019; Brown et al., 2019; Popova et al., 2019; Polykovskiy et al., 2020; Shi et al., 2020; Xie et al., 2021). Zhang et al. (2021) propose to use the number of unique functional groups or ring systems to estimate the chemical space coverage and to compare several recent generative models. Similarly in Blaschke et al. (2020), the number of unique Bemis-Murcko scaffolds is used to measure the variety of drug candidates. Koutsoukas et al. (2014) study the effect of molecular fingerprinting schemes on the internal diversity of compound selection. These measures usually mix the concepts of diversity, coverage, or novelty, and their validity as a measure of exploration is not justified.

To the best of our knowledge, this is the first work that formally investigates the validity of the molecular exploration measures. In particular, axiomatic approaches are used to analytically evaluate various designs of a measurement, such as utility functions (Herstein & Milnor, 1953), cohesiveness (Alcalde-Unzu & Vorsatz, 2013), or document relevance (Fang et al., 2004). Using axiomatic analysis to evaluate the exploration measures in the chemical space is novel. With empirical studies in addition to the axiomatic analysis, we make practical recommendations on effective exploration measures, including two novel measures.

2.3. Encouraging Exploration

A handful of work tries to encourage the model to explore a broader chemical space. For example, Nigam et al. (2020) adds an adversarial loss into the fitness function of GA to avoid the traps of local optima; Blaschke et al. (2020) uses a memory unit to score the novelty of generated molecules.

One of our analyses explicitly formalizes the molecular generation problem as a joint optimization of both molecular properties and a given exploration measure so we can study the behavior of multiple measures in this optimization setting. This formulation also suggests a strong potential of improving drug discovery by encouraging exploration.

3. An Axiomatic Analysis of Molecular Exploration Measures

We first present an axiomatic approach to analytically evaluating various exploration measures of the chemical space.

3.1. Definition of Exploration Measures

To define an exploration measure, we first formalize the notion of chemical space with the following assumption.

Assumption 3.1 (Chemical space)**.** The chemical space $\mathcal{U}$ contains all possible molecules and is a metric space with a distance metric function $d : \mathcal{U} \times \mathcal{U} \to [0, +\infty)$.

This assumption is widely adopted in cheminfomatics, using distance metrics such as the Tanimoto distance (Tanimoto, 1968; Bajusz et al., 2015).

Definition 3.2 (Tanimoto distance)**.** For two molecules $x_1, x_2 \in \mathcal{U}$, whose binary molecular fingerprint vectors are $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \{0, 1\}^n$ where n is the dimensionality of the fingerprint, their Tanimoto distance is defined as

$$d(x_1, x_2) := \frac{\sum_{j=1}^{n} \boldsymbol{x}_{1j} \cdot \boldsymbol{x}_{2j}}{\sum_{j=1}^{n} \max(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j})}.$$

The Tanimoto distance is also referred to as the Jaccard distance (Jaccard, 1912) in other domains. Its range is $[0, 1]$. For finite sets (*e.g.*, molecular fingerprints), the Jaccard distance is a metric function (Kosub, 2019; Lipkus, 1999).

Besides the Tanimoto distance, one can also use the distance of latent hidden vectors (Preuer et al., 2018; Samanta et al., 2020) or the root-mean-square deviation (RMSD) of three-dimensional molecular conformers (Fukutani et al., 2021).

We then define an exploration measure as the following.

Definition 3.3 (Exploration measure)**.** Given the universal chemical space $\mathcal{U}$, an exploration measure is a function that maps a set of molecules to a non-negative real number that reflects to what extent the set spans the chemical space, *i.e.*, $\mu : \mathcal{P}(\mathcal{U}) \to [0, \infty)$, where $\mathcal{P}(\cdot)$ is the notation of power set. In particular, $\mu(\emptyset) = 0$.

3.2. Examples of Exploration Measures

Definition 3.3 is intentionally left general. Many measures related to the coverage, diversity, or novelty of a molecular set fall into this definition. We summarize these various measures used in literature into three categories: *distance-based* measures, *coverage-based* measures, and *locality-based* measures. We are also able to define new exploration measures under this formulation.

3.2.1. Distance-based Measures

For a set $\mathcal{S}$ with n molecules, we can define an exploration measure according to the distances among the molecules:

$$\text{Diversity}(\mathcal{S}) := \frac{2}{n(n-1)} \sum_{\substack{x,y \in \mathcal{S} \\ x \neq y}} d(x, y), \tag{1}$$

$$\begin{aligned} \text{SumDiversity}(\mathcal{S}) &:= \sum_{x \in \mathcal{S}} \frac{1}{n-1} \sum_{\substack{y \in \mathcal{S} \\ y \neq x}} d(x, y) \\ &= n \cdot \text{Diversity}(\mathcal{S}), \end{aligned} \tag{2}$$

$$\text{Diameter}(\mathcal{S}) := \max_{\substack{x,y \in \mathcal{S} \\ x \neq y}} d(x, y), \tag{3}$$

$$\text{SumDiameter}(\mathcal{S}) := \sum_{x \in \mathcal{S}} \max_{\substack{y \in \mathcal{S} \\ y \neq x}} d(x, y), \tag{4}$$

$$\text{Bottleneck}(\mathcal{S}) := \min_{\substack{x,y \in \mathcal{S} \\ x \neq y}} d(x, y), \tag{5}$$

$$\text{SumBottleneck}(\mathcal{S}) := \sum_{x \in \mathcal{S}} \min_{\substack{y \in \mathcal{S} \\ y \neq x}} d(x, y), \tag{6}$$

$$\text{DPP}(\mathcal{S}) := \det(\boldsymbol{S}), \tag{7}$$

where x, y are molecules in $\mathcal{S}$, d is a distance metric as defined previously, and $\boldsymbol{S}$ is the similarity matrix of candidate molecules (*e.g.*, $1 - d(x, y)$ for Tanimoto similarity).

Among these measures, Diversity (also referred to as the *internal diversity*), the average distance between the molecules, is widely used in the literature (You et al., 2018; De Cao & Kipf, 2018; Popova et al., 2019; Polykovskiy et al., 2020; Shi et al., 2020; Xie et al., 2021). We follow the topology theory and also introduce Diameter and Bottleneck as the maximum and minimum distance between any pair of molecules, respectively (Edelsbrunner & Harer, 2010), since they can also reflect the dissimilarity between molecules in $\mathcal{S}$. We further introduce three Sum- variants for the above measures. The Sum- variants will tend to increase when new molecules are added into the set. In addition, the determinant of the similarity matrix of molecules is also a measure of dissimilarity, which is often employed in diverse subset selection as a key concept of the determinantal point processes (DPP) (Kulesza & Taskar, 2011; 2012).

3.2.2. Coverage-based Measures

Another broad category of exploration measures compare the generated molecules $\mathcal{S}$ with a reference set $\mathcal{R}$. In such context, a coverage-based measure can be defined as below:

$$\text{Coverage}(\mathcal{S}, \mathcal{R}) := \sum_{y \in \mathcal{R}} \left(\max_{x \in \mathcal{S}} \text{cover}(x, y) \right), \tag{8}$$

where $\text{cover}(x, y)$ indicates how the molecule x can cover the reference y.

When the reference set $\mathcal{R}$ is taken as a collection of molecular fragments, the coverage function can be written as $\text{cover}(x, y) := \mathbb{I}[\text{molecule } x \text{ contains fragment } y]$, where $\mathbb{I}[\cdot]$ is the indication function. A large body of drug discovery literature uses the number of distinct functional groups

(FG), ring systems (RS), or Bemis-Murcko scaffolds (BM) in $\mathcal{S}$ to gauge the size of explored chemical space (Zhang et al., 2021; Blaschke et al., 2020), corresponding to the cases where $\mathcal{R}$ is the collection of all possible FG, RS, or BM fragments. We denote these specific coverage-based measures as #FG, #RS, and #BM respectively.

3.2.3. Locality-based Measures

Inspired by the sphere exclusion algorithm used in compound selection (Snarey et al., 1997; Gobbi & Lee, 2003), we introduce a new exploration measure that highlights the local neighborhoods covered by a set of molecules:

$$\#\text{Circles}(\mathcal{S}) := \max_{\mathcal{C}\subseteq\mathcal{S}} \ |\mathcal{C}| \quad \text{s.t. } d(x,y) > t, \quad \forall x, y \in \mathcal{C}, \tag{9}$$

where $t \in [0, 1)$ is a distance threshold that corresponds to the diameter of a circle.

Intuitively, #Circles counts the maximum number of mutually exclusive circles that can fit into $\mathcal{S}$ as neighborhoods, with a subset of its members $\mathcal{C}$ as the circle centers. When the threshold $t = 0$, #Circles becomes the richness measure $\text{Richness}(\mathcal{S}) := |\{\mathcal{S}\}|$, which is the number of unique molecules (Moore, 2013; Shi & von Itzstein, 2019).

3.3. Axiomatic Analysis

While all the aforementioned measures can heuristically reflect the degree of exploration, they do not always agree with each other. We need a principled way to select the most suitable measures from a variety of possible choices. Ideally, this criterion should not depend on the particular molecule properties of target or the algorithm used to generate the candidates. We achieve this through an axiomatic approach.

We propose two simple and intuitive principles that a good exploration measure should satisfy: (1) discovering more molecules should not decrease the degree of exploration; (2) molecular sets with more dissimilar molecules have a higher degree of exploration. These two principles are formalized below as two axioms and can be tested analytically.

Axiom 3.4 (Subadditivity)**.** *A good exploration measure μ should be subadditive, i.e., for any two molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$, it holds that*

$$\max(\mu(\mathcal{S}_1), \mu(\mathcal{S}_2)) \leq \mu(\mathcal{S}_1 \cup \mathcal{S}_2) \leq \mu(\mathcal{S}_1) + \mu(\mathcal{S}_2).$$

A subadditive measure tends to increase when more molecules are included. This tendency is very intuitive in drug discovery as testing more molecules means we can have a higher probability to discovery a potentially effective drug. Some direct corollaries are listed in Appendix A.

A subadditive exploration measure is a special case of *outer measures* in the context of mathematical measure theory.

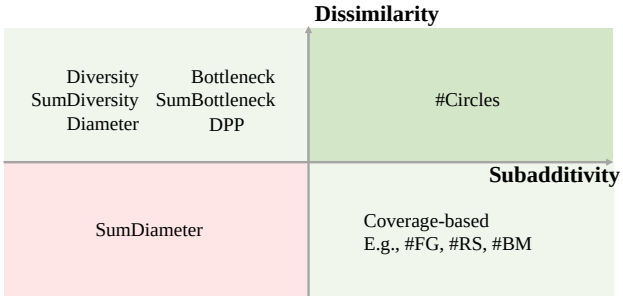


Figure 1. Exploration measures, categorized by whether they satisfy subadditivity and/or dissimilarity. #Circles is the only measure that satisfies both subadditivity and dissimilarity simultaneously.

Outer measures are relaxations of *measures*, where the latter requires a stricter additivity property, *i.e.*, $\mu(\mathcal{S}_1 \cup \mathcal{S}_2) = \mu(\mathcal{S}_1) + \mu(\mathcal{S}_2)$ for any two disjoint sets $\mathcal{S}_1, \mathcal{S}_2$ (Halmos, 2013). We consider additivity to be too strong and can conflict with intuitions in drug discovery: an additive measure defined on a discrete space must take the form of $\mu(\mathcal{S}) = \sum_{x\in\mathcal{S}} w(x)$, in which $w(\cdot)$ is a weight function that independently assigns a score to each element in the space, meaning additive measures defined on the discrete chemical space cannot capture the interrelationship between molecules in a candidate set, which counters the reality.

Axiom 3.5 (Dissimilarity)**.** *A good exploration measure should have a preference to dissimilar elements, i.e., for any two molecules $x_1, x_2 \in \mathcal{U}$ and a molecular set $\mathcal{X}(x_1, x_2) = \{x \mid d(x, x_1) + d(x, x_2) = d(x_1, x_2),\ x \neq x_1, x_2\}$, it holds*

$$\mu(\{x_1, x_2, x^*\}) = \max_{x\in\mathcal{X}} \ \mu(\{x_1, x_2, x\}),$$

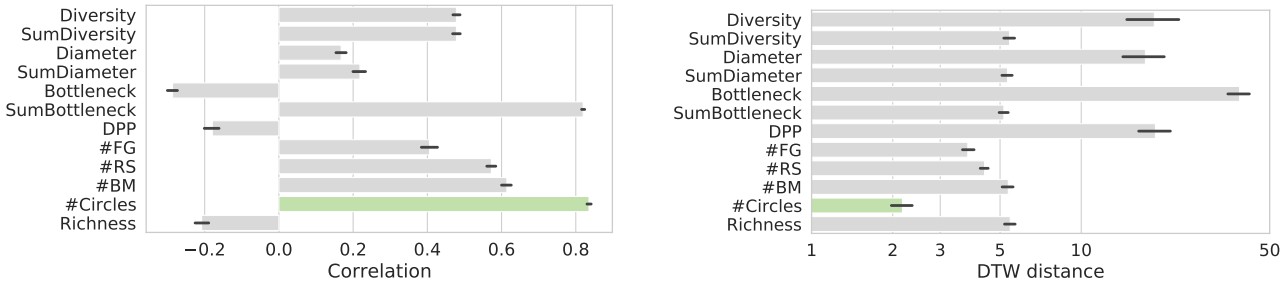
where x^* satisfies $d(x^*, x_1) = d(x^*, x_2) = \frac{1}{2}d(x_1, x_2)$ and d is the distance metric on the chemical space $\mathcal{U}$.

Intuitively, consider two molecules x_1, x_2 that are placed at the two endpoints of a "segment". An exploration measure satisfying the dissimilarity axiom should increase the most when the molecule x^* is added to the middle point of the segment, since x^* is most dissimilar to x_1, x_2 comparing to any other points on the segment.

Depending on whether each exploration measure satisfies *subadditivity* and/or *dissimilarity*, we can put it into one of four quadrants as shown in Figure 1. The formal proofs are provided in Appendix B. Despite that the two axioms are simple and intuitive, surprisingly, the proposed #Circles is the only measure that satisfies both axioms.

4. Correlation with Biological Functionality

While it is encouraging that only one measure stands out in the axiomatic analysis, it does not guarantee its empirical performance. Meanwhile, there are still multiple measures



(a) Correlations between the gold standard GS and exploration measures in the fixed-size setting. The subset size is set as $n = 200$. A higher correlation is better.

(b) DTW distances between the gold standard GS and exploration measures in the growing-size setting. The maximum size is set as $n = 1000$. A smaller distance is better.

Figure 2. Correlation between each exploration measure and biological functionality coverage. The exploration measure with the highest correlation is highlighted in green. Results are aggregated by running experiments independently for ten times.

that satisfy *one* of the two axioms. We further investigate the validity of the exploration measures by testing their correlation with the biological functionality variety of molecules. Such functionalities are expensive to annotate and are biased towards what are already known and tested, but they still provide valuable information in distinguishing molecules and guiding the exploration of the unknown space.

In this section, we correlate the aforementioned exploration measures to the number of unique biological functionalities covered by a molecular set. A better exploration measure should have a higher correlation to the coverage of biological functionalities, even though it is a proxy gold standard.

4.1. Experiment Setup

We base the analysis on the BioActivity dataset that is also used to compare different compound selection algorithms (Koutsoukas et al., 2014). This dataset contains 10,000 compound samples extracted from the ChEMBL database (Gaulton et al., 2017) with bio-activity labels, which annotate 50 activity classes with 200 samples each. Following Koutsoukas et al. (2014), for a subset of this dataset $\mathcal{S}$, we take the number of unique class labels as a proxy "gold standard" of the variety of the molecures in $\mathcal{S}$, *i.e.*,

$$\mathrm{GS}(\mathcal{S}) := \#\text{unique labels in } \mathcal{S},$$

which represents the number of biological functionality types covered by $\mathcal{S}$. We then compare the behavior of the gold standard and the exploration measures in two settings to find out which measures have the highest correlations with the coverage (or variety) of biological functionalities.

4.2. Random Subsets with Fixed Sizes

We first consider randomly sampled molecular subsets of the BioActivity dataset with a fixed size n. We randomly sample n molecules $\mathcal{S}$ from the dataset and compute the biological functionality coverage $\mathrm{GS}(\mathcal{S})$ as well as each exploration measure $\mu(\mathcal{S})$. By repeating the randomization, we can calculate the Spearman's correlation between the gold standard GS and each individual measure μ.

From the correlations shown in Figure 2a, #Circles and SumBottleneck have notably higher correlations to the gold standard than other measures. This indicates the locality information is critical in evaluating exploration, as these two measures both prefer new molecules that are at arm's length from their nearest neighbors.

4.3. Random Subsets with Growing Sizes

To mimic the molecular generation process, we also grow the size of the subsets. Specifically, for a maximum size n, we sequentially sample n molecules without replacement to form n subsets $\{\mathcal{S}_i = \{x_1, \ldots, x_i\}\}_{i=1}^n$. For both the gold standard GS and a exploration measure μ, we record their values as $\mathcal{S}$ grows into a time series, *e.g.*, $\{(i, \mu(\mathcal{S}_i))\}_{i=1}^n$. Comparing the trajectory of an exploration measure with the trajectory of the gold standard, we can observe which measure behaves more similarly to GS. We quantitatively estimate the similarity of their trajectories with the dynamic time warping (DTW) distance of the two time series.

As shown in Figure 2b, #Circles surpasses all other exploration measures. Coverage-based measures such as #FG, #RS, and #BM also perform prominently. Both #Circles and coverage-based measures satisfy the subadditivity property, making them suitable for a "growth" setting. The Sum- variants of distance-based measures outperform their original forms, as they tend to increase when adding new molecules.

4.4. Discussion

The discrepancy of the exploration measures' behavior under the *fixed-size* and the *growth* settings suggests that we might need different measures in different settings. For instance, SumBottleneck performs closely to #Circles in

the fixed-size setting but falls behind remarkably in the growth setting. This might be attributed to the salience of the subadditivity requirement in the growth setting, which SumBottleneck fails to meet. Notably, #Circles stands out in both settings, suggesting its great potential in practice.

The experiment details are listed in Appendix C, where we also discuss the impact of the subset size n, the way of adding new molecules, the choice of the #Circles's threshold t, and the distance metric d.

The empirical analysis shows that the locality-based #Circles measure is a robust choice for all tested scenarios, which reconfirms the conclusion of the axiomatic analysis. SumBottleneck may be an effective choice when a fixed number of candidates are the target. Coverage-based measures may be good alternatives if carefully-designed and comprehensive reference sets are available. Surprisingly, the widely used Diversity measure is rendered inferior both analytically and empirically, casting doubts on its efficiency of measuring and encouraging exploration in molecular generation.

5. Optimizability of Exploration Measures

While satisfying both axioms and having a high correlation with biological functionality coverage assure the quality of an exploration measure in theory, its effectiveness still relies on how it is deployed in practice. In particular, a good exploration measure should be optimizable and actually encourage exploration in the real context of molecular generation. In this section, we further investigate how the selected exploration measures can be optimized in the molecular generation problem, and how different measures behave in the optimization setting and encourage the generative models to explore a wider range of the chemical space.

5.1. Problem Formulation

In drug discovery, the search for active molecules towards a druggable target is often formulated as the following optimization problem (Olivecrona et al., 2017; Gómez-Bombarelli et al., 2018; Liu et al., 2018; You et al., 2018; Jin et al., 2018; De Cao & Kipf, 2018; Popova et al., 2019; Shi et al., 2020; Xie et al., 2021):

$$\operatorname*{argmax}_{x \in \mathcal{U}} \quad \text{Property}(x) \tag{10}$$

where x is a molecule in the chemical space $\mathcal{U}$ and Property : $\mathcal{U} \to \mathbb{R}$ is a function that scores particular biological properties of the molecule. This property term can incorporate bio-activity such as binding affinity to protein targets, drug-likeness, synthesizability, *etc*. (Nicolaou et al., 2012; Jin et al., 2020; Xie et al., 2021). When web-lab experiments are not available, this score is obtained computationally.

However, the computational scoring function is merely an approximation and cannot accurately predict the wet-lab results. As pointed out in the chemistry literature, rather than focusing on optimizing the estimated property scores, it is crucial to generate a variety of compounds that span a wider range of the chemical space (Huggins et al., 2011; Wawer et al., 2014; Ashenden, 2018). We therefore propose the following objective as the goal of molecular generation:

$$\operatorname*{argmax}_{\mathcal{S} \subseteq \mathcal{U}} \quad \mu(\mathcal{S}) \quad \text{s.t.} \quad \text{Property}(x) \geq C,\ x \in \mathcal{S}, \tag{11}$$

where $\mathcal{S} \subseteq \mathcal{U}$ is a molecular candidate set, $\mu(\cdot)$ is an exploration measure, $\text{Property}(x) \geq C$ means the estimated property score of a molecule x is at or above a threshold C.

This constrained optimization problem is very challenging due to its combinatorial characteristics. To this end, we propose a relaxed alternative to Equation 11. We use a greedy strategy to convert the molecular set generation problem into a single molecule generation problem:

$$\begin{aligned}
&\operatorname*{argmax}_{x \in \mathcal{U}} \quad \mu(\mathcal{S} \cup \{x\}) \quad \text{s.t.} \quad \text{Property}(x) \geq C \\
\approx &\operatorname*{argmax}_{x \in \mathcal{U}} \quad \text{Property}(x) + \alpha \cdot \mu(\mathcal{S} \cup \{x\}) \\
= &\operatorname*{argmax}_{x \in \mathcal{U}} \quad \text{Property}(x) + \alpha \cdot [\mu(\mathcal{S} \cup \{x\}) - \mu(\mathcal{S})] \\
= &\operatorname*{argmax}_{x \in \mathcal{U}} \quad \text{Property}(x) + \alpha \cdot \text{Novelty}(x, \mathcal{S}),
\end{aligned} \tag{12}$$

where x is the molecule to be generated and $\mathcal{S}$ is the generated drug candidates. Then to efficiently find the plausible molecule, we can jointly optimize the molecular property and the exploration measure. In the equations, α is a coefficient related to the property threshold C that controls the balance between the property and the exploration. By introducing a constant term $\mu(\mathcal{S})$ and defining the novelty of a molecule x as how much it expands the exploration, *i.e.*, $\text{Novelty}(x, \mathcal{S}) := \mu(\mathcal{S} \cup \{x\}) - \mu(\mathcal{S})$, we obtain a new molecular generation objective as Equation 12.

Like Equation 10, Equation 12 can be optmized with methods like deep generative models, reinforcement learning, Markov chain Monte Carlo (MCMC), or genetic algorithms (GA). However, one should notice that the new objective is always shifting as the set of generated molecules $\mathcal{S}$ grows, making it a much more interesting optimization problem. Besides, there is still a gap between the original objective (Equation 11) and the relaxed objective (Equation 12).

5.2. Experiments

In the experiments, we study which exploration measures are more compatible with the optimization objective Equation 12 and can encourage more chemical space exploration. The implementation details are listed in Appendix D.

Table 1. Exploration measures are optimizable through novelty terms. Adding novelty terms to MARS increases the exploration measures (the larger the better) of generated molecules. Underline indicates the best performance in each measure. Results are aggregated from 5 independent runs of the experiments. Grey highlights #Circles, which is the best exploration measure in theory (among considered).

Model	#Circles(↑)	SumBottleneck(↑)	#FG(↑)	Richness(↑)	Diversity(↑)
MARS (Xie et al., 2021)	380 ± 10	2149 ± 175	346 ± 55	134K ± 14K	0.736 ± 0.007
+ Diversity	514 ± 33	3001 ± 409	441 ± 85	186K ± 29K	0.750 ± 0.006
+ SumBottleneck	709 ± 66	3777 ± 494	868 ± 133	221K ± 30K	0.752 ± 0.003
+ #Circles	706 ± 155	3013 ± 364	601 ± 211	179K ± 31K	0.749 ± 0.006

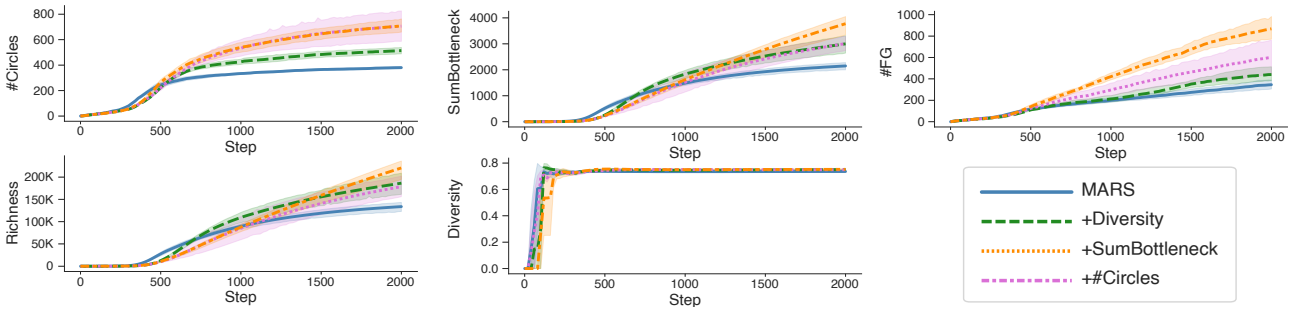


Figure 3. The dynamics of exploration measures over generated molecules. Adding novelty terms increases the exploration in the chemical space. MARS+SumBottleneck achieves the highest degree of exploration over time. All models converge in Diversity quickly.

5.2.1. Experiment Setup

Properties. Following a previous study (Li et al., 2018), we consider the inhibition against an Alzheimer-related target protein c-Jun N-terminal kinase-3 (JNK3) as the biological objective. JNK3 is the neuron-specific isoform of JNK that expresses primarily in brain and heart. Designing the inhibitor of JNK3 signaling has raised interests in science community as it could become a potential drug for neural-degenerative disease such as Alzheimer (Antoniou et al., 2011; Nakano et al., 2020). The JNK3 binding affinity score is predicted by a random forest model[1] based on Morgan fingerprint features of a molecule (Rogers & Hahn, 2010). A higher score indicates that the small molecule design is likely to bind to the JNK3 target. Besides, to obtain drug-like and synthesizable molecules, we also consider the quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012) and the synthetic accessibility (SA) (Ertl & Schuffenhauer, 2009) as suggested in Jin et al. (2020) and Xie et al. (2021). In summary, the overall property scoring function can be written as $\text{Property}(x) = \text{JNK3}(x) + \text{QED}(x) + \text{SA}(x)$, where $\text{JNK3}(\cdot)$, $\text{QED}(\cdot)$, and $\text{SA}(\cdot)$ are all re-scaled to $[0, 1]$ (the larger the better).

Models. We use the recently proposed molecular generation model MARS (Xie et al., 2021) as our baseline. MARS is based on MCMC sampling and aims at optimizing multiple drug discovery objectives simultaneously. We choose MARS mainly because its state-of-the-art performance on multi-objective molecular generation. The model can handily deal with the additional novelty objective in Equation 12.

Novelty terms. According to the axiomatic and empirical analyses in previous sections, we choose the better-performed exploration measures #Circles and SumBottleneck to derive the novelty terms. The Coverage-based measures are not included due to the consideration of computational efficiency. In addition, we include the Diversity measure since it is widely used to evaluate molecular generation models in literature. Specifically, we consider the following three novelty terms[2]:

$$\text{Novelty}_{\text{Diversity}}(x, \mathcal{S}) := \frac{1}{|\mathcal{S}|} \sum_{y \in \mathcal{S}} d(x, y) \tag{13}$$

$$\text{Novelty}_{\text{SumBottleneck}}(x, \mathcal{S}) := \min_{y \in \mathcal{S}} d(x, y) \tag{14}$$

$$\text{Novelty}_{\text{\#Circles}}(x, \mathcal{S}) := \left[\min_{y \in \mathcal{S}} d(x, y) > t \right] \tag{15}$$

We use MARS to optimize both Equation 10 and Equation 12 with different novelty terms as in Equation 13-15.

Evaluation. We compare the molecules generated with or without the novelty term, using multiple exploration measures to validate whether adding the novelty terms indeed increases exploration. We examine the difference in three high-quality exploration measures, #Circles, SumBottleneck, and #FG, since they all have a relatively high correlation with the biological functionality coverage. Note we

[1]The JNK3 prediction model is provided at https://github.com/wengong-jin/multiobj-rationale/blob/master/properties.py.

[2]Equations 13-15 are approximations of $[\mu(\mathcal{S} \cup \{x\}) - \mu(\mathcal{S})]$ defined to avoid numerical issues and for computational efficiency.

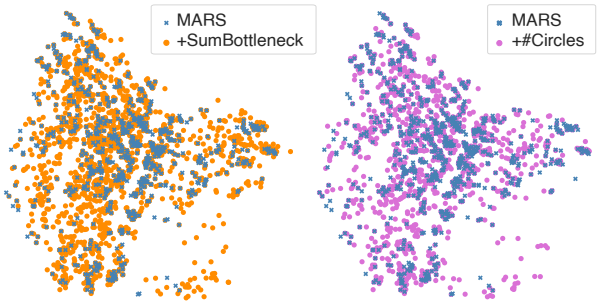


Figure 4. Optimizing exploration measures encourages the molecular generation model to explore a larger span of the chemical space. This figure shows principal component analysis (PCA) of functional groups discovered by different models.

cannot directly measure the biological functionality coverage as in Section 4, as many of the generated molecules are not in the BioActivity database. We also include Richness and Diversity as they are widely used in the literature, even though they perform weakly in the previous analyses (denoted as weak measures). For each model variant, 10M molecules are generated. Duplicated molecules are removed, and the candidates that satisfy JNK3 ≥ 0.5, QED ≥ 0.6, and SA ≥ 0.67 are selected for computing the exploration measures as suggested in Jin et al. (2018); Xie et al. (2021).

5.2.2. Results and Discussion

Table 1 lists the values of exploration measures with or without the novelty terms, estimated for the molecules generated in 2000 sampling steps.

By introducing novelty terms, all exploration measures are significantly improved during molecular generation, showing a strong optimizability of these measures. Comparing model variants implemented with different novelty terms, both MARS+SumBottleneck and MARS+#Circles obtain a higher degree of exploration (in all three high-quality measures) than MARS+Diversity does. In particular, MARS+SumBottleneck leads in all measures and even outperforms the MARS+#Circles variant. We note a larger variance of MARS+#Circle, which indicates although #Circle is a better measure in theory, it may be less efficient to optimize than SumBottleneck. In summary, while all novelty terms can encourage the model to explore a larger chemical space to a certain extent, SumBottleneck and #Circles are much more effective than Diversity.

To better understand the behavior of the four optimization models, we plot their dynamics in each exploration measure over the generated molecules in Figure 3. The MARS+SumBottleneck model converges to a higher degree of exploration in multiple measures, suggesting that it gradually spans a larger portion of the chemical space. MARS+#Circles behaves similarly but with a larger variance. Interestingly, all four models converge in Diversity very quickly, disregarding the large number of new molecules discovered, again suggesting the widely used Diversity might not be suitable for measuring exploration.

To provide a more intuitive view of the effect that adding a joint objective of exploration can encourage the molecular generation model to discover a wider space, we visualize the chemical space explored by the baseline (MARS) and MARS+SumBottleneck as well as MARS+#Circles in Figure 4, where we show the 2D layout of the Morgan fingerprints (Rogers & Hahn, 2010) of the functional groups in the molecules generated through principal component analysis (PCA). The larger number of unique functional groups and the wider spread of the data points obtained by adding exploration-based novelty terms indicate more diverse structures being explored.

In summary, SumBottlenck and #Circles are both well optimizable when they are incorporated into the objective as novelty terms, and they truly encourage molecular generation models to explore broadly in the chemical space.

A few limitations of optimization with novelty terms should be noted. First, the computation of novelty terms on large molecular generation campaigns is significantly more costly. This limitation is to be addressed by more efficient algorithms to compute the measures. Second, the performance of the MARS+#Circles model can be influenced by the selection of the distant threshold t (circle diameter); more sensitivity tests and hyper-parameter tuning strategies are needed for determining the optimal threshold.

6. Conclusion

We present a systematic study on the measurement of chemical space exploration in drug discovery. We formally define the concept of exploration measures and propose a novel framework to validate potential measures from three aspects: analytically validate whether the measures satisfy two axioms, empirically compare the correlations between the exploration measures with the biological functionality coverage of known molecules, and practically benchmark the optimizability of the measures in the real molecular generation task. Overall, we find that the novel #Circles measure is an outstanding choice both theoretically and empirically. A new distance-based measure, SumBottleneck, also demonstrates excellent empirical and practical performance. Diversity, although widely used in literature, is a suboptimal measure of exploration.

For future work, it would be interesting to design and evaluate more exploration measures under this framework and apply them to practical drug discovery scenarios. The optimization of Equation 12 inspires further investigation, since the objective will be consistently shifting. The proposed exploration measures can be applied to other domains.

References

Alcalde-Unzu, J. and Vorsatz, M. Measuring the cohesiveness of preferences: an axiomatic analysis. *Social Choice and Welfare*, 41(4):965–988, 2013.

Antoniou, X., Falconi, M., Di Marino, D., and Borsello, T. Jnk3 as a therapeutic target for neurodegenerative diseases. *Journal of Alzheimer's Disease*, 24(4):633–642, 2011.

Ashenden, S. K. Screening library design. *Methods in enzymology*, 610:73–96, 2018.

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

Bian, Y. and Xie, X.-Q. Generative chemistry: drug discovery with deep learning generative models. *Journal of Molecular Modeling*, 27(3):1–18, 2021.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

Blaschke, T., Engkvist, O., Bajorath, J., and Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. *Journal of cheminformatics*, 12(1):1–17, 2020.

Brown, N., McKay, B., Gilardoni, F., and Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of chemical information and computer sciences*, 44(3): 1079–1087, 2004.

Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

De Cao, N. and Kipf, T. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

Deng, J., Yang, Z., Ojima, I., Samaras, D., and Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics*, 23(1):bbab430, 2022.

Devi, R. V., Sathya, S. S., and Coumar, M. S. Evolutionary algorithms for de novo drug design - a survey. *Applied Soft Computing*, 27:543–552, 2015.

Edelsbrunner, H. and Harer, J. *Computational topology: an introduction*. American Mathematical Soc., 2010.

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4 (4):828–849, 2019.

Ertl, P. and Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8–8, 2009.

Fang, H., Tao, T., and Zhai, C. A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49–56, 2004.

Fukutani, T., Miyazawa, K., Iwata, S., and Satoh, H. G-rmsd: Root mean square deviation based method for three-dimensional molecular similarity determination. *Bulletin of the Chemical Society of Japan*, 94(2):655–665, 2021.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. The chembl database in 2017. *Nucleic Acids Research*, 45, 2017.

Gobbi, A. and Lee, M.-L. Dise: directed sphere exclusion. *Journal of chemical information and computer sciences*, 43(1):317–323, 2003.

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., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Halmos, P. R. *Measure theory*, volume 18. Springer, 2013.

Herstein, I. N. and Milnor, J. An axiomatic approach to measurable utility. *Econometrica, Journal of the Econometric Society*, pp. 291–297, 1953.

Huggins, D. J., Venkitaraman, A. R., and Spring, D. R. Rational methods for the selection of diverse screening compounds. *ACS Chemical Biology*, 6(3):208–217, 2011.

Jaccard, P. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

Jin, W., Barzilay, R., and Jaakkola, T. S. Junction tree variational autoencoder for molecular graph generation.

In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Jin, W., Barzilay, R., and Jaakkola, T. Multi-objective molecule generation using interpretable substructures. In *International Conference on Machine Learning*, pp. 4849–4859. PMLR, 2020.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa971. URL https://doi.org/10.1093/nar/gkaa971.

Kirkpatrick, P. and Ellis, C. Chemical space. *Nature*, 432 (7019):823–824, 2004.

Kosub, S. A note on the triangle inequality for the jaccard distance. *Pattern Recognition Letters*, 120:36–38, 2019.

Koutsoukas, A., Paricharak, S., Galloway, W. R., Spring, D. R., IJzerman, A. P., Glen, R. C., Marcus, D., and Bender, A. How diverse are diversity assessment methods? a comparative analysis and benchmarking of molecular descriptor space. *Journal of chemical information and modeling*, 54(1):230–242, 2014.

Kulesza, A. and Taskar, B. Learning determinantal point processes. 2011.

Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

Li, Y., Zhang, L. R., and Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, 2018.

Lipkus, A. H. A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, 26(1):263–265, 1999.

Liu, Q., Allamanis, M., Brockschmidt, M., and Gaunt, A. L. Constrained graph variational autoencoders for molecule design. *arXiv preprint arXiv:1805.09076*, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Moore, J. C. Diversity, taxonomic versus functional. 2013.

Nakano, R., Nakayama, T., and Sugiya, H. Biological properties of jnk3 and its function in neurons, astrocytes, pancreatic β-cells and cardiovascular cells. *Cells*, 9(8): 1802, 2020.

Nicolaou, C. A., Kannas, C., and Loizidou, E. Multi-objective optimization methods in de novo drug design. *Mini-reviews in Medicinal Chemistry*, 12(10):979–987, 2012.

Nigam, A., Friederich, P., Krenn, M., and Aspuru-Guzik, A. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48–48, 2017.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:1931, 2020.

Popova, M., Shvets, M., Oliva, J., and Isayev, O. Molecularrnn: Generating realistic molecular graphs with optimized properties. *arXiv: Learning*, 2019.

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Samanta, S., O'Hagan, S., Swainston, N., Roberts, T. J., and Kell, D. B. Vae-sim: a novel molecular similarity measure based on a variational autoencoder. *Molecules*, 25(15):3446, 2020.

Schwalbe-Koda, D. and Gómez-Bombarelli, R. Generative models for automatic chemical design. In *Machine Learning Meets Quantum Physics*, pp. 445–467. Springer, 2020.

Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., and Tang, J. Graphaf: a flow-based autoregressive model for molecular graph generation. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Shi, Y. and von Itzstein, M. How size matters: diversity for fragment library design. *Molecules*, 24(15):2838, 2019.

Snarey, M., Terrett, N. K., Willett, P., and Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling*, 15(6):372–385, 1997.

Tanimoto, T. An elementary mathematical theory of classification and prediction, ibm report (november, 1958), cited in: G. salton, automatic information organization and retrieval, 1968.

Wawer, M. J., Li, K., Gustafsdottir, S. M., Ljosa, V., Bodycombe, N. E., Marton, M. A., Sokolnicki, K. L., Bray, M.-A., Kemp, M. M., Winchester, E., et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proceedings of the National Academy of Sciences*, 111(30):10911–10916, 2014.

Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., and Li, L. Mars: Markov molecular sampling for multi-objective drug discovery. *arXiv preprint arXiv:2103.10432*, 2021.

You, J., Liu, B., Ying, Z., Pande, V. S., and Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, 2018.

Yung-Chi, C. and Prusoff, W. H. Relationship between the inhibition constant (ki) and the concentration of inhibitor which causes 50 per cent inhibition (i50) of an enzymatic reaction. *Biochemical pharmacology*, 22(23):3099–3108, 1973.

Zhang, J., Mercado, R., Engkvist, O., and Chen, H. Comparative study of deep generative models on chemical space coverage. *Journal of Chemical Information and Modeling*, 61(6):2572–2581, 2021.

A. Subadditivity Corollaries

Corollary A.1. ***(Subtraction)****. If an exploration measure μ is subadditive, then for any two molecular sets $\forall \mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$, we have*

$$\mu(\mathcal{S}_1) \geq \mu(\mathcal{S}_1 \setminus \mathcal{S}_2) \geq \mu(\mathcal{S}_1) - \mu(\mathcal{S}_2).$$

Corollary A.2. ***(Monotonicity)****. If an exploration measure μ is subadditive, then for any molecular set $\forall \mathcal{S} \subseteq \mathcal{U}$ and a single molecule $x \in \mathcal{U}$, we have $\mu(\mathcal{S} \cup \{x\}) \geq \mu(\mathcal{S})$ and $\mu(\mathcal{S} \setminus \{x\}) \leq \mu(\mathcal{S})$.*

Corollary A.3. ***(Dominance)****. If an exploration measure μ is subadditive, then for any two molecular sets $\forall \mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$, if $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\mu(\mathcal{S}_1) \leq \mu(\mathcal{S}_2)$.*

B. Proofs for Subadditivity and Dissimilarity

In Figure 1 we show the exploration measures according to whether they will satisfy subadditivity and/or dissimilarity. In this section, we provide proofs to verify each measure's subadditivity and dissimilarity.

Note that for distance-based measures, when $|\mathcal{S}| = 1$, $\mu(\mathcal{S})$ is not defined through Equation 1-7. So without loss of generality, we assume $\mu(\{x\}) = w(x),\ x \in \mathcal{U}$ in such cases, where $w : \mathcal{U} \to \mathbb{R}$ is an importance function. However in the following proofs, we can see that the choice of $w(\cdot)$ will not influence our conclusion.

Proposition B.1 (Subadditivity of exploration measures.)**.** *Coverage-based exploration measures and #Circles are subadditive. Distance-based exploration measures including Diversity, SumDiversity, Diameter, SumDiameter, Bottleneck, SumBottleneck, and DPP are not subadditive.*

Proof. For the **coverage-based** exploration measures, we denote the cover$(\cdot,\cdot)$ function as cov$(\cdot,\cdot)$ for short. We prove the subadditivity of coverage-based measures by first proving the subadditivity of the maximum of cover$(\cdot,\cdot)$.

To prove the subadditivity of the maximum of cov$(\cdot,\cdot)$, we consider combining any two molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$. For any reference $y \in \mathcal{R}$, we have

$$\begin{aligned}
&\max\left(\max_{x\in\mathcal{S}_1} \text{cov}(x,y), \max_{x\in\mathcal{S}_2} \text{cov}(x,y)\right) \\
\leq &\max_{x\in\mathcal{S}_1\cup\mathcal{S}_2} \text{cov}(x,y) \\
\leq &\max_{x\in\mathcal{S}_1} \text{cov}(x,y) + \max_{x\in\mathcal{S}_2} \text{cov}(x,y).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\max\left(\text{Coverage}(\mathcal{S}_1), \text{Coverage}(\mathcal{S}_2)\right) \\
\leq &\text{Coverage}(\mathcal{S}_1 \cup \mathcal{S}_2) \\
\leq &\text{Coverage}(\mathcal{S}_1) + \text{Coverage}(\mathcal{S}_1),
\end{aligned}$$

thus proving the Coverage measure is subadditive.

For the **#Circles** exploration measure, we denote #Circles$(\cdot)$ as #C$(\cdot)$ for short and define $\mathcal{C}^*(\mathcal{S}) \subseteq \mathcal{S}$ as an arbitrary set that satisfies $|\mathcal{C}^*(\mathcal{S})| = \#\text{C}(\mathcal{S})$. We prove the subadditivity of #C in two parts.

In the first part, we prove the left-hand side of the subadditivity inequation, *i.e.*, $\max(\#\text{C}(\mathcal{S}_1), \#\text{C}(\mathcal{S}_2)) \leq \#\text{C}(\mathcal{S}_1 \cup \mathcal{S}_2)$. For any two molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$, since $\mathcal{C}^*(\mathcal{S}_1) \subseteq \mathcal{S}_1 \cup \mathcal{S}_2$, according to the definition of #C, we have

$$\#\text{C}(\mathcal{S}_1) \leq \#\text{C}(\mathcal{S}_1 \cup \mathcal{S}_2).$$

Similarly, we also have

$$\#\text{C}(\mathcal{S}_2) \leq \#\text{C}(\mathcal{S}_1 \cup \mathcal{S}_2),$$

proving the left-hand side of subadditivity.

In the second part, we prove the right-hand side of the subadditivity inequation, *i.e.*, $\#\text{C}(\mathcal{S}_1 \cup \mathcal{S}_2) \leq \#\text{C}(\mathcal{S}_1) + \#\text{C}(\mathcal{S}_1)$. We prove this by contradiction. For any two molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$, we assume $\#\text{C}(\mathcal{S}_1 \cup \mathcal{S}_2) > \#\text{C}(\mathcal{S}_1) + \#\text{C}(\mathcal{S}_1)$. Use the notations $\mathcal{C}_1 := \mathcal{C}^*(\mathcal{S}_1 \cup \mathcal{S}_2) \cap \mathcal{S}_1$ and $\mathcal{C}_2 := \mathcal{C}^*(\mathcal{S}_1 \cup \mathcal{S}_2) \cap \mathcal{S}_2$. We have

$$|\mathcal{C}_1| + |\mathcal{C}_2| \geq \#\text{C}(\mathcal{S}_1 \cup \mathcal{S}_2) > \#\text{C}(\mathcal{S}_1) + \#\text{C}(\mathcal{S}_1).$$

Since all values are non-negative, we must have $|\mathcal{C}_1| > \#\text{C}(\mathcal{S}_1)$ or $|\mathcal{C}_2| > \#\text{C}(\mathcal{S}_2)$, contradicting with the definition of $\#\text{C}(\mathcal{S}_1)$ or $\#\text{C}(\mathcal{S}_1)$, thus proving the right-hand side of subadditivity.

For the **Diversity** exploration measure, we denote it as Div$(\cdot)$ for short. We disprove its subadditivity by proving it violates the monotonicity corollary. For a molecule $x \in \mathcal{U}$ and a molecular set $\mathcal{S} \subseteq \mathcal{U}$ with size $n > 1$. If $x \notin \mathcal{S}$, we have

$$
\begin{aligned}
&\text{Div}(\mathcal{S} \cup \{x\}) \\
&= \frac{2}{(n+1)n} \left[\sum_{\substack{y,y' \in \mathcal{S} \\ y \neq y'}} d(y,y') + \sum_{y \in \mathcal{S}} d(x,y) \right] \\
&= \frac{n-1}{n+1} \cdot \text{Div}(\mathcal{S}) + \frac{2}{(n+1)n} \sum_{y \in \mathcal{S}} d(x,y).
\end{aligned}
$$

And the change in Div is

$$
\begin{aligned}
&\text{Div}(\mathcal{S} \cup \{x\}) - \text{Div}(\mathcal{S}) \\
&= \left(\frac{n-1}{n+1} - \frac{n+1}{n+1} \right) \text{Div}(\mathcal{S}) + \frac{2}{(n+1)n} \sum_{y} d(x,y) \\
&= \frac{2}{n+1} \left[-\text{Div}(\mathcal{S}) + \frac{1}{n} \sum_{y} d(x,y) \right].
\end{aligned}
$$

When the average distance of x and $\mathcal{S}$, *i.e.*, $\frac{1}{n}\sum_y d(x,y)$, is less than $\text{Div}(\mathcal{S})$ (*e.g.*, adding an molecule on the "segment" between two existing molecules), Diversity would decrease, thus violating the monotonicity corollary and proving Diversity is not subadditive.

For the **SumDiversity** exploration measure, we denote it as $\text{SD}(\cdot)$ for short. We disprove its subadditivity by proving it violates the right-hand side of the subadditivity, *i.e.*, $\text{SD}(\mathcal{S}_1 \cup \mathcal{S}_2) \leq \text{SD}(\mathcal{S}_1) + \text{SD}(\mathcal{S}_2)$. For two disjoint molecular sets with two molecules in each, *i.e.*, $\{x_1, x_2\}$ and $\{x_3, x_4\}$, we denote $d_{ij} = d(x_i, x_j)$. Then we have

$$
\begin{aligned}
&\text{SD}(\{x_1, x_2, x_3, x_4\}) \\
&= 4 \cdot \text{Div}(\{x_1, x_2, x_3, x_4\}) \\
&= 4 \cdot \frac{2}{4 \cdot 3} (d_{12} + d_{13} + d_{14} + d_{23} + d_{24} + d_{34}),
\end{aligned}
$$

and

$$
\text{SD}(\{x_1, x_2\}) + \text{SD}(\{x_3, x_4\}) = 2 \cdot d_{12} + 2 \cdot d_{34}.
$$

When the inter-set distances are larger than the inner-set distances, *i.e.*, $d_{13}, d_{14}, d_{23}, d_{24} > d_{12}, d_{34}$, we will have $\text{SD}(\{x_1, x_2, x_3, x_4\}) > \text{SD}\{x_1, x_2\} + \text{SD}\{x_3, x_4\}$, violating the right-hand side of the subadditivity, thus proving the SumDiversity measure is not subadditive.

For the **Diameter** exploration measure, we denote it as $\text{Dia}(\cdot)$ for short. We disprove its subadditivity by proving it violates the right-hand side of the subadditivity, *i.e.*, $\text{Dia}(\mathcal{S}_1 \cup \mathcal{S}_2) \leq \text{Dia}(\mathcal{S}_1) + \text{Dia}(\mathcal{S}_2)$. For two disjoint molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$ whose sizes are larger than one, we have

$$
\text{Dia}(\mathcal{S}_1 \cup \mathcal{S}_2) = \max_{\substack{x,y \in \mathcal{S}_1 \cup \mathcal{S}_2 \\ x \neq y}} d(x,y),
$$

and

$$
\text{Dia}(\mathcal{S}_1) + \text{Dia}(\mathcal{S}_2) = \max_{\substack{x,x' \in \mathcal{S}_1 \\ x \neq x'}} d(x,x') + \max_{\substack{y,y' \in \mathcal{S}_2 \\ y \neq y'}} d(y,y').
$$

When the maximum inter-set distance is larger than the maximum inner-set distance, *i.e.*, $\max_{x,y,\in \mathcal{S}_1 \cup \mathcal{S}_2} d(x,y) > \max_{x,x',\in \mathcal{S}_1} d(x,x') + \max_{y,y',\in \mathcal{S}_2} d(y,y')$, we will have $\text{Dia}(\mathcal{S}_1 \cup \mathcal{S}_2) > \text{Dia}(\mathcal{S}_1) + \text{Dia}(\mathcal{S}_2)$, violating the right-hand side of the subadditivity inequation, thus proving the Diameter measure is not subadditive.

For the **SumDiameter** exploration measure, we denote it as $\text{SD}(\cdot)$ for short. We disprove its subadditivity by proving it violates the right-hand side of the subadditivity, *i.e.*, $\text{SD}(\mathcal{S}_1 \cup \mathcal{S}_2) \leq \text{SD}(\mathcal{S}_1) + \text{SD}(\mathcal{S}_2)$. For two disjoint molecular sets with two molecules in each, *i.e.*, $\{x_1, x_2\}$ and $\{x_3, x_4\}$, we denote $d_{ij} = d(x_i, x_j)$. Then we have

$$
\text{SD}(\{x_1, x_2, x_3, x_4\}) = \sum_{i \in [4]} \max_{\substack{j \in [4] \\ j \neq i}} d(x_i, x_j),
$$

and

$$
\text{SD}(\{x_1, x_2\}) + \text{SD}(\{x_3, x_4\}) = 2 \cdot d_{12} + 2 \cdot d_{34}.
$$

When the inter-set distances are larger than the inner-set distances, *i.e.*, $d_{13}, d_{14}, d_{23}, d_{24} > d_{12}, d_{34}$, we will have $\text{SD}(\{x_1, x_2, x_3, x_4\}) > \text{SD}\{x_1, x_2\} + \text{SD}\{x_3, x_4\}$, violating the right-hand side of the subadditivity, thus proving the SumDiameter measure is not subadditive.

For the **Bottleneck** exploration measure, we denote it as $\text{Bot}(\cdot)$ for short. We disprove its subadditivity by proving it violates the monotonicity corollary. Consider adding a molecule x into a molecular set $\mathcal{S}$ with size $n > 1$. If $x \notin \mathcal{S}$, we have

$$
\text{Bot}(\mathcal{S} \cup \{x\}) = \min \left(\text{Bot}(\mathcal{S}),\ \min_{y \in \mathcal{S}} d(x,y) \right).
$$

When x introduces a more restricting bottleneck, *i.e.*, $\min_y\ d(x,y) < \text{Bot}(\mathcal{S})$, we will have $\text{Bot}(\mathcal{S}\cup\{x\}) < \text{Bot}(\mathcal{S})$, violating the monotonicity corollary, thus proving Bottleneck is not subadditive.

For the **SumBottleneck** exploration measure, we denote it as $\text{SB}(\cdot)$ for short. We disprove its subadditivity by proving it violates the monotonicity corollary. Consider adding a molecule x into a molecular set $\mathcal{S}$ with size $n > 1$. If $x \notin \mathcal{S}$, we have

$$\begin{aligned}&\text{SB}(\mathcal{S}\cup\{x\})\\ &= \sum_{y\in\mathcal{S}} \min\left(\min_{\substack{y'\in\mathcal{S}\\ y'\neq y}} d(y,y'),\ d(x,y)\right) + \min_{y\in\mathcal{S}} d(x,y),\end{aligned}$$

and

$$\text{SB}(\mathcal{S}) = \sum_{y\in\mathcal{S}} \min_{\substack{y'\in\mathcal{S}\\ y'\neq y}} d(y,y').$$

When x introduces some more restricting bottlenecks, *i.e.*, for many $y \in \mathcal{S}$, $d(x,y)$ is small (*e.g.*, adding a molecule into a set whose size is two, and the new molecule is added near one of the two molecules), we will have $\text{SB}(\mathcal{S}\cup\{x\}) < \text{SB}(\mathcal{S})$, violating the monotonicity corollary, thus proving SumBottleneck is not subadditive.

Fot the **DPP** exploration measure, we disprove its subadditivity by proving it violates the monotonicity corollary. Consider adding x into $\{x_0\}$ where $x \neq x_0$ $1 - d(x,x_0)$ is denoted as b. We have

$$\text{DPP}(\{x_0,x\}) = \begin{vmatrix} 1 & b \\ b & 1 \end{vmatrix} = 1 - b^2.$$

When $b > 0$ we will have $\text{DPP}(\{x_0,x\}) < \text{DPP}(\{x_0\}) = 1$, violating the monotonicity corollary, thus proving DPP is not subadditive.

□

Proposition B.2 (Dissimilarity property of exploration measures.)**.** *Diversity, SumDiversity, Diameter, Bottleneck, SumBottleneck, DPP, and #Circles have preferences to dissimilarity. The SumDiameter measure does not has the dissimilarity property.*

Proof. In the proof, for adding a molecule x into the molecular set $\{x_1,x_2\}$, we denote $d(x_1,x_2)$ as a, $d(x,x_1)$ as δ, and $d(x,x_2)$ as $a-\delta$ for short.

For the **Diversity** exploration measure, we denote it as $\text{Div}(\cdot)$ for short. For the molecular set $\{x_1,x_2\}$, when adding a new molecule x from $\mathcal{X}(x_1,x_2)$ as defined in the dissimilarity axiom, we have

$$\begin{aligned}&\text{Diversity}(\{x_1,x_2,x\}) = \frac{2}{3\cdot 2}(a+\delta+(a-\delta)) = \frac{2a}{3}\\ &= \text{Diversity}(\{x_1,x_2,x^*\}) = \frac{2}{3\cdot 2}(a+\frac{a}{2}+\frac{a}{2}) = \frac{2a}{3},\end{aligned}$$

which means it can take the optimal value at $x = x^*$, thus proving the dissimilarity property of Diversity.

For the **SumDiversity** exploration measure, we denote it as $\text{SD}(\cdot)$ for short. For the molecular set $\{x_1,x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1,x_2)$, we have

$$\begin{aligned}&\text{SD}(\{x_1,x_2,x\}) = a+\delta+(a-\delta) = 2a\\ &= \text{SD}(\{x_1,x_2,x^*\}) = a+\frac{a}{2}+\frac{a}{2} = 2a,\end{aligned}$$

which means it can take the optimal value at $x = x^*$, thus proving the dissimilarity property of SumDiversity.

For the **Diameter** exploration measure, we denote it as $\text{Dia}(\cdot)$ for short. For the molecular set $\{x_1,x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1,x_2)$, we have

$$\begin{aligned}&\text{Dia}(\{x_1,x_2,x\}) = \max\{a,\delta,a-\delta\} = a\\ &= \text{Dia}(\{x_1,x_2,x^*\}) = \max\{a,\frac{a}{2},\frac{a}{2}\} = a,\end{aligned}$$

which means it can take the optimal value at $x = x^*$, thus proving the dissimilarity property of Diameter.

For the **Bottleneck** exploration measure, we denote it as $\text{Bot}(\cdot)$ for short. For the molecular set $\{x_1,x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1,x_2)$, we have

$$\begin{aligned}&\text{Bot}(\{x_1,x_2,x\}) = \min\{a,\delta,a-\delta\} = \min(\delta,a-\delta)\\ &\leq \text{Bot}(\{x_1,x_2,x^*\}) = \min\{a,\frac{a}{2},a\frac{a}{2}\} = \frac{a}{2},\end{aligned}$$

which means it can take the optimal value at $x = x^*$, thus proving the dissimilarity property of Bottleneck.

For the **SumBottleneck** exploration measure, we denote it as $\text{SB}(\cdot)$ for short. For the molecular set $\{x_1,x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1,x_2)$, we have

$$\begin{aligned}&\text{SB}(\{x_1,x_2,x\}) = \delta+(a-\delta)+\min(\delta,a-\delta)\\ &\leq \text{SB}(\{x_1,x_2,x^*\}) = \frac{a}{2}+\frac{a}{2}+\frac{a}{2} = \frac{3a}{2},\end{aligned}$$

which means it can take the optimal value at $x = x^*$, thus proving the dissimilarity property of SumBottleneck.

For the **DPP** exploration measure, for the molecular set $\{x_1, x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1, x_2)$, we have

$$\begin{aligned}
&\text{DPP}(\{x_1, x_2, x\}) \\
&= \begin{vmatrix} 1 & 1-a & 1-\delta \\ 1-a & 1 & 1-(a-\delta) \\ 1-\delta & 1-(a-\delta) & 1 \end{vmatrix} \\
&= (2a-4)(\delta^2 - a\delta).
\end{aligned}$$

This term will take its largest value when $\delta = \frac{a}{2}$, *i.e.*, when $x = x^*$, thus proving the dissimilarity property of DPP.

For the **#Circles** exploration measure, we denote it as #C$(\cdot)$ for short. we denote #Circles$(\cdot)$ as #C$(\cdot)$ for short and define $\mathcal{C}^*(\mathcal{S}) \subseteq \mathcal{S}$ as an arbitrary set that satisfies $|\mathcal{C}^*(\mathcal{S})| = \#\text{C}(\mathcal{S})$. We prove the dissimilarity property of #C by contradiction.

For the molecular set $\{x_1, x_2\}$, consider adding an arbitrary new molecule $x \in \mathcal{X}(x_1, x_2)$ or the molecule x^*. We assume $\#\text{C}(\{x_1, x_2, x\}) > \#\text{C}(\{x_1, x_2, x^*\})$. This means that we must have $\#\text{C}(\{x_1, x_2, x\}) = 3$ and $\mathcal{C}^*(\{x_1, x_2, x\}) = \{x_1, x_2, x\}$, otherwise according to the monotonicity corollary of #C's subadditivity, $\#\text{C}(\{x_1, x_2, x\}) = \#\text{C}(\{x_1, x_2\}) \leq \#\text{C}(\{x_1, x_2, x^*\})$.

Therefore, according to the definition of #C, we have $\delta > t$ as well as $a - \delta > t$ where t is the distance threshold, indicating that

$$\frac{a}{2} > \min(\delta, a - \delta) > t.$$

This means $\mathcal{C}^*(\{x_1, x_2, x^*\})$ can also simultaneously include x_1, x_2, and x^*, contradicting with our assumption, thus proving the dissimilarity property of #Circles.

For the **SumDiameter** exploration measure, we denote it as SD$(\cdot)$ for short. For the molecular set $\{x_1, x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1, x_2)$, we have

$$\begin{aligned}
&\text{SD}(\{x_1, x_2, x\}) = a + \max(\delta, a - \delta) \\
&\geq \text{SD}(\{x_1, x_2, x^*\}) = a + \max(\frac{a}{2}, \frac{a}{2}) = \frac{3a}{2},
\end{aligned}$$

which means x can be better than x^* when one of δ and $a - \delta$ is larger than $\frac{a}{2}$, thus disproving the dissimilarity property of SumDiameter.

□

Discussing the dissimilarity property of coverage-based measures. We denote the cover$(\cdot, \cdot)$ function as cov$(\cdot, \cdot)$ for short. Then for the molecular set $\{x_1, x_2\}$, when adding a new molecule $x \in \mathcal{X}(x_1, x_2)$, we have

$$\begin{aligned}
&\text{Coverage}(\{x_1, x_2, x\}) \\
&= \sum_{y \in \mathcal{R}} \max\left(\text{cov}(x_1, y), \text{cov}(x_2, y), \text{cov}(x, y)\right).
\end{aligned}$$

The value of Coverage$(\{x_1, x_2, x\})$ will depend on the particular definition of cov$(\cdot, \cdot)$. Generally, an arbitrary coverage function does not necessarily meet the dissimilarity requirement. In our study, we define this function as $\text{cov}(x, y) := \mathbb{I}[\text{molecule } x \text{ contains fragment } y]$, where $\mathbb{I}[\cdot]$ is the indication function, which is loosely correlated with the distance metric d. So the dissimilarity property does not hold for exploration measures #FG, #RS, and #BM.

C. Random Subset Experiment Details

C.1. Bio-Activity Dataset

The 10K BioActivity dataset (Koutsoukas et al., 2014) contains 10,000 compound samples excerpted from the ChEMBL database (Gaulton et al., 2017) with bio-activity labels. These labels are the 50 largest ChEMBL activity classes, including enzymes (*e.g.*, proteases, lyases, reductases, hydrolases, and kinases) and membrane receptors (*e.g.*, GPCRs and non-GPCRs). The label distribution is shown in Figure 5.

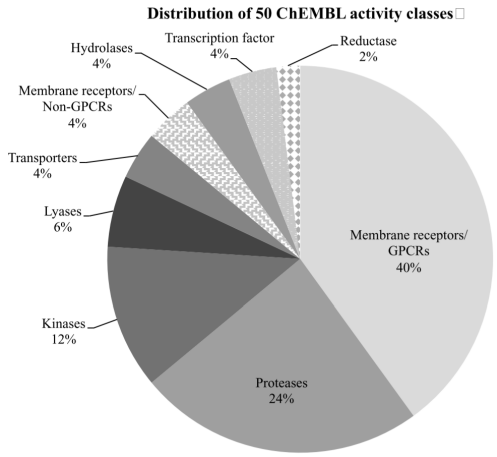


Figure 5. Label distribution of the BioActivity dataset (Koutsoukas et al., 2014). 50 bio-activity functionality classes are included.

We use UMAP (McInnes et al., 2018) to visualize the molecules in this dataset based on their Morgan fingerprints (Rogers & Hahn, 2010) as displayed Figure 6. From the visualization, we can see that the fingerprint similarity is indeed correlated with the bio-activity similarity.

Figure 6. UMAP visualization of compounds in the BioActivity dataset. Different colors stand for different bio-activity labels.

C.2. Random Subsets with Fixed Sizes

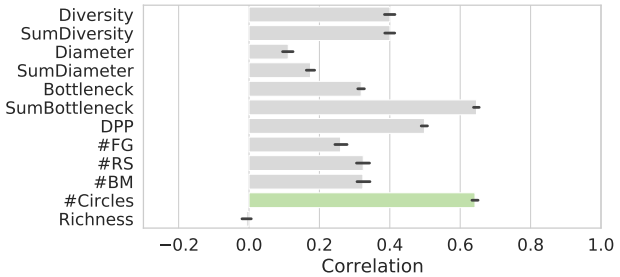
In this experiment, we repeat Algorithm 1 for ten times to obtain reliable correlations.

Algorithm 1 Calculating exploration measures for random subsets with fixed sizes.

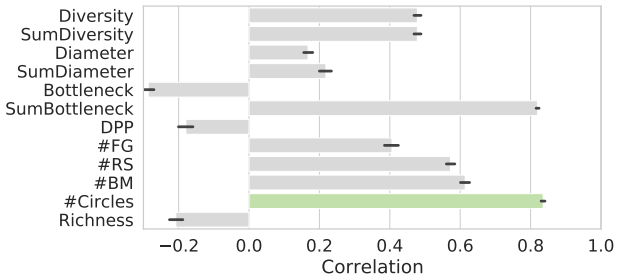
Input: The fixed subset size n; The bio-activity dataset $\{(x_i, y_i)\}_{i=1}^{10\text{K}}$ where $y_i \in \mathcal{Y}$ are bio-activity labels and $|\mathcal{Y}| = 50$; K exploration measures $\{\mu_k\}_{k=1}^{K}$.
repeat
  Sample a number m uniformly from $\{1, \ldots, 50\}$.
  Sample m labels $\mathcal{Y}'$ uniformly from $\mathcal{Y}$.
  Sample n molecules $\mathcal{S}$ with labels in $\mathcal{Y}'$ uniformly.
  Compute $\text{GS}(\mathcal{S})$ and $\mu_k(\mathcal{S})$ for $k \in [K]$.
until repeated for 1000 times
Calculate the correlations between GS and $\{\mu_k\}_{k=1}^{K}$ based on the 1000-times experiment results.

Experiment results. We show the experiment results for different fixed random set size n in Figure 7. We find the #Circles and SumBottleneck measures perform constantly better than all other measures.

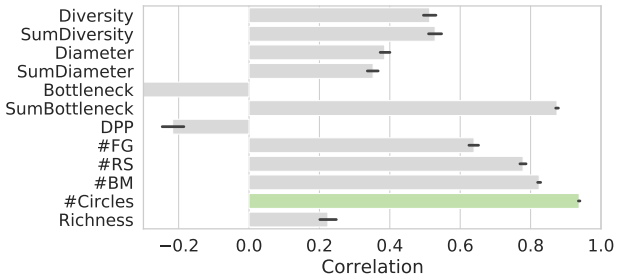
When the fixed size n increases, most exploration measures' performances also increase, except for Bottleneck and DPP, meaning they are not suitable for measuring the variety when the molecules are distributed crowdedly.



(a) Random subsets with a fixed size $n = 50$.



(b) Random subsets with a fixed size $n = 200$.



(c) Random subsets with a fixed size $n = 1000$.

Figure 7. Correlations between the gold standard GS and exploration measures in the fixed-size random subset setting. The fixed size is set as different values. A larger correlation indicates the better. The average results are obtained by running experiments independently for ten times.

Correlation between exploration measures. We also visualize the pairwise correlation between exploration measures in Figure 8. From the figure we can see that, the gold standard GS, #Circles, and SumBottleneck are most similar with each other in the fixed-size setting.

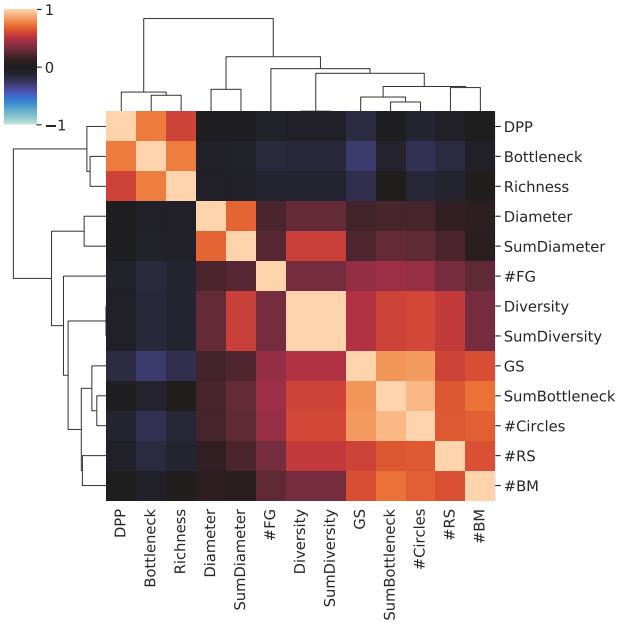


Figure 8. Correlations between exploration measures in the fixed-size random subset setting. The fixed size is set as $n = 200$. A larger correlation indicates the better. The average results are obtained by running experiments independently for ten times.

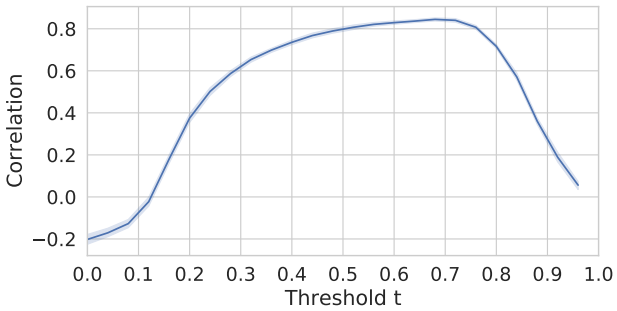


Figure 9. Correlations between the gold standard GS and the #Circles measure in the fixed-size random subset setting with different threshold t. The fixed size is set as $n = 200$. A larger correlation indicates the better. The average results are obtained by running experiments independently for ten times.

Threshold t for #Circles. The #Circles threshold t is selected to maximize the correlation to the gold standard GS. Taking $n = 200$ as an example, we test different t values as Figure 9 displays and select $t = 0.70$ as the threshold. We can see #Circles works well for a wide range of thresholds

like $[0.40, 0.70]$. In Olivecrona et al. (2017), the authors suggest to use a threshold $t = 0.60$ to decide whether two molecules are dissimilar with each other[3], which aligns our results. For $n = 50$ and $n = 1000$, the threshold is set as $t = 0.70$ and $t = 0.65$ respectively.

Distance metric d. We also study the impact of distance metric d. In Table 2, we listed the experiment results for both Tanimoto distance and VAE dissimilarity (Samanta et al., 2020). We find the experiment results obtained with the VAE dissimilarity remain consistent with the results obtained with the Tanimoto distance.

Table 2. Correlations between the gold standard and exploration measures in the fixed-size random subset setting. The fixed size is set as $n = 200$. A larger correlation indicates the better. Results are obtained by averaging ten independent experiments. Italic texts indicate molecular representation and the distance metric. Top three measures are highlighted in green, and the best measure is printed in bold.

	Distance-based	
	Tanimoto	*Latent space*
Diversity	0.478 ± 0.011	0.388 ± 0.040
SumDiversity	0.478 ± 0.011	0.388 ± 0.040
Diameter	0.179 ± 0.031	0.112 ± 0.022
SumDiameter	0.228 ± 0.028	0.201 ± 0.029
Bottleneck	-0.293 ± 0.015	-0.298 ± 0.027
SumBottleneck	0.821 ± 0.010	0.527 ± 0.013
DPP	-0.183 ± 0.021	-0.244 ± 0.030
	Coverage-based	
	Fgragment	
#FG	0.421 ± 0.033	
#RS	0.574 ± 0.025	
#BM	0.610 ± 0.028	
	Locality-based	
	Tanimoto	*Latent space*
#Circles	$\mathbf{0.831 \pm 0.008}$	0.745 ± 0.014
	SMILES	
Richness	-0.207 ± 0.025	

[3]In the original text, the authors suggest a similarity threshold of 0.40 that is equivalent to a distance threshold of 0.60.

C.3. Random Subsets with Growing Sizes

In this experiment, we repeat Algorithm 2 for ten times to obtain reliable DTW distances.

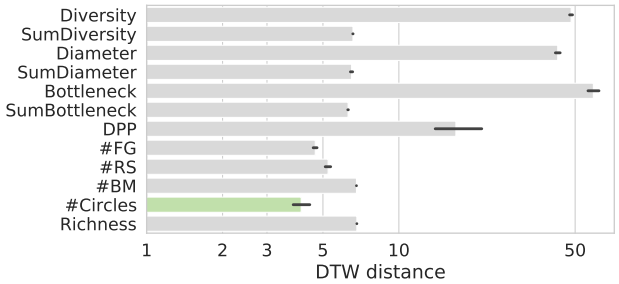
Algorithm 2 Calculating exploration measures for random subsets with growing sizes.

Input: The maximum subset size n; The bio-activity dataset $\{(x_i, y_i)\}_{i=1}^{10\text{K}}$ where $y_i \in \mathcal{Y}$ are bio-activity labels and $|\mathcal{Y}| = 50$; K exploration measures $\{\mu_k\}_{k=1}^{K}$.
Sample a number m uniformly from $\{1, \ldots, 50\}$.
Sample m labels $\mathcal{Y}'$ from $\mathcal{Y}$.
for i **in** $\{1, \ldots, n\}$ **do**
  Sample an unseen molecule x_i whose label is in $\mathcal{Y}'$.
  Set $\mathcal{S}_i := \{x_1, \ldots, x_i\}$
  Compute GS$(\mathcal{S}_i)$ and $\mu_k(\mathcal{S}_i)$ for $k \in [K]$.
end for
Plot exploration measure curves for GS and $\{\mu_k\}_{k=1}^{K}$ where the x axes are $i \in \{1, \ldots, n\}$ and the y axes are exploration measure values GS$(\mathcal{S}_i)$ and $\mu_k(\mathcal{S}_i)$.
Transform the cumulative curves into incremental ones.
Calculate DTW distances between incremental curves.

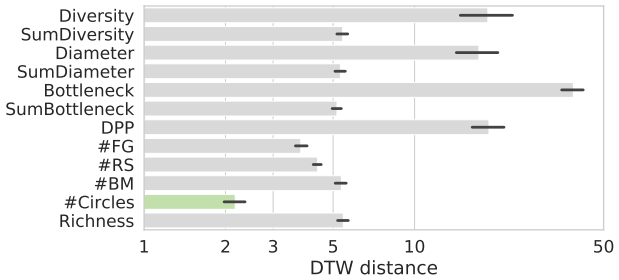
Experiment results. To mimic the way in which generation models propose new molecules, in Algorithm 2, we require the newly sampled molecule x_i to be similar to the already sampled molecules $\{x_1, \ldots, x_{i-1}\}$. The specific implementation can be found in our code[4]. Moreover, we also test the following two cases: (1) All molecules are sampled uniformly; (2) The newly sampled molecule x_i have to be most similar to the already sampled molecules $\{x_1, \ldots, x_{i-1}\}$. The results of DTW distances for these two cases are shown in Figure 10.

From Figure 10 we can see that, #Circles performs the best. Also, as the new molecules to add become more similar to the existing ones, the advantage of #Circles over other measures become larger. This makes #Circles especially suitable for measuring molecular generation models.

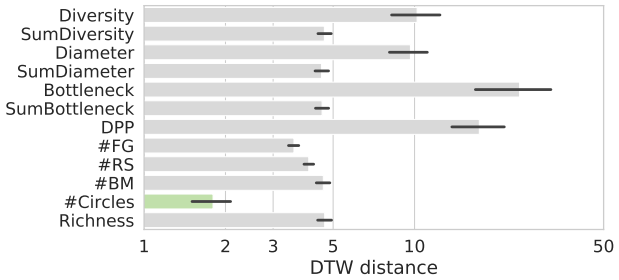
[4]The code will be released after publication.



(a) The new molecule x_i is uniformly sampled from all unseen molecules.



(b) The new molecule x_i needs to be similar to the already sampled ones $x_1, \ldots, x_{i=1}$.



(c) The new molecule x_i needs to be most similar to the already sampled ones $x_1, \ldots, x_{i=1}$.

Figure 10. DTW distances between the gold standard GS and exploration measures in the growing-size random subset setting. The maximum size is set as $n = 1000$. A smaller distance indicates the better. The average results are obtained by running experiments independently for ten times.

DTW distances between exploration measures. We visualize the pairwise DTW distances between exploration measures in Figure 11. From the figure we can see that, the gold standard GS and the #Circles measure are most similar with each other in the growing-size setting.

In addition, we find the Richness, SumDiversity, SumDiameter, SumBottleneck, and #BM are forming a large cluster, while #FG and #RS tend to be similar with each other.

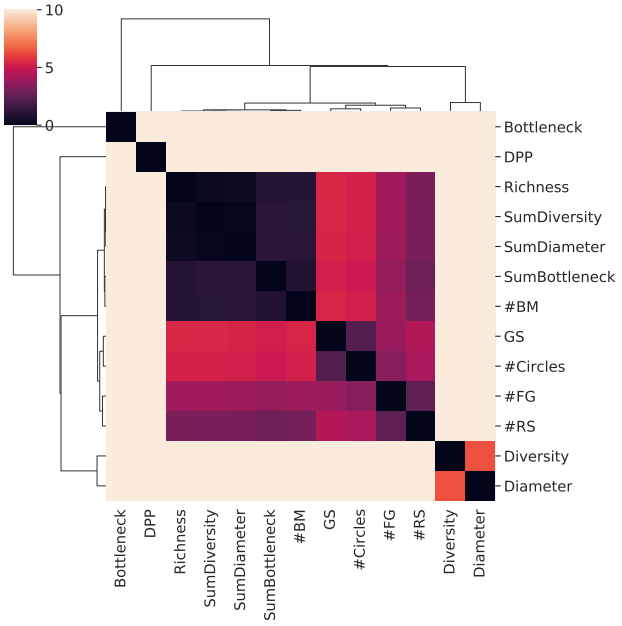


Figure 11. DTW distances between exploration measures in the growing-size random subset setting. The maximum size is set as $n = 1000$, and the new molecule needs to be similar to the already samples ones. A smaller distance indicates the better. The average results are obtained by running experiments independently for ten times.

Threshold t for #Circles. The #Circles threshold t is selected to minimize the DTW distance to the gold standard GS. Taking the second scenario as an example, we test different t values as Figure 12 displays and select $t = 0.76$ as the threshold. For the first and the third scenarios, the threshold is set as $t = 0.84$ and $t = 0.78$ respectively.

Distance metric d. We also study the impact of distance metric d. In Table 3, we listed the experiment results for both Tanimoto distance and VAE dissimilarity (Samanta et al., 2020). We find the experiment results obtained with the VAE dissimilarity remain consistent with the results obtained with the Tanimoto distance.

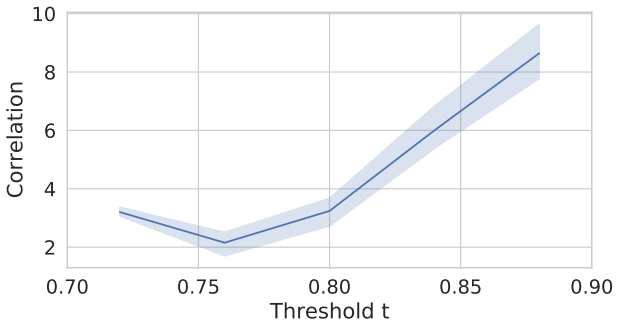


Figure 12. DTW distances between the gold standard GS and the #Circles measure in the growing-size random subset setting with different threshold t. The maximum size is set as $n = 1000$, and the new molecule needs to be similar to the already samples ones. A smaller distance indicates the better. The average results are obtained by running experiments independently for ten times.

Table 3. DTW distances between the gold standard and exploration measures in the growing-size random subset setting. The maximum size is set as $n = 1000$, and the new molecule needs to be similar to the already samples ones. A smaller distance indicates the better.

	Distance-based	
	Tanimoto	*Latent space*
Diversity	18.668 ± 6.973	30.063 ± 4.284
SumDiversity	5.425 ± 0.404	5.484 ± 0.296
Diameter	17.299 ± 4.801	28.071 ± 3.917
SumDiameter	5.328 ± 0.396	5.472 ± 0.297
Bottleneck	38.668 ± 5.769	37.168 ± 5.422
SumBottleneck	5.167 ± 0.353	5.432 ± 0.293
DPP	18.845 ± 3.962	12.052 ± 2.176
	Coverage-based	
	Fgragment	
#FG	3.797 ± 0.295	
#RS	4.382 ± 0.247	
#BM	5.365 ± 0.396	
	Locality-based	
	Tanimoto	*Latent space*
#Circles	$\mathbf{2.173 \pm 0.910}$	2.470 ± 0.629
	SMILES	
Richness	5.454 ± 0.347	

D. Molecular Generation Experiment Details

We implement the MARS baseline (Xie et al., 2021) by following the official code provided by the author and set the hyperparameters as default[5].

For the model variants, we test different α values from $\{0.1, 0.3, 1.0, 3.0\}$ and report the best performance. Specifically, we use $\alpha = 1.0$ for MARS+Diversity, $\alpha = 0.3$ for MARS+SumBottlenck, and $\alpha = 0.1$ for MARS+#Circles. The threshold we use for the #Circles measure is $t = 0.60$.

The computing server has two 2.4 GHz Intel Xeon Gold 6148 CPUs, 192GB memory (about 50G used), and one NVIDIA Tesla V100 GPU with 16G memory. It takes approximately 40 hours for the MARS model to sample in total 10M molecules within 2000 steps.

We show the molecular clustering results in Figure 13. Clusters are calculated based on Morgan fingerprints of the generated molecules and their Tanimoto similarity. Compared to the baseline model, a larger number of clusters can be obtained from MARS+SumBottleneck.

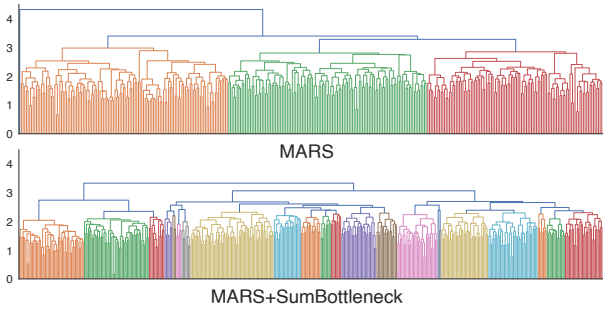


Figure 13. Hierarchical clustering results of generated molecules. Different colors stand for different clusters. The clustering threshold is set as $0.7\times$ the height of the highest linkage in accordance with SciPy's default value.

[5]MARS code: https://github.com/yutxie/mars.