HOW MUCH SPACE HAS BEEN EXPLORED? MEASURING THE CHEMICAL SPACE COVERED BY DATABASES AND MACHINE-GENERATED MOLECULES

Anonymous authors
Paper under double-blind review

ABSTRACT

Forming a molecular candidate set that contains a wide range of potentially effective compounds is crucial to the success of drug discovery. While most databases and machine-learning-based generation models aim to optimize particular chemical properties, there is limited literature on how to properly measure the coverage of the chemical space by those candidates included or generated. This problem is challenging due to the lack of formal criteria to select good measures of the chemical space. In this paper, we propose a novel evaluation framework for measures of the chemical space based on two analyses: an axiomatic analysis with two intuitive axioms that a good measure should obey, and an empirical analysis on the correlation between a measure and a proxy gold standard. Using this framework, we are able to identify a novel chemical space coverage measure, *#Circles*, superior to existing measures both analytically and empirically. We further evaluate how well the existing databases and generation models cover the chemical space in terms of #Circles. The results suggest that many generation models fail to explore a larger space over existing databases, which leads to new opportunities for improving generation models by encouraging exploration.

1 INTRODUCTION

To efficiently navigate through the huge chemical space for drug discovery, machine learning (ML) based approaches have been broadly designed and deployed, especially *de novo* molecular generation methods (Elton et al., 2019; Schwalbe-Koda & Gómez-Bombarelli, 2020; Bian & Xie, 2021; Deng et al., 2022). Such generation models learn to generate candidate drug designs by optimizing various *molecular property scores*, such as the binding affinity scores. 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 obtaining massive labeled training data for machine learning. However, high *in silico* property scores are far from sufficient, as there is usually a considerable misalignment between these scores and the *in vivo* behaviors. Costly wet-lab experiments are still needed to verify potential drug hits, where only a limited number of drug candidates can be tested.

In light of this cost constraint, it is critical to select or generate drug candidates not only with high *in silico* scores, but also covering a large portion of the chemical space. As functional difference between molecules is closely related to their structural difference (Huggins et al., 2011; Wawer et al., 2014), a better coverage of the chemical space will likely lead to a higher chance of hits in wet experiments. For this purpose, quantitative coverage measures of the chemical space become crucial. Such measures can both be used to evaluate and compare the *candidate libraries*[1], and be incorporated into training objectives to encourage ML models better explore the chemical space.

In this paper, we investigate the problem of quantitatively measuring the coverage of the chemical space by a candidate library. There have been a few such coverage measures of chemical space. For example, *richness* counts the number of unique compounds in a molecular set, and it has been used to describe how well a model is able to generate unique structures (Shi & von Itzstein, 2019; Polykovskiy et al., 2020). In addition, molecular fingerprints have been used to calculate pairwise similarity or distance between two compounds, and the average of these pairwise distances has been

[1]A candidate library is a collection of drug candidates filtered by rules or generated by ML models.

used to describe the overall *internal diversity* of a molecular set (Brown et al., 2019; Polykovskiy et al., 2020).

However, most existing coverage measures are heuristically proposed and the validity of these measures is rarely justified. In fact, defining the "right" measure for the coverage of chemical space coverage is challenging. Unlike the molecular property scores, there is no obvious "ground truth" about the coverage of chemical space. Moreover, the chemical space is complex and combinatorial, making the design of a good measure even more difficult.

To address the fundamental problem of properly measuring the coverage of chemical space for drug discovery, we propose a novel evaluation framework with two complementary criteria for evaluating the validity of coverage measures. We first formally define the concept of coverage measures on the chemical space (referred to as *chemical space measures*), where many existing heuristic measures fall into our definition (Section 3). Then we introduce the two criteria and compare various (existing and new) chemical space measures based on the criteria (Section 4). Specifically, the first criterion (Section 4.1) is based on an axiomatic analysis with two intuitive axioms that a good chemical space measure should satisfy. Surprisingly, most heuristic measures that are commonly used in literature, such as *internal diversity*, fail to satisfy these intuitive axioms. The second criterion (Section 4.2) compares the chemical space measures with a proxy of the gold standard: the number of unique biological functionalities covered by the set of molecules. We find that a novel chemical space measure (#Circles, defined in Section 3.2.3) not only satisfies both axioms but also better correlates with the gold standard.

Finally, we apply the #Circles measure to evaluate how well the existing databases and ML models cover the chemical space (Section 5). Interestingly, the evaluation results suggest that many ML models fail to explore a larger portion of chemical space compared to drug candidates obtained from virtual screening over existing databases. We believe these findings lead to a new direction to improve ML-based drug candidate generation models on better exploring the chemical space.

2 Related Work

2.1 Metrics for Evaluating Molecular Databases and Generation Methods

Molecular databases and machine-generated compounds are rich sources of drug candidates for forming a candidate library in drug discovery. To evaluate the quality of molecular databases and molecular generation methods, a variety of metrics are proposed. In general, four categories of evaluation metrics can be identified in the literature, which are related to: (1) bioactivities, (2) molecular properties, (3) data likelihood, or (4) the coverage of the chemical space, respectively.

The first category describes the biological activities of the included or generated molecules towards certain protein targets. Using docking simulators, researchers can calculate docking scores to determine the binding affinity between the given compound structure and the target (Trott & Olson, 2010). However, the simulation usually takes a long time to run. To obtain fast feedback and a sufficient amount of data to train molecular generation models, machine learning (ML) methods are also proposed to predict the bioactivity of molecules (Shoichet et al., 1992; Olivecrona et al., 2017; Li et al., 2018; Ahmadi et al., 2021; Jiang et al., 2022). The second category of metrics evaluates the molecular properties of compounds. Some of them are calculated based on heuristic rules, such as the validity of compounds, molecular weight, octanol-water partition coefficient (logP) (Wildman & Crippen, 1999), drug-likeness score (QED) (Bickerton et al., 2012), and synthetic accessibility (SA) (Ertl & Schuffenhauer, 2009). ML (especially deep learning)-based prediction models are also employed to predict the properties of compounds (Shen et al., 2010; Wang et al., 2019; Gilmer et al., 2017). Another branch of metrics uses data likelihood to evaluate the molecules output by generative models (Gómez-Bombarelli et al., 2018; Liu et al., 2018; Jin et al., 2018; De Cao & Kipf, 2018). A few other metrics are particularly related to the coverage of the chemical space or the extent of exploration in the chemical space. We review these metrics in the subsection below.

2.2 Metrics Related to the Coverage or the Exploration of the Chemical Space

Beyond bio-activity and property optimization, some studies evaluate molecular databases and molecular generation models with measures more or less related to the degree of coverage (or exploration) in the chemical space. Commonly used measures include richness, uniqueness, internal diversity, external diversity, KL divergence, and Fréchet ChemNet Distance (FCD) (Olivecrona et al.,

2017; 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; Jin et al., 2020; Xie et al., 2021). Besides, 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 molecular chemical space 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 chemical space measures in the chemical space is novel. With an empirical analysis in addition to the axiomatic analysis, we make practical recommendations on effective chemical space measures, including two novel measures.

3 Defining Chemical Space Measures

3.1 Definition of Chemical Space Measures

To define a chemical space measure, we first formalize the notion of chemical space by assuming a distance metric exists. This assumption is widely adopted in cheminformatics, using distance metrics such as the Tanimoto distance (Tanimoto, 1968; Bajusz et al., 2015).

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

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

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

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 Tanimoto distance is a metric function (Kosub, 2019; Lipkus, 1999).

Our definition and following analysis are generic to all distance functions and are not limited to 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 a chemical space measure as the following.

Definition 3.3 (chemical space measure)**.** Given the universal chemical space $\mathcal{U}$, a chemical space 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 Chemical Space 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: *reference-based* measures, *aggregation-based* measures, and *locality-based* measures. We are also able to define new chemical space measures under this formulation.

3.2.1 Reference-based Measures

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

$$\text{Coverage}(\mathcal{S}, \mathcal{R}) := \sum_{y \in \mathcal{R}} \left(\max_{x \in \mathcal{S}} \text{cover}(x, y) \right), \tag{2}$$

where cover(x, y) indicates how 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 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 reference-based measures as #FG, #RS, and #BM respectively. Another example is the Richness of the molecular set, *i.e.*, Richness $:= |\mathcal{S}|$ (Shi & von Itzstein, 2019; Polykovskiy et al., 2020), where cover$(x, y) := \mathbb{I}[x = y]$ and $\mathcal{R} = \mathcal{U}$.

3.2.2 AGGREGATION-BASED MEASURES

Though *reference-based* measures estimate chemical space coverage in an intuitive way, such measures highly rely on the characteristics of the reference set and do not consider the relations between the generated compounds. In contrast, the *aggregation-based* measures utilize the pair-wise distances among molecules and reflect the extent of chemical space coverage by aggregating the distances. For a set $\mathcal{S}$ with n molecules, several *aggregation-based* measures can be defined as follow:

$$\begin{aligned}
\text{Diversity}(\mathcal{S}; d) &:= \frac{2}{n(n-1)} \sum_{\substack{x,y \in \mathcal{S} \\ x \neq y}} d(x, y), & \text{SumDiversity}(\mathcal{S}; d) &:= \frac{1}{n-1} \sum_{\substack{x,y \in \mathcal{S} \\ y \neq x}} d(x, y), \\
\text{Diameter}(\mathcal{S}; d) &:= \max_{\substack{x,y \in \mathcal{S} \\ x \neq y}} d(x, y), & \text{SumDiameter}(\mathcal{S}; d) &:= \sum_{x \in \mathcal{S}} \max_{\substack{y \in \mathcal{S} \\ y \neq x}} d(x, y), \\
\text{Bottleneck}(\mathcal{S}; d) &:= \min_{\substack{x,y \in \mathcal{S} \\ x \neq y}} d(x, y), & \text{SumBottleneck}(\mathcal{S}; d) &:= \sum_{x \in \mathcal{S}} \min_{\substack{y \in \mathcal{S} \\ y \neq x}} d(x, y),
\end{aligned} \tag{3}$$

where x, y are molecules in $\mathcal{S}$, d is a distance metric as defined previously.

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[2] 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). The DPP measures is defined as DPP$(\mathcal{S}) := \det(\boldsymbol{S})$, where $\boldsymbol{S}$ is the Tanimoto similarity matrix of candidate molecules (*i.e.*, $1 - d(x, y)$ for Tanimoto distance).

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 chemical space measure that highlights the local neighborhoods covered by a set of molecules:

$$\#\text{Circles}(\mathcal{S}; d, t) := \max_{\mathcal{C} \subseteq \mathcal{S}} \ |\mathcal{C}| \quad \text{s.t.} \quad d(x, y) > t, \quad \forall x \neq y \in \mathcal{C}, \tag{4}$$

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 Richness$(\mathcal{S}) := |\mathcal{S}|$, which is the number of unique molecules (Shi & von Itzstein, 2019; Polykovskiy et al., 2020).

4 SELECTING THE RIGHT MEASURE

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

[2]The SumBottleneck measure is sometimes abbreviated as "SumBot" in the main paper.

variety of possible choices. Ideally, the selection criteria should not depend on the particular molecule properties of target or the algorithm used to generate the candidates. We achieve this through a qualitative axiomatic analysis and a quantitative correlation comparison.

4.1 Criterion #1: An Axiomatic Analysis of Chemical Space Measures

We propose two simple and intuitive principles that a good chemical space measure should satisfy. First, including more molecules should not decrease the degree of chemical space coverage. As it is easy to filter large candidate libraries into smaller subsets, including more molecules (even similar ones) is not harmful, and can render a higher probability of containing drug hits. Second, molecular sets with more dissimilar molecules should have a higher degree of chemical space coverage. These two principles are formalized below as two axioms and can be tested analytically.

Axiom 4.1 (Subadditivity)**.** *A good chemical space measure μ should be subadditive, i.e., for any two molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$, it holds that*

$$\max(\mu(\mathcal{S}_1), \mu(\mathcal{S}_2)) \leq \mu(\mathcal{S}_1 \cup \mathcal{S}_2) \leq \mu(\mathcal{S}_1) + \mu(\mathcal{S}_2). \tag{5}$$

For the first inequality, it states that a subadditive measure tends to increase when more molecules are included. This tendency is intuitive in drug discovery as testing more molecules means we can have a higher probability to discover a drug hit. Some direct corollaries and discussions on how this axiom connects to mathematical measures are in Appendix A.

Axiom 4.2 (Dissimilarity)**.** *A good chemical space measure should have a preference to dissimilar elements, i.e., for any three molecules $x_0, x_1, x_2 \in \mathcal{U}$, if $d(x_0, x_1) \geq d(x_0, x_2)$, it holds*

$$\mu(\{x_0, x_1\}) \geq \mu(\{x_0, x_2\}), \tag{6}$$

where d is the distance metric.

Intuitively, considering adding a new molecule into an existing molecular set $\mathcal{S} = \{x_0\}$, where there are two choices, namely x_1 and x_2, the more dissimilar one x_1 will be preferred.

Depending on whether each chemical space 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.

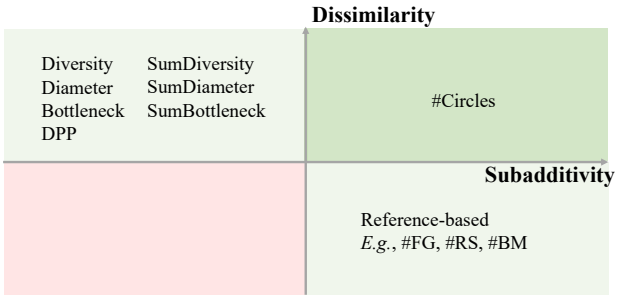


Figure 1: Chemical space measures, categorized by whether they satisfy subadditivity and/or dissimilarity. #Circles is the only measure that satisfies both simultaneously.

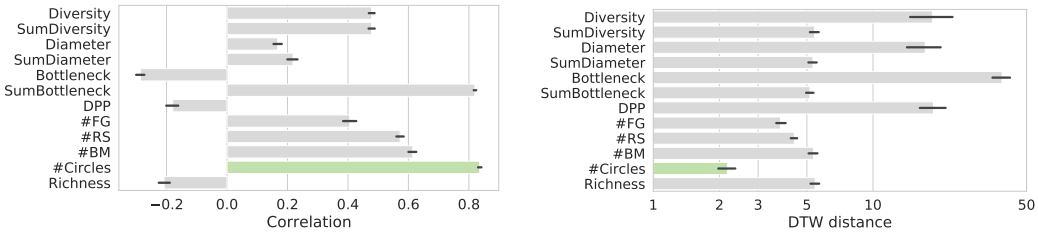
4.2 Criterion #2: Correlation with Biological Functionality

While only one measure stands out in the axiomatic analysis, it does not guarantee its empirical performance. Meanwhile, there are still multiple measures that satisfy *one* of the two axioms.

In this subsection, we further investigate the validity of the chemical space measures by testing their correlations with the biological functionality variety of molecules. Such functionalities can provide valuable information in distinguishing molecules and implicating chemical space coverage. We correlate the chemical space measures to the number of unique biological functionality labels covered by a molecular set. A better chemical space measure should have a higher correlation to the variety of biological functionalities, even though it is a proxy gold standard.

4.2.1 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



(a) Spearman's correlations between the gold standard GS and chemical space measures in the fixed-size setting. A higher correlation is better.

(b) DTW distances between the gold standard GS and chemical space measures in the growing-size setting, the smaller the better. A smaller distance is better.

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

dataset $\mathcal{S}$, we take the number of unique class labels as a proxy "gold standard" of the variety of the molecules in $\mathcal{S}$, *i.e.*, $\mathrm{GS}(\mathcal{S}) := \#$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 chemical space measures in two settings to find out which measures have the highest correlations with the coverage (or variety) of biological functionalities: (1) a *fixed-size* setting where random subsets of the dataset with fixed sizes are measured, and (2) a *growing-size* setting where we sequentially add molecules into a molecular set. We compare the measuring results of GS and chemical space measures with Spearman's correlation coefficient and dynamic time warping (DTW) distance. The experiment details are listed in Appendix C.

4.2.2 Results and Discussion

Experiment results in Figure 2 suggest that #Circles stands out for both empirical experiment settings. For the *fixed-size* setting, #Circles and SumBottleneck have notably higher correlations to the gold standard than other measures. This indicates that the locality information is critical, as these two measures both prefer new molecules that are at arm's length from their nearest neighbors. For the *growing-size* setting, #Circles surpasses all other chemical space measures. Reference-based measures such as #FG also perform prominently. Both #Circles and reference-based measures satisfy subadditivity, making them suitable for a "growth" setting. The Sum- variants of aggregation-based measures outperform their original forms, as they tend to increase when adding new molecules.

Table 1: Regarding the two suggested criteria, **#Circles** is the most recommended chemical space measure.

Measures	*(C1) Axiomatic properties (Sec. 4.1)*		*(C2) Correla -tions (Sec. 4.2)*	
	Subaddi -tivity	*Dissimi -larity*	*Fixed*	*Growing*
Diversity	✗	✓	Med	Low
SumBot	✗	✓	High	Med
DPP	✗	✓	Low	Low
#FG	✓	✗	Med	Med
#Circles	✓	✓	High	High
Richness	✓	✓	Low	Med

More experiment results can be found in Appendix C, where we also discuss the impact of the distance metric d. We also study the sensitivity of the subset size n, the way of adding new molecules, and the choice of the #Circles's threshold t. Results suggest that $t = 0.75$ is a good choice for both two settings. Correlations between chemical space measures are visualized in Figure 7 and 10.

4.3 Which Measure is the Right Choice?

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. Besides, SumBottleneck may be an effective choice when a fixed number of candidates are the target. Reference-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 in measuring and encouraging exploration for molecular generation. Our analysis and comparison results are summarized in Table 1.

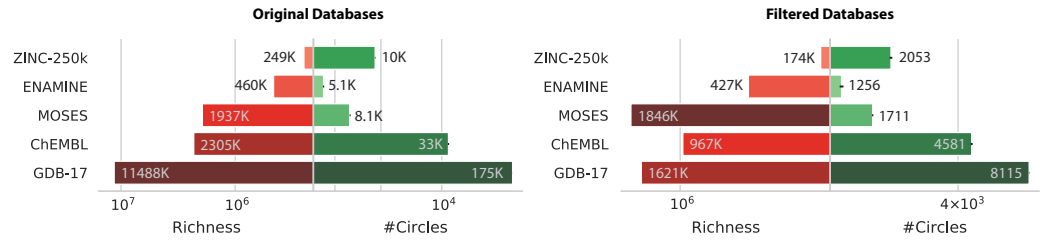


Figure 3: Richness and #Circles of molecular databases. Databases are listed vertically. The horizontal axes are chemical space measures (plotted in logarithmic scales). #Circles has the threshold value ($t = 0.75$). We find a discrepancy between Richness and #Circles, meaning that a larger database does not necessarily span a wider range of the chemical space.

5 Measuring Chemical Space Coverage

In this section, we apply the chemical space measures to evaluate how well the existing databases and ML-based molecular generation models cover the chemical space.

5.1 Measuring Chemical Space Covered by Molecular Databases

Other than being a rich source for virtual screening for drug purposes, molecular databases are also bonanzas of data for training ML models. In both scenarios, the coverage of chemical space is a crucial descriptor for comparing various databases. With a higher coverage of the chemical space, we are able to construct candidate libraries with better varieties. Also, the ML model would benefit from the diverse training data, obtaining a better understanding of the overall chemical space. Here, we compare five commonly used molecular databases with the selected measures.

5.1.1 Experiment Setup

Databases. (1) **ZINC-250k** is a random subset of the ZINC database (Irwin & Shoichet, 2005) and consists of 249K commercially-available compounds from different vendors for virtual screening; (2) **MOSES** (Polykovskiy et al., 2020) is another sub-collection of ZINC molecules. It contains approximately 2M molecules in total, filtered by atom types , ring cycle sizes, medicinal chemistry filters (MCFs), and PAINS filters. (3) **ChEMBL** (Gaulton et al., 2017) is a manually curated database of 2M bioactive molecules with known experimental data, especially with the inhibitory and binding properties against macromolecule targets. (4) **GDB-17** (Ruddigkeit et al., 2012) enumerates chemical structures that contain up to 17 atoms. The lead-like subset filters 11M compounds with lead-like properties. (5) **Enamine Hit Locator Library** contains a diverse collection of molecules with high MedChem tractability. The compounds are guaranteed synthesizable at a reasonable cost by the Enamine company.

Filtering. In drug discovery, drug-likeness and synthesizability of compounds are both important considerations. To this end, we compare both the original databases and the filtered databases. The quantitative drug-likeness (QED) (Bickerton et al., 2012) and the synthetic accessibility (SA) (Ertl & Schuffenhauer, 2009) are computed. The molecules are filtered with QED ≥ 0.6 and SA ≤ 4.

5.1.2 Results and Discussion

Figure 3 compares Richness and Circles of molecular databases. From the results, we conclude the findings as follows. (1) **A larger database does not necessarily span a wider range of the chemical space**: There is a discrepancy between Richness and #Circles, and the #Circles measuring results suggest that, in terms of chemical space coverage, the rank of the databases should be GDB-17 $>$ ChEMBL $>$ ZINC-250k $>$ MOSES $>$ Enamine. (2) **ZINC-250k and GDB-17 are recommended**: With the smallest amount of compounds, ZINC-250k renders a relatively high coverage of the chemical space, which is highly recommended for model training with limited computation resources. Constructed by enumeration, GDB-17 substantially expands the known chemical space to a great level, which is suitable for training ML models for chemical space representations. More details and findings on other chemical measures are provided in Appendix D.

5.2 MEASURING CHEMICAL SPACE EXPLORED BY GENERATION MODELS

5.2.1 PROBLEM FORMULATION

In drug discovery, the search for active molecules towards a druggable target is often formulated as the following unconstrained 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):

$$\arg\max_{x \in \mathcal{U}} \quad s_1(x) \circ s_2(x) \circ \cdots s_K(x), \tag{7}$$

where x is a molecule in the chemical space $\mathcal{U}$, $s_k : \mathcal{U} \to \mathbb{R}$ is a function scoring particular biological properties, and the "$\circ$" operator indicates the combination of multiple scores (*e.g.*, summation or multiplication). This scoring function can be binding affinity to protein targets, drug-likeness, synthesizability, *etc*. When wet-lab experiments are not available, these scores are obtained computationally.

Unfortunately, the majority of the computational scoring functions are merely approximations 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, and use this new objective to compare models:

$$\arg\max_{\mathcal{S} \subseteq \mathcal{U}} \quad \mu(\mathcal{S}) \quad \text{s.t.} \quad s_k(x) \geq C_k, \ \forall k \in [K], \ x \in \mathcal{S}, \tag{8}$$

where $\mathcal{S} \subseteq \mathcal{U}$ is a molecular candidate set, $\mu(\cdot)$ is a chemical space measure, $s_k(x) \geq C_k$ means the estimated property score of a molecule x is at or above a threshold C_k.

5.2.2 EXPERIMENT SETUP

Generation objective. Following Li et al. (2018) and Jin et al. (2020), we consider the inhibition against an Alzheimer-related target protein c-Jun N-terminal kinase-3 (JNK3) as the biological objective. The JNK3 binding affinity score is predicted by a random forest model based on Morgan fingerprint features of a molecule (Rogers & Hahn, 2010). We also consider the drug-likeness (QED) and synthetic accessibility (SA) as they are important in practical drug discovery scenarios.

Models. We employ four recently proposed molecular generation models to optimize the generation objective (JNK3, QED, and SA): (1) **RationaleRL** (Jin et al., 2020) generates molecules by combining rationales with reinforcement learning (RL); (2) **DST** (Fu et al., 2021) supports a gradient-based optimization on chemical graphs; (3) **JANUS** (Nigam et al., 2021) uses the genetic algorithm (GA) to design drugs inversely with parallel tempering; (4) **MARS** (Xie et al., 2021) generates fragment-based molecular structures based on Markov chain Monte Carlo (MCMC) sampling. Due to MARS' great ability to simultaneously optimize multiple drug discovery objectives, we also consider three variants of it (MARS+Diversity, MARS+SumBottleneck, and MARS+#Circles), where the chemical space measures are incorporated into Equation 7 to encourage exploration. In addition, we also include a virtual screening setting as a baseline, where the five **molecular databases** introduced in the previous subsection are combined, forming a candidate set of 11M compounds. Note that many target-irrelevant molecules are included.

Evaluation. To validate the ability in exploring the chemical space, we compare the molecules generated by models using Equation 8 with multiple chemical space measures. The molecules (including molecules generated by models as well as compounds from databases) are filtered with the constraints JNK3 ≥ 0.5, QED ≥ 0.6, and SA ≤ 4, representing the positive candidates. We examine seven different measures, and report #Circles, Richness, and Diversity here; the others are listed in Appendix E). Note we cannot directly measure the biological functionality coverage as in Section 4.2, as many of the generated molecules are not in the database. For RationaleRL and DST, 5000 molecules are generated as suggested in their papers. For JANUS and MARS, the population size and the number of MCMC chains are set as 5000. JANUS evolves for 10 generations, and MARS iterates 2000 steps. More implementation details are listed in Appendix E.

5.2.3 RESULTS AND DISCUSSION

Table 2 lists the chemical space measures estimated for the positive candidate compounds (*i.e.*, JNK3 ≥ 0.5, QED ≥ 0.6, and SA ≤ 4) generated by models or from databases. Based on these

Table 2: Measuring results of the chemical space explored by molecular generation methods. Incorporating chemical space measures into the objectives increases the exploration of the chemical space measures, but none of the tested ML-based models surpasses the virtual screening baseline (**Databases**) in terms of **#Circles** when $t = 0.75$ (as suggested in Sec. 4.2, highlighted in grey). Numbers in the parentheses following **#Circles** are values for the threshold t. In each chemical space measure, the larger value the better. **Bold** indicates the best performance in each measure.

Method	Richness	#Circles (0.70)	#Circles (0.75)	#Circles (0.80)	Diversity
Databases	1,250	78.0 ± 4.4%	**55.7** ± 3.7%	**30.3** ± 5.0%	**0.827**
RationaleRL	526 ± 1.7%	14.2 ± 6.8%	12.1 ± 6.5%	9.9 ± 0.3%	0.762 ± 0.3%
DST	29	2.0 ± 0.0%	1.0 ± 0.0%	1.0 ± 0.0%	0.559
JANUS	261	15.0 ± 11.6%	6.7 ± 8.7%	4.0 ± 25.0%	0.759
MARS	134K ± 10.2%	75.3 ± 4.4%	22.6 ± 13.3%	7.3 ± 16.0%	0.736 ± 0.8%
+Diversity	177K ± 13.7%	93.8 ± 7.6%	24.4 ± 9.8%	8.8 ± 13.9%	0.752 ± 0.2%
+SumBot	**221K** ± 13.4%	128.3 ± 8.1%	29.3 ± 7.6%	8.8 ± 13.0%	0.752 ± 0.3%
+#Circles	179K ± 17.4%	**129.9** ± 19.1%	30.2 ± 18.0%	9.7 ± 16.2%	0.749 ± 0.7%

results, we conclude the followings. (1) **ML-based models fail to explore a larger effectual area compared to databases**: Unexpectedly, we find the molecules obtained by virtual screening over databases span the largest area of the chemical space in terms of #Circles when $t = 0.75$. The most competitive model among the tested ones, MARS+#Circles, explores only about half of the space comparing to virtual screening. (2) **ML-based models like MARS tend to exploit locally instead of exploring globally**: For #Circles $t = 0.75$ and $t = 0.80$, MARS variants fail to surpass virtual screening, whereas $t = 0.70$ tells a different story. With a large t, #Circles computes the coverage of global regions in the chemical space. The results suggest that ML-based models like MARS are more capable of local exploitation rather than global exploration. (3) **Chemical space measures can encourage the model to explore**: By incorporating chemical space measures into the optimization objective of MARS, all measures are significantly improved during molecular generation. (4) **Diversity should be avoided as a descriptor for exploration**: MARS and its three variants all converge quickly in Diversity, despite a large number of new molecules discovered. Again, it suggests that the widely used Diversity is not suitable for measuring chemical space.

A few limitations of the measurement with #Circles should be noted. First, the computation of #Circles is combinatorial and the running time is exponential in theory. We implement a fast approximation by sacrificing accuracy (Appendix F). Second, the measuring results can be sensitive to the selection of the distant threshold t (circle diameter). Different thresholds are investigated in Appendix C, and more sensitivity tests and hyper-parameter tuning strategies are needed for determining the optimal threshold. More details and findings on other chemical measures are provided in Appendix E.

6 Conclusion and Future Work

In this paper, we have presented a systematic study on the coverage measures of chemical space in the context of drug discovery. The core contribution of this study is a novel evaluation framework for selecting proper coverage measures. Using the proposed framework, we have identified a new coverage measure, #Circles, that is superior to existing heuristic measures commonly used in the drug discovery literature. We have also quantitatively compared molecular databases and ML-based generation models in terms of #Circles. The results suggest that many ML-based generation models fail to explore a larger chemical space compared to virtual screening, since they tend to exploit locally instead of exploring widely.

Measuring the chemical space is a fundamental problem in drug discovery. Our work in this direction opens even more research questions that are worth investigating in the future. For example, it would be interesting to design and evaluate more chemical space measures under our framework. Moreover, the newly suggested objective for molecular generation (Equation 8) inspires further study, since it is computationally challenging as a combinatorial constrained optimization problem. In general, the proposed chemical space measures and the generic measure selection framework are not restricted to the application of drug discovery, but can also be used in other domains, such as material design, science of science, and text or image generation.

REFERENCES

Shahin Ahmadi, Zohreh Moradi, Ashwani Kumar, and Ali Almasirad. Smiles-based qsar and molecular docking study of xanthone derivatives as α-glucosidase inhibitors. *Journal of Receptors and Signal Transduction*, pp. 1–12, 2021.

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

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

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

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

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

Thomas Blaschke, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Memory-assisted reinforcement learning for diverse molecular de novo design. *Journal of cheminformatics*, 12(1):1–17, 2020.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019.

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

Jianyuan Deng, Zhibo Yang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics*, 23(1):bbab430, 2022.

Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

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

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

Hui Fang, Tao Tao, and ChengXiang Zhai. 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.

Tianfan Fu, Wenhao Gao, Cao Xiao, Jacob Yasonik, Connor W Coley, and Jimeng Sun. Differentiable scaffolding tree for molecular optimization. *arXiv preprint arXiv:2109.10469*, 2021.

Tomonori Fukutani, Kohei Miyazawa, Satoru Iwata, and Hiroko Satoh. 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.

Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The chembl database in 2017. *Nucleic Acids Research*, 45, 2017.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

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

Rafael Gómez-Bombarelli, Jennifer Nansean Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan Prescott Adams, and Alan Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

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

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

David John Huggins, Ashok Ramakrishnan Venkitaraman, and David Robert Spring. Rational methods for the selection of diverse screening compounds. *ACS Chemical Biology*, 6(3):208–217, 2011.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

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

Huaipan Jiang, Jian Wang, Weilin Cong, Yihe Huang, Morteza Ramezani, Anup Sarma, Nikolay V Dokholyan, Mehrdad Mahdavi, and Mahmut T Kandemir. Predicting protein–ligand docking structure with graph neural network. *Journal of Chemical Information and Modeling*, 62(12): 2923–2932, 2022.

Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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

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

Alexios Koutsoukas, Shardul Paricharak, Warren RJD Galloway, David R Spring, Adriaan P IJzerman, Robert C Glen, David Marcus, and Andreas Bender. 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.

Alex Kulesza and Ben Taskar. Learning determinantal point processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 419–427, 2011.

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

Yibo Li, Liang Ren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, 2018.

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

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L Gaunt. Constrained graph variational autoencoders for molecule design. *arXiv preprint arXiv:1805.09076*, 2018.

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

AkshatKumar Nigam, Robert Pollice, and Alan Aspuru-Guzik. Janus: parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *arXiv preprint arXiv:2106.04011*, 2021.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48–48, 2017.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:1931, 2020.

Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev. Molecularrnn: Generating realistic molecular graphs with optimized properties. *arXiv: Learning*, 2019.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

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

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. 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.

Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E Hubbard, and S David Morley. rdock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology*, 10(4):e1003571, 2014.

Soumitra Samanta, Steve O'Hagan, Neil Swainston, Timothy J Roberts, and Douglas B Kell. Vae-sim: a novel molecular similarity measure based on a variational autoencoder. *Molecules*, 25(15):3446, 2020.

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

Jie Shen, Feixiong Cheng, You Xu, Weihua Li, and Yun Tang. Estimation of adme properties with substructure pattern recognition. *Journal of chemical information and modeling*, 50(6):1034–1041, 2010.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

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

Brian K Shoichet, Irwin D Kuntz, and Dale L Bodian. Molecular docking using shape descriptors. *Journal of computational chemistry*, 13(3):380–397, 1992.

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

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

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 429–436, 2019.

Mathias J Wawer, Kejie Li, Sigrun M Gustafsdottir, Vebjorn Ljosa, Nicole E Bodycombe, Melissa A Marton, Katherine L Sokolnicki, Mark-Anthony Bray, Melissa M Kemp, Ellen Winchester, 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.

Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. Mars: Markov molecular sampling for multi-objective drug discovery. *arXiv preprint arXiv:2103.10432*, 2021.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay S. Pande, and Jure Leskovec. 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.

Jie Zhang, Rocío Mercado, Ola Engkvist, and Hongming Chen. Comparative study of deep generative models on chemical space coverage. *Journal of Chemical Information and Modeling*, 61(6):2572–2581, 2021.

A SUBADDITIVITY COROLLARIES AND DISCUSSIONS

Corollary A.1. ***(Subtraction)***. *If a chemical space 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 a chemical space 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 a chemical space 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 $\mu(\mathcal{S}_1) \leq \mu(\mathcal{S}_2)$.*

A subadditive chemical space measure is a special case of *outer measures* in the context of mathematical measure theory. *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, thus can hardly tell the variety of compounds and how much of the chemical space is been covered.

B PROOFS FOR SUBADDITIVITY AND DISSIMILARITY

In Figure 1 we show the chemical space 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 aggregation-based measures, when $|\mathcal{S}| = 1$, $\mu(\mathcal{S})$ is not defined. 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 chemical space measures.)**.** *Reference-based chemical space measures and #Circles are subadditive. Aggregation-based chemical space measures including Diversity, SumDiversity, Diameter, SumDiameter, Bottleneck, SumBottleneck, and DPP are not subadditive.*

Proof. For the **reference-based** chemical space measures, we denote the cover$(\cdot, \cdot)$ function as cov$(\cdot, \cdot)$ for short. We prove the subadditivity of reference-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** chemical space 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** chemical space 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 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** chemical space measure, we denote it as 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** chemical space 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** chemical space 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** chemical space 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** chemical space 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** chemical space 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 chemical space measures.)**.** *Diversity, SumDiversity, Diameter, SumDiameter, Bottleneck, SumBottleneck, DPP, and #Circles have preferences to dissimilarity.*

Proof. In the proof, for the three molecules $x_0, x_1, x_2 \in \mathcal{U}$ as stated in the dissimilarity axiom in Section 4.1, we assume $d(x_0, x_1) \geq d(x_0, x_2)$.

For the **Diversity** chemical space measure, we have

$$\text{Diversity}(\{x_0, x_1\}) = d(x_0, x_1) \geq d(x_0, x_2) = \text{Diversity}(\{x_0, x_2\}),$$

proving the dissimilarity property of Diversity.

For the **SumDiversity** chemical space measure, we have

$$\text{SumDiversity}(\{x_0, x_1\}) = 2 \cdot d(x_0, x_1) \geq 2 \cdot d(x_0, x_2) = \text{SumDiversity}(\{x_0, x_2\}),$$

proving the dissimilarity property of SumDiversity.

For the **Diameter** chemical space measure, we have

$$\text{Diameter}(\{x_0, x_1\}) = d(x_0, x_1) \geq d(x_0, x_2) = \text{Diameter}(\{x_0, x_2\}),$$

proving the dissimilarity property of Diameter.

For the **SumDiameter** chemical space measure, we have

$$\text{SumDiameter}(\{x_0, x_1\}) = 2 \cdot d(x_0, x_1) \geq 2 \cdot d(x_0, x_2) = \text{SumDiameter}(\{x_0, x_2\}),$$

proving the dissimilarity property of SumDiameter.

For the **Bottleneck** chemical space measure, we have

$$\text{Bottleneck}(\{x_0, x_1\}) = d(x_0, x_1) \geq d(x_0, x_2) = \text{Bottleneck}(\{x_0, x_2\}),$$

proving the dissimilarity property of Bottleneck.

For the **SumBottleneck** chemical space measure, we have

$$\text{SumBottleneck}(\{x_0, x_1\}) = 2 \cdot d(x_0, x_1) \geq 2 \cdot d(x_0, x_2) = \text{SumBottleneck}(\{x_0, x_2\}),$$

proving the dissimilarity property of SumBottleneck.

For the **DPP** chemical space measure defined with the Tanimoto similarity, denoting $d(x_0, x_1)$ and $d(x_0, x_2)$ as d_1 and d_2 respectively, when $d_1, d_2 \in [0, 1]$ (Tanimoto distances), we have

$$\text{DPP}(\{x_0, x_1\}) = \begin{vmatrix} 1 & 1 - d_1 \\ 1 - d_1 & 1 \end{vmatrix} = 2d_1 - d_1^2 \geq 2d_2 - d_2^2 = \begin{vmatrix} 1 & 1 - d_2 \\ 1 - d_2 & 1 \end{vmatrix} = \text{DPP}(\{x_0, x_2\}),$$

proving the dissimilarity property of DPP.

For the **#Circles** chemical space measure, we have

$$\#\text{Circles}(\{x_0, x_1\}) = 1 + \mathbb{I}[d(x_0, x_1) > t] \geq 1 + \mathbb{I}[d(x_0, x_2) > t] = \#\text{Circles}(\{x_0, x_2\}),$$

proving the dissimilarity property of #Circles.

□

Discussion on the dissimilarity property of reference-based measures. For a reference-based measure, considering adding a new molecule x_1 or x_2 to the existing moelculer set $\mathcal{S} = \{x_0\}$, we have

$$\text{Coverage}(\{x_0, x_1\}, \mathcal{R}) = \sum_{y \in \mathcal{R}} \left(\max(\text{cover}(x_0, y), \text{cover}(x_1, y))\right),$$
$$\text{Coverage}(\{x_0, x_2\}, \mathcal{R}) = \sum_{y \in \mathcal{R}} \left(\max(\text{cover}(x_0, y), \text{cover}(x_2, y))\right),$$

where the values of $\text{Coverage}(\{x_0, x_1\}, \mathcal{R})$ and $\text{Coverage}(\{x_0, x_2\}, \mathcal{R})$ will depend on the particular definition of $\text{cover}(\cdot, \cdot)$ and the choice of the reference set $\mathcal{R}$.

Generally, an arbitrary coverage function and an arbitrary reference set do not necessarily meet the dissimilarity requirement. In our study, we define the cover function as $\text{cover}(x, y) := \mathbb{I}[\text{molecule } x \text{ contains fragment } y]$, where $\mathbb{I}[\cdot]$ is the indicator function, and y is a fragment contained in the reference set $\mathcal{R}$. Some counter-examples can be easily constructed, for instance, if x_1 is far away from the reference set $\mathcal{R}$ while x_2 is very close to $\mathcal{R}$ (even contained in $\mathcal{R}$), then the Coverage measure will prefer the molecule x_2 instead of the more dissimilar molecule x_1. Therefore, the dissimilarity property does not hold for chemical space measures #FG, #RS, and #BM.

Discussion on a new definition of reference-based measure. In addition to counting the number of fragments in the reference set $\mathcal{R}$ contained in the molecular set $\mathcal{S}$, another way to define a reference-based measure is to investigate the distance from $\mathcal{S}$ to $\mathcal{R}$. Particularly, we could define such a reference-based measure as below:

$$\text{Coverage}(\mathcal{S}, \mathcal{R}; d, t) := \sum_{y \in \mathcal{R}} \mathbb{I}[\exists x \in \mathcal{S} \text{ such that } d(x, y) < t], \tag{9}$$

where $\mathbb{I}[\cdot]$ is the indicator function, d is the distance metric, and t is the distance threshold.

Similar to the discussion on #FG, #RS, and #BM, the characteristics of the reference-based measure defined by Eq. 9 highly relies on the property of $\mathcal{R}$, and an arbitrary reference set $\mathcal{R}$ can not guarantee the dissimilarity property of this measure.

We now prove that the reference-based measure defined by Eq. 9 satisfies subadditivity. In the proof, $\text{Coverage}(\cdot)$ refers to the definition in Eq. 9.

Proof. We first prove the subadditivity of the indicator function $\mathbb{I}[\cdot]$, where we consider combining any two molecular sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{U}$. For any reference $y \in \mathcal{R}$ and a distance threshold t, we have

$$\begin{aligned}
&\max\left(\mathbb{I}[\exists x \in \mathcal{S}_1 \text{ such that } d(x, y) < t],\ \mathbb{I}[\exists x \in \mathcal{S}_2 \text{ such that } d(x, y) < t]\right) \\
\leq\ &\mathbb{I}[\exists x \in \mathcal{S}_1 \cup \mathcal{S}_2 \text{ such that } d(x, y) < t] \\
\leq\ &\mathbb{I}[\exists x \in \mathcal{S}_1 \text{ such that } d(x, y) < t] + \mathbb{I}[\exists x \in \mathcal{S}_2 \text{ such that } d(x, y) < t].
\end{aligned}$$

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 defined by Eq. 9 is subadditive. □

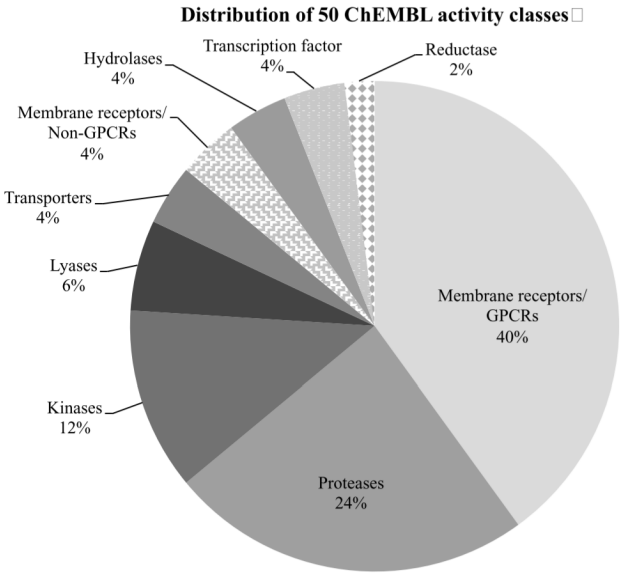


Figure 4: Label distribution of the BioActivity dataset (Koutsoukas et al., 2014). 50 bio-activity functionality classes are included.

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 4.

We use UMAP (McInnes et al., 2018) to visualize the molecules in this dataset based on their Morgan fingerprints as displayed Figure 5. From the visualization, we can see that fingerprint similarity is indeed correlated with bio-activity similarity.

Figure 5: UMAP visualization of compounds in the BioActivity dataset. Different colors stand for different bio-activity labels.

C.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 GS($\mathcal{S}$) as well as each chemical space measure $\mu(\mathcal{S})$. By repeating the randomization, we can calculate Spearman's correlation between the gold standard GS and each individual measure μ. We run the experiment for three different fixed sizes $n = 50$, $n = 200$, and $n = 1000$ to represent different molecular distribution density. The pair-wise correlations between measures are displayed in Figure 7.

Furthermore, when $n = 50$ and the molecules are distributed sparsely, all chemical space measures are positively correlated with the gold standard. However, when the subset size increases to $n = 200$ and $n = 1000$, Bottleneck and DPP becomes negatively correlated, while other measure tend to perform better. This is because these two measures will be bounded by the most similar molecular pair, thus severely conflicting with the subadditivity axiom.

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

Algorithm 1 Calculating chemical space 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 chemical space 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 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 6. We find the #Circles and SumBottleneck measures perform constantly better than all other measures.

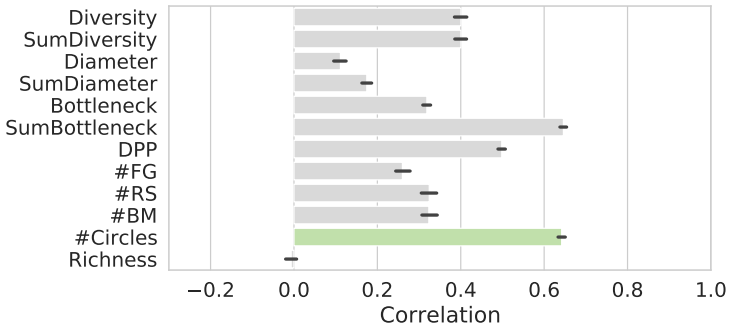
When the fixed size n increases, most chemical space measures' performances also increase, except for Bottleneck and DPP, meaning they are not suitable for measuring the variety when the molecules are distributed crowdedly.

Correlation between chemical space measures. We also visualize the pairwise correlation between chemical space measures in Figure 7. 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.

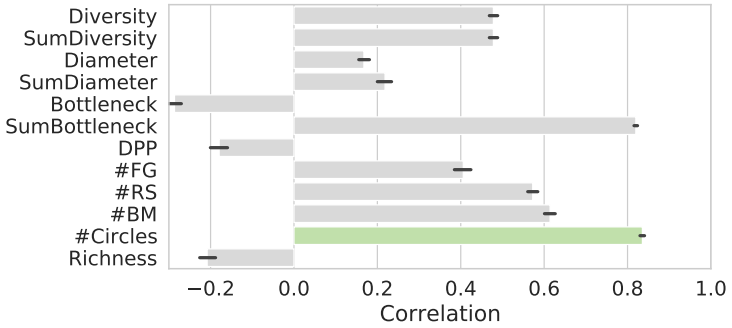
Threshold t for #Circles. The #Circles threshold t is selected to maximize the correlation to the gold standard GS. Taking $n = 200$ as an example, we test different t values as Figure 8 displays and select $t = 0.70$ as the threshold. We can see #Circles works well for a wide range of thresholds like $[0.40, 0.70]$. 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 3, we listed the experiment results for both fingerprint-based Tanimoto distance and VAE-based latent space 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.

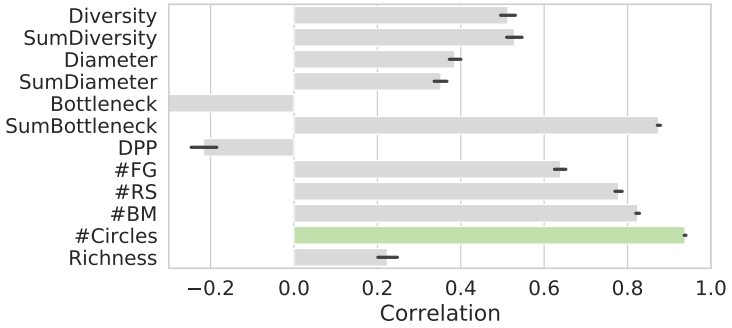
[3]In the original text, the authors suggest a similarity threshold of 0.40 that is equivalent to a distance threshold of 0.60.



(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 6: Correlations between the gold standard GS and chemical space 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.

C.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 chemical space 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 a chemical space 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.

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

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

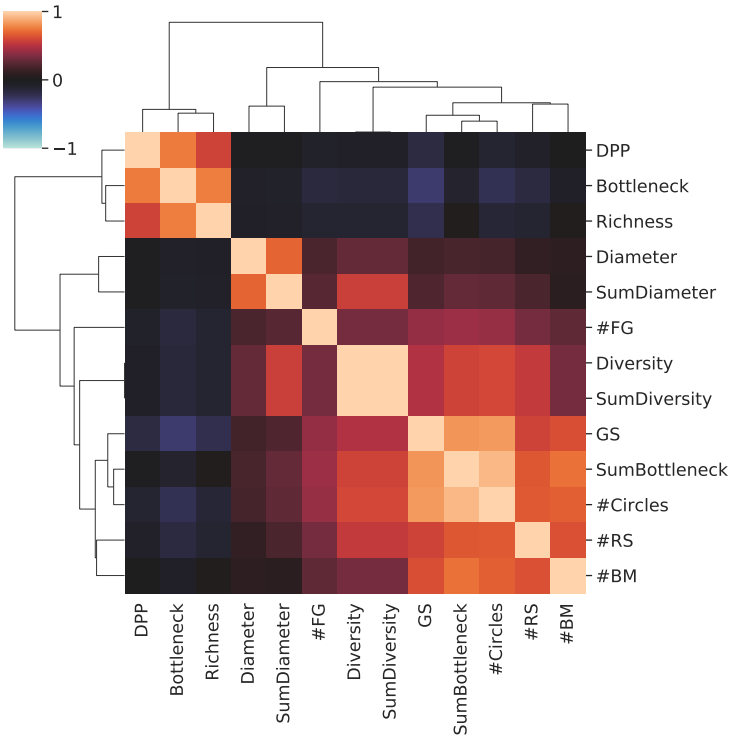


Figure 7: Correlations between chemical space measures in the fixed-size random subset setting. The fixed size is set as $n = 200$. A larger correlation indicates the better. The average results are obtained by running experiments independently for ten times.

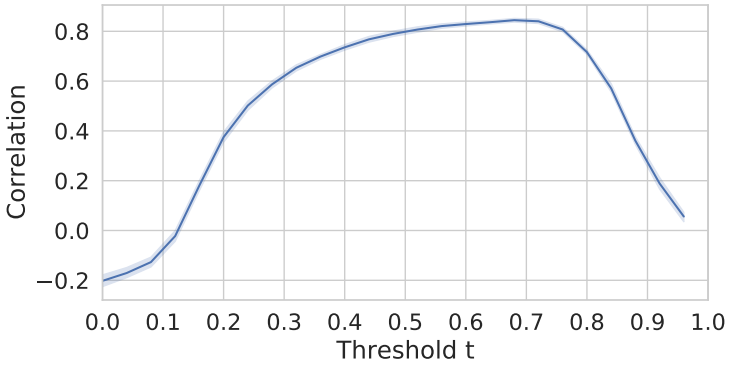
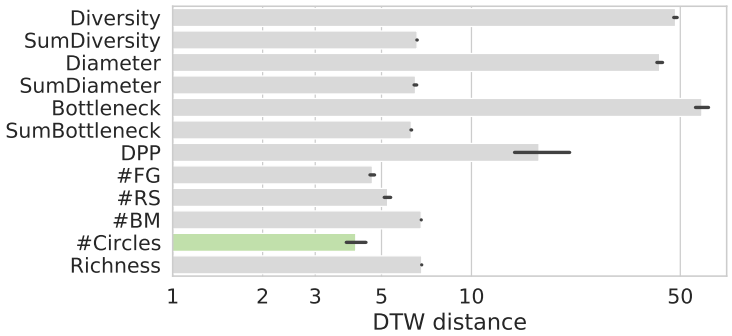


Figure 8: Correlations between the gold standard GS and the #Circles measure in the fixed-size setting with different threshold t. The fixed size is set as $n = 200$. A larger correlation indicates the better. The average results are obtained by running experiments independently for ten times.

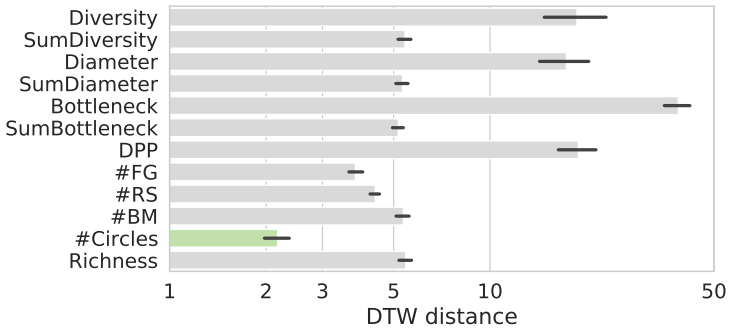
molecules $\{x_1, \ldots, x_{i-1}\}$. The specific implementation can be found in our code[4]. Moreover, we also test the following two cases: (1) All molecules are sampled uniformly; (2) The newly sampled molecule x_i have to be most similar to the already sampled molecules $\{x_1, \ldots, x_{i-1}\}$. The results of DTW distances for these two cases are shown in Figure 9.

From Figure 9 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 becomes larger. This makes #Circles especially suitable for measuring molecular generation models.

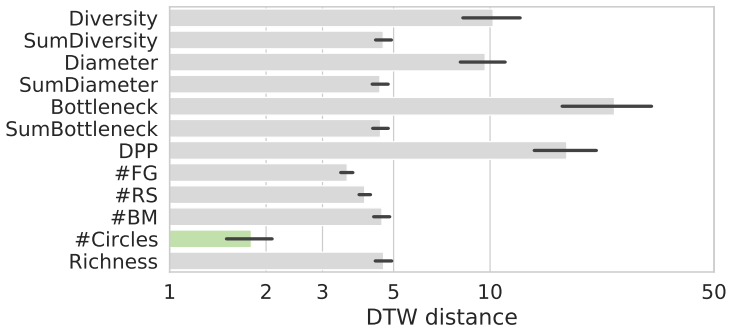
[4]The code will be released after publication.



(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 9: DTW distances between the gold standard GS and chemical space 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.

Table 3: Correlations between the gold standard and chemical space 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. The top three measures are highlighted in green, and the best measure is printed in **bold**.

	Aggregation-based	
	Tanimoto distance	*VAE dissimilarity*
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
	Reference-based	
	Fgragment	
#FG	0.421 ± 0.033	
#RS	0.574 ± 0.025	
#BM	0.610 ± 0.028	
	Locality-based	
	Tanimoto distance	*VAE dissimilarity*
#Circles	$\mathbf{0.831 \pm 0.008}$	0.745 ± 0.014
	SMILES	
Richness	-0.207 ± 0.025	

Algorithm 2 Calculating chemical space 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 chemical space 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 chemical space 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 chemical space 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.

DTW distances between chemical space measures. We visualize the pairwise DTW distances between chemical space measures in Figure 10. 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.

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 11 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 4, we listed the experiment results for both fingerprint-based Tanimoto distance and VAE-based latent space 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.

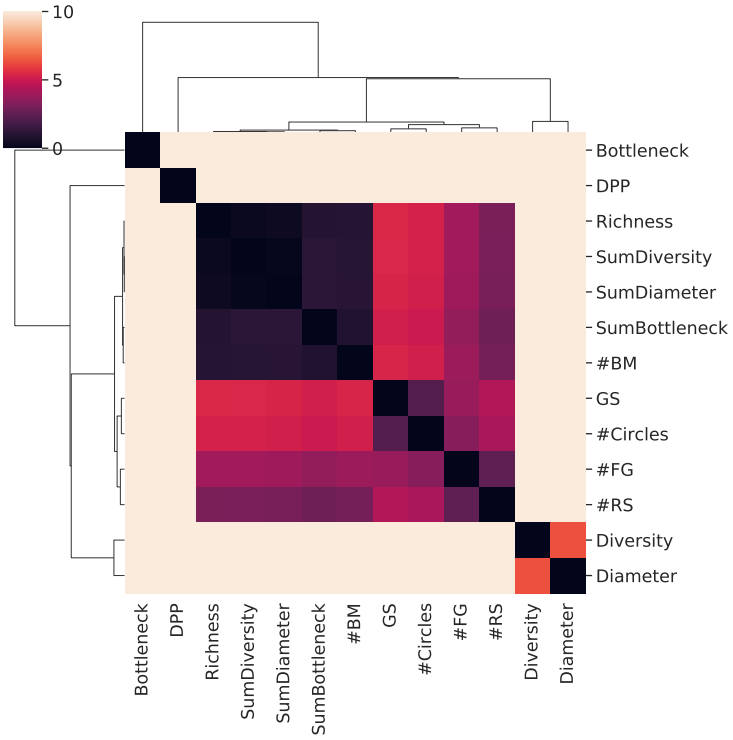


Figure 10: DTW distances between chemical space measures in the growing-size random subset setting. The maximum size is set as $n = 1000$, and the new molecule needs to be similar to the already samples ones. A smaller distance indicates the better. The average results are obtained by running experiments independently for ten times.

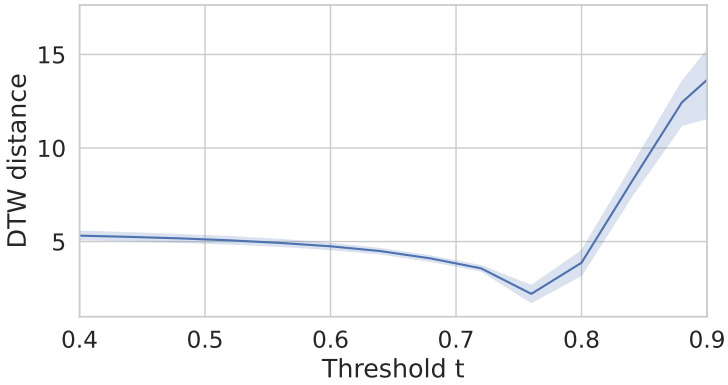


Figure 11: DTW distances between the gold standard GS and the #Circles measure in the growing-size random subset setting with different threshold t. The maximum size is set as $n = 1000$, and the new molecule needs to be similar to the already samples ones. A smaller distance indicates the better. The average results are obtained by running experiments independently for ten times.

D Measuring Molecular Databases

D.1 Molecular Databases

We measure the chemical space coverage for five molecular databases that are commonly used in virtual screening and generative model training:

Table 4: DTW distances between the gold standard and chemical space 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.

	Aggregation-based	
	Tanimoto distance	*VAE dissimilarity*
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
	Reference-based	
	Fgragment	
#FG	3.797 ± 0.295	
#RS	4.382 ± 0.247	
#BM	5.365 ± 0.396	
	Locality-based	
	Tanimoto distance	*VAE dissimilarity*
#Circles	$\mathbf{2.173 \pm 0.910}$	2.470 ± 0.629
	SMILES	
Richness	5.454 ± 0.347	

(1) **ZINC-250k**[5] is a random subset of the ZINC database (Irwin & Shoichet, 2005) and consists of 249K commercially-available compounds from different vendors for virtual screening;

(2) **MOSES** (Polykovskiy et al., 2020) is another sub-collection of ZINC molecules. It contains approximately 2M molecules in total, filtered by molecular weights (ranged range from 250 to 350 Daltons), the number of rotatable bonds (not greater than 7), water-octanol partition coefficient (logP, less or equal than 3.5), atom types (C, N, S, O, F, Cl, Br, and H), ring cycle sizes (no larger than 8), medicinal chemistry filters (MCFs), and PAINS filters.

(3) **ChEMBL** (Gaulton et al., 2017) is a manually curated database of 2M bioactive molecules with known experimental data, especially with the inhibitory and binding properties against macromolecule targets. This database contains not only drug and drug-like molecules, but also natural products and biopolymers.

(4) **GDB-17** (Ruddigkeit et al., 2012) enumerates chemical structures that contain up to 17 atoms of C, N, O, S, and halogens, overlapping with the molecular weight range typical for lead compounds. The lead-like subset filters 11M compounds with lead-like properties (100-350 MW & 1-3 clogP).

(5) Enamine is a company that provides compound libraries for high-throughput screening (HTS). The **Enamine Hit Locator Library**[6] is their largest diversity library with high MedChem tractability, and the compounds in the library are readily available for purchase and are guaranteed synthesizable at a reasonable cost.

D.2 Results on Chemical Space Measures

The measuring results are listed in Figure 12 and 13.

[5]ZINC-250k: `https://www.kaggle.com/datasets/1379f2461e75ef7a11d0c5ff3dd0c2440a6bcf33531bc0d303e8eac81a3a4b17`.

[6]ENAMINE diversity libraries: `https://enamine.net/compound-libraries/diversity-libraries`.

Database	Richness	#Circles (0.7)	#Circles (0.75)	#Circles (0.8)	Diversity	SumBottleneck	#FG	#RS	#BM
ZINC-250k	249455	10281	3683	1035	0.871	1523.23	3490	5075	136439
		10210	3670	1033	0.871	1587.14			
		10295	3653	1020	0.871	1602.87			
ENAMINE	460160	5145	1705	377	0.853	2611.66	150	1435	206497
		5154	1675	387	0.853	2570.52			
		5115	1678	370	0.853	2555.70			
MOSES	1936962	8095	2496	623	0.856	9316.20	1158	5816	449026
		8142	2496	644	0.857	9325.26			
		8178	2496	634	0.857	9264.65			
ChEMBL	2304875	32583	10956	2510	0.879	12003.99	17006	44814	716657
		32656	11173	2674	0.878	12076.51			
		32518	11082	2708	0.879	12019.48			
GDB-17	11487778	175380	29092	3296	0.898	73699.70	82060	1932562	3539100
		175393	29061	2842	0.898	73513.72			
			29116	2848	0.898	73730.85			

Figure 12: Measuring results on original databases.

Database	Richness	#Circles (0.75)	Diversity	SumBottleneck	#FG	#RS	#BM
ZINC-250k	173731	2035	0.861	1106.10	1952	2940	92715
		2084	0.861	1096.43			
		2040	0.863	1115.64			
ENAMINE	427366	1229	0.851	2461.75	142	1268	187742
		1275	0.851	2402.39			
		1263	0.851	2375.03			
MOSES	1846492	1714	0.856	8832.55	938	4513	413521
		1698	0.855	8955.37			
		1720	0.855	8971.22			
ChEMBL	967322	4570	0.872	5423.73	6304	14004	317255
		4631	0.872	5464.92			
		4543	0.873	5368.94			
GDB-17	1621490	8094	0.889	10733.83	5969	167795	425582
		8136	0.889	10670.31			
		8114	0.889	10770.83			

Figure 13: Measuring results on filtered databases.

E Molecular Generation Experiment Details

E.1 Model Implementation

We implement the models with the official repositories[7]. All hyperparameters are set as default.

E.2 MARS Variants

The terms incorporated into objectives are computed as follows[8]:

$$\text{Novelty}_{\text{Diversity}}(x, \mathcal{S}) := \frac{1}{|\mathcal{S}|} \sum_{y \in \mathcal{S}} d(x, y) \tag{10}$$

$$\text{Novelty}_{\text{SumBottleneck}}(x, \mathcal{S}) := \min_{y \in \mathcal{S}} d(x, y) \tag{11}$$

$$\text{Novelty}_{\text{\#Circles}}(x, \mathcal{S}) := \left[\min_{y \in \mathcal{S}} d(x, y) > t \right] \tag{12}$$

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 $t = 0.60$.

[7]RationaleRL: `https://github.com/wengong-jin/multiobj-rationale`.
DST: `https://github.com/futianfan/DST`.
JANUS: `https://github.com/aspuru-guzik-group/JANUS`.
MARS: `https://github.com/bytedance/markov-molecular-sampling`.

[8]Equations 10-12 are approximations of $[\mu(\mathcal{S} \cup \{x\}) - \mu(\mathcal{S})]$ defined to avoid numerical issues and for computational efficiency.

Table 5: Measuring results of the chemical space explored by molecular generation methods. In each chemical space measure, the larger value the better. **Bold** indicates the best performance in each measure.

Method	SumBottleneck	#FG	#RS	#BM
Databases	280	58	132	502
RationaleRL	9 ± 0.3%	39 ± 0.0%	**207 ± 0.0%**	442 ± 2.3%
DST	8	5	9	26
JANUS	76 ±	93 ±	73 ±	133 ±
MARS	535 ± 7.9%	346 ± 16.0%	67 ± 3.5%	19.5K ± 23.1%
+Diversity	715 ± 12.9%	463 ± 17.2%	67 ± 1.9%	**20.1K ± 25.2%**
+SumBot	**926 ± 12.5%**	**868 ± 15.3%**	66 ± 2.3%	14.0K ± 20.8%
+#Circles	742 ± 12.9%	601 ± 35.0%	69 ± 5.5%	18.9K ± 15.3%

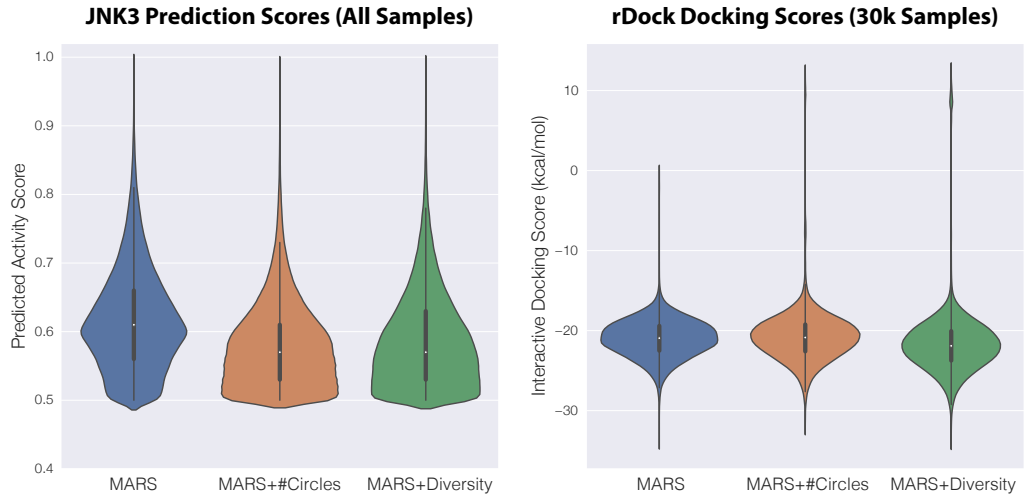


Figure 14: **Left**: Distributions of JNK3 machine learning model predicted score. We observe a downward trend in the MARS variant models, which are jointly optimized with a chemical space measure. However, all generated molecules are above the score threshold, and in practice, due to the limitation of the prediction model, they should be considered equally qualified candidates for further analysis. **Right**: Distributions of interaction scores of the ligands and the receptor docking (lower energy score indicates more favorable interactions). For the docking experiments, we randomly sampled 30K molecules from each generated library and performed molecular docking on the JNK3 target (PDB: 7KSI) using rDock (Ruiz-Carmona et al., 2014). The distributions of docking scores show that the joint optimizations on chemical space measures do not affect the binding affinity significantly when predicted by *in silico* docking.

E.3 Results on Other Measures

The results on other chemical space measures are listed in Table 5.

E.4 Molecular Property and Binding Affinity Distributions

To investigate how incorporating chemical space measures into the objective as Eq. 10-12 can influence the optimization for molecular properties, we examine the property distributions of molecules generated by MARS and its variants.

As shown in Figure 14-15, our resulting libraries from the joint optimizations with chemical space measures (*i.e.*, MARS+#Circles and MARS+Diversity) are not showing a significant downward shift on JNK3 binding affinity as well as QED scores and SA scores. It is expected that compared with vanilla MARS, the property scores of molecules generated by MARS variants would decrease slightly, because instead of solely optimizing the property scores, the model needs to make sacrifices for a more diverse library. However, this sacrifice would be insignificant compared to the diversity gain

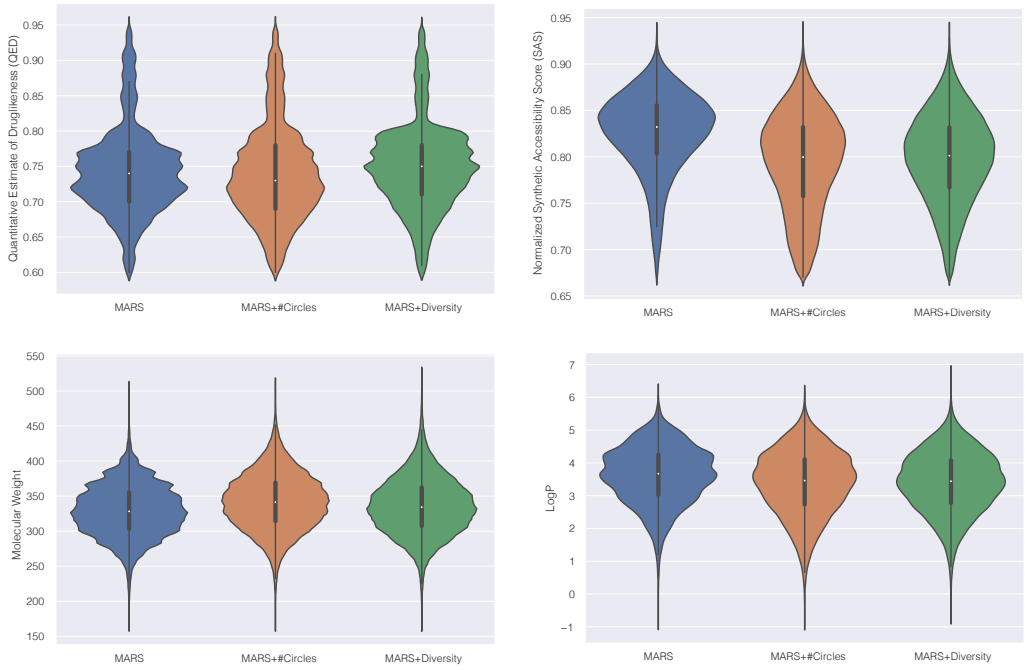


Figure 15: Selected molecular property distributions. Except for SAS (normalized from 0 to 1, the larger the better), all property distributions do not show significant changes across different models. The downward shift of the SAS is likely due to the increased diversity in the molecular libraries.

in the resulting library. In practice, these molecules would pass the scoring threshold and would be considered qualified in these regards. In Table 2, the experiment results show that by incorporating chemical space measures into optimization objectives, the richness of discovered qualified molecules (JNK3 ≥ 0.5, QED ≥ 0.6, and SA ≤ 4) got significantly improved.

E.5 Visualization

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 16, where we show the 2D layout of the Morgan fingerprints 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.

We show the molecular clustering results in Figure 17. 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 18 shows the dynamics of measures for MARS and its variants.

E.6 Curves of the #Circles Measure

We want to stress that the parameter t has an interpretable meaning. When t is smaller, the corresponding #Circles measure tends to reflect the "spread" of molecules at a finer granularity. When t is larger, the corresponding #Circles measure focuses more on the global picture of the molecule distribution. Therefore, we can intuitively characterize different molecular generation methods by examining #Circles values with different thresholds t.

The #Circles curves are displayed in Figure 19. The results suggest that different algorithms tend to show different patterns in exploring the chemical space. We find MARS and JANUS present

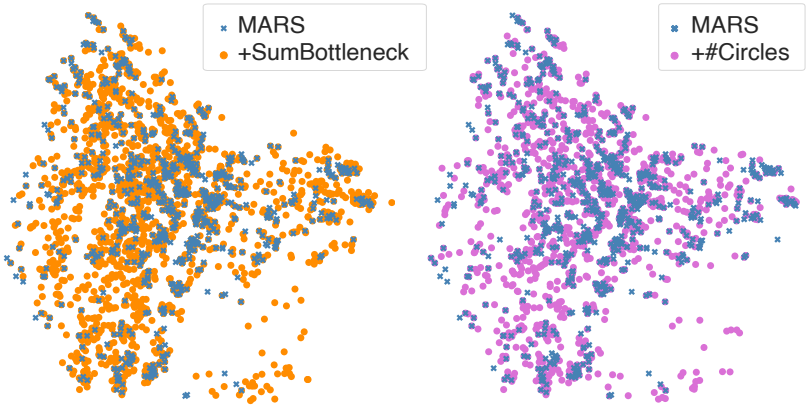


Figure 16: Optimizing chemical space measures encourages the molecular generation model to explore a larger span of the chemical space. This figure shows principal component analysis (PCA) of functional groups discovered by different models.

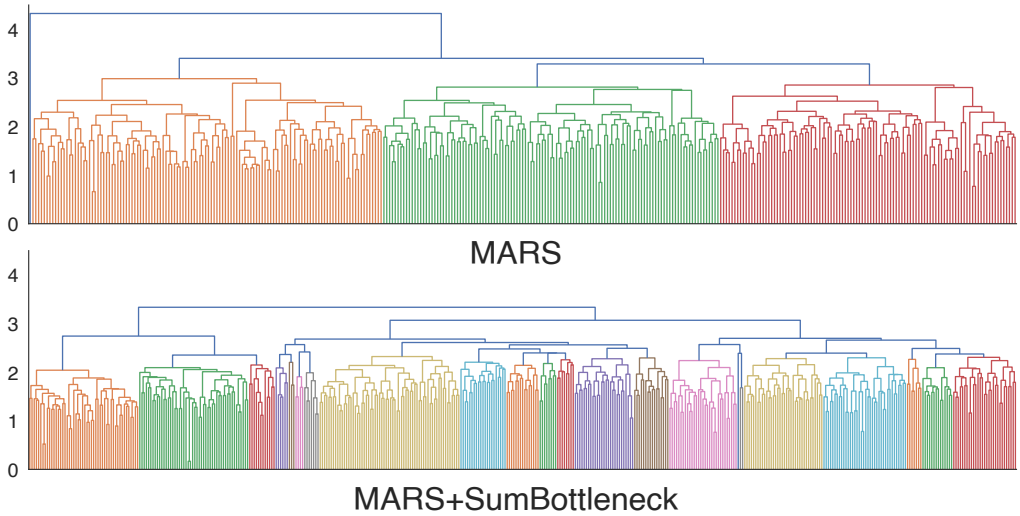


Figure 17: Hierarchical clustering results of generated molecules. Different colors stand for different clusters. The clustering threshold is set as $0.7\times$ the height of the highest linkage in accordance with SciPy's default value.

similar characteristics as they both have a relatively sharp decrease around $t = 0.75$. This might be because both two models are based on sampling algorithms (MARS is based on MCMC sampling while JANUS is based on the genetic algorithm) and tend to exploit locally in the chemical space. In contrast, RationaleRL tends to explore more globally because it is designed to combine active fragments (rationales), so the #Circles will not decrease too much when t gets larger.

We can also compare the models by adopting the concept of Pareto optimization. We find the DST, JANUS, and the vanilla MARS model are completely dominated by the variant MARS+#Circles. Besides, the RationaleRL method is dominated by the virtual screening approach.

F Fast Approximation of #Circles

The computation of #Circles is combinatorial and the running time is exponential in theory. We implement two fast approximation by sacrificing accuracy as Algorithm 3 and Algorithm 4.

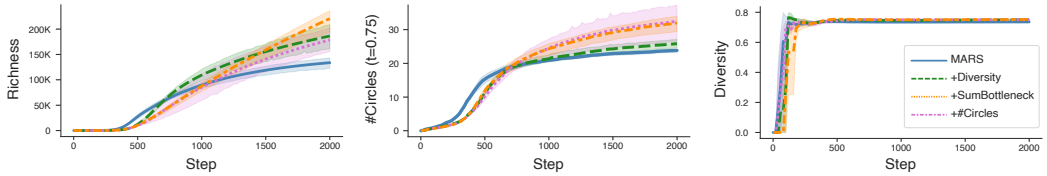


Figure 18: The dynamics of chemical space measures over generated molecules. Incorporating chemical measures into generation objectives increases the exploration of the chemical space.

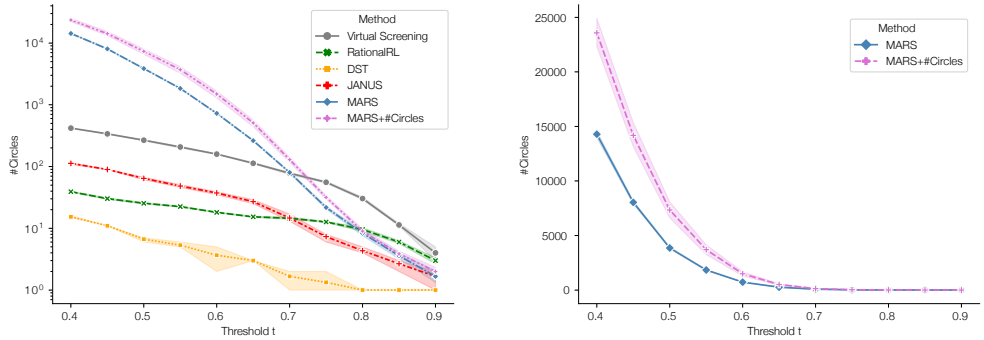


Figure 19: **Left**: #Circles vs. its threshold t selection for various models in the log scale. Curves of the #Circles measure reflects characteristics of molecular generation methods. **Right**: Linear scale shows MARS+#Circles improves significantly over the baseline model MARS. More discussions are in Appendix E.6.

Algorithm 3 Approximation of #Circles (sequential).

Input: The molecular set $\mathcal{S}$ to be measured, the distance metric d, the distance threshold t.
Randomly reorder the molecular set $\mathcal{S}$.
Initialize an empty set $\mathcal{C}$.
Set $n := |\mathcal{S}|$.
for i **in** $\{1, \ldots, n\}$ **do**
  Add $\mathcal{S}_i$ to $\mathcal{C}$ if $\min_{y \in \mathcal{C}} \; d(x, y) > t$.
end for
Return: $|\mathcal{C}|$ and $\mathcal{C}$.

Algorithm 4 Approximation of #Circles (recursive).

Input: The molecular set $\mathcal{S}$ to be measured, the distance metric d, the distance threshold t, maximum number of recursive layers L, number of processors m.
Call Algorithm 3 with arguments $(\mathcal{S}, d, t)$ if $L = 0$.
Randomly reorder the molecular set $\mathcal{S}$.
Evenly split $\mathcal{S}$ into m subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m$.
Call Algorithm 4 with arguments $(S_i, d, t, L-1, m)$ for $i = 1, \ldots, m$ and collect the returns $\mathcal{C}_i$.
Set $\mathcal{C} := \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_m$.
Return: Results obtained by calling Algorithm 3 with arguments $(\mathcal{C}, d, t)$.