Journal of Cheminformatics

METHODOLOGY

Open Access

Hamiltonian diversity: effectively measuring molecular diversity by shortest Hamiltonian circuits

Xiuyuan Hu[1,2], Guoqing Liu[2], Quanming Yao[1], Yang Zhao[1] and Hao Zhang[1*]

Abstract

In recent years, significant advancements have been made in molecular generation algorithms aimed at facilitating drug development, and molecular diversity holds paramount importance within the realm of molecular generation. Nonetheless, the effective quantification of molecular diversity remains an elusive challenge, as extant metrics exemplified by Richness and Internal Diversity fall short in concurrently encapsulating the two main aspects of such diversity: quantity and dissimilarity. To address this quandary, we propose Hamiltonian diversity, a novel molecular diversity metric predicated upon the shortest Hamiltonian circuit. This metric embodies both aspects of molecular diversity in principle, and we implement its calculation with high efficiency and accuracy. Furthermore, through empirical experiments we demonstrate the high consistency of Hamiltonian diversity with real-world chemical diversity, and substantiate its effects in promoting diversity of molecular generation algorithms. Our implementation of Hamiltonian diversity in Python is available at: https://github.com/HXYfighter/HamDiv.

Scientific contribution

We propose a more rational molecular diversity metric for the community of cheminformatics and drug development. This metric can be applied to evaluation of existing molecular generation methods and enhancing drug design algorithms.

Keywords Computer-aided drug design, Molecular generation, Molecular diversity, Shortest Hamiltonian circuit

Introduction

Thanks to the tremendous development in computational tools and machine learning algorithms, computer-aided drug discovery (CADD) has grown considerably in recent years, which can significantly shorten the time of the drug development process [1–3]. De novo molecular design algorithms [4–13] can generate candidate molecules with desirable in silico chemical and biological property scores [14, 15], which can provide meaningful inspirations for downstream preclinical studies and clinical trials.

However, due to the gap between in silico scores and in vivo behaviors of molecules [16], pharmacologists expect the upstream algorithms to provide as diverse a collection of drug candidates as possible, which can increase the probability of them eventually designing a drug to market [17]. Moreover, diverse drug compounds may assist in addressing drug resistance and side effects. Therefore, for molecule generation methods, the diversity of generated candidates is one of their pivotal aspects of performance.

For molecular diversity, besides a large size of the molecular set, we also expect the molecules to be dissimilar to each other, since similar structures cannot

*Correspondence:
Hao Zhang
haozhang@tsinghua.edu.cn
[1] Department of Electronic Engineering, Tsinghua University, Beijing, China
[2] Microsoft Research AI for Science, Beijing, China

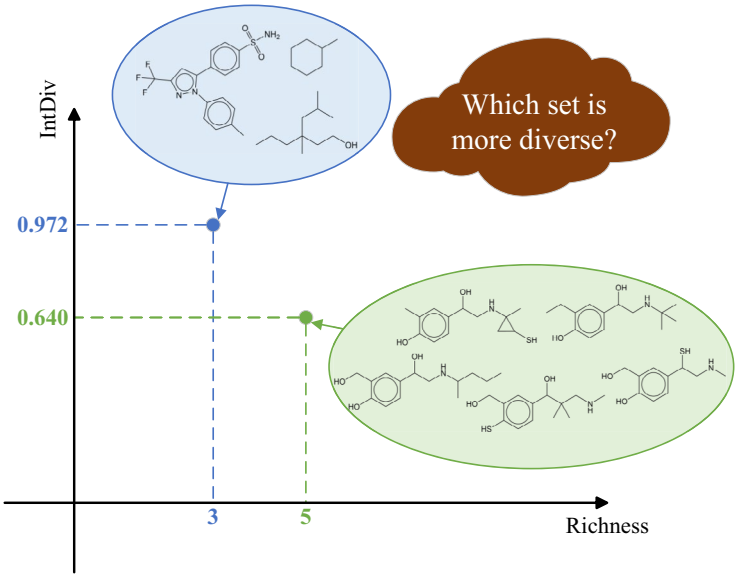

Fig. 1 An example illustrating the problem of using `Richness` and `IntDiv`, which provide contradictory comparison results of molecular diversity. Concretely, compared with the "blue" molecular set, the "green" set has a higher `Richness` and a lower `IntDiv`, so we cannot tell which set is more diverse according to these two metrics

provide much inspiration for pharmacologists. `Richness` [18] and `IntDiv` (internal diversity) [19], currently the two most widely used molecular diversity metrics in CADD, respectively measure the quantity and dissimilarity of molecules, which are two fundamental aspects of molecular diversity. Nevertheless, as illustrated in Fig. 1, higher `Richness` may coincide with lower `IntDiv`, making the comparison of molecular diversity controversial. Hence, we need a more comprehensive metric to assess molecular diversity that reflects both the quantity and dissimilarity of a molecular set.

To address this challenge, we propose Hamiltonian diversity, a novel metric for molecular diversity, in this paper. Specifically, our contributions include:

- We review existing molecular diversity metrics theoretically and formulate two general principles of an ideal molecular diversity metric: monotonicity and dissimilarity. None of the existing metrics satisfy both principles simultaneously.
- We propose a new Hamiltonian diversity based on the shortest Hamiltonian circuit, which adheres to both principles of molecular diversity metrics. We also provide an intuitive explanation and efficient implementation for the novel metric.
- We demonstrate the high consistency between Hamiltonian diversity and real-world chemical diversity. When incorporated into existing molecule generation algorithms, Hamiltonian diversity also helps promote the diversity of generated molecules.

Existing molecular diversity metrics

The drug-like chemical space $\mathcal{S}$ is vast, with an estimated 10^{33} synthesizable molecular structures [20], and it cannot be described in dimensions. Therefore, a metric function is needed in drug design to evaluate the span of a molecular set in $\mathcal{S}$, that is, the molecular diversity.

A molecular diversity metric φ is a function that maps a set of molecules $\mathcal{M} \subseteq \mathcal{S}$ to a non-negative real number which reflects the diversity of the set:

$$\varphi : \mathcal{P}(\mathcal{S}) \to [0, +\infty), \tag{1}$$

where $\mathcal{P}(\cdot)$ denotes the power set. In particular, $\varphi(\emptyset) = 0$.

Some existing metrics for molecular diversity meet the above definition, and they can be divided into two categories: reference-based and distance-based.

Reference-based metrics

Reference-based metrics intuitively compare a molecular set $\mathcal{M}$ with a reference set $\mathcal{R}$:

$$\varphi(\mathcal{M}; \mathcal{R}) := \sum_{r \in \mathcal{R}} \mathbb{I}\big(\exists m \in \mathcal{M},\ m = r \text{ or } m \text{ contains } r\big), \tag{2}$$

where $\mathbb{I}$ is the indicator function. $\mathcal{R}$ can be either a set of molecules or a set of molecular fragments, as a result of

which we use the above formulation which is applicable to different kinds of reference sets.

`Richness` [18] is a widely used reference-base metric where $\mathcal{R} = \mathcal{S}$, counting the number of unique molecules in a set, i.e., $|\mathcal{M}|$.

Distance-based metrics

On the other hand, distance-based metrics quantify molecular diversity based on pairwise distances among molecules instead of depending on given reference sets:

$$\varphi(\mathcal{M}; d) := f\left(\{d(x,y) | \forall x, y \in \mathcal{M}, x \neq y\}\right), \tag{3}$$

where f is a function to be defined, and d is a distance metric between a pair of molecules:

$$d : \mathcal{S} \times \mathcal{S} \to [0, +\infty). \tag{4}$$

The Tanimoto distance [21], denoted as d_T, between extended-connectivity fingerprints (ECFP) [22] of small molecules is a widely used metric function and is considered the most appropriate choice for the distance metric in the chemical space [23, 24]:

$$d_T(x,y) := 1 - \frac{\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i}{\sum_{i=1}^{n} \boldsymbol{x}_i + \sum_{i=1}^{n} \boldsymbol{y}_i - \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i} \in [0,1], \tag{5}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^n$ are n-dimensional binary fingerprints of two molecules $x, y \in \mathcal{S}$.

It is worth noting that our subsequent definition and analysis also apply to other molecular distance functions besides the Tanimoto distance, including Fréchet ChemNet distance [25], VAE-dissimilarity [26] and G-RMSD [27]. The Tanimoto distance is adopted as the default distance function because it is mathematically a metric function, which is theoretically essential for the properties of Hamiltonian diversity. as stated in Section 3.3.

Among distance-based metrics, `IntDiv` [19], which measures the average pairwise distance of a molecular set, is currently the most popular one for the evaluation of molecule generation approaches [1, 14]. The recently proposed `Circles` [28] is defined by the maximum number of mutually exclusive circles of radiuses t that can fit into a molecular set $\mathcal{M}$ as neighborhoods, with a subset of it $\mathcal{C} \subseteq \mathcal{M}$ as the circle centers.

Principles of molecular diversity metrics

The objectives of a molecule generation algorithm include achieving both high `Richness` and high `IntDiv`. However, these two metrics are orthogonal to each other, meaning they represent distinct aspects of the molecular sets. Therefore, it is desirable to have an ideal molecular diversity metric that combines both aspects into a single value. With a given distance metric d, we formulate these two aspects as two principles, respectively:

Principle 1 (Monotonicity). A good molecular diversity metric φ should be monotonic, i.e., for a molecule $x \in \mathcal{S}$ and a molecular set $\mathcal{M} \subseteq \mathcal{S}$, it holds that

$$\varphi(\mathcal{M} \cup \{x\}) \geq \varphi(\mathcal{M}), \tag{6}$$

and the equality holds if the distance between x and a molecule in $\mathcal{M}$ is 0[1]:

$$\exists\, m \in \mathcal{M},\ d(m,x) = 0. \tag{7}$$

Principle 2 (Dissimilarity). A good molecular diversity metric φ should be positively correlated with molecular distances, i.e., for any three molecules $x, y, z \in \mathcal{S}$, it holds that

$$\varphi(\{x,y\}) > \varphi(\{x,z\}),\ \text{if}\ d(x,y) > d(x,z). \tag{8}$$

To ensure the validity of a molecular diversity metric [29], it should satisfy both principles.

As a contrast, the design of `Circles` [28] follows another set of principles for molecular diversity metrics that diverge from our principles in several aspects: (1) their monotonicity principle lacks the inclusion of the equality condition; (2) their dissimilarity principle merely requires a non-strict positive correlation ($\geq$), contrasting with our stricter criterion; (3) they propose a subadditivity principle, derived from the additivity property of outer measures, but we do not accept this principle since no practical insights are provided to support their introduction of "sub". These differences lead to the possibility that the results of the principled analysis of metrics in [28] may differ from ours.

Table 1 provides an overview of existing molecular diversity metrics, demonstrating that none of them fully adhere to both monotonicity and dissimilarity principles. Consequently, the development of a new metric is imperative to tackle this key challenge in drug design.

Hamiltonian diversity

In this section, we present a novel metric for quantifying molecular diversity, Hamiltonian diversity, based on identifying the shortest Hamiltonian circuit within the chemical distribution of a molecular set.

[1] A distance of 0 between two molecules usually indicates that they are almost identical, but they are possible to be slightly different. For example, two slightly different compounds may correspond to the same fingerprint, so the Tanimoto distance d_T between them is 0.

Table 1 Overview of existing molecular diversity metrics

Category	Metric	Definition	Monotonicity	Dissimilarity
Reference-based $\varphi(\mathcal{M};\mathcal{R})$	`Richness` [18]	The number of unique molecules ($\mathcal{R}=\mathcal{S}$)	√	×
	`FG` [30]	The number of unique functional groups ($\mathcal{R}$ = all possible functional groups)	√	×
	`RS` [30]	The number of unique ring systems ($\mathcal{R}$ = all possible ring systems)	√	×
	`BM` [17]	The number of unique Bemis-Murcko scaffolds ($\mathcal{R}$ = all possible Bemis-Murcko scaffolds)	√	×
Distance-based $\varphi(\mathcal{M};d)$	`IntDiv` [19]	$\frac{1}{\lvert\mathcal{M}\rvert(\lvert\mathcal{M}\rvert-1)}\sum_{(x,y)\in\mathcal{M}\times\mathcal{M},x\neq y} d(x,y)$	×	√
	`SumDiv` [28]	$\frac{1}{(\lvert\mathcal{M}\rvert-1)}\sum_{(x,y)\in\mathcal{M}\times\mathcal{M},x\neq y} d(x,y)$	×	√
	`SumDiam` [28]	$\sum_{x\in\mathcal{M}}\max_{y\in\mathcal{M},y\neq x} d(x,y)$	×	√
	`SumBot` [28]	$\sum_{x\in\mathcal{M}}\min_{y\in\mathcal{M},y\neq x} d(x,y)$	×	√
	`Circles` [28]	$\max_{\mathcal{C}\subseteq\mathcal{M}} \lvert\mathcal{C}\rvert$ s.t. $d(x,y)>t, \forall x\neq y\in\mathcal{C}$ $t\in[0,1)$	√	×

Definition

We first formulate a molecular set $\mathcal{M}=\{m_1,m_2,...,m_n\}$ $(n\geq 2)$ as an undirected complete graph $K_n(\mathcal{M})=(\{m_i\},\{d_{ij}\})$, $i,j=1,2,...,n$, $i\neq j$. In this graph, each vertex represents a molecule m_i, and the edge weight between vertices i and j is determined by the pairwise molecular distance: $d_{ij}=d(m_i,m_j)$.

Hamiltonian circuit, a crucial concept in graph theory, is a directed cycle that visits each vertex in a graph exactly once. The Hamiltonian diversity, denoted as `HamDiv`, is defined as the length of the shortest Hamiltonian circuit in the complete graph $K_n(\mathcal{M})$.

Definition (Hamiltonian Diversity).

$$x_{ij}=\begin{cases}1, & \text{if the circuit goes from vertex } i \text{ to } j\\ 0, & \text{otherwise}\end{cases}$$

$$\texttt{HamDiv}(\mathcal{M}):=\min\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n} d_{ij}x_{ij},\ n=\lvert\mathcal{M}\rvert\geq 2$$

$$\text{s.t.}\sum_{i=1,i\neq j}^{n} x_{ij}=1,\ j=1,2,\ldots,n,$$

$$\text{and}\sum_{j=1,j\neq i}^{n} x_{ij}=1,\ i=1,2,\ldots,n,$$

$$\text{and}\sum_{i,j\in s,i\neq j}^{n} x_{ij}\leq\lvert s\rvert-1,\ \forall s\subsetneqq\{1,2,\ldots,n\},\lvert s\rvert\geq 2, \tag{9}$$

where the last constraint ensures the solution is a single circuit rather than the union of smaller circuits. Here, we adopt the widely recognized Dantzig-Fulkerson-Johnson formulation of the Traveling Salesperson Problem (TSP) [31] to establish a mathematically rigorous definition of Hamiltonian diversity.

In particular, if $\mathcal{M}=\{x,y\}$, i.e., $\lvert\mathcal{M}\rvert=2$, $\texttt{HamDiv}(\mathcal{M})=2d(x,y)$. If $\lvert\mathcal{M}\rvert=0$ or 1, $\texttt{HamDiv}(\mathcal{M})=0$.

Explanation

The Hamiltonian diversity adopts the Hamiltonian circuit in a complete graph to measure the diversity of a molecular set, which factors in each molecule equally. Moreover, this approach utilizes insights related to the exploration process within the chemical space, with the shortest Hamiltonian circuit representing the minimal cost of "traveling across" all molecules in a set. Therefore, by employing Hamiltonian diversity, we can not only effectively evaluate the molecular diversity but also obtain interpretability by quantifying each molecule's contribution to the overall diversity.

It is worth emphasizing that with a given molecular distance metric d, Hamiltonian diversity is hyperparameter-free. In contrast to `Circles`, which requires a predefined hyperparameter t that can significantly affect diversity values, `HamDiv` provides a fixed measurement that facilitates fair and uncontroversial evaluations.

Advantages

Hamiltonian diversity adheres to both the monotonicity and dissimilarity principles, and the correctness of the dissimilarity principle is evident since $\texttt{HamDiv}(\{x, y\}) = 2d(x, y)$.

The monotonicity principle is fulfilled when the molecular distance is a metric function, meaning that it satisfies the triangle inequality:

$$d(x, y) + d(y, z) \geq d(x, z)\ \forall x, y, z \in \mathcal{S}. \tag{10}$$

Proof: For a molecule $x \in \mathcal{S}$ and a molecular set $\mathcal{M} \subseteq \mathcal{S}$, we can denote the two molecules near x in the shortest Hamiltonian circuit of $\mathcal{M} \cup \{x\}$ as x_{-1} and x_{+1}. Then we have:

$$d(x_{-1}, x) + d(x, x_{+1}) \geq d(x_{-1}, x_{+1}) \tag{11}$$

A Hamiltonian circuit can be constructed for $\mathcal{M}$ with the edge between x_{-1} and x_{+1}, and other edges are all identical to those in the shortest Hamiltonian circuit of $\mathcal{M} \cup \{x\}$. This Hamiltonian circuit is not longer than the shortest Hamiltonian circuit of $\mathcal{M} \cup \{x\}$, and also not shorter than the shortest Hamiltonian circuit of $\mathcal{M}$. Hence, we have:

$$\texttt{HamDiv}(\mathcal{M} \cup \{x\}) \geq \texttt{HamDiv}(\mathcal{M}). \tag{12}$$

And Hamiltonian diversity also evidently satisfies the equality condition.

In summary, Hamiltonian diversity is, in principle, an ideal metric of molecular diversity. From this perspective, Hamiltonian diversity has advantages over all the existing diversity metrics in Table 1.

Figure 2 and Table 2 demonstrate an example of Hamiltonian diversity and compare it with `Richness` and `IntDiv`. The comparison between {A,B,C,D,E} and {A,B,C,E} shows that `IntDiv` does not satisfy monotonicity, while the comparison between {A,B,C,D} and {A,B,C,E} shows the defects of `Richness` in dissimilarity. By contrast, `HamDiv` meets both principles of diversity metrics in practice.

Table 2 The molecular distance matrix and comparisons among `Richness`, `IntDiv` and `HamDiv`

	A	B	C	D	E
A	0				
B	0.635	0			
C	0.661	0.511	0		
D	0.698	0.609	0.522	0	
E	0.783	0.692	0.691	0.596	0

Molecular set	`Richness`	`IntDiv`	`HamDiv`
{A,B,C,D,E}	5	0.640	3.046
{A,B,C,D}	4	0.606	2.365
{A,B,C,E}	4	0.662	2.619

The molecules A, B, C, D, E are corresponding to those in Fig. 2

Efficient implementation

Following previous choices of the molecular distance metric [19, 28], we calculate the Hamiltonian diversity using the Tanimoto distance between ECFPs of molecules, which satisfies the triangle inequality [32, 33].

The calculation of the Hamiltonian diversity is essentially the solution of the Traveling Salesman Problem (TSP), which is a classic NP-hard problem in combinatorial optimization. Numerous algorithms have been devised to tackle the TSP in a complete graph,

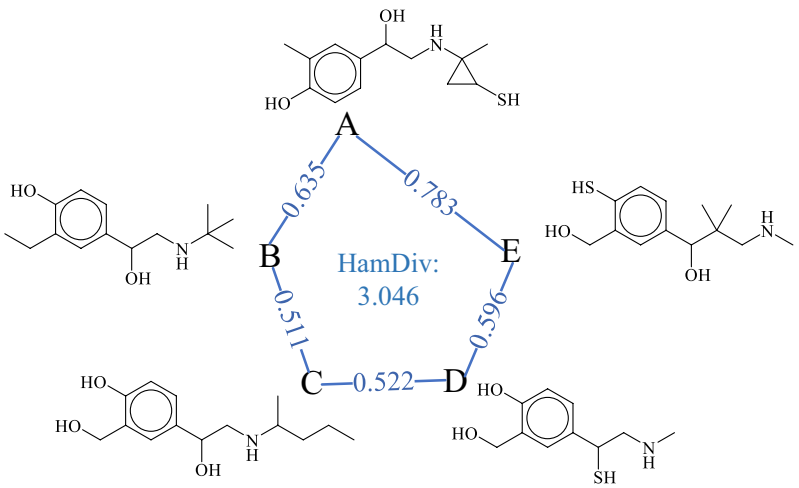

Fig. 2 An example of Hamiltonian diversity of a set of 5 molecules. The molecular distances contributing to the `HamDiv` are labeled, which make up of the shortest Hamiltonian circuit in the "graph" of the molecular set

encompassing exact, approximation, and heuristic methods [34, 35]. To strike a balance between accuracy and efficiency in diversity implementation, we run several popular algorithms for the TSP on molecular sets of different sizes, and compare the results and time costs to determine which algorithm is the best for implementing the Hamiltonian diversity.

Molecular sets

We choose the following real-world drug datasets to test different implementations of Hamiltonian diversity:

1. the 739 inhibitors against the c-Jun N-terminal kinase 3 (JNK3) target [36];
2. the 2664 inhibitors against the Glycogen synthase kinase 3 beta (GSK3β) target [36];
3. the 7218 inhibitors against the Dopamine Receptor D2 (DRD2) target [4].

And smaller sets of sizes 10, 15, 20, and 200 are randomly selected from the JNK3 actives.

Algorithms for solving the TSP

We separately implement Hamiltonian diversity using the following popular algorithms for solving the TSP whose inputs are pairwise distance matrices of the molecular set.

1. Dynamic programming: also called the Bellman-Held-Karp algorithm, giving the exact solution of the TSP with time complexity of $O(2^n n^2)$ [37];
2. Christofides algorithm: providing an approximate solution within a factor of 3/2 of the optimal solution length, with a time complexity of $O(n^3 \log n)$ [38];
3. Greedy algorithm: providing an approximate solution with a time complexity of $O(n^2 \log n)$ (usually as a fast baseline);
4. Simulated annealing algorithm: a metaheuristic algorithm giving an approximate solution [39];
5. Threshold accepting algorithm: another metaheuristic algorithm giving an approximate solution [40];
6. 2-opt Local Search algorithm: a classic heuristic algorithm for the TSP [41].

As shown in Table 3, we report the results and time costs for calculating the Hamiltonian diversity of different molecular sets using different algorithms. The results show that when the molecular set grows above 20, the time consumption of the precise algorithm becomes unacceptable. When the set size is less than 1000, 2-opt local search is the most accurate approximation algorithm and can complete the calculation in a few minutes. When the set size exceeds 1000, the performance of 2-opt local search is severely compromised within a limited running time, and the greedy algorithm is most

Table 3 The results and time costs of different implementations of Hamiltonian diversity tested on multiple molecular sets

Molecular sets		Exact	Christofides	Greedy	Simulated annealing	Threshold accepting	2-opt Local Search
$\|\mathcal{M}\| = 10$	Results	7.908	7.969 ± 0.028	8.023 ± 0.047	7.993 ± 0.028	7.996 ± 0.032	**7.930 ± 0.023**
	Errors	–	0.7% ± 0.4%	1.4% ± 0.6%	1.1% ± 0.4%	1.1% ± 0.4%	**0.3**% ± **0.3**%
	Time (s)	0.02	0.004	0.003	0.02	0.02	0.004
$\|\mathcal{M}\| = 15$	Results	12.21	12.34 ± 0.03	12.33 ± 0.03	12.33 ± 0.05	12.33 ± 0.03	**12.23 ± 0.01**
	Errors	–	1.1% ± 0.2%	1.0% ± 0.2%	1.0% ± 0.4%	1.0% ± 0.2%	**0.2**% ± **0.1**%
	Time (s)	0.9	0.006	0.004	0.02	0.02	0.008
$\|\mathcal{M}\| = 20$	Results	16.21	16.46 ± 0.08	16.43 ± 0.04	16.42 ± 0.05	16.41 ± 0.04	**16.23 ± 0.02**
	Errors	–	1.5% ± 0.5%	1.4% ± 0.2%	1.2% ± 0.3%	1.2% ± 0.2%	**0.1**% ± **0.1**%
	Time (s)	80	0.009	0.006	0.03	0.03	0.02
$\|\mathcal{M}\| = 200$	Results	–	120.0 ± 0.4	114.7 ± 0.2	114.5 ± 0.1	114.7 ± 0.2	**112.5 ± 0.3**
	Time (s)		1	0.1	0.3	0.3	7
JNK3 actives (739)	Results	–	330.0 ± 1.1	313.3 ± 0.8	313.4 ± 0.5	313.1 ± 0.4	**304.0 ± 0.5**
	Time (s)		50	2	3	3	300
GSK3β actives (2664)	Results	–	–	**1105 ± 1**	1105 ± 1	1105 ± 1	*1635 ± 6*
	Time (s)			**20**	30	30	*500*
DRD2 actives (7218)	Results	–	–	**2137 ± 1**	2138 ± 2	2138 ± 2	*5216 ± 18*
	Time (s)			**200**	300	300	*500*

For each set of molecules, the results of the approximate solution with the highest accuracy are bolded, and the results that cannot complete the calculation due to time limits are italicized

Results taking longer than 1000 s are not considered, and the operation time of the 2-opt local search algorithm on the last two molecular sets is limited to 500 s. Since the objective of the TSP is the minimum value, smaller results indicate better performance of algorithms

acceptable in this case. As a consequence, we implement the calculation of Hamiltonian diversity as follows:

- When $|\mathcal{M}| \leq 20$, using dynamic programming;
- When $20 < |\mathcal{M}| \leq 1000$, using 2-opt local search algorithm;
- When $|\mathcal{M}| > 1000$, using greedy algorithm.

Discussions

Richness and internal diversity are currently the two most commonly used metrics of molecular diversity. Compared with them, Hamiltonian diversity has advantages in principle, but there may be inaccuracies in the calculation. So, is there a better alternative in this regard?

First, the calculation of the recently proposed Circles actually also corresponds to the solution of an NP-hard problem, and it is implemented merely using simple greedy algorithms without analysis of accuracies.

In addition, one might want to replace the Hamiltonian circuit with a minimal spanning tree, which can be efficiently solved. But in fact, the molecular diversity metric constructed using a minimum spanning tree does not satisfy the monotonicity principle, because adding one molecule may result in a minimum spanning tree with a smaller total weight.

Above all, since precise molecular diversity values are not required for practical drug design, Hamiltonian diversity is a good metric with superiority in principle.

Experiments

Correlations with biological functionality

We conduct an empirical study comparing molecular diversity metrics by analyzing the correlations between these metrics and a gold standard of biological functionality following the settings presented in [28].

The analysis utilizes the BioActivity dataset [42], which consists of 10,000 compound samples with bioactivity labels sourced from the ChEMBL database [43]. These labels belong to 50 different bio-activity classes, each containing 200 samples. The number of label types covered by a subset of the BioActivity dataset represents its biological functional diversity, which should be reflected by an ideal diversity metric. Therefore, the number of bio-activity labels in a subset is recognized as a proxy gold standard (GS) of molecular diversity.

In order to assess the empirical validity of various molecular diversity metrics, we perform a random sampling of subsets from the BioActivity dataset. The subsets are of fixed sizes, specifically $n = 50, 200, 1000$. Subsequently, we calculate correlations (Spearman's correlation coefficient) between the molecular diversity metrics and the GS. A better diversity metric should have a higher correlation to the GS. The Tanimoto distance between ECFPs of molecules is employed to calculate all the distance-based diversity metrics.

As shown in Fig. 3, Hamiltonian diversity exhibits a better correlation with the GS compared to all other metrics for $n = 50$ and $n = 200$. When $n = 1000$, although slightly lower than `Circles`(0.6) and `Circles`(0.7), Hamiltonian diversity still achieves a high correlation with the GS (> 0.9). These results suggest that Hamiltonian diversity has higher consistency with real-world chemical diversity of molecules than other existing metrics.

Applying Hamiltonian diversity to molecule generation

To demonstrate the potential of using Hamiltonian diversity in real-world drug discovery for promoting molecular diversity, we incorporate it into a reinforcement learning algorithm for molecule generation to encourage diverse exploration in the chemical space.

Scenarios

To simulate practical drug discovery, we consider the following property predictors (also known as oracles): (1) `QED` (Quantitative Estimate of Drug-likeness) [44] and `SA` (Synthetic Accessibility) [45], two commonly used oracles in CADD; (2) `JNK3` (c-Jun N-terminal kinase-3) and `GSK3`β (glycogen synthase kinase-3 beta), two protein targets related to Alzheimer's disease, whose evaluators are random forest models based on Morgan fingerprints [36, 46]. We employ the oracles implemented by [47].

Moreover, we establish three multi-objective molecule generation settings (with constraints) by combinations of these oracles following [8]:

(a) `JNK3`≥ 0.5 and `QED`≥ 0.7 and `SA`≤ 2.5
(b) `GSK3`$\beta \geq 0.5$ and `QED`≥ 0.7 and `SA`≤ 2.5
(c) `JNK3`≥ 0.5 and `GSK3`$\beta \geq 0.5$ and `QED`≥ 0.7 and `SA`≤ 2.5

Under each setting, the generated compounds meeting the constraints are considered desirable drug candidates, and we aim to assess the diversity of this molecular set.

Algorithms

We use the initial version of Reinvent [4, 48] as a baseline, which is a competitive deep reinforcement learning-based approach for goal-directed molecule

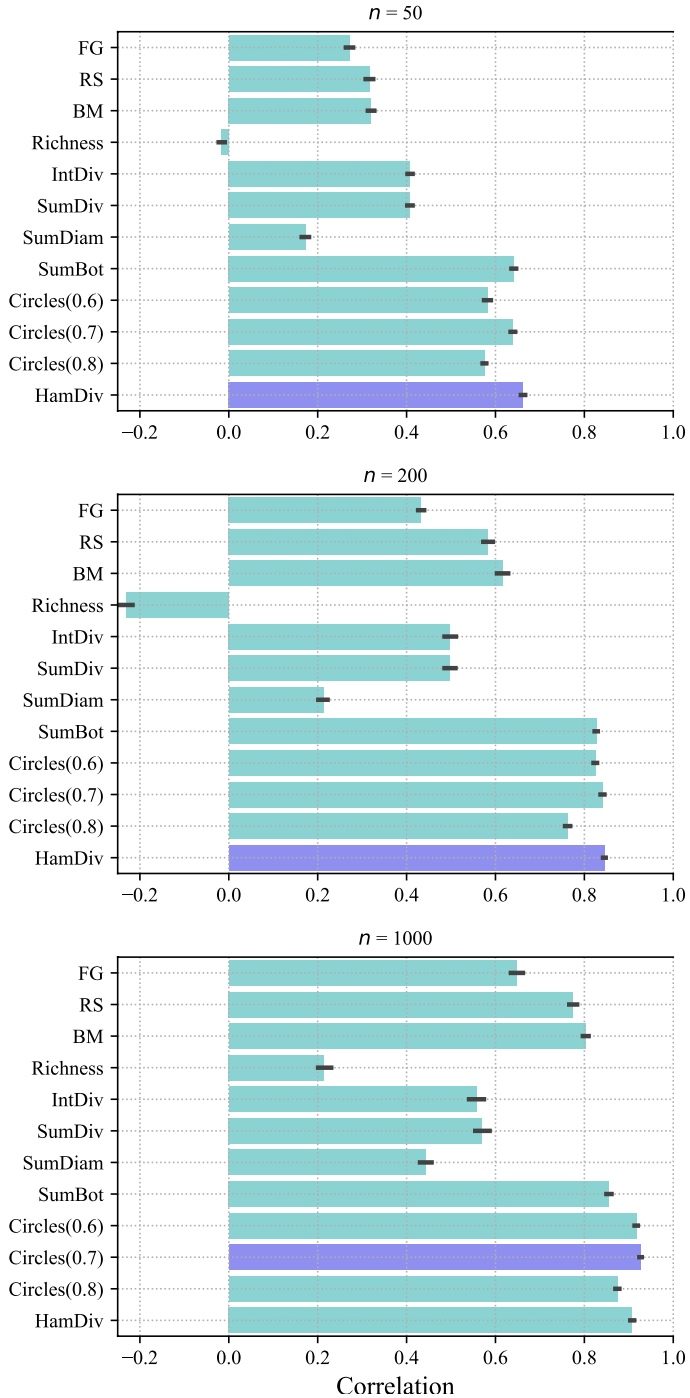

Fig. 3 Correlations between the GS and molecular diversity metrics of fixed-size random subsets. The fixed sizes are set as 50, 200, and 1000. The larger the correlation, the better the diversity metric. The average results and error bars are obtained by running experiments independently ten times

generation [49]. The algorithm uses a combination of recurrent neural networks (RNN) and reinforcement learning (RL) to iteratively optimize the generation of molecules toward the desired properties. The loss function for updating the RNN agent in each RL step is:

$$L(x) = [\log P(x)_{\text{Prior}} - \log P(x)_{\text{Agent}} + \sigma \cdot s(x)]^2 \tag{13}$$

where x denotes a molecule in a SMILES string form, $P(x)_M$ refers to the probability of the model M generating x, and $s(x) \in [0, 1]$ is the predicted property score of x and σ is a coefficient. This function encourages the agent to generate molecules with higher property scores. However, in practice, the agent is likely to converge to a certain local optimum in the chemical space, resulting in the generation of structures with low diversity.

An existing variant of Reinvent [17, 50] enhances molecular diversity by penalizing identical scaffolds in $s(x)$. The scaffolds of molecules generated in each step are stored in a "scaffold memory", and if the scaffold of a newly generated molecule is the same as that of more than k previous molecules, $s(x)$ is set to 0, where k is an integral coefficient.

Inspired by Hamiltonian diversity, we encourage the agent to explore the chemical space more diversely by adding a term on the scoring function:

$$s_{\text{Ham}}(x) = s(x) + \sigma_1 \cdot \min_{m \in \mathcal{M}} d(x, m) \tag{14}$$

where $\mathcal{M}$ is a memory of molecules generated in previous steps. The minimum distance of a newly discovered molecule from a set of previously generated molecules is corresponding to the increment in each step of the greedy algorithm for solving TSP in `HamDiv`. The calculation time of the additional term increases with the number of molecules in $\mathcal{M}$, but it is not computationally expensive in general, because the Tanimoto distances between molecules can be calculated at a speed of 10^6 pairs per second. Furthermore, only one natural way to integrate `HamDiv` into Reinvent is provided here, and we look forward to the exploration of better applications of `HamDiv` in molecular generation algorithms.

Evaluation details

For each scenario, we apply virtual screening on the ExCAPE-DB [51] database and the generated molecular set of each of the three algorithm to obtain the desirable molecules satisfying all the constraints. We use `Richness`, `BM` (number of unique Bemis-Murcko scaffolds), `IntDiv` and `HamDiv` as the metrics for molecular diversity, and report their values on those sets of desirable molecules. The Tanimoto distance between ECFPs of molecules is employed to calculate `IntDiv` and `HamDiv`.

For hyper-parameters, we set $\sigma = 1000, \sigma_1 = 0.1$, and all the algorithms run 2000 steps with a batch size of 128. All the experiments are conducted on a single NVIDIA A100 GPU and each run cost less than 12 hours.

Table 4 The diversity values of molecular sets screened from the database and generated by RL algorithms with different scoring functions

Tasks	Methods / metrics	`Richness`	`BM`	`IntDiv`	`HamDiv`
(a)	Database	103	60	0.836	49.53
	Reinvent	1404	312	0.650	440.42
	Reinvent + Scaffold Memory	4906	494	0.683	1313.51
	Reinvent + `HamDiv`	**6233**	**588**	0.680	**1659.72**
(b)	Database	339	194	0.850	161.72
	Reinvent	16818	649	0.702	4883.68
	Reinvent + Scaffold Memory	7284	435	0.755	2270.98
	Reinvent + `HamDiv`	14062	**691**	**0.759**	**5005.70**
(c)	Database	26	22	0.867	18.18
	Reinvent	258	140	0.604	72.95
	Reinvent + Scaffold Memory	922	252	0.599	230.20
	Reinvent + `HamDiv`	**2486**	**367**	0.601	**569.37**

Values reflecting the superior performance of "Reinvent + \texttt{`HamDiv`}" are bolded

Experimental results

As shown in Table 4, our algorithm designed with Hamiltonian diversity demonstrates a greater capacity for generating diverse molecules in all three scenarios compared with the two existing methods, especially with higher Hamiltonian diversity values. In addition, our algorithm also performs the best in terms of "scaffold richness" shown by `BM`.

Moreover, in the second scenario, Reinvent performs better than "Reinvent + scaffold memory", which is the opposite of the other two scenarios. This is because the second optimization objective is relatively easy, and the diversity constraint added to the baseline will have a large negative effect of "filtering out candidates". This phenomenon suggests that the "scaffold memory" penalty may be excessive for a simple task, whereas `HamDiv` does not have this problem.

In summary, the above results suggest the potential benefits of utilizing Hamiltonian diversity in the design of molecule generation algorithms.

Conclusion

In this paper, we review existing metrics for molecular diversity with a principled analysis. Then, we define and implement the Hamiltonian diversity (`HamDiv`) based on the shortest Hamiltonian circuit and demonstrate its empirical effectiveness through two experiments related to real-world drug discovery. The key advantages of this new molecular diversity metric include:

1. Hamiltonian diversity satisfies both two principles of molecular diversity metrics: monotonicity and dissimilarity. This fundamentally guarantees its higher effectiveness than all previous metrics.
2. Hamiltonian diversity can be interpreted intuitively by directly quantifying the effect of each molecule.
3. Hamiltonian diversity has high consistency with real-world chemical diversity, which reflects its high empirical value.
4. Hamiltonian diversity can assist in enhancing the diversity of molecule generation, which has a good application prospect in drug discovery.

Author contributions
X.H. wrote the main manuscript and all authors have reviewed the manuscript.

Funding
Q. Yao is supported by National Natural Science Foundation of China (under Grant No.92270106) and Beijing Natural Science Foundation (under Grant No.4242039).

Availability of data and materials
Our implementation of Hamiltonian diversity in Python is available at: https://github.com/HXYfighter/HamDiv.

Declarations

Competing interests
The authors declare no competing interests.

Received: 8 February 2024 Accepted: 11 July 2024
Published online: 07 August 2024

References

1. Du Y, Fu T, Sun J, Liu S (2022) Molgensurvey: a systematic survey in machine learning models for molecule design. arXiv preprint arXiv:2203.14500
2. Zhao L, Ciallella HL, Aleksunes LM, Zhu H (2020) Advancing computer-aided drug discovery (cadd) by big data and data-driven machine learning modeling. Drug Discov Today 25(9):1624–1638
3. Bajorath J, Chávez-Hernández AL, Duran-Frigola M, Fernández-de Gortari E, Gasteiger J, López-López E, Maggiora GM, Medina-Franco JL, Méndez-Lucio O, Mestres J et al (2022) Chemoinformatics and artificial intelligence colloquium: progress and challenges in developing bioactive compounds. J Cheminform 14(1):82
4. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. J Cheminform 9
5. Jin W, Barzilay R, Jaakkola T (2018) Junction tree variational autoencoder for molecular graph generation. In: International Conference on Machine Learning, pp. 2323–2332. PMLR
6. Jensen JH (2019) A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. Chem Sci 10(12):3567–3572
7. Ahn S, Kim J, Lee H, Shin J (2020) Guiding deep molecular optimization with genetic exploration. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (Eds.) Advances in neural information processing systems, vol. 33, pp. 12008–12021. Curran Associates, Inc
8. Xie Y, Shi C, Zhou H, Yang Y, Zhang W, Yu Y, Li L (2021) Mars: Markov molecular sampling for multi-objective drug discovery. In: International Conference on Learning Representations (ICLR)
9. Yang S, Hwang D, Lee S, Ryu S, Hwang SJ (2021) Hit and lead discovery with explorative rl and fragment-based molecule generation. Adv Neural Inf Process Syst 34:7924–7936
10. Pereira T, Abbasi M, Ribeiro B, Arrais JP (2021) Diversity oriented deep reinforcement learning for targeted molecule generation. J Cheminform 13(1):21
11. Eckmann P, Sun K, Zhao B, Feng M, Gilson MK, Yu R (2022) Limo: latent inceptionism for targeted molecule generation. In: International Conference on Machine Learning. PMLR
12. Hu X, Liu G, Zhao Y, Zhang H (2023) De novo drug design using reinforcement learning with multiple gpt agents. In: Thirty-seventh Conference on Neural Information Processing Systems
13. Yangyang C, Zixu W, Lei W, Jianmin W, Pengyong L, Dongsheng C, Xiangxiang Z, Xiucai Y, Tetsuya S (2023) Deep generative model for drug design from protein target sequence. J Cheminform 15(38)
14. Brown N, Fiscato M, Segler MH, Vaucher AC (2019) Guacamol: benchmarking models for de novo molecular design. J Chem Inf Model 59(3):1096–1108
15. Trott O, Olson AJ (2010) Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461
16. Benfenati E, Gini G, Hoffmann S, Luttik R (2010) Comparing in vivo, in vitro and in silico methods and integrated strategies for chemical assessment: problems and prospects. Altern Lab Anim 38(2):153–166
17. Blaschke T, Engkvist O, Bajorath J, Chen H (2020) Memory-assisted reinforcement learning for diverse molecular de novo design. J Chem Inf Model
18. Shi Y, Itzstein M (2019) How size matters: diversity for fragment library design. Molecules 24(15):2838
19. Benhenda M (2018) Can ai reproduce observed chemical diversity? bioRxiv, 292177
20. Polishchuk PG, Madzhidov TI, Varnek A (2013) Estimation of the size of drug-like chemical space based on gdb-17 data. J Comput Aid Mol Des 27(8):675–679
21. Tanimoto TT (1958) Elementary mathematical theory of classification and prediction. IBM technical report
22. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
23. Peter Willett, John M, Barnard Geoffrey, Downs M (1998) Chemical similarity searching. J Chem Inf Model 38(6):983–996
24. Bajusz D, Rácz A, Héberger K (2015) Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7(1):1–13
25. Preuer K, Renz P, Unterthiner T, Hochreiter S, Klambauer G (2018) Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. J Chem Inf Model 58(9):1736–1741
26. Samanta S, O'Hagan S, Swainston N, Roberts TJ, Kell DB (2020) Vae-sim: a novel molecular similarity measure based on a variational autoencoder. Molecules 25(15):3446
27. Fukutani T, Miyazawa K, Iwata S, Satoh H (2021) G-rmsd: Root mean square deviation based method for three-dimensional molecular similarity determination. Bull Chem Soc Jpn 94(2):655–665
28. Xie Y, Xu Z, Ma J, Mei Q (2023) How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules. In: International Conference on Learning Representations (ICLR)
29. Fitzner K (2007) Reliability and validity a quick review. Diabetes Educ 33(5):775–780
30. Zhang J, Mercado R, Engkvist O, Chen H (2021) Comparative study of deep generative models on chemical space coverage. J Chem Inf Model 61(6):2572–2581
31. Dantzig G, Fulkerson R, Johnson S (1954) Solution of a large-scale traveling-salesman problem. J Oper Res Soc Am 2(4):393–410
32. Lipkus AH (1999) A proof of the triangle inequality for the tanimoto distance. J Math Chem 26(1–3):263–265
33. Kosub S (2019) A note on the triangle inequality for the jaccard distance. Pattern Recogn Lett 120:36–38
34. Bellmore M, Nemhauser GL (1968) The traveling salesman problem: a survey. Oper Res 16(3):538–558
35. Nemani R, Cherukuri N, Rao GRK, Srinivas P, Pujari JJ, Prasad C (2021) Algorithms and optimization techniques for solving tsp. In: 2021 Fifth

international conference on I-SMAC (IoT in social, mobile, analytics and Cloud) (I-SMAC), pp. 809–814. IEEE
36. Li Y, Zhang L, Liu Z (2018) Multi-objective de novo drug design with conditional graph generative model. J Cheminform 10(1):1–24
37. Bellman R (1962) Dynamic programming treatment of the travelling salesman problem. J ACM (JACM) 9(1):61–63
38. Christofides N (1976) Worst-case analysis of a new heuristic for the travelling salesman problem
39. Skiscim CC, Golden BL (1983) Optimization by simulated annealing: a preliminary computational study for the tsp. Technical report, Institute of Electrical and Electronics Engineers (IEEE)
40. Dueck G, Scheuer T (1990) Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. J Comput Phys 90(1):161–175
41. Croes GA (1958) A method for solving traveling-salesman problems. Oper Res 6(6):791–812
42. Koutsoukas A, Paricharak S, Galloway WR, Spring DR, IJzerman AP, Glen RC, Marcus D, Bender A (2014) How diverse are diversity assessment methods? a comparative analysis and benchmarking of molecular descriptor space. J Chem Inf Model 54(1):230–242
43. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M et al (2019) Chembl: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):930–940
44. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4(2):90–98
45. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 1(1):1–11
46. Jin W, Barzilay R, Jaakkola T (2020) Multi-objective molecule generation using interpretable substructures. In: International Conference on Machine Learning (ICML), pp. 4849–4859. PMLR
47. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, Coley CW, Xiao C, Sun J, Zitnik M (2021) Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. Proceedings of neural information processing systems, neurips datasets and benchmarks
48. Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, Engkvist O (2024) Reinvent 4: modern ai-driven generative molecule design. J Cheminform 16(1):20
49. Gao W, Fu T, Sun J, Coley C (2022) Sample efficiency matters: a benchmark for practical molecular optimization. Adv Neural Inform Process Syst 35:21342–21357
50. Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, Papadopoulos K, Patronov A (2020) Reinvent 2.0: an AI tool for de novo drug design. J Chem Inf Model
51. Sun J, Jeliazkova N, Chupakhin V, Golib-Dzib J-F, Engkvist O, Carlsson L, Wegner J, Ceulemans H, Georgiev I, Jeliazkov V et al (2017) Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. J Cheminform 9(1):1–9

Publisher's Note