JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Letter

Diverse Hits in De Novo Molecule Design: Diversity-Based Comparison of Goal-Directed Generators

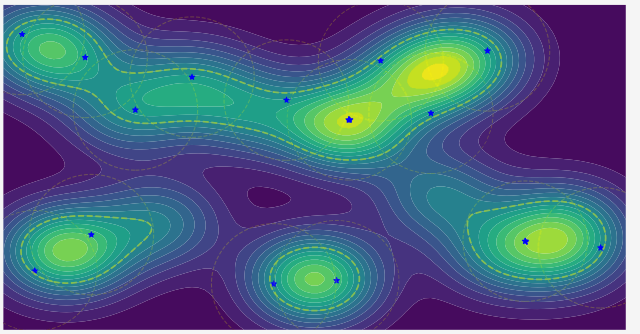
Philipp Renz, Sohvi Luukkonen, and Günter Klambauer*

Cite This: https://doi.org/10.1021/acs.jcim.4c00519

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Since the rise of generative AI models, many goal-directed molecule generators have been proposed as tools for discovering novel drug candidates. However, molecule generators often produce highly similar molecules and tend to overemphasize conformity to an imperfect scoring function rather than capturing the true underlying properties sought. We rectify these two shortcomings by offering diversity-based evaluations using the #Circles metric and considering constraints on scoring function calls or computation time. Our findings highlight the superior performance of SMILES-based autoregressive models in generating diverse sets of desired molecules compared to graph-based models or genetic algorithms.



INTRODUCTION

Goal-directed *de novo* drug design (DNDD) aims to generate small molecules possessing specific properties like efficacy, low toxicity, and drug-likeness,[1] by exploring the vast space of drug-like molecules.[2] This process involves generating novel chemical structures, guided by on-the-fly feedback from a scoring function to incorporate desired properties efficiently. With the surge of generative artificial intelligence, the field has witnessed a surge in interest, leading to the development of numerous new molecule generators, particularly those based on deep learning.[3−6]

Generating diverse sets of high-scoring molecules is essential in drug discovery. While most methods focus on producing individual high-scoring molecules, the reliance on quantitative structure−property relationship (QSPR) models introduces uncertainties and biases due to limited training data.[7] These errors propagate to molecule generators, emphasizing the need for diverse molecule sets to enhance the chances of identifying successful drug candidates.[8−12] Furthermore, diversity in molecule generation can lead to the exploration of novel chemical spaces beyond patented compounds.[13] However, many existing generators suffer from "mode collapse," producing only a limited range of similar molecules.[12,14−17] Various approaches have been proposed to address this issue and improve diversity in generated molecules.[16,18−23]

Previous comparative studies of molecule generators have often used insufficient diversity metrics. Well-known DNDD benchmarking platforms and leaderboards, such as GuacaMol[17] and MOSES,[14] include some classic diversity metrics: uniqueness and/or internal diversity, in the case of nongoal-directed molecule generation. But to our knowledge, there has been no systematic benchmark study of the capacity of different goal-directed molecule generators to generate a diverse set of high-scoring molecules.

Moreover, traditional metrics exhibit significant limitations in accurately characterizing the chemical space represented by a set of molecules.[24,25] For example, in Figure 1 we show how the arguably most commonly used diversity metric, *internal diversity*, fails to capture coverage of chemical space. More simple metrics, like the fraction of unique molecules and unique Bemis-Murcko scaffolds[26] are also inadequate as they can be optimized by generating many highly similar molecules, differing only in minor features. Recently diversity metrics based on sphere exclusion,[27] such as *SEDiv*[28] and *#Circles*,[25] have been proposed to quantify chemical space coverage. These metrics have been shown to align well with chemical intuition regarding the chemical diversity of known libraries and correlate well with the coverage of biological functionalities.

In addition to the aforementioned mode collapse problem, molecule generators tend to overoptimize molecules to their accessed scoring functions rather than to the actual desired properties.[7] A high number of scoring function calls can lead to this overfitting to the biased QSPR models and a decline in molecule quality over the optimization cycles. Therefore, it is essential to evaluate and compare generators within a

Received: March 28, 2024
Revised: July 10, 2024
Accepted: July 11, 2024

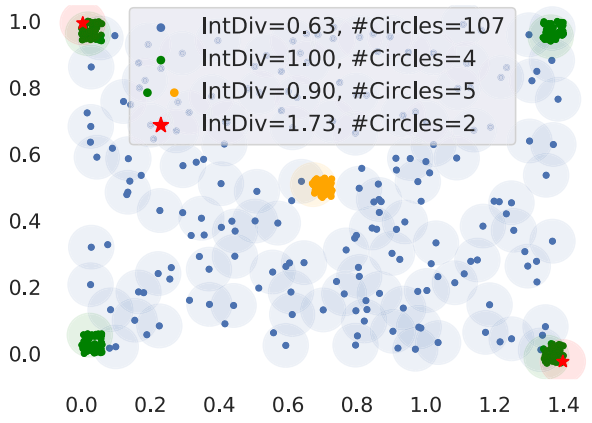


Figure 1. Comparison of internal diversity and #Circles. Internal diversity fails to capture high coverage of chemical space (blue), can be large for a few clusters of very similar molecules (green), and can decrease when adding additional molecules (green/yellow). IntDiv is maximized by two molecules with maximal distance (red). #Circles accurately captures the coverage of all sets.

standardized computational budget, by restricting a) the overall compute time or b) the number of scoring function evaluations. Additionally, there's growing interest in replacing machine learning-based scoring functions with more accurate but computationally expensive physics-based methods like docking.[28−31] Generative methods that can efficiently learn from a few scoring function evaluations are preferable for such costly scoring functions. Gao et al.[32] tested the sample efficiency of a range of generative methods given a constraint on the number of scoring function calls. Still, no studies focus on the generated molecules' diversity under computational constraints.

In this work, we address the two shortcomings of previous comparisons, a) the insufficient diversity metrics and b) the generation without limitations on the computational budget. We systematically benchmark the performance of established molecule generators at generating diverse high-scoring molecules, referred to as *diverse hits*. We evaluate these generators within the framework of goal-directed optimization, where they operate under constraints such as a limited number of scoring function calls or time, emphasizing computational cost. We utilize the *#Circles* diversity metric as a key performance indicator, providing a comprehensive assessment of generative model efficiency in practical scenarios.

BENCHMARK SETUP

Diverse Hits. We evaluate the performance of the tested generators based on the diversity of the generated high-scoring molecules. We define high-scoring molecules as ones with a score above a threshold S and refer to them as *hits*. We use the #Circles[25] metric to measure diversity of the found hits. This metric counts the number of generated hits that are pairwise distinct by a distance threshold D. We refer to this as the number of *diverse hits*.

More specifically, given the set of generated hits, $\mathcal{G}$, the number of diverse hits is given by

$$\mu(\mathcal{G}; D) = \max_{C \in \mathcal{P}(\mathcal{G})} |C| \text{ s.t. } \forall\, x \neq y \in C\colon d(x, y) \geq D$$

where $\mathcal{P}$ denotes the power set and $d(x, y)$ is the distance between molecules x and y. This metric ensures that each found hit that is sufficiently different from those already found adds to the performance, and that hits similar to each other are not double-counted.

Figure 1 illustrates the computation of this metric as finding the largest set of circles centered on the molecules, such that no center lies within another circle. We provide a more detailed description of this metric in Supporting Information Section S1.

Scoring Functions. *Bioactivity Prediction Models.* We evaluate the methods on three well-established molecule binary bioactivity label optimization tasks: JNK3,[33] GSK3β,[33] and DRD2.[16] For each target we train a Random Forest classifier[34] as a basis for our scoring functions. Table S1 provides details on the data sets and the performance of the predictive models. All scoring functions exhibit robust predictive performance, as indicated by their ROCAUC and Average Precision (AP) values.

During optimization, we use the classifier's probabilistic activity output, $p_{\mathrm{RF}}(s)$, as a scoring function. When predicting if a molecule is a hit, we adopt a score threshold of $S = 0.5$. Further details on the QSAR models are given in Supporting Information Section S2.1.

Property Filters. Generative models often generate molecules with high molecular weights (MW) or water-octanol partition coefficients (logP) and may contain idiosyncratic substructures, rendering them impractical for drug discovery projects,[7,35] and often these molecules would be discarded in real-world applications. We address this issue by incorporating lenient property constraints into the scoring functions[35] by defining acceptable ranges for MW ([157,761]Da) and logP ([-2.0,8.3]) values, and the fraction of idiosyncratic substructures ([0.00,0.08]). Further details on these filters are given in Supporting Information Section S2.2. During scoring, molecules violating any of these ranges have their score set to zero.

Diversity Filter. Most goal-directed molecule generators are not suitable for diverse generation out of the box, as they tend to get stuck in local optima of the scoring function.[12,14−17] To enable diverse molecule generation, we enhance the scoring functions with the *diversity filter* (DF) from Blaschke et al.[16] It assigns zero scores to molecules that are within a distance threshold $D_{\mathrm{DF}} = 0.7$ to previously found hits. This approach prevents the optimization process from getting trapped in local optima and promotes the exploration of new chemical space regions. The DF proved to be crucial for performance in preliminary experiments and its use allows for the meaningful inclusion of generative algorithms originally designed for single-molecule optimization. A detailed description of the DF is given in Supporting Information Section S2.3.

The final scoring function is given by the product of the bioactivity model prediction, and the binary property and diversity filters.

Compute Constraints. We evaluate the performance of the generators to create diverse hits under two computational constraint settings: (a) **Sample limit**, we limit the number of scoring function evaluations to 10K as proposed by Gao et al.,[32] and (b) **Time limit**, we limit the time available to the algorithms to 600 s. All algorithms are executed using 8 cores of an AMD Ryzen Threadripper 1920X and a single NVidia RTX 2080 GPU.

Generative Models. We utilize our benchmark setup to assess the following 12 methods. The methods were chosen based on their performance in previous benchmarks[17,32] and to

https://doi.org/10.1021/acs.jcim.4c00519
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

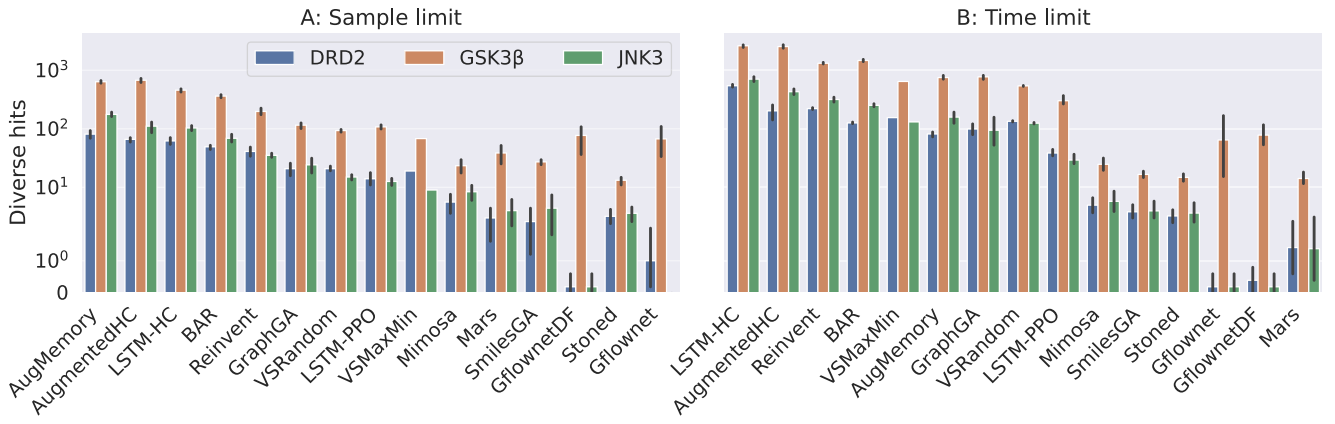


Figure 2. Number of diverse hits found by the tested methods (ordered by average rank) for the three studied optimization tasks. **A.** Results under a constraint of 10K scoring function evaluations. **B.** Results under a time constraint of 600 s. Error bars show the range of the results.

ensure that a range of methodically different approaches is included.

We test six LSTM-based autoregressive models operating on SMILES: **LSTM-HC**[36] optimizing with a hill-climb algorithm, **LSTM-PPO**[37] optimizing with the PPO algorithm, **Reinvent**[38] optimizing with the REINFORCE algorithm,[39] and three extensions of Reinvent: **AugmentedHC** (mixture of Reinvent and hill-climb),[40] **AugMemory**,[41] and BestAgentReminder (**BAR**).[42] We also test three genetic algorithms making use of mutations of different molecular representations: **GraphGA**[43] operating on molecular graphs, **SmilesGA**[44] operating on SMILES, and **Stoned**[45] operating on SELFIES.[46] We further test three models that generate molecules via sequential graph edits: **Mars**,[47] **Mimosa**,[48] and **GFlowNet/GFlowNetDF**,[21] which is tested with and without the DF as it supports diverse generation by default.

We compare these methods against two virtual screening (VS) baselines using the GuacaMol data set[17] as a screening library. VS methods are deemed inefficient as they ignore feedback from already scored molecules but serve as a valuable baseline. The **VS Random** baseline evaluates the molecules with the scoring function in random order from the library. In contrast, the **VS MaxMin** baseline first sorts the molecules in the library with the MaxMin algorithm.[49] This promotes diversity by ensuring that molecules screened first have as large pairwise distances as possible and prevents evaluating redundant molecules. Further details about these methods and the choice to exclude others are discussed in Supporting Information Section S3.

Optimization. We conducted a hyperparameter search to optimize each combination of generative algorithm, scoring function, and computational constraint. Employing a random search with 15 trials per combination, we explored various hyperparameter ranges, and to assess result stability, we executed five independent runs with distinct random seeds. The selected hyperparameters are detailed in Supporting Information Table S2.

Throughout the optimizations, we tracked all generated molecules, their corresponding scores, and the generation time. This comprehensive recording prevents valuable molecules from being discarded unnecessarily. This is particularly crucial when using a diversity filter that steers the search away from already discovered solutions.

RESULTS AND DISCUSSION

We benchmarked the capacity of a wide range of goal-directed molecule generators to design diverse hits under two computational constraint settings for three protein targets. The main results, i.e., the number of diverse hits under the constraints are shown in Figure 2 and discussed here. Extended results with complementary metrics are given in Supporting Information Section S5.

Large Differences in Performance between Models and Tasks. Above all, we observe in Figure 2 a significant difference in the capacity to produce diverse hits between the different algorithms, tasks, and computational constraints. The number of diverse hits ranges from several molecules for Mars (worst) to several thousand molecules for LSTM-HC (best) in the time-constrained setting. We also see that the performance is highly task-dependent: all approaches find ∼10 × more diverse hits for GSK3β than DRD2/JNK3 with the most extreme difference between the GFlowNet generators. Finally, the absolute, as well as relative performance of the algorithms depends strongly on the used computational constraint and on the available compute budget as shown in Supporting Information Sections S5.2 and S5.3.

SMILES-Based LSTM Models Perform Best in Generating Diverse Hits. Generally, the top ranks are dominated by autoregressive SMILES-based models. In the sample limit setting AugMemory performs best, making use of experience replay with selective purge and data augmentation, which leads to high sample efficiency. This allows the model to outperform its parent Reinvent. Similarly, the AugmentedHC model can also outperform its parent models Reinvent and LSTM-HC as shown in the original paper.[40] LSTM-HC attains the third rank and can outperform Reinvent, which is in contrast to results in single molecule optimization tasks.[32] The increase in performance of the extensions compared to their parent methods, comes with a significant computational cost as under the time limit the parent models are more competitive with their extensions, with LSTM-HC and Reinvent taking the top and third rank, respectively. We found that LSTM-HC is the most versatile algorithm achieving on average 84% of the top performance in each of the six settings (see Supporting Information Section S5.4). We think that this model class benefits from their stochastic generation policy which allows them to sample from different regions of chemical space at each optimization step. Thus, they can easily move to new

https://doi.org/10.1021/acs.jcim.4c00519
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

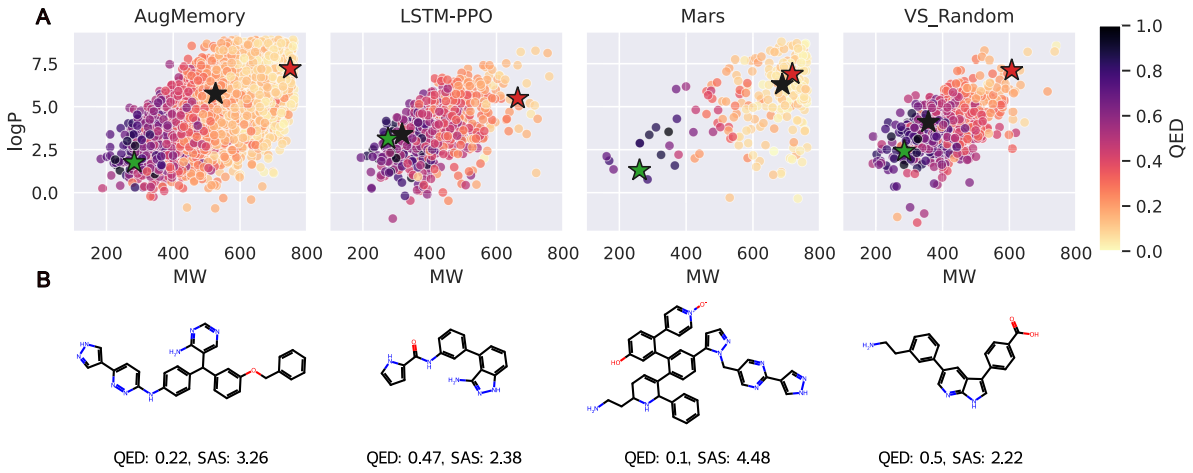


Figure 3. A. Chemical space cover by generated diverse hits by best (AugMemory), median (LSTM-PPO), and worst (Mars) methods, and the VS Random baseline, in the sample-constrained setting (all tasks combined). The green, black, and red stars represent the molecules with the highest, median, and lowest QED, respectively. **B.** Molecules with median QED.

regions once a new diverse hit has been found, resulting in high sample/compute efficiency.

Limited Number of Diverse Hits with Graph-Based and Genetic Algorithms. The graph-based models generally occupy the lower ranks in this comparison. Among them, GraphGA is the only model that outperforms the virtual screening baselines. SmilesGA and Stoned both perform poorly in this comparison. Along with GraphGA, they are not able to match their competitive performance in the single molecule optimization tasks.[32] We also found Mars and GFlowNet to perform poorly in this comparison, despite comparing well in previous diverse optimization studies.[21,25] This discrepancy highlights the importance of a meaningful benchmark setup and comparison to models suited for diverse optimization. We hypothesize that genetic algorithms encounter challenges in diverse generation tasks as they traverse chemical space using incremental modifications to existing molecules. This might cause a slow transition to new high-scoring regions once a hit is found.

Diversity at the Cost of Drug-likeness. We observed that some models achieve high diversity at the cost of drug-likeness. In Figure 3 we illustrate the chemical space covered and drug-likeness of the generated diverse hits by the best (AugMemory), median (LSTM-PPO), and worst (Mars) method, and the VS Random baseline, in the sample-constrained setting. We see that even though LSTM-PPO generates 10 times less diverse hits than AugMemory, these hits are generally more drug-like based on the QED score and overlap better with the VS baseline. In Supporting Information Section S5.5, we present distributions for additional properties of the generated diverse hits for all the methods. These distributions confirm that increased diversity is often achieved by generating larger, less drug-like molecules. We note that also the models achieving the lowest number of diverse hits (Mars and GFlowNet) struggle to generate drug-like molecules.

CONCLUSION

In our study, we rigorously tested a range of molecule generators in diverse *de novo* design tasks, introducing a benchmark setup that overcomes limitations identified in previous studies. Our findings underscore the crucial importance of considering computational resources in the generation of molecules and employing a meaningful diversity measure in the context of DNDD.

We found that SMILES-based autoregressive models perform well, compared to graph-based models and genetic algorithms in generating diverse sets of high-scoring molecules, and that single molecule optimization performance does not necessarily translate to diverse optimization settings. Performance values range over several orders of magnitude for different models and tasks and compute constraints. The latter highlights the importance of considering the specific application and available resources when choosing a generative model for practical applications.

Due to the broad range of possible applications of generative models in drug discovery, this study cannot cover all relevant aspects in detail, such as the used scoring functions, synthesizability issues, or the amount of available compute budgets, which may drastically differ in real-world applications. Nevertheless, we believe that our findings will generalize to other settings, and that our benchmark setup will be useful for future studies in the field.

ASSOCIATED CONTENT

Data Availability Statement

The data, code, and instructions necessary to reproduce the results of this study are available for download at https://github.com/ml-jku/diverse-hits.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00519.

S1 - Diversity metric details, S2 - Scoring function details, S3 - Generative model details, S4 - Hyperparameter optimization, S5 - Extended results (PDF)

Additional drawings of generated structures (ZIP)

AUTHOR INFORMATION

Corresponding Author

Günter Klambauer − *Johannes Kepler University Linz, ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning, Linz, AT 4040, Austria;* orcid.org/0000-0003-2861-5552; Email: klambauer@ml.jku.at

https://doi.org/10.1021/acs.jcim.4c00519
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Authors

Philipp Renz − *Johannes Kepler University Linz, Linz, AT 4040, Austria;* orcid.org/0000-0002-3323-7632

Sohvi Luukkonen − *Johannes Kepler University Linz, ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning, Linz, AT 4040, Austria;* orcid.org/0000-0001-9387-1427

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00519

Author Contributions

P.R. designed the study, incorporating input from G.K. and S.L. P.R. executed the experiments and analyzed the results. P.R., G.K., and S.L. contributed to writing the manuscript, with P.R. taking the lead. All authors reviewed and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The ELLIS Unit Linz, the LIT AI Lab, and the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), AI4GreenHeatingGrids (FFG-899943), INTEGRATE (FFG-892418), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank NXAI GmbH, Audi. JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Laboratories (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo), Software Competence Center Hagenberg GmbH, Borealis AG, TÜV Austria, Frauscher Sensonic, TRUMPF and the NVIDIA Corporation.

■ REFERENCES

(1) Schneider, G. *De Novo Molecular Design*; Wiley-VCH Verlag GmbH & Co., 2013.

(2) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*, 1116−1124.

(3) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360−365.

(4) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design—a Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828−849.

(5) Luukkonen, S.; van den Maagdenberg, H. W.; Emmerich, M. T. M.; van Westen, G. J. P. Artificial intelligence in multi-objective drug design. *Curr. Opin. Struct. Biol.* **2023**, *79*, No. 102537.

(6) Fromer, J. C.; Coley, C. W. Computer-aided multi-objective optimization in small molecule discovery. *Patterns* **2023**, *4*, No. 100678.

(7) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On Failure Modes in Molecule Generation and Optimization. *Drug Discovery Today: Technol.* **2019**, *32*−*33*, 55−63.

(8) Martin, Y. C. Diverse Viewpoints on Computational Aspects of Molecular Diversity. *J. Comb. Chem.* **2001**, *3*, 231−250.

(9) Seneci, P. Trends in Drug Research III. In *Pharmacochemistry Library*; van der Goot, H., Ed.; 2002; Vol. *32*, pp 147−160.

(10) Angeli, P.; Gaviraghi, G. *Pharmacochemistry Library* **2002**, *32*, 95−96.

(11) Gorse, A.-D. Diversity in Medicinal Chemistry Space. *Curr. Top. Med. Chem.* **2006**, *6*, 3−18.

(12) Benhenda, M. ChemGAN Challenge for Drug Discovery: Can AI Reproduce Natural Chemical Diversity? *arXiv*, **2017**.

(13) Shimizu, Y.; Ohta, M.; Ishida, S.; Terayama, K.; Osawa, M.; Honma, T.; Ikeda, K. AI-driven molecular generation of not-patented pharmaceutical compounds using world open patent data. *J. Cheminformatics* **2023**, *15*, 120.

(14) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644.

(15) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 1736−1741.

(16) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-Assisted Reinforcement Learning for Diverse Molecular de Novo Design. *J. Cheminformatics* **2020**, *12*, 68.

(17) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096−1108.

(18) Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D. N. Strategy To Discover Diverse Optimal Molecules in the Small Molecule Universe. *J. Chem. Inf. Model.* **2015**, *55*, 529−537.

(19) Liu, X.; Ye, K.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. An Exploration Strategy Improves the Diversity of de Novo Ligands Using Deep Reinforcement Learning: A Case for the Adenosine A2A Receptor. *J. Cheminformatics* **2019**, *11*, 35.

(20) Chen, B.; Wang, T.; Li, C.; Dai, H.; Song, L. Molecule Optimization by Explainable Evolution. *ICLR*, 2020.

(21) Bengio, E.; Jain, M.; Korablyov, M.; Precup, D.; Bengio, Y. Flow Network Based Generative Models for Non-Iterative Diverse Candidate Generation. *NeurIPS*, **2021**.

(22) Pereira, T.; Abbasi, M.; Ribeiro, B.; Arrais, J. P. Diversity Oriented Deep Reinforcement Learning for Targeted Molecule Generation. *J. Cheminformatics* **2021**, *13*, 21.

(23) Bjerrum, E. J.; Margreitter, C.; Blaschke, T.; Kolarova, S.; de Castro, R. L.-R. Faster and More Diverse de Novo Molecular Optimization with Double-Loop Reinforcement Learning Using Augmented SMILES. *J. Comput. Aided. Mol. Des.* **2023**, *37*, 373−394.

(24) Waldman, M.; Li, H.; Hassan, M. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graph. Model.* **2000**, *18*, 412−426.

(25) Xie, Y.; Xu, Z.; Ma, J.; Mei, Q. How Much Space Has Been Explored? Measuring the Chemical Space Covered by Databases and Machine-Generated Molecules. *ICLR*, **2023**.

(26) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(27) Gillet, V. J.; Willett, P. Dissimilarity-Based Compound Selection for Library Design. In *Combinatorial Library Design and Evaluation: Principles, Software, Tools, and Applications in Drug Discovery*; Ghose, A., Viswanadhan, V., Eds.; CRC Press: Boca Raton, FL, 2001; Chapter 13.

(28) Thomas, M.; Smith, R. T.; O'Boyle, N. M.; de Graaf, C.; Bender, A. Comparison of Structure- and Ligand-Based Scoring Functions for Deep Generative Models: A GPCR Case Study. *J. Cheminformatics* **2021**, *13*, 39.

(29) Guo, J.; Janet, J. P.; Bauer, M. R.; Nittinger, E.; Giblin, K. A.; Papadopoulos, K.; Voronov, A.; Patronov, A.; Engkvist, O.; Margreitter, C. DockStream: a docking wrapper to enhance de novo molecular design. *J. Cheminformatics* **2021**, *13*, 89.

(30) Goel, M.; Raghunathan, S.; Laghuvarapu, S.; Priyakumar, U. D. MoleGuLAR: Molecule Generation Using Reinforcement Learning with Alternating Rewards. *J. Chem. Inf. Model.* **2021**, *61*, 5815−5826.

(31) Elend, L.; Jacobsen, L.; Cofala, T.; Prellberg, J.; Teusch, T.; Kramer, O.; Solov'yov, I. A. Design of SARS-CoV-2 Main Protease Inhibitors Using Artificial Intelligence and Molecular Dynamic Simulations. *Molecules* **2022**, *27*, 4020.

(32) Gao, W.; Fu, T.; Sun, J.; Coley, C. W. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization. *NeurIPS* **2022**, 1.
(33) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective de Novo Drug Design with Conditional Graph Generative Model. *J. Cheminformatics* **2018**, *10*, 33.
(34) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(35) Thomas, M.; O'Boyle, N. M.; Bender, A.; De Graaf, C. Re-Evaluating Sample Efficiency in de Novo Molecule Generation. *arXiv* **2022**, 2212.01385.
(36) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.
(37) Neil, D.; Segler, M.; Guasch, L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design. 2018; https://openreview.net/forum?id=HkcTe-bR-.
(38) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminformatics* **2017**, *9*, 48.
(39) Williams, R. J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach Learn* **1992**, *8*, 229−256.
(40) Thomas, M.; O'Boyle, N. M.; Bender, A.; de Graaf, C. Augmented Hill-Climb Increases Reinforcement Learning Efficiency for Language-Based de Novo Molecule Generation. *J. Cheminformatics* **2022**, *14*, 68.
(41) Guo, J.; Schwaller, P. Augmented Memory: Capitalizing on Experience Replay to Accelerate De Novo Molecular Design. *arXiv* **2023**, 2305.16160.
(42) Atance, S. R.; Diez, J. V.; Engkvist, O.; Olsson, S.; Mercado, R. De Novo Drug Design Using Reinforcement Learning with Graph-Based Deep Generative Models. *J. Chem. Inf. Model.* **2022**, *62*, 4863−4872.
(43) Jensen, J. H. A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chem. Sci.* **2019**, *10*, 3567−3572.
(44) Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-Based de Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **2018**, *47*, 1431.
(45) Nigam, A.; Pollice, R.; Krenn, M.; Gomes, G. d. P.; Aspuru-Guzik, A. Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules Using SELFIES. *Chem. Sci.* **2021**, *12*, 7079−7090.
(46) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
(47) Xie, Y.; Shi, C.; Zhou, H.; Yang, Y.; Zhang, W.; Yu, Y.; Li, L. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. *ICLR*, **2021**.
(48) Fu, T.; Xiao, C.; Li, X.; Glass, L. M.; Sun, J. MIMOSA: Multi-constraint Molecule Sampling for Molecule Optimization. *AAAI*, **2021**.
(49) Sayle, R. *2D Similarity, Diversity and Clustering in RDKit*. 2019; https://www.nextmovesoftware.com/talks/Sayle_2DSimilarityDiversityAndClusteringInRdkit_RDKITUGM_201909.pdf.

https://doi.org/10.1021/acs.jcim.4c00519
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX