D3Rings: A fast and accurate method for ring system identification and deep generation of drug-like cyclic compounds Postprint

Authors: Minfei Ma, Xinben Zhang, Liping Zhou, Zijian Han, Yulong Shi, Jintian Li, Leyun Wu, Zhijian Xu, Weiliang Zhu, Zhijian Xu, Weiliang Zhu

Date: 2024-02-06T19:33:54+00:00

Abstract

Continuous exploration of the chemical space of molecules to find ligands with high affinity and specificity for specific targets is an important topic in drug discovery. A focus on cyclic compounds, particularly natural compounds with diverse scaffolds, provides important insights into novel molecular structures for drug design. However, the complexity of their ring structures has hindered the applicability of widely accepted methods and software for the systematic identification and classification of cyclic compounds. Herein, we successfully developed a new method, D3Rings, to identify acyclic, monocyclic, spiro ring, fused and bridged ring, and cage ring compounds as well as macrocyclic compounds. By using D3Rings, we completed the statistics of cyclic compounds in 3 different databases, e.g., ChEMBL, DrugBank, and COCONUT. The results demonstrated the richness of ring structures in natural products, especially spiro, macrocycles, fused and bridged rings. Based on this, three deep generative models, namely VAE, AAE, and CharRNN, were trained and used to construct two datasets similar to DrugBank and COCONUT but 10 times larger than them. The enlarged datasets were then used to explore the molecular chemical space, focusing on complex ring structures, for novel drug discovery and development. Docking experiments with the newly generated COCONUT-like dataset against three SARS-CoV-2 target proteins revealed that an expanded compound database improves molecular docking results. Cyclic structures were exhibited the best docking scores among the top-ranked docking molecules. These results suggest the importance of exploring the chemical space of structurally novel cyclic compounds and continuous expansion of the library of drug-like compounds to facilitate the discovery of potent ligands with high binding affinity to specific targets. D3Rings is now freely available at http://www.d3pharma.com/D3Rings/.

Full Text

Preamble

D3Rings: A Fast and Accurate Method for Ring System Identification and Deep Generation of Drug-Like Cyclic Compounds

Minfei Ma[1,2], Xinben Zhang[1], Liping Zhou[1,2], Zijian Han[1,2], Yulong Shi[1,2], Jintian Li[1,2], Leyun Wu[1,2], Zhijian Xu[1,2], *Weiliang Zhu*[1,2]

[1]State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

[2]School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China

*To whom correspondence should be addressed.

Phone: +86-21-50805020 (W.Z.), +86-21-50807088 (W.Z.)

Email: zjxu@simm.ac.cn (Z.X.), wlzhu@simm.ac.cn (W.Z.)

ABSTRACT

Continuous exploration of chemical space to identify ligands with high affinity and specificity for specific targets remains a central challenge in drug discovery. Cyclic compounds—particularly natural products with diverse scaffolds—offer valuable insights for novel molecular design, yet the complexity of their ring structures has limited the applicability of conventional identification and classification methods. Herein, we present D3Rings, a novel method that rapidly and accurately identifies and classifies molecules as acyclic, monocyclic, spiro, fused and bridged, cage, or macrocyclic compounds. Applying D3Rings to ChEMBL, DrugBank, and COCONUT databases revealed the rich structural diversity of natural products, especially in spiro, macrocyclic, fused, and bridged ring systems. Building on this analysis, we trained three deep generative models (VAE, AAE, and CharRNN) to construct datasets tenfold larger than DrugBank and COCONUT, specifically enriched in drug-like and natural product-like cyclic compounds. Docking experiments with the expanded COCONUT-like dataset against three SARS-CoV-2 target proteins demonstrated that larger compound libraries improve molecular docking outcomes, with cyclic structures consistently achieving the best docking scores among top-ranked molecules. These results highlight the importance of exploring chemical space for structurally novel cyclic compounds and underscore the value of continuously expanding drug-like compound libraries to facilitate the discovery of potent target-specific ligands. D3Rings is freely available at http://www.d3pharma.com/D3Rings/.

INTRODUCTION

Cyclic compounds constitute a dominant class in modern pharmaceuticals. Among the top 200 drugs by retail sales in 2021 (compiled by M. Haziq Qureshi, University of Arizona), 113 are small-molecule drugs, of which 112 contain cyclic structures. Specifically, 76, 8, and 2 of these cyclic drugs contain fused/bridged ring, macrocyclic, and spiro ring systems, respectively. Cyclic compounds also feature prominently among recently approved small-molecule COVID-19 therapeutics, including Pfizer's Paxlovid (a combination of the 3CLpro inhibitor Nirmatrelvir and Ritonavir), Merck/Ridgeback's RdRp inhibitor Molnupiravir, Shionogi's 3CLpro inhibitor Ensitrelvir, and China's RdRp inhibitor VV116.[**?**]

Advances in computational chemistry and cheminformatics have driven the evolution of molecular scaffold analysis. The foundational approach is the Murcko framework, introduced by Bemis and Murcko,[**?**, **?**] which deconstructs molecules into ring systems, linkers, side chains, and a combined framework of rings and linkers. Building on this, the Scaffold Tree (ST) method employs a hierarchical tree to describe ring systems, iteratively pruning rings according to prioritization rules until only one remains.[**?**] Similarly, SCONP trims terminal side chains to obtain scaffolds, then establishes parent–child relationships to create a classification tree.[**?**]

Graph theory provides a powerful framework for analyzing ring systems by representing compounds as graphs (atoms as nodes, bonds as edges). Graph-based techniques such as the smallest set of smallest rings (SSSR) algorithm and ring perception methods have been instrumental in characterizing ring systems based on topological properties, symmetry, and aromaticity.[**?**, **?**] Herein, we describe a novel strategy for cyclic compound analysis using the SSSR algorithm to identify and classify molecules according to their ring connectivity.

Drug discovery demands novel structures, necessitating the generation of synthetic datasets that mirror the complexity of marketed drugs or natural products for virtual screening. Traditional fragment-based assembly methods struggle with highly complex ring systems. In recent years, deep generative models have emerged as powerful tools for exploring chemical space, offering advantages over conventional methods by operating without direct reliance on structural similarity.[**?**] These models can expand the chemical space of drug-like compounds, enhance structural diversity, and generate large-scale libraries of novel cyclic compounds for future drug research.[**?**]

In this work, we focus on drug-like and natural product-like cyclic compounds. First, we developed D3Rings, a rapid and accurate method for ring system identification and classification. Second, we systematically classified cyclic compounds across multiple databases using D3Rings. Third, we employed three deep generative models—VAE, AAE, and CharRNN—to construct large-scale natural product-like and drug-like molecular datasets. Finally, through virtual screening of millions of compounds, we demonstrated that large-scale databases

enhance the discovery of structurally diverse, high-affinity ligands for drug–target interactions.

MATERIALS AND METHODS

Data Collection

We utilized three molecular databases: ChEMBL30 (drug-like compounds), DrugBank 5.1.9 (approved and investigational drugs), and COCONUT (an extensive open natural product collection).[**?**] These databases exhibit distinct characteristics due to their different data sources. We extracted molecules in SMILES format[**?**] from ChEMBL30, DrugBank 5.1.9, and COCONUT (January 2022), removing the few molecules incompatible with RDKit v.2020.09.1.

Molecular Classification Scheme

Figure 1A–C

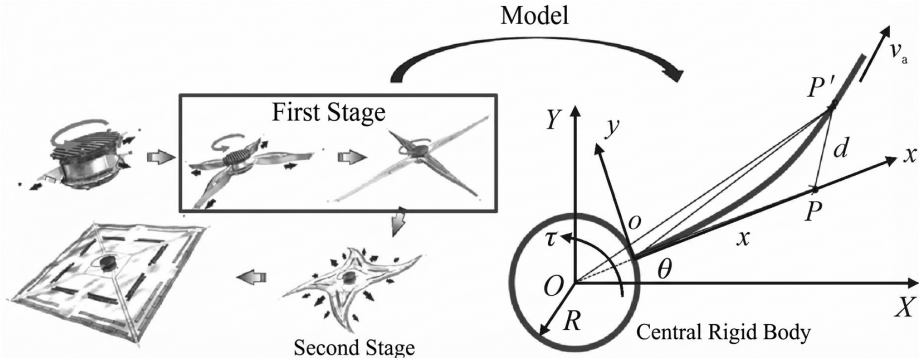

Figure 1: Figure 1

illustrates our classification system based on ring presence and connectivity. Molecules are categorized as: - Acyclic compounds (no ring structures) - Monocyclic compounds (single ring or multiple rings without direct linkages) - Joined ring compounds (two or more directly connected rings)

Joined ring compounds exhibit diverse connectivity patterns (Figure 1C). Spiro rings share a single bridgehead atom with a twisted geometry.[**?**] Fused rings share two adjacent atoms (common ring edges), while bridged rings share two nonadjacent atoms with intervening atoms. Cage rings represent a special class of hollow, three-dimensional structures.[**?**] Macrocyclic compounds contain rings of 12 or more atoms (Figure 1D).[**?**]

Using D3Rings, we classified molecules from ChEMBL30, DrugBank 5.1.9, and COCONUT into these categories based on structural characteristics.

Ring Structure Classification: D3Rings

Figure 2A [FIGURE:2] outlines the D3Rings workflow. The program first identifies ring structures using the SSSR method, then checks for directly connected rings. Subsequent classification leverages unique properties of each ring type: spiro compounds are identified by single bridgehead atoms, macrocycles by rings containing $ $12 atoms, and fused/bridged compounds by the maximum number of directly connected rings.

Figure 2B illustrates the calculation of the maximum number of joined rings. Molecules are fragmented at acyclic bonds, and the ring count for each fragment is calculated as:

$$\text{Number of joined rings} = \text{Number of ring bonds} - \text{Number of ring atoms} + 1$$

The maximum value across all fragments defines the molecule's final classification. For molecules with multiple ring assemblies, classification is based on the maximum number of directly connected rings. The entire process is implemented in Python using RDKit v.2020.09.1. D3Rings is available at http://www.d3pharma.com/D3Rings/.

Halogenated Compound Analysis

To investigate halogen distribution across databases, we analyzed all compounds, spiro ring compounds, fused/bridged ring compounds, and macrocyclic compounds from ChEMBL30, DrugBank 5.1.9, and COCONUT. We quantified the proportions of halogen-containing compounds, single-halogen-type compounds, and the distribution of halogen atom counts within each pool.

Deep Generative Models for Molecular Database Generation

Deep generative models offer powerful approaches for exploring chemical space. We employed three architectures:

Variational Autoencoder (VAE) encodes high-dimensional input into a low-dimensional latent space and reconstructs it through a decoder (Figure 3A

).[**?**, **?**] The loss function combines reconstruction error and Kullback–Leibler divergence.[**?**, **?**] Our implementation uses a bidirectional GRU recurrent neural network with a linear output layer as the encoder and a three-layer GRU with dropout as the decoder. VAE generates novel cyclic drug molecules with properties similar to the training set.

Adversarial Autoencoder (AAE) replaces the KL divergence in VAE with adversarial training for variational inference (Figure 3B).[**?**, **?**] The encoder uses a single-layer LSTM, the decoder a two-layer LSTM, and the discriminator a two-layer fully connected network with exponential linear unit activation. AAE generates molecules with desired properties aligned with training data.

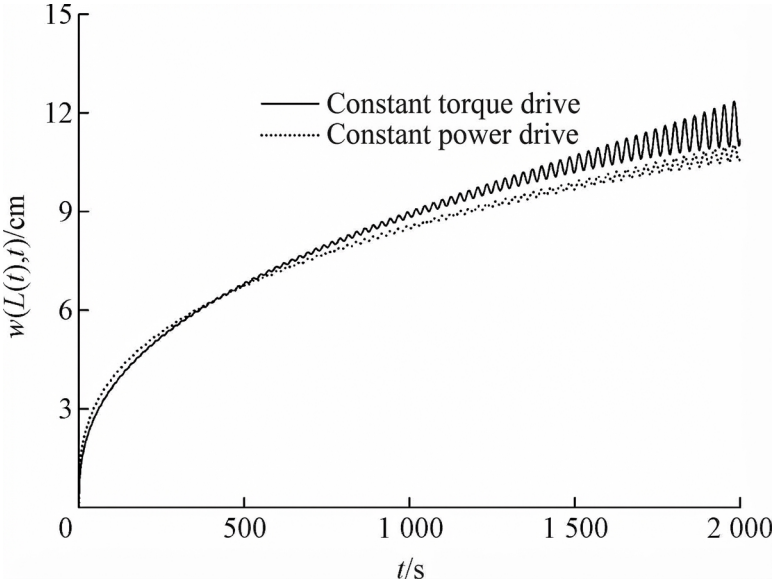

Figure 2: Figure 3

Character-Level Recurrent Neural Network (CharRNN) models character sequences by predicting the next character distribution based on observed characters.[**?**, **?**] We used three-layer LSTM cells with dropout and Softmax output. Trained by maximizing log-likelihood (Figure 3C), CharRNN generates novel cyclic drug molecules and is particularly effective when target properties relate to sequential characteristics like functional groups.

Model Training and Dataset Generation

We split DrugBank (5.1.9) and COCONUT (January 2022) into training and validation sets (9:1 ratio). The DrugBank training set contained 10,154 molecules (1,129 validation), while COCONUT contained 366,228 molecules (40,692 validation). Trained models generated 40,000 valid molecules each for DrugBank and 1,500,000 each for COCONUT. After merging and deduplication, we obtained a DrugBank-like dataset of 119,381 molecules and a COCONUT-like dataset of 4,185,929 molecules—each approximately tenfold larger than the source database. SMILES strings served as input and output representations.

Evaluation of Generated Molecular Databases

We assessed generated molecules using the Molecular Sets (MOSES) metrics:[**?**] - **Validity**: Fraction of chemically valid SMILES - **Novelty**: Fraction not present in training set - **Internal Diversity**: Structural diversity within generated set - **Unique 10k**: Uniqueness among top 10,000 valid molecules - **Frag**: BRICS

fragment similarity to training set - **Scaff**: Bemis–Murcko scaffold similarity - **SNN**: Similarity to nearest neighbor

We also compared distributions of molecular weight (MW), LogP, quantitative estimation of drug-likeness (QED), and synthetic accessibility (SA) between training and generated sets. To verify compositional consistency, we applied D3Rings to classify acyclic, monocyclic, spiro, fused/bridged, cage, and macrocyclic compounds in the generated databases.

Virtual Screening: Impact of Database Size

Molecular docking is a powerful tool for discovering active compounds from chemical libraries.[**?**, **?**, **?**] To evaluate the benefits of large-scale datasets, we performed docking against SARS-CoV-2 conserved proteins (3CLpro, RdRp, and nsp13).[**?**] We created four screening libraries from the COCONUT-like dataset (4,185,929 molecules total): subsets of 0.1%, 1%, 10%, and 100% (4,186, 41,859, 418,593, and 4,185,929 molecules, respectively). All docking calculations used the Glide HTVS program in Schrödinger Release 2020.

RESULTS AND DISCUSSION

Molecular Classification Statistics Across Databases

After initial filtering, our final datasets contained 1,038,551 (ChEMBL30), 11,283 (DrugBank 5.1.9), and 406,920 (COCONUT January 2022) molecules. D3Rings classification revealed striking differences among databases (Figure 4 [FIGURE:4], Table S1).

Overall Ring Distribution: - Acyclic compounds were most prevalent in DrugBank (13.77%) compared to ChEMBL (1.08%) and COCONUT (5.14%) - Monocyclic compounds (single ring or multiple rings without direct linkages) comprised similar fractions in ChEMBL and DrugBank (42.85% vs. 42.15%) but were less common in COCONUT (23.28%) - Joined ring compounds (two or more directly linked rings) dominated COCONUT (71.58%) compared to ChEMBL (56.07%) and DrugBank (44.08%)

In summary, cyclic compounds predominate across all datasets, with ChEMBL showing the fewest acyclic structures, COCONUT the most cyclic structures, and DrugBank the simplest overall ring topology (acyclic + monocyclic >55%). COCONUT exhibits the greatest structural complexity, with joined ring compounds representing over three-quarters of its cyclic molecules.

Detailed Joined Ring Analysis: The right panel of Figure 4 and Table S2 provide deeper analysis across three classification schemes: spiro presence, fused/bridged presence, and maximum number of fused rings.

Key findings: - **Spiro compounds**: COCONUT is notably enriched (5.38%) compared to ChEMBL and DrugBank - **Fused/bridged compounds**: Dom-

inant in all datasets (55.10%, 43.69%, and 71.15% for ChEMBL, DrugBank, and COCONUT, respectively) - **Ring complexity**: COCONUT contains substantially more structures with $ $2 fused rings, particularly those with three or more fused rings - **Cage compounds**: More abundant in COCONUT than in drug or bioactive databases, including alkaloids, terpenoids, xanthones, and unique marine-derived cage structures

Macrocyclic Compounds: COCONUT contains significantly more macrocycles than ChEMBL or DrugBank—approximately 3.6× and 2.3× their proportions, respectively (Figure 5 [FIGURE:5], Table S1).

These results demonstrate that COCONUT compounds possess greater structural complexity than those in ChEMBL and DrugBank, with characteristic features including spiro, fused, bridged, cage, and macrocyclic systems.

Halogenated Compound Distribution

Halogenated compounds follow a consistent trend across databases: bioactive compounds (ChEMBL) show the highest halogenation rates, followed by DrugBank, with natural products showing the lowest (Figures 6 and S1, Tables S3–S4). This pattern holds across all structural classes, including spiro, fused/bridged, and macrocyclic compounds.

Despite the prevalence of halogenated compounds in drug-like libraries, fewer halogenated drugs reach market than expected. Natural products, a key inspiration for drug discovery, contain relatively few halogenated structures.[**?**] While the Dictionary of Natural Products documented only 10,310 halogenated natural products as of March 2021, our COCONUT analysis identified 40,722 halogenated natural products and natural product-like molecules—a substantial increase.[**?**, **?**]

Fluorine- and chlorine-containing compounds dominate across all databases, while bromine and iodine compounds remain scarce (Figures 6 and S1). This reflects the greater stability of fluorine and chlorine compounds compared to iodinated analogs. Notably, most halogenated natural products contain only one halogen atom (up to 61%), whereas halogenated drugs more frequently contain multiple halogen atoms (nearly 50%), highlighting the importance of halogen substitution for modulating lipophilicity, pKa, conformation, and bioavailability.[**?**]

Large-Scale Molecular Dataset Generation

We trained VAE, AAE, and CharRNN models on DrugBank (10,154 molecules) and COCONUT (366,228 molecules) training sets (hyperparameters in Table S6). Each model generated 40,000 valid DrugBank-like molecules and 1,500,000 COCONUT-like molecules. After merging and deduplication, we obtained: - **DrugBank-like dataset**: 119,381 molecules (~10× expansion) - **COCONUT-like dataset**: 4,185,929 molecules (~10× expansion)

To assess the generated molecules, we randomly sampled 10,000 molecules from each training set and generated set, then applied t-SNE dimensionality reduction (Figure 7

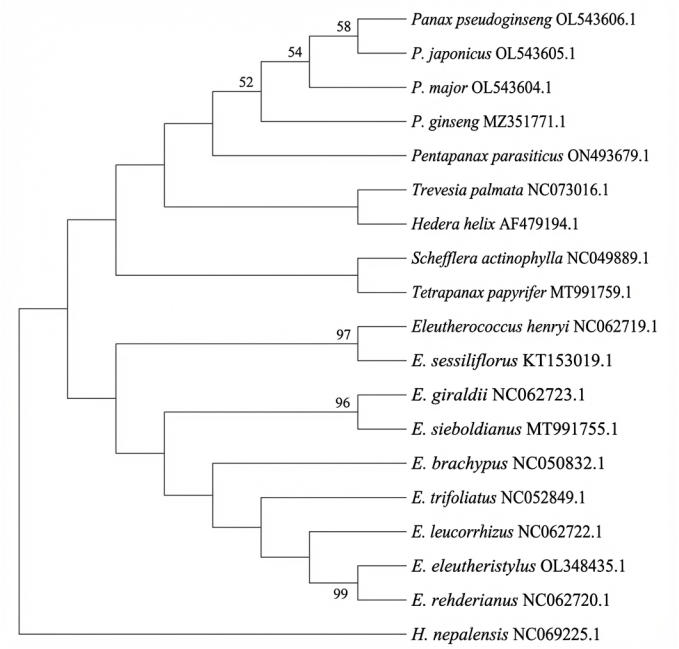

Figure 3: Figure 7

). The substantial overlap between generated and training molecules demonstrates that our models effectively reproduce training set properties while exploring complementary chemical space beyond the original molecules. Figure S3 showcases randomly selected valid, chemically reasonable molecules from both datasets.

Model Performance Evaluation

Table 1 summarizes model performance on MOSES metrics. All models achieved: - **Validity**: Near-perfect SMILES validity (1) - **Novelty**: >99% of generated molecules absent from training sets, indicating no overfitting - **Internal Diversity**: Scores of 0.888–0.902, confirming capacity for novel structure discovery - **Unique 10k**: >0.98 uniqueness among top 10,000 molecules, demonstrating broad generation capability

Substructure analysis revealed that generated molecules preserved BRICS fragment similarity to training sets. COCONUT-trained models produced more novel Bemis–Murcko scaffolds than DrugBank-trained models, suggesting greater scaffold innovation in natural product-like space. SNN values of 0.3–0.5,

combined with high novelty, confirm that models explore new chemical space without overfitting.

Figure 8 [FIGURE:8] compares property distributions (MW, LogP, QED, SA) between training and generated sets. Deep generative models faithfully reproduce training set characteristics, with CharRNN showing particularly close overlap for QED and SA distributions.

We compared our three models against LatentGAN[**?**] and JT-VAE[**?**] using MOSES benchmarks (Table S7, Figure S4). All models were trained on DrugBank and generated 10,000 molecules for evaluation. LatentGAN and JT-VAE performed substantially worse on internal diversity, uniqueness, and similarity metrics. JT-VAE exhibited bias toward lighter, easily synthesizable molecules (low SA) but with poor drug-like properties (low QED). LatentGAN generated excessively heavy molecules (high MW) with low QED and high SA. In contrast, VAE, AAE, and CharRNN closely matched the training data distribution.

Generated Database Composition Analysis

D3Rings analysis of the DrugBank- and COCONUT-like databases (Figure S5, Tables S8–S9) shows that the prevalence of each structural class aligns well with the original databases. However, the proportion of molecules with $ $3 fused rings is lower in generated datasets, suggesting that limited training data for highly complex cyclic structures reduces generation probability.

Halogenated compound statistics for generated databases (Figures S7–S9, Tables S10–S12) mirror the patterns observed in source databases, with COCONUT-like datasets showing greater halogen diversity than DrugBank-like datasets.

Impact of Database Size on Virtual Screening

To evaluate how library size affects hit discovery, we docked COCONUT-like subsets of varying size (0.1%, 1%, 10%, 100%) against SARS-CoV-2 3CLpro, RdRp, and nsp13.

Key findings: - **Hit rates**: The number of molecules with docking scores <-6.0 kcal/mol increased proportionally with library size (Figure 9A [FIGURE:9], Table S13). A 1,000-fold library expansion yielded ~1,000× more high-affinity hits - **Best scores**: Top docking scores improved monotonically with library size (Figure 9B). Improvements ranged from $\Delta = -0.920$ to -2.285 kcal/mol across targets - **Top 100 averages**: Mean scores of top 100 hits improved by $\Delta = -2.227$ to -2.440 kcal/mol when scaling from 0.1% to 100% of the library (Figure 9C), with no saturation observed

Cyclic structures consistently achieved the best docking scores among top-ranked molecules (Figure S10). Monocyclic and fused/bridged compounds performed particularly well across all three targets, with monocyclic

compounds showing optimal RdRp affinity and fused/bridged compounds exhibiting strongest binding to 3CLpro and nsp13.

These results align with Lyu et al., who observed steady improvements in docking scores without saturation when scaling libraries from 10^5 to $>10^9$ molecules.[**?**] Our findings confirm that large-scale virtual screening libraries enhance the discovery of diverse, high-affinity ligands, with cyclic structures—particularly fused/bridged and monocyclic compounds—providing optimal target engagement.

CONCLUSION

Cyclic compounds are ubiquitous in nature and represent a crucial class for drug discovery. We developed D3Rings, a fast and accurate method for identifying and classifying acyclic, monocyclic, spiro, fused/bridged, cage, and macrocyclic compounds. Statistical analysis of ChEMBL, DrugBank, and COCONUT databases revealed that COCONUT is exceptionally rich in complex cyclic architectures, including spiro, fused, bridged, cage, and macrocyclic systems. Using three deep generative models (VAE, AAE, CharRNN), we generated tenfold larger drug-like and natural product-like datasets. Docking against three anti-COVID-19 targets demonstrated that library expansion steadily improves binding affinity for top-ranked molecules. Our findings underscore the practical value of large-scale cyclic compound databases for future drug discovery and highlight the transformative potential of enriched molecular datasets in advancing pharmaceutical innovation.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: XXX.

- Halogenated compound proportions in ChEMBL30, DrugBank, and COCONUT (Figures S1–S2)
- Randomly generated molecules from VAE, AAE, and CharRNN (Figure S3)
- Property distributions for training vs. generated datasets (Figure S4)
- Molecular classification statistics for generated databases (Figures S5–S6)
- Halogenated compound statistics for generated databases (Figures S7–S9)
- Docking performance by structural class (Figure S10)

- Molecular classification statistics (Tables S1–S2)
- Halogenated compound counts and proportions (Tables S3–S5)
- VAE, AAE, and CharRNN hyperparameters (Table S6)
- Model performance comparison (Table S7)
- Generated database classification statistics (Tables S8–S9)
- Halogenated compound statistics for generated databases (Tables S10–S12)
- Docking performance vs. library size (Table S13)

AUTHOR INFORMATION

Corresponding Authors

Zhijian Xu – State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; ORCID: 0000-0002-3063-8473; Email: zjxu@simm.ac.cn

Weiliang Zhu – State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; ORCID: 0000-0001-6699-5299; Phone: +86-21-50805020; Email: wlzhu@simm.ac.cn

Authors

Minfei Ma, Xinben Zhang, Liping Zhou, Zijian Han, Yulong Shi, Jintian Li, Leyun Wu – State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

Author Contributions

Zhijian Xu and Weiliang Zhu designed and supervised the study. Minfei Ma performed the experiments. Xinben Zhang built the website. Liping Zhou, Zijian Han, Yulong Shi, Jintian Li, and Leyun Wu analyzed data. Minfei Ma, Weiliang Zhu, and Zhijian Xu drafted the manuscript.

Conflict of Interest

The authors declare no competing financial interest.

DATA AND SOFTWARE AVAILABILITY

SMILES strings were obtained from ChEMBL (https://www.ebi.ac.uk/chembl/), DrugBank (https://go.drugbank.com/), and COCONUT (https://coconut.naturalproducts.net/). MOSES benchmarking platform is available at https://github.com/molecularsets/moses. Schrödinger Maestro 10.2.010 was provided by Schrödinger.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2022YFA1004304) and the National Natural Science Foundation of China (82273851 and 82322067).

REFERENCES

(1) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; Dantonio, A.; Di, L.; Eng, H.; Ferre, R.; Gajiwala, K. S.; Gibson, S. A.; Greasley, S. E.; Hurst, B. L.; Kadar, E. P.; Kalgutkar, A. S.; Lee, J. C.; Lee, J.; Liu, W.; Mason, S. W.; Noell, S.; Novak, J. J.; Obach, R. S.; Ogilvie, K.; Patel, N. C.; Pettersson, M.; Rai, D. K.; Reese, M. R.; Sammons, M. F.; Sathish, J. G.; Singh, R. S. P.; Steppan, C. M.; Stewart, A. E.; Tuttle, J. B.; Updyke, L.; Verhoest, P. R.; Wei, L.; Yang, Q.; Zhu, Y. *An Oral SARS-CoV-2 Mpro Inhibitor Clinical Candidate for the Treatment of COVID-19.* **Science** 2021, 374 (6575), 1586–1593. DOI: 10.1126/science.abl4784.

(2) Tyndall, J. D. A. *S-217622, a 3CL Protease Inhibitor and Clinical Candidate for SARS-CoV-2.* **J. Med. Chem.** 2022, 10.1021/acs.jmedchem.2c00624.

(3) Sheahan, T. P.; Sims, A. C.; Zhou, S.; Graham, R. L.; Pruijssers, A. J.; Agostini, M. L.; Leist, S. R.; Schäfer, A.; Dinnon, K. H.; Stevens, L. J.; Chappell, J. D.; Lu, X.; Hughes, T. M.; George, A. S.; Hill, C. S.; Montgomery, S. A.; Brown, A. J.; Bluemling, G. R.; Natchus, M. G.; Saindane, M.; Kolykhalov, A. A.; Painter, G.; Harcourt, J.; Tamin, A.; Thornburg, N. J.; Swanstrom, R.; Denison, M. R.; Baric, R. S. *An Orally Bioavailable Broad-Spectrum Antiviral Inhibits SARS-CoV-2 in Human Airway Epithelial Cell Cultures and Multiple Coronaviruses in Mice.* **Sci. Transl. Med.** 2020, 12 (541), eabb5883. DOI: 10.1126/scitranslmed.abb5883.

(4) Thompson, M. G.; Stenehjem, E.; Grannis, S.; Ball, S. W.; Naleway, A. L.; Ong, T. C.; DeSilva, M. B.; Natarajan, K.; Bozio, C. H.; Lewis, N.; Dascomb, K.; Dixon, B. E.; Birch, R. J.; Irving, S. A.; Rao, S.; Kharbanda, E.; Han, J.; Reynolds, S.; Goddard, K.; Grisel, N.; Fadel, W. F.; Levy, M. E.; Ferdinands, J.; Fireman, B.; Arndorfer, J.; Valvi, N. R.; Rowley,

E. A.; Patel, P.; Zerbo, O.; Griggs, E. P.; Porter, R. M.; Demarco, M.; Blanton, L.; Steffens, A.; Zhuang, Y.; Olson, N.; Barron, M.; Shifflett, P.; Schrag, S. J.; Verani, J. R.; Fry, A.; Gaglani, M.; Azziz-Baumgartner, E.; Klein, N. P. *Effectiveness of Covid-19 Vaccines in Ambulatory and Inpatient Care Settings.* **N. Engl. J. Med.** 2021, 385 (15), 1355–1371. DOI: 10.1056/NEJMoa2110362.

(5) Qian, H. J.; Wang, Y.; Zhang, M. Q.; Xie, Y. C.; Wu, Q. Q.; Liang, L. Y.; Cao, Y.; Duan, H. Q.; Tian, G. H.; Ma, J.; Zhang, Z. B.; Li, N.; Jia, J. Y.; Zhang, J.; Aisa, H. A.; Shen, J. S.; Yu, C.; Jiang, H. L.; Zhang, W. H.; Wang, Z.; Liu, G. Y. *Safety, Tolerability, and Pharmacokinetics of VV116, an Oral Nucleoside Analog Against SARS-CoV-2, in Chinese Healthy Subjects.* **Acta Pharmacol. Sin.** 2022, 43 (12), 3130–3138. DOI: 10.1038/s41401-022-00895-6.

(6) Bemis, G. W.; Murcko, M. A. *The Properties of Known Drugs. 1. Molecular Frameworks.* **J. Med. Chem.** 1996, 39 (15), 2887–2893. DOI: 10.1021/jm9602928.

(7) Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T. *Comparative Analyses of Structural Features and Scaffold Diversity for Purchasable Compound Libraries.* **J. Cheminform.** 2017, 9 (1), 25. DOI: 10.1186/s13321-017-0228-8.

(8) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. *The Scaffold Tree—Visualization of the Scaffold Universe by Hierarchical Scaffold Classification.* **J. Chem. Inf. Model.** 2007, 47 (1), 47–58. DOI: 10.1021/ci600338x.

(9) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP).* **Proc. Natl. Acad. Sci. U. S. A.** 2005, 102 (48), 17272–17277. DOI: 10.1073/pnas.0503647102.

(10) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *Review of Ring Perception Algorithms for Chemical Graphs.* **J. Chem. Inf. Comput. Sci.** 1989, 29 (3), 172–187. DOI: 10.1021/ci00063a007.

(11) Lee, C. J.; Kang, Y. M.; Cho, K. H.; *A Robust Method for Searching the Smallest Set of Smallest Rings with A Path-Included Distance Matrix.* **Proc. Natl. Acad. Sci. U. S. A.** 2009, 106 (41), 17355–17358. DOI: 10.1073/pnas.0813040106.

(12) Lavecchia, A. *Deep Learning in Drug Discovery: Opportunities, Challenges and Future Prospects.* **Drug Discov. Today** 2019, 24 (10), 2017–2032. DOI: 10.1016/j.drudis.2019.07.006.

(13) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. *Exploring Chemical Space Using Natural Language Processing Methodologies for*

Drug Discovery. **Drug Discov. Today** 2020, 25 (4), 689–705. DOI: 10.1016/j.drudis.2020.01.020.

(14) Skalic, M.; Sabbadin, D.; Sattarov, B.; Sciabola, S.; De Fabritiis, G. *From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design.* **Mol. Pharm.** 2019, 16 (10), 4282–4291. DOI: 10.1021/acs.molpharmaceut.9b00634.

(15) Vogt, M. *Exploring Chemical Space — Generative Models and Their Evaluation.* **Artif. Intell. Life Sci.** 2023, 3, 100064. DOI: 10.1016/j.ailsci.2023.100064.

(16) Bian, Y.; Xie, X. Q. *Generative Chemistry: Drug Discovery with Deep Learning Generative Models.* **J. Mol. Model.** 2021, 27 (3), 71. DOI: 10.1007/s00894-021-04767-x.

(17) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. *ChEMBL: Towards Direct Deposition of Bioassay Data.* **Nucleic Acids Res.** 2019, 47 (D1), D930–D940. DOI: 10.1093/nar/gky1075.

(18) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. *DrugBank 5.0: A Major Update to the DrugBank Database for 2018.* **Nucleic Acids Res.** 2018, 46 (D1), D1074–D1082. DOI: 10.1093/nar/gkx1037.

(19) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. *COCONUT Online: Collection of Open Natural Products Database.* **J. Cheminform.** 2021, 13 (1), 2. DOI: 10.1186/s13321-020-00478-9.

(20) Weininger, D. *SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules.* **J. Chem. Inf. Comput. Sci.** 1988, 28 (1), 31–36. DOI: 10.1021/ci00057a005.

(21) Zheng, Y.; Tice, C. M.; Singh, S. B. *The Use of Spirocyclic Scaffolds in Drug Discovery.* **Bioorg. Med. Chem. Lett.** 2014, 24 (16), 3673–3682. DOI: 10.1016/j.bmcl.2014.06.081.

(22) Li, Y.; Zhang, L.; Wang, W.; Liu, Y.; Sun, D.; Li, H.; Chen, L. *A Review on Natural Products with Cage-Like Structure.* **Bioorg. Chem.** 2022, 128, 106106. DOI: 10.1016/j.bioorg.2022.106106.

(23) Bai, H.; Wang, J.; Li, Z.; Tang, G. *Macrocyclic Compounds for Drug and Gene Delivery in Immune-Modulating Therapy.* **Int. J. Mol. Sci.** 2019, 20 (9), 2097. DOI: 10.3390/ijms20092097.

(24) Kingma, D. P.; Welling, M. *Auto-encoding Variational Bayes.* arXiv December 10, 2022. DOI: 10.48550/arXiv.1312.6114.

(25) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. *Variational Inference: A Review for Statisticians.* **J. Am. Stat. Assoc.** 2017, 112 (518), 859–877. DOI: 10.1080/01621459.2017.1285773.

(26) Commenges, D. *Information Theory and Statistics: An Overview.* arXiv November 3, 2015. DOI: 10.48550/arXiv.1511.00860.

(27) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. *Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design.* **J. Cheminform.** 2018, 10 (1), 31. DOI: 10.1186/s13321-018-0286-7.

(28) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. *Adversarial Autoencoders.* arXiv May 24, 2016. DOI: 10.48550/arXiv.1511.05644.

(29) Hong, S. H.; Ryu, S.; Lim, J.; Kim, W. Y. *Molecular Generative Model Based on an Adversarially Regularized Autoencoder.* **J. Chem. Inf. Model.** 2020, 60 (1), 29–36. DOI: 10.1021/acs.jcim.9b00694.

(30) Bjerrum, E. J.; Threlfall, R. *Molecular Generation with Recurrent Neural Networks (RNNs).* arXiv May 17, 2017. DOI: 10.48550/arXiv.1705.04612.

(31) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. *Bidirectional Molecule Generation with Recurrent Neural Networks.* **J. Chem. Inf. Model.** 2020, 60 (3), 1175–1183. DOI: 10.1021/acs.jcim.9b00943.

(32) 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. DOI: 10.3389/fphar.2020.565644.

(33) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. *Ultra-Large Library Docking for Discovering New Chemotypes.* **Nature** 2019, 566 (7743), 224–229. DOI: 10.1038/s41586-019-0917-9.

(34) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X. P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; Zvonok, N.; Jain, M. K.; Savych, O.; Radchenko, D. S.; Nikas, S. P.; Petasis, N. A.; Moroz, Y. S.; Roth, B. L.; Makriyannis, A.; Katritch, V. *Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds.* **Nature** 2022, 601 (7893), 452–459. DOI: 10.1038/s41586-021-04220-9.

(35) Lyu, J.; Irwin, J. J.; Shoichet, B. K. *Modeling the Expansion of Virtual Screening Libraries.* **Nat. Chem. Biol.** 2023, 19 (6), 712–718. DOI: 10.1038/s41589-022-01234-0.

(36) Ma, M.; Yang, Y.; Wu, L.; Zhou, L.; Shi, Y.; Han, J.; Xu, Z.; Zhu, W. *Conserved Protein Targets for Developing Pan-Coronavirus Drugs Based on Sequence and 3D Structure Similarity Analyses.* **Comput. Biol. Med.** 2022, 145, 105455. DOI: 10.1016/j.compbiomed.2022.105455.

(37) Gribble, G. W. *Natural Organohalogens: A New Frontier for Medicinal Agents?* **J. Chem. Educ.** 2004, 81 (10), 1441. DOI: 10.1021/ed081p1441.

(38) Lu, Y.; Liu, Y.; Xu, Z.; Li, H.; Liu, H.; Zhu, W. *Halogen Bonding for Rational Drug Design and New Drug Discovery.* **Expert Opin. Drug Discov.** 2012, 7 (5), 375–383. DOI: 10.1517/17460441.2012.678829.

(39) Gribble, G. W. *The Diversity of Naturally Produced Organohalogens.* **Chemosphere** 2003, 52 (2), 289–297. DOI: 10.1016/S0045-6535(03)00207-8.

(40) Sorokina, M.; Steinbeck, C. *Review on Natural Products Databases: Where to Find Data in 2020.* **J. Cheminform.** 2020, 12 (1), 20. DOI: 10.1186/s13321-020-00424-9.

(41) Cochereau, B.; Meslet-Cladière, L.; Pouchus, Y. F.; Grovel, O.; Roullier, C. *Halogenation in Fungi: What Do We Know and What Remains to Be Discovered?* **Molecules** 2022, 27 (10), 3157. DOI: 10.3390/molecules27103157.

(42) Shinada, N. K.; de Brevern, A. G.; Schmidtke, P. *Halogens in Protein–Ligand Binding Mechanism: A Structural Perspective.* **J. Med. Chem.** 2019, 62 (21), 9341–9356. DOI: 10.1021/acs.jmedchem.8b01453.

(43) Prykhodko, O.; Johansson, S. V.; Kotsias, P. C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. *A De Novo Molecular Generation Method Using Latent Vector Based Generative Adversarial Network.* **J. Cheminform.** 2019, 11 (1), 74. DOI: 10.1186/s13321-019-0397-9.

(44) Jin, W.; Barzilay, R.; Jaakkola, T. *Junction Tree Variational Autoencoder for Molecular Graph Generation.* In **Artificial Intelligence in Drug Discovery**; 2020, pp 228–249.

TABLE OF CONTENTS

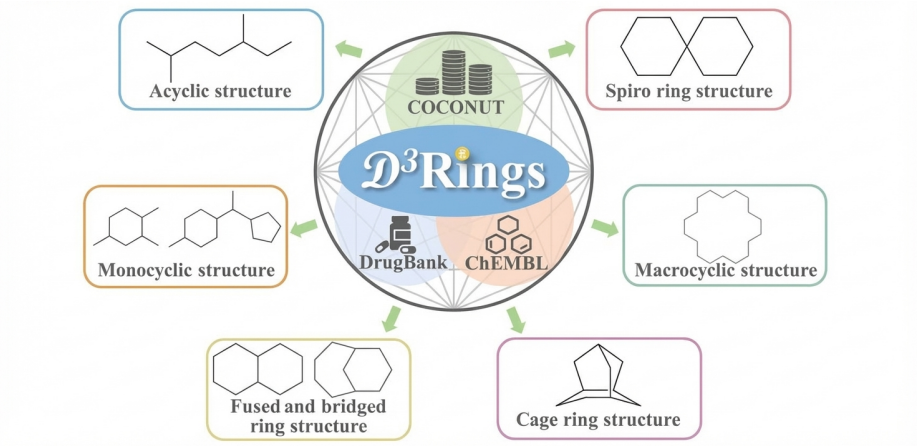

Figure 4: Figure 10

Figures

Source: ChinaXiv — Machine translation. Verify with original.