Journal of
Medicinal
Chemistry

pubs.acs.org/jmc

Article

RediscMol: Benchmarking Molecular Generation Models in Biological Properties

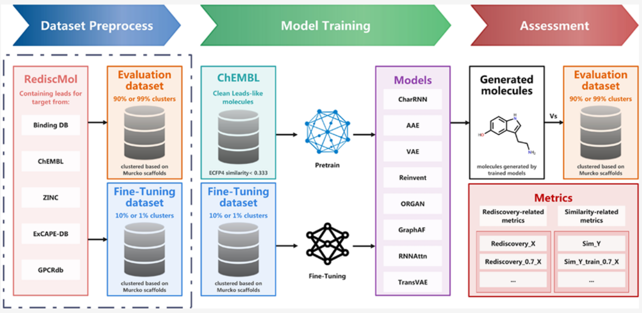
Gaoqi Weng,[§] Huifeng Zhao,[§] Dou Nie, Haotian Zhang, Liwei Liu,* Tingjun Hou,* and Yu Kang*

Cite This: *J. Med. Chem.* 2024, 67, 1533−1543

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Deep learning-based molecular generative models have garnered emerging attention for their capability to generate molecules with novel structures and desired physicochemical properties. However, the evaluation of these models, particularly in a biological context, remains insufficient. To address the limitations of existing metrics and emulate practical application scenarios, we construct the RediscMol benchmark that comprises active molecules extracted from 5 kinase and 3 GPCR data sets. A set of rediscovery- and similarity-related metrics are introduced to assess the performance of 8 representative generative models (CharRNN, VAE, Reinvent, AAE, ORGAN, RNNAttn, TransVAE, and GraphAF). Our findings based on the RediscMol benchmark differ from those of previous evaluations. CharRNN, VAE, and Reinvent exhibit a greater ability to reproduce known active molecules, while RNNAttn, TransVAE, and GraphAF struggle in this aspect despite their notable performance on commonly used distribution-learning metrics. Our evaluation framework may provide valuable guidance for advancing generative models in real-world drug design scenarios.



■ INTRODUCTION

The discovery of potent and safe molecules has been a long-standing challenge in the pharmaceutical domain. Virtual screening (VS) is a computational technique used to search for hit compounds within existing molecule libraries.[1] However, the chemical space of these existing compounds is extremely limited,[2−5] compared to the estimated size of the druglike chemical space ranging from 10^{23} to 10^{60}.[6] More recently, generative models have been proposed to produce novel molecules with desired property profiles from scratch, which are expected to explore the vast chemical space more intelligently, rather than relying solely on screening existing libraries.[7]

Since the first application of an autoencoder in 2016, officially published in 2018,[8] various machine learning (ML)-based generative models have been developed and customized for de novo design in the past few years. Many encouraging results have convincingly demonstrated the potential of these approaches.[9−11] Several ML frameworks and models, such as sequence-based recurrent neural networks (RNN),[7,12,13] variational autoencoders (VAE),[8,9,14] reinforcement learning (RL),[15−17] and generative adversarial networks (GAN)[18,19] were proved to be effective at delivering promising lead compounds. More recently, flow-based and diffusion models have demonstrated remarkable performance in molecular generation.[20−22]

Although a variety of molecular generative models have been developed, evaluating their performance remains a big challenge. Typically, two primary metrics are used for model assessment: distribution-learning and goal-directed metrics. In distribution-learning metrics, validity, uniqueness, and novelty are three major indicators for assessing the performance of generative models. However, these metrics are independent of the biological activities of the generated molecules and can be easily deduced from some simplistic models. For example, the AddCarbon model achieved rather sound performance in these three metrics by just randomly inserting a carbon atom into its simplified molecular-input line-entry system (SMILES) representation.[23] In addition, this model also tricked other distribution-learning metrics,[23] such as Kullback−Leibler (KL) divergence, and Frechet Chemnet distance (FCD).[24] Therefore, the current distribution-learning metrics fall short in providing insights into whether generative models can really produce molecules of practical use. Despite this limitation, the two most widely used benchmarks for evaluating these models, namely MOSES[25] and GuacaMol,[26] rely on these metrics.

In goal-directed metrics, the quality of the generated molecules is usually evaluated by quantitative structure−activity

Received: November 10, 2023
Revised: December 21, 2023
Accepted: December 22, 2023
Published: January 5, 2024

https://doi.org/10.1021/acs.jmedchem.3c02051

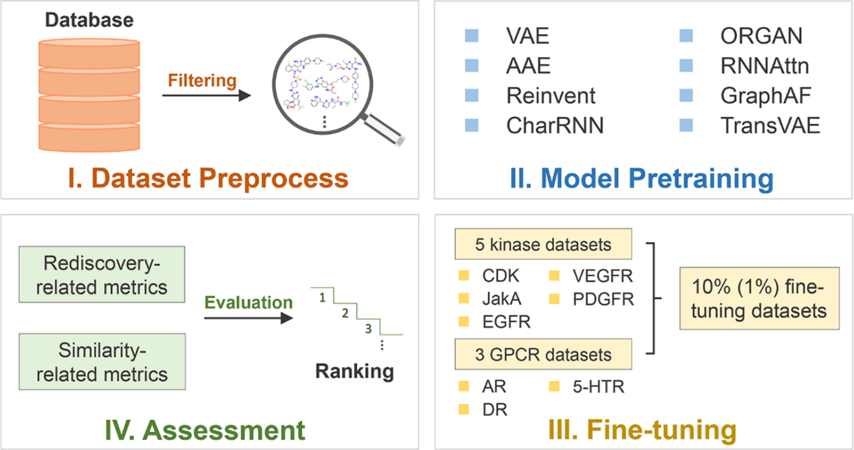


Figure 1. Overview of the performance evaluation of the generative models.

Table 1. Pretraining Results on the Data Set without Similar Compounds of the Kinase Data Sets

							RNNAttn			TransVAE		
pretrain	VAE	AAE	CharRNN	Reinvent	ORGAN	GraphAF	rand	high entropy	*k* high entropy	rand	high entropy	*k* high entropy
Validity	0.867	0.914	0.964	0.944	0.898	**1.000**	0.735	0.738	0.938	0.452	0.446	0.605
Uniqueness	0.998	0.997	0.995	0.997	0.993	0.954	**0.999**	**0.999**	0.905	0.947	0.944	0.898
IntDiv	0.882	0.885	0.883	0.883	0.880	0.908	0.887	0.887	0.839	0.888	0.890	0.848
log *P*	3.20 ± 1.82	3.20 ± 1.92	3.41 ± 1.89	3.40 ± 1.90	3.20 ± 1.81	3.19 ± 2.04	3.05 ± 1.97	3.07 ± 1.97	3.07 ± 2.29	2.75 ± 1.96	2.80 ± 2.06	4.29 ± 2.52
SA	3.11 ± 0.77	3.18 ± 0.81	3.16 ± 0.78	3.17 ± 0.85	2.95 ± 0.75	3.48 ± 1.21	3.73 ± 0.99	3.72 ± 0.98	2.58 ± 0.75	3.72 ± 0.98	3.67 ± 1.01	2.84 ± 0.72
QED	0.58 ± 0.20	0.56 ± 0.21	0.55 ± 0.21	0.56 ± 0.21	0.59 ± 0.20	0.50 ± 0.22	0.54 ± 0.21	0.54 ± 0.20	0.62 ± 0.18	0.55 ± 0.20	0.54 ± 0.21	0.49 ± 0.21
MW	382.84 ± 99.29	382.92 ± 105.87	397.44 ± 104.35	388.34 ± 108.50	369.15 ± 98.80	346.73 ± 153.55	374.79 ± 113.77	371.56 ± 107.08	352.36 ± 68.54	348.65 ± 119.53	348.88 ± 140.70	391.71 ± 142.13
Novelty	0.939	0.911	0.863	0.937	0.930	0.991	**1.000**	**1.000**	**1.000**	0.999	0.999	0.997
SNN/ Gen_train	0.545	0.578	0.645	0.558	0.562	0.423	0.360	0.361	0.438	0.370	0.376	0.425

Bold-faced values represent the model exhibiting the most superior performance in the given row's indicators.

relationship (QSAR) models and scoring functions. However, it is difficult to accurately predict the biological profile of a molecule, such as bioactivities and drug-like properties. Limited by the model architecture and training set, most models used in practice do include specific biases.[23] Taking convolutional neural network (CNN) models trained on the DUD-E data set as an example, the excellent performance of these models is attributed to learning the analogue and decoy bias in the data set rather than the features of protein−ligand interactions.[27] In addition, for models trained on PDBbind consisting of experimentally determined complex structures with known binding affinities, the performance may depend on the similarity of atomic features existing in the test and training data sets.[28] However, the amount of data in PDBbind is still limited and it suffers from the data redundancy caused by the similarity of proteins and ligands.[28]

In GuacaMol,[26] the rediscovery and similarity metrics are employed to describe the similarity between the generated molecules and a set of target molecules comprising marketed drugs. The rediscovery metric suggests that if a model can reproduce the target molecules with experimentally verified biological activities, then it could be considered capable of generating active molecules. However, this benchmark suffers from the very limited number of target molecules, with only three marketed drugs for rediscovery, which may introduce a chance factor. Moreover, in practical scenarios, a pretrained generative model is usually fine-tuned using the active molecules specific to the target protein of interest, suggesting that there is an urgent need for a benchmark to better emulate practical application scenarios.

In this study, considering the accuracy of computational models and the limitation of goal-directed metrics in GuacaMol, the RediscMol benchmark consisting of active molecules from 5 kinase and 3 G protein-coupled receptor (GPCR) data sets was constructed for fine-tuning and evaluation of pretrained models. Moreover, different from the common distribution-learning metrics, computational scores, and goal-directed metrics in GuacaMol, the rediscovery- and similarity-related metrics are proposed and employed to assess the performance of 8 generative models, which consider the generalizability of models and the activity of target molecules. After pretrained in the ChEMBL data set excluding the similar molecules to the RediscMol, the models are fine-tuned on 10% molecules with the same generic Murcko scaffolds from the RediscMol. Then, the remaining molecules in RediscMol are used as the target data set for rediscovery. Additionally, to assess the performance of the generative models for novel targets, the models are fine-tuned on 1% compounds with the same generic Murcko scaffolds from RediscMol, which usually contain around 100 compounds. Based on our evaluation framework, we have observed

https://doi.org/10.1021/acs.jmedchem.3c02051

Table 2. Fine-Tuning Results on the CDK 10%-Fine-Tuning Data Sets

							RNNAttn			TransVAE		
CDK	CharRNN	AAE	VAE	Reinvent	ORGAN	GraphAF	rand	high entropy	*k* high entropy	rand	high entropy	*k* high entropy
Validity	0.973 ± 0.010	0.969 ± 0.007	0.952 ± 0.017	0.983 ± 0.015	0.852 ± 0.027	**1.000 ± 0.000**	0.705 ± 0.026	0.931 ± 0.013	0.714 ± 0.023	0.370 ± 0.022	0.564 ± 0.015	0.364 ± 0.015
Uniqueness	0.517 ± 0.069	0.185 ± 0.012	0.149 ± 0.017	0.135 ± 0.055	0.286 ± 0.029	0.961 ± 0.007	**0.999 ± 0.001**	0.914 ± 0.039	**0.999 ± 0.001**	0.965 ± 0.015	0.869 ± 0.020	0.942 ± 0.010
IntDiv	0.862 ± 0.008	0.863 ± 0.006	0.855 ± 0.010	0.854 ± 0.013	0.862 ± 0.005	0.901 ± 0.001	0.882 ± 0.002	0.829 ± 0.005	0.881 ± 0.002	0.889 ± 0.003	0.852 ± 0.008	0.894 ± 0.004
Novelty	0.929 ± 0.026	0.789 ± 0.050	0.735 ± 0.054	0.659 ± 0.164	0.889 ± 0.036	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
SNN/Gen_train	0.527 ± 0.048	0.630 ± 0.043	0.692 ± 0.042	0.704 ± 0.082	0.545 ± 0.037	0.242 ± 0.009	0.243 ± 0.008	0.290 ± 0.011	0.244 ± 0.008	0.233 ± 0.009	0.264 ± 0.011	0.230 ± 0.010
SNN/Gen_goal	0.477 ± 0.025	0.517 ± 0.022	0.552 ± 0.022	0.556 ± 0.037	0.469 ± 0.022	0.276 ± 0.005	0.277 ± 0.006	0.340 ± 0.007	0.278 ± 0.006	0.266 ± 0.006	0.303 ± 0.007	0.263 ± 0.006
IntDiv_Rediscovery	0.794 ± 0.017	0.775 ± 0.023	0.790 ± 0.016	0.753 ± 0.040	0.756 ± 0.036	0.480 ± 0.054						
SNN/Rediscovery_train	0.792 ± 0.016	0.785 ± 0.021	0.799 ± 0.019	0.797 ± 0.025	0.779 ± 0.026	0.524 ± 0.078		0.405 ± 0.000				
Rediscovery	0.006 ± 0.001	0.006 ± 0.002	0.008 ± 0.002	0.009 ± 0.002	0.004 ± 0.002	0	0	0	0	0	0	0
Rediscovery_number	**87.500 ± 18.954**	26.700 ± 8.259	26.400 ± 6.530	25.800 ± 17.440	27.100 ± 11.309	1.500 ± 1.025	0	0.100 ± 0.300	0	0	0	0

Bold-faced values represent the model exhibiting the most superior performance in the given row's indicators.

significant variations in the performance results of the 8 generative models when compared to previous evaluations.[20,29,30] We postulate that our evaluation framework can provide optional guidance and testing for generative models within a biological context. The benchmark and evaluation code are available at https://github.com/gaoqiweng/RediscMol.

RESULTS AND DISCUSSION

Workflow. The workflow of the performance evaluation is shown in Figure 1, which consists of the data set preprocess, pretraining, fine-tuning, and assessment. The molecules in the ChEMBL database with an ECFP4 similarity higher than 0.333 to any molecule in the RediscMol data set were removed, and the remaining molecules in ChEMBL were used as the pretraining data set. It should be noted that in the case of the kinase and GPCR data sets, two distinct pretraining data sets were developed. The RediscMol data set for fine-tuning comprised 5 kinases (EGFR, CDK, JakA, VEGFR, and PDGFR) and 3 GPCR (AR, 5-HTR, and DR) data sets. Eight generative models, including CharRNN, VAE, AAE, RNNAttn, TransVAE, Reinvent, ORGAN, and GraphAF, were pretrained on the preprocessed ChEMBL data set. The pretraining results are summarized in Tables 1 and S2. A total of 30,000 compounds generated by each model were utilized for the performance assessment. For RNNAttn and TransVAE, three sampling modes, including random, high entropy, and *k*-random high entropy, were used. Among these models, GraphAF performs best in the validity metric, and RNNAttn yields the best results in the uniqueness and novelty metrics, which are consistent with the results in their article.[20] In addition, these models generated molecules with reasonable molecular properties, such as log *P*, SA, QED, and MW. Overall, all generative models have learned to generate valid, unique, and novel molecules.

Subsequently, models were fine-tuned on ten 10%-fine-tuning data sets with the same generic Murcko scaffolds, which increased the structural difference between the training sets and target molecules and improved the difficulty of the reproduction of the target molecules. Each model generated 30,000 molecules for the subsequent evaluation. Since validity, uniqueness, and novelty are related to the number of reproduced target molecules, RDKit was used to filter molecules to maintain the value of these three metrics at 1 for all generative models. Then, the rediscovery- and similarity-related metrics were employed for performance assessment. Considering the limited number of active molecules for novel targets, we also constructed 100 1%-fine-tuning data sets per target protein which usually consisted of around 100 compounds. Similarly, RDKit was employed for the filtering of molecules, and rediscovery- and similarity-related metrics were used to evaluate the performance of the generative models.

Evaluation on 10%-Fine-Tuning Data Sets. Given that the fine-tuning results based on other data sets are consistent with those obtained from CDK, we chose CDK as a representative to show the results. The fine-tuning results on the CDK 10%-fine-tuning data sets are shown in Table 2. GraphAF achieves the best results in the validity metric, and RNNAttn performs best in the uniqueness metric. Compared with the pretraining models, the performance of the fine-tuned CharRNN, VAE, AAE, Reinvent, and ORGAN in the uniqueness metrics drops significantly. For the rediscovery-related metrics, CharRNN performs significantly better than other models and reproduces an average of 87.5 target molecules out of the ten 10%-fine-tuning data sets. Although RNNAttn, TransVAE, and GraphAF show excellent performance in validity, uniqueness, and novelty metrics, they perform poorly in the reproduction of target molecules. According to the SNN/Gen_train and SNN/Gen_goal metrics, the generated molecules from GraphAF, RNNAttn, and TransVAE exhibit limited similarity to the training and target data sets, which raises concerns about whether these models have effectively captured the chemical spatial information on the active molecules present in the CDK 10%-fine-tuning data set.

Considering the influence of the validity, uniqueness, and novelty metrics on the number of reproduced target molecules, RDKit was utilized to filter molecules. The fine-tuning results on the CDK 10%-fine-tuning data sets with RDKit filtering are shown in Figure 2, Tables 3 and S3. The number of reproduced

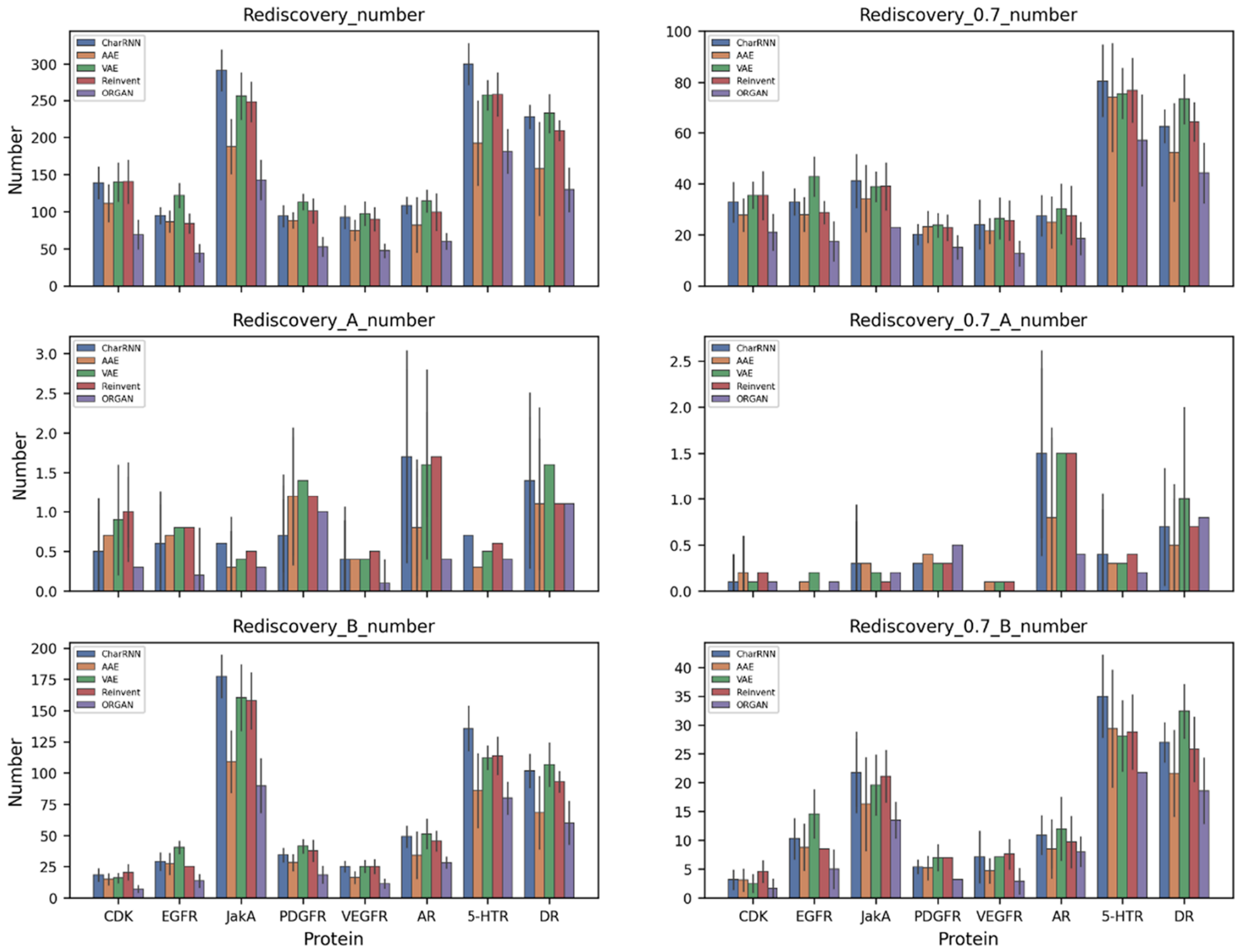


Figure 2. Rediscovery results on the 10%-fine-tuning data sets with RDKit filtering.

target molecules for CharRNN, VAE, AAE, Reinvent, and ORGAN increases significantly. In addition, Reinvent reproduces more molecules than CharRNN and most class A and B molecules. To further evaluate the generalization ability of the generative models, the Rediscovery_0.7_number metric was constructed, which represents the number of the reproduced target molecules with a similarity less than 0.7 to the nearest neighbor molecule in the training data set. Reinvent also performs best in the Rediscovery_0.7_number metric; CharRNN and VAE perform slightly worse than Reinvent, followed by AAE and ORGAN. Similarly, GraphAF, RNNAttn, and TransVAE underperform in reproducing target molecules. Considering the biological activities of the generated molecules, Reinvent also reproduces most class B molecules.

For the similarity-related metrics, VAE generates the most molecules similar to the target data set, and the number of similar molecules greater than 0.7 reaches 3126. Considering the similarity between the generated molecules and the training data set, Reinvent performs slightly better than VAE in the Sim_0.7_train_0.7_number metric. GraphAF, RNNAttn, and TransVAE struggle to generate molecules similar to the target data set.

The fine-tuning results on the EGFR 10%-fine-tuning data sets with RDKit filtering are shown in Table S4 and Figure 2. VAE performs significantly better than the other models with 40.5 class B molecules reproduced, much greater than 29.1 among the other best models. For the generalization ability, VAE also performs best and reproduces 42.8 molecules, of which 14.6 are class B molecules. According to the similarity-related metrics, VAE produces considerably more molecules similar to the target data set than the other models.

For JakA, as shown in Table S1, the data set contains more molecules than the other kinases, and the fine-tuning results are shown in Table S5 and Figure 2. CharRNN shows better performance in the rediscovery and generalization abilities on the fine-tuning data set with more molecules, which reproduces 177.5 class B molecules. VAE and Reinvent perform slightly worse than CharRNN. In the similarity-related metrics, Reinvent produces the most molecules with a similarity greater than 0.7 and 0.8 to the target data set, and CharRNN generates the most with a similarity greater than 0.9.

The PDGFR 10%-fine-tuning results are displayed in Table S6 and Figure 2. VAE achieves the best results and rediscovers 1.4 and 41.4 class A and B compounds, respectively, followed by Reinvent. For the generalization ability, both VAE and Reinvent reproduce 7 class B molecules. In addition, VAE also generates the most molecules similar to the target data set.

For VEGFR, as shown in Table S7 and Figure 2, VAE, CharRNN, and Reinvent perform closely in the rediscovery-related metrics. For the similarity-related metrics, VAE produces the most molecules similar to the target data set.

For GPCR data sets, RNNAttn, TransVAE, and GraphAF were excluded from the evaluation considering their poor performance in the kinase data sets. Other five models were pretrained on a new ChEMBL data set without similar compounds of the GPCR data sets (Table S2). The AR 10%

Table 3. Fine-Tuning Results on the CDK 10%-Fine-Tuning Data Sets with RDKit Filtering

							RNNAttn			TransVAE		
CDK	CharRNN	AAE	VAE	Reinvent	ORGAN	GraphAF	rand	high entropy	*k* high entropy	rand	high entropy	*k* high entropy
IntDiv	0.864 ± 0.006	0.868 ± 0.003	0.861 ± 0.005	0.862 ± 0.005	0.866 ± 0.003	0.901 ± 0.001	1	1	1	1	1	1
IntDiv_Rediscovery	0.803 ± 0.015	0.814 ± 0.005	0.817 ± 0.008	0.810 ± 0.011	0.794 ± 0.021	0.480 ± 0.054						
SNN/Rediscovery_train	0.779 ± 0.016	0.775 ± 0.015	0.779 ± 0.014	0.774 ± 0.009	0.773 ± 0.021	0.524 ± 0.078	0.472 ± 0.000		0.340 ± 0.000			
Rediscovery_number	138.800 ± 21.940	111.400 ± 25.562	140.000 ± 26.226	**140.300 ± 29.564**	69.300 ± 19.920	1.500 ± 1.025	0.100 ± 0.300	0	0.100 ± 0.300	0	0	0
Rediscovery_A_number	0.500 ± 0.671	0.700 ± 0.458	0.900 ± 0.700	**1.000 ± 0.447**	0.300 ± 0.458	0	0	0	0	0	0	0
Rediscovery_B_number	18.500 ± 5.239	15.000 ± 4.940	16.000 ± 4.243	**20.600 ± 6.468**	7.200 ± 3.219	0.200 ± 0.400	0	0	0	0	0	0
Rediscovery_0.7_number	32.800 ± 7.909	27.700 ± 6.543	**35.500 ± 5.334**	35.400 ± 9.604	21.000 ± 7.197	1.200 ± 0.748	0.100 ± 0.300	0	0.100 ± 0.300	0	0	0
Rediscovery_0.7_A_number	0.100 ± 0.300	**0.200 ± 0.400**	0.100 ± 0.300	**0.200 ± 0.400**	0.100 ± 0.300	0	0	0	0	0	0	0
Rediscovery_0.7_B_number	3.200 ± 1.661	3.100 ± 1.972	2.500 ± 1.628	**4.600 ± 1.960**	1.700 ± 1.616	0.200 ± 0.400	0	0	0	0	0	0
Sim_0.7_number	2632.100 ± 383.147	2530.300 ± 303.789	**3126.200 ± 363.647**	2894.700 ± 432.065	1533.500 ± 375.330	18.900 ± 14.591	0.900 ± 0.943	1.100 ± 0.831	0.300 ± 0.458	0.700 ± 0.781	0.800 ± 0.600	1.500 ± 0.806
Sim_0.7_train_0.7_number	527.200 ± 85.040	485.000 ± 103.276	578.700 ± 100.282	**580.500 ± 101.575**	402.800 ± 140.198	10.400 ± 6.037	0.800 ± 0.980	1.100 ± 0.831	0.300 ± 0.458	0.700 ±0.781	0.800 ± 0.600	1.500 ± 0.806
Sim_0.8_number	751.500 ± 90.944	670.300 ± 93.121	**853.600 ± 139.405**	792.700 ± 105.514	404.400 ± 88.550	3.700 ± 2.283	0.100 ± 0.300	0	0.100 ± 0.300	0.300 ± 0.458	0	0
Sim_0.8_train_0.7_number	124.000 ± 24.831	107.300 ± 25.613	128.000 ± 21.194	**133.700 ± 30.932**	83.600 ± 25.660	2.400 ± 2.010	0.100 ± 0.300	0	0.100 ± 0.300	0.300 ± 0.458	0	0
Sim_0.9_number	233.500 ± 32.601	185.900 ± 28.399	**243.600 ± 40.981**	236.200 ± 39.957	117.500 ± 34.106	1.800 ± 1.327	0.100 ± 0.300	0	0.100 ± 0.300	0	0	0
Sim_0.9_train_0.7_number	43.200 ± 9.379	36.500 ± 7.606	**47.300 ± 7.281**	46.000 ± 11.472	28.100 ± 9.944	1.300 ± 0.900	0.100 ± 0.300	0	0.100 ± 0.300	0	0	0

Bold-faced values represent the model exhibiting the most superior performance in the given row's indicators.

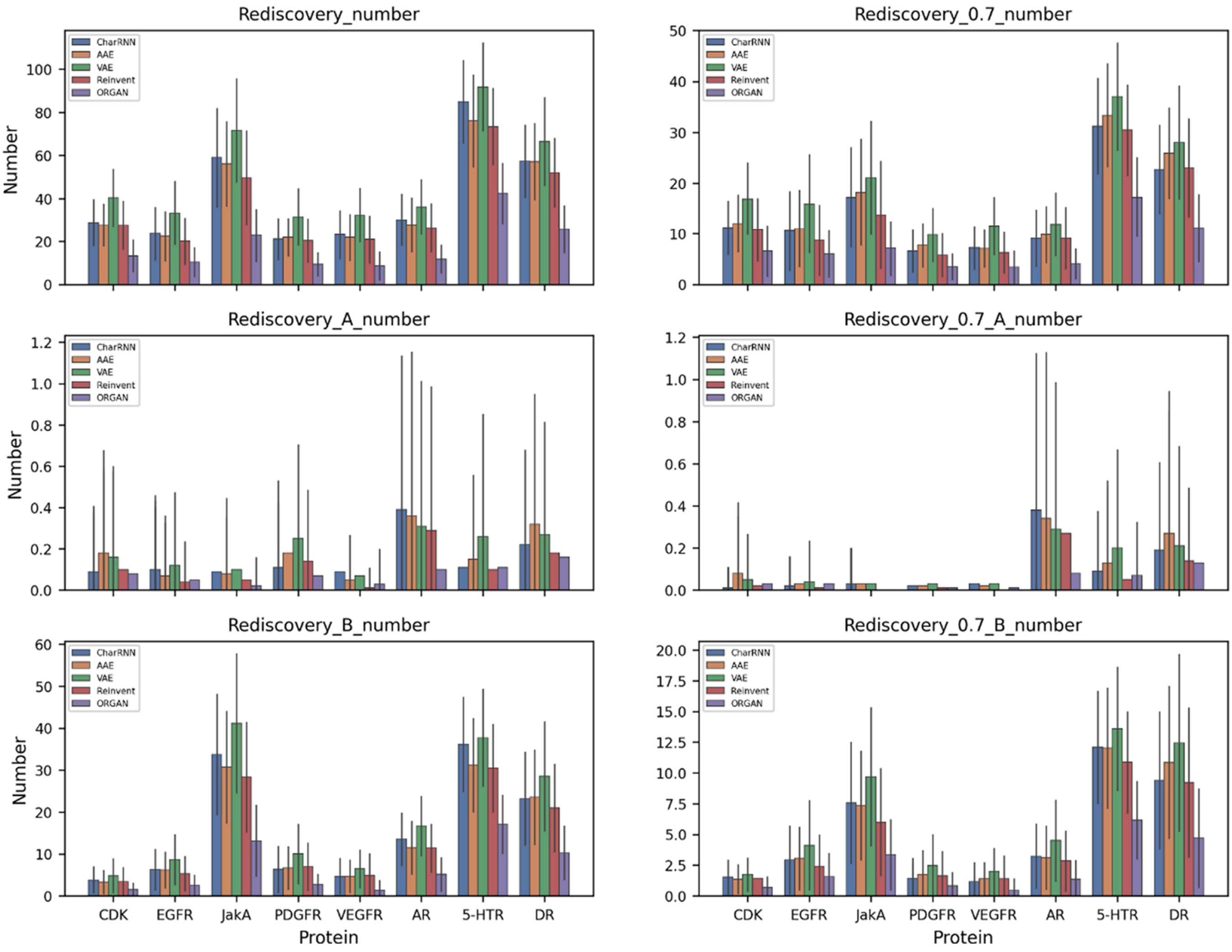


Figure 3. Rediscovery results on the 1%-fine-tuning data sets with RDKit filtering.

Table 4. Fine-Tuning Results on the CDK 1%-Fine-Tuning Data Sets with RDKit Filtering

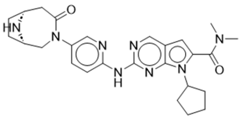
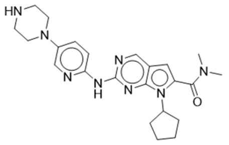
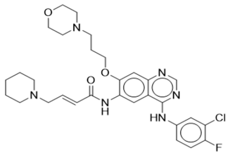
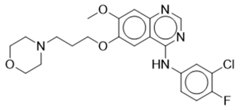
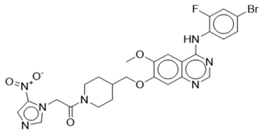
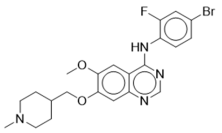
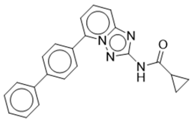
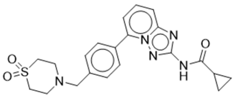
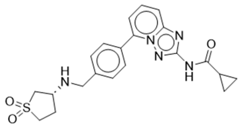
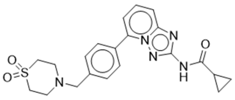
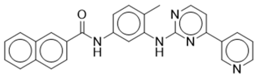
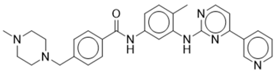
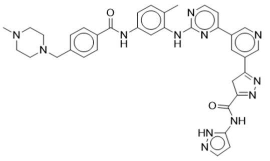
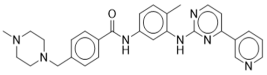
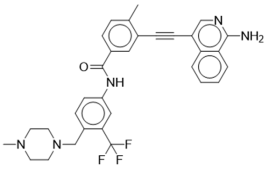
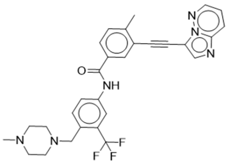
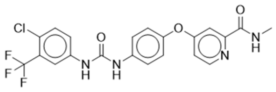
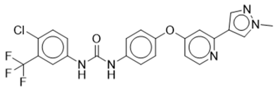
CDK	CharRNN	AAE	VAE	Reinvent	ORGAN
IntDiv	0.863 ± 0.009	0.864 ± 0.011	0.856 ± 0.014	0.861 ± 0.010	0.862 ± 0.009
IntDiv_Rediscovery	0.725 ± 0.051	0.733 ± 0.047	0.739 ± 0.044	0.708 ± 0.058	0.652 ± 0.091
SNN/Rediscovery_train	0.734 ± 0.042	0.727 ± 0.039	0.728 ± 0.036	0.740 ± 0.043	0.710 ± 0.070
Rediscovery_number	28.820 ± 11.018	27.700 ± 10.065	**40.440 ± 13.528**	27.640 ± 11.435	13.320 ± 7.585
Rediscovery_A_number	0.090 ± 0.286	**0.180 ± 0.498**	0.160 ± 0.441	0.100 ± 0.300	0.080 ± 0.366
Rediscovery_B_number	3.820 ± 3.235	3.320 ± 2.881	**4.790 ± 4.134**	3.440 ± 3.471	1.470 ± 1.723
Rediscovery_0.7_number	11.220 ± 5.326	12.080 ± 5.700	**16.940 ± 7.182**	10.820 ± 6.249	6.610 ± 5.048
Rediscovery_0.7_A_number	0.010 ± 0.099	**0.080 ± 0.271**	0.050 ± 0.218	0.020 ± 0.140	0.030 ± 0.171
Rediscovery_0.7_B_number	1.540 ± 1.424	1.370 ± 1.222	**1.740 ± 1.383**	1.420 ± 1.218	0.690 ± 0.902
Sim_0.7_number	935.180 ± 420.562	1236.570 ± 563.006	1476.010 ± 553.351	909.440 ± 424.861	430.420 ± 250.605
Sim_0.7_train_0.7_number	233.040 ± 88.346	287.180 ± 124.353	**383.470 ± 133.998**	209.350 ± 104.085	155.700 ± 92.095
Sim_0.8_number	224.360 ± 121.967	263.450 ± 145.181	**357.650 ± 178.948**	215.270 ± 115.912	94.540 ± 76.383
Sim_0.8_train_0.7_number	46.870 ± 20.807	53.300 ± 24.192	**78.390 ± 32.439**	44.160 ± 22.610	28.380 ± 20.384
Sim_0.9_number	53.240 ± 25.607	53.050 ± 24.058	**77.340 ± 32.271**	50.630 ± 22.797	23.190 ± 14.446
Sim_0.9_train_0.7_number	14.790 ± 6.930	16.020 ± 7.513	**23.110 ± 9.553**	14.450 ± 8.085	8.910 ± 6.782

Bold-faced values represent the model exhibiting the most superior performance in the given row's indicators.

fine-tuning results with RDKit filtering are displayed in Table S8 and Figure 2. VAE performs best and reproduces 1.6 and 51.4 class A and B compounds, respectively, followed by CharRNN and Reinvent. In the similarity-related metrics, Reinvent generates the most molecules with a similarity greater than 0.7 to the target data set, and VAE produces the most with a similarity greater than 0.8 and 0.9.

For 5-HTR, as shown in Table S9 and Figure 2, similar to JakA, CharRNN achieves the best results in the rediscovery and generalization abilities with 135.7 class B molecules reproduced,

Table 5. Class A Compounds Reproduced by VAE and Their Corresponding Nearest Neighbor Molecules in the Kinase Fine-Tuning Data Set

Target	The most similar compounds in the training set	Level	Reproduced class A compounds
CDK		B	
EGFR		C	
EGFR		B	
JakA		C	
JakA		C	
PDGFR		C	
PDGFR		B	
PDGFR		B	
VEGFR		D	

followed by Reinvent and VAE. In the similarity-related metrics, CharRNN and Reinvent produce the most molecules similar to the target data set.

The DR 10% fine-tuning results are displayed in Table S10 and Figure 2. VAE shows the best performance in the rediscovery and generalization abilities, which rediscovers 1.6 and 106.8 class A and B compounds, respectively, followed by CharRNN. In the similarity-related metrics, Reinvent produces the most molecules with a similarity greater than 0.7 to the target data set, and VAE generates the most with a similarity greater than 0.8 and 0.9.

To summarize, CharRNN, VAE, AAE, Reinvent, and ORGAN can reproduce a certain number of target molecules, while RNNAttn, TransVAE, and GraphAF struggle to reproduce the target molecules after fine-tuning. Among them, CharRNN, VAE, and Reinvent show better performance in the rediscovery and generalization abilities.

Evaluation on 1%-Fine-Tuning Data Sets. Considering the limited number of active molecules for novel targets, 1%-fine-tuning data sets were also constructed to evaluate the performance of the generative models in these targets. Since RNNAttn, TransVAE, and GraphAF underperform in 10%-fine-tuning data sets, they are not included in this assessment. The fine-tuning results on the CDK 1%-fine-tuning data sets with RDKit filtering are shown in Figure 3, Tables 4 and S11. Among them, VAE performs significantly better with 40.4 target molecules reproduced, of which 4.8 are class B molecules. For the generalization ability, VAE also performs best and reproduces 16.9 target molecules. In addition, VAE can produce significantly more molecules similar to the target data set than other models.

The fine-tuning results on the EGFR, JakA, PDGFR, VEGFR, AR, 5-HTR, and DR 1%-fine-tuning data sets with RDKit filtering are shown in Tables S12−S18 and Figure 3, respectively, which are consistent with those in CDK. VAE outperforms other models in rediscovery- and similarity-related metrics. In summary, VAE performs best when fine-tuned with only a few active molecules.

Reproduced Class A Compounds by VAE. To further explore the structural information on the generated molecules by the generative models, we investigated the class A compounds reproduced by VAE with a similarity less than 0.7 to the nearest neighbor molecule in the training data set. The kinase results are shown in Table 5. In CDK and EGFR, the molecules in the training set are more complex than those in the corresponding reproduced class A compounds. However, in practical application scenarios, these molecules might be modified based on the class A compounds. For JakA, the substitution of such R groups can provide medicinal chemists with more valuable ideas for structural modification. In particular, for the first compound, the phenyl ring is substituted by the 1,1-dioxidothiomorpholine group, which enhances both binding affinity and solubility. For PDGFR, the modification results based on the class C compound and the class B compound listed in the second row in the training set contribute more inspiration. For VEGFR, VAE generates the class A compounds based on the class D compounds in the training set. As displayed in Tables S19−S21, the GPCR data sets show similar results. Some structural modifications are capable of inspiring medicinal chemists. Overall, although core structures are retained and only some R groups are modified, VAE still can provide some helpful insights into structural modification for medicinal chemists.

■ CONCLUSIONS

Among the existing evaluation protocol, current distribution-learning and goal-directed metrics and benchmarks struggle to evaluate the generated molecules in a biological context. Here, we constructed RediscMol for the fine-tuning of the pretrained generative models and combined the rediscovery- and similarity-related metrics to evaluate the performance of the generative models in biological properties. In the 10%-fine-tuning data sets, CharRNN, VAE, and Reinvent achieve better results, followed by AAE and ORGAN. RNNAttn, TransVAE, and GraphAF underperform in rediscovery- and similarity-related tasks. In addition, considering the limited number of active molecules in novel targets, 1%-fine-tuning data sets were utilized to assess the performance of the generative models in these targets. After fine-tuning with a small number of active molecules, CharRNN, VAE, AAE, Reinvent, and ORGAN can still reproduce some target molecules, and VAE performs best. It is anticipated that our work may offer valuable insights into evaluating the performance of generative models in real-world scenarios.

■ EXPERIMENTAL SECTION

Benchmark. The pretraining data set was collected from the ChEMBL 29 database[31] and the data set was prepared by the following steps: (1) removing salts, (2) removing molecules including any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I, (3) neutralizing charge, (4) keeping molecules with the number of heavy atoms between 10 and 50, molecular weight (MW) not higher than 700, and the number of atoms in the ring not higher than 8, (5) removing the redundant molecules based on International Chemical Identifier (InChI), and (6) removing molecules with an ECFP4 similarity higher than 0.333 compared to a holdout set consisting of 5 preprocessed kinase or 3 GPCR data sets.

The data sets for fine-tuning and evaluation consist of the data sets for 5 kinases and 3 GPCRs, including epidermal growth factor receptors (EGFR), cyclin-dependent kinases (CDK), Janus kinases (JakA), vascular endothelial growth factor receptors (VEGFR), platelet-derived growth factor receptors (PDGFR), adrenergic receptors (AR), 5-hydroxytryptamine receptors (5-HTR), and dopamine receptors (DR), which were derived from BindingDB,[32] ChEMBL, ZINC,[2] GPCRdb,[33] and ExCAPE-DB.[34] The same preprocessing protocol for ChEMBL was also applied to the kinase and GPCR data sets except for step 6. For BindingDB, the molecules with K_i, $IC_{50,}$ and K_d more than 10 μM were removed; for ChEMBL, the molecules with K_i, IC_{50}, and K_d more than 10 μM and the inhibition rate less than 50% at 10 μM were removed. Moreover, compounds with irreversible activity data were removed. For ExCAPEDB, we collected the molecules with pXC50 of more than 5. For GPCR, we further collected the compounds form GPCRdb with pK_i, pIC_{50}, pEC_{50}, and pK_d more than 5. For a molecule with multiple activity metrics in different units, it will be considered to be active if one of the biological activity values meets the above requirements. Then, in terms of the biological activities of the molecules, the data sets for the abovementioned kinase and GPCR targets were classified into five classes of molecules.

Class A: Approved drugs. Besides, for the drug that is not reported mainly targeting the given protein, if it is a class B compound for this protein, it is also marked as a class A compound for this protein.

Class B: The compound with any of the K_i, IC_{50}, or K_d values toward the given protein is less than 10 nM.

Class C: The compound with any of the K_i, IC_{50}, or K_d values toward the given protein is less than 100 nM.

Class D: The compound with any of the K_i, IC_{50}, or K_d values toward the given protein is less than 1 μM.

Class E: The compound with any of the K_i, IC_{50}, or K_d values toward the given protein is less than 10 μM, or it only gives the inhibition rates.

The preprocessed data sets called RediscMol are shown in Table S1. For evaluation, the molecules were clustered based on the generic Murcko scaffolds; 10% of clusters with the same generic Murcko scaffolds were randomly sampled for fine-tuning, and the rest were used as the target data set. The above step was repeated 10 times, and the data sets for fine-tuning were named 10%-fine-tuning data sets. Similarly, 1% of clusters with the same generic Murcko scaffolds were randomly selected for fine-tuning, and the remaining were used as the target data set. This operation was repeated 100 times, and the data sets for fine-tuning were named 1%-fine-tuning data sets.

Baseline Models. In this study, eight generative models that cover different approaches of molecular generation were evaluated, including character-level recurrent neural networks (CharRNN),[25] variational autoencoders (VAE),[25] adversarial autoencoders (AAE),[25] Reinvent,[13] objective-reinforced generative adversarial network (ORGAN),[19,25] attention-based recurrent models (RNNAttn),[30] TransVAE,[30] and GraphAF.[20]

CharRNN uses RNN with long short-term memory (LSTM) to estimate the distribution of the next word given in a sequence of words. The model was trained with SMILES strings.

VAE is composed of two neural networks: an encoder to convert strings into a vector representation and a decoder to convert vector representations back into strings. The model is trained to minimize reconstruction loss and regularization term in the form of Kullback−Leibler divergence with SMILES strings.

ORGAN builds on a sequence-based GAN framework, which contains domain-specific objectives using RL apart from the discriminator reward. GAN usually incorporates at least two networks: a generator to produce indistinguishable examples from the data distribution and a discriminator to learn to distinguish the generated examples from real data samples. Then, both models are trained in alternation. Similar to Reinvent, considering the predictive accuracy of the scoring function, the model was trained using SMILES strings without the RL mode.

AAE replaces the variational inference of the VAE with GAN. A discriminator model is trained to distinguish samples from the hidden code of the autoencoder or from the prior distribution. Similar to VAE, we used SMILES as the input and output representations.

Reinvent includes the RNN and RL agent network. Considering the limited accuracy of the scoring function for the RL approach, the RL mode in Reinvent was not used in this study. This model was also trained with SMILES.

RNNAttn consists of three layers of gate recurrent unit (GRU) networks in both the encoder and decoder. Moreover, a single attention head is added after the final GRU layer in the encoder. Similarly, SMILES was used as the input and output representations.

TransVAE is a molecular generative model that uses the transformer-based VAE with the addition of attention layers to its encoder and decoder. This model was trained with SMILES strings.

GraphAF is a flow-based autoregressive model for molecular generation. We trained this model using SMILES strings, which can be converted into node/adjacency features by the preprocessing module in the model.

Evaluation Metrics. Our evaluation metrics encompass distribution-learning, rediscovery-related, and similarity-related measurements. The distribution-learning metrics align with those of MOSES, including validity, uniqueness, novelty, internal diversity (IntDiv), similarity to the nearest neighbor (SNN), and property distribution. We further assessed the models' performance through the rediscovery- and similarity-related metrics, which involve measuring the similarity between the training data set and the target molecules. The detailed metrics are presented as follows.

Validity reports the validity of the generated SMILES strings, and it was checked using the molecular structure parser of RDKit.

Uniqueness assesses whether models can generate unique molecules.

Novelty represents the ratio of the generated molecules that are not present in the training set.

IntDiv assesses the chemical diversity of the generated molecules in *G*. The last term in the formula calculates the average Tanimoto similarity of the generated molecules.

$$\mathrm{IntDiv}(G) = 1 - \frac{1}{|G|^2} \sum_{m1,m2 \in G} T(m1, m2)$$

log *P* is the octanol/water partition coefficient.

Synthetic Accessibility Score (SA) estimates how difficult it is to synthesize the generated molecule. The SA score was calculated based on the contributions of molecular fragments using RDKit.

Quantitative Estimation of Drug-likeness (QED) measures the drug-like properties of the given molecules, which are between 0 and 1.

MW is the sum of the atomic weights of molecules.

SNN is the average Tanimoto similarity between a molecule from one data set and its nearest neighbor molecule in another data set. In this study, an extended-connectivity fingerprint (ECFP) with a radius of 2 and 1024 bits was used to calculate the Tanimoto similarity by RDKit.[35]

SNN/Gen_train is the average Tanimoto similarity between a molecule from the generated data set and its nearest neighbor molecule in the training data set.

SNN/Gen_goal is the average Tanimoto similarity between a molecule from the generated data set and its nearest neighbor molecule in the target data set.

SNN/Train_goal is the average Tanimoto similarity between a molecule from the training data set and its nearest neighbor molecule in the target data set.

SNN/Goal_train is the average Tanimoto similarity between a molecule from the target data set and its nearest neighbor molecule in the training data set.

SNN/Rediscovery_train is the average Tanimoto similarity between a molecule from the reproduced target molecules and its nearest neighbor molecule in the training data set.

Rediscovery and **Rediscovery_number** are the ratio and number of reproduced target molecules, respectively. The target molecule is considered to be reproduced successfully if the SMILES representations of the two molecules are identical.

Rediscovery_X_number (X: A, B, C, D, E) represents the number of the reproduced class X target molecules.

Rediscovery_0.7 and **Rediscovery_0.7_number** indicate the ratio and number of the reproduced target molecules whose similarity to the nearest neighbor molecule in the training data set is less than 0.7 among all reproduced target molecules, respectively. The generalization abilities of the generative models can be described by both metrics.

Rediscovery_0.7_X_number (X: A, B, C, D, and E) represents the number of the reproduced class X target molecules whose similarity to the nearest neighbor molecule in the training data set is less than 0.7.

Sim_Y and **Sim_Y_number** (Y: 0.7, 0.8, 0.9) are the proportion and number of the generated molecules with a similarity more than Y to the target molecules, respectively.

Sim_Y_train_0.7 and **Sim_Y_train_0.7_number** (Y: 0.7, 0.8, 0.9) are the proportion and number of the generated molecules with a similarity more than Y to the target molecules and less than 0.7 to the nearest neighbor molecule in the training data set, respectively.

Sim_Y_train_0.7_X_number (X: A, B, C, D, E; Y: 0.7, 0.8, 0.9) is the number of the generated molecules with a similarity more than Y to the class X target molecules and less than 0.7 to the nearest neighbor molecule in the training data set.

■ ASSOCIATED CONTENT

Data Availability Statement

The benchmark and evaluation source code are freely available at https://github.com/gaoqiweng/RediscMol.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.3c02051.

> Data statistics of the preprocessed data sets, RediscMol; pretraining results of VAE, AAE, CharRNN, Reinvent, and ORGAN on the data set without similar compounds of the GPCR data sets; fine-tuning results on the 10%-fine-tuning data sets of CDK, EGFR, JakA, PDGFR, VEGFR, AR, 5-HTR, and DR with RDKit filtering; fine-tuning results on the 1%-fine-tuning data sets of CDK, EGFR, JakA, PDGFR, VEGFR, AR, 5-HTR, and DR with RDKit filtering; and class A compounds reproduced by VAE and their corresponding nearest neighbor molecules in the AR, 5-HTR, and DR fine-tuning data set (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Yu Kang − *Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058 Zhejiang, China;* orcid.org/0000-0002-0999-8802;
Email: liuliwei5@huawei.com

Tingjun Hou − *Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058 Zhejiang,*

China; orcid.org/0000-0001-7227-2580;
Email: tingjunhou@zju.edu.cn

Liwei Liu − *Advanced Computing and Storage Laboratory, Central Research Institute, 2012 Laboratories, Huawei Technologies Co., Ltd., Shenzhen 518129 Guangdong, China*; Email: yukang@zju.edu.cn

Authors

Gaoqi Weng − *Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058 Zhejiang, China*

Huifeng Zhao − *Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058 Zhejiang, China*

Dou Nie − *Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058 Zhejiang, China*

Haotian Zhang − *Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058 Zhejiang, China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jmedchem.3c02051

Author Contributions

[§]Equivalent authors.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (2021YFE0206400), the National Natural Science Foundation of China (22220102001, 92370130, 82373791), and the Fundamental Research Funds for the Central Universities (226-2022-00220).

ABBREVIATIONS USED

5-HTR, 5-hydroxytryptamine receptors; AAE, adversarial autoencoders; AR, adrenergic receptors; CDK, cyclin-dependent kinases; CharRNN, character-level recurrent neural networks; CNN, convolutional neural networks; DR, dopamine receptors; ECFP, extended-connectivity fingerprint; EGFR, epidermal growth factor receptors; FCD, Frechet Chemnet distance; GAN, generative adversarial networks; GPCR, G protein-coupled receptor; GRU, gate recurrent unit; InChI, international chemical identifier; IntDiv, internal diversity; JakA, Janus kinases; KL, Kullback−Leibler; LSTM, long short-term memory; ML, machine learning; MW, molecular weight; ORGAN, objective-reinforced generative adversarial networks; PDGFR, platelet-derived growth factor receptors; QED, quantitative estimation of drug-likeness; QSAR, quantitative structure−activity relationships; RL, reinforcement learning; RNN, recurrent neural networks; RNNAttn, attention-based recurrent models; SA, synthetic accessibility; SMILES, simplified molecular-input line-entry system; SNN, similarity to the nearest neighbor; VAE, variational autoencoders; VEGFR, vascular endothelial growth factor receptors; VS, virtual screening

REFERENCES

(1) van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf Model* **2019**, *59*, 644−651.

(2) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf Model* **2020**, *60*, 6065−6073.

(3) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373−D1380.

(4) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, María P.; Mosquera, Juan F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, Chris J.; Segura-Cabrera, A.; Hersey, A.; Leach, Andrew R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

(5) Weng, G.; Cai, X.; Cao, D.; Du, H.; Shen, C.; Deng, Y.; He, Q.; Yang, B.; Li, D.; Hou, T. PROTAC-DB 2.0: an updated database of PROTACs. *Nucleic Acids Res.* **2023**, *51*, D1367−D1372.

(6) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des* **2013**, *27*, 675−9.

(7) 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.

(8) Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci.* **2018**, *4*, 268−276.

(9) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(10) Godinez, W. J.; Ma, E. J.; Chao, A. T.; Pei, L.; Skewes-Cox, P.; Canham, S. M.; Jenkins, J. L.; Young, J. M.; Martin, E. J.; Guiguemde, W. A. Design of potent antimalarials with generative chemistry. *Nat. Mach Intell* **2022**, *4*, 180−186.

(11) Guo, J.; Fialková, V.; Arango, J. D.; Margreitter, C.; Janet, J. P.; Papadopoulos, K.; Engkvist, O.; Patronov, A. Improving de novo molecular design with curriculum learning. *Nat. Mach Intell* **2022**, *4*, 555−563.

(12) Arus-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J. L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **2019**, *11*, 71.

(13) Blaschke, T.; Arus-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf Model* **2020**, *60*, 5918−5922.

(14) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf Model* **2019**, *59*, 43−52.

(15) Wang, J.; Hsieh, C.-Y.; Wang, M.; Wang, X.; Wu, Z.; Jiang, D.; Liao, B.; Zhang, X.; Yang, B.; He, Q.; Cao, D.; Chen, X.; Hou, T. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nat. Mach Intell* **2021**, *3*, 914−922.

(16) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48.

(17) Krishnan, S. R.; Bung, N.; Bulusu, G.; Roy, A. Accelerating De Novo Drug Design against Novel Proteins Using Deep Learning. *J. Chem. Inf Model* **2021**, *61*, 621−630.

(18) Prykhodko, O.; Johansson, S. V.; Kotsias, P. C.; Arus-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*, 74.

(19) Guimaraes, G.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv* **2017**, DOI: 10.48550/arXiv.1705.10843.

(20) Shi, C.; Xu, M.; Zhu, Z.; Zhang, W.; Zhang, M.; Tang, J. GraphAF: a flow-based autoregressive model for molecular graph generation. *arXiv* **2020**, DOI: 10.48550/arXiv.2001.09382.

(21) Schneuing, A.; Du, Y.; Harris, C.; Jamasb, A.; Igashov, I.; Du, W.; Blundell, T.; Lió, P.; Gomes, C.; Welling, M. Structure-based drug design with equivariant diffusion models. *arXiv* **2022**, DOI: 10.48550/arXiv.2210.13695.

(22) Igashov, I.; Stärk, H.; Vignac, C.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv* **2022**, DOI: 10.48550/arXiv.2210.05274.

(23) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov Today Technol.* **2019**, *32*−*33*, 55−63.

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

(25) 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*, No. 565644.

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

(27) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* **2019**, *14*, No. e0220113.

(28) Yang, J.; Shen, C.; Huang, N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Front. Pharmacol.* **2020**, *11*, 69.

(29) Wang, M.; Sun, H.; Wang, J.; Pang, J.; Chai, X.; Xu, L.; Li, H.; Cao, D.; Hou, T. Comprehensive assessment of deep generative architectures for de novo drug design. *Brief. Bioinform.* **2022**, *23* (1), No. bbab544.

(30) Dollar, O.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Attention-based generative models for de novo molecular design. *Chem. Sci.* **2021**, *12*, 8362−8372.

(31) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, 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*, D930−D940.

(32) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−53.

(33) Pándy-Szekeres, G.; Caroli, J.; Mamyrbekov, A.; Kermani, A. A.; Keserű, György M.; Kooistra, Albert J.; Gloriam, D. E. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Res.* **2022**, *51*, D395−D402.

(34) Sun, J.; Jeliazkova, N.; Chupakin, V.; Golib-Dzib, J. F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminform.* **2017**, *9*, 17.

(35) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf Model* **2010**, *50*, 742−54.