Structure-based Drug Design Benchmark: Do 3D Methods Really Dominate?

Anonymous Authors[1]

Abstract

Currently, the field of structure-based drug design is dominated by three main types of algorithms: search-based algorithms, deep generative models, and reinforcement learning. While existing works have typically focused on comparing models within a single algorithmic category, cross-algorithm comparisons remain scarce. In this paper, to fill the gap, we establish a benchmark to evaluate the performance of sixteen models across these different algorithmic foundations by assessing the pharmaceutical properties of the generated molecules and their docking affinities with specified target proteins. We highlight the unique advantages of each algorithmic approach and offer recommendations for the design of future SBDD models. We emphasize that 1D/2D ligand-centric drug design methods can be used in SBDD by treating the docking function as a black-box oracle, which is typically neglected. The empirical results show that 1D/2D methods achieve competitive performance compared with 3D-based methods that use the 3D structure of the target protein explicitly. Also, AutoGrow4, a 2D molecular graph-based genetic algorithm, dominates SBDD in terms of optimization ability.

1. Introduction

Novel types of safe and effective drugs are needed to meet the medical needs of billions worldwide and improve the quality of human life. The process of discovering a new drug candidate and developing it into an approved drug for clinical use is known as *drug discovery* (Sinha & Vohora, 2018). This complex process is fundamental to the development of new therapies that can manage, cure, or alleviate the symptoms of various health conditions.

Structure-based drug design (SBDD) (Bohacek et al., 1996) represents a core strategy within the drug discovery process, which utilizes the three-dimensional (3D) structures of proteins associated with diseases to develop drug candidates, serving as a fundamental method to expedite the drug discovery process through physical simulation and data-driven modeling. Based on the lock and key model (Tripathi & Bankaitis, 2017), molecules that bind more effectively to a disease target tend to inhibit their abnormal activity or modulate their function in a way that contributes to disease treatment, a phenomenon that has been confirmed through experimental studies (Honarparvar et al., 2014; Blundell, 1996; Lu et al., 2022).

Currently, three main algorithmic approaches dominate the drug design field (Brown et al., 2019; Gao et al., 2022; Du et al., 2022): search-based algorithms like genetic algorithms (GA) (Jensen, 2019; Spiegel & Durrant, 2020; Tripp & Hernández-Lobato, 2023; Fu et al., 2022a), deep generative models (a.k.a. generative model) like variational autoencoder (VAE) (Gómez-Bombarelli et al., 2018) and autoregressive models (Luo et al., 2021; Peng et al., 2022; Zhang et al., 2023), and reinforcement learning (RL) models (Olivecrona et al., 2017; Zhou et al., 2018). Also, there is a trend that represents the target protein in 3D format (Zhang et al., 2023; Luo et al., 2021; Fu et al., 2022a; Peng et al., 2022). These models are often regarded as state-of-the-art due to the high validity, diversity, and synthesizability of their generated molecules. However, comparisons among these models remain unclear for several reasons. Firstly, current benchmarks or survey papers tend to compare models within the same algorithmic category, with a particular focus on deep generative models (Du et al., 2022). Secondly, most existing benchmarks emphasize the properties of the molecules themselves, neglecting the evaluation of protein-ligand interactions, which are crucial for real-world applications (Brown et al., 2019; Gao et al., 2022).

To fill this blank, this paper curates a comprehensive benchmark that encompasses sixteen models spanning all three algorithmic approaches. We assess their generated molecules not only through typical heuristic molecular property oracles but also by evaluating docking scores that reflect the quality of interactions between molecules and target proteins (associated to disease). Our analysis of the top-1/10/50/100 scores from each oracle reveals that search-based algorithms, particularly genetic algorithms, generally outperform others. Also, explicit utilization of 3D structure of the target protein has not shown significant improvement compared to 2D methods. While there are some drawbacks in certain aspects, these could potentially be mitigated by integrating other algorithmic strategies.

Table 1. Representative structure-based drug design methods, categorized based on the molecular assembly strategies and the optimization algorithms. Columns are various molecular assembly strategies while rows are different optimization algorithms.

	1D SMILES/SELFIES	2D molecular graph	3D structure-based
Genetic Algorithm (GA)	SMILES-GA (Yoshikawa et al., 2018)	AutoGrow4 (Spiegel & Durrant, 2020), graph GA (Jensen, 2019)	-
Hill Climbing	SMILES-LSMT-HC (Brown et al., 2019)	MIMOSA (Fu et al., 2021)	-
Reinforcement Learning (RL)	REINVENT (Olivecrona et al., 2017)	MolDQN (Zhou et al., 2018)	-
Gradient Ascent (GRAD)	Pasithea (Shen et al., 2021)	DST (Fu et al., 2022b)	
Generative Models	SMILES/SELFIES-VAE-BO (Gómez-Bombarelli et al., 2018)	-	3DSBDD(Luo et al., 2021), Pocket2mol(Peng et al., 2022), PocketFlow(Jiang et al., 2024), ResGen(Zhang et al., 2023)

2. Related Work

There has been significant progress in benchmarking efforts for drug design evaluation (Brown et al., 2019; Tripp et al., 2021; Huang et al., 2021; Gao et al., 2022; Polykovskiy et al., 2020; Harris et al., 2023). Specifically, Guacamol (Brown et al., 2019) encompasses five molecule design algorithms, develops twenty novel objective functions, and assesses their performance comprehensively. Molecular Sets (MOSES) (Polykovskiy et al., 2020) concentrates on five generative-based models (recurrent neural network (RNN), Adversarial Auto-Encoder (AAE) (Makhzani et al., 2015), Variational Auto-Encoder (VAE) (Gómez-Bombarelli et al., 2018)), introducing eleven oracles that primarily evaluate the novelty and uniqueness of the generated molecules. Practical Molecule Optimization (PMO) (Gao et al., 2022) offers a benchmark for twenty-five molecule design models across twenty-three objectives, providing a broad evaluation landscape. POSECHECK (Harris et al., 2023) assesses five generative-based models by employing four physical oracles to gauge the quality of protein-ligand interactions, contributing to the nuanced understanding of model efficacy in simulating realistic biochemical interactions. Recently, (Tripp & Hernández-Lobato, 2023) designed a simple genetic algorithm on molecules based on (Jensen, 2019) and compared it with several other molecule generation algorithms. The results show that genetic algorithms perform at least as well as many more complicated methods in the unconditional molecule generation task.

1D Molecule Design Methods 1D molecule design methods use Simplified Molecular-Input Line-Entry System (SMILES) (Weininger, 1988) or SELF-referencing Embedded Strings (SELFIES) (Krenn et al., 2020) strings as the representation of molecules. Most 1D methods produce molecule strings in an autoregressive manner. In this paper, we discuss several methods that were developed to produce molecule strings, either SMILES or SELFIES strings, including REINVENT (Olivecrona et al., 2017), SMILES and SELFIES VAE (Gómez-Bombarelli et al., 2018), SMILES GA (Yoshikawa et al., 2018), SMILES-LSTM-HC (Brown et al., 2019), and Pasithea (Shen et al., 2021). Although SELFIES string has the advantage of enforcing chemical validity rules compared to SMILES, through thorough empirical studies, (Gao et al., 2022) showed that SELFIES string-based methods do not demonstrate superiority over SMILES string-based ones.

2D Molecule Design Methods Compared to 1D molecule design methods, representing molecules using 2D molecular graphs is a more sophisticated approach. molecular 2D representation, graphs are used to depict molecules, where edges represent chemical bonds and nodes represent atoms. There are two main strategies for constructing these graphs: atom-based and fragment-based. Atom-based methods operate on one atom or bond at a time, searching the entire chemical space. On the other hand, fragment-based methods summarize common molecular fragments and operate on one fragment at a time, which can be more efficient. In this paper, we discuss several methods belonging to this category: MolDQN (Zhou et al., 2018), which uses an atom-based strategy, and Graph GA (Jiang et al., 2024), Multi-constraint Molecule Sampling (MIMOSA) (Fu et al., 2021), Differentiable Scaffolding Tree (DST) (Fu et al., 2022b), and AutoGrow4 (Spiegel & Durrant, 2020), which use fragment-based strategies.

3D Molecule Design Methods Both 1D and 2D molecule design methods are ligand-centric, focusing primarily on designing the molecule itself. In structure-based drug design, as pointed out in (Huang et al., 2021), these models take the

docking function as a black box, which inputs a molecule and outputs the binding affinity score. However, these models fail to incorporate target protein structure information and consequently suffer from high computational time (to find binding pose). In contrast, 3D structure-based drug design methods take the three-dimensional geometry of the target protein as input and directly generate pocket-aware molecules in the pocket of target protein. In this paper, we cover four cutting-edge structure-based drug design methods: PocketFlow (Jiang et al., 2024), 3DSBDD (Luo et al., 2021), Pocket2mol (Peng et al., 2022), and ResGen (Zhang et al., 2023).

3. Models

In this paper, the models we select for evaluation are based on one or a combination of the following algorithms. For ease of comparison, we categorize all the methods based on optimization algorithm and molecular assembly strategy in Table 1.

Screening: Screening (high-throughput screening) is a traditional drug design approach that searches over a library of molecules. However, it is only able to search the known drug molecular space, but is not able to explore unknown chemical space and identify novel/unknown molecules. The known chemical space ($< 10^{15}$) is only taking a tiny fraction of the whole drug-like molecular space (around 10^{60}) (Bohacek et al., 1996). In our evaluation, we use screening as a baseline method, which randomly searches ZINC 250k library (Irwin et al., 2012).

Genetic Algorithm (GA): Inspired by natural selection, genetic algorithm is a combinatorial optimization method that evolves solutions to problems over many generations. Specifically, in each generation, GA will perform crossover and mutation over a set of candidates to produce a pool of offspring and keep the top-k offspring for the next generation, imitating the natural selection process. In our evaluation, we choose three GA models: SMILES GA (Yoshikawa et al., 2018) that performs GA over SMILES string-based space, Graph GA (Jiang et al., 2024) that searches over atom- and fragment-level by designing their crossover and mutation rules on graph matching and AutoGrow4 (Spiegel & Durrant, 2020) which introduce another procedure called elitism that filtering the candidates by pre-defined rules.

Variational Auto-Encoder (VAE): The aim of variational autoencoder is to generate new data that is similar to training data. In the molecule generation area, VAE learns a bidirectional map between molecule space and continuous latent space and optimizes the latent space. VAE itself generated diverse molecules that are learned from the training set. After training VAE, Bayesian optimization (BO) is used to navigate latent space efficiently, identify desirable molecules, and conduct molecule optimization. In our evaluation, we select two VAE-based models: SMILES-VAE-BO (Gómez-Bombarelli et al., 2018) uses SMILES string as the input to the VAE model, and SELFIES-VAE-BO uses the same algorithm but uses SELFIES string as the molecular representation.

Auto-regressive: An auto-regressive model is a type of statistical model that is based on the idea that past values in the series can be used to predict future values. In molecule generation, an auto-regressive model would typically take the generated atom sequence as input and predict which atom would be the next. In our evaluation, we choose four auto-regressive models: PocketFlow (Jiang et al., 2024) is an autoregressive flow-based generative model. 3DSBDD (Luo et al., 2021) based on conventional Markov Chain Monte Carlo (MCMC) algorithms and Pocket2mol (Peng et al., 2022) choose graph neural networks (GNN) as the backbone. Inspired by Pocket2mol, ResGen (Zhang et al., 2023) used a hierarchical autoregression, which consists of a global autoregression for learning protein-ligand interactions and atomic component autoregression for learning each atom's topology and geometry distributions. Also, note that these models use 3D representations of target proteins.

Hill Climbing (HC): Hill Climbing (HC) is an optimization algorithm that belongs to the family of local search techniques (Selman & Gomes, 2006). It is used to find the best solution to a problem among a set of possible solutions. In molecular design, Hill Climbing would tune the generative model with the reference of generated high-scored molecules. In our evaluation, we adopt two HC models: SMILES-LSTM-HC (Brown et al., 2019) uses an LSTM model to generate molecules and uses the HC technique to fine-tune it. MultI-constraint MOlecule SAmpling (MIMOSA) (Fu et al., 2021) uses a graph neural network instead and incorporates it with HC.

Gradient Ascent (GRAD): Similar to gradient descent, gradient ascent also estimates the gradient direction but chooses the maximum direction. In molecular design, the GRAD method is often used in molecular property function to optimize molecular generation. In our evaluation, we choose two GRAD-based models: Pasithea (Shen et al., 2021) uses SELFIES as input and applies GRAD on an MLP-based molecular property prediction model. Differentiable Scaffolding Tree (DST) (Fu et al., 2022b) uses differentiable molecular graph as input and uses a graph neural network to estimate objective and the corresponding gradient.

Reinforcement Learning (RL): In molecular generation context, a reinforcement learning model would take a partially-generated molecule (either sequence or molecular graph) as state; action is how to add a token or atom to the sequence or molecular graph respectively; and reward is the property score of current molecular sequence. In our evaluation, we test on two RL-based models: REINVENT (Olivecrona et al., 2017) is a policy-gradient method that uses RNN to generate molecules and MolDQN (Zhou

et al., 2018) uses a deep Q-network to generate molecular graph.
All the methods used in this paper are summarized in Table 1 for ease of comparison.

4. Experiments

In this section, we demonstrate the experimental results. We start with the description of experimental setup. Then, we present and analyze the experimental results, including protein-ligand bindings, pharmaceutical properties of generated molecules (e.g., drug-likeness and synthetic accessibility), and other qualities of generated molecules (e.g., diversity, validity). The relevant code is available in https://github.com/zkysfls/2024-sbdd-benchmark.

4.1. Experimental Setup

4.1.1. Oracle

In drug discovery, we need to evaluate the pharmaceutical properties of the generated molecules, such as binding affinity to certain target proteins, drug-likeness, synthetic accessibility, and solubility, etc. These property evaluators are also known as *oracle*. In this section, we introduce the oracle we chose to evaluate these models. All our oracle functions come from Therapeutic Data Commons (TDC) (Huang et al., 2022; 2021)[1].

Docking Score: Molecular docking is a measurement of free energy exchange between a ligand and a target protein during the binding process. A lower docking score means the ligand would have a higher potential to pose higher bioactivity with a given target. Compared with other heuristic oracles, such as QED (quantitative estimate of drug-likeness), and LogP (Octanol-water partition coefficient), docking reflects the binding affinities between drug molecule and target (Graff et al., 2021). Our experiments use **TDC.Docking** oracle function, which is based on AutoDock Vina (Eberhardt et al., 2021) to test with these models. We chose seven representative and diverse target proteins in the TDC docking benchmark, which are selected from CrossDock (Francoeur et al., 2020b). The PDBIDs are 1iep, 3eml, 3ny8, 4rlu, 4unn, 5mo4, 7l11. These crystallography structures are across different fields, including virology, immunology, and oncology (Huang et al., 2022; 2021; Lu et al., 2019). They cover various kinds of diseases such as chronic myelogenous leukemia, tuberculosis, SARS-COVID-2, etc. They represent a breadth of functionality, from viral replication mechanisms to cellular signaling pathways and immune responses.

Heuristic Oracles: Although heuristic oracles are considered to be "trivial" and too easily optimized, we still incorporate some of them into our evaluation metrics for comprehensive analysis. In our experiments, we utilize Quantitative Estimate of Drug-likeness (QED), SA, and LogP as our heuristic oracles. QED evaluates a molecule's drug-likeness on a scale from 0 to 1, where 0 indicates minimal drug-likeness and 1 signifies maximum drug-likeness, aligning closely with the physicochemical properties of successful drugs. SA, or Synthetic Accessibility, assesses the ease of synthesizing a molecule, with scores ranging from 1 to 10; a lower score suggests easier synthesis. LogP measures a compound's preference for a lipophilic (oil-like) phase over a hydrophilic (water-like) phase, essentially indicating its solubility in water, where the optimal range depends on the type of drug. But mostly the value should be between 0 and 5 (Krenn et al., 2020).
Molecule Generation Oracles: While docking score oracles and heuristic oracles focus on evaluating individual molecules, molecule generation oracles assess the quality of all generated molecules as a whole. In our experiments, we choose three metrics to evaluate the generated molecules of each model: diversity, validity, and uniqueness. Diversity is measured by the average pairwise Tanimoto distance between the Morgan fingerprints (Benhenda, 2017). Validity is determined by checking atoms' valency and the consistency of bonds in aromatic rings using RDKit's molecular structure parser (Polykovskiy et al., 2020). Uniqueness is measured by the frequency at which a model generates duplicated molecules, with lower values indicating more frequent duplicates (Polykovskiy et al., 2020).

4.1.2. Model Setup

For each model, we generate 1,000 molecules for each given target protein, and each molecule is evaluated by the TDC oracle functions (Huang et al., 2021; 2022). Each experiment is run on one OSC Ascend node (Center, 1987) for 96 hours, which is the maximum time allowed for a single experiment, and we only run each model once. Five models (3DSBDD, PocketFlow, Pocket2mol, ResGen, and AutoGrow4) first generate a certain number of molecules within the given time, and then we run oracle functions on each molecule. All the other models come from the PMO benchmark (Gao et al., 2022), and our experiment follows its setting, where each molecule is first generated, then the oracle function is used to calculate the score, and then the model moves on to generate the second molecule. None of the tested models have prior knowledge of these oracle functions. Among all the models, five of them manage to generate and have been evaluated by the oracle for 1,000 or more molecules within the given time across all target proteins (AutoGrow4, PocketFlow, DST, MIMOSA, and Screening); other models do not generate enough molecules or have not been evaluated for enough molecules within the given 96 hours, mostly because the docking oracle function is time-consuming. The

[1] https://tdcommons.ai/functions/oracles/

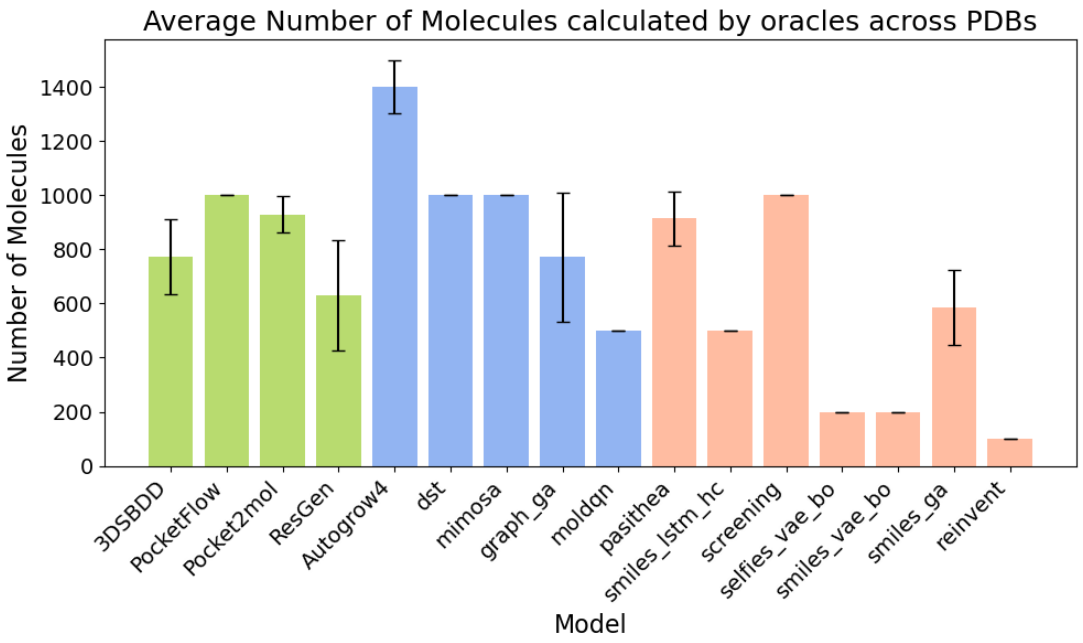

Figure 1. The bar chart of average generated molecules that are calculated by our selected oracles for each model across all target proteins under given time. 1D methods are colored red, blue is used to indicate 2D methods, and green represents 3D methods.

VAE models and REINVENT only generate 200 and 100 molecules, respectively, because we observe that when they generate more than this number of molecules, the models crash. Figure 1 shows the average number of molecules that has been generated and successfully evaluated by oracle functions within experiment time.

4.2. Experimental Results

4.2.1. Results of Binding Affinities

Overall Performance: Overall, search-based algorithms (including Screening, Genetic Algorithm [GA], Hill Climbing [HC], and Gradient-based methods [GRAD]) demonstrate superior performance compared to generative models (like VAE and Auto-regressive) and reinforcement learning-based algorithms. Although generative models have good performance on Top-1 docking score (Table 3), search-based algorithms take advantages on Top-10/50/100 docking score, as shown in Table 2 to 5.

Search-based Algorithms: Among all the search-based algorithms, AutoGrow4 exhibits the best performance. This superiority is not only reflected in its consistently highest scores in the Top 1/10/50/100 categories but also in the outstanding docking scores of the majority of its generated molecules compared to other methods across every target protein, as indicated by Table 3 to 5. We believe that the elitism procedure incorporated in AutoGrow4 enhances its performance by providing better candidates for crossover and mutation. While other search-based methods also perform well overall, no single algorithm category within this group distinctly outperforms the others. It appears that neither the format of the input (such as SMILES/SELFIES or scaffold graph) nor the specific algorithm employed has a significant impact on performance.

Generative Models: In our experiments, we evaluated two categories of generative models: Variational Autoencoders (VAEs) and Auto-regressive models. Firstly, two VAE-based models, SMILES-VAE-BO and SELFIES-VAE-BO, demonstrated consistent performance across all target proteins, with most of their generated molecules achieving docking scores between -6 and -9. However, only a few molecules from these models exceeded a -10 score, and neither model showed a distinct advantage over the other.

Regarding the autoregressive models, among the four models we evaluated, Pocket2Mol demonstrates the best overall performance. This is not only because it achieves the highest scores in Top-1, Top-10, Top-50, and Top-100 rankings but also because its average top docking score remains above -10 across all these rankings. For the remaining models, ResGen ranks second, followed by PocketFlow and 3DSBDD. We also noticed that some of our selected target pockets appeared in Crossdeck (Francoeur et al., 2020a), which is the dataset that our selected autoregressive models used for training and evaluation. However, our experimental results show that these autoregressive models do not always have advantages compared to other models. For example, in target 3EML, AutoGrow4 has the best performance across Top

Table 2. The average of each model's Top-10 Docking score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU	4UNN	5MO4	7L11
3DSBDD	-9.05 ± 0.38	-10.02 ± 0.15	-10.10 ± 0.24	-9.80 ± 0.55	-8.23 ± 0.30	-8.71 ± 0.45	-8.47 ± 0.18
AUTOGROW4	**-13.23** ± 0.11	**-13.03** ± 0.09	-11.70 ± 0.00	-11.20 ± 0.00	-11.14 ± 0.12	**-10.38** ± 0.27	-8.84 ± 0.33
POCKET2MOL	-10.17 ± 0.53	-12.25 ± 0.27	**-11.89** ± 0.16	-10.57 ± 0.12	**-12.20** ± 0.34	-10.07 ± 0.62	**-9.74** ± 0.38
POCKETFLOW	-12.49 ± 0.70	-9.25 ± 0.29	-8.56 ± 0.35	-9.65 ± 0.25	-7.90 ± 0.78	-7.80 ± 0.42	-8.35 ± 0.31
RESGEN	-10.97 ± 0.29	-9.25 ± 0.95	-10.96 ± 0.42	**-11.75** ± 0.42	-9.41 ± 0.23	-10.34 ± 0.39	-8.74 ± 0.24
DST	-10.95 ± 0.57	-10.67 ± 0.24	-10.54 ± 0.22	-10.88 ± 0.37	-9.71 ± 0.19	-10.03 ± 0.36	-8.33 ± 0.41
GRAPH GA	-10.03 ± 0.41	-9.89 ± 0.25	-9.94 ± 0.15	-10.22 ± 0.39	-9.32 ± 0.51	-9.29 ± 0.20	-7.75 ± 0.32
MIMOSA	-10.96 ± 0.57	-10.69 ± 0.24	-10.51 ± 0.23	-10.81 ± 0.39	-9.66 ± 0.25	-10.02 ± 0.36	-8.33 ± 0.41
MOLDQN	-6.73 ± 0.12	-6.51 ± 0.15	-7.09 ± 0.16	-6.79 ± 0.26	-5.92 ± 0.26	-6.27 ± 0.10	-6.87 ± 0.20
PASITHEA	-10.86 ± 0.29	-10.31 ± 0.09	-10.69 ± 0.27	-10.92 ± 0.35	-9.69 ± 0.32	-9.77 ± 0.21	-8.06 ± 0.22
REINVENT	-9.87 ± 0.31	-9.48 ± 0.39	-9.61 ± 0.36	-9.69 ± 0.29	-8.70 ± 0.25	-8.92 ± 0.38	-7.25 ± 0.21
SCREENING	-10.86 ± 0.26	-10.90 ± 0.54	-10.73 ± 0.45	-10.86 ± 0.22	-9.80 ± 0.23	-9.91 ± 0.30	-8.15 ± 0.26
SELFIES-VAE-BO	-10.15 ± 0.60	-9.76 ± 0.12	-9.99 ± 0.28	-10.00 ± 0.23	-9.02 ± 0.33	-9.18 ± 0.39	-7.75 ± 0.22
SMILES GA	-9.56 ± 0.17	-9.56 ± 0.37	-10.00 ± 0.26	-9.61 ± 0.19	-8.80 ± 0.20	-9.21 ± 0.23	-7.54 ± 0.32
SMILES LSTM HC	-10.38 ± 0.21	-10.30 ± 0.15	-10.19 ± 0.12	-10.49 ± 0.49	-9.36 ± 0.17	-9.71 ± 0.43	-7.90 ± 0.26
SMILES-VAE-BO	-9.93 ± 0.22	-9.78 ± 0.10	-9.96 ± 0.29	-10.05 ± 0.20	-9.03 ± 0.30	-9.18 ± 0.39	-7.74 ± 0.25

1 to Top 100, while in target 4UNN, Pocket2mol has the advantage.

Reinforcement Learning: We incorporate two reinforcement learning-based models: MolDQN and REINVENT. Overall, REINVENT demonstrates superior performance compared to MolDQN. The majority of molecules generated by REINVENT have docking scores around -8, whereas those by MolDQN are mostly around -6. This leads us to suspect that the policy-gradient method might be more suitable than the deep Q-network approach for the task of molecular generation. Compared to other models, reinforcement learning models do not exhibit good performance. These two reinforcement learning models have the lowest scores in the docking score oracle. This may suggest that the reinforcement learning algorithm may not have a strong ability to produce molecules with good docking scores.

4.2.2. RESULTS OF PHARMACEUTICAL PROPERTIES

Then, we report and analyze the pharmaceutical properties of the generated molecules.

SA: Overall, most of the models generate molecules with scores between 1 to 3. Notably, 3DSBDD, MolDQN, and REINVENT produce molecules with scores ranging above 3. Additionally, these models exhibit high variance in their scores. For instance, in the case of 3DSBDD, the lowest score observed is 1, yet it can also generate molecules scoring as high as 8 for a specific target protein.

QED: Most of the models generate molecules with scores between 0.8. However, 3DSBDD and REINVENT tend to produce molecules with scores primarily in the range of 0.6 to 0.7, while MolDQN's generated molecules hover around 0.4. Overall nearly all models have the same level of performances except reinforcement learning based models which has worse performances.

LogP: Overall, nearly all the models tested produced the majority of their molecules within the 0 to 3 range, which is deemed suitable for a drug, with the exception of MolDQN. The molecules generated by MolDQN often have a LogP score of less than 0 across all target proteins, indicating a high solubility in water. Furthermore, generative models, particularly 3DSBDD, predominantly generate molecules with scores around 0.

4.2.3. MOLECULE GENERATION QUALITY

In the diversity oracle, all models score above 0.8, with one model from each algorithm category exceeding 0.9: PocketFlow from generative models, Graph GA from search-based methods, and MolDQN from reinforcement learning. This suggests that these three models are particularly effective at generating diverse sets of molecules. In the validity oracle, all models achieve a perfect score of 1, except for 3DSBDD. Similarly, in the uniqueness oracle, all models score 1, except for 3DSBDD, AutoGrow4, and PocketFlow. It is unclear why these models have lower scores in validity and uniqueness, especially when other models from the same algorithm category perform well. One possible explanation for 3DSBDD's low validity and uniqueness scores could be issues with its molecule generation process, such as producing invalid molecular structures or duplicates. Despite their high diversity scores, AutoGrow4 and PocketFlow's lower uniqueness scores might indicate a tendency to generate similar molecules. Further investigation into the specific architectures and training procedures of these models could provide insights into their divergent performance. It may also be valuable to analyze the trade-offs between diversity, validity, and uniqueness in molecule generation and how different models balance these objectives. Overall, while most models demonstrate strong performance across the three oracles, the lower validity and uniqueness scores of 3DSBDD, AutoGrow4, and PocketFlow highlight the importance of evaluating multiple aspects of generated molecules to assess model performance comprehensively.

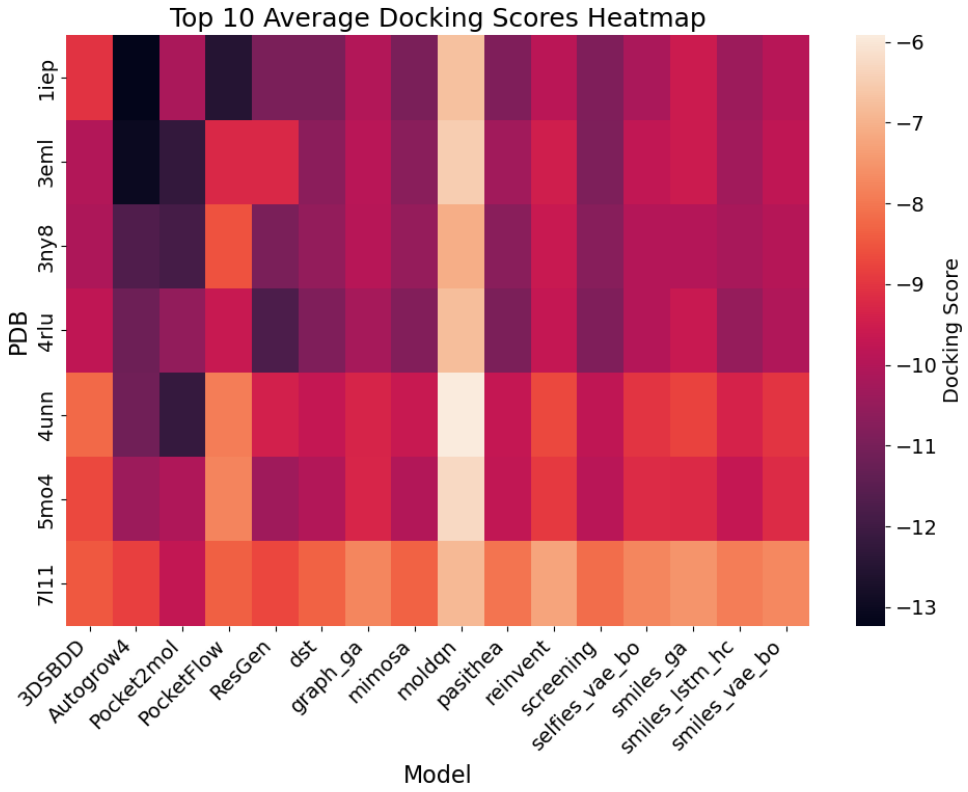

Figure 2. The heatmap based on the average of each model's Top-10 docking score for each target protein.

4.2.4. Key observations

We summarize the following insightful observations drawn from the experimental results, which benefits design of future SBDD models.

- Most structure-based drug design method uses the 3D structure of the target protein explicitly and grow the drug molecules in the pocket of the target protein. We pinpoint another direction that regards the docking function as a black box and uses 1D/2D ligand-centric methods to produce drug molecules, which is usually neglected by the community. In this paper, we empirically prove that this kind of method would achieve superior performance.
- AutoGrow4, a 2D genetic algorithm, exhibits the best optimization performance in terms of top-K docking scores in most target proteins. Also, it owns desirable synthetic accessibility.
- Generally, 3D SBDD algorithms (using 3D target protein structure explicitly) do not demonstrate significant superiority over 2D methods.
- No methods can dominate structure-based drug design in all the evaluation metrics (docking score, SA, QED, diversity, validity, and uniqueness), as shown in Figure 4.

Example of the Generated Molecules Also, we show the 3D poses of molecules that have the best docking score for each target protein in Figure 12 in Appendix. We find that the generated molecules could bind tightly to the pocket of target proteins.

Additional Experimental Results Furthermore, we report the numerical values of top-K docking, QED, SA, LogP score for all the methods across different target proteins in Appendix, as well as their relative ranking on all the metrics.

5. Conclusion and Future Work

Currently, the landscape of structure-based drug design models is vast, featuring various algorithmic backbones, yet comparative analyses across them are scarce. In this study, we design experiments to evaluate the quality of molecules generated by each model. Our experiments extend beyond conventional heuristic oracles related to molecular prop-

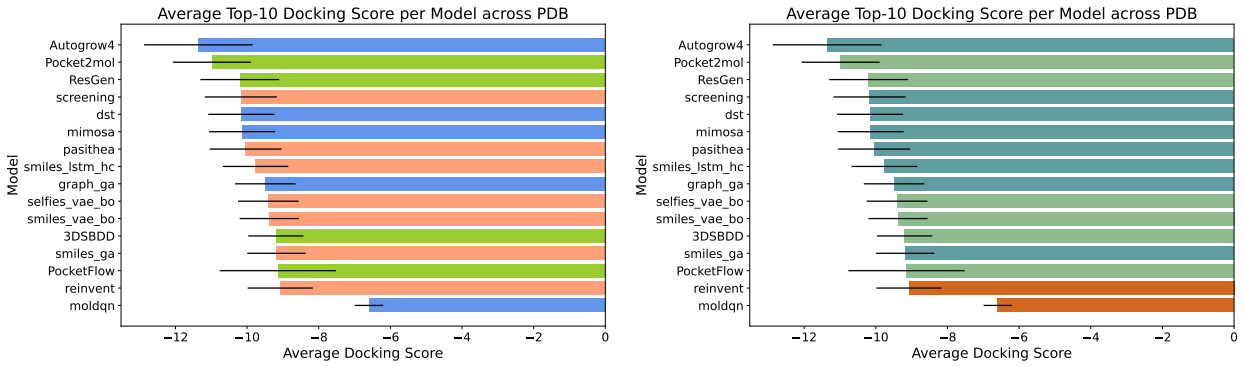

Figure 3. The bar chart is based on the average of each model's Top-10 docking score. On the left, 1D methods are colored red, blue is used to indicate 2D methods, and green represents 3D methods. On the right, generative models are in blue, search-based methods are in green, and reinforcement learning methods are in red.

erties, also examining the affinity between molecules and selected target proteins. Our findings indicate that models based on genetic algorithms exhibit a higher potential for producing molecules that dock effectively with given target proteins. Also, representing target molecules in 3D format does not significantly improve both the molecular quality and blinding affinity. Although we observed that there is no single method that could excel both our two metrics, we suggest that when developing new structure-based drug discovery models in the future, it would be advantageous to integrate genetic algorithms with other computational approaches to enhance both docking scores and molecular properties.

References

Benhenda, M. Chemgan challenge for drug discovery: can ai reproduce natural chemical diversity?, 2017.

Blundell, T. L. Structure-based drug design. *Nature*, 384 (6604):23, 1996.

Bohacek, R. S., McMartin, C., and Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 16(1): 3–50, 1996.

Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, March 2019. ISSN 1549-960X. doi: 10.1021/acs.jcim.8b00839. URL http://dx.doi.org/10.1021/acs.jcim.8b00839.

Center, O. S. Ohio supercomputer center, 1987. URL http://osc.edu/ark:/19495/f5s1ph73.

Du, Y., Fu, T., Sun, J., and Liu, S. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*, 2022.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. doi: 10.1021/acs.jcim.1c00203. URL https://doi.org/10.1021/acs.jcim.1c00203. PMID: 34278794.

Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., and Koes, D. R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, 2020a. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411. PMID: 32865404.

Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., and Koes, D. R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020b.

Fu, T., Xiao, C., Li, X., Glass, L. M., and Sun, J. Mimosa: Multi-constraint molecule sampling for molecule optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 125–133, 2021.

Fu, T., Gao, W., Coley, C. W., and Sun, J. Reinforced genetic algorithm for structure-based drug design, 2022a.

Fu, T., Gao, W., Xiao, C., Yasonik, J., Coley, C. W., and Sun, J. Differentiable scaffolding tree for molecular optimization, 2022b.

Gao, W., Fu, T., Sun, J., and Coley, C. W. Sample efficiency matters: A benchmark for practical molecular optimization, 2022.

Graff, D. E., Shakhnovich, E. I., and Coley, C. W. Accelerating high-throughput virtual screening through

molecular pool-based active learning. *Chemical Science*, 12(22):7866–7881, 2021. ISSN 2041-6539. doi: 10.1039/d0sc06805e. URL http://dx.doi.org/10.1039/d0sc06805e.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, January 2018. ISSN 2374-7951. doi: 10.1021/acscentsci.7b00572. URL http://dx.doi.org/10.1021/acscentsci.7b00572.

Harris, C., Didi, K., Jamasb, A. R., Joshi, C. K., Mathis, S. V., Lio, P., and Blundell, T. Benchmarking generated poses: How rational is structure-based drug design with generative models?, 2023.

Honarparvar, B., Govender, T., Maguire, G. E., Soliman, M. E., and Kruger, H. G. Integrated approach to structure-based enzymatic drug design: molecular modeling, spectroscopy, and experimental bioactivity. *Chemical reviews*, 114(1):493–537, 2014.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 2022.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012. doi: 10.1021/ci3001277. URL https://doi.org/10.1021/ci3001277. PMID: 22587354.

Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

Jiang, Y., Zhang, G., You, J., Zhang, H., Yao, R., Xie, H., Zhang, L., Xia, Z., Dai, M., Wu, Y., Li, L., and Yang, S. Pocketflow is a data-and-knowledge-driven structure-based molecular generative model. *Nature Machine Intelligence*, 6:1–12, 03 2024. doi: 10.1038/s42256-024-00808-8.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100*Machine Learning: Science and Technology*, 1(4): 045024, October 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/aba947. URL http://dx.doi.org/10.1088/2632-2153/aba947.

Lu, Y., Chang, Y.-T., Hoffman, E. P., Yu, G., Herrington, D. M., Clarke, R., Wu, C.-T., Chen, L., and Wang, Y. Integrated identification of disease specific pathways using multi-omics data. *bioRxiv*, pp. 666065, 2019.

Lu, Y., Wu, C.-T., Parker, S. J., Cheng, Z., Saylor, G., Van Eyk, J. E., Yu, G., Clarke, R., Herrington, D. M., and Wang, Y. Cot: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1):vbac037, 2022.

Luo, S., Guan, J., Ma, J., and Peng, J. A 3d generative model for structure-based drug design. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9, 09 2017. doi: 10.1186/s13321-017-0235-x.

Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., and Ma, J. Pocket2mol: Efficient molecular sampling based on 3d protein pockets, 2022.

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., and Zhavoronkov, A. Molecular sets (moses): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11, 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.565644. URL https://www.frontiersin.org/articles/10.3389/fphar.2020.565644.

Selman, B. and Gomes, C. P. Hill-climbing search. *Encyclopedia of cognitive science*, 81:82, 2006.

Shen, C., Krenn, M., Eppel, S., and Aspuru-Guzik, A. Deep molecular dreaming: inverse machine learning for de-novo molecular design and interpretability with surjective representations. *Machine Learning: Science and Technology*, 2(3):03LT02, July 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/ac09d6. URL http://dx.doi.org/10.1088/2632-2153/ac09d6.

Sinha, S. and Vohora, D. Drug discovery and development: An overview. *Pharmaceutical medicine and translational clinical research*, pp. 19–32, 2018.

Spiegel, J. O. and Durrant, J. D. Autogrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Journal of Cheminformatics*, 12(1):25, 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00429-4. URL https://doi.org/10.1186/s13321-020-00429-4.

Tripathi, A. and Bankaitis, V. A. Molecular docking: From lock and key to combination lock. *Journal of molecular medicine and clinical applications*, 2(1), 2017.

Tripp, A. and Hernández-Lobato, J. M. Genetic algorithms are strong baselines for molecule generation, 2023.

Tripp, A., Simm, G. N., and Hernández-Lobato, J. M. A fresh look at de novo molecular design benchmarks. In *NeurIPS 2021 AI for Science Workshop*, 2021.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Yoshikawa, N., Terayama, K., Sumita, M., Homma, T., Oono, K., and Tsuda, K. Population-based De Novo Molecule Generation, Using Grammatical Evolution. *Chemistry Letters*, 47(11):1431–1434, 10 2018. ISSN 0366-7022. doi: 10.1246/cl.180665. URL https://doi.org/10.1246/cl.180665.

Zhang, O., Zhang, J., Jin, J., Zhang, X., Hu, R., Shen, C., Cao, H., Du, H., Kang, Y., Deng, Y., Liu, F., Chen, G., Hsieh, C.-Y., and Hou, T. Resgen is a pocket-aware 3d molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence*, 5(9):1020–1030, September 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00712-7. URL https://doi.org/10.1038/s42256-023-00712-7.

Zhou, Z., Kearnes, S., Li, L., Zare, R., and Riley, P. Optimization of molecules via deep reinforcement learning. 10 2018.

A. Appendix

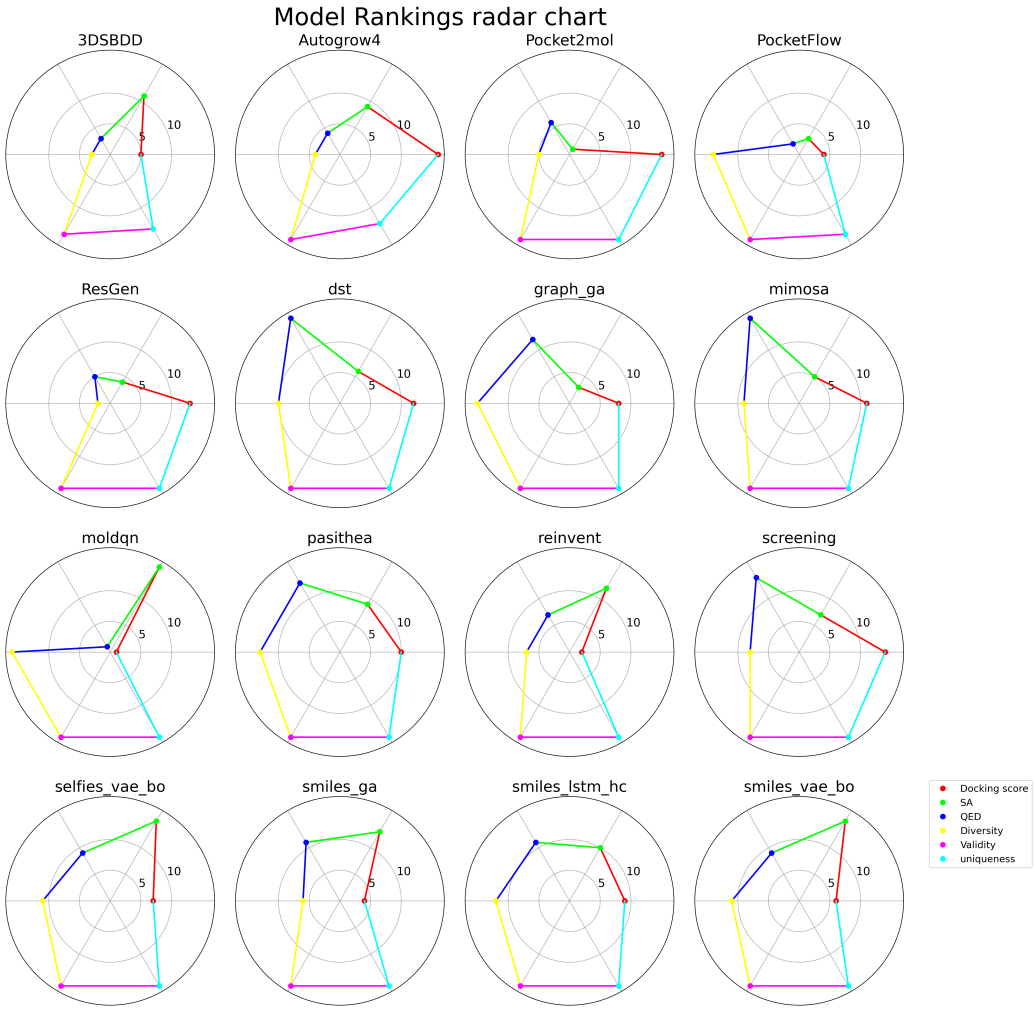

Figure 4. The radar chart is based on the ranking of each model's average performance. The outermost circle represents the best ranking for each metric and vice versa. No methods can dominate structure-based drug design in all the evaluation metrics.

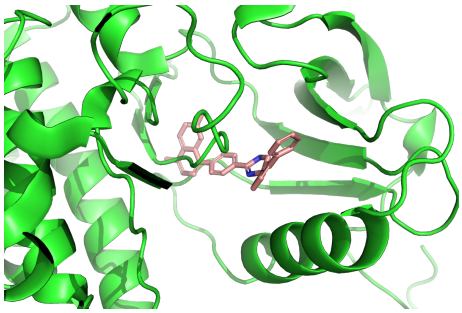

Figure 5. Model: PocketFlow, PDB: 1iep, -13.9 kcal/mol

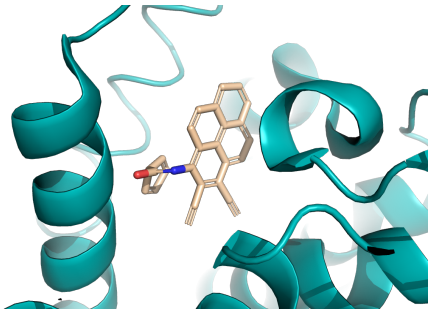

Figure 6. Model: AutoGrow4, PDB: 3eml, -13.3 kcal/mol

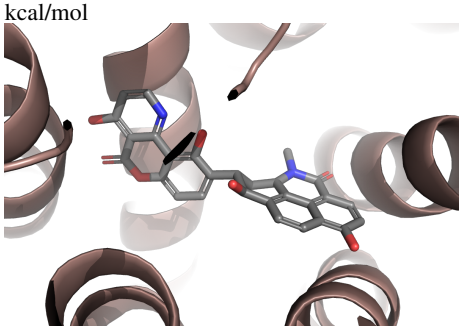

Figure 7. Model: Pocket2mol, PDB: 3ny8, -12.2 kcal/mol

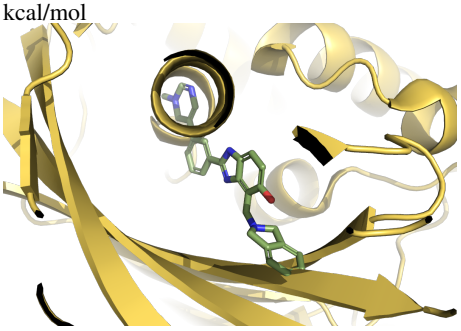

Figure 8. Model: ResGen, PDB: 4rlu, -12.6 kcal/mol

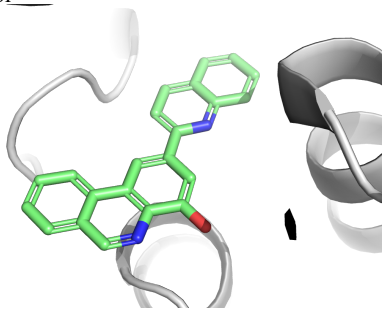

Figure 9. Model: Pocket2mol, PDB: 5mo4, -11.9 kcal/mol

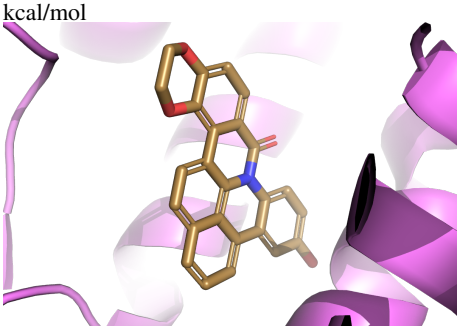

Figure 10. Model: Pocket2mol, PDB: 4unn, -12.8 kcal/mol

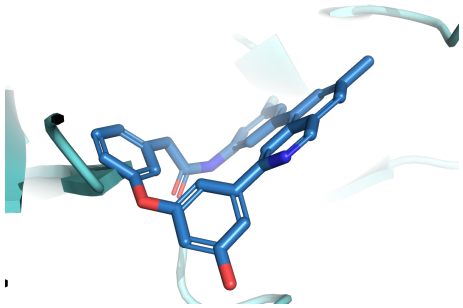

Figure 11. Model: Pocket2mol, PDB: 7l11, -10.4 kcal/mol

Figure 12. Examples of best generated molecules for each PDB

Table 3. Top 1 Docking score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	-10.00	-10.40	-10.60	-11.40
AUTOGROW4	-13.40	**-13.30**	-11.70	-11.20
POCKET2MOL	-11.50	-12.80	**-12.20**	-10.70
POCKETFLOW	**-13.90**	-9.80	-9.20	-10.10
RESGEN	-11.60	-10.80	-11.70	**-12.60**
DST	-12.20	-11.00	-11.00	-11.40
GRAPH GA	-10.80	-10.50	-10.20	-11.10
MIMOSA	-12.20	-11.10	-11.00	-11.40
MOLDQN	-6.90	-6.80	-7.50	-7.50
PASITHEA	-11.60	-10.40	-11.20	-11.60
REINVENT	-10.40	-10.40	-10.30	-10.10
SCREENING	-11.30	-12.20	-11.90	-11.20
SELFIES VAE BO	-11.80	-10.00	-10.70	-10.50
SMILES GA	-9.90	-10.50	-10.60	-9.90
SMILES LSTM HC	-10.70	-10.60	-10.40	-11.30
SMILES-VAE-BO	-10.30	-10.00	-10.70	-10.50

MODEL	4UNN	5MO4	7L11
3DSBDD	-8.90	-9.80	-8.80
AUTOGROW4	-11.20	-10.80	-9.60
POCKET2MOL	**-12.80**	**-11.90**	**-10.40**
POCKETFLOW	-9.50	-8.70	-8.90
RESGEN	-9.90	-11.00	-9.40
DST	-9.90	-11.00	-9.30
GRAPH GA	-10.60	-9.60	-8.60
MIMOSA	-10.00	-11.00	-9.30
MOLDQN	-6.40	-6.50	-7.20
PASITHEA	-10.40	-10.10	-8.60
REINVENT	-9.00	-9.60	-7.70
SCREENING	-10.30	-10.50	-8.70
SELFIES VAE BO	-9.60	-10.10	-8.20
SMILES GA	-9.30	-9.60	-8.10
SMILES LSTM HC	-9.70	-10.60	-8.40
SMILES-VAE-BO	-9.60	-10.10	-8.20

Table 4. Top 50 Docking score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	-8.43 ± 0.39	-9.50 ± 0.34	-9.15 ± 0.61	-9.10 ± 0.47
AUTOGROW4	**-12.47** ± 0.70	**-12.45** ± 0.36	-11.04 ± 0.38	**-10.92** ± 0.19
POCKET2MOL	-9.57 ± 0.40	-11.64 ± 0.37	**-11.24** ± 0.42	-10.18 ± 0.25
POCKETFLOW	-11.60 ± 0.57	-8.52 ± 0.47	-7.97 ± 0.35	-8.96 ± 0.41
RESGEN	-10.36 ± 0.40	-7.22 ± 1.14	-10.35 ± 0.39	-10.79 ± 0.61
DST	-10.07 ± 0.55	-10.03 ± 0.37	-10.04 ± 0.30	-10.24 ± 0.39
GRAPH GA	-9.19 ± 0.53	-9.19 ± 0.43	-9.31 ± 0.40	-9.52 ± 0.45
MIMOSA	-10.09 ± 0.55	-10.01 ± 0.39	-10.04 ± 0.30	-10.23 ± 0.37
MOLDQN	-6.31 ± 0.27	-6.14 ± 0.24	-6.39 ± 0.49	-6.15 ± 0.40
PASITHEA	-10.10 ± 0.48	-9.91 ± 0.25	-10.10 ± 0.36	-10.25 ± 0.42
REINVENT	-8.53 ± 0.81	-8.55 ± 0.58	-8.70 ± 0.56	-8.69 ± 0.63
SCREENING	-10.23 ± 0.40	-10.11 ± 0.50	-10.12 ± 0.40	-10.24 ± 0.38
SELFIES VAE BO	-9.23 ± 0.63	-9.24 ± 0.39	-9.35 ± 0.42	-9.30 ± 0.43
SMILES GA	-8.90 ± 0.41	-8.65 ± 0.53	-9.13 ± 0.52	-9.05 ± 0.33
SMILES LSTM HC	-9.61 ± 0.47	-9.80 ± 0.33	-9.63 ± 0.36	-9.74 ± 0.49
SMILES-VAE-BO	-9.19 ± 0.49	-9.27 ± 0.39	-9.34 ± 0.41	-9.31 ± 0.45

MODEL	4UNN	5MO4	7L11
3DSBDD	-7.45 ± 0.52	-7.98 ± 0.45	-8.08 ± 0.25
AUTOGROW4	-10.89 ± 0.14	**-10.20** ± 0.16	-8.44 ± 0.26
POCKET2MOL	**-11.56** ± 0.40	-9.55 ± 0.40	**-9.11** ± 0.40
POCKETFLOW	-6.91 ± 0.63	-7.06 ± 0.45	-7.85 ± 0.31
RESGEN	-8.63 ± 0.50	-9.39 ± 0.56	-8.28 ± 0.30
DST	-9.13 ± 0.34	-9.30 ± 0.43	-7.81 ± 0.35
GRAPH GA	-8.43 ± 0.54	-8.66 ± 0.41	-7.15 ± 0.38
MIMOSA	-9.11 ± 0.33	-9.32 ± 0.41	-7.83 ± 0.35
MOLDQN	-5.47 ± 0.28	-5.80 ± 0.33	-5.98 ± 0.52
PASITHEA	-9.02 ± 0.41	-9.24 ± 0.33	-7.68 ± 0.23
REINVENT	-7.89 ± 0.56	-7.98 ± 0.58	-6.64 ± 0.40
SCREENING	-9.17 ± 0.38	-9.38 ± 0.34	-7.67 ± 0.30
SELFIES VAE BO	-8.34 ± 0.44	-8.59 ± 0.40	-7.15 ± 0.37
SMILES GA	-8.25 ± 0.35	-8.47 ± 0.45	-7.04 ± 0.30
SMILES LSTM HC	-8.87 ± 0.31	-9.00 ± 0.44	-7.46 ± 0.27
SMILES-VAE-BO	-8.36 ± 0.43	-8.60 ± 0.40	-7.14 ± 0.37

Table 5. Top 100 Docking score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	-8.01 ± 0.52	-9.15 ± 0.42	-8.60 ± 0.72	-8.71 ± 0.53
AUTOGROW4	**-11.80** ± 0.84	**-12.14** ± 0.40	**-10.84** ± 0.34	**-10.73** ± 0.24
POCKET2MOL	-9.36 ± 0.36	-11.30 ± 0.44	-10.73 ± 0.60	-9.95 ± 0.30
POCKETFLOW	-11.20 ± 0.58	-8.03 ± 0.60	-7.69 ± 0.38	-8.64 ± 0.43
RESGEN	-10.03 ± 0.44	-6.43 ± 1.14	-10.04 ± 0.42	-10.27 ± 0.68
DST	-9.72 ± 0.53	-9.74 ± 0.40	-9.77 ± 0.35	-9.91 ± 0.44
GRAPH GA	-8.67 ± 0.65	-8.80 ± 0.51	-8.85 ± 0.55	-9.04 ± 0.60
MIMOSA	-9.74 ± 0.53	-9.73 ± 0.40	-9.77 ± 0.35	-9.90 ± 0.43
MOLDQN	-6.03 ± 0.36	-5.85 ± 0.36	-5.94 ± 0.57	-5.82 ± 0.44
PASITHEA	-9.72 ± 0.52	-9.65 ± 0.32	-9.77 ± 0.43	-9.88 ± 0.48
REINVENT	-7.59 ± 1.23	-7.61 ± 1.50	-7.81 ± 1.16	-7.80 ± 1.16
SCREENING	-9.84 ± 0.50	-9.76 ± 0.51	-9.80 ± 0.44	-9.91 ± 0.43
SELFIES VAE BO	-8.64 ± 0.75	-8.75 ± 0.59	-8.86 ± 0.59	-8.87 ± 0.55
SMILES GA	-8.50 ± 0.50	-8.23 ± 0.57	-8.74 ± 0.54	-8.77 ± 0.38
SMILES LSTM HC	-9.14 ± 0.58	-9.44 ± 0.44	-9.20 ± 0.52	-9.35 ± 0.53
SMILES-VAE-BO	-8.64 ± 0.67	-8.75 ± 0.61	-8.86 ± 0.58	-8.88 ± 0.56

MODEL	4UNN	5MO4	7L11
3DSBDD	-6.71 ± 0.86	-7.50 ± 0.62	-7.86 ± 0.29
AUTOGROW4	-10.81 ± 0.13	**-10.10** ± 0.16	-8.32 ± 0.22
POCKET2MOL	**-11.22** ± 0.45	-9.20 ± 0.48	**-8.78** ± 0.44
POCKETFLOW	-6.58 ± 0.56	-6.70 ± 0.49	-7.62 ± 0.33
RESGEN	-8.13 ± 0.64	-8.88 ± 0.66	-7.97 ± 0.39
DST	-8.83 ± 0.39	-9.03 ± 0.41	-7.55 ± 0.36
GRAPH GA	-8.03 ± 0.56	-8.22 ± 0.53	-6.87 ± 0.40
MIMOSA	-8.81 ± 0.39	-9.05 ± 0.41	-7.56 ± 0.37
MOLDQN	-5.22 ± 0.33	-5.47 ± 0.40	-5.51 ± 0.61
PASITHEA	-8.72 ± 0.43	-8.94 ± 0.38	-7.46 ± 0.29
REINVENT	-7.14 ± 0.97	-7.16 ± 1.06	-6.06 ± 0.75
SCREENING	-8.85 ± 0.43	-9.11 ± 0.36	-7.44 ± 0.32
SELFIES VAE BO	-7.93 ± 0.52	-8.19 ± 0.50	-6.81 ± 0.44
SMILES GA	-7.97 ± 0.37	-8.14 ± 0.47	-6.83 ± 0.31
SMILES LSTM HC	-8.53 ± 0.42	-8.64 ± 0.49	-7.20 ± 0.33
SMILES-VAE-BO	-7.95 ± 0.52	-8.20 ± 0.50	-6.81 ± 0.43

Table 6. Top 1 LogP score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	1.23	2.94	2.66	0.60
AUTOGROW4	3.47	2.95	2.63	2.86
POCKET2MOL	2.32	3.83	4.14	3.50
POCKETFLOW	6.10	4.86	2.57	4.38
RESGEN	3.91	10.72	5.47	3.50
DST	3.72	3.72	3.72	3.72
GRAPH GA	3.06	3.32	3.67	3.70
MIMOSA	3.72	3.72	3.72	3.72
MOLDQN	0.36	0.55	0.36	0.33
PASITHEA	3.22	3.22	3.22	3.22
REINVENT	3.90	3.90	3.90	3.90
SCREENING	3.67	3.67	3.67	3.67
SELFIES VAE BO	3.22	3.22	3.22	3.22
SMILES GA	3.00	3.00	3.40	3.00
SMILES LSTM HC	6.51	6.51	6.51	6.51
SMILES-VAE-BO	3.22	3.22	3.22	3.22

MODEL	4UNN	5MO4	7L11
3DSBDD	2.69	2.10	3.36
AUTOGROW4	2.72	2.93	2.81
POCKET2MOL	2.86	3.66	3.92
POCKETFLOW	2.16	2.26	5.21
RESGEN	3.40	3.54	3.23
DST	3.72	3.72	3.72
GRAPH GA	3.13	3.01	3.76
MIMOSA	3.72	3.72	3.72
MOLDQN	-0.26	1.00	0.50
PASITHEA	3.22	3.22	3.22
REINVENT	3.90	3.90	3.90
SCREENING	3.67	3.67	3.67
SELFIES VAE BO	3.22	3.22	3.22
SMILES GA	3.40	3.00	3.00
SMILES LSTM HC	6.51	6.51	6.51
SMILES-VAE-BO	3.22	3.22	3.22

Table 7. Top 10 LogP score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.16 ± 0.37	1.59 ± 0.67	1.89 ± 0.51	-0.18 ± 0.60
AUTOGROW4	3.36 ± 0.05	2.72 ± 0.13	2.63 ± 0.00	2.86 ± 0.00
POCKET2MOL	2.10 ± 0.13	3.59 ± 0.15	3.66 ± 0.22	3.15 ± 0.19
POCKETFLOW	5.81 ± 0.17	2.21 ± 0.92	2.01 ± 0.32	3.79 ± 0.59
RESGEN	3.58 ± 0.24	6.62 ± 1.55	4.91 ± 0.32	3.05 ± 0.25
DST	3.13 ± 0.21	3.13 ± 0.21	3.13 ± 0.21	3.13 ± 0.21
GRAPH GA	2.79 ± 0.20	2.91 ± 0.23	3.14 ± 0.28	3.00 ± 0.32
MIMOSA	3.13 ± 0.21	3.13 ± 0.21	3.13 ± 0.21	3.13 ± 0.21
MOLDQN	-0.75 ± 0.49	-0.52 ± 0.48	-0.75 ± 0.50	-0.35 ± 0.29
PASITHEA	3.03 ± 0.09	3.03 ± 0.09	3.03 ± 0.09	3.03 ± 0.09
REINVENT	2.50 ± 0.51	2.47 ± 0.53	2.50 ± 0.51	2.47 ± 0.53
SCREENING	3.30 ± 0.18	3.30 ± 0.18	3.30 ± 0.18	3.30 ± 0.18
SELFIES VAE BO	2.68 ± 0.33	2.68 ± 0.33	2.68 ± 0.33	2.68 ± 0.33
SMILES GA	2.32 ± 0.30	2.02 ± 0.36	2.52 ± 0.43	2.17 ± 0.29
SMILES LSTM HC	4.72 ± 0.85	4.72 ± 0.85	4.72 ± 0.85	4.72 ± 0.85
SMILES-VAE-BO	2.68 ± 0.33	2.68 ± 0.33	2.68 ± 0.33	2.68 ± 0.33

MODEL	4UNN	5MO4	7L11
3DSBDD	2.02 ± 0.28	1.40 ± 0.39	2.46 ± 0.59
AUTOGROW4	2.68 ± 0.05	2.86 ± 0.06	2.77 ± 0.02
POCKET2MOL	2.56 ± 0.16	3.23 ± 0.22	3.46 ± 0.19
POCKETFLOW	1.44 ± 0.28	1.00 ± 0.47	4.79 ± 0.26
RESGEN	2.19 ± 0.60	2.71 ± 0.49	2.74 ± 0.27
DST	3.13 ± 0.21	3.13 ± 0.21	3.13 ± 0.21
GRAPH GA	2.84 ± 0.19	2.71 ± 0.20	2.99 ± 0.29
MIMOSA	3.13 ± 0.21	3.13 ± 0.21	3.13 ± 0.21
MOLDQN	-0.95 ± 0.43	-0.12 ± 0.54	-0.23 ± 0.48
PASITHEA	3.03 ± 0.09	3.03 ± 0.09	3.03 ± 0.09
REINVENT	2.50 ± 0.51	2.50 ± 0.51	2.50 ± 0.51
SCREENING	3.30 ± 0.18	3.30 ± 0.18	3.30 ± 0.18
SELFIES VAE BO	2.68 ± 0.33	2.68 ± 0.33	2.68 ± 0.33
SMILES GA	2.48 ± 0.44	2.34 ± 0.30	2.14 ± 0.34
SMILES LSTM HC	4.72 ± 0.85	4.72 ± 0.85	4.72 ± 0.85
SMILES VAE BO	2.68 ± 0.33	2.68 ± 0.33	2.68 ± 0.33

Table 8. Top 50 LogP score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.03 ± 0.18	0.56 ± 0.64	0.53 ± 0.78	-2.01 ± 1.00
AUTOGROW4	2.96 ± 0.22	2.56 ± 0.15	2.49 ± 0.13	2.67 ± 0.16
POCKET2MOL	1.80 ± 0.18	3.12 ± 0.31	3.20 ± 0.30	2.76 ± 0.25
POCKETFLOW	5.38 ± 0.29	1.20 ± 0.70	0.98 ± 0.66	2.19 ± 0.92
RESGEN	2.99 ± 0.37	5.22 ± 1.02	3.99 ± 0.54	2.20 ± 0.51
DST	2.78 ± 0.25	2.78 ± 0.25	2.78 ± 0.25	2.78 ± 0.25
GRAPH GA	2.06 ± 0.42	2.18 ± 0.44	2.22 ± 0.55	2.18 ± 0.49
MIMOSA	2.78 ± 0.25	2.78 ± 0.25	2.78 ± 0.25	2.78 ± 0.25
MOLDQN	-1.73 ± 0.60	-1.51 ± 0.59	-1.68 ± 0.58	-1.20 ± 0.53
PASITHEA	2.69 ± 0.25	2.69 ± 0.25	2.69 ± 0.25	2.69 ± 0.25
REINVENT	1.50 ± 0.65	1.42 ± 0.69	1.46 ± 0.69	1.46 ± 0.66
SCREENING	2.85 ± 0.27	2.85 ± 0.27	2.85 ± 0.27	2.85 ± 0.27
SELFIES VAE BO	1.99 ± 0.42	1.99 ± 0.42	1.99 ± 0.42	1.99 ± 0.42
SMILES GA	1.72 ± 0.39	1.25 ± 0.48	1.81 ± 0.47	1.73 ± 0.32
SMILES LSTM HC	3.51 ± 0.77	3.51 ± 0.77	3.51 ± 0.77	3.51 ± 0.77
SMILES VAE BO	1.99 ± 0.42	1.99 ± 0.42	1.99 ± 0.42	1.99 ± 0.42

MODEL	4UNN	5MO4	7L11
3DSBDD	0.90 ± 0.77	0.30 ± 0.58	0.73 ± 0.98
AUTOGROW4	2.63 ± 0.04	2.70 ± 0.14	2.40 ± 0.31
POCKET2MOL	2.12 ± 0.30	2.80 ± 0.28	2.96 ± 0.32
POCKETFLOW	0.61 ± 0.53	0.57 ± 0.31	4.06 ± 0.44
RESGEN	0.81 ± 0.91	1.68 ± 0.62	1.97 ± 0.47
DST	2.78 ± 0.25	2.78 ± 0.25	2.78 ± 0.25
GRAPH GA	2.15 ± 0.44	2.09 ± 0.37	2.27 ± 0.46
MIMOSA	2.78 ± 0.25	2.78 ± 0.25	2.78 ± 0.25
MOLDQN	-1.88 ± 0.57	-1.46 ± 0.79	-1.34 ± 0.66
PASITHEA	2.69 ± 0.25	2.69 ± 0.25	2.69 ± 0.25
REINVENT	1.47 ± 0.68	1.48 ± 0.67	1.49 ± 0.66
SCREENING	2.85 ± 0.27	2.85 ± 0.27	2.85 ± 0.27
SELFIES VAE BO	1.99 ± 0.42	1.99 ± 0.42	1.99 ± 0.42
SMILES GA	1.77 ± 0.46	1.70 ± 0.40	1.50 ± 0.42
SMILES LSTM HC	3.51 ± 0.77	3.51 ± 0.77	3.51 ± 0.77
SMILES VAE BO	1.99 ± 0.42	1.99 ± 0.42	1.99 ± 0.42

Table 9. Top 100 LogP score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.02 ± 0.13	0.28 ± 0.53	0.26 ± 0.61	-2.58 ± 0.92
AUTOGROW4	2.83 ± 0.21	2.37 ± 0.23	2.20 ± 0.33	2.31 ± 0.40
POCKET2MOL	1.61 ± 0.24	2.79 ± 0.40	2.91 ± 0.37	2.50 ± 0.32
POCKETFLOW	5.02 ± 0.43	0.57 ± 0.82	0.40 ± 0.76	1.42 ± 1.02
RESGEN	2.65 ± 0.44	4.52 ± 1.05	3.59 ± 0.56	1.72 ± 0.61
DST	2.51 ± 0.32	2.51 ± 0.32	2.51 ± 0.32	2.51 ± 0.32
GRAPH GA	1.69 ± 0.49	1.78 ± 0.52	1.76 ± 0.62	1.76 ± 0.56
MIMOSA	2.51 ± 0.32	2.51 ± 0.32	2.51 ± 0.32	2.51 ± 0.32
MOLDQN	-2.21 ± 0.66	-1.97 ± 0.63	-2.19 ± 0.67	-1.72 ± 0.66
PASITHEA	2.42 ± 0.33	2.42 ± 0.33	2.42 ± 0.33	2.42 ± 0.33
REINVENT	0.15 ± 1.82	0.02 ± 1.89	0.05 ± 1.90	0.07 ± 1.88
SCREENING	2.60 ± 0.32	2.60 ± 0.32	2.60 ± 0.32	2.60 ± 0.32
SELFIES VAE BO	1.43 ± 0.68	1.43 ± 0.68	1.43 ± 0.68	1.43 ± 0.68
SMILES GA	1.35 ± 0.47	0.80 ± 0.57	1.41 ± 0.52	1.35 ± 0.45
SMILES LSTM HC	2.96 ± 0.78	2.96 ± 0.78	2.96 ± 0.78	2.96 ± 0.78
SMILES VAE BO	1.43 ± 0.68	1.43 ± 0.68	1.43 ± 0.68	1.43 ± 0.68

MODEL	4UNN	5MO4	7L11
3DSBDD	0.45 ± 0.70	0.15 ± 0.44	-0.05 ± 1.22
AUTOGROW4	2.46 ± 0.21	2.43 ± 0.30	2.10 ± 0.38
POCKET2MOL	1.80 ± 0.39	2.53 ± 0.35	2.50 ± 0.54
POCKETFLOW	0.04 ± 0.69	0.19 ± 0.46	3.56 ± 0.62
RESGEN	0.00 ± 1.05	1.20 ± 0.66	1.48 ± 0.61
DST	2.51 ± 0.32	2.51 ± 0.32	2.51 ± 0.32
GRAPH GA	1.70 ± 0.56	1.70 ± 0.48	1.84 ± 0.55
MIMOSA	2.51 ± 0.32	2.51 ± 0.32	2.51 ± 0.32
MOLDQN	-2.34 ± 0.62	-2.07 ± 0.84	-1.93 ± 0.77
PASITHEA	2.42 ± 0.33	2.42 ± 0.33	2.42 ± 0.33
REINVENT	0.05 ± 1.90	0.08 ± 1.89	0.09 ± 1.89
SCREENING	2.60 ± 0.32	2.60 ± 0.32	2.60 ± 0.32
SELFIES VAE BO	1.43 ± 0.68	1.43 ± 0.68	1.43 ± 0.68
SMILES GA	1.39 ± 0.51	1.37 ± 0.44	1.10 ± 0.52
SMILES LSTM HC	2.96 ± 0.78	2.96 ± 0.78	2.96 ± 0.78
SMILES VAE BO	1.43 ± 0.68	1.43 ± 0.68	1.43 ± 0.68

Table 10. Top 1 QED score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.81	0.93	**0.95**	0.85
AUTOGROW4	0.86	0.85	0.92	0.84
POCKET2MOL	0.92	0.91	0.93	0.92
POCKETFLOW	0.92	0.86	0.81	0.88
RESGEN	0.92	0.80	0.93	0.94
DST	**0.95**	**0.95**	0.95	**0.95**
GRAPH GA	0.94	0.94	0.94	0.94
MIMOSA	**0.95**	**0.95**	0.95	**0.95**
MOLDQN	0.52	0.65	0.62	0.67
PASITHEA	**0.95**	**0.95**	0.95	**0.95**
REINVENT	0.95	0.95	0.95	0.95
SCREENING	0.95	0.95	0.95	0.95
SELFIES VAE BO	0.94	0.94	0.94	0.94
SMILES GA	0.93	0.93	0.94	0.94
SMILES LSTM HC	0.94	0.94	0.94	0.94
SMILES VAE BO	0.94	0.94	0.94	0.94

MODEL	4UNN	5MO4	7L11
3DSBDD	0.88	0.80	0.90
AUTOGROW4	0.89	0.87	0.84
POCKET2MOL	0.88	0.90	0.94
POCKETFLOW	0.83	0.80	0.91
RESGEN	0.93	0.88	0.95
DST	**0.95**	**0.95**	0.95
GRAPH GA	0.94	0.94	**0.95**
MIMOSA	**0.95**	**0.95**	0.95
MOLDQN	0.67	0.52	0.70
PASITHEA	**0.95**	**0.95**	0.95
REINVENT	0.95	0.95	0.95
SCREENING	0.95	0.95	0.95
SELFIES VAE BO	0.94	0.94	0.94
SMILES GA	0.93	0.93	0.94
SMILES LSTM HC	0.94	0.94	0.94
SMILES VAE BO	0.94	0.94	0.94

Table 11. Top 10 QED score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.77 ± 0.02	0.83 ± 0.05	0.88 ± 0.03	0.81 ± 0.02
AUTOGROW4	0.83 ± 0.01	0.79 ± 0.04	0.92 ± 0.00	0.82 ± 0.01
POCKET2MOL	0.90 ± 0.01	0.89 ± 0.01	0.92 ± 0.01	0.90 ± 0.01
POCKETFLOW	0.89 ± 0.01	0.81 ± 0.03	0.77 ± 0.03	0.79 ± 0.03
RESGEN	0.91 ± 0.01	0.73 ± 0.03	0.90 ± 0.02	0.93 ± 0.00
DST	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
GRAPH GA	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.92 ± 0.01
MIMOSA	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
MOLDQN	0.49 ± 0.02	0.55 ± 0.04	0.51 ± 0.05	0.60 ± 0.05
PASITHEA	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
REINVENT	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
SCREENING	**0.94** ± 0.00	**0.94** ± 0.00	**0.94** ± 0.00	**0.94** ± 0.00
SELFIES VAE BO	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
SMILES GA	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.00	0.93 ± 0.01
SMILES LSTM HC	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
SMILES VAE BO	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02

MODEL	4UNN	5MO4	7L11
3DSBDD	0.82 ± 0.03	0.76 ± 0.02	0.86 ± 0.02
AUTOGROW4	0.87 ± 0.01	0.83 ± 0.03	0.80 ± 0.02
POCKET2MOL	0.84 ± 0.02	0.88 ± 0.01	0.94 ± 0.01
POCKETFLOW	0.74 ± 0.04	0.74 ± 0.03	0.90 ± 0.01
RESGEN	0.85 ± 0.04	0.84 ± 0.01	0.92 ± 0.02
DST	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
GRAPH GA	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01
MIMOSA	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
MOLDQN	0.57 ± 0.04	0.48 ± 0.02	0.54 ± 0.07
PASITHEA	0.94 ± 0.00	0.94 ± 0.00	0.94 ± 0.00
REINVENT	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
SCREENING	**0.94** ± 0.00	**0.94** ± 0.00	**0.94** ± 0.00
SELFIES VAE BO	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
SMILES GA	0.93 ± 0.00	0.93 ± 0.00	0.92 ± 0.01
SMILES LSTM HC	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
SMILES VAE BO	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02

Table 12. Top 50 QED score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.69 ± 0.05	0.71 ± 0.07	0.79 ± 0.05	0.77 ± 0.02
AUTOGROW4	0.78 ± 0.03	0.73 ± 0.04	0.88 ± 0.03	0.76 ± 0.04
POCKET2MOL	0.87 ± 0.02	0.85 ± 0.03	0.88 ± 0.03	0.87 ± 0.02
POCKETFLOW	0.85 ± 0.03	0.71 ± 0.06	0.69 ± 0.04	0.69 ± 0.06
RESGEN	0.86 ± 0.04	0.65 ± 0.05	0.85 ± 0.03	0.90 ± 0.02
DST	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
GRAPH GA	0.89 ± 0.02	0.89 ± 0.03	0.89 ± 0.03	0.88 ± 0.03
MIMOSA	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
MOLDQN	0.43 ± 0.03	0.48 ± 0.04	0.44 ± 0.04	0.49 ± 0.06
PASITHEA	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
REINVENT	0.84 ± 0.05	0.84 ± 0.05	0.84 ± 0.05	0.84 ± 0.05
SCREENING	**0.92** ± 0.01	**0.92** ± 0.01	**0.92** ± 0.01	**0.92** ± 0.01
SELFIES VAE BO	0.87 ± 0.03	0.87 ± 0.03	0.87 ± 0.03	0.87 ± 0.03
SMILES GA	0.87 ± 0.03	0.87 ± 0.03	0.89 ± 0.02	0.89 ± 0.02
SMILES LSTM HC	0.88 ± 0.03	0.88 ± 0.03	0.88 ± 0.03	0.88 ± 0.03
SMILES VAE BO	0.87 ± 0.03	0.87 ± 0.03	0.87 ± 0.03	0.87 ± 0.03

MODEL	4UNN	5MO4	7L11
3DSBDD	0.74 ± 0.06	0.64 ± 0.08	0.74 ± 0.07
AUTOGROW4	0.86 ± 0.01	0.75 ± 0.05	0.74 ± 0.04
POCKET2MOL	0.80 ± 0.03	0.84 ± 0.03	0.91 ± 0.02
POCKETFLOW	0.67 ± 0.04	0.68 ± 0.04	0.85 ± 0.03
RESGEN	0.76 ± 0.06	0.78 ± 0.04	0.85 ± 0.04
DST	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
GRAPH GA	0.89 ± 0.03	0.89 ± 0.02	0.89 ± 0.03
MIMOSA	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
MOLDQN	0.49 ± 0.05	0.43 ± 0.03	0.45 ± 0.06
PASITHEA	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
REINVENT	0.83 ± 0.06	0.83 ± 0.05	0.83 ± 0.05
SCREENING	**0.92** ± 0.01	**0.92** ± 0.01	**0.92** ± 0.01
SELFIES VAE BO	0.87 ± 0.03	0.87 ± 0.03	0.87 ± 0.03
SMILES GA	0.90 ± 0.02	0.89 ± 0.02	0.88 ± 0.03
SMILES LSTM HC	0.88 ± 0.03	0.88 ± 0.03	0.88 ± 0.03
SMILES VAE BO	0.87 ± 0.03	0.87 ± 0.03	0.87 ± 0.03

Table 13. Top 100 QED score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	0.64 ± 0.07	0.65 ± 0.08	0.73 ± 0.08	0.75 ± 0.03
AUTOGROW4	0.76 ± 0.03	0.67 ± 0.07	0.84 ± 0.05	0.72 ± 0.05
POCKET2MOL	0.85 ± 0.03	0.82 ± 0.03	0.85 ± 0.04	0.86 ± 0.02
POCKETFLOW	0.80 ± 0.05	0.65 ± 0.07	0.65 ± 0.05	0.64 ± 0.07
RESGEN	0.82 ± 0.05	0.60 ± 0.06	0.83 ± 0.04	0.86 ± 0.04
DST	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
GRAPH GA	0.85 ± 0.04	0.85 ± 0.04	0.85 ± 0.04	0.85 ± 0.04
MIMOSA	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
MOLDQN	0.40 ± 0.04	0.44 ± 0.05	0.41 ± 0.05	0.45 ± 0.06
PASITHEA	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02
REINVENT	0.73 ± 0.14	0.73 ± 0.14	0.73 ± 0.14	0.73 ± 0.14
SCREENING	**0.91** ± 0.02	**0.91** ± 0.02	**0.91** ± 0.02	**0.91** ± 0.02
SELFIES VAE BO	0.83 ± 0.05	0.83 ± 0.05	0.83 ± 0.05	0.83 ± 0.05
SMILES GA	0.84 ± 0.04	0.83 ± 0.05	0.86 ± 0.04	0.86 ± 0.04
SMILES LSTM HC	0.85 ± 0.04	0.85 ± 0.04	0.85 ± 0.04	0.85 ± 0.04
SMILES VAE BO	0.83 ± 0.05	0.83 ± 0.05	0.83 ± 0.05	0.83 ± 0.05

MODEL	4UNN	5MO4	7L11
3DSBDD	0.65 ± 0.10	0.54 ± 0.12	0.66 ± 0.10
AUTOGROW4	0.84 ± 0.02	0.71 ± 0.06	0.69 ± 0.06
POCKET2MOL	0.77 ± 0.04	0.81 ± 0.03	0.88 ± 0.03
POCKETFLOW	0.63 ± 0.05	0.64 ± 0.05	0.82 ± 0.04
RESGEN	0.70 ± 0.07	0.75 ± 0.05	0.81 ± 0.05
DST	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
GRAPH GA	0.85 ± 0.05	0.86 ± 0.04	0.85 ± 0.04
MIMOSA	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02
MOLDQN	0.44 ± 0.06	0.39 ± 0.04	0.41 ± 0.06
PASITHEA	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02
REINVENT	0.73 ± 0.14	0.73 ± 0.13	0.73 ± 0.14
SCREENING	**0.91** ± 0.02	**0.91** ± 0.02	**0.91** ± 0.02
SELFIES VAE BO	0.83 ± 0.05	0.83 ± 0.05	0.83 ± 0.05
SMILES GA	0.88 ± 0.03	0.86 ± 0.03	0.84 ± 0.04
SMILES LSTM HC	0.85 ± 0.04	0.85 ± 0.04	0.85 ± 0.04
SMILES VAE BO	0.83 ± 0.05	0.83 ± 0.05	0.83 ± 0.05

Table 14. Top 1 SA score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	1.37	**1.00**	**1.00**	1.99
AUTOGROW4	**1.00**	**1.00**	1.83	**1.00**
POCKET2MOL	1.05	**1.00**	1.90	**1.00**
POCKETFLOW	**1.00**	**1.00**	**1.00**	1.61
RESGEN	**1.00**	**1.00**	**1.00**	1.16
DST	1.41	1.41	1.41	1.41
GRAPH GA	**1.00**	**1.00**	**1.00**	**1.00**
MIMOSA	1.41	1.41	1.41	1.41
MOLDQN	1.51	1.65	1.51	1.62
PASITHEA	1.41	1.41	1.41	1.41
REINVENT	1.67	1.67	1.67	1.67
SCREENING	1.51	1.51	1.51	1.51
SELFIES VAE BO	1.75	1.75	1.75	1.75
SMILES GA	1.61	1.61	1.60	1.61
SMILES LSTM HC	1.60	1.60	1.60	1.60
SMILES VAE BO	1.75	1.75	1.75	1.75

MODEL	4UNN	5MO4	7L11
3DSBDD	**1.00**	**1.00**	**1.00**
AUTOGROW4	1.74	**1.00**	1.83
POCKET2MOL	**1.00**	**1.00**	**1.00**
POCKETFLOW	1.11	1.54	**1.00**
RESGEN	**1.00**	**1.00**	**1.00**
DST	1.41	1.41	1.41
GRAPH GA	**1.00**	**1.00**	**1.00**
MIMOSA	1.41	1.41	1.41
MOLDQN	1.98	2.04	1.51
PASITHEA	1.41	1.41	1.41
REINVENT	1.67	1.67	1.67
SCREENING	1.51	1.51	1.51
SELFIES VAE BO	1.75	1.75	1.75
SMILES GA	1.60	1.61	1.61
SMILES LSTM HC	1.60	1.60	1.60
SMILES VAE BO	1.75	1.75	1.75

Table 15. Top 10 SA score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	2.20 ± 0.38	1.54 ± 0.24	1.52 ± 0.34	2.80 ± 0.60
AUTOGROW4	1.04 ± 0.13	1.74 ± 0.26	1.91 ± 0.04	1.55 ± 0.31
POCKET2MOL	1.36 ± 0.14	**1.02** ± 0.03	2.07 ± 0.08	**1.09** ± 0.08
POCKETFLOW	**1.00** ± 0.00	1.46 ± 0.19	1.33 ± 0.18	1.61 ± 0.00
RESGEN	1.03 ± 0.04	1.08 ± 0.08	**1.07** ± 0.07	1.40 ± 0.13
DST	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09
GRAPH GA	1.37 ± 0.19	1.26 ± 0.17	1.32 ± 0.21	1.13 ± 0.14
MIMOSA	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09
MOLDQN	2.53 ± 0.37	2.39 ± 0.36	2.61 ± 0.37	2.46 ± 0.32
PASITHEA	1.67 ± 0.10	1.67 ± 0.10	1.67 ± 0.10	1.67 ± 0.10
REINVENT	1.85 ± 0.08	1.88 ± 0.09	1.86 ± 0.09	1.88 ± 0.09
SCREENING	1.63 ± 0.07	1.63 ± 0.07	1.63 ± 0.07	1.63 ± 0.07
SELFIES VAE BO	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06
SMILES GA	1.90 ± 0.11	1.96 ± 0.14	1.86 ± 0.13	1.90 ± 0.11
SMILES LSTM HC	1.77 ± 0.09	1.77 ± 0.09	1.77 ± 0.09	1.77 ± 0.09
SMILES VAE BO	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06

MODEL	4UNN	5MO4	7L11	
3DSBDD	2.20 ± 0.38	1.54 ± 0.24	1.52 ± 0.34	2.80 ± 0.60
AUTOGROW4	1.04 ± 0.13	1.74 ± 0.26	1.91 ± 0.04	1.55 ± 0.31
POCKET2MOL	1.36 ± 0.14	**1.02** ± 0.03	2.07 ± 0.08	**1.09** ± 0.08
POCKETFLOW	**1.00** ± 0.00	1.46 ± 0.19	1.33 ± 0.18	1.61 ± 0.00
RESGEN	1.03 ± 0.04	1.08 ± 0.08	**1.07** ± 0.07	1.40 ± 0.13
DST	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09
GRAPH GA	1.37 ± 0.19	1.26 ± 0.17	1.32 ± 0.21	1.13 ± 0.14
MIMOSA	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09	1.62 ± 0.09
MOLDQN	2.53 ± 0.37	2.39 ± 0.36	2.61 ± 0.37	2.46 ± 0.32
PASITHEA	1.67 ± 0.10	1.67 ± 0.10	1.67 ± 0.10	1.67 ± 0.10
REINVENT	1.85 ± 0.08	1.88 ± 0.09	1.86 ± 0.09	1.88 ± 0.09
SCREENING	1.63 ± 0.07	1.63 ± 0.07	1.63 ± 0.07	1.63 ± 0.07
SELFIES VAE BO	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06
SMILES GA	1.90 ± 0.11	1.96 ± 0.14	1.86 ± 0.13	1.90 ± 0.11
SMILES LSTM HC	1.77 ± 0.09	1.77 ± 0.09	1.77 ± 0.09	1.77 ± 0.09
SMILES VAE BO	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06	1.90 ± 0.06

Table 16. Top 50 SA score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	3.49 ± 0.78	2.28 ± 0.48	2.95 ± 1.00	4.06 ± 0.73
AUTOGROW4	1.53 ± 0.26	2.00 ± 0.20	2.07 ± 0.13	1.93 ± 0.25
POCKET2MOL	1.68 ± 0.19	**1.32** ± 0.19	2.30 ± 0.15	**1.30** ± 0.13
POCKETFLOW	**1.00** ± 0.00	1.88 ± 0.25	1.73 ± 0.28	1.67 ± 0.09
RESGEN	1.36 ± 0.21	1.71 ± 0.41	**1.39** ± 0.21	1.77 ± 0.24
DST	1.82 ± 0.13	1.82 ± 0.13	1.82 ± 0.13	1.82 ± 0.13
GRAPH GA	1.80 ± 0.26	1.77 ± 0.30	1.83 ± 0.30	1.65 ± 0.32
MIMOSA	1.82 ± 0.13	1.82 ± 0.13	1.82 ± 0.13	1.82 ± 0.13
MOLDQN	3.11 ± 0.42	3.07 ± 0.45	3.18 ± 0.41	2.95 ± 0.33
PASITHEA	1.85 ± 0.12	1.85 ± 0.12	1.85 ± 0.12	1.85 ± 0.12
REINVENT	2.32 ± 0.35	2.37 ± 0.35	2.35 ± 0.36	2.38 ± 0.35
SCREENING	1.83 ± 0.12	1.83 ± 0.12	1.83 ± 0.12	1.83 ± 0.12
SELFIES VAE BO	2.15 ± 0.18	2.15 ± 0.18	2.15 ± 0.18	2.15 ± 0.18
SMILES GA	2.27 ± 0.24	2.45 ± 0.30	2.24 ± 0.25	2.23 ± 0.22
SMILES LSTM HC	2.00 ± 0.14	2.00 ± 0.14	2.00 ± 0.14	2.00 ± 0.14
SMILES VAE BO	2.15 ± 0.18	2.15 ± 0.18	2.15 ± 0.18	2.15 ± 0.18

MODEL	4UNN	5MO4	7L11
3DSBDD	2.14 ± 0.73	3.18 ± 0.87	2.52 ± 0.90
AUTOGROW4	2.00 ± 0.14	1.94 ± 0.22	2.06 ± 0.10
POCKET2MOL	**1.60** ± 0.17	**1.26** ± 0.14	1.92 ± 0.26
POCKETFLOW	1.92 ± 0.27	1.72 ± 0.15	**1.43** ± 0.20
RESGEN	2.34 ± 0.46	1.88 ± 0.32	1.57 ± 0.31
DST	1.82 ± 0.13	1.82 ± 0.13	1.82 ± 0.13
GRAPH GA	1.95 ± 0.37	1.76 ± 0.29	1.79 ± 0.35
MIMOSA	1.82 ± 0.13	1.82 ± 0.13	1.82 ± 0.13
MOLDQN	3.24 ± 0.38	3.16 ± 0.36	3.06 ± 0.39
PASITHEA	1.85 ± 0.12	1.85 ± 0.12	1.85 ± 0.12
REINVENT	2.37 ± 0.35	2.36 ± 0.34	2.36 ± 0.34
SCREENING	1.83 ± 0.12	1.83 ± 0.12	1.83 ± 0.12
SELFIES VAE BO	2.15 ± 0.18	2.15 ± 0.18	2.15 ± 0.18
SMILES GA	2.23 ± 0.24	2.30 ± 0.23	2.39 ± 0.27
SMILES LSTM HC	2.00 ± 0.14	2.00 ± 0.14	2.00 ± 0.14
SMILES VAE BO	2.15 ± 0.18	2.15 ± 0.18	2.15 ± 0.18

Table 17. Top 100 SA score for each target protein.

MODEL	1IEP	3EML	3NY8	4RLU
3DSBDD	4.05 ± 0.79	2.85 ± 0.69	3.76 ± 1.09	4.46 ± 0.66
AUTOGROW4	1.72 ± 0.28	2.17 ± 0.23	2.29 ± 0.26	2.11 ± 0.25
POCKET2MOL	1.86 ± 0.23	**1.49** ± 0.22	2.50 ± 0.23	**1.44** ± 0.17
POCKETFLOW	**1.10** ± 0.16	2.11 ± 0.30	2.06 ± 0.40	1.87 ± 0.24
RESGEN	1.64 ± 0.33	2.24 ± 0.65	**1.66** ± 0.32	2.10 ± 0.38
DST	1.93 ± 0.14	1.93 ± 0.14	1.93 ± 0.14	1.93 ± 0.14
GRAPH GA	2.03 ± 0.30	2.02 ± 0.33	2.09 ± 0.34	1.95 ± 0.38
MIMOSA	1.93 ± 0.14	1.93 ± 0.14	1.93 ± 0.14	1.93 ± 0.14
MOLDQN	3.53 ± 0.52	3.49 ± 0.53	3.59 ± 0.51	3.30 ± 0.44
PASITHEA	1.97 ± 0.15	1.97 ± 0.15	1.97 ± 0.15	1.97 ± 0.15
REINVENT	2.96 ± 0.80	3.03 ± 0.85	3.02 ± 0.86	3.03 ± 0.84
SCREENING	1.94 ± 0.14	1.94 ± 0.14	1.94 ± 0.14	1.94 ± 0.14
SELFIES VAE BO	2.42 ± 0.31	2.42 ± 0.31	2.42 ± 0.31	2.42 ± 0.31
SMILES GA	2.52 ± 0.31	2.76 ± 0.38	2.54 ± 0.35	2.51 ± 0.33
SMILES LSTM HC	2.14 ± 0.18	2.14 ± 0.18	2.14 ± 0.18	2.14 ± 0.18
SMILES VAE BO	2.42 ± 0.31	2.42 ± 0.31	2.42 ± 0.31	2.42 ± 0.31

MODEL	4UNN	5MO4	7L11
3DSBDD	3.25 ± 1.29	4.28 ± 1.37	3.49 ± 1.17
AUTOGROW4	2.12 ± 0.16	2.11 ± 0.23	2.19 ± 0.16
POCKET2MOL	**1.76** ± 0.20	**1.40** ± 0.18	2.17 ± 0.31
POCKETFLOW	2.20 ± 0.35	2.04 ± 0.35	**1.67** ± 0.28
RESGEN	2.70 ± 0.49	2.20 ± 0.40	1.97 ± 0.47
DST	1.93 ± 0.14	1.93 ± 0.14	1.93 ± 0.14
GRAPH GA	2.22 ± 0.38	1.99 ± 0.32	2.05 ± 0.36
MIMOSA	1.93 ± 0.14	1.93 ± 0.14	1.93 ± 0.14
MOLDQN	3.60 ± 0.45	3.56 ± 0.48	3.46 ± 0.49
PASITHEA	1.97 ± 0.15	1.97 ± 0.15	1.97 ± 0.15
REINVENT	3.03 ± 0.85	3.01 ± 0.84	3.01 ± 0.84
SCREENING	1.94 ± 0.14	1.94 ± 0.14	1.94 ± 0.14
SELFIES VAE BO	2.42 ± 0.31	2.42 ± 0.31	2.42 ± 0.31
SMILES GA	2.49 ± 0.32	2.54 ± 0.30	2.68 ± 0.36
SMILES LSTM HC	2.14 ± 0.18	2.14 ± 0.18	2.14 ± 0.18
SMILES VAE BO	2.42 ± 0.31	2.42 ± 0.31	2.42 ± 0.31

Table 18. Model Rankings based on QED

MODEL	TOP 1 RANK	TOP 10 RANK	TOP 100 RANK	OVERALL RANK
3DSBDD	13	14	15	14
AUTOGROW4	14	13	12	13
POCKET2MOL	11	11	8	11
POCKETFLOW	15	15	14	15
RESGEN	12	12	11	12
DST	2	2	2	1
GRAPH GA	6	5	6	5
MIMOSA	2	2	2	1
MOLDQN	16	16	16	16
PASITHEA	2	4	4	4
REINVENT	4	10	13	10
SCREENING	5	1	1	3
SELFIES VAE BO	8	8	9	8
SMILES GA	10	6	5	6
SMILES LSTM HC	7	7	7	6
SMILES VAE BO	8	8	9	8

Table 19. Model Rankings based on SA

MODEL	TOP 1 RANK	TOP 10 RANK	TOP 100 RANK	OVERALL RANK
3DSBDD	12	6	1	6
AUTOGROW4	11	9	8	8
POCKET2MOL	14	14	16	16
POCKETFLOW	13	13	15	14
RESGEN	15	16	9	13
DST	9	11	13	11
GRAPH GA	16	15	10	14
MIMOSA	9	11	14	12
MOLDQN	3	1	2	1
PASITHEA	9	8	11	8
REINVENT	4	5	3	5
SCREENING	7	10	12	10
SELFIES VAE BO	1	3	5	2
SMILES GA	5	2	4	4
SMILES LSTM HC	6	7	7	7
SMILES VAE BO	1	3	5	2

Table 20. Model Rankings based on docking score

MODEL	TOP 1 RANK	TOP 10 RANK	TOP 100 RANK	OVERALL RANK
3DSBDD	12	12	13	12
AUTOGROW4	2	1	1	1
POCKET2MOL	1	2	2	2
POCKETFLOW	11	14	14	13
RESGEN	3	3	7	4
DST	6	5	4	5
GRAPH GA	9	9	9	9
MIMOSA	5	6	5	6
MOLDQN	16	16	16	16
PASITHEA	7	7	6	7
REINVENT	15	15	15	15
SCREENING	4	4	3	3
SELFIES VAE BO	10	10	11	10
SMILES GA	14	13	12	13
SMILES LSTM HC	8	8	8	8
SMILES VAE BO	13	11	10	11

Table 21. Model Rankings based on average Molecule Generation Metrics across all target proteins

MODEL	DIVERSITY	VALIDITY	UNIQUENESS
3DSBDD	0.83	0.63	0.63
AUTOGROW4	0.84	1.00	0.29
POCKET2MOL	0.86	1.00	1.00
POCKETFLOW	0.90	1.00	0.87
RESGEN	0.83	1.00	1.00
DST	0.88	1.00	1.00
GRAPH GA	0.91	1.00	1.00
MIMOSA	0.88	1.00	1.00
MOLDQN	0.91	1.00	1.00
PASITHEA	0.89	1.00	1.00
REINVENT	0.88	1.00	1.00
SCREENING	0.88	1.00	1.00
SELFIES VAE BO	0.88	1.00	1.00
SMILES GA	0.88	1.00	1.00
SMILES LSTM HC	0.89	1.00	1.00
SMILES VAE BO	0.88	1.00	1.00

MODEL	DIVERSITY RANK	VALIDITY RANK	UNIQUENESS RANK
3DSBDD	14.00	2.00	3.00
AUTOGROW4	13.00	1.00	4.00
POCKET2MOL	12.00	1.00	1.00
POCKETFLOW	3.00	1.00	2.00
RESGEN	15.00	1.00	1.00
DST	7.00	1.00	1.00
GRAPH GA	2.00	1.00	1.00
MIMOSA	8.00	1.00	1.00
MOLDQN	1.00	1.00	1.00
PASITHEA	4.00	1.00	1.00
REINVENT	10.00	1.00	1.00
SCREENING	9.00	1.00	1.00
SELFIES VAE BO	6.00	1.00	1.00
SMILES GA	11.00	1.00	1.00
SMILES LSTM HC	5.00	1.00	1.00
SMILES VAE BO	6.00	1.00	1.00

Table 22. Number of molecules generated under given 96 hours.

MODEL	1IEP	3EML	3NY8	4RLU	4UNN	5MO4	7L11
3DSBDD	1002	715	753	826	900	616	589
AUTOGROW4	1429	1438	1319	1272	1552	1496	1301
POCKET2MOL	1038	1020	900	900	841	900	900
POCKETFLOW	1000	1000	1000	1000	1000	1000	1000
RESGEN	800	800	800	800	322	369	527
DST	1001	1001	1001	1001	1001	1001	1001
GRAPH GA	700	1001	700	700	300	1001	1001
MIMOSA	1001	1001	1001	1001	1001	1001	1001
MOLDQN	501	501	501	501	501	501	501
PASITHEA	800	1000	800	800	1000	1000	1000
REINVENT	100	100	100	100	100	100	100
SCREENING	1000	1000	1000	1000	1000	1000	1000
SELFIES VAE BO	200	200	200	200	200	200	200
SMILES GA	525	441	615	618	808	710	376
SMILES LSTM HC	501	501	501	501	501	501	501
SMILES VAE BO	200	200	200	200	200	200	200