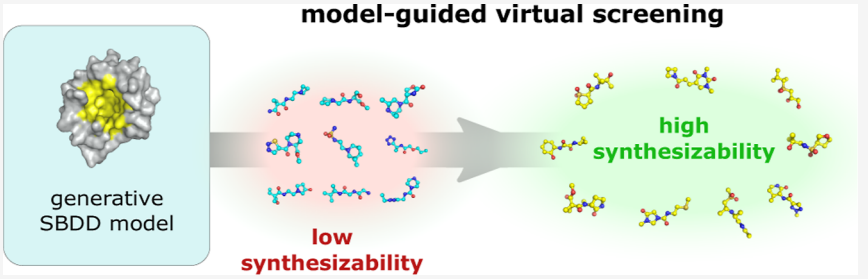
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim
Article

Circumventing the Synthesizability Problem in Generative Molecular Design

Jesse A. Weller, Jinsen Li, Yibei Jiang, and Remo Rohs*

Cite This: https://doi.org/10.1021/acs.jcim.6c00998
Read Online



ABSTRACT: Generative structure-based drug design (SBDD) models have shown great promise for accelerating our ability to discover novel drug candidates. However, these models have been criticized for producing compounds that are not very synthesizable and therefore not practically applicable to drug design. In this work, we propose a way to circumvent the synthesizability issue by introducing a model-guided virtual screening (MGVS) pipeline that pairs SBDD models with efficient chemical similarity search methods to identify synthesizable analogues of generated compounds in existing ultralarge compound databases. Using this approach, we demonstrate that synthesizable analogues of generated compounds with equivalent or better docking scores and similar predicted binding poses can be reliably identified across a wide range of protein targets. We find that MGVS outperforms standard virtual ligand screening (VLS), consistently yielding at least a 25x improvement in docking-based screening efficiency across three different SBDD models. As drug-like chemical spaces continue to grow and standard VLS methods focused on exhaustive screening become increasingly impractical, approaches such as MGVS that effectively narrow the search space will become critical for advancing drug discovery.

INTRODUCTION

Generative deep learning methods for small-molecule drug design have captured widespread attention in both academia and industry. These methods span a range of model classes, including variational autoencoders, autoregressive models, diffusion models, and flow-based models. Generative approaches promise to accelerate early stage drug discovery by identifying novel chemical structures tailored to specific biological targets, potentially unlocking areas of chemical space unexplored by traditional methods. Despite their promise, these models face a critical challenge that limits their practical utility: the synthesizability of the generated compounds. This is a particular issue in the early stages of the drug discovery process, where the ability to rapidly procure and validate potential hit compounds is essential, and custom synthesis of novel chemical structures at scale is impractically slow and expensive.

For this reason, traditional virtual ligand screening (VLS) approaches have focused on screening commercially available libraries of compounds with known synthesis routes.[1−4] Historically, such libraries have been limited to only a few million compounds, whereas drug-like chemical space is on the order of 10^{60} compounds.[5] This limitation has been partially addressed with the development of virtual chemical spaces[1]—enumerated or reaction-defined collections of make-on-demand compounds that can be synthesized reliably at scale. Public databases now include billions[2] to trillions[6] of compounds, constructed by applying known reactions to commercially available building blocks.

While these advances have dramatically expanded the accessible chemical search space, they have also introduced the daunting computational challenge of screening chemical spaces of this scale. Parallel advancements in VLS methods such as synthon-based screening[7,8] have lowered the computational overhead by systematically docking chemical building blocks instead of every reachable compound, thus helping to mitigate the combinatorial explosion. As virtual spaces continue to grow, however, such knowledge-free approaches

Received: April 1, 2026
Revised: June 30, 2026
Accepted: July 2, 2026

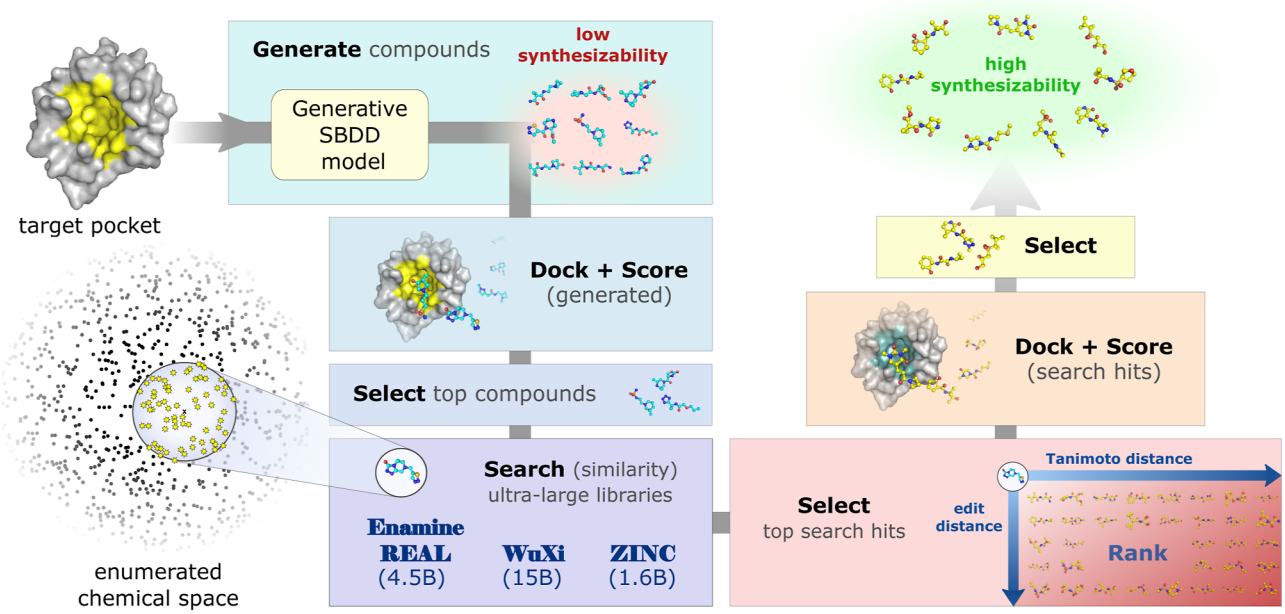


Figure 1. Overview of de novo generation and synthesizable analogue search pipeline. (1) Target pocket information is input into the generative SBDD model and used to generate 1000 compounds. (2) Generated compounds are docked to the target pocket and scored for predicted binding affinity. (3) Compounds are filtered based on chemical structure and properties, and top-10 scoring compounds are selected for analogue search. (4) For each generated query compound, efficient hierarchical graph edit distance (GED) similarity search is performed to identify the 100 most similar compounds from existing ultralarge libraries Enamine REAL, ZINC, and WuXi GalaXi. (5) Top-100 most similar search hits per query are docked to the target pocket and scored for predicted binding affinity. The best synthesizable analogues for each generated query are selected based on docking.

that rely on exhaustively screening the chemical space will become increasingly impractical. In recognition of this challenge, recent efforts have turned toward advanced sampling strategies—ranging from traditional heuristics to modern AI-based methods[1,9,10] to narrow the search space by screening a small fraction of the chemical space and using the most promising candidates to guide further search. However, these approaches remain constrained by their dependence on large-scale enumeration and iterative evaluation, which limits their efficiency and scalability as chemical spaces expand.

Generative deep learning models have recently emerged as a compelling alternative. These models can learn distributions of chemical structures and protein−ligand interactions[11−13] from large chemical data sets, offering the potential to navigate chemical space without explicit enumeration.[11,13] When conditioned on structural information from a protein target, these models can generate compounds that fit within the protein pocket and exhibit favorable binding properties. Despite their promise, generative methods have been criticized for producing compounds that are either synthetically inaccessible or chemically implausible.[14−16] In response, there has been a focus on restricting models to explicitly learn distributions over synthesizable subspaces,[17−19] but these methods often result in making trade-offs in terms of molecular diversity and controllability. In addition, the concept of a "synthesizable" compound is not well-defined ab initio in terms of chemical structure or properties but instead depends on the ever-changing set of available chemical synthesis ingredients and techniques. Restricting models to such arbitrary subspaces is not straightforward and could detract from the primary goal: generating true target-specific binders. Alternatively, we hypothesize that these synthesizability issues could be circumvented by pairing generative models with established chemical similarity search techniques to identify synthesizable analogues. In this way, rather than attempting to supplant the traditional discovery pipeline by directly generating ideal, synthesizable drug candidates, generative methods may be more effectively used to steer discovery toward tractable, high-potential subspaces—within which synthesizable compounds can be identified.

In this work, we introduce a model-guided virtual screening (MGVS) pipeline that involves using structure-based generative models to sample compounds conditioned on a target pocket, selecting the compounds with the strongest binding affinity, and retrieving similar compounds from existing chemical databases. We show that synthesizable compounds with strong predicted binding affinity and similar docking poses can be reliably identified using this generate-then-retrieve approach and that the quality of the discovered candidates represents at least a 25x improvement in docking-based screening efficiency over standard VLS. These results are consistent across three state-of-the-art generative SBDD models with diverse architectures: DrugHIVE[11] (hierarchical VAE), Pocket2Mol[12] (autoregressive), and DiffSBDD[13] (diffusion), exhibiting broad applicability of this approach. In summary, this work demonstrates that (1) existing generative SBDD models reliably identify high-potential chemical subspaces and thus narrow the search for high-quality candidates and (2) though not always directly synthesizable, generated compounds tend to have synthesizable analogues identifiable via similarity search. The synthesizability issue, therefore, does not represent a significant obstacle to their effective application to drug discovery. Improving the capacity of these models to generate high-quality target-specific binders, regardless of synthesizability, could further increase the effectiveness of MGVS over traditional screening methods.

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

METHODS AND DATA

An overview of our MGVS pipeline for de novo molecular generation, followed by a synthesizable analogue search, is shown in Figure 1. Compounds are generated using three top SBDD models with a range of architectures: DrugHIVE,[11] a density-based hierarchical variational autoencoder (VAE) model; Pocket2Mol,[12] a graph-based equivariant autoregressive model; and DiffSBDD,[13] a graph-based equivariant diffusion model. For each model and target, the following five-step MGVS process was used: (1) the target pocket information is provided to the generative SBDD model and used to generate 1000 compounds using default settings. (2) Compounds are then docked to the target pocket and scored using QuickVina2.[20] (3) Compounds are filtered to remove those with PAINS[21] patterns, poor drug-like properties, or irregular structures. The top-10 scoring compounds that remain are selected as queries for analogue search. (4) For each of the top-10 generated query compounds, hierarchical graph edit distance (GED) similarity search is performed using SmallWorld[22] to identify the 100 most similar compounds from existing ultralarge libraries: Enamine REAL, WuXi GalaXi, and ZINC. (5) These compounds are docked to the target pocket and scored for predicted binding affinity using QuickVina2.[20] The best compounds, representing synthesizable analogues, are identified for each generated query based on docking score.

The goal of this workflow is to use generative models to narrow the search to promising regions of chemical space and then identify synthesizable compounds within these regions. In this approach, generated compounds are treated as target-specific hypotheses rather than final candidates. Similarity search is used to find synthesizable analogues with related molecular topology, and docking is then used to test whether these analogues maintain favorable predicted binding scores relative to the generated queries. However, compounds with similar structures and comparable docking scores may still bind to different poses. For this reason, we use protein−ligand interaction analysis to assess whether retrieved analogues tend to preserve similar predicted binding modes. In this way, MGVS does not rely on docking score alone, but instead combines similarity search, docking, and interaction analysis to evaluate whether synthesizable analogues are both high-scoring and structurally related to the generated query compounds.

Compound Filtering

To avoid inaccurate docking results, we filter out molecules with irregular configurations that lead to strained geometries similar to ref 11. We remove molecules with consecutive double bonds, fused ring systems of more than four members or that form loops, and rings other than five- and six-membered rings. For consistency, we apply this filtering to all compound sets.

Synthesizable Analogue Search

We perform a GED search using SmallWorld,[22] which is the default tool used for similarity search in the ZINC database.[3] In principle, MGVS can be implemented with any similarity search method capable of efficiently retrieving compounds from large, synthesizable chemical libraries. In this work, we use SmallWorld because it enables rapid search over ultralarge libraries based on molecular topology. SmallWorld represents molecules as graphs and identifies compounds that can be transformed to the query structure with a small number of graph edits. These graph edits correspond to structural changes, such as atom or bond substitutions, deletions, and other topological modifications. To enable efficient search, SmallWorld uses preindexing of molecular graph space. SmallWorld also reports Daylight and ECFP4 fingerprint distances, which we used as additional similarity metrics for analysis.

The practical accessibility of retrieved analogues depends on the compound database used for similarity search. MGVS can be applied to any sufficiently searchable compound collection, but retrieved compounds should only be interpreted as readily accessible when the search is restricted to purchasable, make-on-demand, or otherwise synthesis-annotated chemical libraries.

Conformer Generation

Starting from the SMILES string representation of a molecule, we generated five conformers (3D coordinates) using the ETKDG algorithm in RDKit.[23] To reduce docking of redundant structures, we performed Butina clustering using RMSD as the distance metric on the five conformers, keeping only the centroid for each cluster. We then dock each remaining conformer to the corresponding target and keep the best-scoring conformer as a representative.

Ligand Docking

Docking compounds to target pockets is performed with QuickVina2[20] (with default settings and a 20 Å wide bounding box), a computationally efficient version of AutoDock Vina.[24] Protein structures are prepared for docking using AutoDockTools.[25] Prior to docking, we perform force field optimization on all molecular structures using the MMFF94 force field[26] in RDKit.[23] This step is especially important with de novo generated structures, which are likely to have unrelaxed conformations. Using unrelaxed structures before docking could lead to significantly exaggerated docking scores.

Evaluation

For the evaluation, we chose 30 protein targets from the PDBbind general set. Our selection process includes (1) clustering all protein sequences by 30% sequence identity, (2) randomly sampling 30 clusters, and (3) choosing one random target per cluster. We generated 1000 molecules for each of the receptors in the test set using each of the three SBDD models. As a reference set for random virtual screening, we randomly sampled 50k compounds from the ZINC20 drug-like subset. We performed force field optimization on each of the molecules using MMFF94 in RDKit[23] before virtually docking to the corresponding receptor. However, the Vina score is strongly correlated with molecule size,[11] which prevents direct comparison between sets of compounds with different size distributions. For this reason, we primarily report Vina efficiency or Vina score divided by the number of heavy atoms, as it corrects for this size bias.

Evaluation Properties and Metrics

Binding affinity is estimated for each ligand with the Vina docking score calculated with QuickVina2,[20] a speed-optimized version of AutoDock Vina.[24] We define ΔVina as the difference between the Vina score of a generated ligand and a reference ligand (e.g., query ligand or ligand from crystal structure). Vina efficiency score (or ligand efficiency) is calculated as the Vina score divided by the number of heavy atoms. Graph Edit Distance (GED) is the number of edits needed to be made to the chemical structure of a compound to make it identical to a reference compound, and in this work is

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

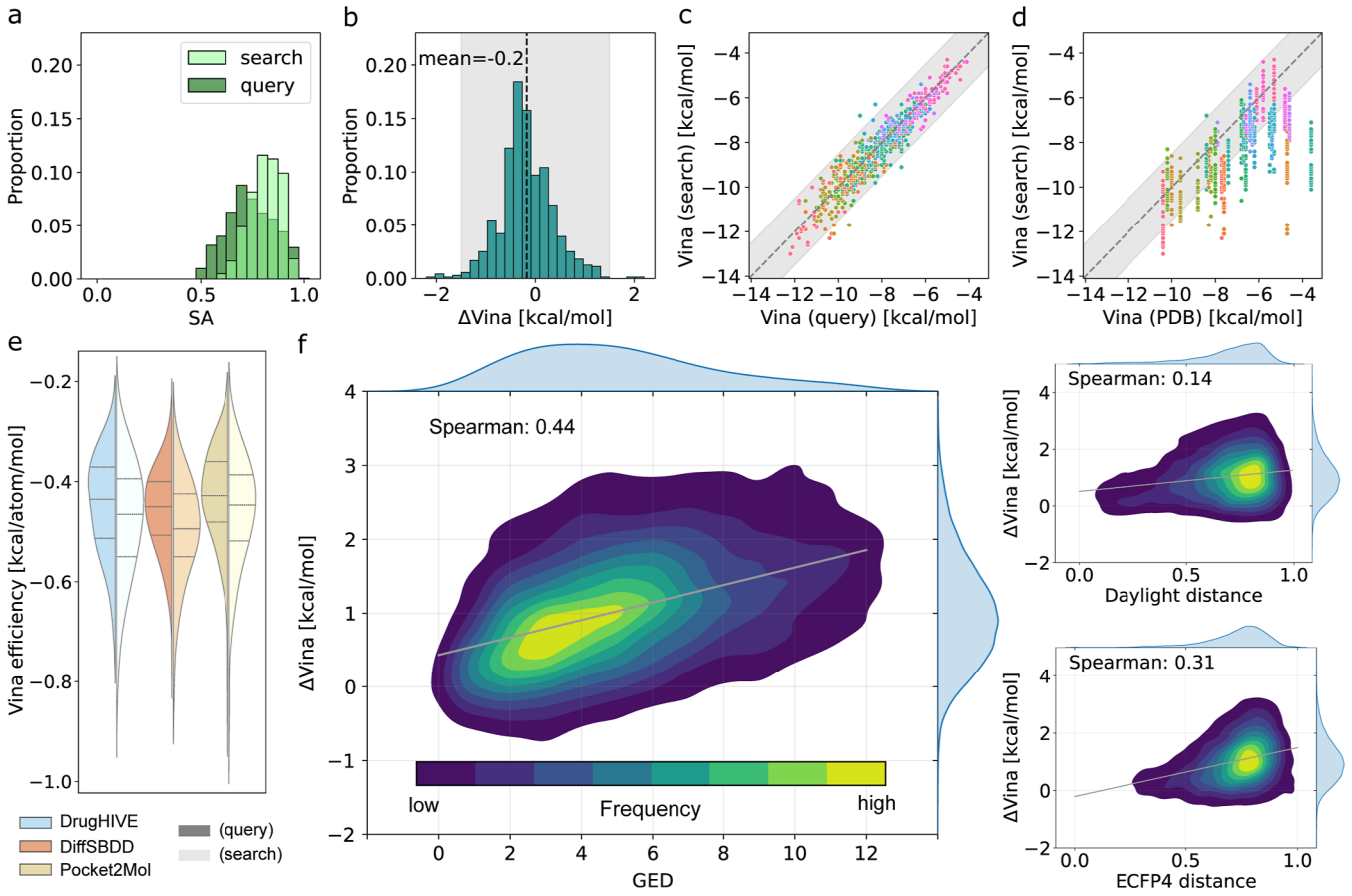


Figure 2. Generated compounds vs search-hit compounds. (a) Histogram showing distribution of Synthetic Accessibility (SA) score for generated query and search-hit compounds. (b) Histogram of ΔVina* scores for top-1 search-hit compounds with mean (black dashed line) and Vina uncertainty (shaded). (c) Vina scores for top-1 search-hit compounds versus the corresponding query compound for each target (colors) with Vina uncertainty (shaded). (d) Vina scores for top-1 search-hit compounds versus the corresponding crystal ligand for each target (colors) with Vina uncertainty (shaded). (e) Distributions of Vina efficiency for queries (dark) and top-1 search-hits (light) for each model (colors). (f) Distribution of ΔVina scores for all (top-100) search hits with respect to graph edit distance (GED), Daylight distance, and ECFP4 distance. Linear regression fit (gray line) shown along with computed Spearman's rank (ρ) correlation coefficient. *ΔVina = $\text{Vina}_{\text{search-hit}} - \text{Vina}_{\text{query}}$.

calculated using SmallWorld.[22] Daylight distance is calculated as one minus the Tanimoto similarity between the Daylight fingerprints of a pair of molecules. ECFP4 similarity is calculated as one minus the Tanimoto similarity between the ECFP4 fingerprints of a pair of molecules. The Quantitative Estimate of Drug-Likeness (QED) score estimates the drug-likeness of a molecule by combining a set of molecular properties.[28] The Synthetic Accessibility (SA) score estimates the ease of synthesis of a molecule.[29]

RESULTS

Generated Compounds Tend to Have Synthesizable Analogues

We start by generating approximately 1000 molecules for each of our test target pockets using three SBDD generative models: DrugHIVE,[11] DiffSBDD,[13] and Pocket2Mol.[12] We then dock each generated molecule into the corresponding target pocket using QuickVina2.[20] To select our final generated set, we filter out molecules with PAINS patterns, unusual structures that could cause inaccurate docking results (see Compound Filtering), or properties outside of the typical drug-like range.[30] We then rank each compound by normalized Vina score and take the 10 best compounds for each model and target. The properties of all generated molecules compared to the final set can be seen in Figure S1.

For each compound in the final generated query set, we perform a similarity search over existing commercially available and readily synthesizable compound libraries. For each generated query compound, we use the SmallWorld[22] API to search for up to 1000 closest compounds by GED within a maximum GED of 12, for each of the ZINC,[3] Enamine REAL,[2] and WuXi GalaXi[3] libraries. We then rank the returned search compounds by GED and then Daylight distance, selecting the top-100 for our final search hit set. We then generate 3D conformers and dock each final search-hit compound to the corresponding protein pocket.

The resulting top search-hit compounds are highly synthesizable and even tend to have improved predicted binding scores compared to the corresponding generated query compounds. As expected, we find a drastic improvement in predicted synthesizability of the top search-hit compounds when compared to the generated query compounds (Figure 2a). We also find an average improvement in predicted binding affinity compared to the query compounds (Figure 2b), with virtually all (98.7%) search-hit compounds within the Vina-estimated margin of error of ±1.5 kcal/mol (Figure 2c). We also see that virtually all of the top search-hit compounds have

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

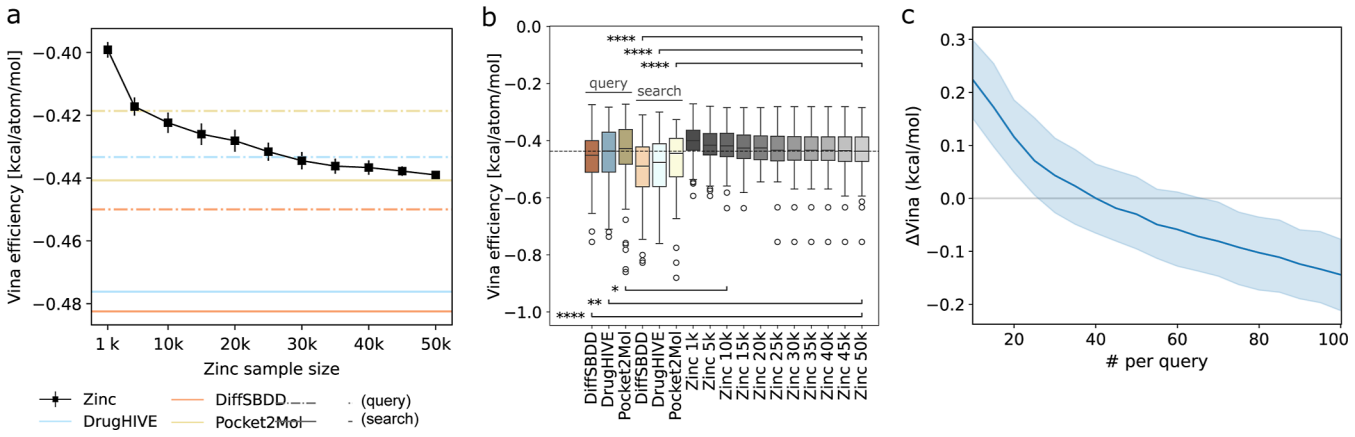


Figure 3. Comparison with random ZINC compounds. (a) Median Vina efficiency of the top-10 compounds for different size subsets of the random ZINC set relative to query (dashed) and search (solid) hit compounds for each model (color). Calculations for ZINC subsets are repeated a total of $n = 20$ times, with median average and std. dev. (b) Boxplots show the distributions of top-10 Vina efficiency for generated query compounds, search hits, and ZINC (random) subsets of varying sizes. For each query or search-hit distribution, the largest ZINC subset distribution with significant Welch's t test result is indicated. Significance denoted as $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****). (c) Average ΔVina between top-10 scoring search-hit compound and query compound vs number of total ranked search hits that were docked to the target pocket, with 95% confidence interval (shaded).

equivalent or better predicted binding affinity compared to the corresponding PDB co-crystal ligand (Figure 2d), with a large number (38.8%) achieving significantly better scores (ΔVina < − 1.5 kcal/mol). For all three models tested, we see a similar distribution shift toward improved predicted binding efficiency in the search-hit compounds, with an average improvement of the median Vina efficiency (Figure 2e).

In Figure 2f, we show a positive Spearman's rank correlation between GED and worse binding affinity scores ($\rho = 0.44$). We observe similar, though weaker, correlations for ECFP4 distance ($\rho = 0.31$) and Daylight distance ($\rho = 0.14$). Since lower similarity search hits from these databases tend to have worse binding scores than the query, this suggests that the generated query compounds indeed represent good templates for a successful predicted binder. Further, GED appears to be a better predictor of binding score than the two fingerprint-based similarity measures. This is likely due to GED being a better measure of topological and structural similarity, which are key determinants of shape complementarity and interaction geometry within the binding pocket.

Generative SBDD Paired with Analogue Search Outperforms Random Screening

To be practically useful, a generative approach combined with similarity search needs to outperform random virtual screening. To assess whether this is the case, we compare our generative approach results with the screening of random compounds from the ZINC[31] database. We randomly sampled 50k compounds from the ZINC drug-like compound subset and docked all compounds into each of our test pockets. In Figure 3a,b, we compare the Vina efficiency of the top-10 compounds from the generated queries, search hits, and different-sized random ZINC subsets (ranging from 1k to 50k). The query compound distributions, which were selected from 1k generated compounds for each target, show significantly better Vina efficiency (Welch's t test $p < 0.05$) than screening 10k (Pocket2Mol), 30k (DrugHIVE), and 50k (DiffSBDD) ZINC compounds. The search-hit compound distributions for all models show significant improvement (Welch's t test $p < 0.05$) compared to their respective query compound distributions. Across all models, the top-10 similarity search hits score better than those from screening up to 50k random ZINC compounds, despite only screening 2k compounds (including queries) per target. This corresponds to a 25x improvement in docking-based screening efficiency relative to the random VLS baseline. The full computational cost of MGVS also depends on generative model inference, similarity search, molecular preparation, and hardware implementation. Using per-step timings measured in this work, we estimate that DrugHIVE-based MGVS provides an approximately 22x runtime improvement relative to docking 50k random ZINC compounds (Table S1).

These results represent selection from a docked candidate pool of the 100 most similar search hits per query compound, however, docking and ranking fewer search hits could lower the computational cost while still achieving favorable results. To assess this, we reran selection using different size search-hit candidate pools (docked search hits) ranging between 1 and 100. In Figure 3c, we show the resulting ΔVina averaged across all models and target ranges from −0.1 to 0.2, with parity (ΔVina = 0) at approximately 40 search hits docked per query. Surprisingly, even when using only a single search hit, the upper limit of the ΔVina 95% confidence interval remains at a modest +0.3 kcal/mol—well within the estimated ±1.5 kcal/mol Vina uncertainty.[24] This demonstrates that synthesizable analogues with comparable docking scores can be reliably identified for generated compounds even with relatively few search hits screened per query.

Shared Intermolecular Interactions between Predicted Binding Poses

Our results so far have shown that using similarity search to identify synthesizable analogues to de novo-generated compounds reliably results in highly similar compounds with strong predicted binding affinity. However, this does not guarantee that the search-hit compounds have a similar predicted binding pose within the target pocket. To assess this, we compared the docked intermolecular interactions made by the top predicted binding pose of the search-hit compounds to those made by the corresponding generated query compound. Interactions were identified using the Protein−Ligand Interaction Profiler (PLIP)[27] on the top-

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

ranking docked pose from QuickVina2[20] for each compound. Because PLIP interactions were used here to assess conservation of predicted binding poses, rather than to exhaustively annotate all possible contacts, we used stricter cutoffs than the PLIP defaults to emphasize higher-confidence, geometrically well-defined interactions and reduce the contribution of weak or marginal contacts in docked poses. In Table 1, we summarize the criteria used to identify each type of interaction.

Table 1. Protein−Ligand Interaction Cutoff Criteria[a]

H-bond	hydrophobic	salt bridge	π-stack (parallel)	π-stack (perpendicular)	π-cation
3.5 Å, 120°	3.8 Å	4.0 Å	4.0 Å	5.5 Å	4.0 Å

[a]For each interaction type, we list the PLIP[27] non-default cutoff criteria that we use for identifying intermolecular interactions. All table values are distance cutoffs, except for hydrogen bonds. For hydrogen bonds, we choose a stricter distance and D-H-A angle cutoff. In the case of π interactions, the table values refer to center−center or center−cation distance.

We consider matching interactions between query compound and search-hit compound at two levels of specificity: (1) same target residue and (2) same specific target atom. In Table 2, we show a summary of the proportion of search hits having shared interactions with the generated query compound for each interaction type (H-bond, hydrophobic, salt bridge, π-cation, and π-stacking) across all models and targets. We observe that a majority of search hits share at least one interaction with the generated query across all interaction types. We also see that search hits are less likely to share the more frequent H-bond and hydrophobic interactions than the less frequent ones, except for π-stacking, which is both frequent and highly conserved.

In Figure 4, we show the proportion of docked search-hit compounds that share at least one or all specific query compound interactions for each generative model. Here we choose to focus on specific nonhydrophobic interactions (hydrogen bonds, π interactions, and salt bridges) because we are primarily interested in assessing how often search hits exhibit the same type of target specificity as the query. The same statistics for all interactions can be found in Figure S2. In Figure S3, we show the distribution of intermolecular interactions made by all generated query compounds. The average number of intermolecular interactions made by query compounds is 2.5 (specific) and 5.2 (all). Across all query compounds, models and targets, an average of 19.5% (atom match) to 31.4% (residue match) of search hits (top-100) share all specific interactions with the query and 76.7% (atom match) to 79.5% (residue match) share at least one (Figure 4a). For top-1 search hits (Figure 4b), we see that a majority of queries have at least one search hit sharing all specific interactions (50.9% and 80.7% for atom and residue match, respectively) and nearly all queries have at least one search hit that shares an interaction (99.2% for both atom and residue match). In Figure 4c, we show that there is some variation in the proportion of queries with search-hit compounds matching all interactions across the different protein targets. However, even for the lowest performing target (PDB ID 5G1P), 10.5% (atom match) to 20.0% (residue match) of query compounds have an exact interaction analogue. For all targets, nearly every query compound has a search hit with at least one interaction.

To give a better sense of what the docked poses of interaction analogues look like, in Figure 4d, we show a selection of docked poses with highlighted intermolecular interactions for query and search-hit pairs with all interactions shared (atom match) for four of the targets. We report the Vina scores of each compound, the number of shared interactions, the database of the search hit, and the molecular distance metrics. We observe that search hits with many shared query contacts tend to also have similar predicted binding poses and a relatively low GED. Interestingly, fingerprint-based distance metrics do not always agree with GED (or with each other). For example, the pair of compounds displayed for target PDB ID3JUZ have clear structural similarities and a relatively low GED (3) but high Daylight distance and ECFP4 distance (0.74 for both). For the pair of compounds displayed for target PDB ID6I18, GED (3) and Daylight distance (0.29) are low, while ECFP4 distance (0.67) is high. To quantify this, we calculated the Spearman's rank correlation between each of the molecular distance metrics and the number of shared interactions (atom match). We found that GED has the highest correlation, especially as the number of protein−ligand interactions made by the query compound increases (Figure S4). These results are consistent with the earlier findings (Figure 2f) that GED is a better predictor of binding affinity score than the fingerprint-based similarity metrics.

DISCUSSION AND CONCLUSIONS

In this work, we have shown that current generative structure-based drug design (SBDD) models paired with similarity-based search can be used to discover synthesizable compounds with high predicted binding affinity for a wide range of protein targets. Generative models tend to produce compounds with low synthetic accessibility scores,[14−16] which is a key limitation to practical drug discovery.[4] To address this, we show that model-guided virtual screening (MGVS)—using generated compounds to guide the search for synthesizable analogues—effectively overcomes this limitation. We apply MGVS across

Table 2. Shared Intermolecular Interactions by Interaction Type[a]

# matched	match type	H-bond (n = 36404)	hydrophobic (n = 62582)	salt bridge (n = 1332)	π-cation (n = 474)	π-stacking (n = 29830)
all	atom	0.17	0.12	0.58	0.79	0.57
	residue	0.27	0.30	0.58	0.79	0.83
at least 1	atom	0.65	0.75	0.73	0.79	0.90
	residue	0.69	0.86	0.73	0.79	0.92

[a]Proportion of search hits having shared interactions with the generated query compound across all models and targets. Statistics are reported for search hits that share all or at least 1 of the query interactions, as well as for different positive match types: matching interactions share the same (1) target atom or (2) target residue. For each interaction type, the total number (n) of search-hit compounds with a query compound predicted to make the corresponding interaction with the target is reported.

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

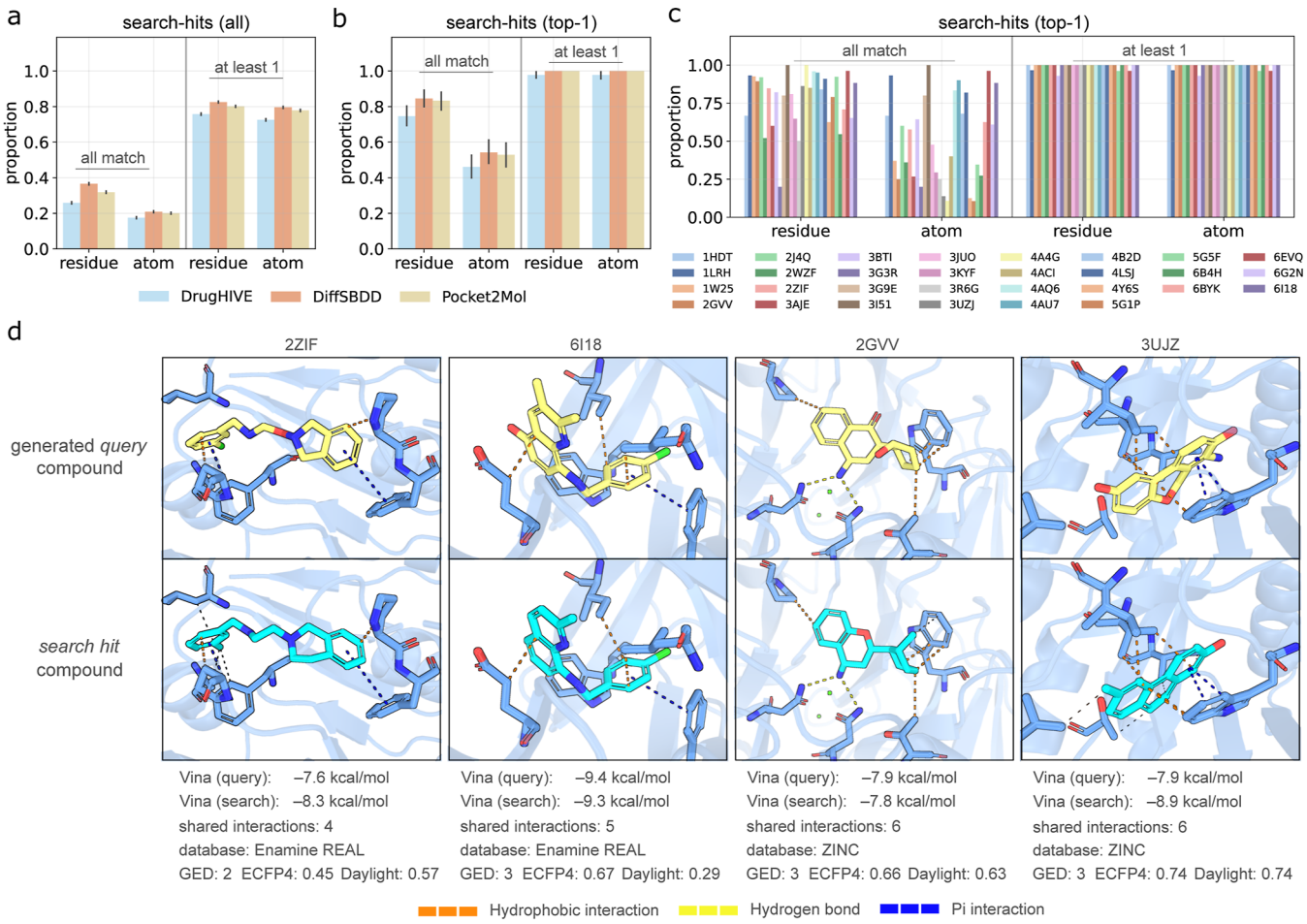


Figure 4. Shared intermolecular interactions between generated queries and search hits. (a–c) Bar plots showing proportion of search hits that share all or at least 1 of query specific protein–ligand interactions. Proportions are shown for both exact residue match (residue) and exact atom match (atom) criteria. (a) Proportion of top-100 search hits for each query with matching specific interactions for each model. (b) Proportion of top-1 search hits with matching specific interactions and query by model. (c) Proportion of top-1 search hits with matching specific interactions for each query by protein target across all models. (d) Docked poses of search hit (yellow carbon) and query (cyan carbon) compound pairs with protein–ligand interactions shown. Vina docking score, number of shared interactions, search-hit database, and chemical distance metrics (GED, ECFP4 distance, and Daylight distance) are shown for each pair.

30 diverse protein targets and observe that for three separate SBDD models (DiffSBDD,[13] DrugHIVE,[11] and Pocket2Mol[12]) screening of only 2k compounds leads to better quality candidates than standard VLS screening of 50k ZINC compounds, representing a 25x improvement in efficiency.

Although current generative models tend to produce compounds with low synthesizability, they are effective at identifying compounds located in promising regions of chemical space. Once these regions have been identified, similarity search can be used to find adjacent synthesizable analogues. Across all models and targets, we find that virtually all (98.7%) generated compounds selected for analogue search yielded a search hit with equivalent or better docking scores. We also found that search hits commonly have similar docking poses to the query compound, indicating that the compounds produced by generative models help identify promising potential binding modes. We observe that ∼50% of query compounds yielded a search hit with a docking pose sharing all specific (nonhydrophobic) residue-level protein–ligand interactions, and virtually all (∼99%) yielded a search hit sharing at least one interaction (Figure 4b). Assuming that the number of shared intermolecular interactions is an indicator of docking pose similarity (Figure 4d), this demonstrates consistent success in identifying predicted binding analogues. In future applications, this initial query selection step could be paired with additional pose-validity filtering before analogue search. For example, tools such as PoseBusters[16] are useful for identifying chemically or physically implausible generated structures or docked poses, and could therefore be applied when the reliability of the initial generative outputs is a primary concern.

Altogether, these results demonstrate that current generative models, despite their limitations, can already be used to improve virtual screening efficiency through MGVS. We show that identifying compounds with strong predicted affinity for a protein target, regardless of synthesizability, is useful for guiding the screening of high-potential subsets of existing ultralarge chemical databases. To be successful, this approach relies on search methods efficient enough to operate on these large and growing chemical spaces,[1] which could be viewed as a limitation. However, improving chemical search methods is an active area of research[32,33] that should enable the efficient querying of expanded chemical databases. As chemical databases grow and search methods improve, we would expect

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

MGVS to become even more effective because of improved retrieval. In contrast, the expansion of chemical spaces poses a real challenge to existing VLS methods that rely on exhaustive screening.[9] Even for the most efficient synthon-based VLS approaches used to screen current ultralarge libraries, scaling to chemical spaces orders of magnitude larger will require innovations that surpass what current methods can support.

MGVS could also be combined with other methods for improving the efficiency of virtual screening, as chemical spaces continue to grow. Pharmacophore- or shape-based screening methods[34,35] can be used to search large libraries when a known ligand, pharmacophore model, or other query representation is available. In MGVS, structure-based generative models provide a way to create target-specific query compounds directly from the protein pocket, which could then be used as inputs to pharmacophore- or shape-based searches. Similarly, surrogate models trained to predict docking scores could be used to prioritize compounds before explicit docking or to guide additional rounds of screening,[10,36] further reducing the computational cost of searching large libraries. These approaches are therefore compatible with MGVS and could be incorporated into future workflows to improve screening efficiency across larger chemical spaces.

In addition to improving the screening workflow around generative models, another important direction is improving the generative models themselves. There are efforts to overcome the generative SBDD synthesizability issue by designing models restricted to synthesizable subspaces.[17−19] However, we observe that even for the current (unrestricted) models that we tested, while they are able to generate high-quality binding candidates, the set of synthesizable analogues identified based on these initial queries often contain compounds with even better predicted binding. This suggests that current models cannot consistently generate the most optimal compounds within these localized regions of chemical space and restricting them further could lead to even worse performance. Efforts to improve the ability of models to generate ideal binders, regardless of synthesizability, seem to be warranted. Applying current models in MGVS already achieves a significant improvement in docking-based screening efficiency over standard virtual screening, so further improvement to models that can be utilized in this paradigm could have significant impact on early drug discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

Data used in evaluation, including molecules, molecular properties, and docking scores, are available in Supporting Information and at 10.5281/zenodo.19377708.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.6c00998.

Property distributions of generated compounds; shared intermolecular interactions between search queries and hits for all interaction types (including hydrophobic); distributions of number of protein−ligand interactions for docked poses across all generated query compounds; correlation of molecular similarity metrics with number of shared interactions between query and search-hit compounds; and all data necessary to reproduce our results including molecular structure files, molecular property files, docking output files, and protein−ligand interaction analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Remo Rohs − *Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California 90089, United States; Department of Physics & Astronomy, Department of Chemistry, Thomas Lord Department of Computer Science, and Alfred E. Mann Department of Biomedical Engineering, University of Southern California, Los Angeles, California 90089, United States; Division of Medical Oncology, Department of Medicine, University of Southern California, Los Angeles, California 90033, United States;* orcid.org/0000-0003-1752-1884; Phone: +1 213 740 0552; Email: rohs@usc.edu

Authors

Jesse A. Weller − *Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California 90089, United States; Department of Physics & Astronomy, University of Southern California, Los Angeles, California 90089, United States;* Present Address: Dyno Therapeutics, Inc., 343 Arsenal Street, Watertown, MA 02472, USA; orcid.org/0000-0001-7184-4241

Jinsen Li − *Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California 90089, United States*

Yibei Jiang − *Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California 90089, United States;* Present Address: Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.6c00998

Author Contributions

Project conception (J.A.W. and R.R.), lead project design (J.A.W.), supporting project design (J.L., Y.J., and R.R.), data generation and analysis (J.A.W.), manuscript writing (J.A.W.), manuscript edits (J.A.W., J.L., Y.J., and R.R.), and project supervision and funding (R.R.).

Funding

This work was supported by the National Institutes of Health [grant R35GM130376 to R.R.], a University of Southern California Office of Research and Innovation SBIR/STTR Planning Award [to R.R.], the Andrea and Mark Gabbay Research Fund [to R.R.] and Tushar Chandra and Rohini Sabikhi Research Fund [to R.R.]. Funding for the open access publication fee was provided by the USC Office of Research and Innovation SBIR/STTR Planning Award [to R.R.].

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge helpful discussions with other members of the Rohs lab.

ABBREVIATIONS

ALogP	octanol−water partition coefficient (Crippen)
HBA	hydrogen-bond acceptors
HBD	hydrogen-bond donors
HTS	high-throughput screening
MGVS	model-guided virtual screening
PAINS	pan-assay interference compounds
PDB	protein data bank
QED	quantitative estimate of drug-likeness
RMSD	root-mean-square distance
SA	synthetic accessibility
SAR	structure−activity relationship
SBDD	structure-based drug design
SMILES	simplified molecular input line entry system
TPSA	topological polar surface area
VAE	variational autoencoder
VLS	virtual ligand screening

REFERENCES

(1) Sadybekov, A. V.; Katritch, V. Computational approaches streamlining drug discovery. *Nature* **2023**, *616*, 673−685.

(2) Grygorenko, O. O.; et al. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23*, 101681.

(3) Tingle, B. I.; et al. ZINC-22−A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J. Chem. Inf. Model.* **2023**, *63*, 1166−1176.

(4) Lam, J. H.; Katritch, V. Navigating structure-based drug discovery with emerging innovations in physics- and knowledge-based approaches. *npj Drug Discovery* **2025**, *2*, 29.

(5) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(6) Neumann, A.; Marrison, L.; Klein, R. Relevance of the Trillion-Sized Chemical Space "eXplore" as a Source for Drug Discovery. *ACS Med. Chem. Lett.* **2023**, *14*, 466−472.

(7) Sadybekov, A. A.; et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601*, 452−459.

(8) Beroza, P.; Crawford, J. J.; Ganichkin, O.; Gendelev, L.; Harris, S. F.; Klein, R.; Miu, A.; Steinbacher, S.; Klingler, F. M.; Lemmen, C. Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nat. Commun.* **2022**, *13*, 6447.

(9) Klarich, K.; Goldman, B.; Kramer, T.; Riley, P.; Walters, W. P. Thompson Sampling−An Efficient Method for Searching Ultralarge Synthesis on Demand Databases. *J. Chem. Inf. Model.* **2024**, *64*, 1158−1171.

(10) Cavasotto, C. N.; Di Filippo, J. I. The Impact of Supervised Learning Methods in Ultralarge High-Throughput Docking. *J. Chem. Inf. Model.* **2023**, *63*, 2267−2280.

(11) Weller, J. A.; Rohs, R. Structure-Based Drug Design with a Deep Hierarchical Generative Model. *J. Chem. Inf. Model.* **2024**, *64*, 6450−6463.

(12) Peng, X.; et al. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets, **2022**, arXiv.2205.07249.

(13) Schneuing, A.; et al. Structure-based Drug Design with Equivariant Diffusion Models. *Nat. Comput. Sci.* **2024**, *4*, 899−909.

(14) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60*, 5714−5723.

(15) Gao, W.; Luo, S.; Coley, C. W. Generative Artificial Intelligence for Navigating Synthesizable Chemical Space. *Proc. Natl. Acad. Sci. U.S.A.* **2025**, *122*, No. e2415665122.

(16) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **2024**, *15*, 3130−3139.

(17) Cretu, M.; et al. SynFlowNet: Design of Diverse and Novel Molecules with Synthesis Constraints. In *Machine Learning for Structural Biology Workshop*; NeurIPS, 2025.

(18) Koziarski, M.; et al. RGFN: Synthesizable Molecular Generation Using GFlowNets. *Advances in Neural Information Processing Systems 37*; Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025.

(19) Seo, S.; et al. Generative Flows on Synthetic Pathway for Drug Design. *arXiv* **2024**, arXiv.2410.04542.

(20) Alhossary, A.; Handoko, S. D.; Mu, Y.; Kwoh, C.-K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics* **2015**, *31*, 2214−2216.

(21) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(22) SmallWorld. NextMove Software. https://www.nextmovesoftware.com/smallworld.html (2025) (accessed 5/7/25).

(23) RDKit: Open-Source Software. https://www.rdkit.org (2022) (accessed 5/7/25).

(24) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(25) Morris, G. M.; et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(26) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(27) Adasme, M. F.; et al. PLIP 2021: expanding the scope of the protein−ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.* **2021**, *49*, W530−W534.

(28) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90−98.

(29) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.

(30) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today:Technol.* **2004**, *1*, 337−341.

(31) 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. Mod.* **2020**, *60*, 6065−6073.

(32) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62*, 2021−2034.

(33) Bellmann, L.; Penner, P.; Gastreich, M.; Rarey, M. Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. *J. Chem. Inf. Model.* **2022**, *62*, 553−566.

(34) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(35) Giordano, D.; Biancaniello, C.; Argenio, M. A.; Facchiano, A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* **2022**, *15*, 646.

(36) Gentile, F.; et al. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6*, 939−949.

https://doi.org/10.1021/acs.jcim.6c00998
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX