JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

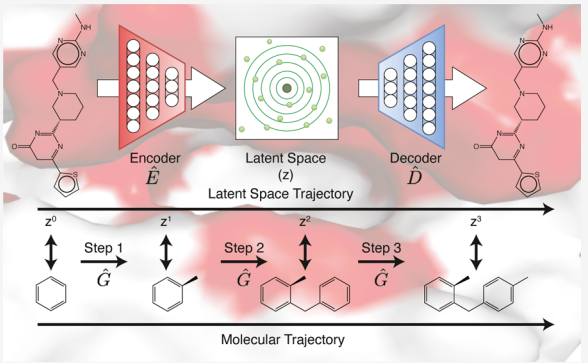
Navigating Chemical Space by Interfacing Generative Artificial Intelligence and Molecular Docking

Ziqiao Xu, Orrette R. Wauchope, and Aaron T. Frank*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 5589−5600 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Here, we report the implementation and application of a simple, structure-aware framework to generate target-specific screening libraries. Our approach combines advances in generative artificial intelligence (AI) with conventional molecular docking to explore chemical space *conditioned* on the unique physicochemical properties of the active site of a biomolecular target. As a demonstration, we used our framework, which we refer to as sample-and-dock, to construct focused libraries for cyclin-dependent kinase type-2 (CDK2) and the active site of the main protease (M^{pro}) of the SARS-CoV-2 virus. We envision that the sample-and-dock framework could be used to generate theoretical maps of the chemical space specific to a given target and so provide information about its molecular recognition characteristics.



INTRODUCTION

Developing new drugs is costly and time-consuming; by some estimates, bringing a new drug to market requires ∼13 years at a cost of ∼US $1 billion.[1] As such, there is a keen interest in accelerating the drug development pipeline and reducing its cost. In principle, computer-aided drug development can both accelerate and lower the cost of identifying viable drug candidates.[2] For instance, during hit identification, where the goal is to identify small molecules that bind to the target, one could restrict testing to only the compounds that are predicted to bind to the target with high affinity. However, the computational cost of screening large chemical libraries can still be burdensome. The computational cost of virtual screening could be reduced by working with smaller and more focused, target-specific virtual libraries, enriched in compounds that are likely to bind to a specific site of the target of interest.[3]

Such target-specific libraries can be constructed by carrying out *de novo* design, during which compounds are designed on-the-fly, guided by the unique physicochemical properties of the active site of the target.[4−17] However, because of the difficulty in formalizing diverse reaction rules and implementing robust algorithms to intelligently apply them, such *de novo* design frameworks often produce synthetically inaccessible compounds. Fortunately, advances in generative artificial intelligence now make it possible to design both chemically novel and synthetically feasible compounds in silico.[18−23] Noteworthy among these frameworks are those based on autoencoders. Autoencoders are unsupervised machine learning models that, by virtue of their architecture, are able to learn a compact, latent space representation of chemical space when trained with molecular data. The task of designing target-specific libraries can, therefore, be cast as one of exploring the latent space of such models, conditioned on the unique properties of the active site of the target of interest. Currently lacking, however, are efficient, structure-aware strategies for sampling the latent space of generative molecular models. Such techniques could find utility in constructing target- and site-specific virtual libraries that are likely to contain hit compounds that, if synthesized, might bind to the targeted site. More immediately, they could be used to construct a ligand-based pharmacophore model that can be used to screen millions of compounds in a matter of seconds.[24]

Here, we tested and implemented a novel, structure-based framework for constructing target-specific libraries by sampling the latent space of generative machine learning models *conditioned* on the unique physicochemical properties of the active site of a target. We reasoned that we could implement such a pipeline by combining generative molecular machine learning models with molecular computer docking algorithms. For instance, one could start by embedding a reference compound, e.g., benzene, into the latent space of a generative autoencoder model and sample latent space points in its neighborhood. The sampled points can then be decoded into their corresponding three-dimensional (3D) representations

Received: June 26, 2021
Published: October 11, 2021

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

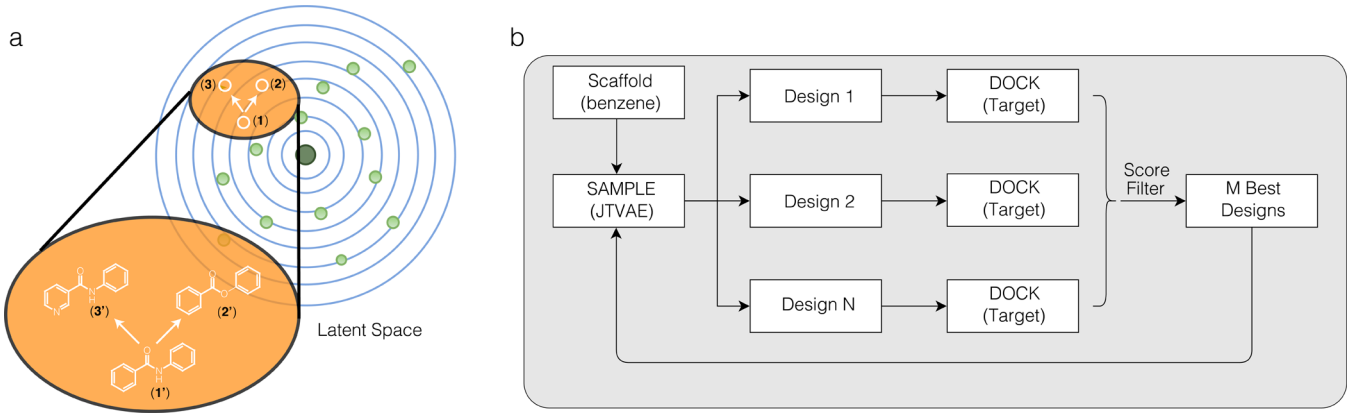


Figure 1. (a) Illustration of how latent spaces can be used to generate new molecules. For a reference point in a learnt latent space (1), which is associated with some reference molecule (1′), additional points within its neighborhood can be sampled (e.g., 2 and 3) and decoded to generate new molecules (e.g., 2′ and 3′). These new molecules can then be used for docking. (b) Illustration of the "sample-and-dock" framework, which we propose for the generation of target-specific libraries. The pipeline is initialized using a starting scaffold (e.g., benzene), which is projected into the latent space of the generative AI model (here, the pretrained junction-tree variational autoencoder (JTVAE) of Jin et al.[25]). Points in its neighborhood are sampled and decoded to generate a set of N designs. The N designs are individually docked onto the target, and the M best designs are selected based on their docking scores and projected back into the latent space to seed the next cycle. In the simplest case, $M = 1$.

and docked onto the target (Figure 1a). The sampled compound that is predicted to interact most favorably with the active site can be identified and then used as the new reference. The process can be repeated (Figure 1b), and in so doing, the latent space of a generative molecular model (and, by extensions, chemical space) can be explored in a semidirected manner, *conditioned* on the properties of the active site of the target. The compounds sampled during this iterative ("sample-and-dock") process can then be used to construct a target-specific virtual library.

METHODS

Sample-and-Dock. To construct structure-aware and target-specific virtual libraries, we implemented a *conditional* generative sampling scheme referred herein to as sample-and-dock, which combines a pretrained generative variational autoencoder (VAE) model with molecular docking (see below). In sample-and-dock, the exploration of the latent space of a pretrained generative model is made *conditional* on the unique physicochemical and biophysical properties of the binding site on a target, $\mathcal{T}$, using a greedy algorithm that optimizes a cost function that is dependent on the properties of the target. Here, we used a molecular docking scoring function as the cost function in our greedy algorithm. Accordingly, given a set of latent space samples, the "best" sample is selected as the one that, when decoded into its molecular representation, yields the highest binding affinities ($\mathcal{A}_i$) or, equivalently, the lowest docking score ($\mathcal{S}_i$), as estimated using a molecular docking scoring function.

Briefly, to construct a customized target-specific library using this conditional sampling scheme, we first embed a reference compound in the latent space of a generative model ($m_i \rightarrow z_i$) and then **sample** a set of points $\{z_{i+1}\}$ in the "neighborhood" of the reference z_i. The $\{z_{i+1}\}$ can then be decoded ($z_{i+1} \rightarrow m_{i+1,D}$), and the resulting molecules are **docked** into the active site of $\mathcal{T}$. Using a greedy optimization strategy, the "best" of these designed molecules can be identified using the docking-derived $\mathcal{S}_{i+1}$ and is then used as the new reference for the next round of **sample**-and-**dock**. The process is then repeated over many cycles to explore the chemical space in a semidirected manner and *conditioned* on the properties of $\mathcal{T}$.

Implementation. Using a set of in-house Python scripts, we implemented a sample-and-dock pipeline that interfaces the junction-tree-variational autoencoder (JTVAE) model of Jin and co-workers[25] and the molecular docking program rDock.[26] The JTVAE model utilizes two latent spaces, referred to as the tree space and molecular graph space. The tree space maps the topological node-and-edge graph of the molecule, and the molecular graph space fits the node-and-edge graph with the appropriate atoms or substructures. The developers of the JTVAE model used a vocabulary of scaffolds extracted from the training set in the SMILES (Simplified Molecular-Input Line-Entry System) format. They then jointly trained two encoders, one for tree space and the other for the molecular graph space. The associated decoder was trained together with the encoders to reconstruct the SMILES from the two latent vectors produced by the encoders. Once trained, the model learned an efficient mapping of molecules in the high-dimensional latent space. During initial tests, we observed that this mapping produced by the JTVAE exhibited the crucial property that the distance between vectors was proportional to the similarity/dissimilarity between the molecules. Therefore, we used this latent space to represent the chemical space in which we perform structure-aware sampling.

All instances of sample-and-dock were initialized using benzene. At the start of each cycle, the seeding molecule was taken in the form of SMILES strings and converted into one-hot encoding according to a set of predefined vocabulary of 531 unique molecular scaffolds SMILES extracted from MOSES data set.[27] The one-hot encoding was first transformed into a set of two 28-dimensional vectors by the encoders of JTVAE as locations in the two latent spaces. The latent vectors were then resampled and reconstructed (decoded to SMILES strings) 20 times by the decoders to generate 20 unique molecules that resemble the seed. For each generated molecule, the SMILES string was converted to a 3D structure using RDKit and then docked into the active site on the target using rDock. Across all of the generated molecules within the cycle, the designed molecule with the highest $\mathcal{A}_i$, as estimated using the rDock scoring function, was used as the

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

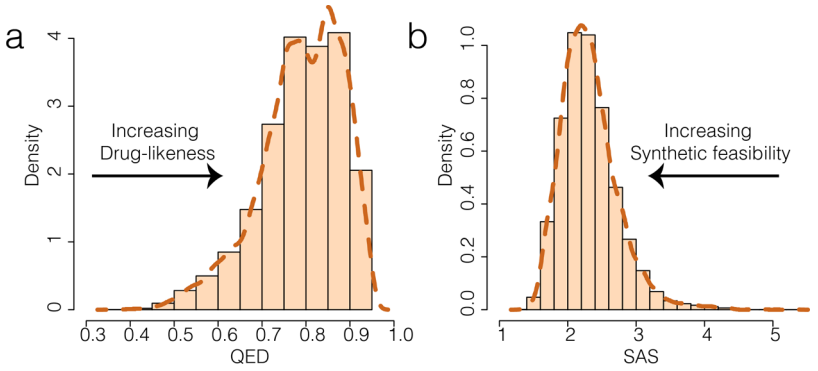


Figure 2. Computed properties of compounds designed to target the active site of CDK2. The distributions of (a) the quantitative estimate of drug-likeness (QED) and (b) the synthetic accessibility scores are shown.

seed molecule for the next cycle. The process is then repeated over many cycles, and the best compound from each cycle is used to construct the final library. We note that for a new generation, it is the latent space representation of the best design that was used to seed the next cycle.

Within this sample-and-dock pipeline, docking is initialized using coordinate files for the receptor and a ligand file. The ligand file is used to identify the precise location of the pocket into which focused docking will be carried out and for which a target-specific library will be generated. To select the best design during each cycle, we used the interaction component of the dock score (SCORE.INTER). After the final cycle, the collection of structures is combined into a single library and saved as an SDF-formatted file.

Target-Specific Libraries. We used our sample-and-dock pipeline to construct target-specific libraries for CDK2 and SARS-CoV-2 M^{pro}. For CDK2, sample-and-dock was carried out using the crystal structure of CDK2 bound with Roniciclib (PDB ID: 5IEV).[28] For SARS-CoV-2 M^{pro}, we used the crystal structure M^{pro} bound to N3 (PDB ID: 6Y2F).[29] From the crystal structures, the ligands and receptors were separated and saved as separate coordinate files using PyMOL. To map the active sites of each target, the PDB-formatted receptor coordinate files were converted to the SYBYL MOL2 format using OpenBabel.[30] Using the MOL2 receptor file, cavity mapping was carried out using the reference-ligand method that is implemented in the `rbcavity` tool from rDock. For the reference-ligand method, the radius of the overlapping sphere was set to 7.0 Å, and the radius of the small probe was set to 1.5 Å. To prepare the designs for docking, the SMILES string was first converted into graph-based representation by RDKit, and the 3D coordinate is generated using ETKDG algorithm implemented with RDKit.[31] Receptor charges are internally assigned in rDock based on substructures and atom names, whereas ligand charges are assigned based on atom number, hybridization, and formal charge.[26] For both CDK2 and M^{pro}, sample-and-dock was initialized using benzene and run for 48 hours on a 16-core x86 Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz node, resulting in 27 810 and 28 134 ligands for CDK2 and M^{pro}, respectively. We note that we chose to generate the libraries over the course of 48 hours because a set of independent preliminary tests, carried out for CDK2, revealed that over the resulting 1600 cycles, the value of the lowest docking score sampled tended to plateau (Figure S1).

Visualizing Target-Specific Libraries. To visualize the libraries, we used tree-maps (TMAPs), a locally sensitive hashing forest constructed with MinHash fingerprints of the molecular structures.[32] Given a collection of high-dimensional items, TMAP generates a two-dimensional (2D) embedding within a tree-like structure, containing branches and sub-branches. Within this tree, items are essentially clustered into branches and then sub-branches based on the pairwise similarity; the distances between clusters are encoded in the distance between the branches and sub-branches. As such, TMAPs can encode the hierarchical, global-to-local structure within a high-dimensional data set.

Starting from the SMILES of our designs, we generated TMAPs by first computing their MHFP6 (512 MinHash permutations) molecular fingerprints of the compounds in the libraries.[32] The individual libraries were then visualized by exporting the 2D, TMAP coordinates to an interactive HTML file using the Python module Faerun,[33] which uses the JavaScript library SmilesDrawer[34] to visualize 2D structures within our library. To aid in the examination of the distributions within the TMAP-space, points were color-coded using docking scores, synthetic accessibility (i.e., SA), and drug-likeness (i.e., QED). The interactive maps can be accessed at https://atfrank.github.io/SampleDock/vis_maps/tmap_CDK2_20designs_all.html (for CDK2) and https://atfrank.github.io/SampleDock/vis_maps/tmap_mpro.html (for M^{pro}).

RESULTS

CDK2 Library. First, we implemented a sample-and-dock pipeline (Figure 1b) by interfacing the junction-tree variational autoencoder (JTVAE)[25] with the docking program, rDock.[26] We then used it to construct a target-specific library for the cyclin-dependent kinase 2 (CDK2) protein. We chose CDK2 because it is an important therapeutic target that has been extensively studied biochemically and structurally and for which a large library of known inhibitors is readily accessible.[35] For this CDK2 test case, we initiated sample-and-dock pipeline by first mapping the CDK2 active site using the crystal structure of CDK2 complexed with Roniciclib (PDB: 5IEV).[28] The JTVAE−rDock sample-and-dock pipeline was then initialized using benzene as a starting scaffold. We chose to use benzene as the starting scaffold because it is small and a chemical motif that is ubiquitous in known drugs. Each sample-and-dock cycle consisted of sampling and docking 20 unique compounds. For each compound, 100 poses were generated during docking. The compound with the lowest docking score was selected and used as the input scaffold for the next cycle of sample-and-dock (Figure 1b).

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

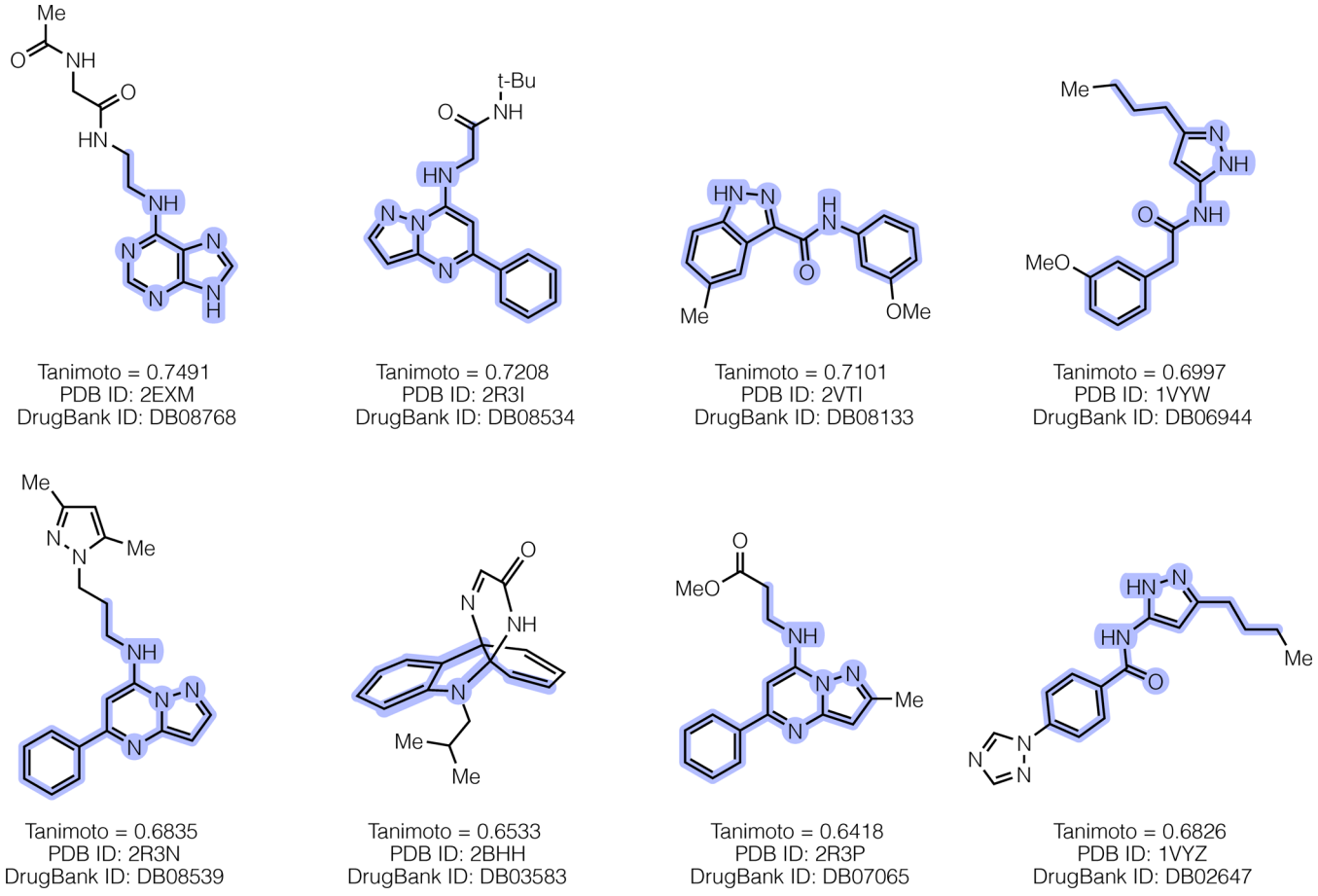


Figure 3. Representative compounds in our sample-and-dock library that closely resemble known inhibitors of CDK2. The substructures that are identical to those found in the most similar compound in our sample-and-dock virtual library are highlighted.

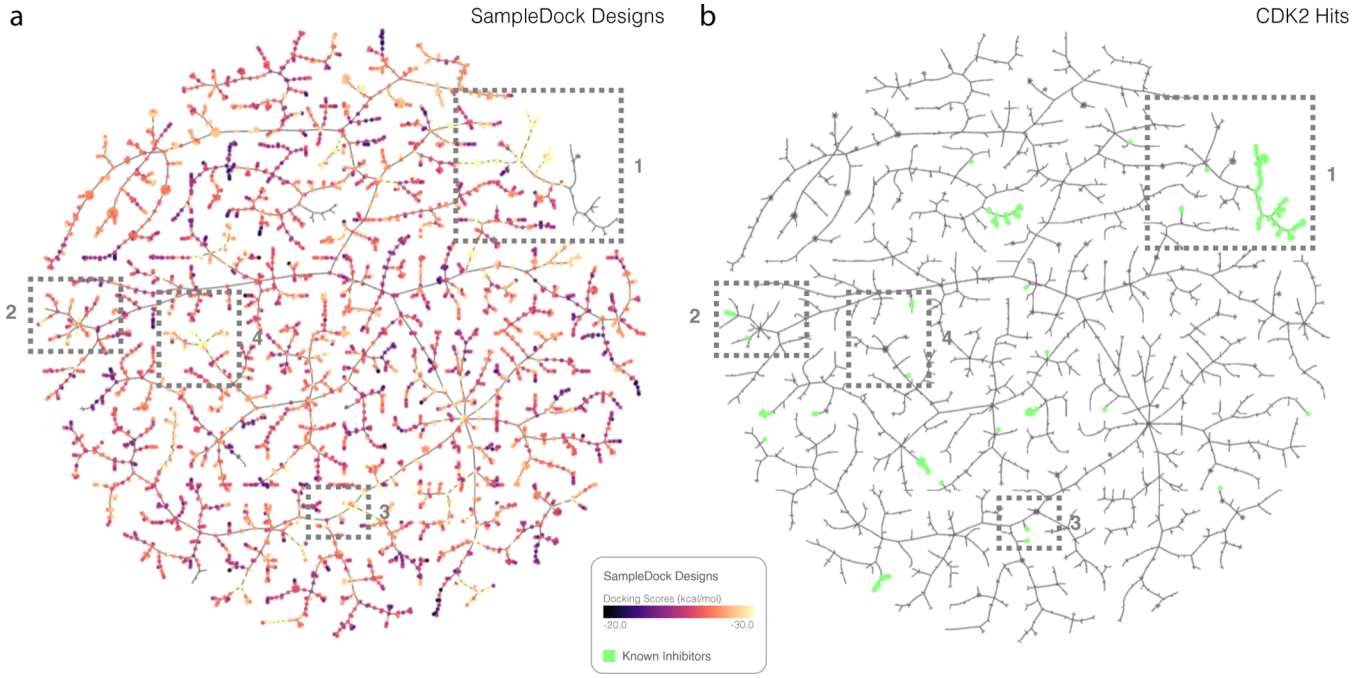


Figure 4. TMAP of best designs against the CDK2 target. TMAPs with (a) our sample-and-dock designs and (b) the top known CDK2 inhibitors found in the DrugBank are shown. The sample-and-dock designs are color-coded based on docking scores. The corresponding interactive TMAP can be accessed via https://atfrank.github.io/SampleDock/. The boxes highlight regions in the TMAP where sample-and-dock designs with low docking scores are near CDK2 inhibitor hits.

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

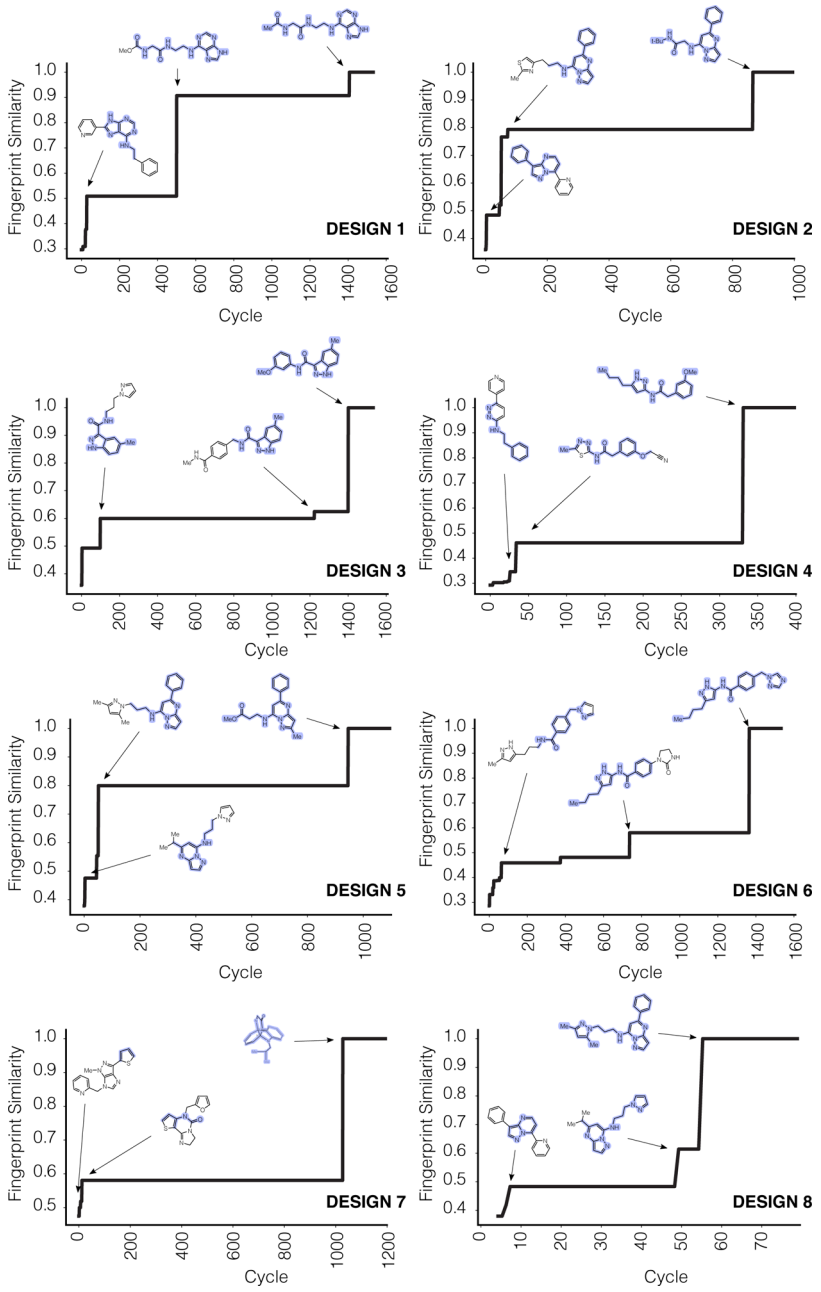


Figure 5. Path to designs resembling known CDK2 inhibitors. For each of the eight designs in our library that most resembled known CDK2 inhibitors, we show the trace of the maximum chemical similarity relative to the design as a function of the cycle. We highlighted the sub-substructures that are identical to those found in the most similar compound in our sample-and-dock virtual library.

To explore the quality of the compounds sampled during sample-and-dock, we computed the distribution of their drug-likeness (estimated using the quantitative estimate of drug-likeness, QED,[36] score) and synthetic feasibility (estimated using their synthetic accessibility score, SAS).[37] Though there are exceptions, compounds suitable for initializing drug development typically exhibit QED and SAS → 1.0. For CDK2, the sample-and-dock molecules exhibited mean QED and SAS values of ∼0.80 (Figure 2a) and 2.5 (Figure 2b), respectively.

To examine whether the compounds explored during sample-and-dock resemble known inhibitors of CDK2, we computed their chemical similarity to known CDK2 inhibitors. Briefly, we identified known CDK2 inhibitors within the DrugBank[35] and calculated the Tanimoto coefficient between the chemical fingerprints (RDKit fingerprint) of the known inhibitors and the sample-and-dock compounds. Despite initializing sample-and-dock with just benzene, we found that we were indeed able to sample designs that closely resemble known CDK2 inhibitors (Figure 3). Interestingly, when we projected the sample-and-dock designs along with known CDK2 inhibitors within a tree-map (TMAP), some of the known CDK2 inhibitors are located in branches that are near low-scoring sample-and-dock designs (Figure 4). Though some of the sample-and-dock designs resembled known inhibitors, the mean Tanimoto coefficient was only ∼0.35 (Figure S2), confirming that the majority of them were indeed

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

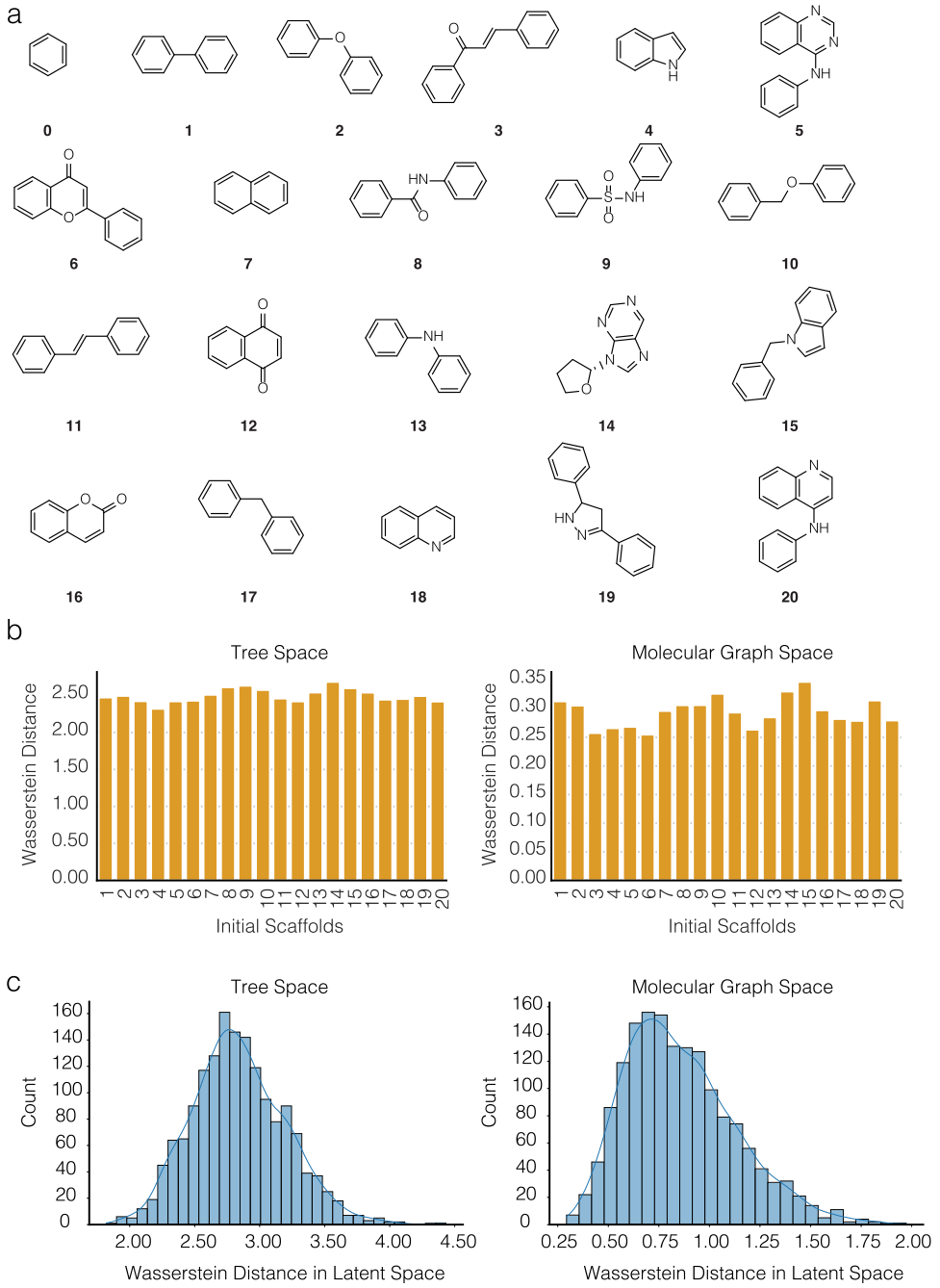


Figure 6. (a) Twenty additional scaffolds we used to constructing CDK2-specific libraries. Panel (b) shows the pairwise Wasserstein distance between the library-seeded benzene and libraries seeded with the other 20 scaffolds (a). The pairwise Wasserstein distances within the tree space and molecular graph space of the JTVAE are shown. (c) For reference, we also show the distribution of the pairwise Wasserstein distance between samples in successive cycles of sample-and-dock.

distinct from previously identified inhibitors and thus our library was indeed populated with novel designs.

To gauge when, during the 1533 design cycles, we sampled the 8 designs that were most similar to known CDK2 inhibitors, we determined their chemical similarity to each compound in the library and plotted the maximum similarity as a function of the design cycle. Figure 5 shows traces of the cumulative maximum chemical similarity for these eight designs as a function of the design cycle. We found that with the exception of designs 4 and 8, ∼1000 cycles (∼224 CPU core-hours) are typically needed to sample the compounds that resemble known CDK2 inhibitors.

To check if, despite using a greedy, local search algorithm, we were able to generate molecules with docking scores on par with known CDK2 inhibitors, we docked the known inhibitors to the same receptor and same pocket we used for sample-and-dock. We then compared the docking score distribution of the known inhibitors to that of our sample-and-dock molecules. For reference, we also docked and compared these distributions to those of a set of molecules randomly sampled from the latent space of the JTVAE and the MOSES data set, respectively. Relative to the JTVAE and MOSES distributions, our sample-and-dock library exhibits better overlap with the docking score distribution of the known CKD2 inhibitors

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

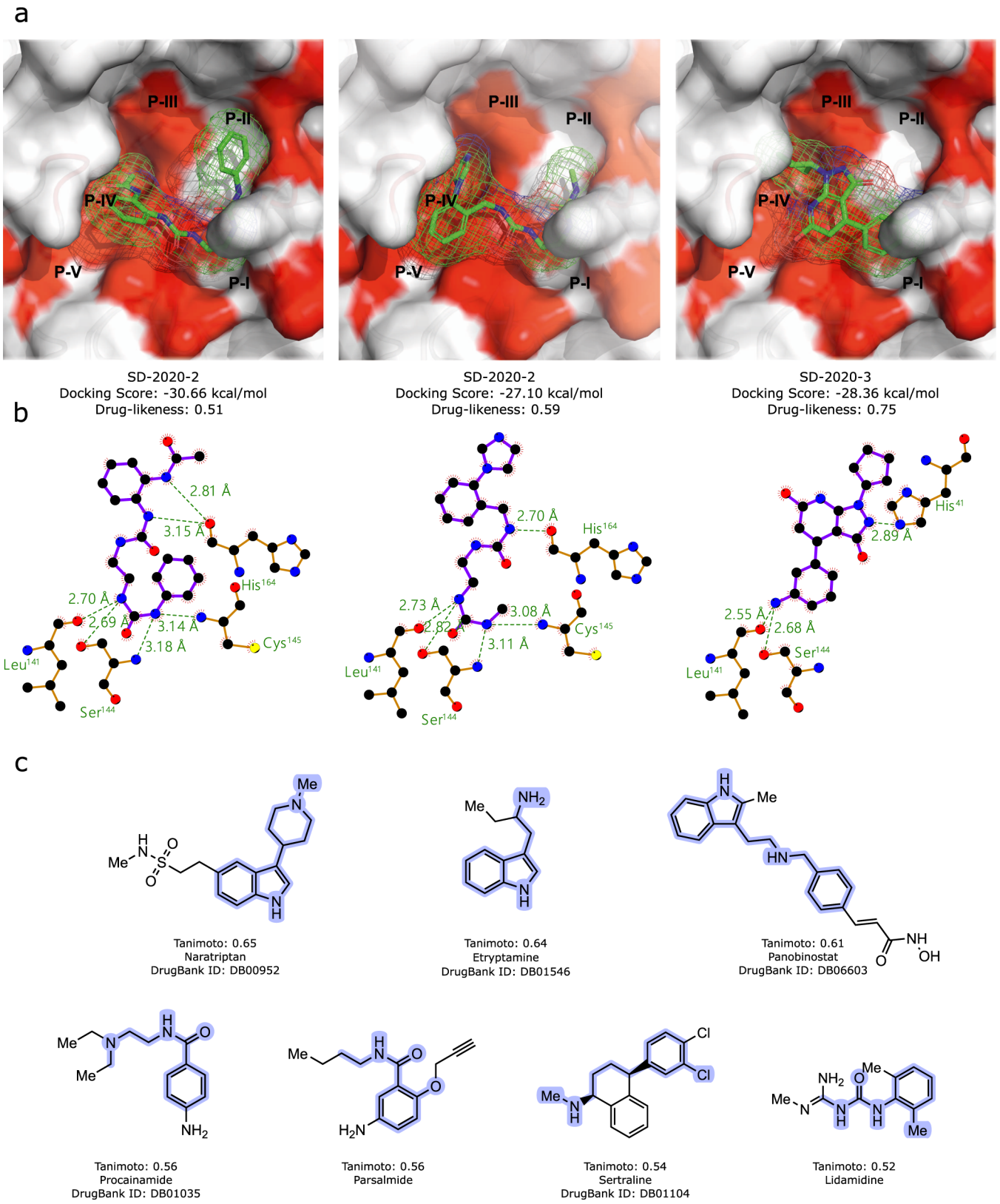


Figure 7. (a) Examples of three potential inhibitors (SD-2020-1, -2, and -3) designed using the sample-and-dock approach. The predicted poses of these compounds, docked into the active site of SARS-CoV-2 M^{pro}, are shown. These three molecules were identified by first clustering the sample-and-dock library and then selecting the top-10 scoring clusters. The synthetic accessibility of the representative design in these 10 clusters was then accessed using expert chemical intuition. The three designed presented here were deemed to be the most synthetically accessible among the 10 representative designs. (b) Ligand interaction maps of the docked poses. Docking predicts that the SD-2020-1 and SD-2020-3 interact with the catalytic residue Cys^{145} and SD-2020-2 interacts with the catalytic residue His^{41}. (c) Known drugs that were most similar to compounds in our sample-and-docking M^{pro} library. Sub-sub-structures that are identical to those found in the most similar compound in the virtual library are highlighted.

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

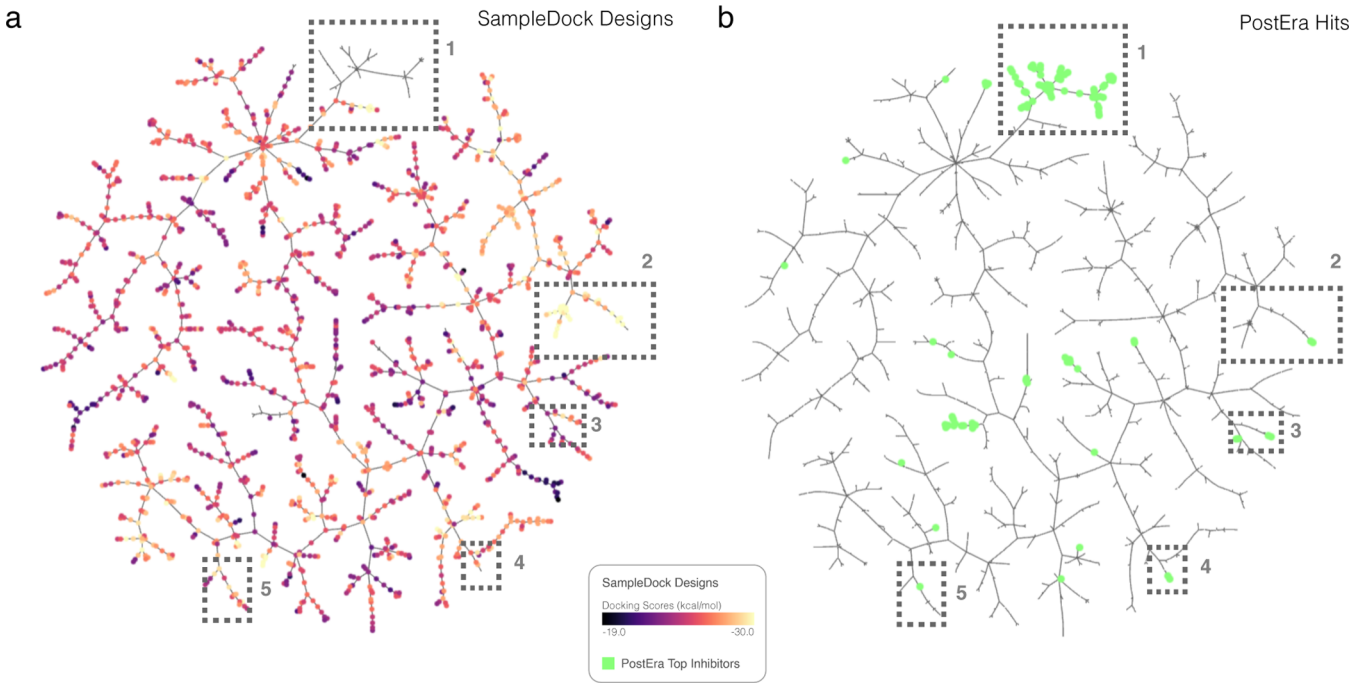


Figure 8. TMAP of best designs against the M^{pro} of SAR-CoV-2 target. The TMAPs with (a) our sample-and-dock designs and (b) the top-100 PostEra hits (bright green points) (https://covid.postera.ai/covid/activity_data) are shown. The sample-and-dock designs are color-coded based on docking scores. The corresponding interactive TMAP can be accessed via https://atfrank.github.io/SampleDock/vis_maps/tmap_mpro.html. The boxes highlight regions in the TMAP where sample-and-dock designs with low docking scores are near PostEra hits.

(Figure S11). We also found that we were about to better recover the molecules that most resemble CDK2 inhibitors from our sample-and-dock library (ROC AUC = 0.63) than either of the set of JTVAE (ROC AUC = 0.49) or MOSES (ROC AUC = 0.53) molecules (Figure S12).

To examine the impact of the initial scaffold choice on the library that emerges from our sample-and-dock pipeline, we repeated designs using a set of 20 unique scaffolds (Figure 6a). The scaffolds were selected by ranking the cumulative occurrences of unique member compounds in ChEMBL20, as reported by Zdrazil et al.[38] The individual libraries generated using these 20 scaffolds were compared to the initial library generated using benzene. For this comparison, we used the Wasserstein distance.[39−41] The Wasserstein distance measures the overlap between two distributions by quantifying the mean distances points in a given sample need to be moved to maximize the overlap between two distributions. As such, we used it to compare the distributions in the latent space of the libraries generated using benzene as the initial scaffold to the libraries generated using 20 additional scaffolds.

Relative to the libraries initialized using benzene, the Wasserstein distances were ∼2.50 and ∼0.30 in the tree space and molecular graph space (Figure 6b), respectively. By comparison, the average Wasserstein distances in tree space and molecular graph space between samples in successive cycles in a typical sample-and-dock run were ∼2.70 and ∼0.60, respectively (Figure 6c). Given that the Wasserstein distances between the benzene library and the other scaffold libraries were on par with the distances between successive cycles of sample-and-dock, we conclude that, for trials of a typical length (at or above 24 hours), the choice of scaffold does not significantly alter the distribution with the final libraries. To visually confirm this, we generated and compared TMAPs of the libraries generated using different scaffolds. In general, we found that the TMAP of individual libraries exhibits a high degree of overlap, consistent with the Wasserstein distance analysis. Similar analysis also showed that independent libraries initialized using the same scaffold result in similar libraries, as do libraries generated using a larger number of designs per cycle.

It is well known that docking scores contain significant errors. As such, bootstrapping sampling of chemical space to docking by identifying the single "best" design cycle based on the lowest docking score can lead to nonoverlap sampling of chemical space. To mitigate this, an ensembling approach in which, instead of a single design, a set or "ensemble" of designs with the lowest docking scores could be selected after each cycle and used to seed the next cycle of designs. To assess the impact of the size of the ensemble on the generated library, we regenerated libraries with the ensemble size $M = 5$ and compared them to the original CDK2 library, i.e., the library generated with the ensemble $M = 1$. Overall, the libraries generated using ensembling at each cycle lead to similar libraries. For example, the Wasserstein distances in the tree space and molecular graph space were similar to the values we obtained when varying the scaffold; they are 2.60 and 0.35, respectively (Table S2). Furthermore, the libraries exhibit several overlapping regions within a TMAP (the interactive TMAP can be accessed here at https://atfrank.github.io/SampleDock/). However, we observed that the $M = 1$ sampling explored a larger region of chemical space in the same number of cycles and, as a result, was able to more efficiently locate hot-spot regions (Figure S5).

Taken together, our results for CDK2 suggest that starting from the simple structure and knowledge of the binding site, generative machine learning models and molecular docking could be seamlessly interfaced to rapidly construct a target-specific screening library composed of compounds that are

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

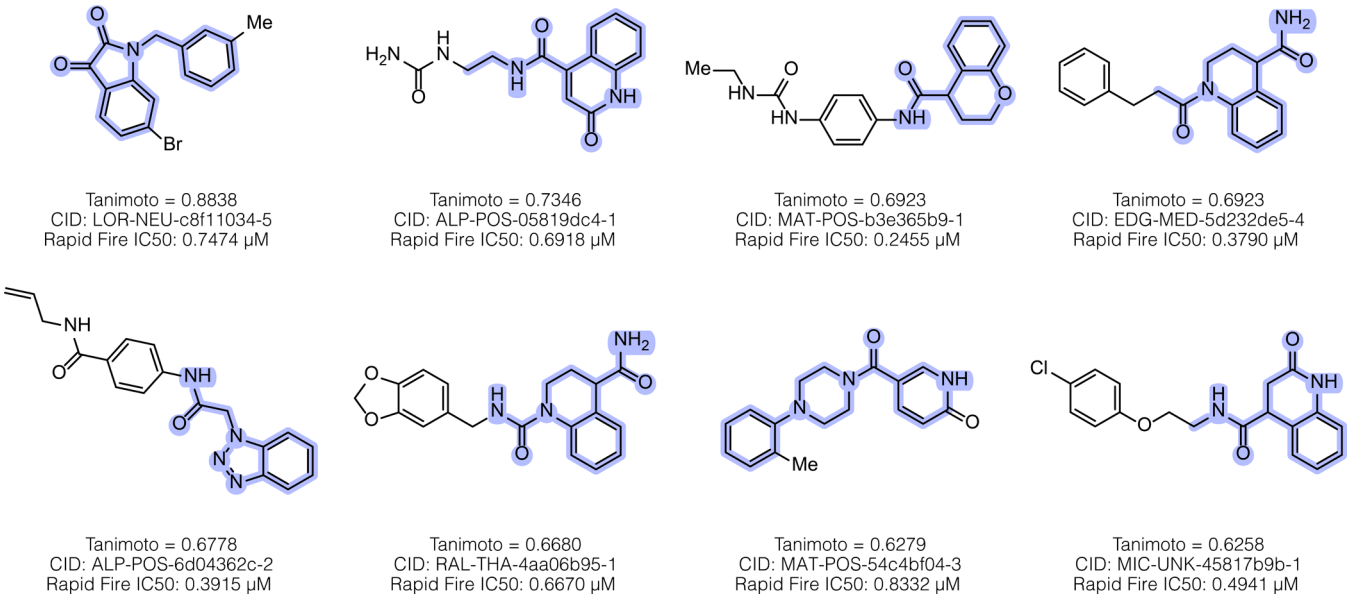


Figure 9. Eight sample-and-dock molecules that were most similar to PostEra hits (https://covid.postera.ai/covid/activity_data). The common sub-sub-structures that were found in the hits and our sample-and-dock designs are highlighted.

predicted to be both drug-like and synthetically feasible and are likely to resemble known small molecule modulators of the target.

SARS-CoV-2 M^{pro} Library. Next, we applied our sample-and-dock framework to the main protease (M^{pro}) of the virus, SARS-CoV-2. SARS-CoV-2 M^{pro} processes polypeptides that are important for viral assembly and, as such, is an attractive target for developing drugs to treat COVID-19.[42,43] To generate a SARS-CoV-2 M^{pro}-specific library, we initiated sample-and-dock calculations starting with benzene as the reference scaffold and targeting the same site on SARS-CoV-2 M^{pro} that is occupied by the inhibitor, N3 (PDB: 6Y2F).[29] The active site of SARS-CoV-2 M^{pro} is composed of five subpockets, which we denote as P-I, P-II, P-III, P-IV, and P-V (Figure 7a). Figure 7 shows three representative compounds within our sample-and-dock SARS-CoV-2 M^{pro} library. Inspection of the predicted poses indicates that the three representative compounds occupied pockets P-I, P-II, and P-IV (Figure 7a) and could form stabilizing contacts with the catalytic residues, His^{41} and Cys^{145} (Figure 7b) Interestingly, both SD-2020-1 (QED = 0.51) and SD-2020-2 (QED = 0.59) contain the peptidyl moiety found in known M^{pro} inhibitors.[29,42] In fact, the presence of the peptidyl moiety was a common feature of many of the compounds that were sampled during sample-and-dock. On the other hand, SD-2020-3 (QED = 0.75), a nucleotide-analogue, is an example of one of the compounds in our library that lacks the peptidyl moiety. The overall drug-likeness and the lack of the reactive peptidyl moiety make SD-2020-3 an interesting lead candidate. On the basis of the results for CDK2, we speculate that the SARS-CoV-2 M^{pro}-specific library we generated using our sample-and-dock approach may contain yet-to-be-explored and promising lead candidates. Accordingly, we make the library of M^{pro} designs fully open to the scientific community with the hope that it will be explored by medicinal chemists to seed the design of novel drugs targeting M^{pro} of SARS-CoV-2 (https://github.com/atfrank/SARS-CoV-2).

To determine whether the compounds in our virtual SARS-CoV-2 M^{pro} library were similar to any known or experimental drugs, we compared the compounds in our library with compounds in SuperDRUG2,[44] a library composed of FDA-approved and experimental drugs. Figure 7c shows the six drugs that exhibited the highest chemical similarity to at least one compound in our SARS-CoV-2 M^{pro} library. Of these six, three are central nervous system drugs, which include sertraline, a selective serotonin reuptake inhibitor (SSRI) that in a recent study has been shown to exhibit antiviral activity against SARS-CoV-2.[45] Intriguingly, another SSRI, fluvoxamine, has recently been identified as a potential COVID-19 drug and is now in clinical trials.[46]

Finally, we compared our sample-and-dock designs to the compounds that others have recently designed, synthesized, and tested against the SAR-CoV-2 M^{pro} as part of the PostEra COVID-19 Moonshot project.[47,48] For this comparison, we generated TMAP using the combined library of our sample-and-dock designs and the 100 PostEra hits with the lowest IC50s. Figure 8 reveals that several branches in the TMAP where low docking scores sample-and-dock designs adjacent to some of the PostEra hits. Moreover, several of the PostEra hits exhibit a striking resemblance to designs in our sample-and-dock libraries (Figure 9).

Though indirect, this result suggests by using the sample-and-dock pipeline and focusing on the regions of chemical space with docking scores, we could have isolated regions of chemical space that contained bioactive small molecule ligand. On the basis of this result, we envision that the TMAP of our sample-and-dock library may serve as a guide to identify new compounds that might also bind to the SARS-CoV-2 M^{pro}.

SampleDock Software. To facilitate the use of our virtual library generation method beyond our research group, we have implemented it as a standalone tool, which we refer to as SampleDock (https://github.com/atfrank/SampleDock). SampleDock can be installed using the conda package management tool. SampleDock takes as input the coordinates of the receptor in the MOL2 format and a pseudo-ligand coordinate in the SD format. The pseudo-ligand file is used to define the location of the binding cavity for which the virtual library should be generated. Users can also specify as a

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

SMILES string of the seed molecule used to initialize library generation and the score term used to guide exploration of chemical space during each cycle.

Future Directions. First, we will implement additional optimization protocols in SampleDock. Currently, SampleDock uses a greedy algorithm; however, other optimization approaches, for instance, simulated annealing, genetic algorithm, and particle swarm, could be implemented.

Second, we will add support for additional generative models so that out-of-the-box, SampleDock users will be able to select from a library of generative models. Finally, we add functionality to SampleDock that makes it easy for users to interface the generative model and docking program of their choice.

Third, while for simplicity, sample-and-dock is currently initialized using a single framework, and though distributions of the libraries generated starting from different fragments show some resemblance (Figure 6), we believe one could further mitigate the effects of any bias due to the starting fragment by initializing sample-and-dock using a library of diverse fragments. Future versions of the SampleDock software tool will include such functionality to initialize the sample-and-dock pipeline using a library for fragments rather than a single fragment.

Fourth, to accelerate sample-and-dock, we will explore the use of an implicit docking strategy that can reproduce docking scores based only on the 2D structure ligands, that is, without explicitly carrying out docking. Such strategies have been recently implemented,[49−51] and they can be used to rapidly generate preliminary libraries that can then be refined using full, explicit (or classical) redocking of the best sample-and-dock designs.

A significant limitation of sample-and-dock is that though the library it generates may contain viable, drug-like ligands that bind to the target, selecting them from the many thousands of compounds in the library remains a nontrivial task. This is partly due to inherent limitations in the scoring functions that are used during docking that result in errors in both pose prediction and binding affinity estimation, both of which can have detrimental impact on one's ability to recover true ligands from decoys. Here, we used the rDock scoring function to guide exploration of "chemical space". Though the scoring function was useful for executing the fast docking calculations that were required, it is not necessarily appropriate to prioritize the compounds in the resulting libraries. To filter the virtual library and prioritize compounds that could be synthesized and then advance the biochemical testing, we envision that one could apply, to a subset of the virtual library, well-established postdocking analysis methods.[52] Such methods may utilize more physically rigorous scoring function,[53,54] quantum-mechanical calculations,[55] or FEP calculations[56] to re-sort and prioritize the compounds and so could be used to identify promising molecules in our sample-and-dock libraries. Alternatively, sample-and-dock libraries could be rescored using simpler yet high-performing scoring functions,[57] such as molecular interaction fingerprinting[58] and interaction graph-matching.[59] These methods are attractive because they are also fast and thus make an excellent complement to our sample-and-dock framework.

CONCLUSIONS

To summarize, we combined emerging generative modeling with well-established biophysical modeling to yield a hybrid, sample-and-dock approach for the structure-guided exploration of chemical space. Instead of exploring chemical space in a large and fixed library, our approach facilitates a directed exploration of the chemical space that is relevant to a specific site on a given target. Instead of attempting to identify a single putative hit, each run of our approach generates a collection/library of compounds that, based on their docking scores, are likely to bind to the target at a user-specified site. One can then use sample-and-dock libraries as the input to more elaborate virtual screening pipelines to select putative hits. The sample-and-dock pipeline we implemented is also modular and flexible. Thus, as more robust generative models are developed, they can be trivially incorporated into our pipeline, as can more accurate and robust scoring functions and postdocking methods. In its current form, our sample-and-dock pipeline can be deployed to construct structure- and target-specific maps of the "privileged" chemical space for therapeutically relevant protein and nucleic acid targets. Applying unsupervised learning to sets of structure- and target-specific chemical space maps could facilitate the emergence of general design principles that can guide the rational structure-guided design of biomolecular probe compounds.[60]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00746.

The Wasserstein distance analysis of the libraries and describing proposed synthetic schemes for the three putative M^{pro} inhibitors (PDF)

AUTHOR INFORMATION

Corresponding Author

Aaron T. Frank − *Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0002-7782-2200; Phone: (734) 615-2053; Email: afrankz@umich.edu

Authors

Ziqiao Xu − *Chemistry Department, University of Michigan, Ann Arbor, Michigan 48109, United States*

Orrette R. Wauchope − *Department of Natural Sciences, City University of New York, Baruch College, New York, New York 10010, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00746

Notes

The authors declare no competing financial interest.

While we were revising this manuscript, we noted that others have reported related approaches.[61,62] Though these methods differ from ours in terms of the details of their implementation, they share a common objective with our technique, namely, that of guiding the exploration of chemical space using specific chemical or biophysical properties.

SampleDock can be accessed as a standalone tool via https://github.com/atfrank/SampleDock or as a webservice at https://smaltr.org/. The sample-and-dock library for SARS-CoV-2 M^{pro} (as well as other COVID-19 targets) can be accessed at https://github.com/atfrank/SARS-CoV-2. The interactive TMAPs for the best designs against CDK2 and SARS-CoV-2

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

M^{pro} can be accessed here at https://atfrank.github.io/SampleDock/.

■ ACKNOWLEDGMENTS

The authors thank the members of the Frank lab for many valuable discussions about this work. The authors were funded by startup funds and a grant from the Michigan Institute for Computational Science and Discovery (MICDE) at the University of Michigan (A.T.F and Z.X.) and the City University of New York (O.W). We acknowledge support for this work through Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562; the allocation identification for this project was TG-MCB190114. We specifically thank the consultants Eroma Abeysinghe, Suresh Marru, and Sudhakar Pamidighantam for their assistance with bringing the SMALTR Science Gateway online and adding SampleDock to SMALTR. The latter was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program. SMALTR gateway is hosted with the SciGaP gateway platform, which is supported by the National Science Foundation award number 1339774.

■ REFERENCES

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203−214.

(2) Willems, H.; De Cesco, S.; Svensson, F. Computational Chemistry on a Budget−Supporting Drug Discovery with Limited Resources. *J. Med. Chem.* **2020**, *63*, 10158−10169.

(3) Clark, D. E. Virtual Screening: Is Bigger Always Better? Or Can Small Be Beautiful? *J. Chem. Inf. Model.* **2020**, *30*, 4120−4123.

(4) Gillet, V. J.; Johnson, A.; Mata, P.; Sike, S. Automated structure design in 3D. *Tetrahedron Comput. Methodol.* **1990**, *3*, 681−696.

(5) Moon, J. B.; Howe, W. J. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 314−328.

(6) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: a program for structure generation. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127−153.

(7) Böhm, H.-J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.

(8) Rotstein, S. H.; Murcko, M. A. GroupBuild: a fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36*, 1700−1710.

(9) Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. PRO_LIGAND: An approach to de novo molecular design. 1. Application to the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 13−32.

(10) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(11) DeWitte, R. S.; Ishchenko, A. V.; Shakhnovich, E. I. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 2. Case studies in molecular design. *J. Am. Chem. Soc.* **1997**, *119*, 4608−4617.

(12) Chéron, N.; Jasty, N.; Shakhnovich, E. I. OpenGrowth: an automated and rational algorithm for finding new protein ligands. *J. Med. Chem.* **2016**, *59*, 4171−4188.

(13) Hoffer, L.; Muller, C.; Roche, P.; Morelli, X. Chemistry-driven Hit-to-lead Optimization Guided by Structure-based Approaches. *Mol. Inf.* **2018**, *37*, No. 1800059.

(14) Daeyaert, F.; Deem, M. W. A Pareto Algorithm for Efficient De Novo Design of Multi-functional Molecules. *Mol. Inf.* **2017**, *36*, No. 1600044.

(15) Bienstock, R. J. Fragment-Based Methods. In *Drug Discovery*; Springer, 2015; pp 119−135.

(16) Li, H.; Leung, K.-S.; Chan, C. H.; Cheung, H. L.; Wong, M.-H. iSyn: WebGL-based interactive de novo drug design, 2014 18th International Conference on Information Visualisation, 2014; 302−307.

(17) Diharce, J.; Cueto, M.; Beltramo, M.; Aucagne, V.; Bonnet, P. In Silico Peptide Ligation: Iterative Residue Docking and Linking as a New Approach to Predict Protein-Peptide Interactions. *Molecules* **2019**, *24*, 1351.

(18) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **2018**, *37*, No. 1700153.

(19) Gupta, A.; Müller, A. T.; Huisman, B. J.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative recurrent networks for de novo drug design. *Mol. Inf.* **2018**, *37*, No. 1700111.

(20) 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.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* **2018**, *4*, 268−276.

(21) Liu, X.; Ye, K.; van Vlijmen, H. W.; IJzerman, A. P.; van Westen, G. J. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A 2A receptor. *J. Cheminf.* **2019**, *11*, No. 35.

(22) Schneider, G.; Clark, D. E. Automated De Novo Drug Design−"Are we nearly there yet?". *Angew. Chem., Int. Ed.* **2019**, *58*, 10792−10803.

(23) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of generative autoencoder in de novo molecular design. *Mol. Inf.* **2018**, *37*, No. 1700123.

(24) Koes, D. R.; Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* **2012**, *40*, W409−W414.

(25) Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. arXiv preprint arXiv:1802.04364 2018.

(26) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.* **2014**, *10*, No. e1003571.

(27) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. arXiv preprint arXiv:1811.12823 2018.

(28) Ayaz, P.; Andres, D.; Kwiatkowski, D. A.; Kolbe, C.-C.; Lienau, P.; Siemeister, G.; Lücking, U.; Stegmann, C. M. Conformational adaption may explain the slow dissociation kinetics of roniciclib (BAY 1000394), a type I CDK inhibitor with kinetic selectivity for CDK2 and CDK9. *ACS Chem. Biol.* **2016**, *11*, 1710−1719.

(29) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **2020**, *368*, 409−412.

(30) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, No. 33.

(31) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**, *60*, 2044−2058.

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600

(32) Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminf.* **2020**, *12*, No. 12.
(33) Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2018**, *34*, 1433−1435.
(34) Probst, D.; Reymond, J.-L. Smilesdrawer: parsing and drawing SMILES-encoded molecular structures using client-side javascript. *J. Chem. Inf. Model.* **2018**, *58*, 1−7.
(35) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.
(36) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90.
(37) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, No. 8.
(38) Zdrazil, B.; Guha, R. The Rise and Fall of a Scaffold: A Trend Analysis of Scaffolds in the Medicinal Chemistry Literature. *J. Med. Chem.* **2018**, *61*, 4688−4703.
(39) Kolouri, S.; Park, S. R.; Thorpe, M.; Slepčev, D.; Rohde, G. K. Optimal Mass Transport. *IEEE Signal Process. Mag.* **2017**, *34*, 43−59.
(40) Kantorovich, L.; Rubinstein, G. S. *On the space of completely additive functions*; Vestnik Leningradskogo Universiteta, 1958; Vol. 3, pp 52−59.
(41) Edwards, D. On the Kantorovich−Rubinstein theorem. *Expo. Math.* **2011**, *29*, 387−398.
(42) Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J. R.; Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **2003**, *300*, 1763−1767.
(43) Goyal, B.; Goyal, D. Targeting the dimerization of main protease of coronaviruses: A potential broad-spectrum therapeutic strategy. *ACS Comb. Sci.* **2020**, *22*, 297−305.
(44) Siramshetty, V. B.; Eckert, O. A.; Gohlke, B.-O.; Goede, A.; Chen, Q.; Devarakonda, P.; Preissner, S.; Preissner, R. SuperDRUG2: a one stop resource for approved/marketed drugs. *Nucleic Acids Res.* **2018**, *46*, D1137−D1143.
(45) Xiao, X.; Wang, C.; Chang, D.; Wang, Y.; Dong, X.; Jiao, T.; Zhao, Z.; Ren, L.; Cruz, C. S. D.; Sharma, L.; Lei, X.; Wang, J. Identification of Potent and Safe Antiviral Therapeutic Candidates Against SARS-CoV-2. *Front. Immunol.* **2020**, *11*, 586572.
(46) Sukhatme, V. P.; Reiersen, A. M.; Vayttaden, S. J.; Sukhatme, V. V. Fluvoxamine: A Review of Its Mechanism of Action and Its Role in COVID-19. *Front. Pharmacol.* **2021**, *12*, No. 652688.
(47) Chodera, J.; Lee, A. A.; London, N.; von Delft, F. Crowdsourcing drug discovery for pandemics. *Nat. Chem.* **2020**, *12*, 581.
(48) Consortium, T. C. M. et al. COVID Moonshot: Open Science Discovery of SARS-CoV-2 Main Protease Inhibitors by Combining Crowdsourcing, High-Throughput Experiments, Computational Simulations, and Machine Learning *bioRxiv* 2020, 2020.10.29.339317.
(49) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Central Science* **2020**, *6*, 939−949.
(50) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Das, K. M. P.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020**, *580*, 663−668.
(51) Berenger, F.; Kumar, A.; Zhang, K. Y. J.; Yamanishi, Y. Lean-Docking: Exploiting Ligands' Predicted Docking Scores to Accelerate Molecular Docking. *J. Chem. Inf. Model.* **2021**, *61*, 2341−2352.
(52) Rastelli, G.; Pinzi, L. Refinement and rescoring of virtual screening results. *Front. Chem.* **2019**, *7*, 498.
(53) Durrant, J. D.; McCammon, J. A. NNScore 2.0: a neural-network receptor−ligand scoring function. *J. Chem. Inf. Model.* **2011**, *51*, 2897−2903.
(54) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: molecular docking with deep learning. *J. Cheminf.* **2021**, *13*, No. 43.
(55) Morao, I.; Fedorov, D. G.; Robinson, R.; Southey, M.; Townsend-Nicholson, A.; Bodkin, M. J.; Heifetz, A. Rapid and accurate assessment of GPCR−ligand interactions Using the fragment molecular orbital-based density-functional tight-binding method. *J. Comput. Chem.* **2017**, *38*, 1987−1990.
(56) Raman, E. P.; Paul, T. J.; Hayes, R. L.; Brooks, C. L., III Automated, Accurate, and Scalable Relative Protein−Ligand Binding Free-Energy Calculations Using Lambda Dynamics. *J. Chem. Theory Comput.* **2020**, *16*, 7895−7914.
(57) Tran-Nguyen, V.-K.; Bret, G.; Rognan, D. True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better. *J. Chem. Inf. Model.* **2021**, *61*, 2788−2797.
(58) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.
(59) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein−Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.
(60) Reymond, J.-L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30−38.
(61) Masuda, T.; Ragoza, M.; Koes, D. R. Generating 3D Molecular Structures Conditional on a Receptor Binding Site with Deep Generative Models. *arXiv* 2020.
(62) Thomas, M.; Smith, R. T.; O'Boyle, N. M.; Graaf, C.; Bender, A. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. *J. Cheminf.* **2021**, *13*, No. 39.

https://doi.org/10.1021/acs.jcim.1c00746
J. Chem. Inf. Model. 2021, 61, 5589−5600