

Reinforcement Learning with Real-time Docking of 3D Structures to Cover Chemical Space: Mining for Potent SARS-CoV-2 Main Protease Inhibitors

Jie Li,[†] Oufan Zhang,[†] Fiona L. Kearns,[‡] Mojtaba Haghghatlari,[†] Conor Parks,[‡]
Xingyi Guan,[†] Itai Leven,[†] Rommie E. Amaro,[‡] and Teresa Head-Gordon^{*,†,¶}

[†]*Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California,
Berkeley, CA, USA*

[‡]*Department of Chemistry and Biochemistry University of California, San Diego, La Jolla,
CA 92093-0340*

[¶]*Departments of Bioengineering and Chemical and Biomolecular Engineering, University
of California, Berkeley, CA, USA*

E-mail: thg@berkeley.edu

Abstract

We propose a novel framework that generates new inhibitor molecules for target proteins by combining deep reinforcement learning (RL) with real-time molecular docking on 3-dimensional structures. We illustrate the inhibitor mining (iMiner) approach on the main protease (MPro) protein of SARS-COV-2 that is further validated via consensus of two different docking software, as well as for druglikeness and ease of synthesis. Ultimately 54 molecules are proposed as potent Mpro inhibitors (7 of which

have better synthetic accessibility), covering a much broader range than crowd-sourced projects like the COVID moonshot, and our generated molecules exhibit an optimized, precise, and energetically consistent fit within the catalytic binding pocket. Moreover, our approach only relies on the structure of the target protein, which means it can be easily adapted for future development of other inhibitors of any protein of disease origin.

Introduction

The COVID-19 pandemic, caused by the spread and infection by a novel betacoronavirus called SARS-CoV-2, has brought immense loss to our global society, causing more than 232 million infection cases and more than 4.7 million deaths globally as of the end of September, 2021.¹ Even though effective vaccines have been developed,² numerous infected patients can still benefit from an effective antiviral drug targeting SARS-CoV-2. While drug repurposing has led to the identification of some drugs as potential treatments for COVID-19 (for example, remdesivir), these have been met with somewhat underwhelming performance in clinical settings.³ Of course, drug re-purposing strategies also suffer the distinct limitation in that they cannot identify novel molecules that would be highly potent for new targets.

Among all the proteins related to the SARS-CoV-2 virus, Mpro has arguably received the most attention with respect to drug re-purposing studies,⁴ in part because it is one of the earliest SARS-CoV-2 proteins in which the 3d structure has been fully determined experimentally.⁵ It is also an attractive target due to its crucial role in the SARS-CoV-2 replication cycle since it is a critical enzyme facilitating the cleavage of non-structural proteins from two polyproteins translated from the SARS-CoV-2 replicase gene (Orf1).^{6,7} The substrate-binding pocket of Mpro is located at a cleft between Domain I (residues 8-101) and Domain II (residues 102-184) of the protein, with a Cys-His catalytic dyad that catalyses the cleavage of the polyprotein 1ab.^{5,8} Molecules inhibiting Mpro can induce a conformational change of the protein which leads to its aggregation,⁹ or occupy the core

of the substrate-binding pocket, blocking substrates from approaching the catalytic site.¹⁰ Besides its crucial biological function for virus viability, Mpro is also a promising target for drug development because it is a cytosolic protein with better accessibility to ingested antivirals, and has a well-defined concave binding pocket that allows for more consistent design of small molecules.¹¹ Furthermore, the cleavage sites Mpro operates at do not have known overlap with human proteases.^{8,12} Therefore, an inhibitor targeting Mpro is unlikely to be toxic to humans.

Traditional methods for identifying specific small molecule inhibitors of a protein target usually start with high throughput virtual screening of massive databases that attempt to capture chemical space and diversity, e.g., ChEMBL,¹³ ZINC,¹⁴ Enamine Diversity Set,¹⁵ and PubChem.¹⁶ Due to the size of such data sets, screening these molecules with sophisticated flexible ligand docking protocols can become intractably expensive. Thus, less sophisticated methods, such as pharmacophore modeling or rigid body docking, are often used to initially screen molecules. Due to the simplicity of these models, the information used to navigate through the small molecule chemical space becomes noisy, and false-positives are ruled in while false-negatives, i.e. potential optimum lead molecules, can be ruled out.^{17,18}

With the advent of modern machine learning, deep learning models have been proposed that can generate new molecules for SARS-COV-2 and other viral diseases,¹⁹⁻²⁹ and the distribution can be skewed towards molecules with specific properties such as drug likeness using techniques such as variational autoencoders (VAE),^{20,21} transfer learning²² and reinforcement learning (RL).²³⁻²⁹ However, most deep learning methods rely on 1-dimensional sequence or 2-dimensional chemical representations of the drug and protein, and do not take full advantage of 3-dimensional structural information of the putative drug, thereby constraining the ability to *generate* drugs with shape and molecular compatibility with the target active site. Recent work has also explored chemical space in the vicinity of some starting molecular scaffold and running docking simulations on these derived molecules,³⁰ however there has been no method that develops new drug molecules with real-time 3d struc-

tural docking to guide the efficient exploration of an immense chemical space with the aid of machine learning.

In this work, we propose a novel workflow dubbed "iMiner" that mines chemical space for new tight binding inhibitors by combining deep RL with real-time flexible ligand docking against a protein binding site (Figure 1). We represent putative inhibitors as Self-Referencing Embedded Strings (SELFIES)³¹ that are generated from an Average Stochastic Gradient Descent Weigh-Dropped Long Short Term Memory (AWD-LSTM)³² recurrent neural network (RNN), allowing wide coverage of chemical space. We illustrate the RL training procedure of iMiner that uses on-the-fly AutoDock Vina³³ with the 3d structures of the Mpro substrate binding pocket and the generated inhibitors. The Vina docking scores are used to adjust the RNN so that the distribution of generated inhibitor molecules are shifted towards those that more strongly interact with the Mpro catalytic site. We perform docking with a second docking software, Schrödinger's Glide SP,³⁴ to build consensus for a drug's strong binding affinity to Mpro, and a final filtering based on synthetic accessibility (SA), druglikeness, and elimination of PAINS³⁵ molecules.

In our Covid-19 relevant example, we propose 54 molecules as potential Mpro inhibitors that are worthy of experimental validation (work in progress). 7 out of the 54 molecules have even better synthetic accessibility. Furthermore, we compare our top hits generated with the iMiner workflow with the molecules submitted to the COVID moonshot project,³⁶ a crowdsourcing effort aimed at developing a novel inhibitor for Mpro, and show that we achieve a broader coverage of the inhibitor drug chemical space. We also find excellent shape and molecular attributes of the inhibitors generated by our model in regard their compatibility with the actual target cavity in Mpro, which is a direct consequence of the real-time docking with actual 3d structures during the training procedure. Further analysis of non-bonded interactions between the found inhibitors with specific binding pocket residues in Mpro also create testable hypotheses in regards their potential role as antivirals to treat COVID-19.

Although we have illustrated the workflow’s first use on a pressing and timely test case – i.e., inhibition of SARS-CoV-2 Mpro due to the desperate need for antiviral treatments of COVID-19 – the iMiner method is highly generalizable. As our workflow only requires a 3D structure of the target protein with a pre-defined binding site, iMiner can be readily adapted to generate small molecules for other protein targets. Thus, we believe our ML algorithm will be of great interest to the drug design community to rapidly screen novel regions of chemical space in real-time for other anti-virals or small molecule therapeutics in the future. All scripts required to run our workflow on an arbitrary protein target can be found on a public GitHub repository¹.

The iMiner Machine Learning Workflow

Figure 1 illustrates the entire iMiner life cycle for generating new inhibitor molecules. Here we describe its components in more detail.

SELFIES representation of inhibitor molecules. An arbitrary molecule can be represented as a topological graph using two main approaches: adjacency matrix based methods and string based methods. The former uses an N by N matrix to encode a molecule, where N is the number of atoms in the molecule, and the values of the matrix are typically bond orders between atoms. An adjacency matrix is not ideal for generative tasks, because the size of the molecule that can be generated should not be fixed, and the learning of chemical knowledge by a ML model through adjacency matrix can be difficult. Instead, string based methods are more suited for molecular generation tasks, and SMILES strings have been the standard for molecular representation due to its conciseness and readability. However, SMILES strings have relatively complex syntax, require matching of open and close brackets for branching, and ring modeling/modification is not trivial. Therefore, generating novel, chemically correct compounds through use of SMILES strings can be challenging.

The SELFIES molecular representation³¹ is specifically designed to ensure that all gen-

¹<https://github.com/THGLab/Covid19>

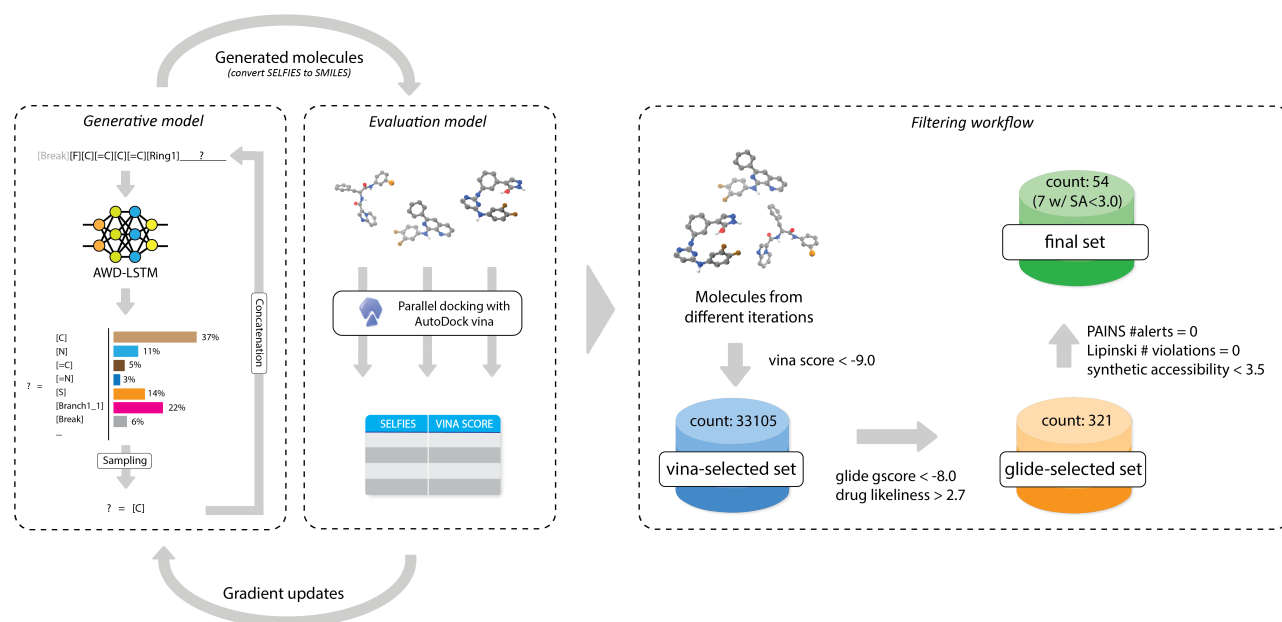


Figure 1: *Illustration of the overall structure of the iMiner workflow which highlights the two major machine learning components, the generative and evaluative models and their interplay, and the subsequent filtering process for chemical and biological feasibility.* (A) The generative model uses SELFIES representations for molecules and a recurrent neural network to “mine” for new molecules that are presented to the evaluative model for 3D docking using AutoDock vina. Vina scores are used in the loss function to drive gradient updates of the neural network. (B) The filtering procedure from molecules collected from intermediate training iterations is based on both favorable Vina and Glide SP docking scores, high drug-likeness scores, no PAINS and no Lipinski’s Rule violations.

erated strings correspond to valid molecules. By utilizing [Branch] and [Ring] tokens with predefined branch lengths and ring sizes, as well as generating symbols using derivation rules, the SELFIES representation guarantees that valence bond constraints are met, and any combinations of tokens from its vocabulary corresponds to a valid molecule. Therefore, we have used SELFIES in our generative model to encode molecules since it does not need to learn chemical syntax rules, and can allocate more of its learning capacity towards generating valid molecules with properties of interest as shown in Figure 1A.

Pre-training the inhibitor molecule generation. Conceptually, generating molecules using string representation is similar to how text is generated in a natural language process-

ing task. Our method starts with a specific [Break] token, and for each molecule, we utilized an RNN that takes in the last token in the string, together with the hidden state from last step to predict a distribution of tokens following the current string. In this work a specific variant of the RNN, known as the AWD-LSTM, was used due to its exceptional performance in similar generative tasks (Figure 1A).³² The network was pre-trained using supervised-learning (SL) of all molecules from the ChEMBL database to learn the conditional probability distributions of tokens that correspond to drug-like molecules. When our trained generative model is used for generating new molecules, a new token is sampled according to the predicted probabilities, and this new token is concatenated to the input string to sample the next token, until the [Break] token is sampled, in which case a complete molecule has been generated.

The performance of our pre-trained generative model was evaluated using the GuacaMol benchmarks,³⁷ which probe 5 different aspects of the distribution of generated molecules with respect to the training dataset (Table 1). Model “validity” reports the proportion of molecules that are syntactically correct. Because we generated molecules via SELFIES representations, we achieved close to 100% validity for all generated molecules. Invalid molecules were either empty strings, or molecules for which the SELFIES package failed to convert into a SMILES string and therefore were discarded before subsequent workflow steps. Model “uniqueness” reports how many generated molecules are duplicates vs. those which are genuinely distinct. Our pretrained models illustrated high uniqueness, indicating the model is able to generate a wide variety of non-redundant molecules. Model “novelty” is defined as the proportion of generated molecules that do not exist in the training dataset. Our model’s high novelty indicates that it is not memorizing molecules from the training dataset, but is indeed generating molecules that it has not seen before. Kullback–Leibler (KL) divergence measures differences in probability distributions of various physicochemical descriptors for the training set and the model generated molecules. As defined by GuacaMol, a high KL divergence benchmark such as predicted for our model suggests that our generated molecules

have similar physicochemical properties to that of training dataset. Finally, Frechet ChemNet Distance (FCD) utilizes a hidden representation of molecules in a previously trained NN to predict biological activities, and thus captures both chemical and biological similarities simultaneously for two sets of molecules.³⁸ Molecules generated by our pre-trained model also have high FCD values, indicating that our molecules are expected to have similar biological activities as molecules from the ChEMBL training dataset.

Table 1: GuacaMol benchmarks for the pretrained generative model and after RL training

Benchmark	Pretrained model	After RL
Validity	0.998	0.998
Uniqueness	0.999	0.983
Novelty	0.867	0.999
KL divergence	0.985	0.791
Frechet ChemNet Distance	0.870	0.007

We then validated our pre-trained distributions using 13 drug-likeness properties between our generated molecules and randomly sampled molecules from ChEMBL database that we used for training. The molecular properties considered are well-recognized chemical features related to the drug-likeness of a molecule which can be obtained through 2D topological connectivity of the molecule: fraction of sp^3 hybridized carbons, number of heavy atoms, fraction of non-carbon atoms in all heavy atoms, number of hydrogen bond donors and acceptors, number of rotatable bonds, number of aliphatic and aromatic rings, molecular weight, quantitative estimate of drug-likeness (QED) value,³⁹ approximate log partition coefficient between octanol and water (alogP),⁴⁰ polarizable surface area (PSA), and the number of structural alerts.⁴¹ Despite the fact that during pre-training only token distributions were used as training targets, all distributions collected from our generated molecules closely follow the distributions from the ChEMBL database (Figure S1). This result suggests our pre-trained model has learned key concepts of “drug-likeness” and provides a good starting point for the RL procedure.

The evaluation module. After our generative model was pre-trained, we employed

an RL workflow to bias the distribution of generated molecules towards specific properties of interest. RL training allows the model to interact with an environment by performing actions according to a policy model, and uses the feedback from the environment to provide training signals to improve the model. In this work, the pre-trained generative model is taken as the policy, and in each iteration 2000 molecules were generated and sent to the evaluation module (Figure 1A).

The central component of our evaluation model is docking with AutoDock Vina executed through cloud computing in parallel with the RL. Within our evaluation model, the Vina score calculator is set up to take a SMILES string representing the ligand, and the 3D structure of the protein target, together with a predefined docking region as input. AutoDock Vina then explores dihedral degrees of freedom and identifies the optimal conformation of the input inhibitor for placement in the designated protein binding site. Finally, AutoDock Vina returns the Vina score as an approximation of the binding energy between the ligand and the protein. Multiple instances of the Vina score calculator tasks were established through Microsoft Azure Batch to allow high-throughput evaluation of the generated molecules. Vina scores were then cycled back to the generative model to improve molecule generation through proximal policy optimization (PPO),⁴² as will be discussed in next section. We emphasize that by using a physics-based docking model which utilizes full 3D structure of our target protein and generated molecules as the critic, the training of the policy model is less likely to be contaminated due to exploiting failure modes of a neural-network based critic, an issue called *wireheading*.⁴³ Instead, we benefit from a more reliable training signal and reduce the false positive and false negative rates of the generated molecules.

Vina scores alone are not sufficient to reliably train a molecule generator, as shown in the Supporting Information (Figure S2) because it will not always satisfy requirements for drug-likeness. To ensure that our generated molecules still bear drug-like properties, we incorporated an additional metric into the reward, S_{DL} , which is a weighted average of the log likelihood for the 13 different drug-like properties used in pre-training assessment.

Formally, our drug-likeness score S_{DL} is defined as:

$$S_{DL}(\mathbf{X}) = \sum_i \sigma_i \log p_i(\text{prop}_i(\mathbf{X})) \quad (1)$$

where $\text{prop}_i(\mathbf{X})$ calculates the i th property of a molecule \mathbf{X} and p_i is defined by the probability distribution of property i by all molecules in the ChEMBL database. The parameter σ_i is defined as:

$$\sigma_i = S_i^{-1} / \sum_j S_j^{-1} \quad (2)$$

where S_i is the entropy of the distribution of property i ,

$$S_i = - \sum_x p_i(x) \log p_i(x) \quad (3)$$

such that a narrower distribution from the ChEMBL database contribute more to the drug likeliness score, and defines the weights for each property as proportional to the inverse of the entropy. Introducing this additional reward ensures our model also accounts for similarity of generated molecules to the drug-likeness present in the ChEMBL database, and ensures that our generated molecules are more likely to be optimal leads for further drug design endeavors.

Reinforcement learning with multiple rewards. Our pretrained policy model defines a probability distribution for an arbitrary sequence of tokens from the SELFIES vocabulary, since the generation of the sequence is a Markovian decision process (MDP), and can be written as:

$$p_{\Theta}(s_T) = p_{\Theta}(s_1|s_0)p_{\Theta}(s_2|s_1)\dots p_{\Theta}(s_T|s_{T-1}) \quad (4)$$

where s_0 corresponds to a starting state with [Break] as the only token in the string,

s_t corresponds to an intermediate state with a finite length string of SELFIES tokens not ended with the [Break] token, and s_T corresponds to the terminal state, with the last token being [Break], or the length of the string exceeding a predefined threshold. $p(s_t|s_{t-1})$ is the transition probability at timestep t , which is the probability distribution of the next token from the generative RNN with network parameters Θ . For each terminal state not exceeding the length limit, a corresponding molecule can be decoded, and the Vina score S_{vina} and drug-likeness score S_{DL} can be calculated. The total reward for a terminal state with a decoded molecule \mathbf{X} is then defined as:

$$r(s_T) = \lambda \max(S_{DL}(\mathbf{X}), 0) - \min(S_{vina}(\mathbf{X}), 0) \quad (5)$$

since the drug-likeness score needs to be maximized and Vina score needs to be minimized. The λ parameter controls the balance between the physical Vina score and the drug-likeness score in the reward function, but in this work we simply used $\lambda = 1$. Negative S_{DL} is upward clipped to 0 and positive S_{vina} is downward clipped to 0 to ensure the reward is non-negative. The expected reward under the MDP is then

$$J(\Theta) = \mathbb{E}_{s_T \sim p_{\Theta}(s_T)}[r(s_T)] \quad (6)$$

Further details of the RL training procedure are given in the Methods section.

Figure 2 compares the distribution of Vina docking scores for molecules generated from the model prior (the pre-trained model) and after RL training which shows a clear shift towards more favorable vina scores. The average Vina score of molecules decreased from -6.95 ± 0.94 kcal/mol to -8.01 ± 0.94 kcal/mol, showing that on average more molecules have stronger interactions with the predefined Mpro docking region. In addition, the GuacaMol benchmarks were also evaluated for the model after RL training, which are also shown in Table 1. Except the Frechet ChemNet Distance (FCD), all other benchmarks are relatively close to the pretrained model, indicating that the RL training does not hurt the quality of

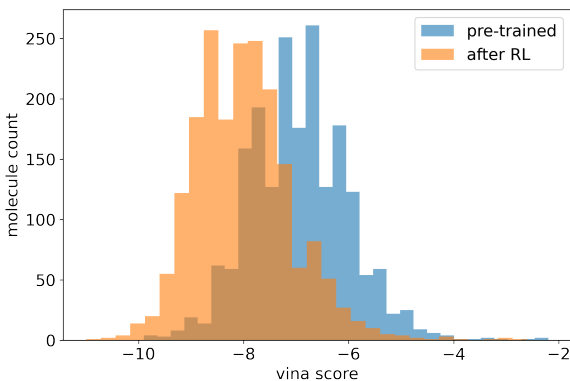


Figure 2: Comparison of AutoDock Vina score distributions for the pre-trained model (blue) and the model trained by reinforcement learning (orange). The mean vina score decreased from -6.95 kcal/mol to -8.01 kcal/mol after RL training.

the generated molecules, and they are still similar to the structures from ChEMBL database. However, FCD has changed significantly, which means the newly generated molecules have different biological activities than the molecules from ChEMBL database. The changes seen in FCD are expected, since, after training, the generated molecules should target a specific cavity of SARS-CoV-2 Mpro, a target for which there are currently no FDA approved treatments. Thus, the FCD differences validate that RL is properly steering the distributions of generated molecules away from its initial distribution.

Validation and filtering of new inhibitor molecules. Validating results from, or checking for consensus between, one docking program with another is often considered standard practice as scoring functions from different programs may have limited accuracy or be parameterized for differing test cases.⁴⁴ Furthermore it is desirable to filter out molecules that are non-specific binders (Pan-assay interference compounds or PAINS) in which we use swissADME⁴⁵ to check for any PAINS alerts,³⁵ as well as Lipinski rule violations,⁴⁶ and to evaluate the synthetic accessibility (SA) scores of these molecules. Figure 1B illustrates the procedures for post-filtering using these additional metrics, which we describe in more detail here.

We start by collecting all molecules from intermediate RL training iterations with a Vina

score < -9.0 , arriving at our “vina-selected set” containing 33,105 molecules (the number of molecules from each training iteration is provided in Figure S3). As expected, more molecules from later iterations were selected, since molecules from later iterations were driven towards having lower Vina scores. Glide Standard Precision (SP)³⁴ docking was performed on all molecules in our vina-selected set with the flexibility to optimize the conformation again with respect to the Glide scoring function. This way we could fully exploit Glide docking as a cross-validation for the generated molecules. Even though the molecules were all good candidates according to Vina score, their glide docking score still showed a wide distribution. We then applied a filter with Glide Gscore (Glide Score) < -8.0 and a drug-likeness score filter of > 2.7 to exclude any structure that is not sufficiently drug-like. After applying these filters we obtained the glide-selected set with 321 molecules in total. The final step was to run these 321 molecules through a final set of filters requiring no PAINS alerts, no Lipinski rule violations and SA scores < 3.5 .

Results

The outcome of the iMiner workflow formulated a final set of 54 molecules shown in Figure 3. These molecules are predicted to be consensus Mpro inhibitors by both AutoDock Vina and Glide SP, they satisfy drug-likeness criterion, and are relatively easy to synthesize due to their predicted low SA scores. The full SMILES representations of these molecules, along with their Vina scores, Glide Gscores, and SA scores are provided in Table S1.

Comparison of chemical diversity of inhibitors discovered by iMiner. Figure 4 compares the total chemical space coverage of molecules generated using iMiner and molecules from ChEMBL and the COVID-moonshot project,³⁶ by performing dimension reduction on the hidden representation of these molecules encoded through ChemNet. ChemNet is a deep network trained on canonized SMILES strings of molecules as input and encodes each molecule into a 512-dimensional latent vector to predict their chemical and biological

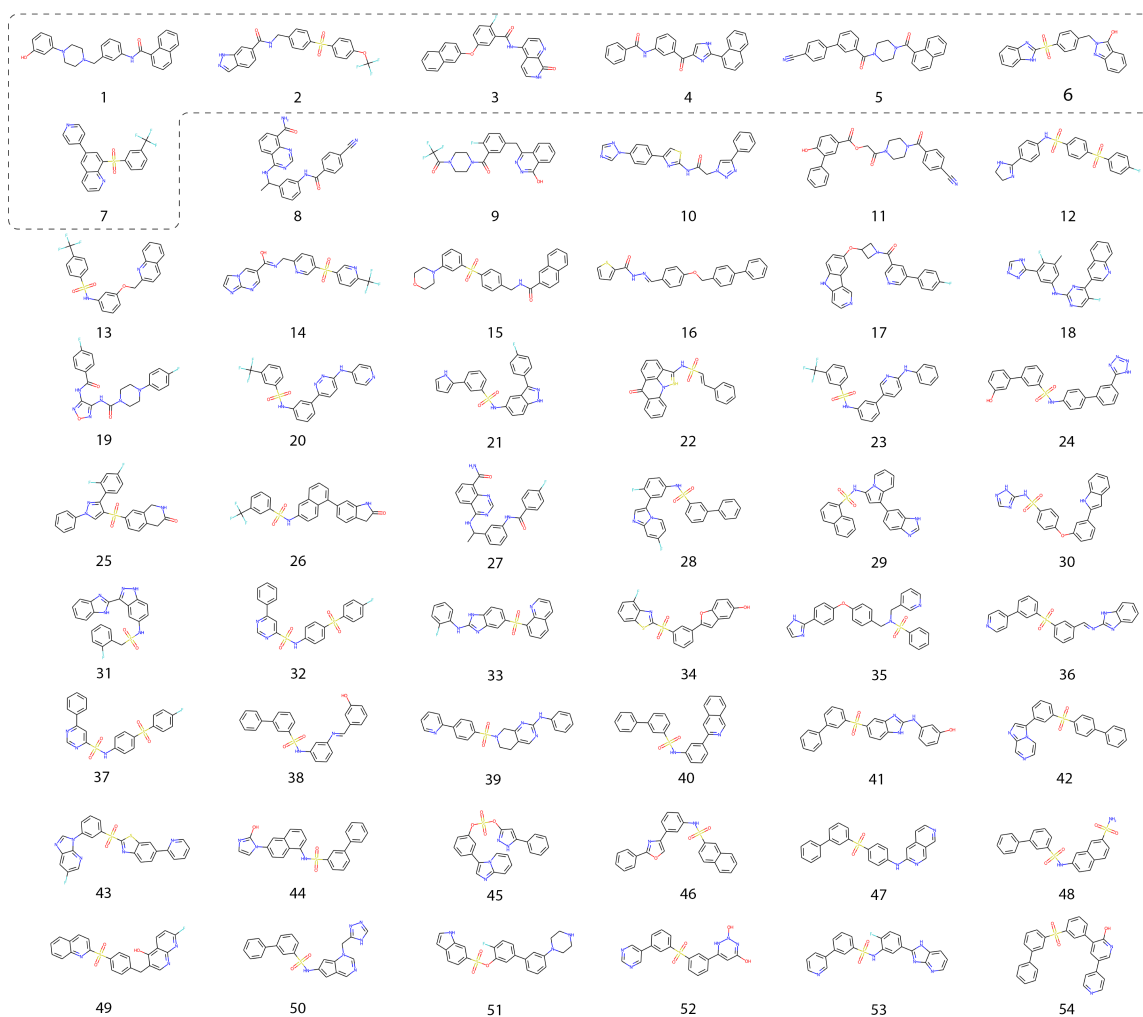


Figure 3: Prediction of 54 molecules that are tight binding inhibitors of Mpro in the final set generated from the *iMiner* workflow. We propose further experimental validations on these molecules as potential SARS-CoV-2 Mpro inhibitors (work in progress). The first 7 molecules in the dashed frame have better synthetic accessibility scores than the rest. The diversity over chemical space of these proposed inhibitors is evident from metrics described in Table 1 and Figure 4.

properties,³⁸ and the dimensions were further reduced to 2 through the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm⁴⁷ for better visualization. Plot points on the resultant figure indicate individual molecules, and points are drawn close to or far from one another based on their degree of chemical similarity: points closer to one another indicate chemically similar molecules and therefore correspondingly low coverage of chemical space, while widely dispersed points indicate dissimilar molecules and therefore broad coverage of chemical space. The nearly 1500 COVID-moonshot molecules are also color-coded with their experimentally determined pIC₅₀ values, and our generated molecules in the vina-selected set are color-coded with their Vina scores.

The visualization clearly shows that the molecules generated by iMiner covers a broader chemical space and are spread evenly within plotting range than those molecules published on the COVID-moonshot website which form several tight clusters. We recognize that one of the reasons for the COVID moonshot molecules to be clustered in chemical space is that many of these molecules are generated through an inspirational approach, i.e., later molecules are borrowing designing ideas and sub-structures from molecules submitted earlier. By comparison, our final-54 set of molecules are dispersed throughout chemical space, which is an important characteristic of our workflow, since it provides a wide variety of structures as candidates for lead optimization. Interestingly, even compared to samples from the training dataset (ChEMBL), the molecules in the vina-selected set are still more diverse, which suggests the model was encouraged to explore more of chemical space during RL training while still reporting low Vina scores. Finally, we also see that the 54 molecules from iMiner are coming from different regions of the chemical space spanned by molecules generated from the model. As drug leads built on single or closely related scaffolds might be ruled out entirely during drug development, a wider coverage of the chemical space gives us a better chance of developing an effective lead for an Mpro inhibitor for treating SARS-CoV-2.

The molecular interactions between generated inhibitors of Mpro’s catalytic site. In this section we analyzed some of the molecules in the final-54 set through vi-

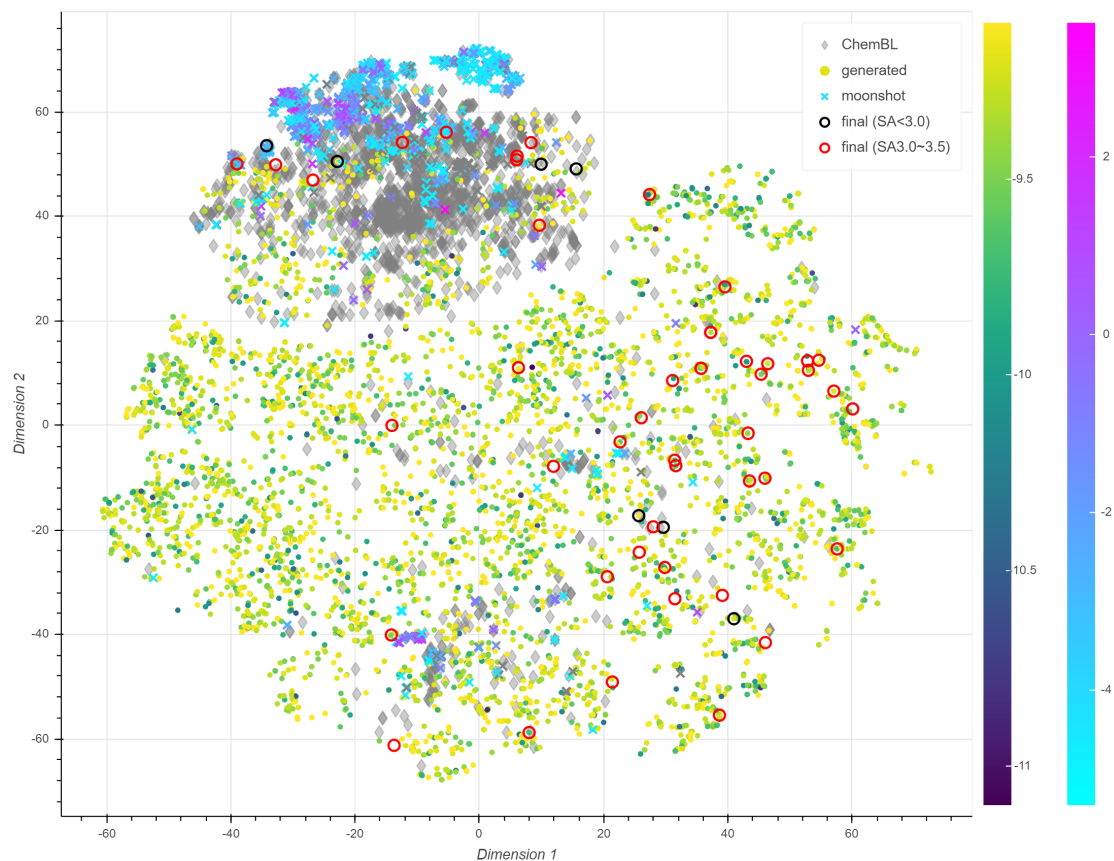


Figure 4: *Dimensionality-reduced latent space scatter plot for molecules from the COVID-moonshot project (crosses), generated molecules from the RL model (dots), molecules randomly sampled from the ChEMBL database (diamonds) and molecules in the final-54 set (circles).* Molecules on the figure are encoded by ChemNet³⁸ and the latent space vectors undergo dimensionality reduction by principal component analysis (PCA)⁴⁸ and t-distributed stochastic neighbor embedding (t-SNE).⁴⁷ Molecules from the COVID-moonshot project are color coded by their experimental pIC_{50} values according to the color bar on the right, and molecules generated by our model are color coded by Vina docking scores according to the color bar on the left.

sualizing their interactions with Mpro’s catalytic site. In Figure 5A, we show an overlay of several iMiner generated molecules in their optimal binding conformations determined through AutoDock Vina with respect to the surface of the binding pocket in which the predicted binding orientations fit nicely into the Mpro’s catalytic pocket. Additionally, ligand functional groups mirror the hydrophobicity requirements imposed by the Mpro binding site topography, meaning the generated molecules indeed have optimized interactions with the pocket. These results are no doubt due to our inclusion of real-time, explicit, flexible ligand docking in our evaluation model as well as a result of requiring minimization of Vina score distributions. Through this visualization we also see an interesting and encouraging result: although our final set of 54 molecules represent vastly different regions of chemical space, these molecules are relatively similar in size (i.e., similar number of heavy atoms), and the optimal docking conformations adopt similar shapes. These results illustrate the true power of our model, that we can quickly enumerate and expand upon the searched chemical space while still ensuring all generated molecules appropriately fit in the target protein pocket.

Figure 5B-E provide two representative examples of the molecular interactions between an iMiner predicted inhibitor and the Mpro binding site residues. Many and various types of favorable ligand-target interactions are observed, including hydrogen bonds, halogen bonds, and different types of π interactions. For example, CYS145 contributes to the π -Sulfur interaction in the first molecule illustrated in 5B and C, but it participates in a conventional hydrogen bond to the SO₂ group in the second molecule illustrated in 5D and E. Furthermore, when comparing the two proposed inhibitors each molecule exhibits unique interaction types to a different or complementary set of MPro protein residues. This variety in intermolecular interaction types stemming from the same protein binding site is a direct result of our enumeration of chemical space and our construction of novel ligand scaffolds.

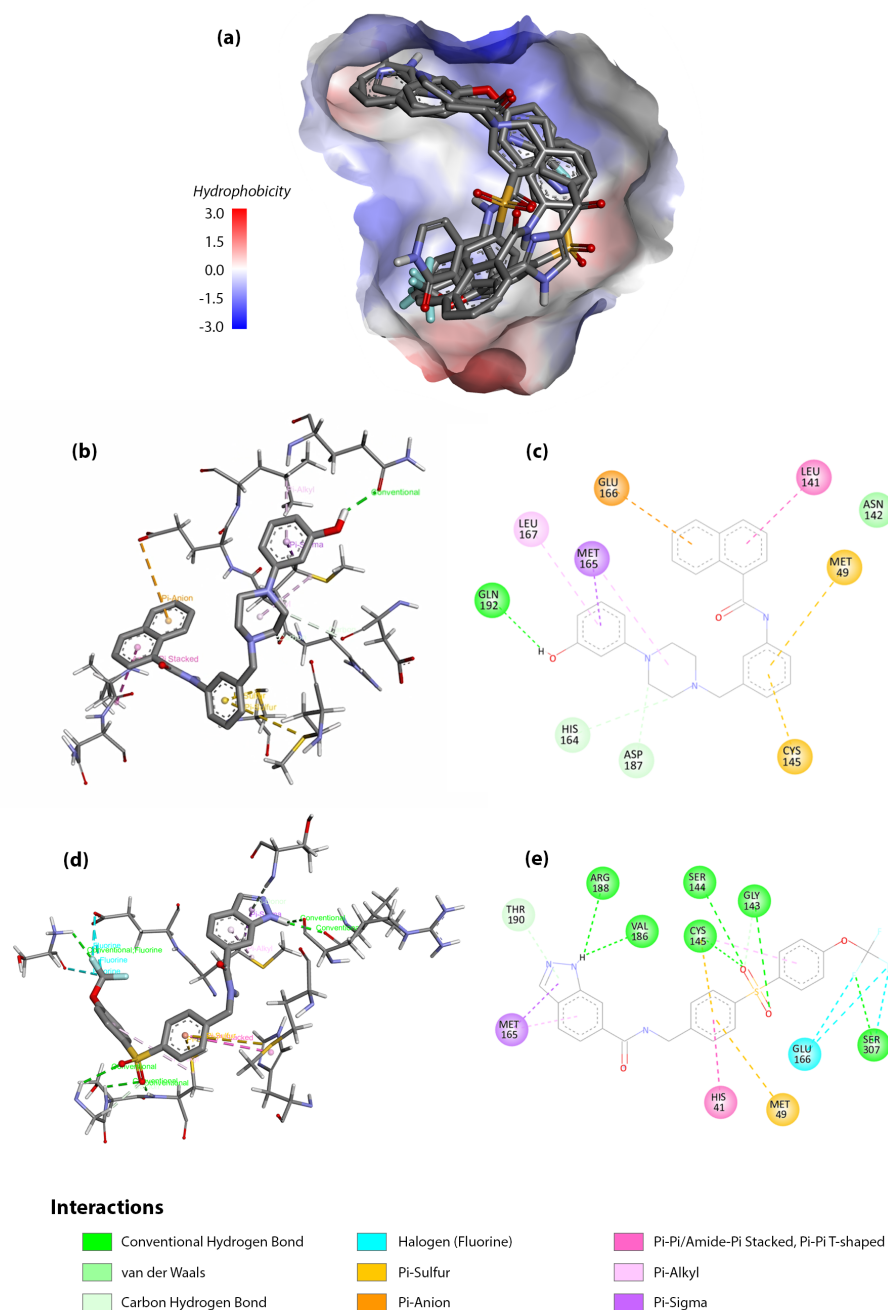


Figure 5: *Conformational and chemical compatibility of inhibitors predicted from iMiner for the MPro catalytic pocket.*(A) Randomly selected molecules from the final set of 54 inhibitors with their docking conformation determined by AutoDock Vina overlaid onto the surface of the binding pocket, with the surface color coded by hydrophobicity. Blue parts are hydrophilic and red parts are hydrophobic. (B) 3D interactions between molecule 1 and residues near the binding pocket (C) 2D illustrations for the interactions between molecule 1 and residues near the binding pocket (D) 3D interactions between molecule 2 and residues near the binding pocket (E) 2D illustrations for the interactions between molecule 2 and residues near the binding pocket. All figures generated by BIOVIA Discovery Studio Visualizer.⁴⁹

Conclusions

In this work we have shown that combining real-time docking of 3D structures with state-of-the-art reinforcement learning algorithms, we can efficiently navigate through uncharted regions of chemical space while maintaining good metrics for synthetic feasibility and drug-likeness. As illustrated using the exemplar target, the Mpro catalytic site, the ultimate final set of 54 inhibitor molecules proposed by our model are optimized with respect to shape and intermolecular interactions to the target protein, but are also diverse enough when compared to other predicted Mpro inhibitor datasets, i.e. molecules submitted to the COVID-moonshot project.³⁶ We understand the true effectiveness of these molecules as Mpro inhibitors can only be determined through experimental screening. Nevertheless, as we have seen agreement between AutoDock Vina and Glide SP results, and since we have visually inspected the predicted binding modes revealing consistency in intermolecular interactions to the Mpro pocket, we strongly believe there is good evidence that these molecules may be potent Mpro inhibitors.

Furthermore, every aspect of this work is generalizable. There are many well defined proteins vital to the replication of SARS-CoV-2 with 3D structures available including the RNA-dependent, RNA polymerase protein (RdRp),⁵⁰ the Papain-like protease (PLpro),⁵¹ and the exonuclease (ExoN).⁵² Although we have focused our current work on targeting SARS-CoV-2’s Mpro, extension of this work to these other targets would be relatively trivial. Although identifying antiviral treatments for SARS-CoV-2 is of pressing concern at the time of this publication, our model could be quickly applied to design novel inhibitors for proteins relevant to other global diseases. For example, bacterial resistance to antibiotics is of preeminent concern in the medical community,⁵³ and our iMiner workflow approach could be used to target novel bacterial biomolecules, such as bacterial Ribosomes, or target resistance conferring bacterial proteins such as β -lactamase.⁵³

Overall, we believe our tool will be of great benefit to the computational and medicinal chemistry fields at large, and potentially aid traditional drug-design workflows as well. For

example, molecules that are experimentally validated through a traditional HTVS approach as good binders could utilize the iMiner algorithm as an optimization or refinement step for elaborating on these existing leads or scaffolds. The potential of the method in this direction will be explored in future work.

Methods

Neural network architecture. The generative model employed in this study was an ASGD Weight-Dropped LSTM (AWD-LSTM),³² which is a specific variant of the Long Short Term Memory (LSTM) recurrent neural network with shared DropConnect for weight regularization, and was trained through a non-monotonically triggered average stochastic gradient descent (NT-ASGD) algorithm.^{32,54} The basic LSTM cell contains two internal states, the hidden state h_t and the cell state c_t , and can be described through the following set of equations:

$$i_t = \sigma(W^i x_t + U^i h_{t-1}) \tag{7}$$

$$f_t = \sigma(W^f x_t + U^f h_{t-1}) \tag{8}$$

$$o_t = \sigma(W^o x_t + U^o h_{t-1}) \tag{9}$$

$$\tilde{c}_t = \tanh(W^c x_t + U^c h_{t-1}) \tag{10}$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \tag{11}$$

$$h_t = o_t \odot \tanh c_t \tag{12}$$

where $[W^i, W^f, W^o, W^c, U^i, U^f, U^o, U^c]$ are the trainable parameters of the model, x_t is the input to the cell at the current timestep, \tilde{c}_t contains the information to be added to the cell state, and i_t, f_t, o_t represent the update gate, forget gate and output gate respectively, which are numbers between (0, 1) that controls how much information should be updated, discarded

or retrieved from the cell state. σ denotes the sigmoid function, and \odot represents element-wise multiplication. The DropConnect mechanism⁵⁵ was applied to the hidden-to-hidden weight matrices $[U^i, U^f, U^o, U^c]$ by randomly zeroing out a small portion of the parameters in these weight matrices to prevent overfitting and ensured that the same positions in the hidden vectors were treated consistently throughout the forward and backward pass in regards to whether or not to be dropped.

The inputs into the RNN cells were tokens embedded as vectors of length 400, and 3 LSTM cells were stacked sequentially, that had 1152, 1152 and 400 units each. The hidden state from the last timestep of the last RNN cell was then connected to a linear decoder with output size of 56 and softmax activation, representing the probabilities of the 56 possible tokens from the vocabulary. The dropout values used in the model were: embedding dropout=0.002, LSTM weight dropout=0.02, RNN hidden state dropout=0.015 and output dropout=0.01. The neural network was implemented using pyTorch⁵⁶ and the fastai package.⁵⁷

Supervised pretraining of the network The generative model was pretrained using molecules from ChEMBL 24,¹³ and a total of 1,440,263 molecules were selected for training. All molecules were first converted to SELFIES strings using the selfies python package,³¹ and the tokens were extracted from the SELFIES strings with fastai language model. We used categorical cross entropy loss:

$$L_{\Theta} = -\frac{1}{N} \sum_{i=1}^N \sum_{t_i} \hat{p}(t_i|t_1, t_2, \dots, t_{i-1}) \log p_{\Theta}(t_i|t_1, t_2, \dots, t_{i-1}) \quad (13)$$

where N represents the number of tokens in a molecule, $\hat{p}(t_i|t_1, t_2, \dots, t_{i-1})$ represents the actual probability of a specific token in the string at position i and with all previous defined tokens t_1 through t_{i-1} , and $p_{\Theta}(t_i|t_1, t_2, \dots, t_{i-1})$ the probability predicted by the neural network with parameters Θ . The model was trained using Adam optimizer⁵⁸ in batches of size 512, and we employed the ‘‘one cycle’’ learning rate policy⁵⁹ with the maximum learning rate

of 0.0005 to achieve superconvergence.⁶⁰ During this pretraining stage we also used weight decay=0.01 and the dropout multiplier of 0.2. The model was pretrained for 30 epochs.

Reinforcement learning procedure. Our RL training target goal is to maximize $J(\Theta)$ from formula(6) by taking steps along $\partial_{\Theta}J(\Theta)$. The exact value for $J(\Theta)$ is intractable to evaluate, but can be approximated through sampling the distribution of s_T , which gives

$$J(\Theta) \approx \sum_{s_T} p_{\Theta}(s_T)r(s_T) \quad (14)$$

and then

$$\partial_{\Theta}J(\Theta) = \sum_{s_T} [\partial_{\Theta}p_{\Theta}(s_T)]r(s_T) \quad (15)$$

$$= \sum_{s_T} p_{\Theta}(s_T) \left[\sum_{t=1}^T \partial_{\Theta} \log p_{\Theta}(s_t|s_{t-1}) \right] r(s_T) \quad (16)$$

Directly taking gradients according to (16) corresponds to the REINFORCE algorithm.⁶¹ In this work we further utilized the PPO algorithm,⁴² which estimated the gradients through a clipped reward and with an extra entropy bonus term:

$$J'(\Theta) = \sum_{s_T} p_{\Theta}(s_T) \left[\sum_{t=1}^T J_t^{\text{CLIP}}(\Theta) + \alpha S[p_{\Theta}(s_t|s_{t-1})] \right] \quad (17)$$

where

$$J_t^{\text{CLIP}}(\Theta) = \min(R_t(\Theta)r(s_T), \text{clip}(R_t(\Theta), 1 - \epsilon, 1 + \epsilon)r(s_T)) \quad (18)$$

with

$$R_t(\Theta) = \frac{p_{\Theta}(s_t|s_{t-1})}{p_{\Theta_{\text{old}}}(s_t|s_{t-1})} \quad (19)$$

denoting the ratio between the probability distribution in the current iteration and the prob-

ability distribution from the previous iteration (the iteration before last gradient update). A PPO algorithm reduces variance in the gradient, stabilizes training runs, and also encourages the model to explore a wider region of the chemical space through the introduction of an entropy bonus term. The two hyperparameters in the algorithm, α and ϵ , were taken as $\alpha = 0.02, \epsilon = 0.1$ in this work.

After the pretraining finished, we copied the weights to a separate model with identical architecture and trained with reinforcement learning using PPO. In each iteration 2000 molecules were sampled, and model weights were updated by taking gradient steps on the target function through formula (17), using a batch size of 1024 and Adam optimizer with fixed learning rate of 0.0001. In each iteration, all collected data were used for training the model for a maximum of 10 epochs. The trainer would continue into next iteration and collect new molecules for training if the K-L divergence between the latest predicted probability and the old probability exceeded 0.03.

The model was trained with RL for 400 iterations, until the mean entropy of the predicted probability of the tokens from the RNN started to decrease drastically. The change of mean entropy and mean vina score during the RL training can be found in Figure S4.

Acknowledgement

The authors thank Greg Merritt, Ian Lumb, and Mike Smith for their generous help in setting up the MicroSoft Azure cloud computing infrastructure and credits for the project. The authors thank the National Institutes of Health for support under Grant No 5U01GM121667. This work was also supported in part by the C3.ai Digital Transformation Institute. This research used Microsoft Azure through C3.ai DTI award, and also computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author Contributions

T.H-G. and J.L. conceived the scientific direction, designed the experiments, analyzed results, and wrote the manuscript. F.L.K. prepared Mpro structures for docking in Glide and AutoDock Vina, conducted Glide SP docking, and read and edited the manuscript. C.P. wrote the code for pre-training the network. O.Z. wrote the code for the AutoDock-vina workflow. N.G., I.L. and M.H. contributed ideas and to discussions to the work. R.E.A. provided drug-design and biophysical guidance, coordinated with experimental/medicinal chemists, read/revised the manuscript.

Supporting Information Available

Methodology details, supporting table S1, supporting figures S1-S4.

Declaration of Interests

CP has equity interest in Athae Bio. The other authors declare no competing interests.

References

- (1) WHO, WHO Coronavirus (COVID-19) Dashboard — WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. <https://covid19.who.int/>.
- (2) Pormohammad, A.; Zarei, M.; Ghorbani, S.; Mohammadi, M.; Razizadeh, M. H.; Turner, D. L.; Turner, R. J. Efficacy and Safety of COVID-19 Vaccines: A Systematic Review and Meta-Analysis of Randomized Clinical Trials. *Vaccine* **2021**, *9*, 467.
- (3) Davies, M.; Osborne, V.; Lane, S.; Roy, D.; Dhanda, S.; Evans, A.; Shakir, S. Remdesivir in treatment of COVID-19: a systematic benefit–risk assessment. *Drug safety* **2020**, *43*, 645–656.

- (4) Cui, W.; Yang, K.; Yang, H. Recent Progress in the Drug Development Targeting SARS-CoV-2 Main Protease as Treatment for COVID-19. *Frontiers in Molecular Biosciences* **2020**, *7*, 398.
- (5) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C., et al. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289–293.
- (6) Hegyi, A.; Ziebuhr, J. Conservation of substrate specificities among coronavirus main proteases. *Journal of general virology* **2002**, *83*, 595–599.
- (7) Fehr, A. R.; Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Coronaviruses* **2015**, 1–23.
- (8) 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.
- (9) Chen, Y.; Yang, W.-H.; Huang, L.-M.; Wang, Y.-C.; Yang, C.-S.; Liu, Y.-L.; Hou, M.-H.; Tsai, C.-L.; Chou, Y.-Z.; Huang, B.-Y., et al. Inhibition of Severe Acute Respiratory Syndrome Coronavirus 2 main protease by tafenoquine in vitro. *Biorxiv* **2020**,
- (10) Su, H.; Yao, S.; Zhao, W.; Li, M.; Liu, J.; Shang, W.; Xie, H.; Ke, C.; Gao, M.; Yu, K., et al. Discovery of baicalin and baicalein as novel, natural product inhibitors of SARS-CoV-2 3CL protease in vitro. *BioRxiv* **2020**,
- (11) Yang, H.; Xie, W.; Xue, X.; Yang, K.; Ma, J.; Liang, W.; Zhao, Q.; Zhou, Z.; Pei, D.; Ziebuhr, J., et al. Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS biology* **2005**, *3*, e324.
- (12) Boopathi, S.; Poma, A. B.; Kolandaivel, P. Novel 2019 coronavirus structure, mecha-

- nism of action, antiviral drug promises and rule out against its treatment. *Journal of Biomolecular Structure and Dynamics* **2021**, *39*, 3409–3418.
- (13) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012**, *40*, D1100–D1107.
- (14) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.
- (15) Shivanyuk, A.; Ryabukhin, S.; Tolmachev, A.; Bogolyubsky, A.; Mykytenko, D.; Chupryna, A.; Heilman, W.; Kostyuk, A. Enamine real database: Making chemical diversity real. *Chemistry today* **2007**, *25*, 58–59.
- (16) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **2021**, *49*, D1388–D1395.
- (17) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function. *ChemMedChem: Chemistry Enabling Drug Discovery* **2008**, *3*, 885–897.
- (18) Duffy, B. C.; Zhu, L.; Decornez, H.; Kitchen, D. B. Early phase drug discovery: cheminformatics and computational techniques in identifying lead series. *Bioorganic & medicinal chemistry* **2012**, *20*, 5324–5342.
- (19) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

- (20) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar variational autoencoder. International Conference on Machine Learning. 2017; pp 1945–1954.
- (21) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786* **2018**,
- (22) Subramanian, A.; Saha, U.; Sharma, T.; Tailor, N. K.; Satapathi, S. Inverse Design of Potential Singlet Fission Molecules using a Transfer Learning Based Approach. *arXiv preprint arXiv:2003.07666* **2020**,
- (23) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **2017**, *9*, 1–14.
- (24) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances* **2018**, *4*, eaap7885.
- (25) Gottipati, S. K.; Sattarov, B.; Niu, S.; Pathak, Y.; Wei, H.; Liu, S.; Blackburn, S.; Thomas, K.; Coley, C.; Tang, J., et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. International Conference on Machine Learning. 2020; pp 3668–3679.
- (26) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A., et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **2019**, *37*, 1038–1040.
- (27) Zhavoronkov, A.; Aladinskiy, V.; Zhebrak, A.; Zagribelnyy, B.; Terentiev, V.; Bezrukov, D. S.; Polykovskiy, D.; Shayakhmetov, R.; Filimonov, A.; Orekhov, P., et al. Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches. **2020**,

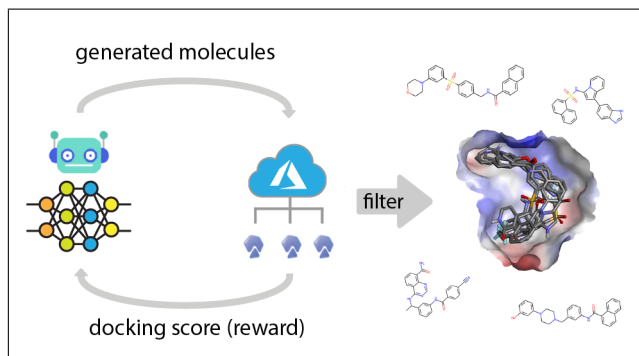
- (28) Bung, N.; Krishnan, S. R.; Bulusu, G.; Roy, A. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future medicinal chemistry* **2021**, *13*, 575–585.
- (29) Born, J.; Manica, M.; Cadow, J.; Markert, G.; Mill, N. A.; Filipavicius, M.; Janakaraman, N.; Cardinale, A.; Laino, T.; Martínez, M. R. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Machine Learning: Science and Technology* **2021**, *2*, 025024.
- (30) Jeon, W.; Kim, D. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific reports* **2020**, *10*, 1–11.
- (31) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100representation. *Machine Learning: Science and Technology* **2020**, *1*, 045024.
- (32) Merity, S.; Keskar, N. S.; Socher, R. Regularizing and Optimizing LSTM Language Models. 2018.
- (33) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010**, *31*, 455–461.
- (34) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *Journal of Medicinal Chemistry* **2006**, *49*, 6177–6196.
- (35) Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. *Journal of medicinal chemistry* **2015**, *58*, 2091–2113.

- (36) Achdout, H.; Aimon, A.; Bar-David, E.; Barr, H.; Ben-Shmuel, A.; Bennett, J.; Bobby, M. L.; Brun, J.; BVNBS, S.; Calmiano, 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**,
- (37) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **2019**, *59*, 1096–1108.
- (38) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling* **2018**, *58*, 1736–1741.
- (39) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **2012**, *4*, 90–98.
- (40) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences* **1999**, *39*, 868–873.
- (41) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **2008**, *3*, 435.
- (42) Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* **2017**,
- (43) Everitt, T.; Hutter, M. Avoiding wireheading with value reinforcement learning. International Conference on Artificial General Intelligence. 2016; pp 12–22.

- (44) Houston, D. R.; Walkinshaw, M. D. Consensus docking: improving the reliability of docking in a virtual screening context. *Journal of chemical information and modeling* **2013**, *53*, 384–390.
- (45) Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports* **2017**, *7*, 1–13.
- (46) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **1997**, *23*, 3–25.
- (47) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
- (48) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* **1987**, *2*, 37–52.
- (49) SYSTEMES, D. BIOVIA Discovery Studio. 2016; <http://accelrys.com/products/collaborative-science/biovia-discovery-studio/>.
- (50) Ahmad, J.; Ikram, S.; Ahmad, F.; Rehman, I. U.; Mushtaq, M. SARS-CoV-2 RNA Dependent RNA polymerase (RdRp)—A drug repurposing study. *Heliyon* **2020**, *6*, e04502.
- (51) Klemm, T.; Ebert, G.; Calleja, D. J.; Allison, C. C.; Richardson, L. W.; Bernardini, J. P.; Lu, B. G.; Kuchel, N. W.; Grohmann, C.; Shibata, Y., et al. Mechanism and inhibition of the papain-like protease, PLpro, of SARS-CoV-2. *The EMBO journal* **2020**, *39*, e106275.
- (52) Moeller, N. H.; Shi, K.; Demir, Ö.; Banerjee, S.; Yin, L.; Belica, C.; Durfee, C.; Amaro, R. E.; Aihara, H. Structure and dynamics of SARS-CoV-2 proofreading exoribonuclease ExoN. *bioRxiv* **2021**,

- (53) Ventola, C. L. The antibiotic resistance crisis: part 1: causes and threats. *P & T : a peer-reviewed journal for formulary management* **2015**, *40*, 277–83.
- (54) Polyak, B. T.; Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* **1992**, *30*, 838–855.
- (55) Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; Fergus, R. Regularization of neural networks using dropconnect. International conference on machine learning. 2013; pp 1058–1066.
- (56) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. **2017**,
- (57) Howard, J.; Gugger, S. Fastai: a layered API for deep learning. *Information* **2020**, *11*, 108.
- (58) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,
- (59) Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820* **2018**,
- (60) Smith, L. N.; Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. 2019; p 1100612.
- (61) Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **1992**, *8*, 229–256.

Graphical TOC Entry



Supplementary Information: Reinforcement Learning with Real-time Docking of 3D Structures to Cover Chemical Space: Mining for Potent SARS-CoV-2 Main Protease Inhibitors

Jie Li,[†] Oufan Zhang,[†] Fiona L. Kearns,[‡] Mojtaba Haghighatlari,[†] Conor
Parks,[‡] Xingyi Guan,[†] Itai Leven,[†] Rommie E. Amaro,[‡] and Teresa
Head-Gordon^{*,†,¶}

[†]*Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California,
Berkeley, CA, USA*

[‡]*Department of Chemistry and Biochemistry University of California, San Diego, La Jolla, CA
92093-0340*

[¶]*Departments of Bioengineering and Chemical and Biomolecular Engineering, University of
California, Berkeley, CA, USA*

E-mail: thg@berkeley.edu

Methodology Details

Tokens in the generative model. Here we provide a complete list of tokens used in the generative model:

- Standard SELFIES tokens: ['#C', '#N', '#O', '#S', '=B', '=C', '=I', '=N', '=O', '=P', '=S', '=Se', '=Si', 'B', 'Br', 'Br+2', 'Branch1_1', 'Branch1_2', 'Branch1_3', 'Branch2_1',

'Branch2_2', 'Branch2_3', 'C', 'Cl', 'Cl+2', 'Cl+3', 'Expl=Ring1', 'Expl=Ring2', 'F', 'I', 'I+2', 'I+3', 'N', 'O', 'P', 'Ring1', 'Ring2', 'S', 'Se', 'Si']

- Modifier tokens: ["H+expl", "H2+expl", "H3+expl", "+expl", "Hexpl", "H2expl", "H-expl", "H2-expl", "H3-expl", "-expl", "expl"]
- Functional tokens: ["Break"]

When sampling molecules represented as SELFIES strings, the first token was always selected as the "Break" token. Then each token was sampled with probability distribution predicted by the generative model. Once the "Break" token was selected again, or the total number of tokens exceeded 500, a single molecule sampling process was considered complete. For each modifier token in the sampled string, it was combined into the previous token and was connected by the "" symbol. For example, ...[C][Hexpl]... would be converted to ...[C^Hexpl] to satisfy SELFIES syntax. If the token before a modifier token could not be modified, the sampled string would be considered invalid and was discarded.

Docking Preparation and Procedures. Stzain et al.² simulated SARS-CoV-2 Mpro (PDB ID 6LU7²) with Gaussian Accelerated Molecular Dynamics to characterize active site and dimer interface dynamics, as well as elucidate the presence of cryptic binding pockets. In total, Stzain et al. produced 6 microseconds of enhanced-sampled Mpro conformations.² These extensive simulations represent an invaluable resource for SARS-CoV-2 antiviral design, and as such Stzain et al. shared their trajectories publicly (<https://amarolab.ucsd.edu/covid19.php>) in accordance with the data sharing philosophy put forth by Amaro and Mulholland.² To ensure we were selecting biologically relevant Mpro conformations for use in our molecule generation workflow, we selected receptor structures from Stzain et al.'s simulations. Selection of each receptor structure and subsequent protein preparation steps are described below.

Mpro Active Site Receptor Selection and Preparation: To generate molecules targeting the Mpro active site, we selected the representative structure from the most populated cluster identified

in Sztain et al.'s enhanced sampling trajectories of Mpro dimer,² simulated with a covalently bound inhibitor called N3. From Sztain et al.'s freely available files, the filename of the selected protease structure was "5.0_2.0_147.0_147.0_295.0_c0.pdb". We deleted the covalently bound N3 from this structure, taking care not to delete the catalytic Cysteine atoms (resids 145 and 451). We then modified the C145 and C451 atom names so that they reflected canonical Cysteine atom names. The Mpro structure was then prepared with AutoDockTools² and Schrödinger's Protein Preparation Wizard for docking in AutoDock Vina and Glide Ligand Docking, respectively (see Protein Preparation section below for more details). The cartesian coordinates for the active site center were found by calculating the center of mass of the C145 bound N3I covalent inhibitor before the inhibitor was deleted ([atomselect top "resname N3I and resid 145"]). This center of mass (x=54.58, y=45.92, z=75.06) was used to define the center of the active site during receptor grid generation steps in AutoDock Vina and Glide docking.

Scoring Generated Molecules with AutoDock Vina. AutoDockTools² was used to convert Mpro .pdb files to AutoDock Vina² compatible .pdbqt files. Additionally, AutoDockTools² was used to convert generated molecule structure files to AutoDock Vina compatible .pdbqt files. Gastieger charges were used for all AutoDock Vina structures. A cubic receptor grid of 30Å x 30Å x 30Å was centered around binding site's central coordinate (listed above), with a grid spacing of 1.0Å.

Re-scoring Generated Molecules with Schrödinger's Glide Ligand Docking. As Schrödinger's Grid-Based Ligand Docking and Energetics (Glide) protocol^{2,2} is one of the most well-trusted docking protocols available, we re-scored all our generated molecules in each Mpro binding site with Glide Standard Precision docking.^{2,2} To do so, we prepared each Mpro protein/receptor structure and all generated molecule structures for Glide docking. Schrödinger's Protein Preparation Wizard^{2,2} was used to prepare the Mpro receptor structures selected from Sztain et al.'s trajectories for Glide docking according to the following settings: Bond orders were calculated, missing hydrogens were added, and disulfide bonds were created all according to default options. Protein protonation states were assigned with PropKa around pH=7.0.^{2,2} A restrained minimization of all

hydrogen atoms was then conducted according to the OPLS4 force field.[?]

Schrödinger's Receptor Grid Generation tool was used to prepare the Mpro structure for Glide docking according to the following settings: The center of the binding site was defined according to the center calculated above. The outer grid box size was set to 30Å x 30Å x 30Å, inner grid box size was set to 10Å x 10Å x 10Å. Grid points were placed every 1.0 Å. Receptor atom van der Waal radii were not scaled (i.e., scaled by a factor of 1.00) and the charge cut off for polarity was set to 0.25. Atom types were assigned according to OPLS 2005 atom types.[?]

To ensure we were utilizing identical molecules for comparison between AutoDock Vina results and Glide SP results, i.e. with respect to stereochemistry, we took output structures from AutoDock Vina (in .pdbqt format) and converted (with Open Babel[?]) first to .pdb files and then (with Open Babel) to .sdf files (SDF files being compatible for Schrödinger's LigPrep). Schrödinger's LigPrep module was then used to prepare all AutoDock Vina output structures for docking with Glide according to the following settings: Max allowed number of atoms per molecule was set to the default of 500. To again ensure that we docked structures identical to those docked with AutoDock, ionization states were not generated, tautomers were not generated, and chiral centers were not varied. Molecules were minimized according to the OPLS3 force field[?] and structures were written to .mae format for docking with Glide.

Schrödinger's Glide Ligand Docking^{???} module was used to re-score all generated molecules according to the Glide SP scoring function. The following settings were used during Glide SP ligand docking: Ligands were docked into each respective receptor according to a flexible ligand/rigid receptor docking protocol in which ligand bonds, angles and dihedral degrees of freedom were explored during docking. The top binding mode per molecule was saved and a Standard Precision Glide score was reported in kcal/mol for each molecule. The OPLS4 force field[?] was used for energetic evaluations and scoring. Glide SP scores were then compared, for each molecule, to AutoDock Vina scores.

All docking input files and protein structures will be shared in conjunction the data sharing philosophy put forth by Rommie E. Amaro and Adrian Mulholland.[?]

Supporting Tables

Table S1: The 54 molecules from the final set

Index	Canonical SMILES	Vina score	Glide gscore	SA score
1	<chem>Oc1cccc(c1)N1CCN(CC1)Cc1cccc(c1)NC(=O)c1cccc2c1cccc2</chem>	-9.10	-8.07	2.85
2	<chem>O=C(c1ccc2c(c1)[nH]nc2)NCc1ccc(cc1)S(=O)(=O)c1ccc(cc1)OC(F)(F)F</chem>	-9.40	-8.14	2.72
3	<chem>Fc1ccc(cc1C(=O)Nc1ccnc2c1cc[nH]c2=O)Oc1ccc2c(c1)cccc2</chem>	-9.40	-8.02	2.88
4	<chem>O=C(c1c[nH]c(n1)c1cccc2c1cccc2)c1cccc(c1)NC(=O)c1cccc1</chem>	-9.10	-8.17	2.97
5	<chem>N#Cc1ccc(cc1)c1cccc(c1)C(=O)N1CCN(CC1)C(=O)c1cccc2c1cccc2</chem>	-9.40	-8.34	2.98
6	<chem>Oc1n(Cc2ccc(cc2)S(=O)(=O)c2[nH]c3c(n2)cccc3)nc2c1cccc2</chem>	-9.10	-8.09	2.85
7	<chem>O=S(=O)(c1cc(cc2c1nccc2)c1ccncc1)c1cccc(c1)C(F)(F)F</chem>	-9.20	-8.20	2.99
8	<chem>N#Cc1ccc(cc1)C(=O)Nc1cccc(c1)C(Nc1ncnc2c1cccc2C(=O)N)C</chem>	-9.10	-8.47	3.34
9	<chem>O=C(C(F)(F)F)N1CCN(CC1)C(=O)c1cc(ccc1F)Cc1nnc(c2c1cccc2)O</chem>	-9.30	-8.32	3.11
10	<chem>O=C(Nc1scc(n1)c1ccc(cc1)n1cncn1)Cn1nnc(c1)c1cccc1</chem>	-9.10	-8.33	3.44
11	<chem>N#Cc1ccc(cc1)C(=O)N1CCN(CC1)C(=O)COC(=O)c1ccc(c(c1)c1cccc1)O</chem>	-9.40	-8.71	3.29
12	<chem>Fc1ccc(cc1)S(=O)(=O)c1ccc(cc1)S(=O)(=O)Nc1ccc(cc1)C1=NCCN1</chem>			

Table S1 (continued)

Index	Canonical SMILES	Vina score	Glide gscore	SA score
		-9.20	-8.10	3.35
13	<chem>FC(c1ccc(cc1)S(=O)(=O)Nc1cccc(c1)OCc1ccc2c(n1)cccc2)(F)F</chem>	-9.10	-8.01	3.07
14	<chem>OC(=NCc1ccc(cn1)S(=O)(=O)c1ccc(nc1)C(F)(F)F)c1cnc2n(c1)ccn2</chem>	-9.10	-8.18	3.47
15	<chem>O=C(c1ccc2c(c1)cccc2)NCc1ccc(cc1)S(=O)(=O)c1cccc(c1)N1CCOCC1</chem>	-9.20	-8.24	3.21
16	<chem>O=C(c1cccs1)NN=Cc1ccc(cc1)OCc1ccc(cc1)c1ccccc1</chem>	-9.10	-8.86	3.34
17	<chem>Fc1ccc(cc1)c1cncc(c1)C(=O)N1CC(C1)Oc1ccc2c(c1)c1cnccc1[nH]2</chem>	-9.50	-8.04	3.31
18	<chem>Fc1cnc(nc1c1cnc2c(c1)cccc2)Nc1cc(C)c(c(c1)c1[nH]cnn1)F</chem>	-10.00	-8.09	3.12
19	<chem>Fc1ccc(cc1)N1CCN(CC1)C(=O)Nc1nonc1NC(=O)c1ccc(cc1)F</chem>	-9.10	-8.10	3.33
20	<chem>O=S(=O)(c1cccc(c1)C(F)(F)F)Nc1cccc(c1)c1ccc(nn1)Nc1cnccc1</chem>	-9.30	-8.25	3.31
21	<chem>Fc1ccc(cc1)c1n[nH]c2c1cc(cc2)NS(=O)(=O)c1cccc(c1)c1ccc[nH]1</chem>	-9.30	-8.03	3.25
22	<chem>O=c1c2cccc3c2n(c2c1cccc2)sc3NS(=O)(=O)C=Cc1ccccc1</chem>	-9.20	-8.20	3.44
23	<chem>O=S(=O)(c1cccc(c1)C(F)(F)F)Nc1cccc(c1)c1ccc(nc1)Nc1ccccc1</chem>	-9.40	-8.54	3.35
24	<chem>Oc1cccc(c1)c1cccc(c1)S(=O)(=O)Nc1ccc(cc1)c1cccc(c1)c1nnn[nH]1</chem>	-9.60	-8.50	3.38

Table S1 (continued)

Index	Canonical SMILES	Vina score	Glide gscore	SA score
25	<chem>O=C1NCc2c(C1)ccc(c2)S(=O)(=O)c1cn(nc1c1ccc(cc1F)F)c1cccc1</chem>	-9.50	-8.29	3.45
26	<chem>O=C1Cc2c(N1)cc(cc2)c1cccc2c1ccc(c2)NS(=O)(=O)c1cccc(c1)C(F)(F)F</chem>	-9.60	-8.02	3.03
27	<chem>Fc1ccc(cc1)C(=O)Nc1cccc(c1)C(Nc1ncnc2c1cccc2C(=O)N)C</chem>	-9.10	-8.18	3.22
28	<chem>Fc1ccn2c(c1)ncc2c1cc(ccc1F)NS(=O)(=O)c1cccc(c1)c1cccc1</chem>	-9.80	-8.23	3.40
29	<chem>O=S(=O)(c1cccc2c1cccc2)Nc1cc(c2n1cccc2)c1ccc2c(c1)[nH]cn2</chem>	-9.30	-8.05	3.25
30	<chem>O=S(=O)(c1ccc(cc1)Oc1cccc(c1)c1cc2c([nH]1)cccc2)Nc1n[nH]1</chem>	-9.20	-8.46	3.29
31	<chem>Fc1cccc1CS(=O)(=O)Nc1ccc2c(c1)c(n[nH]2)c1nc2c([nH]1)cccc2</chem>	-9.20	-8.43	3.06
32	<chem>Fc1ccc(cc1)S(=O)(=O)c1ccc(cc1)NS(=O)(=O)c1ncnc(c1)c1cccc1</chem>	-9.60	-8.32	3.26
33	<chem>Fc1cccc1Nc1nc2c([nH]1)ccc(c2)S(=O)(=O)c1cccc2c1nccc2</chem>	-9.20	-8.02	3.10
34	<chem>Oc1ccc2c(c1)cc(o2)c1cccc(c1)S(=O)(=O)c1nc2c(s1)cccc2F</chem>	-9.40	-8.46	3.49
35	<chem>O=S(=O)(c1cccc1)N(Cc1ccnc1)Cc1ccc(cc1)Oc1ccc(cc1)c1[nH]cn1</chem>	-9.20	-8.36	3.32
36	<chem>O=S(=O)(c1cccc(c1)c1ccncc1)c1cccc(c1)C=Nc1nc2c([nH]1)cccc2</chem>	-9.50	-8.07	3.33
37	<chem>Fc1ccc(cc1)S(=O)(=O)c1ccc(cc1)NS(=O)(=O)c1ncnc(c1)c1cccc1</chem>			

Table S1 (continued)

Index	Canonical SMILES	Vina score	Glide gscore	SA score
38	<chem>Oc1cccc(c1)C=Nc1cccc(c1)NS(=O)(=O)c1cccc(c1)c1cccc1</chem>	-9.60	-8.32	3.26
39	<chem>O=S(=O)(N1CCc2c(C1)nc(nc2)Nc1cccc1)c1ccc(cc1)c1ccccn1</chem>	-9.20	-8.16	3.38
40	<chem>O=S(=O)(c1cccc(c1)c1cccc1)Nc1cccc(c1)c1ncc2c(c1)cccc2</chem>	-9.70	-8.16	3.44
41	<chem>Oc1cccc(c1)Nc1nc2c([nH]1)ccc(c2)S(=O)(=O)c1cccc(c1)c1cccc1</chem>	-9.90	-8.06	3.25
42	<chem>O=S(=O)(c1cccc(c1)c1cnc2n1ccnc2)c1ccc(cc1)c1cccc1</chem>	-9.70	-8.16	3.21
43	<chem>O=S(=O)(c1cccc(c1)c1cnc2n1ccnc2)c1ccc(cc1)c1cccc1</chem>	-9.20	-8.13	3.29
44	<chem>Fc1cnc2c(c1)ncn2c1cccc(c1)S(=O)(=O)c1nc2c(s1)cc(cc2)c1ccccn1</chem>	-9.30	-8.31	3.47
45	<chem>Oc1ncn1c1ccc2c(c1)cccc2NS(=O)(=O)c1cccc(c1)c1cccc1</chem>	-9.60	-8.18	3.19
46	<chem>O=S(=O)(Oc1n[nH]c(c1)c1cccc1)Oc1cccc(c1)c1cnc2n1cccc2</chem>	-9.20	-8.16	3.46
47	<chem>O=S(=O)(c1ccc2c(c1)cccc2)Nc1cccc(c1)c1coc(n1)c1cccc1</chem>	-9.70	-8.46	3.48
48	<chem>O=S(=O)(c1cccc(c1)c1cccc1)c1ccc(cc1)Nc1ncc2c(c1)ccnc2</chem>	-9.20	-8.21	3.24
49	<chem>O=S(=O)(c1cccc(c1)c1cccc1)Nc1ccc2c(c1)cc(cc2)S(=O)(=O)N</chem>	-9.10	-8.03	3.08
49	<chem>Fc1ccc2c(n1)ncc(c2O)Cc1ccc(cc1)S(=O)(=O)c1ccc2c(n1)cccc2</chem>	-9.40	-8.03	3.09

Table S1 (continued)

Index	Canonical SMILES	Vina score	Glide gscore	SA score
50	<chem>O=S(=O)(c1cccc(c1)c1ccccc1)Nc1cc2c(c1)cncn2Cc1[nH]jenn1</chem>	-9.20	-8.22	3.23
51	<chem>Fc1ccc(cc1OS(=O)(=O)c1ccc2c(c1)[nH]cc2)c1cccc(c1)N1CCNCC1</chem>	-9.30	-8.17	3.45
52	<chem>Oc1nn(O)[nH]c(c1)c1cccc(c1)S(=O)(=O)c1cccc(c1)c1cncnc1</chem>	-9.10	-8.71	3.47
53	<chem>Fc1ccc(cc1NS(=O)(=O)c1cccc(c1)c1ccnc1)c1[nH]c2c(n1)nccc2</chem>	-9.60	-8.01	3.21
54	<chem>Oc1ncc(cc1c1cccc(c1)S(=O)(=O)c1cccc(c1)c1ccccc1)c1ccncc1</chem>	-10.00	-8.30	3.44

Supporting Figures

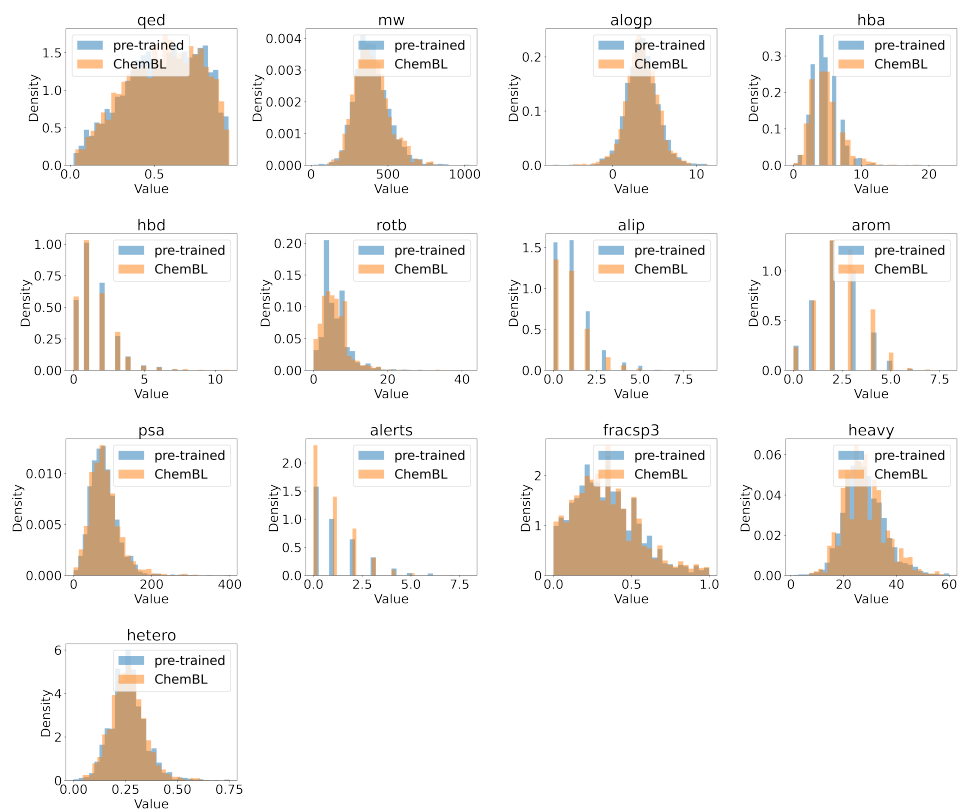


Figure S1: *Distribution comparisons for 13 different properties of the generated molecules from the pretrained model with molecules from the training dataset (ChEMBL).* The molecular properties considered are well-recognized chemical features related to the drug-likeness of a molecule which can be obtained through 2D topological connectivity of the molecule: fraction of sp^3 hybridized carbons(fracsp3), number of heavy atoms(heavy), fraction of non-carbon atoms in all heavy atoms(hetero), number of hydrogen bond donors(hbd) and acceptors(hba), number of rotatable bonds(rotb), number of aliphatic(alip) and aromatic rings(arom), molecular weight(mw), quantitative estimate of drug-likeness (QED) value,² approximate log partition coefficient between octanol and water (alogP),² polarizable surface area (PSA), and the number of structural alerts(alerts).²

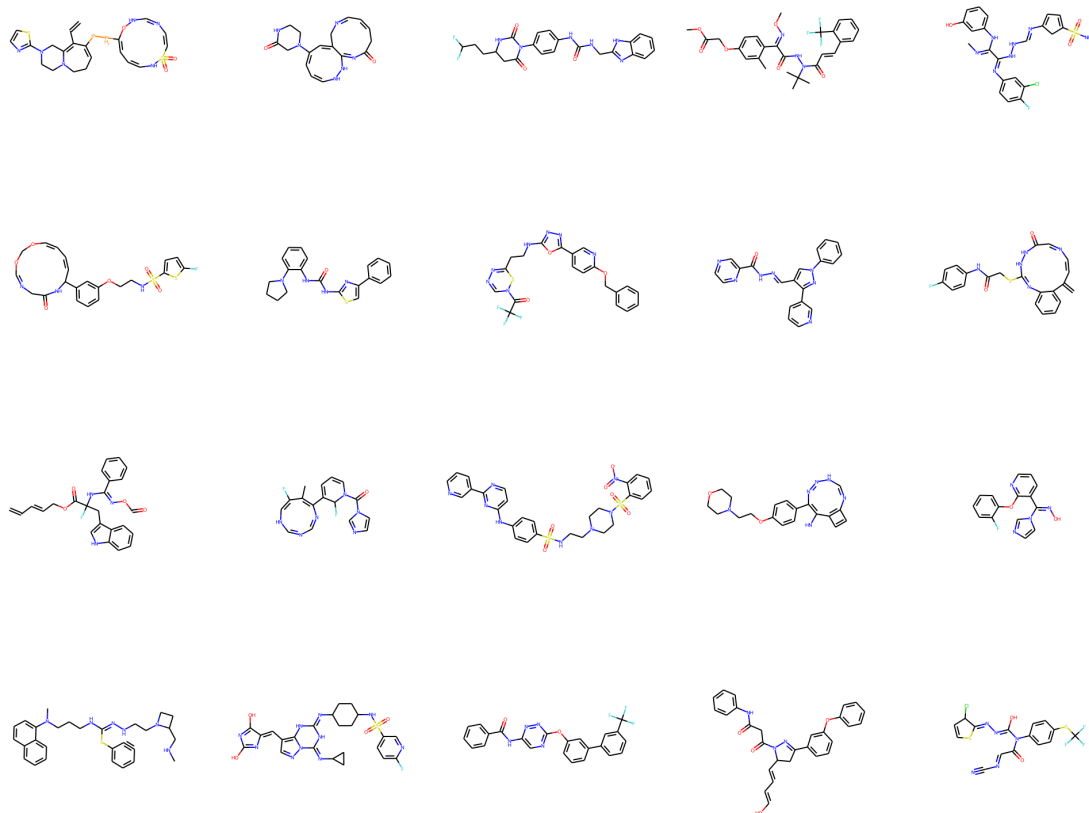


Figure S2: Example molecules generated using reinforcement learning without utilizing the drug-likeness metric as additional reward. Many of these molecules are not drug-like, i.e. having large rings, or having a high proportion of hetero atoms.

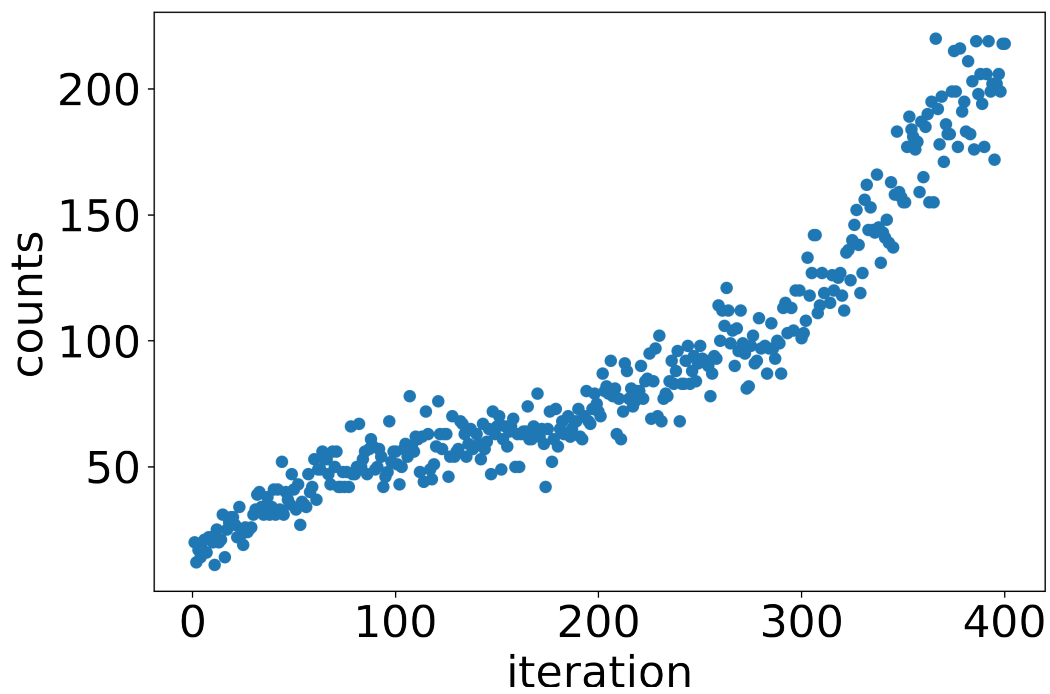


Figure S3: Number of molecules selected into the vina-selected set from each RL training iteration

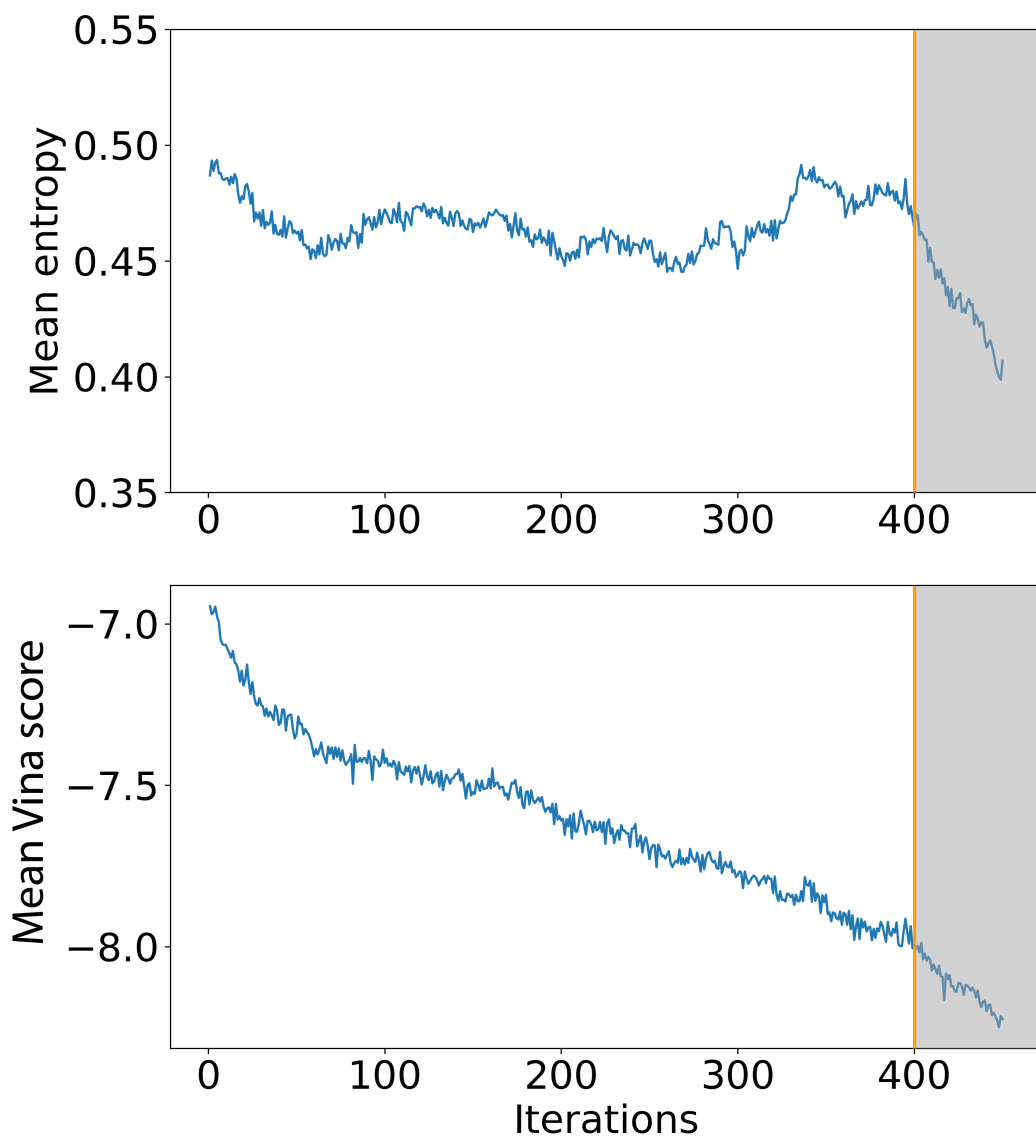


Figure S4: Change of mean entropy of model-predicted token probabilities and mean Vina scores of generated molecules during the training process. The model after RL is the model at iteration 400, and any molecules generated after iteration 400 are not considered for subsequent analysis.