THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

# 3D-Scaffold: A Deep Learning Framework to Generate 3D Coordinates of Drug-like Molecules with Desired Scaffolds

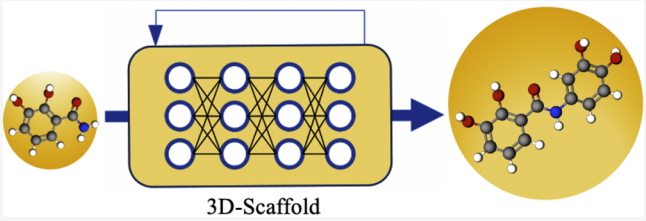
Rajendra P. Joshi, Niklas W. A. Gebauer, Mridula Bontha, Mercedeh Khazaieli, Rhema M. James, James B. Brown, and Neeraj Kumar*

Cite This: *J. Phys. Chem. B* 2021, 125, 12166–12176

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** The prerequisite of therapeutic drug design and discovery is to identify novel molecules and developing lead candidates with desired biophysical and biochemical properties. Deep generative models have demonstrated their ability to find such molecules by exploring a huge chemical space efficiently. An effective way to generate new molecules with desired target properties is by constraining the critical fucntional groups or the core scaffolds in the generation process. To this end, we developed a domain aware generative framework called 3D-Scaffold that takes 3D coordinates of the desired scaffold as an input and generates 3D coordinates of novel therapeutic candidates as an output while always preserving the desired scaffolds in generated structures. We demonstrated that our framework generates predominantly valid, unique, novel, and experimentally synthesizable molecules that have drug-like properties similar to the molecules in the training set. Using domain specific data sets, we generate covalent and noncovalent antiviral inhibitors targeting viral proteins. To measure the success of our framework in generating therapeutic candidates, generated structures were subjected to high throughput virtual screening via docking simulations, which shows favorable interaction against SARS-CoV-2 main protease (Mpro) and nonstructural protein endoribonuclease (NSP15) targets. Most importantly, our deep learning model performs well with relatively small 3D structural training data and quickly learns to generalize to new scaffolds, highlighting its potential application to other domains for generating target specific candidates.




## INTRODUCTION

The discovery and development of a new therapeutic is a long and expensive process with a high degree of uncertainty that sometime takes many years before clinical approval.[1,2] The ongoing novel coronavirus pandemic (COVID-19), caused by SARS-CoV-2, has highlighted the need for novel therapeutics to counter the threat of emerging viral pathogens.[3] One of the challenges in the very early stage of drug design and discovery is to find novel hits with desired functionalities.[4] This is a daunting task with conventional methods, which has slowed the discovery of high impact candidates for diverse applications.[5] Recently, with the rise of deep learning models, several approaches to efficiently explore the astronomically large chemical space of drug-like molecules have been proposed. The majority of existing approaches focus mainly on *de novo* drug design using variational autoencoders, generative adversarial networks, or reinforcement learning to generate molecules mainly in the form of SMILES strings.[6−18]

An alternate and robust way to find compounds of interest is by generating molecules with desired functional groups, core structures, or scaffolds.[19,20] Such scaffolds play an important role in fine-tuning the properties of a generated molecule by reducing the vast chemical space for exploration to local space of interest thus generating targeted molecules. Moreover, the scaffolds can be selected in a way that they influence crucial interactions with a given protein target. Thus, scaffold-based approaches allow the incorporation of such prior knowledge and rule-based learning during the generation process in order to increase the likelihood of identifying molecules with desired properties as compared to simply generating molecules from scratch. Several approaches have been proposed recently to generate compounds of interest built on a core structure.[20−23] Some of these methods are constrained to certain definitions of scaffolds (e.g., Murcko[24] scaffolds) or do not guarantee that the desired scaffold is always preserved during molecule generation, while others do not generalize well for new scaffolds.[21,23] To the best of our knowledge, none of the existing approaches focus on generating 3D coordinates of therapeutic molecules that can be directly tested against the protein target via computational and experimental screening. However, 3D coordinates of generated molecules are required

Received: July 19, 2021
Revised: August 30, 2021
Published: October 18, 2021







https://doi.org/10.1021/acs.jpcb.1c06437
*J. Phys. Chem. B* 2021, 125, 12166−12176

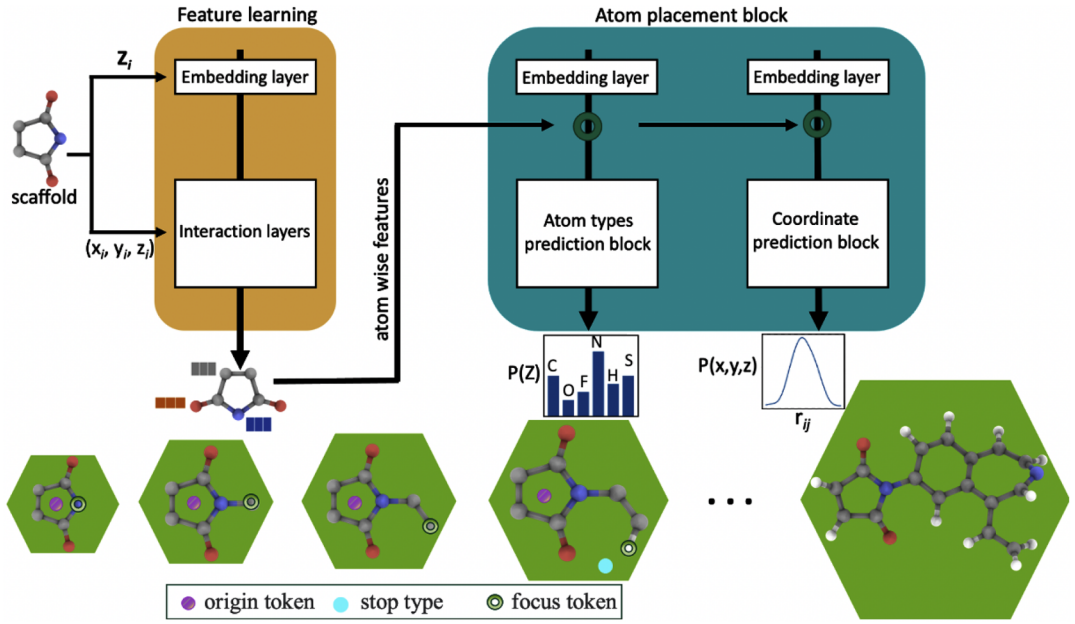



**Figure 1.** 3D-Scaffold framework used as generative model to produce therapeutic molecules with desired functionality. The bottom panel shows the scaffold-based molecular generation scheme, where the origin token, focus token, and stop type aid the generation of the molecules from scaffolds. Our framework generates only atoms in 3D-space which are connected with bonds in bottom panel for visual clarity.

for physics-based simulations as well as for robust graph-based predictive models for estimating drug-like properties. These candidates can be directly used for high-throughput virtual screening through structure-based docking to determine their affinity, activity, and efficacy against a particular disease. It is imperative to have a generative model that quickly generates 3D coordinates of effective therapeutic candidates from the massive drug-like chemical space. This will accelerate hit identification and lead optimization in drug discovery and development.[1,2]

In this work, we propose a deep learning framework called 3D-Scaffold that can generate 3D coordinates of therapeutic candidates given a desired scaffold. It is guaranteed that 100% of the generated molecules contain the desired scaffold. Moreover, our model generalizes well to previously unknown scaffolds that are not included in the training data. Our current framework is different from existing scaffold-based approaches for multiple reasons: (I) In contrast to existing approaches, which generate SMILES strings or molecular graphs, our model generates 3D coordinates of the candidates with a given core structure; (II) It works equally well for all possible scaffold definitions including cyclic skeletons, Bemis−Murcko, or side chains based on SMILES strings; (III) Our model is transferable to generate molecules with new scaffolds; (IV) Without explicitly constraining the model to desired properties, generated molecules show properties similar to the training set.

A few issues arise when constructing physics informed machine learning approaches based on 3D nuclear coordinates in contrast to more abstract molecular representations such as SMILES strings or molecular graphs.[25] The coordinate representation is not invariant to rotation, translation, and indexing of atoms, while most properties of interest (e.g., the potential energy or the log*P* score) are invariant to these transformations or change equivariantly. For instance atomic forces rotate and translate with the coordinates. Our 3D-Scaffold framework systematically obeys these constraints by building on the G-SchNet[26,27] architecture. It allows our model to extract features from the coordinates that capture local symmetries and are invariant to rotation, translation, and indexing of the input coordinates. The distributions it predicts for atom positions equivariantly rotate and translate with respect to the coordinates. Most importantly, we show that our framework designs reasonable molecules even with small training data sets due to the robust architecture of the underlying model. By training it on limited, yet known therapeutic candidates, we aim to generate more and previously unseen novel molecules with desired scaffolds that can be synthesized, which ultimately will contribute toward accelerating the discovery of therapeutic drugs.

In this contribution, we applied our 3D-Scaffold for *de novo* discovery of molecules specifically tailored to bind with given SARS-CoV-2 protein targets. Our methodology is exemplified by the task of designing antiviral candidates to target SARS-CoV-2 related proteins. Using carefully curated covalent and noncovalent antiviral data sets, we were able to constrain the generation space for domain-aware deep generative framework to generate novel covalent and noncovalent inhibitor candidates. The key properties of generated molecules are compared with the molecules in the training set. Generated 3D coordinates of molecules were further examined for their affinity as antiviral inhibitors against SARS-COV-2 main protease (Mpro) and a SARS-CoV-2 nonstructural protein endoribonuclease (NSP15).

## METHODS

**3D-Scaffold Framework.** To build novel therapeutic candidates with key functionalities critical for drug design and development, we developed the 3D-Scaffold framework. It is built on a deep neural network named G-SchNet,[26,27] which generates molecular structures from scratch by iteratively placing one atom after another in 3D space. In 3D-Scaffold, instead of starting from scratch, molecules are built around a desired scaffold.


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

**Table 1. Pseudo Code for Training and Generation Phases in the 3D-Scaffold Framework**

| | |
|---|---|
| training phase | |
| Input: $\mathbf{M}$, $\mathbf{I}_{scaff}$ | ▷training molecule, indices of the atoms in the desired scaffold |
| origin ← get_center_of_mass($\mathbf{M}$, $\mathbf{I}_{scaff}$) | ▷set position of origin token to center of mass of atoms in the scaffold |
| $\mathbf{M}_{part}$ ←{origin, focus} | ▷initialize partial molecule with the two auxiliary tokens |
| $\mathbf{A}$ ← {origin} | ▷initialize set of available atoms with origin token |
| **while** $\mathbf{A} \neq \{\phi\}$ | ▷while set of available atoms is not empty, i.e. not all atoms marked as finished |
| focus ← random($\mathbf{A}$) | ▷randomly select any atom available as focus |
| neighbors ← get_unplaced_neighbors(focus, $\mathbf{M}$, $\mathbf{M}_{part}$) | ▷get all neighbors of focus not in $\mathbf{M}_{part}$ |
| **if** neighbors = $\{\phi\}$ **then** | ▷no neighbors left for the current focus |
| next_atom ← stop | ▷predict stop type to mark current focus as finished |
| $\mathbf{A} \leftarrow \mathbf{A} \setminus$ {focus} | ▷remove focus from the set of available atoms, i.e., mark it as finished |
| **else** | |
| next_atom ← get_closest_atom(origin, neighbors) | ▷find atom in neighbors closest to origin |
| $\mathbf{A} \leftarrow \mathbf{A} \cup$ {next_atom} | ▷add next atom to set of available atoms |
| model.predict_and_backprop($\mathbf{M}_{part}$, next_atom) | ▷predict distributions for type and distances and update model weights |
| **if** next_atom ≠ stop **then** | ▷if the next atom is not the stop type |
| $\mathbf{M}_{part} \leftarrow \mathbf{M}_{part} \cup$ {next_atom} | ▷add next atom to the partial molecule |
| **if** focus = origin **then** | ▷in the very first step (focus is on the origin) |
| $\mathbf{A} \leftarrow \mathbf{A} \setminus$ {origin} | ▷remove origin from the set of available atoms to only focus proper atoms afterward |
| generation phase | |
| Input: model, max_atoms, $\mathbf{A}_{scaff}$ | ▷trained model, maximum number of atoms, atoms in the scaffold |
| origin ← get_center_of_mass($\mathbf{A}_{scaff}$) | ▷set position of origin token to center of mass of atoms in the scaffold |
| $\mathbf{M} \leftarrow$ {origin, focus, $\mathbf{A}_{scaff}$} | ▷initialize molecule with auxiliary tokens and the atoms in the scaffold |
| $\mathbf{A} \leftarrow \{\mathbf{A}_{scaff}\}$ | ▷initialize set of available atoms with atoms in the scaffold |
| t ← 2 | ▷number of tokens (origin and focus) |
| N ← $\lvert\mathbf{A}_{scaff}\rvert$ | ▷number of atoms in the scaffold |
| **for** $i = t + N+1$ **to** $t$ + max_atoms **do** | ▷atom placement loop |
| **while** $\mathbf{A} \neq \{\phi\}$ | ▷type prediction loop |
| focus ← random($\mathbf{A}$) | ▷randomly select an atom to be focused from set of available atoms |
| next_type ← sample(model.predict_type($\mathbf{M}$)) | ▷predict and sample from distribution over type of the next atom |
| **if** next_type = stop **then** | ▷if stop type was sampled |
| $\mathbf{A} \leftarrow \mathbf{A} \setminus$ {focus} | ▷remove current focus from A and repeat type prediction loop |
| **else** | ▷if a proper atom type was sampled |
| **break** | ▷proceed to the actual atom placement |
| **if** $\mathbf{A} = \{\phi\}$ **then** | ▷no atoms in set of available atoms, i.e., all are marked as finished |
| **return** $\mathbf{M} \setminus$ {origin, focus} | ▷return the finished molecule without auxiliary tokens |
| $p(d_{ij})$ = model.predict_dists($\mathbf{M}$, next_type) $\forall j < i$ | ▷predict distributions over pairwise distances $d_{ij}$ to preceding atoms |
| $p(\mathbf{r}_i = \mathbf{r}) = \frac{1}{\alpha}\prod_{j=1}^{i-1} p(d_{ij} = \lVert\mathbf{r} - \mathbf{r}_j\rVert_2)$ | ▷compute probabilities of grid positions $\mathbf{r}$ from distance probabilities |
| next_position ← sample($p(\mathbf{r}_i)$) | ▷sample position of next atom from computed 3d grid distribution |
| $\mathbf{M} \leftarrow \mathbf{M} \cup${(next_type, next_position)} | ▷Add sampled atom to molecule |
| $\mathbf{A} \leftarrow \mathbf{A} \cup${(next_type, next_position)} | ▷Add sampled atom to set of available atoms |
| **del** $\mathbf{M}$ | ▷max_atoms atoms are placed but not all of them marked as finished, thus discard the molecule |

From a computational perspective, the neural network used in our 3D-Scaffold framework for *de novo* therapeutic candidate design can be broken into two major blocks—feature learning and atom placement as shown in Figure 1. In the feature learning block, the embedding and interaction layers of SchNet[28−31] are used to extract and update rotationally and translationally invariant atom-wise features that capture the chemical environment of an unfinished molecule. Here, the neural network utilizes continuous-filter convolution layers as a means to learn robust representations of molecules starting only from the positions of atoms and corresponding nuclear charges. In the atom placement block, the extracted features are used to predict distributions for the type of next atom and its 3D coordinates, where the latter distribution is constructed from predictions of pairwise distances between the next atom and all preceding atoms. In order to do the actual placement of the next atom in 3D space, a distribution on a small grid with candidate positions focused on one of the preceding atoms is constructed from the predicted pairwise distances. The whole procedure is repeated successively to build a complete molecule with the desired scaffold. After the type and position of the next atom has been sampled from the predicted distributions, new atom-wise features incorporating the added atom are extracted in the feature learning block and then used to place the following atom in the atom placement block.

The generation process is aided by two auxiliary tokens with unique, artificial types, namely the origin and focus tokens. At each generation step, one of the already placed atoms is uniformly randomly chosen as the focus token. The origin token, in contrast, stays fixed throughout the entire generation procedure. In previous work with G-SchNet by Gebauer et al.,[26] the origin token marks the center of mass of the molecule.

In our 3D-Scaffold framework, however, we instead use it to mark the center of mass of the scaffold that is the starting point



https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

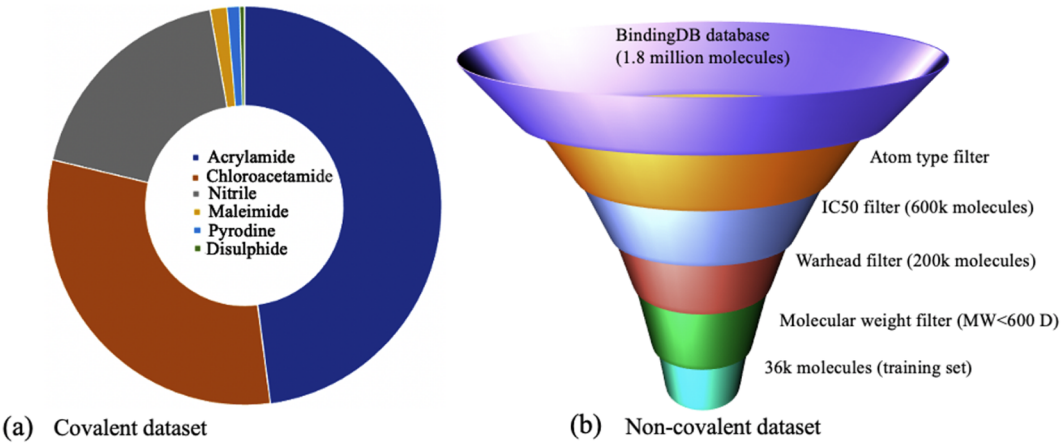



**Figure 2.** (a) Distribution of covalent data set based on scaffolds. (b) Filtering criteron used to generate noncovalent training data set.

of the generation procedure. At each step, the unplaced neighbor of the focus token that is closest to the origin token is supposed to be sampled. This means that while the structure grows around the center of mass of the resulting molecule in the previous G-SchNet model, in our current 3D-Scaffold framework it grows from the center of mass of the desired scaffold given to the model as a starting point. If the currently focused atom has no neighbors left to place, the model should predict the stop type instead of a proper atom type and in this way mark the focused atom as finished. Atoms marked as finished cannot be chosen as focus anymore, and after all atoms have been marked as finished, the generation process terminates. The resulting schemes for training of the model and generation of molecules are summarized as pseudo code in Table 1.

The model is trained end-to-end with backpropagation using the ground truth types and pairwise distances of atoms in training data molecules split into sequential atom placement steps as described in the pseudo code. At each training step, the model predicts the type of the next atom and its distances to all preceding atoms. The distributions predicted by the model are discrete: the type distribution contains a probability value for each atom type occurring in the training data set and the stop type and the distance distributions cover distances between 0 and 15 in 300 equally spaced bins. At any step, let $Z_{\text{next}}$ be the ground truth type of the next atom and $\hat{p}_{\text{type}}^{Z_{\text{next}}}$ the probability that the model assigns to that type at the current step. Then, we use negative log-likelihood as the loss for the type prediction:

$$l^{\text{type}} = -\log(\hat{p}_{\text{type}}^{Z_{\text{next}}}) \tag{1}$$

For the loss on distance predictions, we use the cross-entropy between true and predicted distances

$$l^{\text{dists}} = \sum_{j=1}^{N} \sum_{b \in B} q_j^b \log(\hat{p}_j^b) \tag{2}$$

with Gaussian expanded ground truth distances

$$q_j^b = \frac{e^{-\gamma(\|\mathbf{r}_{\text{next}} - \mathbf{r}_j\|_2 - b)^2}}{\sum_{b' \in B} e^{-\gamma(\|\mathbf{r}_{\text{next}} - \mathbf{r}_j\|_2 - b')^2}} \tag{3}$$

Here $\mathbf{r}_{\text{next}}$ is the ground truth position of the next atom, $\mathbf{r}_j$ is the position of an already placed atom, $N$ is the number of preceding atoms, $\gamma$ determines the width of the expansions, $B$ are the 300 binned distances between 0 and 15, and $\hat{p}_j^b$ is the probability that the model assigns for the distance between $\mathbf{r}_j$ and $\mathbf{r}_{\text{next}}$ to fall into distance bin $b \in B$ at the current step. In steps where the ground truth type is the stop type, the loss on distance predictions is set to zero as no distances are predicted. Descriptions about the hyperparameters used in this work is provided in the Supporting Information.

**Training Data.** Therapeutic candidates interact with target proteins either by forming a covalent bond or noncovalently through nonbonding interactions. Depending on the kind of interaction, the molecule is identified as either a covalent or noncovalent drug candidate.[32] The focus of our study is to develop a general framework capable of producing both covalent and noncovalent novel therapeutic candidates with specific scaffolds, and so we performed experiments on two different data sets.

First, we performed experiments on covalent inhibitor data (hereafter called covalent data set) taken from multiple sources.[33,34] For the covalent data set, we used ∼4000 candidates from a database of FDA approved drugs[33] and cysteine molecules from the enamine database[34] with six different scaffolds namely acrylamides, chloroamides, nitriles, disulfides, maleimides, and pyrodines.[33] These functional groups react with the cysteine residue of the target protein by forming covalent bonds. The distribution of each scaffold in the data set is provided in the pie chart in Figure 2. Nearly 95% of the training set is dominated by three scaffolds. We later show that, irrespective of the fraction of data for each scaffold, our model generalize equally well for all of them. SMILES strings of the molecules are extracted from the respective databases. RDkit[35] with MMFF94[36] force field was used to convert SMILES into the 3D coordinates required as an input for our model.

In addition, for noncovalent inhibitor design, we curated and filtered a large data set of synthesizable molecules from BindingDB[37] to create the noncovalent data set. We used different filtering criteria as shown in Figure 2 for creating the data set. Our noncovalent inhibitor design model is trained with 36k molecules consisting of 10k unique scaffolds. For the noncovalent data set, we use Murcko scaffolds[24] as a definition of scaffolds, which demonstrates the flexibility of our model not only in allowing different scaffold definitions, but also for generating noncovalent inhibitors. We used RDkit to obtain Murcko scaffolds from SMILES strings of molecules in the


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

training set. For generation with this data set, we randomly select 25 out of the 10k scaffolds and generate 1000 molecules for each of them, providing ample generated molecules to assess the performance of the model.

## RESULTS AND DISCUSSION

Accurate prediction of advanced hit candidates and fragment specific lead optimization is crucial for the design and development of novel therapeutics to combat the threat posed by new and emerging viruses.[32,38] There continues to be significant need for the development of small-molecule inhibitors that directly target viral proteins to complement existing therapeutics, not only for SARS-CoV-2, but also for related $\beta$-coronaviruses SARS-CoV and MERS-CoV, which have high mortality rates.[39] To this end, our 3D-Scaffold framework represents a unique opportunity for target-specific noncovalent and covalent drug development and efficient integration of warhead-optimization. We exemplify the application of our scaffold informed ML framework to generate covalent electrophiles and noncovalent inhibitor candidates against the SARS-CoV-2 main protease (Mpro) and SARS-CoV-2 nonstructural protein endoribonuclease (NSP15), essential for viral replication.

**Covalent Antiviral Inhibitor Design for Mpro.** Targeted covalent inhibitors represent a viable strategy to inhibit the main proteases involved in different disease pathologies including SARS-CoV-2.[32] Using the covalent antiviral data set, we first trained the model to generate molecules with six different scaffolds that are common electrophilic warheads for different drug applications. For each of the scaffolds, we generated 2000 molecules and inspected them for their validity, uniqueness, and novelty. To calculate the percentage of valid, unique, and novel molecules, we use

$$\text{validity} = \frac{\text{number of valid molecules}}{\text{number of generated molecules}}$$

$$\text{unique} = \frac{\text{number of unique molecules}}{\text{number of valid molecules}}$$

$$\text{novelty} = \frac{\text{number of generated molecules not in training set}}{\text{number of unique and valid generated molecules}}$$

The validity of generated molecules is examined by converting generated 3D coordinates into canonical SMILES strings using the *xyz2 mol* script from the Jensen group,[40,41] which relies on Rdkit.[35] The conversion can also be accomplished using Rdkit alone or other open source tools like Open Babel, but these tools are less reliable when determining bond orders during conversion. We then used the sanitize functionality of Rdkit to examine the validity of the obtained SMILES strings. Alternatively, the validity of generated molecules can be measured by performing physics-based simulations such as density functional theory. Due to the enormous computational cost required to perform such calculations on thousands of generated molecules, we resort to empirical approaches. To examine the novelty of the generated molecules, we compare the Rdkit topological fingerprint similarity of the molecules in the training set and the generated set. The uniqueness metric is determined similarly by using molecular fingerprints. In addition, to further validate the performance of our model in generating valid and synthesizable molecules, we also query the MCULE database[42] for generated molecules to check how many already exist in the MCULE data set. The performance of our model in terms of these metrics is listed in Table 2.

**Table 2. Table Showing the Statistics of Valid, Unique, and Novel Molecules Generated for Different Scaffolds[a]**

| scaffolds/methods | validity (%) | uniqueness (%) | novelty (%) | known |
|---|---|---|---|---|
| covalent data set | | | | |
| acrylamides | 79 | 96 | 99 | 59 |
| chloroamides | 83 | 93 | 99 | 34 |
| pyrodines | 84 | 83 | 100 | 71 |
| maleamides | 86 | 85 | 99 | 73 |
| nitriles | 81 | 97 | 100 | 59 |
| disulfides | 75 | 98 | 100 | 1 |
| piperazine[b] | 80 | 92 | 100 | 52 |
| noncovalent data set | | | | |
| | 90 | 73 | 100 | |
| literature | | | | |
| G-SchNet[26] | 77 | 92 | 88 | – |
| Lim et al.[21] | 99 | 85 | 99 | – |
| DeepScaffold[23] | 99 | 69 | – | – |
| GraphVAE[43] | 56 | 76 | 62 | – |
| MolGAN[43] | 98 | 10 | 94 | – |

[a]The number of generated molecules that exist in MCULE database (not in the training set) is also listed in the "known" column. For the model trained on the non-covalent dataset, mean values of validity, uniqueness, and novelty for 25 different scaffolds is provided. For comparison, performances of recent methods from the literature are also provided. However, note that literature results stem from experiments with different datasets than the ones used in this work. [b]Novel scaffold.

The performance of our model is similar to existing scaffold-based generative models in terms of generating valid, unique, and novel molecules. For all the scaffolds in the covalent data set, our model performs similarly well, with on average 92% uniqueness among the generated molecules. 81% of generated molecules are valid and ∼100% are novel. These metrics remain similar even for the molecules generated using a novel scaffold (piperazine) as starting point, thus demonstrating the transferability of our model to scaffolds not in the training set. Compared to the existing generative models in the literature, our model shows superior performance in generating unique and novel molecules, while the percentage of valid molecules generated is in general slightly lower than for other generative models. We, however, note that these models were trained on different data sets, making a direct comparison of the reported numbers difficult. Moreover, the performance of our model is especially promising when one takes into account the relatively small amount of training data used (4000) compared to cited models from the literature which were trained on larger training sets. Training our model on larger training sets might further improve the reported statistics as has been reported for other generative models.[44] In addition, compared to ours, models from the literature were trained to generate relatively small molecules with the QM9 data set. Size of molecules generated from our model varies from the size of scaffolds to "*N*" number of atoms provided by user as input. When querying the MCULE database, we found that some of the molecules generated for each scaffold are already known and available in the database, demonstrating the success of our



https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

Table 3. Statistics of Molecules from the Training and Generated Data Set, Respectively, for Each Scaffold[a]

| | | training set | | | generated set | | |
|---|---|---|---|---|---|---|---|
| | | SA | logP | QED | SA | logP | QED |
| chloroamides | mean | 2.55 | 2.30 | 0.84 | 4.59 | 1.73 | 0.65 |
| | std | 0.57 | 1.00 | 0.15 | 1.31 | 1.97 | 0.23 |
| acrylamides | mean | 2.65 | 2.40 | 0.76 | 4.26 | 2.00 | 0.60 |
| | std | 0.55 | 1.33 | 0.21 | 1.42 | 1.99 | 0.25 |
| disulfides | mean | 2.94 | 2.64 | 0.88 | 5.60 | 3.15 | 0.52 |
| | std | 0.89 | 0.82 | 0.22 | 0.95 | 2.15 | 0.26 |
| pyrodines | mean | 2.90 | 1.62 | 0.70 | 4.62 | 0.80 | 0.52 |
| | std | 0.59 | 1.47 | 0.25 | 1.55 | 1.71 | 0.28 |
| maleamides | mean | 2.30 | 1.32 | 0.66 | 4.36 | 0.72 | 0.63 |
| | std | 0.27 | 1.02 | 0.17 | 1.40 | 1.62 | 0.24 |
| nitriles | mean | 2.40 | 3.12 | 0.87 | 4.47 | 2.10 | 0.61 |
| | std | 0.40 | 1.00 | 0.13 | 1.28 | 1.84 | 0.23 |
| piperazine[b] | mean | – | – | – | 4.75 | 1.09 | 0.54 |
| | std | – | – | – | 1.31 | 1.86 | 0.29 |

[a]The mean and standard deviation for each property in each set are provided. [b]Novel scaffold.

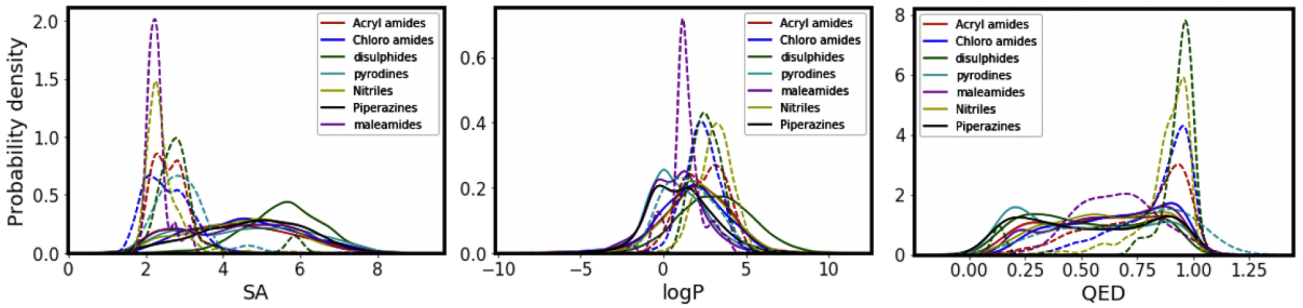



Figure 3. Probability density plots of SA score, logP, and QED for molecules in the training set as well as generated set for each functional group. Solid lines correspond to metrics of data in the generated set, whereas dashed lines of same color correspond to molecules in the training set.

model in generating synthesizable molecules. This also holds for the molecules generated with the novel scaffold piperazine.

An important goal of our work is to generate novel molecules with therapeutic properties while retaining desired scaffolds. To this end, we do not directly condition molecule generation on the desired properties but instead constrain it to the generation of molecules with desired scaffolds. We expect that this will indirectly constrain the properties, as well. The properties of interest are synthetic accessibility (SA) score, quantitative estimation of drug-likeness (QED), and the partition coefficient (logP). The SA score measures the synthesizability of generated molecules and has values in the range 0−10, where the lower end suggests increased accessibility. QED is a useful measure for quantifying and ranking the drug-likeness of a compound. The values range from 0 for unfavorable to 1 for favorable molecules. The partition coefficient, logP, estimates the lipophilicity or hydrophilicity of a compound. It measures the physical nature of a compound and its permeability and ability to reach the target in the body. A positive logP value indicates the compound is lipophilic, and a negative logP value indicates a hydrophilic compound.

We compare the properties of the generated molecules with the ones in the training set to see whether our model can generate new molecules with properties similar to those of the molecules in the training set. Ideally, having similar statistics of properties is an indicator that our model is performing as expected with the constraints imposed upon it. For the statistical analysis, we report the mean and standard deviation of the SA, logP, and QED scores in both the training and the generated sets in Table 3. The mean SA score of both generated and training set molecules falls in the lower half of the SA scale 0−10, implying in general synthesizability of generated molecules. Slight deviation observed between the two sets can be attributed to the lack of explicit conditioning on target properties.[45] The mean value of QED for generated molecules is slightly lower (on average by 0.2 units) compared to molecules from the training set. However, the model also generated molecules with high QED, i.e. strong drug-likeness. Here, logP follows similar trends for its mean value among two sets. We consistently observed relatively large standard deviation for SA, QED, and logP in generated molecules for each scaffolds, reflecting diversity in generated molecules compared to the well curated training data set. To further visualize this data, we display the probability density plots for SA, QED, and logP of the molecules in the training set and the generated set for each scaffolds in Figure 3. Solid lines mark the distributions of generated molecules while dashed lines correspond to molecules in the training set. The distributions of generated molecules with respect to the SA score in Figure 3a show that a good fraction of generated molecules are experimentally synthesizable. Moreover, the distribution of the SA score for the novel functional group, piperazine (not in training set), is similar to other scaffolds in the training set, showing the transferability of our model. This also demonstrate the success of our model in generating


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

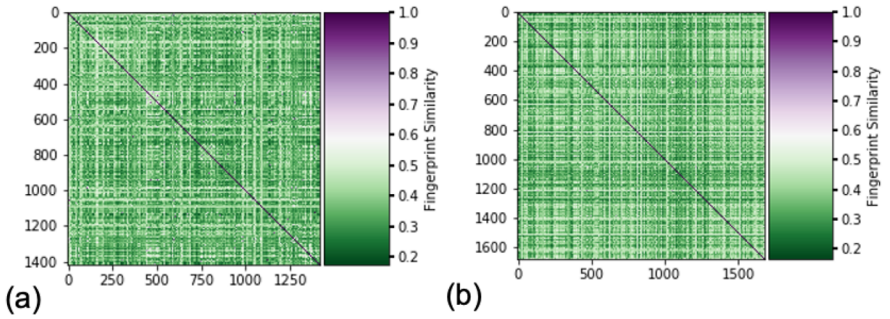



**Figure 4.** Heatmap showing the fingerprint similarity between molecules in the training set (a) and the generated set (b) for the acrylamide scaffold.

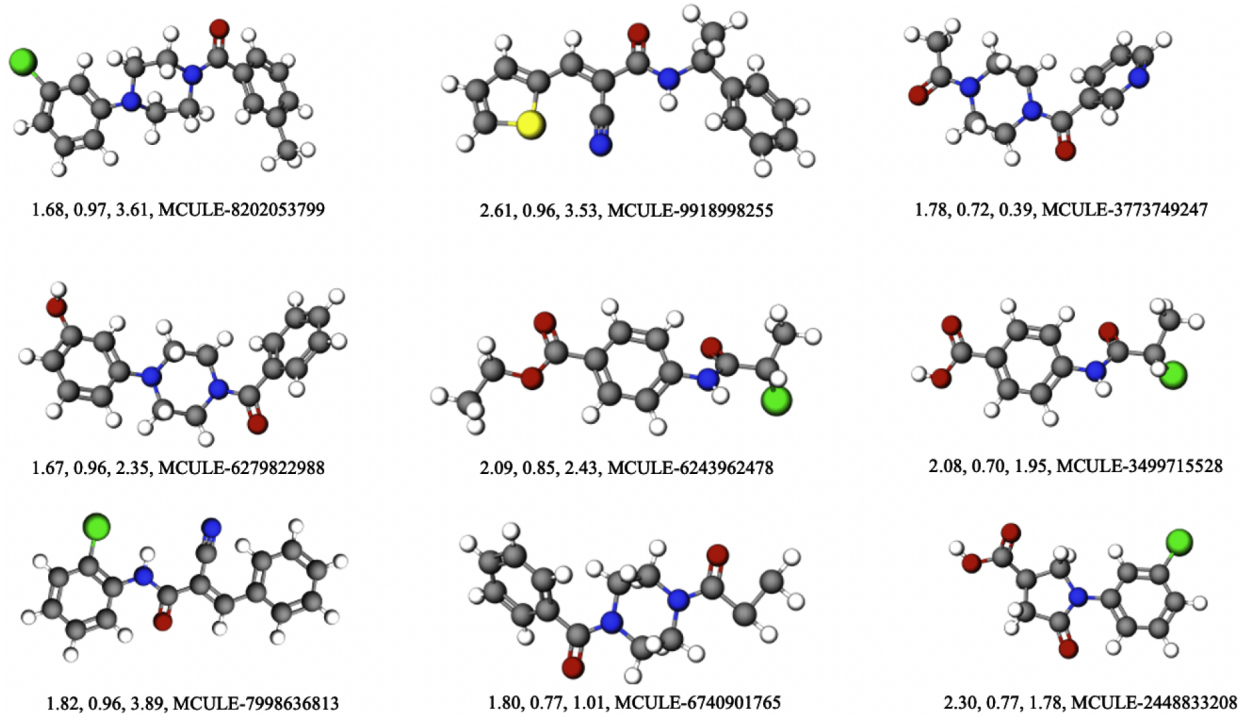



**Figure 5.** Sample of generated candidates along with their SA score, QED, log*P* values, and corresponding MCULE ids. These candidates are synthesizable and available to order from MCULE database.

experimentally synthesizable molecules, which is a big issue with most generative models. For the log*P* metric (see Figure 3b), similar distributions are observed between generated and training molecules. We were able to generate both lipophilic as well as hydrophilic compounds as indicated by positive and negative log*P*, respectively, with the former category being the majority, similar to the molecules in the training set. This again indicates that our model is generating novel molecules with properties similar to the training set. From the QED distribution plot (Figure 3c), we see that the majority (∼60%) of the molecules have a QED score greater than 0.5, with a good chunk of molecules being close to 1 as evident from the peaks of probability distribution curves around 0.9. Minor discrepancies between the properties of generated molecules and the training set may be due to the lack of directly constrained property optimization in our work. Although our model generates molecules with desired properties, it would be interesting to see its performance when explicitly constraining the desired property range. However, this is beyond the scope of our current work and is kept aside for future work.

We further analyzed the diversity of molecules using heatmaps of the Tanimoto coefficient between molecules within the training set (Figure 4a) and within the generated set (Figure 4b). The Tanimoto coefficient is a measure of the similarity of molecules. The heatmap shows that the training set we use is quite diverse as evident by the many green spots (low similarity). A similar heatmap is observed for the generated set, showing that generated molecules are quite different from each other, while predominantly maintaining similar properties (as discussed before). We also note that our model generates diverse molecules in terms of their size, i.e., the number of atoms, while always preserving the given scaffolds.

To check the transferability of our model to generate valid molecules for functional groups that are not in training set, we generated 2000 molecules with piperazine as the starting building block. Generated molecules are checked against the MCULE databases to see if any of the generated molecules are already known. We found that nearly 50 of the molecules generated are available in the MCULE database. This shows


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

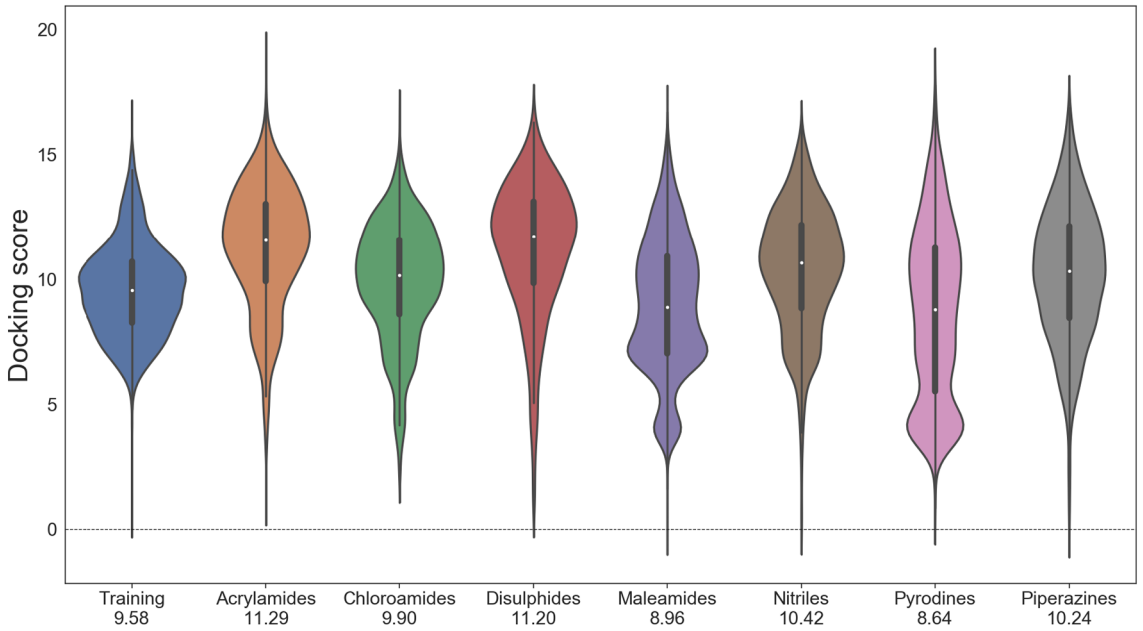



**Figure 6.** Violin plots showing the distribution of the docking score against the MPro protein for generated molecules with different scaffolds and training molecules in the covalent data set. Larger values imply favorable binding.

the capacity of our model to generate valid, synthesizable molecules even for novel scaffolds. The distribution plot for the SA, log*P*, and QED of the molecules generated for piperazine is included in Figure 3. It shows that the properties of molecules generated follow similar distributions as for other functional groups.

We visualize representative generated molecules that we also found in the MCULE database with corresponding SA score, QED, and log *P* values along with the corresponding MCULE ids in Figure 5. Overall, our results show that our model constrained to generating molecules with desired scaffolds indirectly also successfully constrains the properties. Despite the significant variation in the amount of training data for each scaffold, our model consistently generates valid, unique, novel, and experimentally synthesizable molecules with desired therapeutic properties for each scaffold within and outside of the training set.

**Binding Affinities of Covalent Inhibitors against Mpro.** As a proof of concept application for generated molecules, we docked them against main protease (Mpro). Mpro is the key enzyme of SARS-CoV-2 that gets the maximum attention because of its ability to trigger viral replication and transcription.[46,47] Significant effort has been made since the rapid increase of SARS-CoV-2 worldwide to find therapeutic small molecules and vaccines that have desired activity. Most of the early efforts were focused on drug repurposing using already known drug molecules. For future pandemic events, it is possible that an effective drug molecule for repurposing is not yet known. In those scenarios, models that generate novel molecules with certain functionalities as proposed here will be an efficient alternative. First step for such generated molecules is to examine their binding affinity and activity against the target protein using docking simulations. Docking simulations use empirical approaches to determine the favorable/unfavorable binding of ligands with target proteins and numerically rank them using a docking score.

For our molecular docking simulations we utilized the AutoDock for Flexible Receptors (ADFR) package.[48] Generated ligands were covalently bound to Cysteine-145 of the target protein (PDB ID:6WQF), which is part of a catalytic dyad formed with Histidine-41. We compared the docking score of generated molecules against the training molecules in the covalent data set which is shown in Figure 6. A larger magnitude of the docking score implies higher favorability for the docking process. We found that generated molecules show similar docking performance to molecules in the training covalent data set, as illustrated in the violin plots and the corresponding mean docking score noted in the labels of the *x*-axis. For the majority of scaffolds, including the novel scaffold piperazine and the three scaffolds that make up 95% of the training data, the generated molecules on average show higher affinity for docking against the Mpro-target protein than molecules in the covalent data set. The only scaffolds that have a smaller mean docking scores compared to the training molecules are maleamide and pyrodine, with docking scores of 8.96 and 8.64, respectively.

**Noncovalent Antiviral Inhibitor Design for NSP15.** With the goal of generating noncovalent inhibitors for SARS-Cov-2 targets, we trained our model on the noncovalent data set using Murcko scaffolds. The training data consist of 36k molecules with 10k unique scaffolds. The performance of our model trained for generating noncovalent inhibitors is similar (Table 2) to the one for the covalent data set in terms of validity and novelty. However, the percentage of unique molecules generated drops to a mean value of 73% for about 25 different scaffolds. This may be a direct consequence of the limited number of molecules (on average 4) for each scaffold in the noncovalent training set. When generating 1000 molecules for each of the 25 scaffolds, the model repeats some of the generated molecules. However, the absolute amount of uniquely generated molecules per scaffold is still remarkable considering the limited number of training examples per scaffold.

As a part of the DOE National Virtual Biotechnology Laboratories (NVBL) therapeutic design project, we screened millions of compounds in repurposing libraries of drug


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

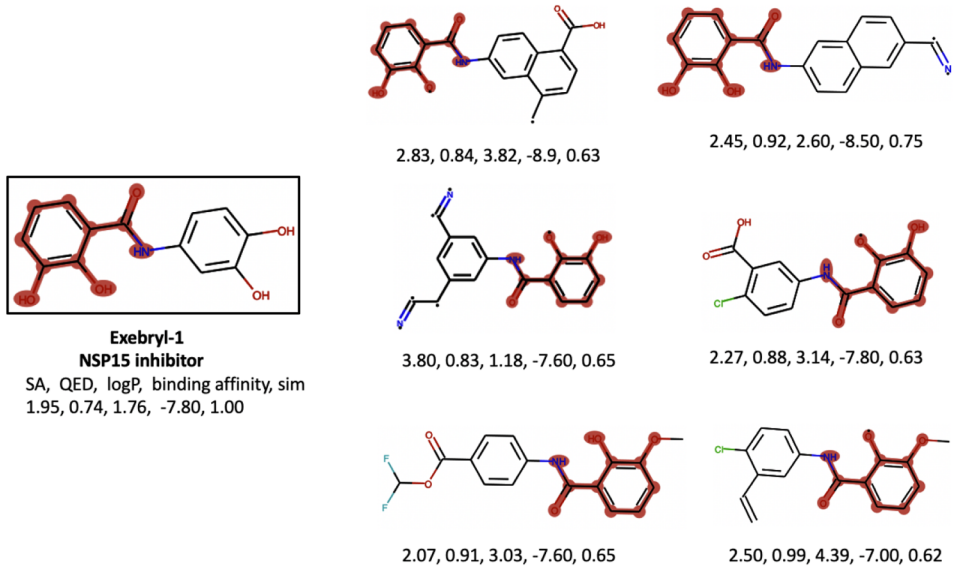



**Figure 7.** Exebryl-1 and representative generated molecules from our 3D-Scaffold framework with high binding affinity against NSP15 protein-target. For each molecule, we list the SA score, QED, log*P*, binding affinity, and fingerprint similarity (labeled sim in the figure) with respect to experimentally known NSP15 inhibitor Exebryl-1. The scaffold used for optimization is highlighted in red in generated molecules.

compounds for activity against NSP1−NSP15 from SARS-CoV-2, followed by experimental validation. In particular, the coronavirus nonstructural protein NSP15 is highly conserved among coronaviruses.[38] It is also a key component for viral replication with no corresponding counterpart in host cells, which makes it an intriguing candidate for drug development. Our recent computational and experimental results demonstrated that Exebryl-1, a $\beta$-amyloid antiaggregation molecule designed for Alzheimer's disease therapy, can bind to NSP15, but it did not have sufficient antiviral activity in cell-based assays for immediate drug repurposing efforts.[49] This provides us an interesting target to optimize the Exebryl-1 hit based on the 3D-Scaffold framework to obtain improved activity and antiviral properties. Our goal is to lead optimization together with *in silico* molecular docking calculations onto the crystal structure of NSP15.

As a test case, we generated noncovalent inhibitors for the SARS-CoV-2 nonstructural protein endoribonuclease (NSP15) target (PDB ID: 6XDH) by optimizing Exebryl-1 based compounds.[49] Exebryl-1 has experimentally been found to be active[49] against NSP15 from high-throughput assay screening from drug and lead repurposing libraries. Our goal is to modify and generate more active compounds against the NSP15 target by building molecules on top of Murcko scaffolds of the Exebryl-1 molecule. When examining the structure−activity relationship, some of the generated molecules (see Figure 7) show good binding activity (docking score) against the NSP15 target. Moreover, these molecules are easily synthesizable (low SA scores) and have desired drug-likeness (large QED values). Generated molecules from our work that showed high activity against NSP15 from docking and molecular dynamics simulations are further being investigated by our experimental collaborators.

## CONCLUSIONS

In this report, we developed a generative framework based on deep neural networks that can generate 3D coordinates of therapeutic candidates with desired scaffolds for covalent and noncovalent drug development. The 3D-Scaffold model is trained end-to-end incorporating robust atomistic representation learning techniques and generates 3D coordinates from the learned probability distributions of atom types and the pairwise distances. Due to the sequential atom-by-atom generation scheme of our framework starting from a given scaffold, the desired scaffold is 100% guaranteed in the generated 3D coordinates. We use covalent and noncovalent antiviral data sets to optimally narrow the search toward novel compounds with therapeutic significance that are reasonable to design as covalent and noncovalent inhibitors. Most importantly, our generated library of covalent warheads with the same scaffold represents a unique opportunity for the efficient integration of warhead optimization into the covalent drug development process. This ligand-based approach that targets desired scaffolds can be used for cell-based probe design.

We demonstrated that our model generates predominantly valid, unique, and novel molecules that have therapeutic drug-like properties similar to the molecules in the training set. The success of our framework lies in generating synthesizable molecules with desired properties without directly constraining on the target properties. Moreover, it performs well for relatively small volumes of training data and generalizes equally well for generating molecules with a new scaffold, which demonstrates the transferability of the proposed framework.

Our framework offers the advantage that the generated 3D coordinates of molecules can be directly used for further simulations such as density function theory, MD, or docking calculations, compared to SMILES strings or graph-based models, where empirical approaches are used to generate 3D coordinates. As an application, the 3D coordinates of generated molecules from our work were examined for their interaction against the Mpro and NSP15 targets of SARS-CoV-2 using docking simulations. Our results show that generated molecules have strong binding affinity against the target protein similar to the molecules in the training set. This holds true for novel scaffolds as well. Although we used our framework to generate covalent and noncovalent inhibitors in this work, our model in principle can be used to generates any


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176

kind of molecules with desired scaffolds, making it applicable to many domains. We believe that the robust performance of our model on relatively small data sets and its generalization on new scaffolds provides an efficient and flexible way of generating new molecules while simultaneously optimizing the functionalities by constraining the types of scaffolds included. Further improvement in the performance of the 3D-Scaffold framework may be observed by generating molecules while also explicitly constraining on the target properties or by generating molecules with more than one critical scaffold. We note that our current framework only generates single conformer of a generated molecules. This framework can be further extended to generate rationally different conformers while also considering desired interaction with the protein receptors.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c06437.

Description of hyperparameters used in this work (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Neeraj Kumar** − *Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0001-6713-2129; Email: neeraj.kumar@pnnl.gov

### Authors

**Rajendra P. Joshi** − *Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

**Niklas W. A. Gebauer** − *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany; BASLEARN − TU Berlin/BASF Joint Lab for Machine Learning, Technische Universität Berlin, 10587 Berlin, Germany; Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*

**Mridula Bontha** − *Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

**Mercedeh Khazaieli** − *Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

**Rhema M. James** − *Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

**James B. Brown** − *Environmental Genomics & Systems Biology, Lawrence Berkeley National Laboratory, Berkley, California 94710, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c06437


### Notes

The authors declare no competing financial interest.

Codes used for this work are available at https://github.com/PNNL-CompBio/3D_Scaffold.

## ■ ACKNOWLEDGMENTS

This research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. Pacific Northwest National Laboratory (PNNL) is a multiprogram national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830. Computing resources was supported by the Intramural program at the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL; grid.436923.9), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and operated under Contract No. DE-AC05-76RL01830. We thank Darin Hauner at PNNL for discussion on the covalent docking simulations. Provisional patent application of small molecule candidates designed from this work is pending (IPID 32215-E).

## ■ REFERENCES


(1) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Diversity* **2021**, *25*, 1315.

(2) Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discovery* **2018**, *17*, 97−113.

(3) Duek, I.; Fliss, D. M. The COVID-19 pandemic - from great challenge to unique opportunity: Perspective. *Ann. Med. Surg.* **2020**, *59*, 68−71.

(4) Chan, H. C. S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **2019**, *40*, 592−604.

(5) Berdigaliyev, N.; Aljofan, M. An overview of drug discovery and development. *Future Med. Chem.* **2020**, *12*, 939−947.

(6) 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 Cent. Sci.* **2018**, *4*, 268−276.

(7) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. *Proceedings of the 34th International Conference on Machine Learning*; 2017; pp 1945−1954.

(8) Guimaraes, G.; Sanchez-Lengeling, B.; Outeiral, C.; Luis, P.; Farias, C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv* 2017; 1705.10843v3.

(9) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. *arXiv* 2018; 1802.08786v1.

(10) Janz, D.; Van Der Westhuizen, J.; Paige, B.; Kusner, M. J.; Miguel Hernández-Lobato, J. Learning a Generative Model for Validity in Complex Discrete Structures. *arXiv* 2018; 1712.01664v4.

(11) Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(12) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885.

(13) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminf.* **2018**, *10*, 31.

(14) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.

(15) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111.

(16) Joshi, R. P.; Kumar, N. Artificial Intelligence based Autonomous Molecular Design for Medical Therapeutic: A Perspective. *arXiv* 2021; 2102.06045.

(17) Ward, L.; Bilbrey, J. A.; Choudhury, S.; Kumar, N.; Sivaraman, G. Benchmarking Deep Graph Generative Models for Optimizing New Drug Molecules for COVID-19. *arXiv* 2021; 2102.04977.

(18) Wu, Y.; Choma, N.; Chen, A.; Cashman, M.; T. Prates, É.; Shah, M.; Melesse Vergara, V. G.; Clyde, A.; Brettin, T. S.; de Jong, W. A. et al. Spatial Graph Attention and Curiosity-driven Policy for Antiviral Drug Discovery. Structure-Based Drug Discovery. *arXiv* 2021; 2106.02190.

(19) Zhang, K. Y. J.; Milburn, M. V.; Artis, D. R. *Structure-Based Drug Discovery*; Springer Netherlands: Dordrecht, 2007; pp 129−153.
(20) Scott, O. B.; Edith Chan, A. W. ScaffoldGraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. *Bioinformatics* **2020**, *36*, 3930−3931.
(21) Lim, J.; Hwang, S.-Y.; Moon, S.; Kim, S.; Kim, W. Y. Scaffold-based molecular design with a graph generative model. *Chem. Sci.* **2020**, *11*, 1153−1164.
(22) Arús-Pous, J.; Patronov, A.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminf.* **2020**, *12*, 38.
(23) Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. *J. Chem. Inf. Model.* **2020**, *60*, 77−91.
(24) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(25) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(26) Gebauer, N.; Gastegger, M.; Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Adv. Neural Inf. Process Syst.* **2019**, 7566−7578.
(27) Gebauer, N. W. A.; Gastegger, M.; Schütt, K. T. Generating equilibrium molecules with deep neural networks. *arXiv* 2018; 1810.11347.
(28) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.-R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.
(29) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(30) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process Syst.* **2017**, 991−1001.
(31) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(32) Tuley, A.; Fast, W. The Taxonomy of Covalent Inhibitors. *Biochemistry* **2018**, *57*, 3326−3337.
(33) *Zinc database*; http://zinc.docking.org/substances/subsets/fda/?page=1 (accessed August 30, 2020).
(34) *Cysteine focused Covalent Fragments*; https://enamine.net/fragments/covalent-fragments/cysteine-focused-covalent-fragments (accessed September 30, 2020).
(35) Landrum, G. *RDKit: Open-Source Cheminformatics Software*; http://rdkit.org/ (accessed September 30, 2020); 2016.
(36) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(37) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−53.
(38) Haque, S. M.; Ashwaq, O.; Sarief, A.; Azad John Mohamed, A. K. A comprehensive review about SARS-CoV-2. *Future Virol.* **2020**, *15*, 625−648.
(39) Ford, N.; Vitoria, M.; Rangaraj, A.; Norris, S. L; Calmy, A.; Doherty, M. Systematic review of the efficacy and safety of antiretroviral drugs against SARS, MERS or COVID-19: initial assessment. *J. Int. AIDS Soc.* **2020**, *23*, D1045−D1053.
(40) Kim, Y.; Kim, W. Y. Universal Structure Conversion Method for Organic Molecules: From Atomic Connectivity to Three-Dimensional Geometry. *Bull. Korean Chem. Soc.* **2015**, *36*, 1769−1777.
(41) *xyz2mol*; https://github.com/jensengroup/xyz2mol.
(42) Kiss, R.; Sandor, M.; Szalai, F. A. http://Mcule.com: a public web service for drug discovery. *J. Cheminf.* **2012**, *4*, P17−P17.
(43) Cao, N. D.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv* 2018; 1805.11973.
(44) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminf.* **2019**, *11*, 71.
(45) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *Npj Comput. Mater.* **2021**, *7*, 84.
(46) Clyde, A.; et al. High Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Non-Covalent Inhibitor. *bioRxiv* 2021.
(47) Ullrich, S.; Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* **2020**, *30*, 127377.
(48) Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Comput. Biol.* **2015**, *11*, e1004586.
(49) Choi, R.; Zhou, M.; Shek, R.; Wilson, J. W.; Tillery, L.; Craig, J. K.; Salukhe, I. A.; Hickson, S. E.; Kumar, N.; James, R. M.; et al. High-throughput screening of the ReFRAME, Pandemic Box, and COVID Box drug repurposing libraries against SARS-CoV-2 nsp15 endoribonuclease to identify small-molecule inhibitors of viral activity. *PLoS One* **2021**, *16*, e0250019.


https://doi.org/10.1021/acs.jpcb.1c06437
J. Phys. Chem. B 2021, 125, 12166−12176