JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 1205−1214

pubs.acs.org/jcim

Shape-Based Generative Modeling for de Novo Drug Design

Miha Skalic,[†] José Jiménez,[†] Davide Sabbadin,[†] and Gianni De Fabritiis*[,†,‡,§]

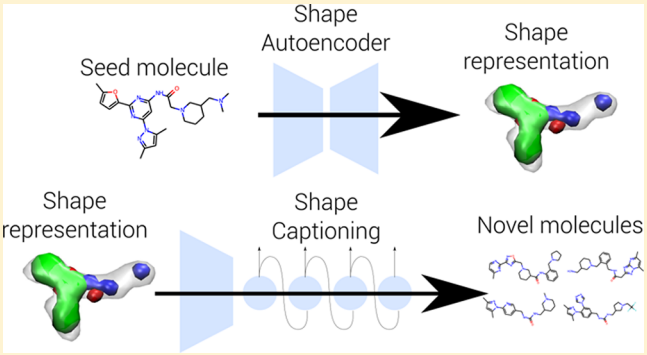
[†]Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C Dr Aiguader 88, 08003 Barcelona, Spain

[‡]Acellera, Barcelona Biomedical Research Park (PRBB), C Dr. Aiguader 88, 08003 Barcelona, Spain

[§]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain

Supporting Information

ABSTRACT: In this work, we propose a machine learning approach to generate novel molecules starting from a seed compound, its three-dimensional (3D) shape, and its pharmacophoric features. The pipeline draws inspiration from generative models used in image analysis and represents a first example of the de novo design of lead-like molecules guided by shape-based features. A variational autoencoder is used to perturb the 3D representation of a compound, followed by a system of convolutional and recurrent neural networks that generate a sequence of SMILES tokens. The generative design of novel scaffolds and functional groups can cover unexplored regions of chemical space that still possess lead-like properties.



1. INTRODUCTION

Neural-network-inspired models have been recently used in computational biology[1] and chemistry to perform protein−ligand activity classification,[2,3] affinity[4] and property prediction,[5] design of synthetic routes,[6,7] and protein folding evaluation.[8] Application of such models to solve common drug discovery problems showed similar or superior predictive performance compared with existing computational methods, thus enabling saving of significant resources when used in a drug discovery program.

In the past years deep generative models have also provided researchers with new unexplored opportunities for applications. These try to learn an underlying data distribution in an unsupervised setting, with two of the most popular methods being generative adversarial networks (GANs)[9] and variational autoencoders (VAEs).[10] Applications of generative models include audio synthesis[11] and image[12] and video[13] generation. Generative models have also been proposed as a viable solution to enable the exploration of the vast drug-like chemical space, estimated to contain approximately 10^{33} compounds.[14]

Most of the previous works focused on generating unseen chemical structures by using recurrent neural networks (RNNs) and a library of compounds satisfying desired properties as input.[15,16] A policy-based reinforcement learning approach to tune RNNs for episodic tasks was proposed by both Olivecrona et al.[17] and Popova et al.[18] Similar models have been proposed for peptide design.[19] A VAE[20] was used by Gómez-Bombarelli et al.[21] to encode a molecule in continuous latent space with the aim of exploring associated properties. Kang and Cho[22] have extended the usage of these models to generate molecules with desired properties.

These approaches are ligand-centric and model compounds as SMILES strings or as primary sequences in the case of peptide design. A sizable list of diverse molecules with known activity at the protein target of interest is also required in order to be able to generate meaningful novel drug candidates. A clear shortcoming of sequence-based generative modeling is that the generated structures are biased toward the design of compounds obtainable by small chemical modifications.[23] On the other hand, shape-based screening methods[24] such as ROCS[25] can screen novel scaffolds, although the outputs are strongly dependent on the input molecular library definition.

Chemical-reaction-driven de novo design of potentially bioactive compounds[26] can improve the output in terms of diversity, but the chemical space explored is limited to the number and variety encoded by the chemical reactions dictionary and building blocks. Other de novo design approaches such as LigBuilder[27] generate a pharmacophore model for a given protein pocket, followed by iterative compound extension through addition of fragments from a predefined library to fit the pharmacophore model and pocket shape constraints. LigVoxel[5] extrapolates a distribution of pharmaco-fields of ligand properties (such as occupancy or aromatics) given the three-dimensional (3D) structure of a protein pocket, but one still needs to screen a large library of

Special Issue: Machine Learning in Drug Discovery

Received: October 10, 2018
Published: February 14, 2019

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

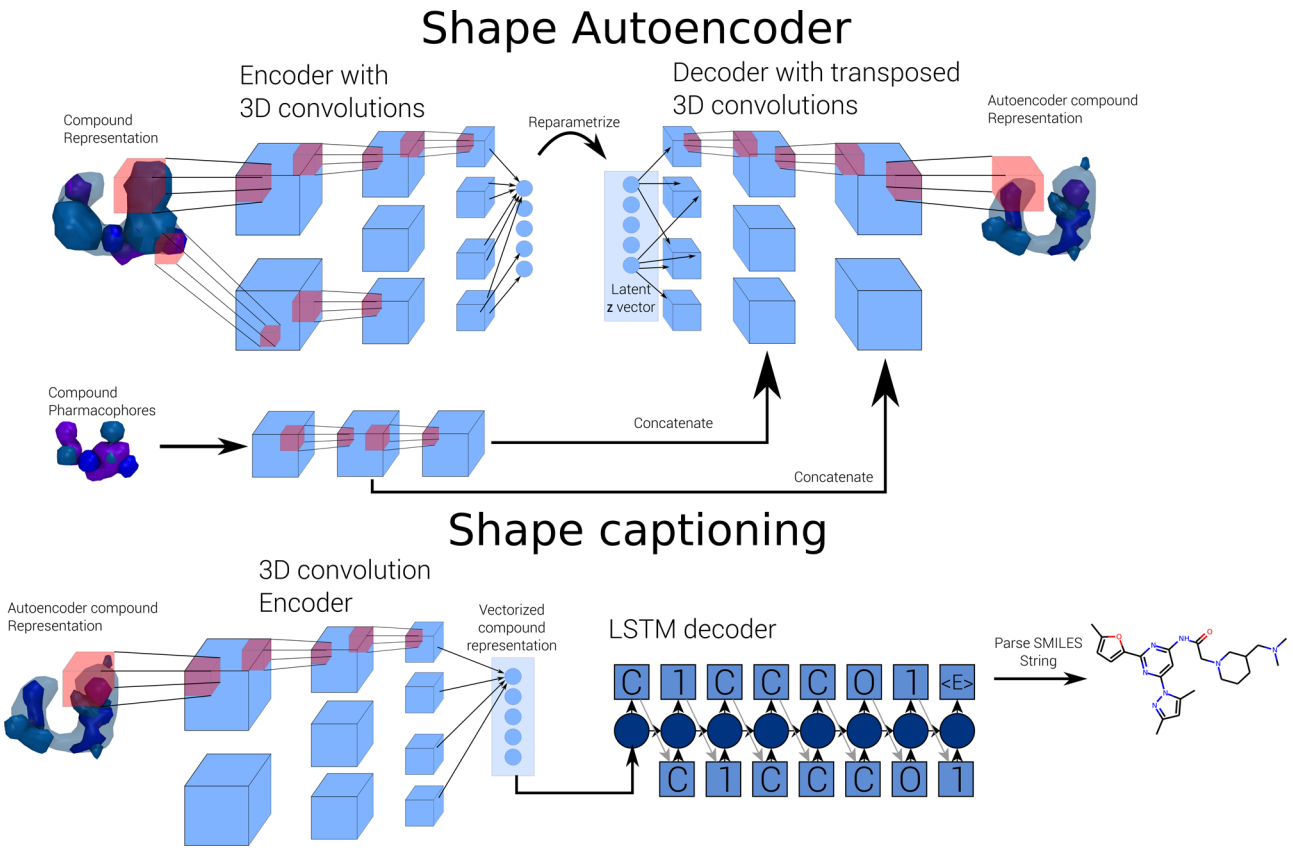


Figure 1. Proposed compound generating pipeline consisting of (top) a shape autoencoder and (bottom) a shape captioning network.

compounds to obtain hit candidates. Partially as a solution of this limitation of LigVoxel, we have developed the method presented here.

It is important to keep in mind that chemical feasibility must be an essential prioritization criterion for de novo-designed potential drug candidates, although synthesis routes can be also assessed for new molecules using automatic planning tools.[7] Furthermore, synthesizing scaffolds that have been completely unexplored in modern drug discovery programs is an essential element to foster innovation in both the pharmaceutical and agrochemical industries.

The method proposed herein takes as input voxelized molecule representations, learns SMILES grammar, and is capable of generating previously unseen compounds in a probabilistic setting where the output SMILES string sequence is extended on the basis of the 3D input shape and previous tokens of the sequence. The pipeline consists of two deep learning models: a conditional variational autoencoder (cVAE)[28] and a captioning network.[29]

2. METHODS AND DATA

The shape-based compound generation pipeline was inspired by structure-based drug design, where spatial information is taken into account, and consists of two main steps: (i) a shape variational autoencoder using convolutional neural networks (CNNs) autoencodes the compound representation and (ii) a combination of CNNs and long short-term memory (LSTM)[30] networks generates SMILES strings. A schematic representation of the process in shown in Figure 1. The role of the shape autoencoder is to generate an imperfect representation of the molecule, where fine details are lost. The pharmacophores, passed into the VAE decoder as conditional input, aid in the shape reconstruction process and prevent major deviations of pharmacophoric points if the VAE latent space is perturbed. This is followed by a captioning network that generates molecules fitting the shape of the ligand representation.

Data. The models were trained using the ZINC 15 database,[31] limiting ourselves to a drug-like subset (logP $\leq$ 5 and molecular weight 250−500 Da). The database was filtered for compounds with SMILES strings of length inferior to 60 characters and unlikely molecules such as those containing radicals. The representations were further simplified by removing characters containing isomer information, preserving strings containing 26 tokens belonging to the following set:

$$\{C, c, N, n, S, s, P, O, o, B, F, I, \text{Cl}, [nH], \text{Br}, 1, 2, 3, 4, 5, 6, \#, =, -, (,)\} \quad (1)$$

The set of molecules was then randomly split into a set and a test set containing 192 813 983 and 192 779 540 molecules, respectively. Because of the large data set size, not all of the molecules available in the database were used in this study.

Finally, three protein targets regulating key biological processes and associated ligands were selected from the DUD-E database[32] in order to compare existing molecule generative methods, understand the potential of the method to generate meaningful molecules, and define its applicability domain. In detail, the adenosine A2A receptor (AA2AR) set (PDB ID 3eml) composed of 464 binders and 31 614 decoys, the thrombin (THRB) set (PDB ID 1ype) with 461 binders and 26 943 decoys, and the stem cell growth factor receptor (KIT) set (PDB ID 3g0e) with 166 binders and 10 450 decoys were selected. Known binders, decoys, and generated molecules were

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

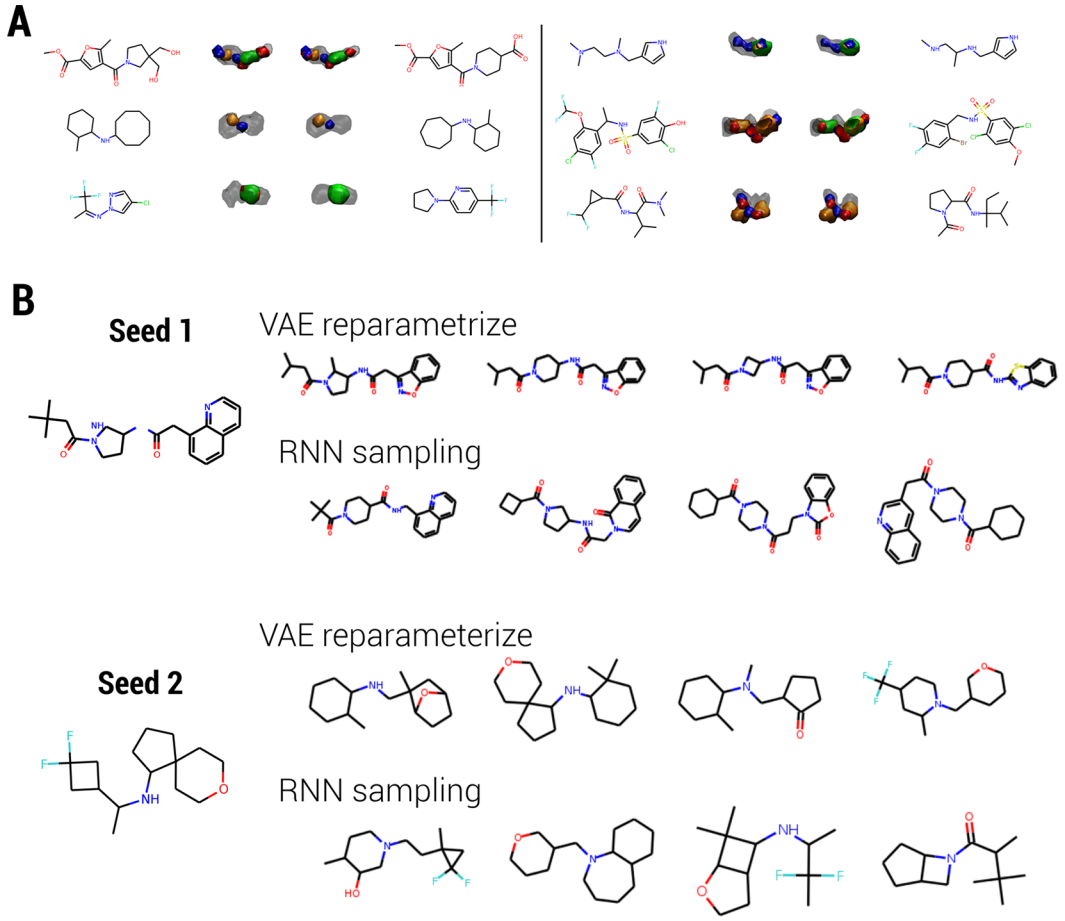


Figure 2. (A) Examples of six generated molecules. Columns (from left to right): input molecule, input volumetric representation, autoencoded volumetric representation, decoded molecule. Colors: transparent gray, occupancy; green, aromatic rings; red, H-bond acceptor; blue, H-bond donor. (B) Examples of de novo-generated molecules starting from two different seed molecules using VAE reparametrization ($\lambda = 1$) and RNN sampling as sources of variability.

docked into the corresponding protein pockets using smina[33] with default parameters and a box side of 20 Å for all three dimensions.

Featurization and Model Training. For each SMILES string, three-dimensional conformers were generated via RDKit and optimized using the MMFF94 force field[34] with default settings. Molecule atoms were then voxelized into a discretized 1 Å cubic grid of side size 24 Å prior to a random rotation and 2 Å translation of the molecule. The value at each voxel is determined by its atom type and the distance r between neighboring atoms and its center:

$$n(r) = 1 - \exp[-(r_{\text{vdw}}/r)^{12}] \tag{2}$$

where r_{vdw} is the corresponding van der Waals radius of a particular atom. An implementation of this procedure is provided in the HTMD[35] Python package, considering five channels: hydrophobic, aromatic, H-bond donors, H-bond acceptors, and heavy atoms (occupancy). Pharmacophoric points, input to the cVAE as conditions, were featurized in a similar way as the atoms by placing property points close to (with 0.5 Å random displacement) atoms with that property. Only pointlike features (H-bond donors, H-bond acceptors, and centers of aromatic rings) were used, while occupancy and hydrophobic regions were ignored. These points were encoded as fixed-size atoms with a radius of 2 Å. During training the canonical version of the SMILES representation was used as the target string.

The shape cVAE was trained simultaneously with the captioning network, feeding the output of the former into latter, although two different optimizers and losses were used for each. The autoencoder was trained to minimize a sum of binary cross-entropy and Kullback–Leibler (KL) divergence:[36]

$$\mathcal{L}_{\text{VAE}} = -\frac{1}{N}\sum_{j=1}^{24^3\times 5} y_j \log(p_j) + \mathcal{L}_{\text{KL}} \tag{3}$$

where $y_i \in Y^{24^3\times 5}$ and $p_i \in P^{24^3\times 5}$ represent ground-truth and model-generated voxel arrays, respectively, and $\mathcal{L}_{\text{KL}}$ for sample i is defined as

$$\mathcal{L}^i_{\text{KL}} = -\frac{1}{2}\sum_{j=1}^{J}[1 - \mu_j^2 - \sigma_j^2 + \log(\sigma_j^2)] \tag{4}$$

where J is the dimensionality of the latent vector and μ and σ are the outputs of the VAE encoder. The captioning network minimized multiclass logloss:

$$\mathcal{L}_{\text{caption}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log(p_{ij}) \tag{5}$$

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

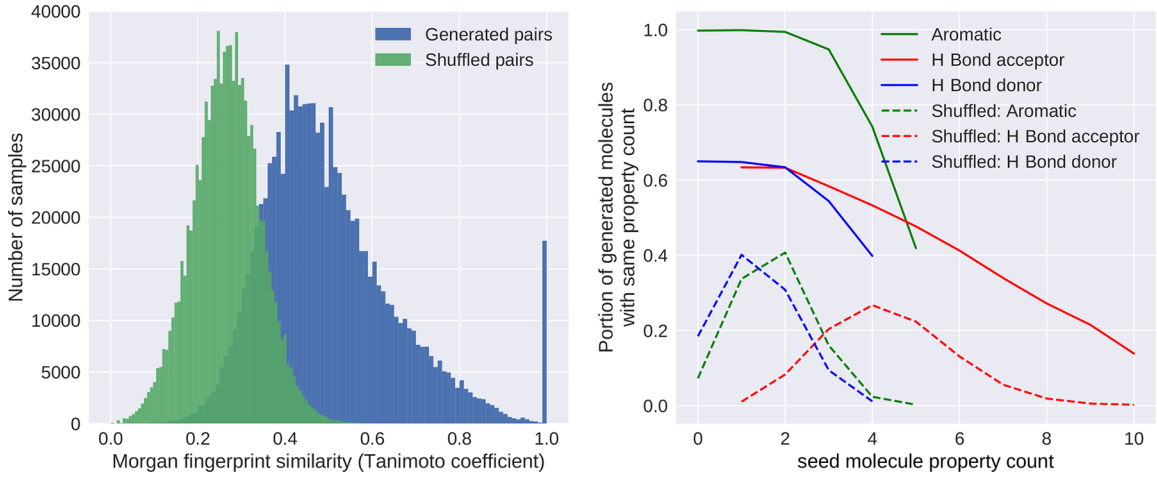


Figure 3. (left) Fingerprint similarity between seed and generated molecules in blue and randomly paired generated and seed molecules as the baseline in green. (right) Portion of generated molecules that have the same property count as the seed molecule. Dotted lines represent random baselines.

where N is the number of samples in a batch and M is the length of the sequence. The training was done in batches of 128 samples.

The model architectures, presented in Figures S1 and S2, were taken from the 2D image analysis domain[29,37] and adapted for 3D inputs, taking into account graphical memory constraints. No explicit hyperparameter optimization was performed. The shape VAE and shape captioning models were trained for 210 000 iterations, until convergence of the captioning network was observed (Figure S3). A total of 26.88 million molecules were randomly drawn without replacement from the set reserved for training and featurized on the fly. Both networks were optimized with the Adam optimizer[38] using standard momentum hyperparameters (β_1 = 0.99, β_2 = 0.999) with starting learning rates of 10^{-4} and 10^{-2} for the VAE and captioning networks, respectively. Every 60 000 iterations the captioning learning rate was reduced by half. The training took approximately 10 days on two NVIDIA GeForce GTX 1080Ti GPUs, written in the PyTorch[39] deep learning framework. The training and generation code and trained models are publicly available.

Generation of Novel Molecules. Once the models were trained, two sources were used to generatively design novel molecules: (i) autoencoder reparametrization and (ii) RNN sampling. In the case of reparametrization, we perturbed the encoded latent vector $\mathbf{z}$ by adding standard Gaussian noise and consequently changing the voxelized representation of the decoding. To control the amount of added noise, we introduced a variability factor λ, thus changing the reparametrization to

$$\mathbf{z} = \boldsymbol{\mu} + \lambda\boldsymbol{\epsilon} \odot \boldsymbol{\sigma} \tag{6}$$

where the vector $\boldsymbol{\epsilon}$ is sampled from a zero-mean, unit-variance multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ given by the VAE encoder and $\odot$ denotes elementwise multiplication. Increasing λ also increases the deviance of $\mathbf{z}$ from $\boldsymbol{\mu}$ and consequently the deviance of the generated molecule representation from its input.

A max-sampling strategy was used in the LSTM, meaning that SMILES strings were grown by selecting the next token on the basis of its highest predicted probability. Alternatively, probabilistic sampling was used, choosing the next token proportionally to its predicted probability. We refer to this process as RNN sampling.

On a single modern GPU-enabled device such as an NVIDIA Titan V, we can decode and caption approximately 250 molecule representations per second.

3. RESULTS

In this section we first analyze the model performance in terms of generating novel compounds from selected seed compounds. We also analyze the compounds generated in an unfocused setting and compare the output of the method described in this work to that of a previously proposed string-based de novo molecule design approach.[21] Finally, we test a structure-based drug design application and use virtual screening to assess the applicability of the de novo design.

Strengths and Limitations of the Shape-Based de Novo Molecular Design Method. One million ZINC compounds from the test set were randomly selected and used in the evaluation. For each SMILES string we generated a 3D conformer using RDKit, voxelized it, autoencoded the voxelized representation using the shape autoencoder, and finally generated the output SMILES strings via the captioning network. The pipeline was able to reproduce 17 402 (1.74%) SMILES strings exactly, while 4513 (0.451%) strings were invalid according to the parser available in RDKit. Out of these, 65% were parsable after disabling sanitation and allowing for nonstandard valence. The remaining strings had an invalid combination of bracket openings and closings (35%) and/or invalid branching (76%). For the rest, in Figure 2A as well as in Figures S4 and S5 we show examples of generated SMILES strings selected according to their seed diversity by max−min picking[40] as implemented in RDKit. Out of the valid generated compounds, about 4.5% were present in the training set, and that number could be reduced to 2.6% by introducing RNN sampling when constructing the SMILES sequence. When comparing reconstructed molecules to their seed counterparts, we often observed changes such as formation, fusion, or opening of ring structures, homologation or shortening of carbon chains, and shifts of functional groups to neighboring atoms. Similar modifications occur when starting from the same compound and using either reparametrization of the latent autoencoder space $\mathbf{z}$ or RNN sampling for diverse generative design (Figure 2B).

In Figure 3 we show the results of Tanimoto similarity of Morgan fingerprints (2 bond radius) and property counts of the

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

output SMILES strings. With a mean Tanimoto similarity of 0.5, the reconstructed SMILES strings show significantly higher similarity than the random baseline (Mann−Whitney U test $p \ll 0.001$) with a mean similarity of 0.27. The random baseline was set by randomly pairing a seed compound with a generated compound. Throughout model captioning loss decreases, producing more similar SMILES strings. In terms of generating valid molecule representations, however, less than 10 000 training steps were needed for the model to start generating valid SMILES strings (Figure S3). By analyzing the properties of the generated SMILES strings (Figure 3, right), we can conclude that the model is efficient at reconstructing molecules with few properties, but the longer the SMILES string, the harder it is to generate a similar molecule with same property count. We attribute this problem to the shortcoming of LSTMs in decoding fixed-length vectors associated with long target sequences, as in language translation tasks,[41] where the performance also drops as the target sequences get longer. Out of the three evaluated properties, aromatic rings seem to be the easiest to reconstruct. When given a molecule containing one or two aromatic rings, in almost all the cases the pipeline will return a molecule with same number of rings. On the other hand, H-bond donors and acceptors are more difficult to recover (Figure 3, right). The same count retrieval probability is greater than 0.6 in the case of two or fewer atoms but drops for more complex molecules. This can be explained by the SMILES representation: aromatic atoms are represented as lower-case letters, but donors and acceptors are not specifically marked. Atoms such as O, N, and S can be considered as H-bond-interacting, but whether they actually are depends on their surroundings. Overall, in 98.08%, 49.38%, and 63.35% of the cases the property count is conserved for aromatics, H-bond donors, and H-bond acceptors, respectively.

There are downsides to the proposed method that limit its usage as a purely pharmacophore-based method without the use of any reference ligand shapes. When using only pharmacophores and randomly sampling $\mathbf{z}$ values of the shape cVAE, the decoder commonly does not generate plausible or desirable shapes. For example, the shape surface can be disjoined, or no area is generated around pharmacophoric points (Figure S6). When these implausible shapes were fed into the captioning network, we did not observe increased generation of invalid SMILES strings, unlike in previous sequence-to-sequence generative models,[21] where random sampling of the latent $\mathbf{z}$ would yield invalid SMILES strings with high probability. We believe that recent methods based on generative adversarial networks[42] can aid in sampling of latent space that is compatible with a desirable shape and pharmacophore constraints. This optimization is outside the scope of this work, and thus, in the following section we use seed molecules to generate the corresponding latent compound $\mathbf{z}$ vector. On the other hand, conditioning the VAE decoding on the pharmacophore location improves the reconstruction of pharmacophore-like channels, especially when the latent representation is perturbed with a bigger λ (Figure S7).

It is also worth mentioning that the model output is not translation-, rotation-, and conformer-invariant, and thus, a different molecule can be generated on the basis of the input conformer and voxelization state (Figure S8). However, we observed that these changes usually have a minor effect on the average similarity to the seed. Similarly, switching the conformer generation force field has a minor effect, while using docked compounds can cause a bigger difference in fingerprint similarities (Figure S9).

Shape-Based Unfocused Compound Generation. To provide insight into the model performance and identify potential strengths and weaknesses compared with other generative models, the method was benchmarked on the MOSES platform.[43]

Following the proposed pipeline, models were trained on 250 000 clean leads from the ZINC data set and tested on 10 000 test and scaffold split test sets. We used the provided models and hyperparameters available in the platform, such as the character-level recurrent neural network (CharRNN),[15] a variational autoencoder (VAE), an adversarial autoencoder (AAE),[44] an objective-reinforced generative adversarial network (ORGAN)[45] and a junction tree variational autoencoder (JTN-VAE).[46] The shape-based model proposed herein was trained on the same data set with a batch size of 64 samples over 50 epochs, with the other hyperparameters set as described in Methods and Data. To make the method work in an unfocused case, the conditional input was dropped, allowing novel compounds to be generated just by sampling from latent VAE vector using a standard normal distribution and decoding using either max sampling (Shape) or probabilistic RNN sampling (ShapeProb).

In an unfocused setting the proposed shape-based model performs similarly to other generative models, with a high percentage of valid and unique SMILES strings (Table 1). As expected, performing RNN sampling increases the variability at the cost of generating more invalid SMILES strings. The experiment also showed that the shaped-based model produces compounds with the highest internal diversity. However, the increased diversity comes at the cost of producing compounds that might be classified as undesired or reactive: 14.4% and 13.5% of max-sampling and RNN sampling molecules, respectively, did not pass the PAINS[47] filters and others reported by Polykovskiy et al.[43] We believe that these are acceptable values in the context of the generation of focused libraries for virtual screening purposes.

Table 1. Properties of the Generated Compounds for Six Generative Models: Properties Include Numbers of Valid and Unique Compounds, Internal Diversity (IntDiv; Average Pairwise Similarity), and Portion of Compounds That Pass Medicinal Chemistry and PAINS Filters (Filters)

model	Valid	Unique@1k	Unique@10k	IntDiv	Filters
AAE	0.937	1.000	1.000	0.857	0.975
CharRNN	0.963	1.000	1.000	0.856	0.992
JTN-VAE	1.000	0.999	0.997	0.851	0.958
ORGAN	0.727	0.960	0.858	0.703	0.888
VAE	0.953	1.000	0.999	0.855	0.993
Shape	0.987	0.989	0.938	0.856	0.856
ShapeProb	0.969	1.000	0.995	0.865	0.865

In terms of the Fréchet ChemNet Distance,[48] a metric that encompasses compound diversity together with chemical and biological properties, the values for our method are on a par with others (Table 2). Fragment[49] and scaffold[50] frequencies as well as similarity to the nearest neighbor are also close to those for other methods.

By analyzing the distributions of lipophilicity,[51] quantitative estimation of drug-likeness,[52] synthetic accessibility,[53] natural-product-likeness,[54] and molecular weight (Figure 4), we observed that the shape-based method produces compounds

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

Table 2. Properties of Generated Compounds for Six Generative Models Calculated in Relation to the Test Set (Test) Compounds and Scaffold-Based Split Set (TestSF): FCD, Fréchet ChemNet Distance; SNN, Nearest-Neighbor Similarity; Frag, Cosine Distance between Vectors of Fragment Frequencies; Scaff, Cosine Distance between Vectors of Scaffold Frequencies

	FCD		SNN		Frag		Scaff	
model	Test	TestSF	Test	TestSF	Test	TestSF	Test	TestSF
AAE	1.711	2.343	0.422	0.412	0.992	0.989	0.758	0.130
CharRNN	0.342	0.841	0.461	0.449	0.998	0.997	0.806	0.123
JTN-VAE	3.924	4.308	0.386	0.386	0.964	0.968	0.375	0.116
ORGAN	44.246	46.014	0.364	0.334	0.765	0.758	0.536	0.007
VAE	0.254	0.696	0.468	0.455	0.998	0.996	0.828	0.093
Shape	1.819	2.472	0.469	0.450	0.979	0.973	0.433	0.046
ShapeProb	1.332	1.850	0.446	0.432	0.984	0.980	0.459	0.066

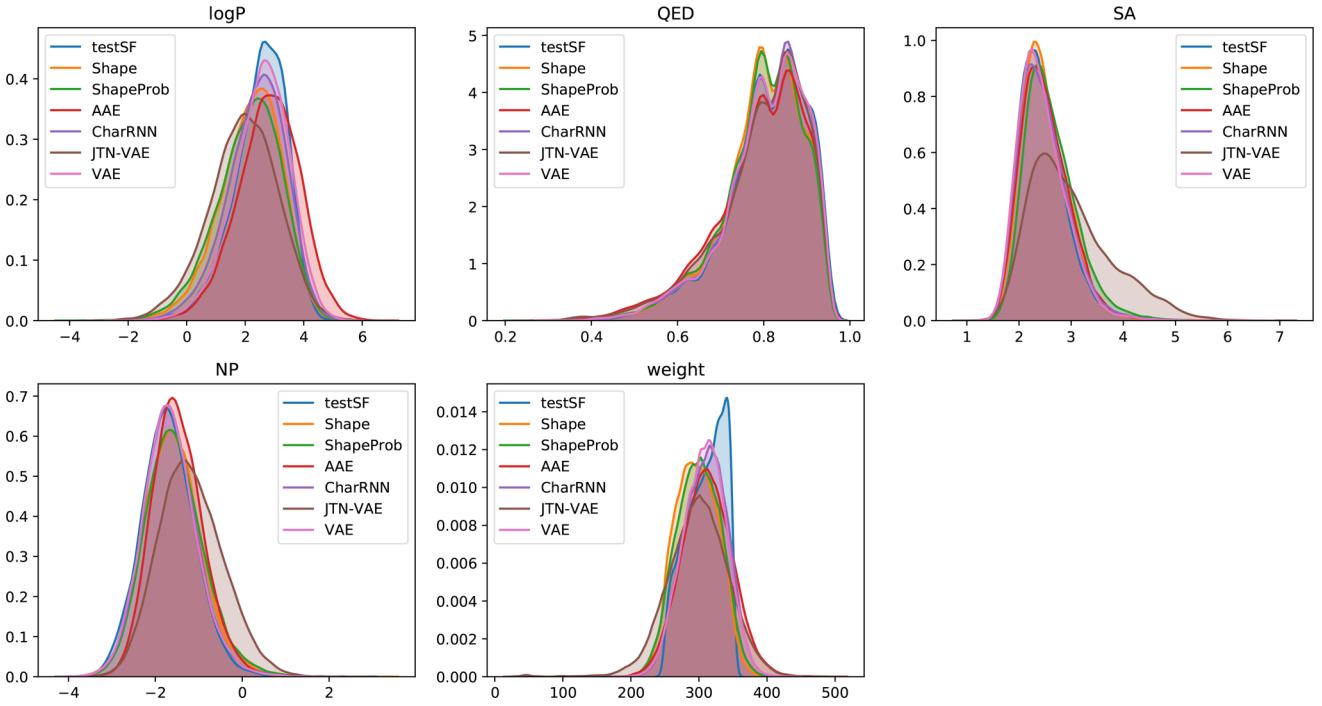


Figure 4. Distributions of property values for the generated compounds. Properties include lipophilicity (logP), quantitative estimation of drug-likeness (QED), synthetic accessibility (SA), natural-product-likeness (NP), and molecular weight. Because of a strong distribution deviation, results for ORGAN are not shown.

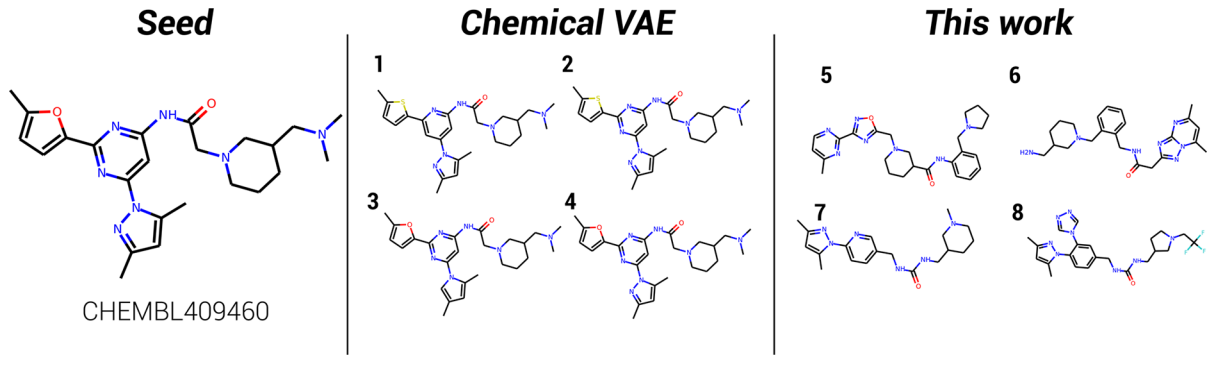


Figure 5. Molecules generated using Chemical VAE and the method proposed herein starting from the potent AA2AR antagonist CHEMBL409460.

with property value distributions similar to the ones produced by other methods.

Applicability and Strengths in a Drug Discovery Project. Exploration of chemical space based on continuous encodings of molecules was recently described by Gómez-Bombarelli et al.[21] In order to understand our performance and highlight what we believe are the strengths of the two methods when applied in a integrated drug discovery project, we implemented their proposed method (here named Chemical VAE; trained on the same data) and compared its output to that of this work (Figures 5 and 6).

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

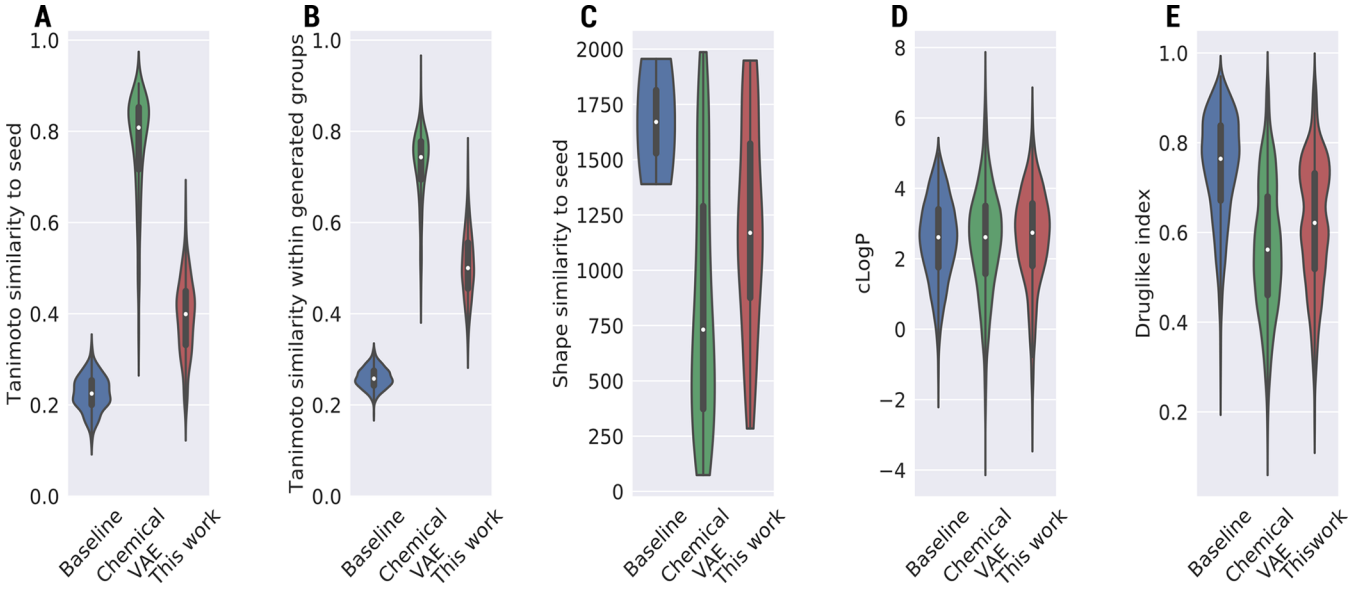


Figure 6. Properties of generated molecules derived from known AA2AR binders. From each of the 461 AA2AR binding compounds, 10 novel compounds were generated and used for property calculations. Baseline represents randomly sampled compounds from the drug-like subset of the ZINC data set.

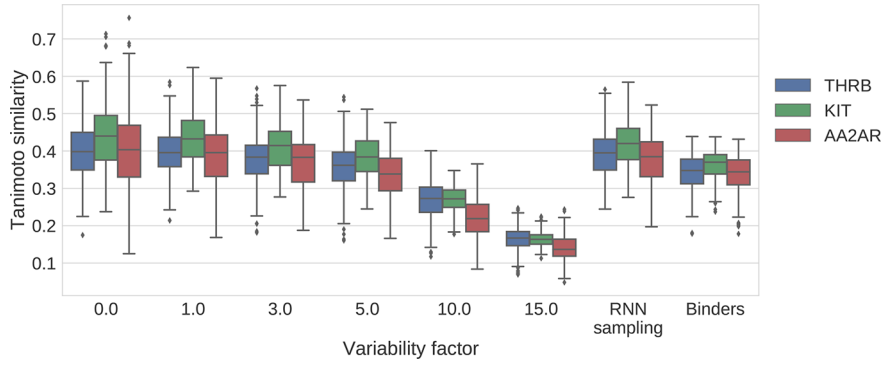


Figure 7. Morgan fingerprint (2 bond radius) similarities between target binders and generated molecules given different variability factors (λ) as well as LSTM transition sampling (RNN sampling). Each point is the average similarity of 50 generated molecules to the seed molecule, except for $\lambda = 0.0$, where a single molecule was generated. The Binders column shows fingerprint similarities of the seed compounds.

The 2-amino-*N*-pyrimidin-4-ylacetamide human AA2AR antagonist ChEMBL409460 was set as the seed molecule, and compounds generated by Chemical VAE display conservative variations of the chemical structure (e.g., from furan to thiophene (compound **2**) and/or pyrrole to pyrazole (compounds **3** and **4**) in Figure 5), which is desirable when embarking on a systematic lead exploration program. Our method, in contrast, enables a bigger structural leap from the seed molecule (Figure 5 compounds **5**−**8**). Interestingly, it is also able to insert a fluorinated alkyl group at the basic nitrogen (compound **8**), which has the effect of lowering the pK_a, indicating the potential to modify even more substantially the properties of the novel designed ligand.

To compare the two methods on a larger scale, novel molecules were generated on the basis of 464 known AA2AR receptor binders; we generated 10 molecules based on each seed molecule. For our method we used $\lambda = 1$, while for Chemical VAE we set the perturbation factor to 10. As a random baseline we randomly picked from the ZINC database an equivalent number of compounds. When comparing similarity to the seed and within its group (Figure 6A−C), we report average per seed values. Shape similarity was computed using the USRCAT method[55] as implemented in the Open Drug Discovery Toolkit,[56] and the shape distance R was calculated as follows:

$$R = \sum_{i=1}^{60} (d_{ri} - d_{qi})^2 \tag{7}$$

where $\mathbf{d}_r$ and $\mathbf{d}_q$ are reference and query USRCAT shape descriptors with length 60, respectively. Both drug-like index[52] and cLogP[51] were computed using RDKit.

Analyzing the results in Figure 6, we observed less diversity in the molecules generated by Chemical VAE in comparison with their seeds (average Tanimoto similarity > 0.8; Figure 6A). Also, the compounds generated by Chemical VAE within the generated series are less diverse than the ones generated by the proposed method (average Tanimoto similarity > 0.7; Figure 6B) while retaining drug-like properties (Figure 6D,E). We have also seen a higher rate of invalid SMILES strings when sampling for higher chemical variability in the latent space for Chemical VAE.

Despite being clearly diverse from the set, the generated molecules have a structural similarity consistently higher than drug-like molecules randomly picked from ZINC (average

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

Tanimoto similarity = 0.2−0.3; Figure 6A,B). The USRCAT shape similarity (Figure 6C) follows the same distribution pattern.

The method for exploration of chemical space reported by Gómez-Bombarelli et al.[21] has good potential in constructing libraries for lead optimization, as the modifications are minor, whereas because of its larger modifications, the approach we propose has its applicability domain in the area of lead discovery, where larger modifications are accepted and in some cases desirable.

Finally we evaluated how changing the variability factor λ and using probabilistic RNN transitions affect molecule generation. To that end, for the three analyzed targets (AA2AR, THRB, and KIT), 50 molecules were generated from each docked binder under eight different conditions ($\lambda = \{1, 2, 3, 4, 5, 10, 15\}$ or using probabilistic RNN sampling). Generated compounds are available in the Supporting Information. Analysis of the compounds (Figure 7) shows that increasing the variability factor causes the generated molecules to have a less similar fingerprint profile to the seed molecule, and thus, this parameter can be tweaked to generate more or less similar compounds. The variability among generated molecules obtained by RNN sampling is similar to the variability introduced by a small λ.

Applicability and Strengths in a Structure-Based Virtual Screening Project. Having demonstrated our ability to generate libraries of diverse de novo-designed molecules from a lead compound, we turned our attention to understanding the applicability of the proposed methodology in virtual screening applications.

In order to evaluate whether the generated compounds could be good binder candidates, we tested the compounds generated for the AA2AR, KIT and THRB target receptors. The compounds, generated as described in the previous section, were docked and scored. We limited the docking to compounds generated using $\lambda = \{1, 5\}$ and RNN sampling. Furthermore, the compounds were filtered for replicates across different seeds. Values of the area under the receiver operating characteristic (ROC) curve (AUC) calculated by Autodock Vina binding affinities and BindScope binding probabilities[57] were chosen as the evaluation metrics. The results are presented in Table 3, and

Table 3. AUC Scores for Two Virtual Screening Tools (smina and BindScope) and Three Targets (AA2AR, THRB, and KIT) When Comparing Decoys versus Actives or Decoys versus Generated Molecules under Different Conditions

	actives	$\lambda = 1$	$\lambda = 5$	RNN sampling
AA2AR−smina	0.700	0.637	0.462	0.619
AA2AR−BindScope	0.859	0.709	0.680	0.727
THRB−smina	0.685	0.550	0.432	0.556
THRB−BindScope	0.970	0.738	0.653	0.742
KIT−smina	0.759	0.654	0.581	0.634
KIT−BindScope	0.986	0.862	0.818	0.864

the ROC curves can be found in Figure S10. For all three targets and both scoring methods, the generated molecules have lower average scores (and AUC values) than the actives from the DUD-E database, meaning that these virtual screening tools are less likely to classify generated molecules as binders than the actual experimentally proven ones. However, the scores are consistently above the 0.5 random baseline, and thus, the screening tools are more likely to classify the generated molecules as binders rather than decoys. The only two exceptions are results from smina AA2AR and THRB given high λ, where the AUC values fall bellow 0.5. As expected, an increase in the λ factor from 1 to 5 generates less similar molecules, causing lower virtual screening scores. Using RNN sampling as the generative source instead yields results comparable to those for the VAE reparametrization with λ = 1, although in three out of six cases better results were obtained.

4. DISCUSSION AND FUTURE WORK

In this work, we have developed a novel method for generating new molecules from three-dimensional representations. The code is available at https://github.com/compsciencelab/ligdream, and a web application is also available at https://playmolecule.org/LigDream/. The approach, which uses autoencoders and captioning networks, can variationally design multiple molecules starting from a single volumetric representation. To the extent of our knowledge, this is the first time that generative captioning-like networks have been applied to generate sequences from 3D input, let alone molecule design. Furthermore, the method described here does not need retraining on focused libraries or sophisticated manipulation in latent space to generate diverse compounds. Instead, it is sufficient to have just one seed molecule (e.g., a patented molecule) to generate similarly shaped compounds.

We have shown that the proposed method is capable of designing novel compounds with desirable characteristics such as drug-likeness and tunable similarity to a seed molecule while at the same time being able to generate a diverse population of novel compounds. However, there are several directions for future improvements to advance generative drug design. For example, tuning the proposed method to introduce smaller changes to the seed molecule would allow it to fit better into the scope of lead optimization. We believe that this could be achieved by directly encoding elements, increasing the resolution of the input representation, preseeding part of the output SMILES sequence,[16] and introducing changes to the model architecture such as an attention mechanism.[58] Another direction for future work could be coupling of generative adversarial networks or reinforcement learning methods[18] with the proposed approach, enabling direct protein to binding compound mapping via generation of a complementary 3D ligand representation from its pocket without the need for a seed molecule. Another direction would focus on the generation of molecules with desirable properties (e.g., favorable ADME) while still maintaining the shape, which could be achieved by adding additional conditions to the decoders.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00706.

Details of model architectures and figures showing generated molecules (PDF)

Files (.sdf) for molecules generated using the method proposed here (chemical VAE method), with randomly selected molecules aligned to the seed molecule from which they were generated (ZIP)

Tables (.csv) with generated molecules given different conditions for the three proteins used in this study (ZIP)

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

AUTHOR INFORMATION

Corresponding Author

*E-mail: gianni.defabritiis@upf.edu.

ORCID

Miha Skalic: 0000-0003-4143-4609
José Jiménez: 0000-0002-5335-7834
Gianni De Fabritiis: 0000-0003-3913-4877

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Acellera for funding. G.D.F. acknowledges support from MINECO (Unidad de Excelencia María de Maeztu MDM-2014-0370 and BIO2017-82628-P) and FEDER. This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 675451 (CompBioMed Project). Special thanks go to Gerard Martínez-Rosell and Alberto Cuzzolin for implementing the tool on the playmolecule.org website.

REFERENCES

(1) Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep Learning for Computational Biology. *Mol. Syst. Biol.* **2016**, *12*, 878.

(2) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.

(3) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neur[a]l Network for Bioactivity Prediction in Structure-Based Drug Discovery. 2015, arXiv:1510.02855 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1510.02855 (accessed Oct 10, 2018).

(4) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287−296.

(5) Skalic, M.; Varela-Rial, A.; Jiménez, J.; Martínez-Rosell, G.; De Fabritiis, G. LigVoxel: inpainting binding pockets using 3D-convolutional neural networks. *Bioinformatics* **2019**, *35*, 243−250.

(6) Segler, M. H.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23*, 5966−5971.

(7) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.

(8) Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. 2018, arXiv:1801.06252 [q-bio.BM]. arXiv.org e-Print archive. https://arxiv.org/abs/1801.06252 (accessed Oct 10, 2018).

(9) Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A. A. *IEEE Signal Process. Mag.* **2018**, *35*, 53−65.

(10) Santerne, A.; Moutou, C.; Tsantaki, M.; Bouchy, F.; Hébrard, G.; Adibekyan, V.; Almenara, J.-M.; Amard, L.; Barros, S. C. C.; Boisse, I.; Bonomo, A. S.; Bruno, G.; Courcol, B.; Deleuil, M.; Demangeon, O.; Díaz, R. F.; Guillot, T.; Havel, M.; Montagnier, G.; Rajpurohit, A. S.; Rey, J.; Santos, N. C. Auto-Encoding Variational Bayes. *Astron. Astrophys.* **2016**, *587*, A64.

(11) van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. 2016, arXiv:1609.03499 [cs.SD]. arXiv.org e-Print archive. https://arxiv.org/abs/1609.03499 (accessed Oct 10, 2018).

(12) Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention Generative Adversarial Networks. 2018, arXiv:1805.08318 [stat.ML]. arXiv.org e-Print archive. http://arxiv.org/abs/1805.08318 (accessed Oct 10, 2018).

(13) Tulyakov, S.; Liu, M.-Y.; Yang, X.; Kautz, J. Mocogan: Decomposing motion and content for video generation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2018**, 1526−1535.

(14) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(15) 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.

(16) 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.

(17) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **2017**, *9*, 48.

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

(19) Müller, A. T.; Hiss, J. A.; Schneider, G. Recurrent Neural Network Model for Constructive Peptide Design. *J. Chem. Inf. Model.* **2018**, *58*, 472−479.

(20) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. 2013, arXiv:1312.6114 [stat.ML]. arXiv.org e-Print archive. https://arxiv.org/abs/1312.6114 (accessed Oct 10, 2018)

(21) 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.

(22) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59*, 43−52.

(23) Schneider, G. *De novo* design—Hop(p)ing against hope. *Drug Discovery Today: Technol.* **2013**, *10*, e453−e460.

(24) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem.* **2018**, *6*, 315.

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

(26) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-driven *de novo* design of bioactive compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380.

(27) Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A practical de novo drug design approach. *J. Chem. Inf. Model.* **2011**, *51*, 1083−1091.

(28) Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, 2015; pp 3483−3491

(29) Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2015)*; IEEE: New York, 2015; pp 3156−3164.

(30) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735−1780.

(31) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(32) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(33) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893−1904.

(34) Halgren, T. A. Merck Molecular Force Field. *J. Comput. Chem.* **1996**, *17*, 490−519.

(35) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845−1852.

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214

(36) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(37) Hou, X.; Shen, L.; Sun, K.; Qiu, G. Deep Feature Consistent Variational Autoencoder. 2016, arXiv:1610.00291 [cs.CV]. arXiv.org e-Print archive. https://arxiv.org/abs/1610.00291 (accessed Oct 10, 2018).
(38) Kingma, D. P.; Ba, J. L. Adam: A method for stochastic optimization. 2014, arXiv:1412.6980 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1412.6980 (accessed Oct 10, 2018).
(39) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in PyTorch. Presented at the NIPS 2017 Autodiff Workshop, Long Beach, CA, Dec 9, 2017; https://openreview.net/pdf?id=BJJsrmfCZ (accessed Oct 10, 2018).
(40) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant. Struct.-Act. Relat.* **2002**, *21*, 598−604.
(41) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014, arXiv:1409.0473 [cs.CL]. arXiv.org e-Print archive. https://arxiv.org/abs/1409.0473 (accessed Oct 10, 2018).
(42) Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. 2016, arXiv:1611.07004 [cs.CV]. arXiv.org e-Print archive. https://arxiv.org/abs/1611.07004 (accessed Oct 10, 2018).
(43) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. 2018, arXiv:1811.12823 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1811.12823 (accessed Oct 10, 2018).
(44) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. 2015, arXiv:1511.05644 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1511.05644 (accessed Oct 10, 2018).
(45) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. 2017, arXiv:1705.10843 [stat.ML]. arXiv.org e-Print archive. https://arxiv.org/abs/1705.10843 (accessed Oct 10, 2018).
(46) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. 2018, arXiv:1802.04364 [cs.LG]. arXiv.org e-Print archive. https://arxiv.org/abs/1802.04364 (accessed Oct 10, 2018).
(47) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(48) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 1736−1741.
(49) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008**, *3*, 1503−1507.
(50) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(51) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Model.* **1999**, *39*, 868−873.
(52) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90−98.
(53) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.
(54) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68−74.
(55) Schreyer, A. M.; Blundell, T. USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminf.* **2012**, *4*, 27.
(56) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. *J. Cheminf.* **2015**, *7*, 26.
(57) Skalic, M.; Martínez-Rosell, G.; Jiménez, J.; De Fabritiis, G. PlayMolecule BindScope: Large scale CNN-based virtual screening on the web. *Bioinformatics* **2018**, DOI: 10.1093/bioinformatics/bty758.
(58) Xu, K.; Ba, J. L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R. S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *J. Mach. Learn. Res.* **2015**, *37*, 2048−2057.

DOI: 10.1021/acs.jcim.8b00706
J. Chem. Inf. Model. 2019, 59, 1205−1214