JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

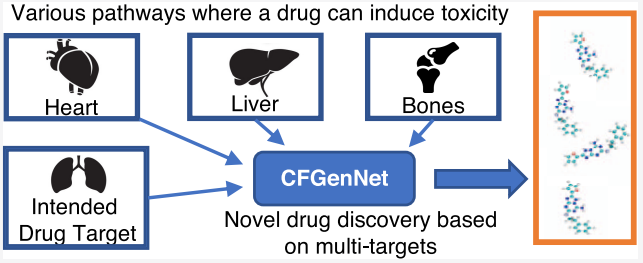
Deep Learning-Based Ligand Design Using Shared Latent Implicit Fingerprints from Collaborative Filtering

Raghuram Srinivas,* Niraj Verma, Elfi Kraka, and Eric C. Larson

Cite This: J. Chem. Inf. Model. 2021, 61, 2159−2174 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In their previous work, Srinivas et al. [*J. Cheminf.* **2018**, *10*, 56] have shown that implicit fingerprints capture ligands and proteins in a shared latent space, typically for the purposes of virtual screening with collaborative filtering models applied on known bioactivity data. In this work, we extend these implicit fingerprints/descriptors using deep learning techniques to translate latent descriptors into discrete representations of molecules (SMILES), without explicitly optimizing for chemical properties. This allows the design of new compounds based upon the latent representation of nearby proteins, thereby encoding druglike properties including binding affinities to known proteins. The implicit descriptor method does not require any fingerprint similarity search, which makes the method free of any bias arising from the empirical nature of the fingerprint models [Srinivas, R.;et al. *J. Cheminf.* **2018**, *10*, 56]. We evaluate the properties of the potentially novel drugs generated by our approach using physical properties of druglike molecules and chemical complexity. Additionally, we analyze the reliability of the biological activity of the new compounds generated using this method by employing models of protein−ligand interaction, which assists in assessing the potential binding affinity of the designed compounds. We find that the generated compounds exhibit properties of chemically feasible compounds and are predicted to be excellent binders to known proteins. Furthermore, we also analyze the diversity of compounds created using the Tanimoto distance and conclude that there is a wide diversity in the generated compounds.



INTRODUCTION

The field of virtual screening, a constituent part of the modern drug discovery process,[1,2] has been entrenched in the pharmaceutical industry for years and has developed into a sophisticated tool.[3−5] A number of successful virtual screening strategies to identify novel hits have been reported, which serves as the starting point for further investigation.[6−8] Many state-of-the-art protein−ligand interaction (PLI) models use machine learning that relies on abstract descriptors of compounds/proteins as input features.[9−14]

Recent years have seen several deep learning techniques applied to various aspects of drug discovery and development process. For example, Wallach et al.[15] introduced AtomNet, a deep convolutional neural network for bioactivity prediction. AtomNet aimed to apply the convolutional concepts of feature locality and hierarchical composition to the modeling of bioactivity and chemical interactions by taking into consideration the targets' structural information. Ragoza et al.[16] demonstrated the abilities of AutoVina, a three-dimensional (3D) convolutional neural network, which outperformed in enrichment performance on DUD-E targets. Stepniewska-Dziubinska et al.[17] demonstrated the abilities of a model named Pafnucy on the PDBbind v2013 core set using Pearson R2 coefficients. In addition, graph neural networks were also leveraged[18] to perform virtual screening.

Virtual screening in drug discovery that uses these models, while high performing, is not free of deficiencies—the limitations of representing drug compounds and targets abstractly also limit our ability to infer their binding properties.[19,20]

We argue that a critical barrier is the lack of a universal fingerprinting model that can amass knowledge about drug compounds, protein targets, and assay characteristics in a shared latent space that can be used by a variety of machine learning models, visualization tools, and compound design tools. We further argue that if the representation is completely abstract, even if it performs well at the PLI prediction, it is fundamentally limited because researchers cannot systematically create candidate compounds based on the featurization of the target.[21−23]

In their previous work, Srinivas et al.[24] proposed the conception and development of an implicit mathematical representation that allows for a more accurate characterization

Received: November 20, 2020
Published: April 26, 2021

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

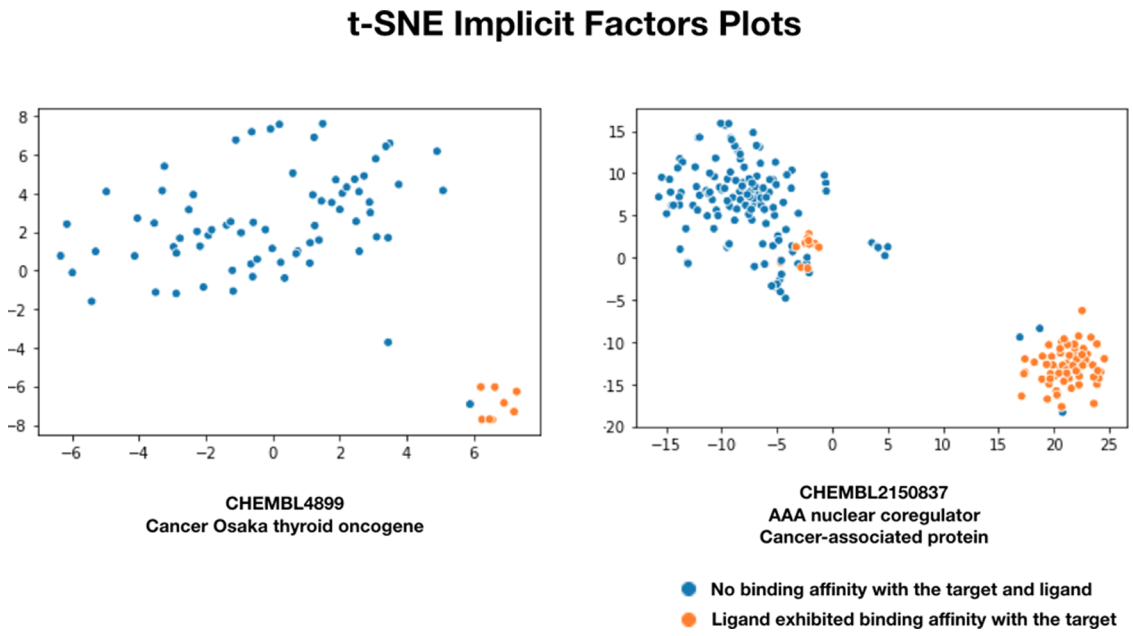


Figure 1. t-SNE plots of implicit ligand fingerprints: plots for two cancer targets are shown, where each point represents a compound assayed from the ChEMBL database. The concentration results of the assays are color-coded. t-SNE plots of the 50-dimensional implicit representations reduced to two dimensions preserving the distance.

of the drug compound and protein target in the same numeric latent space (as opposed to the current practice of separate descriptors for the compound and target), thus narrowing the model-associated bias down to that of the assay, i.e., real clinical (albeit in vitro) environment.[25] Additionally, to facilitate the ability to discover the physical structure of new compounds, in this work, we propose decoding methods that map from the implicit representation of the candidate compounds to their physical structure. We believe that this expanded capacity of the fingerprinting model will have a significant impact on virtual screening and, consequently, drug discovery, as it will render drug discovery less dependent on costly clinical facilities and services. Moreover, the representation will provide new methods for creating and testing candidate drug compounds.

Several recent works have investigated the use of neural embedding on compound structure representations such as SMILES codes, showing that this embedding is effective for exploring the chemical properties[26] and generating novel compounds. Gómez-Bombarelli et al.[27] refer to these embedded fingerprints as implicit representations. However, their methods work upon the raw SMILES textual representation and are therefore limited in their ability to discern more complicated relationships encoded by graphical fingerprints. Recent years have seen a plethora of deep learning-based generative models for de novo drug generation.[28−33] The common theme in these techniques is to provide as input to the deep learning model the molecules only to produce the same or similar molecules as output. The continuous vector representations of the input molecules in the intermediate layers produce a larger chemical property space, which is then sampled to produce novel molecules. In this work, we design and train deep learning methods that leverage the implicit compound fingerprints obtained from collaborative filtering based on the past bioactivity/assay data to map back to the physical structure of compounds. The implicit encoding of compounds is a continuous vector-valued representation and thus lends itself to the use of continuous optimization to generate potentially novel compounds. We further assess the properties of the potentially novel ligands generated in terms of the druglike physical properties of molecules, chemical complexity, and biological activity. We observed that our compounds exhibit properties similar to the known ligands even though our approach does not explicitly train the neural network for optimizing specific properties. Additionally, we compare our work to the prior work of Gómez-Bombarelli et al.[26] on a set of chemical compounds with known binding affinities to cancer targets from the ChEMBL23 database.[34] This comparative analysis investigates not only the potential binding affinity of the generated compounds to selected protein targets but also the diversity of compounds generated. We provide evidence that our method is superior in both binding affinity and compound diversity. Finally, we conclude with a discussion of how our method could be integrated into a compound design tool and explore some of the advantages and limitations that such a tool would provide.

Implicit Fingerprints from Collaborative Filtering. In their previous work, Srinivas et al.[24] investigated implicit fingerprinting models that extend the existing virtual screening mechanisms by incorporating collaborative filtering. Collaborative filtering algorithms are used for designing recommendation systems such as movie recommendation engines.[35,36] In general, collaborative filtering is a method for making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from other users (collaborating). When applied to the field of virtual screening, this approach relies on modeling predictions based on assays measuring the interactions between compounds and targets.[37,38] To intuit this, one can imagine building a recommendation system for matching movies to people. The direct approach might try to extract features specific to the person, like genre preferences and preferred actors, and features of the movie, like genre and runtime, to classify a match. This is the approach that is most similar to virtual screening where researchers directly featurize the compounds and targets based on their geometry and

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

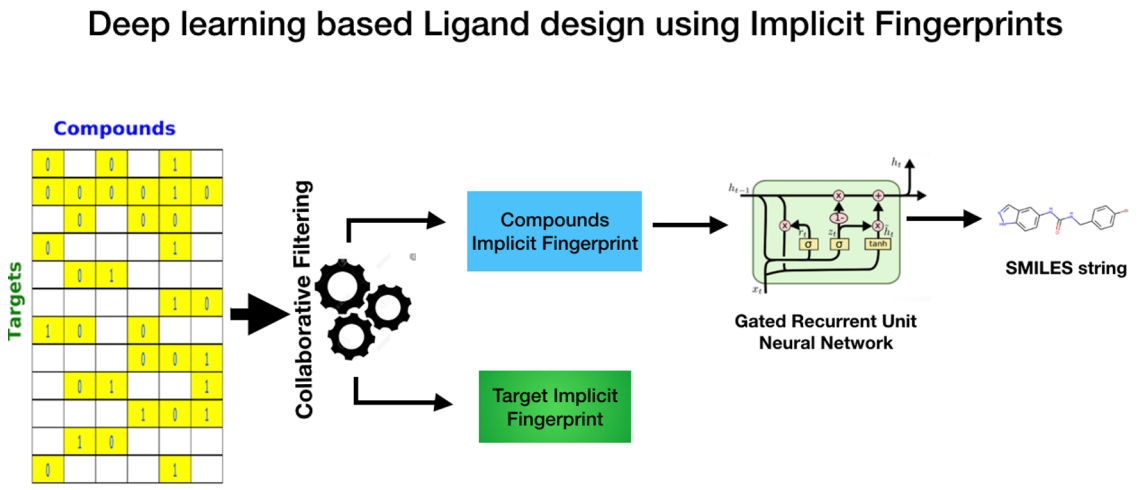


Figure 2. Collaborative filtering-based generative networks (CFGenNets): deep learning-based ligand design using implicit fingerprints from collaborative filtering architecture.

physicochemical properties. A more implicit approach groups users based on the movies they liked and groups movies based on the users that have seen them. Users and movies in similar groups could be implicitly found without attempting to featurize aspects of the users or movies directly.

In their previous work, Srinivas et al.[24] elucidated the performance of the collaborative filtering against the traditional approaches using evaluation criteria such as a 1% enrichment factor (EF1%)[39] for its ability to address specific properties of the early recognition problem specific to virtual screening, Boltzmann-enhanced discrimination of the receiver operating characteristic (BEDROC20),[40] and area under the curve (AUC) of the receiver operating characteristic.[41] The collaborative filtering algorithm was found, at that time, to consistently and significantly outperform all of the other methods using the evaluation criteria. Furthermore, the utility of the implicit continuous representation of the ligands obtained from collaborative filtering was illustrated in an example with cancer-related targets, as described in the next section.

Representation of Ligands in Implicit Fingerprint Space. To help intuit the inherent properties of the implicit latent space, we randomly selected two cancer-related targets from the ChEMBL23 database. We selected targets with ChEMBL23 IDs CHEMBL4899 and CHEMBL2150837 along with all of the ligands with assays for these targets, as shown in Figure 1. The 50-dimensional implicit fingerprints of the compounds are reduced into a two-dimensional space using stochastic neighbor embedding (t-SNE).[42] We visualize all compounds with available assays for the three selected cancer-related protein targets in the ChEMBL23 database. The compounds are color-coded as either having demonstrated binding affinities to the target or not, on the basis of their standardized concentration levels in the assays, where a decreasing concentration level indicates stronger binding affinity. For the t-SNE plots, the ideal result would be perfect clustering for each concentration level, which would indicate that the compounds cluster based on their binding affinity. Interestingly, the implicit ligand fingerprints in Figure 1 demonstrate a very clear separation between the compounds based on the concentration levels required to trigger binding affinities with the respective targets. This visual separation is striking for assays with excellent binding affinity (standard value below 100 nM), indicating that the implicit representation is excellent in its ability to capture properties of similar compounds using a Euclidean distance.

Neural Network Architecture. Our method to generate potentially novel ligands is composed of two steps: (1) we generate implicit ligand and protein fingerprints using collaborative filtering and (2) train a neural network to generate the SMILES string from the implicit representation (i.e., a decoder that can map to a conventional representation from the implicit space). The first step involves generating the implicit fingerprints using known assays by applying the collaborative filtering algorithm.[24] This step yields the implicit fingerprint representations for both ligands and protein targets, as described above. The implicit fingerprints are continuous vectors that represent a point in 50-dimensional space.

The implicit fingerprints of the ligands are then fed into a gated recurrent unit (GRU)[43] neural network to map the corresponding SMILES string encoding. The neural network is trained to minimize the error in reproducing the relevant SMILES string for each input implicit fingerprints of the ligands. The key aspect of the neural network is to learn the function to map the fixed-length continuous vector representation to the SMILES string. This architecture, for what can be described as collaborative filtering-based generative networks (CFGenNets), is illustrated in Figure 2. Additional details of the neural network design are discussed under the Methods section.

As with other methodologies utilizing generative deep learning algorithms,[26] the neural network should ensure that the points in the latent space decode to valid SMILES strings. To avoid the latent space from being sparse and resolve to large "dead-areas" (areas in the space that are never trained to decode from and therefore behave unpredictably), we performed input data augmentation. The data augmentation involved adding randomness to the input layer of the neural network (i.e., adding random perturbations to the implicit vector). The data augmentation incentivizes the decoder to more fully represent the areas in the implicit latent space of the ligands, such that they can successfully resolve to the corresponding SMILES string. The intuition is that adding noise to the encoded molecules forces the decoder to learn how to decode a wider variety of latent points and find more robust representations. This approach follows the intuitions made popular by the variational autoencoders (VAEs)[44] by Bowman et al. The VAEs, instead of decoding from a single point in the latent space, sample from a

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

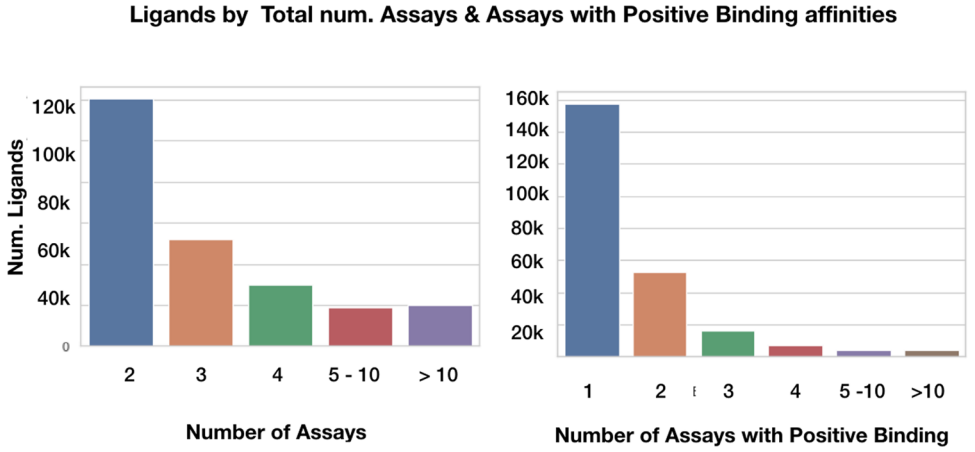


Figure 3. Data distribution: the first figure illustrates the number of molecules against the number of assays, binned at specified values or ranges. Close to 50% of the molecules have only two prior assays. The second figure illustrates the number of ligands against the number of known assays with positive affinities. 62% of the ligands with only one assay with positive binding affinity can be observed.

location centered around the mean value and with spread corresponding to the standard deviation before decoding. This ensures that a sample from anywhere in the area is treated similarly to the original input. Even so, there are differences between the VAE approach and ours. In our approach, the latent space if fixed from the collaborative filtering is not trainable like in the VAE. Importantly, this means that the sampling incentivizes the decoder to reconstruct similar SMILES strings from a given set of similar points. It does not incentivize the collaborative filtering algorithm to change its implicit representation.

The sequential nature of the output SMILES string required us to consider neural network architectures that are adept at handling such data. The application of neural network architectures such as recurrent neural networks and their enhanced variations such as gated recurrent neural networks (GRUs) for problems involving sequential data such as speech recognition and language translation have been very successful.[43,45−47] The GRU neural networks, with their innate abilities to learn long-term dependencies in sequences, are especially useful for handling SMILES strings.

Tools for Bioactivity Prediction. Bioactivity of a drug is of critical importance to highlight the applicability of generated ligands. SSnet[14] and Smina[48] were utilized to obtain a relative bioactivity of ligands toward various targets tested in this work.

SSnet. SSnet is a deep neural network-based framework that requires a protein target in pdb format and a ligand as SMILEs string to predict their bioactivity (probability for binding). SSnet utilizes a protein's fold information extracted as curvature and torsion patterns that hold compact information about potential ligand interaction. SSnet had outperformed state-of-the-art machine learning models like Atomnet,[11] 3D-CNN,[49] and GNN-CNN[12] and classical force field and knowledge-based methods employed by Autodock Vina[50] and Smina,[48] and various standard docking methods such as FRED,[51] Surflex,[52] HYBRID,[53] PLANTS,[54] etc., in identifying positive protein−ligand pairs (protein−ligand complex with high binding affinity). SSnet being pretrained has fast execution time (18 min for 1 million protein−ligand pairs) to curate high affinity ligands from a pool of large libraries such as ZINC database (more than 1 billion ligands).[55] We note that the resource efficiency of SSnet was quickly utilized for the prediction of potential drugs for Covid-19 that shows the biological relevance of the method.[56] SSnet is made available for public use on https://github.com/ekraka/SSnet.

Smina: Scoring and Minimization of Ligand Conformation. To perform virtual screening and docking, Smina[48] was used on a subset of ligands converted to 3D structures via openbabel.[57] The docking on DUD-E proteins was performed utilizing the reference protein structure along with a reference ligand provided by Mysinger et al.[58] The docking was performed on the center of a known ligand in a protein−ligand complex with a box size of 32 Å × 32 Å × 32 Å and an exhaustiveness of 36 on the default scoring function. The box size defines the space to consider in a protein for optimizing a ligand conformation resulting in a binding score. The exhaustiveness is an indication of the computation time for optimization. The exhaustiveness is required as Smina utilizes the Monte Carlo method for optimization.

RESULTS AND DISCUSSION

In this section, we present the details of the experiments conducted with their results. We begin with an exhaustive description of the data used for the experiments.

Data Set Description. Our method involves translating the implicit ligand fingerprints into its corresponding SMILES string. The implicit fingerprints, however, are derived from the ligand−target bioactivity data from the ChEMBL database (Version 23). The bioactivity data, keeping in line with previous studies,[59,60] was focused only on human targets. We restricted bioactivities to three types of binding affinities. This included IC_{50} half-maximal inhibitory concentration, EC_{50} maximal effective concentration, and inhibitory constant (k_i). Following the precedence with previous works,[24,59,60] we converted the data into the binary active−inactive using the following conversation thresholds: lesser than 100 nM for "actives" and greater than 1000 nM as "either weak binders or inactives". Furthermore, to be consistent with Srinivas et al.,[24] we retained only ligands that have at least two prior assays. This resulted in a bioactivity matrix of size 241 260 (ligands) by 2739 (targets). The bioactivity matrix was subjected to the collaborative filtering method, as described in Srinivas et al.[24] The resultant implicit fingerprints were then used as inputs to our deep learning model, with the goal to produce the respective canonical SMILES string as the output. Figure 3 illustrates the data distribution of the number of ligands against the known

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

number of prior assays and known number of prior assays with known positive affinities. As evident from the plots, close to 50% of the ligands have only two prior assays. Additionally, close to 62% of the ligands have only one prior assay with positive affinity. We also wish to note that the number of ligands (241k) used to model the deep learning model is comparable to previous works.[28]

Considering that our approach relies on the prior assay history to determine the implicit ligand fingerprints, having more numerous examples of prior assays for each ligand may also result in better quality implicit fingerprints. This statement is further evidenced by results from the next section: (1) the ability of the decoder to accurately translate implicit fingerprints into the corresponding SMILES and (2) the abilities of the ligands to yield more potentially novel ligands are both influenced by the number of available assays per ligand, as described next.

De Novo Generation of Molecules from Latent Space. In this section, we discuss the outcomes of our method in the context of 5000 randomly selected ligands from the data set for validation purposes. Additionally, we also present the outcomes of a scaffold analysis from the potentially novel ligands generated from ligands with known affinities to cancer targets from the ChEMBL23 database. To further analyze the practical applicability of our approach, the resulting ligands, specifically from approved cancer-related drugs, were further evaluated for their viability to be valid compounds with enhanced biological activities. The complete list of ligands is made available as a part of the Supporting Information (Section 0.4).

Our method samples around the implicit latent space of the known ligands or "anchor−ligands" to generate (potentially novel) compounds. In our testing, we randomly sampled 100 points across the 50 dimensions in the implicit space around our anchor−ligands. Each point was then processed through our neural network to obtain the corresponding SMILES string. The SMILES string was then validated using the RDKIT library. This process is discussed in more detail in the Methods section.

We ran the aforementioned sampling and validation exercise on the 5000 ligands (henceforth referred to as anchor−ligands). A total of 4632 of the 5000 (92.64%) anchor−ligands resolved to at least one valid ligand, although not all resolved ligands were (potentially) novel. As mentioned earlier, 100 points are randomly sampled for each anchor−ligand. Depending on the information encoded in the continuous implicit vector space, multiple points around a given anchor−ligand may resolve to the same ligand. Only those ligands that are generated at least twice, and can be resolved to a valid compound using the RDKIT library, are considered to be "valid" generated ligands. The frequency constraint of "at least twice" is enforced to help ensure that the generated ligand is not generated spuriously.

Novelty among Generated Ligands. The practical applicability of the generative deep learning methods is typically measured by the ability of the methods to generate novel ligands with desirable properties. However, despite the plethora of works in the space, the concept of novelty is loosely defined. Several popular recent works[27,61,62] just validate if the generated ligand was already present in the training data set. If not found, the generated ligands are deemed as novel. Alternately, Popova et al.[63] assessed novelty by checking for the presence of the generated ligands in the training set of 1.5 million ligands from ChEMBL21. Additionally, they also searched for the presence of the generated ligand in the ZINC database for 320 million synthetically accessible druglike molecules. It is to be noted that a difference of even a single atom was deemed as a sufficient condition to term a generated ligand as being novel. With the precedence in the aforementioned prior works serving as a baseline, we adopted a series of multifold conditions to assess the potential of the generated ligands to be deemed novel:

- Is the generated ligand already present in the training data set? A total of 2917 generated from the set of 5k anchor−ligands were not present in the training set.
- Is there any other known ligand in the 1+B ZINC database with a similarity threshold of 0.85 with the generated ligands. We further ran the 2917 ligands against the ZINC database to look for the most similar known ligand from 1.3 billion compounds from the ZINC database, with a similarity threshold of 0.85. The similarity between compounds was measured by the Tanimoto coefficient (TC), which measures a distance between fingerprints resulting in a score ranging from [0,1] (0 corresponds to least similar and 1 to exactly same).[64] We obtained the TC based on Morgan fingerprints[65] of 512 bits vector. This resulted in a total of 2759 ligands from the previous step.
- Among the ligands from the ZINC database with similarity <0.85, are there differences in the scaffolds and/or the number of functional groups[66] between the anchor and generated ligands. To further verify that the generated ligands were meaningfully different from their closest hits from the ZINC database, we further compared the scaffolds and functional groups between the generated ligands and their closest hit from the ZINC database. Of the 2759 ligands from the previous step, only 322 (12%) of ligands had the same functional groups and only 618 ligands (22%) had similar scaffolds as their closest hit from the ZINC database. These numbers are summarized in Table 1 for easier readability. Additionally, Figure 4 plots the distribution of the ligands over the differences in the number of functional groups between each pair of potentially novel ligands and their closest hit from the ZINC database. It is seen that 67% (1831 ligands) of the ligands generated had a difference of at least 2 or more functional groups with their closest hit. Figure 4 also illustrates the TC similarity scores between the potentially novel ligands and their closest ZINC database hits. It is observed that 20% of the potentially novel ligands have a similarity of 0.5 or less. The entire list of the potentially novel ligands along with their closest hit from the ZINC database and the respective scaffold and functional groups is also made available in the Supporting Information (Section 0.4).

Table 1. Novelty Analysis: Summary by Numbers

number of anchor−ligands	5000
number of anchor−ligands yielding at least one valid ligand	4632
number of generated ligands not present in the training data	2917
number of generated ligands with a difference of 2 or more functional groups	1831
number of anchor−ligands yielding at least one ligand outside training data	1332

We wish to note that the references of "novel ligands" in the rest of the sections should be read as being potentially novel and in conjunction with the aforementioned set of conditions. It was also observed that sampling around certain anchor−ligands resulted in numerous potentially novel ligands being generated,

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

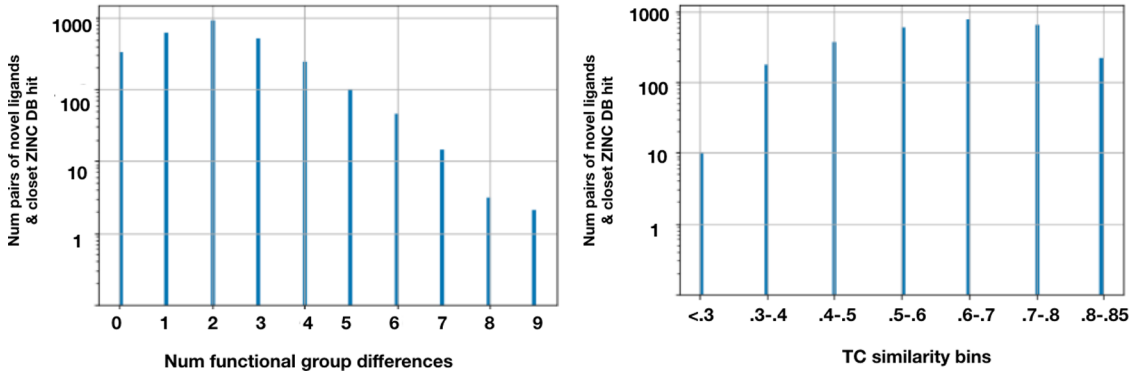


Figure 4. Distribution of the number of potentially novel ligands and their closest hit pairs vs differences in the functional groups and TC similarity ranges. The figure demonstrates that a degree of novelty could be associated with the generated ligands when compared with the 1.3 billion known ligands from ZINC DB. The differences in the number of functional groups between the anchor and generated ligands range from 0 to 9, with at least 67% ligands with two or more differing functional groups. The TC similarity bins help gauge the distribution of the TC similarities between the pairs. It is seen that the lower the similarity, the more likely it is for the pair having varying functional groups.

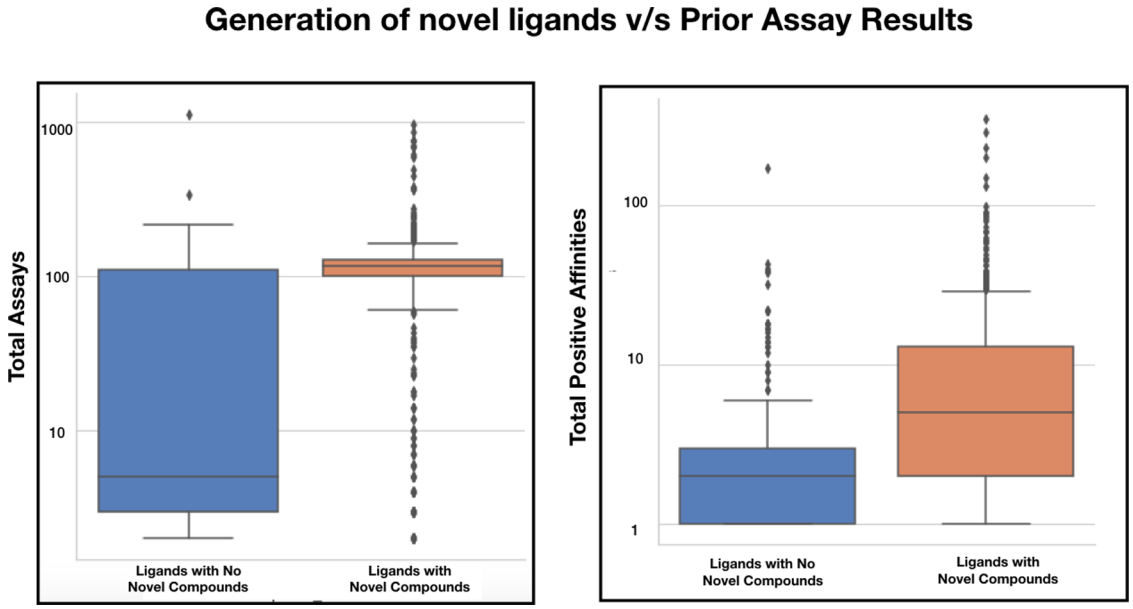


Figure 5. Correlation of the ability to generate potentially novel ligands with prior assays: box plots show the co-relation between the two sets of anchor−ligands—one set from 1 or more potentially novel ligands were generated and the second set that yielded no ligands when sampled in the implicit fingerprint latent space. The first figure visualizes the total number of known assays that exists for each set. The second box plot visualizes the total number of positive binding affinities already recorded for each assay.

while sampling around other anchors did not yield any novel ligands. To further investigate this phenomenon, we analyzed the abilities of the associated anchor−ligands to yield potentially novel compounds by grouping according to known prior assays. Figure 5 (left) illustrates the relationship between the presence of assays of the anchor−ligands and their ability to generate potentially novel ligands. As can be seen, anchors that generated novel ligands tended to have a greater number of known assays. In Figure 5 (right), we can also observe that this relationship holds for the number of anchor−ligands with positive affinities. Anchors with more numerous positive assays also tended to generate potentially novel ligands. This observation is perhaps not surprising considering that the implicit fingerprints are derived from the known assays. This provides evidence that the implicit fingerprints for anchor−ligands encode more meaningful information when there are more numerous assays. One potential explanation for this is that the implicit representation can encode many desired properties that are difficult to measure as the number of assays increases, thereby giving a better blueprint for the decoder to generate potentially novel ligands. However, because the implicit representation does not explicitly model chemical or experimental parameters, this hypothesis must be investigated through the observation of known ligand properties, as discussed next.

Physical Properties of Generated and Anchor−Ligands. To explore the similarity of potentially novel and anchor−ligands, we evaluated the properties of the compounds using a number of scoring measures. More specifically, we used the quantitative estimation of druglikeness (QED), *n*-octanol water partition coefficient (log *P*), and synthetic accessibility score (SAS) and used the number of benzene rings as an indicator of the chemical complexity. Our approach to ascertain the similarities in the physical properties between the anchor−ligands and their corresponding novel ligands considered the following approaches:

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

- Compare the distribution of the populations of property values of the anchor−ligands with the distribution of the generated ligands. This comparison is the standard practice when evaluating the quality of generated ligands.[26,63]
- Additionally, to investigate the similarity of generated ligands with their respective anchor−ligand, we evaluated the magnitude of the difference in the values between the ligands for each of the four aforementioned properties. A residual value, which is the difference between the property values, is calculated for each pair of anchor−ligand and its corresponding generated ligand. A mean residual is then obtained for each anchor−ligand, as described in eq 1. The magnitude of the mean residual value was used as a method to determine the deviation of the properties between anchors and their generated ligands

$$R_{\mathrm{m}} = \frac{\sum_{n=1}^{N} \sqrt{(p_a - p_n)^2}}{N} \quad (1)$$

where R_{m} is the mean residual property value for each anchor−ligand, N is the number of unique potentially novel ligands generated for each anchor−ligand, p_{a} is the property value (QED, log P, SAS, and NumRings) for the anchor−ligand, and p_n is the property value (QED, log P, SAS, and NumRings) for the nth novel ligand for the corresponding anchor−ligand.

The QED ranges between 0 and 1. The ligands with higher values indicate that the molecule is more druglike. Additionally, the method also claims to capture the abstract notion of esthetics in medicinal chemistry.[67] We leveraged the python-based RDKit library to determine the QED scores of the generated novel compounds. As illustrated in Table 2, the average QED score of the novel ligands was found to be 0.57. Figure 6A(i) illustrates the comparison of the distributions of the QED scores from the potentially novel ligands with their anchors. It can be observed that the two distributions are very similar. A 2-sample Student t test statistic of 0.99 with a p-value of 0.35 also confirms that there exists no statistical difference between the two distributions. Table 2 tabulates the mean, standard deviations, and t-test scores of all of the properties calculated as a part of our experiments. Additionally, Figure 6A(ii) illustrates the similarities of the QED scores between the anchor−ligands and their respective generated ligands by measuring the mean residual value, as described in eq 1. It is evident from the plot that a large number of mean residuals are less than 0.1 units. This indicates that the QED scores of close to 80% of the anchor−ligands are within 0.1 units of their generated novel ligands and close to 96% of the anchor−ligands have QED scores within 0.2 units of their generated ligands.

Table 2. Properties of Anchor and Potentially Novel Ligands

		anchor−ligands	potentially novel ligands	t-test
QED	mean	0.69	0.57	t-stat = 0.99
	SD	0.20	0.22	p-value = 0.35
log P	mean	3.41	3.43	t-stat = 0.33
	SD	1.69	1.95	p-value = 0.74
benzene rings	mean	3.45	3.12	MannWhitt stat = 2.1e7
	SD	1.24	1.33	p-value = 4.76e-18
SAS scores	mean	2.67	3.18	t-stat = 21.62
	SD	0.55	0.85	p-value = 6.7e-99

The water−octanal partition coefficient (log P) was another property used to quantify the physical properties of the potentially novel ligands. Log P describes the propensity of ligands to dissolve in an immiscible biphasic system of lipids (fats, oils, organic solvents) and water.[68] A negative value for log P means the ligand has a higher affinity for the aqueous phase (hydrophilic); when log P = 0, the ligand is equally partitioned between the lipid and aqueous phases; a positive value for log P denotes a higher concentration in the lipid phase (lipophilic). The potentially novel ligands tended to be more lipophilic with a mean log P-value of 3.43 with a standard deviation of 1.94. Figure 6B(i) illustrates the distributions of log P scores between the novel and anchor−ligands. The two distributions appear to be visually similar and a 2-sample Student t test score of 0.33 with p-value = 0.74 also confirms the same. Additionally, Figure 6B(ii) illustrates the similarities of the log P scores between the anchor−ligands and their respective generated ligands by measuring the mean residual score, as described in eq 1. It is observed that close to 87% of the anchor−ligands have their log P scores within 1 unit of their generated ligands.

The synthetic accessibility score (SAS), a method that is able to characterize molecule synthetic accessibility as a score between 1 (easy to make) and 10 (very difficult to make),[69] was another property that was evaluated for the potentially novel drugs generated by our method. The mean score was found to be at 3.17 with a standard deviation of 0.85. While the SAS scores between anchors and their novel ligands appear to be similar visually (Figure 6C(i)), a t-statistic score of 21.62 with p-value = 6.7e-99 indicates that the two distributions are statistically different. Nevertheless, a mean score of 3.17 of the potentially novel ligands indicates that the potentially novel ligands are synthesizable to generate valid drugs. Figure 6C(ii) further compares the individual SAS scores between the generated ligands and their respective anchor−ligands. It is observed that 87.3% of anchor−ligands have SAS scores within 1 unit of the generated ligands. This indicates that an overwhelming majority of the anchor−ligands share similar SAS scores with their generated novel counterparts. Additionally, the number of benzene rings was evaluated as a measure of the chemical complexity of the potentially novel ligands. Figure 6D(i) demonstrates that the complexities of the potentially novel drugs are comparable to those of their corresponding anchor−ligands. Figure 6D(ii) compares the similarities in the number of benzene rings between the anchor−ligands with their respective generated novel ligands. From the figure, it is evident that the distribution of the number of rings does not follow a normal distribution. For this reason, we conducted the Mann−Whitney U-nonparametric test[70] to compare the two distributions. The test yielded a statistically significant difference in the two distributions. However, it was observed that approximately 83% of the anchor−ligands had the exact same number of benzene rings as their respectively generated novel ligands.

We further evaluated Lipinski's rule of 5 (LR5) for all of the generated ligands.[71] The LR5 describes critical properties of a ligand in the human body such as absorption, distribution, metabolism, and excretion. The rule states that a ligand to be effective for therapeutics should have less than 5 hydrogen bond donors, less than 10 hydrogen bond acceptors, a molecular mass of less than 500 Da, and the log P less than 5. The LR5 score was computed for all generated ligands based on Yao et al.[72] We observed that 68% of generated ligands completely satisfies the LR5 rule and 22% of generated potentially novel ligands satisfy at least 3 of the 4 rules. This is further illustrated in Figure 7. The

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

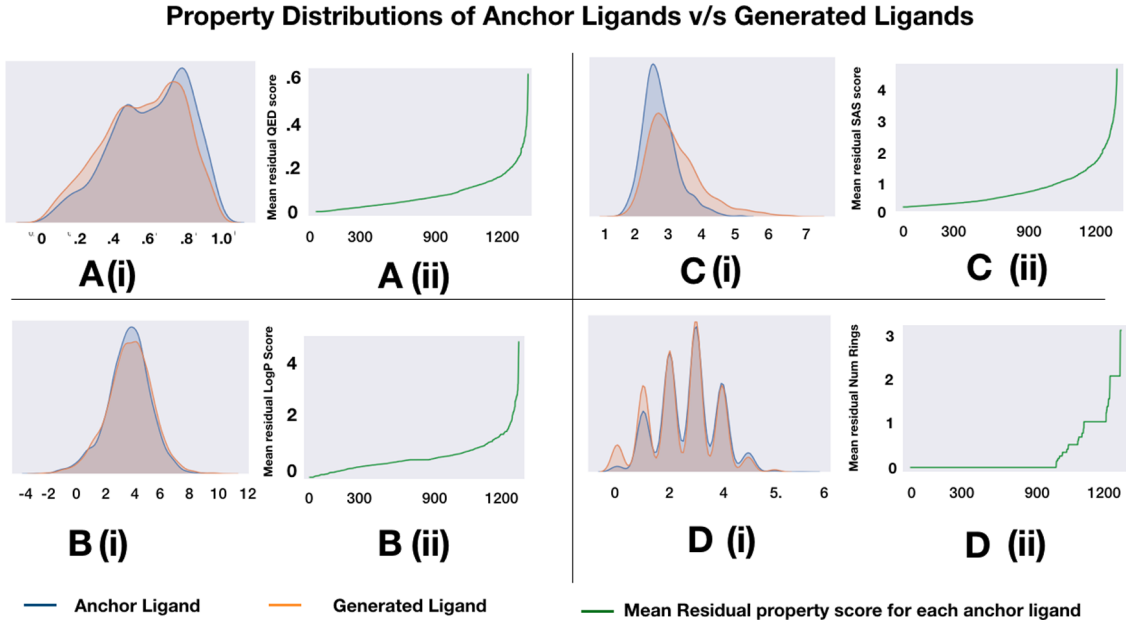


Figure 6. Property distribution between anchor−ligands and generated ligands: (A) quantitative estimate of druglikeness (QED), (B) partition coefficient (log *P*) (C) synthetic accessibility score (SAS), and (D) number of benzene rings. The figure demonstrates that the property distributions of the anchor−ligands are similar to the potentially novel ligands generated from the corresponding anchors across all four properties.

percentage of matches to Lipinski's rule of 5 signifies that the generated ligands have properties to be an effective drug.

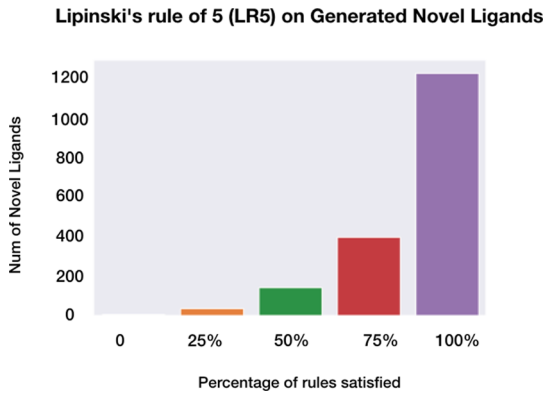


Figure 7. Lipinski's rule of 5 valuated on the potentially novel ligands generated from implicit fingerprints. The figure demonstrates that 80% of the 1831 potentially novel ligands satisfy 3 or more rules, signifying that the generated ligands have properties to be an effective drug.

Now that it is established that the potentially novel and anchor−ligands are likely to have similar and comparable physical properties, we turn our attention to answering whether the novel ligands are also likely to similarly bind to known targets.

Binding Affinity Predictions of the Potentially Novel Ligands. The biological activities of the potentially novel ligands were evaluated by inferring their predicted binding affinities with 102 DUD-E protein targets.[58] The DUD-E targets consist of a variety of proteins exhibiting different mechanisms of protein−ligand interactions. The relationship of bioactivities within the anchor−ligand and generated ligands over the DUD-E targets will highlight the versatility of our model. Thus, we used the anchor−ligands to test their binding affinities with the DUD-E targets.

To validate the similarities of the binding affinity properties of the novel ligands with their respective anchors, the binding affinity scores were determined from SSnet and Smina for the anchor−ligands with the 102 DUD-E proteins. Each ligand (anchor and novel ligands) yielded a distribution of binding affinity scores against each target from the set of 102 DUD-E protein targets. The similarities in the binding affinities of the novel and their respective anchor−ligands were evaluated by comparing the aforementioned binding affinity distributions. Of the total 1332 unique combinations of novel and respective anchor−ligands, approximately 84% demonstrated similar binding affinity behaviors. The similarity score or the measure of intersection over union (IoU)[73] in this exercise is calculated by evaluating the proportion of DUD-E targets to which both the ligands demonstrate binding or lack of binding. An SSnet score of 0.5 or less is considered lack of binding and a score greater than 0.5 as binding. Figure 8 illustrates this for 1332 unique pairs of novel ligands and their anchor−ligands. Each data point on the *x*-axis in Figure 8 represents a unique anchor−novel ligand combination. The *y*-axis represents the intersection over the union score calculated between the two distributions of binding affinity scores, the first distribution being binding affinity indicator of anchor−ligand with 102 DUD-E proteins and the second distribution being the binding affinity indicator of the novel ligand with 102 DUD-E proteins. The figure further illustrates that a large majority of the anchor−ligand pairs exhibit similar binding affinities. A similar observation was made for Smina, as shown in Figure S1, by considering ligand similarity based on −7.5 kcal/mol as a threshold for plotting IoU.

While there is a high coherence of the scores obtained from SSnet, we further evaluated the similarities between the anchor and generated ligands. We calculated the Tanimoto coefficient-based similarity scores between each pair of anchor and generated ligands. Figure 9 plots the IoU scores and the TC similarity scores for each pair. It is evident from the plot that there is no correlation between the IoU scores and the TC similarities. Despite the lack of correlation, the high coherence in

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

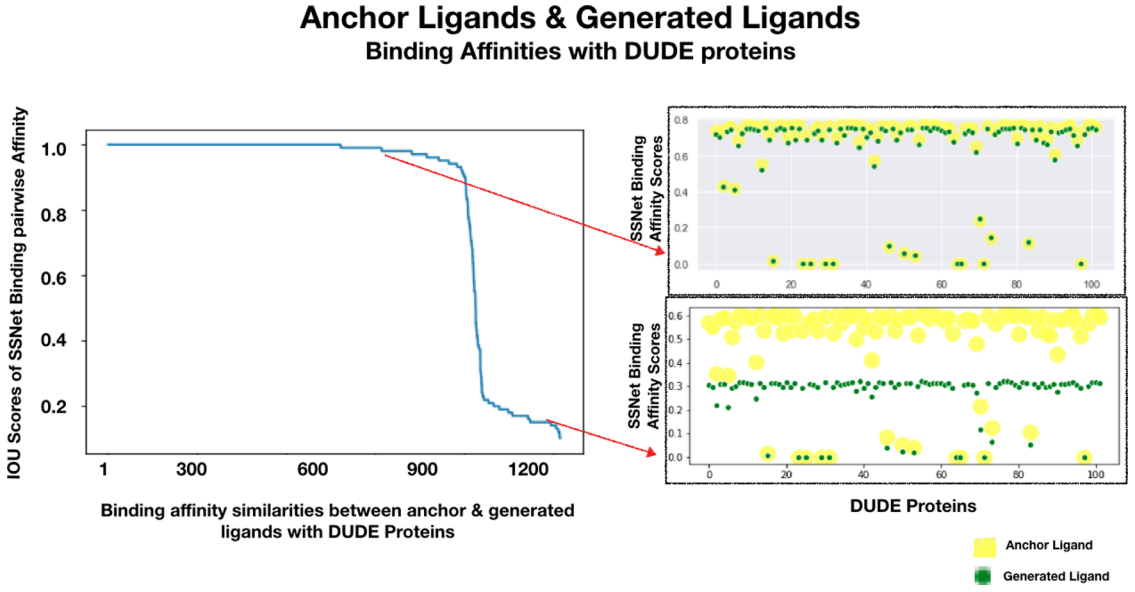


Figure 8. Pairwise binding affinity scores: the plot illustrates the similarities in the bioactivity between each pair of anchor−ligands and their corresponding generated ligands to the 102 DUD-E protein targets. The blue line in the line plot to the left demonstrates a strong co-relation between the binding affinities for most pairs with the DUD-E targets. This is due to very high IoU scores for 84% of anchor and generated ligand pairs. The scatter plots to the right illustrate two sample pairs, with the top right plot representing a pair with very similar affinity scores and the bottom right plot illustrating a pair where the affinities differ between the anchor and generated ligand.

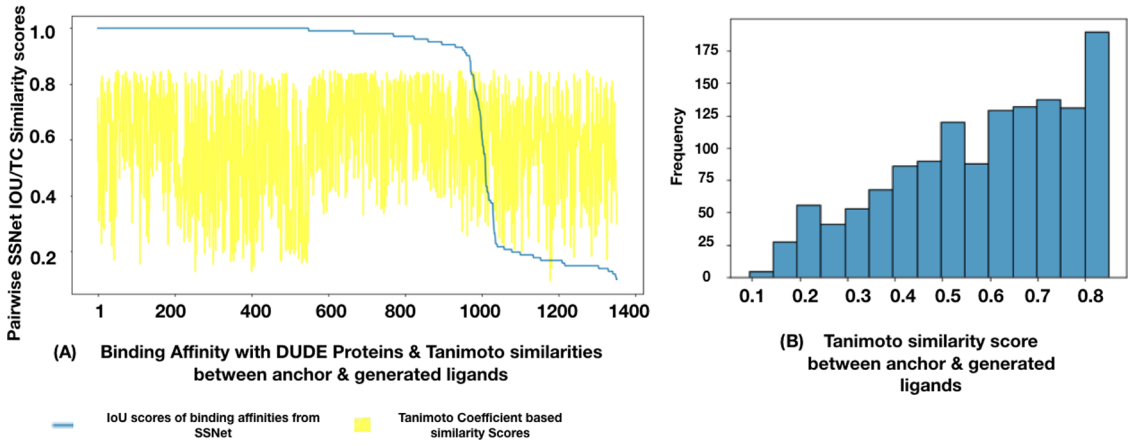


Figure 9. (A) Comparison of the IoU scores and the Tanimoto coefficient scores between the anchor and generated ligands. The figure illustrates that there is no strong correlation between the anchor and generated ligands in terms of TC similarities. (B) Histogram of TC scores across all of the pairs: illustrates the distribution of the similarity scores between the pairs.

the binding affinities could be explained by the scaffold similarities between the anchor and generated ligands. This is further evidenced by the scaffold analysis using the pseudo-Hilbert curve, as described in the subsequent section.

The analysis on QED, log *P*, and SAS provided an intuitive relationship of generated ligands and druglikeness. However, for a drug to be effective for the specific target and show selectivity among other targets, the scaffold should be preserved (core structure of a molecule[74−76]). To analyze if the generated ligands have similar scaffolds, we sorted all of the anchor−ligands by Tanimoto coefficient (TC). The sorting was performed by recursively finding the next most similar ligand from the anchor−ligands starting from a random anchor−ligand. The sorted list was then mapped to a pseudo-Hilbert space-filling curve. The pseudo-Hilbert curve was used to observe molecular scaffolds directly from the map as pseudo-Hilbert curve preserves the spatial proximity of the sorted list. The pseudo-Hilbert map for the generated ligands was made similarly. Each anchor−ligands were repeated to the same number of generated ligands to match one-to-one when comparing the pseudo-Hilbert curve for generated ligands and anchor−ligands. Figure 10a,b shows the pseudo-Hilbert map for anchor−ligands and generated ligands, respectively. The pseudo-Hilbert map is colored based on the SSnet scores obtained by docking the ligands with the DUD-E targets with PDB ID 1B9V and 3KBA. We observe that the clusters are majorly retained for the generated ligands when compared to those for the anchor−ligands. This is further highlighted in Figure 10c, which shows the difference in SSnet scores for generated and anchor−ligands. The map is mostly blue, which represents a mere difference of SSnet scores in generated and anchor−ligands of less than 0.1. The results highlight that the novel molecules generated preserve the scaffold that is essential in protein−ligand binding.

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

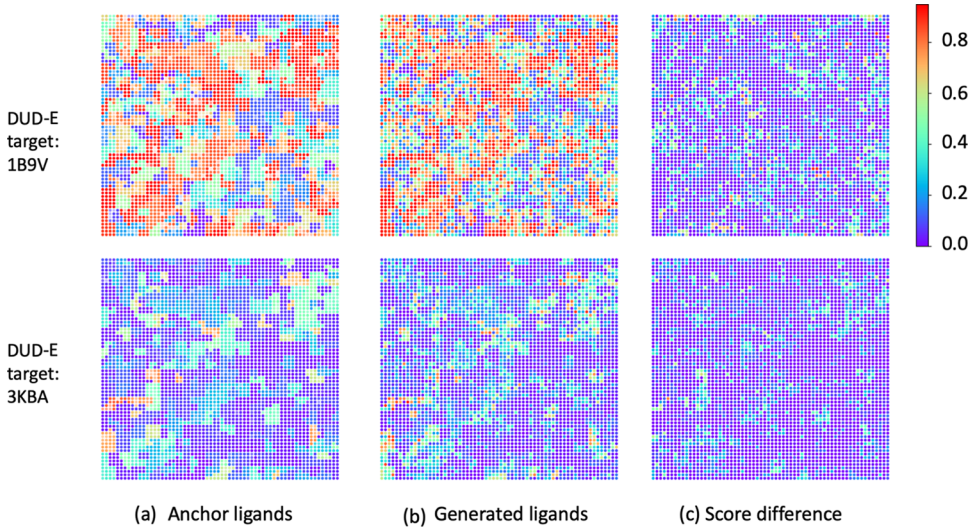


Figure 10. Scaffold analysis. A pseudo-Hilbert curve is plotted for anchor−ligands and generated ligands. The color denotes SSnet scores. Similarity between the anchor and generated pseudo-Hilbert curves and the low difference among them signifies that our method retains scaffolds from the anchor−ligands while also predicting similar bioactivities.

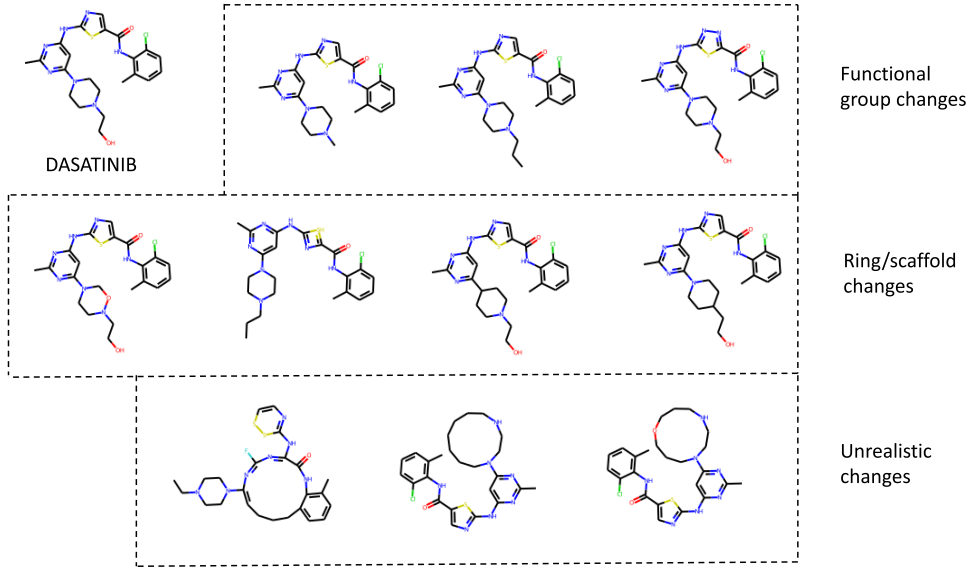


Figure 11. Novel ligands generated around the known cancer drug, DATASINIB: it is observed that the generated ligands have different functional groups and scaffolds. However, it is important to note that some unrealistic compounds are generated. Section 0.2.1 (Supporting Information) enumerates information about the novelty and the binding affinities exhibited by these generated ligands.

Comparison with FDA-Approved Drugs. To further hone in on the practical applicability of our methods and the potentially novel drugs generated, we conducted analysis on novel drugs generated on known cancer-related ligands. For this exercise, we shortlisted 10 drugs approved for treating various forms of cancer also available in the ChEMBL23 database. We present a detailed analysis of the potentially novel ligands generated around a known cancer drug, DASATINIB. Sampling around the implicit fingerprint space of this anchor−ligand yielded 10 novel ligands. Figure 11 illustrates the 10 novel ligands. We observed that new functional groups and scaffolds are generated. It is important to note that 3 of the 10 compounds generated seems to be unrealistic for drug discovery purposes in oncology. Further, screening through the ChEMBL23 database, no subsequence with similar rings as unrealistic labeled compounds in Figure 11 was observed.

The novelty of the compounds was tested from the ChEMBL23 data set (1.4 million compounds) and the ZINC data set (1.3 billion compounds). Across the 10 novel compounds, the maximum similarity score was 0.88 for ligands in the ChEMBL23 data set and 0.92 for ligands in the ZINC data set. Table S3 shows the largest TC obtained for each novel compound. Interestingly, in this particular case, we observe that the scaffold for the anchor−ligand is retained in most of the generated ligands. The results are in line with the scaffold analysis performed for the DUD-E protein targets provided in the previous section (Figure 10). Retention of the scaffold is crucial for ligand binding as the protein pocket, in general, has confined space for docking. The scaffold provides both size and imperative interactions such as hydrogen bonding, π interactions, etc., that contributes to the stability of the protein−ligand complex.

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

To test the bioactivities for the novel ligands generated, we sorted nine known targets for the anchor−ligand, the details of which are provided in Table S2. We conducted a docking method Smina[48] and a deep neural network-based model SSnet[14] for bioactivity score prediction. Figure 12 shows the

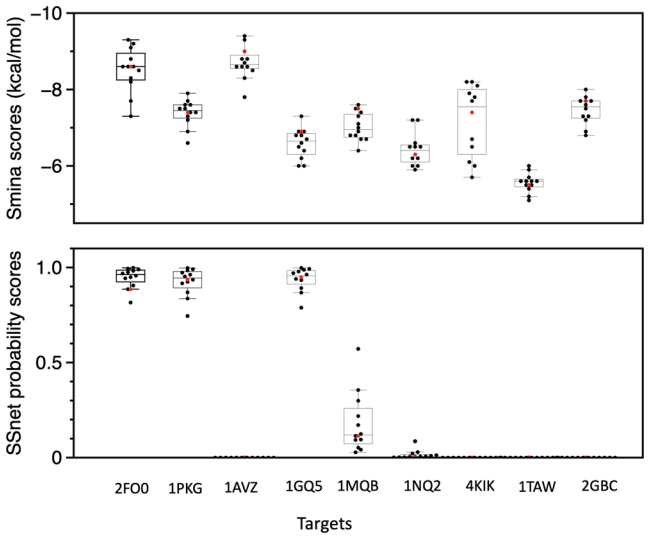


Figure 12. Sample test on various active/inactive targets for anchor−ligands. The first five targets 2FO0, 1PKG, 1AVZ, 1GQ5, and 1MQB are active and the rest are inactive. The red color denotes the anchor−ligands and the black color denotes generated ligands.

results obtained by the two methods. The first five targets are labeled active and the remaining four as inactive for the anchor−ligand used in the ChEMBL data set. We observe that the generated ligands have similar Smina scores as the anchor−ligands. A similar behavior is observed when comparing the SSnet scores for anchor and generated ligands. It is important to note that both Smina and SSnet are sensitive to the ligand and their complex interaction with a protein target. Many factors such as functional group, size of the molecule, molecular weight, etc., govern the bioactivity. The fact that all of the 10 novel generated ligands have similar bioactivities provides evidence that our ligand generation method produces ligands with similar binding characteristics to the anchor−ligand.

We further compared the bioactivities of six FDA-approved drugs and their corresponding generated ligands from the implicit fingerprint and latent space generated from the variational autoencoder (VAE) work from Gómez-Bombarelli et al.,[26] respectively. Each of the six ligands was docked toward its original intended target, the details of which are provided in Table S1. We observe a high similarity in predicted bioactivities for implicit fingerprints compared to the latent space generated from VAE for both the Smina and SSnet scores shown in Figure 13 (Tables S5−S10). A visual inspection of the compounds was generated from our method and the latent space from VAE. Gómez-Bombarelli et al.[27] show that both the scaffolds of the original anchor−ligand (Figures S2−S7) and generated ligands are similar. However, bioactivity is sensitive to small changes in the chemical structure such as a functional group. Our method is perceptive toward functional groups due to the way collaborative fingerprints were modeled, i.e., by considering the bioactivities.

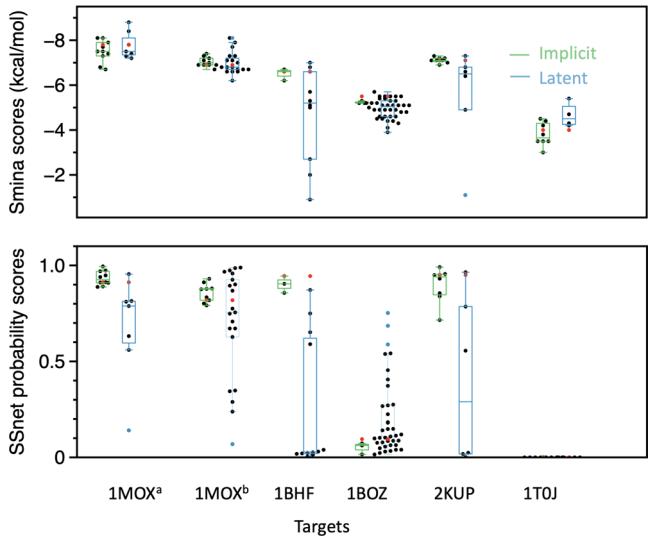


Figure 13. Comparison of implicit and latent fingerprints on FDA-approved drugs and their corresponding targets. The red color denotes the anchor−ligand and the black color denotes generated ligands. The latent label and implicit label show the binding affinities for generated ligands from the method developed by Gómez-Bombarelli et al.[27] (in blue) and our method (in green).

METHODS

The recent years have seen numerous deep learning-based generative models for de novo drug generation. The common theme in these techniques is to provide a deep learning model with anchor−ligands to produce novel ligands with similar properties. Often, the canonical SMILES notation of the ligand or a graphical-based fingerprint is used as the input to these deep learning models. This representation is then translated into a continuous vector representation(s) of the input ligand, whereby the intermediate layers in the deep learning model are slightly perturbed (i.e., with additive noise) to produce novel molecules. Many previous works exist, with the main distinguishing characteristic among the works being the architecture of the deep learning model and classification task employed for training. Some popular methods have been recurrent neural networks,[28] variational autoencoders (VAEs),[29] generative adversarial networks (GANs),[30,31] and graph-based neural networks.[32,33] A survey of recent work is available from Chen et al.[77]

Our approach stands apart in that we use the implicit ligand fingerprints obtained from the prior assay information (collaborative filtering) as inputs to a deep learning model, with the objective of producing the corresponding canonical SMILES representation as the output. This implicit representation can have a number of advantages because it is based solely on the observed behavior of the compound, rather than inherent measures of physical properties. Thus, formulating a decoding procedure from this implicit representation may have distinct advantages over previous methods. The implicit fingerprint, because it is a continuous vector of fixed length (50), also lends itself well to statistical sampling with simple procedures. We employ data augmentation of the input vector by employing a vector of mean μ and another vector of standard deviation, σ. The input vector (implicit fingerprint vector) serves as the vector of means, which is then added to another vector, which is a random normal distribution centered at 0 with standard deviation σ, to yield a statistically sampled point around the

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

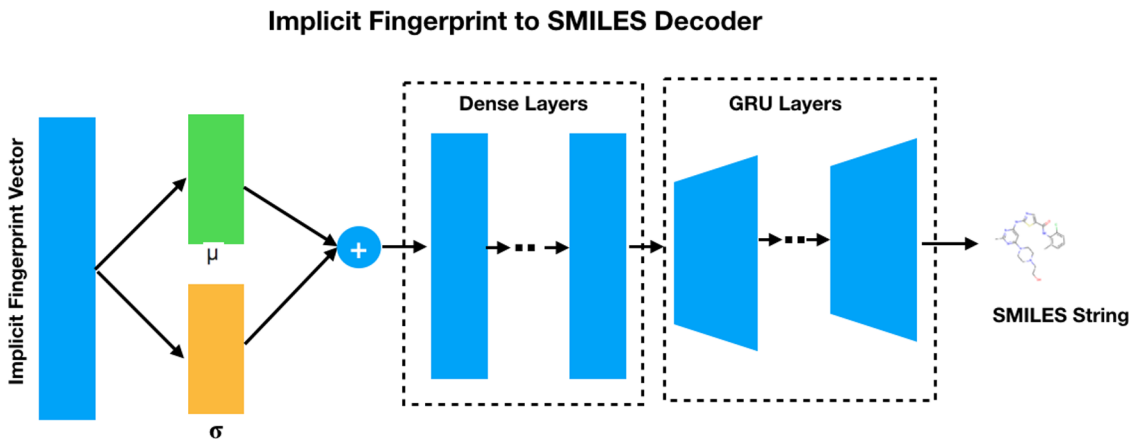


Figure 14. Implicit fingerprints to SMILES decoder using CFGenNets: the deep learning network learns ligand representations by employing the data augmentation technique at the input layer. The continuous representation obtained is then fed into a series of dense layers followed by a gated recurrent unit neural network to obtain the corresponding SMILES string.

implicit fingerprint. The stochastic sampling process ensures that the actual vector will vary on every single iteration due to sampling while keeping the mean and standard deviations the same. Intuitively, the mean vector controls where the implicit fingerprint of a ligand is centered around, while the standard deviation controls the "area", how much from the mean the encoding can vary. The decoder hence learns that not only is a single point in latent space referring to the ligand but all nearby points refer to the same ligand. The decoder is exposed to a range of variations of the encoding of the same input during training. This process is illustrated in the decoder architecture, as shown in Figure 14. The approach adopted here is similar to the data augmentation employed with variational autoencoders.[78]

To generate the SMILES string from the implicit fingerprint, we are motivated to use recurrent neural networks (RNNs) because of their success in modeling sequential data such as natural language. The SMILES strings lend themselves well to this model considering the sequential nature of the notation. Each unit in the RNN attempts to capture state information of the sequence by transforming all of the elements that appeared before it. It does so by encapsulating this information in a hidden state vector, which is passed from one unit back into itself, recurrently. The hidden state h^t of the RNNs can be represented as

$$h^t = \tanh(x^t w^{xh} + w^{hh} h^{t-1})$$

where x^t represents the input at time step t, w^{xh} represents the weight from the input node to the hidden node, w^{hh} represents the weight on the feedback loop from the hidden node to itself, and h^{t-1} represents the previous hidden state. As evident from the equation, the hidden states from the earlier time steps get diluted over long sequences. This problem gets compounded with SMILES considering the long-term dependencies (such as matching brackets, etc.) that need to be maintained to resolve to a valid chemical compound. The gated recurrent neural network attempts to address this problem by introducing two gates called the "update" gate and a "reset" gate along with a memory, which governs how much of the previous state is retained. Each of these units (update gates, reset gates, and memory) has its own trainable weights. The update gate at each unit decides the amount of new information to be added to the hidden states. The reset gate determines the past information to be forgotten or retained at each unit

$$r = \sigma(x^t w^{xr} + h^{t-1} w^{hr})$$

where σ represents a logistic or sigmoid function. These sigmoid values of the reset gate range from 0 to 1 and determine how much of the previous hidden node value is retained. A value of r=0 implies that none of the previous node value is retained and an r=1 ensures that the entirety of the previous node is retained. This memory, m, can be signified by the following equation

$$m = \tanh(x^t w^{xm} + (r \odot h^{t-1}) w^{hm})$$

where $\odot$ represents the Hadamard (or element-wise) multiplication of two vectors. Additionally, the update gate is governed by the following equation

$$u = \sigma(x^t w^{xu} + h^{t-1} w^{hu})$$

The update gate, with values ranging from 0 to 1, determines if the new hidden state should use the previous value or the new value. Tying all of these together, the hidden state is governed by the equation

$$h^t = u \odot m + (1 - u) \odot h^{t-1}$$

Intuitively, the GRUs are better suited than the RNNS to our problem considering the long-term dependencies between symbols that must be maintained in the SMILES string. The ring structures, for example, are represented by matching numeric symbols typically separated by two or more atoms within the SMILES string. The neural network should be able to remember these long-term dependencies for effectively decoding to a valid SMILES string. Figure 14 illustrates multiple GRU layers that make up the decoder to effectively map to the SMILES code. These "layers" shown in the figure are visualized compactly, that is, they are actually two stacked sequences of GRU nodes.

Neural Network Design. We performed extensive training and validation of the vanilla RNN and GRU models with the continuous implicit fingerprint vector as the input and the one hot-encoded SMILES string as the output. We measured the outcome of the models by evaluating the categorical cross-entropy loss and the accuracy. As a part of training, we explored a variety of architecture options with respect to the depth and the width of the deep learning models. We also trained with different compositions of the training sets based on ligands with varying counts of past assay data. Across all training iterations, we

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

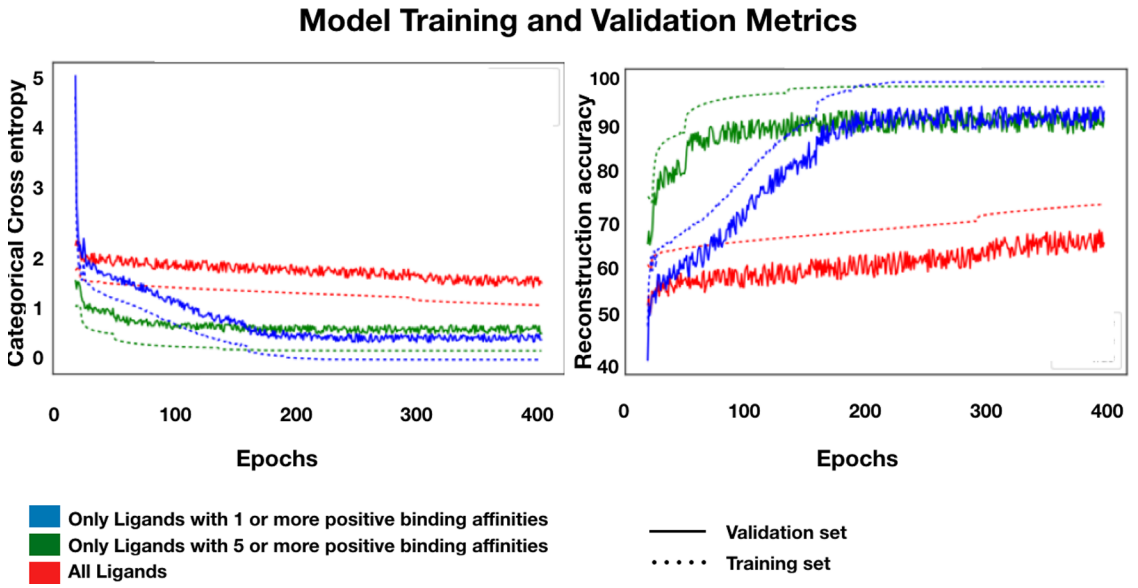


Figure 15. Training and validation losses of the neural network: training and validation losses across multiple runs of the neural network.

noticed that the GRU-based model performed better with lower cross-entropy and higher accuracy. We also noticed that the training loss converged faster compared to the validation loss. Our architecture comprised a series of dense layers, which consume the 50-vector wide implicit fingerprint representation of the ligands, followed by the GRU layers returning sequential information to map to the SMILES representations. The exact makeup of the deep learning architecture with the trainable parameters is provided in the Supporting Information (Section 0.3).

Figure 15 illustrates the performance of the three different models trained with different data sets. The first data set comprised all of the 241k ligands from the data set. We additionally trained with the training set comprising only ligands with at least one positive binding affinity and another iteration comprising ligands with at least five positive binding affinities. As evident from the figure, the training and the corresponding validation loss were lower when trained with the filtered data sets as opposed to the entire population of 241k ligands. This can be attributed to the fact that the implicit fingerprints of the ligands that exhibited positive binding affinities in prior assays tend to encode more information on the ligands and hence decodable into the explicit SMILES representations. While this implies that approximately half the ligands in our data set do not resolve back to their corresponding SMILES representation, it does not, however, dent the utility of our approach. This is due to the fact that our approach is able to resolve the implicit fingerprints of the ligands, which have demonstrated bioactivities in the past, and hence such ligands are more desirable to be used as anchor−ligands from which to generate novel ligands.

CONCLUSIONS

We conclude that our approach of marrying the proven collaborative filtering approach with generative deep learning models is a promising new method for de novo drug generation. Our work shows that the implicit fingerprinting has a number of advantages in terms of encoding the desired properties of the ligands, including binding affinities to known proteins without explicitly optimizing for the said chemical properties. The compounds from the implicit space also demonstrated a wide diversity when measured using the Tanimoto distance. The collaborative filtering approach allows for the implicit fingerprints to be generated for any novel ligand with desired binding affinities to known target proteins. Leveraging these implicit fingerprints with encoded SMILES representations as the basis to generate useful novel druglike compounds could further advance this exciting field of drug discovery using generative deep learning models. We also note that our approach fundamentally relies on having training data for a particular anchor−ligand and particular target. To create an implicit fingerprint, the factorization employed in collaborative filtering requires assay examples. This requirement limits the scalability of the approach to ligands and targets for which assays are available or can be completed. We also point out that our analysis was completed on a large subset of the ChEMBL database, Version 23. Therefore, the consistency of the approach for ligands across different bioactivity databases needs to be further evaluated. Additionally, we note that we considered the implicit fingerprints based on binding affinities alone in this study; there are numerous desired properties (absorption, distribution, metabolism, excretion, toxicity, promiscuity, and pharmacovigilant properties) for which a ligand could be screened. These are typically referred to as secondary screens because they are most often (but not always) screened after affinity has been established. The cumulative generative capabilities by combining implicit fingerprints from these assays could also be evaluated in the future. We also note that further studies need to be conducted on the cumulative generative powers of the SMILES-based generative algorithms and the implicit fingerprints generated from collaborative filtering. One such limitation is the generation of faulty SMILES that cannot be resolved to a valid chemical structure. We also note the limitation in the method to occasionally generate ligands with unrealistic large rings. Future work could study mechanisms to penalize the models for generating ligands with such macrocycles to mitigate such issues. We also note that our work also generates ligands that can sometimes just be simple bioisosteric replacements of other known ligands, similar to other recent works[79] employing deep learning techniques. Future work in this space could leverage the implicit fingerprints with the other

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

popular methods of sampling around the dense continuous representations of the SMILES vectors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01355.

Novel ligands generated around six approved drugs; neural network architecture; list of all novel ligands generated from 5k anchor−ligands; list of all generated ligands and anchor−ligands with IoU scores and TC similarities; details of all the novel ligands generated by the network along with the respective anchor ligands, along with the analysis of the IoU and TC similarities; and details of the network architecture (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Raghuram Srinivas − *Department of Computer Science, Southern Methodist University, Dallas, Texas 75205, United States;* orcid.org/0000-0002-3854-6063; Email: rsrinivas@smu.edu

Authors

Niraj Verma − *Department of Chemistry, Southern Methodist University, Dallas, Texas 75205, United States*

Elfi Kraka − *Department of Chemistry, Southern Methodist University, Dallas, Texas 75205, United States*

Eric C. Larson − *Department of Computer Science, Southern Methodist University, Dallas, Texas 75205, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01355

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Rognan, D. The Impact of in silico Screening in the Discovery of Novel and Safer Drug Candidates. *Pharmacol. Ther.* **2017**, *175*, 47−66.

(2) Tsui, V.; Ortwine, D. F.; Blaney, J. M. Enabling Drug Discovery Project Decisions with Integrated Computational Chemistry and Informatics. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 287−291.

(3) Kitchen, D. B. Computer-aided Drug Discovery Research at a Global Contract Research Organization. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 309−318.

(4) Muegge, I.; Bergner, A.; Kriegl, J. M. Computer-aided Drug Design at Boehringer Ingelheim. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 275−285.

(5) van Vlijmen, H.; Desjarlais, R. L.; Mirzadegan, T. Computational Chemistry at Janssen. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 267−273.

(6) Song, H.; Wang, R.; Wang, S.; Lin, J. A Low-molecular-weight Compound Discovered through Virtual Database Screening Inhibits Stat3 Function in Breast Cancer Cells. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4700−4705.

(7) Allen, B.; Mehta, S.; Ayad, N.; Schürer, S. Ligand-and Structure-based Virtual Screening to Discover Dual EGFR and BRD4 inhibitors. *Neuro-Oncology* **2015**, *16*, No. 60.

(8) Munir, A.; Azam, S.; Mehmood, A.; Khan, Z.; Mehmood, A.; Aqdas, S. Structure-based Pharmacophore Modeling, Virtual Screening and Molecular Docking for the Treatment of ESR1 Mutations in Breast Cancer. *Drug Des.* **2016**, *05*, No. 1000137.

(9) Jacob, L.; Vert, J.-P. Protein-ligand Interaction Prediction: An Improved Chemogenomics Approach. *Bioinformatics* **2008**, *24*, 2149−2156.

(10) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of Drug-target Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* **2008**, *24*, i232−i240.

(11) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv:1510.02855, 2015.

(12) Tsubaki, M.; Tomii, K.; Sese, J. Compound−protein Interaction Prediction with End-to-end Learning of Neural Networks for Graphs and Sequences. *Bioinformatics* **2019**, *35*, 309−318.

(13) Li, L.; Koh, C. C.; Reker, D.; Brown, J. B.; Wang, H.; Lee, N. K.; Liow, H.-H.; Dai, H.; Fan, H.-M.; Chen, L.; Wei, D.-Q. Predicting Protein-Ligand Interactions based on Bow-Pharmacological Space and Bayesian Additive Regression Trees. *Sci. Rep.* **2019**, *9*, No. 7703.

(14) Verma, N.; Qu, X.; Trozzi, F.; Elsaied, M.; Karki, N.; Tao, Y.; Zoltowski, B.; Larson, E. C.; Kraka, E. SSnet: A Deep Learning Approach for Protein-Ligand Interaction Prediction. *Int. J. Mol. Sci.* **2021**, *22*, No. 1392.

(15) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv:1510.02855, 2015.

(16) 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.

(17) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and Evaluation of a Deep Learning model for Protein-ligand Binding Affinity Prediction. *Bioinformatics* **2018**, *34*, 3666−3674.

(18) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R. et al. Relational Inductive Biases, Deep Learning, and Graph Networks. arXiv:1806.01261, 2018.

(19) Schneider, G. Virtual Screening: an Endless Staircase. *Nat. Rev. Drug Discovery* **2010**, *9*, 273−276.

(20) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862−865.

(21) Giordano, A.; Forte, G.; Massimo, L.; Riccio, R.; Bifulco, G.; Di Micco, S. Discovery of New erbB4 Inhibitors: Repositioning an Orphan Chemical Library by Inverse Virtual Screening. *Eur. J. Med. Chem.* **2018**, *152*, 253−263.

(22) Lauro, G.; Romano, A.; Riccio, R.; Bifulco, G. Inverse Virtual Screening of Antitumor Targets: Pilot Study on a Small Database of Natural Bioactive Compounds. *J. Nat. Prod.* **2011**, *74*, 1401−1407.

(23) Xu, X.; Huang, M.; Zou, X. Docking-based Inverse Virtual Screening: Methods, Applications, and Challenges. *Biophys. Rep.* **2018**, *4*, 1−16.

(24) Srinivas, R.; Klimovich, P. V.; Larson, E. C. Implicit-Descriptor Ligand-based Virtual Screening by means of Collaborative Filtering. *J. Cheminf.* **2018**, *10*, No. 56.

(25) Wallach, I.; Heifets, A. Most Ligand-based Benchmarks Measure Overfitting Rather than Accuracy. arXiv:1706.06619, 2017.

(26) 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.

(27) 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.

(28) Bjerrum, E. J.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8*, No. 131.

(29) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, No. 1700123.

(30) 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. arXiv:1705.10843, 2017.
(31) Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A De Novo Molecular Generation Method using Latent Vector Based Generative Adversarial Network. *J. Cheminf.* **2019**, *11*, No. 74.
(32) Li, Y.; Zhang, L.; Liu, Z. Multi-objective De Novo Drug Design with Conditional Graph Generative Model. *J. Cheminf.* **2018**, *10*, No. 33.
(33) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. arXiv:1802.04364, 2018.
(34) ChEMBL23. https://www.ebi.ac.uk/chembl/ (accessed September 30, 2020).
(35) Goldberg, D.; Nichols, D.; Oki, B. M.; Terry, D. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM* **1992**, *35*, 61−70.
(36) Aggarwal, C. C. *Recommender Systems*; Springer, 2016; pp 29−70.
(37) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.* **2006**, *46*, 626−635.
(38) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. arXiv:1406.1231, 2014.
(39) Empereur-Mot, C.; Guillemain, H.; Latouche, A.; Zagury, J.-F.; Viallon, V.; Montes, M. Predictiveness Curves in Virtual Screening. *J. Cheminf.* **2015**, *7*, No. 52.
(40) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(41) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(42) van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(43) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555, 2014.
(44) Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; Bengio, S. Generating Sentences from a Continuous Space. arXiv:1511.06349, 2015.
(45) Tang, Z.; Shi, Y.; Wang, D.; Feng, Y.; Zhang, S. In *Memory Visualization for Gated Recurrent Neural Networks in Speech Recognition*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017; pp 2736−2740.
(46) Santur, Y. In *Sentiment Analysis Based on Gated Recurrent Unit*. International Artificial Intelligence and Data Processing Symposium (IDAP), 2019; pp 1−5.
(47) Zulqarnain, M.; Ishak, S.; Ghazali, R.; Nawi, N. M.; Aamir, M.; Hassim, Y. M. M. An Improved Deep Learning Approach based on Variant Two-state Gated Recurrent Unit and Word Embeddings for Sentiment Classification. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 594−603.
(48) 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.
(49) 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.
(50) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461.
(51) McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578−596.
(52) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(53) Huang, S.-Y.; Li, M.; Wang, J.; Pan, Y. HybridDock: A Hybrid Protein−Ligand Docking Protocol Integrating Protein- and Ligand-Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1078−1087.
(54) Korb, O.; Stutzle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.
(55) Sterling, T.; Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(56) Karki, N.; Verma, N.; Trozzi, F.; Tao, P.; Kraka, E.; Zoltowski, B. Predicting Potential SARS-COV-2 Drugs-In Depth Drug Database Screening Using Deep Neural Network Framework SSnet, Classical Virtual Screening and Docking. *Int. J. Mol. Sci.* **2021**, *22*, No. 1573.
(57) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, No. 33.
(58) 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.
(59) Reker, D.; Schneider, P.; Schneider, G.; Brown, J. Active Learning for Computational Chemogenomics. *Future Med. Chem.* **2017**, *9*, 381−402.
(60) Lenselink, E. B.; Ten Dijke, N.; Bongers, B.; Papadatos, G.; Van Vlijmen, H. W.; Kowalczyk, W.; IJzerman, A. P.; Van Westen, G. J. Beyond the Hype: Deep Neural Networks Outperform Established Methods using a ChEMBL Bioactivity Benchmark Set. *J. Cheminf.* **2017**, *9*, No. 45.
(61) Yasonik, J. Multiobjective De Novo Drug Design with Recurrent Neural Networks and Nondominated Sorting. *J. Cheminf.* **2020**, *12*, No. 14.
(62) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, No. 48.
(63) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.* **2018**, *4*, No. eaap7885.
(64) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.
(65) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(66) Ertl, P. An algorithm to identify functional groups in organic molecules. *J. Cheminf.* **2017**, *9*, No. 36.
(67) 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.
(68) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(69) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, No. 8.
(70) Mann, H. B.; Whitney, D. R. On a Test of whether One of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50−60.
(71) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-five Revolution. *Drug Discovery Today* **2004**, *1*, 337−341.
(72) Yao, Z.-J.; Dong, J.; Che, Y.-J.; Zhu, M.-F.; Wen, M.; Wang, N.-N.; Wang, S.; Lu, A.-P.; Cao, D.-S. TargetNet: A Web Service for Predicting Potential Drug−target Interaction Profiling via Multi-target SAR Models. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 413−424.
(73) Nowozin, S. In *Optimal Decisions from Probabilistic Models: The Intersection-Over-Union Case*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014; pp 548−555.
(74) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(75) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742−1753.
(76) Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *59*, 4062−4076.

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174

(77) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.
(78) Kingma, D. P.; Welling, M. Auto-encoding Variational Bayes. arXiv:1312.6114, 2013.
(79) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

https://doi.org/10.1021/acs.jcim.0c01355
J. Chem. Inf. Model. 2021, 61, 2159−2174