JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

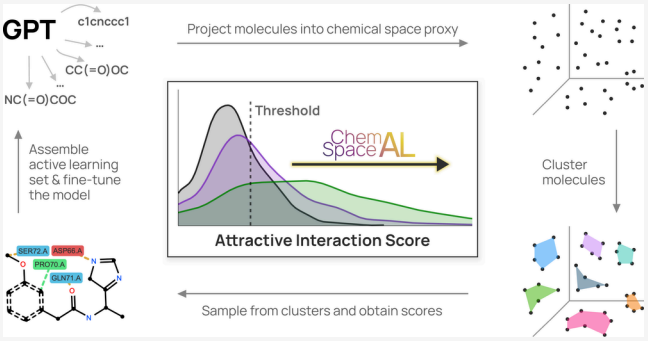
pubs.acs.org/jcim Article

ChemSpaceAL: An Efficient Active Learning Methodology Applied to Protein-Specific Molecular Generation

Gregory W. Kyro, Anton Morgunov, Rafael I. Brent, and Victor S. Batista*

Cite This: J. Chem. Inf. Model. 2024, 64, 653−665 Read Online

ABSTRACT: The incredible capabilities of generative artificial intelligence models have inevitably led to their application in the domain of drug discovery. Within this domain, the vastness of chemical space motivates the development of more efficient methods for identifying regions with molecules that exhibit desired characteristics. In this work, we present a computationally efficient active learning methodology and demonstrate its applicability to targeted molecular generation. When applied to c-Abl kinase, a protein with FDA-approved small-molecule inhibitors, the model learns to generate molecules similar to the inhibitors without prior knowledge of their existence and even reproduces two of them exactly. We also show that the methodology is effective for a protein without any commercially available small-molecule inhibitors, the HNH domain of the CRISPR-associated protein 9 (Cas9) enzyme. To facilitate implementation and reproducibility, we made all of our software available through the open-source ChemSpaceAL Python package.



1. INTRODUCTION

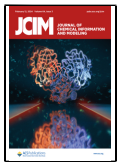
The vast majority of pharmaceutical drugs function by targeting a specific protein.[1] Virtual screening and de novo drug design are popular areas of research aimed at developing effective protein-specific drugs.[2] Molecular generation methods powered by generative artificial intelligence (AI) can advance both of these areas, and there have already been numerous reports of recurrent neural networks (RNNs),[3−25] generative adversarial networks (GANs),[26−39] autoencoders,[40−63] and transformers[64−71] successfully contributing to drug development methods.

Active learning (AL) can be used to fine-tune an AI model with selectively chosen data points, ensuring that the model retains its broad domain knowledge while narrowing its focus toward a more precise objective. In its basic form, AL can be applied by exclusively using data points that have been directly evaluated and satisfy specific criteria. However, within the AL framework, it is feasible to extend traditional methods by not only including directly evaluated data points but also incorporating a mechanism that utilizes unevaluated data points similar to the evaluated ones deemed satisfactory. This approach facilitates the use of resource-intensive scoring functions that otherwise would be too expensive by scoring only a strategically selected subset of data points and extending the insights obtained from the scores to data that have not been evaluated. In this context, the total computational cost is largely dependent on the number of scored molecules necessary to sufficiently represent the search space.

Although there are many notable examples of AL methods for discriminative tasks pertaining to drug discovery,[72−77] the application of AL to molecular generation is comparatively unexplored. Within this domain, recent work has demonstrated the ability of AL to identify molecules with satisfactory in silico binding affinities,[52] yet there remains significant motivation to develop an efficient approach for fine-tuning a molecular generator toward a protein target that minimizes the number of required docking calculations, which are computationally expensive.

In this work, we present a computationally efficient AL methodology that leverages a strategic algorithm for estimating the binding ability of molecules that have not been docked, and requires the evaluation of only a subset of the generated data to successfully align the generated molecular ensemble toward a specified protein target. Specifically, we demonstrate the effectiveness of our methodology by independently aligning a generative pretrained transformer (GPT)-based model to c-Abl kinase and the HNH domain of the CRISPR-associated protein 9 (Cas9) enzyme.[78,79]

Received: September 10, 2023
Revised: December 31, 2023
Accepted: January 3, 2024
Published: January 30, 2024



https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

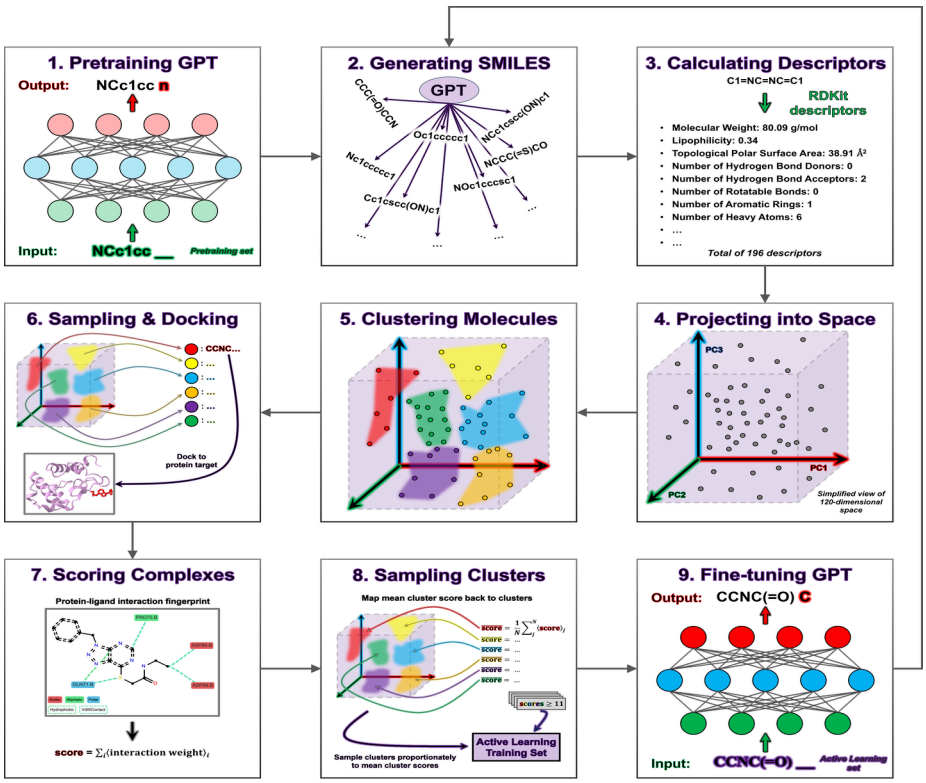


Figure 1. Process flow diagram depicting the complete ChemSpaceAL active learning methodology applied to molecular generation.

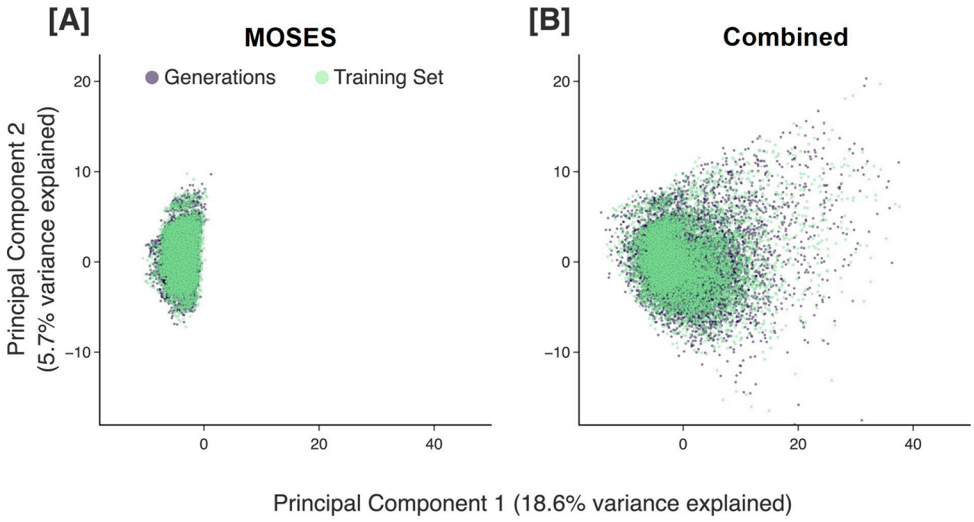


Figure 2. Different pretraining sets (green) plotted with the molecules generated (purple) by the corresponding pretrained model that was trained only on the respective pretraining set. Note that 100 000 data points were randomly sampled from each pretraining set, and 100 000 were generated in each case. The descriptor vectors of the data points are projected into our chemical space proxy, and the first two principal components are shown. Results are displayed for the (A) MOSES and (B) combined pretraining sets.

2. OVERVIEW OF THE CHEMSPACEAL METHODOLOGY

The ChemSpaceAL methodology applied to molecular generation (Figure 1) proceeds as follows:

1) Pretrain the GPT-based model on millions of SMILES (Simplified Molecular Input Line Entry System) strings
2) Use the trained model to generate 100 000 unique molecules (determined by SMILES string canonicalization)
3) Calculate molecular descriptors for each generated molecule
4) Project the descriptor vectors of the generated molecules into a principal component analysis (PCA)-reduced space constructed from the descriptors of all molecules in the pretraining set
5) Use *k*-means clustering on the generated molecules within the space to group those with similar properties
6) Sample about 1% of molecules from each cluster and dock each of them to a protein target (e.g., c-Abl kinase or the HNH domain of Cas9)

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

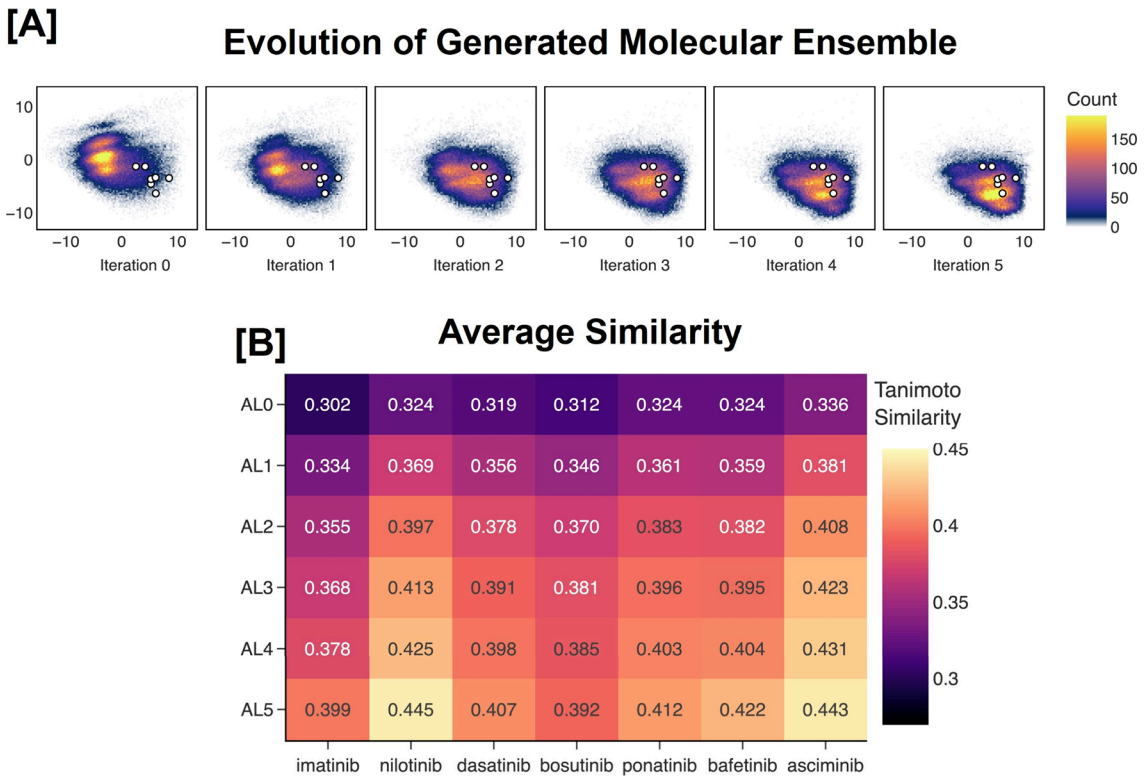


Figure 3. Comparing the evolution of the generated molecular ensemble from the model pretrained on the combined data set to the FDA-approved small-molecule inhibitors of c-Abl kinase. In (A), the descriptor vectors of the generated molecules across each iteration of our methodology are projected into our chemical space proxy and visualized along the first two principal components. The inhibitor descriptor vectors are also projected into the space and are represented by white dots with a black outline. In (B), the average Tanimoto similarities between the RDKit fingerprints of all generated molecules at each iteration and that of each inhibitor are shown. Tanimoto similarities between the inhibitors are reported in Figure S2.1. Iteration 0 refers to the pretraining phase, while later iterations refer to active learning phases.

7) Evaluate the top-ranked pose of each protein−ligand complex with an attractive interaction-based scoring function
8) Construct an AL training set by sampling from the clusters proportionally to the mean scores of the evaluated molecules within each respective cluster and combining the sampled molecules with replicas of the evaluated molecules whose scores meet a specified threshold
9) Fine-tune the model with the AL training set
*) Repeat steps (2)−(9) for multiple iterations

3. ALIGNING THE GENERATIVE MODEL TO SPECIFIED PROTEIN TARGETS

Utilizing a transformer decoder-based GPT model (more details in section 7),[80] our initial goal was to pretrain the model on data that span as much of true chemical space as possible. This approach allows the pretrained model to develop a rich internal representation of SMILES strings, enabling it to generate a diverse array of molecules. To curate an extensive data set for pretraining the model, we combined SMILES strings from four data sets: ChEMBL 33 (about 2.4 million bioactive molecules with drug-like properties),[81] GuacaMol v1 (about 1.6 million molecules derived from ChEMBL 24 that have been synthesized and tested against biological targets),[82] MOSES (about 1.8 million molecules selected from ZINC 15 to maximize internal diversity and suitability for medicinal chemistry),[83,84] and BindingDB 08-2023 (about 1.2 million unique small molecules bound to proteins).[85] After processing, the resulting data set contains about 5.6 million unique and valid SMILES strings and will be referred to as the combined data set. More details regarding the data that we used and the preprocessing methods that we employed are discussed in section 6. To assess the dependence of our methodology on the nature of the pretraining set, we compared two independent models: one pretrained on the combined data set (C model), and one pretrained on the MOSES data set (M model).

In Figure 2, we show 100 000 generated molecules from each model trained solely on either the MOSES data set or the combined data set along the first two principal components of our chemical space proxy. It should be noted that the PCA reduction was performed only once on the molecular descriptors of all molecules in the combined data set, and the obtained principal components are used for all visualizations throughout this work, ensuring fair comparisons between different sets of data points (more details in section 8.1). We see that the pretrained models are able to generate molecules that roughly cover the area spanned by the corresponding pretraining sets (Figure 2).

Using both pretrained models, we independently assessed the ChemSpaceAL methodology with c-Abl kinase and the HNH domain of Cas9. In the first case, we aimed to validate our methodology by showing that the generated molecular ensemble evolves toward the U.S. Food and Drug Administration (FDA)-approved small-molecule inhibitors of c-Abl kinase. In the latter case, we investigated the applicability of

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

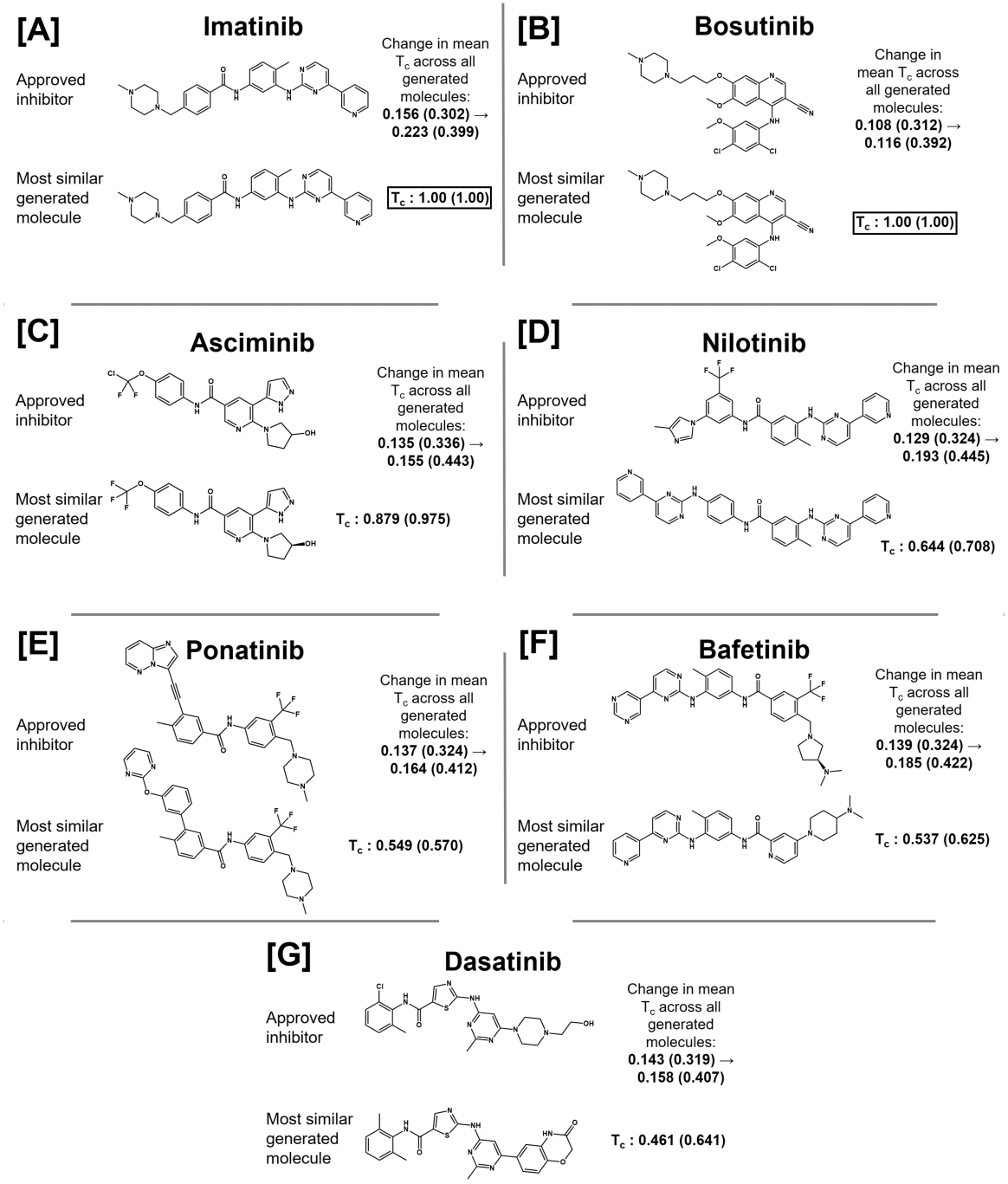


Figure 4. Comparison of the generated molecular ensemble from the model pretrained on the combined data set to the FDA-approved small-molecule inhibitors of c-Abl kinase. For each inhibitor, the most similar generated molecule after five iterations is shown as well as the Tanimoto similarity (T_C) between the two. The change in the mean similarity between each inhibitor and all generated molecules from iteration 0 (pretrained model) to iteration 5 is shown. For all comparisons in this figure, the T_C between extended-connectivity fingerprint 4s is shown, along with the T_C between RDKit fingerprints in parentheses.[89,90] Results are shown for (A) imatinib, (B) bosutinib, (C) asciminib, (D) nilotinib, (E) ponatinib, (F) bafetinib, and (G) dasatinib.

the methodology to a protein without any commercially available small-molecule inhibitors.

In both cases, the generated molecules were filtered based on ADMET (absorption, distribution, metabolism, excretion, and toxicity) metrics and functional group restrictions.[86] ADMET filters were employed to ensure that the molecules possess drug-like properties, and functional group restrictions were used to discard chemical moieties that are less favorable for biological applications. More details regarding the ADMET and functional group filters that we used are reported in Tables S1.1 and S1.2 in the Supporting Information.

3.1. Aligning to c-Abl Kinase. c-Abl kinase (PDB ID: 1IEP)[72] is of significant scientific interest because its dysfunction is associated with the development of chronic myeloid leukemia, making it a vital target for anticancer drugs designed to inhibit its activity and thereby control the proliferation of cancer cells. There are multiple FDA-approved small-molecule inhibitors of c-Abl kinase that have similar

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

structures, including imatinib, nilotinib, dasatinib, bosutinib, ponatinib, bafetinib, and asciminib.[78,87,88] We docked and scored each of the inhibitors using our scoring function, and chose the lowest score among them to be the score threshold for our methodology (more details in section 8.3).

For the C model, the mean Tanimoto similarities between the generated molecular ensemble and each of the seven inhibitors increase at each iteration, indicating a constant evolution toward the inhibitors (Figure 3B). This shift of the distribution toward the region of space that contains the FDA-approved inhibitors can be visualized by projecting the descriptor vectors of the generated ensemble at each iteration of the methodology and those of the inhibitors into the chemical space proxy (Figure 3A). Moreover, the set of generated molecules after five iterations contains imatinib and bosutinib (Figure 4).

We also assessed the performance of the methodology by analyzing the distribution of scores of generated molecules throughout AL iterations. For both the C and M models, the percentage of molecules that reached the scoring threshold significantly increased after five iterations of AL, further validating the applicability of our method to c-Abl kinase; the percentage increased from 38.8% to 91.6% for the C model and from 21.7% to 80.3% for the M model (Table 1). The evolutions of these distributions can be seen in Figure 5.

Table 1. Evolution of Protein−Ligand Attractive Interaction Scores between Molecules in the Generated Ensemble and c-Abl Kinase across Our Complete Active Learning Methodology[a]

iteration[b]	C % ≥ 37	C mean	C max	M % ≥ 37	M mean	M max
0	38.8	32.8	70.0	21.7	30.3	55.5
1	59.3	38.4	74.5	42.1	35.2	57.0
2	70.1	41.4	68.0	59.2	38.0	60.5
3	81.2	44.0	73.5	68.8	39.9	60.0
4	86.6	46.0	77.5	76.2	41.0	59.0
5	91.6	48.5	77.0	80.3	41.8	61.0

[a]The percentage of generated molecules with attractive interaction scores equal to or above our score threshold (% ≥ 37), the mean score, and the maximum score are shown for the model pretrained on the combined data set (C) and the model pretrained on the MOSES data set (M) for five iterations of the methodology. [b]Iteration 0 refers to the pretraining phase, while later iterations refer to active learning phases.

It is worth noting that 38.8% of the molecules generated by the C model reached the score threshold immediately after pretraining, while only 21.7% of the molecules generated by the M model reached the threshold, indicating that our combined pretraining set covers regions of chemical space not spanned by the MOSES data set that contain higher-scoring molecules (Table 1). Moreover, after applying the methodology, the molecular ensemble generated by the C model is more similar to the FDA-approved inhibitors than that generated by the M model (Figure S3.1) and is comprised exclusively of molecules with satisfactory ADMET profiles (Figure S4.1). These results support the notion that our methodology is more effective at generating drug-like molecules specific to a protein target by pretraining on the combined data set and applying filters to the generation stage rather than pretraining on a refined data set, such as the MOSES data set.

3.2. Aligning to the HNH Domain of Cas9. To further evaluate our methodology, we applied it to a protein without any commercially available small-molecule inhibitors, the HNH domain of Cas9 (PDB ID: 6O56).[79] This protein is a nuclease component critical to the function of the CRISPR/Cas9 system and is responsible for cleaving the target DNA strand complementary to the guide RNA, which directs the Cas9 enzyme to the correct sequence for gene modification. The HNH domain of Cas9 is, therefore, particularly interesting because understanding its structure and dynamics can lead to enhancements in the precision and efficiency of CRISPR-based gene editing tools.[91] Furthermore, the ability to develop binders for HNH could offer a direct way to modulate its behavior.

Our methodology requires a score threshold in order to select molecules to be included in the AL training set. In the absence of known small-molecule binders for HNH, we refered to a large database of experimentally determined protein−ligand complexes, the PDBbind v2020 refined set,[92] and selected this threshold to be 11 (more details in section 8.3). This lack of known binders also led us to use the change in the distribution of scores as the primary metric for evaluation. After five iterations of AL, the percentage of generated molecules that reached the score threshold increased from 21.3% to 52.1% for the C model and from 14.3% to 28.2% for the M model (Table 2); the performance differential between the C and M models is commensurate with that observed for c-Abl kinase. The evolutions of these distributions can be seen in Figure S5.1 in the Supporting Information.

4. EVALUATING INDIVIDUAL COMPONENTS OF THE METHODOLOGY

The goal of this section is to isolate and analyze the effectiveness of individual components of our methodology: the chemical space proxy, clustering algorithm, scoring method, and sampling algorithm for constructing AL training sets. For all results presented here, the methodology was applied to the model pretrained on our combined data set for alignment to HNH, without any filters during generation stages. We excluded filters on the generated molecules to probe how the model responds with respect to the scoring function. In addition, we performed analogous analyses of the methodology applied to c-Abl kinase with ADMET and functional group filters applied to the generated molecules, and we observe similar results to those included in this section (see Figures S3.2 and S3.4 in the Supporting Information).

4.1. Naïve Active Learning Control. In order to establish a baseline for comparison to our methodology, we performed a naïve version of AL, where we generated 100 000 unique molecules, randomly selected 1000 of them, docked and scored each of the selected molecules, and then fine-tuned the model with the scored molecules that reached the score threshold. The purpose of this approach is to demonstrate how the fine-tuning would occur if we did not sample from clusters in the chemical space proxy to construct an AL training set. In this case, we constructed the AL training set from N replicas of each molecule that scored equal to or above the score threshold, where N is the smallest integer that achieves a total number of molecules of at least 5000. The model was then further trained on this AL set, and the fine-tuned model was used to generate another 100 000 unique molecules, which were subsequently used for another iteration of the methodology. We repeated this procedure for a total of five iterations

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

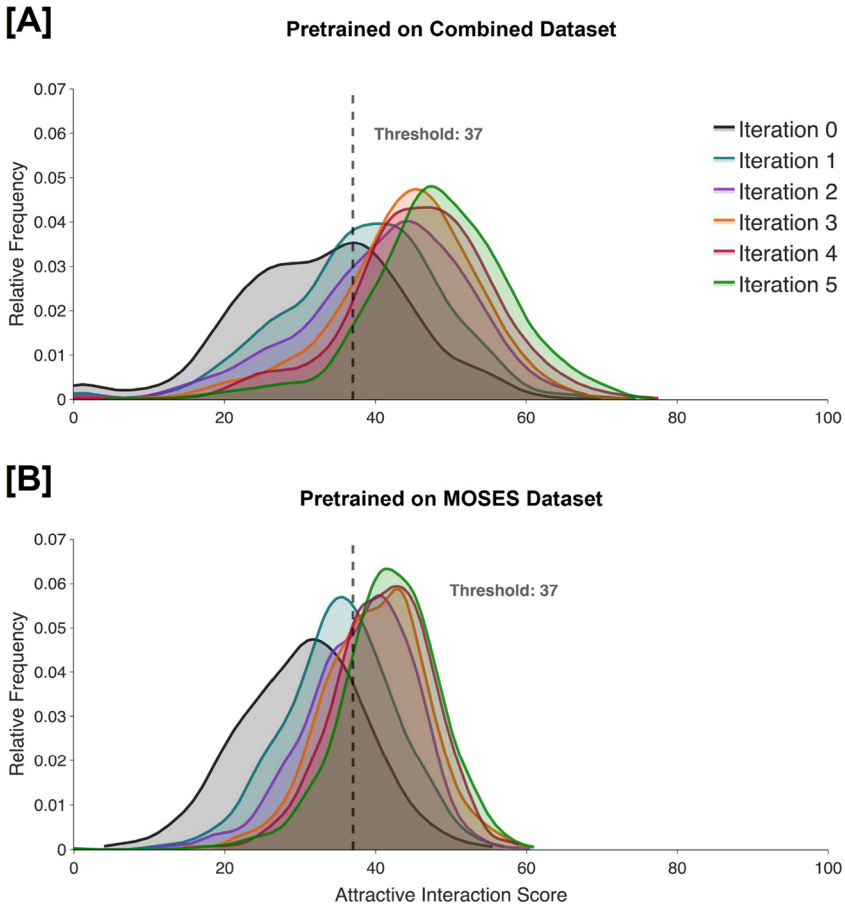


Figure 5. Attractive interaction scores of evaluated molecules across five iterations of active learning forc-Abl kinase. The distributions for the model pretrained on the combined data set are shown in (A), and the distributions for the model pretrained on the MOSES data set are shown in (B). Iteration 0 refers to the pretraining phase, while later iterations refer to active learning phases.

Table 2. Evolution of Protein−Ligand Attractive Interaction Scores between Molecules in the Generated Ensemble and the HNH Domain of Cas9 across Our Complete Active Learning Methodology[a]

iteration[b]	C % ≥ 11	C mean	C max	M % ≥ 11	M mean	M max
0	21.3	7.9	32.5	14.3	7.3	22.5
1	31.9	9.1	26.5	18.9	7.8	21.0
2	39.1	9.8	25.0	22.5	8.2	22.0
3	43.9	10.4	23.0	24.5	8.6	23.0
4	50.1	11.1	33.5	28.7	8.9	21.0
5	52.1	11.5	34.0	28.2	9.0	23.0

[a]The percentage of generated molecules with attractive interaction scores equal to or above our score threshold (% ≥ 11), the mean score, and the maximum score are shown for the model pretrained on the combined data set (C) and the model pretrained on the MOSES data set (M) for five iterations of the methodology. [b]Iteration 0 refers to the pretraining phase, while later iterations refer to active learning phases.

and observe that the percentage of generated molecules that reached the score threshold increased from 26.2% to 44.2% (Figure 6A).

4.2. Chemical Space Proxy and Clustering Algorithm. In order to improve upon naïve AL, we strategically select molecules to be in the AL training set that have not been evaluated. This requires a method for relating molecules that have been scored to those that have not. To achieve this goal, we constructed a proxy for chemical space that is predicated on molecular properties, allowing us to operate within a space where nearby molecules share similar chemical features. More details regarding the construction of our chemical space proxy are discussed in section 8.1.

A correlation must exist between position in the chemical space proxy and the values produced by the scoring function in order to successfully estimate the scores of molecules that have not been evaluated. By visualizing all of the scored molecules from all iterations of the complete methodology (6000 molecules) along the first two principal components of our chemical space proxy, we observe a continuous gradient of scores (Figure 7A), illustrating the relation between position in our chemical space proxy and the values produced by our scoring function. Moreover, when the positions of the scored molecules in the chemical space proxy are reduced to two dimensions using t-distributed stochastic neighbor embedding (t-SNE), a technique that captures nonlinear structures, we also see that the regions containing molecules with higher scores are easily identifiable (Figure 7B; more details in section 6 of the Supporting Information).

Within our chemical space proxy, we utilized k-means clustering with $k = 100$ to group molecules that exhibit similar chemical properties. We also report results for $k = 10$, which proved to be less effective (see Figures S7.1−S7.4 in the Supporting Information). This is likely because much of the diversity in the chemical space is homogenized into clusters that, in the case of $k = 10$, are very large compared to those in

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

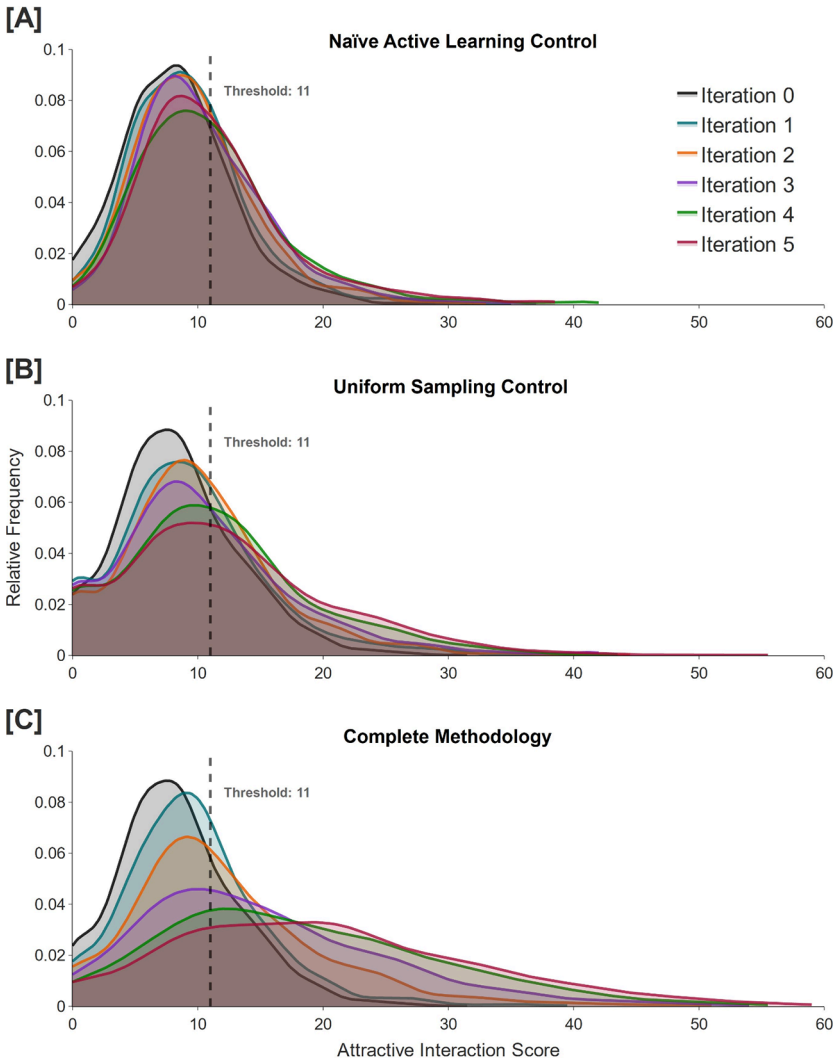


Figure 6. Attractive interaction scores of evaluated molecules across five iterations of active learning. Results for the naïve active learning control are shown in (A), which utilized the random selection of molecules and fine-tuning with only replicas of those that scored equal to or above the score threshold of 11. Results for the uniform sampling control are shown in (B), which used cluster-based sampling, where each cluster was assigned a sampling fraction $f = 0.01$ during the construction of the active learning set. Results for our complete methodology are shown in (C). Iteration 0 refers to the pretraining phase, while later iterations refer to active learning phases.

the case of $k = 100$, and valuable information is lost. In short, we generated 100 clusters and then randomly sampled up to 10 molecules from each cluster, selecting all molecules in cases where a cluster contained fewer than 10 molecules. More details of our clustering method are discussed in section 8.2.

4.3. Docking and Scoring. After strategically selecting 1000 molecules, we docked each of them to a protein target using DiffDock (more details in section 8 of the Supporting Information)[93] and evaluated each top-ranked pose with our scoring function, which is essentially a sum of attractive protein−ligand contact points, each weighted by its interaction type. More details regarding the scoring function we used are discussed in section 8.3.

4.4. Uniform Sampling Control. Because the generated molecules are not evenly distributed in the chemical space proxy, cluster-based sampling introduces a bias in which molecules from less dense regions are sampled more frequently than they would be with random selection. This leads to a score-independent shift in the distribution throughout AL iterations, which we refer to as the diffusion effect. To assess this bias, we constructed AL training sets by randomly selecting 10 molecules from each cluster, scoring each of them, selecting the molecules with scores that reached the score threshold (at least 5000 molecules, including replicas), and sampling from each cluster with the same sampling fraction $f = 0.01$ (about 50 from each cluster for a total of 5000 molecules) for a total of approximately 10 000 molecules. This approach serves as a control for isolating the effectiveness of our algorithm for sampling unscored molecules to be in the AL training set. For this uniform sampling-based approach, the increase in the scores of the molecules in the generated ensemble after five iterations (28.1% to 51.1%) is slightly more pronounced than that achieved via naïve AL (26.2% to 44.2%), as shown in Figure 6B. However, these results are significantly worse than those achieved with our complete methodology (28.1% to 76.0%), indicating that our score-based sampling method is necessary for high performance and aligns the model with the scoring function much more effectively than uniform sampling.

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

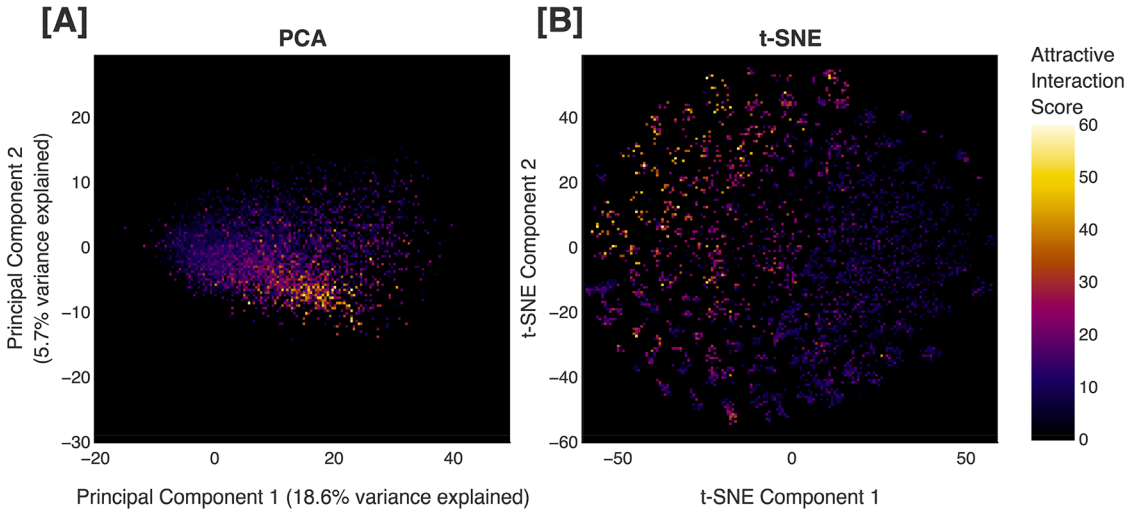


Figure 7. Visualization of scored molecules in the chemical space proxy. All of the scored molecules from all iterations of the complete methodology applied to the model pretrained on our combined data set, for alignment to HNH and with no filters on the generated molecules (6000 molecules), are displayed. (A) Descriptor vectors of the generated molecules projected into the chemical space proxy and shown along the first two principal components. (B) Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) plot of the generated molecules. Plots are colored by score obtained with the scoring function, where black/purple corresponds to lower scores and white/yellow corresponds to higher scores.

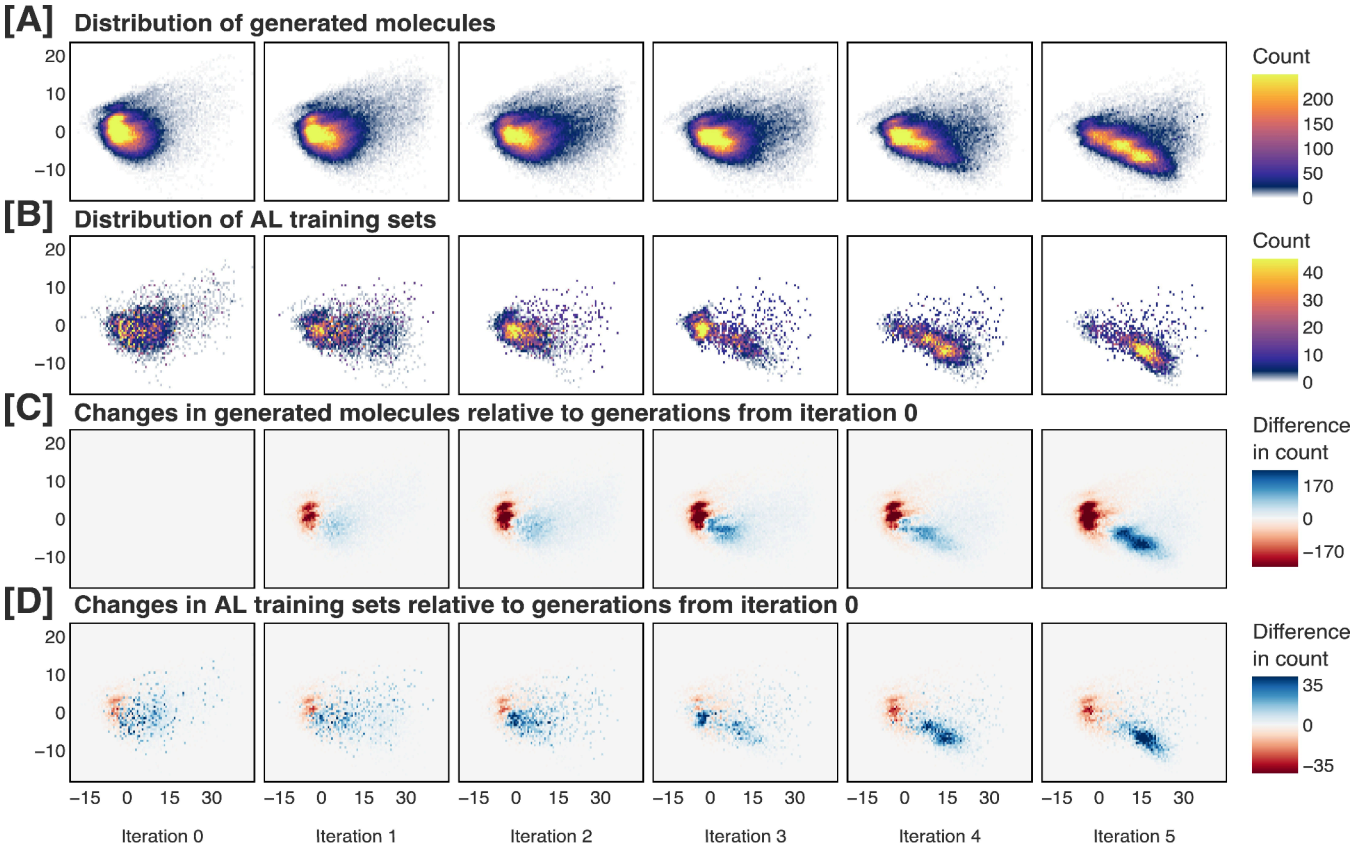


Figure 8. Generated molecules and active learning training sets across five iterations of our complete methodology visualized along the first two principal components of our chemical space proxy. The generated molecular ensembles and active learning training sets at each iteration are shown in (A) and (B), respectively. Changes in the generated molecular ensembles and active learning training sets relative to the molecules generated at iteration 0 are shown in (C) and (D), respectively. In (A) and (C), the 100 000 unique generated molecules from each iteration are used. In (B), the full active learning training sets, each containing approximately 10 000 molecules, are used. In (D), for a proper comparison between the generated molecules at iteration 0 and the active learning training sets, 5000 molecules are randomly sampled from the generated ensemble at iteration 0, and 5000 molecules are randomly sampled from the active learning training set at each iteration. Iteration 0 refers to the pretraining phase, while later iterations refer to active learning phases. More details of this analysis are reported in Figure S9.1 in the Supporting Information.

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

4.5. Sampling from Clusters Proportionally to Their Scores. In order to improve upon uniform sampling, we propose a way to intelligently weight the importance of each cluster when sampling molecules from the chemical space proxy to be in the AL training set. After scoring each of the 1000 protein−ligand pairs, we sampled from the clusters proportionally to the mean scores calculated from the evaluated molecules within each respective cluster. These sampled molecules were then combined with replicas of the evaluated molecules whose scores met the score threshold, forming the AL training set. More details regarding our sampling algorithm are discussed in section 8.4. Our sampling procedure allows us to enrich the AL training set with unscored molecules that would likely obtain high scores, exploiting the fact that position in the chemical space proxy correlates with the scoring function (Figure 7).

Our complete methodology shifted the percentage of generated molecules that reached the score threshold from 28.1% to 76.0% (Figure 6C). This increase can be attributed to the shift of the generated molecular ensemble toward the region of the chemical space proxy associated with higher scores. Figure 8 illustrates this progression, depicting the evolution of the generated ensemble in a constant direction through the chemical space proxy.

5. SUMMARY AND FUTURE OUTLOOK

In this work, we present an efficient AL methodology, and demonstrate its applicability in the context of targeted molecular generation. In particular, we independently enhance attractive interactions between the molecules in the generated ensemble and two protein targets, c-Abl kinase and the HNH domain of Cas9. When aligning toward c-Abl kinase, we were able to shift the distribution of generated molecules toward the region of the chemical space proxy corresponding to several FDA-approved inhibitors for this target. We also showed that our methodology is effective for a protein without any commercially available small-molecule inhibitors, the HNH domain of Cas9. Moreover, we analyzed the effectiveness of individual components of our methodology and showed that the integration of these components in our complete approach aligned the model with the scoring function much more effectively than more naïve AL methods.

The generative model, constructed sample space, and scoring function are all highly substitutable within the framework of our methodology, and we therefore envision that it will be adaptable to future innovations. For instance, the GPT-based model could be replaced with a more capable architecture as soon as one is developed. In addition, any quantifiable features that are correlated with the scoring function can be used to represent the data. In the context of molecular generation, the list of descriptors used to construct our chemical space proxy could be substituted as better molecular descriptors are developed. Moreover, the scoring function that we use can be replaced by a better metric to achieve a closer correspondence with experimental results. The generality of our approach facilitates the applicability and utility of the ChemSpaceAL methodology both at present and as the state of the field inevitably improves.

6. DATA SET COLLECTION AND PREPROCESSING

6.1. Data Collection. We combined all of the SMILES strings from ChEMBL 33, GuacaMol v1, MOSES, and BindingDB, filtered out the strings that were identified as invalid by the RDKit molecular parser, and removed any duplicate strings. The resulting combined data set contained 5 622 772 unique and valid SMILES strings.

6.2. Tokenization. Our combined data set initially had a vocabulary of 196 unique tokens. We found that 148 tokens were represented in the data set fewer than 1000 times; to reduce the size of our vocabulary (from 196 to 48), we removed all SMILES strings containing at least one token that appeared less than 1000 times in the combined data set (details given in Tables S10.1 and S10.2). Most of the SMILES strings that were excluded contain rare transition metals or isotopes.

6.3. Data Preprocessing. The longest SMILES string in the combined data set contained 1503 tokens, while 99.99% of the strings in the data set had 133 or fewer tokens (details given in Figures S11.1 and S11.2). We imposed a SMILES string length cutoff of 133 and removed any string from the data set whose length is greater than this cutoff. All of the remaining SMILES strings were then extended, if necessary, to the length of the longest SMILES string in the data set (133) using a padding token "<", and were augmented with a start token "!" and an end token "∼". The resulting data set contains 5 539 765 SMILES strings, which were randomly split into training (5 262 776 entries; 95.0%) and validation (276 989 entries; 5.0%) sets for pretraining.

7. DETAILS OF THE GENERATIVE MODEL

We utilize a GPT-based model (details of the model architecture can be found in section 12 of the Supporting Information). Our model embeds inputs into a 256-dimensional space and is composed of eight transformer decoder blocks, each of which contains eight attention heads. Dropout with a probability of 10% is applied after each feed-forward network except for the output layer to mitigate overfitting, and gradient clipping with a maximum norm of 1.0 is used in conjunction with layer normalization to stabilize the optimization process and prevent exploding gradients. All weights are initialized according to a Gaussian distribution with a mean of 0 and a standard deviation of 0.02, except for weights involved in layer normalization, which are initialized to 1, and bias parameters, which are initialized to 0. The training process utilizes cross-entropy loss with L2 regularization applied to the linear layers using $\lambda = 0.1$ and the SophiaG optimizer with $\beta_1 = 0.965$, $\beta_2 = 0.99$, and $\rho = 0.04$.[94]

7.1. Pretraining. During pretraining, the learning rate warms up to 3×10^{-4} until the model has been trained on 10% of the total number of tokens in the data set, then decays to 3×10^{-5} using cosine decay. The model was trained with a batch size of 512 for 30 epochs. Learning curves are reported in Figures S13.1 and S13.2.

7.2. Benchmarking. Many generative AI models for molecular discovery have been evaluated with the MOSES benchmark,[77] which constitutes an important standard for the field, with the objective of assessing models' abilities to generate diverse collections of novel and valid molecules. We show that our pretrained model performs among the best in the field (details given in Tables S14.1 and S14.2), establishing its merit as a starting point for AL.

7.3. Fine-Tuning. After compiling a given AL training set, the model is further trained with a batch size of 512 for 10 epochs using a learning rate of 3×10^{-5} with no warmup and a cosine decay to 3×10^{-6}.

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

8. DETAILS OF THE CHEMSPACEAL METHODOLOGY

8.1. Chemical Space Proxy. We first calculated the full set of molecular descriptors that are available through RDKit's CalcMolDescriptors function for each molecule in the combined pretraining set, encompassing a wide range of molecular properties including structural, topological, geometrical, electronic, and thermodynamic characteristics. Among these 209 descriptors, 13 returned NaN (not a number) or infinity for at least one SMILES string in the data set and were consequently discarded (details given in Table S15.1), resulting in 196 descriptors (details in Table S15.2). We used as many RDKit descriptors as possible because this step in the methodology is very fast (see Figure S16.1), enabling us to generate maximally descriptive molecular representations. We also independently investigated the performance of the methodology using only the 42 RDKit molecular quantum numbers (MQNs), which are not included in the CalcMolDescriptors function, and found this representation to yield worse results than those obtained using the PCA-reduced 120-dimensional representation of the 196 descriptors (see Figure S17.1). Additionally, we evaluated our methodology with the logits for the end-of-sequence token as the descriptor vector for each molecule,[95] and the results are reported in Figures S18.1−S18.4. We note that for the proposed methodology to work, the set of descriptors used must satisfy two criteria: (1) position in the chemical space proxy correlates with the scoring function and (2) nearby molecules in the chemical space proxy have similar scores. It is evident that there could exist many sets of descriptors satisfying these criteria; a thorough investigation into the choice of descriptors is outside the scope of this work. After performing PCA using the 196 RDKit descriptors for all molecules in the combined pretraining set, we found that 99% of the variance is explained by the first 113 principal components (details given in Figure S15.3) and used the first 120 principal components throughout the methodology as our chemical space proxy. Our methodology might attain similar results with fewer principal components retained, but this reduction is not necessary since this step is computationally inexpensive.

8.2. Clustering Algorithm. Within our chemical space proxy, we utilized k-means clustering to group molecules that exhibit similar chemical properties with $k = 100$. Given that running k-means is incredibly fast, we performed k-means 100 times to mitigate the potential for poor initialization, seeking to minimize k-means loss and cluster size variance. Initially, we took the five clusterings with the lowest loss, thereby preserving those with more compact clusters. Of these five, we selected the clustering with the lowest variance in cluster size for use in the following stages of the methodology.

After clustering the generated molecules in our chemical space proxy, we randomly selected 10 molecules from each cluster that contained at least 10 molecules and selected all of the molecules from any cluster that contained less than 10 molecules. For AL iterations 1−5, when applying the methodology to the C model for aligning to HNH with no filters on the generated molecules, the number of clusters containing fewer than 10 molecules out of 100 clusters are 4, 3, 5, 2, and 3 for each respective iteration (see Figures S19.1−S19.4). We then randomly sampled from the clusters with more than 10 molecules until we achieved a set of 1000 molecules.

8.3. Scoring Function. Our scoring function considers attractive protein−ligand contact points using the ProLIF software package[96] and assigns handpicked weights for each interaction type: hydrophobic interactions are scored at 2.5, hydrogen-bond interactions at 3.5, ionic interactions at 7.5, interactions between aromatic rings and cations at 2.5, van der Waals interactions at 1.0, halogen-bond interactions at 3.0, face-to-face π-stacking interactions at 3.0, edge-to-face π-stacking interactions at 1.0, and metallic complexation interactions at 3.0.

We assessed our scoring function with the PDBbind v2020 refined set, which contains 5316 unique experimentally determined protein−ligand binding complexes with high-quality labels and structures.[92] We found that there is a positive Pearson correlation of 0.32 between the scores derived from our scoring function and the experimentally determined binding affinities (see Figure S20.1A), supporting that our scoring function is an approximate yet meaningful estimate of binding ability. Furthermore, we found that 99.6% of the complexes achieved the score threshold of 11 (see Figure S20.1B).

The optimal weights for the interaction types may vary significantly depending on the specific target, and therefore, the scoring function employed in this work should be considered a crude estimation. However, the positive correlation with experimentally determined binding affinities supports its utility as a heuristic approximation to binding ability. Moreover, it can be replaced with a more precise metric as long as the replacement metric correlates with the descriptors used to construct the chemical space proxy.

8.4. Sampling Algorithm. After scoring each of the 1000 protein−ligand pairs, we selected N replicas of each molecule that scored equal to or above the score threshold, where N is the smallest integer that achieves a total number of molecules of at least 5000. We then calculated mean cluster scores from the scored molecules, which were converted to sampling fractions with the softmax function. We also considered other methods for converting cluster scores to sampling fractions, and the results are reported for each method attempted in Figures S21.1−S21.17. We then converted $f_i \times 5000$ to an integer (where f_i is the calculated fraction for sampling from cluster i) and sampled the corresponding number of molecules randomly from each respective cluster. If a given cluster had fewer molecules than would satisfy the calculated fraction, we distributed the surplus among the other clusters relative to their sampling fractions. We combined these 5000 molecules with the replicas of molecules that met the scoring threshold to generate an AL training set of approximately 10 000 molecules.

■ ASSOCIATED CONTENT

Data Availability Statement

All of our software is available as open source at https://github.com/batistagroup/ChemSpaceAL. Additionally, the ChemSpaceAL Python package is available via PyPI at https://pypi.org/project/ChemSpaceAL/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01456.

Details regarding the ADMET and functional group filters used, additional analyses pertaining to the FDA-approved inhibitors of c-Abl kinase, the architecture and performance of the pretrained model, and different

https://doi.org/10.1021/acs.jcim.3c01456
J. Chem. Inf. Model. 2024, 64, 653−665

approaches that were investigated for specific components of the methodology (PDF)

AUTHOR INFORMATION

Corresponding Author

Victor S. Batista − *Department of Chemistry, Yale University, New Haven, Connecticut 06511-8499, United States;* orcid.org/0000-0002-3262-1237; Phone: (203) 432-6672; Email: victor.batista@yale.edu

Authors

Gregory W. Kyro − *Department of Chemistry, Yale University, New Haven, Connecticut 06511-8499, United States;* orcid.org/0000-0002-0095-8548

Anton Morgunov − *Department of Chemistry, Yale University, New Haven, Connecticut 06511-8499, United States;* orcid.org/0009-0004-6245-0354

Rafael I. Brent − *Department of Chemistry, Yale University, New Haven, Connecticut 06511-8499, United States;* orcid.org/0000-0002-3233-7914

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01456

Author Contributions

G.W.K., A.M., and R.I.B. contributed to this work equally. G.W.K., A.M., and R.I.B. designed the research, developed the software, published the Python package, and performed the research. G.W.K., A.M., R.I.B., and V.S.B. analyzed data. G.W.K., A.M., and R.I.B. wrote the paper. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge the financial support from the National Institutes of Health under Grant R01GM136815 as well as from the National Science Foundation under Grant DGE-2139841. V.S.B. also acknowledges the high-performance computer time from the National Energy Research Scientific Computing Center and from the Yale University Faculty of Arts and Sciences High Performance Computing Center.

ABBREVIATIONS

AI, artificial intelligence; RNNs, recurrent neural networks; GNNs, generative adversarial networks; AL, active learning; GPT, generative pretrained transformer; Cas9, CRISPR-associated protein 9; SMILES, Simplified Molecular Input Line Entry System; PCA, principal component analysis; ADMET, absorption, distribution, metabolism, excretion, and toxicity; T_C, Tanimoto similarity; LogP, logarithm of the partition coefficient; t-SNE, t-distributed stochastic neighbor embedding; NaN, not a number; MQNs, molecular quantum numbers

REFERENCES

(1) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133*, 9181−9183.

(2) Sadybekov, A. V.; Katritch, V. Computational approaches streamlining drug discovery. *Nature* **2023**, *616*, 673−685.

(3) Segler, M. H. S.; 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.

(4) Urbina, F.; Lowden, C. T.; Culberson, J. C.; Ekins, S. MegaSyn: Integrating Generative Molecular Design, Automated Analog Designer, and Synthetic Viability Prediction. *ACS Omega* **2022**, *7*, 18699−18713.

(5) 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.* **2017**, *37*, 1700111 DOI: 10.1002/minf.201700111.

(6) Xu, M.; Ran, T.; Chen, H. De Novo Molecule Design Through the Molecular Generative Model Conditioned by 3D Information of Protein Binding Sites. *J. Chem. Inf. Model.* **2021**, *61*, 3240−3254.

(7) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 Chemical Space Using Deep Generative Models. *J. Cheminf.* **2019**, *11*, 20.

(8) Yonchev, D.; Bajorath, J. DeepCOMO: From Structure-Activity Relationship Diagnostics to Generative Molecular Design Using the Compound Optimization Monitor Methodology. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 1207−1218.

(9) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 1175−1183.

(10) Zhang, J.; Chen, H. De Novo Molecule Design Using Molecular Generative Models Constrained by Ligand−Protein Interactions. *J. Chem. Inf. Model.* **2022**, *62*, 3291−3306.

(11) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *J. Cheminf.* **2019**, *11*, 71.

(12) Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative Molecular Design in Low Data Regimes. *Nat. Mach. Intell.* **2020**, *2*, 171−180.

(13) Li, X.; Xu, Y.; Yao, H.; Lin, K. Chemical Space Exploration Based on Recurrent Neural Networks: Applications in Discovering Kinase Inhibitors. *J. Cheminf.* **2020**, *12*, 42.

(14) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37*, 1700153 DOI: 10.1002/minf.201700153.

(15) Tan, X.; Jiang, X.; He, Y.; Zhong, F.; Li, X.; Xiong, Z.; Li, Z.; Liu, X.; Cui, C.; Zhao, Q.; Xie, Y.; Yang, F.; Wu, C.; Shen, J.; Zheng, M.; Wang, Z.; Jiang, H. Automated Design and Optimization of Multitarget Schizophrenia Drug Candidates by Deep Learning. *Eur. J. Med. Chem.* **2020**, *204*, 112572.

(16) Bjerrum, E. J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). *ArXiv* **2017**, 1.

(17) Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct Steering of De Novo Molecular Generation with Descriptor Conditional Recurrent Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 254−265.

(18) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design Through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, 48.

(19) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.* **2018**, *4*, No. eaap7885.

(20) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-Assisted Reinforcement Learning for Diverse Molecular De Novo Design. *J. Cheminf.* **2020**, *12*, 68.

(21) Yoshimori, A.; Kawasaki, E.; Kanai, C.; Tasaka, T. Strategies for Design of Molecular Structures with a Desired Pharmacophore Using Deep Reinforcement Learning. *Chem. Pharm. Bull.* **2020**, *68*, 227−233.

(22) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 5918−5922.

(23) Korshunova, M.; Huang, N.; Capuzzi, S.; Radchenko, D. S.; Savych, O.; Moroz, Y. S.; Wells, C. I.; Willson, T. M.; Tropsha, A.; Isayev, O. Generative and Reinforcement Learning Approaches for the Automated De Novo Design of Bioactive Compounds. *Commun. Chem.* **2022**, *5*, 129.

(24) Popova, M.; Shvets, M.; Oliva, J.; Isayev, O. MolecularRNN: Generating Realistic Molecular Graphs With Optimized Properties. *ArXiv* **2019**, 1.
(25) Ghaemi, M. S.; Grantham, K.; Tamblyn, I.; Li, Y.; Ooi, H. K. Generative Enriched Sequential Learning (ESL) Approach for Molecular Design via Augmented Domain Knowledge. In *Proceedings of the 35th Canadian Conference on Artificial Intelligence*, 2022.
(26) Bian, Y.; Wang, J.; Jun, J. J.; Xie, X. Q. Deep Convolutional Generative Adversarial Network (dcGAN) Models for Screening and Design of Small Molecules Targeting Cannabinoid Receptors. *Mol. Pharmaceutics* **2019**, *16*, 4451−4460.
(27) Méndez-Lucio, O.; Baillif, B.; Clevert, D. A.; Rouquié, D.; Wichard, J. De Novo Generation of Hit-Like Molecules From Gene Expression Signatures Using Artificial Intelligence. *Nat. Commun.* **2020**, *11*, 10.
(28) Cao, N. D.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. *ArXiv* **2018**, 1.
(29) Tsujimoto, Y.; Hiwa, S.; Nakamura, Y.; Oe, Y.; Hiroyasu, T. L-MolGAN: An Improved Implicit Generative Model for Large Molecular Graphs. *ChemRxiv* **2021**, 1.
(30) Wang, J.; Chu, Y.; Mao, J.; Jeon, H.-N.; Jin, H.; Zeb, A.; Jang, Y.; Cho, K.-H.; Song, T.; No, K. T. De Novo Molecular Design With Deep Molecular Generative Models for PPI inhibitors. *Briefings Bioinf.* **2022**, *23*, bbac285.
(31) Song, T.; Ren, Y.; Wang, S.; Han, P.; Wang, L.; Li, X.; Rodriguez-Patón, A. DNMG: Deep Molecular Generative Model by Fusion of 3D Information for De Novo Drug Design. *Methods* **2023**, *211*, 10−22.
(32) Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: A Soft tool For 3D Drug Design of Protein Targets by Artificial Intelligence and Classical Algorithm. *Briefings Bioinf.* **2021**, *22* (3), bbaa161 DOI: 10.1093/bib/bbaa161.
(33) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194−1204.
(34) Lee, Y. J.; Kahng, H.; Kim, S. B. Generative Adversarial Networks for De Novo Molecular Design. *Mol. Inf.* **2021**, *40*, 2100045.
(35) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G.; Aspuru-Guzik, A. Optimizing Distributions Over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). *ChemRxiv* **2017**, 1.
(36) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular De Novo Design. *Mol. Pharmaceutics* **2018**, *15*, 4386−4397.
(37) Skalic, M.; Sabbadin, D.; Sattarov, B.; Sciabola, S.; De Fabritiis, G. From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design. *Mol. Pharmaceutics* **2019**, *16*, 4282−4291.
(38) 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*, 74.
(39) Abbasi, M.; Santos, B. P.; Pereira, T. C.; Sofia, R.; Monteiro, N. R. C.; Simões, C. J. V.; Brito, R. M. M.; Ribeiro, B.; Oliveira, J. L.; Arrais, J. P. Designing Optimized Drug Candidates With Generative Adversarial Network. *J. Cheminf.* **2022**, *14*, 40.
(40) 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.
(41) Sousa, T.; Correia, J.; Pereira, V.; Rocha, M. Combining Multi-objective Evolutionary Algorithms with Deep Generative Models Towards Focused Molecular Design. In *Applications of Evolutionary Computation: 24th International Conference, EvoApplications 2021*; Springer, 2021; p 81−96. DOI: 10.1007/978-3-030-72699-7_6.
(42) Chenthamarakshan, V.; Das, P.; Hoffman, S. C.; Strobelt, H.; Padhi, I.; Lim, K. W.; Hoover, B.; Manica, M.; Born, J.; Laino, T.; Mojsilovic, A. CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; NeurIPS, 2020.
(43) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for De Novo Molecular Design. *J. Cheminf.* **2018**, *10*, 31.
(44) Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. In *Artificial Neural Networks and Machine Learning − ICANN 2018*; Springer, 2018; p 412−422. DOI: 10.1007/978-3-030-01418-6_41.
(45) Wang, S.; Song, T.; Zhang, S.; Jiang, M.; Wei, Z.; Li, Z. Molecular Substructure Tree Generative Model for De Novo Drug Design. *Briefings Bioinf.* **2022**, *23* (2), bbab592 DOI: 10.1093/bib/bbab592.
(46) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59*, 43−52.
(47) Samanta, B.; De, A.; Jana, G.; Chattaraj, P. K.; Ganguly, N.; Rodriguez, M. G. *NeVAE: A Deep Generative Model for Molecular Graphs. Proceedings of the AAAI Conference on Artificial Intelligence* **2019**, *33*, 1110−1117.
(48) Lim, J.; Hwang, S.-Y.; Moon, S.; Kim, S.; Kim, W. Y. Scaffold-Based Molecular Design With a Graph Generative Model. *Chem. Sci.* **2020**, *11*, 1153−1164.
(49) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR, 2018; p 2323−2332.
(50) Dollar, O.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Attention-Based Generative Models for De Novo Molecular Design. *Chem. Sci.* **2021**, *12*, 8362−8372.
(51) Krishnan, S. R.; Bung, N.; Vangala, S. R.; Srinivasan, R.; Bulusu, G.; Roy, A. De Novo Structure-Based Drug Design Using Deep Learning. *J. Chem. Inf. Model.* **2022**, *62*, 5100−5109.
(52) Filella-Merce, I.; Molina, A.; Orzechowski, M.; Díaz, L.; Zhu, Y.; Mor, J.; Malo, L.; Yekkirala, A.; Ray, S.; Guallar, V. Optimizing Drug Design by Merging Generative AI With Active Learning Frameworks. *ArXiv* **2023**, 1.
(53) 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.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.
(54) Abeer, A. N. M. N.; Urban, N.; Weil, M. R.; Alexander, F. J.; Yoon, B.-J. Multi-objective Latent Space Optimization of Generative Molecular Design Models. *ArXiv* **2022**, 1.
(55) Nesterov, V.; Wieser, M.; Roth, V. 3DMolNet: A Generative Network for Molecular Structures. *ArXiv* **2020**, 1.
(56) Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59*, 1205−1214.
(57) Hong, S. H.; Ryu, S.; Lim, J.; Kim, W. Y. Molecular Generative Model Based on an Adversarially Regularized Autoencoder. *J. Chem. Inf. Model.* **2020**, *60*, 29−36.
(58) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2017**, *8*, 10883.
(59) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for De Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14*, 3098−3104.
(60) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.;

Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for De Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398−4405.
(61) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient Multi-Objective Molecular Optimization in a Continuous Latent Space. *Chem. Sci.* **2019**, *10*, 8016−8024.
(62) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682−5698.
(63) Sattarov, B.; Baskin, I. I.; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **2019**, *59*, 1182−1196.
(64) Mao, J.; Wang, J.; Zeb, A.; Cho, K.-H.; Jin, H.; Kim, J.; Lee, O.; Wang, Y.; No, K. T. Transformer-Based Molecular Generative Model for Antiviral Drug Design. *J. Chem. Inf. Model.* **2023**, DOI: 10.1021/acs.jcim.3c00536.
(65) Wei, L.; Fu, N.; Song, Y.; Wang, Q.; Hu, J. Probabilistic Generative Transformer Language Models for Generative Design of Molecules. *ArXiv* **2022**, 1.
(66) Wang, J.; Mao, J.; Wang, M.; Le, X.; Wang, Y. Explore Drug-Like Space With Deep Generative Models. *Methods* **2023**, *210*, 52−59.
(67) Grechishnikova, D. Transformer Neural Network for Protein-Specific De Novo Drug Generation as a Machine Translation Problem. *Sci. Rep.* **2021**, *11*, 321.
(68) Kim, H.; Na, J.; Lee, W. B. Generative Chemical Transformer: Neural Machine Learning of Molecular Geometric Structures from Chemical Language via Attention. *J. Chem. Inf. Model.* **2021**, *61*, 5804−5814.
(69) Wang, W.; Wang, Y.; Zhao, H.; Sciabola, S. A Transformer-Based Generative Model for De Novo Molecular Design. *ArXiv* **2022**, 1.
(70) Chen, Y.; Wang, Z.; Wang, L.; Wang, J.; Li, P.; Cao, D.; Zeng, X.; Ye, X.; Sakurai, T. Deep Generative Model for Drug Design From Protein Target Sequence. *J. Cheminf.* **2023**, *15*, 38.
(71) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* **2022**, *62*, 2064−2076.
(72) Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-Based Enumeration, Active Learning, and Free Energy Calculations To Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *J. Chem. Inf. Model* **2019**, *59*, 3782−3793.
(73) Khalak, Y.; Tresadern, G.; Hahn, D. F.; de Groot, B. L.; Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J. Chem. Theory Comput.* **2022**, *18*, 6259−6270.
(74) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **2021**, *12*, 7866−7881.
(75) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Maser, M.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing active learning for free energy calculations. *A. I. Life Sci.* **2022**, *2*, 100050.
(76) Bailey, M.; Moayedpour, S.; Li, R.; Corrochano-Navarro, A.; Kotter, A.; Kogler-Anele, L.; Riahi, S.; Grebner, C.; Hessler, G.; Matter, H.; Bianciotto, M.; Mas, P.; Bar-Joseph, Z.; Jager, S. Deep Batch Active Learning for Drug Discovery. *eLife* **2023**, *12*, RP89679.
(77) Dodds, M.; Guo, J.; Löhr, T.; Tibo, A.; Engkvist, O.; Janet, J. P. Sample Efficient Reinforcement Learning with Active Learning for Molecular Design. *ChemRxiv* **2023**, 1.
(78) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236−4243.
(79) East, K. W.; Newton, J. C.; Morzan, U. N.; Narkhede, Y. B.; Acharya, A.; Skeens, E.; Jogl, G.; Batista, V. S.; Palermo, G.; Lisi, G. P. Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *J. Am. Chem. Soc.* **2020**, *142*, 1348−1358.
(80) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017.
(81) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.
(82) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. Guacamol: Benchmarking Models for De Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096−1108.
(83) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644.
(84) Sterling, T.; Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(85) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−201.
(86) Balani, S. K.; Miwa, G.; Gan, L.-S.; Wu, J.-T.; Lee, F. Strategy of Utilizing In Vitro and In Vivo ADME Tools for Lead Optimization and Drug Candidate Selection. *Curr. Top. Med. Chem.* **2005**, *5* (11), 1033.
(87) Rossari, F.; Minutolo, F.; Orciuolo, E. Past, present, and future of Bcr-Abl inhibitors: from chemical development to clinical efficacy. *J. Hematol. Oncol.* **2018**, *11*, 84.
(88) Schoepfer, J.; Jahnke, W.; Berellini, G.; Buonamici, S.; Cotesta, S.; Cowan-Jacob, S. W.; Dodd, S.; Drueckes, P.; Fabbro, D.; Gabriel, T.; Furet, P.; et al. Discovery of Asciminib (ABL001), an Allosteric Inhibitor of the Tyrosine Kinase Activity of BCR-ABL1. *J. Med. Chem.* **2018**, *61* (18), 8120−8135.
(89) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(90) RDKIT: Open-Source Cheminformatics Software. https://www.rdkit.org.
(91) Wang, J.; Skeens, E.; Arantes, P. R.; Maschietto, F.; Allen, B.; Kyro, G. W.; Lisi, G. P.; Palermo, G.; Batista, V. S. Structural Basis for Reduced Dynamics of Three Engineered HNH Endonuclease Lys-to-Ala Mutants for the Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-Associated 9 (CRISPR/Cas9) Enzyme. *Biochem.* **2022**, *61*, 785−794.
(92) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.
(93) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *ArXiv* **2023**, 1.
(94) Liu, H.; Li, Z.; Hall, D.; Liang, P.; Ma, T. Sophia: A Scalable Stochastic Second-Order Optimizer for Language Model Pre-training. *ArXiv* **2023**, 1.
(95) Chen, D.; Gao, K.; Nguyen, D. D.; Chen, X.; Jiang, Y.; Wei, G.-W.; Pan, F. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* **2021**, *12*, 3521.
(96) Bouysset, C.; Fiorucci, S. ProLIF: A Library to Encode Molecular Interactions as Fingerprints. *J. Cheminf.* **2021**, *13*, 72.