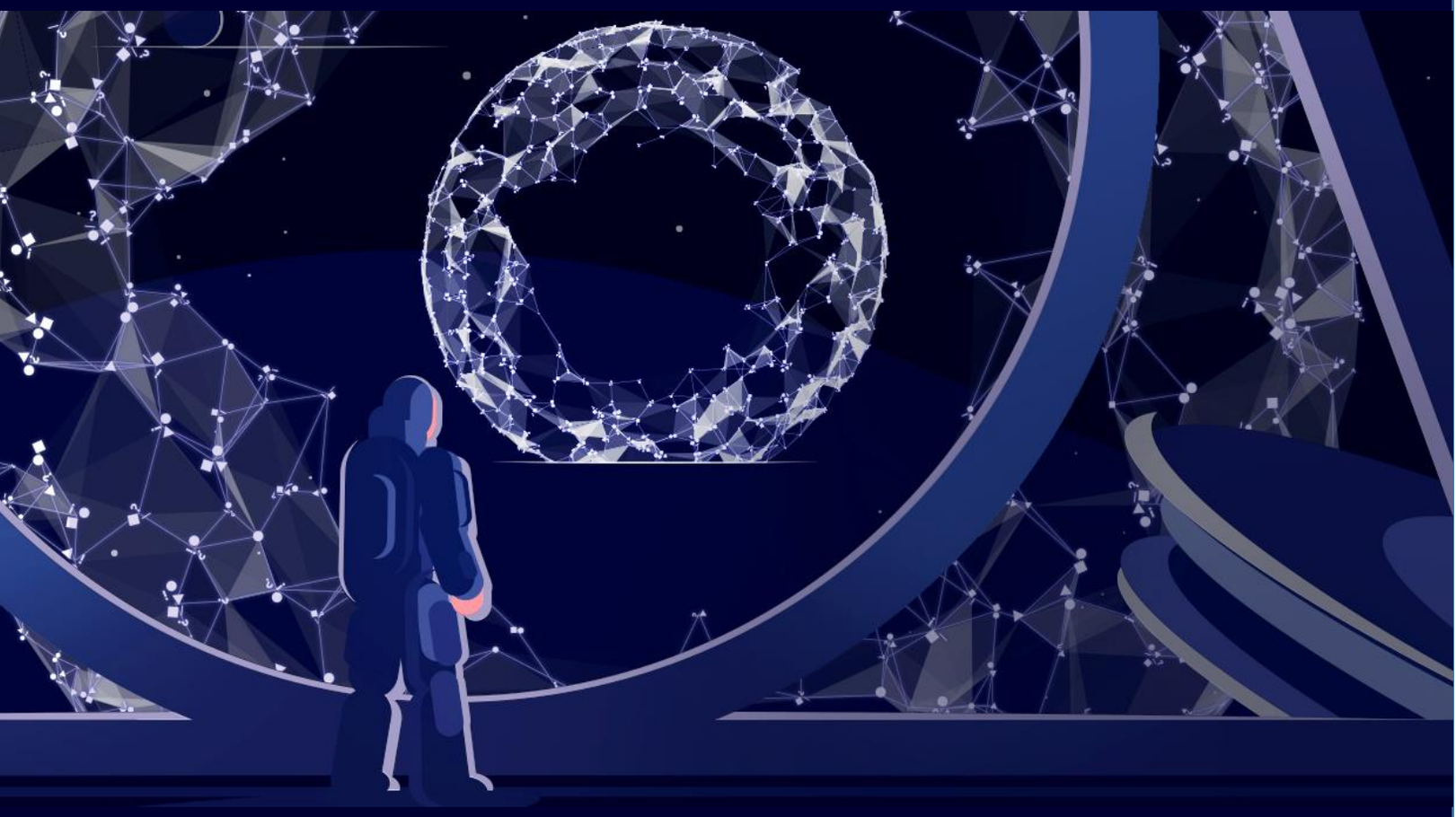
Chapter 3

An exploration strategy improves the diversity of *de novo* ligands using deep reinforcement learning: a case for the adenosine A_{2A} receptor



Xuhan Liu, Kai Ye, Herman W. T. van Vlijmen, Adriaan P. IJzerman and Gerard J. P. van Westen*. Journal of Cheminformatics (2019). Https://doi.org/10.1186/s13321-019-0355-6

Abstract

Over the last five years deep learning has progressed tremendously in both image recognition and natural language processing. Now it is increasingly applied to other data rich fields. In drug discovery, recurrent neural networks (RNNs) have been shown to be an effective method to generate novel chemical structures in the form of SMILES. However, ligands generated by current methods have so far provided relatively low diversity and do not fully cover the whole chemical space occupied by known ligands. Here, we propose a new method (DrugEx) to discover *de novo* drug-like molecules. DrugEx is an RNN model (generator) trained through a special exploration strategy integrated into reinforcement learning. As a case study we applied our method to design ligands against the adenosine A_{2A} receptor. From ChEMBL data, a machine learning model (predictor) was created to predict whether generated molecules are active or not. Based on this predictor as the reward function, the generator was trained by reinforcement learning without any further data. We then compared the performance of our method with two previously published methods, REINVENT and ORGANIC. We found that candidate molecules our model designed, and predicted to be active, had a larger chemical diversity and better covered the chemical space of known ligands compared to the state-of-the-art.

Keywords: deep learning; adenosine receptors; cheminformatics; reinforcement learning; exploration strategy.

3.1. Introduction

G Protein-Coupled Receptors (GPCRs) are the largest family of cell membrane-bound proteins [1], containing about 800 members encoded by approximately 4% of human genes. GPCRs are central to a large number of essential biological processes, including cell proliferation, cell survival, and cell motility [2]. Currently, GPCRs form the main target of approximately 34% of all FDA approved drugs [3]. One of the most extensively studied GPCRs is the human adenosine A_{2A} receptor (A_{2A}AR), which has been shown to be a promising drug target for Parkinson's disease, cardiovascular diseases, and inflammatory disorders [4]. Multiple crystal structures with different ligands have been resolved [5,6], and data on the biological activity of thousands of chemical compounds against the receptor was made available in the public ChEMBL database [7]. Considering the amount of data available and our in-house expertise we exploited machine learning methods to design novel ligands with predicted affinity for the A_{2A}AR.

Over the last years, deep learning (DL) has been at the forefront of great breakthroughs in the field of artificial intelligence and its performance even surpassed human abilities for image recognition and natural language processing [8]. Since then, deep learning is gradually being applied to other data rich fields [9,10]. In drug discovery DL has been used to construct quantitative structure-activity relationship (QSAR) models [11] to predict the properties of chemical compounds, such as toxicity, partition coefficient, affinity for specific targets, *etc.* [12,13]. Most commonly pre-defined descriptors such as Extended Connectivity Fingerprint (ECFP) [14] were used as input to construct fully-connected neural networks [15]. More recently studies were published using other methods wherein neural networks extract the descriptor from chemical structures automatically and directly, such as Mol2Vec [16], DruGAN [17], GraphConv [18], *etc*.

In addition to these *prediction* applications, DL can also be used in chemical structure *generation* [13]. Gupta *et al*. constructed a recurrent neural network (RNN) model to learn the syntax of the SMILES notation and generate novel SMILES for molecules representation [19]. In addition, Olivecrona *et al.* combined RNNs and reinforcement

learning (RL) to generate SMILES formatted molecules that have required chemical and biological properties [20]. RL has been instrumental in the construction of "AlphaGo" designed by DeepMind, which defeated one of the best human Go players [21]. Finally, similar to generative adversarial networks (GANs) for generating images [22], Benjamin *et al.* exploited the GAN for a sequence generation model [23] to generate molecules with multi-objective optimization [24].

In order to maximize the chance to find interesting hits for a given target, generated drug candidates should a) be chemically diverse, b) possess biological activity, and c) contain similar physicochemical properties or chemical scaffolds to already known ligands [25]. Although several groups have studied the application of DL for generating molecules as drug candidates, most current generative models cannot satisfy all of these three conditions simultaneously [26]. Considering the variety of the structure and function of GPCRs and the huge space of drug candidates, it is impossible to enumerate all possible virtual molecules in advance [27]. Here we aimed to discover *de novo* drug-like molecules against $A_{2A}AR$ by our proposed new method DrugEx in which an exploration strategy was integrated into a RL model. The integration of this function ensured that our model generated candidate molecules similar to known ligands of $A_{2A}AR$ with great diversity and predicted affinity for the $A_{2A}AR$. All python code for this study is freely available at http://github.com/XuhanLiu/DrugEx.

3.2. Dataset and methods

3.2.1. Data source

Drug-like molecules were collected from the ZINC database (version 15) [28]. We randomly chose approximately one million SMILES formatted molecules that met the following criteria: $-2 < \log P < 6$ and $200 <$ molecular weight (MW) < 600. The dataset (named *ZINC* hereafter) finally contained 1,018,517 molecules, and was used for SMILES syntax learning. Furthermore, we extracted the known ligands for the $A_{2A}AR$ (ChEMBL identifier: CHEMBL251) from ChEMBL (version 23) [29]. If multiple measurements for the same ligands existed, the average pCHEMBL value (pKi or pIC50 value) was

calculated and duplicate items were removed. If the pCHEMBL value < 6.5 or the compound was annotated as "Not Active" it was regarded as a negative sample; otherwise, it was regarded as a positive sample. In the end this dataset (named as *A2AR*) contained 2,420 positive samples and 2,562 negative samples.

3.2.2. Prediction model (QSAR)

Binary classification through QSAR modelling was used as prediction task. Input data for the model were ECFP6 fingerprints with 4096 bits calculated by the RDKit Morgan Fingerprint algorithm with a three-bond radius [30]. Hence, each molecule in the dataset was transformed into a 4096D vector. Model output value was the probability whether a given chemical compound was active based on this vector. Four algorithms were benchmarked for model construction, Random Forest (RF), Support Vector Machine (SVM), Naïve Bayesian (NB), and Deep Neural Network (DNN). The RF, SVM and NB models were implemented through Scikit-Learn [31], and DNN through PyTorch [32]. In RF, the number of trees was set as 1000 and split criterion was "*gini*". In SVM, a radial basis function (RBF) kernel was used and the parameter space of C and γ were set as $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{5}]$, respectively. In DNN, the architecture contained three hidden layers activated by rectified linear unit (ReLU) between input and output layers (activated by sigmoid function), the number of neurons were 4096, 8000, 4000, 2000 and 1 for each layer. With 100 epochs of training process 20% of hidden neurons were randomly dropped out between each layer. The binary cross entropy was used to construct the loss function and optimized by Adam [33] with a learning rate of 10^{-3}. The AUC of ROC curves was calculated to compare their mutual performance.

3.2.3. Generative model

Starting from the SMILES format, each molecule in the *ZINC* dataset was split into a series of tokens, standing for different types of atoms and bonds. Then, all tokens existing in this dataset were collected to construct the SMILES vocabulary. The final vocabulary contained 56 tokens (Table S3.1) which were selected and arranged sequentially into valid SMILES sequence following the correct grammar.

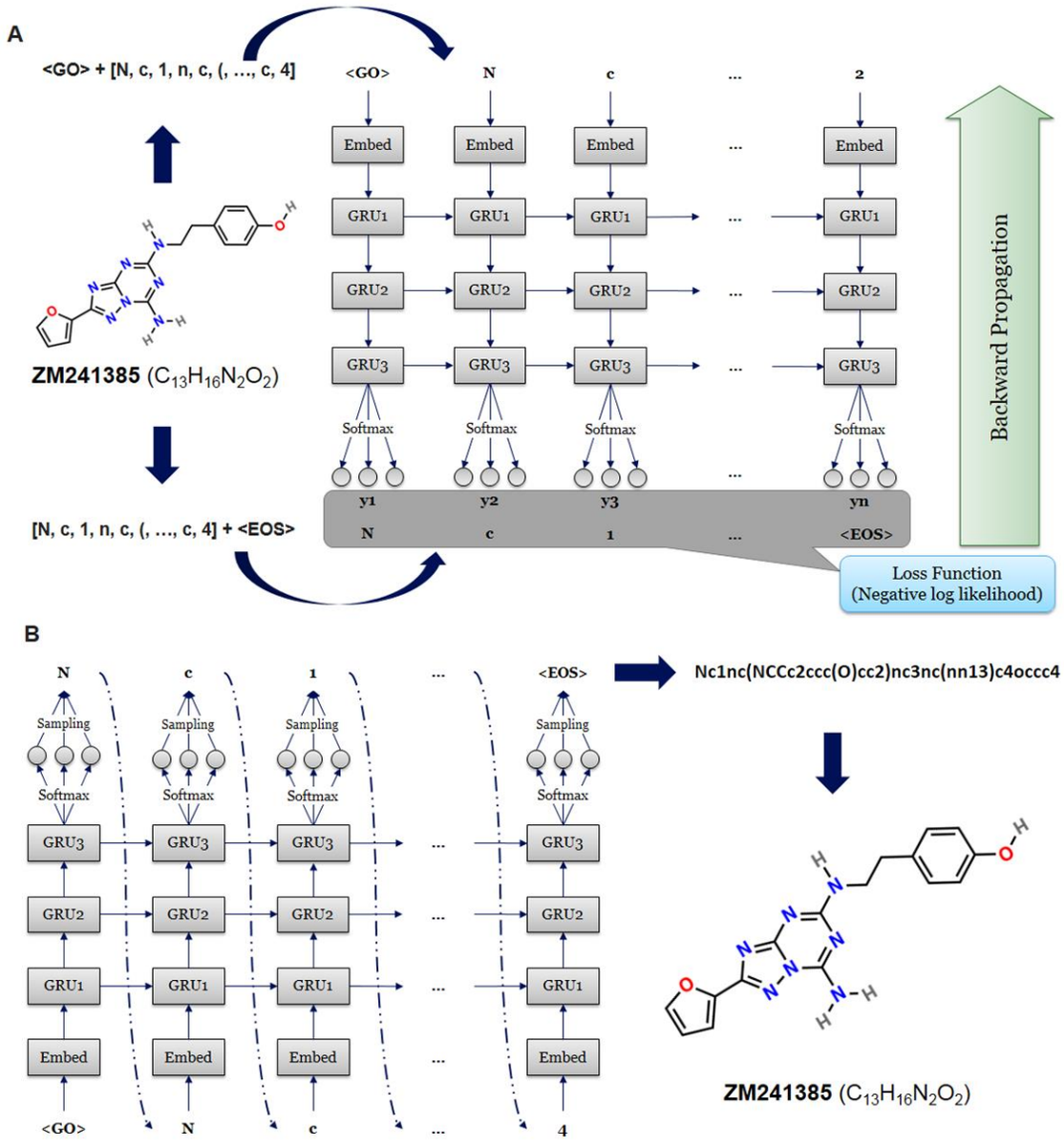


Fig. 3.1: Architecture of recurrent neural networks for training and sampling processes with A_{2A}AR antagonist ZM241385 as an example. (A) In the training process of RNNs, each molecule is decomposed to a series of tokens and then taken as input. Subsequently, the input and output are combined with a start token and an end token, respectively. (B) Beginning with the start token "GO", the model calculates the probability distribution of each token in the vocabulary. For each step, one of the available tokens is randomly chosen based on the probability distribution and is again received by RNNs as input to calculate the new probability distribution for the next step. This process will end if the end token "EOS" is sampled or the number of steps equals 100.

The RNN model constructed for sequence generation contained six layers: one input layer, one embedding layer, three recurrent layers and one output layer (Fig. 3.1). After being represented by a sequence of tokens, molecules can be received as categorical features by

the input layer. In the embedding layer, vocabulary size, and embedding dimension were set to 56 and 128, meaning each token could be transformed into a 128d vector. For the recurrent layer, gated recurrent unit (GRU) [34] was used as the recurrent cell with 512 hidden neurons. The output at each position was the probability that determined which token in the vocabulary would be chosen to construct the SMILES string.

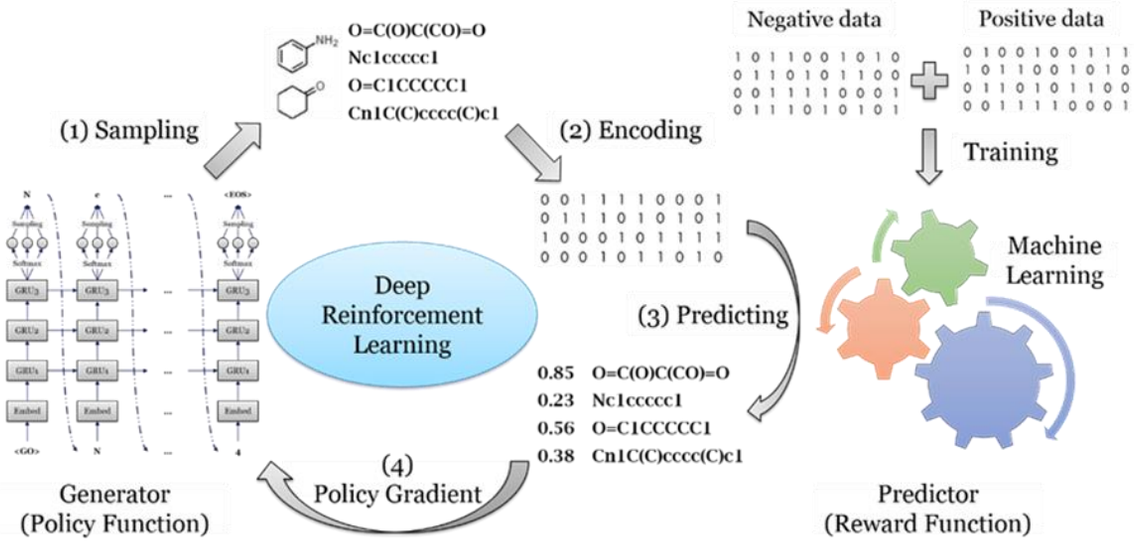


Fig. 3.2: The workflow of deep reinforcement learning. For each loop, it contains several steps: (1) a batch of SMILES sequences was sampled by the generator, which had been initialized by a pre-trained model; (2) each generated molecule represented by this SMILES format was encoded into a fingerprint; (3) a probability score of activity on the $A_{2A}AR$ was assigned to each molecule, calculated by the QSAR model which had been trained in advance; (4) all of the generated molecules and their scores were sent back for training of the generator with the policy gradient method.

During the training process we put the start token at the beginning of a batch of data as input and the end token at the end of the same batch of data as output. This ensures that the generative network could choose correct tokens based on the sequence it had generated (Fig. 3.1A). A negative log likelihood function was used to construct the loss function to guarantee that the token in the output sequence had the largest probability to be chosen after being trained. In order to optimize the parameters of the model, the Adam algorithm [33] was used for optimization of loss function. Here, the learning rate was set at 10^{-3}, batch size was 500, and training steps set at 1000 epochs.

3.2.4. Reinforcement learning

SMILES sequence construction under the RL framework can be viewed as a series of decision-making processes (Fig. 3.2). At each step, the model determines the optimal token from the vocabulary based on the generated sequence in previous steps. However, the pure RNN model cannot guarantee that the percentage of desired molecules (i.e. biologically active on the A_{2A}AR) being generated is as large as possible. To solve this problem RL is an appropriate method because it increases the probability of those molecules with higher rewards and avoids generating those molecules with lower rewards. We regarded the generator as the policy function and the predictor as the reward function. The generator G_θ was updated by employing a policy gradient on the basis of the expected end reward received from the predictor *Q*. The objective function could be designated as generating a sequence from the start state to maximize the expected end reward [23].

$$J(\theta) = E[R(y_{1:T})|\theta] = \sum_{t=1}^{T} G_\theta(y_t|y_{1:t-1}) \cdot (Q(y_{1:T}) - \beta)$$

Here *R* is the reward for a complete sequence which is given by the prediction model *Q*; the generative model G_θ can be regarded as policy function to determine the probability of each token from the vocabulary to be chosen. The parameter β was the baseline of the reward, meaning that if the reward score was not larger than the baseline, the model would take it as a minus score or punishment. The goal of the generative model is to construct a sequence which can obtain the highest score judged by the predictor.

3.2.5. Exploration strategy

In order to improve the diversity of generated molecules, the token selection was not only determined by the generator constructed by the RNN model as described above, but also by a second fixed pre-trained RNN model (Fig. 3.3). The RNN requiring training is deemed the 'exploitation network' (G_θ) and the fixed RNN (not requiring training) is deemed the 'exploration network' (G_φ). Both had an identical network architecture. We define "exploring rate" (ε) in [0.0, 1.0] to determine which fraction of steps was determined by the exploration network. During the training process, each SMILES sequence was generated through the collaboration of these two RNNs. At each step a random number in [0.0, 1.0] was generated. If the value was smaller than ε, the G_φ would determine which

token to be chosen, and vice versa. After the training process was finished, we removed G_φ and only G_θ was left as the final model of DrugEx for molecule generation.

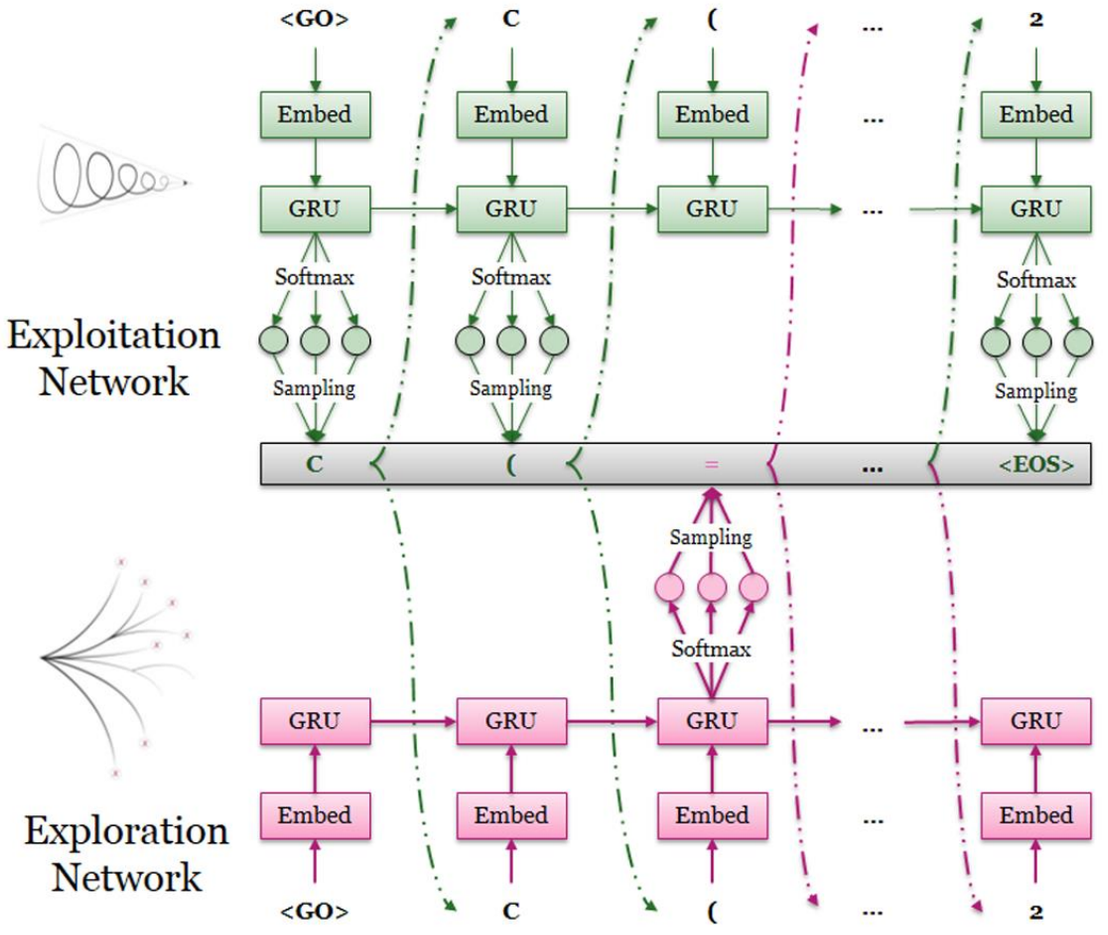


Fig. 3.3: Molecule generation with the assistance of the exploration strategy during the training process. For each step of token selection, a random variable was generated between 0 and 1. If the value is larger than a pre-set threshold (exploring rate, ε), the probability distribution is determined by the current generator (exploitation network, G_θ). Otherwise, it was determined by the exploration network (G_φ).

3.2.6. Molecular diversity

The Tanimoto-similarity was used for measuring the similarity of molecules. Given two compounds a and b and their ECFP6 fingerprints m_a and m_b, the Tanimoto-similarity is defined as:

$$T_s(a,b) = \frac{|m_a \cap m_b|}{|m_a \cup m_b|}$$

where $|m_a \cap m_b|$ represents the number of common fingerprint bits, and $|m_a \cup m_b|$ donates

the total number of fingerprint bits. The Tanimoto-distance is defined as:

$$T_d(a,b) = 1 - T_s(a,b)$$

Similar to Benhenda [26], the diversity I of a set of molecules A (with size of |A|) is defined as the average of the Tanimoto-distance of molecules of every pair of molecules:

$$I(A) = \frac{1}{|A|^2} \sum_{(a,b)\in A\times A} T_d(a,b)$$

In a given set of molecules, the less similar each two molecules are, the larger the value of its diversity will be.

3.3. Results and discussion

3.3.1. Performance of predictors

All molecules in the *A2AR* dataset were used for training the QSAR models, after being transformed into ECFP6 fingerprints. We then tested the performance of these different algorithms with five-fold cross validation of which the ROC curves are shown in Fig. 3.4. The RF model achieved the highest value of AUC, MCC, Sensitivity, and Accuracy, despite its Specificity being slightly lower than DNN. Hence this model was chosen as our predictor whose output would be regarded as the reward for the generator in RL. In our previous study [15], the performance of DNN was better than RF on the chemical space of the whole ChEMBL database. A possible reason for this difference can be that the size of the *A2AR* dataset and chemical diversity was much smaller than ChEMBL as a whole. This has a negative influence on DNN which had more parameters to be optimized than RF. Selecting the predictor was a critical step in this study, as this model would be used to determine whether the following generated molecules were active or inactive.

3.3.2. SMILES syntax learning

For the training of RNNs all molecules in the *ZINC* dataset were used as training set after being decomposed into the tokens which belonged to our vocabulary set. Here, we defined that a SMILES sequence was valid if it could be parsed by RDKit. During the training process, the percentage of valid SMILES sequence through 1,000 times sampling was calculated and was then recorded with the value of the loss function at each epoch (Fig.

3.5A). After about 300 epochs, the loss function had converged, indicating the model was trained well.

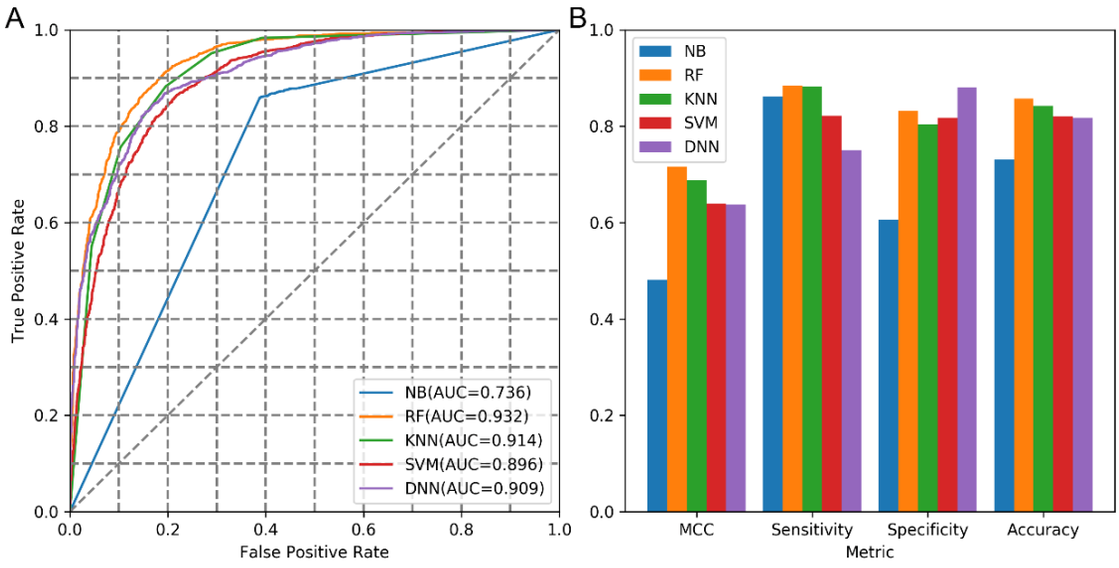


Fig. 3.4: Performance of five different machine learning models based on five-fold cross validation in A2AR dataset with different metrics, including AUC of ROC curve (A), MCC, Sensitivity, Specificity and Accuracy values (B). Except for specificity RF achieved highest scores among these models based on such measurements.

Subsequently, we sampled 10,000 SMILES sequences based on this well-trained model and found that 93.88% of these sequences were grammatically correct SMILES. We then compared some properties of these generated molecules with those in the training set, including number of hydrogen bond donors/acceptors, rotatable bonds, and different kind of ring systems (Fig. 3.6A). The distribution of these properties in the generated molecules highly resembles the molecules in the *ZINC* dataset. The logP ~ MW plot (Fig. 3.7A) shows that most generated molecules were drug-like molecules and cover the vast majority of the square space occupied by the *ZINC* dataset. Besides these eight properties, we also calculated 11 other physicochemical properties (including topological polar surface area, molar refractivity, the fraction of sp^3 hybridized carbon atoms and number of amide bonds, bridgehead atoms, heteroatoms, heavy atoms, spiroatoms, rings, saturated rings, valence electrons) to form 19D physicochemical descriptors (PhysChem). In addition, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were employed for dimension reduction and chemical space visualization with the PhysChem

and ECFP6 of these molecules, respectively. We found that generated molecules covered almost the whole region occupied by molecules in the *ZINC* dataset (Fig. 3.7 B and C) although the number of these generated molecules was less than 1% of the number of molecules in the *ZINC* dataset.

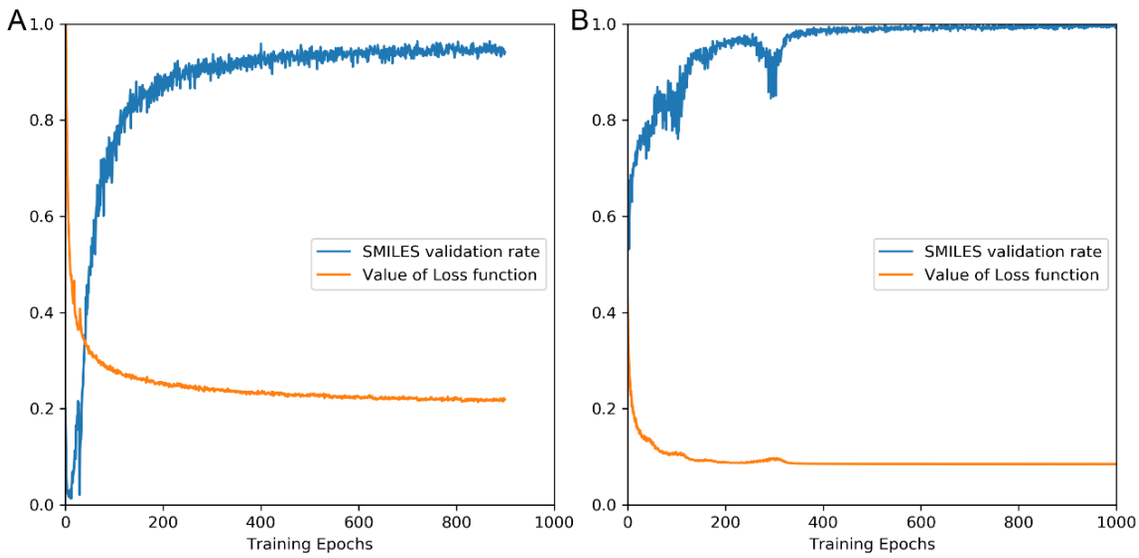


Fig. 3.5: The value of loss function and percentage of valid SMILES sequence during the pre-trained process on *ZINC* dataset (B) and fine-tuned process on *A2AR* dataset (B). The model was well pre-trained after 300 epochs and these two values converged to 0.19 and 93.88%, respectively. The performance of the fine-tuned model converged after 400 epochs with the two values reaching 0.09 and 0.99, respectively.

Subsequently we used the *A2AR* dataset to fine-tune this pre-trained model with 1,000 epochs (Fig. 3.5B). After sampling another 10,000 times, we performed the same comparison with the *A2AR* dataset with respect to the properties mentioned above (Fig. 3.6B) and investigated the chemical space represented by logP ~ MW (Fig. 3.7D), first two components of PCA (Fig. 3.7E) and t-SNE (Fig. 3.7F), yielding results similar to the model without fine-tuning but then focused on the *A2AR* chemical space. These results prove that RNN is an appropriate method to learn the SMILES grammar and to construct molecules similar to the ligands in the training set, which has also been shown in other work [35,19].

3.3.3. Conditional SMILES generation

The RNN model trained on the *ZINC* dataset was used as an initial state for the policy gradient in RL. After the training process of RL converged, 10,000 SMILES sequences

were generated for performance evaluation. However, after removal of duplicates in these sequences, only less than 10 unique molecules were left which were similar to compounds in the *A2AR* dataset. When checking the log file of the training process, we noticed that these duplicated sequences were frequently sampled at each epoch and its duplication rate increased gradually. In order to decrease the bias caused by these molecules with high frequency, we removed all duplicated sequences sampled at each epoch for training with the policy gradient. We found that almost all of the molecules generated according to this procedure were located outside of the drug-like region with regard to logP ~ MW plot (Figure S1). This problem might be caused by the bias of the predictor. ECFP is a substructure-based fingerprint, implying that as long as the molecule contains some critical substructures, it will be prone to be predicted as active. That was the reason why generated SMILES sequences contained a large number of repetitive motifs. Several research groups have made improvements to guarantee that the final model has ability to generate drug-like candidate molecules [24,20]. In the next section, we will describe our proposed method, "DrugEx" by integrating an exploration strategy to solve this problem and compare it to existing methods.

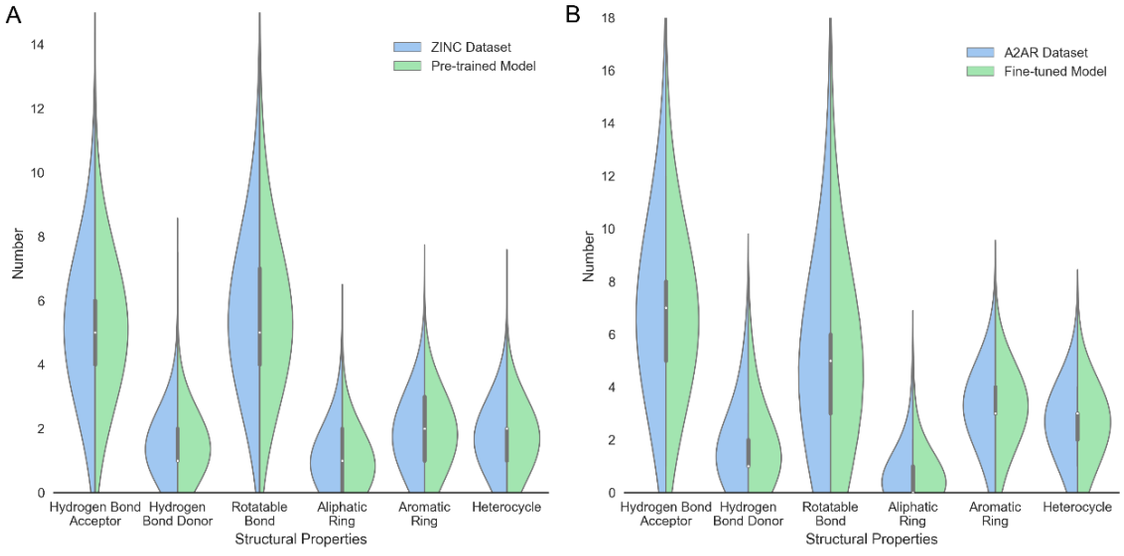


Fig. 3.6: Comparison of the properties of generated molecules by the pre-trained (A) and fine-tuned models (B) and molecules in *ZINC* dataset (A) and *A2AR* dataset (B), respectively. These properties included the number of acceptor/donor of hydrogen bonds, rotatable bonds, aliphatic rings, aromatic rings, and heterocycles.

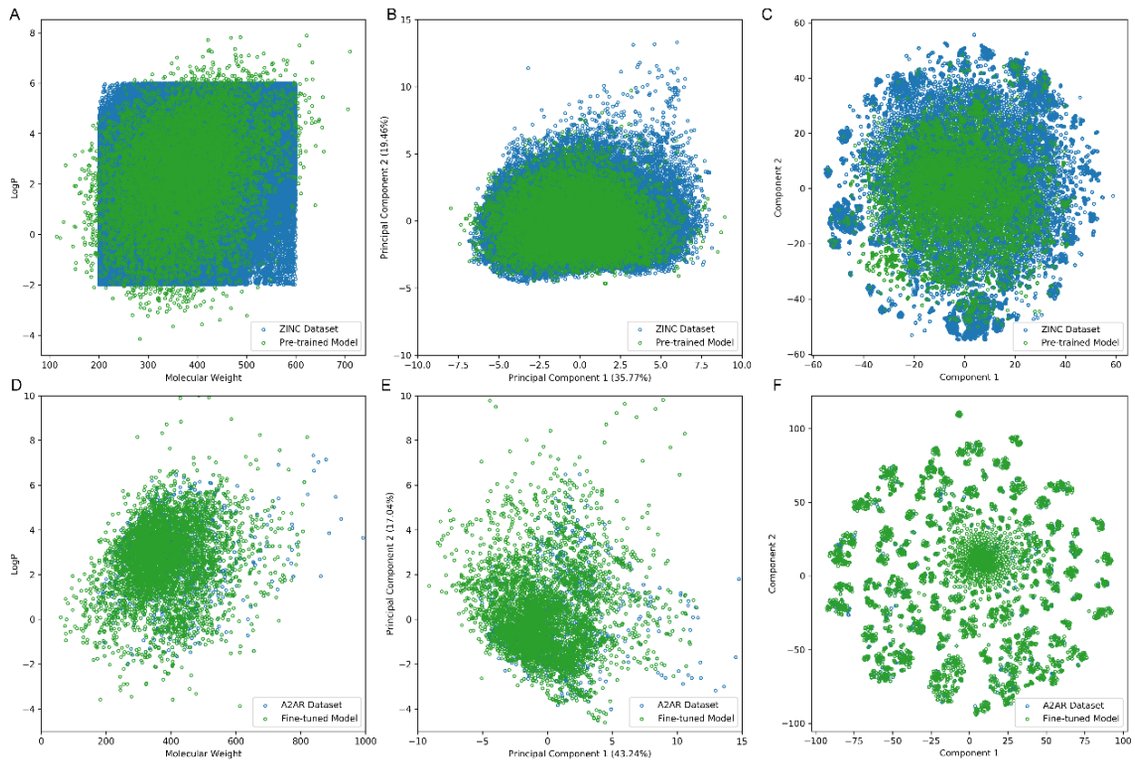


Fig. 3.7: The chemical space of generated molecules by pre-trained models with *ZINC* dataset (A-C) and fine-tuned model with *A2AR* dataset (D-F). The chemical space was represented by either logP ~ MW (A, D) and first two components in PCA on PhysChem descriptors (C, E) and t-SNE on ECFP6 fingerprints (D, F).

3.3.4. Exploration strategy

During the training process, the generated sequence is determined by both the G_θ and the G_φ where ε determines how many contributions the G_φ made. The G_φ and G_θ were both initialized by the pre-trained RNN model on the *ZINC* dataset. The G_φ was fixed and only parameters in the G_θ were updated. In order to optimize parameters, the parameter space was designated [0.01, 0.1] and [0.0, 0.1] for ε and β, respectively. After the model converged at 200 epochs (Fig. 3.8A), the performance of these models was evaluated subsequently based on 10,000 sampled sequences. Firstly, it was found that the number of duplicate SMILES notations was reduced dramatically and almost all SMILES notations represented drug-like molecules (Fig. 3.9A, 10D). Table 3.1 shows that when ε was increased, the model generated fewer active ligands for $A_{2A}AR$ but the diversity of generated molecules (represented as unique desired SMILES) increased significantly. It was also observed that with higher ε, the distribution of different kinds of ring systems in the generated desired molecules became more similar to the known active ligands in the

A2AR dataset (Fig. 3.9A). In order to further investigate the effect of ε, we also tested a range of different values as [0.01, 0.05, 0.10, 0.15, 0.20, 0.25] and the results are shown in Figure S2. The G_φ can hence help the model produce more molecules similar to known active ligands of the given target but not identical to them. At higher ε, the baseline can help the model improve the average score and generate more desired molecules. However, this effect was not significant at lower values of ε. It is worth noticing in this study that if the baseline was larger than 0.1, the training process of the generative model did not converge.

Table 3.1: Comparison of the performance of the different methods

	DrugEx (Pre-trained)				**DrugEx (Fine-tuned)**				**REINVENT**	**ORGANIC**	**Pre-trained**	**Fine-tuned**
ε	0.01	0.01	0.1	0.1	0.01	0.01	0.1	0.1	--	--	--	--
β	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	--	--	--	--
Valid SMILES	98.3%	98.9%	95.9%	98.8%	99.1%	99.0%	98.2%	97.5%	98.8%	99.8%	93.9%	96.2%
Desired SMILES	97.5%	98.0%	74.6%	80.9%	98.3%	98.5%	94.4%	94.5%	98.2%	99.8%	0.7%	47.9%
Unique SMILES	96.5%	96.3%	73.0%	80.0%	96.5%	96.6%	84.8%	86.0%	95.8%	94.8%	0.7%	22.7%
Diversity	0.74	0.75	0.80	0.80	0.75	0.74	0.80	0.80	0.75	0.67	0.83	0.82

The pre-trained network, fine-tuned network, REINVENT, ORGANIC and DrugEx with different G_φ (shown in the parentheses), ε and β were compared.

Subsequently, the fine-tuned network was used as G_φ to be involved in our proposed training method of RL. After the training process converged at 200 epochs (Fig. 3.8B), 10,000 SMILES were generated. Compared to the pre-trained network, there were more unique molecules generated (Table 4.1), most of which were drug-like compounds (Fig. 3.9B and 10A). However, with appropriate ε the fine-tuned network helped the model generate more valid desired SMILES than with the pre-trained network. At the same time the duplication rate was also increased and there were more repetitive molecules being generated. A possible reason is that the percentage of active ligands was higher in the *A2AR*

dataset than in the *ZINC* dataset, while the size of *A2AR* dataset was much smaller than *ZINC* dataset, causing a higher number of duplicated samples generated by the fine-tuned model. In addition, a PCA showed that the fine-tuned network was more effective than the pre-trained network as $G_{\varphi,}$ as it helped the model in generating molecules with larger diversity and higher similarity to the known active ligands (Fig. 3.9 and 10). These results prove that the exploration strategy is an effective way to assist the model training for generating molecules with similar chemical and biological properties to existing molecules in a specific part of chemical space.

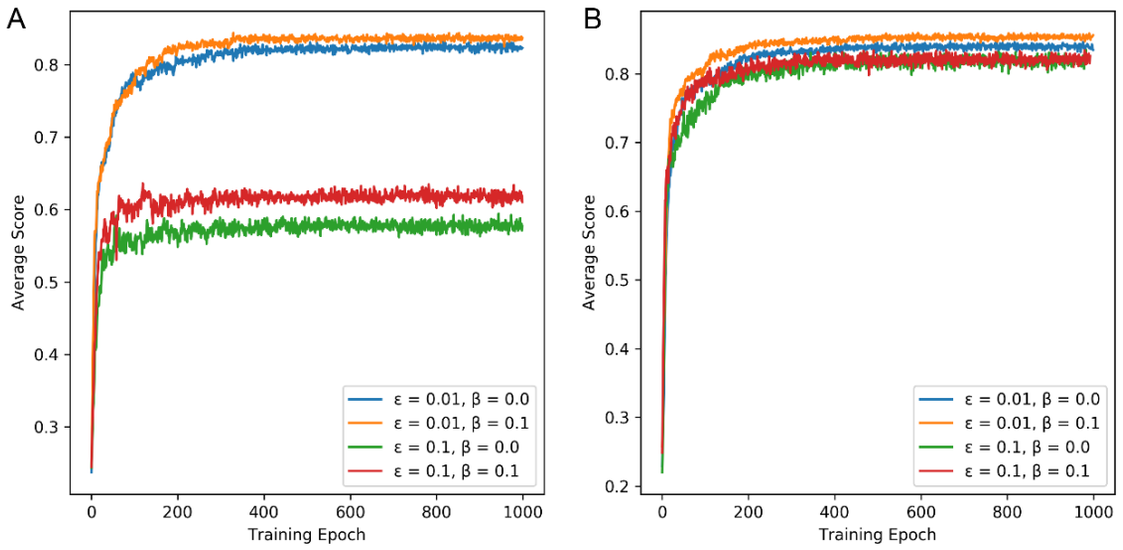


Fig. 3.8: The average score of generated SMILES sequences during the training processes of deep reinforcement learning with different *ε, β and* G_{φ}. The pre-trained model on *ZINC* dataset (A) and the fine-tuned model on *A2AR* set (B) were used as G_{φ}. After 200 epochs, the average scores for all training processes converged and all of these models were well trained.

3.3.5. Comparison with other methods

Several papers on SMILES generation using deep learning have been published. Olivecrona *et al.* proposed a method named "REINVENT" [20], in which a new loss function was introduced based on the Bayesian formula for RL,

$$L(\theta) = \left[logP_{Prior}(y_{1:T}) + \sigma R(y_{1:T}) - logP_{Agent}(y_{1:T})\right]^2$$

The authors used all molecules in the ChEMBL database to pre-train an RNN model as the *Priori*. With the parameter σ, they integrated the reward R of each SMILES into the loss

function. The final *Agent* model was regarded as the *Posteriori* and trained with the policy gradient. Finally, they successfully identified a large number of active ligands against the dopamine D2 receptor (DRD2).

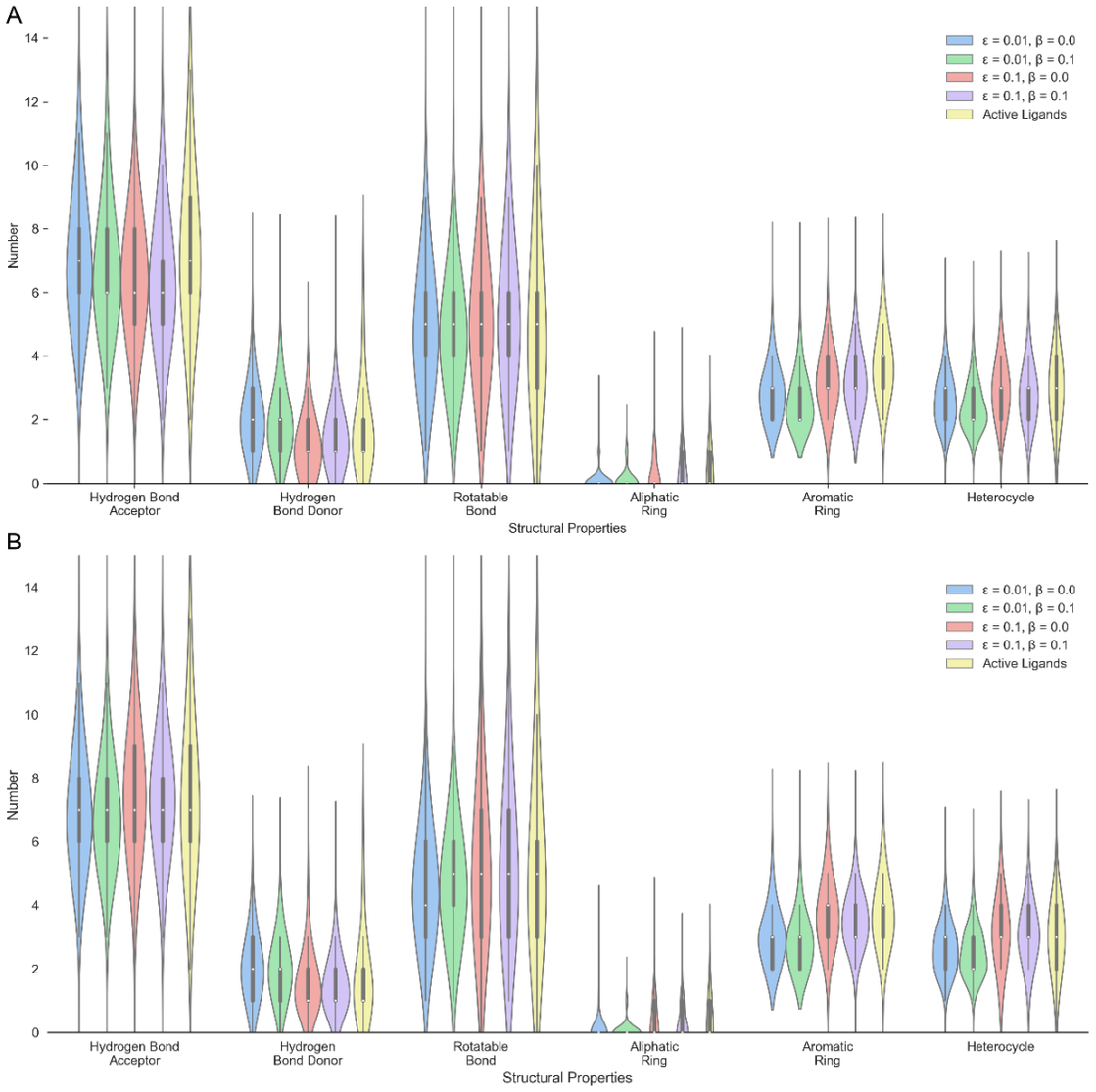


Fig. 3.9: Comparison of the properties of generated molecules by RL models with different ε, β and G_φ. The pre-trained model on *ZINC* dataset (A) and the fine-tuned model on *A2AR* dataset (B) were used as G_φ. These properties included the number of hydrogen bond donors/acceptors, rotatable bonds, aliphatic rings, aromatic rings, and heterocycles.

Likewise, Benjamin *et al.* proposed another method named "ORGANIC" [24] by combining a GAN model for sequence generation and a prediction model to form a comprehensive reward function for RL.

$$R(y_{1:t}) = \lambda R_d(y_{1:T}) + (1-\lambda)R_c(y_{1:T})$$

Here, the reward is represented as the weighted sum of two parts determined by parameter λ: 1) the reward R_c was provided by the prediction model, and 2) the reward R_d was calculated by discriminator neural network D, which was trained with generator simultaneously by minimizing the following loss function:

$$L(\theta) = \sum_{y \in Real} (logD(y_{1:T})) + \sum_{y \in Fake} (log(1 - D(y_{1:T})))$$

With the policy gradient optimization, the final model generated many different desired molecules which were predicted as active ligand against a given target and were similar to the chemical compounds in the ligands dataset. In the following section DrugEx and its performance is compared with these two methods.

The code of REINVENT and ORGANIC was downloaded from GitHub and executed with default parameters ($\sigma = 60$ in REINVENT and $\lambda = 0.5$ in ORGANIC). The prior network of REINVENT and generator network of ORGANIC were initialized with the pre-trained model, and the agent network of REINVENT was started from the fine-tuned model. The RF-based predictor with ECFP6 was exploited as reward function for both methods identical to our own implementation. After these models were trained, 10,000 SMILES sequences were generated for performance comparison with each other (Table 4.1). Our method generated molecules that had the larger diversity at ε = 0.1. While DrugEx did not outperform REINVENT based on the percentage of unique desired SMILES, this value was improved dramatically and closely resembled that of REINVENT when ε was set to 0.01. In addition, although most of the molecules generated by these methods were drug-like molecules (Fig. 3.10A-D), through PCA on PhysChem descriptors (Fig. 3.10E-H) and t-SNE on ECFP6 fingerprints (Figrue 10I-L), we found that molecules generated by our method covered the whole region of chemical space occupied by known active ligands. Conversely, molecules generated by both REINVENT and ORGANIC only covered a small fraction of the desired chemical space and were mostly centered in Rule-of-5 compliant chemical space even though the chemical space for the $A_{2A}R$ transcends this region of space. Moreover, we also used a k-means algorithm to cluster the active ligands

in the *A2AR* dataset and generated molecules into 20 groups with the ECFP6 fingerprints of the full compound structure, Murcko scaffold and topological Murcko scaffold. The results indicated that the generated molecules by DrugEx covered all of the clusters that contain active ligands in the *A2AR* dataset, but some of these clusters cannot be covered by REINVENT and ORGANIC (Figure S3). The distribution of the molecules in each cluster generated by DrugEx was closer to active ligands in the *A2AR* dataset than REINVENT and ORGANIC.

Previous work on the binding mechanism between the $A_{2A}AR$ and its ligands identified a number of critical substructures that play an important role to improve binding affinity [36]. For example, the oxygen in the furan ring of ZM241385 and related ligands can form a hydrogen bond with residue N253, the purine ring acts as hydrogen bond donor to N253 and forms π-π interaction with F168 [6]. However, molecules containing such a furan ring tend to be blocking the receptor (antagonists) rather than activating it (agonists). Hence, while the furan ring is common in the set of known $A_{2A}R$ ligands, its presence might not always be favorable for generated ligands. Moreover, in general fused rings have been shown to be important in the chemical structure of drugs [37]. Therefore, we compared the percentage of molecules containing furan rings, fused rings, and benzene rings. Only 0.20% of the desired molecules generated by REINVENT contained a fused ring (Table 4.2) while they were present in 79.09% of active ligands in the A2AR set. Similarly, ORGANIC only generated a very low percentage of molecules containing a fused ring system (0.02%).

With the pre-trained network as G_{φ}, DrugEx produced 9.12% of molecules containing fused rings, while the fine-tuned network improved the percentage of molecules containing fused rings up to 60.69%. Moreover, 95.26% and 99.96% of molecules generated by REINVENT and ORGANIC contained a furan ring, respectively, while this percentage was only 40.29% for known active ligands. In DrugEx, 82.32% of molecules contained a furan ring under the pre-trained network as G_{φ}, similar to the other two methods. However, when the fine-tuned network was used this rate decreased substantially to 66.35%.

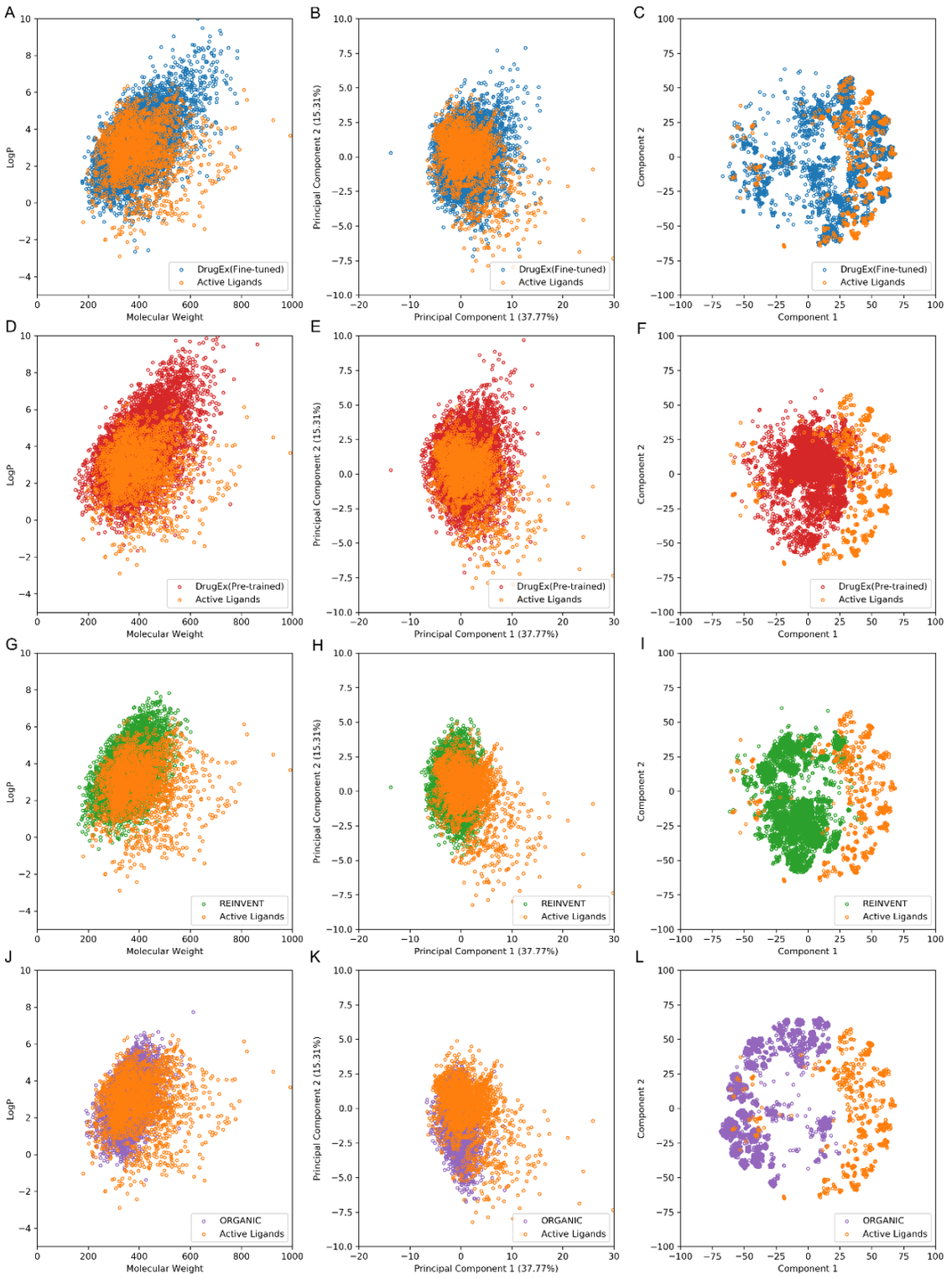


Fig. 3.10: Comparison of the chemical space of active ligands in the *A2AR* set and generated molecules by REINVENT, ORGANIC and DrugEx with different G_φ (shown in parentheses). Chemical Space was represented by logP ~ MW (A, D, G and J) and first two components in PCA on PhysChem descriptors (B, E, H and K) and t-SNE on ECFP6 fingerprints (C, F, I and L).

REINVENT and ORGANIC have been reported to generate various molecules containing different fused ring structures against DRD2 [24,20]. One possible reason they were not

able to do so here might lie in the bias of *A2AR* dataset. In table 4.2, we noticed that there were more active ligands containing a furan ring than inactive ligands (four fold difference). This led to both methods only generating molecules containing a furan ring which were prone to be predicted as active. However, both methods neglected to construct more complicated fused rings which is a decisive difference between active and inactive ligands in *A2AR* dataset. These results indicate that DrugEx is more robust to overcome the bias of the training set to generate more similar compounds to known A_{2A}AR ligands (tuned for the target chemical space) and less generic SMILES sequences. Hence, we consider these molecules more appropriate drug candidates against A_{2A}AR than the molecules produced by REINVENT and ORGANIC. As an example, 24 candidate molecules generated by DrugEx were selected and are shown in Fig. 3.11 ordered by the probability score and Tanimoto-distance to the *A2AR* dataset.

Table 4.2: Comparison of the percentage of important substructures contained in the molecules generated by the methods.

		Fused Ring	Furan Ring	Benzene Ring
DrugEx (Pre-trained)		9.12%	82.32%	61.48%
DrugEx (Fine-tuned)		60.69%	66.35%	65.62%
REINVENT		0.20%	95.26%	61.98%
ORGANIC		0.02%	99.96%	39.45%
Pre-trained		24.22%	4.51%	63.31%
Fine-tuned		76.33%	23.82%	72.85%
ZINC		26.66%	3.86%	63.97%
A2AR	**Active**	79.09%	40.29%	75.33%
	Inactive	76.73%	9.33%	70.88%

The table compares DrugEx with pre-trained and fine-tuned model as different G_φ (in the parentheses), REINVENT, ORGANIC, Pre-trained model, Fine-tuned model and the molecules in *ZINC* and *A2AR* dataset.

In REINVENT, the pre-trained model acted as a "priori" in the Bayesian formula to ensure that the generated SMILES are drug-like molecules. The final model was trained by improving the probability of desired generated SMILES while maintaining the probability of undesired generated SMILES similar to the pre-trained model. In DrugEx the pre-trained

model was *only* used for initialization and did not directly affect the training process and performance evaluation. The mechanism of DrugEx appears quite similar to a genetic algorithm (GA) previously developed in our group for *de novo* drug design [38]. The exploration strategy can be regarded as "random mutation" in a GA context for sequence generation. Instead of changing the token selection directly, this manipulation just changed the probability distribution of each token in the vocabulary. Furthermore, although "crossover" manipulation was not implemented here, such mutations can still help the model search the unfamiliar chemical space in which the molecules do not have a high probability to be sampled. In contrast to ORGANIC, there was no need to construct another neural network specifically to measure the similarity between generated and real molecules, saving valuable time and resources required to train and select appropriate parameters. Despite the inevitable introduction of some duplicates the molecules generated by DrugEx can be regarded as reasonable drug candidates for $A_{2A}AR$.

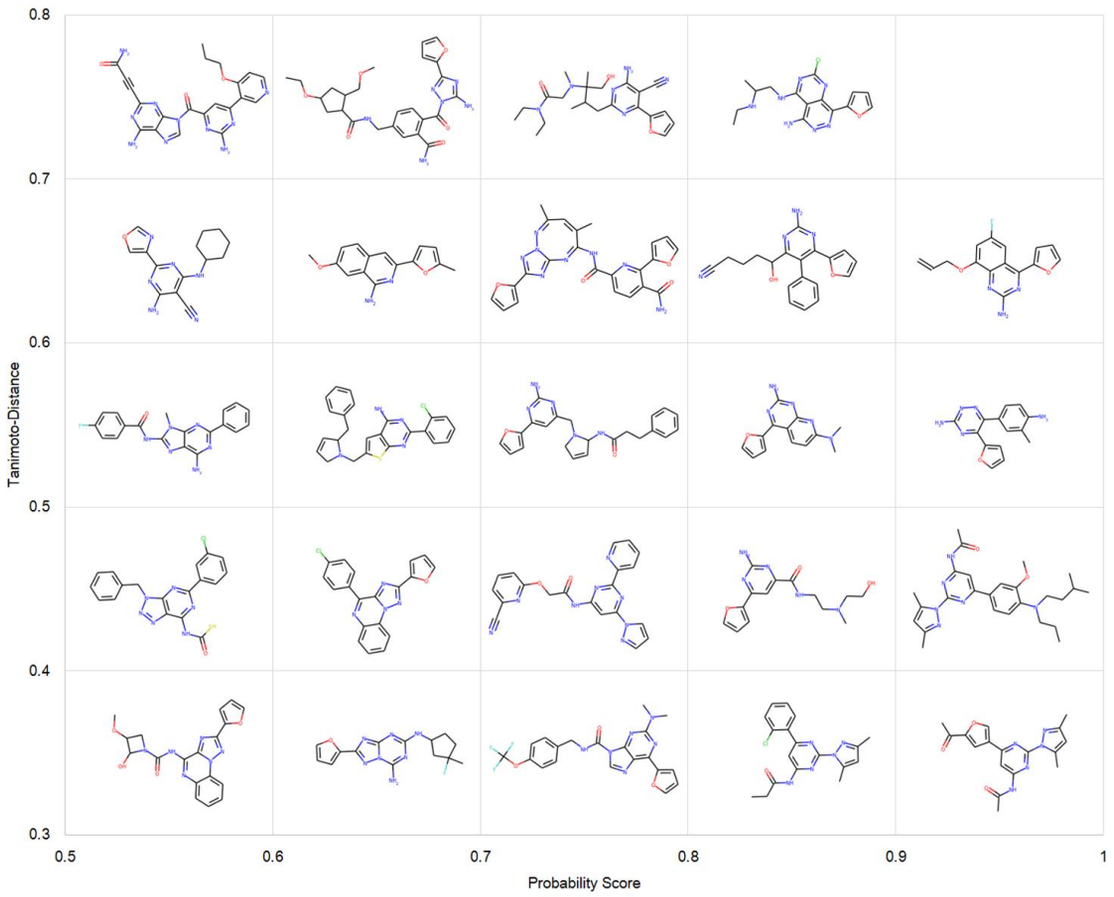


Fig. 3.11: 24 Candidate molecules were selected from 10,000 SMILES sequences generated by DrugEx. These molecules were ordered by the probability score given by the predictor and Tanimoto-distance to *A2AR* dataset.

3.4. Conclusion and future prospect

In this study a new method is proposed to improve the performance of deep reinforcement learning to generate SMILES based ligands for targets of interest. Applied to the $A_{2A}AR$, generated molecules had high diversity combined with chemical and predicted biological properties similar to known active compounds. Previous work has shown that RL cannot guarantee the model to generate molecules distributed over chemical space comparable to ligands of a target of interest. To solve this problem, another pre-trained RNN model was included as exploration strategy to force the model to enlarge the chemical space of the generated molecules during the training process of RL. Compared with other DL-based methods, DrugEx generated molecules with larger diversity and higher similarity to known active ligands, albeit at the expense of more inactive or duplicated molecules.

In future work, the aim is to update DrugEx with multi-objective optimization. As a given drug (candidate) likely binds some other targets (*i.e.* off-target efficacy) which can cause side-effects [39]. Incorporating multiple objectives in SMILES generation will allows the search for ways to eliminate potential off-target affinity.

Declarations

Availability of data and materials

The data used in this study is publicly available ChEMBL data, the algorithm published in this manuscript is made available at https://github.com/XuhanLiu/DrugEx.

Authors' Contributions

XL and GJPvW conceived the study and performed the experimental work and analysis. KY, APIJ and HWTvV provided feedback and critical input. All authors read, commented on and approved the final manuscript.

Acknowledgements

XL thanks Chinese Scholarship Council (CSC) for funding, GJPvW thanks the Dutch Research Council and Stichting Technologie Wetenschappen (STW) for financial support (STW-Veni #14410).

Competing Interests

The authors declare that they have no competing interests

References

1. Lv X, Liu J, Shi Q, Tan Q, Wu D, Skinner JJ, Walker AL, Zhao L, Gu X, Chen N, Xue L, Si P, Zhang L, Wang Z, Katritch V, Liu ZJ, Stevens RC (2016) In vitro expression and analysis of the 826 human G protein-coupled receptors. Protein & cell 7 (5):325-337. doi:10.1007/s13238-016-0263-8
2. Dorsam RT, Gutkind JS (2007) G-protein-coupled receptors and cancer. Nature reviews Cancer 7 (2):79-94. doi:10.1038/nrc2069
3. Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE (2017) Trends in GPCR drug discovery: new agents, targets and indications. Nature reviews Drug discovery 16 (12):829-842. doi:10.1038/nrd.2017.178
4. Chen JF, Eltzschig HK, Fredholm BB (2013) Adenosine receptors as drug targets--what are the challenges? Nature reviews Drug discovery 12 (4):265-286. doi:10.1038/nrd3955
5. Liu W, Chun E, Thompson AA, Chubukov P, Xu F, Katritch V, Han GW, Roth CB, Heitman LH, AP IJ, Cherezov V, Stevens RC (2012) Structural basis for allosteric regulation of GPCRs by sodium ions. Science 337 (6091):232-236. doi:10.1126/science.1219218
6. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science 322 (5905):1211-1217. doi:10.1126/science.1164772
7. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research 40 (Database issue):D1100-1107. doi:10.1093/nar/gkr777
8. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521 (7553):436-444. doi:10.1038/nature14539
9. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of Deep Learning in Biomedicine. Molecular pharmaceutics 13 (5):1445-1454. doi:10.1021/acs.molpharmaceut.5b00982
10. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2017) Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics. doi:10.1093/bib/bbx044
11. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin, II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry 57 (12):4977-5010. doi:10.1021/jm4004285
12. Ekins S (2016) The Next Era: Deep Learning in Pharmaceutical Research. Pharmaceutical research 33 (11):2594-2603. doi:10.1007/s11095-016-2029-7
13. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug discovery today. doi:10.1016/j.drudis.2018.01.039
14. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. Journal of chemical information and modeling 50 (5):742-754. doi:10.1021/ci100050t
15. Lenselink EB, Ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, AP IJ, van Westen GJP (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of cheminformatics 9 (1):45. doi:10.1186/s13321-017-0232-0
16. Jaeger S, Fulle S, Turk S (2018) Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. Journal of chemical information and modeling 58 (1):27-35. doi:10.1021/acs.jcim.7b00616
17. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A (2017) druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. Molecular pharmaceutics 14 (9):3098-3104.

doi:10.1021/acs.molpharmaceut.7b00346

18. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional Networks on Graphs for Learning Molecular Fingerprints. ArXiv:1509.09292
19. Gupta A, Muller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G (2018) Generative Recurrent Networks for De Novo Drug Design. Molecular informatics 37 (1-2). doi:10.1002/minf.201700111
20. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. Journal of cheminformatics 9 (1):48. doi:10.1186/s13321-017-0235-x
21. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529 (7587):484-489. doi:10.1038/nature16961
22. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative Adversarial Networks. ArXiv:1406.2661
23. Yu L, Zhang W, Wang J, Yu Y (2016) SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. ArXiv:1609.05473
24. Benjamin S-L, Carlos O, Gabriel L. G, Alan A-G (2017) Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). doi:10.26434/chemrxiv.5309668.v3
25. Preuer K, Renz P, Unterthiner T, Hochreiter S, Klambauer GUN (2018) Frechet ChemNet Distance: A metric for generative models for molecules in drug discovery. Journal of chemical information and modeling. doi:10.1021/acs.jcim.8b00234
26. Benhenda M (2017) ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? ArXiv:1708.08227
27. Schneider G, Fechner U (2005) Computer-based de novo design of drug-like molecules. Nature reviews Drug discovery 4 (8):649-663. doi:10.1038/nrd1799
28. Sterling T, Irwin JJ (2015) ZINC 15--Ligand Discovery for Everyone. Journal of chemical information and modeling 55 (11):2324-2337. doi:10.1021/acs.jcim.5b00559
29. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. Nucleic acids research 45 (D1):D945-D954. doi:10.1093/nar/gkw1074
30. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org.
31. Reddy AS, Zhang S (2013) Polypharmacology: drug discovery for the future. Expert review of clinical pharmacology 6 (1):41-47. doi:10.1586/ecp.12.74
32. PyTorch. https://pytorch.org/.
33. Kingma DP, Ba J (2014) Adam: A Method for Stochastic Optimization. arXiv:1412.6980
34. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv:1412.3555
35. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. Acs Central Sci 4 (1):120-131. doi:10.1021/acscentsci.7b00512
36. Jaakola VP, Lane JR, Lin JY, Katritch V, Ijzerman AP, Stevens RC (2010) Ligand binding and subtype selectivity of the human A(2A) adenosine receptor: identification and characterization of essential amino acid residues. The Journal of biological chemistry 285 (17):13032-13044. doi:10.1074/jbc.M109.096974

37. Feher M, Schmidt JM (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. Journal of chemical information and computer sciences 43 (1):218-227. doi:10.1021/ci0200467
38. Lameijer EW, Kok JN, Back T, Ijzerman AP (2006) The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. Journal of chemical information and modeling 46 (2):545-552. doi:10.1021/ci050369d
39. Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, Nakamura Y (2007) When good drugs go bad. Nature 446 (7139):975-977. doi:10.1038/446975a

Table S3.1: All tokens in vocabulary for SMILES sequence construction with RNN model.

Atoms				Bonds	Controls		
Common Atoms			Aromatic Atoms	--	Rings	Branchs	On-Off
B	[B-]	[O-]	[cH-]	-	1	(	GO
C	[BH-]	[O]	[n+]	=	2	)	EOS
F	[C+]	[P+]	[nH]	#	3		
I	[C-]	[PH]	[s+]		4		
Cl	[CH-]	[Br+]	c		5		
N	[CH]	[S+]	n		6		
O	[C]	[SH]	o		7		
P	[N+]	[SiH3]	p		8		
Br	[NH+]	[SiH]	s		9		
S	[N]	[Si]					
		[Sn]					

Considering that there are no drug-like molecules containing more than 10 rings, we omitted the token “0” and “%” for the construction of more than 10 rings. In addition, we ignored the isomerism of molecules and ionic bond, therefore we removed the “@”, ‘\’, ‘/’, ‘.’ and all metal ion.

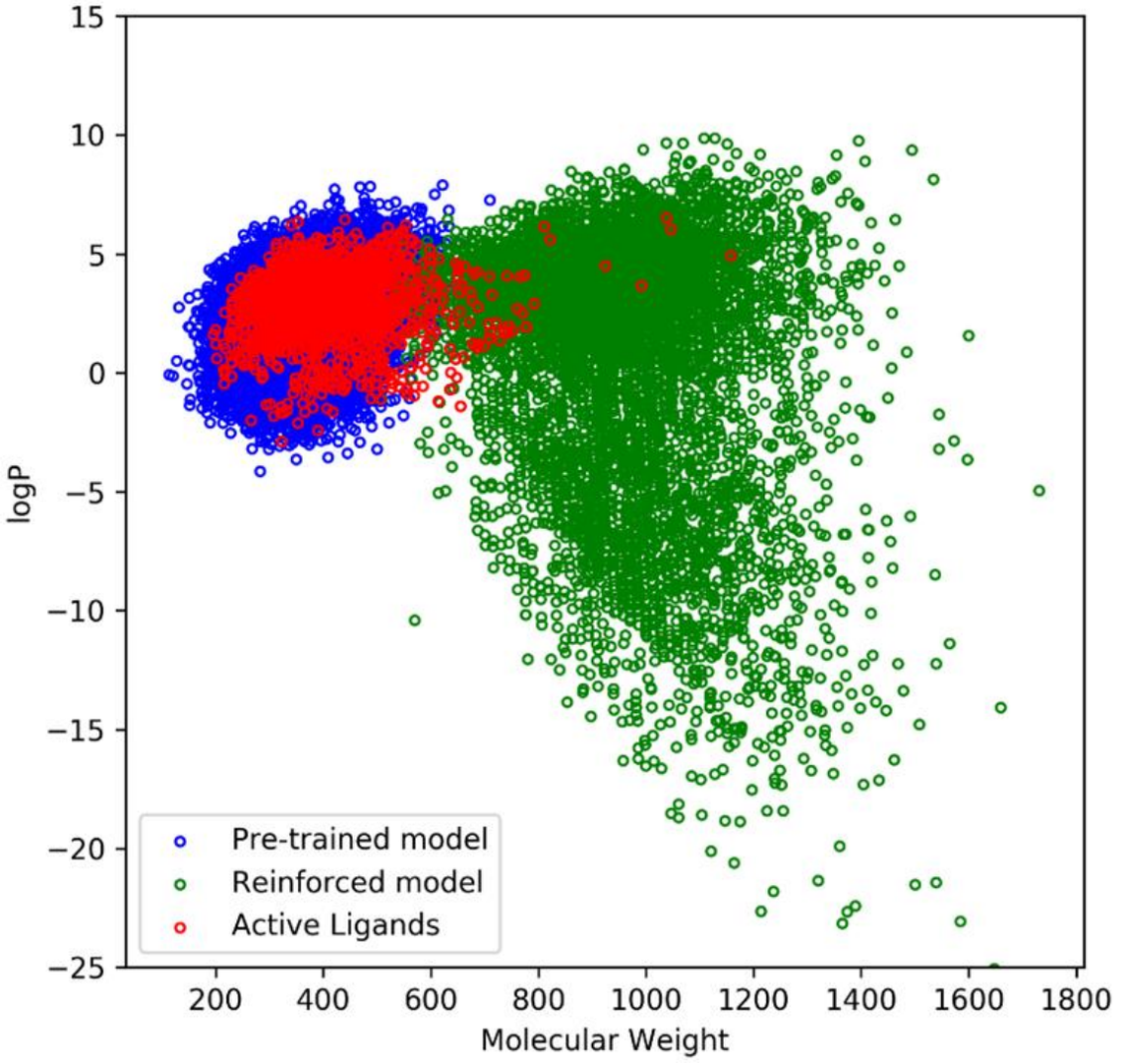
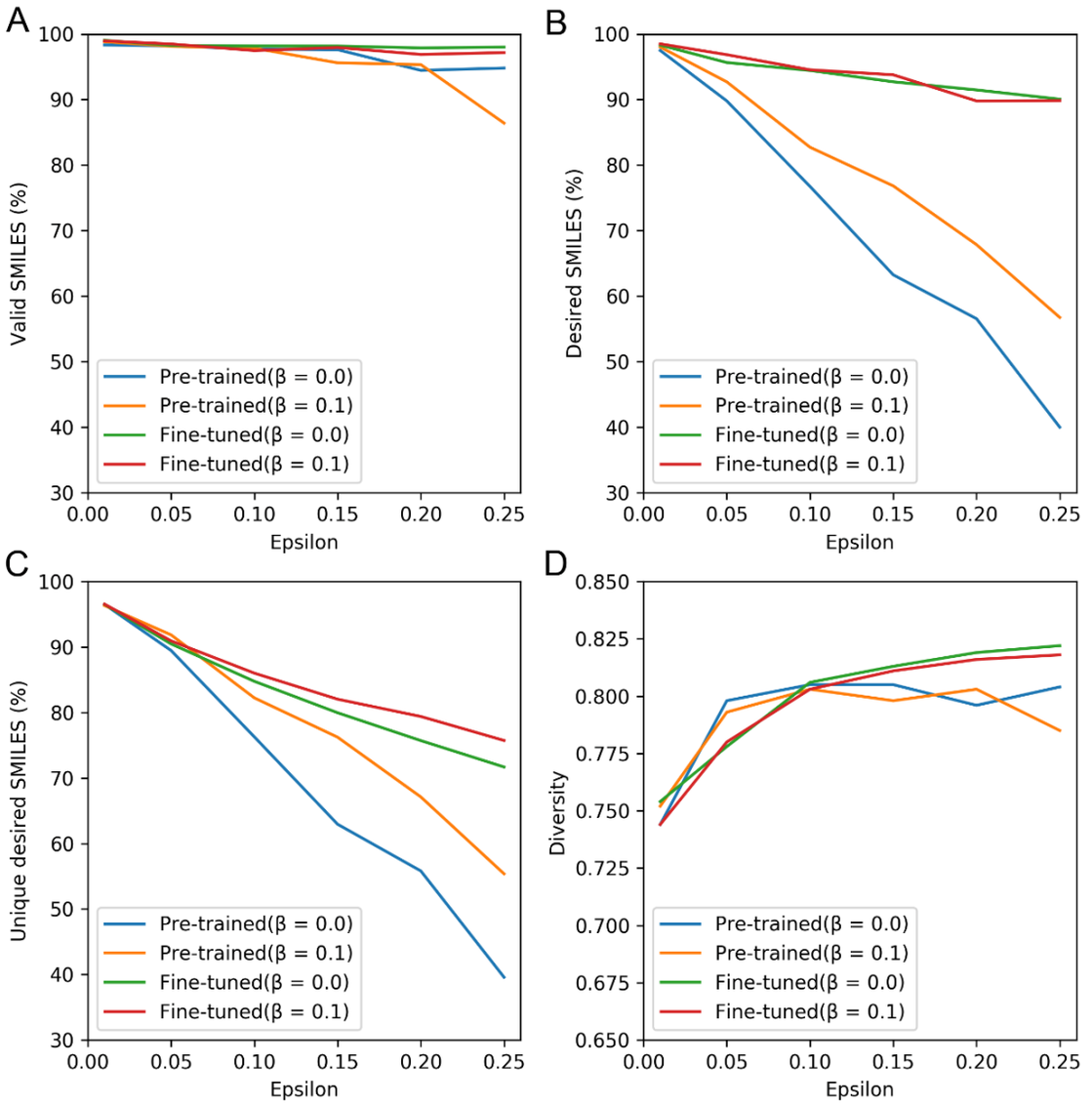


Figure S3.1: The chemical space of generated molecules by pre-trained models, traditional reinforced model and active ligands in the *A2AR* set. The chemical space was represented as logP ~ MW. The generated molecules by pre-trained model covered the greater part of space of known active ligands, while the molecules generated by reinforced model were distributed in a distinct region, which cannot be regarded drug-like although the compounds were predicted as active ligands.



TableS3.2: The performance of DrugEx with different G_φ (pre-trained and fine-tuned model) and hyperparameters (including ε and β). These performance included the percentage of valid SMILES (A), desired SMILES (B) and unique desired SMIIES (C) and diversity (D).

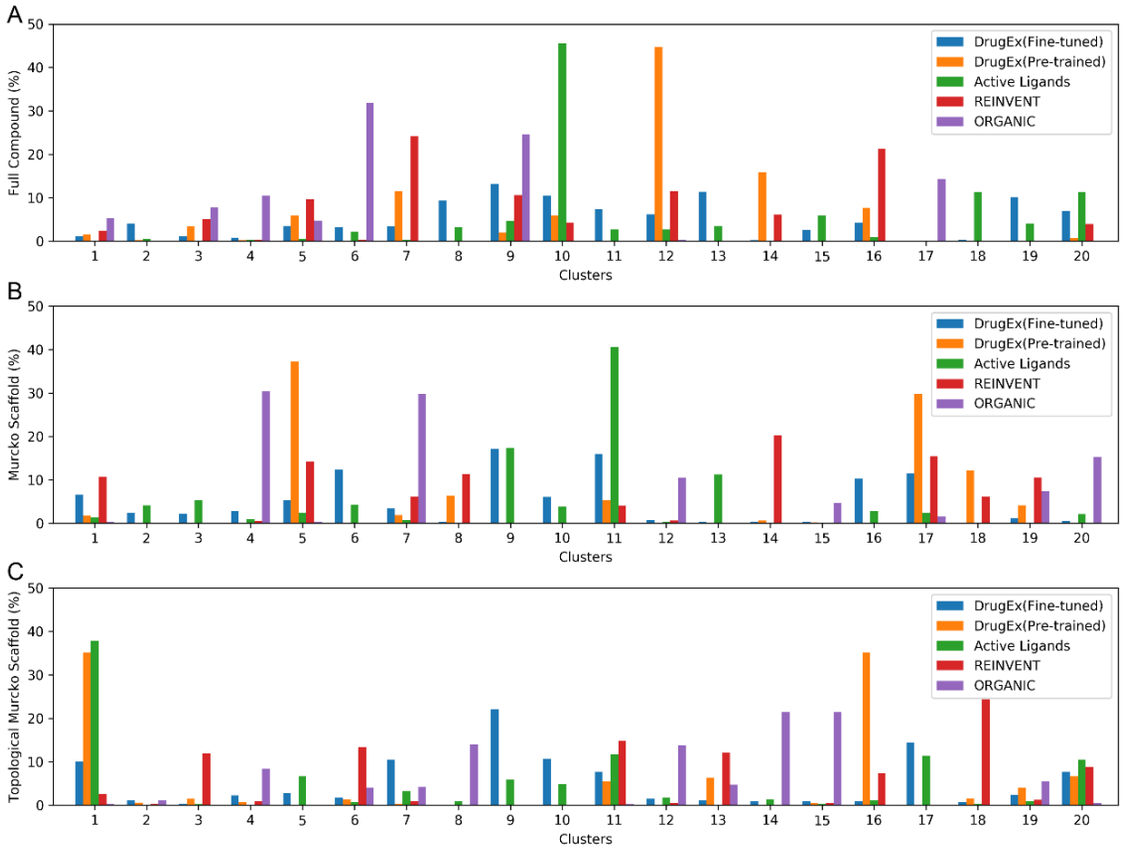


Figure S3.3: The percentage of molecules in 20 groups clustered by k-means algorithm on ECFP6 fingerprints of generated molecules with full compound (A), Murcko scaffold (B) and topological Murcko scaffold (C). These molecules included active ligands in *A2AR* dataset and molecules generated by REINVENT, ORGANIC and DrugEx with different G_{φ} (shown in the parentheses)