CNSMolGen: a bidirectional recurrent neural networks based generative model for de novo central nervous system drug design

Rongpei Gou[a], Jingyi Yang[a], Menghan Guo[a], Yingjun Chen[a,*], Weiwei Xue[a,*]

[a] *Chongqing Key Laboratory of Natural Product Synthesis and Drug Research, School of Pharmaceutical Sciences, Chongqing University, Chongqing 401331, China*

Abstract

Central nervous system (CNS) drugs have had a significant impact on human health, e.g., treating a wide range of neurodegenerative and psychiatric disorders. In recent years, deep learning-based generative models, particularly those for designing drugs from scratch, have shown great potential for accelerating drug discovery, reducing costs and improving efficacy. However, specific applications of these techniques in CNS drug discovery have not been widely reported. In this study, we developed the CNSMolGen model, which uses a bidirectional recurrent neural networks (Bi-RNNs) system for *de novo* molecular design of CNS drugs by learning from compounds with CNS drug properties. Result shown that the pre-trained model was able to generate more than 90% of completely new molecular structures, and these new molecules possessed the properties of CNS drug molecules and synthesizable. In addition, transfer learning was performed on small datasets with specific biological activities to evaluate the potential application of the model for CNS drug optimization. Here, we used drugs against the classical CNS disease target serotonin transporter (SERT) as a fine-tuned dataset and generated a Focused database against the target protein. The potential biological activities of the generated molecules were verified using the physics-based induced fit docking study. The success of this model demonstrates its potential in CNS drug design and optimization, which provides a new impetus for future CNS drug development.

Keywords: Bidirectional recurrent neural network; Transfer learning; De novo drug design; Central nervous system drug; Selective serotonin reuptake inhibitor

1. Introduction

Along with the social construction of population aging, the incidence of CNS disorders, including Alzheimer disease (AD), Parkinson disease (PD), and depression, is continuously increasing[1]. However, the number of approved drugs targeting these diseases remains limited, partially attributed to the blood-brain barrier (BBB) that prevents the access of drug molecules into the CNS[2]. Consequently, discovery and development of novel and effective drugs for the treatment of human CNS disorders remains a crucial responsibility in the field of pharmaceutical sciences[3].

* Corresponding Authors
Email: xueww@cqu.edu.cn (Weiwei Xue) and cxue2006@126.com (Yingjun Chen)

To address this problem, several innovative approaches have been developed and used in CNS drug discovery, including molecular drug targets identification and drug design aided by either physics-based modeling or deep-learning based artificial intelligence (AI)[4, 5]. For example, monoamine transporters (MATs), including dopamine transporter (DAT), norepinephrine transporter (NET), and serotonin transporter (SERT), play a crucial role in regulating neurotransmitters through reuptake and have been identified as important molecular targets of CNS disorders' medications (e.g. anti-neurodegenerative drugs and antidepressants)[6]. Nonetheless, currently approved MATs-targeted drugs still exhibit certain side effects, including addiction and drug resistance[5]. Therefore, rational design of new drugs acting on MATs are urgently needed[7]. For the traditional computer-aided drug design (CADD) approach, it is often necessary to conduct virtual screening on molecular libraries with large size to identify molecules with specific characteristics[8-10]. Despite significant progress has achieved in the computational power for processing these databases, the enormous size of the chemical space 10^{23}-10^{60} drug-like molecules[11], continues to pose challenges in efficiently identifying molecules that specifically bind to molecular drug target[12]. While the recent advent of deep learning-based molecular generation techniques may offer viable solutions to this predicament[13].

In 2016, Gómez-Bombarelli et al. first introduced a method which converts discrete representation of molecules into a multi-dimensional continuous representation, representing an early application of the variational autoencoder (VAE) model in the field of molecular generation[14]. Subsequently, an array of different deep learning models has been developed in this field[15-17]. Among them, recurrent neural networks (RNNs) (**Fig. 1**) have shown remarkable performance. This is because RNN can obtain high precision and good performance when processing time series predictions based on a large number of data sets[18]. In particular, Bjerrum and Threlfall introduced RNN to generate chemically plausible and novel molecules by incorporating the Long Short-Term Memory (LSTM) network to overcome the challenges associated with vanishing gradient or explosion problems[19]. In addition, Wu et al. proposed bidirectional long short-term memory (BiLSTM) attention network (BAN) to enhance the extraction of crucial features from the simplified molecular input line entry specification (SMILES) strings, leading to improved performance in molecular property prediction[20]. However, no deep learning-based model specifically for generating molecules with CNS drug properties has been reported yet.

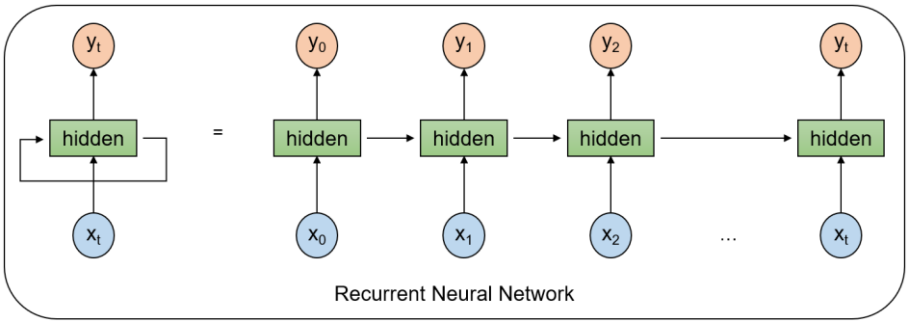

Fig. 1. Schematic diagram of a typical recurrent neural network (RNN) structure and the loop unrolled. xt denotes the input at moment t, yt is the output corresponding to xt, and hidden denotes the hidden layer, which is related to both the input xt and the previous state (at t-1).

In this work, we present CNSMolGen, a generative model specifically constructed for *de novo* design of CNS drugs. The model employs a bidirectional RNNs (Bi-RNNs) based on SMILES string representation[21-23] to design target molecules from scratch. First, an RNN model based on LSTM units to generate small molecule datasets with specific properties was pre-trained using the large dataset of ChemBridge-MPO which contains 504,853 compounds with good CNS drug

properties and synthetic accessibilities. Then, using the representative CNS diseases molecular drug target SERT as an example[24-26], transfer learning technique was employed to fine-tune the pre-trained model with the aim of generating target-specific small molecule datasets under the condition of limited data size. Finally, the physics-based induced fit docking (IFD) was used to verify the binding mode and binding affinities of *de novo* designed molecules to SERT.

2. Materials and methods

2.1. Data preparation

Pre-training dataset

In this work, we selected the ChemBridge-MPO database for use in training pre-trained models in CNSMolGen. This database was created by ChemBridge to provide a library of small molecule compounds for the CNS drug discovery and contains 504,853 high quality, PAINS-free small molecule compounds (https://chembridge.com/targeted-and-specialty-libraries/cns/). The CNS MPO is a well- recognized algorithm that assigns a score based on six key physicochemical properties (cLogP, cLogD, MW, TPSA, HBD and pKa) related to the blood-brain barrier (BBB) penetration, which is critical for CNS drug development[27]. The CNS MPO score with scores ≥ 4.0 was widely used as a cut-off to select compounds for hit discovery in drug discovery programs in the CNS therapeutic area.

Fine-tuning dataset

In the meantime, we have screened 44 small molecule compounds from the most recent review literature on monoamine transporter protein drugs[7]. All of these compounds are selective serotonin reuptake inhibitors (SSRIs) either on the market or under clinical investigation. To meet the training needs of the CNSMolGen model, we removed the stereochemical information of these compounds and finalized 36 compounds as the fine-tuned training dataset.

RDKit calculation

The molecules in each dataset were processed using RDKit (https://www.rdkit.org/) and saved as canonical SMILES strings. During this processing we removed salts and stereochemical information from the molecules. In addition, to reduce the degree of data heterogeneity, we only retained compounds with SMILES strings between 10 and 74 characters in length. After completing these steps, we obtained a pre-trained dataset containing 504,853 SMILES strings and a fine-tuned dataset containing 36 SMILES strings.

2.2. Architecture of CNSMolGen

CNSMolGen consists of two basic models (**Fig. 2**), the first model is mainly a bidirectional RNNs model based on SMILES[22].

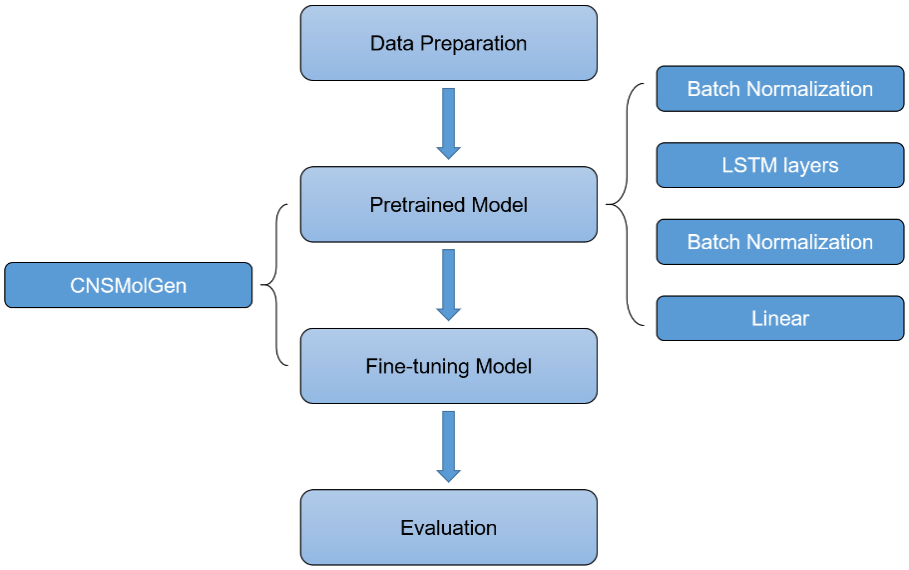

Fig. 2. The workflow of model training and compound design. CNSMolGen consists of two basic models, the pre-trained model and fine-tuning model. Each model contains four groups of layers, including Batch Normalization, LSTM layers, Batch Normalization, linear.

Its bidirectionality and holistic nature makes it more suitable for molecular generation process compared to traditional RNNs, also known as pre-trained model. And the second one is the transfer model based on the same architecture as the previous model and used to migrate the general knowledge to learn the focused knowledge by sharing the previous network and reweighting the layers. Here, to illustrate the difference between the pre-trained model in CNSMolGen and the traditional RNN, we compare their objective functions. The objective function of the classical RNN is given by

$$\mathrm{P}(x_{t+1} = k|x1, \ldots, x_t) = \frac{\exp(y_t^k/T)}{\sum_{i=1}^{K} \exp(y_t^i/T)} \tag{1}$$

where y_t^k represents the model output for the k-th token at time t, and i iterates over the set K, which encompasses all tokens. The token sampling process is regulated by the temperature parameter T. In this study, the start and end tokens for SMILES strings were denoted by "G" and "E" respectively. The key difference between the pre-trained model and traditional RNNs is that the pre-trained model consists of two RNNs that estimate two conditional probability distributions for the forward and backward directions as follows

$$\mathrm{P}(x_{m+t'+1} = k|x_{m-t'}, \ldots, x_{m+t'} = \frac{\exp(y_{m+t'}^k/T)}{\sum_{i=1}^{K} \exp(y_{m+t'}^i/T)} \tag{2}$$

$$\text{where } t' = \frac{t-1}{2} \text{(odd } t \text{ values)}$$

$$\mathrm{P}(x_{m-t^*-1} = k|x_{m-t^*}, \ldots, x_{m+t^*} = \frac{\exp(y_{m-t^*}^k/T)}{\sum_{i=1}^{K} \exp(y_{m-t^*}^i/T)} \tag{3}$$

$$\text{where } t^* = \frac{t}{2} \text{(even } t \text{ values)}$$

where $y_{m+t'}^k$ and $y_{m+t^*}^k$ are the model output for the k-th token at the considered time step ($m - t'$ and $m + t^*$, respectively). At any given t-th temporal interval, the pre-trained model interprets the sequence x = $\{x_m, x_{m+1}, \ldots, x_t\}$ in a bidirectional manner forward from x_m to x_t and

backward from x_t to x_m. The construction of the sequence commences with the initial token "G" and advances in both orientations until the concluding token "E" is produced.

The pre-trained model for bidirectional RNNs is consistent with the non-univocal nature of SMILES coding. Although most of the time we choose the standard SMILES encoding for the application, the molecule itself does not have a starting point and direction, i.e., the SMILES encoding can be written in any direction starting from any non-hydrogen atom. Therefore, it is more beneficial to generate molecules with a non-directional CNSMolGen model when performing deep learning[22].

2.3. Model evaluation

The evaluation of molecular generation models involves the analysis of numerous parameters[28]. Here, we will primarily focus on cross-entropy loss and molecular properties to evaluate the model CNSMolGen trained on the datasets of ChemBridge-MPO Library.

Cross-entropy loss

The cross-entropy loss is a loss function that measures the discrepancy between the predicted results and the actual outcomes of a model[29]. It serves as a crucial metric for optimizing model parameters. A lower value of L is indicative of a narrower gap between the model's predictions and the actual outcomes, thus reflecting superior model performance. The value of L typically declines within the initial epochs, yet prolonging the training duration can further ameliorate the model. Caution is advised as neural network models may reach a saturation point, at which the learning curve will exhibit overfitting.

Number of valid, unique, and novel smiles strings

The efficacy, novelty, and uniqueness of generated molecules are the most critical criteria for evaluating molecular generation models[30]. Efficacy refers to the percentage of generated SMILES strings that are chemically valid; uniqueness pertains to the percentage of unique SMILES strings (after canonicalization) within the generated set; and novelty denotes the percentage of generated SMILES strings that represent molecules not included in the training set. These three metrics collectively reflect the model's capacity to learn molecular structures, explore chemical space, and innovate in molecular design.

Molecular property

Molecular property predictions are conducted for small molecules generated by the model before and after fine-tuning. The properties of LogP, LogD, polar surface area, molecular weight, HBD, and pKa are calculated using ChemAxon[31], while the synthetic accessibility score (SAScore) is predicted using RDKit (http://www.rdkit.org/).

2.4. Induced fit docking

Molecular docking is one of the most esteemed and efficacious structure-based computational methodologies to predict the interactions between drugs and targets at the atomic level[32]. In recent years, researches have substantiated the significant impact of protein conformational flexibility, especially the ligand binding site, on the formation of protein-ligand complex[33]. Hence, we utilize the Induced Fit Docking (IFD) method for predicting the binding mode to estimate the binding energy between target protein and *de novo* designed molecules[34].

Prior to IFD, the designed compounds were energy optimization using LigPrep in Schrödinger software with OPLS3 force field[35], resulting in the generation of corresponding 3D conformations. Subsequently, the Epik algorithm[36] was employed for ionization handling at a pH value of 7.0 ± 2.0. The target protein structure was retrieved from PDB database[37]. The grid

box on SERT was defined using the co-crystalized ibogaine. During the IFD, 20 poses were generated for each compound. The IFD complexes were evaluated and ranked based on energy using the Prime[38] and Glide XP scoring function[39]. All graphical images are generated using the PyMOL software[40].

3. Results and discussion

3.1. Pre-trained model of CNSMolGen

As shown in **Fig. 3**, the pre-trained model of CNSMolGen network was composed of four groups of layers, including Batch Normalization[41], LSTM layers (each LSTM layer group consisted of a forward and a backward LSTM layer), Batch Normalization, linear. For the pre-trained model of CNSMolGen, the parameters including the number of epochs as well as the size of the hidden units and the number of layers of LSTM layers were important and were explored.

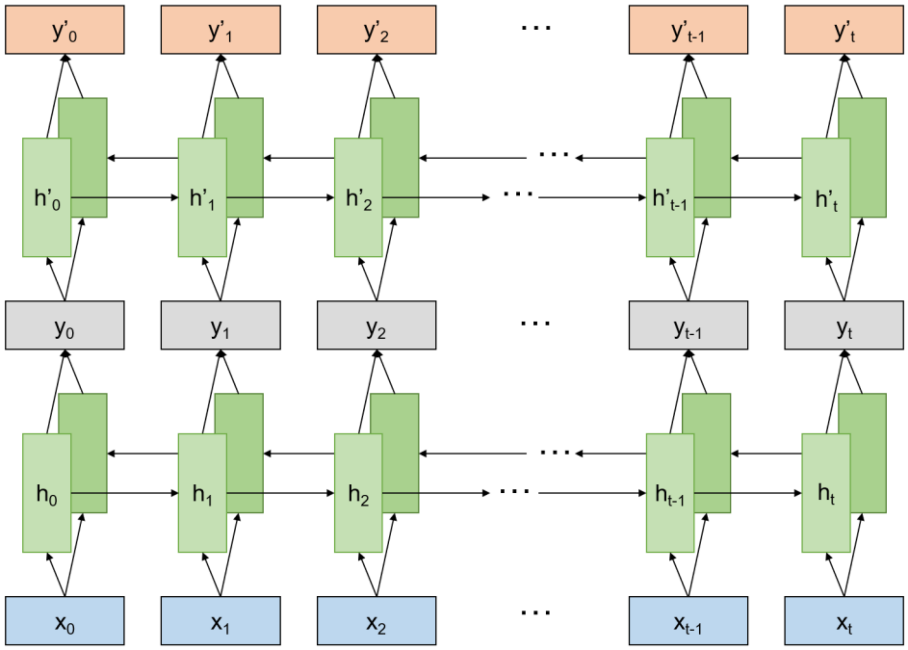

Fig. 3. Schematic diagram of two layers of bidirectional LSTM. The output result of the first layer LSTM, *y*, is used as input to the second layer LSTM. Each layer of bidirectional LSTM (green) consists of two independent LSTM layers, a forward layer and a reverse layer. These two layers process the forward and reverse versions of the input sequence, respectively.

The pre-trained model training: 20 epochs

In this study, we used cross-validation to assess the performance of the models and compared them by calculating the average results of three cross-validations[42]. The performance of the model is measured by the loss values on the validation set (**Fig. 4**). During the initial 20 epochs of training, the loss values on the validation set continued to decrease and stabilize, indicating that neither overfitting nor underfitting occurred during model training. Therefore, in subsequent experiments, we chose the training results of the 20th epoch as the evaluation criterion.

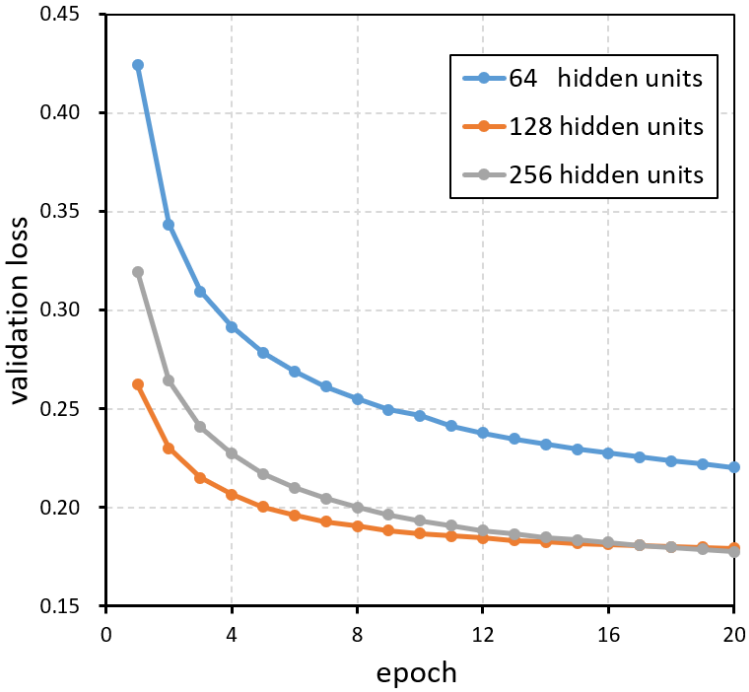

Fig. 4. Cross-entropy loss of the pre-trained model of CNSMolGen. As the number of hidden units increases, the complexity of the model increases and more epochs are required to learn the patterns in the data. At the 20th epoch the loss curve levels off, which means that the model has started to converge.

Parameter selection: hidden units and LSTM layers

We then adjusted two key parameters in the pre-trained model: the size of the hidden units and the number of LSTM layers[43]. This was done to evaluate and compare the model's ability to learn to generate SMILES strings under different configurations and to determine the optimal parameter settings.

At the 20th training epoch, the models with different parameter settings were able to generate more than 70% of novel, unique and valid SMILES strings. This result indicates that by pre-trained on a large set of molecules, the model has effectively mastered the SMILES syntax and is able to reliably generate molecules with potential drug properties **(Fig. 5A**). Adjusting the size of the hidden layer had a significant effect on training speed and model performance. For example, when the hidden layer size was set to 64, the ability to generate new molecules was relatively weak despite the shortest training time. In contrast, when the hidden layer size was set to 128 or 256, the model showed very similar and positive results, with around 90% of the molecules generated being valid and completely new. At the same time, increasing the number of LSTM layers resulted in the model requiring more training cycles to converge (**Fig. 5B**). Considering the training effect and the required computational resources, we found that the model with two LSTM layers showed an excellent ability to generate molecules at the 20th training epoch, successfully synthesizing more than 90% of the novel molecules.

Therefore, we chose the model with 2 sets of LSTM layers and each LSTM layer consisting of 128 hidden units, giving a total of 512 hidden units, as the default pre-trained model for the subsequent study.

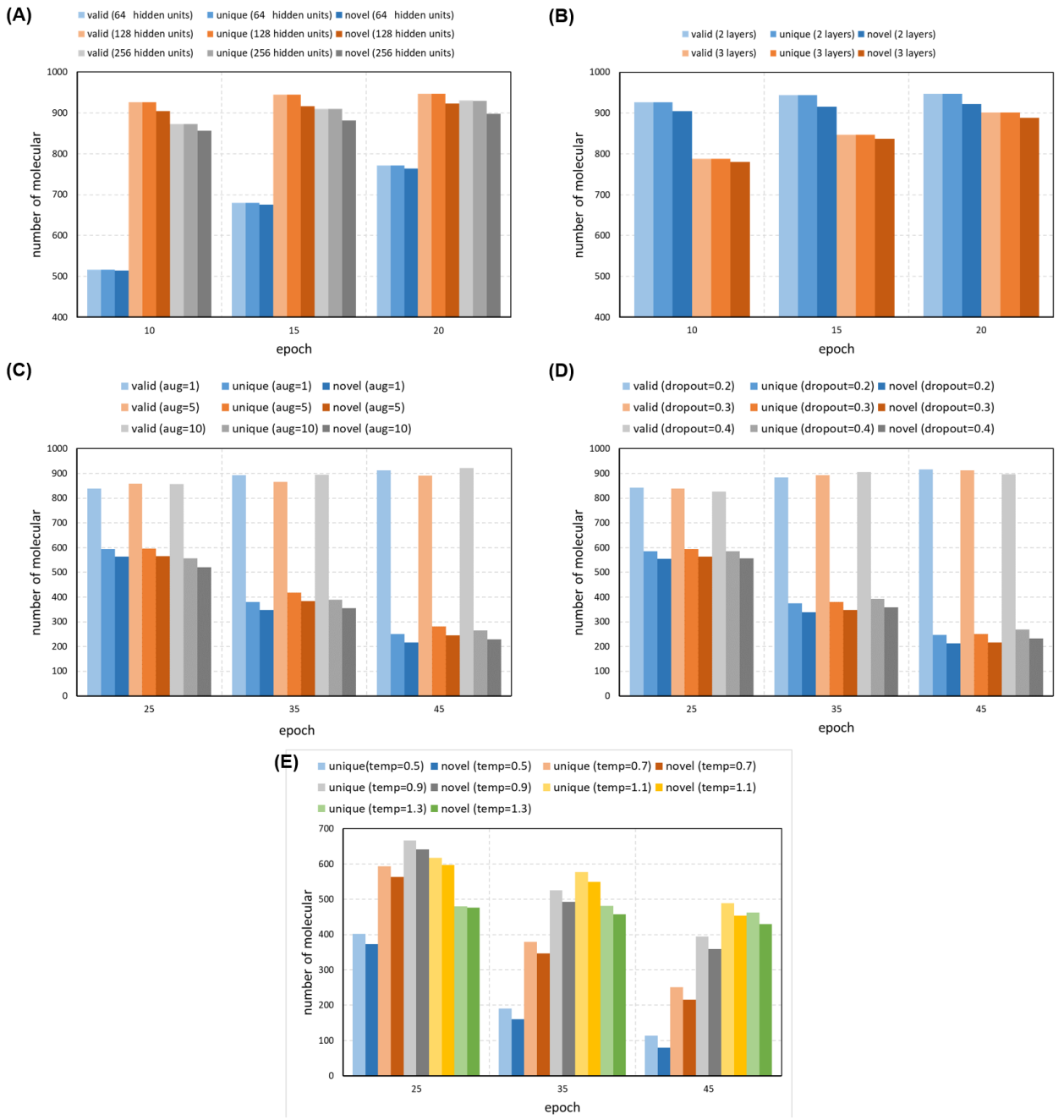

Fig. 5. Statistics on the number of molecules generated by the model under different parameter settings. (A) Number of molecules generated by pre-trained models with different sized hidden layers. (B) Number of molecules generated by pre-trained models with different number of recurrent neural network layers. (C) Number of molecules generated by the fine-tuning model with different data augmentations. (D) Number of molecules generated by the fine-tuning model with different dropout. (E) Number of molecules generated by the fine-tuning model with different sampling temperature.

3.2. Molecular generation ability of the pre-trained model

The molecular generation ability of the pre-trained model of CNSMolGen was further investigated by evaluate the MPO scores and SAScore.

Evaluation of CNS properties: MPO scores

To verify whether the properties of the generated molecules match those of the training data, we randomly selected 100 molecules from the 20th epoch of the ChEMBL-MPO training set and calculated their attributes, including LogP[44], LogD, polar surface area[45], molecular weight, HBD, and pKa[46], along with their average MPO scores[31]. The high degree of similarity between the two sets indicates that the generated molecules have successfully replicated the characteristics of the training molecules and conform to the physicochemical properties of CNS drugs (**Table 1**).

Table 1. Comparison of molecular properties between generated molecules by pre-trained model and pre-training datasets (ChemBridge-MPO)

Property	Generated	ChemBridge-MPO
Log_P	2.21	2.55
Log_D	1.89	2.16
MW	345.94	339.59
TPSA	57.90	61.23
HBD	0.71	1.1
PEA	4.55	5.66
CNS	5.35	5.27
SAScore	2.59	2.79

Evaluation of synthesizability: SAScore

We used the SAScore to assess the synthetic difficulty of the generated molecules. This method describes the accessibility of molecule synthesis as a score between 1 (easy to make) and 10 (very difficult to make) based on molecular complexity and fragment contribution.

The average SAScore of molecules in ChemBridge-MPO dataset is 2.79, while molecules generated by the pre-trained model have a lower average SAScore of 2.59. This suggests that molecules generated by CNSMolGen's pre-trained model generally have a lower synthetic difficulty and, compared to molecules in the pre-trained set, the synthetic accessibility scores of these new molecules have decreased. This result confirms that the CNSMolGen model performs well in generating molecules that are easy to synthesize.

3.3. De novo design of focused library by fine-tuning model

The second part of CNSMolGen contains a transfer learning model. In this part, transfer learning updates the pre-trained model using a specific data set. In our study, we chose the fine-tuning dataset containing 36 compounds to fine-tune the pre-trained model.

Accurate parameter settings are critical for training deep learning models, as these choices impact the learning duration and the quality of the generated outcomes. During molecular generation, the greater the number of SMILES strings produced, the higher the probability of exploring the chemical space. Here, we opted to generate 1,000 SMILES strings per fine-tuning cycle.

The fine-tuning model training: epoch

Transfer learning models have additional constraints on the properties of the molecules generated during training compared to pre-trained models. Extending the fine-tuning epoch of the model may result in the generation of molecules that are closer to those in the fine-tuned dataset, but it may also result in a decrease in the diversity and uniqueness of the molecules generated. Therefore, we chose to perform 45 epochs of fine-tuning as the loss values converged, and used the molecules generated in the last 5 epochs for subsequent parameter analyses. To optimize the training results, we will also tune and analyze the following three parameters: argument, dropout and temperature.

Parameter selection: data augmentation, dropout and temperature

Data augmentation, the practice of artificially increasing the volume of training data to improve model performance, was implemented by altering the starting position of the marker, allowing the model to learn from any point within the SMILES encoding. Comparative analysis showed that quintupling the data augmentation in the last five training cycles yielded the best effectiveness and novelty in the generated compounds (**Fig. 5C**).

Dropout, a technique to combat overfitting by deactivating a subset of neurons during training, slows down the training rate as the dropout value increases. After altering the dropout value, the overall model showed no significant difference in molecular generation outcomes, leading to the selection of an intermediate value of 0.3 (**Fig. 5D**).

The model, trained in this study, employed temperature sampling for SMILES encoding. At lower temperature values, token generation is primarily based on estimated probabilities. However, as the temperature value increases, token generation tends towards uniform probability selection. The effectiveness of generating new molecules decreases with higher temperature values due to the increased randomness of token generation, leading to errors in the SMILES strings, such as missing or mismatched parentheses or ring closures. Nonetheless, higher temperature values exhibited superior performance in terms of the novelty and effectiveness of generated molecules. This is attributed to the increased randomness of token generation, which enhances molecular diversity and expands the explorable chemical space. In this experiment, the model demonstrated optimal uniqueness and novelty in generated molecules at a temperature value of 1.1 (**Fig. 5E**).

Taking all of these factors into account, we determined that setting the data augmentation value to 5, the temperature parameter to 1.1 and the dropout value to 0.3 would serve as the default parameter settings for the fine-tuning model in subsequent experiments.

3.4. Comparison of molecules generated by pre-trained model with molecules generated by fine-tuning model

In this section we compare the properties of the molecules generated by the pre-trained and fine-tuned models, as well as the properties of the molecules in the pre-trained and fine-tuned datasets. We randomly select 100 molecules from each molecule source to analyse (if the number of molecules from a source is less than 100, all molecules are used) and compare the properties of these molecules separately. In **Fig. 6**, the molecules in the ChemBridge-MPO dataset have the highest average MPO scores, while the molecules generated by the pre-trained model have the next highest MPO scores. This indicates that CNSMolGen as a molecule generation model has a strong learning capability to effectively master the rules of molecular SMILES representation and accurately capture the features of CNS molecules. In addition, the average MPO scores of the model-generated molecules decreased after transfer learning, which may be related to the lower average MPO scores of the molecules in the transfer learning dataset. Nevertheless, the MPO scores of the molecules generated after transfer learning were still high and met the MPO criteria for CNS

drugs, indicating that they are potential CNS drug candidates. It is important to note that of all the parameters affecting CNS scores, the molecules generated by deep learning had lower scores for LogP and LogD, suggesting future improvement of molecule generation models.

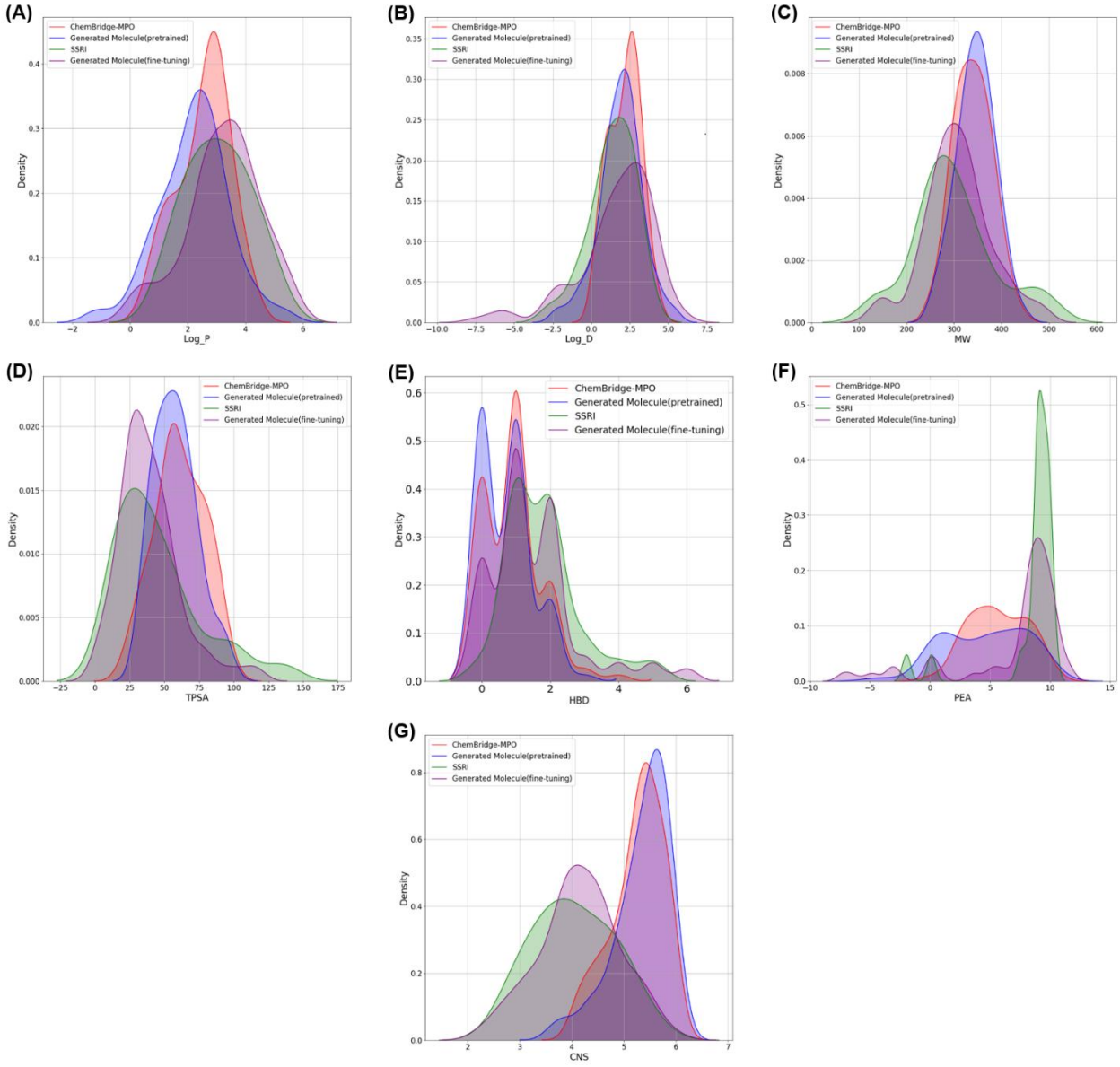

Fig. 6. Quantitative evaluations for molecules generated by fine-tuning model of CNSMolGen. (A-F) Distributions of properties, including LogP, LogD, Molecular weight (MW), total polar surface area (TPSA), number of hydrogen bond donors (HBD) and pKa. (G) The CNS-MPO of the generated molecule was calculated based on the above six properties. When the CNS-MPO score of the molecule is greater than or equal to 4, the molecule can be considered to be more compatible with the properties of CNS drugs.

3.5. CNSMolGen based on ChemBridge-MPO dataset show greatest performance

To investigate how the selection of pre-trained datasets affects the molecular design capability of CNSMolGen, two additional datasets were selected for analysis, the ChemBL-MPO dataset and the ChemBL-ACT dataset. A detailed description of the ChemBL-MPO dataset and the ChemBL-ACT dataset collection methods can be found in the Supporting Information. We trained these two datasets using the same training procedure and default parameter settings and compared the results

of the molecules generated. The results show that CNSMolGen based on the ChemBridge-MPO dataset performs best.

In this study, we used different pre-training datasets as well as a unified transfer learning dataset to train the CNSMolGen model and compared the results for the compounds generated (**Table 2**). The results showed that the model generated the highest number of new compounds when using the ChemBridge-MPO dataset, followed by the use of the ChemBL-smiles dataset, while the lowest number of new compounds was generated when using the ChemBL-MPO dataset. This phenomenon may be related to the number and diversity of compounds in the datasets. The SAScore of the three datasets were further analyzed, and it was found that the average SA scores of the compounds when using the ChemBridge-MPO, ChemBL-MPO and ChemBL-smiles datasets were 2.77, 2.69 and 2.63, respectively, which were relatively close and with low mean values, suggesting that different training datasets have a synthetic approachability is limited. Subsequently, we compared the CNS scores of the three datasets and found that the average CNS scores corresponding to the ChemBridge-MPO, ChemBL-MPO and ChemBL-smiles datasets were 4.15, 4.19 and 4.35, respectively, which suggests that differences in the training datasets do not affect the chemical properties of compounds generated by the model. This suggests that the change in the pre-trained dataset does not have an impact on the final properties of molecule generation, but it is worth noting that the largest number of molecules were generated after transfer learning when ChemBridge-MPO was selected as the pre-trained dataset, followed by ChemBL-MPO. We analyzed this to be due to the fact that ChemBridge-MPO contains more data, while the molecules in ChemBL-MPO are more consistent with the results of transfer learning. This result suggests that the size of the original dataset will have a more direct impact on the final generated results during the transfer learning process. The selection of suitable similar datasets will also have a good impact on the generation of molecules.

Table 2. Comparison of generation results of three different pre-training datasets corresponding to CNSMolGen model

	ChemBL-ACT	ChemBL-MPO	ChemBridge-MPO
CNS-MPO	4.35	4.19	4.15
SAScore	2.63	2.69	2.77
No. of molecules	44	65	215

3.6. Verification of the de novo designed molecules targeting SERT

In order to entirely evaluate the promising application of the CNSMolGen in the field of CNS drug design, the 215 candidate molecules generated from the fine-tuned model were submitted for a comparative analysis of their binding modes and binding affinities to target protein SERT[47]. Notably, four approved drugs that cocrystalized with SERT were used as controls. They are escitalopram-SERT (PDB ID: 5I71)[48], sertraline-SERT (PDB ID: 6AWO)[49], fluvoxamine-SERT (PDB ID: 6AWP)[49], and paroxetine (PDB ID: 6W2C)[50], respectively. Here, all of the 215 molecules were first docked into the ibogaine binding pocket of SERT by considering the pocket flexibility. And the ibogaine binding site was selected because of the relatively large volume of the pocket[47], which were more suitable for IFD study of molecules with different scaffolds.

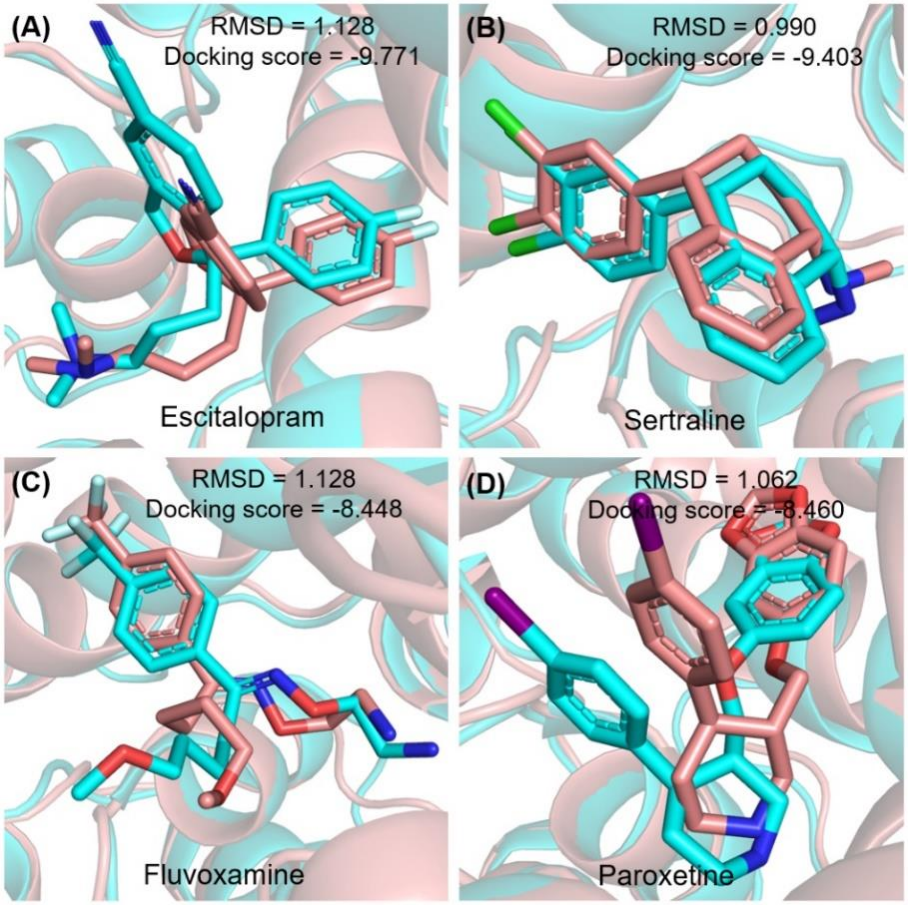

Fig. 7. The conformational superimposition between the crystal structure (cyan) and induced fit docking poses (salmon) of (A) escitalopram, (B) sertraline, (C) fluvoxamine, and (D) paroxetine.

As a result, the redocking scores of escitalopram, sertraline, fluvoxamine, and paroxetine were -9.771 kcal/mol, -9.403 kcal/mol, -8.448 kcal/mol, and -8.467 kcal/mol, respectively (**Fig. 7**). For all of the 215 generated molecules, using the docking score of fluvoxamine (-8.448 kcal/mol) and escitalopram (-9.771 kcal/mol) as a reference, the number of generated molecules that have potential higher binding affinities were 129 (60.00%) and 89 (41.40%), respectively. Thus, compared to the known SSRI molecules, the generated molecules (e.g., compounds **1** (-10.435 kcal/mol) and **2** (-10.453 kcal/mol)) may have comparable or even better binding affinities. The detailed binding mode of the compounds **1** and **2** in SERT were shown in **Fig. 8A** and **8B**. Among them, the ligand-receptor interaction of compound 1 is mainly due to the hydrogen bond interaction formed by amino acid residues Y95, A96, S336 with the ligand; compound 2 is mainly due to the hydrogen bond interaction of D98, F335 with the ligand. In addition, it was found that although the generated molecules have different structures from the approved drugs, they share similar binding conformations at the pocket (**Fig. 8C** and **8D**). Despite the inherent limitations of molecular docking techniques, the results of this study suggest that the CNSMolGen opens up new possibilities for drug design targeting SERT. However, the actual biological effects of the generated molecules on SERT need to be further verified by subsequent experimental studies.

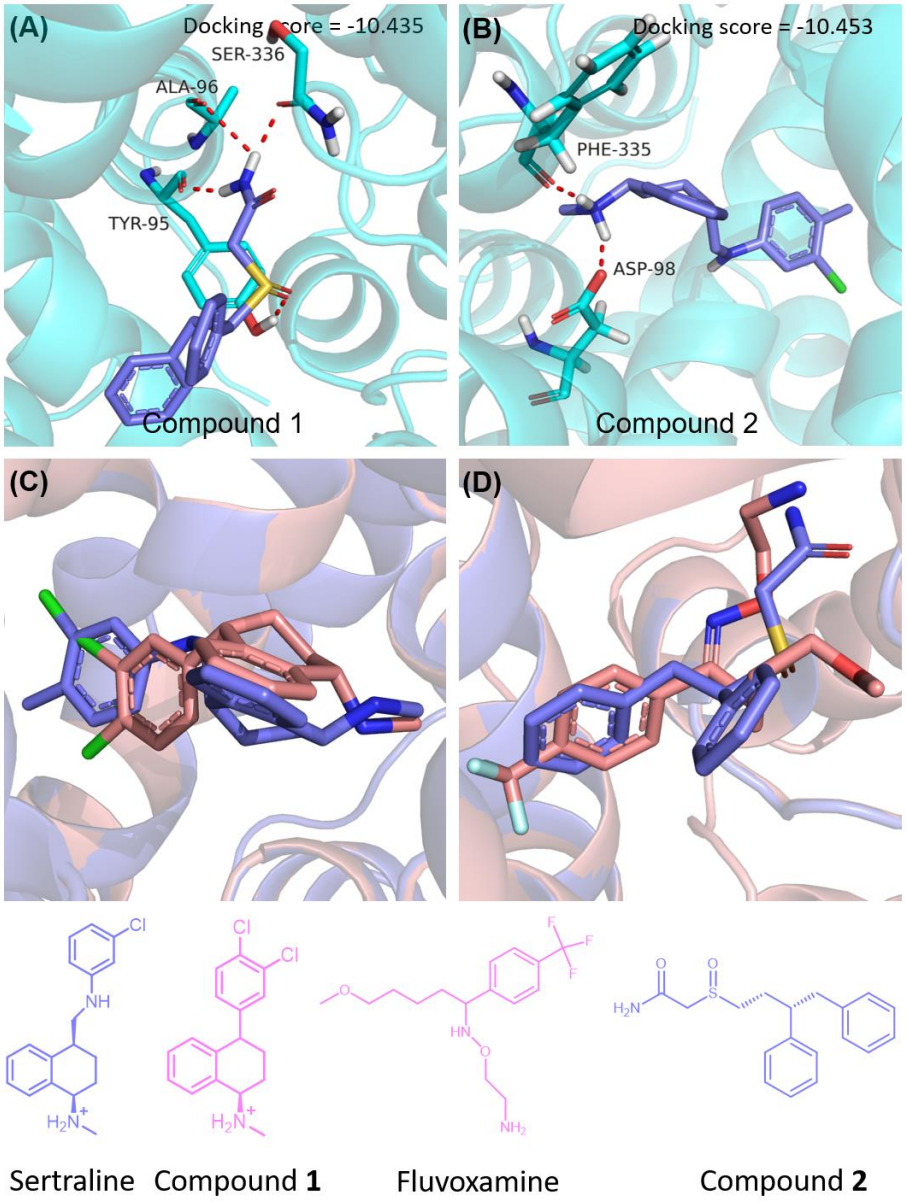

Fig. 8. Induced fit docking poses of the generated compounds in SERT. (A) the interactions of compound **1** (slate) and SERT (cyan). (B) the interactions of compound **2** (slate) and SERT (cyan). (C) Structural superimposition of sertraline (slate) and compound **1** (salmon). (D) Structural superimposition of sertraline (slate) and compound **2** (salmon). The hydrogen bond was represented as a red dashed line.

4. Conclusions

In this study, we introduce CNSMolGen - a novel molecular generation model designed to design central nervous system (CNS) drugs from scratch. The model combines pre-training and transfer learning to generate new compounds with CNS activity by learning the SMILES strings of the molecules and shows superior performance on small datasets, opening up new avenues for drug design. Further, we investigated the effect of using different pre-training datasets on the model performance and found that pre-training on large datasets helps to improve the performance of the generated molecules, while pre-training on CNS-specific datasets is more favorable for generating molecules with CNS properties. The application of CNSMolGen in the actual generation of CNS drug molecules was validated using SSRI as an example. The model generated a total of 215

molecules, 129 of which showed good binding affinity in molecular docking. In conclusion, the CNSMolGen model demonstrated significant effectiveness in generating CNS drugs, especially in handling small datasets, and also demonstrated its potential to provide valuable support for future drug design and discovery. With the increasing role of AI tools in drug design and optimization, experimental validation of their results remains indispensable.

Data and Code Availability

The training datasets used in this work were provided as csv files in https://github.com/xueww/CNSMolGen/data. The code of CNSMolGen and other in-house script for data analysis are open for academic usage and available in the https://github.com/xueww/CNSMolGen/, and the source code partially referenced from https://github.com/ETHmodlab/BIMODAL/. Models in this work were trained on a NIVDIA GTX 3080. Programming language: Python. Other requirements: Python3.7, Tensorflow, RDKit.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (21505009), the Natural Science Foundation of Chongqing (2023NSCQ-MSX0140), the Open Project of Central Nervous System Drug Key Laboratory of Sichuan Province (230012-01SZ).

References

[1] Y. Hou, X. Dan, M. Babbar, Y. Wei, S.G. Hasselbalch, D.L. Croteau, V.A. Bohr, Ageing as a risk factor for neurodegenerative disease, Nature reviews. Neurology, 15 (2019) 565-581.
[2] V.K. Gribkoff, L.K. Kaczmarek, The need for new approaches in CNS drug discovery: Why drugs have failed, and what can be done to improve outcomes, Neuropharmacology, 120 (2017) 11-19.
[3] J.J. Danon, T.A. Reekie, M. Kassiou, Challenges and Opportunities in Central Nervous System Drug Discovery, Trends in Chemistry, 1 (2019) 612-624.
[4] M. Charveriat, V. Lafon, F. Mouthon, L. Zimmer, Innovative approaches in CNS drug discovery, Therapie, 76 (2021) 101-109.
[5] G. Tu, T. Fu, G. Zheng, B. Xu, R. Gou, D. Luo, P. Wang, W. Xue, Computational Chemistry in Structure-Based Solute Carrier Transporter Drug Design: Recent Advances and Future Perspectives, Journal of chemical information and modeling, (2024).
[6] M.H. Cheng, I. Bahar, Monoamine transporters: structure, intrinsic dynamics and allosteric regulation, Nature structural & molecular biology, 26 (2019) 545-556.
[7] W. Xue, T. Fu, G. Zheng, G. Tu, Y. Zhang, F. Yang, L. Tao, L. Yao, F. Zhu, Recent Advances and Challenges of the Drugs Acting on Monoamine Transporters, Current medicinal chemistry, 27 (2020) 3830-3876.
[8] W.P. Walters, R. Wang, New Trends in Virtual Screening, Journal of chemical information and modeling, 60 (2020) 4109-4111.
[9] S. Deng, H. Zhang, R. Gou, D. Luo, Z. Liu, F. Zhu, W. Xue, Structure-Based Discovery of a Novel Allosteric Inhibitor against Human Dopamine Transporter, Journal of chemical information and modeling, 63 (2023) 4458-4467.
[10] C.J. Chen, C. Jiang, J. Yuan, M. Chen, J. Cuyler, X.Q. Xie, Z. Feng, How Do Modulators Affect the Orthosteric and Allosteric Binding Pockets?, ACS chemical neuroscience, 13 (2022) 959-977.
[11] P.G. Polishchuk, T.I. Madzhidov, A. Varnek, Estimation of the size of drug-like chemical space based on GDB-17 data, Journal of computer-aided molecular design, 27 (2013) 675-679.
[12] A.A. Sadybekov, A.V. Sadybekov, Y. Liu, C. Iliopoulos-Tsoutsouvas, X.P. Huang, J. Pickett, B. Houser, N. Patel, N.K. Tran, F. Tong, N. Zvonok, M.K. Jain, O. Savych, D.S. Radchenko, S.P. Nikas, N.A. Petasis, Y.S. Moroz, B.L. Roth, A. Makriyannis, V. Katritch, Synthon-based ligand discovery in virtual libraries of over 11 billion compounds, Nature, 601 (2022) 452-459.
[13] S. Vatansever, A. Schlessinger, D. Wacker, H.U. Kaniskan, J. Jin, M.M. Zhou, B. Zhang, Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions, Medicinal research reviews, 41 (2021) 1427-1473.

[14] R. Gomez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernandez-Lobato, B. Sanchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, ACS central science, 4 (2018) 268-276.
[15] D.M. Anstine, O. Isayev, Generative Models as an Emerging Paradigm in the Chemical Sciences, Journal of the American Chemical Society, 145 (2023) 8736-8750.
[16] M. Ma, X. Zhang, L. Zhou, Z. Han, Y. Shi, J. Li, L. Wu, Z. Xu, W. Zhu, D3Rings: A Fast and Accurate Method for Ring System Identification and Deep Generation of Drug-Like Cyclic Compounds, Journal of chemical information and modeling, 64 (2024) 724-736.
[17] M. Wang, Z. Wang, H. Sun, J. Wang, C. Shen, G. Weng, X. Chai, H. Li, D. Cao, T. Hou, Deep learning approaches for de novo drug design: An overview, Current opinion in structural biology, 72 (2022) 135-144.
[18] Y. Chen, Q. Cheng, Y. Cheng, H. Yang, H. Yu, Applications of Recurrent Neural Networks in Environmental Factor Forecasting: A Review, Neural computation, 30 (2018) 2855-2881.
[19] E.J. Bjerrum, R. Threlfall, Molecular Generation with Recurrent Neural Networks (RNNs), ArXiv, abs/1705.04612 (2017).
[20] C.K. Wu, X.C. Zhang, Z.J. Yang, A.P. Lu, T.J. Hou, D.S. Cao, Learning to SMILES: BAN-based strategies to improve latent representation learning from molecules, Briefings in bioinformatics, 22 (2021).
[21] B. Sattarov, Baskin, II, D. Horvath, G. Marcou, E.J. Bjerrum, A. Varnek, De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping, Journal of chemical information and modeling, 59 (2019) 1182-1196.
[22] F. Grisoni, M. Moret, R. Lingwood, G. Schneider, Bidirectional Molecule Generation with Recurrent Neural Networks, Journal of chemical information and modeling, 60 (2020) 1175-1183.
[23] X.C. Zhang, J.C. Yi, G.P. Yang, C.K. Wu, T.J. Hou, D.S. Cao, ABC-Net: a divide-and-conquer based deep learning architecture for SMILES recognition from molecular images, Briefings in bioinformatics, 23 (2022).
[24] F. Li, J. Yin, M. Lu, M. Mou, Z. Li, Z. Zeng, Y. Tan, S. Wang, X. Chu, H. Dai, T. Hou, S. Zeng, Y. Chen, F. Zhu, DrugMAP: molecular atlas and pharma-information of all drugs, Nucleic acids research, 51 (2023) D1288-D1299.
[25] Y. Zhou, Y. Zhang, D. Zhao, X. Yu, X. Shen, Y. Zhou, S. Wang, Y. Qiu, Y. Chen, F. Zhu, TTD: Therapeutic Target Database describing target druggability information, Nucleic acids research, 52 (2024) D1465-D1477.
[26] W. Xue, P. Wang, B. Li, Y. Li, X. Xu, F. Yang, X. Yao, Y.Z. Chen, F. Xu, F. Zhu, Identification of the inhibitory mechanism of FDA approved selective serotonin reuptake inhibitors: an insight from molecular dynamics simulation study, Physical chemistry chemical physics : PCCP, 18 (2016) 3260-3271.
[27] T.T. Wager, X. Hou, P.R. Verhoest, A. Villalobos, Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties, ACS chemical neuroscience, 1 (2010) 435-449.
[28] J. Wang, Y. Chu, J. Mao, H.N. Jeon, H. Jin, A. Zeb, Y. Jang, K.H. Cho, T. Song, K.T. No, De novo molecular design with deep molecular generative models for PPI inhibitors, Briefings in bioinformatics, 23 (2022).
[29] S.G. Zadeh, M. Schmid, Bias in Cross-Entropy-Based Training of Deep Survival Networks, IEEE transactions on pattern analysis and machine intelligence, 43 (2021) 3126-3137.
[30] L. Yao, M. Yang, J. Song, Z. Yang, H. Sun, H. Shi, X. Liu, X. Ji, Y. Deng, X. Wang, Conditional Molecular Generation Net Enables Automated Structure Elucidation Based on (13)C NMR Spectra and Prior Knowledge, Analytical chemistry, 95 (2023) 5393-5401.
[31] T.T. Wager, X. Hou, P.R. Verhoest, A. Villalobos, Central Nervous System Multiparameter Optimization Desirability: Application in Drug Discovery, ACS chemical neuroscience, 7 (2016) 767-775.
[32] L. Pinzi, G. Rastelli, Molecular Docking: Shifting Paradigms in Drug Discovery, International journal of molecular sciences, 20 (2019).
[33] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, Journal of molecular graphics & modelling, 21 (2003) 289-307.
[34] W. Sherman, T. Day, M.P. Jacobson, R.A. Friesner, R. Farid, Novel procedure for modeling ligand/receptor induced fit effects, Journal of medicinal chemistry, 49 (2006) 534-553.
[35] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J.Y. Xiang, L. Wang, D. Lupyan, M.K. Dahlgren, J.L. Knight, J.W. Kaus, D.S. Cerutti, G. Krilov, W.L. Jorgensen, R. Abel, R.A. Friesner, OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins, Journal of chemical theory and computation, 12 (2016) 281-296.

[36] R.C. Johnston, K. Yao, Z. Kaplan, M. Chelliah, K. Leswing, S. Seekins, S. Watts, D. Calkins, J. Chief Elk, S.V. Jerome, M.P. Repasky, J.C. Shelley, Epik: pK(a) and Protonation State Prediction through Machine Learning, Journal of chemical theory and computation, 19 (2023) 2380-2388.
[37] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic acids research, 28 (2000) 235-242.
[38] M.P. Jacobson, D.L. Pincus, C.S. Rapp, T.J. Day, B. Honig, D.E. Shaw, R.A. Friesner, A hierarchical approach to all-atom protein loop prediction, Proteins, 55 (2004) 351-367.
[39] R.A. Friesner, R.B. Murphy, M.P. Repasky, L.L. Frye, J.R. Greenwood, T.A. Halgren, P.C. Sanschagrin, D.T. Mainz, Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes, Journal of medicinal chemistry, 49 (2006) 6177-6196.
[40] M.A. Lill, M.L. Danielson, Computer-aided drug design platform using PyMOL, Journal of computer-aided molecular design, 25 (2011) 13-19.
[41] N. Talat, A. Alsadoon, P.W.C. Prasad, A. Dawoud, T.A. Rashid, S. Haddad, A novel enhanced normalization technique for a mandible bones segmentation using deep learning: batch normalization with the dropout, Multimedia Tools Appl., 82 (2022) 6147–6166.
[42] S. Kucheryavskiy, S. Zhilin, O. Rodionova, A. Pomerantsev, Procrustes Cross-Validation-A Bridge between Cross-Validation and Independent Validation Sets, Analytical chemistry, 92 (2020) 11842-11850.
[43] F. Arifin, H. Robbani, T. Annisa, N.N.M.I. Ma'arof, Variations in the Number of Layers and the Number of Neurons in Artificial Neural Networks: Case Study of Pattern Recognition, Journal of Physics: Conference Series, 1413 (2019) 012016.
[44] F. Csizmadia, A. Tsantili-Kakoulidou, I. Panderi, F. Darvas, Prediction of distribution coefficient from structure. 1. Estimation method, Journal of pharmaceutical sciences, 86 (1997) 865-871.
[45] P. Ertl, B. Rohde, P. Selzer, Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, Journal of medicinal chemistry, 43 (2000) 3714-3717.
[46] B. Zdrazil, E. Felix, F. Hunter, E.J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D.M. Lopez, J.F. Mosquera, M.P. Magarinos, N. Bosc, R. Arcila, T. Kiziloren, A. Gaulton, A.P. Bento, M.F. Adasme, P. Monecke, G.A. Landrum, A.R. Leach, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, Nucleic acids research, 52 (2024) D1180-D1192.
[47] J.A. Coleman, D. Yang, Z. Zhao, P.C. Wen, C. Yoshioka, E. Tajkhorshid, E. Gouaux, Serotonin transporter-ibogaine complexes illuminate mechanisms of inhibition and transport, Nature, 569 (2019) 141-145.
[48] J.A. Coleman, E.M. Green, E. Gouaux, X-ray structures and mechanism of the human serotonin transporter, Nature, 532 (2016) 334-339.
[49] J.A. Coleman, E. Gouaux, Structural basis for recognition of diverse antidepressants by the human serotonin transporter, Nature structural & molecular biology, 25 (2018) 170-175.
[50] J.A. Coleman, V. Navratna, D. Antermite, D. Yang, J.A. Bull, E. Gouaux, Chemical and structural investigation of the paroxetine-human serotonin transporter complex, eLife, 9 (2020) e56427.

Appendix A. Supplementary data

Details of CHEMBL-MPO and CHEMBL-ACT datasets preparation

ChEMBL-MPO: First, we collected 504,228 molecules in ChEMBL database (https://www.ebi.ac.uk/chembl/) annotated with CX LogP, CX LogD, polar surface, molecular weight, HBD, and CX Basic pKa. Then, the MPO values of these molecules were calculated and filtered using the rules: (1) RO5 violations = 0; (2) 250 ≤molecular weight ≤ 450; (3) 0 ≤ Polar surface area ≤ 100; and (4) MPO ≥ 4. As a result, a total of 412,154 molecules were obtained.

ChEMBL-ACT: a selection of 271,914 compounds with annotated Kd/Ki/IC50/EC50 < 1 μM from the ChEMBL22 database (10.6019/CHEMBL.Database.22.), while removing salts and stereochemical information.