Universiteit Leiden

Master Computer Science

Application of deep learning in lead optimization for improving activity of drug molecules

Name:	Youliang Luo
Student ID:	s2587491
Date:	20/01/2022
Specialisation:	Data Science
1st supervisor:	Dr.Bas van Stein
2nd supervisor:	Dr.Gerard J.P. van Westen

Master's Thesis in Computer Science Leiden Institute of

Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Application of deep learning in lead optimization for improving activity of drug molecules

ABSTRACT

Machine learning techniques have been applied to optimize properties of potential drug-like molecules, such activity, fat solubility and toxicity. Machine translation models, especially recurrent neural networks (RNNs), sequence-to-sequence (seq2seq), and the transformer model, have been shown to have a great capacity of generating novel drug-like compounds or optimizing properties of a promising candidate drug. Improving the activity of a promising candidate drug is an essential task in lead optimization. Inspired by successful applications of deep learning approaches in machine translation, we innovatively regarded the task of improving the activity of a candidate drug as a machine translation problem by applying the seq2seq model and the transformer model, trained end-to-end and totally data-driven. We demonstrate that the transformer model outperforms the seq2seq model in the task of improving the activity of a candidate compound. It is reasonable to infer that the transformer model has great potential to design better drugs.

Keywords: Drug activity, Seq2seq, The transformer model

1. INTRODUCTION

The human adenosine A_{2A} receptor ($A_{2A}R$), a member of the G-protein-coupled receptor (GPCR), has been widely investigated over the last several decades because it has been shown to be a attractive and promising therapeutic target for regulating myocardial oxygen demand and increasing coronary circulation by vasodilation [1]. Due to the enormous body of knowledge and multiple research methods accessible, the adenosine A_{2A} receptor is an enticing topic for medicinal chemists to investigate. Considering the data accessible in the public ChEMBL database, deep learning methods were employed to improve biological activity of drug-like molecules.

During the last decade, deep learning approaches have gained significant progress in a variety of artificial intelligence research fields. This technique, which evolved from prior research on artificial neural networks, has better performance than existing traditional machine learning algorithms in image and speech recognition, natural language processing, and other applications [2]. In several fields, applications of cutting-edge deep learning models outperform humans. [3]. New drug development, which aims to introduce a new effective pharmacological molecule into clinical practice, is costly and time-consuming. Every drug company with an research and development department has taken a variety of steps to accelerate the drug development process [4]. The successful applications of deep learning in these data-rich fields have also sparked chemists' imagination in pharmaceutical research. As a highly imaginative landing scenario for deep learning, drug development is being continuously changed with the effort of scientists. Diverse challenging problems were addressed with the help of deep learning, including compound property and activity prediction, generation of new chemical structures, reactions and retrosynthetic analysis and so on [2].

In drug discovery, A lead compound is a chemical molecule that exhibits pharmacological or biological action that has the potential to be therapeutically effective, but whose structure is suboptimal and requires change in order to better match the target. It may not be developed directly as a new drug due to limitations, such as low activity, low selectivity, poor pharmacokinetics, or high toxicity. The chemical structure of a lead compound serves as a starting point for modifying it chemically to enhance its qualities. [5].

Optimization of the lead structure is a way to improve its activity strength. Deep learning methods can be employed to modify the chemical structures of existing compounds, aiming to improve the chemical properties of drug-like molecules. Deep learning approach has been proven to be a novel method to generate chemical structures [6, 7]. The seq2seq model and the transformer model, mapping a sequence to another sequence, have had significant success in foreign speech translation [8], paraphrase generation [9] and retrosynthetic reaction prediction [10]. Considering the powerful learning capability of the seq2seq model and the transformer model, we exploited these two models as an editing approaches to tweak and optimize the activity of existing compounds.

In our work, improving the activity of candidate drugs is considered a machine translation problem, translating an existing nonactive compound to a new active drug-like compound. The seq2seq model and the transformer model were employed to improve the activity of candidate drugs. Each compound was represented in the Simplified Molecular-Input Line-Entry system (SMILE). The seq2seq model and the transformer model were constructed to learn the syntax of SMILES notation and recognize patterns between inactive compounds and active drug-like molecules. Meanwhile, traditional machine learning methods were used to construct quantitative structure-activity relationship (QSAR) models [11] to predict the activity of output molecules of generative models. We investigated Random Forest (RF) [12], K nearest neighbors (KNN) [13], Support Vector Machine (SVM) [14] and lightGBM [15] to construct a high-performance predictor. For hyperparameter optimization of these four models, grid search and bayesian optimization were employed to

tune these models.
This thesis makes the following technological contributions: 1) The seq2seq model was employed to optimize the biological activity of existing inactive drug-like compounds. Several modifications are made to the original seq2seq model learning in order to make it more suitable for our task.

- Dropout layer was included in order to prevent overfitting during training sessions,
- the Teacher Forcing approach was applied to speed the training of the model.

2) The transformer model with multi-head scaled dot-product attention mechanism was constructed to optimize the activity of existing compounds. 3) Traditional machine learning models were built to predict the activity of output compounds. Meanwhile, grid search and bayesian optimization were used to tune these models.
The remainder of the thesis is structured as follows. In Section 2, we summarise several related work, including sequence to sequence learning in machine translation and deep learning applications in drug discovery. In Section 3, we detail our entire pipeline, from datasets to neural network architectures and metrics for algorithm evaluation. In Section 4, we summarize and analyze the outcomes of our experiments, confirming our models' higher performance. In Section 5, we conclude and discuss the future direction of our work.

2. RELATED WORK

This section discusses the seq2seq and transformer machine translation models, as well as their applications in drug development.

2.1 The seq2seq model and the transform model in machine translation

Natural language processing (NLP) is a broad field, and numerous techniques and algorithms have been developed for text interpretation. In 2014, ilya sutskever et al. [16] proposed a novel neural network, using a multilayered Long Short-Term Memory (LSTM) to map the sequence to a vector, and then another deep LSTM to decode the target sequence from the vector. This sequence to sequence learning method demonstrated superior performance on an English to French translation task. Later on, this novel model was applied in many fields, such machine translation, text summarization, and Chatbot. However, this model has very limited memory, and it does not have sufficient capacity to deal with long sequences. To sidestep the limitation of the seq2seq model, in 2017, Vaswani et al. [17] proposed a novel network architecture, based solely on attention mechanisms, not using RNNs. This novel model was a new state-of-the-art model for English-to-French translation tasks, and it was shown to perform well in a variety of fields. The transform model has a great capacity for dealing with long sequences.

2.2 The seq2seq model and the transform model in drug development

More recently, many state-of-the-art deep learning models have also attracted the attention of researchers in drug discovery. To explore the wide chemical space of drug-like molecules and manufacture novel drug-like chemicals, generative models based on sequence-to-sequence autoencoders have been developed. Winter er al. [18] proposed to exploit the powerful ability of the seq2seq model to learn continuous and data-driven molecular descriptors by translating equivalent chemical representations. Zheng Xu et al. [19] exploited the seq2seq model to provide a continuous feature vector for each molecule for many downstream tasks and demonstrated its superior performance on the classification task. The transformer model was employed to predict retrosynthetic reaction in Pavel Karpov et al [10]. Lukasz Maziarka et al. customized transformer model and proposed Molecule Attention Transformer (MAT) and their experiments showed it is effective in a wide range of molecular prediction tasks. [20].
When training all of the models proposed in these studies, the input and output sequences were identical. The input sequence was `CN1CCC[C@H]1c2cccnc2` (Nicotine), for example, the output sequence was also `CN1CCC[C@H]1c2cccnc2`. Winter er al. failed to train models on translating from canonical SMILES to the International Chemical Identifier (InChI) representations. Their models were unable to learn anything, and the reason they gave might be the higher complexity of the InChI format [18].
The input sequence and the output sequence were different when training our models. Our models' neural network architectures are similar to those in these related work. We trained similar models on different datasets for different purposes.

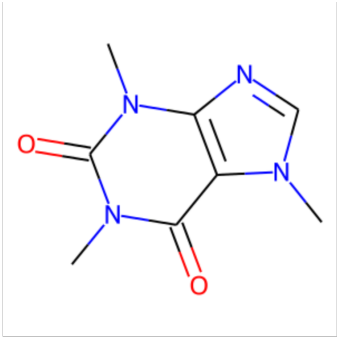
3. DATASET AND METHODS

3.1 Dataset

ChEMBL is a database of bioactive compounds with drug-like characteristics that has been carefully selected. It combines chemical, bioactivity, and genetic data to facilitate the translation of genomic information into successful novel medications [21]. The known ligands for the A_{2A} (ChEMBL identifier: CHEMBL251) from ChEMBL (version 23) were retrieved as the drug discovery target. These known active ligands of the $A_{2A}R$ were used as output molecules when training our generative models. The pChEMBL value allows a number of roughly comparable measures of half-maximal response concentration/potency/affinity to be compared on a negative logarithmic scale [22]. pChEMBL is currently defined as [22]:

$$-\log_{10}(molar IC50, XC50, EC50, AC50, Ki, Kd\ or\ Potency)$$

The compound with the pCHEMBL value ≥ 6.5 was considered as "Active", and others were annotated as "Not Active" [23]. Input molecules for training our generative models were selected from the ChEMBL database (version 28), containing 2,066,376 molecules. Output molecules were known active ligands for target CHEMBL251 from the ChEMBL23 database. All molecules collected from the ChEMBL database were represented by a linear form as a SMILES string. Figure 1 is a 2D depiction of the caffeine molecule and its SMILES string.



CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Figure 1: An example of caffeine molecule and its SMILES string

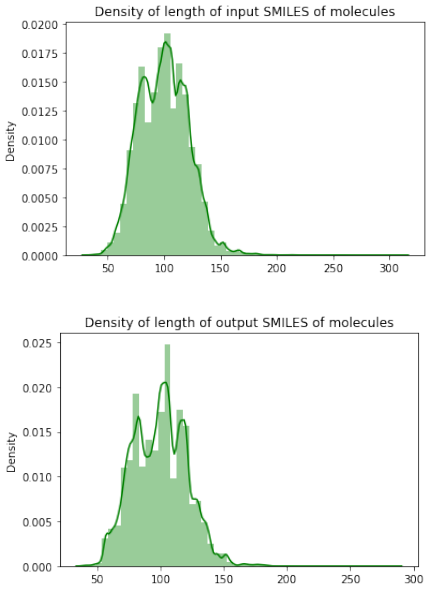


Figure 2: The top plot is the length of general drug-like chemical space (i.e. ChEMBL28) and the bottom the length of chemicals that have been tested/designed for the adenosine a2a receptor (i.e. target ChEMBL251 in ChEMBL)

The Tanimoto coefficient is the most commonly used measure of similarity when comparing chemical structures represented by fingerprints [24], and thus Tanimoto similarity was used to measure the similarity of the input compound and the output active compound. When training our generative models, a molecule from the ChEMBL database (version 28) is the input of generative models, and an active molecule for target CHEMBL251 from the ChEMBL database is the output of generative models when the Tanimoto-similarity score of these two molecules is ≥ 0.4. All the SMILES strings of these compounds were transformed into a series of tokens. The final dataset for our generative models was pairs of molecules represented by SMILES format, including 220,460 samples. And then it was split into a training set containing 170,000 pairs and a testing set containing 50,460 pairs.

Figure 2 shows that the length of the vast majority of input and output molecules is less than 150, which is an important hyperparameter in our generative models. When training the seq2seq model and the transformer model to generate output drug-like molecules, we need to specify the maximum length of output sequence in these two models.
The dataset for training prediction models was also from the known ligands for the $A_{2A}R$ (ChEMBL identifier: CHEMBL251), including 5,157 compounds. The compound was labeled as "Active" when its pCHEMBL value ≥ 6.5, and it was regarded as a positive sample; others were "Not Active" and were viewed as negative samples. This dataset was split balanced as active and inactive compounds, containing 3120 positive samples and 2037 negative samples, respectively. Therefore, the prediction of active or inactive chemicals may be considered a binary classification issue.

3.2 Prediction model (QSAR)

Machine Learning:RF, KNN, SVM and lightGBM
The prediction goal was binary categorization using QSAR modeling. Featurizing molecules is the first and essential step when performing machine learning algorithms on molecular data. Although there is a wide variety of machine-readable chemical representations, a molecule is commonly represented by linear notations as a SMILES string, or by graph form as an adjacent matrix [25]. In our custom dataset, a molecule was already represented with a SMILES string, and its SMILES string could be further converted into its molecular fingerprint. Extended-connectivity fingerprints (ECFPs), which are created using a variation of the Morgan algorithm, are a newly discovered fingerprint technology that is specifically designed to capture molecular properties associated with molecular activity. [26]. In RDKit, Extended-connectivity fingerprints are called Morgan Fingerprints. When generating Morgan fingerprints, the radius of the fingerprint, namely number of iterations, must also be specified. Generally, radius 2 and 3 are commonly used. Taking a cue from Xuhan Liu's paper, the parameter radius was set to 3 in our work. The input data transformed from SMILES of compounds were Extended Connectivity Fingerprint 6 (ECFP6) fingerprints with 4096 bits, which were calculated by the RDkit Morgan Fingerprint algorithm with three-bound radius [23]. Each compound was represented by a 4096 vector of 1s and 0s, indicating the presence or absence of chemical substructures in a compound. The output of the prediction model was the predicted probability of whether an input compound was active. Here, traditional machine learning algorithms, including Random Forest(RF), K nearest neighbors (KNN) and Support Vector Machines (SVMs), were benchmarked by using scikit-learn. Additionally, LightGBM was used since it is a rapid and high-performance gradient boosting framework based on decision tree techniques that has demonstrated outstanding performance in the data science sector. [15].
The options of grid search and bayesian optimization for model optimization were considered to find the optimal hyperparameters.
GridSearchCV, a function from Scikit-learn's package, was used to hypertune our three traditional model parameters.

Compared with RF, KNN and SVM, lightGBM model has more parameters and hyperparameters. Bayesian optimization is more efficient than grid search in obtaining the optimum collection of parameters. As a result, bayesian optimization was applied to optimize lightGBM.

For classification tasks, the random forest output, which is the class picked by the majority of trees, has demonstrated exceptional performance in scenarios when the number of variables exceeds the number of observations [27]. RandomForestClassifier from scikit-learn was employed to construct the model and GridSearchCV was applied to optimise this model. The number of trees in the forest was set 600 and the split criterion was *"gini"*. Other parameters are default. KNN is a non-parametric supervised machine learning approach that is capable of dealing with classification and regression problems. It's simple to set up and comprehend, but it can give highly competitive results. The disadvantage of this simple algorithm is that it becomes noticeably slower as the amount of data increases.
KNeighborsClassifier was used and the optimal value of k was determined by grid search. SVMs are supervised learning models, aiming to find a hyperplane in an N-dimensional space that clearly classifiers the input multi-dimensional data points. Advantages of these methods are that they are effective in high dimensional spaces even in situations where number of features of input data is greater than the number of samples.
In SVM, implemented through scikit-learn, the radial basis function (RBF) kernel was selected. γ was searched between 2^{-15} and 2^{5} and parameter C was set as $\left[2^{-5}, 2^{15}\right]$. Other parameters are default [23].

LightGBM is a framework for gradient boosting that makes use of tree-based learning techniques. This framework is designed with low memory usage, high accuracy, fast training speed and high efficiency and capability of handing large-scale data [15]. LightGBM allows for extensive customization through a range of hyper-parameters. While certain hyper-parameters have a recommended default value that produces generally acceptable outcomes.

Validation metrics
Model validation metrics are needed to measure agreement between a predictive model and real observations. Compared with other common evaluation metrics for binary classification, such as accuracy and F_1 score, the Matthews correlation coefficient (MCC) provides a reliable and informative score when carrying out a proper evaluation of binary classifications [28]. And thus MCC was used to measure and compare the results of these four machine learning models, which could produce an informative and reliable score in evaluating performance of binary classification models [28]. MCC is defined as the following formula

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)\,(tp + fn)(tn + fp)(tn + fn)}}$$

Here, *tp*, *tn*, *fp* and *fn* are the number of true positives, true negatives, false positives and false negatives,respectively.
Meanwhile, the area under the receiver operating characteristics (AUROC) was also applied to compare the performance of these four prediction models.

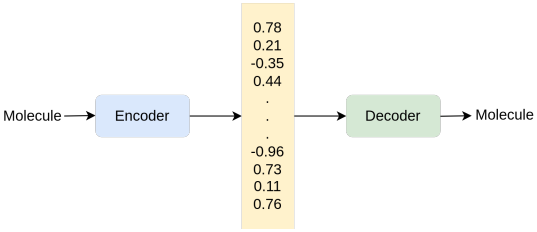


Figure 3: The Encoder-Decoder

3.3 Generative model (seq2seq)

The seq2seq model is a machine translation method based on encoder-decoder. This method maps an input sequence of variable length to an output sequence of another length, and the length of the two sequences may not be equal [16]. This model has been widely used in applications, such as Chatbot, speech recognition and speech generation, machine language translation, text summarization, etc [29, 30, 31]. The encoder block and the decoder block are essential parts of the seq2seq model (Figure 3). The encoder block receives the input sequence and processes a symbol at each time step. It is finally converted into a fixed-length feature vector (context vector). In this process, the encoder block will encode important information in the sequence and lose less important information. The context vector can be regarded as a summary of the entire input sequence. The decoder block gradually generates another output sequence, generating an output symbol at each time step. When the decoder is initialized, it receives the hidden state (context vector) at the last moment and the special symbol of word segmentation <SOS> (the symbol to start decoding). Each subsequent time step receives the hidden state and symbol output at the previous time.

Here, the seq2seq model acted as a baseline and it is a basic model for better understanding the transformer model. In our work, each molecule from ChEMBL28, represented by SMILES format, was featurized as a series of tokens. Then, all these tokens were gathered to construct our input SMILES vocabulary. In our custom dataset, there are 57 and 36 tokens in the input vocabulary set and the output vocabulary set, respectively.

Figure 4 shows the architecture of the seq2seq network. Each molecule was featurized as a series of tokens and then each token was transformed into a 128 dimensional vector. For encoder layer and decoder layer, Long short-term memory (LSTM) was used as the recurrent cell. The output sequence was the selection of tokens from output vocabulary with the maximum probability.
During the training process of this baseline, 3-layer LSTM and 4-layer LSTM were used to compare the performance of different settings of this model.
In the training of the seq2seq model, the Teacher Forcing approach was applied to converge faster. This approach for training RNNs provides observed sequence values as inputs during training and doing multi-step sampling with the network's one-step-ahead predictions [32]. The disadvantage of

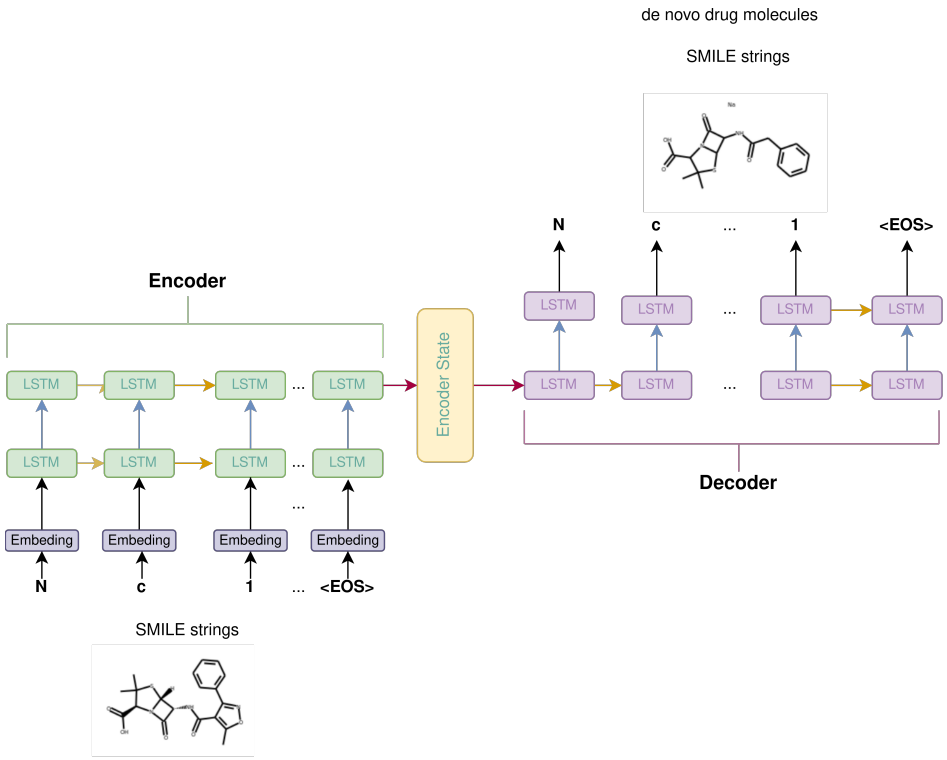


Figure 4: Architecture of seq2seq network for generating drug-like active molecules

this training approach is a discrepancy between training and inference, possibly resulting in poor model performance and instability [33].
The main drawbacks of the seq2seq model are the following.

1. The content vector does not fully represent the information in the entire input sequence, which is equivalent to "lossy compression" of the information.

2. Any token in an input sequence has the same impact on generating an output token without any difference [34].

3.4 Generative model (transformer)

The transformer model was expected to generate drug-like active molecules by modifying the chemical structure of existing compounds in our work. At a high level, this model is based on the encoder-decoder structure (Figure 5). Figure 5 shows that the transformer model, like the seq2seq model, consists of two parts, namely the encoder block and the decoder block, which can transform one sequence into another. But, it differs from the previously described the seq2seq model. The transformer model abandons the RNNs used in the preceding seq2seq model in favor of self-attention or multi-head self-attention, which allows for simultaneous processing of the incoming data and increases operational efficiency.
Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence [17]. An attention mechanism allows the transformer model's encoder and decoder to simultaneously observe the whole input sequence, directly representing these dependencies. The input of Encode block of the transformer model is a series of vectors and its output is a series of different vectors. In practice, the transformer model has a stack of encoder blocks and decoder blocks of the same number. In our work, 3 and 4 encoder blocks were experimented with to examine their effects on our custom dataset.
Unit of multi-head self-attention mechanism is a major component in the transformer and it consists of several scaled-dot attention layers. Attention mechanism with Query-Key-Value (QKV) was adopted in the transformer model. Given the packed matrix representations of queries Q, keys K, and values V, The score of scaled dot-product attention was calculated as follows.

$$attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The closeness between the keys and the queries is indicated by the dot product of the keys and the queries. The value of d_k, a scaling factor, depends on the dimension of the layer [17, 35].
Rather of computing attention once, the multi-head method iteratively computes the scaled dot-product attention. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions [17].

$$MultiHeadAttn(Q, K, V) = Concat(head_1, ..., head_H)W^O$$

$$where \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Here, W_i^Q, W_i^K, W_i^V, W^O are parameter matrices to be learned through training models on datasets.

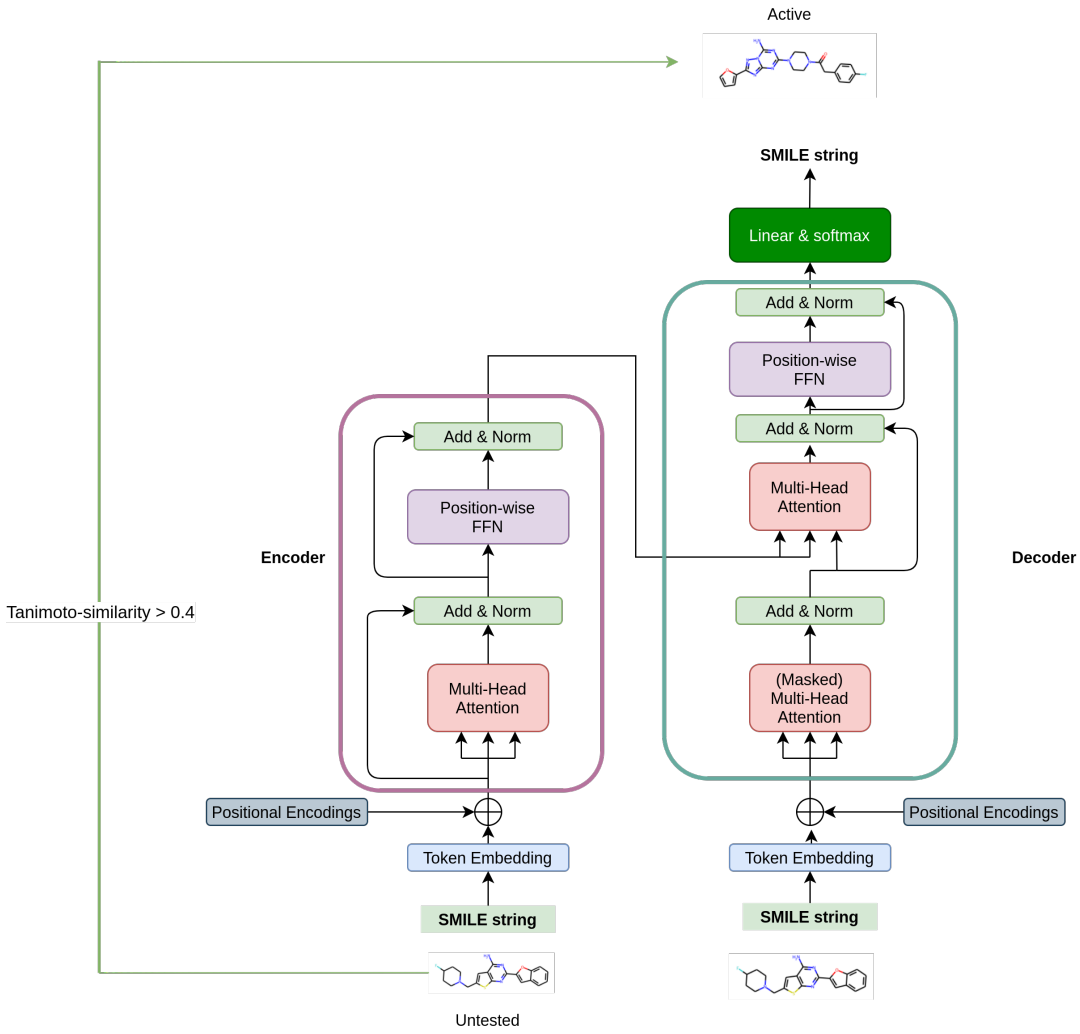


Figure 5: Architecture of the transformer model for generating drug-like active molecules

3.5 Molecular similarity and visualization of chemical space

In chemoinformatics, molecular similarity refers to the structural or functional resemblance of chemical elements, molecules, or chemical compounds. Although there are several chemical similarity methods, Tanimoto-similarity was employed to indicate the similarity of two compounds [36, 37]. Compounds are featurized in ECFP6 fingerprints (Figure 6) and Tanimoto-similarity is defined as the following.

$$T_s(A, B) = \frac{c}{a + b - c}$$

where, c indicates the number of common features of two molecules, $a + b - c$ represents the total number of features of these two molecules. The range of $T_s(A, B)$ is between 0 and 1. When two molecules are identical, its Tanimoto-similarity is 1.

In 1996, Patterson et al. concluded that two compounds have a high probability of having the same activity when their Tanimoto-similarity score is ≥ 0.85 [38]. When we built our custom dataset for training our generative models, the Tanimoto-similarity value of the input compound and the output compound was < 0.85. Meanwhile, to get enough data and ensure the chemical similarity of two compounds, we set the Tanimoto-similarity value ≥ 0.4.

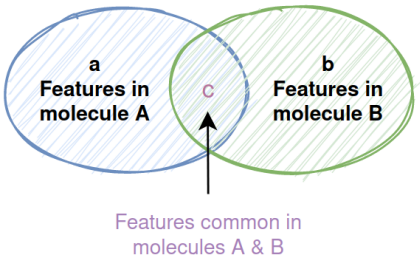


Figure 6: Tanimoto-similarity of two molecules

PCA (Principal component analysis) and t-SNE (t-distributed stochastic neighbor embedding), dimensionality reduction techniques, were used to evaluate chemical space coverage of input data and output data of our generative model. T-SNE

is proven to be a great way to compare the chemical space covered by different datasets [39]. For the input dataset and output dataset, the following processes were conducted:

1. Each molecule, represented in the formation of SMILES, was transformed into ECFP6 fingerprints(4096-dimension).
2. 4096 dimensions of each molecule were reduced to 2 dimensions with PCA and t-SNE [40].
3. The reduced features of each molecule were plotted into 2 dimensional space.

When plotting reduced features with PCA, the amount of variance accounted for by two principal components was checked to evaluate the performance of PCA. The small amount of variance indicates that to preprocess data to obtain a distance or similarity matrix should be considered. Meanwhile, other techniques for data-dimensionality reduction should be considered. For example, t-SNE may be a better approach for visualizing high-dimensional data.

4. EXPERIMENTS AND RESULTS

The known ligands for the $A_{2A}R$ (ChEMBL identifier: CHEMBL251) were featurized with the package RDkit for machine learning models. Each molecule was transformed into ECFP6 fingerprints with 4096 bits by RDkit Morgan Fingerprint algorithm. Here a three-bond radius was set to generate 4096 features that indicate the presence or absence of a particular molecular feature, defined by some local arrangement of atoms.

Grid search with 5-fold cross-validation was employed to search the optimum values of hyperparameters of models, namely RF, SVM and KNN. Table 1 displays optimal parameters of RF and KNN in bold text by applying grid search. In the RF model, the number of trees in the forest was set 600 and the criterion was *"entropy"*. In the SVM model, a radial basis function *"rbf"* was used as the kernel type. Regularization parameter *"C"* was set to 32768.0 and kernel coefficient *"gamma"* was *"auto"*. In KNN model, the value of *"k"* experimented with 3, 5, 7 and 9. The optimal value of *"k"* was set to 7.

In our work, the python package `bayesian-optimization` was used to optimize the optimal values of hyperparameters of lightGBM [41]. In our experiments, the cross-validation AUCROC function was to be maximized. The hyperparamter search space was specified as in Table 2. The optimal hyperparameters are in bold. The other parameters are default. Some other important hyperparameters were tuned, such max number of leaves in one tree, the max depth for tree model, subsample ratio of the training instance.

After the optimal value of hyperparameters of these four models were obtained, we compared the performance of these supervised machine learning algorithms with 5-fold cross validation of which the ROC curves are displayed in Figure 7. As Figure 7 illustrates, the RF model outperforms the other three models, having the highest value of AUC (0.93). Figure 8 shows the RF model performs best among these models in terms of metric MCC on our small dataset. It was chosen

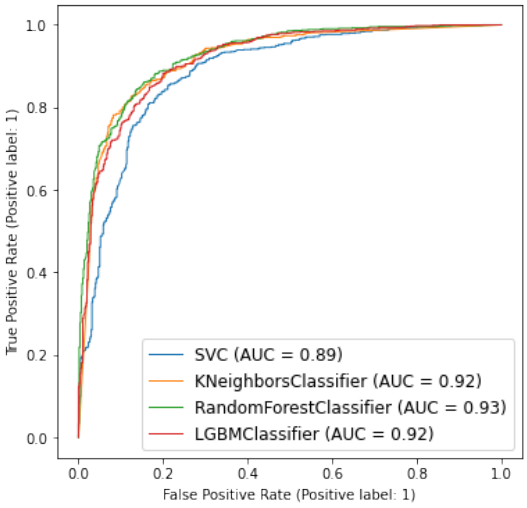


Figure 7: AUC of ROC curve of four machine learning models

to predict the activity of drug-like molecules generated by our generative models due to its the highest value of AUC and MCC.

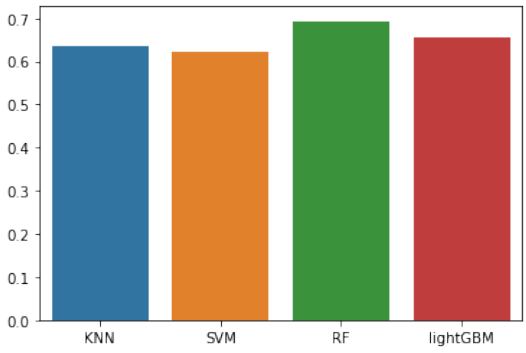


Figure 8: MCC of four machine learning models

4.1 Performance of the seq2seq model

In our work, the seq2seq model was expected to act as a baseline to improve the activity of molecules. The input molecule is from the ChEMBL28 dataset. The test set was used to evaluate the grammatical correctness of drug-like molecules while training the model. The grammatical correctness of a generated SMILES sequence was checked by package `RDKit` during the process of training. In each epoch of training, the percentage of grammatically correct SMILES and the value of the loss function were regarded as metrics for evaluating the model's performance. They were calculated and recorded for visualization.
Throughout the seq2seq model experiments, three parameters settings were experimented on GPU clusters, and some experimental setups are reported in Table 3.

Figure 9 shows the values of loss function and the percentage of valid SMILES sequences in the training process of the seq2seq (M1), seq2seq (M2) and seq2seq (M3). The or-

Table 1: Parameter settings to determine the optimal hyperparameters for RF, KNN, and SVMs

Model	Hyperparameter	Grid Search Parameter Space
RF	number of trees (nt)	300, 400, 500, **600**
	criterion	"gini", **"entropy"**
	bootstrap	True, **"False"**
	minimum samples to split	2, 3, **5**, 7, 10 15
KNN	K	3, 5, **7** , 9, 11, 13, 15, 17, 19
	weight options	uniform, **distance**
SVM	regularization parameter C	$[2^{-5}, 2^{-15}]$
	gamma(γ)	$[2^{-15}, 2^{5}]$

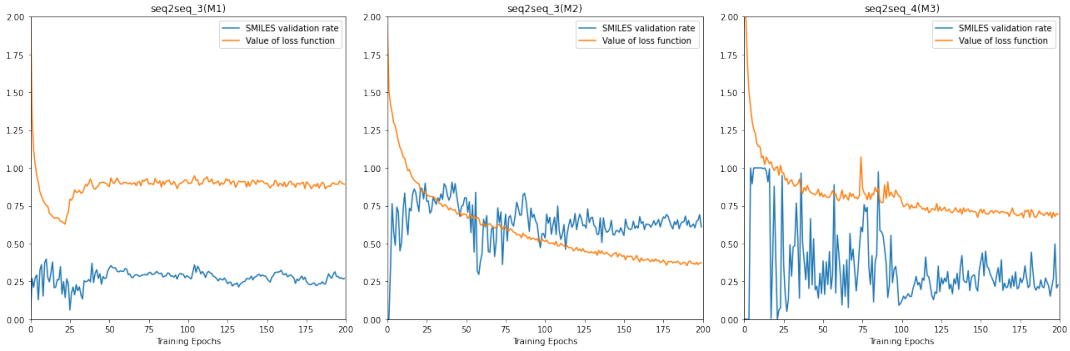


Figure 9: The value of loss function and the percentage of valid SMILES sequences of the seq2seq M1, M2 and M3 during the training process

Table 2: Parameter Search Space of bayesian optimization of lightGBM

Hyperparameters	Parameter Search Space
num_ leaves	(10, 100), **68**
max_depth	(2, 50), **49**
subsample	(0.7, 1.0), **0.8**
colsample_bytree	(0.5, 1), **0.445**
min_child_samples	(3, 30), **13**

ange learning curve was value of loss function on the training set in the process of training. The blue learning curve was the percentage of valid SMILES sequences generated by the seq2seq model when the input data was the test data set.
Model seq2seq (M1) and seq2seq (M2) have 3 layers in encoder block and decoder block while model seq2seq (M3) has 4 layers in both blocks and their other values of parameters are some. Figure 9 shows that after around 50 epochs, the loss function of seq2seq (M1) had converged, while the percentage of valid SMILES sequences was around 32.6%. Model seq2seq (M1) has only 3 layers in encoder block and decoder block and its performer is bad, indicating this model was underfitted. To obtain a more powerful seq2seq model, we considered increasing the number of hidden layers in encoder and decoder blocks. Model seq2seq (M3) has 4 layers in encoder block and decoder block. With the convergence of value of loss function after about 100 epochs, the percentage of valid SMILES sequences of this model is still not stable, fluctuating between 25% to 58%. Here we conclude that the seq2seq model with more layers could not perform better on our custom dataset.
Figure 9 also shows the performance of 3-layer seq2seq (M2) with dropout value 0. The value of the loss function was declining stably after 100 epochs, and then the percentage of valid SMILES sequences was around 78%. Here it is concluded that 3 layers are an optimal setting of seq2seq model on our custom dataset. Due to the limitation of time and computing power, we did not optimize this 3-layer seq3seq model.

4.2 Performance of the transformer model

For the training of the transformer model, all compounds in our custom dataset, in formation of SMILES, were decomposed into the tokens. `RDKit` was used to check the validity of generated SMILES sequence. Some training parameters of this model are shown in Table 4.
This model was obtained by training about 26h on GPU cluster (Nvidia Titanium). As Table 4 shows, the learning rate was set 0.0001 to avoid unstable training, whereas, learning rate was set 0.01 in previous seq2seq model. The learning rate was the most important hyperparameter when we attempt to tune this model [42]. When the learning rate was 0.01, the model could not converge stably. Tiny rates, like 0.00001, resulted in a failure to train. Dropout rate, a regularization method, was set 0.1 to prevent our model from overfitting. The original 2017 transformer [17] consists of 6 encoder layers and decoder layers, whereas there are 2 encoder layers and decoder layers in this transformer model. The main reason of this setting is due to our small custom dataset.
Dot-product attention is a family of attention mechanisms, and it was applied in our transformer model.The unit of multi-head self-attention mechanism was the major component in the transformer model. The number of multi-header is a hyperparameter that needs to be tuned. In our exper-

Table 3: The training parameters of seq2seq models

Parameter	value		
	Seq2seq_3 (M1)	Seq2seq_3 (M2)	Seq2seq_4 (M3)
Batch size	256	256	256
Learning rate	0.01	0.01	0.01
Optimiser	Adam	Adam	Adam
Layers	3	3	4
Hidden dimension	128	128	128
Encoder dimension	256	256	256
Decoder dimension	256	256	256
Encoder dropout	0.5	0	0.5
Decoder dropout	0.5	0	0.5
Trainable parameters	952,100	952,100	1,216,292
SMILES validation rate	0.42	0.76	0.53

iment, the number of encoder headers and decoder headers was 8.

Figure 10 shows learning curves of the transformer model in the training process. The orange learning curve was the value of loss function on the training data set, and the blue learning curve was the percentage of valid SMILES sequences when the input data set was the test data set. Two learning curves indicates that the convergence speed of the loss function changed quickly and this model was trained well and converged after around 50 epochs. After about 50 epochs, the percentage of valid SMILES sequences fluctuated between 83% to 87%. Then we evaluated the quality of drug-like molecules generated by this model.

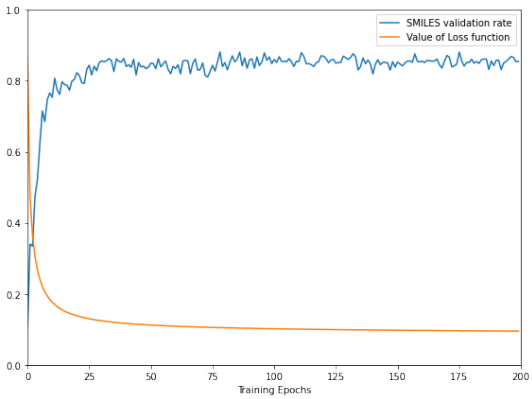


Figure 10: The value of loss function and the percentage of valid SMILES sequences in the training process of transformer model

Visualization of chemical space

Molecular weight (MW) and logP were calculated to explore the chemical space of these two datasets. Figure 11a is the logP∼MW plot, indicating that the vast majority of generated molecules were drug-like, and these two datasets of molecules seem to share the same chemical space.

Subsequently, PCA and t-SNE, were employed for dimensionality reduction and evaluating chemical space coverage of these two datasets. In Figures 11b, and c, 10,000 molecules from the ChEMBL28 dataset, as the input data of the transformer model, are shown in purple, while the generated drug-like molecules are shown in orange. Figures 11b and c were the visualization of two datasets by applying PCA and t-SNE, respectively. Figure 11b displays that there are two distinguishable clusters in PCA space. A possible reason for the two clusters observed here could be that there are two different groups in the test set or the size of the dataset is too small. Figure 11f displays three distinguishable clusters in the generated inactive compounds, indicating that they may have similar characteristics.

The second row in Figure 11 is the visualization of generated molecules categorized as inactive ones and active ones. As Figure 11f shows the generated inactive molecules cluster in 3 cluster in t-SNE space.

Random forest classifier was used to predict the activity of whole molecules in the ChEMBL28 dataset, and only 9.62% of molecules are active. In contrast, the majority of generated drug-like molecules are active, up to approximately 76.39% of the whole output compounds. As an example, five candidate molecules generated by the transformer model were selected, shown in Figure 12. The top row molecules are input data for the transformer model, and the bottom row molecules are generated drug-like molecules. Numbers in the middle of the arrow are the Tanimoto-similarity scores of the input molecule and its generated drug-like molecule. The Tanimoto-similarity of active and inactive molecules generated by the transformer model to every ligands for $A_{2A}R$ (ChEMBL identifier: CHEMBL251) were calculated. And then PCA was used on this similarity matrix 14.

The distribution of Tanimoto-similarity scores of input and its output compounds is shown in Figure 13. The great majority of input molecule and its generated molecule have chemical similarities between 0.2 and 0.6, as can be observed. However, the vast majority of the chemicals produced are active. The transformer model has dramatically improved the bioactivity of existing compounds.

5. CONCLUSION AND FUTURE WORK

The purpose of this study was to compare the performance of the seq2seq and transformer models on the objective of enhancing the activity of existing compounds. This work was motivated by a recent success in English-to-French language translation, dubbed the sequence to sequence learning model. Our generative models translated the inactive molecular SMILES string to another molecular SMILES string, with the expectation that the input molecule and its out-

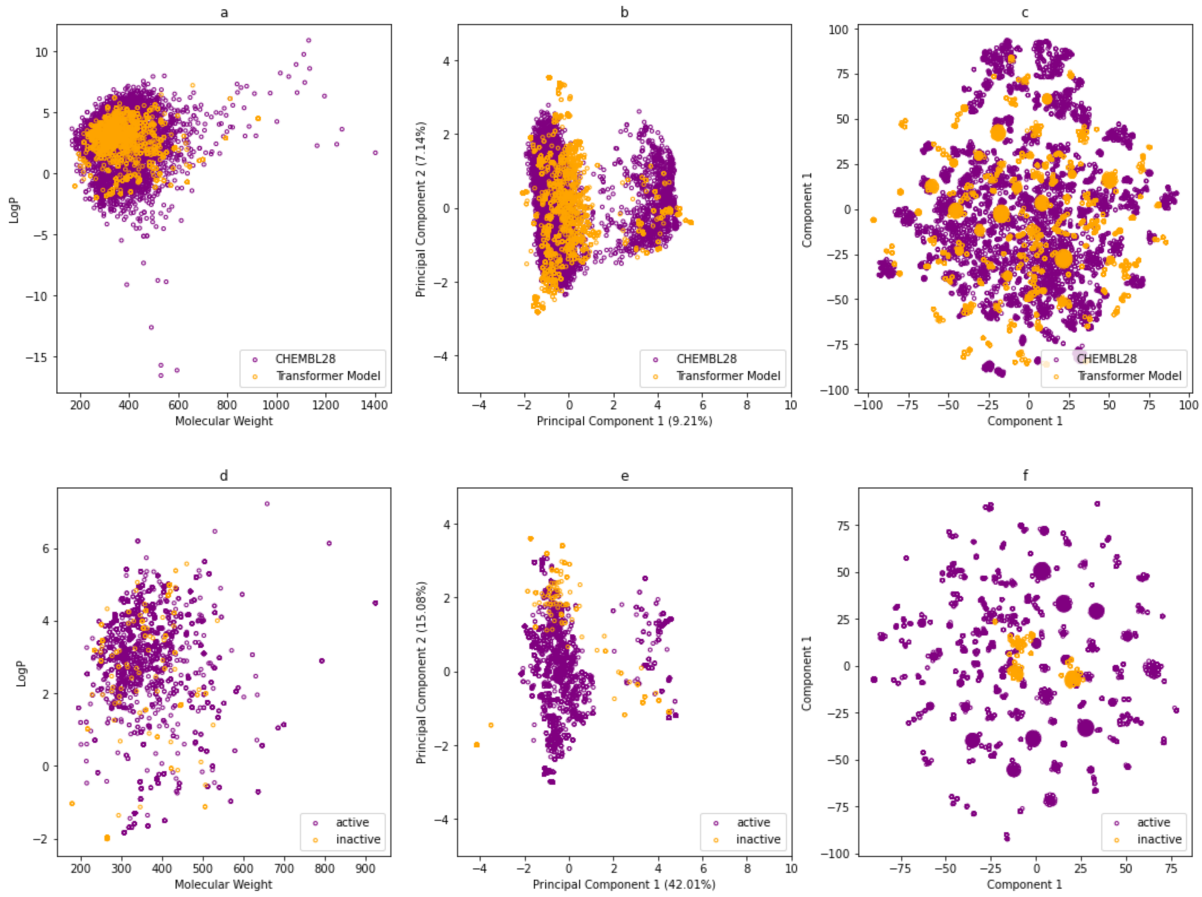


Figure 11: The chemical space of generated molecules by the transformer model with the ChEMBL28 (Only compounds used for training) (a-c). The chemical space was represented by either logP∼MW (a), first two components in PCA on PhysChem descriptors (b), and t-SNE on ECFP6 fingerprints (c). The second row is the visualization of generated inactive and inactive molecules

put drug-like molecule to be as similar as possible in chemical structure. Based on visualization and exploitation of the chemical space of generated drug-like compounds, the transformer model outperform the seq2seq model in terms of improving the activity of compounds by modifying the chemical structure of existing molecules. Our work indicates the transformer model with a multi-head attention, a data-driven sequence to sequence learning method, has a potential to optimize the property of compounds.
In future work, data-driven molecular fingerprints may be applied for activity prediction since a SMILES string could be further converted into different formats. Many sophisticated deep learning models have been constructed to generate molecular representation for the downstream tasks. Molecular fingerprint, such as seq2seq fingerprint or mol2vec [43], maybe used as molecular representation for approaching molecular machine learning tasks in drug discovery.

6. ACKNOWLEDGEMENTS

Youliang thanks Dr.Gerard J.P.van Westen for providing me with this interesting research topic and many valuable suggestions. Thank Dr. B.van stein for helping me with the project management of my master thesis. Last but not least, Many thanks to Dr.Xuhan Liu for providing expert guidance and evaluating the results of experiments.
In practical work, `RDkit`, `Scikit-learn` and `Pytorch` were used in our experiments. Here we thank the contributors of these packages.

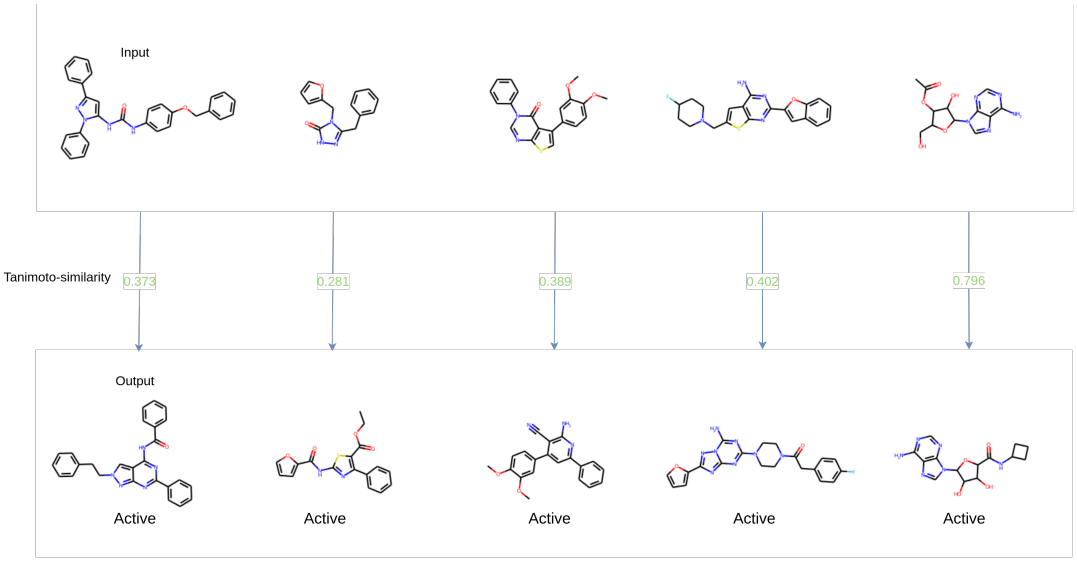


Figure 12: 5 input compounds and output drug-like active compounds generated by the transformer model

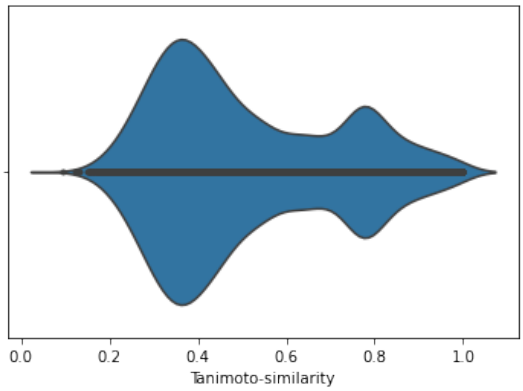


Figure 13: Distribution of Tanimoto-similarity scores of input and output compounds of the transformer model

Table 4: The training parameters of transformer model

Parameter	Value
Batch size	256
Learning rate	0.0001
Dropout	0.1
Encoder layers	2
Decoder layers	2
Encoder headers	8
Decoder headers	8
Epoch	200
Optimizer	Adam
Loss function	Cross entropy
Attention Mechanisms	Dot product
Trainable parameters	1,970,468
SMILES validation rate	0.87

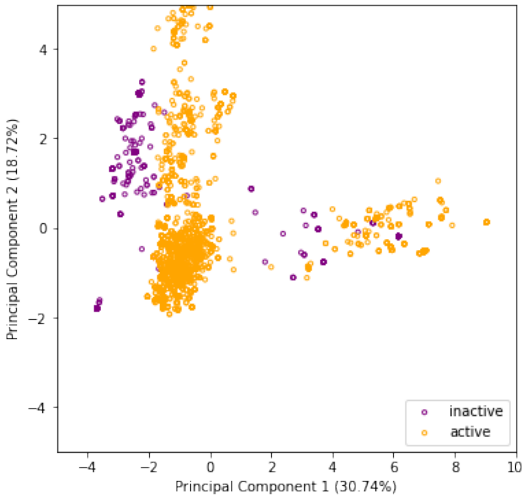


Figure 14: Visualization of inactive and active molecules on a similarity matrix (ChEMBL251)

7. REFERENCES

[1] Manuel de Lera Ruiz, Yeon-Hee Lim, and Junying Zheng. Adenosine a2a receptor as a drug discovery target. *Journal of medicinal chemistry*, 57(9):3623–3650, 2014.

[2] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[4] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 285–294, 2017.

[5] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[6] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[7] MH Segler, T Kogej, C Tyrchan, and MP Waller. Generating focussed molecule libraries for drug discovery with recurrent neural networks, 2017. *arXiv preprint arXiv:1701.01329*.

[8] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*, 2017.

[9] Elozino Egonmwan and Yllias Chali. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, 2019.

[10] Pavel Karpov, Guillaume Godin, and Igor V Tetko. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*, pages 817–830. Springer, 2019.

[11] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.

[12] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[13] Min Shen, Yunde Xiao, Alexander Golbraikh, Vijay K Gombar, and Alexander Tropsha. Development and validation of k-nearest-neighbor qspr models of metabolic stability of drug candidates. *Journal of medicinal chemistry*, 46(14):3013–3020, 2003.

[14] Ryszard Czermiński, Abdelaziz Yasri, and David Hartsough. Use of support vector machine in pattern classification: Application to qsar studies. *Quantitative Structure-Activity Relationships*, 20(3):227–240, 2001.

[15] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[18] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

[19] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 285–294, 2017.

[20] Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

[21] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

[22] A Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J Bellis, Jon Chambers, Mark Davies, Felix A Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, et al. The chembl bioactivity database: an update. *Nucleic acids research*, 42(D1):D1083–D1090, 2014.

[23] Xuhan Liu, Kai Ye, Herman WT Van Vlijmen, Adriaan P IJzerman, and Gerard JP Van Westen. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine a 2a receptor. *Journal of cheminformatics*, 11(1):1–16, 2019.

[24] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[25] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):1–22, 2020.

[26] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[27] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

[28] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[29] Abonia Sojasingarayar. Seq2seq ai chatbot with attention mechanism. *arXiv preprint arXiv:2006.02767*, 2020.

[30] Yu Zhang, William Chan, and Navdeep Jaitly. Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4845–4849. IEEE, 2017.

[31] Yong Zhang, Dan Li, Yuheng Wang, Yang Fang, and Weidong Xiao. Abstract text summarization with a convolutional seq2seq model. *Applied Sciences*, 9(8):1665, 2019.

[32] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[33] Anirudh Goyal, Alex Lamb, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. NIPS'16, page 4608–4616, 2016.

[34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[35] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv preprint arXiv:2106.04554*, 2021.

[36] Ajay Kumar. Chemical similarity methods-a tutorial review. *Chem Educator*, 16:1, 2011.

[37] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[38] David E Patterson, Richard D Cramer, Allan M Ferguson, Robert D Clark, and Laurence E Weinberger. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of medicinal chemistry*, 39(16):3049–3059, 1996.

[39] Xuanyi Li, Yinqiu Xu, Hequan Yao, and Kejiang Lin. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *Journal of cheminformatics*, 12:1–13, 2020.

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[41] bayesian-optimization 1.2.0. `https://pypi.org/project/bayesian-optimization/`.

[42] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[43] Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.