Deep Recurrent Networks for Molecular Drug Design

Maryam Abbasi
https://eden.dei.uc.pt/~maryam
Bernardete Ribeiro
https://eden.dei.uc.pt/~bribeiro/pagina/
Joel P. Arrais
http://eden.dei.uc.pt/~jpa

CISUC,
Department of Computer Engineering,
University of Coimbra

Abstract

In drug discovery, deep learning algorithms have emerged to be an effective method to generate novel chemical structures. They can speed up this process and decrease expenditure. We propose a computational model for molecular *de novo* drug design that is able to produce new drug compounds. This computational model based on the recurrent neural network (RNN) can learn the syntax of molecular representation in terms of Simplified Molecular Input Line Entry Specification (SMILES) strings. The model and its generated SMILES are evaluated using MolVS tool syntactically and biochemically. We analyze the best recurrent network and the parameters. The network that reaches the best result, 98% of valid SMILES, was an RNN containing long short term memory(LSTM) cells.

1 Introduction

Creating novel drugs is a remarkably hard and complex problem. Drug design can be considered as a sampling problem. One of the main challenges in drug design is the cardinality of the search space for novel molecules. It has been estimated that over 10^{60} drug-like molecules could be synthetically accessible [1]. Researchers have to select and analyze molecules from this vast space to find molecules that are active towards a biological target. This process is prohibitively expensive. It is desirable to have computational tools to narrow down the search space.

Searching can be carried out using similarity-based metrics, which provides a quantifiable statistical indicator of closeness between molecules. Heuristics and modern virtual screening techniques can help to narrow the space of possibilities, but the task remains daunting. In contrast, in *de novo* drug design, the practitioners try to directly design novel molecules that are active towards the desired biological target [2].

Machine Learning techniques are lately applied to generate molecule libraries for drug discovery [3]. Here, we present generative deep learning networks based on *Recurrent Neural Networks* (RNNs) for molecular drug design. These models capture the syntax of molecular representation in terms of SMILES string. In particular, we train different RNNs model architecture in order to obtain the maximum possible valid SMILES. Moreover, a comprehensive study on the network's parameters is done in order to optimize the results.

2 Methods

We explored four different types of Recurrent Neural Networks(RNN) for the structure of our model. Simple RNN, Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and finally, Bidirectional LSTM (BLSTM) are applied with different parameters. Figure 1 describes the general schema of the proposed model in order to generate valid drugs.

2.1 Recurrent Neural Networks

RNN models [4] can be used to generate sequences one token at a time, as these models can output a probability distribution over all possible tokens at each time step. Typically, the RNN generator aims to predict the next token based on the all the tokens seen so far. In particular, RNNs process a sequence of data $S = s_1 s_2 \dots s_\ell$ by taking as input each item s_i in the sequence. The RNN passes the input through a series of gates and returns some hidden state h_i and an output vector $\tau = \tau_1 \tau_2 \dots \tau_n$. The hidden state h_i is passed from cell to cell and reflects which information the RNN has seen previously. Additional recurrent connections allow RNNs to learn complex temporal dependencies. The target vector t_i is an array of one-hot encoded vectors, where each vector represents one token. The output

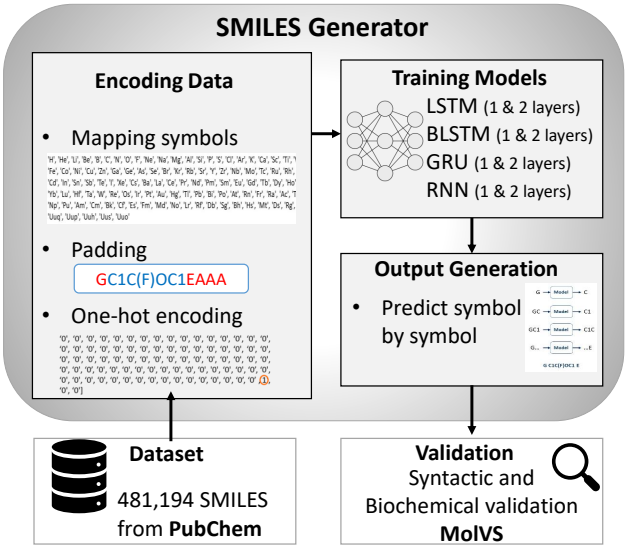

Figure 1: The General schema of the proposed model

vector τ_i is a probability distribution over the possible tokens in one hot encoding. The model aims to maximize the probability assigned to the correct token for every vector in the array. The methodology used to generate SMILES divided into three main parts: encoding data, training model and the output generation.

- **Encoding Data**

  The SMILES generation components consist of three primary steps: mapping symbols, padding and one-hot encoding. Mapping symbols rely on a dictionary of all possible characters included in SMILES. Each SMILES in the dataset is kept in a string format, and each different symbol is tokenized into a char type. The second step is to padding the SMILES. Character 'G'," meaning "go is added at the start of each SMILES, and character 'E' is added to the end, meaning "end". Each SMILES is padded to the length of the longest SMILE string. Padding denoted by the character 'A'. The third step is one-hot encoding. In this step, each character in each SMILES is transformed into a one-hot encoded array. In one-hot encoding, only one bit of a zero vector of the length of the number of tokens in the dataset is set(HOT).

- **Training Models**

  We analysed eight different model structure consist of simple RNN, LSTM, GRU and BLSTM. Each of them is considered with 1 and 2 layers. Figure 2 shows the main structure of LSTM model with 2 layers. In this example, there are 2 LSTM layers followed by a dense layer and a neuron unit with a *softmax* activation function. A dense layer is a linear operation which every input is connected to every output by weight. The other models are implemented in the same way, with different numbers of layer, 1 or 2 and different types of neural network.

- **Output Generation**

  RNN models can be used to generate sequences one token at a time, as these models can output a probability distribution over all possible tokens at each time step. Typically, the RNN aims to predict the next token of a given input. It is worth noting that the input can be one or more tokens in length; if the input has x tokens, then the model predicts the $x+1$st token.

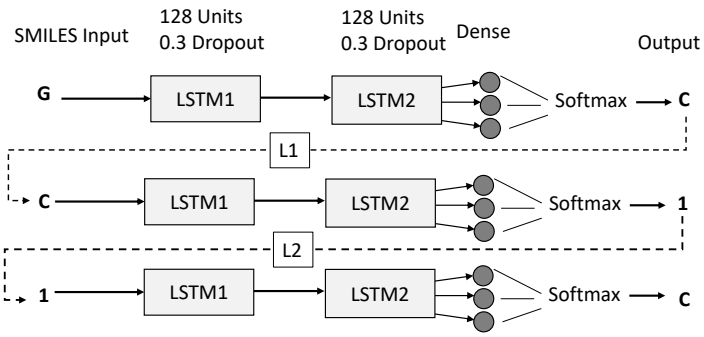

Figure 2: Model of LSTM producing SMILES strings symbol by symbol

2.2 Validation and Datasets

The SMILES generated by the proposed model are syntactically and biochemically validated in RDkit (www.rdkit.org).

For training the RNN models, we compiled a dataset of 481,194 SMILES strings with annotated nanomolar activities as $K_{d/i/B}$,IC/EC_{50} from PubChem (https://pubchem.ncbi.nlm.nih.gov/). The dataset was then pre-processed to remove duplicates, salts and stereochemical information. In addition, pre-processing filtered out nucleic acids and long peptides which lay outside of the chemical space from which we sought to sample.

3 Experimental Results and Discussion

The primary purpose of these tests was to evaluate the set of parameters of the networks that reaches the higher ratio of valid SMILES. Table 1 shows the preliminary results for the 8 proposed models. It contains the percentage of valid and unique generated SMILES. The "valid" SMILES are the ones that passed through the validation process (see Section 2.2), and the "unique" are the ones that are not repeated. After training for twenty-two epochs, the *Model 4* produced the maximum average of 71.03% valid and 71.02% unique SMILES strings. This model consists of two LSTM layers, each with a hidden state vector of size 256, regularized with dropout 0.3. At this stage, we consider the ADAM optimizer and softmax function with the temperature equal to 1.0.

Model	Description	Valid(%)	Unique(%)
Model 1	RNN - 1 Layer	1.55%	1.53%
Model 2	RNN - 2 Layer	2.5%	2.23%
Model 3	LSTM - 1 Layer	60.99%	60.98%
Model 4	LSTM - 2 Layer	**71.03%**	**71.02%**
Model 5	GRU - 1 Layer	60.79%	60.79%
Model 6	GRU - 2 Layer	67.75%	67.74%
Model 7	BLSTM - 1 Layer	70.97%	0.19%
Model 8	BLSTM - 2 Layer	1.54%	0.04%

Table 1: The percentage of valid and unique SMILES generated with 8 different models

We analyzed the effect of the different number of epochs. The numbers 4, 8, 12, 16, 20, 24 and 28 are examined for all eight models. Figure 3 shows the models 3,4,5 and 6 and the percentages of valid and unique molecules generated by considering the different number of epochs. The 1st and 2nd models reached 3.24% and 2.32% of valid SMILES in epoch eight.

We tested the batch size of 64, 128, 256 and 512 for the models using LSTM and GRU. It is observed that the best result obtained by using the model 4, and the 2nd best was model 6 with the size 128. Thus these two models were selected for the rest of the experiment. The next parameter to test during this study was the optimizer. Figure 4 shows the results obtained for different optimizer for models 4 and 6. The two optimizers that presented similar results, 78.14% and 78.69% of valid SMILES were Adam and RMSprop, respectively.

Softmax function normalizes, at each iteration, the candidates by establishing that the network's outputs should all be between one and zero. Temperature is a variable of this function used to control the randomness of predictions. We consider the values of temperature as 0.3, 0.5, 0.8, 1 and 1.5. Using Model 6, at $T = 0.5$ 95.46% of the SMILES were valid. Based on the results of our experiment, we exclusively relied on LSTM Model with 2 layers with Adam optimizer, and dropout equals to 0.3 for

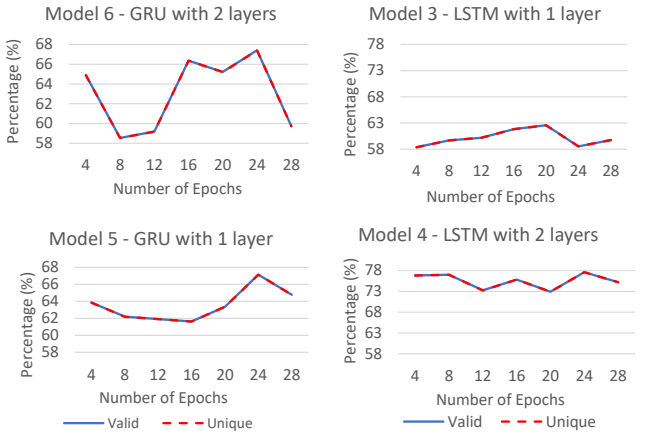

Figure 3: Traning epoches and percentage of valid and unique molecules for Models 3, 4,5 and 6

Figure 4: The percentage of valid and unique SMILES for Models 4 and 6 with different optimizer

the production runs. After analyzing all the tests, it is possible to conclude that the best recurrent network to achieve the main goal is an LSTM with two layers, using Adam optimizer. The best percentage of valid SMILES obtained was 98.04%, with 97.5% $\pm$ 0.31% of valid SMILES. The parameter that influenced more the results is softmax temperature.

4 Conclusions

Deep learning adoption on biomedicine has been slow, and this work intends to contradict this resistance by showing its potential in the field of drug discovery. Furthermore, big data investments in the pharmaceutical industry will reach 4.7 Billion in 2018, which reinforces the need for study and development of novel and better approaches for this field. Everything indicates that the future of computer-aided drug discovery will be promising. The strategy applied in this study gave rise to 97.5% of valid SMILES, on average. The obtained results prove that this deep model can be further used to generate new molecules using the SMILES format. The choice of the dataset and the validation of the model are fundamental for the triumph of the model. It is necessary to continue exploring these techniques possibilities and how they could be applied to drug discovery.

Acknowledgments

This research has been funded by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266)

References

[1] Rodrigues, Tiago, et al. "Counting on natural products for drug design." Nature chemistry 8.6 (2016): 531.

[2] Schneider, Gisbert, and Uli Fechner. "Computer-based *de novo* design of drug-like molecules." Nature Reviews Drug Discovery 4.8 (2005): 649.

[3] Segler, Marwin HS, et al. "Generating focused molecule libraries for drug discovery with recurrent neural networks." ACS central science 4.1 (2017): 120-131.

[4] Mikolov, Tomas, et al. "Recurrent neural network based language model." Eleventh annual conference of the international speech communication association. 2010.

*/