CISUC
CENTRO DE INFORMÁTICA E SISTEMAS DA UNIVERSIDADE DE COIMBRA

# Deep Reinforcement Learning for Optimized Drug Design

**Tiago Pereira** · top@dei.uc.pt
**Maryam Abbasi** · maryam@dei.uc.pt
**Bernardete Ribeiro** · bribeiro@dei.uc.pt
**Joel P. Arrais** · jpa@dei.uc.pt

**CISUC, Department of Informatics Engineering**
**University of Coimbra**
**D4 - Deep Drug Discovery and Deployment Project**
https://www.cisuc.uc.pt/projects/show/259

## Vision

Deep learning (DL) has been an increasingly explored tool in the development of new drugs. These methods efficiently handle huge amounts of data and, therefore, DL models can learn the syntax of the SMILES notation and generate novel SMILES representing valid compounds. Our purpose is to explore deep learning methods to generate valid molecules, with desired physicochemical and biological characteristics, to be used as viable drugs. To do this, we use two interdependent neural networks, one that acts as a generator of new compounds and the other acting as its evaluator.

Hence, these two models need to learn hidden rules of forming sequences of letters that correspond to legitimate SMILES strings. Therefore, both models use Recurrent Neural Networks (RNN). For the Generator, we use a Long Short-Term Memory (LSTM) architecture with two layers that are trained to generate valid molecules. Afterward, a Quantitative Structure-Activity Relationship model (QSAR) is explored to predict the properties we want to optimize. Typically, the approach is to relate a set of descriptor variables to the response variable. In our case, the only descriptor is the SMILES string and the response is the desired property. Subsequently, a Deep Reinforcement Learning framework is employed, through a policy gradient method, to make the model produce fine-tuned molecules towards the desired biological purpose.

## Objectives

- To inter-connect the Generator and the Predictor models, by exploring a reinforcement learning strategy to optimize the generation of candidate drugs.
- To use a QSAR model to predict the desired property of the new generated SMILES and then assign a numerical reward that will be used to update the Generator according to a policy gradient algorithm.
- To optimize the compound's aqueous solubility, by searching for a specific range of the logarithm of the partition coefficient (logP) of each molecule.
- To minimize the half-maximal inhibitory concentration (pIC50) for a cytosolic tyrosine kinase (JAK2). JAK2 is required for cell proliferation, survival, and myeloid development. However, mutations in this kinase are related to breast cancers, B-cell leukemias, and lymphomas. The goal is to minimize the pIC50 of the generated molecules which means to minimize the necessary concentration to inhibit JAK2
- To analyze different deep learning architectures and optimize the hyperparameters of both models.

## Datasets

For training the Generator model, we compiled a dataset of 481,194 SMILES strings from PubChem (https://pubchem.ncbi.nlm.nih.gov/).

The octanol/water partition coefficient was extracted of PHYSPROP database (https://www.srcinc.com). Experimental IC50 data tested against JAK2 was extracted from ChEMBL, PubChem and Eidogen-Sertanty KKB.

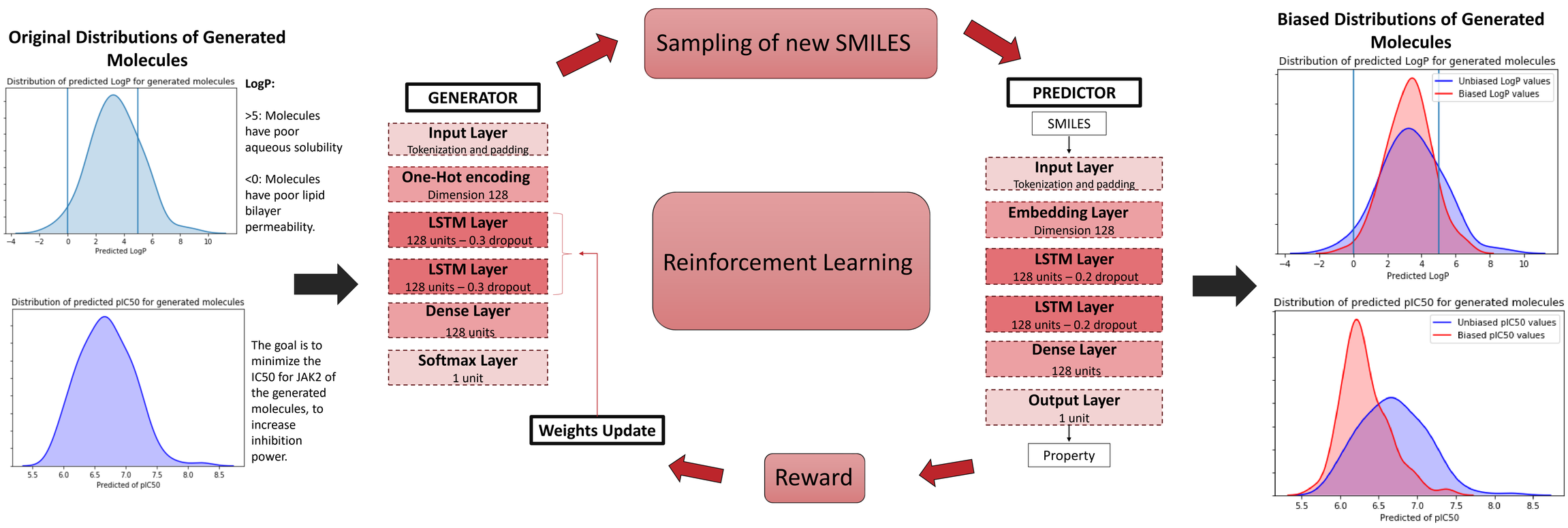


Fig 1: General pipeline of training procedure using Reinforcement Learning

## Generator

As described in figure 2, the two LSTM layers are followed by a dense layer and a neuron unit with a SoftMax activation function. A dense layer is a linear operation in which every input is connected to every output by weight.

To feed this architecture, it's necessary to encode data in the right format. We perform the tokenization of each SMILES in the dataset is tokenized into a char type. Each SMILES starts with a 'G' and ends with an 'E'. The padding is performed in the molecules that are smaller than the longest SMILE string so that all the input data has the same size. Then, 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).

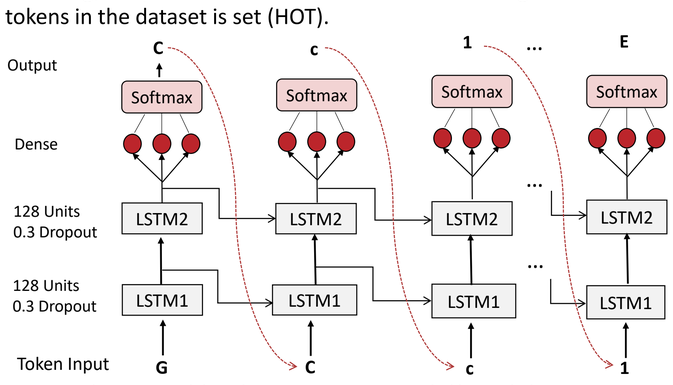


Figure 2: The LSTM model producing SMILES strings symbol by symbol

The SMILES generated by the proposed model are syntactically and biochemically validated in RDkit (www.rdkit.org), showing that 70% of the generated SMILES were valid.

## Predictor

It's a QSAR model that predicts properties of chemical compounds, such as the partition coefficient and affinity for a specific target. The particularity of this approach is that it doesn't need any descriptors besides the SMILES strings. Hence, we perform the tokenization and padding of each SMILE before doing the 5-fold Cross-Validation and data splitting for training, testing and validating.

For the encoding of SMILES, we use an embedding layer transforming the SMILES molecules into a vector of 256 continuous numbers. The model was tested with LSTM layers and Gated Recurrent Units (GRU) with different parameters. After the recurrent layers, there is a two dense layer with 128 units and another with 1 unit, which is the output layer.

The model's performance was evaluated with the metrics described in figure 3 and it shows that we achieve the best result with 128 units LSTM architecture.

| Metrics | Mean square error | Root mean square error | R-Squared error (%) |
|---|---|---|---|
| LSTM / 128 units | $8.889 \times 10^{-3}$ | $6.828 \times 10^{-2}$ | 90.46 |
| GRU / 128 units | $9.730 \times 10^{-3}$ | $7.048 \times 10^{-2}$ | 87.22 |
| LSTM / 256 units | $4.596 \times 10^{-2}$ | 0.153 | 46.27 |
| GRU / 256 units | $4.455 \times 10^{-2}$ | 0.154 | 52.36 |

Fig 3: Summary table with its performance for the logP prediction.

## Reinforcement Learning

We explore an RL approach to bias the Generator to produce molecules towards the desired target properties. The reward function's design depends on if the goal is to minimize, maximize or optimize a specific range of the generated molecule's property. This function accepts the property computed by the Predictor and, for each SMILES string, outputs a higher reward as better the property fits our goal. The RL framework increases the probability of molecules with higher rewards and avoids generating those molecules with lower rewards. To do that, it is necessary to maximize the objective function.

$$J(S|\Theta) = -\frac{1}{n}\sum_{i=1}^{|S|}\sum_{j=1}^{length(S_i)} R_i \cdot \gamma^i \cdot log[p(S_i|S_0 \dots S_{i-1}\theta)]$$

Furthermore, we update the weights of the Generator model with the following rule to maximize the expected reward, i.e., applying the policy gradient algorithm:

$$\theta = \theta + \alpha \frac{\partial}{\partial \theta} J(\Theta)$$

We tested different parameters such as:

- Discount factor ($\gamma$) - to give more importance to the recent rewards;
- Learning rate ($\alpha$);
- Several types of reward functions to optimize the model;

| Unbiased | | | | Biased | | | |
|---|---|---|---|---|---|---|---|
| Log P | | JAK2 | | Log P | | JAK2 | |
| % inside drug-like region | % Valid | $pIC_{50}$ median value | % Valid | % inside drug-like | % Valid | $pIC_{50}$ mean value | % Valid |
| 77.78 | 70.00 | 6.68 | 70.00 | 89.58 | 72.50 | 6.25 | 71.00 |

Fig 4: Generator's results for both properties, before and after applying RL, producing new 200 SMILES.

## References

[1] Liu, X., Ye, K., van Vlijmen, H. W. T., IJzerman, A. P., & van Westen, G. J. P. (2019). An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor. *Journal of Cheminformatics*. https://doi.org/10.1186/s13321-019-0355-6

[2] Popova, Mariya, Olexandr Isayev, and Alexander Tropsha. "Deep reinforcement learning for de novo drug design." *Science advances* 4, no. 7 (2018)

[3] Chakravarti, S. K., & Alla, S. R. M. (2019). Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2019.00017

[4] Sutton, R. S., & Barto, A. G. (2017). Reinforcement learning: an introduction 2018 complete draft. In *UCL,Computer Science Department, Reinforcement Learning Lectures*. https://doi.org/10.1109/TNN.1998.712192

[5] Kesarwani, M., Huber, E., Kincaid, Z., Evelyn, C. R., Biesiada, J., Rance, M., … Azam, M. (2015). Targeting substrate-site in Jak2 kinase prevents emergence of genetic resistance. *Scientific Reports*. https://doi.org/10.1038/srep14538

UNIVERSIDADE D COIMBRA
FCT Fundação para a Ciência e a Tecnologia
Cofinanciado por: COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA