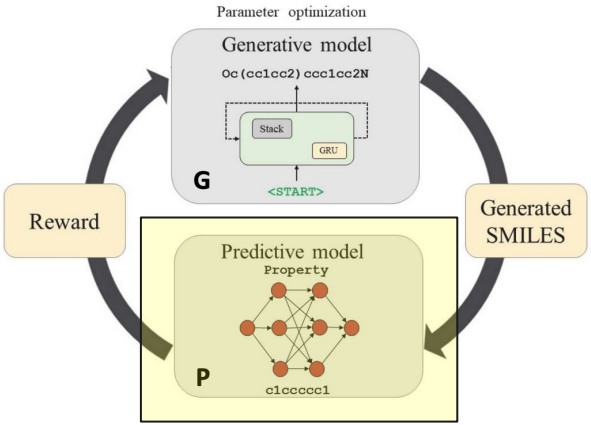
Exploration and Comparison of Modern AI Algorithms to Predict Drug Efficacy

Paper ID: 491
By Anish Kasi, Manish Shetty M, Roshan Neil, Vidhya Murali, Dr. Prashanth Athri, Dr. Gowri Srinivasa

Background and Motivation

- **De-Novo Drug Design**
- **ReLeaSE (Reinforcement Learning for Structural Evolution)**, a recent novel methodology proposes the generation of chemical compounds with desired physical, chemical, and/or biological properties, using deep reinforcement learning(RL).



Reward influences the generation of favorable molecules under the ReLeaSE architecture.

The reward is generally a function of the target property eg: pIC50

Background and Motivation

- The ReLeaSE framework and most recent work in de-novo drug design have made use of a string representation of molecules called **SMILES : Simplified molecular-input line-entry system** and it's one-hot encoded vector embeddings.

- **But we make the following observations/proposal that are novel**
  a. Tap into Hierarchical substructures that are built into SMILES representations
  b. Propose to use more complex embedding techniques, when compared to just one-hot encodings
  c. Propose the use of simpler machine learning models like Random Forest, when compared to LSTMs

Problem Statement

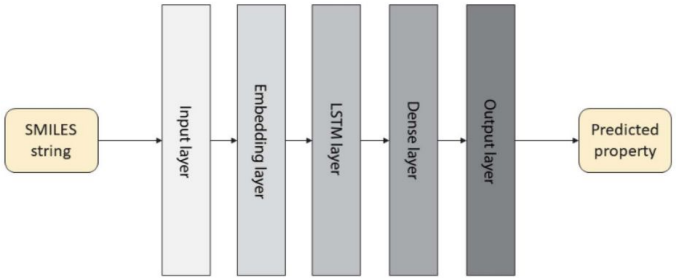
- Exploring modern AI algorithms to predict drug efficacies: Comparing abilities of various **machine learning variants and input embeddings** to **predict inhibitory concentration (pIC50)** of organic molecules against **JAK2 protein target.**

- Our work involves improving the existing predictor component in the **ReLeaSE framework** by comparing various models such as Random Forest Regressor and Convolutional Neural Network for prediction.

Baselines

Our baseline predictor, implemented as in ReLeaSE

a. Uses an LSTM based Recurrent neural network that transformed SMILES tokens into vectors.
b. We implemented this model and observed an **R2 score of 0.56** on the 2000 data points.



Data Augmentation

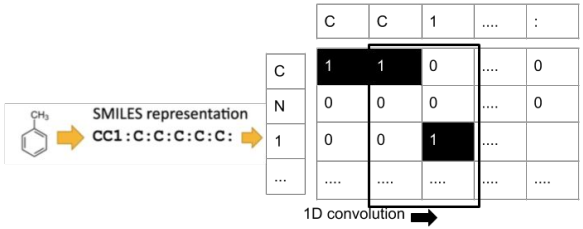
- One molecule can have multiple SMILES strings, which is a reason that canonical SMILES have been defined, which ensures a one to one correspondence between SMILES string and molecule.
- Here the fact that multiple SMILES represent the same molecule is explored as a technique for data augmentation of a molecular dataset.
- We extend this feature to improve input data for our models, by enumerating the 2000 data-points in the JAK2 dataset by 10x, creating a 20000 data-point set.

Research Methodology

1. *Convolutional Neural Network (with auto learnt encoded SMILE strings)*

    a. One hot encoded SMILE strings acting like a binary image.

    b. This model gave us a **R2 score of 0.66 on the 2000** point dataset and a **R2 score of 0.81** on the augmented 20000 point dataset.



Research Methodology

2. *Random Forest (with OpenBabel fingerprint)*
    a. FP2, a path based fingerprint that indexes small molecules based on linear fragments of varying sizes(upto 7 atoms). Each pattern gets bit

    b. This model gave us a **R2 score of 0.71** on the 2000 point dataset and a **R2 score of 0.97** on the augmented 20000 point dataset.

The molecule OC=CN would generate:

0-bond paths:	C	O	N
1-bond paths:	OC	C=C	CN
2-bond paths:	OC=C	C=CN	
3-bond paths:	OC=CN		

Research Methodology

3. Random Forest (with PPMI embedding)

1. We make use of skip-grams for calculation of PMI since the locality of influence for a character(atom) spans across a larger range. In our implementation, we have chosen a window of size +-2. The PMI values calculated are with reference to characters (atoms). In order to encapsulate the mutual information of a SMILE string, we sum up individual PMI values for a character(atom)

2. This model gave us an **R2 score of 0.41 on the 2000 point dataset** and a **R2 score of 0.74 on the augmented 20000** point dataset.

Results of Research Study

Embedding	*Model*	*R^2*	*Adjusted R^2*	*MSE*
OneHot SMILES	LSTM	0.56	0.50	0.38
PPMI SMILES	Random Forest	0.41	0.40	0.76
CNN Feature Map	Neural Network	0.66	0.58	0.36
Open Babel Fingerprints	**Random Forest**	**0.71**	**0.35**	**0.29**

Results for 2000 point dataset with 80:20 split

Embedding	*Model*	*R^2*	*Adjusted R^2*	*MSE*
OneHot SMILES	LSTM	0.68	0.61	0.31
PPMI SMILES	Random Forest	0.74	0.72	0.36
CNN Feature Map	Neural Network	0.81	0.79	0.25
Open Babel Fingerprints	**Random Forest**	**0.97**	**0.96**	**0.18**

Results for 20000 point dataset with 80:20 split