Using Artificial Intelligence To Generate Naturally Diverse Molecules For Drug Discovery

Tshepo Kenneth Moagi

Explore Data Science Academy

INTRODUCTION

Drug discovery has followed three notable periods, in the nineteenth century it was all based on the serendipity of the chemists. Second period in the early twentieth century when new drug structures were found, coupled with new techniques such as molecular modeling, combinatorial chemistry. This period also was revolutionized by the emergence of recombinant DNA technology which made it easier to develop drug target candidates. Now in recent years machine learning has made a significant contribution to drug discovery, as the field progresses it is becoming possible to create new molecules with desired chemical properties.

OBJECTIVES

- To train a model to generate naturally diverse molecules.
- Generate models with desired specific properties to aid drug discovery.

Models to train from DeepChem

- One-shot Learning [Limited data]
- Complete Multitask Deep Neural Network

Other Model

- ORGANIC (Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry)

METHODS

Packages used:

- ORGANIC -https://github.com/aspuru-guzik-group/ORGANIC
- DeepChem - https://github.com/deepchem/deepchem

Data Source:

- Pubchem

Metric	Process of Measuring
drugs_lipinski	Model assigns 0.25 for every rule of the RO5 observed.
Drug Creativity	Computes the Tanimoto distance of a SMILE.
batch_novelty	Assigns 1.0 if the molecule is not in the training set, otherwise 0.
batch_diversity	Compares the Tanimoto distance of a given molecule with a random sample
batch_SA	Checks synthesizability of a given molecule.

MODELS

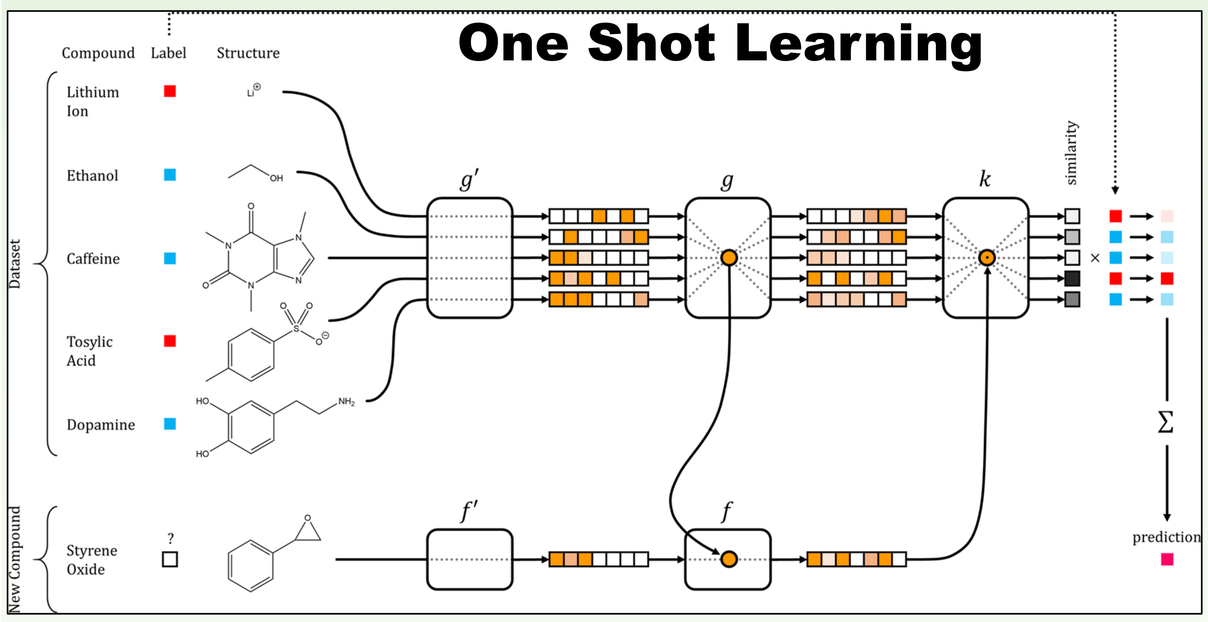

Figure 1: One-shot Learning [source: Altae-Tran et al. 2017]

Siamese one-shot learning and Matching-networks provide ways to learn meaningful metrics from embedded versions of data points and queries. Let f' and g' be graph convolutions. Then matching networks define context aware embeddings.

$$g(x \mid S) = \mathrm{BiLSTM}([g'(x_1)| \cdots |g'(x_m)])$$

$$f(x \mid S) = \mathrm{attLSTM}(f'(x), \{g(x_i \mid S)\})$$

BiLSTM used in Matching networks to define context embedding g is an order-dependent primitive.

attLSTM for f is order independent (better), but since query embedding f depends on g, we can't directly replace BiLSTM for g.(Altae-Tran et al. 2017)

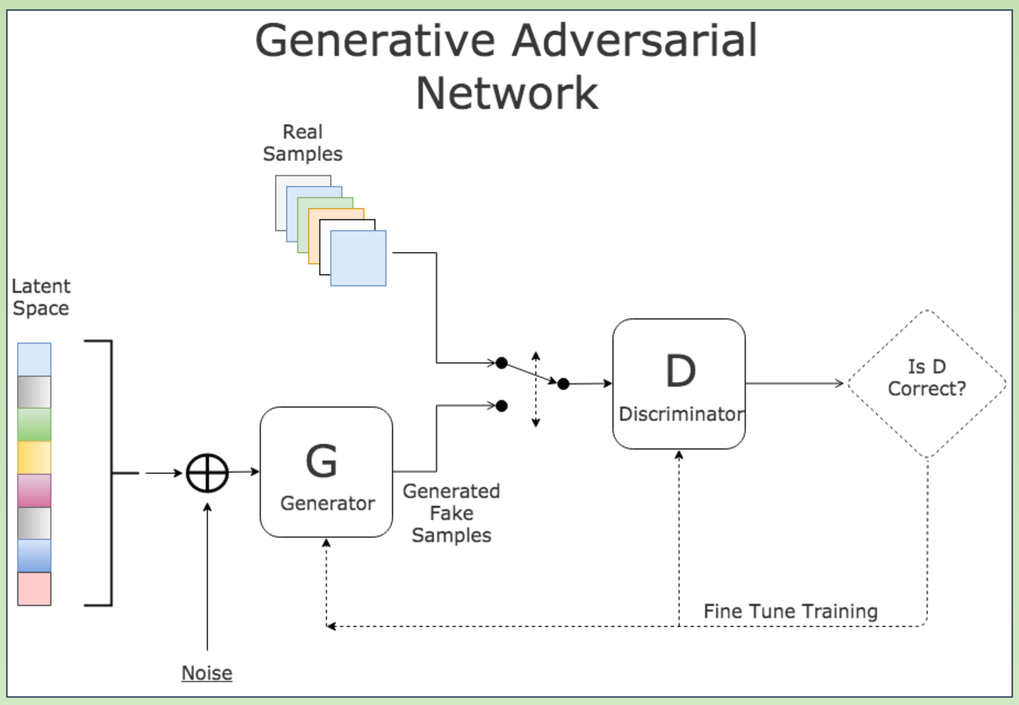

Figure 2: Generative Adversarial Network [source: https://i.stack.imgur.com/UnKny.png]

GANs introduced by I. Goodfellow in 2014 are very powerful in generating realistic outputs which can not be distinguished from a real data, they can also be used to generate new molecule SMILES.

GENERATOR PROCESS

➔ Recurrent process in the LSTM
➔ Processes the batch
➔ Performs Unsupervised training

GENERATOR PROCESS

➔ Uses CNN for text classification (SMILES)
➔ Uses an embedding layer
➔ Followed by a convolutional layer
➔ Then a max-pooling and softmax layer

PRELIMINARY RESULTS [ORGANIC]

Summary of the epoch [λ=0.5, Generator epoch = 1]

Total samples : 31980
Unique : 10271 (32.12%)
Unverified : 6783 (21.21%)
Verified : 25197 (78.79%)

Real Examples:

✓ O=Cc1ccon1 - ***($C_4H_3NO_2$)***
✓ [NH]C1OC=NO1 - ***($C_2H_4N_2O_2$)***
✓ On1cnnn1 - ***(CH_2N_4O)***
✓ N#Cc1nccc(F)n1 - ***($C_5H_2FN_3$)***
✓ C#CCC#N - ***(C_4H_3N)***
✓ COC - ***(C_2H_6O)***
✓ C#CC(C=O)C=O - ***($C_5H_4O_2$)***
✓ c1cnn[nH]1 - ***($C_2H_3N_3$)***
✓ Fc1cc(F)cnc1F - ***($C_5H_2F_3N$)***
✓ [NH][C]1OCOC1=O - ***($C_3H_5NO_3$)***

Fake Generated Examples:

×NH]C1NC=N1CO1
×N#CCC#CC(=O)N#N
×N#CC1(CO)CC#N
×O=Cn1nn[nH]n1
×O=CC1OC2CC=O
×Nc1noc(O)n1C
×O=C1CCO2
×Nc1nc2ocno1
×O=c1ncc[nH]c1
×N#CCN=c1nonno1

CONCLUSIONS

- Though preliminary results show diversity amongst fake generated samples using Pubchem data, this research requires more time.
- Second phase will be generate molecules with desired chemical properties

REFERENCES

1. [Altae-Tran et al. 2017] Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; and Pande, V. 2017. Low data drug discovery with one-shot learning. ACS central science 3(4):283–293.

2. Mostapha Benhenda, V. 2017. ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity. arXiv preprint arXiv:1708.08227v3

ACKNOWLEDGEMENTS

Computations were performed with 1 GPU on Amazon Web Services