Utilization of Generative Neural Networks to Develop Novel Molecules to Target Leishmaniasis

M/CS1001

Background

Almost **a billion people are affected by neglected tropical diseases**, leading to thousands of deaths per year. Development of medicines to address these tropical diseases has been limited since drug research and development is expensive, and the target population cannot afford new drugs.

One of the most prevalent and deadly of these tropical diseases is **Leishmaniasis affecting 1.6 million people** with 300,000 deaths per year; this project will focus on finding a treatment for this disease.

Artificial intelligence has been shown to be a promising for new medicine discovery. Artificial intelligence has been successful in categorizing complex spaces and generating novel data thus has the potential to be a cost-effective approach to generate novel molecules.

Such an **AI guided new molecule drug discovery approach could be applied to find treatments** for other neglected diseases and help **mitigate this devastating cycle of market driven drug development.**

Engineering Goal

My engineering goal was to create an **open-source neural network to generate *de novo* molecules** that can become treatments for leishmaniasis.

One approach to achieving this goal is to utilize existing information relating to a specific target that is important in killing *leishmania donovani*. The reason for focusing on a single target is to increase the likelihood of success and allow the network to learn from *a priori* data. Through an iterative process, we looked at different neural network model structures to see which models can create molecules that fit the constraints.

By making an open-source neural network developed for leishmaniasis available for use in other neglected diseases, **the drug discovery cost typically used by pharma could be substantially reduced** in this setting. This also allows for **community contribution and widespread use**, unlike the current models for drug development.

Procedure

1. Search for molecular targets that can be used to kill the leishmania parasite.
2. Target selection is based on the effectiveness and quality of data available.
3. Determine what neural network architecture to use, such as a Variational Auto-Encoder or Generative Adversarial Neural Networks.
4. Search for past data (libraries of molecules and their effect on target) in open source databases to use, specifically related to the target selected.
   - Gather data from broad to more specialized to make sure the neural network doesn't get confounded.
5. Decide on reward functions and ways to check for loss.
   - This includes looking into what properties should be optimized: IC50's, novelty, druggability etc.
6. Run through the neural network and check the novel generated molecules against prior data categorizations.
7. Go back to step 1 if results are unsatisfactory.
8. Choose top de novo molecules to analyze structures, as well as chemical synthetic feasibility.
9. Open source the developed neural network and curated datasets.

Target: Methionyl-tRNA synthetase, putative

Selecting the Target

Targets can be generally lethal to the organism or can be to a specific pathway or protein that is essential to the organism. The criteria to choose a target include - IC50 data on a diverse set of compounds and the lethality of the target. At first, I looked at the general *Leishmania donovani* organismal target, which affected a stage in the parasite's lifecycle. This target had the majority of data which is a key to training a successful neural network. However, I steered away from this target due to its nonspecificity.

The target chosen is a specific protein target, Methionyl-tRNA synthetase, having 70,331 molecules' data, all being towards that specific feasible target. Since Methionyl-tRNA synthetase is a common target with trypanosomatid parasites (parasites having singular flagela), it has a rich prior dataset, with many different properties. **From there, I collected multiple datasets of the target from ChemBL** for the neural network.

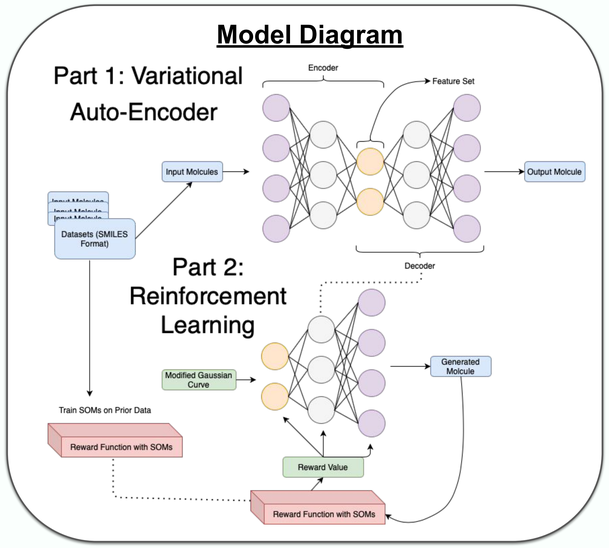
Data

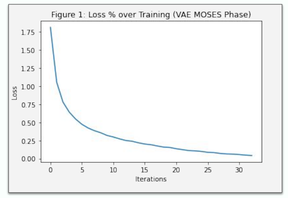
Architecture

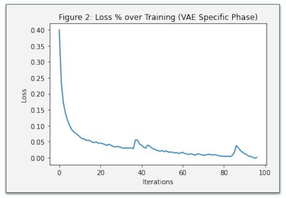
After reviewing multiple models for drug discovery with high probability of success, the model chosen was from Alex Zhavoronkov's research article on using neural networks to generate an inhibitor of DDR1 kinase. This research group has **utilized this model to successfully generate and synthesize novel molecules** for several targets in cancer and infectious diseases, using generous amounts of unlabeled data.

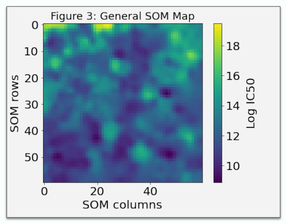
The model, named GENTRL (Generative Tensorial Reinforcement Learning), **utilizes a multi pronged approach that simulates the drug discovery process.** First, it trains a variational auto-encoder (VAE), a common generative network that uses compression of the given data (molecules in this case) to discover generalized features of the data set, and uses the features to generate new and unique data. The second step uses reinforcement learning to fine tune the network to certain rewards. These rewards can be generated through a self organizing maps that are especially adept at dealing with unlabeled data, along with other values to optimize synthesizability.

For the first part with the VAE, it **utilizes datasets from very generalized biological molecules, and moves on to specific molecules to their target**. I emulate this by using the same MOSES biological molecule dataset they use, my manually collected general ligase (the parent category of synthetase) dataset, and another manually collected specific dataset particular to Methionyl-tRNA synthetase. The latter two had a mix of molecules with and without IC50's, which is where **GENTRL's focus on semi-unsupervised learning** with sparse data came in handy.









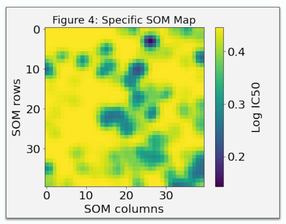
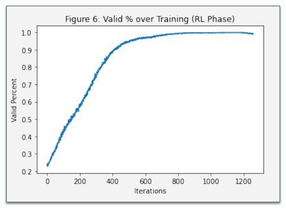


Figure 5: Reward Function

$$R = logP - 1.2 * SA - numRings + SOM_{specific} + SOM_{general}$$



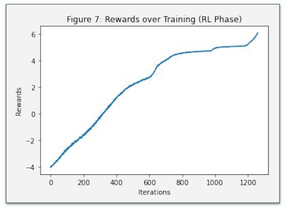
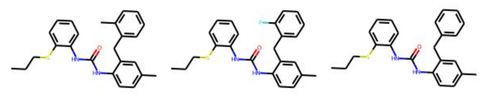


Figure 8: Generated Molecules



5.772022324190315 5.72381501347731 5.710093922618852

Training

The first step was passing the MOSES "general" molecule dataset through the neural network. During this time, the loss, which was generated through finding the quality of the compression, decreased at an exponential pace over the time trained ending at 0.01 (Figure 1). Then I trained it through the general and specific ligase datasets (Figure 2). Overall, **the loss for this phase was 0.008.**

For the reinforcement learning phase, we **needed to create reward functions** to fine tune the generation of molecules, by using the general and specific datasets, and finding correlations between the molecules and IC50s. We chose self organizing maps (SOMs) for this, because they **map high dimensions (input molecules), to low dimensional values (such as IC50s)** in a semi-supervised way, because a majority of the data did not contain IC50 values. SOMs are also efficient in the way that they determine data through competition, and not a form of error correction like regular neural networks.

After training the general ligase SOM, we ended up with an **R squared score of 55%, which showed moderate causation**, which was expected for such a diverse dataset (Figure 3). The specific SOM had an **R squared score of 67%**, which was significantly higher, and showed significant correlation (Figure 4) - yellow is hot and blue is cold.

For the other parts of the reward function, we **took into account synthetic accessibility, log P, and number of rings.** This was to check that the molecule would be able to be produced en masse, and be able to be absorbed by cells. If the molecule had too many rings, or wasn't synthetically accessible (calculated through the synthetic accessibility score in Peter Ertl's paper), this implied that it would be impossible to use in the long term. It was also given a reward of -5 if the molecule could not be parsed (ie. not a valid sequence). The full reward function can be seen in Figure 5

The **results for training the model in the second phase were promising.** The percentage of valid formatted molecules increased exponentially (Figure 6) as well as the rewards (Figure 7). During the first few iterations, reward functions were not calibrated properly, and molecules were not coming out as feasible (mostly carbon strains, which were the most accessible to make, but not effective). After tuning it, the reward topped off at around 5.8, reaching the peaks of both SOMs. The generated molecules can be seen in Figure 8.

Analysis

The model generated molecule (Figure 8 #1) **scored well in the SOMs** with a 0.42 score or about 93% of the maximum of the specific SOM, which shows that it would be predicted to have a low IC50 against the chosen target. The synthetic accessibility score was 2.8, which as seen in the Ertl study, is a common range for bioactive molecules. There are 3 aromatic rings in most of the molecules, indicating easier synthesis. The logP was higher - 7 - above the average 2-4, which would make it harder to disperse.

I showed the model generated molecules to an expert drug discovery chemist, Dr. Pasit Phiasivongsa, to check if they pass the straight face test. This is commonly used to see if there are any glaring errors in the molecules, which could happen due to applying AI without domain expertise. He said the outputs look like real synthesizable molecules, and were sufficiently complex enough to warrant bioactivity - **the druggability and synthesizability look promising.**

Conclusion

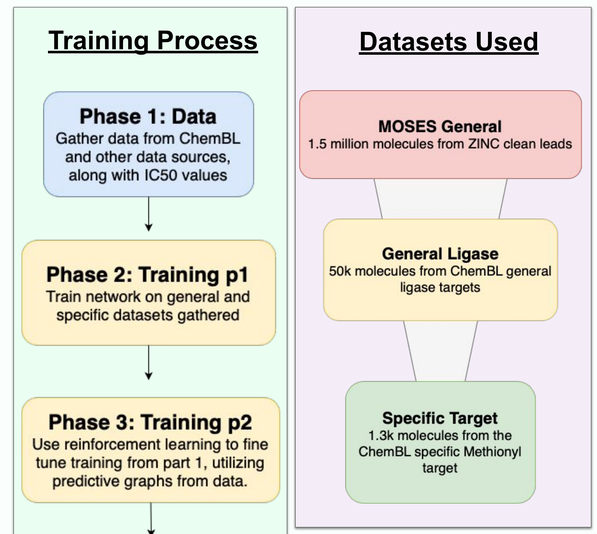
The generated molecules look promising, after passing the straight faced test by having a expert chemist take a look at them. They would be effective with a 93% on the specific SOM reward and be relatively easy to synthesize — scoring a 2.8 — compared to the average 3-5 for bioactive molecules. Taking these into account, we can say that **we achieved our engineering goal.**

Future Work

Since the molecule generation was successful, and looks effective when checking past data, the next step would be to try and partner with laboratories to synthesize the molecule/s and test for activity. I'm currently in contact with Dr. Benjamin Perry from DNDI, where we are discussing a collaboration.

Bibliography

1. Ertl, Peter, and Ansgar Schuffenhauer. "Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions." *Journal of Cheminformatics*, vol. 1, no. 1, June 2009, p. 8. *BioMed Central*, doi:10.1186/1758-2946-1-8.
2. Mendez, David, et al. "ChEMBL: Towards Direct Deposition of Bioassay Data." *Nucleic Acids Research*, vol. 47, no. D1, Jan. 2019, pp. D930–40. *DOI.org (Crossref)*, doi:10.1093/nar/gky1075.
3. *Neglected Tropical Diseases*. 1 Jan. 2001, https://www.gatesfoundation.org/What-We-Do/Global-Health/Neglected-Tropical-Diseases
4. Torrie, Leah S., et al. "Chemical Validation of Methionyl-TRNA Synthetase as a Druggable Target in Leishmania Donovani." *ACS Infectious Diseases*, vol. 3, no. 10, Oct. 2017, pp. 718–27. *PubMed Central*, doi:10.1021/acsinfecdis.7b00047.
5. *Visceral Leishmaniasis | DNDi*. 1 Jan. 2020, https://dndi.org/diseases/visceral-leishmaniasis/.
6. Zhavoronkov, Alex, et al. "Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors." *Nature Biotechnology*, vol. 37, no. 9, Sept. 2019, pp. 1038–40. *www.nature.com*, doi:10.1038/s41587-019-0224-x.



2-minute video presentation