

Variational autoencoders (VAEs) as a generative tool to produce de-novo lead compounds for viral Neuraminidase enzyme

Varun Ullanat (varunullanat.bt17@rvce.edu.in)

Abstract

A variational autoencoder (VAE) is a generational deep learning model based on encoding a particular observation into a latent space and then decoding it while incorporating some random noise with the intuition of being able to generate slightly different forms of the input observation. Here, I propose a VAE that is able to generate de-novo drug compounds by feeding in a known set of drug compounds against a particular biological target. To demonstrate the ability of the proposed VAE as a generative model, known drug molecules belonging to the class of Neuraminidase (NA) inhibitors were taken from the ZINC database and fed into the VAE model as one-hot encoded SMILES strings. Similarly, active NA inhibitors and decoy molecules together were also fed to compare efficiency. The generated molecules were then screened to remove impractical structures. Next, a drug-likeness (QED) score was computed for each candidate molecule and a cutoff of 0.5 was used to extract viable candidates. To ensure that the generated drug compounds were active NA inhibitors, a series of Artificial Neural Networks (ANNs) classifiers based on three different featurization techniques, namely chemical fingerprinting, molecular descriptions and graph convolutions, were developed to identify active NA inhibitors from decoy molecules. The feature candidate data were then fed into the three designed ANNs to obtain the final set of novel, viable and active NA inhibitors. Fifty-six new NA inhibitors were obtained after three runs of the VAE model under different parameterizations. The proposed VAE can hence be used to generate de-novo drug compounds for a wide variety of biological targets.

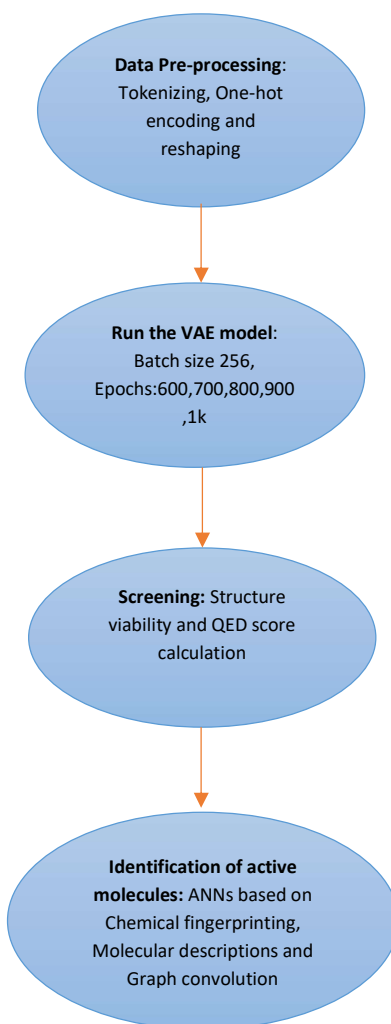
Keywords: Variational Autoencoder, ANNs, ZINC Database, SMILES strings, QED score, Neuraminidase

1. Introduction

An autoencoder is an unsupervised deep learning technique that aims to extract representations of the input data. A typical AE contain three layers: an encoder which compresses the input data, a latent space which contains the representations of the encoded data and the decoder which decompresses the representations to give the output. The objective of an autoencoder is to make the output as close as to the input as possible. Since the past 10 years, autoencoders have been used in many fields, such as dimensionality reduction, anomaly detection, image processing. A type of an autoencoder known as a variational autoencoder (VAE) has been particularly gaining popularity as an efficient tool for new data generation. VAEs consist of a conventional autoencoder with a couple of tweaks that enable it to generate new data rather than simply reconstructing the input space. VAEs employ a loss function which ensures that the latent representations follow a normal distribution with mean zero and variance one. The regularization term usually employed is the Kulback-Leibler (KL) divergence. A reconstruction loss is also added to the loss term to ensure that the data classes are separated in the latent space. Hence, VAEs are able to generate new data which closely resembles the input data but is not entirely identical to it.



Modern drug discovery is a long and expensive process. Generating de-novo drug compounds is usually done manually by humans, with small structural changes applied to pre-existing drug compounds while also ensuring that drug activity and potency is retained. Generally, these methods depend upon the availability of senior chemists and is a limiting step in new lead development. Here, I have proposed a VAE model that is able to streamline the lead development process and rapidly generate de-novo drug compounds. For demonstration of the model, it is fed with one-hot encoded SMILES strings of pre-existing drug compounds which target and inhibit the enzyme Neuraminidase (NA). These drugs, also called NA inhibitors (NAIs) are used as anti-viral drugs to treat diseases such as influenza. Univariate (only active NA inhibitors) and Bivariate (active NA inhibitors and decoy molecules) were both fed into the model as the two experimental conditions. The generated SMILES strings are then screened in three different steps, namely: structure viability, drug-likeness and classification using Artificial Neural Networks (ANN's) for different featurization. The designed process can be extended to generate new drug compounds for various bacterial and viral targets.

2. Experimental Design



3. Results

Table 1. Generated active NA inhibitors for univariate case (First four)

S. No	SMILES	Structure	MW	logP	Charge
1	<chem>CCCCC1CCCCN1[C@@H]1C=C(C(=O)O)C[C@H](N)[C@H]1NC(C)=O</chem>		337.464	1.64640	0
2	<chem>CC(=O)N[C@@H](CC(C)C)[C@@H]1N[C@@H](C(=O)O)C[C@H]1C1=CCC1</chem>		294.395	1.68870	0

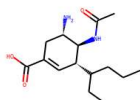
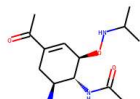
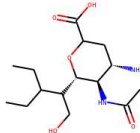
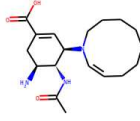
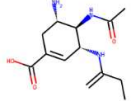

3	<chem>CCCC(CC)[C@@H]1C=C(C(=O)O)C[C@H](N)[C@H]1NC(C)=O</chem>		282.384	1.67560	0
4	<chem>CC(=O)N[C@@H]1[C@@H](N)CC(C(C)=O)=C[C@H]1ONC(C)C</chem>		269.345	0.03570	0

Table 2. Generated active NA inhibitors for the bivariate case (First four)

S.No	SMILES	Structure	MW	logP	Charge
1	<chem>CCC(CC)C(CO)[C@@H]1OC(C(=O)O)C[C@H](N)[C@H]1NC(C)=O</chem>		316.398	0.10520	0
2	<chem>CC(=O)N[C@@H]1[C@@H](N)CC(C(=O)O)=C[C@H]1N1C=CCCCCCCC1</chem>		335.448	1.77160	0
3	<chem>C=C(CC)N[C@@H]1C=C(C(=O)O)C[C@H](N)[C@H]1NC(C)=O</chem>		267.329	0.11500	0
4	<chem>C/C=C/[C@@H]1C[C@H](C(=O)O)N[C@H]1[C@@H](NC(C)=O)[C@@H]1CCC=CO1</chem>		308.378	1.19120	0

4. Discussion

To evaluate the generation process, we can define a formula for the efficiency as:

$$Efficiency = Final\ number\ of\ active\ molecules \div Total\ number\ of\ inputs\ to\ the\ VAE\ model$$

This comes out to be 0.014 for the univariate case and 0.006 for the bivariate case. It was seen that ten molecules were common between the two cases, so the final number of generated active NA inhibitors across both the cases was 56 with an efficiency of 0.007. The efficiency value is still pretty low for each case, as most of the structures generated by the model is impracticable. This is a bottleneck in this approach.

5. Conclusion

The proposed method was able to generate 56 new NA inhibitors using a VAE trained on two different conditions. The entire method is quick and requires only about 10-20 minutes while running on a good GPU. Training the VAE required about 7 minutes for the entire set of epochs mentioned in Section 3. Training and identification using the three ANNs requires even lesser time, and the preprocessing steps are fast as well. The proposed method can be extended to produce more active molecules by either, 1) Training for a different number of epochs 2) Using

different parameterizations of ϵ_{std} . The method ultimately depends on the quality of the training set itself, and can produce better results if there is an ample number of active molecules to begin with. The efficiency of the method is appreciable, but it can be improved by using memory-based components such as LSTMs in the VAE which give importance to the position of each character as well.