ABIGAIL Report

Siddharth Yenamandra

Amyloid-Beta Inhibition Via AI Learning

Abstract:

Amyloid-Beta is a protein that aggregates to form AB plaques. These plaques are one of the biggest causes of Alzheimer's disease. The current framework to finding inhibitors of amyloid-beta is extremely inefficient and inaccurate. Hence, I developed ABIGAIL, a model that uses a 3D latent space to generate novel amyloid-beta inhibitors. Testing ABIGAIL against a classification model I built, resulted in a molecule screening to HITS rate of 61.77%. After chemical analysis (testing for molecule stability, inhibition ability and blood brain barrier permeability) I found the HITS to leads rate to be 30.262%. Compared to the 0.1% HITS to leads rate of current molecule discovery methods, ABIGAIL shows incredible promise for the use of multidimensional latent space models to generate amyloid-beta inhibitors, potentially leading to more clinical trials and available treatments.

Introduction:

Molecule generation is the first step of drug discovery. Rather than using trial and error, current day drug development often uses algorithms to create new molecules. Called de novo methods, these algorithms aren't accurate enough to greatly increase the speed of drug development. The primary problem with de novo algorithms is that they do not utilize patterns of already existing drugs to create molecules. Furthermore, even when provided with prior molecular data, de novo

algorithms fail to recreate similarly functioning drugs as they lack the depth to understand hidden and deep patterns within molecule datasets. In this paper, we will discuss using a graph neural network (ABIGAIL 1.0) to detect patterns within an Alzheimer's drug treatment dataset, identify whether a molecule holds promise for further drug testing, and lastly, create novel molecules.

Technical Background:

Graph neural networks are optimal for this task as they allow a model to receive and analyze molecules with different numbers of atoms. Simply put, we can graph out the molecule by defining nodes as atoms and edges as bonds between the atoms. This is a new approach to detecting and discovering different drugs and hasn't been explored fully as the field of neural networks is quite new.

Nodes are defined as points in space, with each node having its own set of node features. In the context of drug classification and development, these features include atom number, number of valence electrons, electronegativity, ionization energy etc. Each node's/atom's node features are groups into a matrix which is assigned to its respective node. Similarly, edges also have their own features, called edge features: type of bond, is the bond in a ring, bond energy etc. Since edges/bonds represent properties of their respective nodes/atoms, the edge features are concatenated into the two nodes that the edge in question connects. Specifically, the edge features are concatenated into the node feature matrices.

Biochemistry Background:

Starting out, I knew that I wanted to use a GNN to help classify and create Alzheimer's drugs. To do this, I had to understand the causes of Alzheimer's. Alzheimer's is caused primarily by two things: the buildup of amyloid-beta plaques, and the formation of tau protein tangles. I decided to

focus on the inhibition of amyloid-beta plaques for this paper as there is much more data on prevention and inhibition of amyloid-beta plaques. Furthermore, amyloid-beta plaques appear 20+ years before Alzheimer's diagnosis, so inhibition of these plaques at an early stage can aid greatly in the prevention of Alzheimer's.

Design Intent:

The intent of my research project to code a Generative Variational Autoencoder (GVAE) model that can create novel inhibitors of amyloid beta (Aβ).. By using already known data and mechanisms of Aβ formation, my GVAE aims to efficiently select and design small molecules as potential inhibitors of Aβ aggregation. The intent is to accelerate the discovery of potential therapeutic candidates for Alzheimer's disease.

Methods:

Drug Targets:

Before compiling a database of potential inhibitors relating to the formation of amyloid-beta plaques, I created a list of drug targets regarding amyloid-beta plaque aggregation. This list has two tiers: primary targets that have a clear connection to AB development and supplementary targets whose increased presence have a correlation to AB concentration. Furthermore, I developed a list of negative controls (drugs that don't inhibit amyloid-beta), so their respective inhibitors could be used to train ABIGAIL.

Primary Drug Targets:

Looking into the specific pathway in which AB aggregation occurs allowed me to develop a list of primary drug targets. Amyloid-beta plaques are developed in two major steps: N-terminal truncation and cyclization.

N-terminal truncation is responsible for the initial production of amyloid-beta. It occurs via proteolytic processing of amyloid precursor protein. During amyloid precursor protein proteolytic processing, ß-secretase (BACE1) cleaves the N-terminus of amyloid precursor protein, removing the aspartic acid and alanine amino acids from the n-terminus. Amyloid precursor protein proteolytic processing is also aided by the enzyme y-secretase, which also help cleave the N-terminus. Thus, BACE1 and y-secretase are prime drug targets for Alzheimer's.

Cyclization occurs when glutamate in the N-terminal of amyloid beta converts to pyroglutamate, creating pyroglutamate amyloid beta (pGlu-Aβ). During cyclization, the α amino group of glutamate, NH_2, loses a proton, and the R-group of glutamate, CH_2CH_2COOH, loses a hydroxide. The lost proton and hydroxide combine to form H_2O, a side process of cyclization. The remaining glutamate is known as pyroglutamate. Cyclization is catalyzed by the enzyme glutaminyl cyclase (QC), hence making QC another optimal drug target. These drug targets are all included in Figure 1.

Supplementary Drug Targets:

To identify supplementary drug targets, I researched drugs that had a higher concentration in brains with higher concentrations of amyloid-beta plaques. These are also summarized in Figure 1.

Negative Control Drug Targets:

To provide ABIGAIL with negative control data (drugs that don't inhibit amyloid-beta) while still making these negative control drugs similar enough to ensure ABIGAIL is accurate, I used

inhibitors of other drug targets relating to Alzheimer's. These included Tau protein, Acetylcholinesterase (AChE) and Butyrylcholinesterase (BuChE).

Inhibitor Data Collection:

To utilize the list of drug targets, I had to compile a large dataset of inhibitors pertaining to each of these drug targets. I used the Inhibitors of Proteins associated with Alzheimer's Disease Database (IPAD-DB) to generate a dataset of inhibitors associated with each drug target identified in Figure X. I used PubChem to obtain SMILES from these molecules, as SMILE representations of the inhibitors helped when creating the GNN.

With this dataset, I wanted to ensure that there was no bias. Hence, I also ensured that each type of inhibitor was equally represented. For instance, as seen in Figure Y, only 26% of the dataset belonged to the negative control. Hence, I duplicated the negative control twice to ensure that the GNN wouldn't be biased towards classifying drugs as AB inhibitors. Lastly, to simplify the data analysis and decrease model training times, I simplified the dataset to two

Figure 1

No.	Drug Target	Source	Role in Aβ Formation
1.	β-Secretase	Astrocyte	Formation of Aβ
2.	γ-Secretase: Presenilin I	Medial temporal lobe cortex	Formation of Aβ
3.	Dopamine 2 receptor	Central nervous system	Aβ plaques
4.	Amyloid protein precursor (APP)	Hippocampus, olfactory bulb	Formation of Aβ
5.	Non-amyloid-beta component precursor	Neocortex, hippocampus, olfactory bulb, Striatum, thalamus, and cerebellum	Amyloid genesis and plaque formation
6.	Glutaminyl Cyclase	pituitary, hypothalamus, adrenal medulla	Cyclization in amyloid-beta production

columns: SMILES and a binary value of whether the drug was an AB inhibitor.

GNN Architecture:

For ABIGAIL, I decided on using a GVAE, which incorporates both a GNN (Graph Neural Network) and a VAE (Variational Autoencoder). The GNN aspect allows ABIGAIL to process graph-like molecular data while the VAE aspect allows ABIGAIL to map probabilities. ABIGAIL utilizes four different TransformerConv layers. Each of these layers compute query, key, and value projections for each node, calculate attention scores between connected nodes, apply SoftMax to get attention weights, aggregate neighboring node features weighted by attention, and apply a final linear transformation. Furthermore, since ABIGAIL uses four attention heads to analyze different aspects of node relationships. Each of the four TransformerConv layers passes inputs to next layer utilizing batch normalization to stabilize training. The last TransformerConv layer passes inputs to a Set2Set pooling layer, which takes in all node/edge feature information and aggregates it into a fixed-size representation. Hence, regardless of the size of the molecule, the final output will always be the same size.

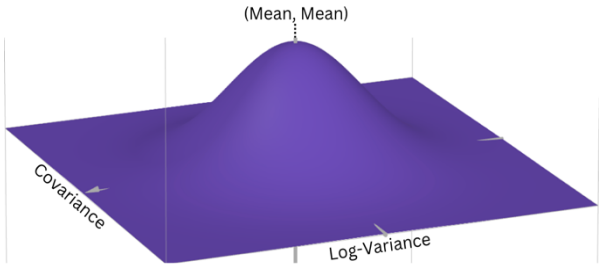
Latent Variables:

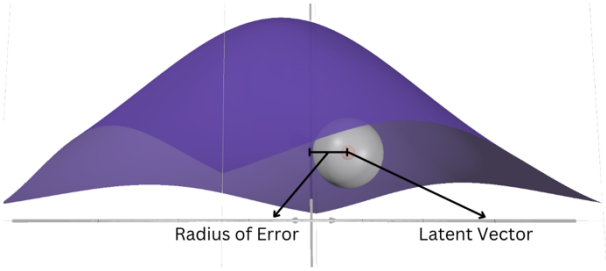
The fixed-size representation, also referred to as pooled graph features, is transformed into two latent variables, μ (mean) and σ^2 (standard deviation). These latent variables create a Gaussian probability distribution where μ determines the average of the distribution while $\log(\sigma^2)$ determines the spread of the distribution. I used log-variance instead of variance to keep numerical stability. Since each molecule has unique pooled graph features, they will have unique Gaussian probability distributions.

Latent Vectors:

Each molecule's mean and variance are applied into the formula $\mathbf{z = \mu + \epsilon \cdot \sigma}$. ABIGAIL uses this formula to determine each molecules' latent vector. Z is the latent vector; μ is the mean latent variable; ϵ is a random noise vector sampled from a standard distribution; σ is standard deviation. The random noise vector is included to ensure diversity in outputs.

The latent vector generated by ABIGAIL represents a point on the Gaussian distribution defined by the latent variables mentioned above. This vector is extremely useful for testing ABIGAIL's ability to generate molecules. I utilized reparameterization and decoding where ABIGAIL would take in the latent vector of a molecule and with that singular point as an input, ABIGAIL would attempt to reconstruct the graph structure, predicting atom and bond types. Furthermore, I attempted to have ABIGAIL take in latent vectors close to the latent vector of the actual molecule to produce novel molecules. Next, by applying a hybrid reconstruction method, ABIGAIL is able to produce molecules, and by checking whether they are inhibitors or not, ABIGAIL can verify that it is creating novel amyloid beta inhibitors.





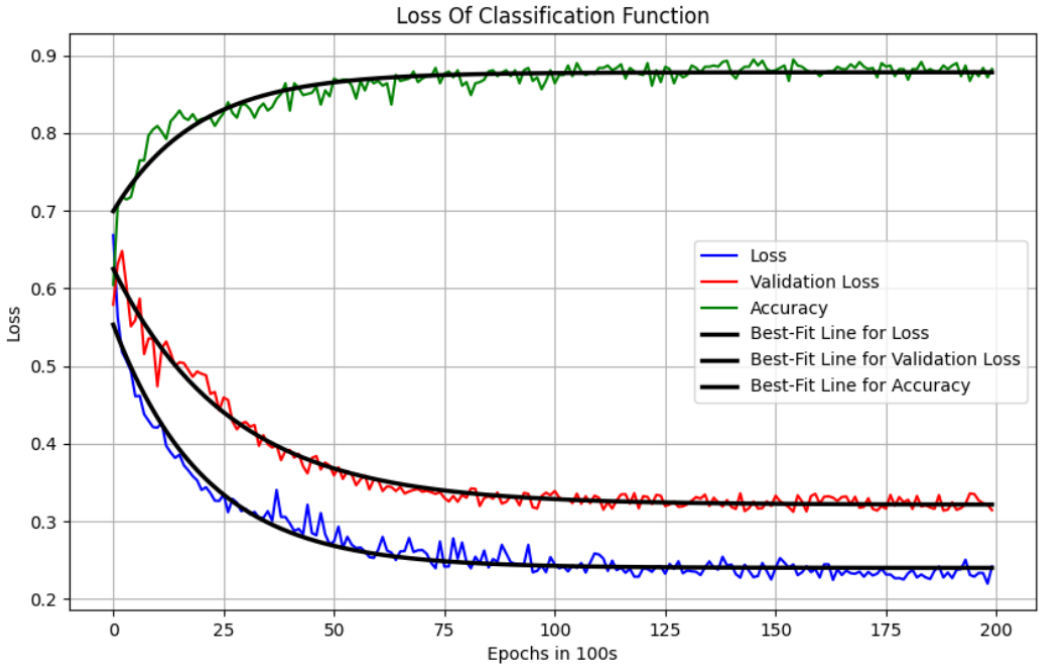
Results:

Performance Criteria:

I measured my model's success on two criteria. Firstly, I measured the classification model accuracy and loss (loss function was binary-entropy). Secondly, I measured the loss of the reconstruction part of ABIGAIL.

Prediction:

Currently, my model can predict the amyloid-beta inhibition of potential inhibitors with an ~80-85% accuracy. Although this is lower than the industry standard of 90% for classification models regarding molecule property prediction, to see the full picture, we must look at the context in which my model operates.



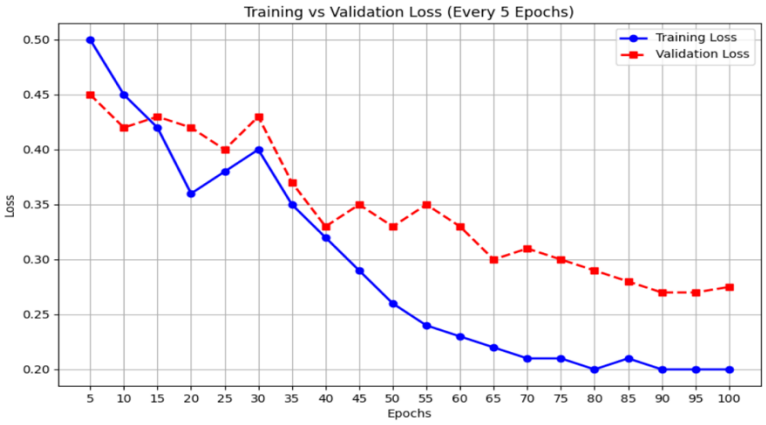
ABIGAIL is built to increase the prediction accuracy of potential amyloid-beta inhibitors. Current algorithms have an accuracy of ~75% in this department, thus showing how an ~80-85% average accuracy is an improvement on current algorithms. These algorithms, and my model, have such low accuracy rates due to the limited data available regarding amyloid-beta inhibition, and the complex nature of classifying inhibitors. Classifying inhibitors is extremely tough as it requires models and algorithms to simultaneously consider several different features of a molecule, thus resulting in lower accuracy rates when compared to other classification tasks such as inhibitor type grouping, or blood-brain-barrier permeability. Furthermore, there is little data regarding amyloid-beta inhibitors. Typical GNN's regarding molecule property prediction are trained with tens of

thousands of molecules. ABIGAIL is trained with only 1268 molecules. Hence, the use of a graph neural network does improve classification of drugs regarding amyloid-beta inhibition.

Generation of Novel Molecules:

Firstly, we measured results of ABIGAIL's loss when reconstructing molecules from their latent vectors. To measure total loss, we used the formula **Total Loss = Node Reconstruction Loss + Edge Reconstruction Loss + KL Divergence Loss.** To measure node reconstruction loss and edge reconstruction loss, we used cross-entropy loss to determine how different the original nodes and edges were from the reconstructed nodes and edges.

KL Divergence Loss is used to measure the difference between the standard distribution made by the original molecule's latent variables and the standard distribution made by the reconstructed molecule's latent variables.



Currently, my reconstruction loss is ~2.0. I hope to further the model and fine tune it to reach a reconstruction loss of ~1.5.

Right now, my model can't generate novel molecules due to shape errors. The new molecules' latent vectors become arrays of node and edge features. However, this array is a different shape and can't fit into my molecule construction method. Hence, I'm planning to implement dynamic size inputs to my molecule construction method.

Discussions/Conclusions:

Looking at the results, we can conclude that the use of GNN models to predict drug amyloid-beta inhibition abilities is greatly beneficial. This is especially clear when considering the application of ABIGAIL in drug development. By running ABIGAIL's predictions before developing a theoretical drug, companies could save invaluable money, time and resources.

Furthermore, if the shape-to-shape error in the molecule construction method is resolved, ABIGAIL will be able to construct and predict the efficacy of novel molecules, thus greatly reducing the time needed to develop potential amyloid-beta inhibitors.

Acknowledgements:

I would like to thank IPAD-DB and PubChem for allowing all their data to be accessible online. Without their data, I wouldn't have been able to conduct my research.

References:

Murphy, M. P., & LeVine, H. (2010). Alzheimer's disease and the amyloid-B peptide. Journal of Alzheimer S Disease, 19(1), 311–323. https://doi.org/10.3233/jad-2010-1221

Paik, S. R., Lee, J., Kim, D., Chang, C., & Kim, Y. (1998). Self-oligomerization of NACP, the precursor protein of the non-amyloid β/A4 protein (Aβ) component of Alzheimer's disease amyloid, observed in the presence of a C-terminal Aβ fragment (residues 25–35). FEBS Letters, 421(1), 73–76. https://doi.org/10.1016/s0014-5793(97)01537-8

Peng, C., Liu, X., Meng, X., Chen, C., Wu, X., Bai, L., Lu, F., & Liu, F. (2024). IPAD-DB: a manually curated database for experimentally verified inhibitors of proteins associated with Alzheimer's disease. Database, 2024. https://doi.org/10.1093/database/baae048

PubChem. (n.d.). PubChem. PubChem. https://pubchem.ncbi.nlm.nih.gov/

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., Van Hoesel, C., Schopmans, H., Sommer, T., & Friederich, P. (2022). Graph neural networks for materials science and chemistry. Communications Materials, 3(1). https://doi.org/10.1038/s43246-022-00315-6

Tamo, W., Imaizumi, T., Tanji, K., Yoshida, H., Mori, F., Yoshimoto, M., Takahashi, H., Fukuda, I., Wakabayashi, K., & Satoh, K. (2002). Expression of α-synuclein, the precursor of non-amyloid β component of Alzheimer's disease amyloid, in human cerebral blood vessels. Neuroscience Letters, 326(1), 5–8. https://doi.org/10.1016/s0304-3940(02)00297-5