Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/33/20240821

Transfer learning-enabled classification of drugs for Alzheimer's Disease

Ruoning Gu

Shanghai Pinghe School, Shanghai, China

guruoning@shphschool.com

Abstract. Alzheimer's disease (AD) is a neurodegenerative disease with a worldwide prevalence of 760.5 per 100,000 inhabitants. Environment factors and genetic predispositions can influence an individual's risk of developing AD. Recent research provided insights into the genetic mechanisms of AD, but a comprehensive database of drug-like molecules that can exacerbate or reduce AD risk remains unavailable. Here, we use machine learning to create a similarity map that reveals new putative drugs structurally similar to existing compounds known to be targeting pathways relevant in AD. We trained an autoencoder on a large drug database of over 14,000 drugs, with features derived from several modalities including molecular fingerprints. We then computed similarity scores based on these reduced dimensions. We show that our model is able to identify new compounds structurally similar to existing drugs linked to AD. We conclude that our model holds the potential to elucidate new compounds based on structural similarity and can be used to identify new drugs that affect critical pathways in AD.

Keywords: Alzheimer's Disease, drugs, risk factors, machine learning.

1. Introduction

Alzheimer's disease (AD) is a form of dementia that slowly destroys memory and thinking skills. The disease now has a worldwide prevalence of 760.5 per 100,000 inhabitants [1]. AD is characterized by neurofibrillary tangles of tau protein and Aβ plaque [2]. Tau protein promotes microtubule assembly and is important for the functions of neurons [3] while Aβ plaques can accumulate extracellularly and affect interneuronal communication. The role of pharmacological agents in either exacerbating or mitigating the risk of AD has emerged as a domain of critical importance.

Gallacher and colleagues (2012) showed that benzodiazepines, widely prescribed for anxiety and insomnia, could contribute to AD pathogenesis through mechanisms like neurotransmitter disruption, impairment of brain plasticity, and interference with memory formation [4]. Anticholinergic drugs, which block the action of the neurotransmitter acetylcholine, are also implicated due to their potential impact on cognitive function. Furthermore, the long-term use of certain antipsychotics has raised concerns, given their complex interactions with brain neurotransmitters. There is thus an urgent need to leverage our current understanding of molecular pathways related to AD to inform drug development and safety.

High-throughput screening of molecular bioactivity is a commonly-used technique to elucidate the connection between drugs and a protein important in AD progression. However, the application of traditional chemistry or high-throughput approaches is time-consuming, costly and also carries a high failure rate. Machine learning is emerging as a powerful, reliable and cost-effective approach to drug development, which can accelerate discovery and decision making for predefined questions with precise data. Machine learning has been used in pharmaceutical development, bioactivity prediction, de novo molecular design, synthesis prediction and biological image analysis, among other applications. It is a popular tool in drug discovery when using a large arsenal of compounds; however, the limitation of broad application is the necessity for a sizable number of labeled data points to ensure model generalizability and avoid overfitting.

Here, we leverage existing knowledge on drug-protein interactions in AD to develop a machine learning model that can identify potentially harmful drugs that may cause AD based on chemical structures of previously discovered drugs related to the disease. A dataset of 14610 common drugs and their chemical structures were downloaded from Kaggle and used as the pool for screening. To find drugs of similar features to known drugs causing AD, we used packages such as “Chem” to extract the molecular features. To make the data processing possible via computer, feature reduction techniques were also used including “Principal Component Analysis” and “autoencoder”. Finally, the relationship between the molecules were expressed using UMAP and the possible molecules’ localization was discovered. Our research aims to screen other possible AD-causing drugs so that the risk factors for AD may be reduced.

2. Methods

2.1. Feature Extraction

The dataset contained the molecular structure of 14610 compounds that are drugs. All the coding was done using Python. RDKit package was downloaded, which was developed as an analyzer of chemistry data [5]. The most used module within the RDKit package was the “Chem” module, with which we converted the original structure of the molecules (“mol”) to another expression of the structure that could be understood by the computer (“smiles”). The various functions in the “Chem” module were also used to extract structural features, such as the number of valence electrons and the number of heteroatoms. The function “Morgan Fingerprint” was also obtained through the module. This function served to convert certain features of the molecular structure to numbers of 0 and 1, similar to the binary code and easily processed by the computer [6]. Mol2vec package was also added as another way to extract structural features through unsupervised machine learning and the features were converted to vectors for further analysis [7].

2.2. Feature Reduction

The technique Principal Component Analysis (PCA) was used to reduce the total of 20 features extracted from the chemical compounds. To reduce the features the multiple characteristics need to be combined into a variable that can express the features as accurately as possible. PCA reduces the features to values on the x and y axis and forms a map of the data as dots on a graph [8]. To evaluate the accuracy of the reduction of features, a test of initial variance was calculated. The reduced features needed to be further processed to show relationships of the most similar compounds to the target drug.

Another feature reduction technique employed was the autoencoder. It is a form of unsupervised artificial neural network that reduces the features into fewer dimensions [9]. It had an advantage of self-evaluation, which meant that when the features were reduced they would be expanded again to automatically test if the expanded features fitted the original ones. We also ran the autoencoder on all the data and features, reducing the features into two dimensions.

2.3. Compounds’ Similarity

The technique of the UMAP was used to show the relationship between chemicals based on their features using a graph. Every dot on the UMAP graph represented a compound. The results obtained from PCA and autoencoder were very different. Without the automatic evaluation process, the UMAP graph from PCA showed a much looser relationship between individual compounds, making the similar compounds unclear. However, the graph obtained from the autoencoder showed clear clusters of dots representing compounds, indicating that the chemicals within the clusters showed great similarity to each other. If the target drug was in the midst of one cluster, then the other compounds within the cluster would also be likely to cause Alzheimer's.

2.4. Drug Localization

To search for the most similar drugs and locate them on the dataset, PCA's data was used in an Eucilidian metric to determine the drugs that have the closest features to the target drug C_3H_6NBr. Then the location of these drugs were identified in the list of the dataset.

3. Results

3.1. Generation of molecular features for model training

First of all, the information of drugs provided in the database needs to be extracted by Python. The features were extracted by importing the "chem" package. As we concentrated on matching the chemicals' structures, features extracted included the number of atoms, number of heavy atoms, and number of atoms frequently seen in organic compounds, such as oxygen (O), nitrogen (N), carbon (C), hydrogen (H), and chlorine (Cl). They are extracted using Morgan Fingerprint, a technique used in coding to convert chemical structures into mathematical representation [6]. The theory is that in chemistry, compounds with similar structures would often have more or less similar chemical properties as in reaction it is the bonds and the atoms that really matter. We extracted 20 features in total. Then we used "mol2vec" to make the list of features analyzable for the computer.

Table 1. Extracted features of possible Alzheimer-causing drugs.

index	smiles	logP	num_of_atoms	num_of_heavy_atoms	num_of_C_atoms	num_of_O_atoms	num_of_N_atoms	num_of_Cl_atoms	num_of_carbon_atoms
0	C[C@H]([C@@H](C)Cl)Cl	2.3	14	6	4	0	0	2	4
1	C(C=CBr)N	0.3	11	5	3	0	1	0	3
2	CCC(CO)Br	1.3	15	6	4	1	0	0	4
3	[13CH3][13CH2][13CH2][13CH2][13CH2][13CH2]O	2.0	21	7	6	1	0	0	6
4	CCCOCCP	0.6	20	7	5	1	0	0	5
5	C(C(F)(F)F)F	1.7	8	6	2	0	0	0	2
6	[2H]C([2H])C(C)(C)Cl	1.8	14	5	4	0	0	1	4
7	CCCC(CI)O	2.0	18	7	5	1	0	0	5
8	CCCCCC[CH+]C	3.9	25	8	8	0	0	0	8
9	C(CO)NCCO	-1.4	18	7	4	2	1	0	4
10	CCCCP(C)P	1.1	21	7	5	0	0	0	5
11	C(O)OOOCO	-1.0	13	7	2	5	0	0	2
12	C1CC1CCP	1.4	17	6	5	0	0	0	5
13	COCCC(Cl)Cl	1.8	15	7	4	1	0	2	4
14	[2H]C1=CSC=C1[2H]	1.8	9	5	4	0	0	0	4

This table is generated using Python. The column of "smiles" shows the chemical structures of the drugs. "logP" represents the partition coefficient solubility. Other columns such as "num_of_atoms" and "num_of_heavy_atoms" are specific features of the chemical structures of the drugs. There are 14610 drugs in total in the database. The data shown in the table is a sample and contains 15 drugs of the database.

To train a model to recognize similar structures of the drugs with those that have the potential to cause AD, we need to decompose the number of feature dimensions of the various chemical characteristics of each drug. Otherwise, there would be too many features to analyze. We did this by using Principle Component Analysis (PCA), and reduced the 20 features to 2 combined features.

3.2. Dimensionality reduction of molecular fingerprint data

To train a model to recognize similar structures of the drugs with those that have the potential to cause AD, we need to decompose the number of feature dimensions of the various chemical characteristics of each drug. Otherwise, there would be too many features to analyze. We did this by using Principle Component Analysis (PCA), and reduced to 2 combined features (Figure 2).

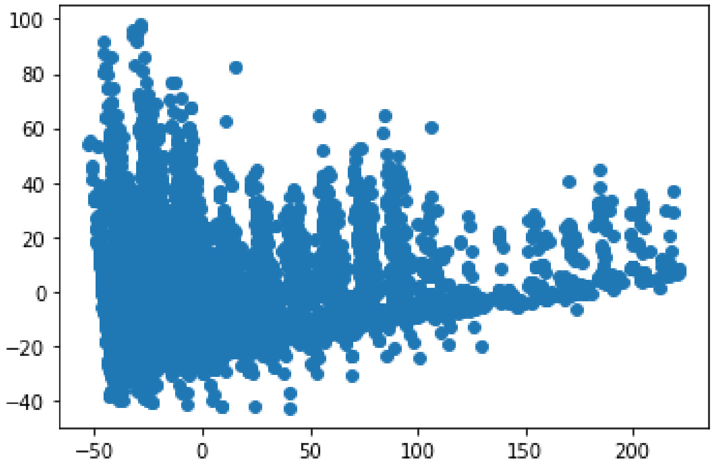

Figure 2. PCA dimensionality reduction of molecular fingerprint data.

Each dot represents a single molecule. The x and y-axes are derived from latent PCA loadings.

The variance from the original data needs to be checked to make sure that the reduced dimensions are a good representation for the original data. Afterwards, Uniform Manifold Approximation and Projection (UMAP) is used to show the relation between compounds: the closer the dots are to each other, the more similar their structures are (Figure 3).

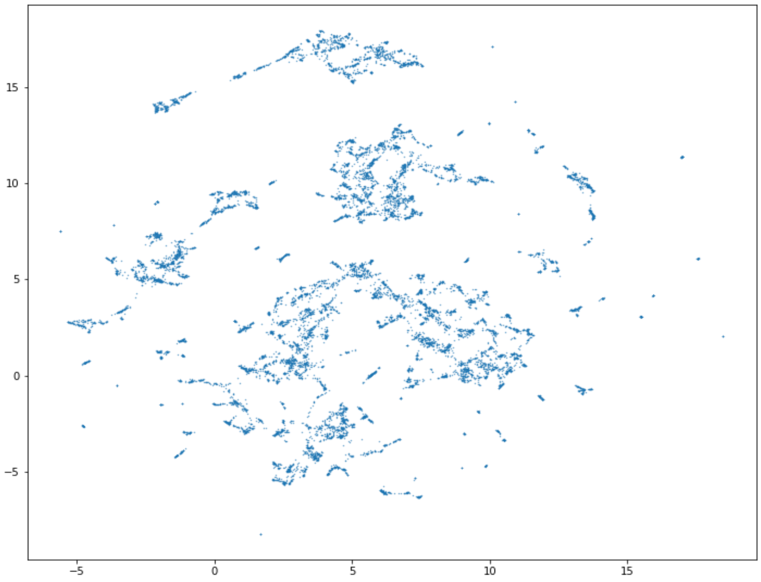

Figure 3. Uniform Manifold Approximation and Projection (UMAP) of molecular fingerprint data.

Each dot represents a single molecule. The x and y-axes are derived from latent UMAP loadings.

3.3. Autoencoder-based feature reduction for molecular group classification

The relationship obtained from PCA was loosely presented. Next we also tried using an alternative method for feature reduction. Autoencoder is a common process in machine learning. It encodes features into summarized fewer components, and it also has the advantage of encoding those features back to the original state to check the accuracy of the reduction automatically. From this process we obtained a much more condensed and clear relationship between the molecules (Figure 4). Clusters are formed, so if there are potential AD-causing drugs within the clusters, the other compounds within the same clusters would also be very likely to be AD-causing compounds.

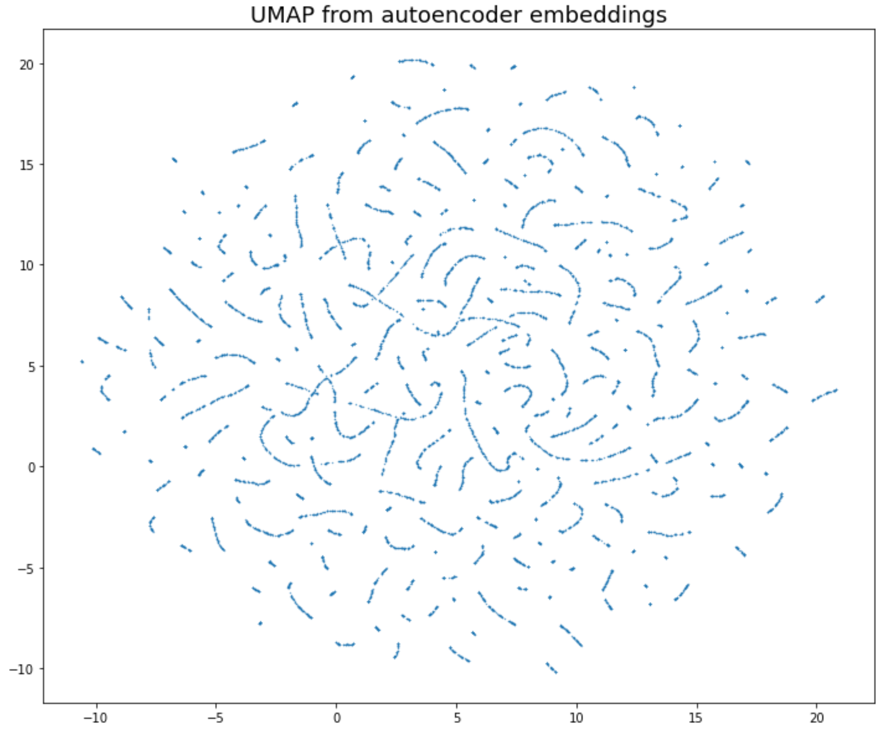

Figure 4. Uniform Manifold Approximation and Projection (UMAP) of the encoded latent features.

Each dot represents a single molecule. The x and y-axes are derived from latent UMAP loadings based on the encoded variables of the autoencoder.

4. Discussion

In conclusion, the paper used feature extraction, reduction and comparison techniques to find the most possible drugs that may be a risk factor for Alzheimer's Disease. The coding was generally successful in finding the drugs that may have the closest relationship with known AD-causing drugs. However, there are still many things that remain uncertain.

First, the extraction of features was based on the theory that the more similar the molecules were, the more similar their chemical properties would be. While this is true in theory, it cannot be guaranteed to be completely accurate in real life. Special cases should always be considered. It would have been better if the features could have a variety of sources, not only their molecular structure, but also their real-life experimental properties. However, as computer coding could only take the absolute data recognized by the computer, the choice of information was limited.

What's more, there were multiple possible errors brought to the results. Because the feature dimensions needed reducing, the reduced results might not have been an entirely accurate representation of the original data, as demonstrated by the PCA technique. Also, the relationship provided by the UMAP was ambiguous as there was only the graph to show the possible related molecules.

5. Conclusion

The methods applied for screening in this research paper are the generation of molecular features and feature extraction, dimensionality reduction using principal component analysis, and the use of uniform manifold approximation and projection to analyze the relationships between the drugs being screened and the target drug. The paper has provided a convenient process of drug screening, but further precision is needed. Perhaps a more precise dimensionality reduction method and a more direct method for investigating correlation would be better for the results of screening.

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

References

[1] Olazarán J., Carnero-Pardo C., Fortea J., Sánchez-Juan P., García-Ribas G., Viñuela F, Martínez-Lage P, Boada M (2023) Prevalence of treated patients with Alzheimer’s disease: current trends and COVID-19 impact. Alzheimer’s Research & Therapy 15:130.

[2] Goedert, M. (1993). Tau protein and the neurofibrillary pathology of Alzheimer’s disease. Trends in neurosciences, 16(11), 460-465.

[3] Baas, P. W., & Qiang, L. (2019). Tau: it’s not what you think. *Trends in cell biology*, *29*(6), 452-461.

[4] Gallacher, J., Elwood, P., Pickering, J., Bayer, A., Fish, M., & Ben-Shlomo, Y. (2012). Benzodiazepine use and risk of dementia: evidence from the Caerphilly Prospective Study (CaPS). *J Epidemiol Community Health*, *66*(10), 869-873.

[5] Landrum G. (2006). RDKit: Open-source cheminformatics. 2006. Google Scholar.https://cir.nii.ac.jp/crid/1370004237630036224 [Accessed October 2, 2023].

[6] Morgan H.L. (1965). The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J Chem Doc 5:107–113.

[7] Jaeger S., Fulle S., Turk S.. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model 58:27–35.

[8] Abdi H., Williams L. J. (2010). Principal component analysis. WIREs Computational Statistics 2:433–459.

[9] Niki K. (1989). Layered dynamic auto-associative memory with auto-encoder and feedback. In: International 1989 Joint Conference on Neural Networks, pp 571 vols.2- Available at: https://ieeexplore.ieee.org/abstract/document/118308 [Accessed October 4, 2023].