Predictive Modeling of Natural Medicinal Compounds for Alzheimer's Disease Using Cheminformatics

Hafiza Syeda Yusra Tirmizi[1], Syed Ibad Hasnain[1], Muhammad Faris[1], Rabail Khowaja[1], Saad Abdullah[2]

[1] Faculty of Engineering Science and Technology, Department of Biomedical Engineering, Hamdard University, Karachi, Pakistan

[2] School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden

Abstract

The most conventional cause of dementia is that of a progressive neurodegenerative disorder: that of Alzheimer disease, which is a disease affecting older adults and slowly worsening memory, thinking and behaviour. It is typified by deposition of maladaptive protein in the brain i.e. amyloid-b plaques and neurofibrillary tangles of tau protein, which interfere with neuronal communication and cause brain cells to die. Early symptoms usually involve a slight loss of memory and the inability to acquire new information and later in the disease, the sufferers may experience severe impairment of cognition, loss of autonomy and personality and behavioural changes. Even though the precise cause of Alzheimer disease is not clearly known, age, genetics, lifestyle and cardiovascular health are some of the major points of concern in the development of this disease. No cure exists currently, although the timely diagnosis, pharmacological therapy and supportive care can be applied to slow down the advancement of the symptoms and enhance the quality of life of both patients and their caregivers. This paper is a predictive chemo metallurgy-based model, which predicts natural medicinal compounds with possible therapeutic benefit in the treatment of Alzheimer Disease (AD). The model is an effective drug screening system based on molecular descriptors and machine learning to discover anti-Alzheimer active natural compounds. ChEBI, SynSysNet and INDOFINE provided more than 7000 compounds, which were pre-processed using Open Babel and analyzed using Dragon descriptors. A classifier based on FDA-approved treatments trained as a Random Forest yielded some level of precision (5970) and recall (6590), and identified 73 promising compounds. The major descriptors were polarizability of atoms, multiplicity of bonds and the number of non-hydrogen bonds. Findings show that cheminformatics is useful in early drug discovery, which is a safer, more approachable route to AD treatments.

Keywords: Cheminformatics; Alzheimer disease; Random Forest; QSAR; Natural compounds; Drug discovery

1. Introduction

Alzheimer disease (AD) is a neurodegenerative, progressive and irreversible disorder and the most prevalent cause of dementia in the entire world. It is marked by the slow deterioration of memory, language, executive functions among other cognitive functions which highly disrupt everyday life. Neuropathologically, AD is found to be characterized by the presence of extracellular amyloid plaques that are predominantly composed of amyloid- amyloid-β (Aβ) peptides and intracellular neurofibrillary tangles made up of hyper phosphorylated tau protein[1] . In spite of the extensive literature, AD still remains an important unsatisfied medical need with no curative or strongly disease-modifying therapy avenues. As the approved pharmacological therapy of AD is currently in existence, it only gives the symptomatic relief [2]. They are acetylcholinesterase inhibitors- donepezil, rivastigmine, and galantamine, and an NMDA receptor antagonist memantine. Although the drugs can be used to temporarily enhance cognitive functioning or delay the onset

of symptoms, they fail to modify the disease pathology. The pipeline of drug development in the Alzheimer disease case has been reviewed indicating that the rate of drug wastage is very high with a number of the drugs being terminated in late clinical trials phases because they were not effective or safe enough [3]. The major pathological finding in AD is the deposition of amyloid-B peptides, and especially the $A\beta_{1-42}$ isoform that is highly aggregable and neurotoxic. They are peptides, owing to amyloid precursor protein (APP), which build up, owing to the lack of equilibrium between production and clearance. There is developing evidence to suggest that sporadic AD cases are more definitely connected to the failure of the A -clearance mechanisms as opposed to production, and thus there is a need to use therapeutic interventions that increase clearance and lessen neurotoxicity [4]. Single-target drugs have not been very successful because of the multifactorial pathology of AD, which incorporates oxidative stress, neuroinflammation, dysfunction of synapses and neuronal loss. Therefore, multitarget therapeutic approaches are becoming more and more popular. In that regard, the natural compounds appear as potential candidates because of the structural diversity, biological activity, and the generally positive safety profile. Natural products have also played a major role in the discovery of drugs, some of which are used to treat complex chronic illnesses. The development of cheminformatics and machine learning has revolutionized the contemporary drug discovery process because it allows the rapid screening and prediction of bioactive substances. Included in computational methods to identify promising natural compounds are QSAR modeling, molecular similarity and analysis, and virtual screening, which are less cost and time-consuming than the traditional method of experimental procedures [5]. Thus, this research aims to recognize natural compounds that have a possibility of being anti-Alzheimer using predictive methodologies based on cheminformatics, to find safer, cheaper, and more effective therapeutic agents. Effective disease-modifying therapies of Alzheimer disease (AD) development has been extremely difficult, in spite of decades of extensive research. Surveys of current drug development pipelines reveal that the majority of investigational therapies focus on amyloid- 2 or tau pathology, but only a small number of them have been approved by the regulators. Consecutive failures of amyloid-targeting drugs in late-stage clinical trials have raised doubts about single-target strategies and have resulted in a shift in strategy toward early-intervention and multitarget therapy mechanisms and modalities that more effectively represent the multifactorial nature of AD [6]. Natural products still are considered a significant component in the present-day drug discovery and still find use as an alternative source of structurally diverse and pharmacologically active compounds. An extensive review of the history of drug development during close to forty years showed that a good percentage of the drugs approved are either natural products or analogues of natural products [7]. Recent investigations also highlight the applicability of natural product chemistry to the context of responding to unmet medical demands, especially in more complex diseases including neurodegeneration, where synthetic methods have been less effective in general [8] . Multitarget drug discovery has become a growing trend in the context of Alzheimer disease research, with the idea that drugs which have the ability to influence a variety of pathological pathways have the potential to provide a better therapeutic effect. A number of studies have demonstrated that natural and synthetic compounds have potential to inhibit both of major AD-associated targets, such as β-secretase (BACE-1) and glycogen synthase kinase-3b (GSK-3b), which can justify the viability of this strategy. The development of cheminformatics and quantitative structure activity relationship (QSAR) modeling has become a meaningful contribution to drug discovery at an early stage. Virtual screening using QSAR makes it possible to predict biological activity on the basis of the molecular structure and to efficiently identify promising compounds provided by large chemical libraries [9]. The techniques are very dependent on the molecular descriptors, which are quantitative measurements of physicochemical and structural characteristics of chemical compounds and form the foundation of computational modeling. Chemical data processing, the generation of descriptors, and the standardization of compounds have also been supported with the help of open-source software like Open Babel, which allows large-scale computational studies to be carried out. Machine-learning algorithms, especially random forest models, have been demonstrated to have good predictive results in cheminformatics because it can deal with complex nonlinear data and biological

variability [10] . Other recent studies have also reported the usefulness of constrained machine-learning methods to enhance the overall model robustness and interpretability of biomedical models. The machine learning field has gained more and more popularity in biomedical studies, especially those that involve neurological and psychiatric disorders, where high-dimensional data, including complex ones, are prevalent [11] . The analysis of heterogeneous data in large amounts has shown a significant potential of identifying disease-related patterns and detecting mental illnesses using ML-based techniques. Machine-learning methods have been extensively used in drug discovery, particularly in target identification, screen of compounds, lead optimization, and toxicity prediction, saving a lot of time and money that is usually costly in conventional experimental procedures. In the current investigation, machine-learning algorithms and cheminformatics were combined to evaluate and rank natural compounds that can have anti-Alzheimer-like activity [12]. Before analysis, all chemical structures were standardized with Open Babel including addition of missing hydrogen atoms, geometry optimization, and conversion of structure files to formats that are compatible with DRAGON software, to reduce structural inconsistencies and artifacts in computation [13]. After the preprocessing of the structures, molecular descriptors were then computed with DRAGON molecular descriptor calculators to provide a numerical expression of the structural and physicochemical characteristics of each compound. Over 3, 200 descriptors have been produced including constitutional descriptors, topological indices, geometrical descriptors, atom-type descriptors, and three dimensional (3D) molecular descriptors [14]. Overly correlated, redundant, non-informative descriptors were removed to reduce dimension to enhance the quality of data. Statistica software was then used to build a RF classifier, where data was partitioned into training and testing sets of 70 percent and 30 percent, respectively, and compounds that were considered active or inactive RF was picked because it has a high level of robustness, the capacity to work with high-dimensional data, and, features with importance are estimated automatically [15]. Simultaneously, the molecular similarity analysis was utilized in order to justify the prioritization of compounds due to the structural similarity with the known anti-Alzheimer drugs. This mathematical model complies with the current evidence of therapeutic value of nutraceuticals and phytochemicals which tend to have antioxidant, anti-inflammatory, and anti-amyloidogenic effects. The current developments in computational modeling, biomarker identification, and artificial intelligence are still advancing to improve the study of neurodegenerative diseases and speed up the process of identifying safe and effective therapeutic candidates in the case of Alzheimer disease [16].

2. Methodology

2.1. Study Workflow

Data Collection: over 7000 natural compounds were found in the ChEBI, SynSysNet and INDOFINE databases. Description Calculation: Descriptors obtained with Dragon software. Preprocessing Structures with Open Babel structure standardization. Model Development: Statistica random forest classification that was trained using known FDA-approved AD drugs. Validation: Measured using precision and recall, with strong predictions of the model. Figure 1 shows the general cheminformatics workflow that was used in the present study that includes data collection, preprocessing, calculating descriptors, model development, and predicting activities.

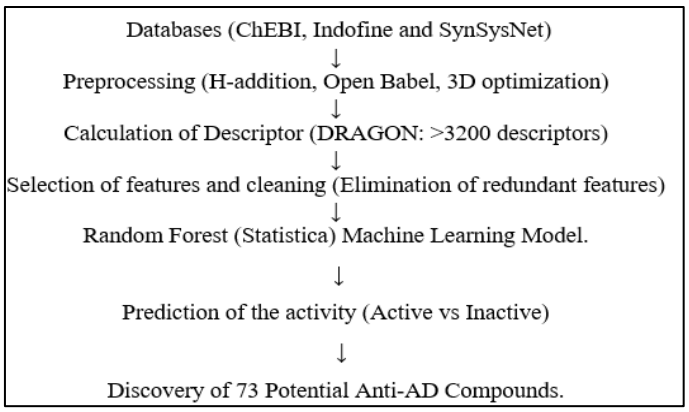

Fig. 1. Anti-Alzheimer workflow Cheminformatics workflow

2.2 Data Collection

The collection of natural compounds into publicly available chemical databases consisted of 7,004 compounds, of which the Chemical Entities of Biological Interest (ChEBI) database provided approximately 3,000 of the compounds, the SynSysNet provided about 3,885 compounds and the Indofine Chemical Company provided 119 compounds. Alongside these natural compounds, the set of 20 FDA-approved anti-Alzheimer drugs or the list of anti-Alzheimer drugs reported in clinical trials served as an active reference dataset to train and test the predictive model.

2.3 Structure Preprocessing

The chemical structures were analyzed by using Open Babel to standardize them before analysis. The preprocessing involved the insertion of missing hydrogen atoms, geometry optimization of the molecular structures and converting the structure files into a format that is readable by DRAGON. The molecular descriptors were computed after preprocessing using Molecular Descriptor Calculators to express a quantitative measure of the structural and physicochemical properties of the compounds. To compute the molecular descriptors, over 3,200 descriptors were calculated with the help of DRAGON software. These descriptors covered a large variety of molecular properties, such as constitutional descriptors, topological indices, geometrical descriptors, atom-type descriptors, and three-dimensional (3D) molecular descriptors and gave a complete numerical description of a given compound. Descriptions that were highly correlated and non-informative were eliminated before model development.

2.4 Development of machine learning model

Random Forest (RF) classifier was performed with the help of Statistica software. The data was divided into a training and testing set of 70 and 30 percent, respectively. The classification criterion was binary; active (anti-Alzheimer) and inactive. RF was chosen because it is robust, handles high-dimensional data, is not prone to overfitting, and has in-built feature significance estimation.

3. Results

This prediction system indicates how cheminformatics can be powerful in the identification of drug-like natural compounds. The findings have offered a useful screening tool in the discovery of early-stage AD drugs, which is the background to laboratory validation developmental work. Table 1 is a summative table of the most effective molecular descriptors based on the random forest model. These descriptions are important physicochemical and structural features like electronic distribution, bonding pattern and hydrogen-bonding capacity. Combined, they indicate molecular

characteristics essential to the ligand–receptor interactions, vascular permeability across the blood brain barrier and binding affinity of the Alzheimer disease-related targets.

Table 1. Major Molecular Descriptors of Model Development

Descriptor	Description	Relevance to AD Activity
sp	Sum of atomic polarizabilities	Influences electronic distribution and ligand–receptor interactions
nBo	Number of non-hydrogen bonds	Related to molecular rigidity and binding stability
nBM	Number of multiple bonds	Indicates degree of unsaturation and π–π interactions
MW	Molecular weight	Associated with drug-likeness and blood–brain barrier permeability
HBD	Hydrogen bond donors	Affects enzyme and receptor binding
HBA	Hydrogen bond acceptors	Influences binding affinity

Descriptor Description Relevancy to AD Activity is the sum of atomic polarizabilities It plays a role in electronic distribution and ligand receptor interactions. nBo non-hydrogen bonds number Molecular rigidity and binding stability. nBn Multiplicity of multiple bonds shows degree of unsaturation. MW (Molecular weight) Drug-likeness and blood-brain barrier permeability. HBA Hydrogen bond acceptors Affects binding affinity.

Table 2. Performance Measures

Metric	Value Range
Precision	59–70%
Recall (Sensitivity)	65–90%
Specificity	~80%
Overall Accuracy	~75–85%

The performance of the Random Forest model in classification is shown in Table 2. The discrimination values of balanced precision and recall are the indication of reliable discrimination of active and inactive compounds. The high specificity and the out-of-bag (OOB) error are also indicators of the soundness of the model and its minimal vulnerability to overfitting.

Table 3. Compound Matching on The Basis of Molecular Weights

Reference Compound (Training Set)	Molecular Weight (g/mol)	Test Compound (CID)	Molecular Weight (g/mol)
Donepezil	**379.49**	10319750	416.59
		10409683	416.59
		10000637	393.28
		10000636	393.28
		10133297	363.19
Rivastigmine	**250.33**	10489322	224.25
		10106813	263.30
		446475	246.26
Tacrine	**198.26**	10313432	186.23
		10487582	179.22
		10081162	187.21
		10081163	187.22
		10081053	182.26
Δ9-tetrahydrocannabinol (THC)	**314.46**	10064386	317.16
		10157035	297.41
		10214826	290.12
		72276	290.27
		5280701	318.45
		10064445	318.10
Physostigmine	**275.34**	10067250	363.19
		10106813	263.30
		10133297	363.19
		102175	281.20
		10214826	290.12
		72276	290.27
Vitamin E	**430.70**	10389037	449.72
		10048025	425.11
		5459811	426.72
		638072	410.72
		5289598	414.35
Curcumin	**368.37**	10067250	363.19
		10133297	363.19
		10000106	384.55
		10090714	373.22
		10361722	369.24
		65728	386.65
Memantine	**179.30**	10313432	186.23
		10487582	179.22
		104766	173.16
		1050	167.16
		1005	168.04
		10081053	182.26
		10081162	187.21
		10313301	170.20
		79025	180.16
		445929	194.14
Huperzine A	**242.32**	10490043	237.29
		10269	236.35

Table 4. ***Best Predicted Anti-Alzheimer Natural Compounds*.**

PubChem ID	Compound Name	RF Probability Score	Source Database
10419163	(E)-3-aminoprop-1-enyl-hydroxy-oxo-phosphanium	0.909	SynSysNet
10419374	(3R)-3-amino-2-hydroxybutyl-hydroxy-oxophosphanium	0.892	SynSysNet
1005	Phosphoenolpyruvate	0.846	ChEBI
10219727	Amino-triazolyl propanoic acid	0.846	SynSysNet
10219992	Cyclopentylidene-pentanedioic acid	0.846	SynSysNet
10297	Phenylpropanolamine	0.846	ChEBI

The important features are analyzed according to their significance to the study. Figure 2 demonstrates the relative significance of the most significant molecular descriptors found by the Random Forest model. The graph indicates the importance of the molecular descriptors in the model. The strongest impact is made by atomic polarizability, then the bonding-related features and finally the impact is made by molecular weight and hydrogen bonding. The model uses electronic properties and bond structure as the key factors in making its predictions, as opposed to the size of the molecule and capability of hydrogen-bonding.

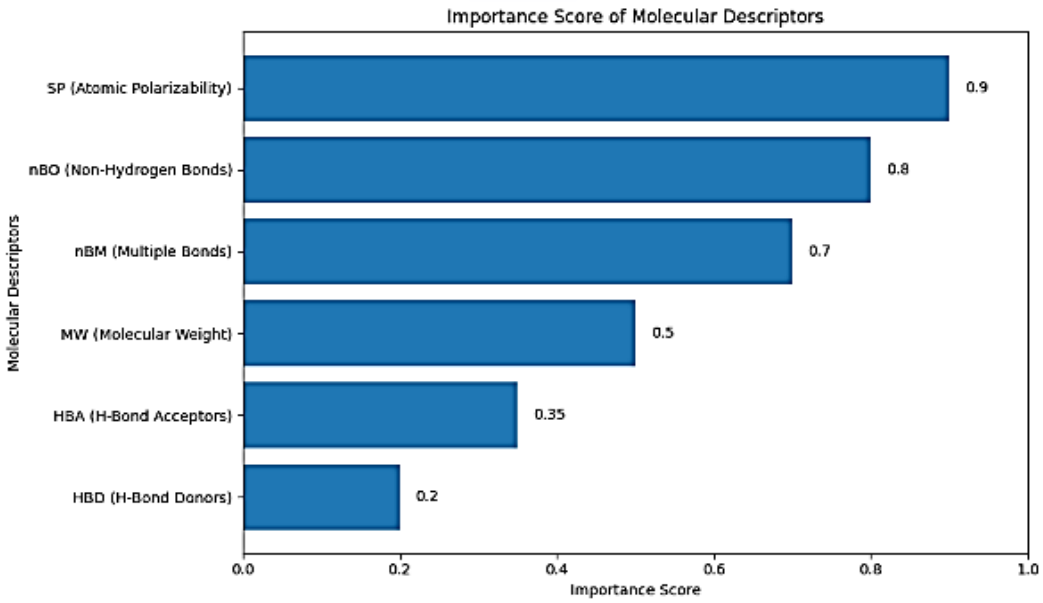

Fig 2. Relative significance of the best molecular descriptors

This confusion matrix indicates the level of accuracy of the model in classifying compounds into Active or Inactive. Of the really inactive compounds, 56 were rightly identified as inactive and 7 were rightly identified as active. This indicates that the model is highly effective in determining inactive compounds. In the case of actually active compounds, 13 correctly predicted active, and 4 incorrectly predicted inactive, that is, a few active compounds have been missed. The error analysis (out-of-bag (OOB) and confusion matrix) (refer to Figure 3) demonstrates the fact that overfitting is minimal, and generalization is accurate.

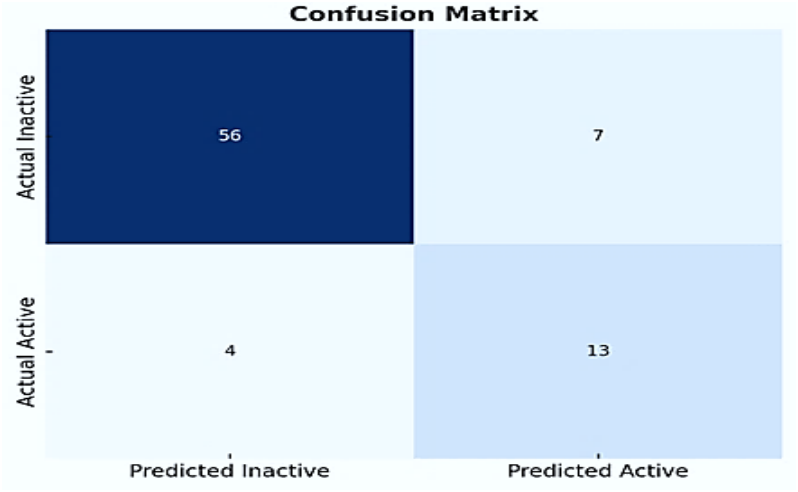

Fig 3. Confusion Matrix of the Random Forest Model

The model had forecasted 73 natural compounds that could be anti-Alzheimer. Most of them were based on SynSysNet and ChEBI data bases. A number of compounds had high RF probability scores (>0.85) showing high predicted activity. The predicates of the active compounds are shown in Figure 4, as they are distributed among the various sources databases. The proportion of compounds in each of the source databases is depicted in this bar graph. The most prominent contributor of compounds is SynSysNet (56), which implies that it is the leading data source. ChEBI contributes less (17 compounds), and Indofine contributes none of the compounds (0). Most of the predicted actives came in via SynSysNet with ChEBI coming in next as occurred as a result of the greater chemical diversity in these repositories.

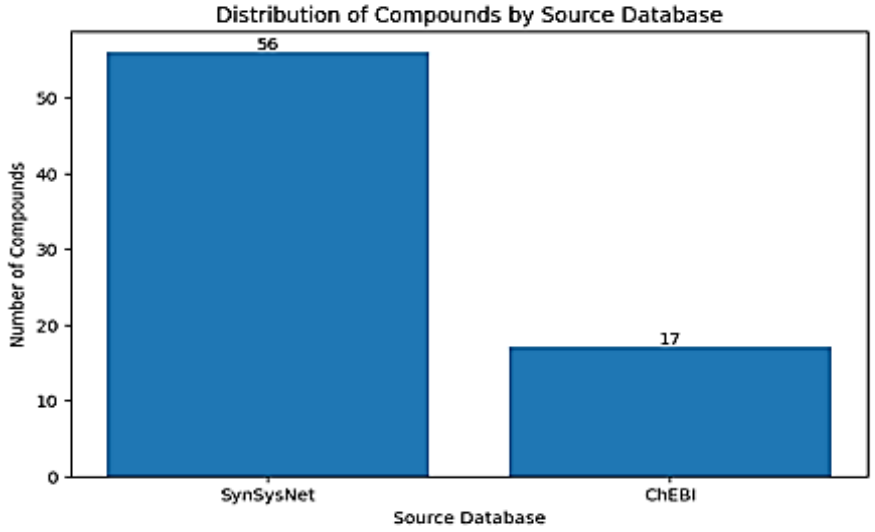

Fig 4. Database dispensing of active compounds of interest.

The suggested compounds had structural characteristics that are attributable to the known mechanisms in anti-Alzheimer such as cholinesterase, glutamatergic signalling, antioxidant, and A-beta aggregation. The significance of the electronic distribution and molecular flexibility in the interactions between ligands and target molecules is indicated by the popularity of the atomic polarizability and bonding-related descriptors. These findings confirm that the random forest types of cheminformatics models are applicable to conducting large scale virtual screening and that more experiments should be conducted on the identified candidates.

4. Discussion

The proposed study proves that a cheminformatics paradigm based on random forests is highly effective to screen natural compounds with anti-Alzheimer potential activity. The proposed compounds have predicted structural characteristics, which are similar to established pharmacological processes, such as cholinesterase inhibition, antioxidant activity, glutamatergic signalling and amyloid-2 aggregation. The analysis of feature importance revealed that such characteristics as atomic polarizability, the number of non-hydrogen bonds, multiplicity of bonds were important descriptors, which should be considered when studying the interaction of ligands and targets, focusing on the importance of the electronic properties of compounds and their ability to be flexed. The model demonstrated a good level of predictive performance that was balanced with consistent out-of-bag error, which indicated a high level of robustness and less overfitting. These findings justify the validity of ensemble machine learning techniques in conducting large-scale virtual screening in the study of neurodegenerative diseases.

5. Conclusion and Future Work

This paper establishes that predictive modeling involving cheminformatics is a viable approach to finding promising natural compounds as a therapeutic approach to Alzheimer disease. The screening of more than 7000 compounds resulted in the identification of 73 most promising anti-Alzheimer structure. The technology provides a low-cost and high-throughput drug discovery solution that is used during early drug discovery stages and forms a good basis to conduct in-vitro and in-vivo validation studies. The next round of research will aim at testing the hypothesized compounds during molecular docking with the most pivotal targets of Alzheimer disease such as acetylcholinesterase, butyrylcholinesterase, NMDA receptors, and amyloid-2 aggregates. Pharmacokinetic and toxicity Pharmacokinetic and toxicity profiling (ADMET) will be done to determine drug-likeness and blood-brain barrier permeability. To verify anti-Alzheimer in activity of high-scoring compounds in-vitro biological assays will be conducted. Also, it is possible to construct multi-target and deep learning-based QSAR models which could be used to enhance the accuracy of prediction and therapeutic relevance further.

6. Acknowledgements

We express our sincere gratitude to Hamdard University and Mälardalen University for supporting throughout the research.

References

[1] D. V. Parums, "A review of the current status of disease-modifying therapies and prevention of Alzheimer's disease," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research,* vol. 30, pp. e945091-1, 2024.

[2] J. Cummings, G. Lee, A. Ritter, M. Sabbagh, and K. Zhong, "Alzheimer's disease drug development pipeline: 2019," *Alzheimer's & Dementia: Translational Research & Clinical Interventions,* vol. 5, pp. 272-293 %@ 2352-8737, 2019.

[3] P. Van Bokhoven *et al.*, "The Alzheimer's disease drug development landscape," *Alzheimer's research & therapy,* vol. 13, no. 1, pp. 186 %@ 1758-9193, 2021.

[4] D. J. Newman and G. M. Cragg, "Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019," *Journal of natural products,* vol. 83, no. 3, pp. 770-803 %@ 0163-3864, 2020.

[5] M. S. Butler and J. J. La Clair, "The Role of Natural Product Chemistry in Drug Discovery: Two Decades of Progress and Perspectives," *Journal of Natural Products %@ 0163-3864,* 2025.

[6] F. Prati *et al.*, "Multitarget drug discovery for Alzheimer's disease: triazinones as BACE‐1 and GSK‐3β inhibitors," *Angewandte Chemie International Edition,* vol. 54, no. 5, pp. 1578-1582 %@ 1433-7851, 2015.

[7] B. J. Neves, R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade, "QSAR-based virtual screening: advances and applications in drug discovery," *Frontiers in pharmacology,* vol. 9, pp.

1275 %@ 1663-9812, 2018.
[8] Y. Qiu and F. Cheng, "Artificial intelligence for drug discovery and development in Alzheimer's disease," *Current opinion in structural biology,* vol. 85, pp. 102776 %@ 0959-440X, 2024.
[9] O. V. Tinkov, V. N. Osipov, A. V. Kolotaev, D. S. Khachatryan, and V. Y. Grigorev, "HT_PREDICT: a machine learning-based computational open-source tool for screening HDAC6 inhibitors," *SAR and QSAR in Environmental Research,* vol. 35, no. 6, pp. 505-530 %@ 1062-936X, 2024.
[10] S. I. Hasnain, H. Israr, M. Faris, R. Kamal, and H. S. Y. Tirmiz, "Evaluation of Machine Learning–Based Methods to Detect Bipolar Disorder in Individuals With Mental Health Conditions," *VFAST Transactions on Software Engineering,* vol. 13, no. 3, pp. 129-139, 2025.
[11] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *International Journal of Data Science and Analytics,* vol. 20, no. 2, pp. 197-211, 2025.
[12] T. T. Bui and T. H. Nguyen, "Natural product for the treatment of Alzheimer's disease," *Journal of basic and clinical physiology and pharmacology,* vol. 28, no. 5, pp. 413-423 %@ 2191-0286, 2017.
[13] R. Artiñano-Muñoz, L. Prieto-Santamaría, A. Pérez-Pérez, and A. Rodríguez-González, "DRAGON: drug repurposing via graph neural networks with drug and protein embeddings as features," 2024: IEEE, pp. 170-175 %@ 9798350384727.
[14] A. Ion, M. Praisler, and S. Gosav, "Molecular descriptors–an useful tool for assessing the physico-chemical properties of hallucinogenic drugs of abuse," *Analele Universităţii" Dunărea de Jos" din Galaţi. Fascicula II, Matematică, fizică, mecanică teoretică/Annals of the" Dunarea de Jos" University of Galati. Fascicle II, Mathematics, Physics, Theoretical Mechanics,* vol. 44, no. 1, pp. 26-29 %@ 2668-7151, 2021.
[15] G. Cano *et al.*, "Automatic selection of molecular descriptors using random forest: Application to drug discovery," *Expert Systems with Applications,* vol. 72, pp. 151-159 %@ 0957-4174, 2017.
[16] K. Skolariki, T. P. Exarchos, and P. Vlamos, "Computational models for biomarker discovery," 2022: Springer, pp. 289-295.