COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Computational Intelligence and Machine Learning
Volume 6, Issue 1, April 2025

Quantitative Structure-Activity Relationship (QSAR) Modeling and Similarity Search for Virtual Screening of Acetylcholinesterase (AChE) Inhibitors in Alzheimer's Disease

Leandro Pedrosa[1*], Renata Dutra Braga[2], Tatiane Nogueira Rios[3]

[1, 2, 3] MBA in Artificial Intelligence and Big Data, Department of Computer Science, Institute of Mathematics and Computer Sciences (ICMC), University of São Paulo, São Carlos-SP, Brazil
*Corresponding Author: leandropedrosalp@gmail.com

Abstract

This study explores the combined use of virtual screening, QSAR modeling, and machine learning to identify potential acetylcholinesterase (AChE) inhibitors for Alzheimer's disease. Virtual screening enables the analysis of large molecular databases, while QSAR modeling predicts the biological properties of compounds based on their chemical structures. By integrating these approaches with molecular descriptors (Morgan, RDKit, and SiRMS), the research improves predictive accuracy and efficiency. Machine learning and deep learning models, including Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP), and TensorFlow, were applied, demonstrating how these techniques complement each other to enhance drug discovery. We applied a structured process to identify promising compounds through virtual screening. Initially, thousands of potential acetylcholinesterase (AChE) inhibitors were identified. To enhance reliability, consensus modeling was used, integrating predictions from four algorithms and three molecular descriptor sets. Of these, consensus analysis selected 37 compounds based on predictions from four algorithms and three descriptor sets, further refined by similarity searches using the Tanimoto coefficient. Compounds with more than 50% similarity to reference molecules such as tacrine were prioritized, highlighting the robustness of the approach. These results underscore the potential of computational approaches to accelerate drug discovery and improve therapeutic outcomes for neurodegenerative diseases.

Keywords

Alzheimer's Disease, AChE Inhibitors, Machine Learning, Virtual Screening

INTRODUCTION

Alzheimer's disease is one of the most common and debilitating neurodegenerative disorders worldwide, characterized by progressive memory loss, cognitive decline, and a substantial burden on patients, families, and healthcare systems. Affecting millions of people worldwide, Alzheimer's disease represents a significant public health challenge, particularly as the population ages and the prevalence of dementia increases [1]. The pathology of Alzheimer's disease is characterized by the accumulation of amyloid beta plaques and neurofibrillary tangles in the brain, which disrupt neuronal communication and lead to cognitive impairment [2]. This deterioration leads to a gradual but severe loss of independence, culminating in total dependence on caregivers and a dramatic reduction in quality of life [3].

Despite significant progress in research, there is still no definitive cure for Alzheimer's disease, and existing treatments provide only symptomatic relief and do not slow or halt the progression of the disease. This pressing challenge underscores the urgent need for innovative therapeutic approaches that address the underlying mechanisms of the disease. Among these, inhibition of acetylcholinesterase (AChE) - an enzyme responsible for the breakdown of acetylcholine in the brain - has emerged as a promising strategy, since acetylcholine levels can be increased, thereby improving neuronal transmission and potentially alleviating the cognitive symptoms associated with Alzheimer's disease [4]. Therefore, the discovery and development of potent and selective AChE inhibitors is a critical step in advancing more effective treatment options for this debilitating disease.

In recent years, the use of machine learning and deep learning techniques has emerged as a promising approach in the search for new compounds to treat Alzheimer's disease [1, 2, 3]. These computational methods enable the rapid and cost-effective identification of potential drug candidates by analyzing complex data sets and predicting biological activity. However, the accuracy and reliability of these models is largely dependent on the careful selection of algorithms and molecular descriptors, which are numerical or categorical representations of a molecule's chemical and structural features, used to predict physicochemical, biological or toxicological properties. These elements are critical to building robust and accurate Quantitative Structure-Activity Relationship (QSAR) models [4, 5]. By optimizing them, it is possible to improve the efficiency and reliability using machine learning, paving the way for innovative therapeutic solutions for Alzheimer's disease.

The primary research problem was to identify compounds that could effectively inhibit AChE with high specificity and potency, thereby offering therapeutic potential for Alzheimer's disease. Traditional drug discovery methods are time-consuming and resource-intensive, often requiring extensive in vitro and in vivo testing before viable candidates are identified. Computational approaches, such as QSAR modeling [5], offer a promising alternative by predicting biological activity based on molecular structure. However, achieving both accuracy and efficiency in QSAR models remains a challenge, largely due to the complexity of molecular interactions and the need for high-quality data and robust algorithms [4].

The hypothesis of this research was that the development of QSAR models using a combination of machine learning and deep learning approaches along with diverse molecular descriptors would yield predictive models with greater generalizability and accuracy. By selecting and combining models based on their hyperparameters and domain applicability, it was hypothesized that QSAR models could outperform conventional methods in predicting AChE inhibitory activity.

The main contributions of the paper are:

1 - The use of consensus modeling, which combines the predictions of multiple models, to enhance the accuracy of virtual screening.

2 - By improving virtual screening methods, the study contributes to the identification of new potential acetylcholinesterase (AChE) inhibitors for Alzheimer's disease.

3 - The research demonstrates how integrating various computational approaches, such as QSAR modeling and machine learning, can lead to more reliable drug discovery processes.

BACKGROUND

This section will explore the role of Quantitative Structure-Activity Relationship (QSAR) modeling and chemoinformatics in the identification of AChE inhibitors for Alzheimer's disease. It will cover the application of Artificial Intelligence (AI) by machine learning methods and molecular descriptors used in QSAR modeling to predict biological activity. Additionally, the section will discuss the integration of virtual screening techniques and how AI is enhancing the accuracy and efficiency of the drug discovery process for Alzheimer's disease.

QSAR and Artificial Intelligence in Chemoinformatics

QSAR modeling is a cornerstone of chemoinformatics, providing a computational framework for predicting the biological or physicochemical properties of compounds based on their chemical structures [6]. Central to this methodology is the use of molecular descriptors that quantitatively encode chemical properties, allowing the construction of mathematical models that correlate these descriptors with experimentally observed biological activities. This approach is based on the hypothesis that the biological activity of a molecule is intrinsically linked to its structural properties, allowing predictions to be made for untested compounds [7].

The integration of artificial intelligence and machine learning into QSAR modeling has revolutionized the field, significantly improving the accuracy, scalability and efficiency of predictive models [4, 8]. AI techniques facilitate the analysis of large chemical data sets, enabling rapid screening and prioritization of potential drug candidates, reducing the need for extensive laboratory experimentation. By employing algorithms capable of handling high-dimensional data and identifying non-linear relationships, artificial intelligence-based QSAR models have expanded the scope of virtual screening and drug discovery [9].

One of the advances brought by AI is the ability to integrate different types of molecular characteristics, known as descriptors, into predictive models. These descriptors can be simple properties, such as molecular weight (which indicates the size of a molecule) and logP (which measures how well a molecule dissolves in fat versus water). They can also be more complex, such as van der Waals volume, which describes how much space a molecule occupies in three dimensions [10]. Additionally, AI enables the combination of data from multiple sources, creating a more detailed picture of how molecules interact. By leveraging these capabilities, scientists can build more accurate models that address specific challenges in drug discovery [4, 5, 6].

The synergy between QSAR and AI represents a pivotal shift in chemoinformatics, providing a means to efficiently navigate chemical space and accelerate therapeutic discovery. This convergence is not only transforming the field, but also addressing the challenges of modern drug discovery, such as reducing cost and time while maintaining accuracy and reliability [11].

AChE Inhibition and Alzheimer’s Disease

AChE plays a central role in the breakdown of acetylcholine, a neurotransmitter essential for cognitive function in the human brain. In Alzheimer's disease, reduced levels of acetylcholine are associated with cognitive decline, memory loss and other symptoms of dementia [12]. By inhibiting AChE, acetylcholine levels can be maintained, potentially improving neuronal transmission and alleviating cognitive symptoms. AChE inhibition has therefore become a strategic target for the development of Alzheimer's disease therapeutics [13]. The aim of our study was to identify new AChE inhibitors that could be used in the treatment of

Alzheimer's disease, contributing to a growing field of research focused on restoring cognitive function in patients through enzyme modulation.

Machine Learning Methods and Molecular Descriptors for QSAR Modeling

As a subarea of AI, machine learning techniques form the backbone of modern QSAR modeling, providing the computational power necessary to derive meaningful predictions from complex chemical data sets. This study used advanced machine learning methods, including Support Vector Machine (SVM) [14], Random Forest (RF) [15], and Multilayer Perceptron (MLP) [16], each selected for their ability to handle intricate molecular data. In addition, deep learning frameworks such as TensorFlow Keras were used to capture non-linear relationships to further improve predictive performance [17].

A key aspect of our research was the selection of diverse molecular descriptors that encode chemical structures into quantitative features essential for activity prediction. In contrast to the general discussion in QSAR theory, our study emphasized the practical application of descriptors such as Morgan fingerprints, RDKit descriptors, and SiRMS. Each provided unique information into molecular similarity and structure-activity relationships, enabling the development of accurate and generalizable models [10, 11].

To improve predictive performance, the machine learning models underwent extensive hyperparameter tuning and validation. This included the optimization of feature selection procedures and the integration of advanced descriptor combinations, which contributed to the efficiency of the virtual screening for AChE inhibitors. By leveraging these tailored approaches, the study demonstrated significant progress in identifying promising drug candidates for Alzheimer's disease.

By focusing on the operational application of QSAR modeling, this research demonstrates how the integration of specific machine learning techniques and molecular descriptors can directly improve the accuracy and reliability of predictions, ultimately streamlining the drug discovery process.

Virtual Screening

Virtual screening is a computational approach used to analyze large molecular databases in search of compounds with desired chemical or biological properties for potential experimental testing. This method allows the in silico evaluation of hundreds or thousands of compounds against biological targets, significantly increasing the efficiency and success rate of identifying promising drug candidates [8].

In our study, virtual screening was used to identify novel AChE inhibitors, which are critical to Alzheimer's disease treatment strategies. Machine learning techniques were instrumental in supporting the virtual screening process by predicting biological activities and filtering chemical libraries for compounds with high therapeutic potential [4].

Virtual screening approaches have been categorized into ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). LBVS relied on the structural similarity of compounds to known active ligands, while SBVS used molecular docking techniques to predict interactions between compounds and biological target structures [11].

The primary goal of virtual screening was not to replace in vitro or in vivo testing, but to accelerate the drug discovery process, reduce the number of compounds requiring experimental testing, and optimize the selection process. This resulted in significant time, cost and resource savings. QSAR models, which were widely used in the early stages of virtual screening, proved to be valuable tools due to their rapid and accurate predictions of biological activity.

Machine learning-based virtual screening was one of the most computationally efficient techniques developed in recent decades. Researchers built and validated machine learning models that were then used to screen unseen chemical libraries to identify compounds that were likely to have the desired biological activity. Compounds with the highest predicted activity, called "hits", underwent further in vitro testing to confirm their biological effects. The most promising candidates, or "leads," were then optimized and evaluated as potential drugs.

Several machine learning algorithms have been used in virtual screening studies, including Naïve Bayes, k-Nearest Neighbors, SVM, RF, and Artificial Neural Networks. These algorithms have been effective in predicting protein-compound interactions, discovering therapeutic inhibitors, and optimizing compound properties. For example, Naïve Bayesian classifiers and RF were used to predict AChE inhibitors and other therapeutics. Deep learning approaches have further improved the accuracy and scalability of virtual screening efforts. Unlike previous studies, our approach integrates multiple machine learning models (SVM, RF, MLP, and TensorFlow) with diverse molecular descriptors (Morgan, RDKit, and SiRMS), leveraging consensus modeling and similarity searches to enhance predictive reliability and compound selection [14, 15, 16].

Given the urgent need for innovative treatments for Alzheimer's disease, the application of virtual screening represents a significant advance in drug discovery, providing a focused and efficient computational framework for identifying and optimizing new therapeutic candidates.

METHODS

The methodology employed in our study was structured into four main stages: data preparation, QSAR model construction, model validation and virtual screening [9]. Each stage of this methodology involved a series of carefully designed steps to ensure reliable results in the identification of potential AChE inhibitors.

Stage 1 - Data Preparation

The first stage focused on data set preparation, which included defining the chemical, biological, and molecular targets relevant to AChE inhibition.

1) Target definition: The primary focus was on compounds with known or potential inhibitory effects on AChE, a key enzyme involved in Alzheimer's disease.
2) Data set organization: An original dataset was organized that included chemical compounds with relevant structural and biological activity data. This dataset served as the basis for the subsequent modeling steps.
3) Dataset accuracy evaluation: The accuracy and integrity of the dataset were evaluated to ensure high quality data, and the dataset was refined to improve the reliability of the modeling process.
4) Calculate molecular descriptors: Molecular descriptors, or predictive attributes, were calculated for each compound in the dataset. Descriptors such as Morgan Fingerprints, SiRMS, and RDKit were used to capture key structural properties and provide a quantitative basis for the QSAR modeling phase.

Stage 2 - QSAR Model Construction

In the second stage, QSAR models were constructed to predict AChE inhibition potential based on the prepared data set.

1) Dataset splitting: The dataset was divided into training and test sets to enable machine learning model development and performance evaluation. The training set was used to build the models, while the test set provided an independent measure of their predictive accuracy.
2) Model training: QSAR models were developed from the training data using machine learning algorithms such as SVM, RF and MLP. Hyperparameter tuning was performed to optimize the predictive performance of the models.
3) Model validation with test set: The models were validated using the test set, which allowed evaluation of the generalizability and accuracy of each model in predicting AChE inhibition activity.
4) Model selection for external validation: Models with the highest performance metrics were selected for external validation to ensure their reliability across diverse data.
5) Permutation testing: Permutation testing was performed to assess the robustness of the models and to confirm that their predictive power was statistically significant and not a result of chance.

Stage 3 - Model Validation

The third stage involved a comprehensive validation of the models.

1) Consensus prediction for external evaluation: The selected models were applied to an external evaluation dataset using consensus prediction to define domains of applicability. This consensus approach was designed to improve the reliability and robustness of the predictive models in identifying new AChE inhibitors.

Stage 4 - Virtual Screening in Chemical Databases

In the final stage, virtual screening was performed to identify compounds with structural and functional similarity to known AChE inhibitors.

1) Consensus prediction for compounds: The QSAR models generated consensus predictions for compounds screened from large chemical databases, identifying those with a high likelihood of AChE inhibitory activity.
2) Execution of the virtual screening: The virtual screening process was performed using the consensus model predictions to prioritize compounds with the most promising activity profiles. This targeted approach enabled efficient identification of potential therapeutic candidates for Alzheimer's disease.

RESULTS

This experimental evaluation study involved data preparation, QSAR model development and virtual screening to identify potential AChE inhibitors for Alzheimer's disease. The full source code used in this study is available at: https://github.com/leandropedrosa/virtual-screening-qsar-alheimer-acetylcholinesterase, allowing for reproducibility and further exploration of the methodologies applied.

Data Preparation

The data preparation phase was essential to ensuring the quality and relevance of chemical data for modeling construction and virtual screening. AChE was selected as the primary target due to its critical role in the pathophysiology of Alzheimer's disease. As the enzyme responsible for breaking down acetylcholine, its inhibition can help alleviate symptoms of the disease, making it a well-established biomarker for drug discovery.

To build a reliable dataset, data organization and accuracy were prioritized. The initial dataset included 8,832 compounds obtained from the ChemBL database, available at: https://www.ebi.ac.uk/chembl/. Each compound was assessed for its AChE inhibitory activity, and the dataset was refined through preprocessing steps, such as removing

missing values, standardizing activity measurements, and ensuring compliance with Lipinski's Rule of Five, which evaluates drug-likeness.

Following exploratory data analysis and feature selection, the dataset was filtered to retain only compounds classified as active or inactive. This refinement resulted in a final dataset of 4,829 compounds, which was then split into training (80%) and test (20%) subsets to facilitate robust model development and validation.

For virtual screening, compounds were prioritized based on structural similarity to known AChE inhibitors, enabling a more targeted approach in identifying promising drug candidates. Additionally, molecular descriptors were calculated to provide a comprehensive representation of each compound's structure. Three descriptor sets were used: Morgan fingerprints (2048 features), RDKit descriptors (209 features), and SiRMS descriptors (1764 features). These descriptors were selected because they complement each other in capturing different aspects of molecular properties. Morgan fingerprints, based on circular substructures, effectively encode molecular similarity, making them highly useful for ligand-based virtual screening. RDKit descriptors provide a diverse set of physicochemical properties, aiding in a more nuanced understanding of molecular characteristics relevant to bioactivity. SiRMS descriptors incorporate fragment-based representations, which enhance the predictive power of QSAR models by considering structural variations at multiple levels. Together, these three descriptor sets offer a well-rounded approach, balancing structural, physicochemical, and fragment-based information, thus improving the accuracy and generalizability of predictive modeling.

QSAR Model Construction

Using the prepared dataset, machine learning and deep learning algorithms, including SVM, RF, MLP and TensorFlow, were applied to predict AChE inhibitory activity. Hyperparameter optimization using RandomizedSearchCV and Keras Tuner improved model performance.

The models were evaluated using metrics such as accuracy, sensitivity, specificity, and F1-score:

- Morgan descriptors: SVM achieved the highest accuracy (0.87), followed by RF (0.84) and MLP (0.86).
- RDKit descriptors: MLP outperformed with an accuracy of 0.90, while RF and SVM scored 0.85 and 0.92, respectively.
- SiRMS descriptors: Both SVM and RF showed strong results, with accuracies of 0.90 and 0.91.

Permutation tests confirmed the statistical robustness of the models, with p-values below the 0.05 threshold for most configurations, suggesting that the observed performance was not due to chance.

Model Selection and Validation

The best performing models were identified through cross-validation and hyperparameter optimization to ensure their suitability for predicting AChE inhibitory activity. External validation was then performed using independent test data to confirm their generalizability and robustness [18].

Permutation testing was used to assess the statistical significance of model performance (Figure 1) [19]. By shuffling the labels and recalculating the evaluation scores multiple times, these tests generated a distribution of random scores to compare with the true model scores. The resulting p-values, which were consistently below the 0.05 threshold for most configurations, provided strong evidence that the predictive power of the models was not the result of chance.

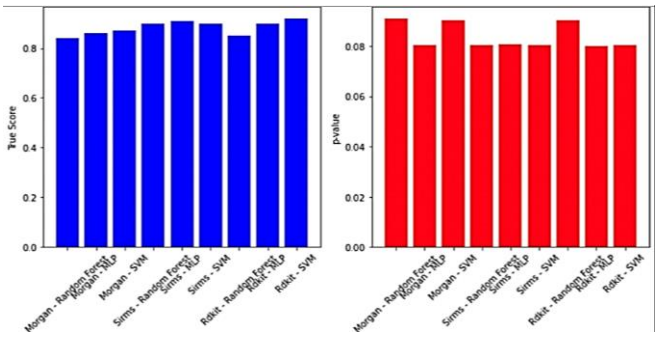


Figure 1. Graphs illustrating permutation tests.

This evaluation confirmed that the selected models, particularly those using the RDKit and SiRMS descriptors, exhibited high accuracy, sensitivity, and specificity, making them reliable tools for virtual screening and further experimental validation [8, 10].

Virtual Screening in Chemical Databases

A virtual screening of 101,097 compounds from PubChem identified potential AChE inhibitors:

1) Screening execution: Each descriptor set was applied using the best performing QSAR models (SVM, MLP, RF, TensorFlow). The screening prioritized compounds within the applicability domain (AD) for higher reliability.

2) Consensus analysis: Integrating the predictions of all models provided a consensus of active compounds. For example:

- Morgan descriptors identified 6,455 consensus hits.
- RDKit descriptors produced 3,773 consensus hits.
- SiRMS descriptors produced 3,629 consensus hits.

3) Similarity search: Tanimoto similarity [20] was used to identify compounds structurally similar to known AChE inhibitors (e.g. Tacrine - Table 1). A total of 5,837 compounds showed significant similarity, narrowing down the potential candidates.

Table 1. Similarity consensus.

ID	CID	Canonical SMILES	Nearest Neighbor	Similarity (%)
0	165748451	CC(C)(C)c1ccccc1C(=O)C(F)F	Donepezil	0.262530
1	126973612	CC(C)c1ccccc1C(=O)C(F)F	Donepezil	0.253333
2	118729284	CN(CCCCCCN1C(=O)c2ccccc2C1=O)Cc1ccccc1	Donepezil	0.314754
3	21994169	O=C1NC(=O)c2c(CCCN3CCC(Cc4ccccc4)CC3)cccc21	Donepezil	0.385714
4	22132546	Nc1ccc2c(c1)CN(CCCCN1C(=O)c3ccccc3C1=O)CC2	Donepezil	0.347181
5	12004040	c1ccc2ncc(Nc3ncnc4c3CCN(CC3CCCCC3)C4)cc2c1	Galantamine	0.412822
6	54542240	O=C1NC(=O)c2c(CCCN3CCc4ccccc4C3)cccc21	Donepezil	0.369673
7	54403061	O=C1NC(=O)c2c(CCCCN3CCc4ccccc4C3)cccc21	Donepezil	0.395980
8	60259671	CC1CCCN(Cc2ccc(CNC(=O)c3ccc4c(c3)C(=O)NC4=O)cc...	Donepezil	0.332971
9	119536775	O=C(NCCC1CCNC1)c1cccc(CN2C(=O)c3ccccc3C2=O)c1	Donepezil	0.322513
10	66587765	CCN(CC)c1cccc(-c2cc(C(=O)NC3CCCc4ccccc43)ccn2)c1	Galantamine	0.399876
11	22588138	CCCCC(CC)CNC(=O)CCCCCn1c(=O)[nH]c2ccccc2c1=O	Galantamine	0.326582
12	120179720	CNCCC1CCN(C(=O)c2ccc3c(c2)CCC(=O)N3)CC1	Donepezil	0.380859
13	17956492	Cc1cc(N(C)C(=O)NCCN2CCC(Cc3ccccc3)CC2)c2ccccc2n1	Tacrine	0.465433
14	647903	CCCCc1nc2ccccc2c(NC(=O)CN2CCN(C)CC2)c1CCC	Tacrine	0.714721
15	1099160	CCCc1nc2ccccc2c(NC(=O)CNC2CCCCC2)c1CC	Tacrine	0.674723
16	4218057	CCCc1nc2ccccc2c(NC(=O)C[NH+]2CCCCCC2)c1CC	Tacrine	0.675076
17	4990629	CCCc1nc2ccccc2c(NC(=O)C[NH+]2CCCCC2)c1CC	Tacrine	0.677126
18	6966754	CCCc1nc2ccccc2c(NC(=O)C[NH2+]C2CCCCC2)c1CC	Tacrine	0.674723
19	133412317	CCCc1cc(NCCC2CCN(C(C)=O)CC2)c2ccccc2n1	Tacrine	0.548712
20	55964361	CC1CC(C)CN(Cc2ccc(CNC(=O)C=Cc3cccnc3)cc2)C1	Donepezil	0.302455
21	119438861	CCNCc1ccccc1NC(=O)C1CCCN(C(=O)c2ccncc2)C1	Donepezil	0.331806
22	121108747	CCCc1ccc(C(=O)NCC2CCN(c3ccncc3)CC2)cc1	Donepezil	0.307607
23	14783862	Cc1ccccc1C(=O)NCCC1CCN(Cc2ccccc2)CC1	Donepezil	0.345588
24	56396266	CC1CC(C)CN(Cc2ccc(CNC(=O)c3cccc(F)c3F)cc2)C1	Donepezil	0.309979
25	38401175	Cc1ccc(F)cc1C(=O)NCc1ccc(CN2CCC(C)CC2)cc1	Donepezil	0.332594
26	46465375	Cc1ccccc1C(=O)NCC(=O)NCc1ccc(CN2CCCC(C)C2)cc1	Donepezil	0.331089
27	55714142	CNC(=O)c1ccc(C=CC(=O)NCc2ccccc2CN2CCCC(C)C2)cc1	Donepezil	0.329361
28	84422326	CN(C)CCCNC(=O)c1ccc(CNC(=O)c2ccc(C(C)(C)C)cc2)cc1	Donepezil	0.242925
29	95809500	O=C(c1ccncc1)N1CCC(c2cccc(Cc3ccccc3)n2)CC1	Galantamine	0.307245
30	95816347	Cc1cc(Cc2ccccc2)cc(C2CCCN(C(=O)c3ccncc3)C2)n1	Galantamine	0.348428
31	110249502	O=C(c1ccncc1)N1CCCC(c2cccc(Cc3ccccc3)n2)C1	Galantamine	0.346926
32	109236248	CC1CCCN(c2cncc(C(=O)NCCc3ccccc3F)c2)C1	Galantamine	0.317481
33	109227211	CC1CCN(c2cncc(C(=O)NCCc3ccccc3F)c2)CC1	Galantamine	0.311528
34	109103313	O=C(NCCC1=CCCCC1)c1cncc(C(=O)NCc2ccc(F)cc2)c1	Galantamine	0.255319
35	37027068	O=C(Nc1cccc(F)c1)C1CCN(C(=O)c2cccnc2)CC1	Donepezil	0.324111
36	46547516	O=C(NCc1cccnc1)C1CCCN(C(=O)Cc2ccccc2)C1	Donepezil	0.327818

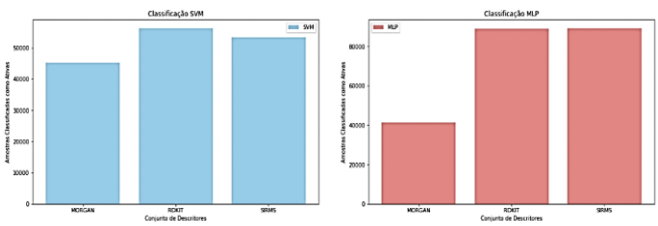
The analysis highlighted the complementary strengths of different molecular descriptors and machine learning models in predicting AChE inhibitory activity. These conclusions were drawn from performance metrics across models, which should be explicitly presented in tables and figures to reinforce these findings. Morgan descriptors, based on circular substructures, provided balanced performance across algorithms but identified fewer active compounds, indicating a more conservative approach in similarity-based screening. In contrast, RDKit descriptors, which encode a broad set of physicochemical properties such as hydrogen bond donors, rotatable bonds, and topological polar surface area, excelled at capturing chemical nuances, resulting in higher accuracy and broader identification of active compounds. SiRMS descriptors, which rely on a fragment-based approach to represent molecular structures at multiple levels, showed strong predictive power, although the consensus rates between models were moderate. This suggests that while fragment-based descriptors enhance prediction capabilities, their variability may introduce inconsistencies across

different machine learning models. To further strengthen these conclusions, it is essential to explicitly reference the corresponding performance data in tables and graphical visualizations.

The integration of QSAR modeling with well-curated descriptor selection and virtual screening proved to be a highly effective strategy for identifying potential drug candidates. Permutation testing confirmed the statistical robustness of the models, supporting their scalability for large-scale chemical screening and their potential application in Alzheimer's disease drug discovery.

In our study, machine learning and deep learning models-SVM, MLP, RF, and TensorFlow-were applied to three descriptor sets (Morgan, RDKit, and SiRMS) to classify compounds as active or inactive. Key findings include:

- The models exhibited variable performance across different descriptor sets. For instance, Morgan descriptors resulted in fewer samples classified as active, while RDKit descriptors identified a larger number of active compounds.
- Each model performed differently depending on the descriptor set, highlighting the importance of selecting appropriate models tailored to specific descriptor sets.
- Across all descriptor sets, a significant number of samples were consistently classified as active by all models. This consistency suggested the robustness of these samples and their potential significance.
- The choice of descriptors proved to be critical to model performance, as each descriptor set captured chemical information uniquely, leading to distinct results.
- The trained models and obtained results demonstrated potential applications in chemical compound screening, drug discovery, and pharmaceutical research, where accurate classification of compounds as active or inactive is essential (Figure 2).
- While the models performed well on internal validation data, external validation using independent datasets was necessary to truly assess their robustness and generalizability.
- The application of these models in the pharmaceutical and chemical industries required careful consideration of ethical and safety factors to ensure that compounds identified as active were both safe and effective.



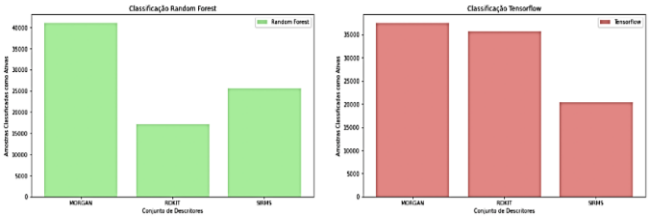


Figure 2. Classification using SVM, MLP, RF and TensorFlow

Future directions should focus on refining consensus modeling techniques to improve predictive reliability, incorporating diverse datasets for external validation, and exploring novel descriptors or hybrid modeling approaches to better capture complex molecular interactions. In addition, the integration of advanced deep learning frameworks with explicable AI could shed light on the molecular determinants of AChE inhibition and further advance drug discovery efforts.

In summary, this study demonstrated the potential of machine learning and deep learning models to enhance the screening and classification of chemical compounds. Careful selection of descriptors and rigorous validation processes were identified as critical steps in using these models for practical applications in pharmaceutical and chemical research. Continued advances in these areas will play an important role in accelerating the discovery of effective treatments for neurodegenerative diseases.

DISCUSSION

Analysis of the results highlighted several important points. The permutation tests performed provided insight into the statistical significance of the models' performance, with p-values generally below 0.05, indicating that the observed results were unlikely to have occurred by chance. Performance varied significantly between models and descriptor sets. For example, the SVM model achieved high accuracy with RDKit descriptors, while MLP performed best with Morgan descriptors. The inclusion of the Applicability Domain (AD) concept, which defines regions within the feature space where predictions can be considered reliable, played a critical role. Models outside the AD threshold were considered unreliable and excluded from further analysis [21].

The inclusion of AD had a significant impact on the performance metrics of the algorithms, improving sensitivity and specificity by an average of 20-25%. Models such as RF and SVM benefited the most from this improvement, with performance gains of 25-30%, demonstrating the importance of AD in optimizing model reliability within defined ranges of applicability. These results underscore AD as a critical tool for ensuring the accuracy and reliability of predictions,

especially in domain-specific contexts [22].

TensorFlow also demonstrated robust performance, achieving metrics comparable to or better than RF, SVM, and MLP, with average improvements of 10-15%. This highlighted TensorFlow as a competitive choice for classification tasks. In addition, the inclusion of AD further improved TensorFlow's sensitivity and specificity by 15-20%, underscoring its ability to accurately identify relevant samples within AD [17].

The models demonstrated their ability to effectively classify large datasets when applied to a large external database of 101,097 samples. The consensus approach, which integrates predictions from all four algorithms (SVM, MLP, RF, and TensorFlow) across three descriptor sets, identified compounds that were consistently classified as active by all models (Table 2) [23].

Table 2. Model consensus (number of compounds).

Descriptor Set	SVM	MLP	RF	TensorFlow	AD Consensus
Morgan	45.152	41.198	41.060	37.505	6.455
RDKit	56.229	89.058	17.183	35.748	3.773
SiRMS	53.447	89.156	25.636	20.438	3.629

This consensus method proved critical in highlighting reliable and consistent results. In addition, similarity searches using the Tanimoto coefficient identified compounds with structural similarity to reference compounds such as Tacrine, expanding the scope for discovering novel drug candidates [20].

The practical implications of these findings are significant. The machine learning and deep learning models developed in our study demonstrated their value in large-scale compound screening, reducing the time and resources required by prioritizing promising compounds for subsequent laboratory testing [23]. The similarity search strategy using the Tanimoto coefficient proved valuable in identifying chemical compounds with common features to reference compounds, providing a robust tool for pharmaceutical research and medicinal chemistry [20].

In summary, incorporating AD significantly improved model performance, with average gains of 20-25% in sensitivity and specificity. TensorFlow proved to be a strong alternative to traditional algorithms, achieving consistent results with 10-15% improvements in metrics. Consensus and similarity search methods further improved the reliability of results and aided in the identification of new compounds with potential pharmacological activity. These methods hold great promise for accelerating the drug discovery process by enabling more efficient prioritization of candidate compounds for experimental validation.

CONCLUSION

This study investigated the application of machine learning and deep learning models to classify chemical compounds as active or inactive within their range of applicability, using three different sets of molecular descriptors: Morgan, RDKit, and SiRMS. A variety of models, including SVM, MLP, RF, and TensorFlow, were trained and validated for each descriptor set, providing valuable insight into their performance and potential applications.

The results showed that the performance of the models varied depending on the descriptor set. For example, models using Morgan descriptors classified fewer compounds as active, while those using RDKit descriptors identified a greater number of active compounds. Each model also exhibited different performance depending on the descriptor set, highlighting the need for careful selection of models suited to specific types of descriptors. Across all descriptor sets, certain compounds were consistently classified as active by all models, highlighting their robustness and potential importance for further exploration.

The study highlighted the critical role of descriptor selection in determining model performance. Different descriptors captured unique aspects of chemical information, leading to different results. This highlighted the importance of aligning descriptor selection with the specific goals of a given application. The trained models and insights gained from this study demonstrated promising applications in chemical compound screening, drug discovery, and pharmaceutical research, where accurate classification of active and inactive compounds is essential.

While the models performed well on internal validation data sets, external validation using independent data sets was necessary to truly assess their robustness and generalizability. Internal datasets refer to the data used during model training and validation, typically derived from the same source or preprocessed under similar conditions. These datasets help optimize model performance but may not fully represent the diversity of real-world chemical space. In contrast, external datasets consist of completely independent data, often obtained from different sources or experimental conditions, and serve as a critical benchmark for evaluating how well the model generalizes to unseen compounds. In addition, application of these models in industrial contexts, such as pharmaceutical or chemical development, required attention to ethical and safety considerations to ensure that the identified compounds were both effective and safe.

In conclusion, this research demonstrated the potential of machine learning and deep learning models to improve the screening and classification of chemical compounds in large datasets. However, careful selection of descriptors and models remains fundamental, as does the need for external

validation prior to practical application. Continued advances in the field are expected to further improve the accuracy and efficiency of these models, paving the way for significant contributions to the fields of chemistry and pharmacology.

Acknowledgements

The authors would like to express their gratitude to everyone who contributed to this research, including colleagues, mentors, and institutions that provided support throughout the study. Special thanks go to family and friends whose encouragement and inspiration have been invaluable.

This study was also driven by a deeply personal motivation. One of the authors has a father who suffers from Alzheimer's disease, which inspired the pursuit of new and effective treatments. Beyond being a scientific endeavor, this research represents a personal commitment to contributing to advancements that could one day improve the lives of patients and their families affected by this devastating condition.

REFERENCES

[1]. Dai, R., et al., 2022, Anti-Alzheimer's disease potential of traditional Chinese medicinal herbs as inhibitors of BACE1 and AChE enzymes. *Biomedicine & Pharmacotherapy*, 154, 113576.

[2]. Hu, Y., et al., 2019, Identify compounds' target against Alzheimer's disease based on in-silico approach. *Current Alzheimer Research*, 16(3), 193–208.

[3]. Das, S., Chakraborty, S., and Basu, S., 2019, Hybrid approach to sieve out natural compounds against dual targets in Alzheimer's disease. *Scientific Reports*, 9, 3714.

[4]. Dhamodharan, G., and Mohan, C. G., 2022, Machine learning models for predicting the activity of AChE and BACE1 dual inhibitors for the treatment of Alzheimer's disease. *Molecular Diversity*, 26, 1501–1517.

[5]. Neves, B. J., et al., 2018, QSAR-based virtual screening: Advances and applications in drug discovery. *Frontiers in Pharmacology*, 9, 1275–1282.

[6]. Sharma, S., and Sharma, D., 2018, Intelligently applying artificial intelligence in chemoinformatics. *Current Topics in Medicinal Chemistry*, 18(20), 1804–1826.

[7]. Chen, H., Kogej, T., and Engkvist, O., 2018, Cheminformatics in drug discovery, an industrial perspective. *Molecular Informatics*, 37(9–10), e1800041.

[8]. Carpenter, K. A., and Huang, X., 2018, Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: A review. *Current Pharmaceutical Design*, 24(28), 3347–3358.

[9]. Tropsha, A., et al., 2017, Predictive QSAR modeling: Methods and applications in drug discovery and chemical risk assessment. In: *Handbook of Computational Chemistry* (Springer), pp. 1–48.

[10]. Xue, L., and Bajorath, J., 2000, Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial Chemistry and High Throughput Screening*, 3, 363.

[11]. Saha, B., et al., 2024, Identification of coumarin derivatives targeting acetylcholinesterase for Alzheimer's disease by field-based 3D-QSAR, pharmacophore model-based virtual screening, molecular docking, MM/GBSA, ADME and MD Simulation study. *Current Research in Structural Biology*, 7, 100124.

[12]. Walczak-Nowicka, L. J., and Herbet, M., 2021, Acetylcholinesterase inhibitors in the treatment of neurodegenerative diseases and the role of acetylcholinesterase in their pathogenesis. *International Journal of Molecular Sciences*, 22(17), 9290.

[13]. Jang, C., et al., 2018, Identification of novel acetylcholinesterase inhibitors designed by pharmacophore-based virtual screening, molecular docking and bioassay. *Scientific Reports*, 8, 14921.

[14]. James, G., et al., 2017, *An Introduction to Statistical Learning with Applications in R*, 8th ed., Springer Texts in Statistics.

[15]. Horvath, T., and Aldahdooh, J., 2017, Investigating the importance of meta-features for classification tasks in meta-learning. Ph.D. dissertation, Faculty of Informatics, Eötvös Loránd University, Budapest.

[16]. Quadri, T. W., et al., 2022, Multilayer perceptron neural network-based QSAR models for the assessment and prediction of corrosion inhibition performances of ionic liquids. *Computational Materials Science*, 214.

[17]. Abadi, M., et al., 2016, Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *Distributed, Parallel, and Cluster Computing*, 1.

[18]. Wu, J., et al., 2019, Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1).

[19]. Ojala, M., and Garriga, G. C., 2010, Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11, 1833–1863.

[20]. Maggiora, G., et al., 2014, Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), 3186–3204.

[21]. Alamro, H. M., et al., 2023, Exploiting machine learning models to identify novel Alzheimer's disease biomarkers and potential targets. *Scientific Reports*, 13, 4979.

[22]. Sushko, I., 2011, Applicability domain of QSAR models. M.S. thesis, Chair of Genome-Oriented Bioinformatics, Technical University of Munich, Ludwig Maximilian University of Munich.

[23]. Bao, L. Q., et al., 2023, Development of activity rules and chemical fragment design for in silico discovery of AChE and BACE1 dual inhibitors against Alzheimer's disease. *Molecules*, 28(8), 3588.