universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Machine-learning based approaches for predicting cholinesterase inhibition

verfasst von | submitted by

Mariia Demidova BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Magistra pharmaciae (Mag. pharm.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 605
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Pharmazie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker

Acknowledgments

The past semester has been a highly challenging, but also an extremely valuable time period, offering me an introduction to the field of pharmacoinformatics.

Firstly, I would like to express my sincere gratitude to my supervisor Univ.-Prof. Mag. Dr. Gerhard Ecker for his constant guidance and feedback. His expertise in this area and insights have been crucial in navigating the challenges of this thesis.

Also, I would like to thank the Pharmacoinformatics Research Group for their help, but also moral support throughout the practical parts of this work.

Lastly, I am deeply grateful to my parents for their unwavering support and encouragement, especially during the most difficult moments. It would have never been possible for me to come this far without them.

Table of Contents

Abstract

The goal of this thesis was to build robust computational methods for predicting the activity of potential inhibitors targeting human acetylcholinesterase (hAChE), human butirylcholinesterase (hBChE) and pacific ray acetylcholinesterase (TcAChE). These enzymes are critical to neurological function: hAChE primarily regulates cholinergic neurotransmission and is a key target in treating nerological disorders like Alzheimer's disease, while hBChE plays a complementary role that becomes increasingly significant as the disease progresses. Due to its availability and high himology to the human enzyme, TcAChE has been frequently used as a model for hAChE inhibition studies.

Regression models were built for each enzyme using a diverse set of molecular features, including fingerprint-based and physichochemical descriptors to capture the diversity of the compounds. Additionally, structural information on key binding site residues was integrated into a proteochemometric model to enhance the predictive power by also considering the enzyme-ligand interactions. Finally, a dual-output model was developed and trained on shared molecular features of compounds with experimentally determined activities against both hAChE and hBChE.

The bioactivity data for the three targets was retrieved from the assays available in the ChEMBL34 database. While all the models have demonstrated a reasonable predictive performance, further validation on larger datasets would be necessary to evaluate the generalizability of the model to unseen compound data and assess its applicability in real-world settings.

Zusammenfassung

Das Ziel dieser Masterarbeit war es, robuste in-silico Methoden zur Vorhersage der Aktivität potenzieller Inhibitoren zu entwickeln, die auf humane Acetylcholinesterase (hAChE), humane Butyrylcholinesterase (hBChE) und die Acetylcholinesterase des kalifornischen Zitterrochens (TcAChE) abzielen. Diese enzyme sind essenziell für die neurologische Funktion: hAChE reguliert primär die cholinerge Neurotransmission und ist ein zentraler Angriffspunkt für die Behandlung neurologischer Erkrankungen wie Morbus Alzheimer, während hBChE eine komplementäre Rolle spielt, die mit dem Fortschreiten der Krankheit zunehmend an Bedeutung gewinnt. Aufgrund ihrer Verfügbarkeit und hohen Homologie zum humanen Enzym wird TcAChE häufig als Modell für Inhibitionsstudien zur hAChE verwendet.

Für jedes Enzym wurden Regressionsmodelle unter Verwendung einer Vielzahl molekularer Deskriptoren entwickelt, einschließlich auf fingerprint-basierender und physikochemischer Deskriptoren, um die Vielfalt der molekularen Strukturen abzubilden. Zusätzlich wurde strukturelle Information über die wichtigsten Aminosäuren der Bindungsstelle in ein proteochemometrisches Modell integriert, um die Vorhersagekraft zu verbessern, indem auch die Enzym-Ligand-Interaktionen berücksichtigt wurden. Schließlich wurde ein Dual-Output-Modell entwickelt und auf gemeinsamen molekularen Merkmalen von Verbindungen trainiert, deren experimentelle Aktivitäten sowohl gegen hAChE als auch gegen hBChE bestimmt wurden.

Die Bioaktivitätsdaten für die drei Enzyme wurden aus den in der ChEMBL34-Datenbank verfügbaren Assays abgerufen. Während alle Modelle eine vernünftige Vorhersageleistung zeigten, wäre eine weitere Validierung auf größeren Datensätzen erforderlich, um die Generalisierbarkeit des Modells auf bisher nicht gesehene Verbindungsdaten zu bewerten und seine Anwendbarkeit in realen Szenarien zu prüfen.

1. Introduction

1.1. Alzheimer's Disease

Alzheimer's disease (AD) is one of the most prevalent health issues affecting the aging population. It accounts for over two-thirds of dementia cases in individuals over 65, making it the most common form of dementia, and affecting more than 50 million people worldwide. This number is projected to keep growing, with estimates suggesting it could reach up to 152 million cases by 2050. AD imposes a heavy burden on the patients and their families, charazterized by a progressive loss of cognitive function and a diminishing quality of life. (Breijyeh & Karaman, 2020)

The exact mechanisms underlying AD remain unclear, and currently there is no cure or viable preventive treatment. The drugs currently available on the market are limited to symptomatic treatment without the ability to alleviate the disease or slow down the progression. As of now, only two classes of small molecule drugs have been approved by the FDA for AD treatment: acetylcholinesterase (AChE) inhibitors and N-methyl d-aspartate (NMDA) receptor antagonists. (Zuliani et al., 2024) For over twenty years, no further treatments have been approv
ed, with the exception of the two monoclonal antiboduies Aducanumab and Lecanemab. These antibodies were designed to slow down the disease progression, but their clinical implementation has been highly controversial due to limited efficacy and considerable side effects. Consequently, AChE inhibitors remain as the dominant class of drugs used in AD treatment. (Espay et al., 2024)

A key feature of AD is the cholinergic deficit in the brain. AChE plays a vital role in regulating acetylcholine (ACh) levels within both the peripheral and central nervous systems. Therefore, drugs targeting AChE have been developed to restore normal ACh levels and alleviate some symptoms of AD. AChE, an enzyme belonging to the serine hydrolase family, plays a critical role in neurotransmission, terminating the signal by cleaving acetylcholine into acetic

acid and choline in the synaptic cleft. (Akhoon et al., 2020) Recently, pseudocholinesterase or butyrylcholinesterase (BChE) has been gaining increased attention in regards to its role in the AD pathology. Like AChE, BChE is also present in the brain and can hydrolyse ACh. While BChE also hydrolyzes ACh, it differs from AChE in its widespread expression in the body and broader substrate range, including the metabolism of substances such as aspirin and other xenobiotics. Although BChE hydrolyzes ACh more slowly in healthy individuals, it plays a more significant role in ACh hydrolysis during the later stages of AD, when ACh levels increase dramatically. (Mushtaq et al., 2014)

1.2. Cholinesterase family

1.2.1. Structures

AChE exhibits an exceptionally high turnover rate, nearing that of a diffusion-controlled reaction. This extraordinary efficiency has made the enzyme a subject of extensive structural research. The first high-resolution crystal structure of AChE, obtained in 1991 from an electric ray species *Torpedo californica* (TcAChE), enabled the identification and first-time visualization of its binding pocket at an atomic level. Subsequent studies of TcAChE crystal structures in complexes with various inhibitors has allowed further insights into the enzyme structure and binding mode. The choice of TcAChE as the first reliable model system was likely driven by the enzyme's abundance in the electric organs of the ray, as well as a high homology with hAChE (Dvir et al., 2010). Notably, the binding pocket was located at the bottom of a narrow gorge around 20 Å deep, lined with 14 conserved aromatic residues that make up most of the gorge surface (Silman & Sussman, 2008).

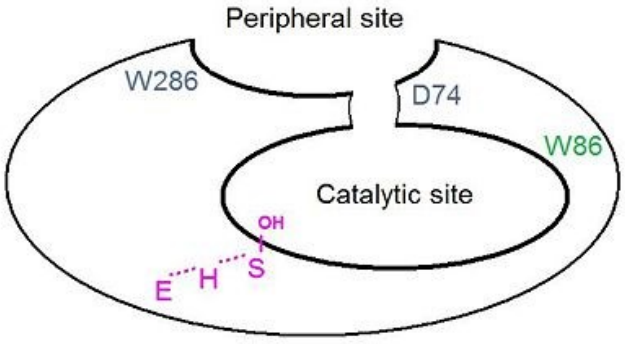


Figure 1: Schematic view of the AChE catalytic and peripheral sites (Proteopedia, 2023)

Despite the high degree of homology between TcAChE and hAChE, there are notable structural differences between the two. In hAChE, there is an additional peptide sequence that partially obstructs the entrance into the gorge. It was later discovered that the stabilization of the ligand is partially facilitated by a ‘breathing’ movement of the peptide chains, as well as a stabilization at the peripheral anionic site (PAS) rich in aromatic amino acids. Located close to the mouth of the gorge, it interacts with the substrate and guides it into the catalytic anionic site (CAS) (Fig.1). The CAS itself comprises two sub-sites: an ‘esteratic’ and an ‘anionic’ site. The esteratic side consists of the catalytic triad characteristic of other serine hydrolases (E334, H447, S203), while the anionic site specifically interacts with the charged quaternary ammonium group of choline (Silman & Sussman, 2008).

Human BChE, although structurally similar to hAChE and sharing the same catalytic triad in the esteratic side, differs in its structure and function. The active site gorge of BChE has a bowl-like shape and is significantly wider than in AChE. The BChE active site gorge has a broader, bowl-like shape, and its surrounding residues, including those in the PAS, differ notably from those in AChE. BChE has approximately 40% fewer aromatic residues than AChE. (Silman & Sussman, 2017).

1.2.2. Inhibitors

A deeper structural understanding of cholinesterases is crucial for the development of effective inhibitors. The existing cholinesterase inhibitors are a very diverse group, ranging from medicinal drugs to pesticides and nerve agents. Those inhibitors can be broadly categorized as reversible, pseudo-irreversible or irreversible. Most pharmacologically active drugs, like donepezil, galantamine or rivastigmine fall into the reversible or pseudo-irreversible categories (Fig.2), while irreversible inhibitors, including organophosphates like sarin and parathion, bind covalently to the serine residue within the catalytic triad and are primarily used as pesticides or nerve agents (Moss & Perez, 2021). In order to enhance the binding

affinity to the enzyme, attempts have been made at designing dual-binding inhibitors that would bind to both CAS and PAS simultaneously. The development of such inhibitors started with amalgamating tacrine with other pharmacophores to be able to bridge the two sites (Alonso et al., 2005).

The currently available cholinesterase inhibitors are short-acting and target primarily AChE, which limits their efficacy and creates a dose-dependent gastrointestinal toxicity that limits their application to suboptimal dosages. Consequently, more and more current research is starting to consider BChE as a target and also focus on developing and testing more selective, irreversible inhibitors (Moss, 2020).

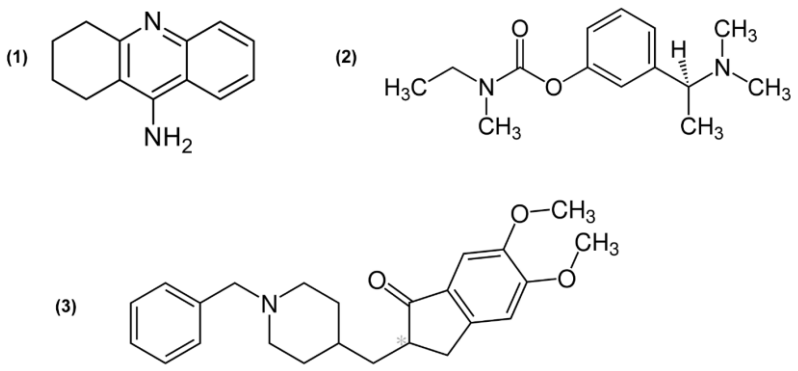


Figure 2: Some of the FDA-approved AChE-inhibitors. (1) Tacrine, (2) Rivastigmine and (3) Donepezil

1.3. CADD in AD therapy

For over twenty years, no new AD drugs have been introduced to the market, with the exception of Aducanumab and Lecanemab monoclonal antibodies that were designed to slow down the disease progression. However, while displaying some minor efficacy, the side effects have been considerable and have made the implementation of the drugs in current therapies highly debatable (Salman et al., 2021). This situation highlights the urgent need for innovative approaches in drug development.

The drug development process is notoriously lengthy and expensive, and is riddled with high failure rates, particularly in Phase II and III trials. Recent advancements like high-throughput-screening (HTS) have significantly sped up the early-stage drug discovery by enabling the rapid screening of vast chemical libraries. This, coupled with the accumulation of large volumes of experimental data as well as advancements in available computational power has given rise to computer-aided-drug design (CADD). By leveraging data from the vast chemical databases, CADD aims to effectively reduce the immense cost and time strain in all stages of drug development, from target identification up to preclinical studies. It is estimated that CADD could potentially cut drug development costs by up to 50% (Salman et al., 2021).

The backbone of many ligand-based CADD strategies is quantitative-structure-activity-relationship (QSAR) modelling. These models utilize biological data from databases such as ChEMBL, PubChem or UniProt to establish correlations between molecular descriptors and biological endpoints, such as bioactivity. QSAR models can be broadly divided into classification and regression models, with the latter being particularly critical for predicting continuous bioactivity values. The effectiveness of QSAR models depends on several factors, including the quality of the input data to the choice of molecular descriptors and the evaluation metrics that assess the model performance. (Zhao et al., 2020)

1.4. QSAR models of Cholinesterases

Numerous QSAR models have been developed for cholinesterases, with the majority focusing on acetylcholinesterase. However, many of these models were trained on small, homogenous datasets containing only a few classes of compounds, such as models for piperidine or N-benzylpyrrolidine derivatives. (Chekmarev et al., 2009; El Khatabi et al., 2021)

A recent comprehensive study by Vignaux et al. aimed to develop robust classification and regression models for AChE from nine different species. Using data from ChEMBL30, BindingDB, and Tox21, this study included 3652 compounds in the hAChE regression models. While primarily focused on

classification, the study achieved strong regression performance, with the best model, a Support Vector Regression algorithm, yielding a training R^2 of 0.81 and a test R^2 of 0.58. (Vignaux et al., 2023)

Building upon these efforts, this thesis aims to expand the scope of regression models for cholinesterases by developing robust regression models for hAChE, hBChE and TcAChE based off large, structurally diverse datasets. These models aim to explore existing limitations and provide deeper insights into the bioactivity predictions across multiple cholinesterase targets.

2. Methods

2.1. Software and databases

2.1.1. ChEMBL

ChEMBL is a large-scale, manually curated chemical database of bioactivity data for small molecules, managed by EMBL-EBI. It provides information about FDA-approved drugs, compounds undergoing clinical trials, bioactivity assays and molecular targets. ChEMBL assigns unique identifiers, ChEMBL IDs, to compounds, and allows to distinguish between different salt complexes and stereoisomers. The data can be accessed via a web interface, ChEMBL API, or by downloading the full database in various formats (Zdrazil et al., 2024).

For this thesis, ChEMBL version 34 was used. A manual search through the database for bioactivity assays for hAChE, hBChE and TcAChE was conducted using the built-in search bar. The unfiltered bioactivity data was downloaded directly in CSV format for subsequent processing.

2.1.2. Protein Data Bank

The Protein Data Bank (PDB) is a publically accessible archve for 3D structural data for macromolecules, managed by the Worldwide Protein Data Bank (wwPDB). The database contains structures stemming from experimental

methods like X-Ray crystallography, NMR spectroscopy or cryo-electron microscopy. Each structure is assigned a unique four-character PDB ID for easy identification and referencing (Fig.3). PDB also offers an advanced search system that allows users to filter data based on resolution, organism of origin, mutations and other specific parameters.

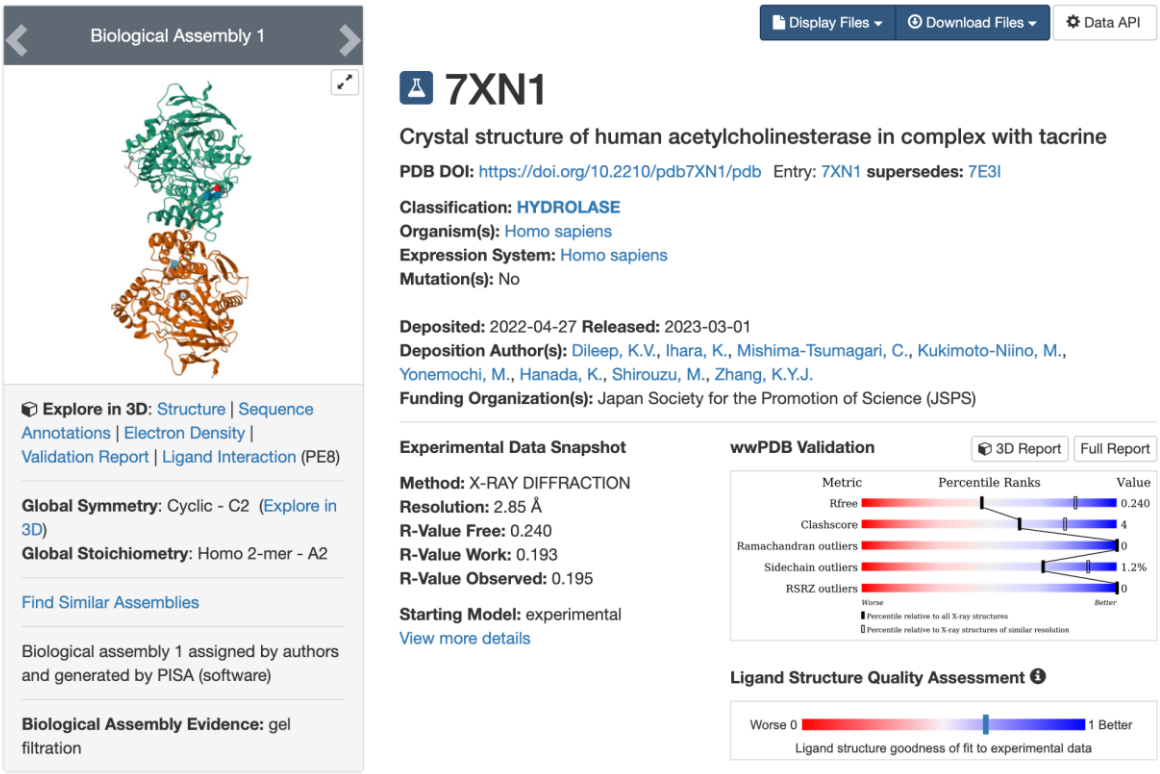


Figure 3: An example of a PDB entry

2.1.3. KNIME

KNIME, the Konstanz Information Miner is a no-code, open-source tool for data analytics and data science. It operates using a node-based workflow system, with each node being configurable to perform a specific action on the data. KNIME supports a wide range of extensions for chemoinformatics, including RDKit or CDK for handling molecular data.

KNIME also enables a to run Python scripts within the workflows with a Python Script (Labs) node. For this thesis, KNIME 5.3.0 was used with an Anaconda environment set up to be used both within the KNIME integration to maintain all of the preprocessing steps within a single workflow.

2.1.4. Python

The machine learning models were built in Python 3.6.13, within the same Anaconda environment integrated with KNIME. The key libraries in this environment included:

- **RDKit:** An open-source chemoinformatics software that also allows to use it's functionality in Python. It offers a wide range of functions for handling chemical structures, molecular descriptor calculation and chemoinformatics tasks.
- **Mordred:** An additional descriptor calculation library that allows to calculate more than 1800 2D and 3D molecular descriptors.
- **Pandas and NumPy:** Two essential libraries for data manipulation and numerical computations.
- **Scikit-learn:** A machine learning model providing tools for model building, hyperparameter tuning, model validation and performance evaluation.

Additionally, CDDD package was used to generate Continuous Data-Driven Descriptors. Unlike classical 1D or 2D descriptors, these are vector embeddings mapped by processing the molecular structures with a pre-trained neural network model. Since this model relies on TensorFlow to work, a separate Anaconda environment with TensorFlow 1.10 without GPU was used for the task, as well as for building a separate neural network for the cholinesterase datasets. Most of the figures and visualizations were done using MatPlotLib and Seaborne libraries.

2.1.5. PyMol

PyMol is a widely used software that allows for 3D visualization, editing and structure analysis of both small molecules and proteins. It features an advanced graphical user interface, as well as a built-in command line functionality that uses internal text commands (Fig.4), which are all documented and can be referenced on the official PyMol website. For more advanced or automated workflows, PyMol allows the user to run their own Python scripts in the command line.

Since 2009, the PyMol license has been commercialized by Schroedinger so any new updates are no longer freely available, but the original open-source version is still accessible free of charge. For the purposes of this thesis, the open-source version 3.0 was used to handle protein data. PyMol allows to import PDB structures directly by typing the 'fetch' command into the command line along with the PDB ID of interest.

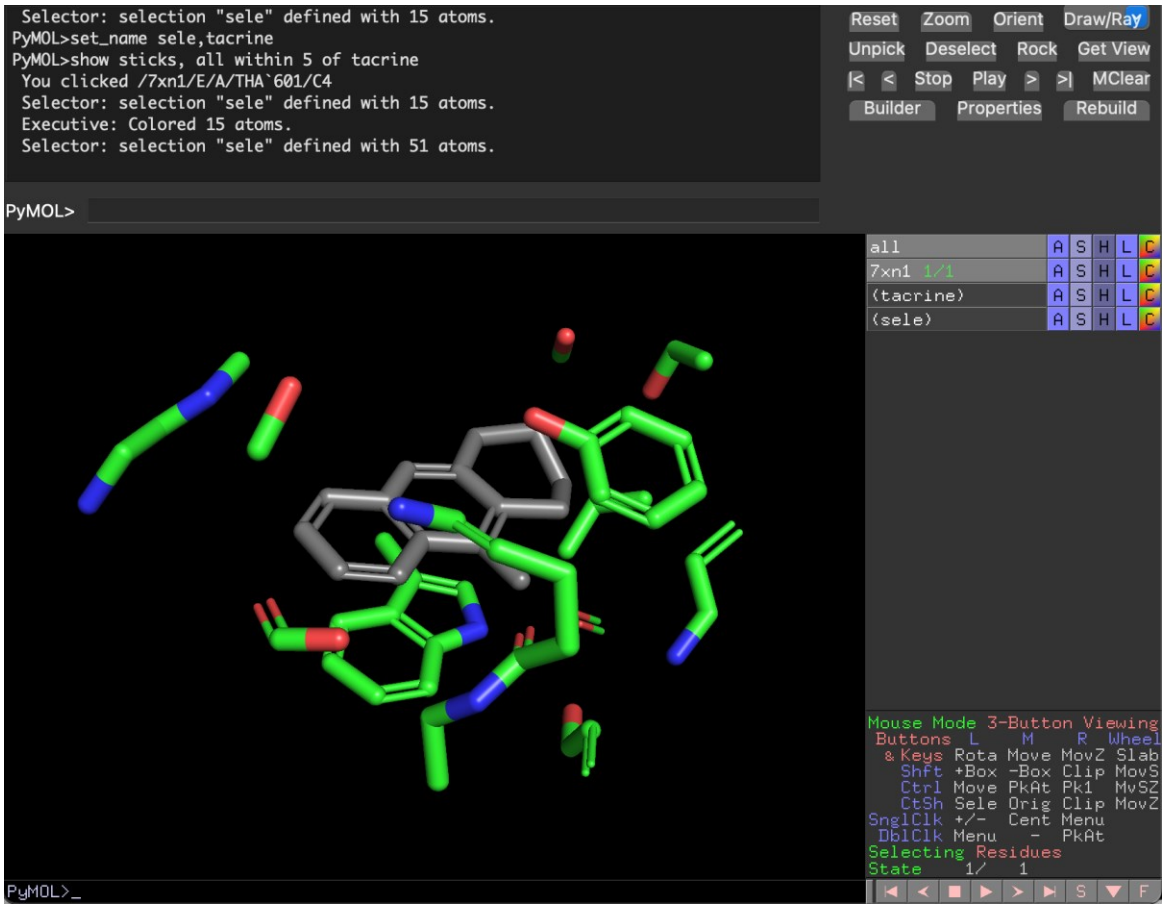


Figure 4: A demonstration of the PyMol interface

2.1.6. PaDEL-Descriptor

PaDEL-Descriptor is an open-source software for calculating a wide variety of molecular descriptors including topological, physichochemical as well as ten types of fingerprint-based descriptors. It is available in a graphical interface (Fig.5), as well as a PaDELpy Python library, and supports many types of input molecular file formats including SMILES, SDF and Mol2. Version 2.21 was used to calculate 881-bit PubMed fingerprint descriptors for the molecular datasets (Yap, 2011).

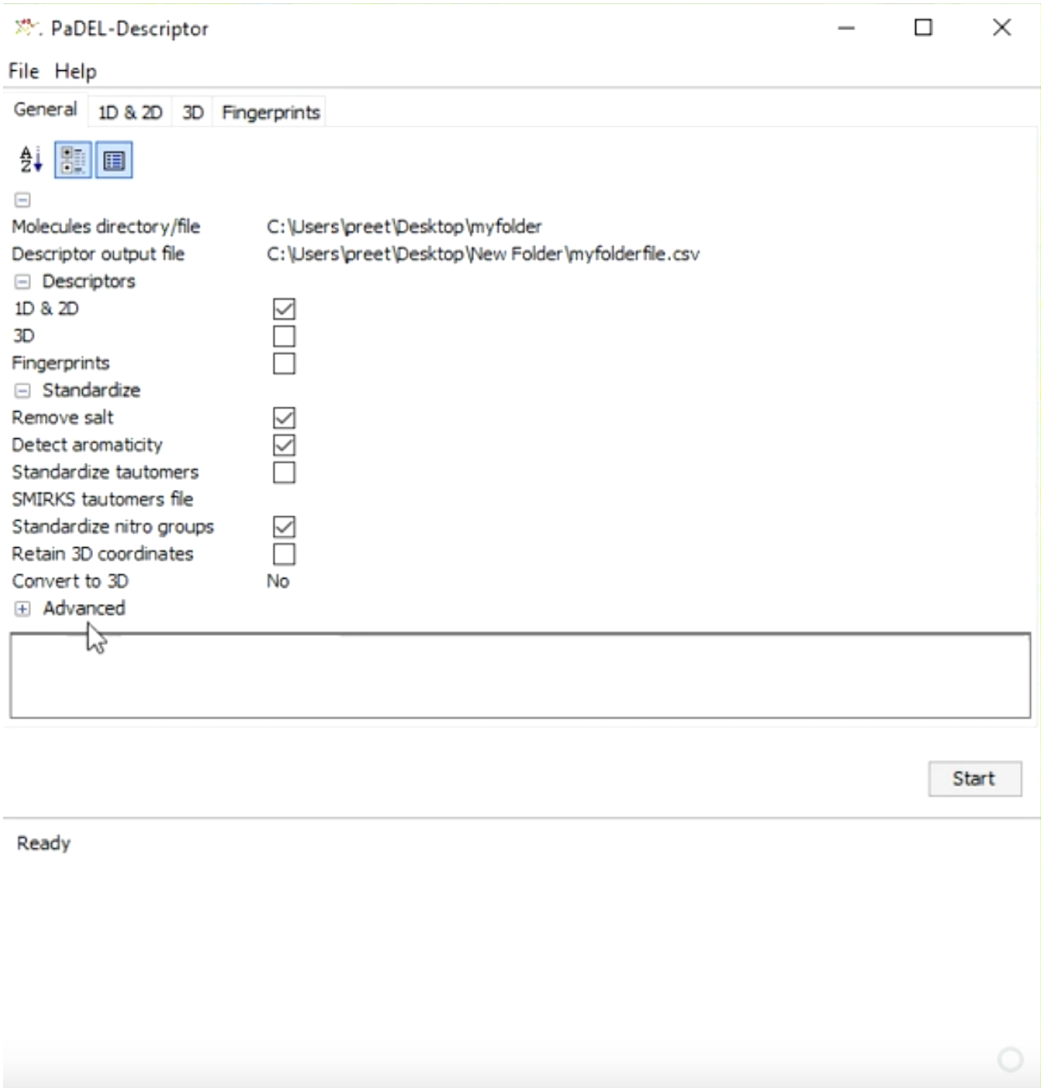


Figure 5: A demonstration of the PaDEL-Descriptor software interface

2.2. Building the datasets

2.2.1. Data collection

Bioactivity data for hAChE (CHEMBL220), hBChE (CHEMBL1914), and TcAChE (CHEMBL4780) was retrieved from ChEMBL version 34, forming the primary input for the regression models. The data was retrieved via the ChEMBL web interface, with no initial filters applied, as a rigorous curation process was performed later.

Additional bioactivity data was collected from PubChem for hAChE in order to validate performance on an external dataset unseen by the models during training. Only assays not originating from ChEMBL were included to avoid redundancy, yielding additional data from Tox21 and BindingDB.

2.2.2. Data curation

Before training the machine learning models, it is necessary to ensure that the data is clean, consistent and appropriate for agorithms to work effectively. The preprocessing of the ChEMBL datasets began with filtering the data to retain only entries with IC50 as their bioactivity metric and a pChEMBL value present. Entries with missing critical information, such as standard units, standard relation, standard type or SMILES were excluded. Data entries with standard relations other than '=' and those with standard values equal to zero were also removed (Fig.6).

For compounds with identical ChEMBL IDs, duplicates were handled based on the coefficient of variation (CV) between their pChEMBL values. Compounds with a CV greater than 20% were removed, while those with a CV of 20% or less were averaged to their mean to retain a single representative value.

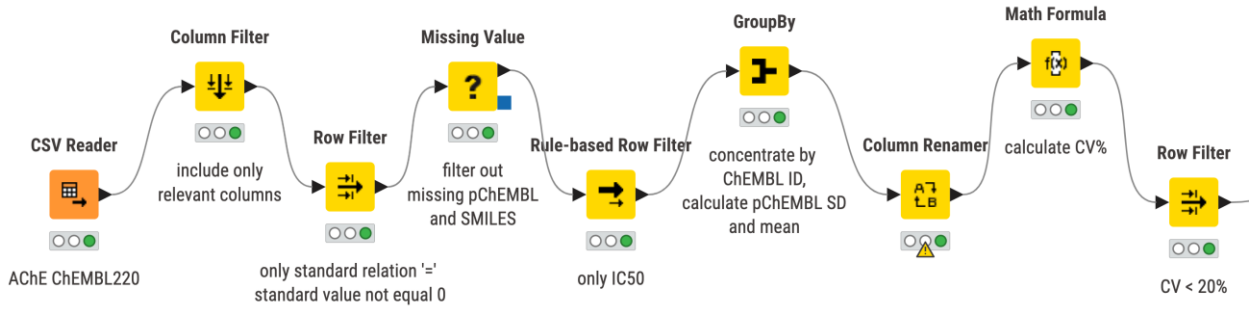


Figure 6: An excerpt from the preprocessing workflow in KNIME

From the validation datasets for hAChE collected from PubChem, only compounds flagged as 'active' were used. Since the Tox21 datasets included extrapolated activity values, filters were set to only keep the compounds with a full titration curve to avoid using unreliable data. The rest of the preprocessing for the validation datasets underwent the same steps as the primary datasets.

2.2.3. Standardisation

Following curation, compounds were standardized using a PharmInfo Vienna KNIME workflow based on the standardization protocol proposed by Francis Atkinson from EMBL-EBI. Mixtures of molecules, salts and molecules with nonorganic atoms were removed. For each target, two separate standardized datasets were saved: one retaining stereochemistry and hydrogens, and one without these features in order to then evaluate the importance of stereochemistry in the datasets. After standardisation, the duplicate removal procedure was repeated, but now using InChIKeys instead of ChEMBL IDs to handle identical structures arising from salt removal or stereochemistry standardization.

2.3 Molecular descriptors

Molecular descriptors are regularly used in QSAR studies as a representation of the physicochemical properties of small molecules. They are calculated from a chemical structure representation like SMILES and include, among others, 1D or

constitutional descriptors, which are based solely on the sequence information, and 2D descriptors which also consider the connectivity of a molecule. There are also 3D descriptors, which are more sensitive to the conformation of the molecules and require more preparation (Bahia et al., 2023).

Apart from continuous descriptors, the structural information of the molecules can also be represented by binary fingerprint descriptors. They usually come in form of bit strings that represent the presence or absence of certain molecular features with 1 or 0 respectively. Key-based fingerprints like 166-bit MACCS or 881-bit PubChem contain information about predefined substructures. On the other hand, circular fingerprints like Morgan or ECFP construct the hashed bitvectors by picking an initial identifier on one atom and then extending it to the neighbouring atoms by a certain radius. This radius usually extends to around 2 bonds, but can be modified if necessary. Such fingerprints usually differ in the properties that they use for the atom identifiers. While they are a widely used, powerful tool for representing the physicochemical information of molecules, they are limited in their interpretability since they cannot be easily decoded to deduce the original substructures they represent (Boldini et al., 2024).

A new way to represent the molecular structures emerged with continuous data-driven descriptors (CDDD). CDDD use a pretrained unsupervised neural network to encode molecular features in fixed-length vectors compressed and generated through deep learning models. Unlike classical 1D and 2D descriptors, CDDD adapt to the specific dataset presented to the model and can capture dataset-specific patterns. It has been demonstrated that models based on such descriptors can perform just as well as, and even sometimes outperform, the classical fingerprints (Winter et al., 2019).

For this thesis, six types of descriptors and fingerprints were calculated for the three cholinesterase datasets without stereochemical information (Table 1). MACCS and FeatMorgan fingerprints were calculated using the RDKit Fingerprint node in KNIME. For FeatMorgan, it was chosen to limit the vector length to 1024-bit as to limit the size of the descriptor space relative to the number of bioactivities. Mordred and RDKit were calculated using the respective Python

libraries, and PubChem fingerprints were generated with PaDEL-Descriptor software. The MACCS Keys and RDKit descriptors were used both separately and combined together.

Additionally, to investigate the impact of stereochemistry on predictive performance, continuous data-driven descriptors (CDDD) were computed twice for datasets with and without stereochemical information.

2.4. Individual QSAR models

2.4.1. Model building

Machine learning is routinely used in drug discovery to generate predictions for various crucial endpoints like toxicity, bioactivity and physicochemical properties. Classification models place the output variable into a discrete category like active or inactive, toxic or not toxic. On the other hand, regression models predict continuous outcomes, like pIC_{50}. Both approaches are viable and widely used depending on the specific objectives of the research.

An ideal model would be one that could capture the patterns in the training data while generalizing them well to unseen data. However, achieving this perfect balance is most often impossible, especially in the context of biological data since it is inherently noisy and complex due to variability and errors in experimental conditions and the nature of biological systems themselves. A key concept in building robust models is bias-variance tradeoff. Here, bias refers to the difference between the observed values and the predictions of the model, and variance refers to how sensitive the model is to the fluctuations within the training data. A model that is too simple would have a high bias and underfit the data, while a model that is too complex would have high variance and fit the training data perfectly, but generalize poorly. (Vabalas et al., 2019)

In order to try and develop robust QSAR models for each of the three targets that would be neither too simple nor too complex, on top of the different descriptor sets three different supervised regression algorithms were chosen: Elastic Net

regression, tree-based XGBoost regression, and Support Vector Regression (SVR). An overview of the models finds itself below:

- **Elastic Net:** A model built upon the multiple linear regression (MLR) algorithm. An MLR model simply calculates a linear equation from the given data by finding the intercept and assigning coefficients to the features, and fits a straight line through the data. Elastic Net adds two regularization parameters: L1 (Lasso) and L2 (Ridge). L1 helps remove useless features by shrinking some coefficients to zero, while L2 minimizes the sum of squared coefficients and stabilizes the model if there are many correlated features. Since it is a linear model, it does not represent nonlinear relationships very well. In this thesis, Elastic Net was primarily used as a baseline model for performance comparison.
- **XGBoost**: A decision-tree-based ensemble algorithm that uses gradient boosting to optimize model performance. Gradient boosting involves iteratively building a sequence of decision trees, where each following tree corrects the errors of the previous ones. XGBoost also includes regularization, parallel processing, and pruning to enhance efficiency and accuracy. While XGBoost is suitable for capturing nonlinear relationships in data, it is also prone to overfitting, especially on small or noisy datasets, so a careful hyperparameter tuning is necessary to control the model complexity.
- **SVR:** An algorithm that uses special kernel functions to map features into a higher-dimensional space, and finds a hyperplane that best fits the data within a specified tolerance margin (ϵ-tube). By focusing on points near the margin, SVR can handle nonlinearity and complex datasets through the choice of kernel functions (e.g. radial basis function, polynomial) and is less prone to overfitting.

SVR and XGBoost used the calculated feature spaces described in 2.3 without any selection. Since Elastic Net models assume a lack of multicollinearity between descriptors, an additional preprocessing step was done to remove features with a Pearson's correlation coefficient $r > 0.75$.

2.4.2. Model validation

Datasets are routinely split into several subsets to avoid overfitting. A model is overfit when it simply memorises the data and does not learn any actual trends, and performs well on the dataset it has been trained on but exhibits a drop in performance when tested on new data. Currently, the best practice is to have a corresponding train set from which the model learns, a validation set which is used to adjust the hyperparameters of the specific model to achieve best performance, and a test set to check how the model performs on new data unseen by the model.

To find the optimal hyperparameters for each model, a grid-search was conducted for each model using the GridSearchCV class from Scikit-learn. The dataset was randomly split into ten cross-validation (CV) folds, and the best-performing model was chosen based on the highest average R2 among the folds. In cases where there was a significant improvement in the training set performance but minimal gain in the CV score, models with a smaller gap between training and CV performance were preferred. The hyperparameters to be tuned were chosen as follows:

- **SVR:** kernel = ‘rbf’, ‘poly’, ‘sigmoid’
  Gamma = ‘scale’, ‘auto’
  C = 0.1, 2, 5, 8, 10, 20, 50
- **XGBoost:** max_depth= 4, 5, 6, 7, 8, 10
  n_estimators = 100, 200, 300, 400
  Subsample: 0.7, 0.8, 0.9, 1
  Colsample_bytree: 0.8, 0.9, 1
  Min_child_weight: 1, 2, 3
- **Elastic Net:** optimal regularization parameters were selected using ElasticNetCV class.

After the optimal hyperparameters were determined through 10-fold CV, the model was retrained on the entire dataset. For hAChE, external validation was performed to assess generalizability to unseen data. Performance metrics

included coefficient of determination R^2 and mean squared error (MSE), denoted as R^2 and MSE_{train} for training data, and R^2_{CV} and MSE_{CV} for cross-validation. The metrics for the external hAChE dataset were recorded as R^2_{test} and MSE_{test}.

2.5. Dual-output MLP

The availability of biological and chemical data is growing exponentially, driving the development of increasingly complex predictive algorithms. Artificial neural networks (ANNs) were developed to loosely mimic the structure of the brain and consist of interconnected neurons, each with a nonlinear activation function. These neurons are grouped into layers, allowing the network to learn intricate patterns through weighted connections and iterative adjustments. Such neural networks allow for more flexibility in the model structure, and can help capture more complex patterns in the data.

To capture the shared structural features of ligands for both AChE and BChE, a dual-output model was developed to predict bioactivity against both enzymes simultaneously. A multi-layer perceptron (MLP) model was chosen due to its simplicity and suitability for datasets of moderate size and complexity. MLP is a feedforward ANN that consists of fully connected layers of neurons, and it was one of the first and simplest types of deep learning models.

2.6. Proteochemometrics

While traditional QSAR modelling focuses solely on predicting ligand activity based on the calculated molecular descriptors, it completely overlooks the influence of the target proteins on the activity. Proteochemometrics (PCM) is a different approach that allows to combine both ligand and target protein information in a single predictive model. This allows PCM to model interactions across multiple targets, and to assess how specific residues in the target proteins contribute to ligand binding and activity. This approach is also appropriate for

enzymes like AChE and BChE, which share many structural similarities but still exhibit different substrate preferences.

2.6.1. Target protein information

To incorporate protein information into the regression model, it was first necessary to define what would be considered as binding pockets of the targets. Tacrine, a well-characterized inhibitor, was chosen as the reference ligand due to its established binding mode and availability of its protein-ligand complexes. The binding pocket was then defined as residues within 5Å vicinity of tacrine. Complexes with tacrine were identified for hAChE and hBChE at resolutions of 2.85 Å and 2.10 Å, respectively. Since no tacrine complex was available for TcAChE, the native enzyme structure at resolution of 2.50 Å was used. These structures were then aligned in PyMol 3.0.0, and the divergent residues within the binding pocket were identified.

2.6.2. Descriptor space

To build a comprehensive dataset suitable for PCM, bioactivity data for all three targets was merged into a single dataset. This integration resulted in a total of 7,997 individual bioactivity entries. To differentiate between ligands that had recorded activity against multiple targets, three binary columns (‘hAChE’, ‘hBChE’ and ‘TcAChE’) were included to indicate whether a particular compound had recorded bioactivity data against each respective target (1 = yes, 0 = no).

2.6.3. Model evaluation

The PCM models were trained and evaluated across same eight descriptor sets and machine learning algorithms (Elastic Net, SVR and XGBoost) that were applied to ordinary QSAR models. The hyperparameters were individually optimized for each model using the same grid search procedure described above to find the best-performing 10-fold CV model score, and then retrain it on the full dataset.

3. Results and discussion

3.1. Exploring the dataset

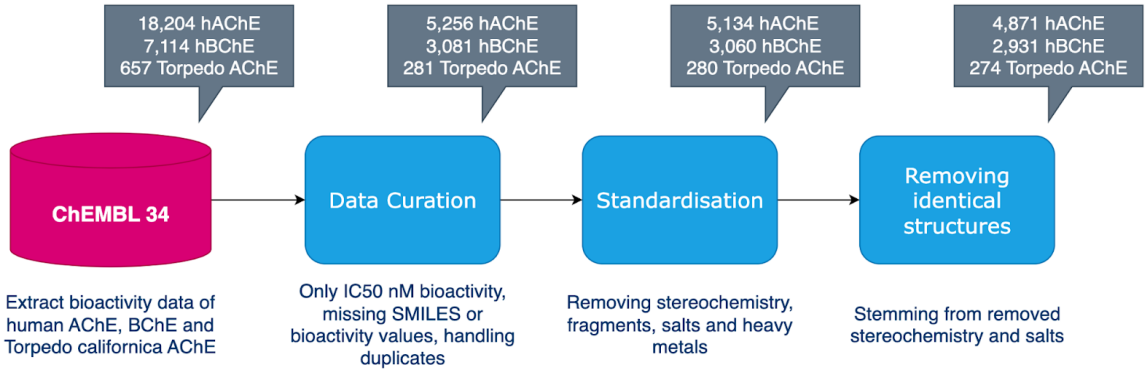


Figure 7: A visualization of the data remaining after the main preprocessing steps

After curation and standardization, the datasets have been significantly reduced in size. A visualization of the general data preprocessing workflow with the resulting number of individual compounds is shown in Fig.7. The largest reduction in raw ChEMBL data occurred when bioactivities recorded as anything other than IC50 with a standard relation of "=" were filtered out (Fig.8).

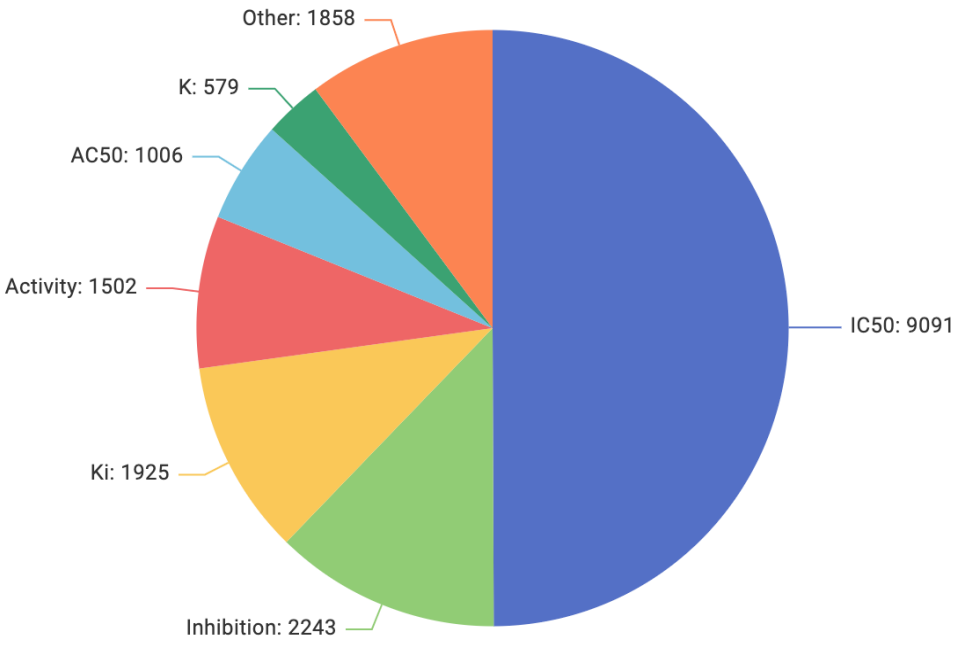


Figure 8: A pie chart of the bioactivity recordings distribution in the raw hAChE dataset

The inclusion of Ki bioactivities was initially considered. To evaluate their impact, compounds with IC50 and Ki bioactivities were processed separately, standardized, and then compared using a correlation plot (Fig. 9). This analysis was conducted on the hAChE dataset, as it was the largest and most relevant. After standardization, 2,871 IC50 and 498 Ki compounds remained, with 262 identical compounds present in both datasets. A Pearson's correlation coefficient of 0.872 indicated a strong relationship between IC50 and Ki values, though some outliers were observed. Given the minimal loss of only 236 unique compounds when excluding Ki bioactivities, it was decided to retain only IC50 data entries to ensure consistency and reduce potential fluctuations in bioactivity data measurements.

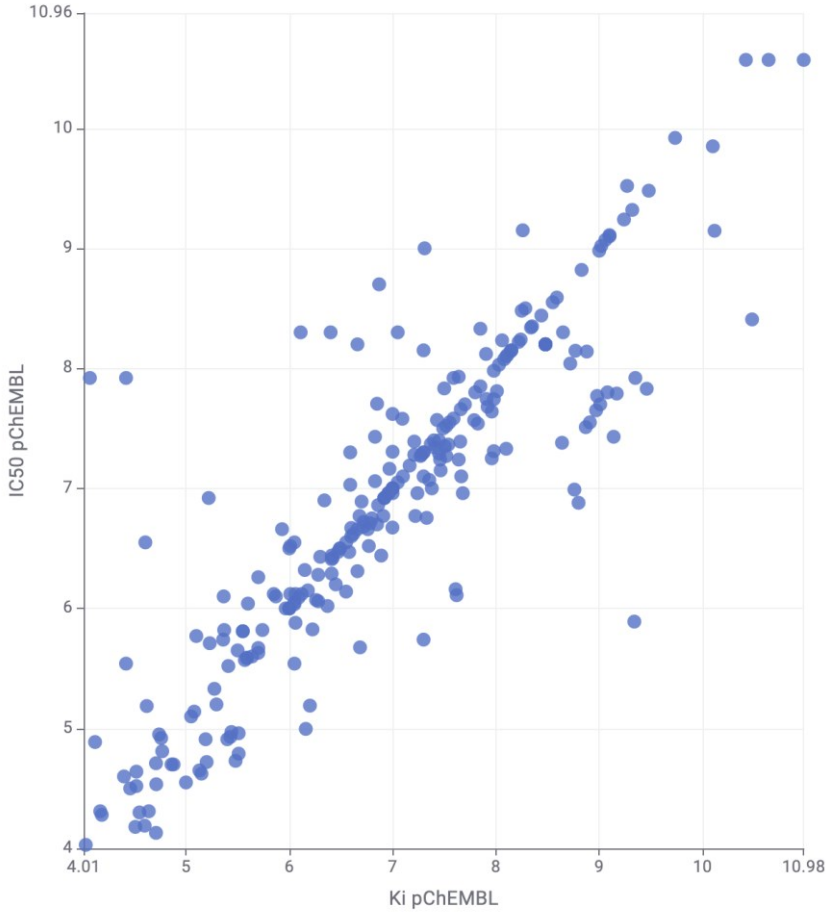


Figure 9: Scatter plot of Ki against IC50 pChEMBL in the hAChE dataset

3.1.1. Covalent and noncovalent binding compounds

Classical QSAR models assume that structurally similar compounds exhibit similar bioactivities. However, covalent and noncovalent binding compounds often differ in structure and activity profiles, making it difficult to model them simultaneously. Including both types in the same model could potentially introduce noise and reduce predictive accuracy, as molecular descriptors relevant to noncovalent interactions may not be as informative for covalent interactions, and vice versa.

In order to improve model accuracy, an initial attempt was made to separate covalent and noncovalent binding compounds in the datasets. However, distinguishing between them proved challenging. To filter the compounds suspected to bind covalently to the targets, KNIME's substructure search node was used. This node allows to filter molecules with SMARTS queries, which are line notations describing substructures in the molecule with specific atom and bond types.

Firstly, organophosphates, which are known to bind covalently to cholinesterases, were identified using a simple SMARTS query ‘P=O’. The resulting compounds were checked manually to confirm the presence of an organophosphate moiety, and such compounds were removed from the datasets. The second group, carbamates, posed a greater challenge, as both covalent and noncovalent (reversible) binding carbamates are known. Filtering carbamates was done with the query “[CX3](=[O])[NX3]” (Table 2). Due to carbamates comprising relatively large chunks of their respective datasets and the inability to definitively separate the covalent from noncovalent binding carbamates, it was decided to retain in the dataset without further selection.

Table 2: Number of presumably covalent substructures in the datasets

Dataset	Number of organo-phosphates	Number of carbamates	Total number of compounds
hAChE	17	1889	4,871
hBChE	48	1353	2,931
TcAChE	0	7	274

3.1.2. Enantiomers in the dataset

Stereochemistry plays a critical role in the develoment of a robust QSAR model, as different enantiomets can exhibit vastly different bioactivities. Removing hydrogens and stereochemistry from the dataset during standardization could potentially lead to significant information loss, in case there were many stereoisomers with different bioactivities present. To investigate how impactful stereochemistry was on the three cholinesterase datasets, the sizes of the preprocessed and standardized datasets with and without stereochemical information were compared (Table 3). There difference in dataset sizes appeared to be minimal, suggesting that removing stereochemistry didn't have a significant impact on the overall dataset composition.

Table 3: Composition of the datasets with and without stereochemistry

Dataset	No. of compounds when stereo info is kept	No. of compounds when stereo info is removed
hAChE	4,962	4,791
hBChE	3,015	2,931
TcAChE	276	274

Furthermore, in order to estimate how large the bioactivity differences between stereoisomers were, InChIKeys were calculated for the ‘non-stereo’ dataset, and only the compounds that had identical InChIKeys coming up in the set more than once were kept. The respective stereisomers were then added to the set by cross-referencing with the ChEMBL IDs from the ‘stereo’ dataset. By calculating the pChEMBL range for each individual InChIKey, only 26 compounds showed pChEMBL differences in the range >1 in the largest hAChE dataset. This suggested that removing stereochemical data did not substantially alter dataset composition.

Since there wasn’t a large amount of enantiomers within the dataset and most of them did not have large pChEMBL differences between each other, it was decided to primarily use the dataset with the stereochemical information and hydrogens removed for subsequent molecular descriptor calculation and machine learning. This approach allowed to simplify the dataset and the process of descriptor calculation.

3.2. Calculating descriptors

The sample to feature ratio is also important to consider when constructing the descriptor space for the dataset. If there are too many features relative to samples, the model performance might be overestimated, and the computational time for the models also increases. Therefore, some additional preprocessing was done on all descriptors sets to reduce the number of features fed into the models. The only exception were the 512-bit CDDD descriptors, which were left untouched due to their unique nature as learned embeddings.

To remove redundancy, constant features with zero variance were excluded from all datasets. Additionally, for the smaller TcAChE dataset Pearson correlation coefficients were calculated for descriptor pairs, and if the coefficient (r) exceeded 0.70, one descriptor from the pair was randomly removed. The resulting dimensions of the feature spaces can be seen in Table 1.

3.3. Individual QSAR models

3.3.1. Overall model performance

The regression model results for the three cholinesterase targets across the eight descriptor sets are summarized in Tables 4-6. XGBoost and SVR demonstrated comparable performance across most descriptors, while Elastic Net showed higher variability and consistently lower predictive accuracy. This was expected, as Elastic Net assumes a linear relationship between the descriptors and the target and is more sensitive to noise, making it less suited for complex datasets.

The best-performing models for each dataset can be seen in Table 7. For both hAChE and hBChE, the best performers were the SVR models trained on CDDD calculated from datasets without stereochemical information. Models trained on CDDD with stereochemical performed similarly, indicating no significant improvement from including stereochemistry in the datasets. For TcAChE, the combination of RDKit and MACCS descriptors yielded the best performance. Although the XGBoost model showed slightly better cross-validation performance, it had a far larger gap to the training score, so the SVR model was chosen as the best model for TcAChE. Overall, the models performed slightly better on the hBChE dataset despite its smaller size. This could potentially indicate that the hAChE dataset consists of more noisy or irrelevant data, though it is also possible that the models simply overfit on the hBChE datasets due to a larger feature to sample ratio.

Table 7: Best-performing regression models

Target	Best-performing model	Q2	R2
hAChE	SVR with CDDD	0.720	0.902
hBChE	SVR with CDDD	0.744	0.938
TcAChE	MACCS + RDKit	0.623	0.922

Generally, the models have exhibited quite large gaps between cross-validation and train scores. This may suggest that the models were overfitting, and may not have captured the relationships within the datasets properly. An additional validation on external datasets would allow to gain more insight into whether the patterns learned by the best-performing models could generalize well to unseen data.

3.3.2. Residual analysis

To further evaluate the best-performing hAChE model (i.e. SVR with CDDD descriptors without stereochemical information), a residual plot was generated (Fig.10). Residuals were defined and calculated as the difference between observed and predicted pChEMBL values. In an ideal non-linear model, residuals should be scattered randomly around zero with no systematic patterns. Otherwise, the latter may indicate that the model has failed to capture nonlinear relationships in the dataset. Residual plots are also useful for identifying poorly predicted outlier values, which can be further analyzed to gain more insight into the limitations of the model or the dataset.

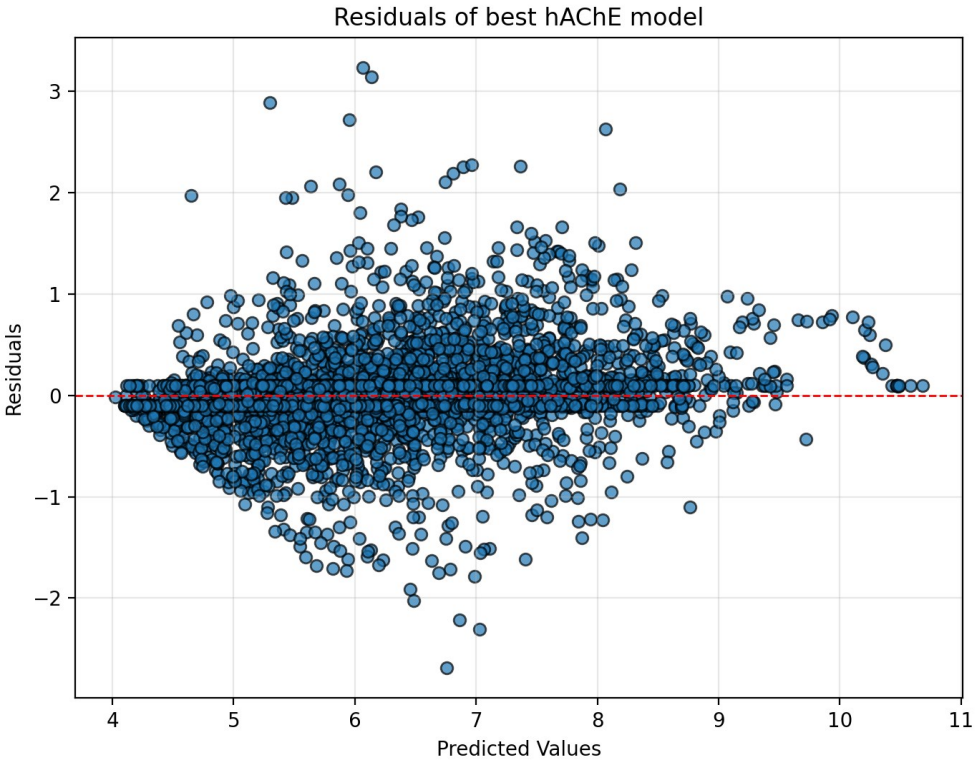


Figure 10: A residual plot for the hAChE SVR model trained on CDDD descriptors

Outliers in the best-performing hAChE model were identified as compounds with an absolute difference between actual and predicted pChEMBL values ≥ 2.5. Six such outliers were identified in the dataset.

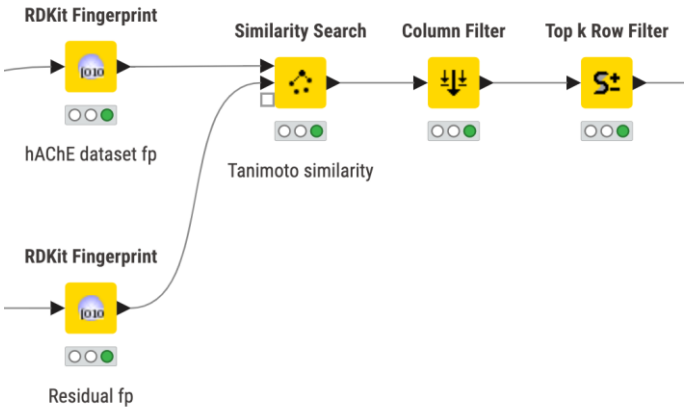
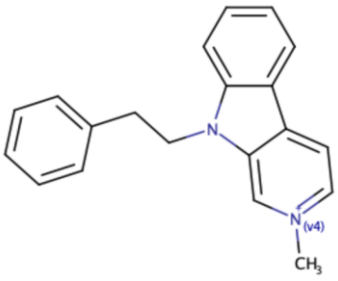
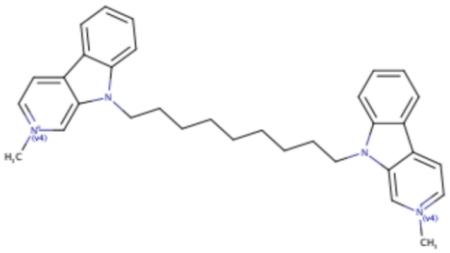
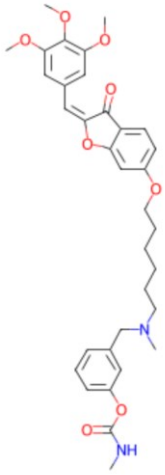
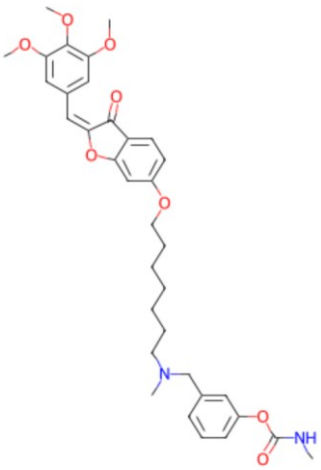
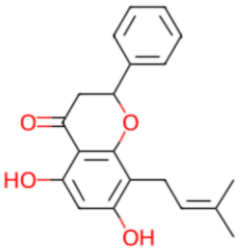
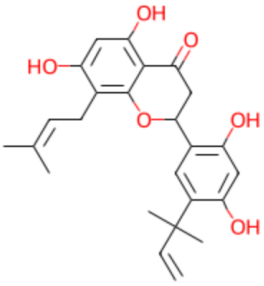
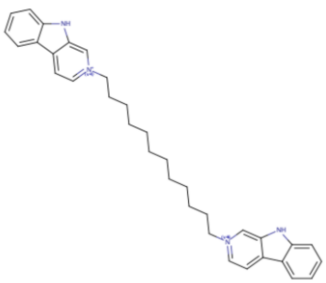
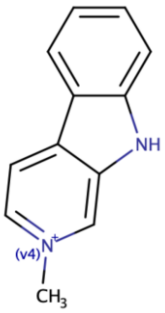
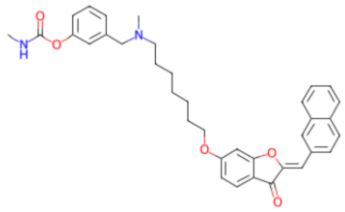
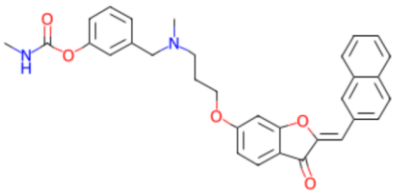
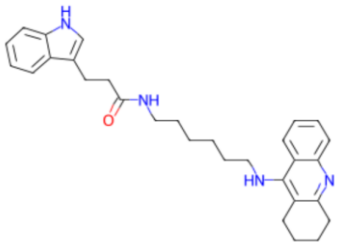
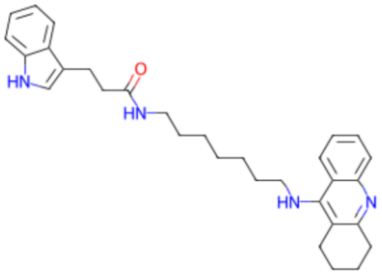


Figure 11: An excerpt from the KNIME workflow for retrieving molecules similar to the large residuals in the best hAChE model

To further explore why these particular compounds were predicted poorly, structurally similar compounds were searched within the train dataset. This was done in KNIME by first calculating 2048-bit FeatMorgan fingerprints for the compounds, and then searching for molecules similar to the outliers by calculating the Tanimoto similarity, ranked in a descending order (Fig.11). FeatMorgan fingerprints were used for this task because as circular fingerprints, they can capture structural features of the molecules especially effectively. The results of the search can be seen in Table 8.

Table 8: Largest residuals and compounds most similar to them in the hAChE dataset

Residual molecule		SImilar molecules	
	Observed pChEMBL: 9.30 Predicted pChEMBL: 6.06		Observed pChEMBL: 9.22 Predicted pChEMBL: 8.90
	Observed pChEMBL: 9.28 Predicted pChEMBL: 6.14		Observed pChEMBL: 6.28 Predicted pChEMBL: 6.02
	Observed pChEMBL: 8.19 Predicted pChEMBL: 5.29		Observed pChEMBL: 4.54 Predicted pChEMBL: 4.44
	Observed pChEMBL: 4.07 Predicted pChEMBL: 6.76		Observed pChEMBL: 5.39 Predicted pChEMBL: 5.32

Residual molecule		SImilar molecules	
	Observed pChEMBL: 8.68 Predicted pChEMBL: 5.95		Observed pChEMBL: 5.47 Predicted pChEMBL: 5.57
	Observed pChEMBL: 10.7 Predicted pChEMBL: 8.07		Observed pChEMBL: 10.22 Predicted pChEMBL: 8.18

The analysis revealed that most of the outliers represented activity cliffs. Structurally similar compounds generally exhibited moderate pChEMBL values (5–6) and were predicted accurately by the model. The assays associated with the outlier compounds were reviewed manually, confirming that all represented real and valid molecules with correctly recorded bioactivity values.

3.3.3. Evaluation of the external datasets

Following curation, the bioactivity data from the two external sources, BindingDB and Tox21, was evaluated for its overlap and compatibility with the primary training ChEMBL dataset. This step was important to assess the reliability of the external datasets for model validation.

The BindingDB dataset contained 120 compounds, of which 84 overlapped with the ChEMBL training set. A correlation analysis of pChEMBL values between shared compounds in BindingDB and ChEMBL yielded an excellent correlation coefficient of r = 0.965, confirming dataset reliability.

In contrast, the Tox21 dataset included 170 compounds, only 15 of which overlapped with the ChEMBL training dataset. A correlation analysis on this limited subset yielded a correlation coefficient r = 0.74, with several visible

outliers (Fig.12). Given the small overlap and moderate correlation, the reliability of this dataset for model evaluation was uncertain.

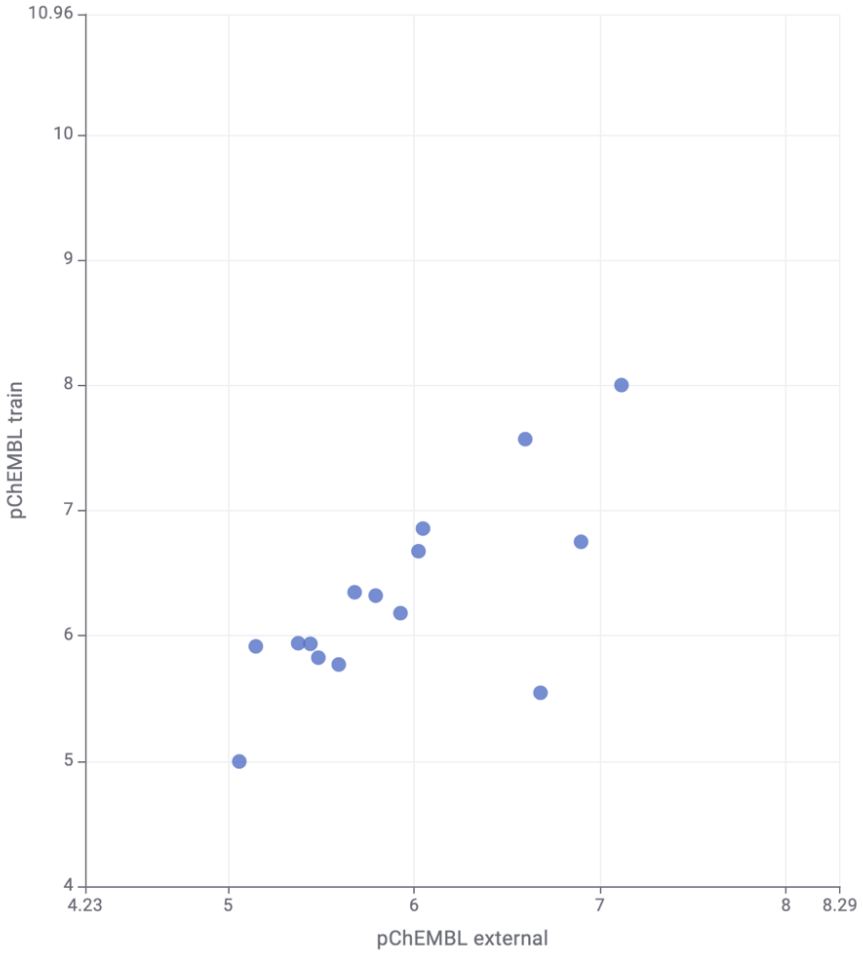


Figure 12: A correlation plot between shared structures in the train ChEMBL and external PubChem datasets for hAChE

In order to visualize the chemical space of the compounds in a graph, dimensionality reduction techniques like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation (UMAP) are often used. These unsupervised algorithms allow to transform the data into a lower-dimensional chemical space while still capturing as much original information as possible. For representing the chemical spaces of the train and external hAChE datasets, 2048-bit FeatMorgan fingerprints with radius 2 were chosen. However, reducing the fingerprint space to two dimensions

with PCA resulted in less than 25% of information retained. Therefore, it was chosen to do a UMAP projection since it is a more modern method that can compress the feature space without losing much information, and is claimed to preserve more global structure in the data than t-SNE (Orlov et al., 2025). A visual inspection of the chemical spaces implied that the external compounds were within the original dataset's chemical space, since there were seemingly no outliers within them (Fig.13).

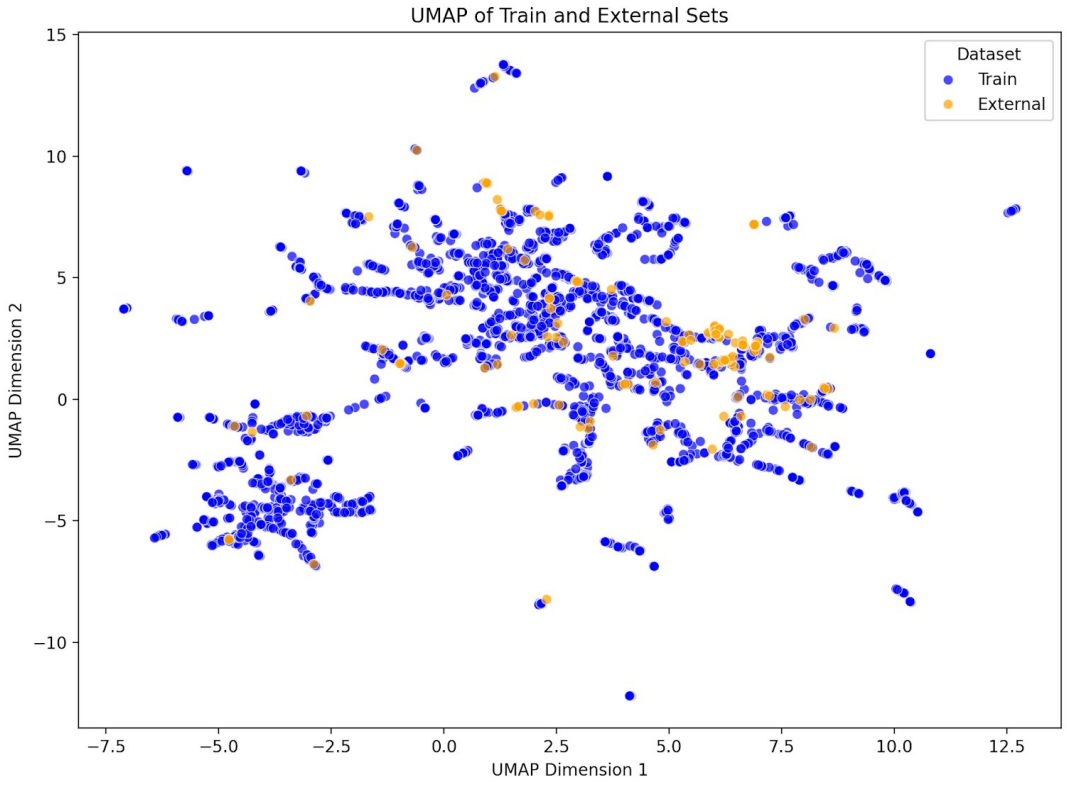


Figure 13: A UMAP projection of the chemical space structures in the train ChEMBL and external PubChem datasets for hAChE

Despite the similarity of the chemical space, testing the hAChE model on the external dataset (191 compounds in total) yielded poor predictive performance, with an R^2 of 0.2, despite a mean squared error comparable to the cross-validation set. A closer examination of the predicted against observed pChEMBL values revealed that most predictions fell within a ±1 pChEMBL window, but were broadly distributed (Fig.14). This dispersion has likely contributed to the low R^2. The small size of the external dataset may have further worsened this issue.

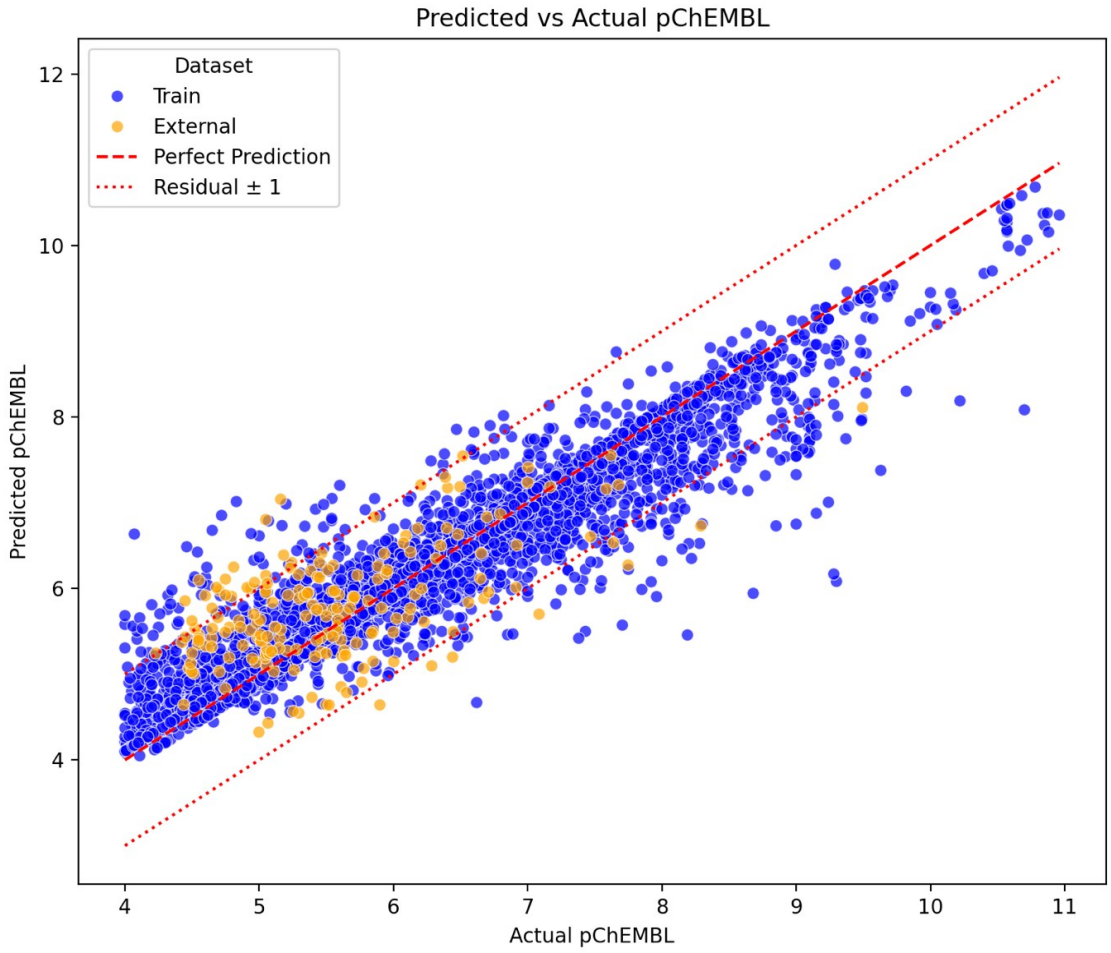


Figure 14: A scatter plot comparing distribution of predicted against observed pChEMBL values for the train ChEMBL and external PubChem datasets for hAChE

The observed gap between the model's performance on training set and on the externnal set raises questions about the generalizability of the model. The high R^2 performance of the training and cross-validation sets may be miselading and capture the noise in the data instead of the underlying patterns. However, since the correlation of pChEMBL values between shared compounds in train and external datasets had distinct outliers and was only performed on 15 compounds, this external data may not be a reliable reference for model validation. A proper, larger external dataset with reliable bioactivity data would be more reliable for estimation of the model performance.

3.4. Multi-Output MLP Model

To create a suitable input for the dual-output model, the AChE and BChE datasets were filtered in KNIME to only include molecules with dual activity data, resulting in a total of 1993 unique compounds. Identical molecular structures were handled as described earlier through identification with InChIKeys and concentration to a pChEMBL mean value. This ensured the integrity of the data and avoided redundancy in the training process.

Unlike the classical QSAR models for hAChE and hBChE, only three sets of descriptors were calculated, since the dataset of dual bioactivity structures was smaller in size. A large number of descriptors relative to the bioactivity samples could likely lead to an overtrained model. Therefore, additional feature selection steps were also applied. Features with Pearson correlation coefficients > 0.75 were removed to reduce redundancy, and features with variance < 0.15 were also removed (Table 9).

Table 9: Descriptor spaces used for the MLP model

Descriptor	Standard number generated	Number selected for the models
MACCS keys	166	41
RDKit 1D- and 2D	~200	25
MACCS+RDKit	~366	75

A preliminary analysis of the dual-activity dataset demonstrated a general positive correlation, as expected from the two structurally and functionally similar targets (Fig.15). The data seemed to spread a little more towards higher pChEMBL values, suggesting the presence of compounds highly selective for one target over the other, introducing more variability into the dataset.

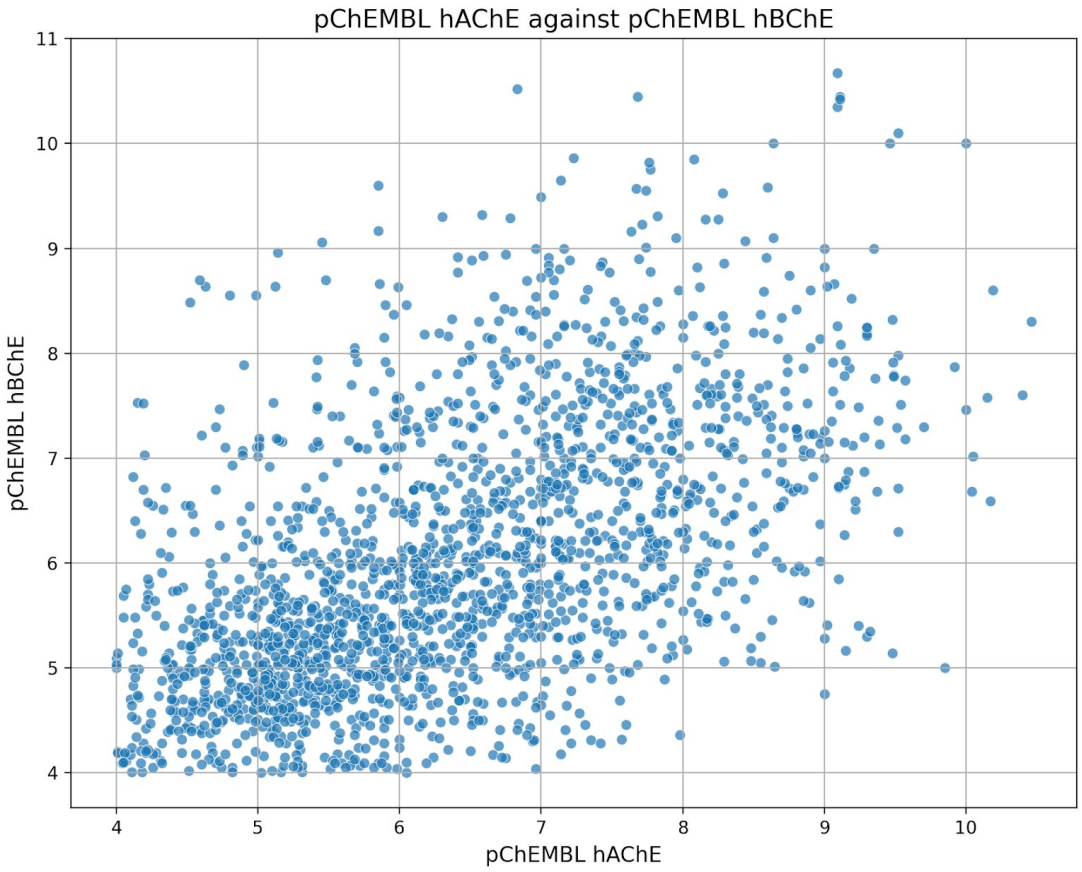


Figure 15: A scatter plot for visualizing the correlation between hAChE and hBChE pChEMBL values

The MLP model consisted of a single hidden layer with ReLU activation function, followed by an output layer with 2 neurons and a linear activation function to predict pChEMBL for both hAChE and hBChE. The dataset was split into training (70%), validation (15%) and test (15%) sets to ensure a proper evaluation of the model performance. An L2 regularization was applied to the hidden layer to mitigate overfitting. The optimal number of neurons in the hidden layer was determined separately for each model using the validation set. The model was trained using the Adam optimizer with a mean squared error (MSE) loss function. Early stopping was implemented with a patience of 10 epochs to prevent overfitting, and the best weights were restored based on the validation loss.

The model performance comparable to the individual regression models (Table 10). The gaps between the training and validation performances suggest some degree of overfitting despite the implementation of early stopping and regularization parameters. The BChE predictions performed consistently better

than the AChE predictions, aligning with the results from the previous classical QSAR models.

Table 10: Performance of the dual-output MLP model

Descriptor type	$R2_{train}$		$R2_{val}$		$R2_{test}$	
	AChE	BChE	AChE	BChE	AChE	BChE
MACCS	0.864	0.856	0.602	0.679	0.577	0.617
RDKit	0.777	0.773	0.593	0.686	0.547	0.625
MACCS +RDKit	0.880	0.877	0.614	0.704	0.609	0.639

Overall, the model demonstrated acceptable performance, highlighting a potential for multi-target activity prediction in cholinesterases. However, it should be noted that the model's performance was limited by the relatively small number of compounds available.

3.5. Proteochemometrics

After aligning the three target proteins in PyMol, differences in amino acid residues across them were observed at three key positions: 83, 337 and 446 (Table 11).

To quantify the difference in properties of the amino acids at these positions, three Z-scale descriptors were generated (Table 12). Similarly to molecular descriptors, Z-scales are computed to describe the physicochemical properties of amino acids. Originally, three Z-scales were introduced by Hellberg et al. based on NMR and thin-layer chromatography data, but later two more were added by Sandberg et al. (Hellberg et al., 1987; Sandberg et al., 1998). Tentatively, Z1 can be interpreted as hydrophobicity of the amino acid, Z2 as the 3D space it takes up, and Z3 as its electronic properties. The last two can be more difficult to interpret, so within this thesis only Z1, Z2 and Z3 were used.

The models were built using the same eight descriptor sets used for the individual QSAR models. The performance results of all the constructed models can be seen in Table 14. Overall, the model performance was consistent with the individual QSAR models for the targets, with Elastic Net consistently underperforming. The best-performing model was XGBoost using PubChem descriptors, which was subsequently used to assess the relevance of individual residue properties using a feature importance ranking (Table 15).

The feature importance values were derived based on the Gain metric, which represents the relative contribution of each feature in the splits across each tree in the model. The importance scores were normalized to add up to 1, and ranked in descending order to estimate the relative significance of individual descriptors within the models.

Position 446, which contains proline in hAChE, was the most important in the ranking extracted from the XGBoost model, differing substantially from methionine/isoleucine at the same position in hBChE/TcAChE. In contrast, position 83, corresponding to threonine in hAChE and serine in hBChE/TcAChE, was less important (Table 13).

Table 15: Best-performing models of the PCM model with and without protein descriptors

Model used	With protein descriptors		Without protein descriptors	
	R^2	Q^2	Q^2	Q^2
PubChem fingerprints XGBoost	0.899	0.692	0.815	0.554
SVR with CDDD	0.870	0.658	0.796	0.528

To evaluate the contribution of protein descriptors, models were retrained on the same datasets, but excluding the z-scale descriptors and only keeping the binary

target identifiers and ligand descriptors. This led to a noticeable decrease in model performance for the two best-performing models (Table 15), underlying the contribution of protein descriptors to improve predictive accuracy.

To visualize the distribution of bioactivity data across the three targets, a Venn diagram was made to represent their respective sizes and overlaps with each other in terms of individual molecular structures. (Fig. 16) The diagram demonstrates that the bioactivities and by extension the protein descriptors in the dataset are not evenly represented. While there was a relatively large overlap between hAChE and hBChE bioactivities, it was very limited with the TcAChE dataset. Since randomized 10-fold CV was used to tune the model, this unequal representation may have provided misleading results and overestimated the model's generalizability. To mitigate this effect, one possibility would be to use stratified sampling in future studies to ensure a more equal representation across the target and minimized selection bias.

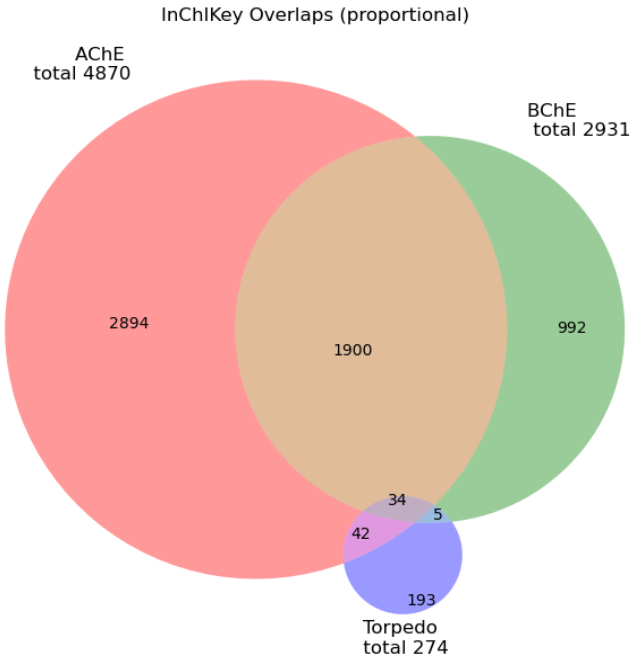


Figure 16: A Venn diagram demonstrating the distribution of bioactivities across the three datasets available for PCM

4. Conclusion

Within this thesis, various QSAR models for predicting the bioactivity of compounds targeting cholinesterases (hAChE, hBChE and TcAChE) were developed, including a PCM model. Regression models using the SVR and XGBoost algorithms demonstrated consistent predictive performance, particularly for hBChE, although concerns regarding overfitting and overall generalizability of the models were raised due to large gaps between training and cross-validation scores. A residual analysis revealed that the models struggled to predict activity peaks. The suboptimal performance across external validation datasets also underlined the necessity of using larger, high-quality external datasets for reliable model validation. The dual-output MLP model showed promising results for multi-activity prediction across cholinesterase, and PCM indicated that the addition of target protein information could improve performance. However, there were also distinct limitations:

- **Data quality and consistency:** Since cholinesterase inhibitors are a heterogenous group with both covalent and noncovalent binding compounds, separating the two could possibly create a more consistent input for the machine-learning models.
- **Overfitting and generalizability:** The large gaps between training and cross-validation scores, as well as the poor performance on the external datasets, raised concerns about the model's ability to generalize to unseen data. A larger, high-quality external dataset would allow for a more robust validation of the models.
- **Uneven representation of bioactivities:** The unequal distribution of bioactivities across the three targets might have biased the proteochemometric models, suggesting the need for a proper stratified sampling in future studies.

If these limitations are addressed, the approaches presented in this thesis could prove to be useful tools for predicting cholinesterase inhibition and advancing the drug-discovery efforts.

List of abbreviations

hAChE	Human Acetylcholinesterase
hBChE	Human Butyrylcholinesterase
TcAChE	Torpedo californica Acetylcholinesterase
QSAR	Quantitative Structure-Activity-Relationship
CDDD	Continuous Data-Driven Descriptors
MLP	Multi-Layer Perceptron
SVR	Support Vector Regression
UMAP	Uniform Manifold Approximation and Projection
MSE	Mean Squared Error
PCM	Proteochemometrics

List of tables

Table 1: Size of descriptor spaces for each dataset

Descriptor type	Number of descriptors			
	Standard	hAChE	hBChE	TcAChE
MACCS keys	166	151	151	43
RDKit 1D- and 2D	~200	184	187	52
MACCS+RDKit	~366	335	338	78
Mordred 1D and 2D	~1200	1057	1144	-
FeatMorgan	1024	1015	971	-
PubChem	881	633	601	-

Table 4: Performance metrics for the hAChE models

Descriptors	Elastic Net				XGBoost				SVR			
	Q2	MSE_{CV}	R2	MSE_{train}	Q2	MSE_{CV}	R2	MSE_{train}	Q2	MSE_{CV}	R2	MSE_{train}
MACCS	0.311	1.153	0.357	1.079	0.672	0.547	0.869	0.220	0.629	0.623	0.794	0.347
RDKit	0.341	1.117	0.400	1.006	0.661	0.565	0.872	0.214	0.668	0.551	0.883	0.195
MACCS +RDKit	0.430	0.957	0.507	0.827	0.661	0.565	0.863	0.230	0.698	0.503	0.891	0.182
Mordred	0.482	0.927	0.588	0.691	0.677	0.538	0.915	0.142	0.691	0.518	0.928	0.122
Feat Morgan	0.555	1.037	0.678	0.540	0.696	0.508	0.869	0.219	0.685	0.527	0.943	0.096
PubChem	0.488	0.873	0.583	0.699	0.699	0.502	0.898	0.171	0.646	0.591	0.832	0.282
CDDD no stereo	0.495	0.846	0.590	0.688	0.663	0.562	0.946	0.091	**0.720**	**0.468**	**0.902**	**0.165**
CDDD stereo	0.481	0.832	0.569	0.636	0.652	0.580	0.939	0.101	0.714	0.476	0.912	0.147

Table 5: Performance metrics for the hBChE models

Descriptors	Elastic Net				XGBoost				SVR			
	Q2	MSE_{cv}	R2	MSE_{train}	Q2	MSE_{cv}	R2	MSE_{train}	Q2	MSE_{cv}	R2	MSE_{train}
MACCS	0.443	0.979	0.490	0.865	0.653	0.580	0.791	0.354	0.670	0.551	0.874	0.213
RDKit	0.464	0.904	0.518	0.816	0.680	0.535	0.844	0.264	0.700	0.501	0.884	0.196
MACCS +RDKit	0.540	0.812	0.613	0.657	0.686	0.526	0.848	0.258	0.708	0.489	0.885	0.194
Mordred	0.546	0.776	0.664	0.570	0.700	0.501	0.907	0.157	0.717	0.473	0.919	0.137
Feat Morgan	0.627	0.906	0.749	0.425	0.722	0.465	0.889	0.187	0.704	0.495	0.923	0.129
PubChem	0.566	0.767	0.673	0.554	0.732	0.447	0.921	0.135	0.683	0.530	0.910	0.153
CDDD no stereo	0.575	0.716	0.696	0.515	0.676	0.543	0.914	0.145	**0.744**	**0.429**	**0.938**	**0.105**
CDDD stereo	0.571	0.729	0.681	0.542	0.677	0.542	0.914	0.146	0.735	0.445	0.909	0.154

Table 6: Performance metrics for the TcAChE models

Descriptors	Elastic Net				XGBoost				SVR			
	Q2	MSE_{cv}	R2	MSE_{train}	Q2	MSE_{cv}	R2	MSE_{train}	Q2	MSE_{cv}	R2	MSE_{train}
MACCS	0.481	1.221	0.586	0.937	0.583	0.890	0.903	0.220	0.609	0.844	0.894	0.239
RDKit	0.499	1.130	0.652	0.788	0.625	0.804	0.989	0.025	0.601	0.859	0.897	0.232
MACCS +RDKit	0.560	0.991	0.743	0.582	0.635	0.797	0.990	0.023	0.623	0.829	0.922	0.176

Table 11: Differing residues across the binding sites of the three targets

Position	hAChE	hBChE	TcAChE
83	THR	SER	SER
337	TYR	ALA	PHE
446	PRO	MET	ILE

Table 12: Z-scale descriptors generated for the residues

	83z1	83z2	83z3	337z1	337z2	337z3	446z1	446z2	446z3
AChE	0.92	-2.09	-1.40	-1.39	2.32	0.01	-1.22	0.88	2.23
BChE	1.96	-1.63	0.57	0.07	-1.73	0.09	-2.49	-0.27	-0.41
Torpedo	1.96	-1.63	0.57	-4.92	1.30	0.45	-4.44	-1.68	-1.03

Table 13: Z-scale descriptors from the XGBoost model trained on PubChem fingerprints, ranked in order of descending gain

Importance ranking	Protein descriptor
1	446z3
2	337z2
3	83z2
4	446z1
5	337z1
6	83z3
7	83z1
8	446z2
9	337z3

Table 14: Performance metrics for the PCM models

Descriptors	Elastic Net			XGBoost				SVR			
	Q2	R2	MSE_{train}	Q2	MSE_{cv}	R2	MSE_{train}	Q2	MSE_{cv}	R2	MSE_{train}
MACCS	0.310	0.336	1.138	0.661	0.580	0.880	0.206	0.641	0.614	0.875	0.214
RDKit	0.325	0.350	1.114	0.639	0.619	0.934	0.114	0.650	0.600	0.850	0.257
Mordred	0.432	0.494	0.868	0.621	0.648	0.899	0.173	0.623	0.647	0.854	0.251
Feat Morgan	0.510	0.584	0.713	0.684	0.541	0.864	0.233	0.634	0.625	0.924	0.129
PubChem	0.457	0.518	0.825	**0.692**	**0.527**	**0.899**	**0.173**	0.620	0.649	0.826	0.297
CDDD no stereo	0.459	0.523	0.818	0.606	0.674	0.962	0.066	**0.658**	**0.586**	**0.870**	**0.223**
CDDD stereo	0.448	0.509	0.840	0.599	0.685	0.956	0.075	0.647	0.604	0.863	0.234

Bibliography

Akhoon, B. A., Choudhary, S., Tiwari, H., Kumar, A., Barik, M. R., Rathor, L., Pandey, R., & Nargotra, A. (2020). Discovery of a New Donepezil-like Acetylcholinesterase Inhibitor for Targeting Alzheimer's Disease: Computational Studies with Biological Validation. *Journal of Chemical Information and Modeling*, *60*(10), 4717–4729. https://doi.org/10.1021/acs.jcim.0c00496

Alonso, D., Dorronsoro, I., Rubio, L., Muñoz, P., García-Palomero, E., Del Monte, M., Bidon-Chanal, A., Orozco, M., Luque, F. J., Castro, A., Medina, M., & Martínez, A. (2005). Donepezil–tacrine hybrid related derivatives as new dual binding site inhibitors of AChE. *Bioorganic & Medicinal Chemistry*, *13*(24), 6588–6597. https://doi.org/10.1016/j.bmc.2005.09.029

Bahia, M. S., Kaspi, O., Touitou, M., Binayev, I., Dhail, S., Spiegel, J., Khazanov, N., Yosipof, A., & Senderowitz, H. (2023). A comparison between 2D and 3D descriptors in QSAR modeling based on bio-active conformations. *Molecular Informatics*, *42*(4), 2200186. https://doi.org/10.1002/minf.202200186

Boldini, D., Ballabio, D., Consonni, V., Todeschini, R., Grisoni, F., & Sieber, S. A. (2024). Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *Journal of Cheminformatics*, *16*(1), 35. https://doi.org/10.1186/s13321-024-00830-3

Breijyeh, Z., & Karaman, R. (2020). Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules*, *25*(24), 5789. https://doi.org/10.3390/molecules25245789

Chekmarev, D., Kholodovych, V., Kortagere, S., Welsh, W. J., & Ekins, S. (2009). Predicting Inhibitors of Acetylcholinesterase by Regression and Classification Ma-

chine Learning Approaches with Combinations of Molecular Descriptors. *Pharmaceutical Research*, *26*(9), 2216–2224. https://doi.org/10.1007/s11095-009-9937-8

Dvir, H., Silman, I., Harel, M., Rosenberry, T. L., & Sussman, J. L. (2010). Acetylcholinesterase: From 3D structure to function. *Chemico-Biological Interactions*, *187*(1–3), 10–22. https://doi.org/10.1016/j.cbi.2010.01.042

El Khatabi, K., El-Mernissi, R., Aanouz, I., Ajana, M. A., Lakhlifi, T., Khan, A., Wei, D.-Q., & Bouachrine, M. (2021). Identification of novel acetylcholinesterase inhibitors through 3D-QSAR, molecular docking, and molecular dynamics simulation targeting Alzheimer's disease. *Journal of Molecular Modeling*, *27*(10), 302. https://doi.org/10.1007/s00894-021-04928-5

Espay, A. J., Kepp, K. P., & Herrup, K. (2024). Lecanemab and Donanemab as Therapies for Alzheimer's Disease: An Illustrated Perspective on the Data. *Eneuro*, *11*(7), ENEURO.0319-23.2024. https://doi.org/10.1523/ENEURO.0319-23.2024

Hellberg, S., Sjoestroem, M., Skagerberg, B., & Wold, S. (1987). Peptide quantitative structure-activity relationships, a multivariate approach. *Journal of Medicinal Chemistry*, *30*(7), 1126–1135. https://doi.org/10.1021/jm00390a003

Moss, D. E. (2020). Improving Anti-Neurodegenerative Benefits of Acetylcholinesterase Inhibitors in Alzheimer's Disease: Are Irreversible Inhibitors the Future? *International Journal of Molecular Sciences*, *21*(10), 3438. https://doi.org/10.3390/ijms21103438

Moss, D. E., & Perez, R. G. (2021). Anti-Neurodegenerative Benefits of Acetylcholinesterase Inhibitors inAlzheimer's Disease: Nexus of Cholinergic and Nerve Growth Factor Dysfunction. *Current Alzheimer Research*, *18*(13), 1010–1022. https://doi.org/10.2174/1567205018666211215150547

Mushtaq, G., Greig, N., Khan, J., & Kamal, M. (2014). Status of Acetylcholinesterase and Butyrylcholinesterase in Alzheimer's Disease and Type 2 Diabetes Mellitus. *CNS & Neurological Disorders - Drug Targets*, *13*(8), 1432–1439. https://doi.org/10.2174/1871527313666141023141545

Orlov, A. A., Akhmetshin, T. N., Horvath, D., Marcou, G., & Varnek, A. (2025). From High Dimensions to Human Insight: Exploring Dimensionality Reduction for Chemical Space Visualization. *Molecular Informatics*, *44*(1), e202400265. https://doi.org/10.1002/minf.202400265

Salman, M. M., Al-Obaidi, Z., Kitchen, P., Loreto, A., Bill, R. M., & Wade-Martins, R. (2021). Advances in Applying Computer-Aided Drug Design for Neurodegenerative Diseases. *International Journal of Molecular Sciences*, *22*(9), 4688. https://doi.org/10.3390/ijms22094688

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., & Wold, S. (1998). New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *Journal of Medicinal Chemistry*, *41*(14), 2481–2491. https://doi.org/10.1021/jm9700575

Silman, I., & Sussman, J. L. (2008). Acetylcholinesterase: How is structure related to function? *Chemico-Biological Interactions*, *175*(1–3), 3–10. https://doi.org/10.1016/j.cbi.2008.05.035

Silman, I., & Sussman, J. L. (2017). Recent developments in structural studies on acetylcholinesterase. *Journal of Neurochemistry*, *142*(S2), 19–25. https://doi.org/10.1111/jnc.13992

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Vignaux, P. A., Lane, T. R., Urbina, F., Gerlach, J., Puhl, A. C., Snyder, S. H., & Ekins, S. (2023). Validation of Acetylcholinesterase Inhibition Machine Learning Models for Multiple Species. *Chemical Research in Toxicology*, *36*(2), 188–201. https://doi.org/10.1021/acs.chemrestox.2c00283

Winter, R., Montanari, F., Noé, F., & Clevert, D.-A. (2019). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science*, *10*(6), 1692–1701. https://doi.org/10.1039/C8SC04175J

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474. https://doi.org/10.1002/jcc.21707

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, *52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004

Zhao, L., Ciallella, H. L., Aleksunes, L. M., & Zhu, H. (2020). Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discovery Today*, *25*(9), 1624–1638. https://doi.org/10.1016/j.drudis.2020.07.005

Zuliani, G., Zuin, M., Romagnoli, T., Polastri, M., Cervellati, C., & Brombo, G. (2024). Acetyl-cholinesterase-inhibitors reconsidered. A narrative review of post-marketing studies on Alzheimer's disease. *Aging Clinical and Experimental Research*, *36*(1), 23. https://doi.org/10.1007/s40520-023-02675-6