ORIGINAL ARTICLE

A C G publications *Bioorg. Med. Chem. Rep.* X:X (202X) X-X

Bioorganic & Medicinal Chemistry Reports

Comparing machine learning models for acetylcholine esterase inhibitors

Mehmet Ali Yucel [1,2*]

[1]*Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Erzincan University, 24100, Erzincan, Turkiye*
[2]*Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Mersin University, 33169, Mersin, Turkiye*

(Received June 06, 2022; Revised July 05, 2022; Accepted July 07, 2022)

Abstract: Acetylcholinesterase is the main neurotransmitter in the cholinergic system. Impairment of the cholinergic system can be a reason for Alzheimer's and multiple sclerosis. Alzheimer's disease and multiple sclerosis affect patients and their relatives' daily lives enormously. New therapies with more benefits than current therapies for these diseases would facilitate patients' lives. In this respect, discovering novel acetylcholine esterase inhibitors with more effective and fewer side effects is highly important. Machine learning algorithms are very useful to predict the activity of molecules for a biological target. In this study, our classification models were built with Deep Neural Networks (DNN), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) to predict molecules as active or inactive for acetylcholinesterase inhibitors. These models were evaluated with various metrics. As a result, The DNN model showed a better ability to classify (accuracy=0.93, F1 score=0.88, MCC=0.8, Roc-Auc=0.89 in the test set) molecules than the other models.

Keywords: Alzheimer's disease; machine learning; deep learning; acetylcholine esterase inhibitors; multiple sclerosis.

1. Introduction

Neurodegenerative diseases such as Parkinson's and Alzheimer's are often related to primary neuronal degeneration. This degeneration often depends on dysregulation of the cholinergic system. The major neurotransmitter of the cholinergic system is acetylcholine (ACh) which controls via acetylcholinesterase (AChE), and dysregulation of ACh can cause neurodegenerative diseases. Furthermore, AChE can be responsible for inflammation, cell apoptosis, morphogenic and adhesion functions, as well as participation in oxidative stress.[1]

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases. The prevalence of AD recently continues to increase. According to the Turkish Neurological Society, there are at least 600 thousand people who suffer from AD in Turkiye. In the United States, it is estimated that the number of people with AD will reach 13.8 million by 2050.

One of the crucial steps in the progression of AD is the accumulation of amyloid-beta (Aß) peptides. This progress causes several conditions such as tauopathy, neuroinflammation, neurofibrillary tangles, and synaptic injury. The cumulative Aß causes cholinergic dysfunction, decreasing ACh activity and increasing AChE activity. All of these compose memory loss, confusion, and other clinical symptoms. The popular treatment of AD is the using AChE inhibitors such as donepezil, galantamine, and rivastigmine.[1,2]

* E-Mail: mehmet.yucel@erzincan.edu.tr

The article was published by ACG Publications
https://acgpubs.org/journal/bioorganic-and-medicinal-chemistry-reports © Month-Month 202X EISSN:1307-6175
DOI: http://doi.org/10.25135/bmcr.29.22.06.2483
Available online: August 14, 2022

Multiple sclerosis (MS) is an autoimmune disease of the central nervous system characterized by inflammatory demyelinating and secondary axonal degeneration. The incidence of MS disease varies according to geographical regions. In epidemiological studies conducted in Turkiye, the prevalence of MS was found to be 0.4-1 in 1000 young adults. The prevalence in Europe and the United States is 1 in 1000 young adults.[1]

AChE has a key role in producing cholinergic inflammation. The increased AChE level provokes inflammation and stimulates the production of proinflammatory cytokines. Also, the Immune cells (T lymphocytes, B lymphocytes, macrophages, and dendritic cells) can produce AChE. These reasons led to studies about the effect of cholinergic systems on MS disease. One of these studies showed lower ACh levels in MS patients' sera and cerebrospinal fluid compared with healthy subjects. However, these studies are not sufficient yet and need to be further investigations of the cholinergic mechanisms in MS disease. [3,4]

Discovering new molecules is a highly costly and time-consuming process. In recent years, machine learning (ML) methods show this process can be more efficient. One of the reasons machine learning has become very popular recently is that a lot of data has been produced in every field today and it is still being produced. ML is an effective way of turning this data into knowledge. Since drug research and development studies are essential in terms of people's quality of life, the number of data obtained in this field is high. The success of ML methods in computer vision and natural language processing has also shed light on drug development. One of the first notable examples is the success of deep neural networks in the Kaggle competition held by Merck in 2012. Another example study in 2019, discovered Potent inhibitors for "discoidin domain receptor 1 (DDR1)" by In silico Medicine researchers in as little as 21 days.[5]

ML can be used in distinct stages of drug development such as the estimation of the chemical properties of the molecule and toxicity profiles. With ML, more rational drug design becomes achievable, and as a result, molecules with high activity, low toxicity, and fewer side effects profiles can be obtained in less time, at less cost. In this study, we explored the data and built three different machine models (DNN, XGBoost and SVM) to classify molecules for AChE inhibition.

2. Experimental

2.1. Data Curation and Preparation

The experimental IC_{50} values (only nM) of acetylcholinesterase inhibitors were collected from an online ChEMBL database (release 28, Feb 2021). After preliminary model exploration, it was found that the original IC_{50} data distribution was particularly discrete, and the numerical difference was quite large, resulting in a poor model effect. Therefore, we took a negative logarithm of IC_{50} from the bottom of 10 to obtain pIC_{50}. (Figure 1) The mean of pIC_{50} values is 5.82 and the standard deviation is 1.56. We set the threshold pIC_{50} value as 7.0. Molecules are labelled active if their pIC_{50} value is higher than 7 otherwise, they are labeled inactive. (Figure 2) Before splitting the dataset, checked "NaN" values in smiles and activity columns. Duplicated compounds and compounds without IC_{50} values were removed. The duplicated molecule with a low activity value is preserved. 5328 compounds for AChE were used to construct and validate models. After that, the dataset has split into training and test datasets. Validation set obtains from train dataset with 5-fold cross-validation. Finally, there are 4051 molecules in the training dataset, 1013 molecules in the validation set, and 563 molecules in the test dataset.

Also, the principal component analysis was created for data exploration with some molecular descriptors (hydrogen bond acceptor, hydrogen bond donor, molecular weight, log P, and polar surface area) (Figure 3).

2.2 Machine Learning Models

Support vector machines (SVM) is one of the widely used machine learning algorithms which can perform both classification and regression problems. The purpose of SVM is to find a hyperplane that can divide samples from different classes. If the data can be separated linearly, it is called linear SVM. However, in real-life problems, it is difficult to find a linear relationship. When data cannot be linearly separable, a non-linear kernel function of SVM can be used. Kernel functions can be linear, polynomial, RBF (radial basis function), and sigmoid. SVM is an efficient way for high-dimensional spaces. One of the disadvantages

of SVM is that in cases where the number of features is much greater than the number of samples, it can cause overfitting.[6] The data can be divided by some contamination samples, it is called soft margin. Soft margin provides better generalization. One of the crucial parameters is the C value which relates to the penalty for incorrect predictions. Lower C values indicate small penalty or large margins for support vectors. Another parameter of SVM is gamma which can be beneficial for model performance.[7]

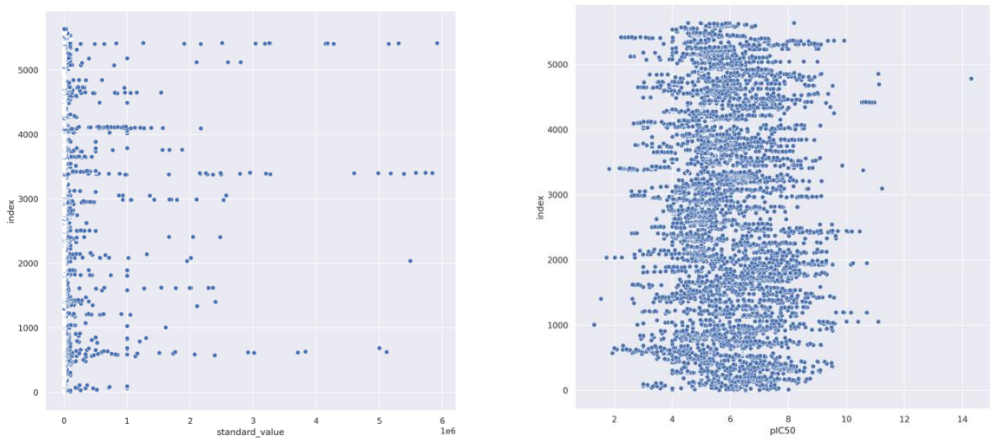


Figure 1. Distribution of IC_{50} and pIC_{50} values

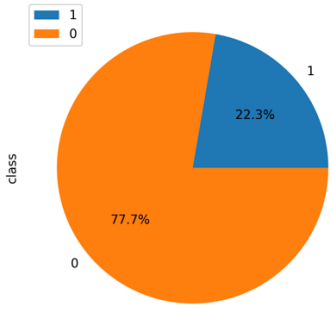


Figure 2. Percentage of active and inactive molecules in the dataset

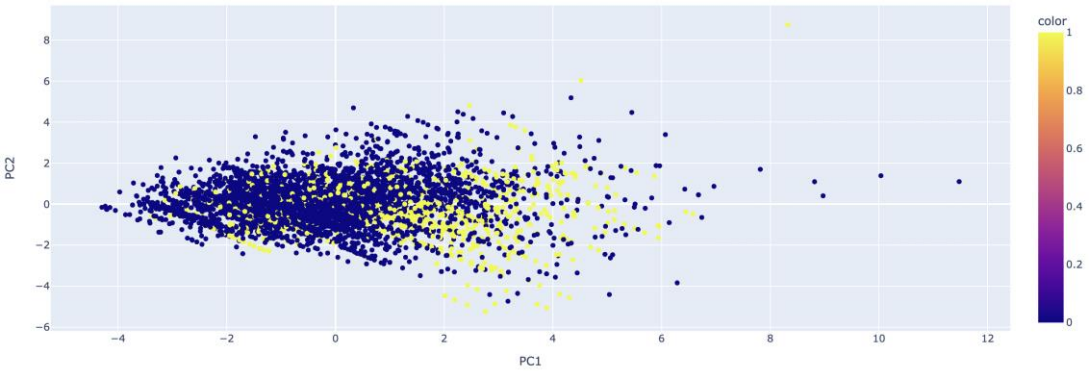


Figure 3. PCA result of molecule descriptors. Explained variance ratio of PC1:0.52 and PC2:0.21. Yellow: active molecules, blue: inactive molecules

Deep learning is a subfield of machine learning that uses algorithms called neural networks (NN). Neural networks are information processing systems with interconnected neurons. Using certain algorithms, it can recognize, cluster, and classify hidden patterns and correlations in raw data. In generally, NN are composed of three types of layers: input, hidden, and output layers. Neurons are nodes in the network

through which data and computations flow. The input, hidden and output layers consist of neurons. The input layer feeds by data although data must be numerical. Layers after the input layer are called hidden layers because they don't have a direct the data connection. Hidden layers have neurons which use activation functions. The activation function is the function that decides how a linear function calculated using the inputs of a neuron will exit the neuron. It is used to distort the linearity and determines how the linearity will be distorted. Artificial neural network (ANN), recurrent neural network (RNN) and convolution neural network (CNN) are the most popular types of neural networks.[8,9]

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework which builds a sequential series of smaller trees where each tree corrects for the residuals in the predictions made by all the previous trees. XGBoost provides a parallel tree boosting (also known as GBDT, GBM).[10] XGBoost, is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges.[11] XGBoost has various parameters which crucial for model building. Eta, max_depth, colsample_bytree, colsample_bylevel and objective function are some of these parameters.[12]

2.3 Molecular Descriptors

Molecules need to be transformed into mathematical expressions for machine-readable representations. Thus, molecules have an ideal form for computation by machines. There are different molecule representations such as fingerprints and SMILES string.[13] These representations can be categorized by their dimensionality. The information maintains from empirical formulas such as molecular weight; atom number are 0D descriptors. They don't provide sufficient knowledge of chemical structure. Fingerprints are more useful for providing knowledge about chemical structure. Fingerprints are binary vectors with each dimension in the vector indicating the existence or nonappearance of a particular substructure. Substituent atoms, chemical bonds, and functional groups that represent molecules are examples of 1D descriptors. Moreover, there are 2D descriptors to represent the atom connectivity and molecular topology, for instance, Molecular ACCess System (MACCS), Path-based fingerprints and Extended Connectivity Fingerprints (ECFPs).[14]

In this study, ECFP6 (2048 bits) were used for molecular representations. They are among the most popular fingerprints in drug discovery, and they are effectively used in a wide variety of applications. There are four steps involved in generating ECFP features for a molecule. First, each atom in the molecule is assigned a unique integer identifier. This identifier is generated by hashing a combination of properties like atomic numbers, atomic mass, etc. In the beginning, the initial atom identifier only represents information about the atom itself and its attached bonds. Next, each atom collects its identifier and the identifiers of its immediately neighbouring atoms, into an array. A hash function is applied to reduce this array back into a new, single-integer identifier. This step is done to capture the neighbourhood of the atom. Once all atoms have generated their new identifiers, the old identifiers are replaced with the new ones. This updating process is done multiple times iteratively: In the initial iteration, capturing individual atomic properties, the next iteration, accounting for its neighbours, and the next one, neighbours of neighbours. The third stage removes the duplicate features from our generated feature list. Once all the identifiers are calculated for a specified number of iterations, the final step is to reduce these identifiers into a bit vector.[15]

One of the advantages of ECFPs is, that they are not predefined and can represent a huge number of different molecular features (including stereochemical information). Also, they can be computed swiftly.

3. Results and Discussion

In this study, the DNN model was built with an input layer, 3 hidden layers (1024, 512, 16) and an output layer. ReLu function was used for hidden layers and the sigmoid function was used for the output layer. There are batch normalization and dropout after all hidden layers (Figure 4) (Table 1). XGboost model was built with these parameters; n_estimators=300, eval_metric=auc, booster=gbtree, learning_rate=0.01, max_depth=10. (Table 2) C and gamma parameters in SVM are crucial for training the model. In this study, C value was set 10, and gamma was set 0.4. (Table 3)

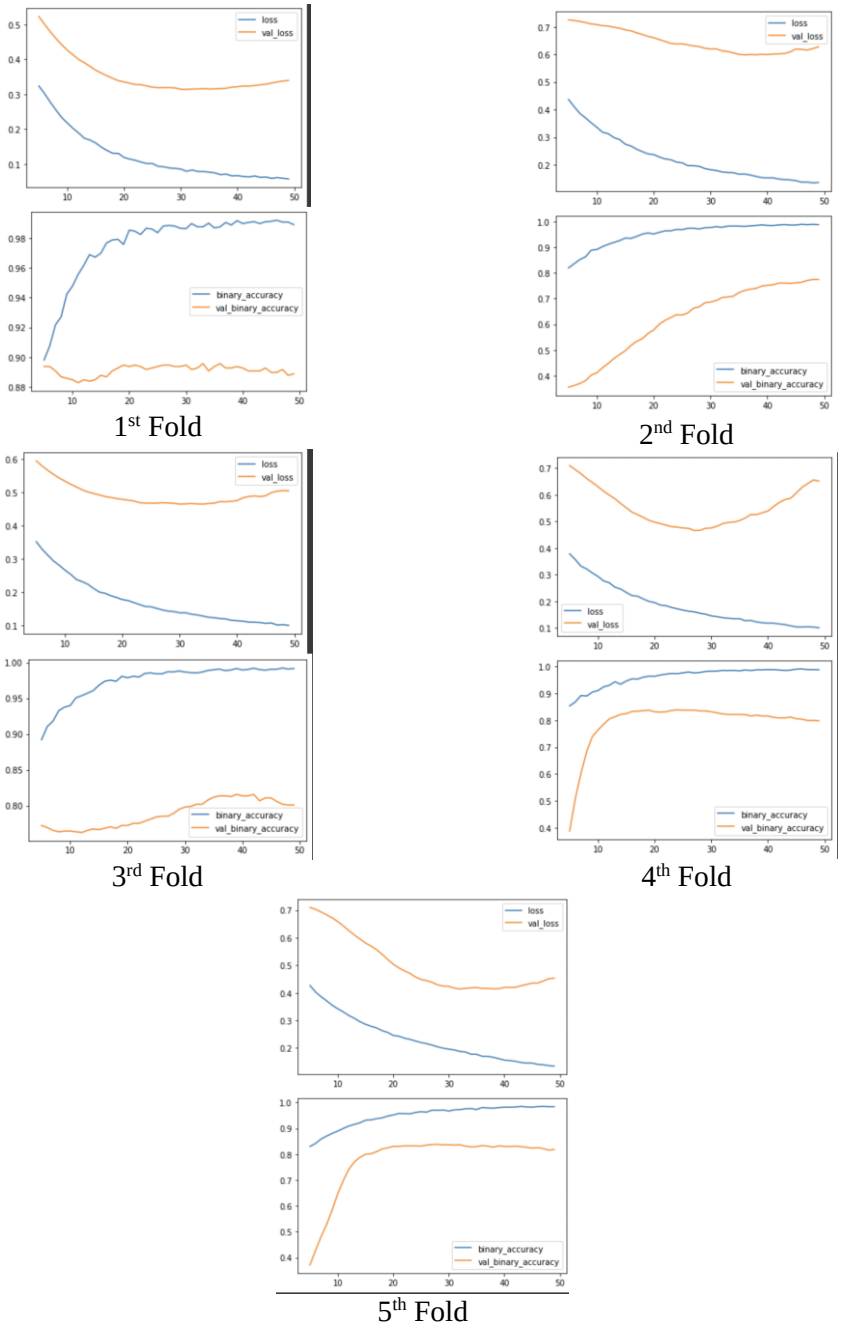


Figure 4. Binary cross entropy (loss function) and binary accuracy (metric) values for each fold

Table 1. Evaluation of the DNN model in the validation set for each fold

Validation sets	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold
Accuracy	0.80	0.77	0.88	0.79	0.81
Roc-Auc	0.66	0.79	0.74	0.76	0.73
F1 Score	0.48	0.73	0.62	0.55	0.58
MCC	0.38	0.56	0.58	0.44	0.46

Comparing the evaluation of the test set and validation set, the discrepancy is obvious. These results show that dataset splitting has a significant role in models' performance. Also, the numbers of active and inactive molecules in the training set, validation set, and test set should be carefully arranged to avoid imbalance. Finally, the DNN model showed slightly better performance than the other models.

Table 2. Evaluation of the XGBoost model in the validation set for each fold

Validation sets	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold
Accuracy	0.76	0.74	0.85	0.79	0.80
Roc-Auc	0.53	0.67	0.73	0.68	0.68
F1 Score	0.51	0.68	0.74	0.67	0.69
MCC	0.13	0.42	0.49	0.35	0.39

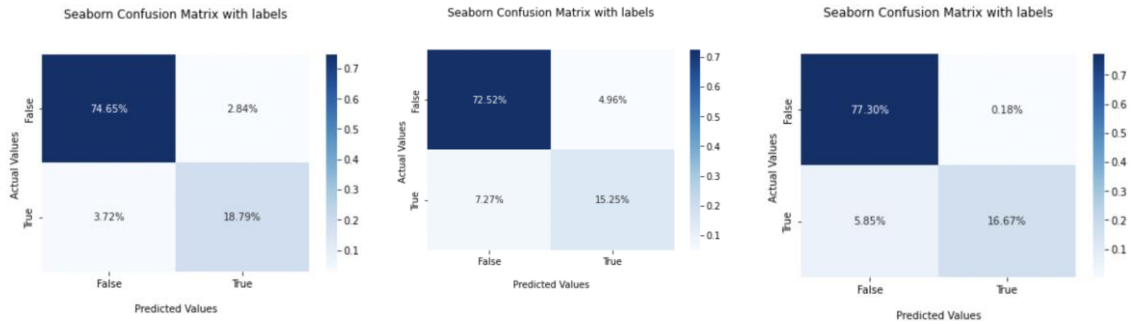
Table 3. Evaluation of the SVM model in the validation set for each fold

Validation sets	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold
Accuracy	0.79	0.69	0.84	0.83	0.78
Roc-Auc	0.56	0.57	0.58	0.53	0.51
F1 Score	0.79	0.69	0.84	0.83	0.78
MCC	0.32	0.28	0.36	0.23	0.14

Table 4. Evaluation of the models in test set

Models	Accuracy	F1 score	MCC	ROC-AUC
DNN	0.93	0.88	0.80	0.89
XGBoost	0.87	0.71	0.63	0.80
SVM	0.93	0.84	0.82	0.86

After training models and evaluating validation sets, the test set was evaluated. Evaluation scores for each model showed in Table 4 and also confusion matrix can be seen in Figure 5.



A B C

Figure 5. A) Confusion matrix of DNN model B) Confusion matrix of XGBoost model C) Confusion matrix of SVM

4. Conclusion

Diseases because to dysregulation of the cholinergic system such as Alzheimer's and multiple sclerosis are increasing recently. For this purpose, machine learning is a critical tool to discover more effective molecules with low cost and fewer side effects. Also, discovering new molecules can be democratized with machine learning all around the world.

In this study, three different machine models were developed and, these models (DNN, SVM, XGBoost) can be used for predicting AChE inhibitors as active or inactive. There is another study in 2021, Prabha Garg et al. developed SVM, k-nearest neighbour, and random forest to predict AChE inhibitors. They downloaded 5692 molecules from BindingDB (we downloaded from ChEMBL). They decided on 1000 nM as the activity threshold IC50 value (we converted pIC_{50} value). Their best accuracy scores are 0.84 for the validation set and 0.85 for the test set.[16] Our best accuracy scores are 0.88 for the validation set and 0.93 for the test set. The important differences between these studies are descriptors and machine learning algorithms.

Our results showed that novel molecules can be designed effectively with fewer experimental structure-activity relationship studies. These models will use in further works to discover novel acetylcholinesterase inhibitors.

Acknowledgements

I would like to express my very great appreciation to Prof. Dr. Oztekin Algul for his valuable and constructive suggestions during the planning and development of this research work.

ORCID
Mehmet Ali Yucel: 0000-0003-2880-7992

References

[1] Walczak-Nowicka, ŁJ.; Herbet M. Acetylcholinesterase inhibitors in the treatment of neurodegenerative diseases and the role of acetylcholinesterase in their pathogenesis. *Int. J. Mol. Sci.* **2021**, *22*(*17*), 9290, doi: 10.3390/ijms22179290.

[2] Chen, Z.R.; Huang, J.B.; Yang, S.L.; Hong, F.F. Role of cholinergic signaling in Alzheimer's disease. *Molecules* **2022**, *27*(*6*), 1816, doi: 10.3390/molecules27061816.

[3] Gatta, V.; Mengod, G.; Reale, M.; Tata, A.M. Possible correlation between cholinergic system alterations and neuro/inflammation in multiple sclerosis. *Biomedicines* **2020**, *8*(*6*), 153, doi: 10.3390/biomedicines8060153.

[4] Bari, M.; Pinto, G.; Reale, M.; Mengod, G.; Tata, A.M. Cholinergic system and neuroinflammation: implication in multiple sclerosis. *Cent. Nerv. Syst. Agents Med. Chem.* **2017**, *17*(*2*), 109-115.

[5] Deng, J; Yang, Z; Ojima, I; Samaras, D; Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Brief. Bioinform.* **2022**, *23*(*1*), 430, doi: 10.1093/bib/bbab430.

[6] Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. *Expert Opin. Drug Discov.* **2014**, *9*(*1*), 93-104.

[7] Maltarollo, V.G.; Kronenberger, T.; Espinoza, G.Z., Oliveira, P.R.; Honorio, K.M. Advances with support vector machines for novel drug discovery. *Expert Opin. Drug Discov*. **2019,** *14*(*1*), 23-33.

[8] Zhenqin, W.; Bharath, R.; Evan, N. F.; Joseph, G.; Caleb, G.; Aneesh, S. P.; Karl, L.; Vijay, P. MoleculeNet: benchmark for molecular machine learning, *arXiv* **2017**, 1703.00564, doi: 10.1039/C7SC02664A.

[9 Lane, T; Russo, D.P.; Zorn, K.M. Comparing and validating machine learning models for mycobacterium tuberculosis drug discovery. *Mol Pharm.* **2018**, *15*(*10*), 4346-4360.

[10] Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme gradient boosting as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model*. **2016**, *56*(*12*), 2353-2360.

[11] Tianqi, C.; Carlos, G. XGBoost: a scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA*, **2016**, 785–794.

[12] Minjian, Y.; Bingzhong, T.; Chengjuan, C.; Wenqiang, J.; Shaolei, S.; Tiantai, Z.; Xiaojian, W. Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 inhibitors. *J. Chem. Inf. Model.* **2019**, *59*(*12*), 5002-5012.

[13] David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminfor*. **2020**, *12*(*1*), 1–22.

[14] Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Brief. Bioinform*. **2022,** *23*(*1*), 430, doi: 10.1093/bib/bbab430.

[15] David, R.; Mathew, H. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*(5), 742-754.
[16] Sandhu, H.; Kumar, R.N.; Garg, P. Machine learning-based modeling to predict inhibitors of acetylcholinesterase. *Mol. Divers.* **2022**, *26*(*1*)**,** 331–340.