**Fall 2023 CS/BIOL 123A Bioinformatics Project Report**

**PROJECT TITLE**: Comprehensive Quantitative Structure-Activity Relationship model utilizing Machine Learning to predict the bioactivity of molecular inhibitors against Thrombin.

**NAME TEAM MEMBER 1** Ananya Gupta, MS Bioinformatics

**NAME TEAM MEMBER 2** Shwethal Sayeeram Trikannad, MS Bioinformatics

**NAME TEAM MEMBER 3** Manh Tuong (Eddie) Nguyen, BS Data Science

**NAME TEAM MEMBER 4** Jennifer Melendez, BS Molecular Biology

# **ABSTRACT**

The results of predictive accuracy of a quantitative structure-activity relationship (QSAR) random forest regressor model against other regression models for identifying and classifying thrombin antagonists are declared. Performance was compared and reported for 1064 inactive and 921 active potential lead drug candidates obtained from the ChEMBL database. The DecisionTreeRegressor performed the best with the RMSE value of 0.605590054218893 on the training dataset. It also had an R-squared value of 0.870199002143443. The NuSVR model performed the best on the test dataset whose RMSE and R-squared values were 1.41171639451711, 0.353901904966875 respectively. The RMSE value of Random Forest was around 1.47.This work clarifies the complexities of QSAR modeling by emphasizing the value of careful model evaluation on various test datasets. The results suggest that more investigation is needed to improve the interpretability and accuracy of the model, maybe by adding more features or using different modeling strategies, including using particular chemical descriptors for thrombin interaction.

**CONTENTS**

[**ABSTRACT**](#_sqs5e9rqt3lu) **1**

[**LIST OF FIGURES**](#_iy3wjtuqgec) **3**

[**LIST OF TABLES**](#_5ucxvzvrf1pf) **4**

[**INTRODUCTION**](#_vybxhgiyofi3) **5**

[**BACKGROUND**](#_5psb6h2e2n2q) **8**

[**DATA COLLECTED/ACCESSED**](#_uyh1icze8pi4) **10**

[**APPROACH AND METHOD**](#_xqr8nx6z4n94) **11**

[**EVALUATION OF RESULTS**](#_86g45eji4hgq) **15**

[**CONCLUSION AND DISCUSSION**](#_5e85geakz0iy) **22**

[**FUTURE WORK**](#_eazpy5d8qi3r) **23**

[**REFERENCES**](#_x4ch91qkk2t) **24**

[**TEAM MEMBER CONTRIBUTION:**](#_qmnh6w7e4w2b) **29**

# **LIST OF FIGURES**

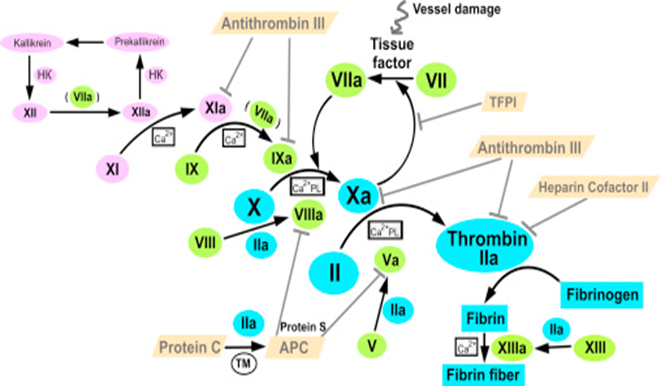
|  |  |  |
| --- | --- | --- |
| **SERIAL NUMBER** | **FIGURE** | **PAGE NUMBER** |
| 1. | Coagulation cascade in the human body | 5 |
| 2. | Typical flowchart depicting steps in drug development and discovery | 7 |
| 3. | Plot depicting relationship between logP and molecular weight of compounds | 12 |
| 4. | Plot showcasing standard value before and after transformation | 13 |
| 5. | A section of the final usable data frame | 13 |
| 6. | Mann–Whitney U test for pIC50, MW, NumHdonors, NumHacceptors and LogP | 16 |
| 7. | Barchart for RMSE of regression models | 19 |
| 8. | Barchart for R-square of regression models | 19 |
| 9. | Barchart for Time of regression models | 20 |

# **LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **SERIAL NUMBER** | **DESCRIPTION** | **PAGE NUMBER** |
| 1. | Results of Mann-Whitney U test for inactive and active groups | 14 |
| 2. | Predictions on train | 17 |
| 3. | Predictions on test | 18 |

# **INTRODUCTION**

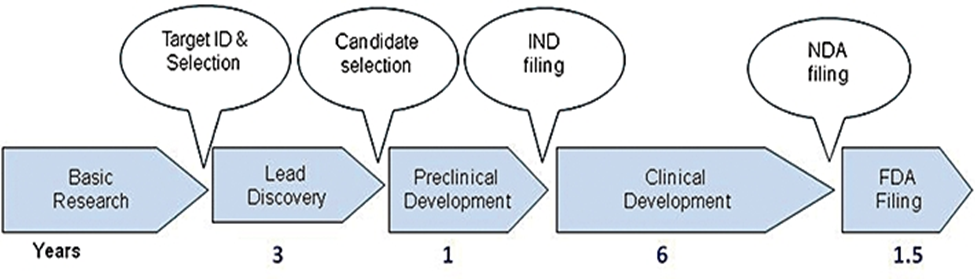
For a long time, Vitamin K antagonists were the sole oral anticoagulants available commercially but their limitations like narrow therapeutic window, delayed onset of action among others render them tedious to administer [1]. A protein molecule thrombin arose as the alternative target molecule for anticoagulant medication. The function of thrombin is to help in the clotting mechanism, as shown in Fig.1, of the human body by converting fibrinogen to fibrin and activating clotting factors XIII, V, VIII,XI and platelets [2][3] rendering it essential in hemostasis



**Fig 1. Coagulation cascade in the human body Source: DOI:**[**10.1016/j.bbrep.2015.11.011**](http://dx.doi.org/10.1016/j.bbrep.2015.11.011)

The therapeutic window of a drug is specific to its concentration and is the product of a delicate balance between thrombosis and bleeding [4]. Drug discovery and associated development are convoluted, painstaking procedures. Immense advances in technology and the advent of artificial intelligence (AI) provide the pharmaceutical industry with a plethora of opportunities to ease this process. The world of AI has had a significant shift from studies to actual industrial applications. Machine learning (ML) in its essence, is the utilization of algorithms created by programmers to parse through, comprehend, master, and make predictions off data. The quality and accuracy of results are directly proportional to the amount of training data input into the model. ML finds application in pharmaceuticals especially in the identification of drug targets [5]. Quantitative structure activity relationship (QSAR) is a statistical process of applying models that relate the chemical structure of compounds to their bioactivity. Variables in the form of numerical descriptors are used to represent the chemical structure and then

analyze the relationship between the same with the bioactivity of the compound in question using predictive models [6]. QSAR models grant us the power to create safe, potent, and efficacious drugs replacing the need to create new testable compounds cutting labor, time, and monetary losses [7]. Hence, our project chose to apply this important, interesting, high impact bioinformatics tool to identify drug targets for thrombin in the hopes of collecting and categorizing compounds as active or inactive which could then serve as the basis for potential anticoagulant medication in the pharmaceutical space. Thrombin, though having wound healing properties by partaking in the coagulation cascade, also has numerous pathogenicities linked to it. Widespread accumulation of thrombin in the form of plaques, amyloid deposits, and neurofibrillary tangles is noticed in the brains of patients suffering from Alzheimer's disease [8]. Animal models of Multiple Sclerosis displayed axonal loss associated with fibrin accumulation in cranial vessels [9]. Low quantities of the compound protect the neurons and astrocytes from stressors like hypoglycemia while higher concentrations promote death of the hippocampal and motor neurons [10,11,12]. This compound also makes the body more susceptible for central nervous system (CNS) infections like meningitis and human immunodeficiency virus related nervous complications, as it increases the ability of pathogens to cross the blood brain barrier [13]. Thrombosis, formation of an endogenous blood clot, when coupled with atherosclerotic plaques, especially unstable ones, lead to an increased predisposition to myocardial infarctions (MI) [14]. Thrombosis affects both veins (Deep Vein Thrombosis) and arteries (Ischemic stroke, Acute limb Ischemia) alike. Hormonal therapy, pregnant and postpartum women, long hours of travel, pre-existing conditions like Diabetes, Hypertension, obesity are associated with thrombosis in the deep veins of the leg especially the calf [15,16]. Conditions like DVT increase the predilection of ischemic stroke due to transfer of a dislodged clot into the cerebral vasculature. Thrombin inhibition is crucial to ameliorate such disastrous events. Thrombin inactivation is either direct or indirect based on the involved binding site. Heparin induces indirect activation by binding at the exosite no. 2 of thrombin but requires cofactors like antithrombin and additional saccharides to maximize its effect. Direct thrombin inactivators can bind to both free and fibrin-bound thrombin molecules. They also do not require any cofactor like antithrombin to function. Both parenteral (Bivalirudin, Lepirudin) and oral (Dabigatran etexilate) variants of anticoagulant medication are available. As with any drug the process involved in discovery and development goes through many phases namely target discovery, target validation, lead compound identification, lead compound optimization, preclinical development, and clinical trials as shown in Fig. 2. Depending on the type of drug and the disease in question these processes could sometimes span over a decade. Lead compound identification takes a long time and requires screening hundreds of potential candidates to find a few molecules that could then progress to the next phase. The expense incurred runs into millions.



**Fig 2. Typical flowchart depicting steps in drug development and discovery.**

**Source: doi:** [**10.1111/j.1476-5381.2010.01127.x**](https://doi.org/10.1111%2Fj.1476-5381.2010.01127.x)

Addressing this problem with the help of artificial intelligence is thus necessary to cut costs while also ensuring the quicker release of a much-needed drug to combat the conditions linked to thrombosis. Machine learning is increasingly being used in numerous stages of drug design such as identifying targets [5], improving small-molecule compound design and optimization and developing new markers for drug efficacy [17]. There are two main types of ML techniques: supervised and unsupervised learning. Supervised learning methods develop training models for predictive purposes on a known input dataset, whereas unsupervised methods are used for exploration and clustering of the data. This project aims to create a QSAR model with supervised machine learning to identify and categorize potential drug candidates for thrombin. One of the major benefits of this proposed solution is significant enhancement in efficiency, particularly when dealing with substantial volumes of data. Robust machine learning models can analyze thousands of data points in a condensed time frame enabling high throughput screening of potential drug candidates against thrombin. Compounds with a reduced probability of success are eliminated at a very early stage leading to huge economic gains. QSAR models provide the researcher with a comprehensive insight into the relationship between bioactivity of a compound and its structure which is key in designing new drug candidates. ML also eases the burden on the researchers by alleviating the need to create physical testable samples. Another benefit of utilizing ML is the fact that databases everywhere are curated with the latest advancements ensuring researchers access to state-of-the-art developments in real time. This tool also behaves as an aid to in vitro testing by providing data-driven actionable insights, thus optimizing, and enhancing the whole drug development process from the initial stage. QSAR models also considerably increase the chances of identifying novel drug candidates by allowing scientists to explore structurally eclectic compounds which might be difficult to find traditionally. Machine learning also assists in the idea of making personalized precision medication a reality as these algorithms function on a diverse range of input variables. There is no restriction on the number and type of data and accuracy essentially improves to provide the best possible results. Medication can be tailor made to suit the patient’s specific genetic and molecular makeup thus increasing chances of better drug interaction and improved efficacy in treatment of disease. QSAR models can be integrated with differing data sources such as omics data improving the understanding of the complex biochemical interactions involved in thrombin inhibition and guides in identifying multi-target drug candidates.

# **BACKGROUND**

Quantitative structure activity relationship (QSAR) was developed over 60 years ago and continues to be a very important aspect of drug discovery and development till date [18]. QSAR has persisted as an efficacious methodology for constructing mathematical models. These models aim to establish a statistically meaningful correlation between the chemical structure and continuous attributes (such as pIC50, pEC50, Ki, etc.) or categorical/binary descriptors (including active, inactive, toxic, nontoxic, etc.) of biological or toxicological relevance. This correlation is pursued through the application of regression techniques for continuous properties and classification techniques for categorical attributes [19]. In machine learning, a diverse array of methodologies is generally contemplated, encompassing techniques such as Random Forest, Naive Bayesian Classification (NBC), Multiple Linear Regression (MLR), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Probabilistic Neural Networks (PNN), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), among others [20]. Computational intelligence provides several methods of analysis and learning in the context of drug development, highlighting the AI driven procedures used to find a range of drugs methodically and seamlessly [21]. In recent times, QSAR modeling encompasses the adept application of various machine learning techniques to model and conduct virtual screening on extensive datasets, featuring a multitude of diverse chemical structures [22,23]. Modern QSAR models implement auxiliary features like set of empirical rules (eg. Lipinski’s rules) [24], chemical feasibility [20] among others. Random forest QSAR models were built to predict ligand activity toward targets and rank the targets for a specific ligand [25]. Antagonists and agonists of Epidermal growth factor receptor, a cancer drug target, were categorized utilizing random forest QSAR models for a vast set of compounds spanning diverse classes [26]. 4-aminopyrimidine-5-carbaldehyde oxime with potent inhibition of vascular endothelial growth factor receptor 2 (VEGFR2) has been found utilizing QSAR associated with an SVM model [27]. QSAR models utilizing SVM have also been implemented to predict HIV protease inhibitors and were revealed to be superior to Multiple linear regression (MLR) models [28]. A gradient boosting algorithm for QSAR models to predict drug blockade of the Human Ether a-go-go related Gene (hERG1) channel has been built [29]. Comparative analysis of various machine learning QSAR models has been carried out to predict the inhibitory constant of thrombin antagonists among which SVM emerged superior [30]. A two-stage machine learning model composed of several classifier and regression models have been deployed to predict peptide thrombin inhibitors resulting in the creation of a dataset of potential direct thrombin inhibitor drug candidates [31]. Our project differs in the fact that we are utilizing QSAR along with machine learning to create a random forest regressor model and then comparing its performance against 41 other regression models to predict and classify thrombin inhibitors accessing the CHEMbl database, an integrative repository which incorporates the chemical, bioactivity, and genomic data of drug-like compounds under a singular framework.

# **DATA COLLECTED/ACCESSED**

For a project with the objective of finding thrombin antagonists with QSAR machine learning models we decided on a few key features for the type of dataset to be used. One of them included comprehensive molecular descriptors. These descriptors represent the chemical structure of the potential drug candidates and the set of descriptors we chose for this purpose was the Lipinski descriptor array. Biological activity denoting structure-activity relationship within the dataset displayed as quantitative units was the second feature we decided on. For our project IC50 (half maximal inhibitory concentration) was chosen as the bioactivity measure. A third feature we believed was crucial to forming our dataset was consistency of input data into the QSAR models. This would establish the fact that same molecular structures would deliver the same input ensuring the reliability, accuracy and reproducibility of the machine learning models in use. To fulfill this need we resolved to choose compounds with canonical smiles representations. Finally, we wanted to incorporate a chemically diverse set of compounds that would improve the model’s predictive abilities while also recording a broad spectrum of biochemical compounds. Considering the complexity of our models, number of input features, diversity in chemical structures, bioactivity measure and limited computational resources our team believed a few thousand compounds would be sufficient to conduct the tests on. The Lipinski rule of 5 is recognized as the standard for medicinal chemists and is especially relevant to our project as factors like molecular weight, lipophilicity, hydrogen bond donors, and acceptors help assess drug-likeness. These properties guide researchers in the direction of compounds that are suitable for administration. Compounds that flout the Lipinski rule tend to have poorer absorption and distribution within the human body. IC50 is relevant to our objective as it indicates the concentration of the compound required to antagonize 50% of thrombin activity. It also helps map out the fluctuations in inhibitory effects versus substrate concentration and is pivotal in predicting the therapeutic benefits a specific compound may potentially have. Canonical SMILES (Simplified Molecular Input Line Entry System) is a standard representation of the structure of any molecule using ASCII characters. As compared to SMILES, canonical SMILES certify a unique representation of each molecule thus contributing to consistency in the data. This in turn, is essential in maintaining the standardization of input data into QSAR machine learning models and thus is relevant to our project. The ChEMBL database [32] is a directory created by manually curating compounds from peer reviewed scientific papers in research journals. The substrates here matched our requirements hence, ChEMBL was our repository of choice to query, collate and create the dataset utilized in this project.

# **APPROACH AND METHOD**

The workflow was inspired by the approach of utilizing QSAR modeling and molecular docking on human acetylcholinesterase [33]. The ultimate objective is to build a machine-learning model that can work with a molecular data set and output the classified potential drug candidates against thrombin. The approach we used is Data Collection -> Data Processing -> Exploratory Data Analysis -> Training Model -> Comparing model [34].This project uses Python 3.9 and the following Python libraries were installed:

· [NumPy]

· [Pandas]

· [matplotlib]

· [seaborn]

· [chembl\_webresource\_client]

· [scikit-learn]

· [rdkit]

· [LazyPredict]

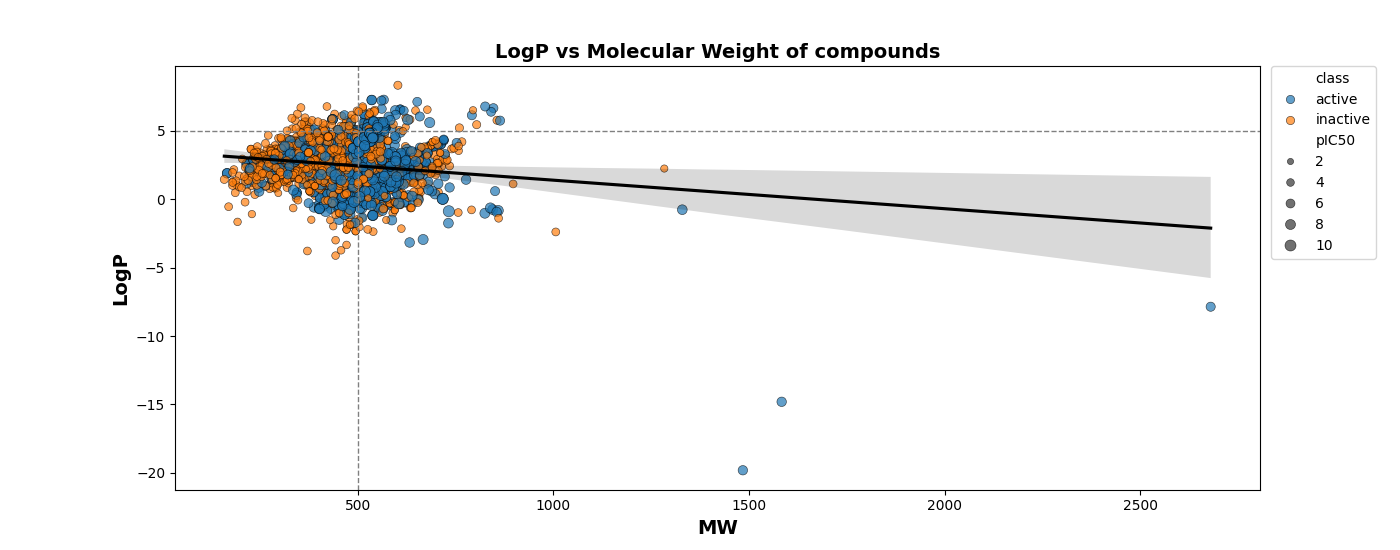
The Anaconda distribution of Python was used, which already has the above packages and more included. Pycharm IDE was used to collaborate and write code. A more detailed description of the steps are provided in the README file. ChEMBL is a unique chemical database emphasizing all fronts of drug discovery. ChEMBL includes information on the interaction of small molecules with respective protein targets, effects on the cells and organism involved and the pharmacokinetics of the substrates stored in this database. Additionally ChEMBL holds molecular properties like Lipinski’s parameters, bioactivity data like IC50 and standardized two-dimensional structures represented by canonical SMILES. Hence, we chose ChEMBL as our database of choice. After a programmatic search of the ChEMBL database for our target of interest we isolated bioactivity data of the thrombin target with ChEMBL ID `ChEMBL204` reported as IC50 values in nM (nanomolar) units. We dropped null and duplicate rows for canonical smiles. On the basis of IC50 standard values the compounds were classified as inactive, active and intermediately active. For the active group, candidates must have a standard value less than or equal to 1000 nM. and substrates with a standard value greater than 10000 nM are deemed inactive. In general, the values that we were interested in were molecule ID, canonical SMILES, and standard value. In this project we will only consider active and inactive compounds. After dropping null and duplicate rows for canonical smiles, the final data frame consisted of 3034 potential drug leads which were binned into active, inactive and intermediate compounds on the basis of IC50 values. In this project we will only consider active and inactive compounds. Lipinski’s rule of 5 or Lipinski’s law states the following:

· The number of hydrogen bond donors (OH, NH) is 5 or less.

· The number of hydrogen bond acceptors (N, O, etc.) is 10 or less.

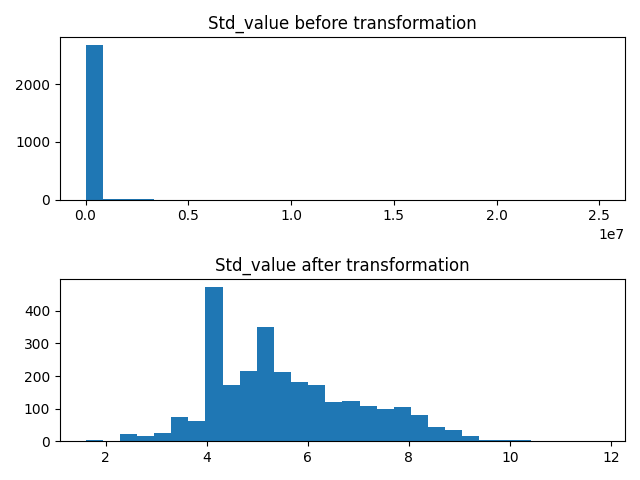
· Octanol-water partition coefficient (LogP) is 5 or less.

· Molecular weight is 500 or less.



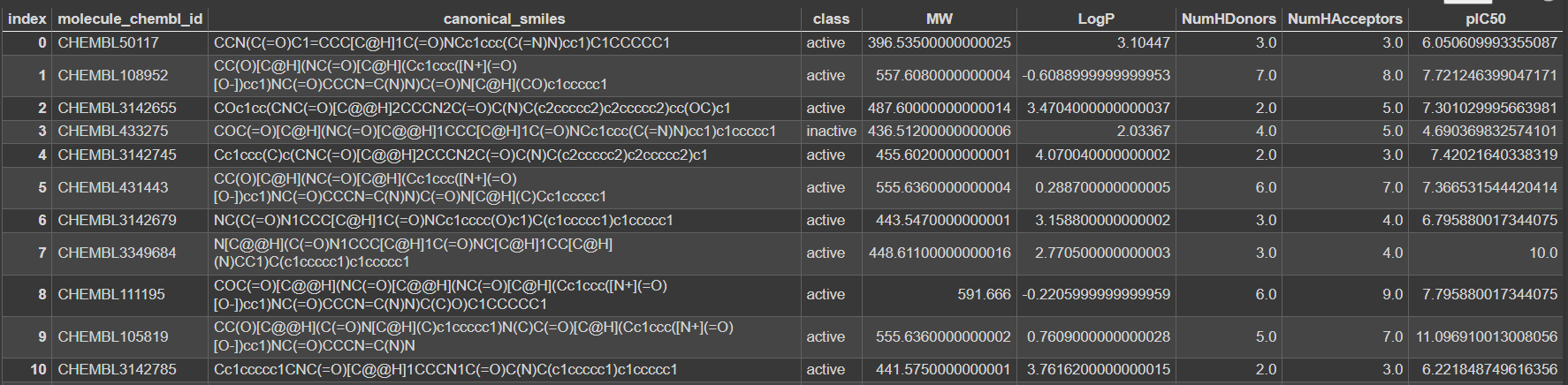
**Fig 3. Plot depicting relationship between logP and molecular weight of compounds**

Fig. 3 depicts the relationship between logP and molecular weight of compounds and there are a wide range of LogP and MW values within each class, but generally, active compounds cluster towards lower MW and a LogP closer to zero, whereas inactive compounds are spread out over a broader range of LogP values and generally have higher MW. The calculation of Lipinski’s descriptors required only a part of the canonical SMILES of the molecules which was filtered out with the help of a simple loop. A custom function for extracting Lipinski’s descriptor of each element was employed in the dataset using rdkit. Normalization of IC50 to pIC50 was carried out to ensure uniform distribution and improve the readiness of the standard value.



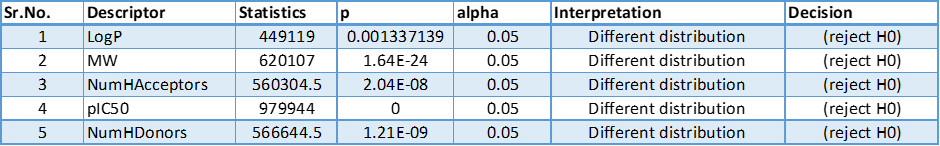
**Fig 4. Plot showcasing standard value before and after transformation**

The IC50 values from the standard value column were converted from nM to M by multiplying the value by 10-9. Then, the molar value underwent a -log10 transformation.Note that all IC50 values larger than 108 were set as 100000000 to prevent negative values. Fig 3. Shows the plot for the normalization. Fig. 4. depicts a portion of the final data frame displaying molecule id, canonical smiles, lipinski’s molecular descriptors and pIC50 values of the testable drug candidates. Pandas’ DataFrame is an effective solution for temporarily storing data due to its compact and easy-to-use methods to work with tables that we can utilize to clean null values, drop/add rows or columns, imperatively, it is the universal data type that most Python packages require, and thus was utilized.



**Fig. 5. A section of the final usable data frame**

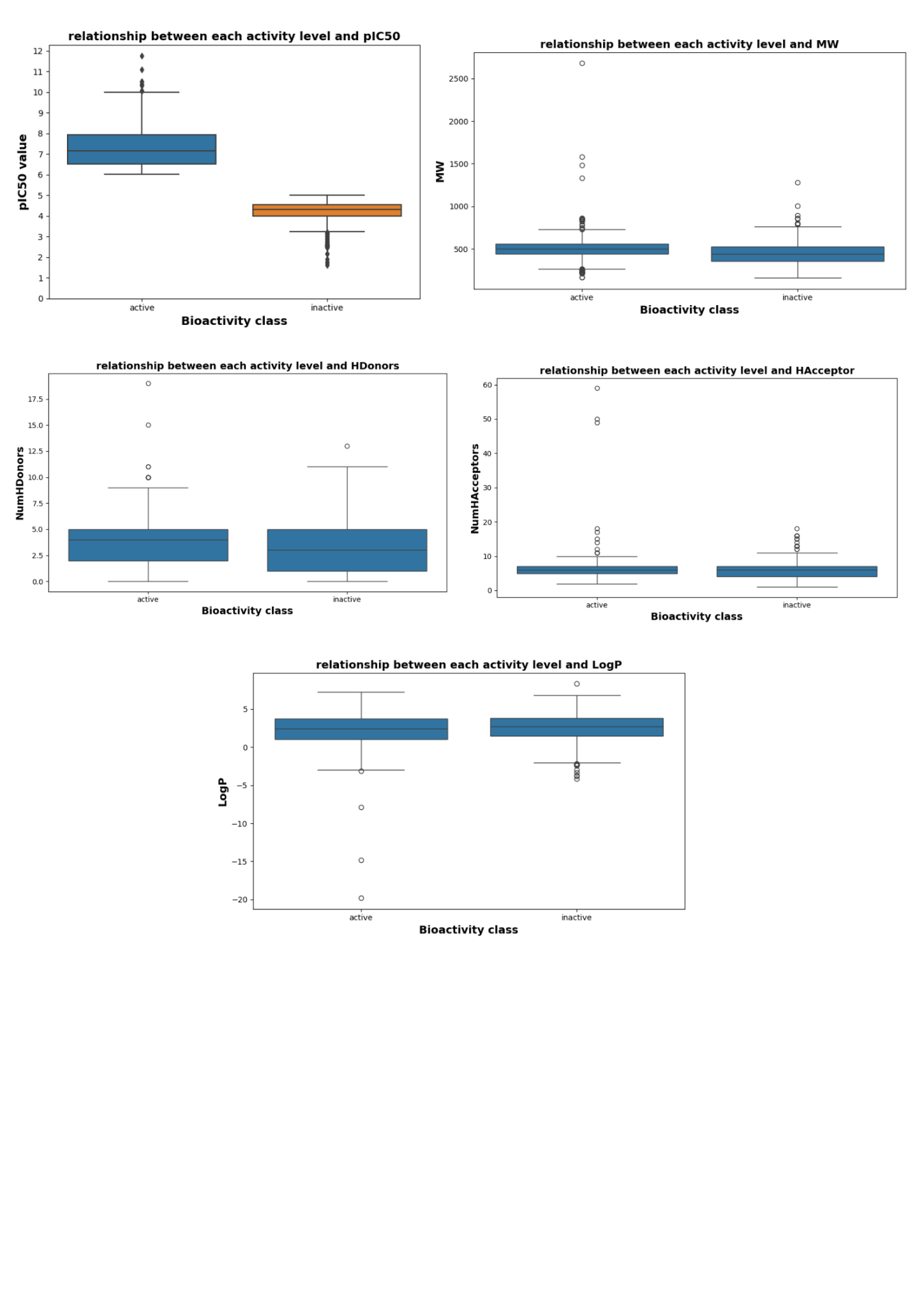
To discern the impact of each descriptor on bioactivity, statistical tests were conducted. The Mann-Whitney U test was particularly employed to compare the distribution of descriptors between active and inactive compounds, thus identifying descriptors with significant discriminatory power.This nonparametric test ascertains whether there exists a significant disparity in the dependent variable between two independent groups. Its purpose is to assess whether the distribution of the dependent variable is uniform across both groups, thereby indicating conformity in the populations from which the respective groups are drawn. The results of the Mann-Whitney U test are reported in table 1.

 **Table 1. Results of Mann-Whitney U test for inactive and active groups**

The model utilized fingerprint PaDEL as they are a representation of the molecule’s geometry and chemical structure of the substrate as compared to Lipinski’s descriptors which describe the molecule’s physicochemical properties. The PaDEL descriptor [35] is composed of 1875 distinct types of descriptors. These can be categorized as one-dimensional (1D) descriptors (which count the number of specific groups or atoms), two-dimensional (2D) descriptors (which are graph invariants and measure molecular properties), and three-dimensional (3D) descriptors (which are based on geometry). By definition, the PaDEL descriptor contains more information about the spatial geometry and atomic connections of the molecules in question [36]. These descriptors included, but were not limited to, molecular weight, logp, and counts of hydrogen bond donors and acceptors. Each descriptor quantified a particular aspect of the chemical structure that could influence the interaction with biological targets. The main objective was to build a model to predict the potential drug candidate which warranted a regression model. Its good performance, scalability, and ease of use made random forest regressor, a great algorithm for machine-learning tasks. The algorithm is flexible and can naturally assign feature importance scores, allowing it to handle redundant feature columns. It can scale to large datasets with ease and is generally robust to overfitting. Additionally, the algorithm does not require the data to be scaled and can model a nonlinear relationship [37]. We believed this was the best model for this project.We used the PaDEL calculation package [38] to calculate the descriptors by feeding it the ChEMBL ID of the molecules and their canonical SMILES formula. Then we used a simple shell script to activate the calculation and store it in a different file without the column name. This data was the input of the algorithm and the output would be the pIC50 value. We desired to eliminate entries with low variance as they do not provide the model with useful information to learn patterns [39]. So, the last step before training was dropping low variance entries by using the VarianceThreshold method in Scikit-learn with the threshold (0.8 \* (1 – 0.8)) and using fit transform() to convert the data into a format that the model understands. The curated and feature-selected dataset was then split into training and test sets. An 80:20 split ratio was maintained, ensuring a substantial training dataset for model development while reserving enough data for validation purposes. The model was trained on the training dataset, fine-tuning its parameters to maximize predictive performance. Predictions were made on the test set, and the model's performance was evaluated using R-squared and Root Mean Square Error (RMSE) metrics. R-squared provided insight into the proportion of variance explained by the model, while RMSE offered a measure of prediction error. However, a comparison with other models was important to validate the best model for the objective of identifying potential lead thrombin antagonists.The model's performance was compared against other regression models, automated by using the LazyRegressor tool. This facilitated a comprehensive comparison based on RMSE, computation time, and R-squared values and adjusted R-squared values (modified version of R-squared value that has been adjusted for the number of predictors in the model) aiding in the selection of the most suitable model for bioactivity prediction of lead candidates for thrombin molecule thus helping achieve outline project objective and goal.

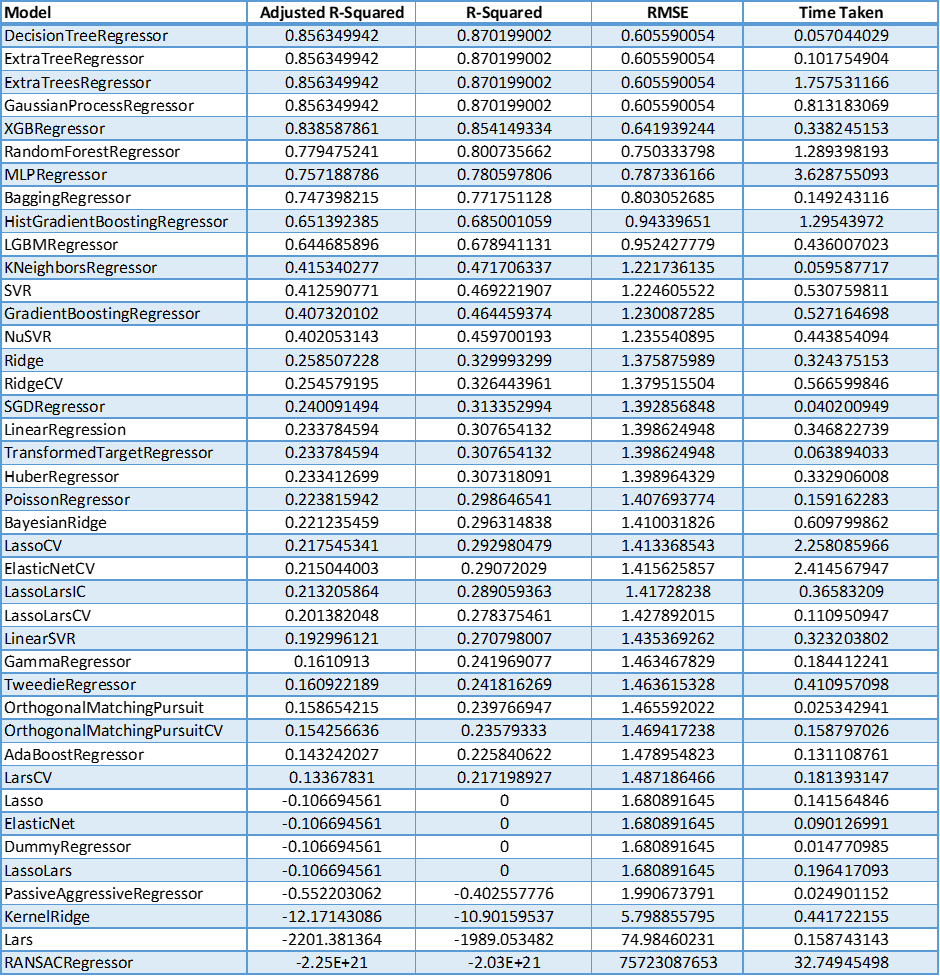
# **EVALUATION OF RESULTS**

Within the provided dataset, the first batch of compounds comprised 3330 examples, each with 881 characteristics. Following feature selection, which eliminated low variance descriptors, 154 descriptors were still included in the dataset, guaranteeing a more targeted and computationally effective modeling procedure.The results of the Mann-Whitney U and the plots suggest that active compounds are distinct from inactive compounds in their molecular descriptors. Fig. 6 depicts the results of the statistical test.



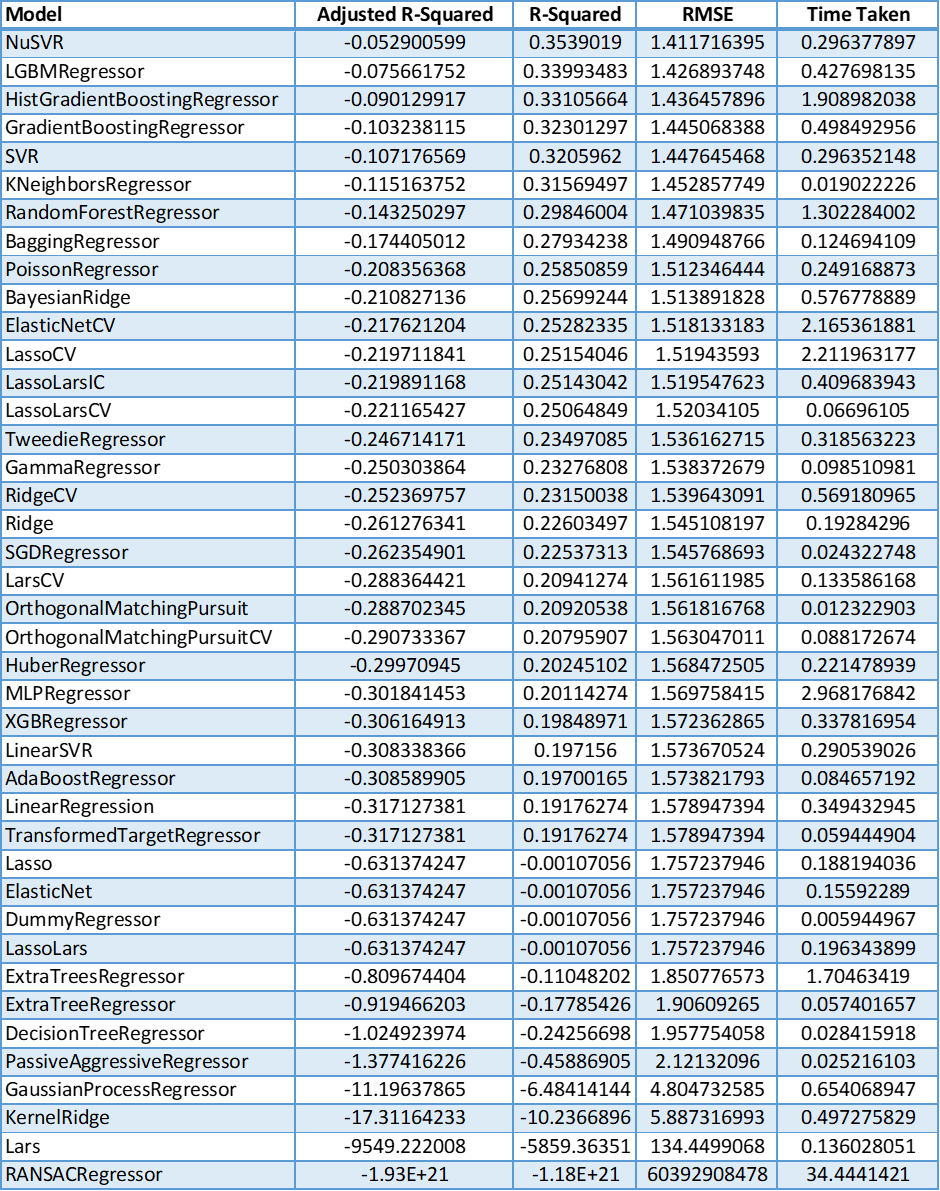
**Fig 6: Mann–Whitney U test for pIC50, MW, NumHdonors, NumHacceptors and LogP**

These differences are indicative of the molecular properties that influence a compound's bioactivity. The significant p-values across all tests reinforce the conclusion that the observed differences are unlikely to be due to random chance. ​Comparative QSAR modeling using PaDEL descriptors was carried out for 41 distinct regression models. The Root Mean Squared Error (RMSE) was used to help to measure the model's prediction error on the tested set. A lower RMSE value shows lesser variation from the actual value and thus indicates a more accurate model. Additionally, the R-squared (R²) metric provided a measure of the goodness of fit, reflecting the proportion of variance in the dependent variable that can be predicted from the independent variable(s). This measure was considered for both the training and test datasets to assess the models' performance and generalizability. The time taken for model training and evaluation was also recorded, reflecting the models' computational efficiency. The DecisionTreeRegressor was the model that performed the best with the RMSE value of 0.605590054218893 when predictions were made on the training dataset, demonstrating its ability to successfully identify patterns in the training data. It also had an R-squared value of 0.870199002143443. When this was adjusted for predictors the adjusted R-squared value also came up to be 0.85634994170268. The time taken by it was 0.0570440292358398.

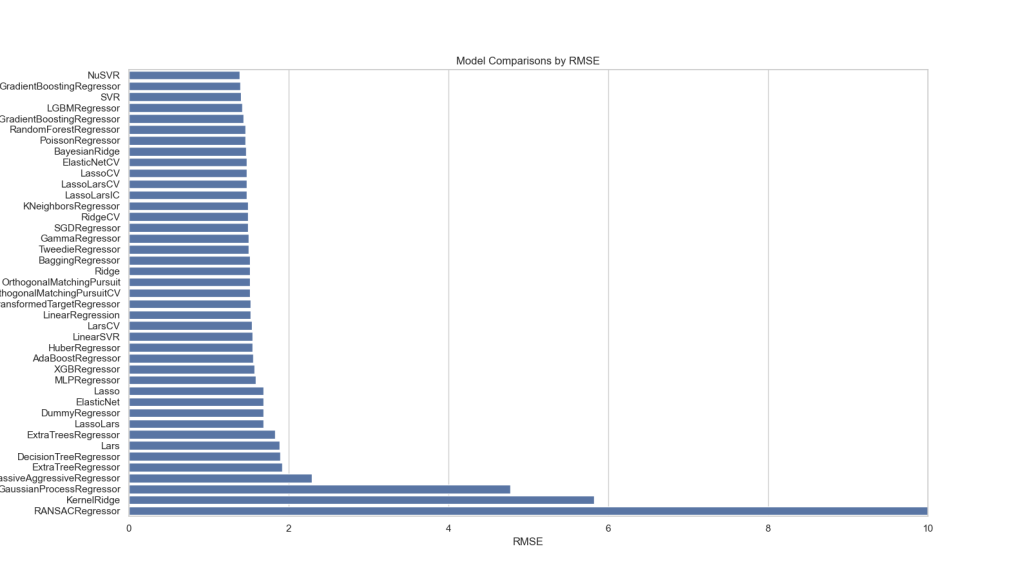


**Table 2: Predictions on train**

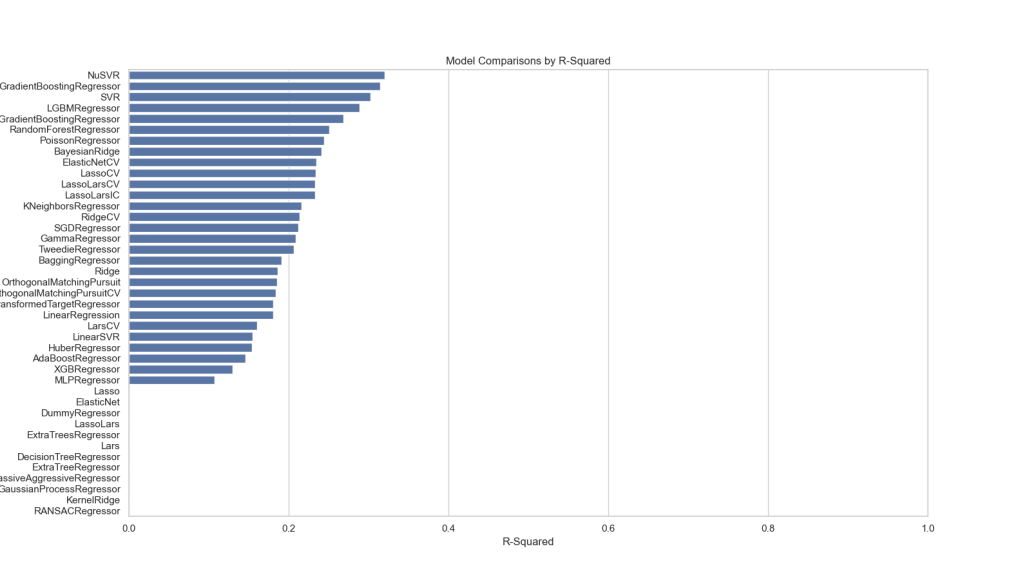
However, testing accuracy on the same data the models were trained on is not a good measure of performance. The NuSVR model performed the best on the test dataset and thus predicts bioactivity the best. The model’s RMSE and R-squared values were 1.41171639451711, 0.353901904966875 respectively. Time taken by the model was 0.296377897262573. Renowned for its resilience, the Random Forest classifier performed consistently in both datasets. Even the best-performing models on the test set, though, only produced R-squared values of about 0.3, meaning that only roughly 30% of the variance in the bioactivity data was explained. Comparison plots of the performance of these regression models when used on the testing data are added as well.



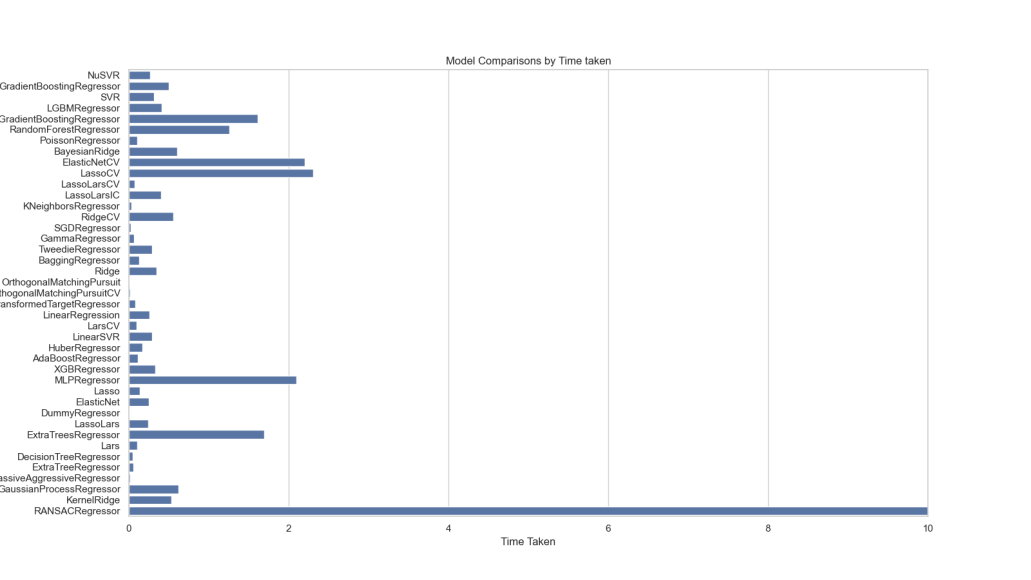
**Table 3: Predictions on test**



**Fig 7: Barchart for RMSE of regression models**



**Fig 8: Barchart for R-square of regression models**



**Fig 9: Barchart for Time of regression models**

# 

# **CONCLUSION AND DISCUSSION**

To sum up, this study included a thorough investigation of QSAR modeling using PaDEL descriptors, including feature selection and analysis of different regression models. After a rigorous feature selection process, the original dataset—which included 3330 compounds with 881 features each—was modified to include 154 descriptors. The purpose of this careful selection was to improve the computational efficiency and specificity of the modeling process.This comparison study included 41 regression models, and the prediction accuracy and goodness of fit were measured using assessment metrics including Root Mean Squared Error (RMSE) and R-squared (R²). With an R-squared value of 0.870199002143443 and an RMSE value of 0.605590054218893, the DecisionTreeRegressor was the most successful model during training, demonstrating its ability to identify patterns in the training data.The study did find, though, that assessing accuracy on the same dataset that was used for training might not accurately represent the performance of the model. With an RMSE value of 1.41171639451711, an R-squared value of 0.353901904966875, and a computational time of 0.296377897262573, the NuSVR model notably showed better performance on the test dataset.It is remarkable that even the most successful models on the test set only explained about 30% of the variance in the bioactivity data, despite the impressive performance of individual models, such as the robust Random Forest classifier. This highlights the bioactivity prediction task's intrinsic complexity and multivariate character, pointing to possible directions for future research that could be expanded upon and explored. More broadly, this work highlights the significance of rigorous model evaluation on separate test datasets and offers insightful information on the complexities of QSAR modeling. The results suggest that more research should be done to improve the interpretability and forecast accuracy of the models, maybe by adding more characteristics or using different modeling strategies.

# **FUTURE WORK**

There are various avenues for future work and improvements in a project of this scale and type. Additional molecular descriptors that might encompass unique aspects of thrombin antagonism including descriptors specific to thrombin interaction could be utilized to get more efficacious and potent lead candidates. An exhaustive exploration of hyperparameter tuning for the Random Forest Regressor and optimization of the model's performance by systematically tuning parameters such as the number of trees and minimum samples per leaf could be performed. Imbalances as seen in the present dataset with inactive compounds surpassing the active ones could be corrected by applying techniques such as oversampling, under sampling, or utilizing advanced algorithms designed for imbalanced datasets like balanced random forests can be investigated. Cross validation of the machine learning models can be carried out to reduce bias, overfitting and estimate performance. Experimental validation of the identified thrombin antagonists with the help of biological assays performed by skilled researchers will definitely reinforce the reliability of the QSAR model's predictions. Surveying data augmentation techniques to increase the diversity of the dataset could enhance model generalization. This could be achieved by generating synthetic data points or introducing variations to present data.

# **REFERENCES**

1.Weitz J. I. (2007). Factor Xa or thrombin: Is thrombin a better target?, *Journal of*

*Thrombosis and Haemostasis,* *Volume 5*, *Supplement 1*, 65-67. ISSN 1538-7836,<https://doi.org/10.1111/j.1538-7836.2007.02552.x>. (https://www.sciencedirect.com/science/article/pii/S1538783622176051)

2. Orcutt SJ, Krishnaswamy S. (2004). Binding of substrate in two conformations to human

prothrombinase drives consecutive cleavage at two sites in prothrombin. *J Biol Chem*, 279:54927–5493

3. Lane DA, Philippou H, Huntington JA. (2005). Directing thrombin. *Blood*,106:2605–2612.

4. Zavyalova E., Kopylov A. (2016). Exploring potential anticoagulant drug formulations using

thrombin generation test, *Biochemistry and Biophysics Reports, Volume 5,* 111-119. ISSN 2405-5808,<https://doi.org/10.1016/j.bbrep.2015.11.011>.

5. Jeon J et al. (2014). A systematic approach to identify novel cancer drug targets using machine

learning, inhibitor design and high-throughput screening. *Genome Med*. 6, 57

6. Ghamali M. , Chtita S., Ousaa A., Elidrissi B., Bouachrine M.,Lakhlifi T. (2017). QSAR

analysis of the toxicity of phenols and thiophenols using MLR and ANN, *Journal of Taibah University for Science*, *11:1,* 1-10. DOI: [10.1016/j.jtusci.2016.03.002](https://doi.org/10.1016/j.jtusci.2016.03.002)

7. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH (2018).

QSAR-based virtual screening: Advances and applications in drug discovery. *Front Pharmacol.*9:1275*.* doi: 10.3389/fphar.2018.01275. PMID: 30524275; PMCID: PMC6262347

8. Akiyama H., Ikeda K., Kondo H., McGeer P.L. (1992). Thrombin accumulation in brains of

patients with Alzheimer’s disease. *Neurosci. Lett.*146:152–154. doi: 10.1016/0304-3940(92)90065-F.

9. Paterson P., Koh C., Kwaan H.(1987). Role of the clotting system in the pathogenesis of

neuroimmunologic disease. *Fed. Proc.,*46:91

10.Vaughan P.J., Pike C.J., Cotman C.W., Cunningham D.D. (1995).Thrombin receptor

activation protects neurons and astrocytes from cell death produced by environmental insults. *J. Neurosci,*15:5389–5401. doi: 10.1523/JNEUROSCI.15-07-05389.1995

11. Striggow F., Riek M., Breder J., Henrich-Noack P., Reymann K.G., Reiser G.(2000). The

protease thrombin is an endogenous mediator of hippocampal neuroprotection against ischemia at low concentrations but causes degeneration at high concentrations. *Proc. Natl. Acad. Sci. USA,*97:2264–2269. doi: 10.1073/pnas.040552897.

12.Wu X., Zhang W., Li J.Y., Chai B.X., Peng J., Wang H., Mulholland M.W. (2011). Induction

of apoptosis by thrombin in the cultured neurons of dorsal motor nucleus of the vagus. *Neurogastroenterol. Motil,*23:279.e124. doi: 10.1111/j.1365-2982.2010.01641.x.

13. Shlobin NA, Har-Even M, Itsekson-Hayosh Z, Harnof S, Pick CG. (2021). Role of thrombin

in central nervous system injury and disease. *Biomolecules.*11(4):562. doi: 10.3390/biom11040562. PMID: 33921354; PMCID: PMC8070021.

14. Foley C.J., NICHOLS L., JEONG K., MOORE C.G., RAGNI M. V. (2010). Coronary

atherosclerosis and cardiovascular mortality in hemophilia, *Journal of Thrombosis and Haemostasis, Volume 8, Issue 1*, 208-211, ISSN 1538-7836,<https://doi.org/10.1111/j.1538-7836.2009.03669.x>.

15. Wattanakit K., Lutsey P.L., Bell E.J., Gornik H., Cushman M., Heckbert S.R., Rosamond

W.D., Folsom A.R.(2012) Association between cardiovascular disease risk factors and occurrence of venous thromboembolism. *Thromb. Haemost,*108:508–515.

16.Stone J., Hangge P., Albadawi H., Wallace A., Shamoun F., Knuttien M.G., Naidu S., Oklu R.

(2017). Deep vein thrombosis: Pathogenesis, diagnosis, and medical management. *Cardiovasc. Diagn. Ther. 7(Suppl. 3):*276–284. doi: 10.21037/cdt.2017.09.01.

17. Mamoshina P et al. (2018).Machine learning on human muscle transcriptomic data for

biomarker discovery and tissue-specific drug target identification. *Front. Genet*. 9, 242.

18. Hansch C., Fujita T. (1964). p -σ-π analysis. A method for the correlation of biological

activity and chemical structure. *J. Am. Chem. Soc.* 86 1616–1626. 10.1021/ja01062a035

19. Cherkasov A., Muratov E. N., Fourches D., Varnek A., Baskin I. I., Cronin M., et al. (2014).

QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* 57 4977–5010. 10.1021/jm4004285

20.Lavecchia A, Di Giovanni C. (2013). Virtual screening strategies in drug discovery: A critical

review. *Curr Med Chem,*20(23):2839–2860. doi: 10.2174/09298673113209990001.

21. Duch W, Swaminathan K, Meller J. (2007). Artificial intelligence approaches for rational

drug design and discovery. *Curr Pharm Des,*13(14).1497–1508. doi: 10.2174/138161207780765954.

22. Mitchell J. B. O. (2014). Machine learning methods in chemoinformatics. *Wiley Interdiscip.*

*Rev. Comput. Mol. Sci.* 4 468–481. 10.1002/wcms.1183

23. Ekins S., Lage de Siqueira-Neto J., McCall L.-I., Sarker M., Yadav M., Ponder E. L., et al.

(2015). Machine learning models and pathway genome data base for *Trypanosoma cruzi* drug discovery. *PLoS Negl. Trop. Dis.* 9:e0003878. 10.1371/journal.pntd.0003878

24. Lipinski C. A., Lombardo F., Dominy B. W., Feeney P. J. (1997). Experimental and

computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23 3–25. 10.1016/S0169-409X(96)00423-1

25. Lee, K., Lee, M. & Kim, D. (2017).Utilizing random Forest QSAR models with optimized

parameters for target identification and its application to target-fishing server. *BMC Bioinformatics 18* (*Suppl 16),* 567.<https://doi.org/10.1186/s12859-017-1960-x>

26. Singh, H., Singh, S., Singla, D. et al. (2015). QSAR based model for discriminating EGFR

inhibitors and non-inhibitors using random forest. *Biol Direct 10*,

<https://doi.org/10.1186/s13062-015-0046-9>

27. Nekoei, M., Mohammadhosseini, M. & Pourbasheer, E. (2015). QSAR study of VEGFR-2

inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): A comparative approach. *Med Chem Res 24,* 3037–3046.<https://doi.org/10.1007/s00044-015-1354-4>

28. Darnag, R., Minaoui, B., & Fakir, M. (2017). QSAR models for prediction study of HIV

protease inhibitors using support vector machines, neural networks and multiple linear regression. *Arabian Journal of Chemistry*, *10*, S600-S608.

29.Wacker, S., & Noskov, S. Y. (2018). Performance of machine learning algorithms for

qualitative and quantitative prediction drug blockade of hERG1 channel. Computational Toxicology, 6, 55–63. doi:10.1016/j.comtox.2017.05.001

30.Zhao, J., Zhu, L., Zhou, W., Yin, L., Wang, Y., Fan, Y., ... & Liu, H. (2018). In silico

prediction of inhibitory constant of thrombin inhibitors using machine learning. Combinatorial Chemistry & High Throughput Screening, 21(9), 662-669.

31. Balakrishnan N, Katkar R, Pham PV, Downey T, Kashyap P, Anastasiu DC,

Ramasubramanian AK. (2023). Prospection of Peptide Inhibitors of Thrombin from Diverse Origins Using a Machine Learning Pipeline. *Bioengineering. 10(11):*1300.<https://doi.org/10.3390/bioengineering10111300>

32.Gaulton A., Bellis L.J., Bento A.P.,Chambers J., Davies M.,Hersey A., Y. Light, McGlinchey

S., Michalovich D.,Al-Lazikani B., Overington J.P., (2012).ChEMBL: a largescale bioactivity database for drug discovery, *Nucleic Acids Res. 40,* D1100–D1107.

33. Simeon S, Anuwongcharoen N, Shoombuatong W, Malik AA, Prachayasittikul V, Wikberg JES,

Nantasenamat C.(2016)Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking*. PeerJ 4,* e2322. <https://doi.org/10.7717/peerj.2322>

34.Liu, G., Lu, J., Lim, H. S., Jin, J. Y., & Lu, D. (2022). Applying interpretable machine

learning workflow to evaluate exposure–response relationships for large‐molecule oncology drugs. *CPT: Pharmacometrics & Systems Pharmacology, 11(12),* 1614-1627.

35.Yap, C. W. (2010). 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>

36.Sun, W., Zheng, Y., Yang, K., Qi, Z., Shah, A. A., Wu, Z., Sun, Y., Feng, L., Chen, D., Xiao,

Z., Lu, S., Li, Y., & Sun, K. (2019). Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances*, *5*(11). <https://doi.org/10.1126/sciadv.aay4275>

37.Tatsat, H., Puri, S., & Lookabaugh, B. (2020). *Machine learning and data Science blueprints*

*for finance*. O’Reilly Media.

38.Das, S., & Cakmak, U. M. (2018). *Hands-On Automated Machine Learning: A beginner’s*

*guide to building automated machine learning systems using AutoML and Python*. Packt Publishing Ltd.

# **TEAM MEMBER CONTRIBUTION:**

**Programmer types:**

The development of the code transpired through collective efforts of Ananya, Manh (referred to as Eddie), and Shwethal who jointly engaged in tasks pertaining to function definition, error bugging, and the deliberation on statistical tests and comparative models. Shwethal and Ananya predominantly collaborated on the processing of the dataset, computation of descriptors, and formulation of corresponding functions. Eddie handled the computation of Padel descriptors utilizing a shell script. Ananya undertook the responsibility of generating graphical representations and tabular columns. Evaluation of performance of machine learning models was carried out by everyone. Shwethal along with Jennifer worked on studying, compiling and collecting information for the introduction, background and data accessed sections of the project report. Ananya and Eddie were responsible for the remaining sections of the project report. References were worked on by all the teammates. Final editing of the report was carried out by Shwethal. ReadMe was written by Ananya and all the files were pushed to GitHub by Eddie. All teammates were present for both in person and online meetings and engaged actively to ensure meeting the project’s objectives.

**Non-programmer types:**

Jennifer was responsible for studying, collecting, collating and compiling information for the introduction, background and data accessed sections of the project report along with Shwethal. She was responsible for the PowerPoint presentation as well as the hands on exercise. She was present for all meetings both in person and online and contributed to the project efficiently.

**Meeting times and Work done:**

10/4/2023 - Project Proposal (in person)

10/14/2023 - Defined objectives, started literature review (in person)

10/21/2023 - Identify drug target, continue literature review (in person)

10/28/2023 - Prepare dataset, finish literature review (online)

11/1/2023 - Project Progress Report (in person)

11/11/2023 - Completion of Exploratory Data Analysis (in person)

18/11/2023 - Model development and Comparison (online)

12/2/2023 - Model deployment, begin report writing (in person)

12/9/2023 - Powerpoint presentation, continue report writing, begin work on hands on exercise, debugging (online)