Fall 2023 CS123A 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 (Eddie) Tuong Nguyen, BS Data Science

NAME TEAM MEMBER 4 Jennifer Melendez, BS Molecular Biology

ABSTRACT

Our research project aims to develop a Quantitative Structure-Activity Relationship (QSAR) model using machine learning algorithms to predict the bioactivity of molecular inhibitors against Thrombin, an endogenous procoagulant protein whose increase is linked to Coronary Artery Disease, Deep Vein Thrombosis and Stroke among others. We will curate a robust dataset of the bioactivity of compounds that target Thrombin. We will attempt to identify Lipinski molecular descriptors and structural features using advanced machine learning techniques. Our goal is to create a predictive model that not only speculates the bioactivity of untested compounds but also provides valuable insights into the rational design of more effective thrombin inhibitors. This interdisciplinary study integrates computational methods with medicinal chemistry, paving the way for advancements in targeted drug design for cardiovascular diseases. We compare various ML regressor models based on RMSE and R squared measure and time against our RF model.

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| SERIAL NUMBER | TOPIC | PAGE NUMBER |
| 1. | INTRODUCTION | 3 |
| 2. | BACKGROUND | 5 |
| 3. | DATA | 6 |
| 4. | APPROACH AND METHODS | 7 |
| 5. | ANALYSIS OF RESULTS | 9 |
| 6. | CONCLUSION AND DISCUSSION | 10 |
| 7. | FUTURE WORK | 10 |
| 8. | REFERENCES | 11 |

LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| SERIAL NUMBER | FIGURE | PAGE NUMBER |
| 1. | Fig.1 | 3 |
| 2. | Fig 2. | 4 |

INTRODUCTION

A diagram of a cell block

Description automatically generatedFor 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. 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, 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. 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. A diagram of a process

Description automatically generated

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)

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)

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 develop 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 timeframe 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 invitro 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 fulfil 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 ultimate objective is to build a machine-learning model that can work with a molecular data set and output the classified potential drug candidates. The approach we used is Data Collecting -> Data Processing -> Exploratory Data Analysis -> Training Model -> Comparing model [33].

This project uses Python 3.9 and the following Python libraries installed:

* [NumPy]
* [Pandas]
* [matplotlib]
* [seaborn]
* [chembl\_webresource\_client]
* [scikit-learn]
* [rdkit]

A Pycharm IDE was used to collaborate and write code. The Anaconda distribution of Python was used, which already has the above packages and more included. After a programmatic search of the ChEMBL database for our target of interest we isolated bioactivity data of thrombin target with ChEMBL ID `CHEMBL204` reported as IC50 values in nM (nanomolar) unit. We dropped null and duplicate rows for canonical smiles. The final dataframe composed of 1985 potential drug leads is saved as a csv file which has the following columns: -

* `molecule\_chembl\_id`: Unique ChEMBL ID of the molecule
* `canonical\_smiles`: Information about the chemical and molecular structure
* `standard\_value`: quantitative IC50 measures.
* `class`: IC50 values binned into `active`, `inactive` and `intermediate`

In this project we will only consider active and inactive compounds. 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., and most importantly, it is the universal data type that most Python packages required. 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. The next step was to perform exploratory data analysis. 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.

The calculation of Lipinski’s descriptors requires only a part of the canonical SMILES of the molecules which is filtered out with the help of a simple loop. A custom function for extracting Lipinski’s descriptor of each element is employed in the dataset using rdkit. Normalization of IC50 to pIC50 is carried out to ensure uniform distribution and improve the readiness of the standard value. The IC50 values from the standard value column are converted from nM to M by multiplying the value by 10-9. Then, the molar value undergoes a -log10 transformation. 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. Noted that all IC50 values larger than 108 will be set as 100000000 to improve runtime.The model utilized fingerprint PaDEL instead of Lipinski’s descriptors . The PaDEL descriptor [34] is comprised 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 molecule than the Lipinski one and gives better accuracy for the random forest regressor model to be deployed. [35]. 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 is to build a model to predict the potential drug candidate which warrants a regression model. Its good performance, scalability, and ease of use make it 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 [36]. We believe this is the best model for this project. However, a comparison with other models is important to validate the best model for this task. We used the PaDEL calculation package [37] 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. These data would be the input of the algorithm and the output would be the pIC50 value. Note that we want to remove the entries with low variance as they do not provide the model with useful information to learn patterns [38]. So, the last step before training will be 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 adjust the data to be in the format that the model understand. 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 purpose. 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. The model's performance was benchmarked against other regression models using the LazyRegressor tool. This facilitated a comprehensive comparison based on RMSE, computation time, and R-squared values, aiding in the selection of the most suitable model for bioactivity prediction.

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 DecisionTreeRegressor was the model that performed the best in the training dataset, demonstrating its ability to successfully identify patterns in the training data. On the other hand, the NuSVR model performed the best on the test dataset, indicating that it can effectively generalize to new data.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. 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 DecisionTreeRegressor was the model that performed the best in the training dataset, demonstrating its ability to successfully identify patterns in the training data. On the other hand, the NuSVR model performed the best on the test dataset, indicating that it can effectively generalize to new data.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.The evaluation of regression models in our study utilized the Root Mean Squared Error (RMSE) to measure the model's prediction error on the test set. A lower RMSE value 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. Among the regressors evaluated, the DecisionTreeRegressor for traning and NuSVR models stood out based on these criteria.

CONCLUSON AND DISCUSSION

The research used a rigorous computational method to predict the bioactivity of thrombin inhibitors, as well as a suite of statistical and machine learning technologies. Despite the intrinsic complexity of biological interactions, the use of several regression models, such as the well-known DecisionTreeRegressor and NuSVR, provides useful insights into the prediction of molecular bioactivity for train and test respectively. The study highlights the difficulties of predictive modelling in bioinformatics, but it also lays the door for future advancements through methodological modifications and the incorporation of more advanced biological datasets. This initiative is a first step towards developing more precise and computationally efficient models in the search for new medicinal drugs.

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. 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.
34. 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>
35. 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>
36. Tatsat, H., Puri, S., & Lookabaugh, B. (2020). *Machine learning and data Science blueprints for finance*. O’Reilly Media.
37. 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.

TEAMMEMBER CONTRIBUTION:

Programmer types:

The code was not worked on in individual modules. The programmer types Ananya, Manh (eddie) and Shwethal worked together on defining the functions, error bugging and deciding the statistical test and models to compare. Changes and additions were made along the way equally by all. Shwethal and Ananya primarily worked together on dataset processing, calculating descriptors and defining the functions for the same. Eddie worked with the calculation of padel descriptors using shell script. Ananya worked on making the plots. All worked together on comparing ML models. In total 4 meetings were in person. The rest of the meetings were online and everyone attended all the meetings and worked on their parts.

Non-programmer types:

Jennifer incorpated in analyzing the research and background information to our main focus to the project. As well as doing the research for the project, she incorporated and did the powerpoint for the project as well as doing hands excercise.