Machine Learning for Cancer Drug Discovery

Machine Learning Modelling to Predict the Efficacy of Cancer Treatment Drugs

Edicon Tze Shun Chan\*, Leon Edmiidz\*, Jacob Hambly\*, Zach Ribau\*

\*Department of Animal Biosciences

University of Guelph, Guelph, Ontario, N1G 2W1, Canada

# ABSTRACT

The drug development process poses significant challenges, including lengthy timelines, high experimental costs, and low success rates in clinical trials. Despite advancements in high throughput screening (HTS), the abundance of false positives associated with this method requires additional experimental validation. To alleviate this burden, there is growing interest in applying machine learning (ML) models to HTS, enhancing accuracy and efficiency throughout the drug discovery stage. Quantitative structure-activity relationship (QSAR) models can utilize ML algorithms to predict compound bioactivity against specific targets. This study investigates various QSAR models for their ability to identify important structural and physiochemical properties relevant to target inhibition and their feasibility in screening potential bioactive molecules of interest. Results obtained from this study show that an optimized QSAR model exhibits less overfitting with comparable predictive performance. Key physiochemical features identified by the models, including VSA descriptors and *Max\_estate.* Furthermore, a pyrimidine ring structure, recognized for its prominence in oncological therapeutics, was identified as a key structure within the fingerprinting features. These findings exhibit the optimized QSAR model’s effectiveness in uncovering biologically significant features. More importantly, this study suggests that ML holds promise in predicting IC50 values and could further enhance specificity across a variety of biological targets. Optimizing the compound screening and selection stage shows ML can be a valuable approach in advancing the drug discovery process.

**Key words:**

machine learning, random forest, QSAR, drug discovery, cancer, EGFR

**List of abbreviations:**

ADA: AdaBoost Regression

DT: Decision Tree

EGFR: Epidermal Growth Factor Receptor

GBR: Gradient Boost Regression

HTS: High throughput screening

IC50: Half-maximal inhibitory concentration

KNN: K-Nearest Neighbours

ML: Machine Learning

AI: Artificial Intelligence

MLP: Multi-Layer Perceptron

QSAR: Quantitative Structure-Activity Relationship

RF: Random Forest

SMILES: Simplified molecular-input line-entry system

SVR: Support Vector Regression

RFE: Recursive Feature Elimination

MSE: Mean Squared Error

RMSE: Root Mean Squared Error

MAE: Mean Absolute Error

# INTRODUCTION

In the drug development process, from drug screening to market for a single drug, it takes on average a decade and can cost up to a billion dollars (Paul *et al*., 2021). Additionally, only about 10-20% of drugs that go to clinical trials are approved for patient use (Yamaguchi *et al*., 2021). These statistics are made worse when looking at certain diseases such as cancer, where 95% of compounds showing promise in inhibiting cancer cells *in vitro* or in animal trials fail in human clinical trials due to ineffectiveness (Kunnumakkara *et al*., 2019). The fact that there has not been much improvement in the success rate of developing novel therapeutics indicates novel innovations are needed. Efforts to incorporate artificial intelligence (AI) to increase the efficiency and effectiveness of the drug development cycle have been on the rise (Vamathevan *et al*., 2019). Particularly, due to the explosion of available data as a result of innovations in high throughput screening (HTS). HTS has allowed for large databases of bioactive compounds for training machine learning (ML) models (Gaulton *et al*., 2012). Excitement over the potential incorporation of AI in drug development can be partly attributed to its proficiency in finding relationships and patterns within highly complex bioactivity data (Vamathevan *et al*., 2019). This is of particular interest in the screening of potential bioactive compounds against a biological target of interest.

Contemporary, high throughput screening (HTS) is used for discovering novel active compounds (Blay *et al*., 2020). HTS is an automated process where a large library of compounds is tested for their ability to inhibit a biological target in an assay (Blay *et al*., 2020). HTS has accelerated the pace and lowered the cost at which bioactive compounds are discovered (Boldini *et al*., 2023). However, HTS often produces large numbers of false positives which subsequently leads to extensive experimental validation post-HTS (Boldini *et al*., 2023). To help solve this issue, there has been attention towards utilizing machine learning models to aid HTS during the discovery stage of drug development (Boldini *et al*., 2023). ML algorithms can be employed to build a model that learns the relationships between the structural and physicochemical properties of a compound and its ability to inhibit a certain biological target (Spiegel and Senderowitz, 2020). This type of ML model is called a quantitative structure-activity relationship (QSAR) model and several studies have demonstrated QSAR’s strong performance in the in-silico screening and identification of bioactive compounds that inhibit a biological target (Speck-Planche *et al*., 2012; Zhao *et al*., 2017; Capuzzi *et al*., 2018). ML models can also be used for discovering important structural features of an inhibitor by determining which particular subset of features were most important for the ML model in predicting the bioactive status of a molecule (Wang *et al*., 2015). This can help guide researchers to synthesize stronger and more effective inhibitory compounds (Zhao *et al*., 2017).

The employment of an ML predictive model in conjunction with HTS has been shown to be particularly useful for narrowing down the list of compounds to include only those that show the most promise for the inhibition of a biological target (Neves *et al*., 2018). Additionally, compounds identified as bioactive against the biological target by HTS can be verified using an ML predictive model which would thus decrease the rate of false positives identified and reduce the need for extensive experimental validation (Boldini *et al*., 2023).

This project's objective will be to test the proof of concept that machine learning models trained using data on the bioactive ability of compounds against a particular target can be used to accurately screen for novel bioactive molecules. Additionally, we will extract which structural/physicochemical properties were most important in predicting the bioactivity of compounds and thus inferring which features could be important for the inhibition of a particular biological target.

# MATERIALS AND METHODS

## 2.1. Data collection

The data for this project was obtained from ChEMBL, a manually curated database managed by the European Bioinformatics Institute. This high-quality, open and large-scale database contains data on bioactive compounds and is a widely used drug-discovery platform (Zdrazil *et al*., 2024). Through the ChEMBL web resource client Python package, drugs that inhibited the Erbb1 protein were filtered for and downloaded. The dimensions of the data before any processing were 17,286 rows by 46 columns. The IC50s of the compounds were measured in two different assay types, single protein-based assay and cell-based assay, therefore the data was split by assay type and then cleaned. Rows with duplicate ChEMBL IDs, duplicate SMILES, missing values, missing canonical SMILES, and outliers were removed. The IC50 values were then transformed to -log10(M) to represent the values on a more interpretable scale. The only columns of interest were canonical smiles and IC50 values, therefore every other column was removed. The two datasets were then concatenated back together while including the assay type as a new column. The resulting dataset consisted of 8521 rows of data and 3 columns; SMILES, -log10(M) IC50 values, and assay type.

2.2. Feature Engineering

RDkit is an open-source cheminformatics toolkit accessible through a Python package (Bento *et al*., 2020).Using the canonical SMILES of the compounds in the downloaded dataset, RDkit allowed for the construction of various features of the compounds. Two different types of features were constructed; molecular descriptors and Morgan fingerprints. Molecular descriptors include 208 physiochemical properties of the compounds such as molecular weight, valence electrons information, halogen frequencies, different functional groups, and so on. These physiochemical properties included many columns measured by continuous values; therefore, they were scaled before training the models. For the Morgan fingerprints, 3072 different features were engineered, representing different chemical structures of the compounds such as various functional groups, chains of carbon atoms, and the types of bonds between various atoms. In the dataset, these fingerprints are measured by 1 or 0, 1 representing the presence of the fingerprint and 0 representing the absence. The length of 3072 molecular fingerprint bits was selected based on the protocol found by Tang et al*.* (2024).

## 2.3. Data processing

After obtaining the molecular descriptors and Morgan fingerprints, additional data processing steps were conducted to reduce the number of features. This included dropping columns containing single values or NAs and columns of low variance. For the molecular descriptors, columns with a variance of less than 1 were dropped. A lower variance threshold of 0.1 was set for Morgan fingerprint columns to account for their binary values. After processing, the dataset retained 8522 rows with 186 features: 95 molecular descriptors, 89 fingerprints, the assay type, and –log10(M) IC50 values.

## 2.4. Problem definition

The general goal of the project is to train regression models to predict the -log(M) IC50 values of the drugs based on the features that were engineered by RDkit. Instead of running classification algorithms, the aim is to improve the predictive resolution through the use of regression models. Given that the IC50 values represent the efficacy of the cancer-treating drugs in inhibiting the EGFR protein, the idea is to identify which drugs are likely to be effective and are worth exploring for further testing in their application for treating patients with cancer.

## 2.5. Statistical analyses

To analyze the performance of the models, the mean absolute error, mean squared error, root mean square error and R² values were calculated. These scoring metrics allowed for determining which models performed best and which parameter settings were optimal for the Random Forest model during hyperparameter optimization.

## 2.6. Machine learning modelling

### 2.6.1. Algorithms Evaluated

Seven different regression-based supervised machine learning algorithms were evaluated: Random Forest regressor (RF), K-Nearest Neighbours (KNN), Gradient Boost Regression (GBR), Support Vector Regression (SVR), XGBoost (XGB), AdaBoost (ADA), Decision Tree (DT) and Multi-Layer Perceptron (MLP).

DT facilitates the analysis of data by organizing it into a structured format. A DT comprises decision nodes, which evaluate attribute values; edges, representing the outcomes of these evaluations and connecting subsequent nodes; and leaf nodes, responsible for predicting outcomes (Singh Kushwah *et al*., 2022). Within the decision tree framework, each attribute within a dataset is represented as a node, with a designated root node serving as the starting point for analysis (Singh Kushwah *et al*., 2022). The decision-making process within a decision tree commences at the root node and progresses downwards, with each node representing a specific attribute evaluation. This iterative process continues until a terminal node, signifying a prediction or output, is reached.

RF is an ensemble method which predicts the value of a target variable by combing the outputs from various decision tree algorithms (Breiman, 2001; Rodriguez-Galiano *et al*., 2015). RF builds numerous decision trees on different data subsets by bagging (Rodriguez-Galiano *et al*., 2015). To make a final decision, RF averages the results produced from the numerous decision trees (Rodriguez-Galiano *et al*., 2015).

KNN stands as an intuitive and non-parametric approach, which stores all the observations from the training set to predict outcomes (Ortiz-Bejar *et al*., 2018). This prediction process relies on similarity or distance metrics. In the context of KNN regression, given an input, the output is determined to be the same value as its nearest neighbours (Ortiz-Bejar *et al*., 2018). The parameter *k* denotes the number of neighbours considered in the prediction process (Ortiz-Bejar *et al*., 2018). When *k* is greater than 1, the output can be computed either as the average of the outputs from the *k* nearest neighbours (Ortiz-Bejar *et al*., 2018). Each neighbour can carry equal weight, or the weighted contributions of neighbours may be considered based on their respective distances (Ortiz-Bejar *et al*., 2018).

In solving binary classification dilemmas, Support Vector Machines (SVMs) identify the maximum margin hyperplane, ensuring accurate classification of as many training points as feasible (Awad and Khanna, 2015). The SVM's capacity for sparse solutions and robust generalization renders it adaptable to regression challenges. Transitioning from classification to regression, SVMs extend to SVR by introducing an ε-insensitive region, known as the ε-tube, around the function (Awad and Khanna, 2015). A multi-objective function emerges from the interplay between the loss function and the geometric properties of the tube. Subsequently, employing appropriate numerical optimization algorithms, the unique solution to the convex optimization problem is derived (Awad and Khanna, 2015).

AdaBoost is a sequential ensemble technique designed to amalgamate multiple weak learners sourced randomly from the dataset to forge a robust learner (Shanmugasundar *et al*., 2021). These weak learners are crafted through the application of various ML algorithms (Shanmugasundar *et al*., 2021). In each iteration of the training process, weights are assigned to individual sample observations, shaping the learning process for each hypothesis. Subsequently, false predictions are identified and redirected to subsequent base learners, with heightened emphasis placed on rectifying these inaccuracies (Shanmugasundar *et al*., 2021). This iterative procedure persists until the algorithm achieves accurate outputs.

The fundamental components of MLP comprise neurons, also referred to as nodes. These nodes are organized into layers within the neural network, interconnected in a manner that facilitates the unidirectional flow of information (Rodriguez-Galiano *et al*., 2015). Specifically, data moves from the input nodes, traverses through the hidden layers, and eventually reaches the output layer. Input units transmit signals to the hidden nodes within subsequent layers. Each node undertakes a process involving linear regression followed by the application of a nonlinear function (Rodriguez-Galiano *et al*., 2015). Interconnections between neurons across different layers are established through corresponding links, commonly known as weights (Rodriguez-Galiano *et al*., 2015).

GBR is an ensemble regression tree-based gradient-boosting algorithm (Natekin and Knoll, 2013). It sequentially trains a series of regression trees, where each subsequent tree corrects the errors (residuals) of the previous trees to minimize a specified loss function, such as mean squared error (MSE) (Natekin and Knoll, 2013).

XGBoost regression is an advanced algorithm based on the Gradient Boosting framework, offering additional features such as regularization to control model complexity and mitigate overfitting (Chen and Guestrin, 2016). XGBoost also incorporates optimizations like handling missing values, parallel processing, and efficient tree pruning techniques for improved efficiency and scalability (Chen and Guestrin, 2016).

### 2.6.2. Feature Selection and Optimization

To address potential overfitting in the RF model selected for further analysis, feature selection and hyper-parameter tuning were implemented to improve model generalizability while retaining performance comparable to initial cross validation results. Recursive Feature Elimination (RFE) was used to identify the most influential molecular fingerprint and physicochemical descriptors for the RF model. RFE involves iteratively fitting a model, assessing feature importance, and removing the least important features. Feature importance was quantified with the mean decrease Gini score, which reflects an average of how each feature's presence decreases the impurity across the decision nodes it was used in. A greater mean decrease in Gini reflects a greater importance for predictive accuracy. In this analysis, RFE was set to iteratively remove features one at a time from the initial 184 features to 50. The choice to reduce the space to 50 features was made based on selecting the highest adjusted R2 statistic for RFE with cross-validation completed with feature sizes of 20, 50 and 80, on the training dataset (results not reported). Due to computational limitations, a more comprehensive selection of feature size was not feasible.

Hyperparameter optimization was carried out with the RandomizedSearchCV function from the Scikit-Learn library, which facilitated computationally efficient evaluation of the max\_features, max\_depth and min\_samples\_leaf parameters. The max\_depth parameter restricts tree complexity by setting growth limits, helping to avoid overfitting in high-dimensional data by preventing excessively complex decision boundaries that fail to generalize (Biau, 2012). Simultaneously, max\_features regulates the number of features considered for splits, enhancing robustness through enforcing diversity, which is imperative for data with large feature sizes (Louppe *et al*., 2013). Finally, min\_samples\_leaf sets the minimal number of samples required to be present at a leaf, further restricting tree complexity; larger values prevent overfitting by ensuring adequate learning at terminal nodes, promoting a more generalizable model (Svetnik *et al*., 2003). Accordingly. the optimization grid was set with max\_features at [5, 10, 15, 20, 25, 30, 'sqrt', 'log2'], max\_depth at [5, 10, 15, 20, 30] and min\_samples\_leaf at [5, 10, 15, 20, 25, 30]. These ranges ensure the selection of parameters within the constraints, ensuring that generalizability is prioritized. One hundred parameter settings were sampled, and the process aimed to minimize the negative root mean squared error with 5-fold cross validation. The optimal parameter combination was identified, and the corresponding metrics were dynamically used to retrain a final RF model on the entire training dataset prior to validation with the held-out testing set.

Feature importance was evaluated by extracting feature column codes along with their respective mean decrease Gini scores from the final RF model. Scores for all 50 features were obtained and the top 15 are reported (Table 4). Scores for all 50 features are available within the supplemental data files.

3. RESULTS

*K-Fold Cross Validation Comparison of Algorithms*

From 5-fold cross validation, the RF regressor algorithm demonstrated best results with a MAE of 0.598, a mean squared error MSE of 0.651, a RMSE of 0.807 and a R2 of 0.684. The KNN has a MAE of 0.609, an MSE of 0.700, an RMSE of 0.836 and a R2 of 0.660. GBR had a MAE 0.778, an MSE of 0.971, a RMSE of 0.985, and a R2 of 0.529. SVR had a MAE of 0.845, a MSE of 1.204, a RMSE of 1.097 and a R2 of 0.416. DT had a MAE of 0.775, an MSE of 1.287, an RMSE of 1.139 and R2 of 0.375. ADA had a MSE of 0.999, an MSE of 1.435, an RMSE of 1.198 and a R2 of 0.304. MLP had a MAE of 1.216, an MSE of 2.064, an RMSE of 1.436 and a R2 of 0.001.

*Learning Curve Comparison of Algorithms*

Based on the learning curves generated on the algorithms used during cross-validation we see an overall trend of the model's overfitting on the training data. MLP and DT demonstrated high scores on the training data with low cross-validation scores. This indicates that these models effectively memorized the training data and were unable to properly generalize unseen data. ADA had low training and cross validation scores which indicates that the model is not adequately capturing the relationships in the data and is underfitting the data set. KNN, GBR, XGB, SVR and RF appeared to be capturing the relationship in the data with higher scores during cross-validation however we still see overfitting on the training set. Particularly the Random Forest model as seen in Figure 1d where we see quite a large gap between the training and validation and no apparent trend for convergence after including all the training data. The XGB Regressor model shows less overfitting as the training score and the cross-validation are closer and appear to be converging after all the training data has been included.

*Feature Selection Hyper-Parameter Tuning of Random Forest Model*

The resulting hyper-parameters identified from the randomized cross validation optimization with the 50 RFE selected features are shown in Table 3. The best model corresponding to the selected parameters achieved an RMSE of 0.865, which exceeded that of the initial RF model trained with default parameters, with an RMSE of 0.807 (Table 2). The resulting learning curve computed for the final RF can be seen in Figure 3, where a greater degree of convergence is evident as compared to the initial learning curve for RF (Figure 1). This indicates a reduction in overfitting as the training score is notably decreased on average, with comparable predictive performance as evidenced by the CV score.

*Test Set Validation*

The final RF model achieved comparable test set performance to the initial non-optimized RF model evaluated with cross validation and default parameters (Table 2). Notably, the final RF model was trained on only 50 select features, whereas the non-optimized RF model was trained on all initial 184 features. Thus, the final model, while achieving lower performance than the initial non-optimized model, retains comparable accuracy with a sparser set of features and less evidence of overfitting.

*Feature Importance*

Table 4 shows the top 15 most important features based on mean decrease Gini scores from the final RF model. The most important feature, *col\_343*, was ranked considerably more important than all other selected features. The identity for this feature was a Morgan fingerprint corresponding to a pyrimidine ring structure. The majority of the remaining top-selected features belonged to the family of RDkit’s physiochemical volume and surface area, VSA descriptors. Among the entire RFE selected feature set, energy state descriptors, denoted with trailing *Estate* in the name, were another prominent grouping identified. Of the energy state descriptors, only one appeared in the top 15. A complete table of all 50 selected features and their importance scores is available in the supplemental files

# 4. DISCUSSION

*Model Performance*

The final RF model achieved an R2 value of 0.622, effectively capturing a considerable degree of variance within the IC50 values in the dataset. Thus, the optimized RF model, trained on a reduced feature space of important molecular features, presents as a viable screening tool for the prediction of continuous IC50 values. Consequently, the model can be used to evaluate the potential inhibitory capacity of novel compounds towards the EGFR protein, thereby contributing to the identification of candidate anti-cancer compounds. Furthermore, the model was able to perform comparably well when tested on the external validation split, to the initial RF model trained on all 184 features; accordingly, a sparse set of important features has been identified that can further inform future studies. To this end, a detailed analysis of the ten most important features (Table 4) can be particularly informative for experimental research seeking a better understanding of the complex molecular interactions underlying binding affinity towards the EGFR protein.

The performance of the final RF model is competitive with an alternative model that has been trained to evaluate inhibitors of EGFR in the literature. For instance, Chauhan and colleagues (2014) developed models trained with small quinazoline and imidazothiazoles/pyrazolopyrimidines inhibitors of EGFR exclusively and reported correlations ranging from 0.761 to 0.850. While these indicate greater variance explained than this current study, it is important to note that the limited diversity in small molecule inhibitors, and the small sample size of 128 inhibitors, limits direct comparability. Outside of this study, the majority of ML predictions for EGFR inhibition have performed classification tasks, which lack the granularity often required to transition candidate compounds to real-world competitors. Thus, our analysis offers a unique contribution to the literature, for which future work should build upon with the aim of uncovering EGFR-specific predictors of inhibition.

*Final Random Forest Model Selects Biologically Relevant Descriptor*s

The significance of VSA-related physicochemical features in predicting interactions at protein binding sites is widely recognized in the literature (Segura *et al*., 2011), supporting the efficacy of the optimized Random Forest model and RFE in identifying biologically relevant features. RDkit's physiochemical VSA descriptors, reflecting atom-level contributions to stearic qualities like molecular volume and surface area, are fundamentally related to substrate-ligand compatibility (Das *et al*., 2010; Segura *et al*., 2011). RDkit VSA features have been correlated with binding affinity in other QSAR studies. Cai and colleagues, for example, showed that VSA features were abundant among top predictors of binding strength and kinetic rate constants for ligands targeting small RNAs, modelling with multiple linear regression (Cai *et al*., 2022). Furthermore, electrostatic factors also play a significant and possibly more specific role in driving substrate-ligand binding affinity (Das *et al*., 2010; Ehrt *et al*., 2018). Accordingly, the presence of *Max\_Estat*e properties in the selected feature set further supports the validity of the final RF model in selecting biologically relevant features representing diverse chemical mechanisms.

*Pyrimidine Ring Structure*

The final RF model identified the Morgan fingerprint corresponding to the pyrimidine ring structure as the most important feature. This is particularly significant due to the functional relevance of pyrimidine derivatives in pharmaceutical inhibitors of oncogenic protein targets (Matada *et al*., 2020; Mandour *et al*., 2022), which have been further validated in molecular docking studies (Patil and Kumbhar, 2023) and in inhibition studies of EGFR specifically (Matada *et al*., 2020). Pyrimidines are a class of heterocyclic aromatic organic compounds analogous to benzene but with nitrogen atoms replacing carbon atoms at substituent positions 1 and 3. In the context of drug design, modifications to the pyrimidine ring greatly influence pharmacological properties (Matada *et al*., 2020), including its ability to interact with biological targets. Thus, this finding further highlights the final model's efficacy in identifying biologically relevant features for the inhibition of biological targets. Future work should evaluate the potential of RF regression modelling for resolving specificity in QSAR analysis, and the potential for machine learning in general to resolve differences in specificity across diverse biological targets.

*Limitations*

A limitation of this project was the fact that we only optimized the Random Forest model after initial cross-validation. Due to time and computational constraints, we chose to only focus on optimizing the Random Forest model as it showed the most promise in successfully fitting the data based on cross validation results. We were confident in this decision as other researchers have pointed to Random Forest-based QSAR models as being a ‘gold standard’ for modelling a compound's structural/physicochemical properties and bioactivity (Kwon *et al*., 2019). However, optimization for each algorithm used in cross validation might have pointed toward a different optimized model with better results on unseen data compared to the Random Forest model. Particularly, the XGB regressor algorithm showed cross-validation results comparable to Random Forest with less overfitting during cross validation. Recently, the XGBoost algorithm has garnered attention for its predictive capabilities in QSAR modelling (Boldini *et al*., 2023). It has also been shown that hyperparameter optimization for XGBoost models is particularly important for QSAR modelling (Boldini *et al*., 2023). Thus, a comparison of optimized models would be key to further enhancing this project and determining the best model for predicting a compound's ability to inhibit EGFR.

*Next steps*

Further work for this project would also include molecular docking in our analysis. Many molecular docking software exist and can predict the interactions of a compound to the binding site of a biological target (Torres *et al*., 2019). Molecular docking can be used as an additional step in this project by estimating the binding affinities of the strongest inhibitors predicted in the test set with the active site of EGFR. This would help in validating the model as strong inhibitors identified in our model should have strong binding affinities with EGFR. Additionally, molecular docking would aid in identifying 3D features important in blocking the active site of EGFR which is important in the synthesis of novel inhibitors of EGFR (Zhao *et al*., 2017).

# CONCLUSIONS

This analysis presents a unique application of ML methods to enhance the drug discovery process for EGFR inhibition and consequent cancer therapeutics, specifically with regression modelling. After a brief evaluation of 8 different ML regression algorithms was completed, with cross validation and learning curve analysis leading to the selection of RF as the model for further consideration. A final RF model was optimized with randomized cross validation optimization, and trained with 50 physiochemical and Morgan fingerprint descriptors, pre-selected with RFE. This model was able to predict IC50 values for unseen test data with an R2 of 0.622. Thus, this analysis concludes that RF regression models trained on molecular features demonstrated the potential to reduce the time and cost associated with traditional drug development methods, with additional implications for accelerating the arrival of new therapeutic agents for treating cancer.

# ACKNOWLEDGEMENTS

The authors would like to thank Dr. Dan Tulpan for his guidance and the RDkit development team for their work.

# AUTHORS’ CONTRIBUTIONS

All authors improved and contributed to the editing of the manuscript. All authors read and approved the final manuscript.

# DISCLOSURES

The authors declare no real or perceived conflicts of interest.

# LITERATURE CITED

Awad, M., and R. Khanna. 2015. Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA. p. 67–80.

Bento, A. P., A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij, and A. R. Leach. 2020. An open source chemical structure curation pipeline using RDKit. J Cheminform. 12. doi:10.1186/s13321-020-00456-1.

Blay, V., B. Tolani, S. P. Ho, and M. R. Arkin. 2020. High-Throughput Screening: today’s biochemical and cell-based approaches. Drug Discov Today. 25. doi:10.1016/j.drudis.2020.07.024.

Boldini, D., F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber. 2023. Practical guidelines for the use of gradient boosting for molecular property prediction. J Cheminform. 15. doi:10.1186/s13321-023-00743-7.

Breiman, L. 2001. Random Forests. Mach Learn. 45:5–32. doi:10.1023/A:1010933404324.

Cai, Z., M. Zafferani, O. M. Akande, and A. E. Hargrove. 2022. Quantitative Structure-Activity Relationship (QSAR) Study Predicts Small-Molecule Binding to RNA Structure. J Med Chem. 65. doi:10.1021/acs.jmedchem.2c00254.

Capuzzi, S. J., W. Sun, E. N. Muratov, C. Martínez-Romero, S. He, W. Zhu, H. Li, G. Tawa, E. G. Fisher, M. Xu, P. Shinn, X. Qiu, A. García-Sastre, W. Zheng, and A. Tropsha. 2018. Computer-Aided Discovery and Characterization of Novel Ebola Virus Inhibitors. J Med Chem. 61. doi:10.1021/acs.jmedchem.8b00035.

Chauhan, J. S., S. K. Dhanda, D. Singla, S. M. Agarwal, and G. P. S. Raghava. 2014. QSAR-based models for designing quinazoline/imidazothiazoles/ pyrazolopyrimidines based inhibitors against wild and mutant EGFR. PLoS One. 9. doi:10.1371/journal.pone.0101079.

Chen, T., and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Vol. 13-17- August-2016.

Das, S., M. P. Krein, and C. M. Breneman. 2010. Binding affinity prediction with property-encoded shape distribution signatures. J Chem Inf Model. 50. doi:10.1021/ci9004139.

Ehrt, C., T. Brinkjost, and O. Koch. 2018. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). PLoS Comput Biol. 14. doi:10.1371/journal.pcbi.1006483.

Gaulton, A., L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. 2012. ChEMBL: A large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40. doi:10.1093/nar/gkr777.

Kunnumakkara, A. B., D. Bordoloi, B. L. Sailo, N. K. Roy, K. K. Thakur, K. Banik, M. Shakibaei, S. C. Gupta, and B. B. Aggarwal. 2019. Cancer drug development: The missing links. Exp Biol Med. 244. doi:10.1177/1535370219839163.

Kwon, S., H. Bae, J. Jo, and S. Yoon. 2019. Comprehensive ensemble in QSAR prediction for drug discovery. BMC Bioinformatics. 20. doi:10.1186/s12859-019-3135-4.

Mandour, A. A., I. F. Nassar, M. T. Abdel Aal, M. A. E. Shahin, W. A. El-Sayed, M. Hegazy, A. M. Yehia, A. Ismail, M. Hagras, E. B. Elkaeed, H. M. Refaat, and N. S. M. Ismail. 2022. Synthesis, biological evaluation, and *in silico* studies of new CDK2 inhibitors based on pyrazolo[3,4-d]pyrimidine and pyrazolo[4,3-e][1,2,4]triazolo[1,5-c]pyrimidine scaffold with apoptotic activity. J Enzyme Inhib Med Chem. 37. doi:10.1080/14756366.2022.2086866.

Matada, G. S. P., N. Abbas, P. S. Dhiwar, R. Basu, and G. Devasahayam. 2020. Design, Synthesis, *In Silico* and *In Vitro* Evaluation of Novel Pyrimidine Derivatives as EGFR Inhibitors. Anticancer Agents Med Chem. 21. doi:10.2174/1871520620666200721102726.

Natekin, A., and A. Knoll. 2013. Gradient boosting machines, a tutorial. Front Neurorobot. 7. doi:10.3389/fnbot.2013.00021.

Neves, B. J., R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade. 2018. QSAR-based virtual screening: Advances and applications in drug discovery. Front Pharmacol. 9. doi:10.3389/fphar.2018.01275.

Ortiz-Bejar, Jose, M. Graff, E. S. Tellez, Jesus Ortiz-Bejar, and J. C. Jacobo. 2018. k-Nearest Neighbor Regressors Optimized by using Random Search. In: 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). IEEE. p. 1–5.

Patil, P. B., and S. T. Kumbhar. 2023. Identification of potential CDK 8 inhibitor from pyrimidine derivatives via In-Silico approach. Journal of Medical Pharmaceutical and Allied Sciences. 12. doi:10.55522/jmpas.V12I5.4784.

Paul, D., G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade. 2021. Artificial intelligence in drug discovery and development. Drug Discov Today. 26. doi:10.1016/j.drudis.2020.10.010.

Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev. 71. doi:10.1016/j.oregeorev.2015.01.001.

Segura, J., P. F. Jones, and N. Fernandez-Fuentes. 2011. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. BMC Bioinformatics. 12. doi:10.1186/1471-2105-12-352.

Shanmugasundar, G., M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran. 2021. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. Processes. 9. doi:10.3390/pr9112015.

Singh Kushwah, J., A. Kumar, S. Patel, R. Soni, A. Gawande, and S. Gupta. 2022. Comparative study of regressor and classifier with decision tree using modern tools. Mater Today Proc. 56. doi:10.1016/j.matpr.2021.11.635.

Speck-Planche, A., V. V. Kleandrova, F. Luan, and M. N. D. S. Cordeiro. 2012. Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the *in silico* discovery of anti-breast cancer agents. European Journal of Pharmaceutical Sciences. 47. doi:10.1016/j.ejps.2012.04.012.

Spiegel, J., and H. Senderowitz. 2020. Evaluation of qsar equations for virtual screening. Int J Mol Sci. 21. doi:10.3390/ijms21217828.

Tang, T., D. Song, J. Chen, Z. Chen, Y. Du, Z. Dang, and G. Lu. 2024. Utilizing Machine Learning Models with Molecular Fingerprints and Chemical Structures to Predict the Sulfate Radical Rate Constants of Water Contaminants. Processes. 12. doi:10.3390/pr12020384.

Torres, P. H. M., A. C. R. Sodero, P. Jofily, and F. P. Silva-Jr. 2019. Key topics in molecular docking for drug design. Int J Mol Sci. 20. doi:10.3390/ijms20184574.

Vamathevan, J., D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao. 2019. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 18. doi:10.1038/s41573-019-0024-5.

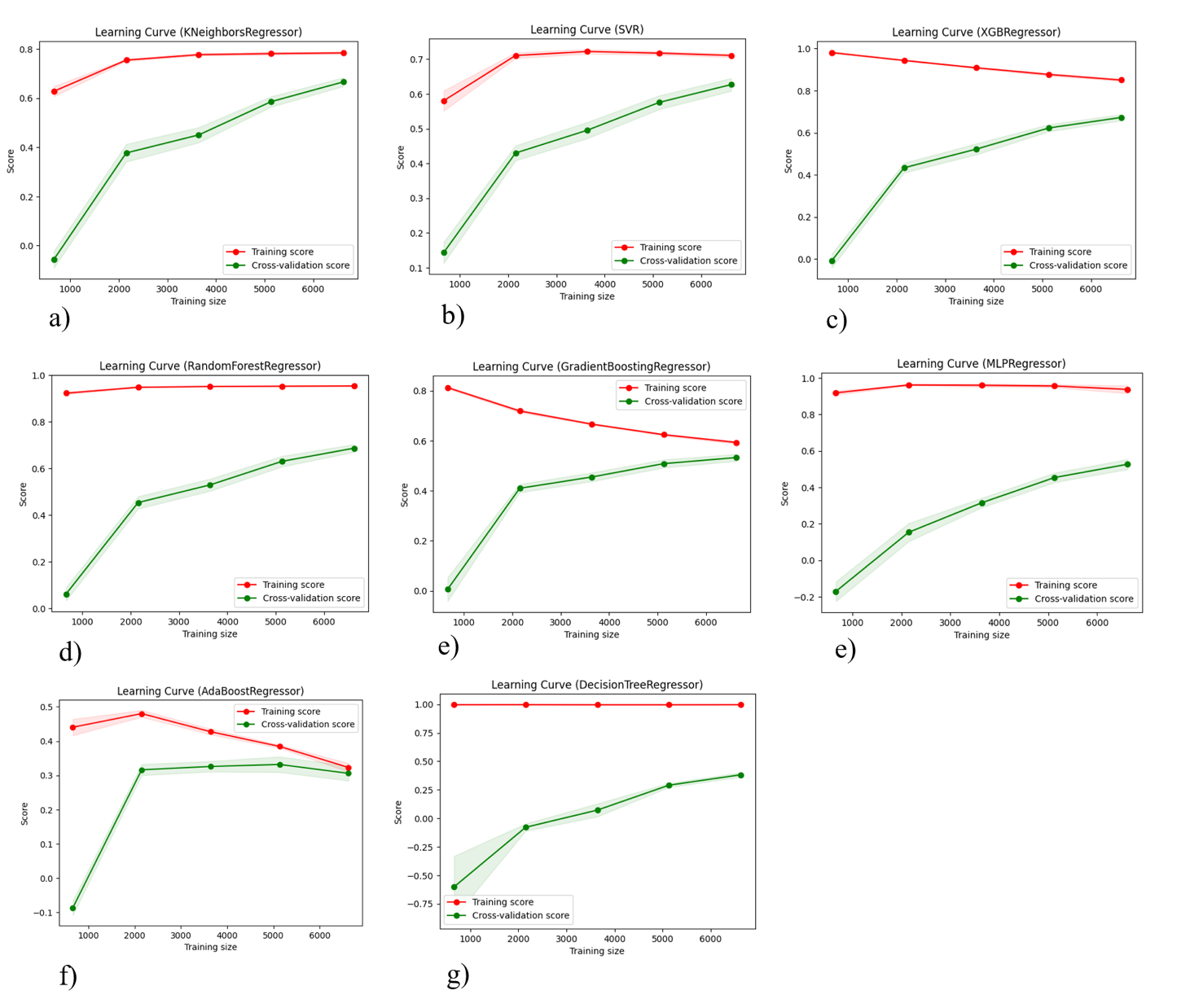
Wang, Y., Y. Guo, Q. Kuang, X. Pu, Y. Ji, Z. Zhang, and M. Li. 2015. A comparative study of family-specific protein-ligand complex affinity prediction based on random forest approach. J Comput Aided Mol Des. 29. doi:10.1007/s10822-014-9827-y.

Yamaguchi, S., M. Kaneko, and M. Narukawa. 2021. Approval success rates of drug candidates based on target, action, modality, application, and their combinations. Clin Transl Sci. 14. doi:10.1111/cts.12980.

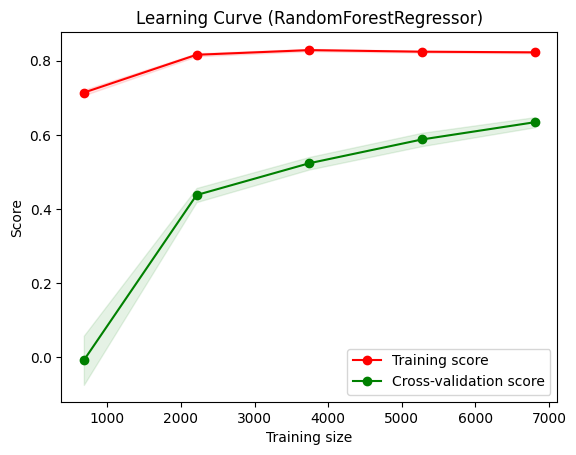
Zdrazil, B., E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum, and A. R. Leach. 2024. The CHEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 52. doi:10.1093/nar/gkad1004.

Zhao, M., L. Wang, L. Zheng, M. Zhang, C. Qiu, Y. Zhang, D. Du, and B. Niu. 2017. 2D-QSAR and 3D-QSAR Analyses for EGFR Inhibitors. Biomed Res Int. 2017. doi:10.1155/2017/4649191.

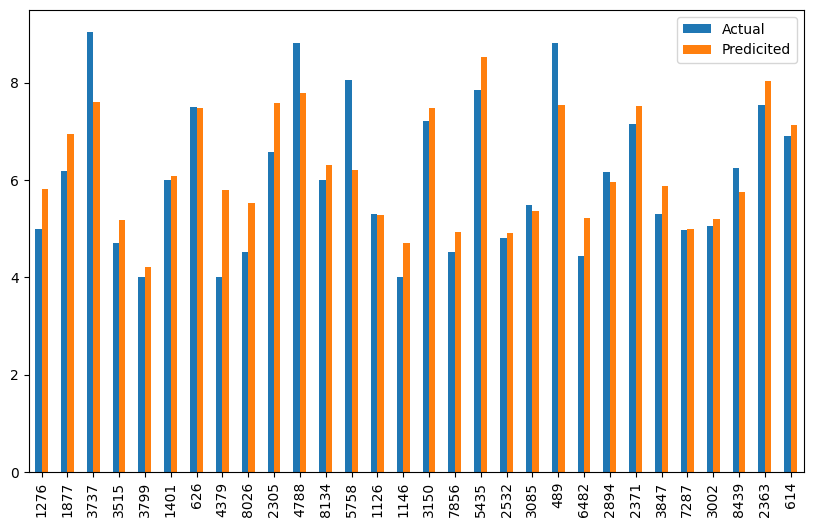
# FIGURES



**Figure 1:** Learning curves for 8 regression algorithms evaluated. The y axis corresponds to R2 statistic, and the figures are not normalized to display the same range; accordingly, values should not be directly compared between algorithms visually. Rather, overall learning trends should be assessed.



**Figure 2**: Learning curve for Random Forest with optimized hyper-parameters fit to training data set with RFE selected top 50 features.

**Figure 3**: These are from test set predictions with optimized model!!!

# TABLES

**Table 1**. Cross validation results before feature selection and hyper-parameter optimization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | MAE | MSE | RMSE | R2 |
| RF | 0.598 | 0.651 | 0.807 | 0.684 |
| XGB | 0.638 | 0.701 | 0.837 | 0.659 |
| KNN | 0.609 | 0.700 | 0.836 | 0.660 |
| GBR | 0.778 | 0.971 | 0.985 | 0.529 |
| SVR | 0.845 | 1.204 | 1.097 | 0.416 |
| DT | 0.775 | 1.287 | 1.139 | 0.375 |
| ADA | 0.999 | 1.435 | 1.198 | 0.304 |
| MLP | 1.216 | 2.064 | 1.436 | 0.001 |

**Table 2.** Mean RF cross validation results before vs. after feature selection and hyper-parameter optimization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | MAE | MSE | RMSE | R2 |
| RF | 0.598 | 0.651 | 0.807 | 0.684 |
| RFE & Optimized RF | 0.670 | 0.748 | 0.865 | 0.622 |

**Table 3.** RF hyper-parameters selected during randomized cross validation search with 100 iterations, scoring by RMSE.

|  |  |
| --- | --- |
| Parameter | Selected Value |
| min\_samples\_leaf | 5 |
| max\_features | 15 |
| max\_depth | 20 |

**Table 4.** Feature importance scores, based on mean decrease Gini scores, for final RF model.

|  |  |
| --- | --- |
| Features | Feature importance |
| Col\_343 | 0.13345 |
| PEOE\_VSA3 | 0.050568 |
| PEOE\_VSA10 | 0.045092 |
| SMR\_VSA3 | 0.04338 |
| SMR\_VSA7 | 0.032921 |
| Col\_1366 | 0.029502 |
| SlogP\_VSA10 | 0.025785 |
| BertzCT | 0.025776 |
| SMR\_VSA10 | 0.023012 |
| SlogP\_VSA6 | 0.021486 |
| EState\_VSA8 | 0.020074 |
| VSA\_EState10 | 0.018901 |
| SMR\_VSA6 | 0.018882 |