Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777218

A QSAR STUDY AND MODEL DEVELOPMENT FOR TYROSINE KINASE INHIBITORS

[*1]Sharav A. Desai, [2]Deepak S. Musmade, [1]Ravi S. Khandare, [3]Asit R. Sahu,[4]Vijay K. Patel, [5]Ankita S. Patel,[5]Ronak Patel, [5]Janki Patel

[1]Department of Pharmaceutics, Sanjivani College of Pharmaceutical Education and Research, Kopargaon, Maharashtra, India

[2]Department of Pharmaceutical Chemistry, Sanjivani College of Pharmaceutical Education and Research, Kopargaon, Maharashtra, India

[3]Department of Pharmaceutics,
Dr. Ambedkar Institute of Pharmaceutical Science, Rourkela, Odisha 769042

[4]Department of Pharmaceutics
Sharda School of Pharmacy, Pethapur, Gandhinagar,Gujarat,India

[5]Department of Pharmaceutical Chemistry
Sharda School of Pharmacy, Pethapur, Gandhinagar,Gujarat,India

***Dr. Sharav Desai,**
Email: - sharavdesai@gmail.com

ABSTRACT

The enzyme kinase is a member of a broad family that catalyses the transfer of highly energetic phosphate molecules to substrates like protein, lipids, carbohydrates, and nucleic acid. Protein tyrosine kinase is a target for therapeutic intervention since it is crucial for several immunological and signal transduction processes. Protein kinase is dysregulated, overexpressed, and mutated in a variety of disorders, including cancer and immunopathological conditions. Drug discovery techniques developed in-silico are far more affordable and quicker than those used now. The utility of the QSAR model in the current study is demonstrated in the search for novel tyrosine kinase inhibitors. 100 highly powerful compounds were chosen out of a total of 7226 compounds that were pulled from the ChEMBL database. For each chemical, more than 2000 descriptors of various classes were computed. Techniques for feature selection and outlier reduction were employed to cut down on the quantity of unimportant characteristics. The final QSAR model is created using the SVR, MLR, RF, and RT machine learning techniques. We also used the internal and external assessment criteria to assess the model's predictability and stability. For the training set, the four created models all displayed acceptable R2 values of 85, 81, 91, and 94 for MLR, SVR, RF, and RT, respectively. Apart from the RT technique, the test dataset's evaluation used the same matrix and showed virtually identical values to those of the train set. A Y-randomization test was also conducted, and the results showed that the model was not generated randomly.

Key Words: Protein tyrosine kinase, QSAR, Cancer, Machine learning, MLR, SVR, RF, RT

INTRODUCTION

A class of proteins called kinases catalyses the transfer of a phosphate group from highly energetic molecules that donate phosphate to a particular substrate. More than 500 kinase-

encoding genes may be found in the human genome. A kinase can be categorised based on the substrates it reacts with. They may be of the lipid, protein, or carbohydrate kinase types.(Manning et al., 2002; Scheeff & Bourne, 2005). Since protein kinases do have many substrates and proteins might act as substrate for different protein kinases, protein kinases are called based on the regulators of their activity(Krebs, 1985). On their serine, threonine, tyrosine, or histidine residues, proteins are phosphorylated by protein kinases. Protein structure and function are altered in a variety of ways by this phosphorylation. This alteration may lead to an increase or reduction in activity, stability, destruction marking, localization, and other outcomes (Manning et al., 2002). The human genome contains 90 protein tyrosine kinases, which phosphorylate tyrosine residues on the target protein. This phosphorylation is significant because it controls the majority of cellular metabolism, differentiation, and proliferation. There are two different receptor types that can activate tyrosine kinase. The first kind is one in which the polypeptide chain of the receptor contains the tyrosine kinase enzyme. They are referred to as RTKs, or receptor tyrosine kinases. Another type of receptors includes cytokine receptors, which are closely linked together despite being expressed by distinct genes. Similar to cytokines, cytokine receptors have a common ancestry and developed from them. The intrinsic activity is not present in cytokine receptors. Rather, the cytosolic domain of cytokine receptors is firmly linked to JAK. JAK kinase is also referred to as "just another kinase" since, at the time of its discovery, it was unclear what it did. The Jak2 was initially cloned by Wilks and colleagues more than 20 years ago (Harpur et al., 1992). This Janus kinase or another kinase family of tyrosine kinase enzymes shared hallmark characteristics with the Jak2 protein. Almost all of the body's cells express this protein, which is broadly distributed. For many ligands, including those that bind cytokines, tyrosine kinase receptors, and G-protein coupled receptors, Jak2 is a crucial downstream signalling molecule. Jak2 is a member of the Jak family and is involved in a number of activities, including cell formation, differentiation, and histone modifications. Additionally, it regulates crucial signalling activities in innate and adaptive immunity. Its related type 1 receptors, such as growth hormone receptors (GHR), prolactin receptors (PRLR), and leptin receptors (LEPR), or type II receptors, such as IFN-alpha, IFN-beta, and IFN-gamma, and various interleukins, play a crucial role in the cytoplasm in signal transduction (Sakatsume et al., 1995). Activation of Jak Kinase leads to phosphorylation of tyrosine residues in cytokine receptors. This will create docking site for signal transducers and activators of transcription (STATs)(Saltzman et al., 1998). Once the STATs protein is needed by the receptors, it will phosphorylate it and travel to the nucleus to activate gene transcription (Berry et al., 2011). Additionally, Jak2 mediates the phosphorylation of ARHGEF1 caused by angiotensin II (Guilluy et al., 2010). Jak2 also plays a significant role in cell cycle by phosphorylation of CDKN1B(Jäkel et al., 2011) Tyrosine kinase activities are often tightly regulated and strongly regulated by antagonising the impact. Tyrosine kinase can gain transforming activities in a variety of circumstances, which can impair the regulatory mechanisms that control cellular responses such cell division, development, and death (Bertram, 2000). Mutation in tyrosine kinase is associated with glioblastoma, ovarian tumours, non-small cell lung carcinoma, multiple myeloma, human bladder and cervical carcinoma(Nishikawa et al., 1994; Paul & Mukhopadhyay, 2012; Zwick et al., 2002).

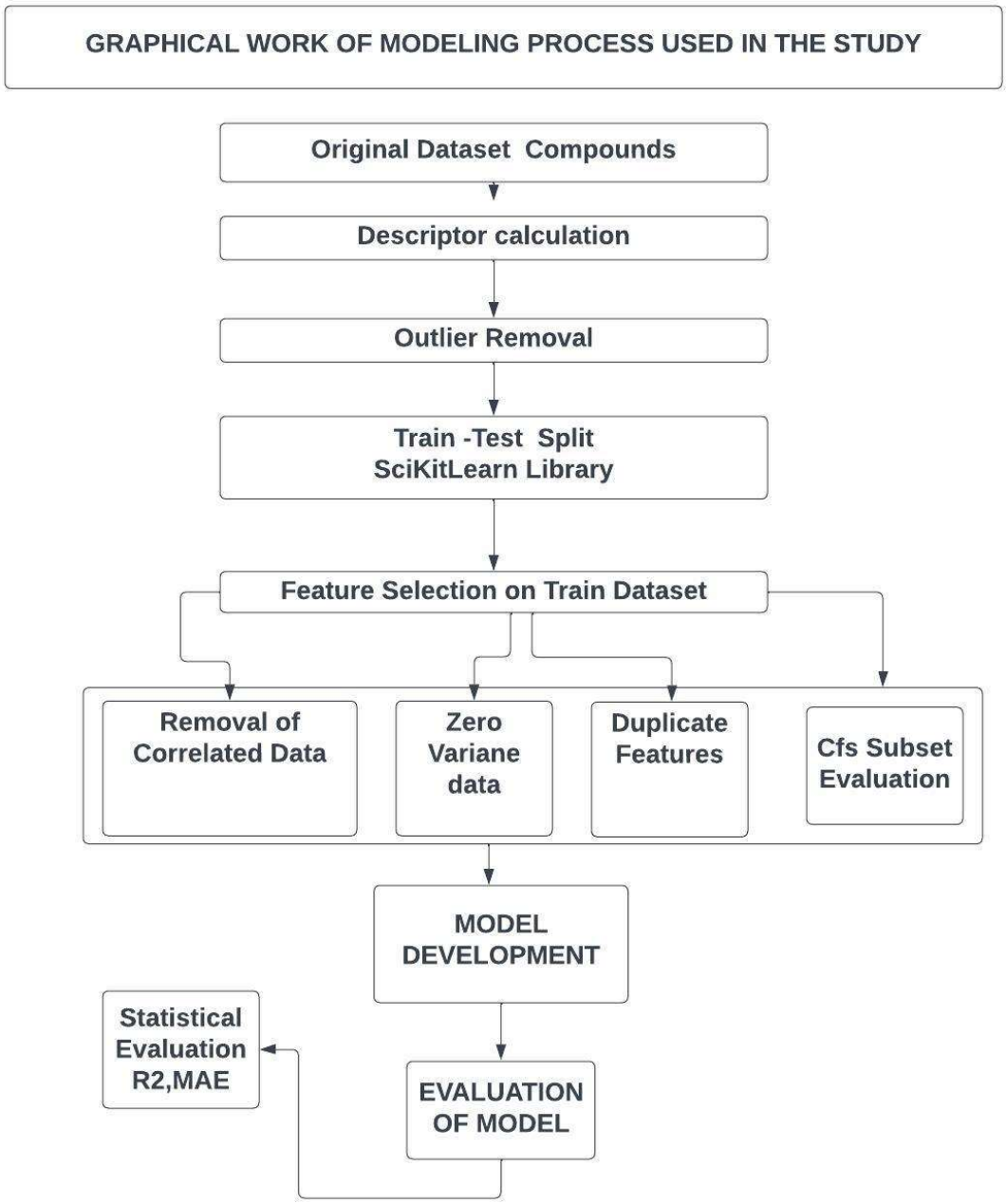


Figure 1. Graphical Work flow used in the development.

There are several more reasons to think about Jak as a possible therapeutic target in addition to using cancer as one of the key justifications. Numerous studies indicate that JaK-dependent cytokines are a key contributor to immunopathology, and that inhibiting them with biologics can improve immune-mediated responses (Schwartz et al., 2017). The current, conventional approach to drug research and design takes a lot of time and money. A breakthrough therapeutic agent's introduction to the market is predicted to take 10 to 15 years and 800 million dollars in funding. Pharmaceutical firms now prioritise saving time and money while still developing new drugs without sacrificing medication quality (Dickson & Gagnon, 2004; Pan et al., 2013). We

used a technology that allows us to screen a large number of chemicals at once to reach this high throughput. HTS approaches were quite helpful, however the level of significance of achievement at the conclusion of the development phase was relatively low (Clark et al., 2010; Szymański et al., 2012). The present discovery process can be facilitated by a mix of advanced computer tools, biological research, and chemical synthesis. The use of computers in the drug discovery process is referred to as computer-aided drug design. This field focuses on developing drugs using a variety of methods, including molecular docking, simulation, machine learning, and drug receptor interactions. The author of the current paper employed the QSAR approach to forecasting the new Tyrosine Kinase inhibitor. The QSAR approach to drug design is ligand-based. The foundation of the QSAR model is the idea that a compound's biological properties are closely tied to its structural characteristics. The QSAR model entails building a mathematical equation based on the computed structural properties and their biological activity (Dong & Zheng, 2008; Karelson et al., 1996). The QSAR model presupposes that substances with comparable structural characteristics have comparable biological activity. By gathering the lead chemicals with recognized biological action, a model is initially created. Utilizing a number of machine learning techniques developed during the development process, a model is utilized to forecast the activity of an unidentified molecule. QSAR model now a days are widely used to modify existing molecule to enhance its biological activity(Acharya et al., 2010; Yang, 2010).

In the current study, the author developed a QSAR model for the prediction of inhibitory activity against tyrosine kinase using a supervised machine learning technique. Statistical parameters and tests are used to train, test, and verify a model.

MATERIAL AND METHODS

Dataset collection and preparation

The ChEMBL web server was utilized to download dataset required to build QSAR model. The compounds with known biological activity (IC50) values were downloaded by searching the target section for tyrosine kinase enzyme. Initial dataset comprises of total 7226 SMILES with their inhibitory activities was downloaded to local computer(Gaulton et al., 2017; Mendez et al., 2019) **(Suplimenatary_1)**. From downloaded database top 100 compounds with potent inhibitory activity were selected for the development of QSAR model. The IC50 values were converted in to pIC5 values (-Log of IC50 values). The structures were available in to SMILE format and were converted into sdf format using OpenBabel software (O'Boyle et al., 2008a, 2008b) **(Suplimenatry_2)**. During conversion hydrogen atom was added to make compounds explicit and to mimic the real situations(Kumar & Roy, 2020).

Descriptor calculation

In the present work, total 8 groups of descriptors namely 'constitutional indices', 'Ring Descriptors', 'Topological indices', 'connectivity indices', 'Functional group counts', 'Atom-type-E-state indicies','2D-atom pairs' and 'molecular properties' were calculated using Alvadesc V. 2.0.10(Mauri, 2020). A total of 2438 descriptors were calculated for the given 100 'sdf' structures. The database with 2438 descriptors and 100 compounds was checked for the presence of outliers **(Suplimentary_3)**. In the present work, both structural and activity outliers were considered insignificant for model development. Inter quartile range was used to calculate the outliers and total of 64 structural and functional outliers were removed from the database

(Suplimentary_4). Final data base of 36 compound was then used for selection of significant descriptors based on several statistical calculations(Tropsha, 2010) **(Suplimentary_5)**.

Descriptor selection and removal

It is important for any machine learning model to have a smaller number of significant features to develop the prediction model. First, the dataset was divided in to training and testing dataset using 70 % and 30 % ration **(Suplimentary_6)**. Several feature selection techniques were used to reduce the number of descriptors to develop final QSAR model. All the techniques were applied on the training set. Initially, the training dataset was filtered to remove the highly correlated descriptors (R>0.8) **(Suplimentary_9)** and descriptors with low variance (<1%) **(Suplimentary_7)**. Both of which will not provide enough prediction power to the model. Further, descriptors with the duplicate values were also removed from the dataset(Pedregosa et al., 2011). All the features removal techniques were applied using sklearn package of python programming language. For the identification of relevant descriptors, CfsSubsetEval and locally predictive features were then applied. CfsSubsetEval generates subsets of the characteristics with low intercorrelation and strong correlation to the class/activity. If there isn't already an attribute in the subset that has a stronger correlation with the class/activity, the locally predictive attribute detects locally predictive qualities and iteratively adds attributes with the greatest correlation with the class (Hall et al., 2009; Kotthoff et al., 2017; Li et al., 2017).s

Regression algorithms

The following algorithms support machine, multiple linear regression, random forest regressor and random tree regressor were used to build QSAR model. All the algorithms were implemented using WEKA.

Multiple linear regression

It is also called as multiple regression and it uses several explanatory variables to predict the outcomes of response or target variables. M5 method for linear regression was adopted to build a regression equation. This equation is developed by removing smallest standardize coefficient until no improvement is observed in estimate of error given by Akaike information criterion(Pandis, 2016; Uyanık & Güler, 2013).

Support vector machine

Support vector regression uses the same principle as the support vector machine. In the present work poly kernel was used to developed to best fit line called hyperplane to predict the continuous discrete variable(Smola & Schoelkopf, 1998).

Random Forest

For regression, the supervised machine learning algorithm random forest employs ensemble learning. The ensemble technique proposes training numerous models or trees on the same data and averaging their outputs. For regression task in random forest, mean prediction of individual tree was used(Breiman, 2001).

Random tree regression

For evaluating the model's predictive power, a different supervised machine learning tree-based algorithm called random tree was applied. To assess the model's effectiveness, iterative portioning was used to split the data, and the optimal split was chosen using a reduction in the impurity index.(Holmes et al., 1999; Wang & Witten, 1997).

Evaluation of Model

Many statistical validation criteria were used to assess the resilience and prediction capacity of the QSAR model created by four distinct algorithms. To determine the model's importance, we employed both internal and external validation techniques in this study. Pearson's correlation coefficient, coefficient of determination, mean absolute error, and root mean squared error were calculated for each model. The model's performance was also examined using an external dataset or test dataset to determine the model's prediction power. Moreover, we used the Y-randomization or Y-Scrambling approach to test the model's resilience.(Christopher Rucker, Gerta Rucker, 2007; Rücker et al., 2007). This test is employed to figure out whether the model was created by chance or not. It was carried out by first developing the model on the original dataset, and the model's quality was recorded. The performance of the model was assessed after we randomized the target or activity column in the next step to swap out the right target-feature combination with the erroneous one. This scrambling was carried out 50 times, and the output of 50 QSAR models was recorded and evaluated.

RESULT AND DISCUSSION

We created a QSAR model to forecast the suppressive efficacy of drugs against the Tyrosine Kinase enzyme, which is involved in a variety of malignancies and other immunological disorders. In order to identify the substances with demonstrated inhibitory action against the chosen target, we used the Chembl database web server. For the purpose of creating a QSAR model, 100 compounds with strong inhibitory action were downloaded to a local computer. Structures in the SMILE format were available and were transferred into structure data format (sdf). We have determined about 2438 descriptors of the specified categories as discussed in materials and method part of the article. Descriptors are the chemical identifiers in the numerical form for the given set of the compounds.

The advancing database was later explored to eliminate outliers. Outlier is the datapoint possessing inconsistent quantity compared to other datapoint. Existence of outlier in the dataset can seriously affect the quality of model. Outlier detection done by utilizing inter quartile range technique. This technique is accepted widespread to filter outliers from the vast dataset. The dataset was separated into four equal parts, or quartiles, and the IQR was calculated using the distance between the quartiles. The values outside from 1.5* IQR below Q1 and 1.5* IQR above Q3 are what we have labelled as outliers. The evolving database was cleared of a total of 64 structural and functional outliers.

The leftover 36 structures were divided amongst training and testing database having random split of 70 % and 30 %. This was performed to evaluate the produced model on the hidden data with known target values.

To increase the estimator accuracy scores and their functionality on high-dimensional databases, we have adopted feature selection, also referred as dimensionality reduction approaches. To exclude the descriptors with zero variance included inside the dataset, we utilised the scikit-learn 1.1.2 library's variance threshold technique. As a result, the prediction power of the final model cannot be significantly increased by these. The number of descriptors was decreased by removing the 1936 constant columns that were discovered in. The duplicate

transform function in the python scikit-learn 1.1.2 package was used to eliminate 121 duplicate columns.

To eliminate the multicollinear characteristics, we manually constructed a function. High correlation features are linearly dependent and affect the target in the same way. We maintained a cut off of 0.8, meaning that characteristics with 80% similarity were eliminated from the dataset in order to remove such features. We deleted 374 strongly linked characteristics from the database because of our study.

CfsSubsetEval and locally predictive characteristics were used to choose the final features from an existing database that had 74 different features. The Weka machine learning program's choose attribute function was utilised to use it. We have found a total of 09 characteristics using CfsSubsetEval that may be utilised to build a QSAR model for the prediction of inhibitory activity against a chosen target Tyrosine kinase. To develop the model, we employed supervised machine learning. Known activity levels and existing data from the training dataset are both utilised in this method. The model was constructed using a total of four separate algorithms: MLR, SVR, RF, and RT. The performance of the model is validated using all the statistical parameters provided in the **(Table 1).**

Table 1. Statistical results for QSAR model.

Algorithm	**Statistical Parameters**		
MLR	**R^2**	**MAE**	**RMSD**
Training set	0.85	0.06	0.0936
Test Set	0.82	0.15	0.201
SVR			
Training set	0.81	0.04	0.106
Test Set	0.93	0.09	0.1577
RF			
Training set	0.92	0.03	0.0704
Test Set	0.85	0.1	0.184
RT			
Training set	0.94	0.01	0.0619
Test Set	0.96	0.1	0.1775

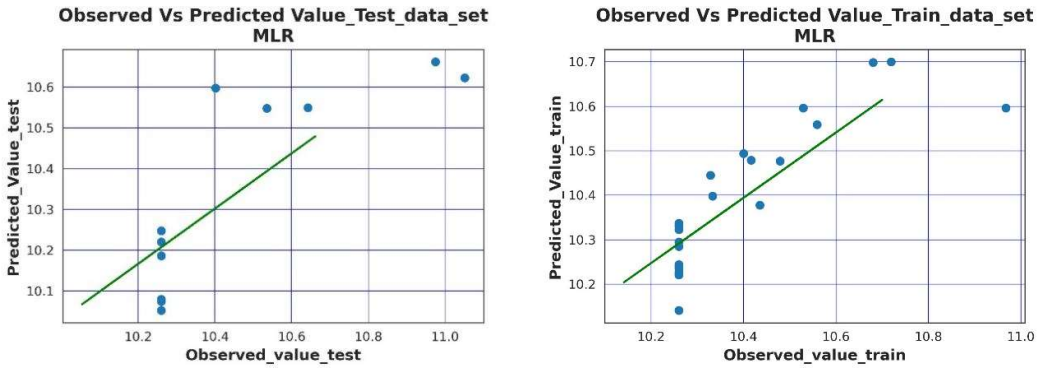


Figure 2. Scatter plot of observed and predicted values for training and testing dataset from the developed Multiple linear regression model.

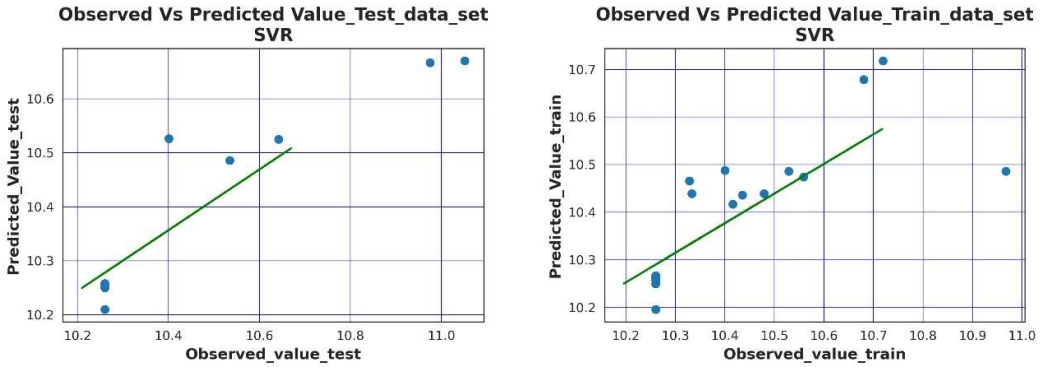


Figure 3.Scatter plot of observed and predicted values for training and testing dataset from the developed Support Vector regression model

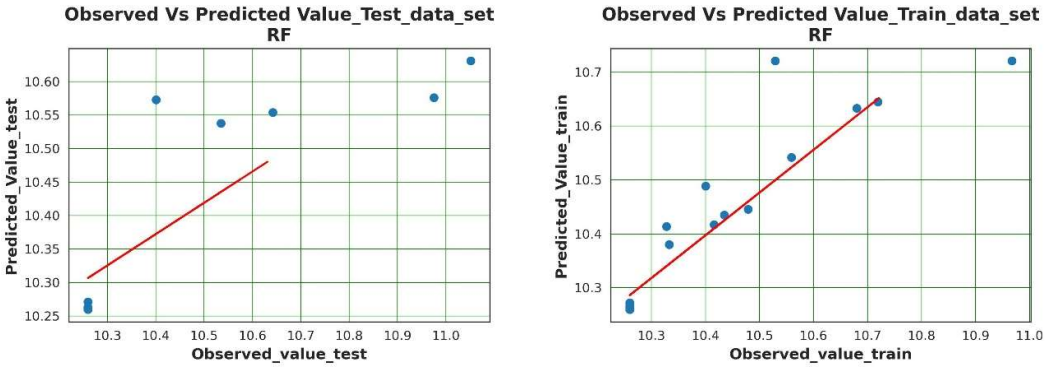


Figure 4. Scatter plot of observed and predicted values for training and testing dataset from the developed Random Forest model.

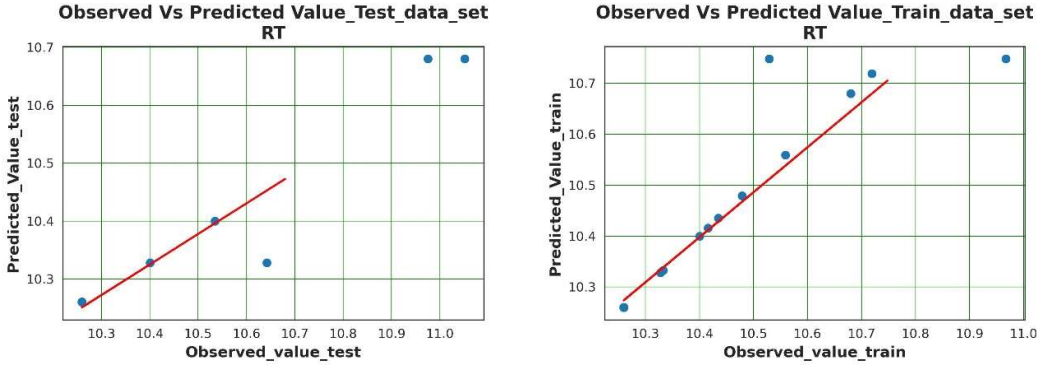
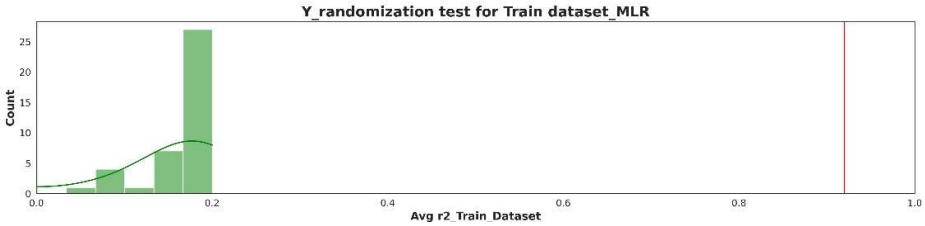


Figure 5. Scatter plot of observed and predicted values for training and testing dataset from the developed Random trees model.



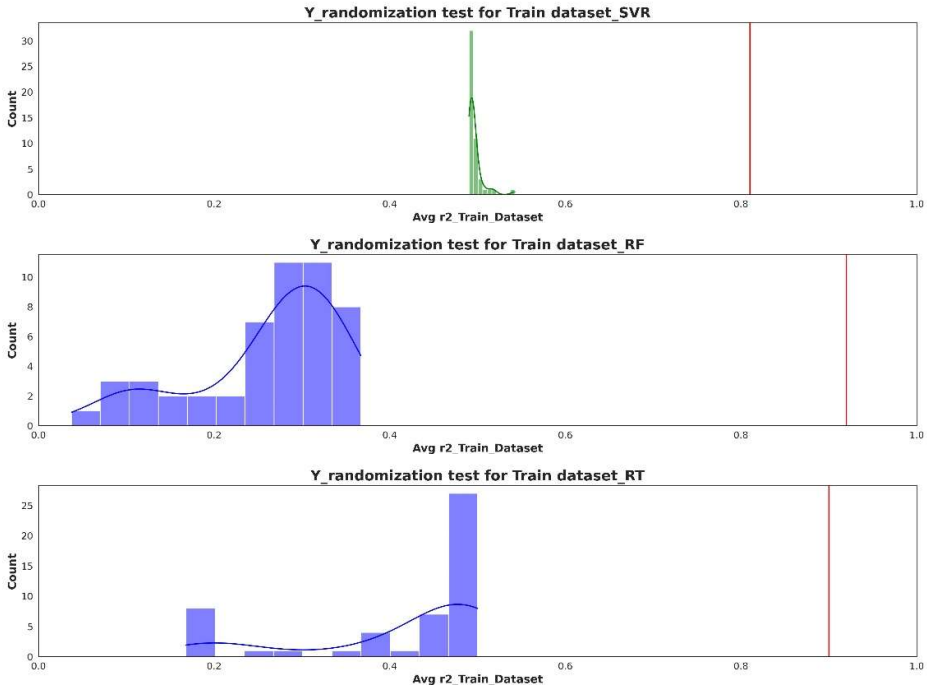


Figure 6. The performance of all four QSAR models built with randomized data is inferior to that of the QSAR model developed with non-permuted data set. (a); Y-randomization test for train dataset for MLR, (b); Y-randomization test for train dataset for SVR, (c); Y-randomization test for train dataset for RF, (d); Y-randomization test for train dataset for RT.

The R2 values for all of the models are above 90% according to the data in the table, and the highest R2 values are found for the RF and RT algorithms. 09 relevant descriptors were selected for the model from a huge pool of 2439 descriptors that were initially generated for chemicals. The model's equation can be used to predict the chemicals' inhibitory action. Also, the scatter plot created between observed and anticipated values validates the similarity and suggests the model's resilience and stability **(Figure 2,3,4 & 5)**.

pIC50 = 0.0217 * max_conj_path + 0.1892 * nR07 + -0.0654 * SsCH3 + -0.108 * B06[N-N] + -0.0821 * ALOGP + 10.4397

Equation 1. Multiple linear regression prediction equation.

weights (not support vectors): + 0.2336 * (normalized) max_conj_path - 0.0734 * (normalized) nCIC + 0.2731 * (normalized) nR07 + 0.0957 * (normalized) nRCONR2 - 0.014 * (normalized) SsCH3 - 0.0299 * (normalized) SsssCH - 0.0769 * (normalized) MaxsCH3 - 0.0047 * (normalized) B06[N-N] + 0.0024 * (normalized) ALOGP + 0.0886

Equation 2. Support Vector regression equation for prediction.

Table 2. Y-Randomization test results.

Algorithm	Actual R^2	Y-randomized R^2
MLR	0.85	0.11
SVR	0.81	0.49
RF	0.92	0.25
RT	0.94	0.41

Table 3. Significant descriptors involved in model formation.

Descriptors	Description	Group
max_conj_path	maximum number of atoms that can be in conjugation with each other	Constitutional indices
nCIC	number of rings (cyclomatic number)	Ring descriptors
nR07	number of 7-membered rings	Ring descriptors
nRCONR2	number of tertiary amides (aliphatic)	Functional group counts
SsCH3	Sum of sCH3 E-states	Atom-type E-state indices
SsssCH	Sum of sssCH E-states	Atom-type E-state indices
MaxsCH3	Maximum sCH3	Atom-type E-state indices
B06[N-N]	Presence/absence of N – N at topological distance 6	2D Atom Pairs
ALOGP		

```
RandomTree
==========

MaxsCH3 < 1.78
|   SsCH3 < 1.24 : 10.75 (2/0.05)
|   SsCH3 >= 1.24
|   |   nR07 < 0.5
|   |   |   SsCH3 < 1.41
|   |   |   |   MaxsCH3 < 1.33
|   |   |   |   |   SsssCH < 0.29 : 10.4 (1/0)
|   |   |   |   |   SsssCH >= 0.29 : 10.42 (1/0)
|   |   |   |   MaxsCH3 >= 1.33 : 10.56 (1/0)
|   |   |   SsCH3 >= 1.41 : 10.33 (1/0)
|   |   nR07 >= 0.5
|   |   |   SsCH3 < 1.45 : 10.68 (1/0)
|   |   |   SsCH3 >= 1.45 : 10.72 (1/0)
MaxsCH3 >= 1.78
|   SsssCH < 0.16
|   |   ALOGP < 1.93 : 10.48 (1/0)
|   |   ALOGP >= 1.93
|   |   |   ALOGP < 2.04 : 10.33 (1/0)
|   |   |   ALOGP >= 2.04 : 10.44 (1/0)
|   SsssCH >= 0.16 : 10.26 (14/0)
```

Figure 7. Random tree created for significant descriptors.

For the 11 compounds that were kept apart prior to model creation, we conducted external validation. **(Supplementary_10)**.

The test dataset, which was kept secret during the model-development process, was utilised to predict activity using the generated model. In the equation, the descriptors stand in for both negative and positive values. These show the contributions they made to the compounds' ultimate activity. Positive contribution shows that adjectives contribute to the activity's positive, whereas negative contributions do the opposite. The supplemental file that is provided provides a detailed description of the kind of descriptor and its role in the final QSAR model.

CONCLUSION

Tyrosine protein kinase plays a significant role in development of several types of cancer along with it is also involved in severe immunopathological conditions. several studies are published showing the use of computer aided drug design and its uses in the development of novel inhibitors against tyrosine kinase. Currently we have high computational power with super-fast super vised machine learning algorithms available. QSAR model is getting more popularity in the novel drug discovery and design as they are relatively easy to develop and if evaluated and validated correctly can produce nearly true predictions. In the present investigation, QSAR model for prediction of inhibitory activity against tyrosine kinase enzyme was developed. Once the model is validated through internal and external parameters, it can be used to screen a very large database within a very short period. Such a model was developed using supervised machine learning technique and under those four different algorithms were applied. A very

large database of compounds with known inhibitory activity used for model development. Before development, several tools were applied to filter the insignificant features. Remaining significant features were divided in training and testing dataset. training model was statistically evaluated and was tested on test dataset. QSAR model developed can be further tested through invitro and in vivo activity.

Data availability statement: Data supporting the result of this paper is supplied as supplementary file. More information can be obtained from corresponding author.

Conflict of interest: The author declares no conflict of interest.

Abbreviation:

MLR: Multiple Linear regression
SVR: Support vector regression
RF: Random Forest
RT: Random tree
IC50: Inhibitory concentration
QSAR: Quantitative structural activity relationship
RMSD: Root Square Mean Deviation

REFERENCES

Acharya, C., Coop, A., E. Polli, J., & D. MacKerell, A. (2010). Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Current Computer Aided-Drug Design*, *7*(1). https://doi.org/10.2174/157340911793743547

Berry, D. C., Jin, H., Majumdar, A., & Noy, N. (2011). Signaling by vitamin A and retinol-binding protein regulates gene expression to inhibit insulin responses. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(11). https://doi.org/10.1073/pnas.1011115108

Bertram, J. S. (2000). The molecular biology of cancer. In *Molecular Aspects of Medicine* (Vol. 21, Issue 6). https://doi.org/10.1016/S0098-2997(00)00007-8

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Christopher Rucker, Gerta Rucker, M. M. (2007). Y-Randomization-A useful tool in QSAR Validation,or Folklore? *Journal of Chemical Information and Modeling*, *47*(6).

Clark, R. L., Johnston, B. F., Mackay, S. P., Breslin, C. J., Robertson, M. N., & Harvey, A. L. (2010). The Drug Discovery Portal: A resource to enhance drug discovery from academia. In *Drug Discovery Today* (Vol. 15, Issues 15–16). https://doi.org/10.1016/j.drudis.2010.06.003

Dickson, M., & Gagnon, J. P. (2004). Key factors in the rising cost of new drug discovery and development. In *Nature Reviews Drug Discovery* (Vol. 3, Issue 5). https://doi.org/10.1038/nrd1382

Dong, X., & Zheng, W. (2008). A New Structure-Based QSAR Method Affords both Descriptive and Predictive Models for Phosphodiesterase-4 Inhibitors. *Current Chemical Genomics*, *2*. https://doi.org/10.2174/1875397300802010029

Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The

ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1). https://doi.org/10.1093/nar/gkw1074

Guilluy, C., Brégeon, J., Toumaniantz, G., Rolli-Derkinderen, M., Retailleau, K., Loufrani, L., Henrion, D., Scalbert, E., Bril, A., Torres, R. M., Offermanns, S., Pacaud, P., & Loirand, G. (2010). The Rho exchange factor Arhgef1 mediates the effects of angiotensin II on vascular tone and blood pressure. *Nature Medicine*, *16*(2). https://doi.org/10.1038/nm.2079

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1). https://doi.org/10.1145/1656274.1656278

Harpur, A. G., Andres, A. C., Ziemiecki, A., Aston, R. R., & Wilks, A. F. (1992). JAK2, a third member of the JAK family of protein tyrosine kinases. *Oncogene*, *7*(7).

Holmes, G., Hall, M., & Frank, E. (1999). Generating Rule Sets from Model Trees. *Twelfth Australian Joint Conference on Artificial Intelligence*, 1–12.

Jäkel, H., Weinl, C., & Hengst, L. (2011). Phosphorylation of p27Kip1 by JAK2 directly links cytokine receptor signaling to cell cycle control. *Oncogene*, *30*(32). https://doi.org/10.1038/onc.2011.68

Karelson, M., Lobanov, V. S., & Katritzky, A. R. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical Reviews*, *96*(3). https://doi.org/10.1021/cr950202r

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2017). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, *18*. https://doi.org/10.1007/978-3-030-05318-5_4

Krebs, E. G. (1985). The phosphorylation of proteins: a major mechanism for biological regulation. *Biochemical Society Transactions*, *13*(5). https://doi.org/10.1042/bst0130813

Kumar, V., & Roy, K. (2020). Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel 3C-like protease (3CLpro) enzyme inhibitors against SARS-CoV diseases. *SAR and QSAR in Environmental Research*, *31*(7). https://doi.org/10.1080/1062936X.2020.1776388

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. In *ACM Computing Surveys* (Vol. 50, Issue 6). https://doi.org/10.1145/3136625

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., & Sudarsanam, S. (2002). The protein kinase complement of the human genome. In *Science* (Vol. 298, Issue 5600). https://doi.org/10.1126/science.1075762

Mauri, A. (2020). alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In *Methods in Pharmacology and Toxicology* (pp. 801–820). Humana Press Inc. https://doi.org/10.1007/978-1-0716-0150-1_32

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., de Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., … Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky1075

Nishikawa, R., Ji, X. D., Harmon, R. C., Lazar, C. S., Gill, G. N., Cavenee, W. K., & Huang, H. J. S. (1994). A mutant epidermal growth factor receptor common in human glioma confers

enhanced tumorigenicity. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(16). https://doi.org/10.1073/pnas.91.16.7727

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008a). Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, *2*(1). https://doi.org/10.1186/1752-153X-2-5

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008b). Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, *2*(1). https://doi.org/10.1186/1752-153X-2-5

Pan, S. Y., Zhou, S. F., Gao, S. H., Yu, Z. L., Zhang, S. F., Tang, M. K., Sun, J. N., Ma, D. L., Han, Y. F., Fong, W. F., & Ko, K. M. (2013). New perspectives on how to discover drugs from herbal medicines: CAM'S outstanding contribution to modern therapeutics. *Evidence-Based Complementary and Alternative Medicine*, *2013*. https://doi.org/10.1155/2013/627375

Pandis, N. (2016). Multiple linear regression analysis. In *American Journal of Orthodontics and Dentofacial Orthopedics* (Vol. 149, Issue 4). https://doi.org/10.1016/j.ajodo.2016.01.012

Paul, M. K., & Mukhopadhyay, A. K. (2012). Tyrosine kinase – Role and significance in Cancer. *International Journal of Medical Sciences*. https://doi.org/10.7150/ijms.1.101

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*.

Rücker, C., Rücker, G., & Meringer, M. (2007). Y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, *47*(6). https://doi.org/10.1021/ci700157b

Sakatsume, M., Igarashi, K., Winestock, K. D., Garotta, G., Larner, A. C., & Finbloom, D. S. (1995). The Jak Kinases Differentially Associate with the α and β (Accessory Factor) Chains of the Interferon γ Receptor to Form a Functional Receptor Unit Capable of Activating STAT Transcription Factors. *Journal of Biological Chemistry*, *270*(29), 17528–17534. https://doi.org/10.1074/jbc.270.29.17528

Saltzman, A., Stone, M., Franks, C., Searfoss, G., Munro, R., Jaye, M., & Ivashchenko, Y. (1998). Cloning and characterization of human Jak-2 kinase: High mRNA expression in immune cells and muscle tissue. *Biochemical and Biophysical Research Communications*, *246*(3). https://doi.org/10.1006/bbrc.1998.8685

Scheeff, E. D., & Bourne, P. E. (2005). Structural Evolution of the Protein Kinase–Like Superfamily. *PLoS Computational Biology*, *1*(5). https://doi.org/10.1371/journal.pcbi.0010049

Schwartz, D. M., Kanno, Y., Villarino, A., Ward, M., Gadina, M., & O'Shea, J. J. (2017). JAK inhibition as a therapeutic strategy for immune and inflammatory diseases. In *Nature Reviews Drug Discovery* (Vol. 16, Issue 12). https://doi.org/10.1038/nrd.2017.201

Smola, A. J., & Schoelkopf, B. (1998). *A tutorial on support vector regression*.

Szymański, P., Markowicz, M., & Mikiciuk-Olasik, E. (2012). Adaptation of high-throughput screening in drug discovery-toxicological screening tests. In *International Journal of Molecular Sciences* (Vol. 13, Issue 1). https://doi.org/10.3390/ijms13010427

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. In *Molecular Informatics* (Vol. 29, Issues 6–7). https://doi.org/10.1002/minf.201000061

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*. https://doi.org/10.1016/j.sbspro.2013.12.027

Wang, Y., & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. *Poster Papers of the 9th European Conference on Machine Learning*.

Yang, S. Y. (2010). Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. In *Drug Discovery Today* (Vol. 15, Issues 11–12). https://doi.org/10.1016/j.drudis.2010.03.013

Zwick, E., Bange, J., & Ullrich, A. (2002). Receptor tyrosine kinases as targets for anticancer drugs. In *Trends in Molecular Medicine* (Vol. 8, Issue 1). https://doi.org/10.1016/S1471-4914(01)02217-1