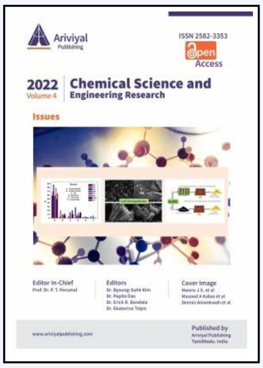
Chemical Science &
Engineering Research

Ariviyal Publishing

Research Article

DOI: 10.36686/Ariviyal.CSER.2022.04.10.056

Chem. Sci. Eng. Res., 2022, 4(10), 46-53.



Ariviyal Publishing

ISSN 2582-3353
Open Access

2022 Volume 4 | Chemical Science and Engineering Research

Issues

Editor In-Chief
Prof. Dr. P. T. Perumal

Editors
Dr. Byoung-Suhk Kim
Dr. Papita Das
Dr. Erick R. Bandala
Dr. Ekaterina Tsipis

Cover Image
Meera J.S. et al
Masood A Kaloo et al
Dennis Amankwah et al

www.ariviyalpublishing.com

Published by
Ariviyal Publishing
TamilNadu, India

Development of Interpretable QSAR Model for Quick Screening of Inhibitors against Tyrosine Protein Kinase JAK-2

Sharav Desai*[a] and Dhananjay Meshram[b]

[a]*Department of Pharmaceutical Microbiology and Biotechnology, Pioneer Pharmacy Degree College, Vadodara-390019, Gujarat, India.*
[b]*Department of Quality assurance, Pioneer Pharmacy Degree College, Vadodara-390019, Gujarat, India.*

***Corresponding author E-mail address**: Sharavdesai@gmail.com (Sharav Desai)

ISSN: 2582-3353

Publication details
Received: 01st May 2022
Revised: 15th June 2022
Accepted: 15th June 2022
Published: 04th July 2022

Abstract: Kinase belongs to large family of enzymes that catalyse transfer of high energy phosphate molecule to substrates like protein, lipids, carbohydrates and nucleic acid. Protein tyrosine kinases are becoming therapeutically active target as It plays a significant role in several signal transduction and immunological reactions. Dysregulation, overexpression and mutation of protein kinase found in many diseases including cancer and immunopathological conditions. InSilico methods of drug discovery are considerably cheaper and faster compared to traditional methods available today. In the present work, the use of QSAR model is shown in the discovery of new tyrosine kinase inhibitors. Total of 7226 compounds retrieved from the ChEMBL database and were used after manual curation. More than 2000 descriptors of different class were calculated for individual compounds. Manual curation, outlier removal and feature selection techniques were used to reduce the number of insignificant features. Four machine learning algorithms called SVR, MLR, RF and RT are used to build the final QSAR model. We also have applied the internal and external evaluation parameters to check the model stability and its prediction power. All the four models developed were showing acceptable range of R^2 like 59.40, 58.84, 97.1, and 99.32 for MLR, SVR, RF and RT respectively on training set. Similarly test dataset was evaluated with the same matrix and showing nearly similar values to train set except RT algorithm. Y-randomization test also performed and confirmed that model is not produced by chance.

Keywords: Protein tyrosine kinase; QSAR; Cancer; Machine learning; MLR; SVR; RF; RT

1. Introduction

Kinases are grouping of proteins those catalyses the transfer of phosphate group from high energy phosphate donating molecules to specific substrate. Human genome contains more than 500 kinase encoding genes. Kinases can be classified according to substrates they act upon. They can be of lipid kinase, protein kinase or carbohydrate kinase.[1,2] Protein kinase are named based on the regulators of their activity as it is being observed that protein kinases do have a multiple substrates and proteins can serve as substrate for more than one protein kinase.[3] Protein kinase does phosphorylation of proteins on their serine, threonine, tyrosine or histidine residues. Such phosphorylation modifies the structure and function of proteins in many ways. This modification can result in increase or decrease in the activity, stabilization, marking for destruction, localization and others. Protein kinase covers majority of the all kinases and are also well studied. These kinases play a major role in signalling in the cell.[1] Protein tyrosine kinase, of which 90 there are in human genome, phosphorylates tyrosine residue on the target protein. This phosphorylation is noteworthy because they regulate most aspects of cell proliferation, differentiation and cell metabolism. We have two types of receptors those can activate tyrosine kinase, first is a type, in which tyrosine kinase enzyme is integral part of receptor's polypeptide chain. These are called as receptor tyrosine kinase (RTKs). In another class where receptor such as cytokine receptors, receptors and kinase are encoded by the different genes yet bound together tightly. Like cytokines, cytokine receptors are also evolved from the common ancestors and have a common structure. Cytokine receptors do not possess the intrinsic activity. Rather the tightly bound JAK to cytosolic domain of cytokine receptors. JAK kinase is also known as just another kinase, because when they were discovered, their function was unknown. It has been over more than 2 decades when the Jak2 was first cloned by Wilks and colleagues.[4] This Jak2 protein shared a hall mark features with Janus kinase or just another kinase family of tyrosine kinase enzymes. This protein is widely expressed and virtually found in every cell of the body. Jak2 is an important downstream signalling molecule for number of ligands including those that binds cytokine, tyrosine kinase receptor and G-protein coupled receptors. Among Jak family Jak2 is involved in various processes such as cell growth, development, differentiation or histone modifications. It also mediates essential signalling events in adaptive and innate immunity. In the cytoplasm it plays an essential role in signal transduction via its associated type 1 receptors such as growth hormone receptors (GHR), Prolactin (PRLR), Leptin (LEPR) or type II receptors including IFN-alpha, IFN-beta, IFN-gamma and multiple interleukins.[5] 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).[6] Subsequently, it will phosphorylate STATs protein once they are required to the receptors and will move to nucleus for the activation of gene transcription.[7] In addition to that Jak2 mediates angiotensin-2 induced ARHGEF1 phosphorylation[8] Jak2 also plays a significant role in cell cycle by phosphorylation of CDKN1B.[9] Normally the functions of tyrosine kinase are highly regulated and tightly controlled by antagonizing the effect. There are several instances where tyrosine kinase acquires transforming functions and it will hamper the regulatory functions in cellular responses like cell division, growth and death.[10] Mutation in tyrosine kinase is associated with glioblastoma, ovarian tumours, non-small cell lung carcinoma, multiple myeloma, human bladder and cervical carcinoma.[11–13] Apart from considering cancer as a main reason to use Jak as a target, there are several other reasons also to consider it as a potential therapeutic target. Plenty of literature suggest that JaK dependent cytokines are major cause of immuno pathology and blocking such cytokines with biologics can be beneficial in immune mediated responses.[14] Current traditional methodology adopted for drug discovery and design requires a long time and huge amount of money. It has been estimated that to introduce a novel therapeutic agent in to market requires around 10-15 years and around US 800$ million of investments. Today, pharmaceutical companies are focusing on reducing the time and money in development of new drug without affecting the quality of drug.[15,16] To achieve this high through put technique was adopted, in which we can screen a huge number of compounds at one time. HTS techniques helped a lot but very low significant success was found at the end stage of development process.[17,18] Today, we have combination of high computational techniques, biological science, chemical synthesis which can facilitate the current discovery process. The term computer aided drug design is adopted for the use of computer in drug discovery process. This branch focuses on drug design based on drug receptor interactions, molecular docking, simulation, machine learning and many other techniques. In the present work, the author has used QSAR (Quantitative Structural Activity Relationship) method to predict the novel Jak-2 inhibitor. QSAR method is ligand-based drug design. The principle of QSAR model is based on the belief that biological property of the compound is directly correlated to its structural features. QSAR model involves the construction of mathematical equation based on the structural features calculated and its biological activity.[19,20] QSAR model assumes that compounds with similar structural properties have similar biological activities. A model is developed first by collecting the lead compounds with known biological activity. A model is used to predict the activity of unknown compound using several machine learning algorithms used in the development process. QSAR model now a days are widely used to modify existing molecule to enhance its biological activity.[21,22]

In the present work, the author has used supervised machine learning approach to develop the QSAR model for prediction of inhibitory action against Jak-2 tyrosine kinase. A model is trained, tested and validated using statistical parameters and tests.

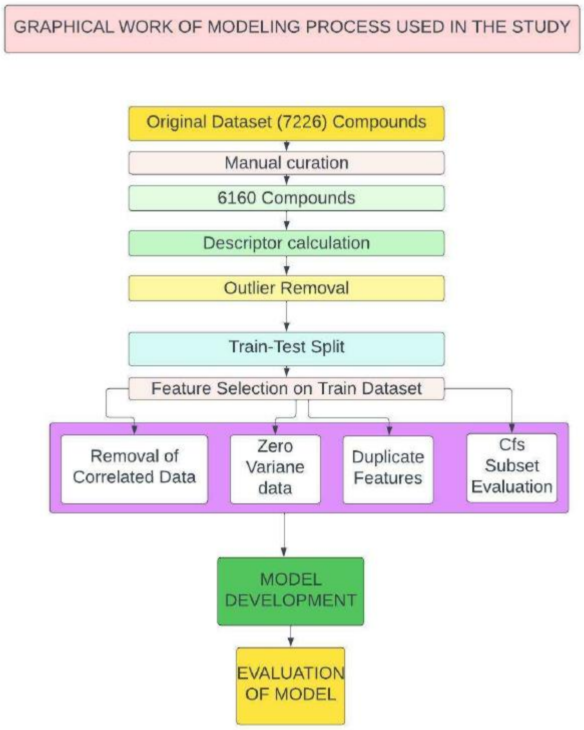


Fig. 1. Graphical Work of modelling process used in the study

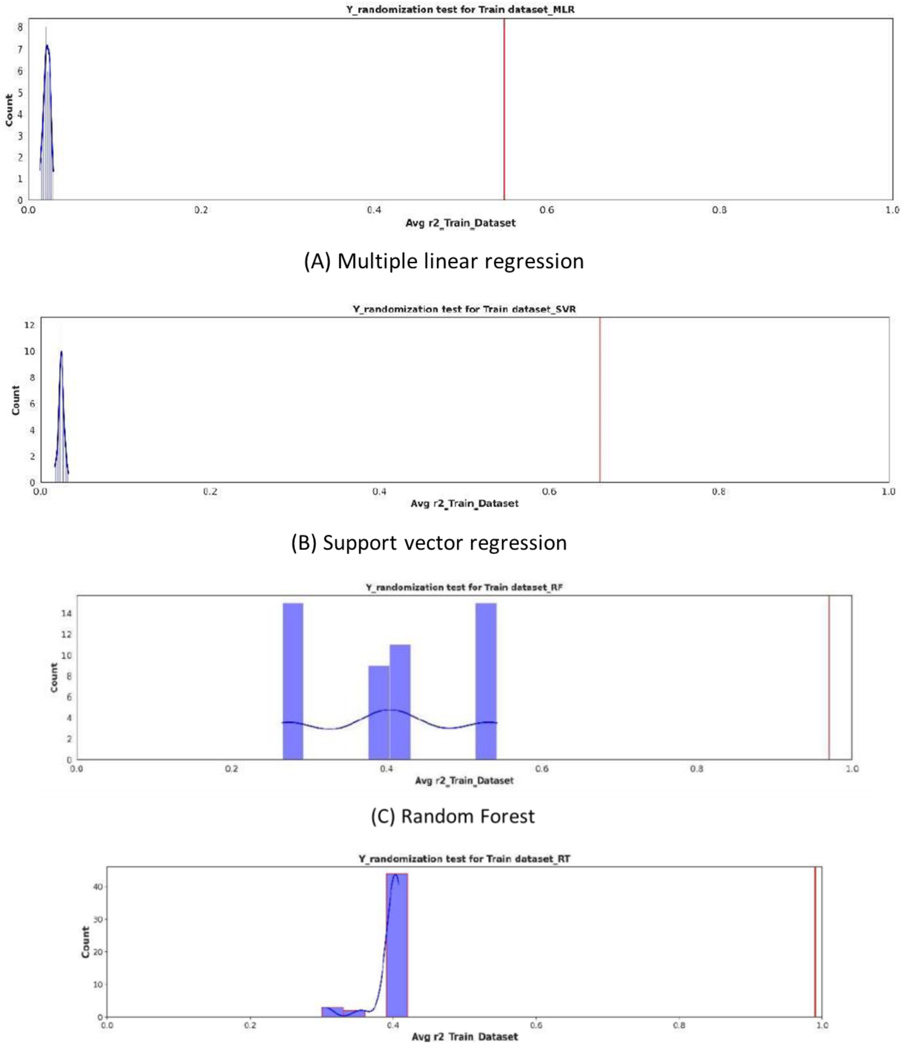
2. Materials and Methods

2.1. The data set

The ChEMBL web server was used to download dataset required to build QSAR model (Fig. 1). 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[23,24] (Supplementary_1). At first the dataset was manually curated for 'NIL', 'BLANK', 'ZERO' and invalid ChEMBL IDs. Total of 1066 compounds were removed from the downloaded database. The IC50 values of the remaining 6160 compounds was converted in to pIC50 (-log of IC50 values) and used for model development. Next, 6160 SMILES were converted in to sdf format using OpenBabel software[25] (Supplementary_2). During conversion hydrogen atom was added to make compounds explicit and to mimic the real situations.[26]

2.2. 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.[27] A total of 2323 descriptors were calculated for the given 6160 'sdf' structures. The descriptors having 'na' values were removed from the progressing database. The database with 2226 descriptors and 6160 compounds was checked for the presence of outliers (Supplementary_3). In the present work, both structural and activity outliers were considered insignificant for model development. Inter quartile range was used to calculate the



(A) Multiple linear regression

(B) Support vector regression

(C) Random Forest

(D) Random tree

Fig. 2. Y-randomization test for train data set. (A) Multiple linear regression, (B) Support vector regression, (C) Random Forest, (D) Random tree

Table 1. QSAR model with R^2.

Algorithms	Correlation Coefficient (R2)			
	MLR	SVR	RF	RT
Training set	59.40	58.84	97.1	99.32
Test Set	55.42	55.65	72.94	52.09
Cross validation	56.82	55.98	73.21	55.26

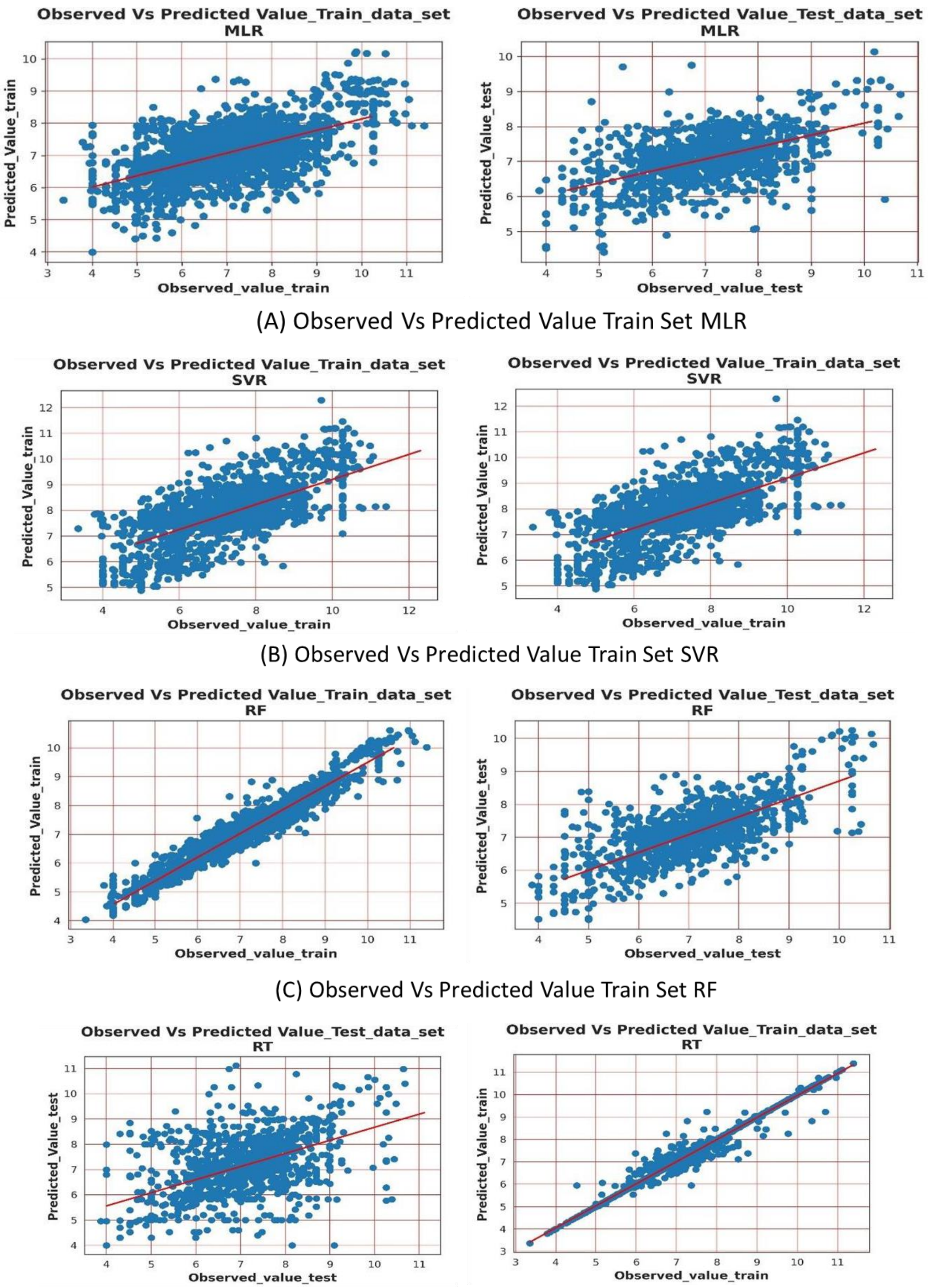
Table 2. Y-Randomization test results

Algorithm	Actual R^2	Y-randomized R^2
MLR	0.55	0.019
SVR	0.66	0.26
RF	0.97	0.4
RT	0.99	0.4

outliers and total of 2310 outliers were removed from the database (Supplementary_4). Final data base of 3850 compound was then used for selection of significant descriptors based on several statistical calculations[28] (Supplementary_5).

2.3. 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 (Supplementary_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>0.9) (Supplementary_8) and descriptors with low variance (<1%) (Supplementary_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.[29] All the features removal techniques were applied using sklearn package of python programming language. Subsequently, CfsSubsetEval along with locally predictive attributes was used for significant descriptor selection. CfsSubsetEval produces subsets of the features that are highly correlated to the class/activity while having low intercorrelation. Locally predictive attribute identifies locally predictive attributes and iteratively add attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the class/activity.[30–32]



(A) Observed Vs Predicted Value Train Set MLR

(B) Observed Vs Predicted Value Train Set SVR

(C) Observed Vs Predicted Value Train Set RF

(D) Observed Vs Predicted Value Train Set RT

Fig. 3. Observed vs. Predicted values of train data set. (A) Multiple linear regression, (B) Support vector regression, (C) Random Forest, (D) Random tree.

Table 3. Significant descriptors with their groups

S.No	Descriptors	Description	Group
1.	nB	number of Boron atoms	Functional Group Counts
2.	nBridgeHead	number of bridgehead atoms	Functional Group Counts
3.	nCconj	number of non-aromatic conjugated C(sp2)	Functional Group Counts
4.	nR=Cp	number of terminal primary C(sp2)	Functional Group Counts
5.	nArCOOH	number of carboxylic acids (aromatic)	Functional Group Counts
6.	nRCONHR	number of secondary amides (aliphatic)	Functional Group Counts
7.	nArCONHR	number of secondary amides (aromatic)	Functional Group Counts
8.	nRNH2	number of primary amines (aliphatic)	Functional Group Counts
9.	nArNH2	number of primary amines (aromatic)	Functional Group Counts
10.	nN-N	number of N hydrazines	Functional Group Counts
11.	nArCN	number of nitriles (aromatic)	Functional Group Counts
12.	nRNO2	number of nitro groups (aliphatic)	Functional Group Counts
13.	nN(CO)2	number of imides (-thio)	Functional Group Counts
14.	nSO3	number of sulfonates (thio-/dithio-)	Functional Group Counts
15.	nP(=O)R3/nPR5	number of phosphoranes (thio-)	Functional Group Counts
16.	nR=CRX	number of R=CRX	Functional Group Counts
17.	nCHRX2	nCHRX2	Functional Group Counts
18.	nOxiranes	number of Oxiranes	Functional Group Counts
19.	nOxetanes	number of Oxetanes	Functional Group Counts
20.	nOxolanes	number of Oxolanes	Functional Group Counts
21.	nPyrroles	number of Pyrroles	Functional Group Counts
22.	nImidazoles	number of Imidazoles	Functional Group Counts
23.	nPyridines	number of Pyridines	Functional Group Counts
24.	nBridgeHead	number of bridgehead atoms	Functional Group Counts
25.	nCconj	number of non-aromatic conjugated C(sp2)	Functional Group Counts
26.	nR=Cp	number of terminal primary C(sp2)	Functional Group Counts
27.	nArCOOH	number of carboxylic acids (aromatic)	Functional Group Counts
28.	nRCONHR	number of secondary amides (aliphatic)	Functional Group Counts
29.	nArCONHR	number of secondary amides (aromatic)	Functional Group Counts
30.	nRNH2	number of primary amines (aliphatic)	Functional Group Counts
31.	nArNH2	number of primary amines (aromatic)	Functional Group Counts
32.	nN-N	number of N hydrazines	Functional Group Counts
33.	nArCN	number of nitriles (aromatic)	Functional Group Counts
34.	nRNO2	number of nitro groups (aliphatic)	Functional Group Counts
35.	nN(CO)2	number of imides (-thio)	Functional Group Counts
36.	nSO3	number of sulfonates (thio-/dithio-)	Functional Group Counts
37.	nP(=O)R3/nPR5	number of phosphoranes (thio-)	Functional Group Counts
38.	nR=CRX	number of R=CRX	Functional Group Counts
39.	B06[O-Cl]	Presence/absence of O – Cl at topological distance 6	2D Atom Pairs
40.	B07[C-N]	Presence/absence of C – N at topological distance 7	2D Atom Pairs
41.	B07[N-N]	Presence/absence of N – N at topological distance 7	2D Atom Pairs
42.	B07[F-Cl]	Presence/absence of F – Cl at topological distance 7	2D Atom Pairs
43.	B08[S-S]	Presence/absence of S – S at topological distance 8	2D Atom Pairs
44.	B08[S-Cl]	Presence/absence of S – Cl at topological distance 8	2D Atom Pairs
45.	B08[Cl-Cl]	Presence/absence of Cl – Cl at topological distance 8	2D Atom Pairs
46.	B09[C-P]	Presence/absence of C – P at topological distance 9	2D Atom Pairs
47.	B09[S-Cl]	Presence/absence of S – Cl at topological distance 9	2D Atom Pairs
48.	B10[C-S]	Presence/absence of C – S at topological distance 10	2D Atom Pairs
49.	B10[N-Br]	Presence/absence of N – Br at topological distance 10	2D Atom Pairs
50.	B10[O-S]	Presence/absence of O – S at topological distance 10	2D Atom Pairs
51.	B10[O-Cl]	Presence/absence of O – Cl at topological distance 10	2D Atom Pairs
52.	B10[S-F]	Presence/absence of S – F at topological distance 10	2D Atom Pairs
53.	B10[Cl-Cl]	Presence/absence of Cl – Cl at topological distance 10	2D Atom Pairs
54.	F06[S-F]	Frequency of S – F at topological distance 6	2D Atom Pairs
55.	F06[F-Cl]	Frequency of F – Cl at topological distance 6	2D Atom Pairs
56.	F07[N-N]	Frequency of N – N at topological distance 7	2D Atom Pairs
57.	F10[C-S]	Frequency of C – S at topological distance 10	2D Atom Pairs
58.	SAscore	Synthetic Accessibility score	Molecular properties

2.4. Regression algorithms

The following algorithms support machine, multiple linear regression, random forest regressor and random tree regressor were used to build QSAR model (Fig. 2). All the algorithms were implemented using Python () using Scikit-learn package.

2.5. Multiple linear regressions

It is also called as multiple regressions 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.[33,34]

2.6. Support vector regression

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.[35]

2.7. Random Forest

Random forest is also supervised machine learning algorithm uses ensemble learning method for regression. Ensemble method suggests multiple models/trees trained over same data and averaging the results of each tree. For regression task in random forest, mean prediction of individual tree was used.[36]

2.8. Random tree regression

Another supervised machine learning tree-based algorithm called random tree for examining the prediction capability of the model was used. Recursive portioning to split the data and finding the best split using reduction in impurity index was used to evaluate the model performance.[37,38]

2.9. Evaluation of QSAR Model

For any machine learning model, statistical evaluation is most important criterion to check the robustness and stability of the prediction model. In the present work, internal and external both validation methods used to check the quality of the model. Several statistical values like correlation coefficient and cross validated R^2 values were calculated. The model was also evaluated on external dataset/test dataset to check the accuracy of model on data unknown to it. The same metrics were used to assess the model. Additionally, to check whether the model is produced by chance or not, Y-randomisation test was also performed. In this test, all the activity values were shuffled and for these values QSAR model was model was developed.[39,40] These shuffling and development were done for 50 times to build 50 QSAR model. Mean R^2 was calculated and was compared with actual R^2.[41, 42]

3. Results and Discussions

In the current research work, QSAR model has been developed using four supervised machine learning algorithms including MLR (Multiple Linear regression), SVR (Support vector regression), RF (Random Forest) and RT (Random tree). The QSAR model has been developed to predict inhibitors for Tyrosine kinase enzyme responsible for several types of cancer and other immunopathological conditions. Total of 2323 types of descriptors belongs to different classes were calculated for inhibitors known with biological activity. After applying several feature selection and removal techniques the QSAR model has been developed with 58 significant descriptors (Table 3) having structural relationship in relation to biological activity. The scatter plot (Fig. 3) drawn between observed and predicted values for all four QSAR models, clearly indicates the closeness between them. From the Table 1 it is clear that the R^2 score obtained for all four-machine learning algorithm is showing strength in prediction but more confidence is observed in the random forest and random tree algorithms. The trained model was also evaluated on the testing set, which was kept hidden to training set during model development. In all models except random tree, test R^2 was also showing good strength. It is also possible that training dataset every time we split will have different compounds within and can produce different R^2 every time. To overcome this, the training dataset was cross validated using cross validation for 5 times and average R^2 was calculated. In all four models, cross validated R^2 is less because train dataset was shuffled five times and average was calculated. In the present investigation, mathematical equations were developed for MLR and SVR model. These equations (Equation 1 & Equation 2) were clearly showing the contributions made by each significant descriptor towards the final biological activity of the compounds. From both the equation it was very much clear that functional group counts and 2D atom pairs were the group of descriptors playing great role in describing biological activity of the compounds. Some of the descriptors were also showing negative contribution to the activity which can be modified to enhance the final biological activity.

Equation 1. Multiple linear regression prediction equation

pIC50 = -2.19 * nB + -0.6238 * nBridgeHead + 0.0528 * nCconj + -0.8149 * nR=Cp + -0.4847 * nArCOOH + -0.5916 * nRCONHR + -0.3679 * nArCONHR + -0.4427 * nRNH2 + 0.1915 * nArCN + -1.5682 * nN(CO)2 + 0.5348 * nOxetanes + 0.2972 * nPyrroles + -0.37 * nImidazoles + 0.2366 * nPyridines + 0.4248 * nPyrazines + 0.7299 * n124-Triazines + -0.1956 * SdsN + 0.3236 * MaxssCH2 + -0.1238 * T(P..F) + -1.7037 * B02[S-S] + 0.5305 * B04[F-Cl] + -0.3369 * B05[C-O] + -0.3642 * B05[F-F] + -0.3304 * B06[O-Cl] + 0.8899 * B07[C-N] + 0.2789 * B07[N-N] + 3.1084 * B07[F-Cl] + 1.1434 * B08[S-S] + 1.2059 * B08[S-Cl] + 1.1645 * B08[Cl-Cl] + -2.2158 * B09[S-Cl] + 0.6183 * B10[C-S] + -0.4178 * B10[O-S] + -1.2293 * B10[S-F] + -0.542 * F06[S-F] + 0.843 * F06[F-Cl] + 0.1523 * F07[N-N] + 0.3151 * SAscore + 4.3003

Equation 2. Support Vector regression equation for prediction

weights (not support vectors): - 0.2802 * (normalized) nB-0.1501 * (normalized) nBridgeHead + 0.0088 * (normalized) nCconj - 0.1888 * (normalized) nR=Cp - 0.0502 * (normalized) nArCOOH - 0.1791 * (normalized) nRCONHR -0.1387 * (normalized) nArCONHR - 0.079 * (normalized) nRNH2 - 0.0106 * (normalized) nArNH2 - 0.1145 * (normalized) nN-N + 0.0406 * (normalized) nArCN - 0.1456 * (normalized) nRNO2 - 0.1683 * (normalized) nN(CO)2 - 0.165 * (normalized) nSO3 + 0.0875 * (normalized) nP(=O)R3/nPR5 - 0.0039 * (normalized) nR=CRX + 0.0851 * (normalized) nCHRX2 - 0.1067 * (normalized) nOxiranes + 0.0808 * (normalized) nOxetanes + 0.0369 * (normalized) nOxolanes + 0.0862 * (normalized) nPyrroles - 0.051 * (normalized) nImidazoles + 0.0504 * (normalized) nPyridines + 0.0724 * (normalized) nPyrazines + 0.0932 * (normalized) n124-Triazines - 0.1983 * (normalized) SdsN + 0.1001 * (normalized) SssssSi + 0.0532 * (normalized) MaxssCH2 - 0.326 * (normalized) T(P..F) + 0.0227 * (normalized) B01[C-N] + 0.0242 * (normalized) B02[N-Cl] - 0.1687 * (normalized) B02[S-S] - 0.001 * (normalized) B04[O-Br] + 0.0588 * (normalized) B04[F-Cl] + 0.0346 * (normalized) B04[Br-Br] + 0.0445 * (normalized) B05[C-N] - 0.0127 * (normalized) B05[C-O] + 0.0016 * (normalized) B05[F-F] - 0.0419 * (normalized) B06[O-Cl] + 0.0356 * (normalized) B07[C-N] + 0.0428 * (normalized) B07[N-N] + 0.3286 * (normalized) B07[F-Cl] + 0.0917 * (normalized)

B08[S-S] + 0.0583 * (normalized) B08[S-Cl] + 0.1164 * (normalized) B08[Cl-Cl] + 0.1039 * (normalized) B09[C-P] - 0.2306 * (normalized) B09[S-Cl] + 0.0457 * (normalized) B10[C-S] + 0.0526 * (normalized) B10[N-Br] - 0.0599 * (normalized) B10[O-S] + 0.0476 * (normalized) B10[O-Cl] - 0.0486 * (normalized) B10[S-F] - 0.0767 * (normalized) B10[Cl-Cl] - 0.2211 * (normalized) F06[S-F] + 0.3497 * (normalized) F06[F-Cl] + 0.1379 * (normalized) F07[N-N] + 0.0909 * (normalized) F10[C-S] + 0.0791 * (normalized) SAscore + 0.1205

The model's performance was found to be considerably worse than the original model. The values of R^2 calculated for all four machine learning algorithms. Total 50 randomized models were developed and their average R^2 was considered (Table 2). These results are clearly indicating the models are not obtained by chance. There is fundamental relationship exist between the structural features calculated and final biological activity.

4. Conclusions

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 of time. 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 is 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 in-vitro 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.

Conflicts of Interest

The authors declare no conflict of interest.

References

1 Manning G.; Whyte D.B.; Martinez R.; Hunter T.; Sudarsanam S. The Protein Kinase Complement of the Human Genome. *Science*, 2002, **298**, 1912-1934. [CrossRef]

2 Scheeff E.D.; Bourne P.E. Structural Evolution of the Protein Kinase–like Superfamily. *PLoS Comput. Biol.*, 2005, **1**, e49. [CrossRef]

3 KREBS E.G. The Phosphorylation of Proteins: A Major Mechanism for Biological Regulation. 1985. [CrossRef]

4 Harpur A.G.; Andres A.C.; Ziemiecki A.; Aston R.R.; Wilks A.F. JAK2, A Third Member of the JAK Family of Protein Tyrosine Kinases. *Oncogene*, 1992, **7**, 1347-1353. [Link]

5 Sakatsume M.; Igarashi K.I.; Winestock K.D.; Garotta G.; Larner A.C.; Finbloom D.S. The Jak Kinases Differentially Associate with the A and B (Accessory Factor) Chains of the Interferon Γ Receptor to Form A Functional Receptor Unit Capable of Activating STAT Transcription Factors. *J. Biol. Chem.*, 1995, **270**, 17528-17534. [CrossRef]

6 Saltzman A.; Stone M.; Franks C.; Searfoss G.; Munro R.; Jaye M.; Ivashchenko Y. Cloning and Characterization of Human Jak-2 Kinase: High mRNA Expression in Immune Cells and Muscle Tissue. *Biochem. Biophys. Res. Commun.*, 1998, **246**, 627-633. [CrossRef]

7 Berry D.C.; Jin H.; Majumdar A.; Noy N. Signaling by Vitamin A and Retinol-Binding Protein Regulates Gene Expression to Inhibit Insulin Responses. *Proc. Natl. Acad. Sci.*, 2011, **108**, 4340-4345. [CrossRef]

8 Guilluy C.; Brégeon J.; Toumaniantz G.; Rolli-Derkinderen M.; Retailleau K.; Loufrani L.; Henrion D.; Scalbert E.; Bril A.; Torres R.M.; Offermanns S. The Rho Exchange Factor Arhgef1 mediates the Effects of Angiotensin II on Vascular Tone and Blood Pressure. *Nat. Med.*, 2010, **16**, 183-190. [CrossRef]

9 Jäkel H.; Weinl C.; Hengst L. Phosphorylation of p27Kip1 by JAK2 Directly Links Cytokine Receptor Signaling to Cell Cycle Control. *Oncogene*, 2011, **30**, 3502-3512. [CrossRef]

10 Bertram J.S. The Molecular Biology of Cancer. *Mol. Aspects Med.*, 2000, **21**, 167-223. [CrossRef]

11 Zwick E.; Bange J.; Ullrich A. Receptor Tyrosine Kinases as Targets for Anticancer Drugs. *Trends Mol. Med.*, 2002, **8**, 17-23. [CrossRef]

12 Nishikawa R.; Ji X.D.; Harmon R.C.; Lazar C.S.; Gill G.N.; Cavenee W.K.; Huang H.J. A Mutant Epidermal Growth Factor Receptor Common in Human Glioma Confers Enhanced Tumorigenicity. *Proc. Natl. Acad. Sci.*, 1994, **91**, 7727-7731. [CrossRef]

13 Paul M.K.; Mukhopadhyay A.K. Tyrosine Kinase–Role and Significance in Cancer. *Int. J. Med. Sci.*, 2004, **1**, 101. [CrossRef]

14 Schwartz D.M.; Kanno Y.; Villarino A.; Ward M.; Gadina M.; O'Shea J.J. JAK Inhibition as a Therapeutic Strategy for Immune and Inflammatory Diseases. *Nat. Rev. Drug Discov.*, 2017, **16**, 843-862. [CrossRef]

15 Dickson M.; Gagnon J.P. Key Factors in the Rising Cost of New Drug Discovery and Development. *Nat. Rev. Drug Discov.*, 2004, **3**, 417-429. [CrossRef]

16 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. New Perspectives on how to Discover Drugs from Herbal Medicines: CAM's Outstanding Contribution to Modern Therapeutics. *Evid.-Based Complementary and Alternative Medicine*, 2013, **2013**. [CrossRef]

17 Szymański P.; Markowicz M.; Mikiciuk-Olasik E. Adaptation of High-throughput Screening in Drug Discovery—Toxicological Screening Tests. *Int. J. Mol. Sci.*, 2011, **13**, 427-452. [CrossRef]

18 Clark R.L.; Johnston B.F.; Mackay S.P.; Breslin C.J.; Robertson M.N.; Harvey A.L. The Drug Discovery Portal: A Resource to Enhance Drug Discovery from Academia. *Drug Discov. Today*, 2010, **15**, 679-683. [CrossRef]

19 Dong X.; Zheng W. A New structure-based QSAR Method Affords both Descriptive and Predictive Models for Phosphodiesterase-4 Inhibitors. *Curr. Chem. Genom.*, 2008, **2**, 29. [CrossRef]

20 Karelson M.; Lobanov V.S.; Katritzky A.R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.*, 1996, **96**, 1027-1044. [CrossRef]

21 Yang S.Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discov. Today*, 2010, **15**, 444-450. [CrossRef]

22 Acharya C.; Coop A.; E Polli J.; D MacKerell A. Recent Advances in Ligand-based Drug Design: Relevance and Utility of the

Conformationally Sampled Pharmacophore Approach. *Curr. Comput.-Aided Drug Des.*, 2011, **7**, 10-22. [CrossRef]

23 Gaulton A.; Hersey A.; Nowotka M.; Bento A.P.; Chambers J.; Mendez D.; Mutowo P.; Atkinson F.; Bellis L.J.; Cibrián-Uhalte E.; Davies M. The ChEMBL Database in 2017. *Nucleic Acids Res.*, 2017, **45**(D1), D945-D954. [CrossRef]

24 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. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.*, 2019, **47**(D1), D930-D940. [CrossRef]

25 O'Boyle N.M.; Morley C.; Hutchison G.R. Pybel: a Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chem. Cent. J.*, 2008, **2**, 1-7. [CrossRef]

26 Kumar V.; Roy K. 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 QSAR Environ. Res.*, 2020, **31**, 511-526. [CrossRef]

27 Mauri A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In: Roy, K. (eds) Ecotoxicological QSARs. Methods in Pharmacology and Toxicology, 2020. Humana, New York, NY. [CrossRef]

28 Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.*, 2010, **29**, 476-488. [CrossRef]

29 Pedregosa F.; Varoquaux G.; Gramfort A.; Michel V.; Thirion B.; Grisel O.; Blondel M.; Prettenhofer P.; Weiss R.; Dubourg V.; Vanderplas J. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830. [Link]

30 Kotthoff L.; Thornton C.; Hoos H.H.; Hutter F.; Leyton-Brown K. Auto-WEKA: Automatic Model Selection and Hyperparameter Optimization in WEKA. In *Automated Machine Learning,* 2019, 81-95. Springer, Cham. [CrossRef]

31 Hall M.; Frank E.; Holmes G.; Pfahringer B.; Reutemann P.; Witten I.H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 2009, **11**, 10-18. [CrossRef]

32 Li J.; Cheng K.; Wang S.; Morstatter F.; Trevino R.P.; Tang J.; Liu H. Feature Selection: A Data Perspective. *ACM Comput. Surv. (CSUR)*, 2017, **50**, 1-45. [CrossRef]

33 Pandis N. Multiple Linear Regression Analysis. *Am. J. Orthod. Dentofac. Orthop.*, 2016, **149**, 581. [CrossRef]

34 Uyanık G.K.; Güler N. A Study on Multiple Linear Regression Analysis. *Procedia-Soc. Behav. Sci.*, 2013, **106**, 234-240. [CrossRef]

35 Smola A.J.; Schölkopf B. A Tutorial on Support Vector Regression. *Stat. Comput.*, 2004, **14**, 199-222. [CrossRef]

36 Breiman L. Random Forests. *Mach. Learn.*, 2001, **45**, 5-32. [CrossRef]

37 Holmes G.; Hall M.; Prank E. Generating Rule Sets from Model Trees. In *Australasian Joint Conference on Artificial Intelligence,* December 1999, 1-12. Springer, Berlin, Heidelberg. [CrossRef]

38 Wang Y.; Witten I.H. Induction of Model Trees for Predicting Continuous Classes. 1996. [Link]

39 Rücker C.; Rücker G.; Meringer M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.*, 2007, **47**, 2345-2357. [CrossRef]

40 Rücker C.; Rücker G.; Meringer M. Y-Randomization–A Useful Tool in QSAR Validation, or Folklore. 2007, **47**. [Link]

41 Adler J.; Parmryd I. Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient is Superior to the Mander's Overlap Coefficient. *Cytometry Part A*, 2010, **77**, 733-742. [CrossRef]

42 Nakagawa S.; Johnson P.C.; Schielzeth H. The Coefficient of Determination R^2 and Intra-Class Correlation Coefficient From Generalized Linear Mixed-Effects Models Revisited and Expanded. *J. R. Soc. Interface*, 2017, **14**, 20170213. [CrossRef]