ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

Research Article

open access

Maiden Drug Designing Pipeline in R, using Open-source Tools, R Packages, and Machine Learning QSAR Models to Counter COVID

Shradheya R.R. Gupta[1], Subham Verma[2], Rakesh Sharma[3], Sumita Kachhwaha[4*]

[1,2,3]DBT-Bioinformatics Infrastructure Facility, University of Rajasthan, India
[2]Center of Converging Technology, University of Rajasthan, India
[4]Department of Botany, University of Rajasthan, India

***Address for Correspondence:** Dr. Sumita Kachhwaha, Professor, Department of Botany, University of Rajasthan, India
E-mail: sumitakachhwaha@gmail.com

Received: 26 Jun 2021/ Revised: 23 Aug 2021/ Accepted: 03 Oct 2021

ABSTRACT

Background: R language is highly cited for QSAR model generation because of its capacity to solve statistical modelling problems but gets negligible credit for its potential to work as a complete tool for cheminformatics. The vision of the work is to boost the confidence in R for developing a structure-based computer-aided drug-designing pipeline using the available packages and open-source tools and give researchers a tool to counter COVID.

Methods: The amalgamation of public tools with R shows us its adaptability and flexibility. To show its working JAK 1 protein is selected to find the suitable inhibitor from the dataset of 10 million compounds. For fast and effective filtering QSAR modelling is done using the Random Forest algorithm.

Results: Employing Python and Linux scripts within R Studio for QSAR in the present study, three models were generated: the Kinase inhibitor model, the Amine mutagen model and the BBB model. Using these models, significant 25 structural features of distinct chemicals were identified and each of the three models surpassed the required accuracy.

Conclusion: Since the FDA has not given a green signal to any of the drugs so far which can effectively be used to cure COVID-19 infections so it is need of the hour that we identify some potent molecules which can be employed to inhibit coronaviruses. Our findings can be instrumental for screening new chemicals for COVID essential regulatory purposes.

Key-words: COVID, Drug designing, JAK inhibitors, QSAR, R-language

INTRODUCTION

Over 3 million pieces of literature prove that the process of drug development is a very costly and time-consuming business [1]. A successful drug takes more than 15 years and 500 million US dollars before its large-scale production. The condition was much more disappointing, tedious, costly and high on failures before the early nineties due to the hit and trial approach. However, the development of HTML in 1990 and microprocessor in 1993 became the cornerstone for getting data and a high processing system to analyze that data respectively.

How to cite this article
Gupta SRR, Verma S, Sharma R, Kachhwaha S. Maiden Drug Designing Pipeline in R, using Open-source Tools, R Packages, and Machine Learning QSAR Models to Counter COVID. SSR Inst. Int. J. Life Sci., 2021; 7(6S): 95S-104S.

Access this article online
https://iijls.com/

Additionally, parallel processing in 1995 and 64-bit processor in 2003 changed the way we see computer intelligence. Now, exascale computing speeding every bit of a process and merges all the known fields. Imprinting of these advancements is visible in the drug development process also. The process took advantage of this growing hybrid science and unfolds a new field, Computer-Aided Drug Designing (CADD) [2-5]. Today it successfully overcame the disadvantages and provided significant insights on drug development. Contrary, the use of statistical tools like SAS and R gain popularity in preclinical and clinical data analysis. R dominates the other tools because of its lexical scoping, openness, active community support, binding to ancillary software, customization properties, developing interface, and availability of more than 50 thousand packages [6-8]. CADD has reduced the cost of drug screening to half but commercial software makes chemoinformatics and drug

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

discovery a costly affair. In contrast to it, open-source tools believe in free access to data, large peer collaboration, reduction in duplicate research, and supporting developing countries by saying no to the patents. Hereto, there are examples where these tools have gained commercial success, like the Firefox browser and Linux operating system [9-11].

R language [12] is highly cited for QSAR model generation because of its statistical modelling problem but gets negligible credit for its potential to work as a complete tool for chemoinformatics. QSAR is a ligand-based drug designing process [5]. The process converts the chemical structure information into the mathematical model to find out the statistically significant correlation between the lead and biological properties, using regression and classification algorithms (machine learning). The process is highly adopted due to its fast and diverse screening, identification, and optimization of lead. It is highly used for identifying BBB, ADME, toxic and mutagenic like properties in a lead. The efficacy of the QSAR model depends on the selection of uncorrelated descriptors, quality of training dataset and regularization (avoiding overfitting) process [13-15]. Furthermore, machine learning has emerged as a solid pillar in decreasing the rate of drug failures by ninety per cent. This happens due to the availability of "big data," it is estimated that 2.5 quintillion bytes are created every single day. According to SAS, "the learning is based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention." From various machine-learning algorithms, Random Forest is considered a gold standard in drug discovery due to its high level of predictability, robustness, and simplicity [16-18]. The vision of the work is to boost the confidence of researchers in R by developing a drug-designing pipeline based on R using the available packages and open-source tools. The amalgamation of public tools with R shows us its adaptability and flexibility. The pipeline is validated by finding the kinase inhibitors from the dataset of 10 million chemical compounds.

MATERIALS AND METHODS

The study was conducted using a computational facility established by the DBT-Bioinformatics Infrastructure Facility, University of Rajasthan, Jaipur, India from September 2019 onwards till March 2020.

For the development of pipeline (Fig. 1), the study referred to Talevi [4] and Yu *et al.* [5] work. The first study provided an overview of CADD whereas the second work provided us with a detailed methodology. In the study, JAK 1 is selected for the testing of the pipeline. It is a well-established fact that the kinases are one of the well-studied enzymes that transfer the high-energy phosphate group from the ATP to the substrate. As it is involved in various pathways, we were able to find out the research papers and lists of available inhibitors that can affect the activity of the protein [19,20]. Additionally, there is much evidence of activation/involvement of JAK in COVID [20-22]. As the availability of a decent amount of significant data in a useful format is crucial for digesting the data and learning. We downloaded the nucleotide sequence in FASTA format (Table 1). The sequence was then read and converted to the amino acid sequence by the package Seqinr (Table 2). The amino acid sequence is BLAST against the PDB database [23,24]. The best hit was selected and downloaded in the PDB format. The PDB was sent to bio3d package for Ramachandran validation and then converted to pdbqt format using Open Babel package [25] (Table 3).

Table 1: File formats used in the pipeline

S.No.	File Format	Explanation
1	FASTA	It represents the nucleotide and protein sequence in a single letter code. It is simple, easy to manipulate and universal accepted.
2	BLAST	It is an alignment search algorithm to find the similarity between the sequences. It uses scoring matrix, positive score for every match, and negative score for every mismatch. Based on the total score best hit is arranged.
3	PDB	It is a standard representation for macromolecular structure data, derived from X-ray diffraction and NMR studies.
4	PDBQT	It stores the atomic coordinates, partial charges, and AutoDock atom types, for

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

		both the receptor and the ligand.
5	URI	It stores the address of the data from where it will be fetched
6	SMILES	It uses line notation for describing the structure of chemical species using short ASCII strings.
7	CSV	Comma-separated file format stores the data in tabular form.

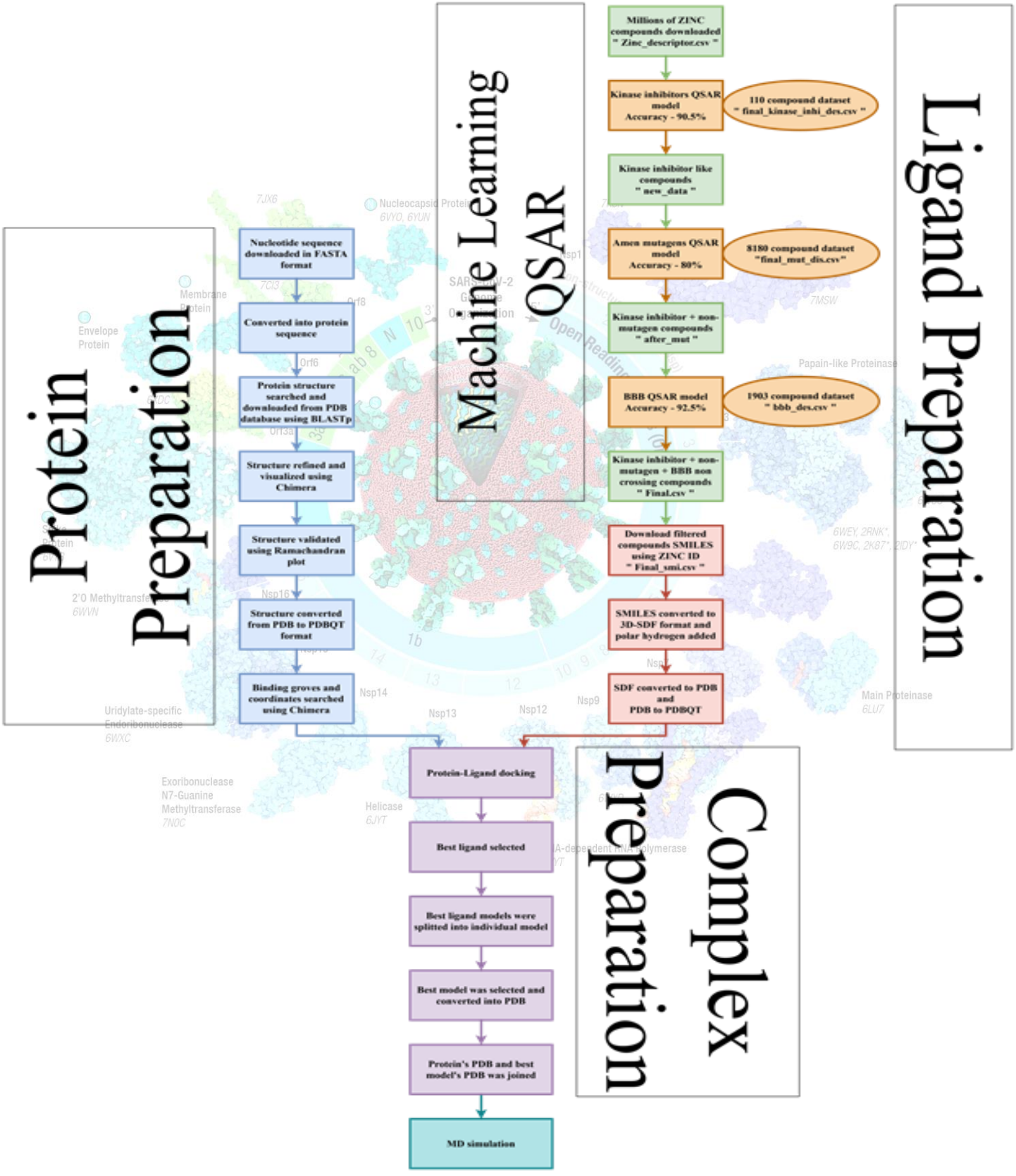

Fig. 1: A drug-designing pipeline based on R

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

Table 2: Packages used in the development of pipeline

S.No.	Packages	Use
1	BiocManager	Install and manage packages from the Bioconductor project.
2	Bio3d	For the analysis of protein structure, sequence and trajectory data.
3	Seqinr	Exploratory data analysis and data visualization for biological sequence.
4	Dplyr	Data manipulation.
5	Devtools	Provide functions that simplify common tasks.
6	MDplot	Provides automation for plot generation succeeding common molecular dynamics analyses.
7	ChemmineR	Analyze drug-like small molecule data.
8	Filesstrings	Manipulate files and strings.
9	Stringr	Data cleaning and prepare tasks with string objects.
10	Reader	Read data from files.
12	Randomforest	Implements Breiman's random forest algorithm for classification and regression.
13	Caret	Streamline the process for creating predictive models.
14	Readr	Fast and friendly way to read rectangular data.

Table 3: Open source tools used in the development of pipeline

S.No.	Tools	Use
1	OpenBabel	Search, convert, analyze and store data in different formats.
2	PaDEL-Descriptor	Calculate molecular descriptors and fingerprints.
3	Smilite-master	Python module to download and analyze SMILES strings of chemical compounds from ZINC
4	AutoDock Vina	Molecular modeling simulation, designed for protein-ligand docking.

In structure-based virtual screening, the ligand library preparation is the next step. To test the machine models, we selected diverse 3D tranches from the ZINC database [26] in SMILES format and downloaded them using the URL method (Table 2). The 'uri' file contains the address of all the tranches. The downloaded tranches were split in sequential order and stored in folders using the base commands. Further, PaDEl-Descriptor tool [27] (Table 3) is used; it is a java-based tool for generating compound descriptors using their SMILES. Descriptors or variables are the properties of a compound in a machine model that forms columns. The selection of independent variables is based on the principle of data compression, where a minimum number of variables are selected to generate sufficient information about the response variable.

We employed a random forest algorithm, part of the supervised machine-learning program for QSAR model creation [7,13,28]. All model dataset is split into training and test dataset in 80:20 ratio, model fitting is done on the training dataset. In random forest decision trees are formed, which are split into the nodes and further to the leaf nodes when the covariant reaches the given threshold. Here each leaf node represents the specific decision outcome [16,18]. To validate the model, the test dataset is supplied to it. For getting the most appropriate cut-off ROC curve is considered, as the best cut-off has the highest true positive rate with the lowest false positive rate (1-Specificity). Whereas the confusion matrix helped in knowing the accuracy of the models before the real dataset is supplied to the models.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{False Positive} + \text{False Negative}}$$

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

After filtering 10 million compounds, the sieved compounds' IDs were downloaded using the Smilite tool (Table 3). The tool picks the SMILES, searches the same SMILES in the ZINC database and grep its ID. In the backend, the tool uses the ZINC 15 database in place of the default ZINC 12 database. The downloaded IDs with their respective SMILES were then converted to pdbqt format. For docking, AutoDockVina [29], Table 3 is used in a loop with the base system command. The best-docked model is then used to generate the protein-ligand complex for molecular simulations.

RESULTS

For the study, human JAK 1 nucleotide sequence from the NCBI, I.D. NP_001307852.1 was downloaded and searched using the pipeline. JAK 1 is a human tyrosine kinase protein involved in important cell signalling by interacting with IL-2, IL-4 and gp130 receptor families. The pipeline searched out PDB ID 5L04. A similar search was done manually on the NCBI website which fetched the same result. The pipeline PDB was piped for Ramachandran plot validation. The plot showed only 3.3% outliers. The PaDEl-Descriptor tool generated 1875 variables from them only 25 uncorrelated variables were selected (Table 4).

The selection of variables is based on important structural and physicochemical properties of compounds. These variables were able to predict the maximum properties of an unknown compound that are important to be preliminarily called a lead. For QSAR three models were generated. First, the Kinase inhibitor model contains 110 compounds of which 37 are kinase inhibitors [19,20]. Second, the Amine mutagen model has 8180 compounds of which 4753 are mutagenic [30]. Third, the BBB model has 1473 compounds, from them, 1903 can cross the blood-brain barrier [31,32]. For the kinase model, we achieved Accuracy: 0.9048, Sensitivity: 1.0000, Specificity: 0.6667 (Fig. 2, Table 5). For Mutagen model Accuracy: 0.7983, Sensitivity: 0.8653, Specificity: 0.6997 (Fig. 3, Table 6). For BBB model Accuracy: 0.9253, Sensitivity: 0.7160, Specificity: 0.9830 (Fig. 4, Table 7). Finally, these models were able to filter out 12390 compounds from the given 10 million compounds. The filtered compounds were then pushed for docking with the protein. The docking process is set to run on a loop that takes the compound, docks with the protein stores the result, and generates a list of best-docked compounds based on binding energies.

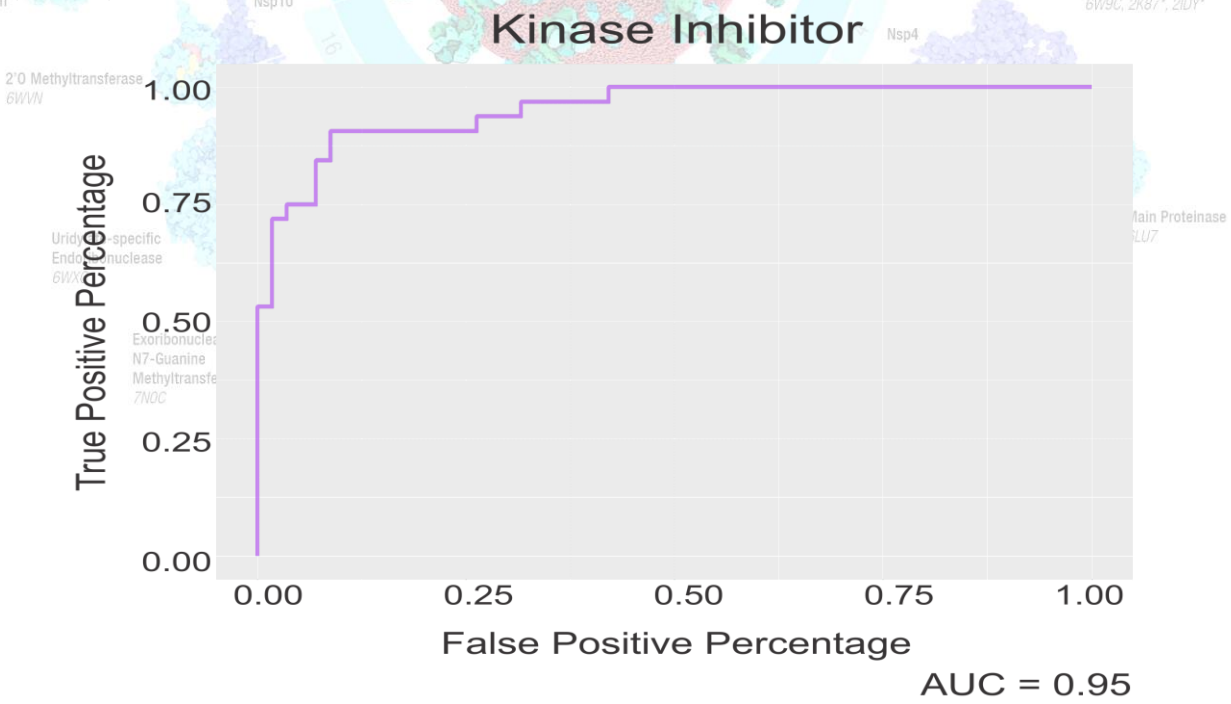

Fig. 2: ROC curve for Kinase Inhibitor model
(The area under the curve is 0.95 or 95%)

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

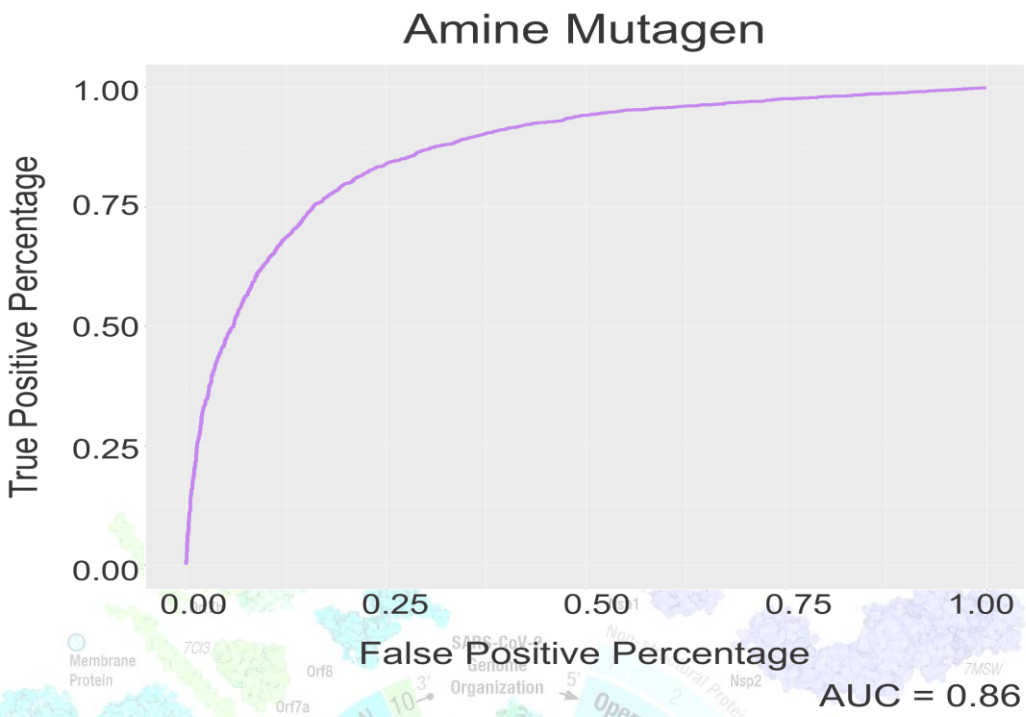

Fig. 3: ROC curve for Amen mutagen
(The area under the curve is 0.86 or 86%)

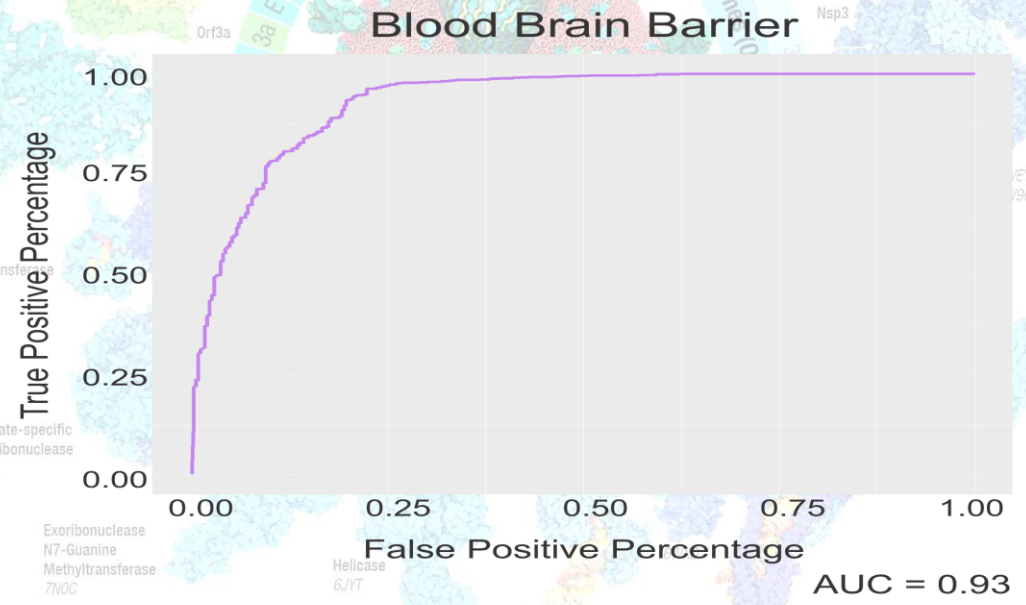

Fig. 4: ROC curve for Blood-Brain Barrier model
(The area under the curve is 0.93 or 93%)

Table 4: List of variables (features) used in the machine model development and for the real dataset

S.No.	Descriptors	Name
1	AMR	Molar refractivity (Polar Ability)
2	naAromAtom	Number of aromatic atoms
3	nAtom	Number of atoms
4	nHeavyAtom	Number of heavy atoms (i.e. not hydrogen)
5	nH	Number of hydrogen atoms
6	nB	Number of boron atoms
7	nC	Number of carbon atoms

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

8	nN	Number of nitrogen atoms
9	nO	Number of oxygen atoms
10	nS	Number of sulphur atoms
11	nP	Number of phosphorus atoms
12	nF	Number of fluorine atoms
13	nCl	Number of chlorine atoms
14	nBr	Number of bromine atoms
15	nI	Number of iodine atoms
16	nX	Number of halogen atoms (F, Cl, Br, I, At, Uus)
17	nHBAcc_Lipinski	Number of hydrogen bond acceptors (using Lipinski's definition: any nitrogen; any oxygen)
18	nHBDon_Lipinski	Number of hydrogen bond donors (using Lipinski's definition: Any OH or NH. Each available hydrogen atom is counted as one hydrogen bond donor)
19	nRing	Number of rings
20	nHeteroRing	Number of rings containing heteroatoms (N, O, P, S, or halogens)
21	nRotB	Number of rotatable bonds, excluding terminal bonds
22	LipinskiFailures	Number failures of Lipinski's Rule Of 5
23	TopoPSA	Topological polar surface area
24	MW	Molecular weight
25	XLogP	XLogP

Table 5: Kinase Inhibitor model Confusion Matrix and Statistics

Confusion Matrix		
	Reference	
Prediction	No	Yes
No	15	2
Yes	0	4
	Statistics	
Accuracy	0.9048	
95% CI	(0.6962, 0.9883)	
No Information Rate	0.7143	
p-value (Acc>NIR)	0.03671	

Table 6: Amine mutation model Confusion Matrix and Statistics

Confusion Matrix		
	Reference	
Prediction	Mutagen	Nonmutagen
Mutagen	848	200
Nonmutagen	132	466
	Statistics	
Accuracy	0.7983	
95% CI	(0.7781, 0.8174)	
No Information Rate	0.5954	
p-value (Acc>NIR)	< 2.2e-16	

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

Table 7: Blood Brain Barrier model Confusion Matrix and Statistics

Confusion Matrix		
	Reference	
Prediction	No	Yes
No	58	5
Yes	23	289
	Statistics	
Accuracy	0.9253	
95% CI	(0.8939, 0.9498)	
No Information Rate	0.786	
p-value (Acc>NIR)	9.539e-14	

DISCUSSION

There is no doubt that R is highly recommended for academia and researchers dealing with data, statistical calculations, and machine models. Furthermore, it believes in the vision where the user can transform into the programmer. With it, it provides better Object-Oriented Programming and GUI (R Studio) compared to other languages. Due to these factors, the language is gaining popularity among researchers, corporate and even drug designers. On the other hand, its effectiveness in filtering tremendous chemo-informatics data, smooth visualization, research reproducibility, and the capacity to customize systems and work is not utilized satisfactorily in drug designing. Python, which has a large community and library support, is another popular language outside R. In the current work, we employed Python and Linux scripts within R Studio, demonstrating that we can use multiple languages depending on our usability and that other languages may be used in R Studio. Mente and Kuhn [8] and Talevi [4] completely focused on the use of R in QSAR modelling. On the other hand, the current study moved a step forward and a step backwards from QSAR to provide the whole picture so the researchers can get a comprehensive understanding of CADD in R. Many studies have been conducted in the last few months which emphasize the importance of machine learning models in pin-pointing appropriate proteins which can be selectively employed as potential drug candidates to combat COVID-19 epidemic [33,34] but unlike the present study, different machine learning tools have been used. Wet lab studies using three tyrosine kinase inhibitors was initiated to evaluate their efficacy against viruses and one of them was found to be very potent, giving leads that clinical trials can further confirm their candidature as drugs against combat SARS-CoV-2 [35]. In the present study, the Kinase inhibitor model displayed high accuracy, along with two other models. Already, Baricitinib, which was initially used for rheumatoid arthritis and is an oral Janus kinase (JAK)1/JAK2 inhibitor has been recommended to be utilized for COVID-19 patients [36], this is a good example, which emphasizes that using machine learning tools we can restrict SARS-CoV-2. Hence, the work focuses on showing its flexibility with open-source tools and examples of packages that are designed for comprehensive computer-aided drug-designing pipelines.

CONCLUSIONS

We have successfully generated the models that have higher accuracy than the other models using the same data by improving the variable selection pattern. These results and pipelines will help in finding new treatments for COVID. Admittedly, coming plans are to reduce the number of packages to make it more compact and to introduce GUI so that the general audience can use it.

Subsequently, improvement in data preparation and variable selection will improve accuracy, sensitivity, and specificity. Summing up, from the results we are assured that the R will find its place in the CADD workflow in the coming years.

ACKNOWLEDGMENTS

Computational facility provided by DBT- Bioinformatics Infrastructure Facility at the University of Rajasthan is gratefully acknowledged.

CONTRIBUTION OF AUTHORS

Research concept- Shradheya R.R. Gupta
Research design- Shradheya R.R. Gupta, Subham Verma
Supervision- Sumita Kachhwaha

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al*., 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

Materials- Shradheya R.R. Gupta, Subham Verma, Rakesh Sharma
Data collection- Shradheya R.R. Gupta, Subham Verma
Data analysis and Interpretation- Shradheya R.R. Gupta, Subham Verma
Literature search- Shradheya R.R. Gupta, Subham Verma
Writing article- Shradheya R.R. Gupta
Critical review- Sumita Kachhwaha
Article editing- Shradheya R.R. Gupta, Sumita Kachhwaha
Final approval- Sumita Kachhwaha

REFERENCES

[1] Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA, 2020; 323: 844-53.
[2] Macalino SJ, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. Arch Pharm Res., 2015; 38: 1686-701.
[3] Singh DB. Success, Limitation and Future of Computer-Aided Drug Designing. Transl Med., 2014; 04.
[4] Talevi A. Computer-Aided Drug Design: An Overview. Methods Mol Biol., 2018; 1762: 1-19.
[5] Yu W, MacKerell AD, Jr. Computer-Aided Drug Design Methods. Methods Mol Biol., 2017; 1520: 85-106.
[6] Acharya C, Coop A, Polli JE, Jr ADM. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore. Current Computer-Aided Drug Design, 2011; 7: 10–22.
[7] Carpenter KA, Huang X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. Curr Pharm Des., 2018; 24: 3347-58.
[8] Mente S and Kuhn M. The Use of the R Language for Medicinal Chemistry Applications. Curr Top Med Chem., 2012; 12: 1957-64.
[9] Ardal C, Rottingen JA. Open source drug discovery in practice: a case study. PLoS Negl Trop Dis., 2012; 6: e1827.
[10] Robertson MN, Ylioja PM, Williamson AE, Woelfle M, Robins M, et al. Open source drug discovery-a limited tutorial. Parasitol., 2014; 141: 148-57.
[11] Singla D, Dhanda SK, Chauhan JS, Bhardwaj A, Brahmachari SK, et al. Open Source Software and Web Services for Designing Therapeutic Molecules. Curr Top Med Chem., 2013; 13.
[12] Morandat F, Hill B, Osvald L, Vitek J. Evaluating the Design of the R Language. ECOOP, 2012; 104-31.
[13] Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, et al. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Front Pharmacol., 2018; 9: 1275.
[14] Winkler DD. The role of quantitative structure activity relationships (QSAR) in biomolecular discovery. Brief Bioinform., 2001; 3: 73.
[15] Kwon S, Bae H, Jo J, Yoon S. Comprehensive ensemble in QSAR prediction for drug discovery. BMC Bioinform., 2019; 20: 521.
[16] Stephenson N, Shane E, Chase J, Rowland J, Ries D, et al. Survey of Machine Learning Techniques in Drug Discovery. Curr Drug Metab., 2019; 20: 185-93.
[17] Talevi A, Morales JF, Hather G, Podichetty JT, Kim S, et al. Machine Learning in Drug Discovery and Development Part 1: A Primer. CPT Pharmacometrics Syst Pharmacol., 2020; 9: 129-42.
[18] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov., 2019; 18: 463-77.
[19] Bechman K, Yates M, Galloway JB. The new entries in the therapeutic armamentarium: The small molecule JAK inhibitors. Pharmacol Res., 2019; 147: 104392.
[20] Fragoulis GE, McInnes IB, Siebert S. JAK-inhibitors. New players in the field of immune-mediated diseases, beyond rheumatoid arthritis. Rheumatol., (Oxford). 2019; 58: i43-i54.
[21] Mehta P, Ciurtin C, Scully M, Levi M, Chambers RC. JAK inhibitors in COVID-19: the need for vigilance regarding increased inherent thrombotic risk. Eur Respir J., 2020; 56.
[22] Seif F, Aazami H, Khoshmirsafa M, Kamali M, Mohsenzadegan M, et al. JAK Inhibition as a New Treatment Strategy for Patients with COVID-19. Int Arch Allergy Immunol., 2020; 181: 467-75.
[23] McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res., 2004; 32: W20-25.
[24] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. Nucleic Acids Res., 2000; 28: 235-42.
[25] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, et al. Open Babel: An open chemical toolbox. J Cheminformatics, 2011; 3: 33.

ISSN (O): 2581-8740 | ISSN (P): 2581-8732
Gupta *et al.*, 2021
DOI: 10.21276/SSR-IIJLS.2021.7.6S.17

[26]Sterling T, Irwin JJ. ZINC 15--Ligand Discovery for Everyone. J Chem Inf Model., 2015; 55: 2324-37.

[27]Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem., 2011; 32: 1466-74.

[28]Breiman L. Random Forests. Machine Learning, 2001; 45: 5-32.

[29]Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem., 2010; 31: 455-61.

[30]Hansen K, Mika S, Schroeter T, Sutter A, Laak At, et al. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. J Chem Inf Model., 2009; 49.

[31]Daina A, Zoete V. A boiled-Egg To Predict Gastrointestinal Absorption and Brain Penetration of Small Molecules. Chem Med Chem., 2016; 11: 1117-21.

[32]Gao Z, Chen Y, Cai X, Xu R. Predict drug permeability to blood-brain-barrier from clinical phenotypes: drug side effects and drug indications. Bioinformatics, 2017; 33: 901-08.

[33]Ivanov J, Polshakov D, Kato-Weinstein J, Zhou Q, Li Y, et al. Quantitative Structure–Activity Relationship Machine Learning Models and their Applications for Identifying Viral 3CLpro- and RdRp-Targeting Compounds as Potential Therapeutics for COVID-19 and Related Viral Infections. ACS Omega, 2020; 5: 27344-58.

[34]Kumar A, Loharch S, Kumar S, Ringe RP, Prakash R. Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2. Comput Struct Biotechnol J., 2021; 19: 424-38.

[35]Cagno V, Magliocco G, Tapparel C , Daali Y.The tyrosine kinase inhibitor nilotinib inhibits SARS-CoV-2 in vitro. Basic Clin Pharmacol Toxicol., 2021; 128: 621–24.

[36]Stebbing J, Krishnan V, de Bono S, Ottaviani S, Casalini G, et al. Mechanism of baricitinib supports artificial intelligence-predicted testing in COVID-19 patients. EMBO Mol Med., 2020; 12(8): e12697.