Artificial Intelligence Suggested Repositionable Therapeutics for Managing COVID-19: An Investigation with Machine Learning Algorithms and Molecular Structures

Dr. T.V. Sundar (sunvag@gmail.com)

PG & Research Dept. of Physics, National College (Autonomous), affiliated to Bharathidasan University, Tiruchirappalli-620001, India. https://orcid.org/0000-0001-5186-5029

Dr. K. Menaka

Dept. of Computer Science, Urumu Dhanalakshmi College, affiliated to Bharathidasan University, Tiruchirappalli-620019, India.

G. Vinotha

PG & Research Dept. of Physics, National College (Autonomous), affiliated to Bharathidasan University, Tiruchirappalli-620001, India.

Research Article

Keywords: Artificial Intelligence, Chemoinformatic Descriptors, COVID-19, Drug Repositioning, Free energy, Machine Learning Algorithms, Parthasarathi quantifier, ANOVA and Protein-small molecule binding

Posted Date: July 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-40988/v1

Abstract

The COVID-19 pandemic is almost half year old now and is still tormenting the humans to unimaginable extent with deeper interference to their routine life and peace. As approved vaccines are yet to be synthesized and standard therapeutic procedures are awaiting establishment for fighting against new Corona virus, several treatment modalities are being suggested and tried out by scientific community. Many of such approaches follow a drug repurposing approach as a possible remedy could prevent a great amount of loss in a shorter span of time. In this background, we report our attempt made for identifying a solution to this malady with a similar strategy. We used machine learning algorithms and the structural information of already approved drugs to identify potential therapeutics for managing the Covid-19 crisis. The experiments have been done with a group of 77 antiviral molecules (for the training phase of machine learning) and another group, comprising 9 antivirals and 11 antimalarials (meant for the testing phase). All the chosen molecules are approved category drugs and have significant drug action against the viruses. The identified molecules are subjected to validation by making docking studies with recently released crystal structures of Corona Virus. The binding affinity of the tested small molecules with three selected severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) structures are computed and compared with the affinity scores of five other medications viz. Hydroxychloroquine, Favipiravir, Dexamethasone, Dichlorobenzyl alcohol and Amyl metacresol followed by subjecting the results to the statistical test of ANOVA. The predicted therapeutics in conjunction with their already established characteristics could be further put to evaluation by approved clinical trials towards determining the efficiency of them against COVID-19 infection.

Introduction

The present Coronavirus disease 2019, COVID-19, having its origin in the Chinese city of Wuhan has been declared as pandemic by the World Health Organization (WHO)[1] and is nearly six months old now. It is creating devastating effects globally over the health and routine life of people, especially tormenting the middle and lower income group of people. The disease is observed in humans for the first time and no standard therapeutic procedure is existing till date to encounter it. Frequent hand wash with sanitizers, disinfecting the surroundings, maintaining social distance and self-quarantine by the affected persons, lockdown of geographical areas are deemed as effective measures to curtail the disease spread and its automatic stoppage. These suggested measures by WHO and faithful implementation by the multitude of nations help to fight against COVID-19 on one side but has started ruining the economy of many nations on the other side. In this work, we report our humble attempt made to identify a potential drug molecule for fighting against COVID-19. We adopted a drug repurposing approach and employed Machine Learning Algorithms (MLAs) to identify potential therapeutic molecule as it could greatly minimize time and loss in arriving at a solution to the problem. MLAs aim to optimize the performance of a certain task by using examples and/or past experience[2]. In general, the initial stages of drug discovery process involves the virtual screening of a large number of molecules possessing potential drug activity. This stage involves the use of rigorous computational techniques. Employing machine learning is a

viable strategy as it utilizes the intelligence captured (identifying a data pattern) by the algorithms while processing the supplied data. The present attempt of trying to identify a suitable drug for curing the pandemic virus involves a drug repurposing approach as it eliminates the laborious initial process of screening hundreds of candidate molecules and selection of a few out of them for further processing. Also, the approach fine tunes the focus of the experimental objectives. The main notion of employing MLAs in the present problem is to create a training model which could be used to identify a drug which could cure the flu symptoms and interact with the structural proteins of the Corona virus by learning decision rules inferred from the training data. For this, we derived descriptor data using the geometrical configuration of the molecules. The molecular descriptors are predominantly numerical in nature. According to Todeschini & Consonni[3], "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment". Thus a molecular descriptor is simply a numerical value that could be derived from molecular information sources like chemical formula, molecular structure, its interaction with other molecules etc. and are descriptors of the molecular details. Also the status of crystallography, which is the science of molecular structure elucidation, with the aid of computer automation for huge data processing, has been elevated from a probable molecular lead compound identifier to assist a fast, accurate, and efficient method for possible drug molecule identifier[4]. After identifying the potential molecules for the treatment of COVID-19, we have tested the binding of these molecules by performing docking studies with the recently released crystal structure of COVID-19 protease. The results could be put to further verification, laboratory evaluation, clinical trials and other evaluation protocols before concluding the suitability of the drug for Covid-19 treatment.

Material & Methods

Classification

The term, classification refers to the process of categorizing objects or ideas in the manner in which they are recognized and differentiated. An algorithm that implements classification with a rule or mathematical function by going through a set of data is known as a classifier and could map the input data to a category. In the terminology of machine learning, classification is considered as an instant of supervising learning procedure, where a training set of correctly identified observation is available. The corresponding unsupervised procedure is termed as clustering or cluster analysis. It involves grouping the data into categories based on some measure of inherent similarity. A classification algorithm in data mining creates a step by step guide for determining the output of a data instance. For classification experiment with machine learning, training examples are used to generate a model that can learn and classify the data samples into known classes. The Classification process involves the following sequence of operations: (i) Creation of training data set, (ii) Identification of class attributes and classes (iii) Relevance of useful attributes for classification (iv) Generating learning model with optimum tuning of algorithmic parameters for accuracy (v) Validation of the model with a known set of data and (vi)

Testing the model with independent set of data. The details of the algorithms used in this work are given in Table-1.

The principle of SVM is supervised learning. Its learning strategy tries to keep the error to the minimum values the method can be reliably used even for scrambled data sets. LIBSVM[5] is a popular Support Vector Machine (SVM) algorithm used in many classification problems with options to change the kernels (problem dependent) for effective classification. This kernel dependence has been illustrated in our earlier work[6] of druglikeness prediction of molecules with SVM. The Decision Tree (DT) classifiers create trees in the decision making process. In a created tree, after learning, each node represents a spot where, the decision had been taken based on the input. Also, it gives the direction of the test node and further nodes until the appearance of a leaf which predicts the output. A decision stump is a machine learning model consisting of a one-level decision tree[7]. It is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Usually it used in conjunction with a boosting algorithm. It performs regression (based on mean-squared error) or classification (based on entropy). The missing element is treated as a separate value. AdaB boosts nominal class classification with decision stump classifier. LogiB could perform additive logistic regression with decision stump classifier.

RC is capable of building an ensemble of randomizable base classifiers. Each base classifiers is built using a different random number seed (but based on the same data). The final prediction is a straight average of the predictions generated by the individual base classifiers.

Table-1 List of MLAs used and their learning scheme

S.No.	Algorithm	Scheme
1	LIBSVM with Linear kernel (LIBSVM-L)	Function
2	LIBSVM with Polynomial kernel (LIBSVM-P)	Function
3	AdaboostM1 with Decision Stump classifier (AdaB)	Meta
4	Logiboost with Decision Stump classifier (LogiB)	Meta
5	Random Committee (RC)	Meta
6	PART	Rule
7	J48	Tree
8	Random Tree (RT)	Tree

J48 is an algorithm for generating a pruned or unpruned decision tree[8]. It builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class. RT is supervised Classifier and is an ensemble learning algorithm[9]. It generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In a standard tree, each node is split using the best split among all variables. It constructs a tree that considers K randomly chosen attributes at each node. It performs no pruning. It also has an option to allow estimation of class probabilities (or target mean in the regression case) based on a hold-out set (back fitting).

Feature selection and test data

The data set is generated from the geometrical structure of each molecule using SwissADME[10] for building the training set. The numerical and categorical values computed under the categories viz. Physicochemical Properties, Lipophilicity, Water Solubility, Pharmacokinetics, Druglikeness and Medicinal Chemistry are taken as the data attributes. The categorical attributes with values 'Yes' or 'High' or numerically encoded as 1 and 0 respectively. The solubility class attributes viz. Highly soluble, Very Soluble, Moderately soluble, Soluble, Poorly soluble and Insoluble are encoded in a six point scale from 5 to 0. Prediction class of the molecule is added as the 48^{th} attribute in the data set. A total of 98 Molecules, comprising 87 antiviral and 11 antimalarial drug molecules belonging to approved drugs are taken from the DrugBank[11] molecular repository for the present study. The details are given in Table 2.

As Drug repurposing (repositioning) strategy is capable of identifying novel uses from existing approved and investigational drugs outside of their original indication, we have adopted that method for selecting the test data. The significant advantage in this approach is the absence of higher failure ratio in the clinical trials as the drug information and safety records are already available for consideration. Thus the method is more efficient and cost effective than traditional drug development since pre-clinical and early-stage clinical trials do not need to be repeated[12]. There are some explorations in trying out antivirals, used as HIV and influenza therapeutics[13, 14], on COVID-19 affected patients along with antimalarials [15, 16, 17]. In view of these and to expedite the process of getting some solution to fight COVID-19 we restricted our training and test set data to the approved antimalarials and antivirals (Table-2).

Classification of data sets

The nominal items 'Potential Drug' and 'Rejected' formed the classifier set elements in the last column of the training set data fed to the algorithms. As the training set must contain both positive examples and negative samples for arriving at a learning pattern, the antivirals approved for flu treatment are put under "Potential Drug' class and those used in HIV, hepatitis and HPV infections under 'Rejected' categories.

This training scheme resulted in the assignment of 3696 chemoinformatic descriptors to the algorithms for learning.

Table-2 Details of Molecules used in Training and Test Phases of the Experiment

Phase	Drug Category	Molecules*	Molecular Count	Total Number of Descriptors
Training	Antivirals	DB09296, DB11574, DB09372, DB08934, DB09027, DB13908, DB00426, DB13269, DB06290, DB01601, DB12020, DB00198, DB13997, DB00900, DB11799, DB00724, DB00718, DB09274, DB00879, DB00220, DB00529, DB08873, DB13156, DB08930, DB01048, DB13421, DB09101, DB00915, DB09102, DB09299, DB00432, DB00503, DB00558, DB00787, DB01004, DB00442, DB01179, DB06817, DB12301, DB13878, DB01319, DB11995, DB13879, DB00932, DB06414, DB00701, DB12070, DB12466, DB00249, DB00478, DB13288, DB05521, DB09183, DB11575, DB00632, DB00299, DB11613, DB00300, DB00369, DB12026, DB11586, DB00987, DB00577, DB00625, DB06614, DB03312, DB01265, DB00507, DB04835, DB00649, DB09297, DB00224, DB00709, DB01072, DB08864,	77	3696

		DB00705, DB01232.		
Test	Antivirals	DB00126, DB00943, DB14511, DB00495, DB01264, DB00238, DB01610, DB00811, DB00194, DB13751.	10	480
	Antimalarials	DB00468, DB00358, DB01218, DB01131, DB06697, DB09241, DB01087, DB06708, DB00205, DB01117, DB00613.	11	528

*** DrugBank codes**

Machine Learning Experiments

The injection of artificial intelligence to the machine learning algorithms can be done with the proper set of preprocessed data. The learning experiments are performed using WEKA[18] (Waikato Environment for Knowledge Analysis) (Version 3.8.3) which is a popular suite of machine learning software written in JAVA. It contains a collection of visualization tools and algorithm for data analysis with graphical user interfaces for easy access to its functions. All the experiments are carried out with a 2.4 GHz, i5 processor in a 64-bit operating system with 8 GB RAM.

Results

All the computational experiments are performed by using the default parameters provided in WEKA for the respective algorithms. The outcome of the classification trials are presented in Table 3. All the 21 test set molecules were presented to the learned models as potential drugs to COVID-19 treatment. Each algorithm identified the true cases and rejected the rest as false positives. The antiviral Zidovudine and the antimalarial Methylene blue are identified as potential therapeutics, sharing four votes each, by the eight algorithms testing the molecules. Proguanil (antimalarial) comes next with a selection frequency of three. The antiviral Nevirapine and the antimalarial Artemether are identified one time each while the rest of the tabulated antivirals find a prediction frequency of two times.

Table-3 List of Potential Drugs identified by the MLAs to act against COVID-19

S.No.	Algorithm	Probable Therapeutics
1	LIBSVM-L	Zidovudine, Ribavirin, Vidarabine, Methylene blue
2	LIBSVM-P	Zidovudine, Ribavirin, Vidarabine
3	AdaB	Ascorbic acid, Proguanil, Methylene blue
4	LogiB	Ascorbic acid, Proguanil, Artemether, Methylene blue
5	RC	Zidovudine, Methylene blue
6	PART	Zidovudine, Primaquine
7	J48	Nevirapine
8	RT	Proguanil, Primaquine

(Ascorbic acid DB00126; Zidovudine DB00495; Nevirapine DB00238; Ribavirin DB00811

Vidarabine DB00194; Proguanil DB01131; Artimether DB06697; Methylene blue DB09241

Primaquine DB01087)

The reliability of the above results could be accepted by computing some parameters with which we can evaluate the performance of the above classification algorithms. The commonly used ones are TP rate, FP rate, Precision, Recall, F- Measure, MCC and ROC area and are explained below. The overall accuracy of a classifier ($Accuracy_{total}$) on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The Error Rate (E_{rate}) is the misclassification rate of a classifier. Matthews Correlation Coefficient (MCC) is a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes The Confusion Matrix consisting of TP, TN, FP and FN as its elements is a useful tool for analyzing how well the classifier can recognize tuples of different classes. The sensitivity and specificity measures can be used to calculate accuracy of classifiers. Sensitivity is also referred to as the true positive rate (the proportion of positive tuples that are correctly identified), while Specificity is the true negative rate (that is, the proportion of negative tuples that are correctly identified). The definitions of these measures are given below.

$Accuracy_{total} = (TP+TN)/(TP + FP + TN + FN)$

$E_{rate} = (FP + FN) / (TP + TN + FN + FP)$

Sensitivity or TP rate or Recall = TP/ (TP + FN)

Specificity or FP rate = TN/ (TN + FP)

Precision or Positive Predictive Value (PPV) = TP/ (TP + FP)

Negative Predicted Value (NPV) = TN/(TN + FN)

$MCC = (TP*TN-FP*FN)((TP+FP)(TP+FN)(TN+FP)(TN+FN))^{1/2}$

where, True Positive (TP) and True Negative (TN) are correctly predicted Class-I and Class-II molecules, respectively. Similarly, False Positive (FP) and False Negative (FN) are wrongly predicted Class-II and Class-I molecules, respectively. To check whether the present classification scheme is better than a random prediction, a reliability factor (R) can be computed. It is given as

R = [((TP + FN) * (TP + FP)) + (TN + FN) * (TN + FP))] / (TP + TN + FP + FN)

This will give an anticipated number of molecules that could be correctly classified by random prediction. A factor **S** which is independent of the total number of samples in the data set can also be computed as

S = ((TP + TN) - R) / [((TP + TN + FP + FN) - R) x 100]

and it gives the normalized percentage of correctly classified Class-II molecules better than random classification. A value of S = 100% stands for a perfect classification and S = 0% for a poor classification. Apart from these parameters, the overall performance may be revealed out by a static called F parameter. It is the harmonic mean of precision and recall (or between sensitivity and positive predictive value). It is given as F = (2 x TP) / (2 x TP + FP + FN). The computed parameters are pictorially illustrated in the figures 1-3.

Interpretation Of The Results

Interpretation of Results

The total accuracy of each of the eight algorithms being over 96% , and that too with four of them scoring a perfect centum in the training phase indicates a cohesive learning of the presented descriptor data set by all of them. A very high as well as consistent sensitivity and specificity scores, over 87.5% and 95.5% respectively, obtained for the imbalanced training set (only 9 ‘potential drug’ class and 68 ‘rejectable’ drug class) is a measure of the reliability of the training phase. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. Similarly, the modelling performance indicators viz. F-ratio and MCC, both being in the range 0.8-1 indicate the good degree of prediction performance by all the algorithms. Generally, the area under precision-recall curve shows the tradeoff between precision and recall for different thresholds. In the Receiver Operating Characteristic (ROC) curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The ROC and PRC values also show a good degree of prediction performance by all the algorithms except for J48. As precision is considered as a direct and intuitive measure of performance in numerical experiments, the least values, PRC (0.708) and Precision

(PPV) (66.7%), for J48 algorithms suggest that Nevirapine, the mono molecular prediction by this algorithm, could be given least priority in the further investigations.

Molecular Binding Profiles:

To validate the results obtained and to explore the possibility of trying out the predicted molecules against the SARS-CoV-2, the two highly voted therapeutic candidates viz. Zidovudine and Methylene blue are subjected to binding affinity studies. For this, molecular docking experiments are performed using SwissDock[19,20]. The facility is a protein-small molecule docking web service based on EADock Dihedral Space Sampling (DSS) and makes fast docking using the CHARMM force field with EADock DSS. Three SARS-CoV-2 proteins are used in the analysis: The crystal structure of COVID-19 main protease in complex with an inhibitor N3[21] (6lu7), Structure of post fusion core of 2019-nCoV's S2 subunit[22] (6lxt) and Crystal structure of the free enzyme of the SARS-CoV-2 main protease[23] (6y2e). All the three proteins were retrieved from the RCSB Protein Data Bank (http://www.rcsb.org). The PDB codes are given in the parenthesis.

Oxford *et al.* [24] had earlier reported *in vitro* studies for the inactivation of three respiratory viruses (syncytial virus, influenza A and SARS-CoV). The research group used a lozenge formulation of amyl meta cresol and dichlorobenzyl alcohol at low pH, dissolved in artificial saliva and observed immediate clumping of virions occurred in the presence of the lozenge solution with Transmission Electron Microscopy studies and thereby established a virucidal destructive effect. After the outbreak of COVID-19, many drugs are being explored to reposition against SARS-CoV-2, some of them after clinical trials and after due statutory approval from Government bodies are being tried as therapeutics in the COVID-19 treatment. We have selected three such drugs viz. Hydroxychloroquine, Dexamethasone and Favipiravir along with amyl metacresol and dichlorobenzyl alcohol for making comparison with Zidovudine and Methylene blue in the docking experiments. While performing the experiments the default parameters offered by the SwissDock server is used for all the molecules. The results of the experiments are tabulated in Table-4.

Table 4: Profile of estimated free energies in the docking trials

Drug Molecule	DG (kcal/mol)			SARS-CoV-2
	min	max	mean	proteins
Zidovudine	-7.43	-5.53	-6.48	**6lu7**
Methylene blue	-7.63	-5.74	-6.69	
Dexamethasone	-7.21	-5.31	-6.26	
Favipiravir	-6.59	-4.69	-5.64	
Hydroxychloroquine	-9.20	-5.77	-7.49	
Dichlorobenzyl alcohol	-6.31	-5.30	-5.81	
Amylmetacresol	-6.48	-5.40	-5.94	
Zidovudine	-7.51	-4.54	-6.03	**6lxt**
Methylene blue	-8.18	-3.03	-5.61	
Dexamethasone	-7.64	-4.97	-6.31	
Favipiravir	-6.55	-4.53	-5.54	
Hydroxychloroquine	-8.24	-5.74	-6.99	
Dichlorobenzyl alcohol	-6.39	-4.83	-5.61	
Amylmetacresol	-6.18	-5.15	-5.67	
Zidovudine	-7.52	-5.77	-6.65	**6y2e**
Methylene blue	-8.37	-6.22	-7.30	
Dexamethasone	-8.47	-6.00	-7.24	
Favipiravir	-6.45	-5.15	-5.80	
Hydroxychloroquine	-9.06	-5.82	-7.44	
Dichlorobenzyl alcohol	-6.45	-5.33	-5.89	
Amylmetacresol	-6.69	-5.72	-6.21	

For these seven investigated molecules, the docking trials predicted 252- 257 best possible full fitness binding conformations, spread over eight clusters in each case. The mean values of the estimated free energy in binding (DG, in kcal/mol) hovers between -5.54 kcal/mole (for Favipiravir with 6lxt) and -7.49kcal/mol (for Hydroxychloroquine with 6lu7). In the results of the docking experiments, consistency in the computed values of binding free energy of the ligand is often expected. This is due to the fact that the binding mode of lead compounds is not determined experimentally but is essential for structure-

based lead optimization. The estimation of the binding free energy requires that the predicted and the native binding modes are as close as possible. Hence, to extract out the better ligand-receptor interaction information from the computed values of DG we propose and have calculated a new quantifier viz. '~DG' and call it as "*Parthasarathi Quantifier*". The symbol "~" indicates the difference between the maximum and minimum values of DG determined in the various binding conformations. The computed values of this parameter are given in Table-5.

Table 5: Values of the Parthasarathi Quantifier (~DG) for the investigated molecules

Drug Molecule	~DG values in kcal/mol for the SARS-CoV-2 proteins		
	6lu7	6lxt	6y2e
Zidovudine (Z)	1.90	2.97	1.75
Methylene blue (M)	1.89	5.15	2.15
Dexamethasone (D)	1.90	2.67	2.47
Favipiravir (F)	1.90	2.02	1.30
Hydroxychloroquine (H)	3.43	2.50	3.24
Dichlorobenzyl alcohol (DA)	1.01	1.56	1.12
Amylmetacresol (A)	1.08	1.03	0.97

A perusal of the above listed ~DG values show a consistent pattern except for Hydroxychloroquine. If we expect a consistent and minimal ~DG value for a specific molecule when it is docked with target proteins, for better ligand–target interactions, then Methylene blue which has a ~DG value of 5.15 kcal/mol with the protein 6lxt eludes from the tested group of seven molecules. Interestingly, it is one of the two candidate molecules predicted at par with Zidovudine in the AI based ML experiments. Moreover, the recent reports in the media regarding the inconclusive efficiency of Hydroxychloroquine for COVID-19 treatment and wide spread use of Dexamethasone and Favipiravir with relatively better performance also be kept in mind. Hence, before making a conclusion we performed the statistical test of ANalysis Of VAriance (ANOVA)[25] to determine the presence or absence of ligand-protein binding variation among a group of these seven molecular data sets. It is one of the most elegant, powerful and useful techniques for studying total variation in a set of data which may be reduced to components associated with possible scores of variability whose relative importance is needed. It involves the calculation of variance values of every member of the group elements from the mean value of that group and subsequently

analyzing the total variation between inter (variation between molecules) and intra (variation within a particular molecule's values) groups under analysis.

Table 6: ANOVA test of the binding interactions of the investigated molecules using Parthasarathi Quantifier dat

Groups	Members	Group Size	Degrees of freedom V1	Degrees of freedom V2	Table F-Ratio	Cal. F-Ratio	Variation in degree of protein-ligand interaction
1	Z, M, D, F, DA, A	6	5	12	3.11	2.59	No
2	Z, D, F, H, DA, A	6	5	12	**3.11**	**8.04**	**Yes**
3	M, D, F, H, DA, A	6	5	12	**3.11**	**3.48**	**Yes**
4	Z, D, F, DA, A	5	4	10	**3.48**	**5.09**	**Yes**
5	M, D, F, DA, A	5	4	10	3.48	2.65	No
6	Z, M, DA, A	4	3	8	4.07	2.70	No
7	Z, M, D, F,	4	3	8	4.07	0.83	No
8	D, F, DA, A	4	3	8	4.07	3.76	No

If we make all the seven molecules are equally good as potential therapeutics in the treatment of COVID-19, then according to ANOVA method the two mean squares (that of within molecules and between molecules) estimate the same quantity (error variance), and should be of approximately equal magnitude . Then the tabulated F-values (here, taken at five percent level) and the computed F- values are compared and used for arriving at a decision. If Table F-ratio is less than the computed F-ratio, then there is a scope for difference in the binding pattern of ligand-protein interactions by the selected ligands with a chosen target. Otherwise, the interactions could be almost similar. The results of the ANOVA test by altering the molecular groups is presented in Table-6. Here the degrees of freedom v1 and v2, used to compute the F values, denote number of molecules taken for the experiment minus one and number of ~DG values in all molecules minus number of molecules.

From the results of the ANOVA tests, we see that, if Hydroxychloroquine is excluded from the test group null hypothesis is validated and when included gets violated. Also, Zidovudine when grouped with Dexamethasone, Favipiravir, Dichlorobenzyl alcohol and Amylmetacresol null hypothesis is violated but this could be due to the statistical bias arising out of the binding score of Zidovudine's 2.97 kcal/mol when bound with 6lxt. The best binding positions of the ligands to the target, drawn using Chimera[26] are shown in Figure 4.

Conclusion

The performed machine learning experiments suggest that the dideoxynucleoside compound, Zidovudine (1-[(2R,4S,5S)-4-azido-5-(hydroxymethyl)oxolan-2-yl]-5-methyl-1,2,3,4-tetrahydropyrimidine-2,4-dione) and the oxidation-reduction agent Methylene blue (7-(dimethylamino)-N,N-dimethyl-3H-phenothiazin-3-iminium trihydrate chloride) could be two potential repositionable drug candidates for effectively fighting against the COVID-19 pandemic. The structure based descriptor tests of Zidovudine and Methylene blue show that both these molecules have high Gastro Intestinal (GI) absorption and fulfill all the criteria posed by Lipinski for druglikeness behavior. Also DrugBank annotations report that Methylene blue as a multipurpose drug finding several therapeutic and diagnostic applications and Zidovudine is capable of preventing DNA replication and improve immunological function. However, it is also reported that Zidovudine is found to initiate anemia and liver damage in certain cases which necessitates adjunct dosage if Zidovudine serves the purpose. In the light of these and based on the small molecule-protein docking studies and subsequent ANOVA analysis of the derived Parthasarathi Quantifiers we suggest that both Zidovudine and Methylene blue could be subjected to clinical trials similar to Favipiravir and Dexamethasone and actual efficacy of these two drug molecules could be estimated against SARS-CoV-2. A successful result of such clinical experiments could add additional warrior/s in the fight against COVID-19.

Declarations

Acknowledgments

TVS thanks Dr. V. Parthasarathi, Professor (Retired), School of Physics, Bharathidasan University, Tiruchirappalli-620024, India, Dr. S. Thamotharan, Senior Associate Professor, School of Chemical & Biotechnology, SASTRA Deemed University, Thanjavur - 613401, India and Dr. P.R. Ratnam M.D. (Ped) D.C.H., Tiruchirappalli for fruitful discussions. TVS, KM and GV thank their respective College Managements for encouraging research endeavours.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. European Center for Disease Control. “Covid-19.”https://www.ecdc.europa.eu/en/covid-19-pandemic. Accessed 2020 Apr 7.
2. Alpaydin, E. Introduction to Machine Learning. The MIT Press, Cambridge, MA, 2014.
3. Todeschini, R. and Consonni, V. Handbook of molecular descriptors. WileyVCH, Weinheim, 2000, Pub-2008.
4. Maveyraud, L and Mourey, L. Protein X-ray Crystallography and Drug Discovery. Molecules, 25, 1030, 2020; doi:10.3390/molecules25051030
5. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.
6. Vinotha, G. and Sundar, T. V. Drug Likeness Prediction Using Structure Based Molecular Descriptors and Support Vector Machines. International Conference on Advanced Materials for Technological Applications – ICAM 18. Materials Today: Proceedings 18, 1658–1669, 2019.
7. Wayne Iba; and Pat Langley,(1992); Induction of One-Level Decision Trees, in ML92: Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 1–3 July 1992, San Francisco, CA: Morgan Kaufmann, P.233–240.
8. Leo Breiman ,Random Forests. Machine Learning. 45(1):5-32; 2001.
9. Cutler.A, “Fast Classification Using Perfect Random Trees,”TechnicalReport5/99, Department of Mathematics and Statistics, Utah State University,May 1999.
10. Daina, A., Michielin, O., Zoete and V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Scientific Reports. 2017; 7:42717.see, http://www.swissadme.ch/index.php
11. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.
12. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M: Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019 Jan;18(1):41-58. doi: 10.1038/nrd.2018.168. Epub 2018 Oct 12.
13. Cao B, Wang Y, Wen D et al. A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19. N Engl J Med 2020; doi: 10.1056/NEJMoa2001282.
14. Blaising J, Polyak SJ, Pecheur EI: Arbidol as a broad-spectrum antiviral: an update. Antiviral Res. 2014 Jul;107:84-94. doi: 10.1016/j. antiviral.2014.04.006. Epub 2014 Apr 24
15. Liu J, Cao R, Xu M, et al. Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. Cell Discov. 2020;6:16. [PMID: 32194981] doi:10.1038/s41421-020-0156-0

16. Yao X, Ye F, Zhang M, et al. In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Clin Infect Dis. 2020. [PMID: 32150618] doi:10.1093/cid/ciaa237
17. Gautret P, Lagier JC, Parola P, et al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. Int J Antimicrob Agents. 2020:105949. [PMID: 32205204] doi:10.1016/j.ijantimicag.2020.105949
18. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
19. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W270-7. doi:10.1093/nar/gkr366. Epub 2011 May 29.
20. Grosdidier A, Zoete V, Michielin O. Fast docking using the CHARMM force field with EADock DSS. J Comput Chem. 2011 Jul 30;32(10):2149-59. doi: 10.1002/jcc.21797. Epub 2011 May 3.
21. Liu, X.; Zhang, B.; Jin, Z.; Yang, H.; Rao, Z., The crystal structure of COVID-19 main protease in complex with an inhibitor N3. 2020.
22. Zhu, Y., Sun, F. Structure of post fusion core of 2019-nCoV S2 subunit. (2020) Cell Res 30: 343-355
23. Crystal structure of the free enzyme of the SARS-CoV-2 (2019-nCoV) main protease Zhang, L., Sun, X., Hilgenfeld, R. (2020) Science
24. John S Oxford, Robert Lambkin, Iain Gibb, Shobana Balasingam, Charlotte Chan and Andrew Catchpole. A throat lozenge containing amyl meta cresol and dichlorobenzyl alcohol has a direct virucidal effect on respiratory syncytial virus, influenza A and SARS-CoV. Antiviral Chemistry & Chemotherapy 16:129–134,2005. International Medical Press 0956-3202.
25. NIST/SEMATECH e-Handbook of Statistical Methods, https://doi.org/10.18434/M32189.
26. UCSF Chimera–a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. *J Comput Chem.* 2004 Oct;25(13):1605-12. Home page http://www.rbvi.ucsf.edu/chimera

Figures

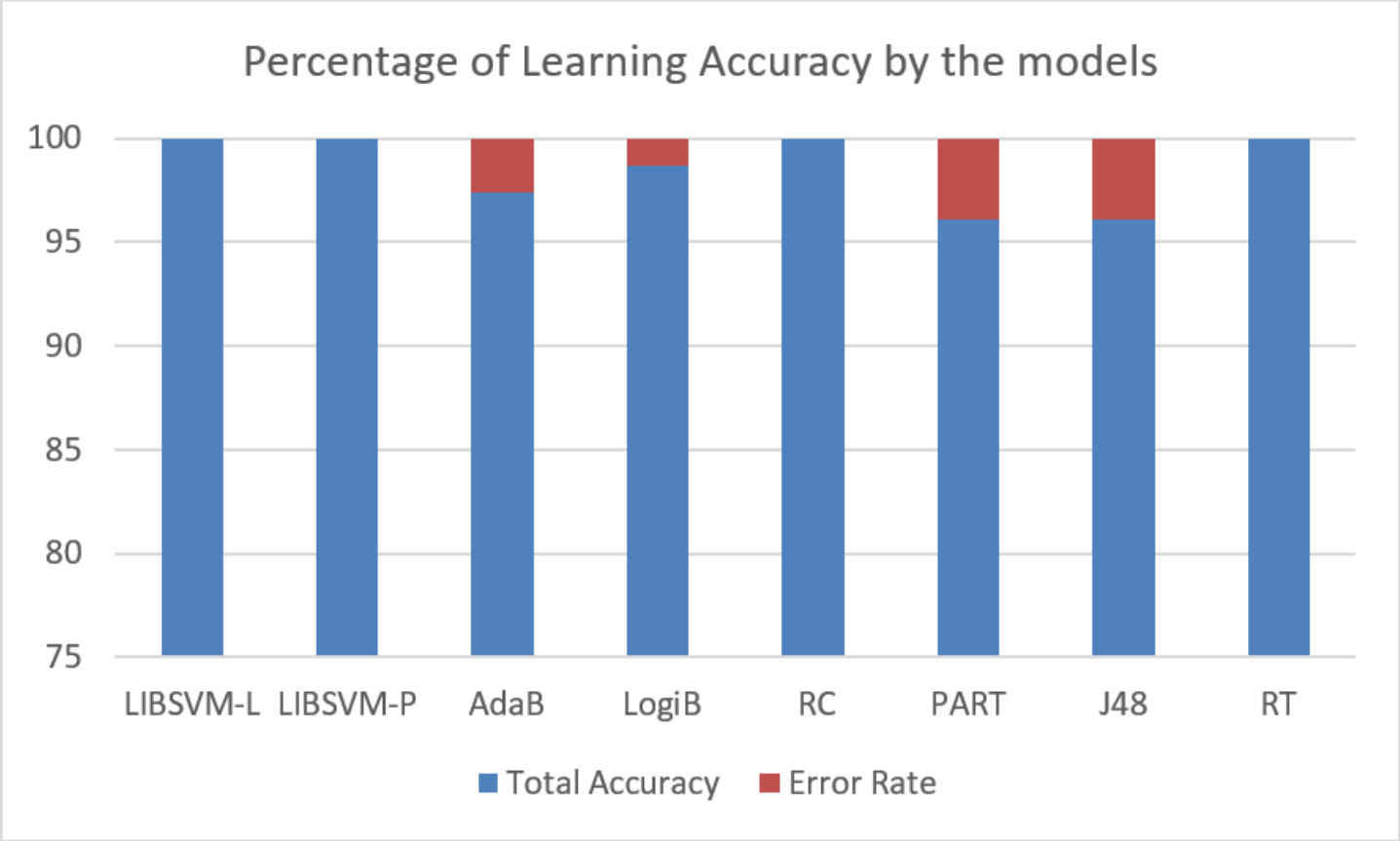

Figure 1

A comparative view of the accuracy values of learning rate by the MLAs

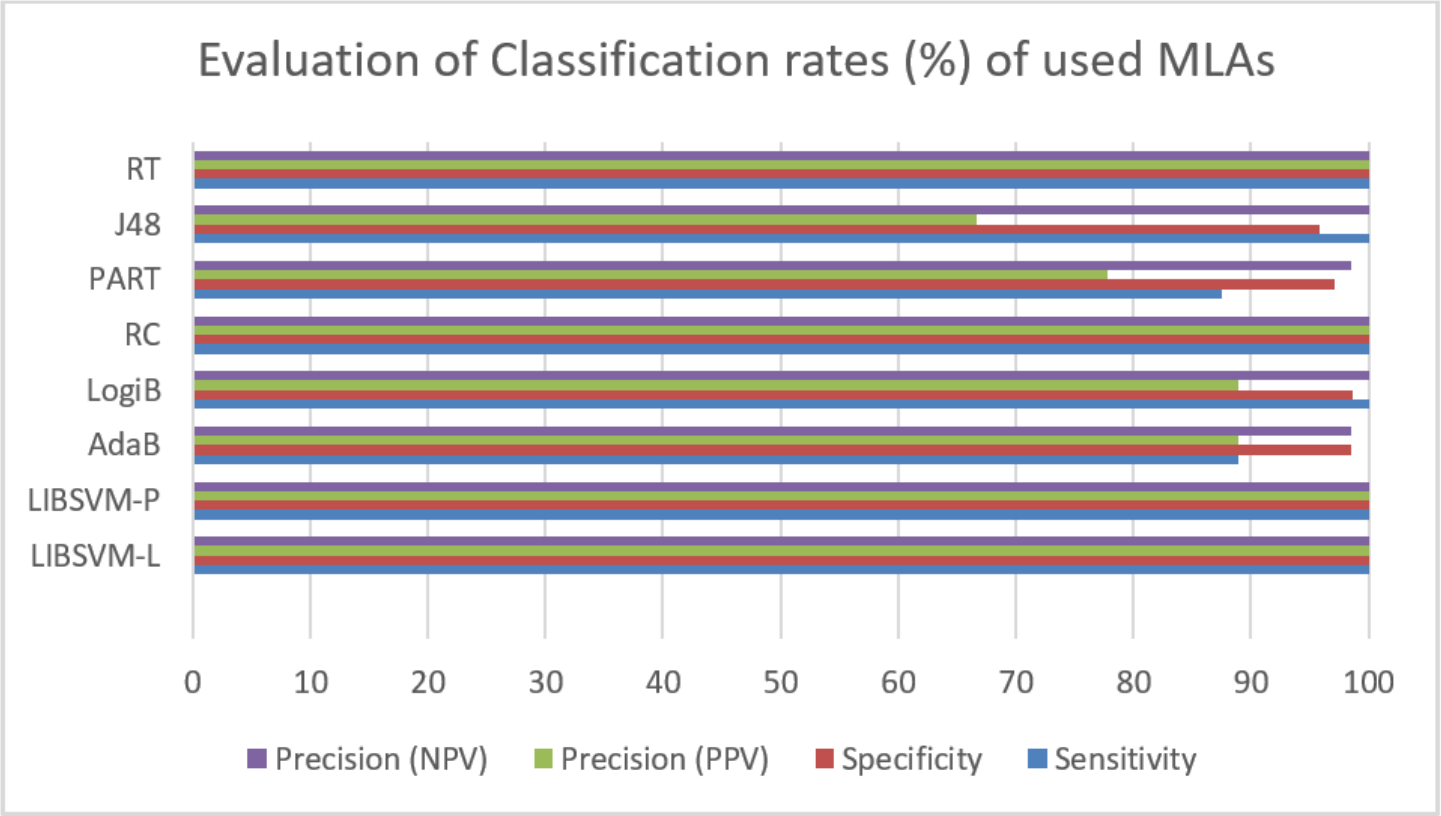

Figure 2

Comparison of the classification performance of the MLAs

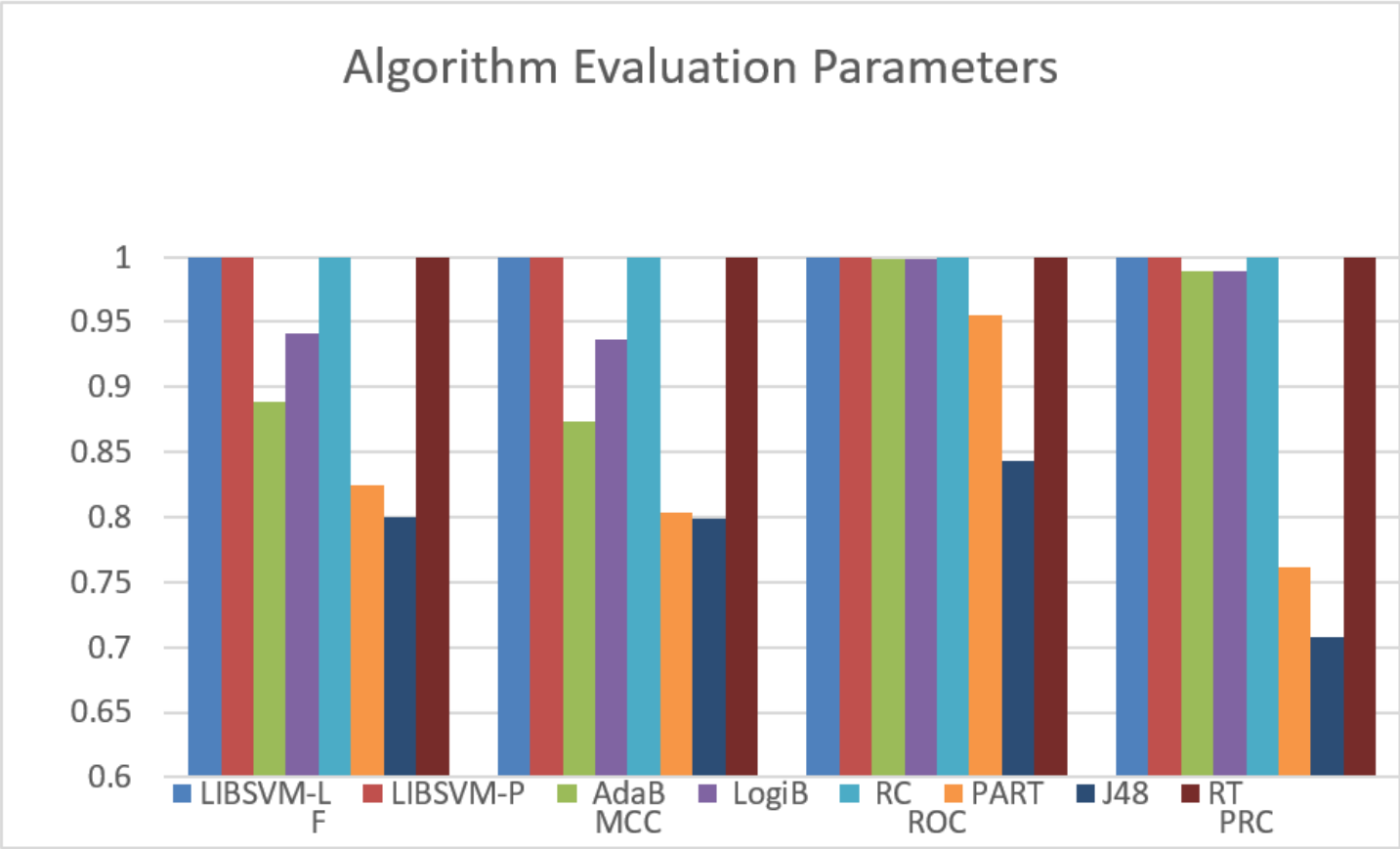

Figure 3

Range of Model evaluation parameters

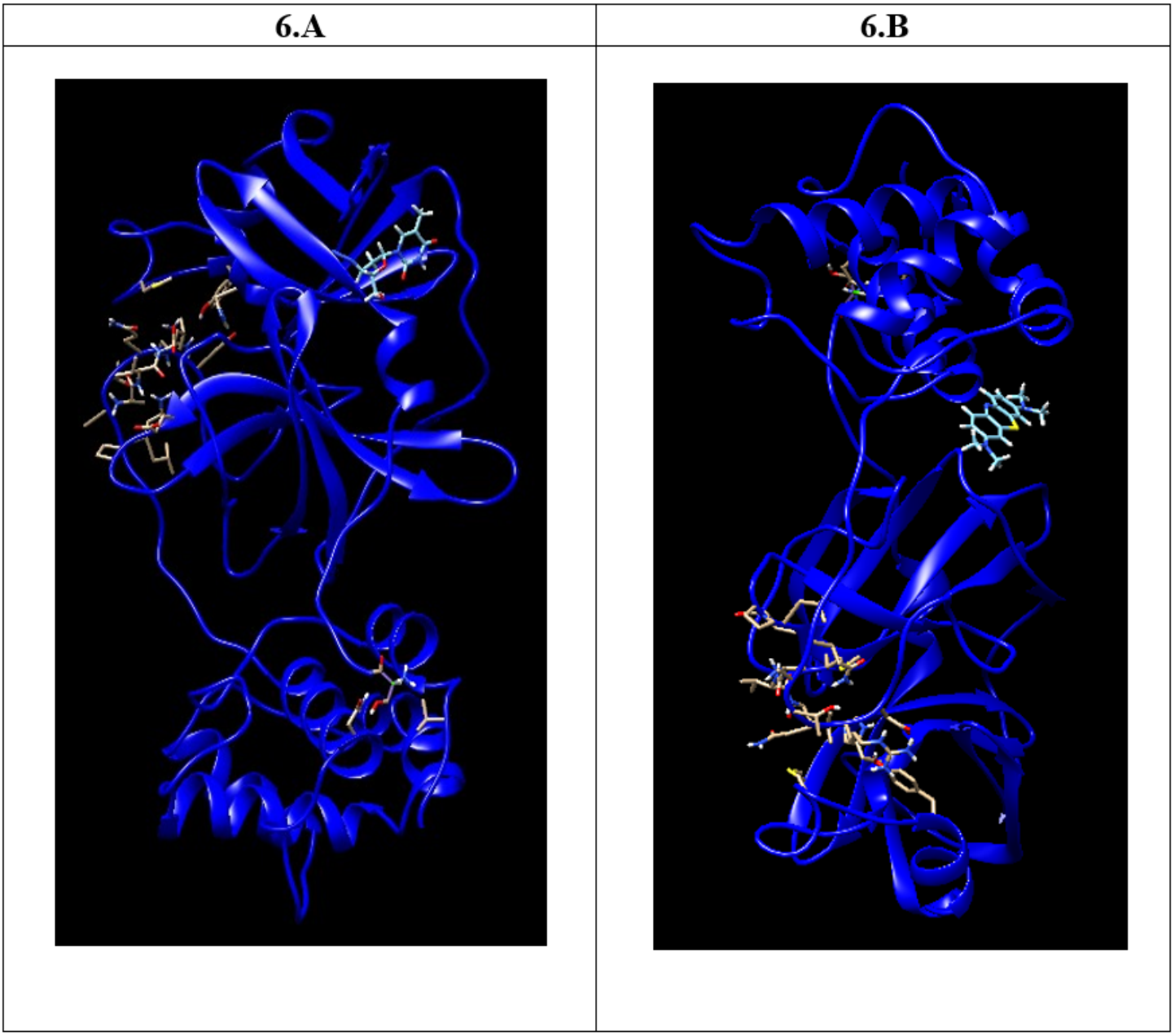

Figure 4

Best conformation for free energy of binding in 6lu7 with (A) Zidovudine (in the top right, -7.43 kcal/mol) (B) Methylene blue (in the middle right, -7.63 kcal/mol)