S. Afr. J. Chem., 2026, 80, S37–S54
https://journals.co.za/content/journal/chem/

Predicting the Inhibitory Potentials of 1-heteroaryl-2-alkoxyphenyl Analogues against SARS-CoV-2 virus via QSAR Models and Molecular Modelling

Oluwatoba Emmanuel Oyeneyin[1,2]*, Segun Michael Orimoloye[3,4], Ogunyemi O. Oderinlo[5], Nureni Ipinloju[1], Blessing Lully Isaiah[1], Abdul Rashid Issahaku[6], Damilola Alex Omoboyowa[7], Edmore Farayi Kangara[2] and Matshawandile Tukulula[2]

[1]*Theoretical and Computational Chemistry Unit, Department of Chemical Sciences, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria.*
[2]*School of Chemistry and Physics, College of Agriculture and Engineering, University of KwaZulu Natal, Westville, Private Bag X54001, Durban, 4000, South Africa.*
[3]*Faculty of Computing, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria.*
[4]*School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield S1 2NU, United Kingdom.*
[5]*Department of Chemistry, School of Mineral and Physical Sciences, University of Limpopo, Limpopo, Sovenga 0727, South Africa.*
[6]*Department of Chemistry, University of the Free State, 205, Nelson Mandela Avenue, Bloemfontein 9301, South Africa.*
[7]*Department of Biochemistry, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria.*

ABSTRACT

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has led to the hospitalization of over 287 million individuals worldwide and resulted in more than 7.1 million deaths globally. This pandemic has sparked a significant surge in medical research globally. Computer-aided drug design enables the study of the behavior of existing molecules. It makes predictions on the properties of new derivatives, which could show promise as safer and more effective alternatives. Therefore, this work considered machine learned quantitative structure-activity modeling in predicting the bioactivity of some 1-heteroaryl-2-alkoxyphenyl compounds. The grid-search support vector regression (GS-SVR) model was adopted as the representative machine learned QSAR model as it outperformed other models, $R^2 = 0.924$ (≥ 0.6) and $Q^2cv = 0.858$ (≥ 0.5). Furthermore, structural activity relationship studies generated 102 new compounds which were evaluated using the GS-SVR model. Lead compounds were selected and docked at the active sites of SARS-CoV-2 main protease (M^{pro}), papain-like protease (PL^{pro}), and spike glycoprotein. These compounds showed visible interactions with the amino acid residues at the active sites of these receptors. However, the docking scores of the compounds with the spike glycoprotein were higher than those of the standard drug, CP-100356, a spike glycoprotein inhibitor, unlike in the M^{pro} and PL^{pro}. The stability and flexibility of the complexes were determined *via* molecular dynamics simulation. The average RMSD values of the unbound protein and the complexes were below 2.50 Å, confirming their stability. Given the promising results, these molecules are currently being considered for synthesis and further investigation against SARS-CoV-2 cell lines.

KEYWORDS

Spike-glycoprotein, Quantitative structure-activity relationship, Machine learning, Molecular docking, Molecular dynamics simulation

Received: 18 July 2025 Revised: 14 January 2026 Accepted: 15 January 2026

INTRODUCTION

The positive-sense RNA virus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of COVID-19, which has caused millions of deaths and morbidity globally.[1] SARS-CoV-2 belongs to the large, enveloped, single-stranded RNA virus family *Coronaviridae* and subfamily *Coronavirinae,* which have a diameter ranging from 65 to 125 nm. They appear spherical with finger-like protrusions (glycoproteins) termed spikes on their surface.[2,3] The *Coronavirinae* subfamily is categorized into four groups (alpha, beta, gamma and delta) that have been reported to infest mammals, causing varying clinical syndromes.[4]

As a capped and polyadenylated viral RNA, SARS-CoV-2 has a 70-base 5′ leader sequence that encodes 13–15 open reading frames (ORFs), with 7–10 of those bases being transcription regulatory sequences (TRS-L). ORF1a, the longest region of the genome, encodes two genes; the papain-like protease and the 3CL protease. In contrast, ORF1b encodes enzymes critical for viral replication, including RNA-dependent RNA polymerase (RdRp), endonuclease and a helicase. Other regions encode structural proteins such as the spike (S) protein, envelope (E) protein, membrane (M) protein, nucleocapsid (N) protein, and accessory proteins. These proteins play crucial roles in the replication and survival of the virus.[5,6] The spike protein is a 150 kDa transmembrane protein embedded into the surface of the viral envelope and contains subunits S1 and S2 that moderate the anchoring and penetration of the virus *via* binding with the host cell protein angiotensin-converting enzyme 2 (ACE2).[7,8]

The M-protein is the most abundant structural protein of coronaviruses. The position of this protein enables it to interact with other structural proteins, such as the S-, E- and N-proteins[9], thus facilitating viral assembly. It also aids in S-protein attachment and is therefore culpable in immune evasion and pathogenicity.[9-11]. Papain-like protease (PLpro), also known as Nsp3, is the biggest Nsp and has a variety of functions. It oversees the cleavage of viral polyproteins into functional components along with Nsp5.[12] In addition, NNsp3 contributes to the assembly of viral particles, the start of cytokine storms, and the creation of double-membrane vesicles through interactions with Nsp4. Lastly, Nsp3 decreases the production of interferon genes, which are essential for the controllin viral infections, as well as the host enzyme poly-[ADP-ribose] polymerase (PARP), whose function is to impede viral replication.[13] The crucial roles of these proteins in the replication and survival of SARS-CoV-2 virus have made them attractive therapeutic targets for inhibiting and reducing the deleterious effects of the virus in humans.

Since the outbreak of the COVID-19 pandemic, several efforts have been made to contain the spread, as well as to treat and manage COVID-19 conditions. Although a level of success has been achieved in the vaccine industry[14,15], the demand for small molecules with the potential to target and inhibit the virus remains high on the agenda. Critical viral proteins such as the spike protease[16], the M-protease[17], and the Papain-like protease[18], among others, have been targeted with novel structurally diverse small molecules, including the repurposing of existing drugs[19,20]; however, small molecules with high selectivity

*To whom correspondence should be addressed
Email: emmanueltoba90@gmail.com; OyeneyinEmmanuelO@ukzn.ac.za; oluwatoba.oyeneyin@aaua.edu.ng

ISSN 1996-840X Online / South African Chemical Institute / http://saci.co.za/journal
© The Author(s) Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0)
https://doi.org/10.17159/0379-4350/2026/v80se1a04

and potency are in demand. Recently, a target-based screening of Medicines for Malaria Venture (MMV) open-access compound libraries against SARS-CoV-2 proteases,specifically the papain-like protease (PLpro), was conducted. From this screening campaign, followed by subsequent hit-optimisation, three heterocyclic small molecules emerged as promising PLpro inhibitors, exhibiting IC_{50} values in the range of 0.06–0.16 μM. Furthermore, the most potent compound in this series inhibited SARS-CoV-2 viral replication in HeLa cells and demonstrated favorable pharmacokinetic properties.[21]

Computer-aided drug design is crucial in the search for new and more efficient drugs than existing ones. It also speeds up the drug discovery process and has become popular among drug developers.[22] In drug optimization and modification, one of the techniques involves studying the behavior of existing drugs and derivatizing them for the purpose of improving their bioactivity with improved safety or for repurposing. Derivatives of 1-heteroaryl-2-alkoxyphenyl analogues have been reported to possess inhibitory potentials against SARS-CoV-2 replication.[23] However, further improvements and insights into their specific targets are required. This work utilises quantitative structural activity relationship (QSAR) to predict more potent derivatives, as well as other techniques including molecular docking and dynamics simulation, to study the interactions at the molecular level and their impact on the conformational dynamics of their targets. We further developed mathematical models, combiningtraditional QSAR and machine learning models to predict the activity of these compounds. This work thus provides novel derivatives with improved inhibitory prowess against SARS-CoV-2 targets (Spike protein).

MATERIALS AND METHODS

Data Collection

Some anti-COVID 1-heteroaryl-2-alkoxyphenyl analogues from antiviral assay against SARS-CoV-2, named A1-A21 here, with their EC_{50} in **μM** (Table 1), from the literature[23] were modeled and imported into Padel descriptor software[24] to generate their molecular descriptors. The descriptors were generated and saved in a .csv file for further processing. This process was repeated to generate the descriptors of 102 (B1-B102) newly derivatized molecules *via* structural-activity relationship (SAR).

Data Pre-processing

Normalization and Data Pre-treatment

The calculated descriptors were normalized to within the range of 0 to 1 using Equation 1. The Min-Max normalization scaler algorithm in Python's machine learning framework: scikit-learn was deployed. This gives each variable the same opportunity at the onset to influence and develop a good model.[25]

$$X = \frac{X_1 - X_{min}}{X_{max} - X_{min}} \quad (1)$$

X_1 is the descriptor's value for each molecule, χ_{min} and χ_{max} are the minimum and maximum values for each descriptor, respectively. The first phase of data pre-treatment was performed using the DTC Lab software[26] before data division to remove inter-correlated descriptors, using a Pearson correlation coefficient threshold of 0.85. A total of 199 such descriptors were removed from the original 1445 descriptors in the dataset.

Data Division

The Kennard and Stone's algorithm[27] method was used to divide the data into a training set (70% of the dataset, comprising 15 compounds), to build the model and a test set (30% of the dataset, comprising 6 compounds), for external validation of the built model. The second phase of data pre-treatment was done after data division to remove an additional 46 inter-correlated descriptors, resulting in 1200 descriptors for the actual model development. The Kennard Stone algorithm partitions the dataset based on samples with the maximum Euclidean distance between them. Threshold used for the statistical significance tests – R^2, LOF and Q^2 were included in the discussion of the results.

GFA-QSAR Model Development

The genetic function approximation (GFA) approach was used to develop the first QSAR model, from Material Studio software.[28] GFA is a computational method that models evolution in humans.[29] The population and generation of the model were set as large as possible for adequate convergence of the model, to 1000 and 2000, respectively. The biological activity (pEC_{50}) is the dependent (target) variable, while the physicochemical properties (molecular descriptors) are the independent variables. Four optimal descriptors (SpMax8_Bhm, SpMin4_Bhp, CIC2 and MDEC-22) obtained from GFA were used in building the model.

Internal validation of the GFA-QSAR model

Standard internal validation parameters from Material Studio were used to internally validate the training set. They include the following:

Friedman's lack of fit (LOF)

Friedman's LOF (Equation 2) was used to measure the fitness score of the models. LOF is defined as:[30]

$$LOF = \frac{SEE}{\left(1 - \frac{C + dp}{M}\right)^2} \quad (2)$$

where SEE, p, d, c, and M are the standard error of estimation, the total number of descriptors in the model, the user-defined smoothing parameter, the number of terms in the model, and the number of compounds in the training set, respectively. A model is described as good if it has a low SEE value, it is described in Equation 3 as:

$$SEE = \sqrt{\frac{\left(Y_{exp} - Y_{pred}\right)^2}{N - P - 1}} \quad (3)$$

The Coefficient of Determination (R^2)

This was used for the internal assessment of the QSAR model derived. $\mathbf{R^2}$ is expressed as (equation 4):

$$r^2 = 1 - \frac{\Sigma\left(Y_{exp} - Y_{pred}\right)^2}{\Sigma\left(Y_{exp} - Y_{training}\right)^2} \quad (4)$$

where $Y_{training}$ is the mean of the experimental bioactivity, Y_{exp} is experimental bioactivity and Y_{pred} is the predicted bioactivity in the training set.

Adjusted (R^2)

The adjusted R^2 is defined as (equation 5):

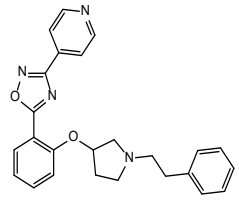
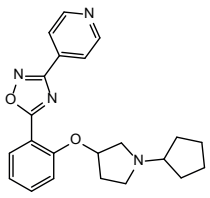
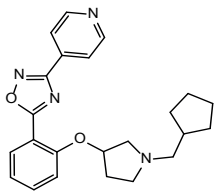
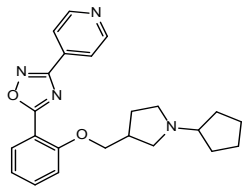
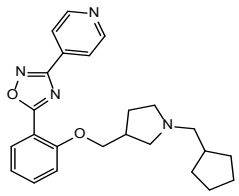
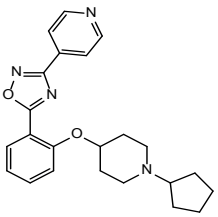
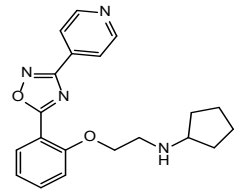
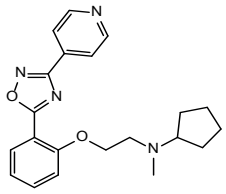
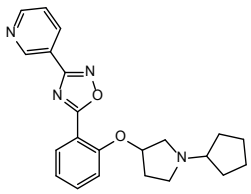
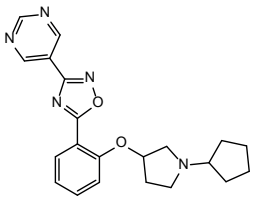
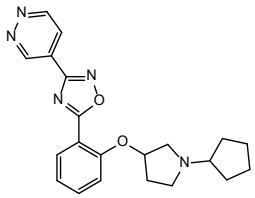
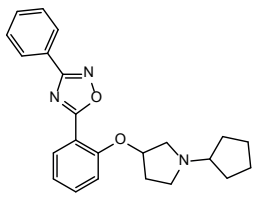
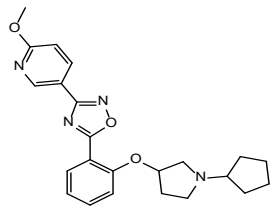
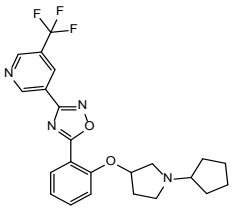
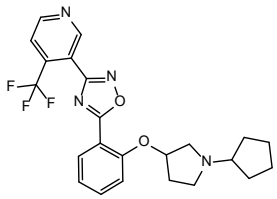
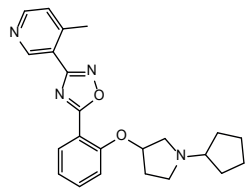
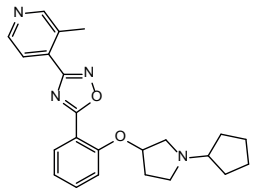
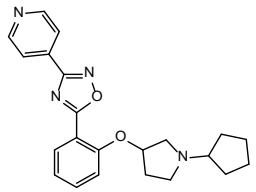
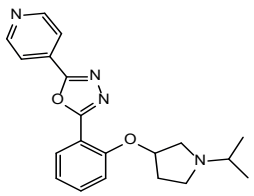
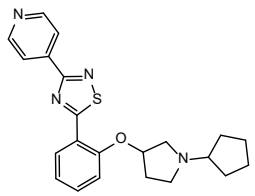
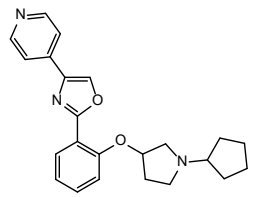
$$R^2_{adj} = \frac{R^2 - P(n-1)}{n - p + 1} \quad (5)$$

where p and n are the number of descriptors in the model and the number of compounds that make up the training set.

The cross-validation coefficient $\left(Q^2_{cv}\right)$

The strength of the QSAR model in predicting the activity of a new compound was determined using a cross-validation test. The cross-

Table 1: The anti-COVID compounds and their pEC_{50}, converted from their experimental EC_{50} (using $pEC_{50} = -\log EC_{50}$).

 **A1** (pEC_{50} = 5.3279)	 **A2** (pEC_{50} = 5.2676)	 **A3** (pEC_{50} = 5.7959)
 **A4** (pEC_{50} = 5.7696)	 **A5** (pEC_{50} = 4.3979)	 **A6** (pEC_{50} = 5.7696)
 **A7** (pEC_{50} = 5.0757)	 **A8** (pEC_{50} = 5.0706)	 **A9** (pEC_{50} = 5.0757)
 **A10** (pEC_{50} = 4.8239)	 **A11** (pEC_{50} = 4.3010)	 **A12** (pEC_{50} = 4.3010)
 **A13** (pEC_{50} = 5.3872)	 **A14** (pEC_{50} = 5.6778)	 **A15** (pEC_{50} = 5.5086)
 **A16** (pEC_{50} = 5.8539)	 **A17** (pEC_{50} = 5.7447)	 **A18** (pEC_{50} = 5.2076)
 **A19** (pEC_{50} = 4.8239)	 **A20** (pEC_{50} = 5.6383)	 **A21** (pEC_{50} = 5.2676)

*Experimental activity $pEC_{50}(\mu M)$.[23]

validation coefficient $\left(Q^2_{cv}\right)$ is defined as (equation 6):

$$\left(Q^2_{cv}\right) = 1 - \left\{ \frac{\Sigma\left(Y_{pred} - Y_{exp}\right)^2}{\Sigma\left(Y_{exp} - Y_{training}\right)^2} \right\} \quad 6$$

where $Y_{training}$ is the mean of the experimental bioactivity, Y_{exp} is experimental bioactivity and Y_{pred} is the predicted bioactivity in the training set.

External validation of GFA-QSAR model

The built model was first validated externally using the $R^2_{predicted}$ value. The $R^2_{predicted}$ value is defined as (equation 7):

$$R^2 = 1 - \frac{\Sigma\left(Y_{pred\,test} - Y_{exp\,test}\right)^2}{\Sigma\left(Y_{pred\,test} - \bar{Y}_{training}\right)^2} \quad 7$$

Standard model performance evaluation metrics: correlation coefficient (R), mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE) and root mean squared error (RMSE) were also used to externally validate the GFA-QSAR model. They are defined in equations 8, 9, 10, 11 and 12, respectively.

$$R = \sqrt{1 - \frac{\Sigma\left(Y_{pred\,test} - Y_{exp\,test}\right)^2}{\Sigma\left(Y_{pred\,test} - \bar{Y}_{training}\right)^2}} \quad 8$$

$$MAE = \frac{\Sigma(|\,Y_{pred\,test} - Y_{exp\,test}\,|)}{n} \quad 9$$

$$MAPE = \frac{\Sigma\left|\frac{Y_{pred\,test} \quad Y_{exp\,test}}{exp\,test}\right|}{} \times 100 \quad 10$$

$$MSE = \frac{1}{n}(Y_{pred\,test} - \bar{Y}_{pred_{test}})^2 \quad 11$$

$$RMSE = \sqrt{\frac{1}{n}(Y_{pred\,test} - \bar{Y}_{pred_{test}})^2} \quad 12$$

where n, $Y_{predtest}$, $Y_{exptest}$, $\bar{Y}_{predtest}$ and $\bar{Y}_{training}$ are the number of compounds, predicted bioactivity, experimental bioactivity and the mean of the predicted bioactivity for the test set.

Evaluation of the applicability domain of the QSAR model

The leverage approach was employed here[31,32] to verify the QSAR model's suitability for screening new compounds in its domain of application, and it is given as (equation 11):

$$hi = X_i\left(X^TX\right)^{-1}X_i^T \quad 13$$

where hi is the leverage of a chemical compound, X_i is the training compounds matrix of i. X is the m x k descriptor matrix of the training set compound and X^T is the transpose matrix of X. As a prediction tool, the warning leverage (h*) is the limit of normal values for X outliers and is defined as follows (equation 12):

$$h^* = 3\frac{(k+1)}{n} \quad 14$$

where n and k are the descriptors and the training set compounds, respectively.

Machine learning-based QSAR models

Four prominent machine learning algorithms that have been widely used for QSAR modeling and have yielded promising results were deployed to build QSAR models in order to compare their performance with the baseline GFA-QSAR model. They include multiple linear regression (MLR)[33], random forest (RF) regression[34], support vector regression (SVR)[35] and a variant of SVR that uses the Grid-search algorithm with cross-validation for hyper-parameter optimization (Grid-SVR).[36]

Multiple Linear Regression

Multiple (multivariate) linear regression (MLR) is one of the most basic supervised machine learning algorithms. MLR came about as a result of the limitations of the simple linear regression model, where the prediction of the response (target) variable is dependent on a single independent variable (predictor). In real-life scenarios, there are many factors that determine the value of a continuous-valued response variable, hence the need for MLR.

Given *p* distinct predictor (input) variables: $X_1, X_2, \ldots, X_p$, a multiple linear regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p + \varepsilon \quad 15$$

where X_j represents the j^{th} predictor and β_j is a regression coefficient of j^{th} predictor variable, which connotes a quantitative relationship between that variable and the target variable; ε is the error term.[37]

The least-squares approach is used to fit a line of fit between the predicted response and actual response, and estimate the parameters of the MLR model – residual sum of squares (RSS) and residual standard error (RSE). The predicted response for the i^{th} predictor is termed $\hat{y}_i$, while the actual value is y_i. We therefore estimate the values of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$ that would minimize RSS.[33]

$$RSS = \sum_{i=1}^{n}\left(yi - \widehat{yi}\right)^2 \quad 16$$

Random Forest Regression

Random forests are a special category of computational intelligence techniques that belong to the ensemble-based machine learning models group, where multiple models are combined into one to give more accurate predictions. Ensemble models can be either homogeneous (same type of individual models) or heterogeneous (different types of machine learning models combined). In the case of random forests, the models are homogeneous - multiple decision trees.[38] Random forests as originally introduced by[34] were meant to be classifiers but random forest regression is a generalization of random forest classifiers for solving regression problems. Random forests prominence was born out of the need to significantly reduce the bias and variance associated with individual classification and regression (decision) trees by using the divide and conquer approach to grow randomized tree predictors from data samples and aggregating their results.[39]

Cutler and Cutler[40] gave a mathematical basis for random forest regression, thus:

Let $X = (X_1, X_2, \ldots\ldots, X_p)^T$ be a p-dimensional random vector representing the real-valued input or predictor variables and Y be a random variable representing the real-valued target, we assume an unknown joint distribution $P_{XY}(X,Y)$. The objective is to find a prediction function $f(X)$ for predicting Y such that a loss function L(Y, $f(X)$) minimizes the expected value of the loss given as:

$$E_{XY}\,(L(Y, f(X))) \quad 17$$

L(Y, $f(X)$) is a measure of the closeness of $f(X)$ to Y and it penalizes values of $f(X)$ that are far off from Y. A typical choice of L is *squared error loss* L(Y, $f(X)$)

$$L(Y, f(X)) = (Y - f(X))^2 \quad 18$$

Minimizing $E_{XY}\,(L(Y, f(X)))$ for squared error loss gives the condition expectation:

$$f(X) = E(\,Y|\,X = x) \quad 19$$

also called the *regression function.*

The function *f* is constructed as a collection of each decision tree known as base learners, denoted as $h_1(X, \Theta_1)$, $h_2(X, \Theta_2)$,….. $h_j(X, \mathbf{\Theta}_j)$ where $\Theta_1, \Theta_2, \ldots \mathbf{\Theta}_j$ are independent random variables. They are averaged and combined to give the ensemble predictor *f*(X). They are averaged thus:

$$f(X) = \frac{1}{J}\sum_{j=1}^{J} hj(X) \quad 20$$

Support Vector Regression

The Support Vector Machine (SVM) as a statistical learning method was originally developed for classification problems where the values of the target variable to be predicted belong to a finite or discrete set, and thus formulates the classification problem as a convex optimization problem.[41] This optimization problem is aimed at finding the maximum margin separating the hyperplane, where as many training examples (points) as possible are correctly classified. The optimal hyperplane selected is represented with support vectors. However, in regression problems, as in this work, the classification problem is generalized such that the model predicts a continuous-valued output. The sparseness of the solution and the good generalization ability of SVMs are made possible through their adaptation to SVRs (Support Vector Regression) by introducing the ε-insensitive loss function.[42]

The main idea of SVR is to map the support vectors into a higher-dimensional feature space, and then construct a regression function *f*(*x*) that can accurately predict future values.[43] Given a set of training data: $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_n, y_n)\}$ where $x_i \in R^N$ and $y_i \in R$, our objective is to produce a regression function f(x) that approximates the non-linear relationship in (18) such that f(x) is as close as possible to y to avoid over-fitting.

$$f(x) = \omega^T.\Phi(x_i) + b \quad 21$$

where ω^T is the weight vector, b is the bias and Φ(x) represents the transformation function that maps the lower-dimensional input space to a higher-dimensional space. The optimization problem is presented in a more formal way, thus:

$$\text{Min } \|\omega\|^2 + C\sum_{i=i}^{n}(\xi + \xi^*)$$

$$\text{s.t} \quad \begin{aligned} & y_i - \omega^T \varphi(x_i) - b \le \varepsilon + \xi \\ & \omega^T \varphi(x_i) + b - y_i \ge \varepsilon + \xi^* \\ & \xi, \xi^* \ge 0, i = 1,2,\ldots n \end{aligned} \quad 22$$

where ε is the difference between the predicted value, f(x) and actual value y, n is the number of training examples and ξ, ξ^*, *i* are additional slack variables introduced by[44], which determines that deviations of magnitude ξ above ε error are tolerated. The constant *C* is referred to as the penalty (or regularization) parameter, which determines the penalty associated with the misclassifications based on the hyperplane adopted by the model.

Lagrangian multipliers, which are reliant on the dot product of Φ(x) are adopted in SVRs and are accomplished through kernel functions defined as $K(x_i,x_j) = (\Phi(x_i), \Phi(x_j))$. Common kernel functions used in SVRs are Radial Basic Function or Gaussian (RBF), linear, polynomial and sigmoid.[35]

Grid-Search Based Support Vector Regression

It has been established that the level of accuracy of a machine-learning model is dependent on the parameter settings. Therefore, developing an accurate model requires the selection of optimal values of the model's hyperparameters. The hyperparameters of SVR algorithm include the type of kernel function and its parameter (gamma), the kernel's degree (for polynomial kernel), the insensitive loss function parameter (ε) which is the maximum tolerance level allowed for the model's residual error; and the penalty/regularization parameter (C) which is a penalty that makes the model avoid large prediction residual error, by penalizing predictions that are larger than ε – the larger the penalty, the more difficult it is for the model to produce large errors.[45]

Different machine learning QSAR modelling techniques, such as genetic algorithms[46] and Particle Swarm Optimization[47], have been used to select the optimal values of SVR's hyperparameters. However, the Grid search still remains one of the most widely used. The Grid search algorithm with cross-validation employs a brute-force approach to select the optimal value of each hyperparameter from a range of values specified apriori, such that the prediction performance criterion, which in most cases, the mean squared error (MSE) is minimized.[36] It finds the best combination from every possible combination of hyperparameter values in the search space. Its major advantage lies in its high learning accuracy and parallel processing on each iteration of the cross-validation procedure. It is, however, limited by the computational overhead incurred in large datasets, which makes it more ideal for small datasets.[43]

SAR on the existing compounds

A structure-activity relationship (SAR) study was carried out on existing compounds to get new derivatives whose bioactivity can be predicted using the developed QSAR models. The derivatives were predicted by considering various factors. Sheathing the amine group, for example, could provide many derivatives. In total, 102 new compounds were derivatized from the existing compounds. Some derivatives were obtained by incorporating cyclohexyl linkers into compound A6, for instance.

Molecular docking

In order to investigate the inhibitory potentials of these compounds against SARS-CoV-2, the lead compounds were docked at the binding pockets of SARS-CoV-2 main protease (5R7Y, resolution 1.65 Å), spike glycoprotein (6M0J, resolution 2.45 Å) and a papain-like protease (3E9S, resolution 2.70 Å) using Schrödinger suite.[48] The structures of the proteins were downloaded from an online protein database (www.rcsb.org). For 6M0J, it is a SARS-CoV-2 spike protein subunit S1 that was complexed with an angiotensin-converting enzyme 2 (ACE2). The ACE2 and its accompanying ions and/metals were removed to retain the SARS-CoV-2 spike protein subunit S1. The protein and ligand structures were prepared using the protein preparation wizard and LigPrep[49] module using the OPLS4 forcefield.[50] From the proteins, water molecules and other non-binding atoms/molecules/metals were removed. Additionally, missing loops, hydrogen atoms, cofactors and ions were also added during protein preparation to ensure the full function of the binding domain. The active sites (x = 10.59, y = -2.29, z = 24.17 for 5R7Y; x = -30.99, y = 21.75, z = 30.12 for 3E9S and x = -32.48, y = 12.64, z = 32.5 for 6M0J) of the proteins were determined using the site map and receptor grid generation modules before docking the molecules at the active sites.

Molecular Dynamic Simulation

MD simulations for six (6) systems comprising the unbound protein(Apo), A16 complex, A17 complex, B92 complex, B101 complex, and CP10035 complex were performed using the Graphics Processing Unit (GPU) version of Particle Mesh Ewald Molecular Dynamics (PMEMD) of AMBER18.[51] The FF14SB force field was used to parameterize the complex systems.[52] ANTECHAMBER was used to create atomic partial charges for the compounds by using the Restrained Electrostatic Potential (RESP) and the general Amber Force Field (GAFF) protocols.[53] The systems were then suspended in Transferable Intermolecular Potential with 3 Points (TIP3P) solvent and then neutralized with the addition of 0.15M Na^+ and Cl^- counter ions. Before MD production, the systems were initially minimized with the steepest gradient for 10,000 steps and then finally minimized with a conjugate gradient for 5,000 steps using a restraint potential of 500 kcal/molA^2.

The systems were then heated slowly from 0 K to 300 K for 5ps in an isothermal-isobaric NPT ensemble using a Langevin thermostat.[54] To remove bad energy and contacts, the systems were equilibrated at 300 K for 1ps without energy restraint. MD production was then performed for 200ns, with the SHAKE algorithm, which restrained all hydrogen bonds at 2fs time step. The post-MD production analysis was performed through the CPPTRAJ and PTRAJ modules[55] of AMBER 18. The Origin data tool was used to plot graphs, while Schrödinger Maestro was used for visualization.

Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) calculations

The MM-PBSA method was employed to calculate the total free binding energies averaged over 50,000 snapshots. The calculated free binding energies help in understanding the ligand(compound)-receptor relationship. This method has proven effective and reliable in predicting the free binding energies of complexes[56,57] as seen in equations 23-27.

$$\Delta E_{bind} = E_{complex} - (E_{protein} + E_{ligand}) \quad 23$$

$$\Delta G_{bind} = E_{MM} + G_{sol} - T\Delta S \quad 24$$

where ΔG_{bind} indicates the free energy of the ligand-protein complex total binding, minus conformational binding entropy ($T\Delta S$).

$$\Delta E_{MM} = \Delta E_{int} + \Delta E_{elec} + \Delta E_{vdw} \quad 25$$

where ΔE_{MM} shows the total gas phase energy contributed by sum of internal energy, electrostatic and van der Waals energy components. The internal energy indicates the energy that arises from various bonds, angles and torque.

$$E_{int} = E_{bond} + E_{angle} + E_{torsion} \quad 26$$

The sum of the polar $\Delta G_{PB/GB}$ and non-polar ΔG_{SA} contributions to solvation are also represented as follows:

$$\Delta G_{sol} = \Delta G_{PG/GB} + \Delta G_{SA} \quad 27$$

The polar solvation contribution is denoted as G_{PB} and $G_{non\text{-}polar}$ represents the non-polar contribution energy and is computed from the solvent assessable surface area (SASA), which is obtained by using a 1.4 Å water probe radius.

RESULTS AND DISCUSSION

GFA-QSAR Model

Four models were generated using the Material Studio 2017 software based on the GFA approach. The model with the most optimal set of parameter values suitable for a reliable and stable model, which also satisfied the test of statistical significance, was selected (Table S1). The list of descriptors, their descriptions and the dimensions of the descriptors used in building the model are reported in Table S2. They descriptors are: largest absolute eigenvalue of Burden modified matrix-n8/ weighted by relative mass (SpMax8_Bhm), smallest absolute eigenvalue of Burden modified matrix-n4/weighted by relative polarizabilities (Smallest absolute eigenvalue of Burden modified matrix-n4/weighted by relative polarizabilities), complementary information content index/ neighborhood symmetry of 2-order (CIC2) and molecular distance edge between all secondary carbons (MDEC-22). The model generated by GFA is presented in equation 28.

$$pEC_{50} = 1.128099888 * \mathbf{SpMax8_Bhm} - 2.101497100 * \mathbf{SpMin4_Bhp} - 1.558563054 * \mathbf{CIC2} + 1.863917815 * \mathbf{MDEC\text{-}22} + 4.860691421 \quad 28$$

The comparison of experimental bioactivity, predicted bioactivity and residuals of the developed GFA-QSAR model was presented in Table 2.

The graphical representations in Figures 1 and 2 depicts the comparison of the experimental and predicted bioactivity and the residual plot respectively.

Table 2: Comparison of experimental bioactivity, predicted activity and residuals of GFA-QSAR model.

S/N	Experimental Bioactivity	Predicted Bioactivity	Residual Values
A1	*5.32790200	5.39808500	-0.07018200
A2	5.26760600	4.99249800	0.27510800
A3	5.79588000	5.64036300	0.15551800
A4	*5.76955100	5.75999200	0.00955900
A5	4.39794000	4.34194800	0.05599200
A6	5.76955100	5.75999200	0.00955900
A7	5.07572100	5.09947600	-0.02375500
A8	*5.07058100	5.03378200	0.03679900
A9	5.07572100	5.09947600	-0.02375500
A10	4.82390900	4.66317700	0.16073100
A11	4.30103000	4.63411100	-0.33308100
A12	4.30103000	4.37018100	-0.06915100
A13	5.38721600	5.20068500	0.18653100
A14	5.67778100	5.70092400	-0.02314300
A15	5.50863800	5.78809700	-0.27945900
A16	*5.85387200	5.74587300	0.10799900
A17	5.74472700	5.74707200	-0.00234400
A18	*5.20760800	5.40840700	-0.20079800
A19	4.82390900	4.89146700	-0.06755800
A20	5.63827200	5.40840700	0.22986500
A21	*5.26760600	5.40204000	-0.13443400

*test set

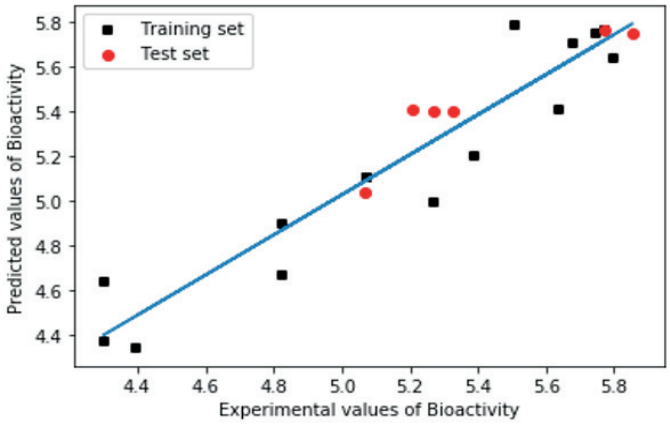


Figure 1: Graph of Experimental vs. Predicted Bioactivity for GFA-QSAR model.

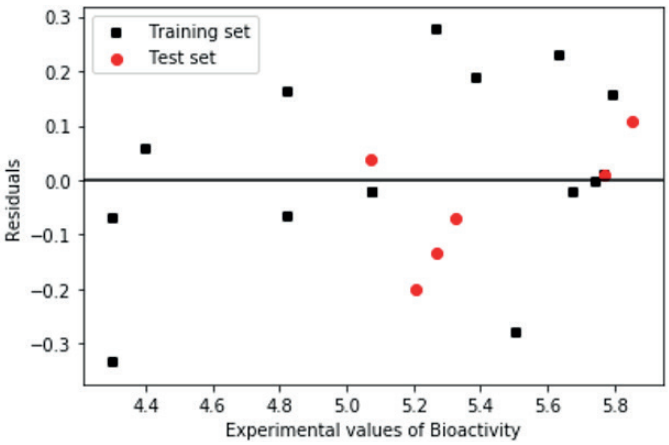


Figure 2: Residual Plot for GFA-QSAR model.

Validation results for GFA-QSAR model

The results for the internal validation of the GFA-QSAR model are presented below (Table S3). LOF = 0.139184, R^2 training set = 0.898489, R^2 adjusted = 0.873112, Q^2_{cv} = 0.89833, N training = 15. External validation results for the GFA-QSAR model are: correlation coefficient (R) = 0.9377, $R^2_{predicted}$ = 0.8792, mean absolute error (MAE) = 0.093295, mean absolute percentage error (MAPE) = 1.74%, mean squared error (MSE) = 0.012738, root mean squared error (RMSE) = 0.11286.

Applicability Domain Analysis

The warning leverage (h*) was computed as h* = (3*(4 + 1)) / 5 =1.0. The Williams Plot of leverage and Standardized Residuals is presented in Figure 3.

The results from both internal and external validation of the GFA-QSAR model show that it is a statistically valid and reliable model as the values of all validation parameters passed the statistical test of significance and exceeded the threshold values - correlation coefficient (R) = 0.9377, R^2 = 0.898489, R^2 adjusted = 0.873112, R^2 predicted = 0.8792 (all ≥ 0.6, and the closer to 1.0 the better); R^2 - Q^2 = 0.159 (≤ 0.3), Q^2_{cv} = 0.89833 (≥ 0.5); LOF = 0.139184 (the closer to 0 better); MAE = 0.093295, MAPE = 1.74%, MSE = 0.012738, RMSE = 0.11286 (the closer the values of the model performance evaluation metrics are to 0, the better).[58] The high correlation coefficient, R value for the training set (0.9377) confirmed that the model can successfully predict the activity of a new compound, as the correlation between the predicted bioactivity and experimental bioactivity is quite strong.

The GFA-QSAR model, as shown in the from the Williams Plot in Figure 3, indicates that there was only one structurally influential chemical – a compound from the test set (compound 16), which had a leverage value higher than the warning leverage (h = 2.18, where h* = 1.0). There were no outliers in the model as all compounds (both in the training and test sets) had standardized residuals within the allowable ± 3.0 standard deviation limit.[59] Since all but one of the leverage values fell within the chemical space and there are no outliers, we can therefore conclude that the developed QSAR model effectively predicts that a new chemical will lie within its structural domain and is hence considered reliable.

Machine-learning based QSAR Models

Codes for the machine-learning models were written in Python from the Scikit-learn framework. The equation for predicting the bioactivity of chemical compounds using multiple linear regression is presented in equation 29.

$$pEC_{50} = 1.036156 * \mathbf{SpMax8_Bhm} - 2.413685 * \mathbf{SpMin4_Bhp} - 2.140968 * \mathbf{CIC2} + 2.808991 * \mathbf{MDEC\text{-}22} + 4.8997081 \quad (29)$$

A range of values was used for each of the hyper-parameter of the support vector regression (SVR) model, which are:

1. Kernel function: Radial Basis Function (RBF), Polynomial (poly), Sigmoid and Linear
2. e-insensitive loss function parameter (ε): 0.001, 0.01,0.1 and 1
3. Gamma parameter of kernel function, g: 0.01, 0.001 and 0.0001
4. Penalty parameter, C: 4, 5 and 6

The optimal values of the hyper-parameters for SVR obtained using the Grid-search algorithm are:

Kernel function: Linear, ε = 0.1, g = 0.01 and C = 6.

The equation for predicting the bioactivity of molecules using the support vector regression model with grid-search hyper-parameter optimization (Grid-SVR model) is given in equation 30.

$$pEC_{50} = 1.10083009 * \mathbf{SpMax8_Bhm} - 1.72257249 * \mathbf{SpMin4_Bhp} - 1.35899954 * \mathbf{CIC2} + 1.35135594 * \mathbf{MDEC\text{-}22} + 4.93987096 \quad (30)$$

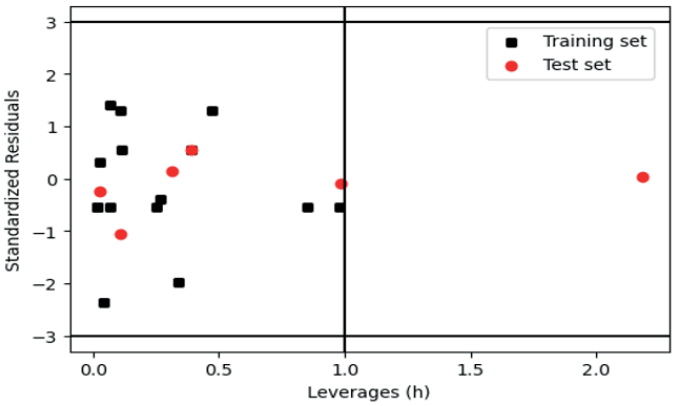


Figure 3: Williams Plot of standardized residuals vs. leverage with a warning leverage of h* = 1.0

The predicted bioactivity obtained from the Machine learning QSAR models were compared to the experimental bioactivity as presented in Table 3.

Graphs showing comparison of experimental and predicted bioactivity and Residual plots of the machine learning based QSAR models are presented in Figures 4 and 5.

External Validation of Machine learning based QSAR models

External validation of the machine learning-based QSAR models was done on the external (test) set through a performance comparison of the models with the baseline GFA-QSAR based on the standard metrics: correlation coefficient, coefficient of determination (predicted), mean absolute error, mean absolute percentage error, mean squared error and root mean squared error. The results are presented in Figures 6, 7, 8 and Table 4, respectively.

Internal Validation of Machine learning based QSAR models.

The Machine learning based QSAR models were validated internally using the leave-one-out-cross validation coefficient $\left(Q^2_{cv}\right)$ of the training set. Q^2_{cv} gives an idea of the predictive power of the QSAR model in predicting the bioactivity of a new compound. Values of Q^2_{cv} for the four machine-learning models are presented in Table 5 and the plot comparing these values with those of the baseline GFA-QSAR model is also presented in Figure 9.

The cross validation coefficient (Q^2_{cv}) results of the machine learning models show that all four were able to build valid models from the training set with $Q^2cv \geq 0.5$, although the random forest Regression model performed the poorest with $Q^2cv = 0.797$ (Table 5). However, the machine learning models, save for Grid-search based SVR, were unable to generalize properly on an external dataset that was not included in training the models (the test set). In other words, they suffered from *overfitting* problems. The Grid-search-based SVR model outperformed the baseline GFA-QSAR model in this regard - R^2 predicted = 0.9235 compared to R^2 predicted = 0.8792 for GFA. All three other machine learning models had R^2 predicted values ≤ 0.6. Regarding performance evaluation metrics of the QSAR models as presented in Figures 7, 8 and Table 4. The Grid-Search-based SVR models also outperformed other QSAR models with a performance enhancement of 43.1% over GFA model, 291.9% over MLR, 281.6% over RF and 314.3% over SVR, respectively, based on MAE evaluation criteria. Considering MAPE, Grid-search SVR recorded a performance improvement of 43.8% over GFA, 293.4% over MLR, 248.8% over RF and 315.7% over the SVR model, respectively. A similar trend was also observed in the RMSE values of the QSAR models with the Grid-search SVR model outperforming others. It is not by happenstance that the SVR model with grid-search optimization performed better than all others, especially in comparison with the normal SVR model that used the default hyper-parameter values and settings to build its

Table 3: Comparison of Experimental and Predicted Bioactivity using Machine learning models.

	Bioactivity of compounds				
		Predicted Bioactivity			
S/N	Experimental Bioactivity	MLR	RF	SVR	Grid-SVR
A1	*5.327902	5.803247	5.140861	4.68275	5.32016
A2	5.267606	5.07403	5.029921	5.006586	5.012286
A3	5.79588	5.835405	5.698683	5.69552	5.55748
A4	*5.769551	5.866569	5.680253	5.669284	5.669292
A5	4.39794	4.404238	5.140861	4.498131	4.497693
A6	5.769551	5.866569	5.680253	5.669284	5.669292
A7	5.075721	5.119062	4.992259	4.975696	5.172917
A8	*5.070581	5.028713	4.967078	4.965952	5.115567
A9	5.075721	5.119062	4.992259	4.975696	5.172917
A10	4.823909	4.671934	4.66512	4.724266	4.764941
A11	4.30103	4.63363	4.455968	4.710897	4.733229
A12	4.30103	4.28204	4.651209	4.401401	4.400915
A13	5.387216	5.146439	5.416273	5.47446	5.287953
A14	5.677781	5.638011	5.650647	5.650793	5.774943
A15	5.508638	5.699539	5.589685	5.598023	5.867789
A16	*5.853872	5.66089	5.660511	5.617467	5.826474
A17	5.744727	5.685171	5.677426	5.64446	5.817732
A18	*5.207608	5.48062	5.607622	5.549991	5.401324
A19	4.823909	4.933179	4.90473	4.963278	4.923793
A20	5.638272	5.48062	5.607622	5.549991	5.401324
A21	*5.267606	5.720233	4.747866	5.075904	5.28467

*test set

model. This demonstrates the effect of hyper-parameters optimization on the performance of the machine learning models, with an exponential improvement in their performance level.

Predicting the bioactivity of derivatized compounds.

The molecules derivatized using SAR were presented (Figures 10, 11 and 12). For instance, the sheathing of the amine in A7 enabled the introduction of a wide range of cycloalkyl amines (Figure 10, **B1–B9**). These analogues were incorporated to evaluate the influence of the secondary amine in A7 on the activity. Cyclic amines such as pyrrolidine, piperidine, and piperazine are valuable building blocks in the design of pharmacologically active molecules, and many approved drugs feature these motifs. Their inclusion often imparts pharmacokinetic benefits,[60,61] and their inherent conformational constraint can reduce susceptibility to metabolic enzymes. The inclusion of the primary amine B11 is to assess the contribution of this less hindered amine to the overall activity.

Similarly, the eastern phenyl ring of A1 was substituted with various bioisosteres to afford the **B12–B25** derivatives (Figure 11). The concept of bioisosterism is well established in medicinal chemistry and is frequently used to enhance potency or address physicochemical liabilities[62,63]. For example, previous studies have shown that cyclohexyl rings interact with aromatic amino-acid side chain residuesin proteins in a manner comparable to phenyl rings, thereby serving as effective non-aromatic bioisosteres[64]. Trimming and/or elongating the piperidyl ring gave new derivatives (Figure 12), others are presented in Figures S1-5. The descriptors of the derivatized compounds were generated, and the Grid-SVR model equation was used as the representative QSAR model to estimate their bioactivities, as it performed the best. The bioactivities of the newly designed B1-B102 are presented (Table S4).

Molecular docking

Lead compounds (A14, A16 and A17), (B76, B92, B94 and B101) and some inhibitors of SARS-CoV-2 main protease, M^{Pro} (nirmatrelvir/PF-07321332), spike glycoprotein/p-glp (CP-100356) and papain-like protease, PL^{Pro} (GRL0617) were docked at the binding pockets of M^{Pro}, PL^{Pro} and p-glp proteins. The co-ligand of the M^{Pro} was docked, for comparison purposes. The docking scores and post-docking MM/GBSA binding free energies of the complexes are presented in Table 6. The compounds' interactions at the active sites of the receptors are presented in (Figures S6-S30).

These molecules (A1-A21) were initially synthesized and screened against the spike-glycoprotein.[23] The conducted ML and SAR studies showed that A14, A16 and A17 from the synthesized compounds are lead compounds, while B76, B92, B94 and B101 are lead compounds from the modeled ones. To extend the anti-COVID-19 potentials of the compounds beyond inhibiting the spike protein, they were also investigated against the SARS-CoV-2 main protease and papain-like protease. The molecules exhibited notable binding affinities with all the receptors, with the best results observed with the spike glycoprotein receptor (Table 6). However, with M^{Pro} and PL^{Pro}, the compounds generally had lower docking scores and binding free energies than Nirmaltrelvir. Apart from A14, which did not exhibit visible interaction, all the compounds interacted with the two proteases *via* hydrogen bond and π-π stacking (supplementary material, Figures S6-S30). The compounds had better docking scores and binding free energy than the co-ligand (118569) accompanying the M^{pro}. With the spike glycoprotein, all compounds except B94 (-4.767 kcal/mol) showed lower docking scores than CP-100356 (-4.782 kcal/mol) with SP docking, while all the compounds had higher docking scores than CP-100356 (-3.810 kcal/mol) with XP docking. The compounds

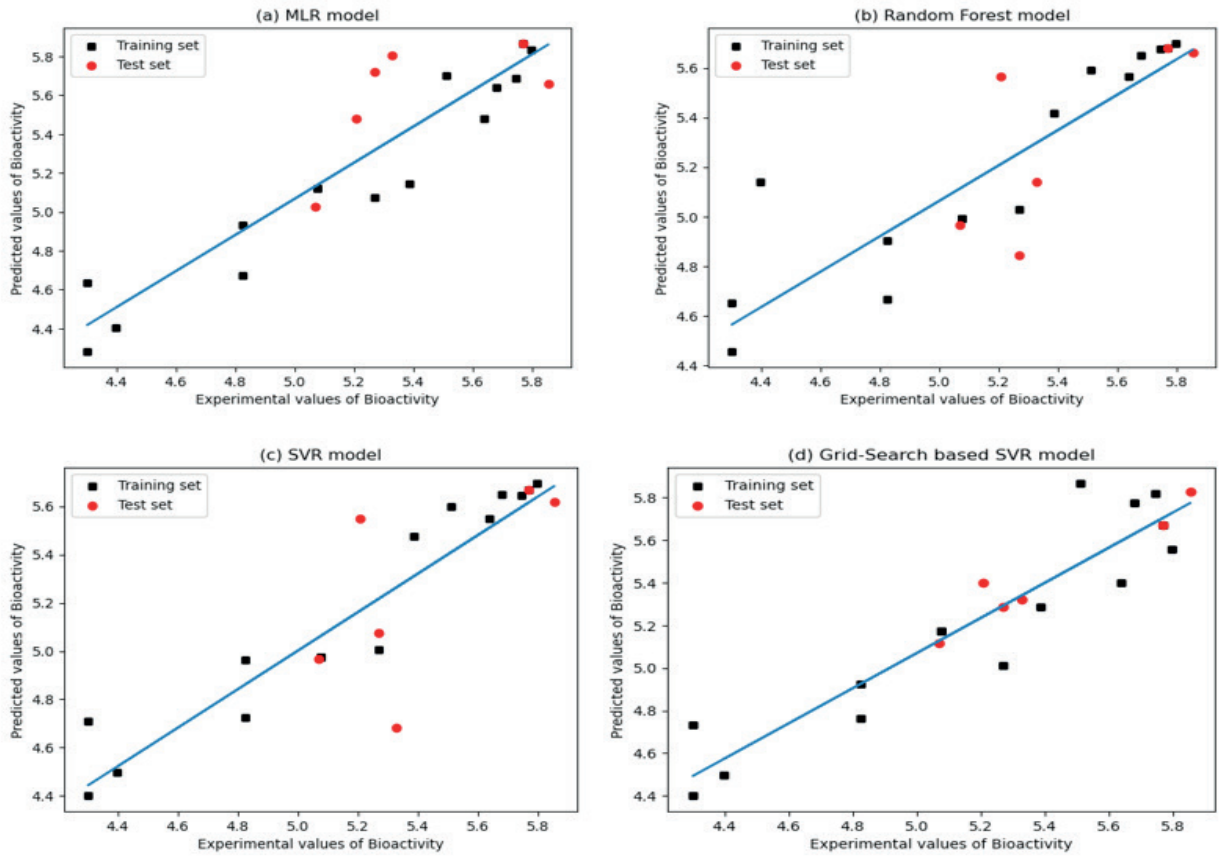


Figure 4: Experimental vs. Predicted Bioactivity of Machine learning QSAR models

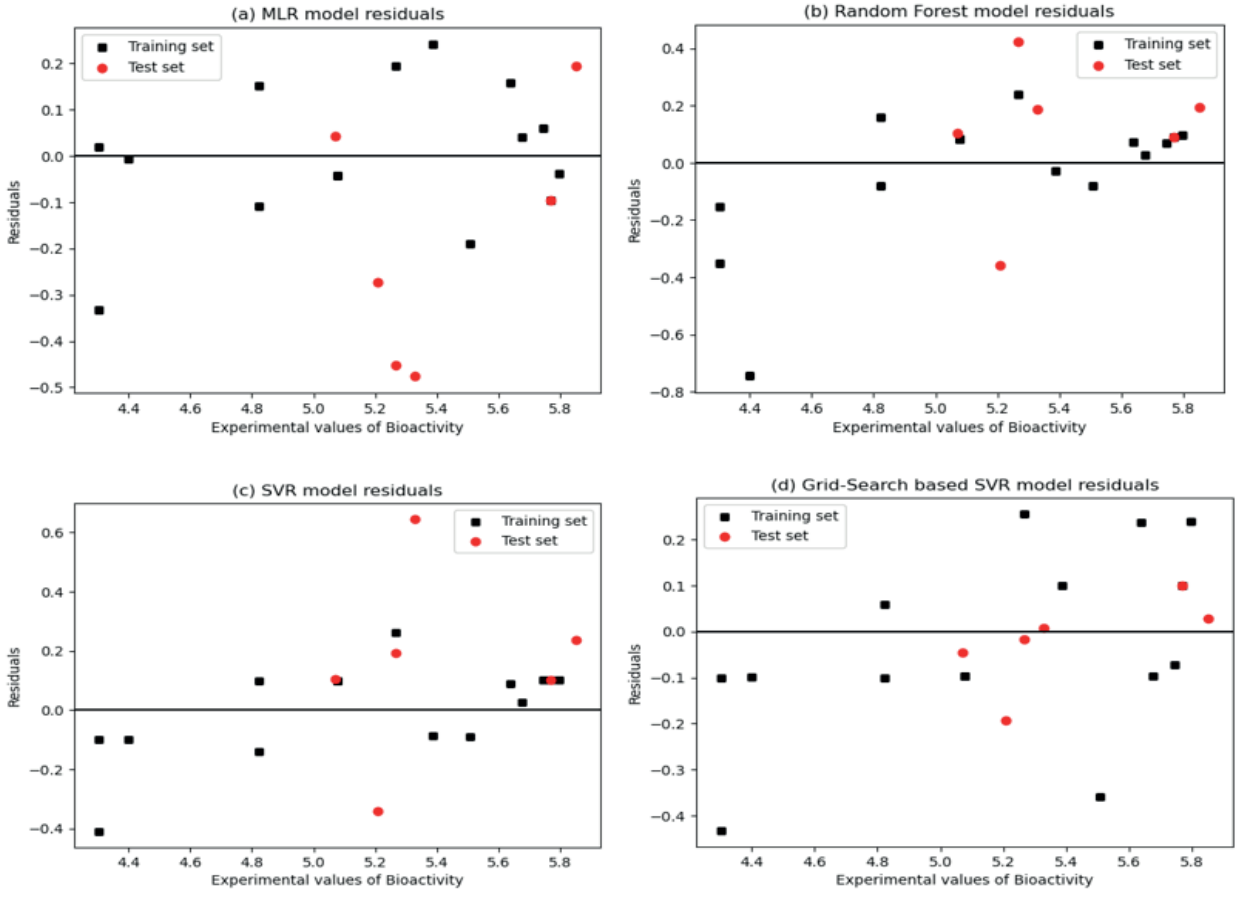


Figure 5: Residual Plots of Machine learning based QSAR models.

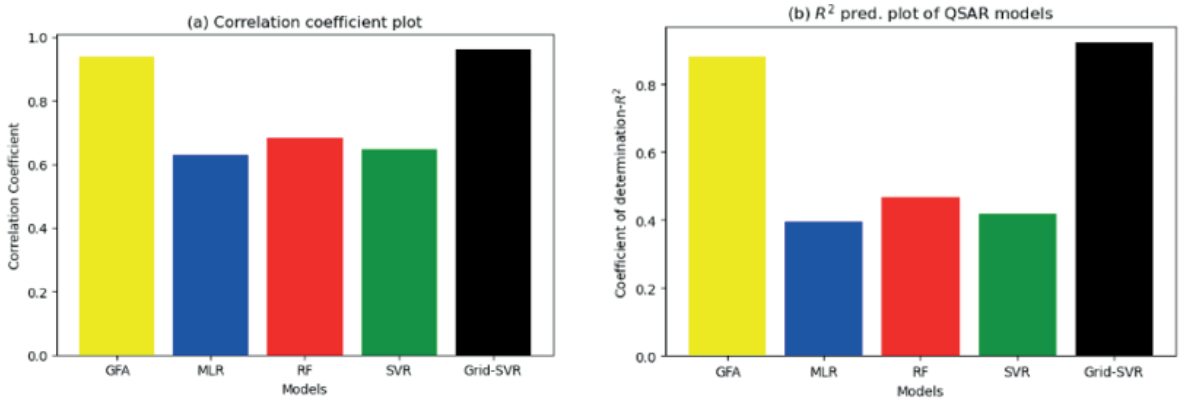


Figure 6: Performance comparison of Machine learning QSAR models based on R and R^2.

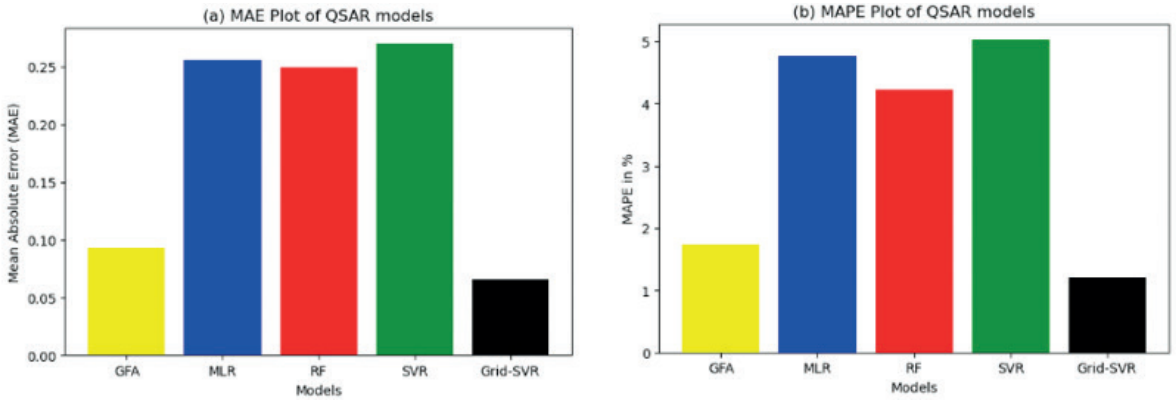


Figure 7: Performance comparison of Machine learning QSAR models based on MAE and MAPE.

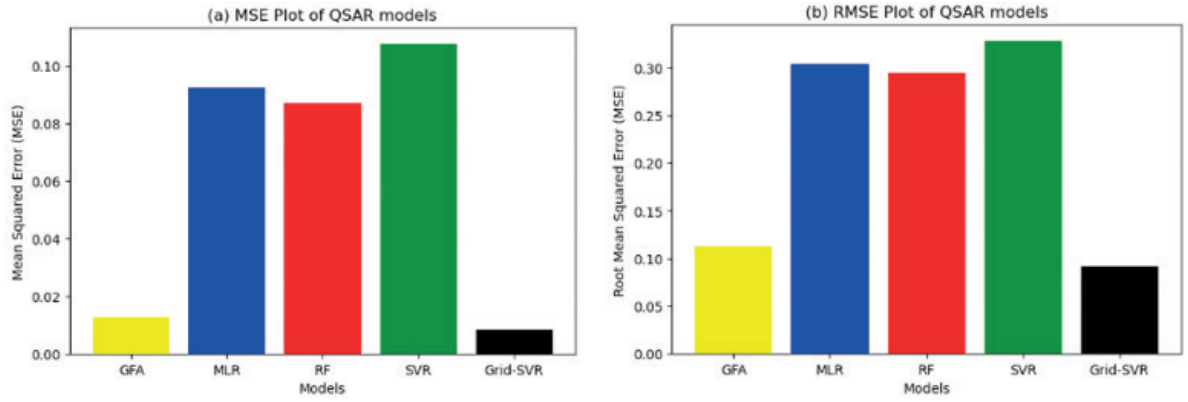


Figure 8: Performance comparison of Machine learning QSAR models based on MSE and RMSE.

Table 4: Table showing performance comparison of Machine learning QSAR models.

	MLR	RF	SVR	Grid-SVR
R	0.6277	0.6837	0.6457	0.961
R^2	0.394	0.467	0.417	0.924
MAE	0.2555	0.2488	0.2701	0.0652
MAPE (%)	4.76	4.22	5.03	1.21
MSE	0.0923	0.0869	0.1078	0.0085
RMSE	0.3038	0.2947	0.3284	0.0919

Table 5: Internal validation results of Machine learning based QSAR models.

	MLR	RF	SVR	Grid-SVR
Q^2	0.921	0.797	0.912	0.858

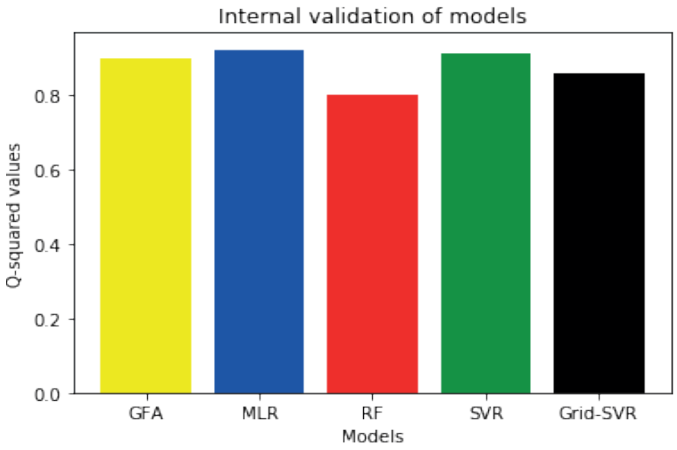


Figure 9: Graph of Q^2 values for machine-learning QSAR models.

also had higher binding free energies than CP-100356 (-48.53 kcal/mol) except B94 (-46.71 kcal/mol) and B76 (-46.53 kcal/mol). A14 (Figure S23) interacted with GLY339 *via* its oxadiazole nitrogen (hydrogen bonding). A16 and A17 (Figures S24 and S25) interacted with GLY339 *via* their pyrrolidine oxygen linking the phenyl ring (hydrogen bonding). This is because they are both isomers of each other. B76 and B92 (Figures S26 and S27) interacted with GLY339 *via* their oxadiazole nitrogen (hydrogen bonding). B94 (Figure S28) interacted with GLY339 *via* its phenoxy oxygen (hydrogen bonding) and ASN334 *via* its sulfoxide group (hydrogen bonding). B101 (Figure S29) interacted with ASN334 *via* its sulfoxide group (hydrogen bonding) while CP-100356 showed no visible interaction (Figure S30). The interactions of inhibitors of the SARS-CoV-2 spike protein with amino acid residues of the protein, such as GLY339 and ASN334, have been reported several times.[65-67]. Therefore, 6M0J-A16, 6M0J-A17, 6M0J-B92, 6M0J-B101 and 6M0J-CP-100356 complexes were processed for MD simulation.

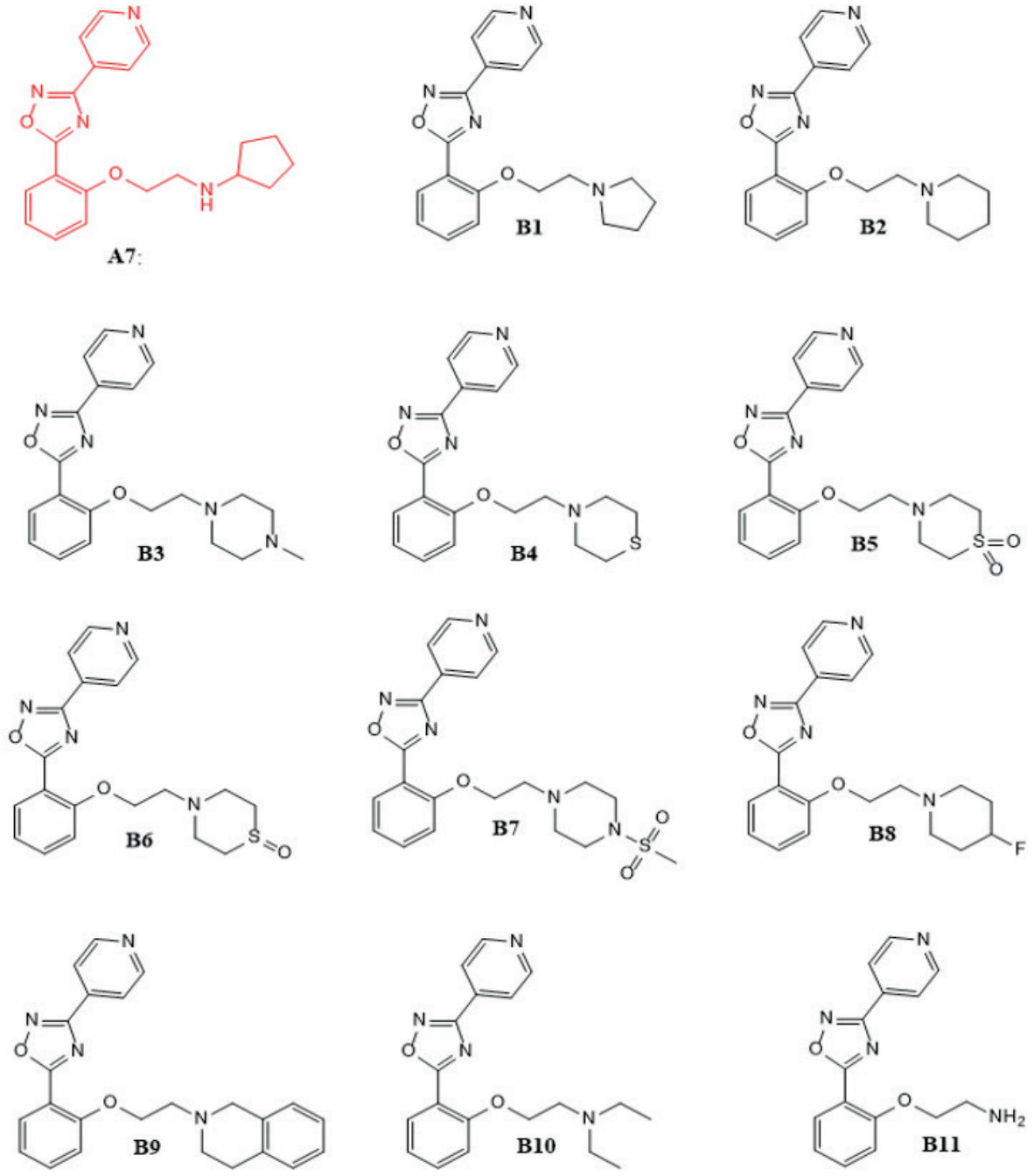


Figure 10: Derivatives from sheathing of the amine group.

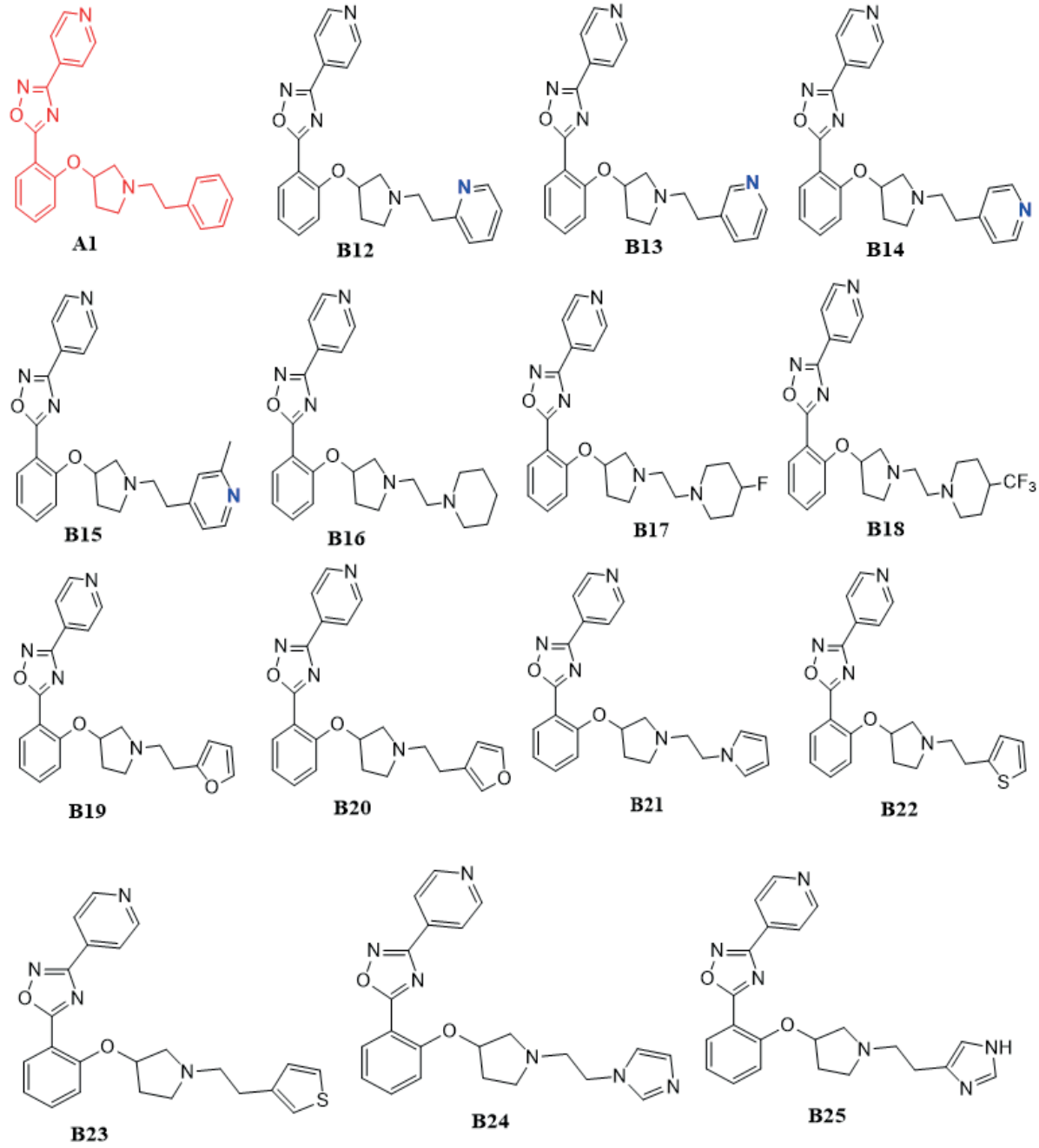


Figure 11: Derivatives from replacement of the eastern phenyl ring.

Stability and flexibility of the complexes

The stability and flexibility of the complexes were determined by calculating the RMSD and RMSF of the spike-glycoprotein c-α. The complexes of Spike-glycoprotein were selected because they possess higher docking scores and MM/GBSA binding free energies than the standard drug CP-100356, unlike those in the main protease and papain-like protease. The average RMSD, RMSF, RoG, SASA and number of hydrogen bonds values and their corresponding standard deviations are presented (Table 7), while their plots are presented in Figure 13.

The average RMSD were 2.28±0.60 Å, 2.32±0.49 Å, 1.88±0.41 Å, 1.49±0.24 Å, 1.78±0.24 Å and 1.57±0.19 Å for unbound protein, A16, A17, B92, B101 and CP-100356 complexes respectively. This is indicative that all the complexes, apart from A16, stabilized the protein[68], with B92 and B101 displaying lower deviations of the c-α atoms relative to the other complexes. The average RMSF were 1.44±1.01 Å, 1.30±0.98 Å, 1.09±0.66 Å, 1.10±0.50 Å, 1.04±0.53 Å and 1.01±0.44 Å for unbound protein, A16, A17, B92, B101 and CP-100356 complexes respectively. Higher RMSF values indicate higher flexibility of the complexes and vice-versa. The compounds reduced the flexibility of the domain upon binding, as the unbound protein displayed a higher RMSF value.[69]

The calculation of the solvent accessibility surface area (SASA) for the domain was performed to determine the surface area of the compound available for solvent interaction. The inner core of the protein is hydrophobic in nature, therefore providing insight into the folding and unfolding of the protein.[70] The average SASA values of the complexes were 10011.22±251.39 A^2, 9746.57±243.90 A^2, 9898.22±227.76 A^2, 9714.11±217.78 A^2, 9919.77±232.96 A^2 and 9777.47±234.97 A^2. B92 and unbound protein showed the least and highest available surface areas, respectively. Generally, the compounds induced folding of the protein binding domain, with B92 showing the highest folding, followed by A16, while B101 showed the least folding among the compounds. The compactness of the protein was investigated by

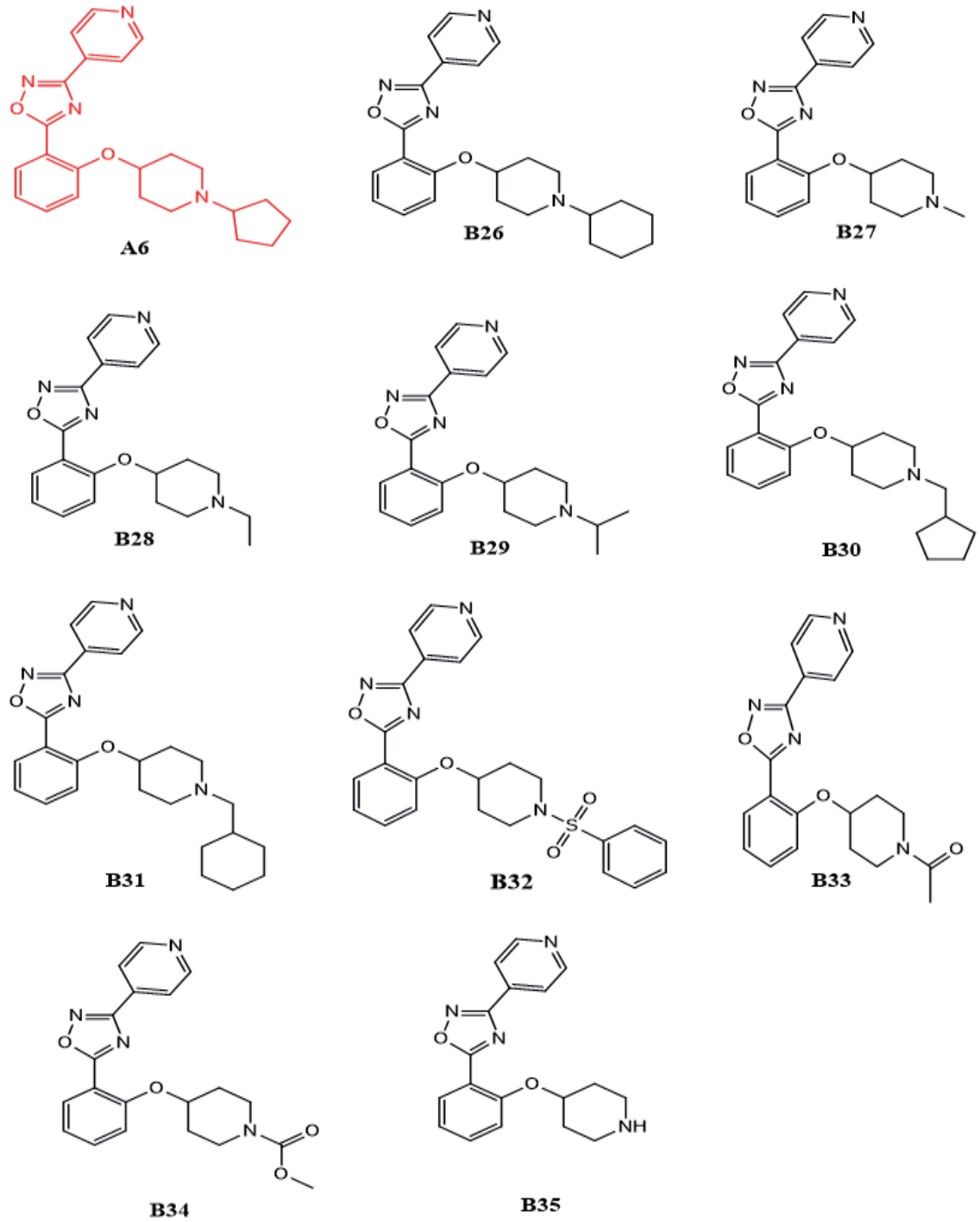


Figure 12: Derivatives from elongation and trimming of piperidyl ring.

computing the radius of gyration (RoG). High RoG is associated with less compactness and vice versa. The average RoG values of c-α atoms were 18.51±0.16, 18.45±0.17, 18.39±0.13, 18.40±0.10, 18.52±0.13 and 18.45±0.12. Generally, the compounds and unbound protein have almost similar average RoG values. Compared with the unbound protein, the compounds increased the number of hydrogen bonds in the protein matrix. Hydrogen bond interactions have been reported to play a major role in the binding affinities of ligands.[71]

Hydrogen bond interactions of protein-ligand complexes

The dynamic hydrogen bond occupancy of the complexes for the 200ns simulation was computed and presented (Table 8). The percentage occupancy recorded is the fraction of the simulation time occupied by the hydrogen bonds. The O2 atom of A16 and the ND2 of ASN343 displayed the highest occupancy of 9.07 % at 2.90 Å and an angle of 150.24° followed by the N4-A16 and OG of SER373 (4.21 %) at 2.83 Å and an angle of 160.35°. The O2 atom of A17 and the ND2 of ASN343 displayed the highest occupancy of 9.20 % at 2.90 Å and an angle of 147.75°, followed by the N3-A17 and ND2 of ASN343 (8.74 %) at 2.92 Å and an angle of 156.00°. The N4 atom of B92 and OG of SER371 displayed the highest occupancy of 4.80 % at 2.87 Å and angle of 157.88°, followed by the N3-B92 and OG of SER371 (4.24 %) at 2.84 Å and angle of 159.39°. The N4 atom of B101 and N of ASP364 displayed the highest occupancy of 20.04 % at 2.92 Å and angle of 158.98°, followed by the N3-B101 and OD1 of ASN343 (11.53 %) at 2.82 Å and angle of 156.67°. The N4 atom of CP-100356 and OG of SER371 displayed the highest occupancy of 4.80 % at 2.87 Å and angle of 157.88°, followed by the N3-CP-100356 and OG of SER371 (4.24 %) at 2.84 Å and angle of 159.39°.

Table 6: Docking scores and MM/GBSA energies of the molecules against SARS-CoV-2 receptors.

Compounds	SP docking score (kcal/mol)	XP docking score (kcal/mol)	MM/GBSA
MPro (5R7Y)			
A14	-7.52	-5.84	-40.21
A16	-6.42	-5.29	-23.83
A17	-5.84	-4.72	-37.51
B76	-5.69	-6.10	-35.35
B92	-6.35	-5.25	-44.46
B94	-6.52	-5.83	-43.16
B101	-6.43	-5.44	38.84
Nirmatrelvir	-6.84	-6.39	-57.82
118569	-5.13	-3.86	-27.78
PLPro (3E9S)			
A14	-5.12	-3.92	-47.48
A16	-4.32	-2.59	-32.81
A17	-5.77	-5.89	-51.40
B76	-7.27	-7.63	-57.59
B92	-4.27	-3.53	-48.71
B94	-5.59	-5.89	-61.40
B101	-6.32	-5.94	51.76
GRL0617	-8.28	-7.55	-68.88
Spike Glycoprotein (6M0J)			
A14	-4.83	-4.35	-50.93
A16	-4.81	-4.32	-52.77
A17	-4.82	-4.54	-54.27
B76	-5.39	-3.97	-46.53
B92	-5.74	-4.11	-54.69
B94	-4.77	-4.59	-46.71
B101	-5.71	-5.32	-48.18
CP-100356	-4.78	-3.81	-48.53

Table 7: Average RMSD, RMSF, radius of gyration (RoG), solvent assessable surface area (SASA), hydrogen bonds of the complexes and average RMSD and RMSF of the compounds, all with their corresponding standard deviations.

	RMSD (Å)	**RMSF** (Å)	**SASA** (A^2)	**RoG**	**HB**
Apo	2.28±0.60	1.44±1.01	10011.22±251.39	18.51±0.16	81.25±6.12
A16	2.32±0.49	1.30±0.98	9746.57±243.90	18.45±0.17	84.85±6.36
A17	1.88±0.41	1.09±0.66	9898.22±227.76	18.39±0.13	84.09±6.21
B92	1.49±0.24	1.10±0.50	9714.11±217.78	18.40±0.10	82.29±6.01
B101	1.78±0.24	1.04±0.53	9919.77±232.96	18.52±0.13	84.68±6.20
CP	1.57±0.19	1.01±0.44	9777.47±234.97	18.45±0.12	84.02±6.40

Post-MD MM-PBSA

Table 9 shows the MM-PBSA calculations, wherein the enthalpy contributions are considered while calculating the binding free energies (ΔG_{bind}) of the simulation trajectory. The energies represent the total free binding energies that characterize the complexing of the compounds.[57] These energies were averaged over 50,000 snapshots. The computed energies indicate that the control (CP-100356) showed the highest total free energy relative to the other compounds. The higher the free energy, the lesser the reaction kinetics and the more spontaneous the complexing.[57] Other energies contributing significantly to the ΔG include the electrostatic, van der Waals, and the gas phase.

CONCLUSION

This study employed a quantitative structure-activity relationship using machine learned QSAR modelling techniques to model the bioactivities of some 1-heteroaryl-2-alkoxyphenyl compounds obtained from the literature. Genetic function approximation was used as the baseline QSAR method to first perform feature selection of the most relevant molecular descriptors and to obtain a preliminary model. Four machine-learning-based QSAR models, namely, multi linear regression, random forest regression, support vector regression and grid search-based support vector regression (GS-SVR), were in turn adopted in developing predictive models of the compounds' bioactivities. The GS-SVR was found to outperform all other models in estimating the bioactivities, owing to the hyper-

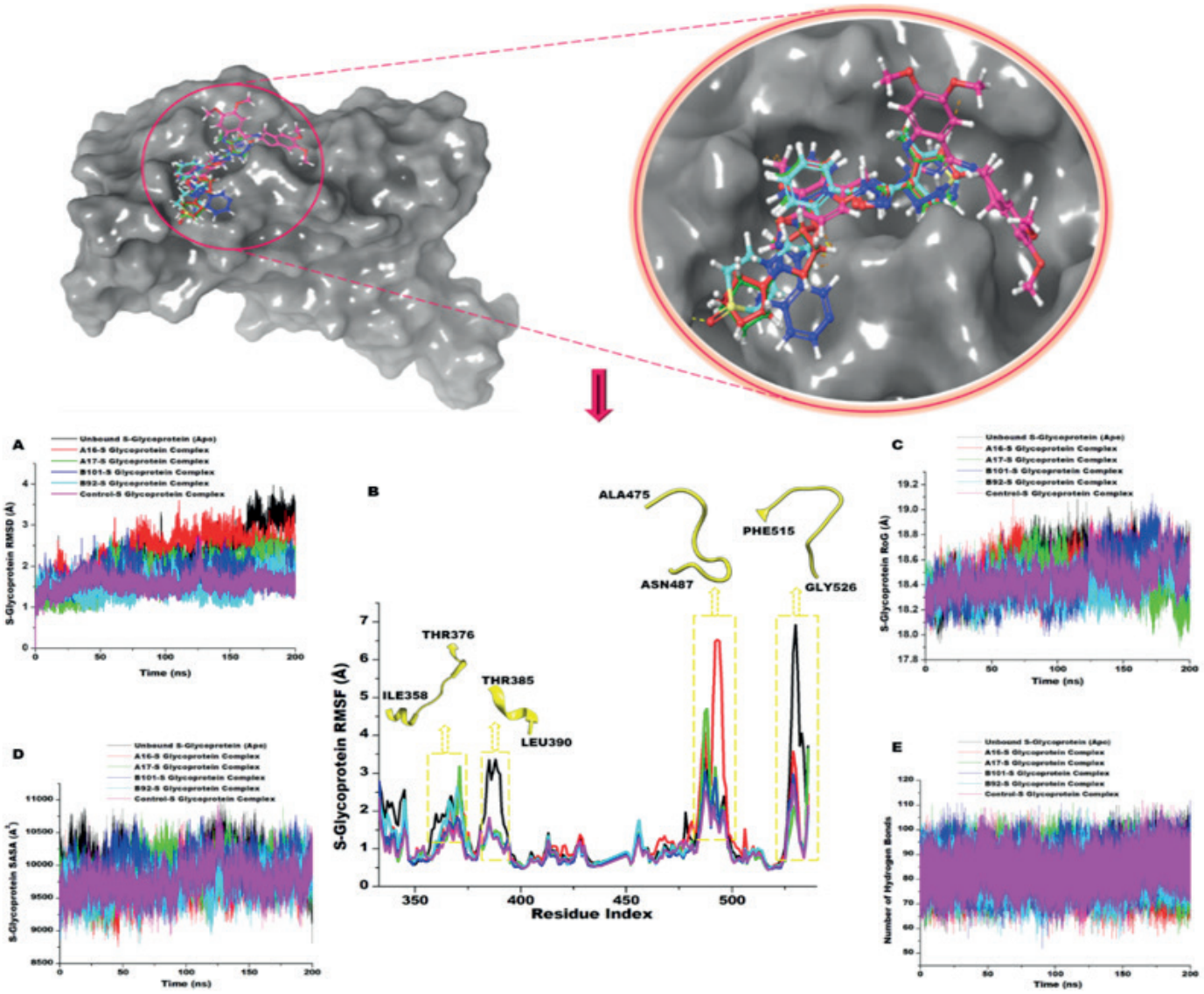


Figure 13: Comparative plots (a) RMSD (b) RMSF (c) RoG (d) SASA and (e) number of hydrogen bonds showing the dynamics of the complexes over the simulation periods. Inserts show the regions of the protein that display conspicuous fluctuations.

parameters optimization in the machine learning models. This is a result of the optimization of hyper-parameters of machine-learning models. A structural activity relationship study was performed on some molecules to generate new derivatives, resulting in 102 new compounds. The GS-SVR model was used in predicting the bioactivity of these new compounds. The compounds with the highest bioactivity were selected for molecular docking and dynamic simulations within the active site of SARS-CoV-2 receptors. The compounds showed the best results with a spike glycoprotein receptor, 6M0J, as they had higher docking scores and MM/GBSA binding free energies than CP-100356, a p-glycoprotein inhibitor. The compounds interacted with the protein's amino acid residues, especially with GLY339 and ASN334, which is consistent with the interactions of reported spike glycoprotein inhibitors. The average RMSD of the complexes and unbound protein was below 2.5 Å, indicating the stability of the compounds at the active site of the protein. The compounds reduced the flexibility of the domain upon binding, as the unbound protein displayed a higher RMSF value. From the solvent accessible surface area calculations, the compounds induced folding of the protein binding domain, with B92 showing the highest degree of folding. The computed energies indicate that all complexing were spontaneous in the systems. Investigating these compounds further would help in the discovery of more potent inhibitors of SARS-CoV-2 spike glycoprotein.

SUPPLEMENTARY MATERIALS

Supplementary material available online as a separate pdf.

ACKNOWLEDGEMENTS

Oyeneyin Oluwatoba Emmanuel acknowledges the Theoretical and Computational Chemistry Unit, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria and the School of Chemistry and Physics, University of KwaZulu-Natal, Westville campus, South Africa for creating enabling environments, funding and infrastructure to carry out part of this research work. Oyeneyin Oluwatoba Emmanuel acknowledges computational cluster resources at the Center for High Performance Computing (CHPC, www.chpc.ac.za), Cape Town, South Africa for molecular dynamic simulations. Oyeneyin Oluwatoba Emmanuel appreciates the funding received from the Tertiary Education Trust Fund (TETFUND), Nigeria with award number TETF/ES/UNI/ONDO/TSAS/2021. Matshawandile Tukulula acknowledges the South African National Research Foundation (NRF) and UKZN for financial assistance.

Table 8: Hydrogen bond occupancy, distance and angle formed between the host residues and the compounds (ligands).

Compounds	H-Acceptor	H-Donor	Occupancy (%)	Distance (Å)	Angle (°)
A16	A16-O2	ASN343-ND2	9.07	2.90	150.24
	A16-N4	SER373-OG	4.21	2.83	160.35
	A16-O1	ASN343-ND2	2.81	2.89	153.34
	A16-N3	ASN343-ND2	2.73	2.93	155.83
	A16-N2	SER371-OG	0.13	2.85	157.33
A17	A17-O2	ASN343-ND2	9.20	2.90	147.75
	A17-N3	ASN343-ND2	8.74	2.92	156.00
	A17-O1	ASN343-ND2	0.79	2.89	153.96
	A17-N4	SER373-OG	0.41	2.84	160.41
	A17-O2	ASN343-ND2	0.39	2.89	149.91
	A17-N3	ASN343-ND2	0.14	2.93	158.07
B92	SER371-OG	B92-N4	4.80	2.87	157.88
	B92-N3	SER371-OG	4.24	2.84	159.39
	ASN343-OD1	B92-N4	2.16	2.81	159.96
	B92-O1	ASN343-ND2	1.74	2.89	160.10
	B92-N3	SER373-OG	0.63	2.83	161.13
	SER371-O	B92-N4	0.33	2.91	151.22
	B92-N2	ASN343-ND2	0.29	2.92	153.54
	B92-N1	GLY7-N	0.23	2.92	153.23
	B92-N1	ASN343-ND2	0.17	2.92	158.00
B101	B101-N4	ASP364-N	20.04	2.92	158.98
	ASN343-OD1	B101-N3	11.53	2.82	156.67
	B101-O3	GLY339-N	7.64	2.87	150.59
	B101-O1	ASN343-ND2	6.14	2.88	158.23
	B101-O2	PHE338-N	1.12	2.87	156.06
	B101-N2	ASN343-ND2	0.96	2.91	157.64
	GLU340-OE1	B101-N3	0.68	2.76	163.36
	B101-O2	ASN343-ND2	0.21	2.92	153.48
	B101-O3	ASN334-ND2	0.15	2.88	158.91
	B101-O3	ASN370-ND2	0.14	2.86	157.65
	B101-N2	GLY339-N	0.07	2.94	151.68
CP-100356	SER371-OG	CP-N4	4.80	2.87	157.88
	CP-N3	SER371-OG	4.24	2.84	159.39
	ASN343-OD1	CP-N4	2.16	2.81	159.96
	CP-O1	ASN343-ND2	1.74	2.89	160.10
	CP-N3	SER373-OG	0.63	2.83	161.13
	SER371-O	CP-N4	0.33	2.91	151.22
	CP-N2	ASN343-ND2	0.29	2.92	153.54
	CP-N1	GLY339	0.23	2.92	153.23
	CP-N1	ASN343-ND2	0.17	2.92	158.00

Table 9: MMPBSA energy profiles of the complexes averaged over 50,000 snapshots

System	Energy components (Kcal/mol)				
	ΔE_{vdw}	ΔE_{ele}	ΔG_{gas}	ΔG_{sol}	ΔG_{bind}
A16	-41.22±3.87	-10.68±8.58	-51.90±10.70	18.60±9.04	-33.30±3.56
A17	-33.42±2.98	1.26±9.05	-32.15±10.13	4.40±9.02	-27.75±2.85
B92	-33.32±5.26	-6.97±10.25	-26.35±10.38	1.00±9.25	-25.34±5.06
B101	-41.11±4.56	-4.31±15.37	-45.43±17.07	-7.71±13.25	-37.71±6.56
CP-100356	-63.01±5.01	-9.93±4.62	-72.95±6.06	-25.57±4.46	-47.37±4.46

FUNDING

This work was supported by the Tertiary Education Trust Fund (TETFUND), Nigeria with award number TETF/ES/UNI/ONDO/TSAS/2021, received in 2021 by O.E.O., the South African National Research Foundation (NRF CSUR: 116282) by M.T and the University of KwaZulu-Natal (UKZN) for funding.

AUTHOR CONTRIBUTIONS

Conceptualization: O.E.O., S.M.O.; Methodology: O.E.O., S.M.O., O.O.O., D.A.O., N.I., B.L.I., A.R.I.; Investigation: O.E.O., S.M.O., D.A.O., O.O.O., A.R.I.; Visualization: O.E.O., S.M.O., O.O.O., A.R.I., E.F.K.; Supervision: O.E.O., M.T.; Writing – original draft: O.E.O., S.M.O., A.R.I., N.I., B.L.I.; Writing – review & editing: O.E.O., S.M.O., O.O.O., D.A.O., N.I., B.L.I., A.R.I., E.F.K., M.T. All authors have read and agreed to the published version of the manuscript.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The data are presented in the supplementary information.

CONFLICTS OF INTEREST

The other authors declare no conflicts of interest.

REFERENCES

1. Jamison DA Jr, Anand Narayanan S, Trovão NS, Guarnieri JW, Topper MJ, Moraes-Vieira PM, Zaksas V, Singh KK, Wurtele ES, Beheshti A. A comprehensive SARS-CoV-2 and COVID-19 review, Part 1: intracellular overdrive for SARS-CoV-2 infection. Eur J Hum Genet. 2022;30(8):889–898. https://doi.org/10.1038/s41431-022-01108-8.
2. Cortese M, Lee J, Cerikan B, Neufeldt CJ, Oorschot VMJ, Köhrer S, Hennies J, Schieber NL, Ronchi P, Mizzon G, et al. Integrative imaging reveals SARS-CoV-2-induced reshaping of subcellular morphologies. Cell Host Microbe. 2020;28(6):853–866.e5. https://doi.org/10.1016/j.chom.2020.11.003.
3. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, Katzourakis A. The evolution of SARS-CoV-2. Nat Rev Microbiol. 2023;21(6):361–379. https://doi.org/10.1038/s41579-023-00878-2.
4. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17(3):181–192. https://doi.org/10.1038/s41579-018-0118-9.
5. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol. 2021;19(3):155–170. https://doi.org/10.1038/s41579-020-00468-6.
6. Emrani J, Ahmed M, Jeffers-Francis L, Teleha JC, Mowa N, Newman RH, Thomas MD. SARS-COV-2, infection, transmission, transcription, translation, proteins, and treatment: A review. Int J Biol Macromol. 2021;193:1249–1273. https://doi.org/10.1016/j.ijbiomac.2021.10.172.
7. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science. 2020;367(6485):1444–1448. https://doi.org/10.1126/science.abb2762.
8. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C, Abiona O, Graham BS, McLellan JS (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science (80) 367:1260–1263. https://doi.org/10.1126/science.abb2507.
9. Hu Y, Wen J, Tang L, Zhang H, Zhang X, Li Y, Wang J, Han Y, Li G, Shi J, et al. The M protein of SARS-CoV: basic structural and immunological properties. Genomics Proteomics Bioinformatics. 2003;1(2):118–130. https://doi.org/10.1016/S1672-0229(03)01016-7.
10. de Haan CA, Vennema H, Rottier PJ. Assembly of the coronavirus envelope: homotypic interactions between the M proteins. J Virol. 2000;74(11):4967–4978. https://doi.org/10.1128/JVI.74.11.4967-4978.2000.
11. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–220. https://doi.org/10.1038/s41586-020-2180-5.
12. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. Antiviral Res. 2018;149:58–74. https://doi.org/10.1016/j.antiviral.2017.11.001.
13. Claverie J-M. A putative role of de-mono-ADP-ribosylation of STAT1 by the SARS-CoV-2 Nsp3 protein in the cytokine storm syndrome of COVID-19. Viruses. 2020;12(6):646. https://doi.org/10.3390/v12060646.
14. Liu Y, Li D, Han J. COVID-19 vaccines and beyond. Cell Mol Immunol. 2024;21(3):207–209. https://doi.org/10.1038/s41423-024-01132-2.
15. Brice Y, Morgan L, Kirmani M, Kirmani M, Udeh MC. COVID-19 Vaccine evolution and beyond. Neurosci Insights. 2023;18: 26331055231180543. https://doi.org/10.1177/26331055231180543.
16. Choudhary V, Gupta A, Sharma R, Parmar HS (2021) Therapeutically effective covalent spike protein inhibitors in treatment of SARS-CoV-2. J Prot Proteom. 12:257–270. https://doi.org/10.1007/s42485-021-00074-x.
17. Huff S, Kummetha IR, Tiwari SK, Huante MB, Clark AE, Wang S, Bray W, Smith D, Carlin AF, Endsley M, et al. Discovery and mechanism of SARS-CoV-2 main protease inhibitors. J Med Chem. 2022;65(4):2866–2879. https://doi.org/10.1021/acs.jmedchem.1c00566.
18. Shan H, Liu J, Shen J, Dai J, Xu G, Lu K, Han C, Wang Y, Xu X, Tong Y, et al. Development of potent and selective inhibitors targeting the papain-like protease of SARS-CoV-2. Cell Chem Biol. 2021;28(6):855–865.e9. https://doi.org/10.1016/j.chembiol.2021.04.020.
19. Xiu S, Dick A, Ju H, Mirzaie S, Abdi F, Cocklin S, Zhan P, Liu X. Inhibitors of SARS-CoV-2 entry: current and future opportunities. J Med Chem. 2020;63(21):12256–12274. https://doi.org/10.1021/acs.jmedchem.0c00502.
20. Rudrapal M, Issahaku AR, Agoni C, Bendale AR, Nagar A, Soliman MES, Lokwani D. In silico screening of phytopolyphenolics for the identification of bioactive compounds as novel protease inhibitors effective against SARS-CoV-2. J Biomol Struct Dyn. 2022;40(20):10437–10453. https://doi.org/10.1080/07391102.2021.1944909.
21. Gawriljuk VO, Noske GD, Fernandes RS, Nakamura AM, Freire MCLC, Godoy MO, Bonatto V, Chelucci R, Andricopulo A, Bakowski MA, et al. Screening of medicines for malaria venture open boxes identifies potent SARS-CoV2 papain like protease (PLpro) inhibitors. ACS Omega. 2025;10(37):42711–42724. https://doi.org/10.1021/acsomega.5c04858.
22. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616(7958):673–685. https://doi.org/10.1038/s41586-023-05905-z.
23. Vangeel L, Koukni M, Arzel P, Zwaagstra M, Lyoo H, Wanningen P, Ahmad S, Zhang L, Sun X, Delpal A, et al. Profiling of 1-heteroaryl-2-alkoxyphenyl analogs as inhibitors of SARS-CoV-2 replication. Molecules. 2022;27:1–26. https://doi.org/https://doi.org/10.3390/molecules27031052.
24. Yap CW. PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–1474. https://doi.org/DOI10.1002/jcc.21707. https://doi.org/10.1002/jcc.21707.
25. Choi JS, Trinh TX, Yoon TH, Kim J, Byun HG. Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. Chemosphere. 2019;217:243–249. https://doi.org/10.1016/j.chemosphere.2018.11.014.
26. Roy K. Quantitative-structure-activity relationships (QSAR): A few validation methods and software tools developed at the DTC laboratory. J Indian Chem Soc. 2018;95:1497–1502.
27. Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics. 1969;11(1):137–148. https://doi.org/10.1080/00401706.1969.10490666.
28. Accelrys. Materials Studio; 2016. http://accelrys.com/products/collaborative-science/biovia-materials-studio/.
29. Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci. 1993;33:854–866.
30. Friedman JH. Multivariate Adaptive Regression Splines. Ann Stat. 1991;19:1–141.
31. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci. 2003;22(1):69–77. https://doi.org/10.1002/qsar.200390007.
32. Gajewicz A. How to judge whether QSAR/read-across predictions can be trusted: A novel approach for establishing a model's applicability domain. Environ Sci Nano. 2018;5(2):408–421. https://doi.org/10.1039/C7EN00774D.
33. Daniya T, Geetha M, Kumar BS, Cristin R. Least Square Estimation of Parameters for Linear Regression. Int J Control Autom. 2020; 13(2):447-452.
34. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
35. Smola AJ, Scholkopf B. A tutorial on support vector regression*. Stat Comput. 14:199–222. https://doi.org.10.1023/B:STCO.0000035301.49549.88
36. Ma X, Zhang Y, Wang Y. Performance evaluation of kernel functions based on grid search for support vector regression. In: 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics Automation and Machatronics (RAM). Angkor Wat, Cambodia. IEEE; 2015;283–288. https://doi.org/10.1109/ICCIS.2015.7274635.
37. Hastie T, Tibshirani R, James G, Witten D. An introduction to statistical learning: with applications in R. (2nd ed). Springer Texts in Statistics; 102. New York: Springer; 2021. https://doi.org/10.1007/978-1-0716-1418-1.
38. Polikar R. Ensemble based systems in decision making. IEEE Circuit Syst. Mag. 2006;6:21–45. https://doi.org/10.1109/MCAS.2006.1688199.

39. Biau G, Scornet E. A random forest guided tour. Test. 2016;25(2):197–227. https://doi.org/10.1007/s11749-016-0481-7.
40. Cutler A, Cutler DR, Stevens JR. Random Forests. In: Zhang C, Ma Y (editors), Ensemble Machine Learning. New York: Springer. pp 157–175. https://doi.org/10.1007/978-1-4419-9326-7.
41. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995. https://link.springer.com/book/10.1007/978-1-4757-2440-0.
42. Vapnik V, Golowich SE, Smola AJ. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer MC, Jordan MI, Petsche T, editors. Advances in Neural Information Processing Systems 9 (NIPS 1996); 1996 Dec 2–5; Denver, CO. Cambridge, MA: MIT Press; 1997. p. 281–287.
43. Zakrani A. (2018) Support Vector Regression Based on Grid-Search Method for Agile Software Effort Prediction. 2018 IEEE 5th Int Congr Inf Sci Technol. pp 1–6. https://doi.org/10.1109/CIST.2018.8596370.
44. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20:273–297. https://doi.org/10.1023/A:1022627411411.
45. Shim J, Seok KH, Hwang C, Cho D. Support Vector Quantile Regression Using Asymmetric e-Insensitive Loss Function. Commun Kor Stat Soc. 2011;18(2):165–170. https://doi.org/10.5351/CKSS.2011.18.2.165.
46. Oliveira ALI, Braga PL, Lima RMF, Cornélio ML. GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. Inf Softw Technol. 2010;52(11):1155–1166. https://doi.org/10.1016/j.infsof.2010.05.009.
47. Huang S, Tian L, Zhang J, Chai X, Wang H, Zhang H. Support vector regression based on the particle swarm optimization algorithm for tight oil recovery prediction. ACS Omega. 2021;6(47):32142–32150. https://doi.org/10.1021/acsomega.1c04923.
48. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem. 2006;49:6177–6196.
49. Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des. 2013;27(3):221–234. https://doi.org/10.1007/s10822-013-9644-8.
50. Lu C, Wu C, Ghoreishi D, Chen W, Wang L, Damm W, Ross GA, Dahlgren MK, Russell E, Von Bargen CD, et al. OPLS4: improving force field accuracy on challenging regimes of chemical space. J Chem Theory Comput. 2021;17(7):4291–4300. https://doi.org/10.1021/acs.jctc.1c00302.
51. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Gilson MK, et al. AMBER 2018. University of California, San Francisco; 2018.
52. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput. 2015;11(8):3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.
53. Sprenger KG, Jaeger VW, Pfaendtner J. The General AMBER Force Field (GAFF) can accurately predict thermodynamic and transport properties of many ionic liquids. J Phys Chem B. 2015;119(18):5882–5895. https://doi.org/10.1021/acs.jpcb.5b00689.
54. Larini L, Mannella R, Leporini D. Langevin stabilization of molecular-dynamics simulations of polymers by means of quasisymplectic algorithms. J Chem Phys. 2007;126(10):104101. https://doi.org/10.1063/1.2464095.
55. Roe DR, Cheatham TE 3rd. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. J Chem Theory Comput. 2013;9(7):3084–3095. https://doi.org/10.1021/ct400341p.
56. Miller BR 3rd, McGee TD Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA.py: an efficient program for end-state free energy calculations. J Chem Theory Comput. 2012;8(9):3314–3321. https://doi.org/10.1021/ct300418h.
57. Issahaku AR, Aljoundi A, Soliman MES. Establishing the mutational effect on the binding susceptibility of AMG510 to KRAS switch II binding pocket: computational insights. Inform Med Unlocked. 2022;30:100952. https://doi.org/10.1016/j.imu.2022.100952.
58. Oyeneyin OE, Obadawo BS, Metibemu DS, Owolabi TO, Olanrewaju AA, Orimoloye SM, Ipinloju N, Olubosede O. An exploration of the antiproliferative potential of chalcones and dihydropyrazole derivatives in prostate cancer via androgen receptor: combined QSAR, machine learning, and molecular docking techniques. Phys Chem Res. 2022;10:211–223. https://doi.org/10.22036/pcr.2021.293051.1932.
59. Gramatica P. Principles of QSAR models validation: internal and external. 2007;26(5):694–701. https://doi.org/10.1002/qsar.200610151.
60. Kalgutkar AS, Driscoll JP. Is there enough evidence to classify cycloalkyl amine substituents as structural alerts? Biochem Pharmacol. 2020;174:113796. https://doi.org/10.1016/j.bcp.2020.113796.
61. Grygorenko OO, Melnykov KP, Holovach S, Demchuk O. Fluorinated cycloalkyl building blocks for drug discovery. ChemMedChem. 2022;17(21):e202200365. https://doi.org/10.1002/cmdc.202200365.
62. Meanwell NA. The Influence of Bioisosteres in Drug Design: Tactical Applications to Address Developability Problems. In: Meanwell N, editor. Tactics in Contemporary Drug Design. Topics in Medicinal Chemistry, 9. Berlin, Heidelberg: Springer; 2013. https://doi.org/10.1007/7355_2013_29.
63. Meanwell NA. Applications of bioisosteres in the design of biologically active compounds. J Agric Food Chem. 2023;71(47):18087–18122. https://doi.org/10.1021/acs.jafc.3c00765.
64. Subbaiah MAM, Meanwell NA. Bioisosteres of the phenyl ring: recent strategic applications in lead optimization and drug design. J Med Chem. 2021;64(19):14046–14128. https://doi.org/10.1021/acs.jmedchem.1c01215.
65. Biswas S, Mahmud S, Mita MA, Afrose S, Hasan MR, Sultana Shimu MS, Saleh MA, Mostafa-Hedeab G, Alqarni M, Obaidullah AJ, et al. Molecular docking and dynamics studies to explore effective inhibitory peptides against the spike receptor binding domain of SARS-CoV-2. Front Mol Biosci. 2022;8:791642. https://doi.org/10.3389/fmolb.2021.791642.
66. Lin Y-C, Chen W-Y, Hwu E-T, Hu W-P. In-Silico selection of aptamer targeting SARS-CoV-2 spike protein. Int J Mol Sci. 2022;23(10):5810. https://doi.org/10.3390/ijms23105810.
67. Nedyalkova M, Vasighi M, Sappati S, Kumar A, Madurga S, Simeonov V. Inhibition ability of natural compounds on receptor-binding domain of SARS-CoV2: an in silico approach. Pharmaceuticals (Basel). 2021;14(12):1328. https://doi.org/10.3390/ph14121328.
68. Issahaku AR, Salifu EY, Agoni C, Alahmdi I, Abo-Dya NE, Soliman MES, Rudrapal M, Podila N. Discovery of potential KRAS-SOS1 inhibitors from South African natural compounds: an in silico approach. ChemistrySelect. 2023;8(24):e202300277. https://doi.org/10.1002/slct.202300277.
69. Issahaku AR, Salifu EY, Soliman MES. Inside the cracked kernel: establishing the molecular basis of AMG510 and MRTX849 in destabilising KRASG12C mutant switch I and II in cancer treatment. J Biomol Struct Dyn. 2022;41:4890–4902. https://doi.org/10.1080/07391102.2022.2074141.
70. Issahaku AR, Ibrahim MAA, Mukelabai N, Soliman MES. Intermolecular and dynamic investigation of the mechanism of action of reldesemtiv on fast skeletal muscle troponin complex toward the treatment of impaired muscle function. Protein J. 2023;42(4):263–275. https://doi.org/10.1007/s10930-023-10091-y.
71. Oyeneyin OE, Ibrahim A, Ipinloju N, Ademoyegun AJ, Ojo ND. Insight into the corrosion inhibiting potential and anticancer activity of 1-(4-methoxyphenyl)-5-methyl-N'-(2-oxoindolin-3-ylidene)-1H-1,2,3-triazole-4-carbohydrazide via computational approaches. J Biomol Struct Dyn. 2024;42(20):11149–11166. https://doi.org/10.1080/07391102.2023.2260491.