BMC Chemistry

RESEARCH **Open Access**

Prospective virtual screening combined with bio-molecular simulation enabled identification of new inhibitors for the KRAS drug target

Amar Ajmal[1], Hind A Alkhatabi[2], Roaa M. Alreemi[2], Mubarak A. Alamri[3], Asaad Khalid[4*], Ashraf N. Abdalla[5], Bader S. Alotaibi[6] and Abdul Wadood[1*]

Abstract
Lung cancer is a disease with a high mortality rate and it is the number one cause of cancer death globally. Approximately 12–14% of non-small cell lung cancers are caused by mutations in KRAS^{G12C}. The KRAS^{G12C} is one of the most prevalent mutants in lung cancer patients. KRAS was first considered undruggable. The sotorasib and adagrasib are the recently approved drugs that selectively target KRAS^{G12C}, and offer new treatment approaches to enhance patient outcomes however drug resistance frequently arises. Drug development is a challenging, expensive, and time-consuming process. Recently, machine-learning-based virtual screening are used for the development of new drugs. In this study, we performed machine-learning-based virtual screening followed by molecular docking, all atoms molecular dynamics simulation, and binding energy calculations for the identifications of new inhibitors against the KRAS^{G12C} mutant. In this study, four machine learning models including, random forest, k-nearest neighbors, Gaussian naïve Bayes, and support vector machine were used. By using an external dataset and 5-fold cross-validation, the developed models were validated. Among all the models the performance of the random forest (RF) model was best on the train/test dataset and external dataset. The random forest model was further used for the virtual screening of the ZINC15 database, in-house database, Pakistani phytochemicals, and South African Natural Products database. A total of 100 ns MD simulation was performed for the four best docking score complexes as well as the standard compound in complex with KRAS^{G12C}. Furthermore, the top four hits revealed greater stability and greater binding affinities for KRAS^{G12C} compared to the standard drug. These new hits have the potential to inhibit KRAS^{G12C} and may help to prevent KRAS-associated lung cancer. All the datasets used in this study can be freely available at (https://github.com/Amar-Ajmal/Datasets-for-KRAS).

Keywords KRAS, Machine-learning, External validation, Molecular docking, Molecular dynamics simulation

*Correspondence:
Asaad Khalid
akahmed@jazanu.edu.sa
Abdul Wadood
awadood@awkum.edu.pk
[1] Department of Biochemistry, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan
[2] Department of Biochemistry, College of Science, University of Jeddah, Jeddah, 21959, Saudi Arabia
[3] Department of Pharmaceutical Chemistry, College of Pharmacy, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[4] Substance Abuse and Toxicology Research Center, Jazan University, P.O. Box: 114, Jazan 45142, Saudi Arabia
[5] Department of Pharmacology and Toxicology, College of Pharmacy, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[6] Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Shaqra Univesity, Al- Quwayiyah, Riyadh, Saudi Arabia

Introduction

The most frequently mutated gene family in human malignancies is RAS including KRAS, NRAS, and HRAS. KRAS is one of the most prevalent isoforms among the RAS family, being found in 85% of pancreatic, 45% of colo-rectal, and 30% of lung cancer [1]. KRAS is a member of the broad family of proteins known as GTPases. KRAS is a plasma membrane-bound protein. KRAS functions as a molecular switch for downstream signal transduction by cycling between the on (GTP) and off (GDP) states [2]. Every year, about 50,000 people in the United States alone receive a new diagnosis of lung cancer caused by KRAS mutations [3]. Moreover, a history of smoking is frequently related to KRAS-driven lung malignancies [4]. The KRAS gene has six exons and is found on chromosome 12p11.1–12.2. KRAS is a small protein that is 21 KDa in size. The two domains of KRAS G and C domains, which are made up of six beta strands encircled by five alpha-helices, are the protein's two functional domains. The C terminal domain is lipid-modified which helps in the membrane anchoring [5]. The G domain of KRAS is one of the functional biological areas which has 1-166 residues. Other crucial KRAS functional regions include the switch I and switch II domains, which operate as a binding interface for effector proteins. The Walker A motif, a short P-loop component with 10–14 residues, is present in the KRAS structure. In the P-loop or switch −2 region the cancer mutation hotspots are found predominantly [6, 7]. The G-domain, a highly conserved region that contains switch I and switch II loops and is involved in GDP-GTP exchange, is one of the protein's three primary domains [4]. KRAS transmits signals from the cell membrane to the nucleus when it is active, activating a variety of signaling pathways after receptor tyrosine kinase (RTK) activation (EGFR, ALK, or cMET) and ultimately causing the activation of transcription factors that regulate cell growth (cell proliferation and cell survival) and differentiation [8]. KRAS is activated when GTP binds to KRAS and causes alterations in the switch I and switch II loops of the G-domain. "KRAS-GTP directly interacts with and activates a number of downstream effector proteins in the active state, including RAF and PI3K". A-RAF, B-RAF, and C-RAF are the three subtypes of serine-threonine kinase. The RAF in an active state activates MEK which activates ERK which promotes cell growth and proliferation [9]. The GTP and GDP-bound forms of the KRAS protein cycle during its 24-hour half-life with resynthesis [10]. KRAS was first considered undruggable due to lack of binding pocket which can be accessible to small molecules. However, sotorasib and adagrasib which just have recently been discovered and selectively target KRAS^{G12}, offer new treatment approaches to enhance patient outcomes. Due to the development of sotorasib and adagrasib KRAS^{G12C} is now druggable [11]. Early adaptive feedback reactivation of signaling pathways has impeded prior attempts to target the RAS-RAF-MEK pathway and resulted in treatment resistance. Currently, in vitro and the clinical setting, have described that secondary KRAS mutations confer acquired resistance to KRAS^{G12} inhibitors [12]. Drug design and development is a challenging, expensive, and time-consuming procedure. It involves the discovery of promising targets and the design of therapeutically effective and safe drugs against promising targets. Computer-aided drug design (CADD), employs a number of computational and statistical techniques to efficiently assess biological target selection and hit identification [13, 14]. The process of drug development can be sped up by using advanced computational techniques. For the purpose of drug development, CADD can further make use of the integrated biochemical space to improve safety, and efficacy [15]. A number of machine learning algorithms are increasingly used in the drug development process. Only when reliable and accurate pre-processed data are combined with efficient computational methods and tools successful applications in the drug-designing process can be achieved [16] In this study, machine learning-based virtual screening was performed for the identification of new inhibitors against KRAS^{G12}.

Materials and methods

Dataset preparation

The active compounds against KRAS^{G12} mutant with experimentally determined IC50 values were retrieved from the Binding DB [17]. The compounds in SDF format were imported to MOE software. Moreover, the DUD-E web database was accessed and the corresponding decoys were generated [18]. The class label was added to the dataset all the inactive compounds were labeled as 0 and the active compounds were labeled as 1. The entire dataset was split into train and test sets (70% and 30%) respectively [19]. Prior to train and test set splits 20% of the data was separated from the whole dataset which was used as an independent dataset for external validation.

Molecular descriptors calculation and features selection

MOE software was used for 2D feature calculation [20, 21]. A total of 208 2D features were calculated. In order to avoid overfitting and to enhance the generalizability of the models, the dataset was preprocessed which included the removal of zero and not available (NA) values. To develop a model that is easy to understand and computationally cheap, it is essential to choose feature subsets that are most relevant to predictive targets. We used support SVM-RFE to choose features in order to collect as useful information as possible [22].

Machine learning models

Four models including k-nearest neighbors, support vector machine, naïve Bayes, and random forest models were developed using open-source python v3.9. The Scikit-learn library of python was used for model development [23].

K-nearest neighbor (kNN)

k-Nearest Neighbors (KNN) is a binary classification algorithm that classifies the data by calculating the distance between the nearest neighbors [24]. The number of neighbors considered for classification is represented by the parameter n_neighbors [25]. The best k value in this investigation was found as 11.

Naïve Bayesian (NB)

The naive Bayesian is a reliable classification algorithm that is based on the Bayes theorem. A data set can be classified using the NB model under the suppositions that each attribute contributes equally and independently to a dataset [26]. In this work, python v3.9 was used for NB model generation.

Random forest model

Breiman introduced the categorization algorithm known as RF [27]. Random Forest (RF) is a popular model that can be employed for data classification or regression tasks [28]. The tree of the random forest is trained using a bootstrap sample and the majority vote of the trees determines the predictions. The two primary hyperparameters that were optimized during model construction were max_features and n_estimators, which represent the number of trees constructed prior to predictions [25]. The number of estimations from 100 to 500 was taken into consideration.

Support vector machine (SVM)

The SVM model is frequently used to solve the problems of classification, pattern recognition, and regression [29]. The multiclass classification problems can also be solved by SVM. SVM draws margin lines (support vectors) parallel to a separable hyper-plane between the data classes. To transform the low dimensional data to higher dimensional space SVM model use different kernel trick, which includes the linear, polynomial, sigmoid, and radial base function [30]. We used RBF and the grid search approach to determine the best values for the C and γ parameters. Finally, the optimal values were determined to be $C=1000$ and $\gamma=1$.

Performance evaluation of models

Different performance evaluation parameters are used in machine learning to evaluate how effective an intelligent model is [31]. When a classification system generates true and false predictions, they are kept in a confusion matrix [32]. In many classification models, accuracy is used to evaluate the quality of the classification algorithm, but in some cases (such as with imbalanced datasets), accuracy alone is insufficient to assess a prediction model's overall effectiveness [33]. The MCC parameter is considered an important indicator for measuring the performance of binary classification. The highest MCC value is an indication of the good performance of the model [34].

The receiver operating characteristic (ROC) curve is also effective to evaluate the performance of the models. A ROC curve can visually represent the true positive rate against the false positive rate [35]. In this study, various parameters such as sensitivity, specificity, accuracy, and MCC were calculated for the developed models. To further evaluate the performance of the best model the area under the ROC curve (AUC), which is used to rate the models, was also calculated. The perfect model has an AUC value of 1 while a value of 0.5 indicates the random performance of the model [36].

Models validation

The three methods of validation most frequently employed by researchers are the independent tests, k-fold CV, and jackknife CV [37]. In order to evaluate the effectiveness of our models, we used five-fold cross-validation.

Virtual screening and molecular docking study

The model with the best accuracy and MCC value was further used to screen a total of four databases including the Zinc database, the South African natural product database, Pakistani phytochemicals, and the in-house database. The hits predicted by ML algorithms were further docked against the KRAS^{G12C} mutant. For the molecular docking study, the PDB structure of the receptor KRAS^{G12C} mutant (PDB ID 6OIM) was retrieved from the RCS PDB database. As the target protein structure may be coupled to heavy atoms, water, ligands, and cofactors, it cannot be directly employed for molecular docking. Polar charges were added to the structure and the water molecules were removed [38]. The energy of the receptor was minimized using a gradient of 0.05. A total of 10 conformations were generated for each ligand. After docking completion, the best ligand conformations were evaluated for their binding interactions using PyMol software [39]. Furthermore, for covalent ligands, we employed the covalent docking protocol of MOE (2016) software. The Cys12 residue of the KRAS^{G12C} was defined as the reactive residue for covalent docking and Michael's adduct reaction was used as the suitable reaction type for covalent docking [40].

All atoms MD simulation

AMBER version 20 was used to perform the simulation of the top best docking score complexes. The ff14SB was used as the force field for the protein while the General Amber force field (GAFF) parameters were assigned to the ligands [41]. A TIP3P cubic box with an 8 Å distance around the protein complex was used for the MD simulations study. For the systems to be neutralized, counter ions like Na+or Cl ions were added. Energy minimization was done in two steps prior to MD simulations. In the first stage, 5000 steps of steepest descent were applied then conjugate gradient minimization was performed to gradually minimize the whole system. The system's temperature was increased from 0 to 300 K during MD simulation with constant volume and periodic boundary conditions. All the systems were equilibrated for 3 ns with constant pressure and constant temperature [42]. Finally, a total of 100 ns MD simulation was performed for the top four protein-ligand complexes. Using the cpptraj module of AMBER 20 software, all of the generated MD trajectories were analyzed. The post-simulation analysis such as root mean square deviations (RMSD), root mean square fluctuations (RMSF), the radius of gyration (RoG), and dynamics cross-correlation map (DCCM) were performed using the CPPTRAJ module of Amber 20 after the completion of MD simulations [43].

Binding energy calculation

The MMGBSA is the most significant approach in reranking the binding conformations [44]. In order to calculate the binding free energy of the KRAS^{G12C}-ligand complexes by taking into account 2500 snapshots, we employed the MMPBSA.py script [45]. To estimate the binding free energy the following equation was used:

$$\Delta Gbind = \Delta Gcomplex - [\Delta Greceptor + \Delta Gligand]$$

ΔGreceptor, ΔGligand, and ΔGcomplex represent the binding energies of proteins, drugs, and complexes, respectively while the ΔGbind represents the total binding energy.

The individual binding energies that make up the overall binding free energy, such as those that are bonded (Gbond), electrostatic (Gele), polar (Gpol), and nonpolar (Gnpol), were estimated using the following equation.

$$G = G_{bond} + G_{ele} + G_{vdW} + G_{pol} + G$$

Results

Dataset preparation

A total of 386 active compounds for KRAS^{G12C} with reported IC50 values were retrieved from the binding databank database. The DUDE database [45] was accessed to generate the inactive compounds. A total of 1608 decoys were generated. By combining the active compounds and the decoys a dataset of 1994 compounds was prepared. The dataset was labeled with 1 and 0 indicating the active and inactive compounds respectively. From the whole dataset, 20% of the data was separated which was further used as an independent dataset for external validations of the ML models.

Features calculation

MOE software was used to calculate a total of 208 2D descriptors. In order to avoid overfitting and enhance the generalizability of the model the dataset was preprocessed by removing zero and NA values present in the dataset. The number of features was reduced to 172 after preprocessing.

Optimum features selection

Filter, wrapper, and embedding approaches are the three types of methods currently used by the SVM to evaluate the significance of variables in the dataset. In the present study, we used recursive feature elimination (RFE), for the optimum features selection. The RFE is a gold standard method among wrapper techniques [46]. Out of 172 features, a total of 13 optimum features were selected. Figure 1 shows the optimum feature selection curve. The optimum features including PEOE_VSA+2, PEOE_VSA_POS, PEOE_VSA+0, PEOE_VSA-0, SlogP_VSA3, SMR_VSA6, vsa_hyd, PEOE_VSA_NEG, Weight, PEOE_VSA_HYD, Q_VSA_HYD, Q_VSA_POS, and vdw_area were selected using the SVM RFE technique. To improve each model's performance, selected subsets of features were used to train all machine learning models.

Chemical space and diversity

The reliability of ML algorithm depends on the chemical diversity of a dataset. To execute the models perfectly, substantial chemical space is required. Figures 2 and 3 displays the significant chemical space between logP and molecular weight (MW) for the train and test set respectively. A significant chemical gap between inhibitors that are active and those that are not, with logP and MW varying from 4 to 8 and 250–600 Da was found for both the train and test datasets.

Performance of machine learning models

The dataset was split into train and test sets with 70% of the data considered as a train set while 30% of the data was selected as a test set. Open-source python v3.9 was used for model generation. A number of supervised machine learning models including KNN, SVM, GNB, and RF were applied. To evaluate the model performance different parameters such as accuracy, sensitivity, specificity, and MCC were calculated. The RF model

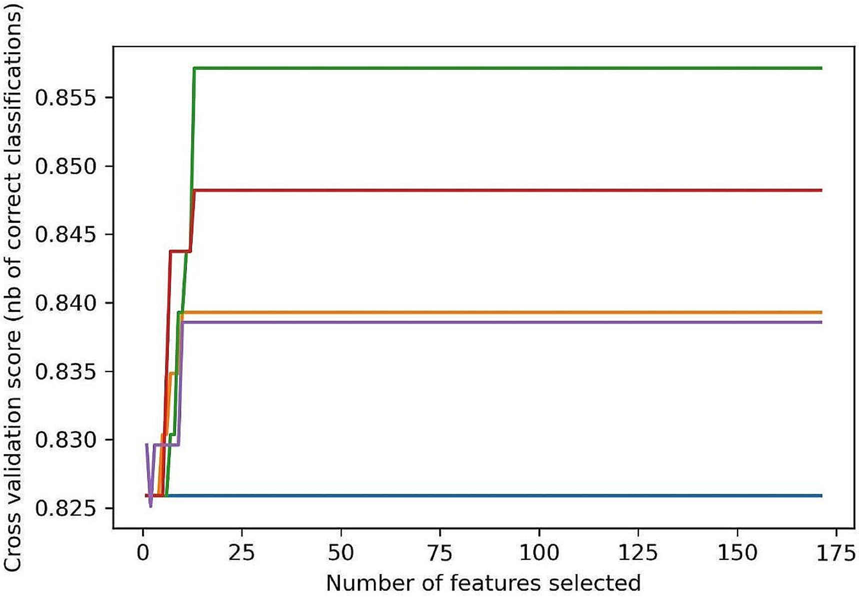

Fig. 1 The feature selection curve for 2D molecular descriptors and the number of optimum features selected were 13

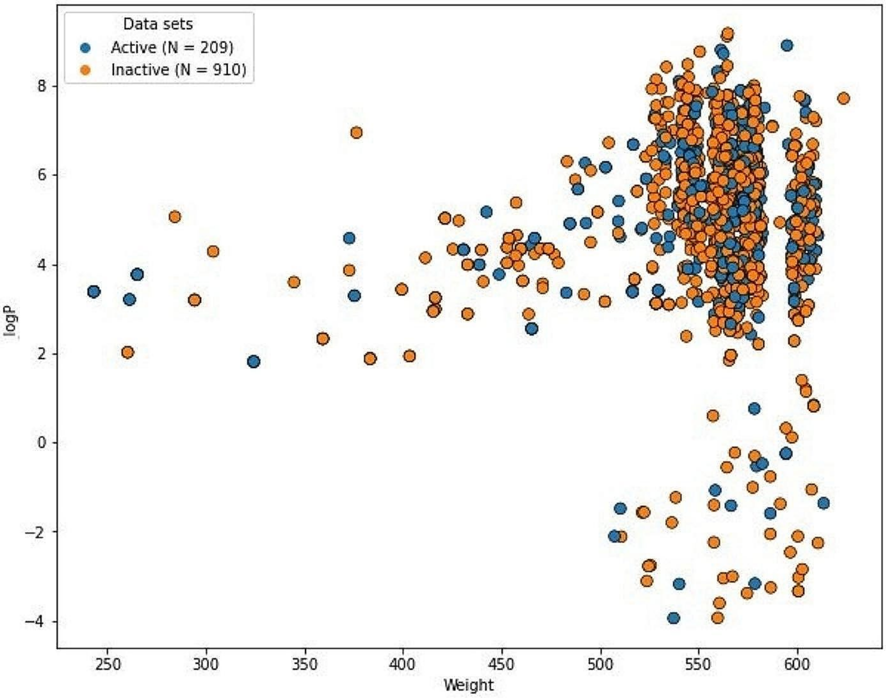

Fig. 2 Chemical diversity distribution of the training set. The X-axis defined the molecular weight and Y-axis shows logP

was found as the best model its performance was best on both the train and test set. The accuracy of RF model was 98% on both the train and test sets. The MCC value of the RF model was 0.95 on the train set and 0.97 on the test set. The accuracy of the KNN and SVM models was 94% on both the train and test sets while the MCC of the KNN model was 0.82 on the train set and 0.87 on the test set. The accuracy of the GNB model was 92% on the train set and 89% on the test set. The overall performance of the four generated models on the train set is summarized in Table 1 while Table 2 describes the performance of the four models on the test set. One of the most reliable methods for evaluating the model performance is the analysis of the ROC-AUC curve. The RF model has

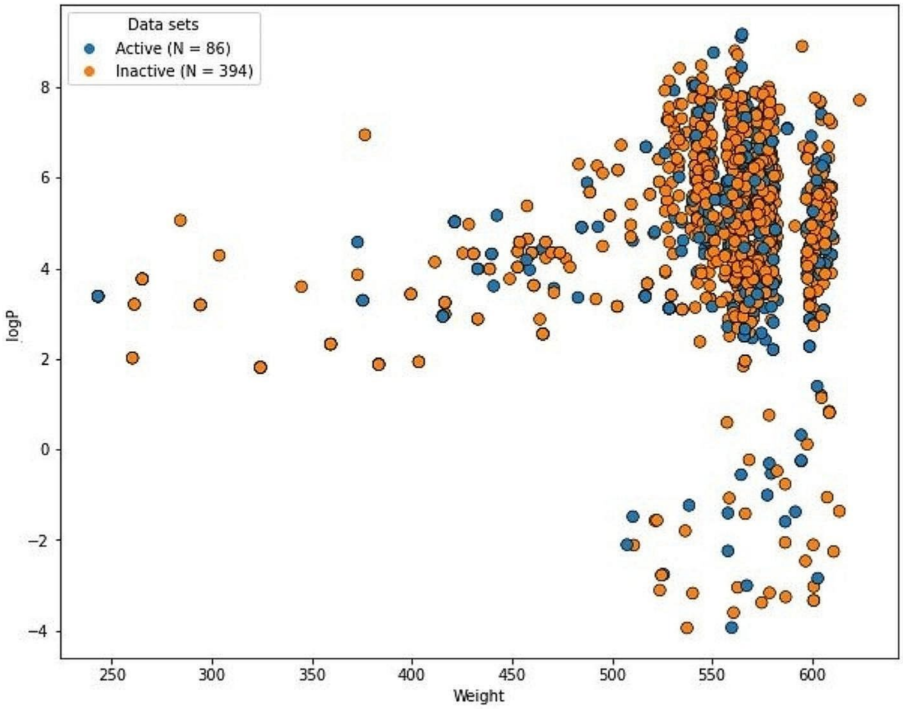

Fig. 3 Chemical diversity distribution of the test set. The X-axis defined the molecular weight and Y-axis shows logP

Table 1 Performance of ML models on the train set

Model	Accuracy	Sensitivity	Specificity	MCC	Precision	Recall
KNN	94%	0.95	0.94	0.82	0.74	0.86
SVM	94%	0.92	0.95	0.84	0.83	0.90
RF	98%	0.96	0.99	0.95	0.97	0.97
GNB	92%	0.85	0.94	0.77	0.70	0.90

Table 2 Performance evaluation of ML models on the test dataset

ML model	Accuracy	Sensitivity	Specificity	MCC	Precision	Recall
KNN	96%	0.87	0.98	0.87	0.86	0.89
SVM	94%	0.85	0.96	0.81	0.81	0.86
RF	98%	0.95	0.99	0.97	0.93	0.95
GNB	89%	0.85	0.90	0.68	0.80	0.93

achieved the highest area under the curve (AUC) value of 0.99 on both the train and test set, followed by KNN with an AUC value of 0.94 on the train and 0.93 on the test set. Figure 4 represents the ROC-AUC curve on the train set while Fig. 5 represents the ROC-AUC curve on the test set.

Models validation by independent dataset

A classification model's predictive power is not only assessed by its MCC score and accuracy during internal validation. As a result, any machine learning-based model's performance is dependent both on internal and external validation. For external validation of the models, an independent dataset was prepared. A total of 397 compounds were added to the dataset with a total of 99 compounds as active against KRAS^{G12C} while the remaining were inactive compounds in the independent dataset. The inhibitors of the external dataset were not present in the train or test sets. As compared, other ML algorithms the RF model revealed the highest accuracy, sensitivity, specificity, and MCC values on the independent dataset (Table S1, Table S2). Since the RF model performs the best among all of the models, so it was used for virtual screening to find potent KRAS^{G12C} inhibitors. The ROC-AUC curve for the independent dataset is presented in Figures S1 and S2.

Virtual screening

Four lakh compounds retrieved from the ZINC15 database were passed from the Lipinski rule of 5 before the

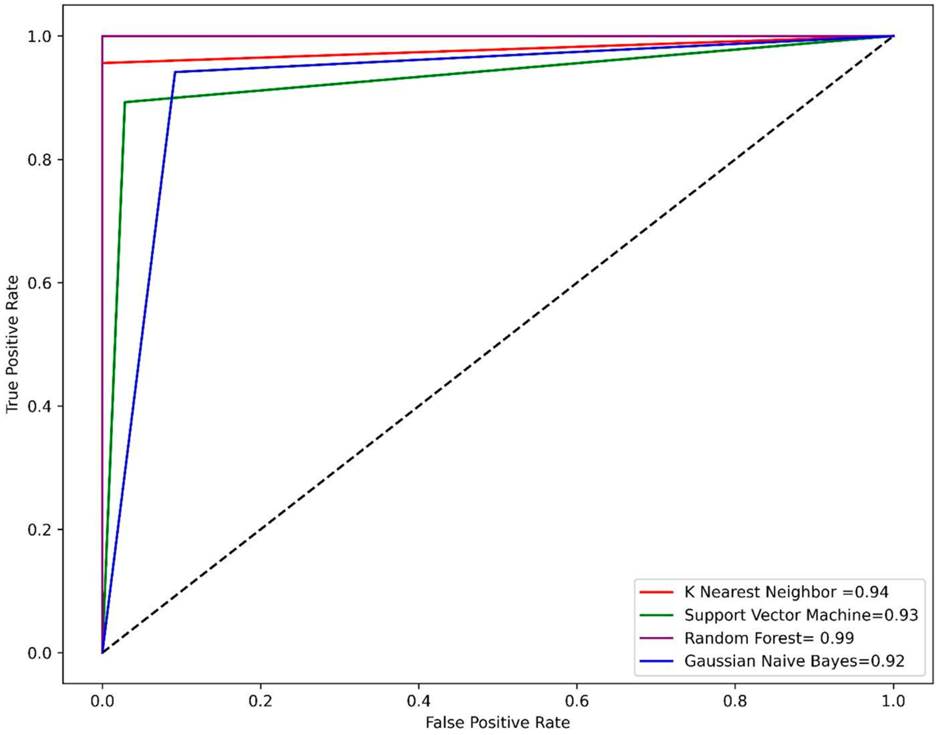

Fig. 4 The AUC-ROC curve on the train set for all four models

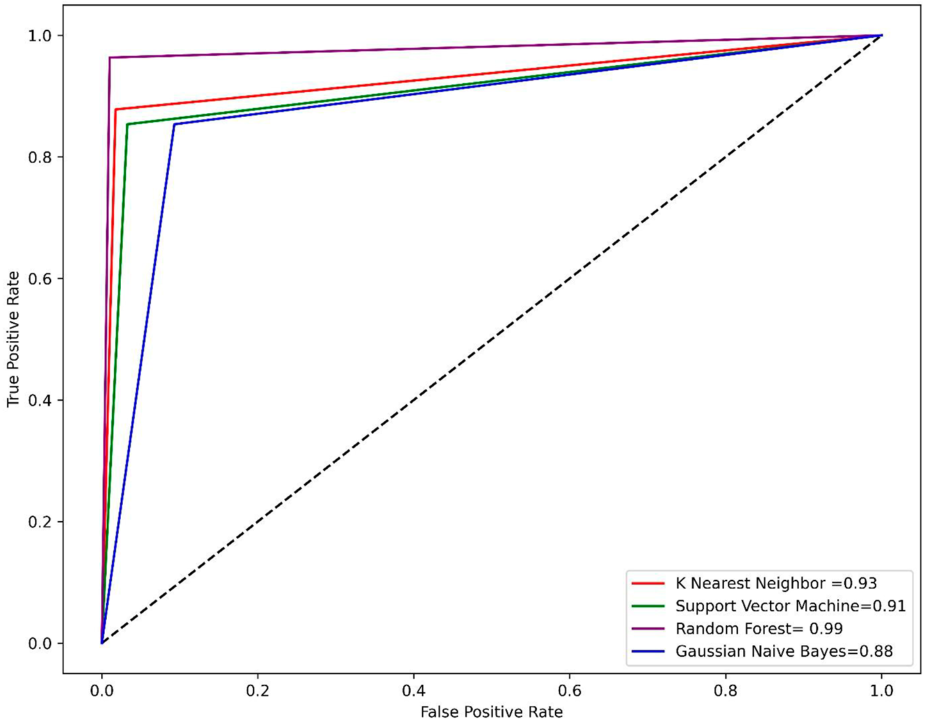

Fig. 5 The AUC-ROC curve on the test set for all four models

virtual screening. Among the 4 lakh compounds, only 1 lakh compounds were found to obey the Rule of five. Only sixty thousand were non-toxic so only these compounds were selected for the virtual screening. The updated version of the South African natural product database contained a total of 1012 compounds that were used for virtual screening. Furthermore, a total of five thousand Pakistani phytochemicals retrieved from the PubChem database and a total of 2 thousand compounds from the in-house database were used for the virtual screening. A total of 101 compounds from the ZINC15 database, 42 compounds from the Pakistani phytochemicals, 23 compounds from the in-house database, and 19 compounds from the SANCDB (South African Natural Compounds Data Base) were predicted as active by the RF (Random Forest) model.

Molecular docking analysis

The compounds predicted as active by the RF model were docked against the KRAS^{G12C} mutant. The docking study revealed that among the 101 active compounds of the ZINC database, the docking score of most of the compounds was good. The compound ZINC001458505494 was predicted as the covalent inhibitor of KRAS G12C. The docking score of compound ZINC001458505494 was found as -7.80. The compound ZINC001436082395 established a total of four conventional hydrogen bond interactions with Pro34, Tyr32, Lys16, and Ala59 residue of the receptor while one covalent bond with Cys12 and one Pi-H contact with Gln61 was also observed. Compound ZINC001436082395 with an S score of -12.15 formed one five H-bond interactions with Ser65, Glu62, Gly60, Gln61, and Ala66. Table S3 describes the interaction pattern of the top four best compounds along with their docking score while Table S4 describes the drug-like properties of the top five best docking-scored compounds of the ZINC database.

Twenty-three hits were found as active against KRAS^{G12C} from the virtual screening of the in-house database. The docking analysis of these 23 compounds revealed that H-209 is the most potent with an S score of -16.16. The compound H-209 formed four H-bond interactions with the Cys12, Asn86, Lys88 residues, and one Pi-H contact with the Cys12 residue of KRAS^{G12C}. The compound H-164 with a docking score of -14.14 was found as the second potent. The compound H-164 formed hydrogen bond interactions with Gln99, Glu91, Glu62, Arg102, and Lys88. Compound H-237 formed a similar pattern of interaction with Arg102, Glu98, and Glu91 as established by the compound H-164. However, compound H-237 formed two pi-pi interactions with His95 cryptic pocket residue and His94. The interaction pattern of the top five best docking score compounds is present in Table S5 while the drug-like properties of the top 5 best compounds of the in-house database are present in Table S6.

The virtual screening of the South African Natural Products database revealed a total of 19 compounds out of 1012 compounds as active against the KRAS^{G12C}. Among the 19 docked compounds the compound SANC00905 was the most promising covalent inhibitor with an S score of -9.61. The most potent compound SANC00905 revealed a total of five hydrogen bond interactions with Cys12, Glu62, Gln99, and Arg68 residues of the KRAS^{G12C}. It was also found that the SANC00905 formed one covalent bond with the mutated Cys12 residue of KRAS G12C and one H-pi contact with the His95 cryptic pocket residue of the KRAS G12C protein. The interaction pattern of the top six best docking-scored compounds is present in Table S7 while the properties of the best docking-scored compounds of the SANCDB are present in Table S8. Furthermore, 42 compounds were identified as active out of the total five thousand Pakistani phytochemicals. Among the docked compounds the compound PubChem ID 11,968,893 was predicted as the most promising with a docking score of -18.58. The potent compound PubChem ID 11,968,893 formed a total of seven hydrogen bond interactions with Cys12, Gln61, Glu62, Asn86, and Lys88 residues while one Pi-H contact was also observed with the Lys88. The interactions and docking scores of the most promising Pakistani phytochemicals are present in Table S9 and their drug-like properties are present in Table S10. Overall our molecular docking study revealed two covalent inhibitors (ZINC001458505494 and SANC00905) and two non-covalent inhibitors (H-209, and PubChem ID 11,968,893) for the KRAS G12C drug target. The 3D interactions of the covalent inhibitors in complex with KRAS^{G12C} are present in Fig. 6 (A) while 3D interactions of the non-covalent inhibitors are shown in Fig. 6 (B). The 2D interactions of the best docking score compounds from all databases are shown in Figure S3. Table S11 displays the docking result of the covalent inhibitors.

Post simulation analysis

RMSD analysis

The stability of KRAS^{G12} in complexes with the top four ligands was evaluated from the RMSD analysis during the 100ns MD simulation. The stability of complexes and details regarding the structural conformation during the simulation can be provided by the RMSD value. Figure S4 represent the RMSD plots for the top four best docking-score protein-ligand complexes and the standard drug sotorasib in the complex with the KRAS^{G12C}. All four complexes were stable during the MD simulation. In Replica-1 no significant deviations were observed in the complex H-209- KRAS^{G12C} during the simulation. Some minor fluctuations were observed

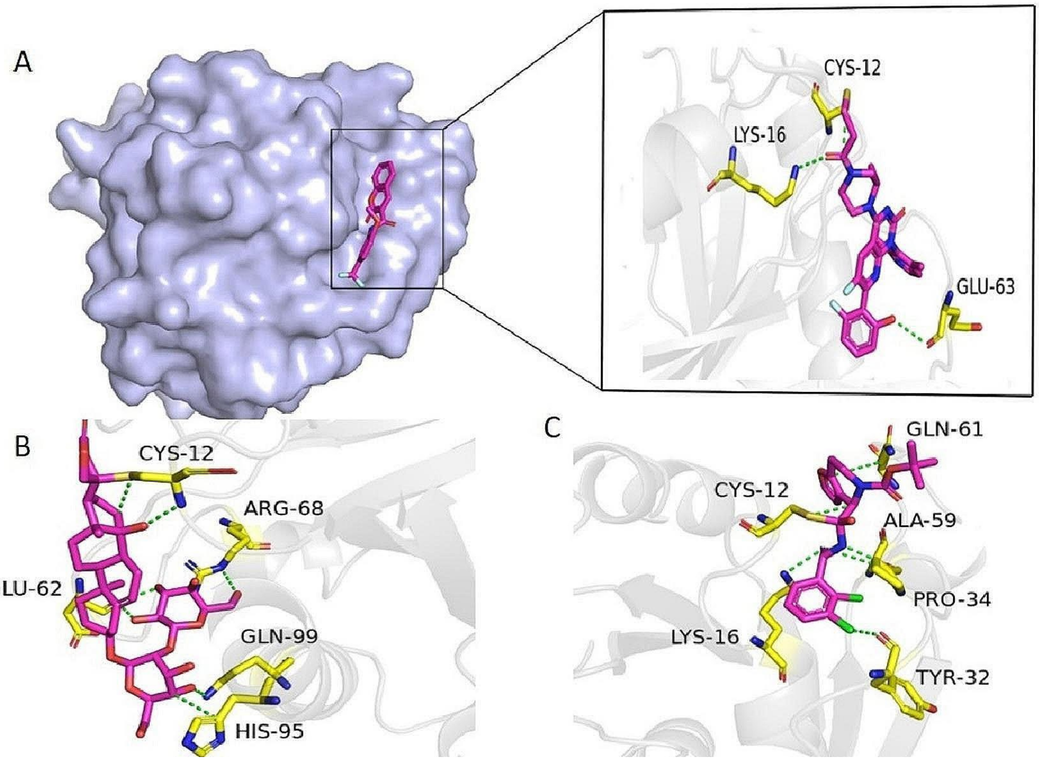

Fig. 6A 3D interactions of covalent inhibitors (**A**) standard drug- $KRAS^{G12C}$ complex (**B**) SANC00905- $KRAS^{G12C}$ (**C**) ZINC001458505494- $KRAS^{G12C}$

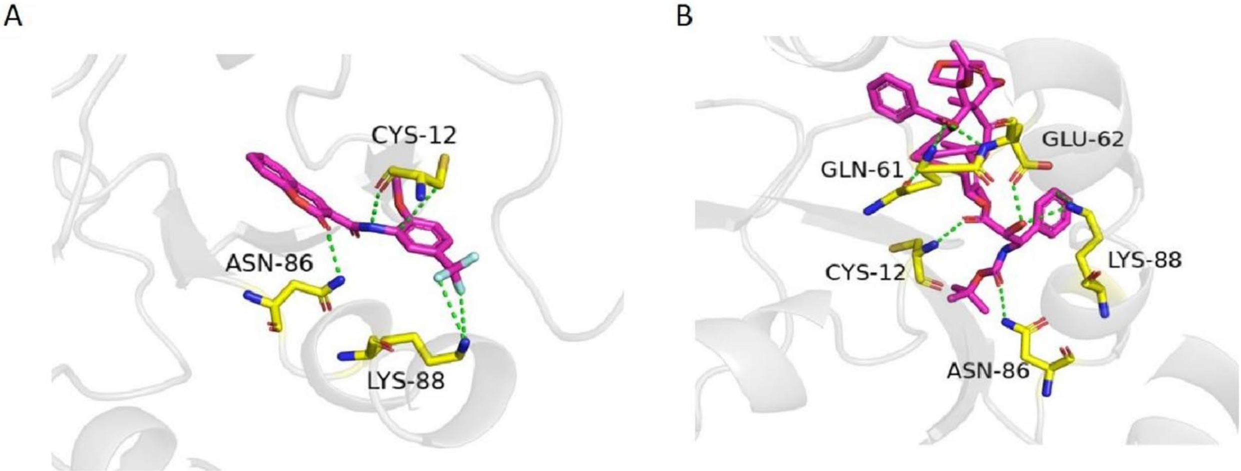

Fig. 6B 3D interactions of non-covalent inhibitors (**A**) H-209- $KRAS^{G12C}$ (**B**) PubChem-CID11968893-$KRAS^{G12C}$

during 5–15 ns and 20–35 ns but after that, the complex remained stable till the 100 ns simulation (Figure. S4). In Replica-2 only minor deviations were seen during 20–25 ns in the H-209 complex after that the RMSD converged and a stable behavior was observed during the entire 100 ns MD simulation. The RMSD of complex PubChem-CID11968893-$KRAS^{G12C}$ was highly stable in Replica-1, and only minor fluctuations during the 25–30 ns and 45–50 ns were observed after that the complex reached stability and remained stable till 100 ns. A similar pattern of RMSD was observed for PubChem-CID11968893-$KRAS^{G12C}$ in Replica-2. The system was found to be highly stable in both runs (Figure. S4). The complex SANC00905- $KRAS^{G12C}$ revealed stable behavior during the simulation. However, some minor deviations during the 23–38 ns were reported in Replica-1, and after that the system gained stability and remained stable during the 100 ns MD simulation. In

Replica-2 the complex SANC00905- KRAS^{G12C} revealed a more stable behavior and only minor deviations were seen during 25–38 ns after that no major or minor deviations were observed till 100 ns. Initially, the RMSD of complex ZINC001458505494- KRAS^{G12C} was high during the first 10 ns afterward the RMSD decreased and reached stability but only minor deviations during 42–45 ns and 94-98ns were found in Replica-1 and the overall RMSD was stable. In Replica-2 the RMSD of complex ZINC001458505494- KRAS^{G12C} was highly stable except 80–90 ns. The RMSD of the standard drug in complex with KRAS^{G12C} was initially stable in Replica-1 but then some major fluctuations were found during the 60-65ns and 70–85 ns after that the system attain stability till 100ns in both the Replica.

RMSF analysis

Flexibility at the residue level can be analyzed from the RMSF profile. The flexible and stable regions are represented by a higher and lower RMSF value respectively [47]. The complexes almost showed the same pattern of residual flexibility in Replica-1 and Replica-2. The residues 1–29, 69–100, and 105–170 revealed high stability during the MD run. A total of three different peaks were found at three different time periods (Figure S5). Initial peak was observed with ASP30, GLU31, TYR32, PRO34 and THR 35, second with GLU62, GLU63, TYR64, SER65, ALA66, MET67, and ARG68 and third with ARG102, VAL103 and LYS104 residues. The residues that revealed great fluctuations were not the active site residues except Glu63. On the other hand, the residues lying in the active site were found as highly stable. Figure S5 display the RMSF pattern of Replica 1 and Replica 2 for all the systems.

Compactness analysis

To comprehend the degree of compactness of each ligand-bound system the Gyration radius (RoG) was determined. To evaluate how these ligands remained intact with the KRAS^{G12C} during the 100 ns MD simulation the RoG was calculated. Moreover, the compactness of systems represents stability. A lower RoG value denotes greater stability and a high RoG value denotes an unstable system [48]. In Replica-1 and Replica-2 the average RoG value was determined as 15.6–15.9 Å for the H-209-KRAS^{G12C} complex. The average RoG of the PubChem-CID11968893-KRAS^{G12C} system was 15.3–15.8 Å in Replica-1 while in Replica-2 the RoG was found to be 15.2–15.5 Å and the complex was observed as highly compact with no major deviations. The average RoG of the SANC00905-KRAS^{G12C} complex was 15.2–15.6 Å in Replica-1 while in Replica-2 a similar pattern of the RoG was found. The RoG of the ZINC001458505494-KRAS^{G12C} was found as 15.6–16.6 Å in Replica-1 while in Replica-2 the average RoG was found to be 15.4–16.4 Å. The RoG of the standard drug-KRAS^{G12C} complex was initially compact but increased from 15.2 to 16.4 Å during the 80-90ns then decreased after 90ns and remained compact till 100 ns. As compared to all other complexes the standard drug in complex with KRAS^{G12C} revealed a little unstable behavior. Among all the complexes the compound SANC00905 found as the covalent inhibitors for KRAS G12C was more compact during MD simulation. Figure S6 displays the RoG plots of all the ligand-bound complexes.

Dynamic cross-correlation map (DCCM)

The negative correlations imply that residues move in the opposite direction, and positive correlations show that residues are moving in the same direction i.e. anti-parallel and parallel direction. The residues displayed a positive correlation suggesting that the positive correlation may be caused by ligands interactions with the active site residues of KRAS. The green color indicates positive correlation while the dark brown color revealed negative correlation among the residues (Fig. 7). Among all the simulated systems PubChem-CID11968893-KRAS^{G12C} complex and SANC00905-KRAS^{G12C} have the highest positive correlation motions followed by the complex ZINC001458505494-KRAS^{G12C}. Additionally, the H-209-KRAS^{G12C} complex exhibits greater correlated and anti-correlated motion as compared to the control.

Binding energy calculation

It was found that all of the predicted active ligands in the complex with KRAS^{G12C} displayed strong binding affinity as compared to the control system, indicating that all of the systems are more likely to be stable. Table 3 provides an overview of the binding free energy and its components determined by the MM-GBSA calculation. According to the findings, the SANC00905-KRAS^{G12C} complex had the highest negative binding free energy value (-53 kJ/mol), followed by the PubChem-CID11968893-KRAS^{G12C} complex (-50 kJ/mol). As demonstrated in Table 3, all four predicted ligand complexes had very good binding free energy values, compared to the control indicating that they were more stable during the simulation. The total binding energy for sotorasib- KRAS^{G12C} on the other hand, was predicted to be around −45 kcal/mol. These findings clearly imply that these ligands have more potent inhibitory potential than the standard drug.

Discussion

Approximately 1.8 million people die each year due to lung cancer and lung cancer is considered the number one cause of cancer death worldwide with a high mortality rate. Approximately 84% of all lung malignancies are

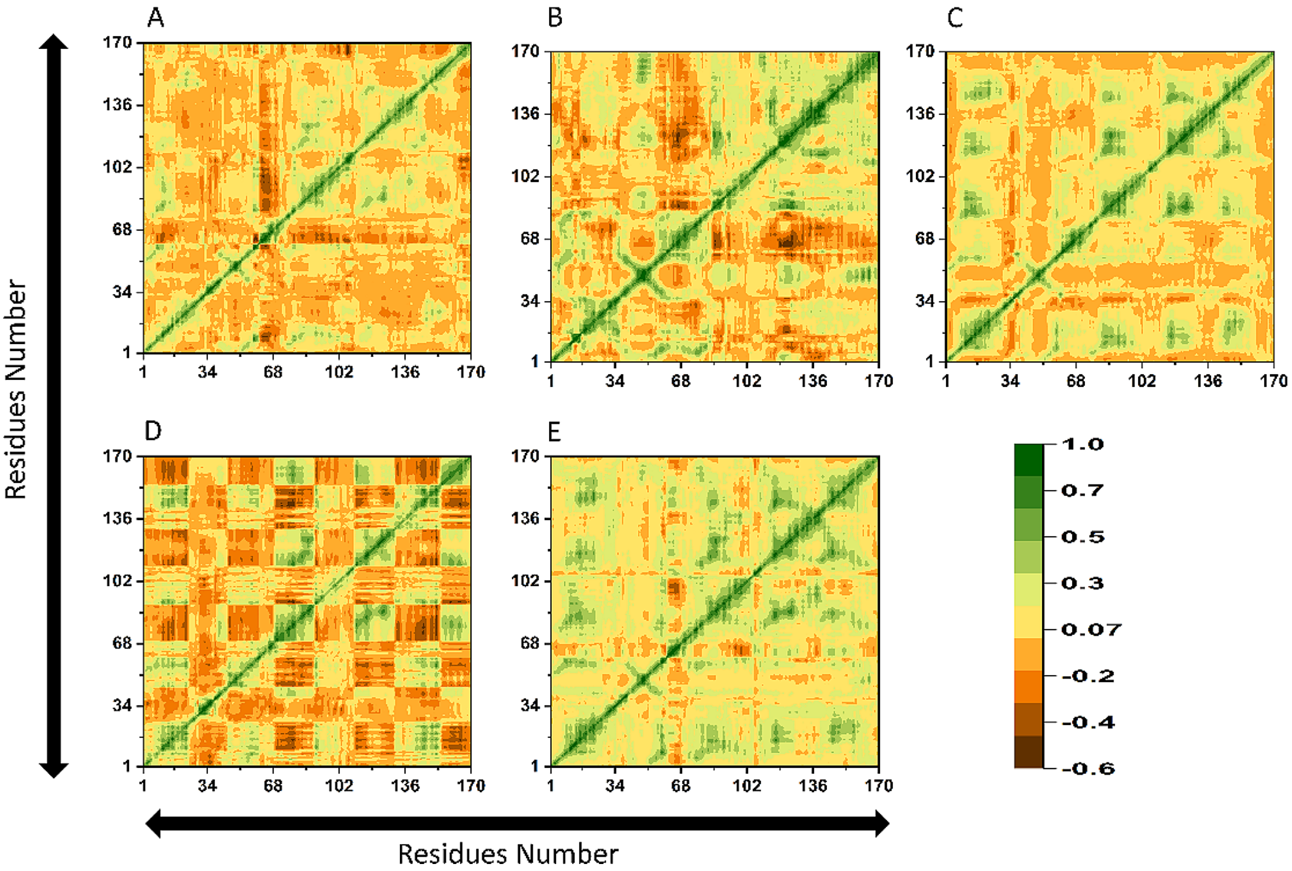

Fig. 7 The DCCM map of (**A**) H-209- KRAS^{G12C} (**B**) PubChem-CID11968893-KRAS^{G12C} (**C**) SANC00905-KRAS^{G12C} (**D**) ZINC001458505494-KRAS^{G12C} (**E**) standard drug-KRAS^{G12C} complex. The X-axis and Y-axis display number of residues

Table 3 MMGBSA analysis indicating binding energy of all the complexes

Complex (Replica-1)	**vdW**	**EEL**	**ESURF**	**EGB**	**ΔG TOTAL**
H-209-KRAS^{G12C}	-55.26	-9.38	-4.51	28.34	-45.82
ZINC001458505494-KRAS^{G12C}	-61.41	-11.74	-6.01	26.56	-47.61
SANC00905-KRAS^{G12C}	-71.52	-20.78	-9.10	18.70	-53.71
PubChem-CID11968893-KRAS^{G12C}	-69.60	-17.03	-8.53	23.89	-50.28
Sotorasib-KRAS^{G12C}	-52.47	-7.00	-4.06	29.21	-45.42
Complex (Replica-2)	**vdW**	**EEL**	**ESURF**	**EGB**	**ΔG TOTAL**
H-209-KRAS^{G12C}	-48.78	-9.82	-5.86	25.38	-39.08
ZINC001458505494-KRAS^{G12C}	-42.78	-5.46	-4.81	17.82	-35.24
SANC00905-KRAS^{G12C}	-66.44	-4.81	-6.86	20.80	-57.33
PubChem-CID11968893-KRAS^{G12C}	-66.58	-10.66	-6.66	32.41	-51.50
Sotorasib-KRAS^{G12C}	-51.98	-11.82	-6.09	29.10	-40.80

vdW = van der Waals energy, EEL = electrostatic energy, ESURF = surface areas energy, EGB = the electrostatic contribution to the solvation free energy

non-small cell lung cancers (NSCLC). The five-year survival rate of NSCL cancer patients is only 25% [49]. The KRAS^{G12C} gene mutation is a key initiator of NSCLC [50]. The clinical response rates of the developed drugs sotorasib and adagrasib are high, and toxicity from these drugs is low. However, resistance often develops after a few months of treatment [50]. Researchers are now searching for new effective drugs. The process of finding new drugs has substantially advanced through the use of ML algorithms. The use of multiple ML algorithms in drug discovery has considerably benefited pharmaceutical industries. These algorithms are frequently used in predicting the bioactivity of molecules, predicting drug-protein interactions, and optimizing the bioactivity and safety profile of the molecules [51]. A number of studies have been conducted on ML-based virtual screening [52, 53]. For instance, in our previous studies, we carried out ML-based virtual screening for the identification of new inhibitors against STAT3 a cancer drug target, and the Main protease drug target in the SARS CoV-2 [20, 54]. Similarly, some researchers used ML-based models for the identification of the functional groups responsible for binding [55]. A previous study used various ML algorithms for the identification of the new inhibitors from the mangrove secondary metabolic natural products database against KRAS^{G12C} protein [56].

In this study, we also used ML-based virtual screening to predict new inhibitors against KRAS^{G12C}. Different machine-learning models including the ensemble RF model, KNN, SVM, and GNB were used for the classification purpose. The performance of all the proposed algorithms was evaluated by the parameters such as accuracy, sensitivity, specificity, and MCC. The performance of all the models revealed that the ensemble RF model was the best by achieving an accuracy of 98%. The RF model was further used for the virtual screening of in-house, ZINC, Pakistani phytochemicals, and South African Natural Products databases. Phytochemicals, or naturally occurring plant molecules, are important sources of new drug discovery and are also used to treat cancer. These phytochemicals frequently work by controlling molecular pathways that are connected to the development and spread of cancer. The precise processes include boosting antioxidant status, inhibiting carcinogens, reducing proliferation, and inducing cell cycle arrest and apoptosis [58]. The previous study reported that the phytochemicals are effective against a variety of diseases including, diabetes, TB, skin infections, malaria, anemia, and epilepsy [57]. The hits predicted by the RF model were further docked against the KRAS^{G12C}. The docking results revealed a number of compounds with good docking scores and interactions with KRAS^{G12C} as compared to the standard drug sotorasib. Most of our predicted compounds revealed a similar pattern of interactions to the previous molecular docking studies carried out for KRAS^{G12C}. In the previous study two most promising compounds such as compound 14 and compound 44 revealed interactions with Cys12, Lys16, Pro34, Gly60, and Arg68 residues of the receptor [56]. Our compounds also established interactions with Cys12, Lys16, Pro34 and Gly60. Similarly, in another study compound CID_146235508 was found the most potent compound against KRAS^{G12C}. This compound made hydrogen bond and hydrophobic contacts with Cys12, Glu63, Lys16, Met72, Arg68, Ala59, and Tyr96 residues of the receptor [39] and a similar pattern of interactions was found in our study. After molecular docking, 100 ns MD simulation was performed for the top four complexes to reveal dynamic changes and the stability of the complexes. The RMSD analysis indicated the stable binding of the predicted compounds with the protein indicating these compounds as suitable inhibitors against the KRAS^{G12C}. The RoG analysis which are in line with the RMSD profile, further supported the complex SANC00905- KRAS^{G12C} stability compared to all complexes. The calculated binding free energy for the four complexes and the control revealed that the binding energy of all the complexes was lower as compared to the standard drug which clearly indicates that these compounds can bind strongly with the receptor and can make more stable complexes with the KRAS^{G12C}.

Conclusion

Targeting KRAS^{G12C} is found to be viable in anti-cancer research. In this study, both synthetic and natural compounds were screened against KRAS^{G12C} using machine-learning algorithms. To find the new hit compounds with the strongest anti-cancer potential, four databases such as ZINC, in-house, Pakistani phytochemicals, and South African Natural Products databases were screened. Using the molecular docking study the hits predicted from virtual screening were further analyzed for interactions with KRAS^{G12C}. Based on the interaction, the compounds that revealed good binding interactions were selected for MD simulation and binding energy calculation. As compared to the standard drug sotorasib the four predicted compounds such as ZINC001458505494, H-209, SANC00905, and PubChem CID: 11,968,893 revealed great stability and strong binding affinity for KRAS^{G12C}. We hope our virtual screening protocol can be helpful to find new inhibitors against the KRAS mutants and other drug targets in the future.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13065-024-01152-z.

Supplementary Material 1

Acknowledgements

The authors exten their appreciation to the Deputyship of Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project with number: ISP23-81.

Author contributions

Conceptualization, Abdul Wadood and Asad Khalid Methodology, Amar Ajmal, Hind A Alkhatabi, Mubarak A Aalmri and Roaa M Alreemi Software Mubarak A Aalmri and Bader S Alotaibi Data curation, Hing A Alkhatibi, Roaa M Areemi, Ashraf N Abdalla and Mubarak A Alamri Writing- Original draft preparation. Amar Ajmal, Asaad Kkhali and Abdul Wadood Visualization, Hind A. Alkhatabi, Ashraf N Abdalla and Bbader S Alotaibi Investigation. Abdul Wadood, Asaad Khalid, Mubarak A. Alamri, and Ashraf N Abdalla: Supervision. Abdul Wadood and Asaad Khalid, Validation, Hind A Alkhatabi, Roaa M Alreemi, and Bader S Alotaibi Writing- Reviewing and Editing Abdul Wadood, Asaad Khalid, Mubarak A Alamri and Ashraf N Abdalla.

Funding

The study was funded by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia (Project number: ISP23-81).

Data availability

All data is available in manuscript.

Declarations

Ethical approval
This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent
We have not performed work on human/animal, therefore, there is no need of informed consent.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 October 2023 / Accepted: 26 February 2024
Published online: 25 March 2024

References

1. Skoulidis F, Li BT, Dy GK, Price TJ, Falchook GS, Wolf J, Italiano A, Schuler M, Borghaei H, Barlesi F. Sotorasib for lung cancers with KRAS p. G12C mutation. N Engl J Med. 2021;384(25):2371–81.
2. Luo Y, Li Z, Kong Y, He W, Zheng H, An M, Lin Y, Zhang D, Yang J, Zhao Y. KRAS mutant–driven SUMOylation controls extracellular vesicle transmission to trigger lymphangiogenesis in pancreatic cancer. J Clin Investig. 2022;132(14).
3. Segal NH, Gada P, Senzer N, Gargano MA, Patchen ML, Saltz LB. A phase II efficacy and safety, open-label, multicenter study of imprime PGG injection in combination with cetuximab in patients with stage IV KRAS-mutant colorectal cancer. Clin Colorectal Cancer. 2016;15(3):222–7.
4. Canon J, Rex K, Saiki AY, Mohr C, Cooke K, Bagal D, Gaida K, Holt T, Knutson CG, Koppada N. The clinical KRAS (G12C) inhibitor AMG 510 drives anti-tumour immunity. Nature. 2019;575(7781):217–23.
5. Gerber M, Goel S, Maitra R. In silico comparative analysis of KRAS mutations at codons 12 and 13: structural modifications of P-loop, switch I&II regions preventing GTP hydrolysis. Comput Biol Med. 2022;141:105110.
6. Sayyed-Ahmad A, Gorfe AA. How to make an undruggable enzyme druggable: lessons from ras proteins. Adv Protein Chem Struct Biology. 2020;122:181–202.
7. Pantsar T. The current understanding of KRAS protein structure and dynamics. Comput Struct Biotechnol J. 2020;18:189–98.
8. Martin P, Leighl NB, Tsao M-S, Shepherd FA. KRAS mutations as prognostic and predictive markers in non–small cell lung cancer. J Thorac Oncol. 2013;8(5):530–42.
9. Reita D, Pabst L, Pencreach E, Guérin E, Dano L, Rimelen V, Voegeli A-C, Vallat L, Mascaux C, Beau-Faller M. Direct targeting KRAS mutation in non-small cell lung cancer: focus on resistance. Cancers. 2022;14(5):1321.
10. Désage A-L, Léonce C, Swalduz A, Ortiz-Cuaran S. Targeting KRAS mutant in non-small cell lung cancer: novel insights into therapeutic strategies. Front Oncol. 2022;12:796832.
11. Punekar SR, Velcheti V, Neel BG, Wong K-K. The current state of the art and future trends in RAS-targeted cancer therapies. Nat Reviews Clin Oncol. 2022;19(10):637–55.
12. Koga T, Suda K, Fujino T, Ohara S, Hamada A, Nishino M, Chiba M, Shimoji M, Takemoto T, Arita T. KRAS secondary mutations that confer acquired resistance to KRAS G12C inhibitors, sotorasib and adagrasib, and overcoming strategies: insights from in vitro experiments. J Thorac Oncol. 2021;16(8):1321–32.
13. Rajkishan T, Rachana A, Shruti S, Bhumi P, Patel D. Computer-aided drug designing. Adv Bioinf. 2021:151–82.
14. Niazi SK, Mariam Z. Computer-aided drug design and drug discovery: a prospective analysis. Pharmaceuticals. 2023;17(1):22.
15. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev. 2019;119(18):10520–94.
16. Katsila T, Spyroulias GA, Patrinos GP, Matsoukas M-T. Computational approaches in target identification and drug discovery. Comput Struct Biotechnol J. 2016;14:177–84.
17. Machado MR, Karray S. Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Syst Appl. 2022;200:116889.
18. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem. 2012;55(14):6582–94.
19. Kuldeep J, Sharma SK, Sharma T, Singh BN, Siddiqi MI. Targeting mycobacterium tuberculosis enoyl-acyl carrier protein reductase using computational tools for identification of potential inhibitor and their biological activity. Mol Inf. 2021;40(5):2000211.
20. Wadood A, Ajmal A, Junaid M, Rehman AU, Uddin R, Azam SS, Khan AZ, Ali A. Machine learning-based virtual screening for STAT3 anticancer drug target. Curr Pharm Design. 2022;28(36):3023–32.
21. Yu L, Sun X, Tian S, Shi X, Yan Y. Drug and nondrug classification based on deep learning with various feature selection strategies. Curr Bioinform. 2018;13(3):253–9.
22. Akbar S, Ahmad A, Hayat M, Rehman AU, Khan S, Ali F. iAtbP-Hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. Comput Biol Med. 2021;137:104778.
23. Hao J, Ho TK. Machine learning made easy: a review of scikit-learn package in python programming language. J Educational Behav Stat. 2019;44(3):348–61.
24. Zhang Z. Introduction to machine learning: k-nearest neighbors. Annals Translational Med. 2016;4(11).
25. Di Stefano M, Galati S, Ortore G, Caligiuri I, Rizzolio F, Ceni C, Bertini S, Bononi G, Granchi C, Macchia M. Machine learning-based virtual screening for the identification of CDK5 inhibitors. Int J Mol Sci. 2022;23(18):10653.
26. Akcay M, Etiz D, Celik O, Ozen A. Evaluation of prognosis in nasopharyngeal cancer using machine learning. Technol Cancer Res Treat. 2020;19:1533033820909829.
27. Bozorgmehr A, Thielmann A, Weltermann B. Chronic stress in practice assistants: an analytic approach comparing four machine learning classifiers with a standard logistic regression model. PLoS ONE. 2021;16(5):e0250842.
28. Galasso J, Cao DM, Hochberg R. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. Chaos Solitons Fractals. 2022;156:111779.
29. Ahmad I, Basheri M, Iqbal MJ, Rahim A. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access. 2018;6:33789–95.
30. Chu M, Gong R, Gao S, Zhao J. Steel surface defects recognition based on multi-type statistical features and enhanced twin support vector machine. Chemometr Intell Lab Syst. 2017;171:140–50.
31. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29:685–93.
32. Khan MA, Khan R, Algarni F, Kumar I, Choudhary A, Srivastava A. Performance evaluation of regression models for COVID-19: a statistical and predictive perspective. Ain Shams Eng J. 2022;13(2):101574.
33. Ahmad A, Akbar S, Tahir M, Hayat M, Ali F. iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. Chemometr Intell Lab Syst. 2022;222:104516.
34. Cai Y, Guan K, Peng J, Wang S, Seifert C, Wardlow B, Li Z. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. Remote Sens Environ. 2018;210:35–47.
35. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74.
36. Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JM, Moni MA. Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Comput Biol Med. 2021;136:104672.
37. Qazi S, Das S, Khuntia BK, Sharma V, Sharma S, Sharma G, Raza K. In silico molecular docking and molecular dynamic simulation analysis of phytochemicals from Indian foods as potential inhibitors of SARS-CoV-2 RdRp and 3CLpro. Nat Prod Commun. 2021;16(9):1934578X211031707.
38. Hari S. In silico molecular docking and ADME/T analysis of plant compounds against IL17A and IL18 targets in gouty arthritis. J Appl Pharm Sci. 2019;9(7):018–26.
39. Ogunlana AT, Oyedele A-QK, Boyenle ID, Ayoola SO, Ajibare AC, Adeyemi AO, Jinadu LA, Adenrele OT, Alausa AO, Adelusi TI. Computer-aided drug design of some KRAS G12C inhibitors: targeting the covalent and allosteric binding site for cancer therapy. Inf Med Unlocked. 2022;32:101032.
40. Khan M, Ahad G, Alam A, Ullah S, Khan A, Salar U, Wadood A, Ajmal A, Khan KM, Perveen S. Synthesis of new bis (dimethylamino) benzophenone hydrazone for diabetic management: in-vitro and in-silico approach. Heliyon. 2023.

41. Halder AK, Honarparvar B. Molecular alteration in drug susceptibility against subtype B and C-SA HIV-1 proteases: MD study. Struct Chem. 2019;30:1715–27.
42. Sahihi M, Gaci F, Navizet I. Identification of new alpha-synuclein fibrillogenesis inhibitor using in silico structure-based virtual screening. J Mol Graph Model. 2021;108:108010.
43. Ajmal A, Ali Y, Khan A, Wadood A, Rehman AU. Identification of novel peptide inhibitors for the KRas-G12C variant to prevent oncogenic signaling. J Biomol Struct Dynamics. 2023;41(18):8866–75.
44. Mahmood A, Samad A, Bano S, Umair M, Ajmal A, Ilyas I, Shah AA, Li P, Hu J. Structural and dynamics insights into the GBA variants associated with Parkinson's disease. J Biomol Struct Dynamics. 2023:1–13.
45. Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, Koes DR, Kurtzman T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS ONE. 2019;14(8):e0220113.
46. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
47. Khan AA, Baildya N, Dutta T, Ghosh NN. Inhibitory efficiency of potential drugs against SARS-CoV-2 by blocking human angiotensin converting enzyme-2: virtual screening and molecular dynamics study. Microb Pathog. 2021;152:104762.
48. Emmanuel IA, Olotu FA, Agoni C, Soliman ME. In silico repurposing of J147 for neonatal encephalopathy treatment: exploring molecular mechanisms of mutant mitochondrial ATP synthase. Curr Pharm Biotechnol. 2020;21(14):1551–66.
49. Mugarza E, van Maldegem F, Boumelha J, Moore C, Rana S, Llorian Sopena M, East P, Ambler R, Anastasiou P, Romero-Clavijo P. Therapeutic KRASG12C inhibition drives effective interferon-mediated antitumor immunity in immunogenic lung cancers. Sci Adv. 2022;8(29):eabm8780.
50. Xu K, Park D, Magis AT, Zhang J, Zhou W, Sica GL, Ramalingam SS, Curran WJ, Deng X. Small molecule KRAS agonist for mutant KRAS cancer therapy. Mol Cancer. 2019;18(1):1–16.
51. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine learning methods in drug discovery. Molecules. 2020;25(22):5277.
52. Sharma G, Shukla R, Singh TR. Identification of small molecules against the NMDAR: an insight from virtual screening, density functional theory, free energy landscape and molecular dynamics simulation-based findings. Netw Model Anal Health Inf Bioinf. 2022;11(1):31.
53. Zhu J, Wu Y, Wang M, Li K, Xu L, Chen Y, Cai Y, Jin J. Integrating machine learning-based virtual screening with multiple protein structures and bioassay evaluation for discovery of novel GSK3β inhibitors. Front Pharmacol. 2020;11:566058.
54. Samad A, Ajmal A, Mahmood A, Khurshid B, Li P, Jan SM, Rehman AU, He P, Abdalla AN, Umair M. Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation. Front Mol Biosci. 2023;10:1060076.
55. Hadfield TE, Scantlebury J, Deane CM. Exploring the ability of machine learning-based virtual screening models to identify the functional groups responsible for binding. bioRxiv. 2023:2023.2004. 2029.538820.
56. Luo L, Zheng T, Wang Q, Liao Y, Zheng X, Zhong A, Huang Z, Luo H. Virtual screening based on machine learning explores mangrove natural products as KRASG12C inhibitors. Pharmaceuticals. 2022;15(5):584.
57. Choudhari AS, Mandave PC, Deshpande M, Ranjekar P, Prakash O. Phytochemicals in cancer treatment: from preclinical studies to clinical practice. Front Pharmacol. 2020;10:1614.

Publisher's Note