ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

Virtual Screening, Molecular Docking, and Molecular Dynamic Simulation Methods for the In Silico Prediction of Novel Inhibitors against Kirsten Rat Sarcoma G12D Cancer Drug Target

[1.] **Ms. Sk. Salma Sultana,** Asso Professor, DEPT OF Pharmacology,
[2.] **Y. RATNAKUMARI,** ASSO. PROFESSOR, DEPT OF PHARMACOGNOSY,
[3.] **P. Padmavathi Devi**, Asst. Professor, dept of Pharmaceutics,
[4.] **Mrs. K Venkata Swapna,** asst professor,dept of Pharmacognosy, SWATHI COLLEGE OF PHARMACY,NELLORE
[5.] **Dr. M. SREENIVASULU**, PRINCIPAL,
[1,2,3,5] NARAYANA PHARMACY COLLEGE, CHINTHA REDDY PALEM, NELLORE

Abstract: The vast majority of human cancers are caused by mutations in the Kirsten rat sarcoma (KRAS) viral proto-oncogene. A significant portion—approximately 30%—of lung, pancreas, and colon cancers in humans are triggered by oncogenic KRAS mutations. An appealing therapeutic target is one of the most common mutant KRAS G12D mutations, which causes pancreatic cancer. There are currently no medications that have been authorised for use by the FDA that target the KRAS G12D mutation. In light of this, research towards a viable treatment for KRAS G12D must proceed. Discovering new medications is a laborious and costly procedure. Alternatively, in silico drug development approaches save time and money. In this study, we used ML methods including K-nearest neighbour (KNN), support vector machine (SVM), and random forest (RF) to find novel inhibitors for the KRAS G12D mutant. Based on the predictions, 82 hits were active against the KRAS G12D mutant. Docking the active hits into the KRAS G12D mutant's active site was the process. In addition, the stability of the compounds with strong docking scores was assessed by running 200 ns MD simulations on the top two complexes and the reference complex (MRTX-1133). As compared to the conventional compound, the top two hits demonstrated great stability. In comparison to the gold standard compound, the binding energies of the top two hits were respectable. Our discovered hits may aid in the fight against cancer by blocking the KRAS G12D mutation. We are unaware of any previous research that has used molecular docking, molecular dynamics simulation, virtual screening based on machine learning, and the KRAS G12D mutant to find potential novel inhibitors.

Keywords: KRAS G12D; machine learning-based virtual screening; molecular docking; MD simulations

1. Introduction

Among the leading causes of death on a worldwide scale, cancer ranks high [1]. The United States is expected to have 1,958,310 new cases of cancer and 609,820 cancer-related deaths in 2023 [2]. Worldwide, around 7% of cancer cases are caused by radiation, germs, or viruses [3]. To produce oncogenes, one has to make certain genetic changes, such as a point mutation, deletion, or amplification [4]. The majority of cancers are caused by mutations in genes that oversee cell division and differentiation. Another factor that may lead to cancer is a mutation in the KRAS gene [5]. Located on chromosome 12, KRAS is a gene that belongs to the RAS superfamily. By alternating between its active (GTP-bound) and inactive (GDP-bound) states, KRAS regulates a number of signal transduction pathways. Among these signal transduction cascades is the RAF-MEK-ERK pathway [6]. The three RAS proteins—KRAS4A, KRAS4B, HRAS, and NRAS—are encoded by the three genes (HRAS, NRAS, and KRAS) [7]. Because of a discrepancy in the C-terminal region, the alternative splicing of exon 4 produces two isoforms, KRAS4A and KRAS4B [8]. In contrast to viral KRAS, which is expressed at lower levels, KRAS4B is the most abundant isoform in human cells [9]. Cancer in humans is most often caused by single-point mutations in the KRAS gene. Thirty percent or more of human malignancies in the liver, colon, pancreas, thyroid, and lungs are attributed to oncogenic KRAS mutations [10]. G12 accounts for the vast majority of cancer-promoting KRAS mutations (89%), whereas codons 12, 13, and 61 are common locations for these changes. The three most common KRAS mutations are KRAS G12D (36%), KRAS G12V (23%), and KRAS G12C (14%). Drug research attempts aim to target the G12D mutation, which is responsible for pancreatic cancer [12]. The structural resistance of KRAS to small-molecule alteration has been shown to be high due to the absence of binding pockets [13]. Medications that target the KRAS G12D mutation have not been authorised by the FDA as of yet. Patients with advanced solid tumours linked to the KRAS G12D mutation are being studied in clinical trials for one of Mirati MRTX1133's medicines, albeit [14]. Developing new drugs takes a lot of effort and money. A budget of $2 billion and a time frame of 10-15 years are possible [14]. On the other hand, in silico methods for drug design are quick and cheap [15]. The use of computer-assisted drug discovery (CADD) technologies has had a substantial impact on the drug development process [16]. The efficiency of lead finding in pharmaceutical research has been greatly enhanced by these in silico methods and the development of supercomputing capabilities [17]. In order to find novel lead compounds, artificial intelligence (AI) and machine learning methods are often used [18,19]. Using AI and ML techniques considerably improves the screening and development of novel lead compounds that attach to therapeutic drug targets [20]. New inhibitors that show promise for the KRAS G12D mutant are the focus of the current research. To find more potential hits in the ZINC database for the KRAS G12D cancer treatment target, we used several machine learning techniques. From the ZINC database, drug-like molecules were chosen using Lipinski's rule of five. The KRAS G12D mutant was docked with the drug-like compounds. We ran a 200 ns simulation on the complexes that had the best docking scores. According to the results of the MD simulation, the recently discovered hits were more stable. These novel hits may be inhibitors of the KRAS G12D protein, according to the results, which might have significant implications for cancer therapy.

2. Results

2.1. *Preparation of Dataset*

From the binding databank database, a total of 2526 compounds with reported IC50 values for KRAS G12D were obtained. Those compounds for which the IC50 value was not reported were removed from the dataset. The compounds were labeled as active or inactive based on the IC50 value of the standard compound MRTX1133 (6.1 nM) [21]. The active and inactive compounds in the dataset were denoted by the labels 1 and 0, respectively. The compound with an IC50 value less than or equal to the reference was labeled as active while the compound with an IC50 value higher than the reference was labeled as inactive. In our dataset, 422 compounds were found as active while the remaining were labeled as inactive. MOE (2016) software was employed to compute 208 2D descriptors in total. To prevent overfitting and improve the model's generalizability, the dataset underwent preprocessing to eliminate any zero and NA values. After preprocessing, there were only 172 descriptors left.

2.2. *Optimum Features Selection*

Currently, the SVM uses three different sorts of methods—filter, wrapper, and embedding approaches—to determine the importance of variables in the dataset. Within the realm of wrapper methods, RFE stands as the gold standard [22]. In order to pick the most relevant features for our investigation, we used recur-sive feature elimination (RFE). The following 57 characteristics out of 172 were determined to be optimal: weinerPath, PEOE_VSA+2, weight, Q_VSA_HYD, Q_VSA_POS, vdw_area, vdw_vol, vsa_hyd, and 57 more. The following datasets were chosen: SlogP_VSA0, PEOE_VSA+0, SMR_VSA6, SlogP_VSA3, Zagreb, TPSA, SMR_VSA1, SlogP_VSA7, PEOE_VSA-4, a_IC, SMR_VSA5, PEOE_VSA-0, vsa_pol, b_single, b_heavy, bpol, PEOE_VSA-1, a_heavy, SMR_VSA2, diameter, logP, weinerPol, and others. The ideal curve for selecting features is shown in Figure 1. In order to improve the performance of each machine learning model, we trained them using optimal feature subsets.

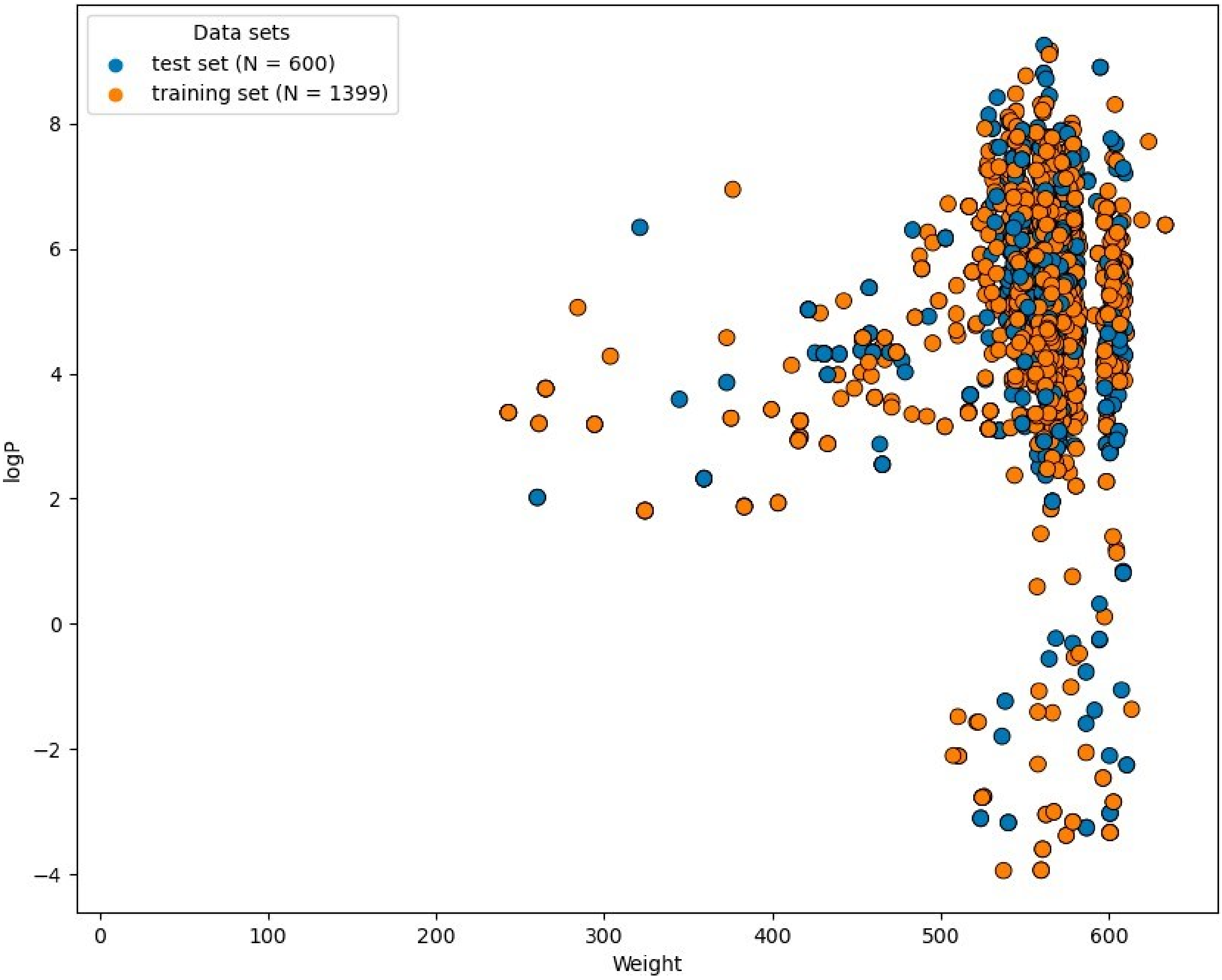


Figure 1. The chemical space and diversity distribution of the dataset. The scatter plot indicates the average results from the cross-validation. The molecular weight and LogP are shown on the X and Y axes, respectively.

2.3. *Chemical Space and Diversity*

The chemical diversity of a dataset significantly affects the reliability of the ML algo- rithm. Adequate chemical space is needed for model performance [23]. The significant chemical gap between logP and molecular weight (MW) is shown in Figure 1. A substantial chemical gap between active and inactive inhibitors was observed, with logP ranging from −4 to 8 and MW ranging from 250–600 Da, respectively.

2.4. *Performance Evaluation of Models*

A number of supervised ML models were trained using Python v3.9, including KNN, SVM, and RF. To evaluate the efficacy of each model, many metrics were calculated, including accuracy, sensitivity, specificity, and MCC. With an MCC value of 0.96 and an accuracy of 99%, the RF model was determined to be the best model out of all of them. Second place went to the KNN model in terms of accuracy and MCC value. The KNN model achieved an accuracy of 98% and an MCC of 0.94. Third place went to the SVM model, which achieved 96% accuracy and an MCC value of 0.90. All of the models' performance evaluations are shown in Table 1. We employed five-fold cross-validation to ensure the reliability of the findings. When evaluating the efficacy of a model, one of the most trustworthy approaches is to examine the ROC-AUC curve. As shown in Figure 2, the RF model achieved an area under the curve (AUC) value of 0.99, which was higher than the KNN and SVM models, which had AUC values of 0.98 and 0.95, respectively.

Table 1. Performance evaluation of machine-learning models.

Models	Accuracy	Sensitivity	F1 Score	MCC
KNN	98	0.99	0.95	0.94
SVM	96	0.93	0.92	0.90
RF	99	0.94	0.96	0.96

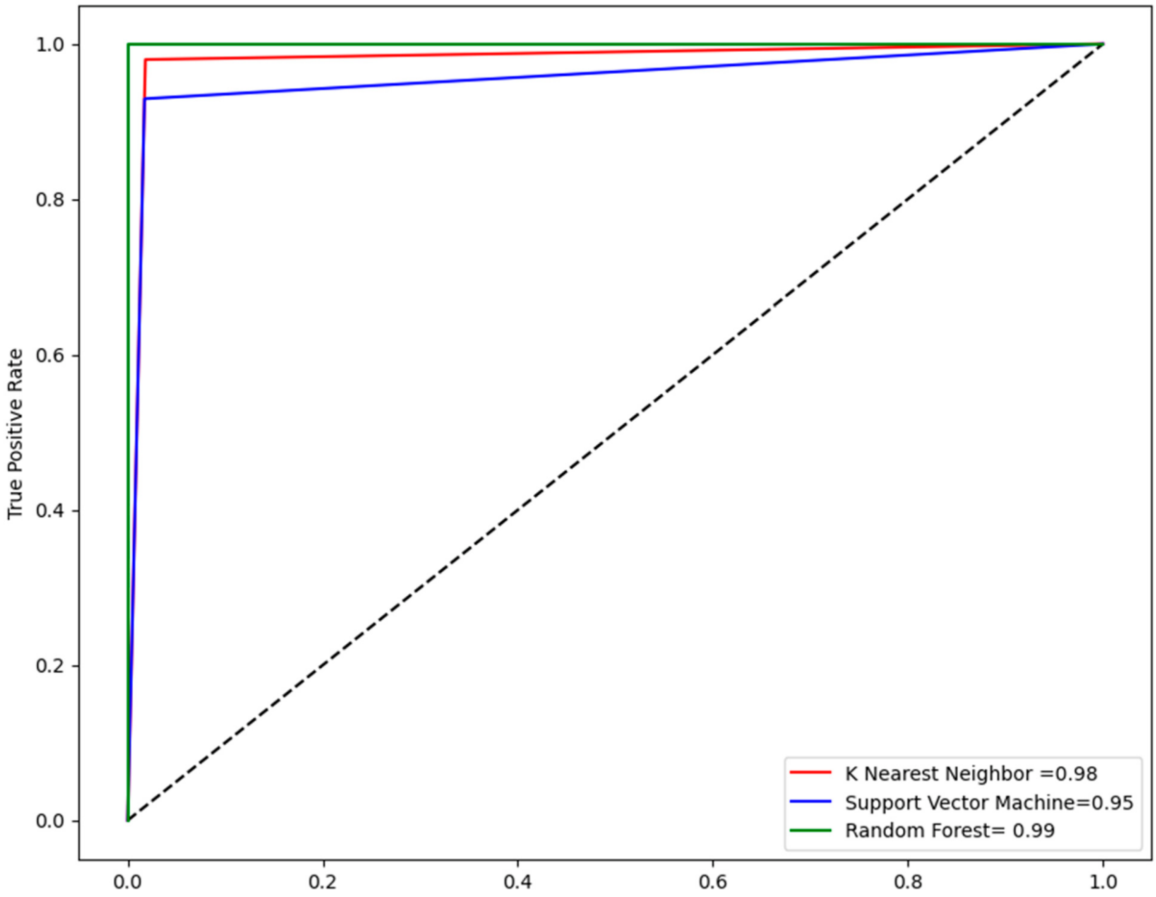


Figure 2. The ROC-AUC curve developed in Python v3.9 shows the TP against the FP rate on the cross-validation.

2.5. Virtual Screening

A total of twenty thousand drug-like compounds were virtually screened using the RF model, which stood out among the ML algorithms due to its high accuracy and MCC score. The compounds were collected from the ZINC database. Based on the predictions, 82 hits were active against the KRAS G12D mutant. We deleted 10 of the 82 hits that were shown to be harmful from the database and docked the remaining chemicals that were determined not to be toxic against the KRAS G12D mutant.

2.6. Molecular Docking Study

A total of seventy-two hits docked with the KRAS G12D mutant's active site. The docking study showed that the majority of the newly found hits interacted well with the KRAS G12D mutant and had excellent docking scores. For the docking investigation, the control molecule was chosen as MRTX-1133. With a docking score of -7.91 (kcal/mol), compound ZINC05524764 was determined to be the most promising. There is one ionic contact with the Glu62 residues of KRAS G12D and five hydrogen bonds with Asp92, Asp12, His95, and Gly60 that are formed by compound ZINC05524764. With a docking score of -6.85 (kcal/mol), compound ZINC05828661 was determined to be the second most potent of the compounds tested. The active site residues Asp12, Lys16, Ala59, and Arg68 were all contacted by compound ZINC05828661 via six hydrogen bond interactions. The chemical ZINC05725307 was anticipated to have a docking score of -6.70 (kcal/mol). The KRAS G12D receptor residues Asp12, Arg102, Lys16, Ala59, and Arg68 were all involved in interactions with compound ZINC05725307, which included three hydrogen bond contacts, one ionic interaction, one arene-H interaction, and one arene-cation interaction. Among the residues found in the active site of KRAS G12D, the control compound MRTX1133 formed four hydrogen bonds with Asp12, Glu62, and His95, and one arene-cation interaction with Arg68. The most promising findings from the ZINC database are shown in Table 2 together with their docking scores and interactions. In Figure 3, we can see the three-dimensional interactions between the reference chemical and the most interesting compounds.

Table 2. Docking score and interactions of the most potent compounds of ZINC database.

Zinc ID	Interacting Residues	Interaction Type	Distance (Å)	Energy (kcal/mol)	S Score (kcal/mol)
ZINC05524764	GLU 62	H-bond	3.30	−2.0	−7.91
	ASP 92	H-bond	3.13	−1.8	
	ASP 12	H-bond	3.02	2.1	
	HIS 95	H-bond	2.96	−2.8	
	GLY 60	H-bond	3.23	−3.5	
	GLU 62	Ionic	3.72	−1.1	
ZINC05828661	ASP 12	H-bond	3.01	−2.6	−6.85
	LYS 16	H-bond	3.15	−1.7	
	Ala 59	H-bond	3.25	0.8	
	ASP 12	H-bond	3.30	−0.5	
	ARG 68	H-bond	3.20	−2.6	
	ARG 68	H-bond	3.23	−1.5	
ZINC05725307	ASP 12	H-bond	2.88	−1.6	−6.70
	ARG 102	H-bond	2.88	−5.1	
	LYS 16	H-bond	3.33	0.9	
	LYS 16	Ionic	2.78	−6.2	
	ALA 59	Arene-H	4.12	−0.6	
	ARG 68	Arene-cation	4.83	−0.8	
ZINC17004657	GLN 61	Arene-H	3.88	−1.1	−5.68
	ASP 12	H-bond	2.98	1.0	
	ASP 12	H-bond	3.05	−1.2	
	LYS 16	H-bond	3.30	−1.0	

Table 2. *Cont.*

ZINC18169629	GLN 61	H-bond	3.09	−0.6	−6.19
	HIS 95	H-bond	2.91	−6.2	
	GLY 60	H-bond	3.26	−1.0	
	LYS 16	H-bond	3.13	−3.0	
	ALA 59	Arene-H	4.03	−1.2	
	GLY 60	Arene-H	4.39	−0.6	
	THR 58	Arene-H	4.02	−0.8	
ZINC22760692	GLU 63	H-bond	3.20	−1.1	−6.51
	HIS 95	H-bond	3.24	−0.8	
	ARG 68	H-bond	3.12	−0.5	
	GLY 10	H-bond	3.10	−0.5	
	LYS 16	H-bond	3.16	−0.8	
	MET 72	Arene-H	4.17	−0.6	
Control	GLU 62	H-bond	3.29	−1.4	−5.39
	GLU 62	H-bond	3.30	−0.7	
	ASP 12	H-bond	2.64	−3.1	
	HIS 95	H-bond	2.77	−3.0	
	ARG 68	Arene-cation	4.72	−0.7	

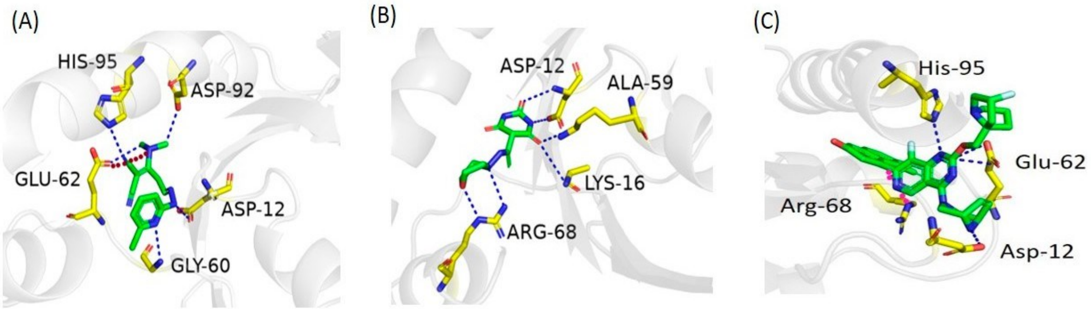


Figure 3. Three-dimensional interactions of (**A**) ZINC05524764, (**B**) ZINC05828661, and (**C**) the control compound with the KRAS G12D mutant. The blue dotted lines indicate hydrogen bonds, the red dotted line indicates the ionic bond, and the pink dotted line indicates the arene-cation bond, while ligands are shown as green sticks.

2.7. *Docking Validation*

The docking procedure was validated by removing the co-crystal ligand (PDB ID: 7RPZ) and then re-docking it into the active site using MOE (2016) software [23]. The RMSD value between the top-ranked docked conformation and the co-crystallized ligand was predicted to be 0.148 Å (Figure 4), revealing the validity of the MOE docking protocol.

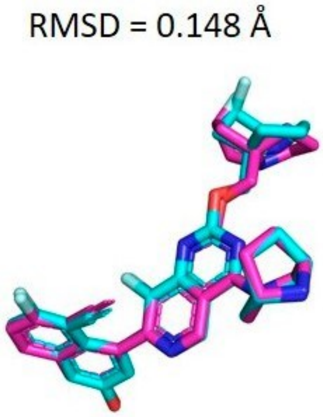


Figure 4. Superposition of co-crystallized and docked conformations of the ligand. The magenta color represents the native co-crystallized ligand and the cyan color is the docked ligand.

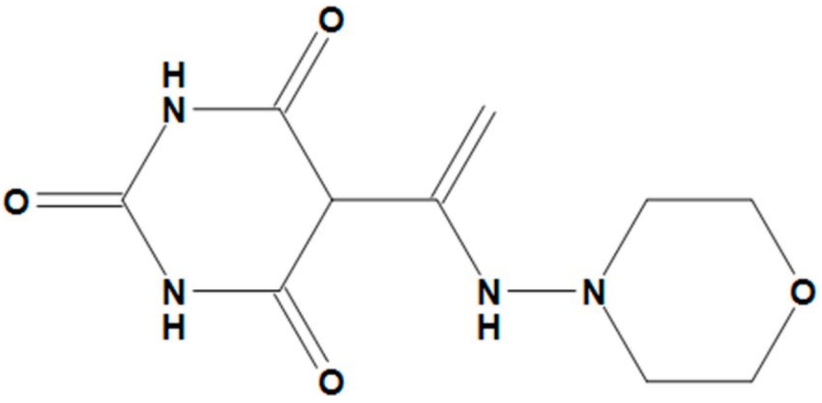
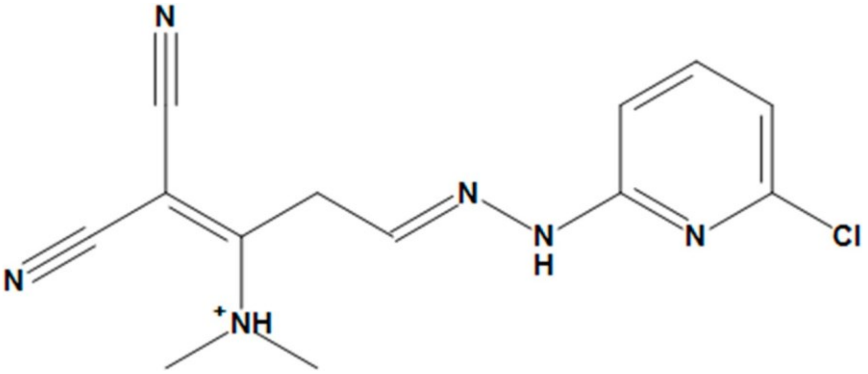
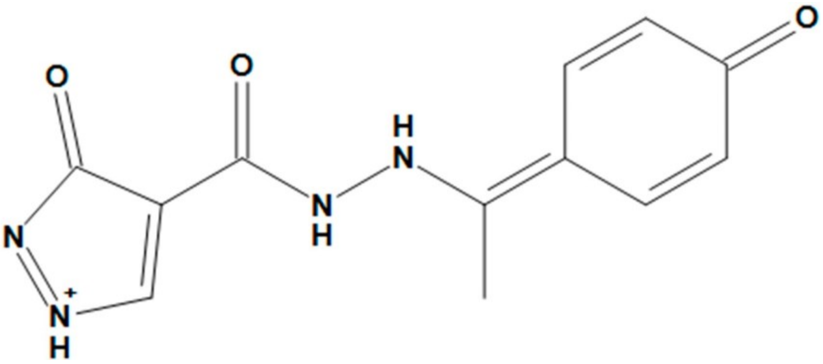
2.8. Drug-Likeness and Toxicity Analysis of the Compounds

In evaluating the drug-likeness of the compounds, one widely accepted criterion is the Lipinski rule of five. In this study, the MOE software was employed to calculate the drug-likeness of the compounds. The Lipinski rule of five for the most promising compounds is present in Table 3. All the compounds obeyed the Lipinski rule of five. Our newly identified compounds against the KRAS G12D target possess drug-likeness. Furthermore, the virtual toxicity of the compounds was evaluated by using the MOE software. All the compounds were predicted non-toxic as presented in Table 4.

Table 3. Drug-likeness of the compounds.

Compound ID	M-Weight	HB-Donor	HB-Acceptor	logP
ZINC05524764	254.25	3	5	−1.41
ZINC05828661	289.75	2	4	0.13
ZINC05725307	259.24	3	4	0.41

Table 4. Two-dimensional structures and toxicity analysis of the most promising compounds.

Compound ID	2D Structure	Toxicity
ZINC05828661		No
ZINC05524764		No
ZINC05725307		No

2.9. *Post-Simulation Analysis*

2.9.1. RMSD Analysis

Performing MD simulations is one of the most reputable ways to investigate the fundamental stability of protein-ligand interactions. We used root-mean-square (RMSD) analysis to check how stable the complexes were. The RMSD of the KRAS G12D was shown for the 200 ns production simulations and compared to the control complex. At first, the ZINC05524764 complex's RMSD was steady up to 50 ns, but then there were small oscillations from 50 to 55 ns, after which the system converged and remained stable up to 120 ns. The system reached stability at 120 ns, and it stayed that way for the next 200 ns, even though the RMSD climbed progressively until it reached 170 ns. After a steady initial 50 ns, the ZINC05828661 complex's RMSD showed some small fluctuations between 70 and 100 ns, but after that, the system stabilised and stayed that way all the way up to 200 ns, with the exception of 125–175 ns. The two systems' RMSDs were determined to be quite stable during the 200 ns MD simulation, in contrast to the control system. All systems showed a steady RMSD, while the control system exhibited erratic behaviour from 60 to 125 ns. On average, the ZINC05524764, ZINC05828661, and control systems were determined to have RMSD values of 2 Å, 2.1 Å, and 2.5 Å, respectively. All of the complicated systems' RMSD charts are shown in Figure 5. After binding to the KRAS G12D protein, the ligand maintains a constant location inside the binding site, as seen by the low variation of the ligand RMSD. A stable complex that is less prone to dissociate under physiological settings is shown by the low departure of the RMSD ligand from the RMSD complex, which shows a synergistic stability between the ligand and the protein. This finding provides further evidence that ZINC05524764 may inhibit the KRAS G12D protein. Figure S2 displays the complex systems both before and after MD simulation, whereas Figure S1 displays the RMSD ligand plots.

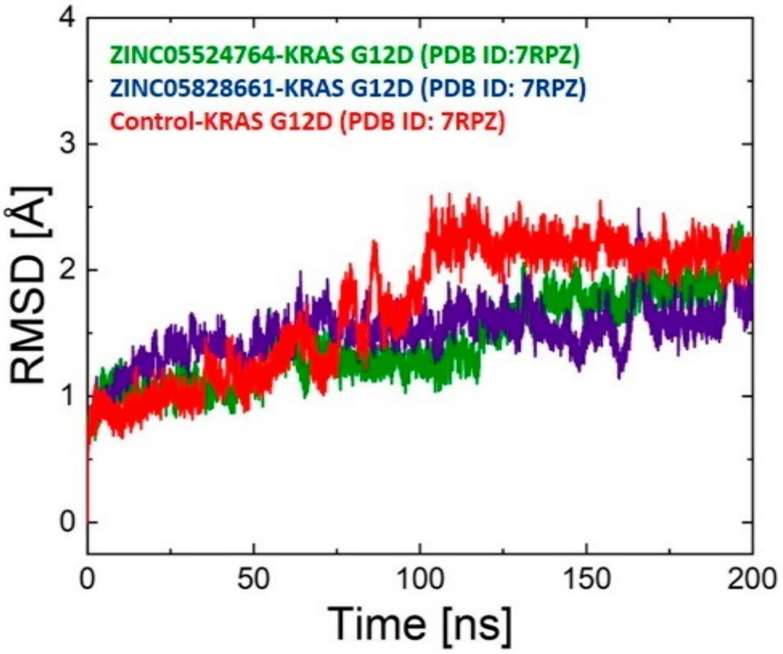


Figure 5. RMSD plot for ZINC05524764 (green), ZINC05828661 (purple), and the control (red) systems. Time in ns is shown on the X-axis and the RMSD value of each system is shown on the Y-axis.

2.9.2. RMSF Analysis

The root mean square fluctuation (RMSF) allowed for a more thorough examination of the protein's backbone flexibility. The RMSF plots for all the complexes are shown in Figure 6. The loop regions had the highest variations, with an overall comparable tendency in the fluctuations. Residues Asp30, Glu31, Tyr32, Asp33, Pro34, Thr35, Ile36, Ser65, Ala66, Met67, Arg68, and Asp69 revealed high fluctuations during MD simulation. Conversely, a decrease in flexibility was noted in the region where the inhibitor was bound, indicating the impact of inhibitor interactions with the active site residues of KRAS G12D.

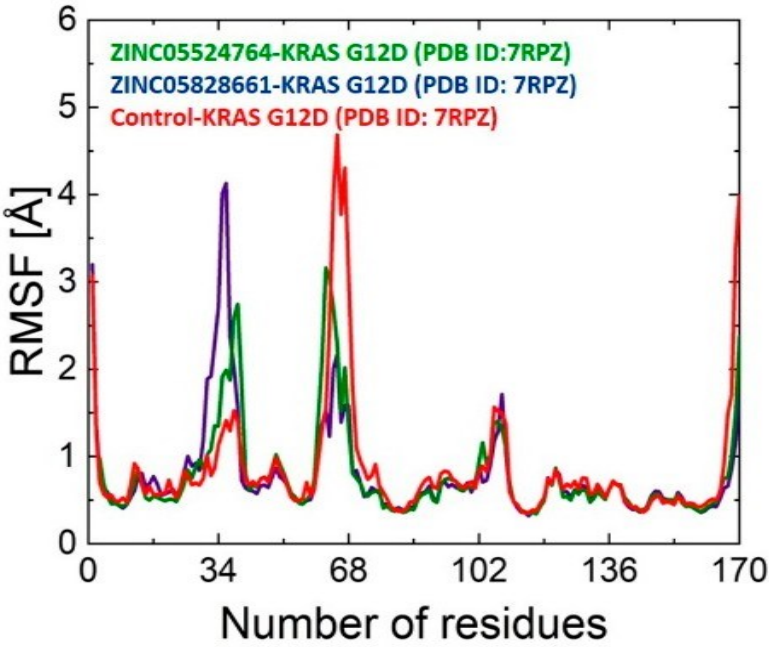


Figure 6. RMSF plot for ZINC05524764 (green), ZINC05828661 (purple), and the control (red) systems. The number of residues is displayed on the X-axis and the RMSF value of each system is present on the Y-axis.

2.10. Structure Compactness Analysis

We determined the binding and unbinding processes that occurred throughout the simulation by calculating the structural compactness in a dynamic situation. The structural compactness was assessed by plotting the radius of gyration (Rg) against time. Figure 7 shows that the Rg of ZINC05828661 followed a pattern comparable to that of RMSD. The complex initially showed low Rg values for a brief duration in the first 50 ns. Following then, the Rg value rose to 15.9 Å, fell back down, and maintained a steady pattern all the way up to 200 ns. The green ZINC05524764 system had an average Rg value of 15.2-15.6 Å, the ZINC05828661 system had a Rg value of 15.1-15.8 Å, and the control system had a Rg value of 15.3-15.7 Å. All of the systems' Rg charts are shown in Figure 6.

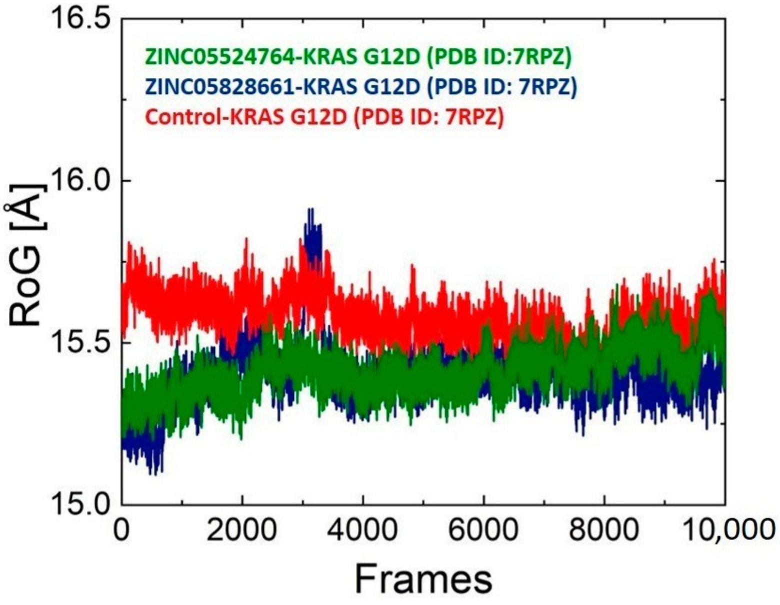
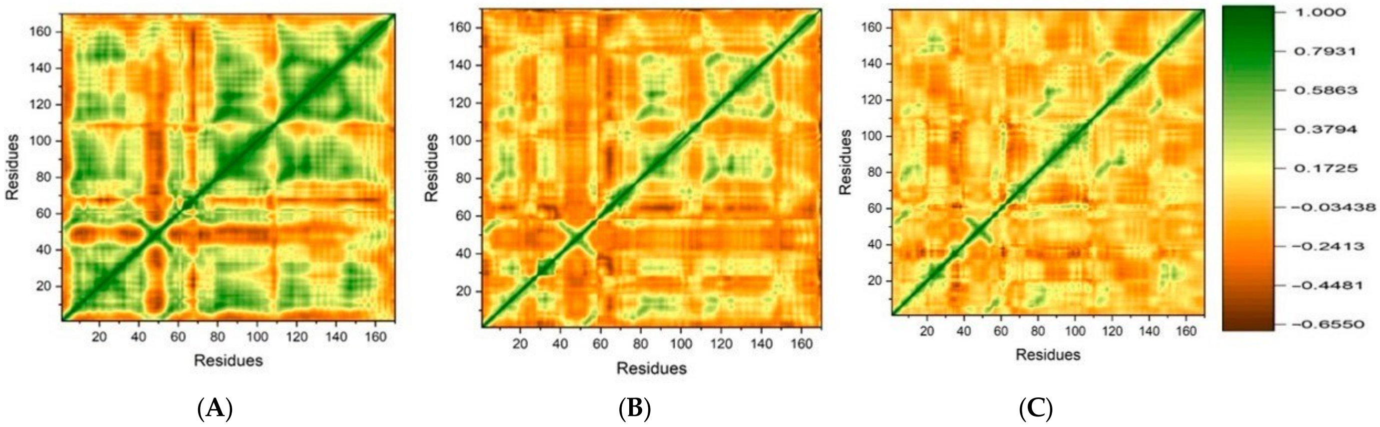


Figure 7. RoG plot for ZINC05524764 (green), ZINC05828661 (purple), and the control (red) systems. The number of frames and the RoG value are presented on the X and Y axis.

DCCM Analysis

In order to learn about correlated movements in the MD simulation, we used the dynamic cross-correlation map (DCCM) to calculate the correlation among receptor residues. The relationships among the residues in the systems were investigated using DCCM, an inter-residue correlation study. The DCCM findings for each of the complicated systems are shown in Figure 8. The amino acid movements seemed to be positively linked, suggesting a significant association with correlated motions. If the amino acids' motions are going in the other way, we have anti-correlations of motion. The positive correlations between the systems' residues are represented by the anti-parallel direction, whereas the negative correlations are represented by the parallel direction [24]. The graphs reveal that the residues are negatively correlated (dark brown area) and positively correlated (green sections). When compared to the control system, ZINC05524764 and ZINC05828661 showed more favourable relationships.



(A) (B) (C)

Figure 8. DCCM plot for the (**A**) ZINC05524764, (**B**) ZINC05828661, and (**C**) control systems. The X and Y axis shows the number of residues.

2.11. *Binding Energy Calculation*

Using the binding free energy method, or MM-GBSA, to measure the binding strength of small molecules is a frequently used technique to confirm the ligand binding and docking stability. In terms of calculation, the MM-GBSA approach which was previously reported is less expensive and, as compared to the rational scoring functions, is one of the most accurate techniques [25]. We also used this method to determine the binding free energy for the ZINC05524764, ZINC05828661, and control complexes, keeping in mind its applicability. Total binding free energy (TBFE) estimates for the ZINC05524764 complex were −39 kcal/mole, for the ZINC05828661 complex the binding energy was calculated as −35 kcal/mole, and for the control system, the binding free energy was found as
−30 kcal/mole. Table 5 shows the results of the MMGBSA analysis.

Table 5. MMGBSA analysis indicating the binding energy of all the complexes.

Complex	vdW	EEL	ESURF	EGB	∆*G* TOTAL
ZINC05524764-KRAS^{G12D}	−48.7803	−9.8255	−5.8669	25.3835	−39.0880
ZINC05828661-KRAS^{G12D}	−42.7893	−5.4652	−4.8129	17.8249	−35.2418
Control-KRAS^{G12D}	−26.6921	−29.9760	−4.5080	30.4723	−30.7021

3. Discussion

In the United States, pancreatic ductal adenocarcinoma (PDAC) is ranked as the second leading cause of cancer-related mortality. Because there aren't many treatment options for metastatic PDAC, the 5-year survival rate is around 5% [26,27]. Oncogenic transformation relies on RAS gene activation, which is often associated with human malignancies due to missense mutations in KRAS, HRAS, and NRAS [28]. It was previously believed that oncogenic RAS proteins were unfixable because they lacked binding sites that small-molecule inhibitors might bind to [29]. G12D mutations account for 35% of KRAS mutations at codon 12, with G12V variants accounting for 20-30%, G12R mutations for 10-20%, Q61 mutations for ~5%, G12C mutations for 1-2%, and other unusual mutations accounting for the remaining mutations. [30] Both sotorasib (AMG510) and adagrasib (MRTX849) have been given the green light by the FDA to treat advanced lung cancer patients with a KRASG12C mutation. Furthermore, a KRAS G12D inhibitor, MRTX 1133, is already participating in a phase 1 clinical study after showing promising results in preclinical testing. There are currently no medications for the KRAS G12D mutation that have been authorised by the FDA. Research into a novel, effective treatment for KRAS G12D is, hence, urgently required [31]. The use of various machine learning algorithms for the purpose of drug development has been tremendously beneficial to the pharmaceutical sector. Common applications of these algorithms include bioactivity prediction, drug-protein interaction analysis, and compound safety and bioactivity improvement [32]. A lot of research has employed ML-based virtual screening to find novel inhibitors for various pharmacological targets [33,34].

This research employed a variety of machine learning techniques to scour the ZINC database for novel, promising hits targeting the KRAS G12D cancer treatment target. Ten hits were determined to be harmful out of eighty-two that were anticipated as active. After these dangerous substances were extracted, the remaining hits were docked into the KRAS G12D active site. The molecular docking study validated six compounds as the top contenders for KRAS G12D inhibitors. Inhibitors of the KRAS G12D mutant that showed promise in an earlier investigation were resveratrol, quercetin, and psoralidin. Hydraulic bonds were established by these inhibitors with the Gly10, Thr58, Asp69, Tyr96, Gln61, Glu62, Tyr64, Met72, and Arg68 residues of the KRAS G12D active site [35]. In addition to interacting with Gly10, Asp12, Lys16, Thr58, Glu62, Gly60, Arg68, Met72, and His95, our inhibitors showed promise in binding to the active site residues. In order to ascertain the stability of the top two complexes and the standard complex, a 200 ns MD simulation was conducted after molecular docking. According to the results of the RMSD study, these compounds are effective inhibitors of KRAS G12D, as the discovered hits showed stable binding to the protein. In line with the RMSD profile, the RoG analysis further supported the stability of the ZINC05524764 complex compared to all other complexes. In addition, the two complexes had much higher binding energies than the control complex, as shown by the MMGBSA study.

4. Materials and Method

4.1. Dataset Preparation

A total of 2526 compounds for the KRAS G12D mutant found in the Binding DB were extracted. MRT1133 was considered as the standard compound. The standard compound's IC50 value was found to be 6.1 nM [21]. Based on the IC50 value, the compounds were divided into active and inactive categories. For 526 compounds, the IC50 value was not reported so these were removed. A total of 1578 compounds were categorized as inactive because their IC50 value exceeded that of the reference compounds, while 422 compounds were considered active because their IC50 value was equal to or less than that of the reference compound. In the target class, the active and inactive compounds were indicated by 1 and 0, respectively.

4.2. Features Extraction and Dataset Cleaning

The experimentally validated compounds against the KRAS G12D mutant were obtained from Binding DB. Then, descriptors were calculated in MOE (2019) software [36]. A total of 206 features were computed by MOE software. All the 0 and null (NA) values were removed from the dataset using python v3.9. The dataset cleaning was carried out using the pandas library of python [37]. Then, the dataset was split into training (70%) and test (30%) subsets. The train_test_split function was used to divide the dataset into training and test sets [38].

4.3. Feature Selection

To develop a computationally inexpensive model and to improve model performance, optimum features selection was carried out. We employed SVM-RFE to choose optimum features for model development [39].

4.4. ML Models

Using open-source Python v3.9, three models such as the k-nearest neighbors, support vector machine, and random forest models were developed. All the models were developed using the scikit-learn package of the Python software v3.9 [23].

4.5. K-Nearest Neighbor (kNN)

The k-nearest neighbors (KNN), also known as a lazy algorithm, can solve the problems of classification as well as regression. First, the distance between the nearest neighbors in the data can be measured [40]. The parameter n_neighbors can be used to select the nearest neighbors [41]. The optimal k value was found to be 11.

4.6. Support Vector Machine (SVM)

The SVM model can tackle the problems of regression and classification [42]. Apart from binary classification, SVM can address multiclass classification problems. SVM classifies data with the help of an optimum hyper-plane. Various kernel functions (linear, polynomial, sigmoid, and radial base functions) are used to convert low-dimensional data into a higher dimensional space [43]. The grid search method and RBF were employed to predict the optimal values for the C and γ parameters. Finally, C = 1000 and γ = 1 were found to be the ideal values.

4.7. Random Forest (RF)

The RF algorithm was first presented by Breiman [44]. It is a favored model for data categoriza- tion or regression tasks. A bootstrap sample is used to train the random forest tree, and predictions are made by the majority vote of the trees. Max_features and n_estimators, which indicate the number of trees built before predictions, were the two main hyperparameters that were optimized during model development [41]. Some 100 to 500 estimates were taken during model generation.

4.8. Models Validation and Performance Evaluation

In the case of unbalanced datasets, accuracy alone is not sufficient to access the strength of a classification model [45]. In the case of binary classification problems, the MCC parameter can be used to evaluate the performance of a model. The receiver operating characteristic (ROC) curve is another useful tool for evaluating the models' performance. A ROC curve can be used to visually represent the true positive rate against the false positive rate [46]. For ML model evaluation, several parameters were calculated, including accuracy, F1 score, MCC score, and ROC curves. We employed five-fold cross-validation in this study.

4.9. Virtual Screening and Molecular Docking Study

Virtual screening of the 20,000 drug-like chemicals in the ZINC database was conducted using the model that demonstrated good accuracy and MCC values [47]. Docking the RF model's hits onto the KRAS G12D mutant was the next step. We used the PDB database to get the 3D model of the KRAS G12D mutant (PDB ID: 7RPZ). Prior to docking, the structure had its water molecules removed [48]. Reduced maximum power was achieved by applying a 0.05 root-mean-square (RMS) gradient. Chemical Computing Group's (MCG) MOE version 2016 software's protein preparation module was used to get the structure ready. The three-dimensional protonation of the KRAS structure was observed. For every hit, a total of ten conformations were produced [49]. Lastly, the docking analysis was conducted using the program PyMOL version 2.5 from Schrödinger in the United States and MOE version 2016 from Chemical Computing Group in Montreal, Quebec, Canada.

4.10. MD Simulation

For 200 ns, the stability and dynamic evaluation of the best complexes were examined using MD simulation using the AMBER version 2022 software (Schrödinger, San Francisco, CA, USA) [24]. The FF19SB force field was used for protein molecules and the GAFF for ligand molecules, respectively, according to [50]. The addition of Na+ ions mitigated the impact of any charge, and the energy reduction process was carried out in two stages, using the conjugate gradient and steepest descent techniques, respectively [51]. The further steps of heating and equilibration were then performed. The next step was to execute the 200 ns production run for every complex. A cutoff distance of 10.0 Å was used to apply the particle mesh Ewald algorithm to the long-range electrostatic interactions [52]. Finally, PMEMD.cuda was used to run the simulations, and the CPPTRAJ package was used to analyse the trajectories [53].

4.11. Binding Free Energy Calculations

The most frequently utilized method in various research studies is the assessment of the potency of small molecule binding by calculating the binding free energy (BFE) using the MM/GBSA approach [54]. We employed the MMPBSA.py script to calculate the binding free energy of the protein–ligand complexes by taking into account 2500 snapshots. To calculate the BFE, the following formula was applied:

$$\Delta G\ bind = \Delta G\ complex - [\Delta G\ receptor + \Delta G\ ligand]$$

The binding energy of a protein, drug, or complex is represented by the symbols ΔG *receptor*, ΔG *ligand*, and ΔG *complex*, respectively, while the overall binding energy is represented by the symbol ΔG *bind* [25].

5. Conclusions

Pancreatic cancer is caused by the KRAS G12D mutation, which is being targeted by efforts to create drugs for cancer. This research aimed to find novel inhibitors of the KRAS G12D mutant using several computational methods. The KARS G12D mutant showed the highest promise for two compounds, ZINC05524764 and ZINC05828661, out of 72 active hits against KRAS G12D. Our compounds showed very good stability throughout the 200 ns MD simulation, especially when compared to the reference compound MRTX 1133. Our discovered hits may aid in the fight against cancer by blocking the KRAS G12D mutation. Results from this research give promise for future medication development targeting the KRAS G12D mutation and its associated cancers. The groundwork is laid for future breakthroughs in drug discovery by this effort. It is also suggested to use in vitro and in vivo methods to assess these drugs' inhibitory capability.

References

1. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef] [PubMed]
2. Parkin, D.M. The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* **2006**, *118*, 3030–3044. [CrossRef] [PubMed]
3. Almasmoum, H. Characterization of Mucin 2 Expression in Colorectal Cancer with and without Chemotherapies. *Vivo Vitr. Study. JUQUMS* **2021**, *7*, 18–22. [CrossRef]
4. Meng, M.; Zhong, K.; Jiang, T.; Liu, Z.; Kwan, H.Y.; Su, T. The c urrent understanding on the impact of KRAS on colorectal cancer. *Biomed. Pharmacother.* **2021**, *140*, 111717. [CrossRef] [PubMed]

5. Chen, J.; Zhang, S.; Wang, W.; Pang, L.; Zhang, Q.; Liu, X. Mutation-induced impacts on the switch transformations of the GDP-and GTP-bound K-ras: Insights from multiple replica Gaussian accelerated molecular dynamics and free energy analysis. *J. Chem. Inf. Model.* **2021**, *61*, 1954–1969. [CrossRef] [PubMed]
6. Favazza, L.A.; Parseghian, C.M.; Kaya, C.; Nikiforova, M.N.; Roy, S.; Wald, A.I.; Landau, M.S.; Proksell, S.S.; Dueker, J.M.; Johnston, E.R. KRAS amplification in metastatic colon cancer is associated with a history of inflammatory bowel disease and may confer resistance to anti-EGFR therapy. *Mod. Pathol.* **2020**, *33*, 1832–1843. [CrossRef] [PubMed]
7. Chakrabarti, M.; Jang, H.; Nussinov, R. Comparison of the conformations of KRAS isoforms, K-Ras4A and K-Ras4B, points to similarities and significant differences. *J. Phys. Chem. B* **2016**, *120*, 667–679. [CrossRef]
8. Cox, A.D.; Der, C.J. Ras history: The saga continues. *Small GTPases* **2010**, *1*, 2–27. [CrossRef]
9. Lam, K.K.; Wong, S.H.; Cheah, P.Y. Targeting the 'Undruggable'Driver Protein, KRAS, in Epithelial Cancers: Current Perspective. *Cells* **2023**, *12*, 631. [CrossRef]
10. Shen, H.; Lundy, J.; Strickland, A.H.; Harris, M.; Swan, M.; Desmond, C.; Jenkins, B.J.; Croagh, D. KRAS G12D Mutation Subtype in Pancreatic Ductal Adenocarcinoma: Does It Influence Prognosis or Stage of Disease at Presentation? *Cells* **2022**, *11*, 3175. [CrossRef] [PubMed]
11. Hofmann, M.H.; Gerlach, D.; Misale, S.; Petronczki, M.; Kraut, N. Expanding the reach of precision oncology by drugging all KRAS mutants. *Cancer Discov.* **2022**, *12*, 924–937. [CrossRef] [PubMed]
12. Nagasaka, M.; Li, Y.; Sukari, A.; Ou, S.-H.I.; Al-Hallak, M.N.; Azmi, A.S. KRAS G12C Game of Thrones, which direct KRAS inhibitor will claim the iron throne? *Cancer Treat. Rev.* **2020**, *84*, 101974. [CrossRef]
13. Kargbo, R.B. Targeting KRAS^{G12D} Mutations: Discovery of Small Molecule Inhibitors for the Potential Treatment of Intractable Cancers. *ACS Med. Chem. Lett.* **2023**, *14*, 1041–1042. [CrossRef] [PubMed]
14. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26*, 80. [CrossRef]
15. Noor, F.; Noor, A.; Ishaq, A.R.; Farzeen, I.; Saleem, M.H.; Ghaffar, K.; Aslam, M.F.; Aslam, S.; Chen, J.-T. Recent advances in diagnostic and therapeutic approaches for breast cancer: A comprehensive review. *Curr. Pharm. Des.* **2021**, *27*, 2344–2365. [CrossRef] [PubMed]
16. Noor, F.; Tahir ul Qamar, M.; Ashfaq, U.A.; Albutti, A.; Alwashmi, A.S.; Aljasir, M.A. Network pharmacology approach for medicinal plants: Review and assessment. *Pharmaceuticals* **2022**, *15*, 572. [CrossRef]
17. Floresta, G.; Zagni, C.; Gentile, D.; Patamia, V.; Rescifina, A. Artificial intelligence technologies for COVID-19 de novo drug design. *Int. J. Mol. Sci.* **2022**, *23*, 3261. [CrossRef] [PubMed]
18. Sadaqat, M.; Qasim, M.; ul Qamar, M.T.; Masoud, M.S.; Ashfaq, U.A.; Noor, F.; Fatima, K.; Allemailem, K.S.; Alrumaihi, F.; Almatroudi, A. Advanced network pharmacology study reveals multi-pathway and multi-gene regulatory molecular mechanism of Bacopa monnieri in liver cancer based on data mining, molecular modeling, and microarray data analysis. *Comput. Biol. Med.* **2023**, *161*, 107059. [CrossRef]
19. Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov. Today* **2022**, *27*, 103356. [CrossRef]
20. Tang, D.; Kang, R. Glimmers of hope for targeting oncogenic KRAS-G12D. *Cancer Gene Ther.* **2023**, *30*, 391–393. [CrossRef] [PubMed]
21. Lin, X.; Yang, F.; Zhou, L.; Yin, P.; Kong, H.; Xing, W.; Lu, X.; Jia, L.; Wang, Q.; Xu, G. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J. Chromatogr. B* **2012**, *910*, 149–155. [CrossRef] [PubMed]
22. Samad, A.; Ajmal, A.; Mahmood, A.; Khurshid, B.; Li, P.; Jan, S.M.; Rehman, A.U.; He, P.; Abdalla, A.N.; 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. [CrossRef]
23. Ajmal, A.; Ali, Y.; Khan, A.; Wadood, A.; Rehman, A.U. Identification of novel peptide inhibitors for the KRas-G12C variant to prevent oncogenic signaling. *J. Biomol. Struct. Dyn.* **2023**, *41*, 8866–8875. [CrossRef] [PubMed]
24. Khan, A.; Randhawa, A.W.; Balouch, A.R.; Mukhtar, N.; Sayaf, A.M.; Suleman, M.; Khan, T.; Ali, S.; Ali, S.S.; Wang, Y. Blocking key mutated hotspot residues in the RBD of the omicron variant (B. 1.1. 529) with medicinal compounds to disrupt the RBD-hACE2 complex using molecular screening and simulation approaches. *RSC Adv.* **2022**, *12*, 7318–7327. [CrossRef] [PubMed]
25. Mizrahi, J.D.; Surana, R.; Valle, J.W.; Shroff, R.T. Pancreatic cancer. *Lancet* **2020**, *395*, 2008–2020. [CrossRef] [PubMed]
26. Rahib, L.; Wehner, M.R.; Matrisian, L.M.; Nead, K.T. Estimated projection of US cancer incidence and death to 2040. *JAMA Netw. Open* **2021**, *4*, e214708. [CrossRef] [PubMed]
27. Moore, A.R.; Rosenberg, S.C.; McCormick, F.; Malek, S. RAS-targeted therapies: Is the undruggable drugged? *Nat. Rev. Drug Discov.* **2020**, *19*, 533–552. [CrossRef] [PubMed]
28. Akkapeddi, P.; Hattori, T.; Khan, I.; Glasser, E.; Koide, A.; Ketavarapu, G.; Whaby, M.; Zuberi, M.; Teng, K.W.; Lefler, J. Exploring switch II pocket conformation of KRAS (G12D) with mutant-selective monobody inhibitors. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2302485120. [CrossRef]
29. Waters, A.M.; Der, C.J. KRAS: The critical driver and therapeutic target for pancreatic cancer. *Cold Spring Harb. Perspect. Med.* **2018**, *8*, a031435. [CrossRef]
30. Yousef, A.; Yousef, M.; Chowdhury, S.; Abdilleh, K.; Knafl, M.; Edelkamp, P.; Alfaro-Munoz, K.; Chacko, R.; Peterson, J.; Smaglo, B.G. Impact of KRAS mutations and co-mutations on clinical outcomes in pancreatic ductal adenocarcinoma. *NPJ Precis. Oncol.* **2024**, *8*, 27. [CrossRef] [PubMed]

31. Patel, L.; Shukla, T.; Huang, X.; Ussery, D.W.; Wang, S. Machine learning methods in drug discovery. *Molecules* **2020**, *25*, 5277. [CrossRef] [PubMed]

32. Sharma, G.; Shukla, R.; Singh, T.R. 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 Inform. Bioinform.* **2022**, *11*, 31. [CrossRef]

33. 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 bio-assay evaluation for discovery of novel GSK3β inhibitors. *Front. Pharmacol.* **2020**, *11*, 566058. [CrossRef]

34. Oyedele, A.-Q.K.; Owolabi, N.A.; Odunitan, T.T.; Christiana, A.A.; Jimoh, R.O.; Azeez, W.O.A.; Titilayo, M.B.-H.; Soares, A.S.; Adekola, A.T.; Abdulkareem, T.O. The discovery of some promising putative binders of KRAS G12D receptor using computer-aided drug discovery approach. *Inform. Med. Unlocked* **2023**, *37*, 101170. [CrossRef]

35. Wadood, A.; Ajmal, A.; Junaid, M.; Rehman, A.U.; Uddin, R.; Azam, S.S.; Khan, A.Z.; Ali, A. Machine learning-based virtual screening for STAT3 anticancer drug target. *Curr. Pharm. Des.* **2022**, *28*, 3023–3032. [CrossRef] [PubMed]

36. Sahoo, K.; Samal, A.K.; Pramanik, J.; Pani, S.K. Exploratory data analysis using Python. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 4727–4735. [CrossRef]

37. Datta, R.; Das, D.; Das, S. Efficient lipophilicity prediction of molecules employing deep-learning models. *Chemom. Intell. Lab. Syst.* **2021**, *213*, 104309. [CrossRef]

38. Akbar, S.; Hayat, M.; Tahir, M.; Chong, K.T. cACP-2LFS: Classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. *IEEE Access* **2020**, *8*, 131939–131948. [CrossRef]

39. Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 218. [CrossRef]

40. 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*, 10653. [CrossRef]

41. Ahmad, I.; Basheri, M.; Iqbal, M.J.; Rahim, A. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. *IEEE Access* **2018**, *6*, 33789–33795. [CrossRef]

42. Halwani, A.A. Development of pharmaceutical nanomedicines: From the bench to the market. *Pharmaceutics* **2022**, *14*, 106. [CrossRef]

43. Denisko, D.; Hoffman, M.M. Classification and interaction in random forests. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 1690–1692. [CrossRef]

44. Akbar, S.; Rahman, A.U.; Hayat, M.; Sohail, M. cACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. *Chemom. Intell. Lab. Syst.* **2020**, *196*, 103912. [CrossRef]

45. Jiao, Y.; Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* **2016**, *4*, 320–330. [CrossRef]

46. Alotaibi, B.S.; Ajmal, A.; Hakami, M.A.; Mahmood, A.; Wadood, A.; Hu, J. New drug target identification in Vibrio vulnificus by subtractive genome analysis and their inhibitors through molecular docking and molecular dynamics simulations. *Heliyon* **2023**, *9*, e17650. [CrossRef] [PubMed]

47. Qazi, S.; Das, S.; Khuntia, B.K.; 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*, 1934578X211031707. [CrossRef]

48. Ullah, H.; Nawaz, A.; Rahim, F.; Uddin, I.; Hussain, A.; Hayat, S.; Zada, H.; Khan, M.U.; Khan, M.S.; Ajmal, A. Synthesis, in vitro β-glucuronidase inhibitory potential and molecular docking study of benzimidazole analogues. *Chem. Data Collect.* **2023**, *44*, 100996. [CrossRef]

49. Ajmal, A.; Mahmood, A.; Hayat, C.; Hakami, M.A.; Alotaibi, B.S.; Umair, M.; Abdalla, A.N.; Li, P.; He, P.; Wadood, A. Computer-assisted drug repurposing for thymidylate kinase drug target in monkeypox virus. *Front. Cell. Infect. Microbiol.* **2023**, *13*, 618. [CrossRef]

50. Muhammad, N.; Khan, R.; Seraj, F.; Khan, A.; Ullah, U.; Wadood, A.; Ajmal, A.; Ali, B.; Khan, K.M.; Nawaz, N.U.A. In vivo analgesic, anti-inflammatory and molecular docking studies of S-naproxen derivatives. *Heliyon* **2024**, *10*, e24267. [CrossRef] [PubMed]

51. He, Y.; Liu, K.; Cao, F.; Song, R.; Liu, J.; Zhang, Y.; Li, W.; Han, W. Using deep learning and molecular dynamics simulations to unravel the regulation mechanism of peptides as noncompetitive inhibitor of xanthine oxidase. *Sci. Rep.* **2024**, *14*, 174. [CrossRef] [PubMed]

52. Korlepara, D.B.; Vasavi, C.S.; Srivastava, R.; Pal, P.K.; Raza, S.H.; Kumar, V.; Pandit, S.; Nair, A.G.; Pandey, S.; Sharma, S.; et al. PLAS-20k: Extended Dataset of Protein-Ligand Affinities from MD Simulations for Machine Learning Applications. *Sci. Data* **2024**, *11*, 180. [CrossRef] [PubMed]

53. Khan, H.; Waqas, M.; Khurshid, B.; Ullah, N.; Khalid, A.; Abdalla, A.N.; Alamri, M.A.; Wadood, A. Investigating the role of Sterol C24-Methyl transferase mutation on drug resistance in leishmaniasis and identifying potential inhibitors. *J. Biomol. Struct. Dyn.* **2023**, 1–14. [CrossRef] [PubMed]