Heliyon 8 (2022) e10495

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

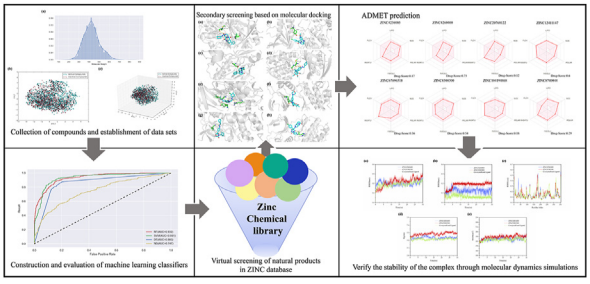
Research article

Multi-stage virtual screening of natural products against p38α mitogen-activated protein kinase: predictive modeling by machine learning, docking study and molecular dynamics simulation

Ruoqi Yang *, Xuan Zha, Xingyi Gao, Kangmin Wang, Bin Cheng, Bin Yan **

Shandong University of Traditional Chinese Medicine, Jinan, 250355, China

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
p38α Mitogen-activated protein kinase
Natural product
Virtual screening
Machine learning
Molecular docking
Molecular dynamics simulation

ABSTRACT

p38α is a mitogen-activated protein kinase (MAPK), and the signaling pathways involved are closely related to the inflammation, apoptosis and differentiation of cells, which also makes it an attractive target for drug discovery. With the high efficiency and low cost, virtual screening technology is becoming an indispensable part of drug development. In this study, a novel multi-stage virtual screening method based on machine learning, molecular docking and molecular dynamics simulation was developed to identify p38α MAPK inhibitors from natural products in ZINC database, which improves the prediction accuracy by considering and utilizing both ligand and receptor information compared to any individual approach. Ultimately, we screened out two candidate inhibitors with acceptable ADMET properties (ZINC4260400 and ZINC8300300). Among the generated machine learning models, Random Forest (RF) and Support Vector Machine (SVM) performed better, with the area under the receiver operating characteristic curve (AUC) values of 0.932 and 0.931 on the test set, as well as 0.834 and 0.850 on the external validation set. In addition, the results of molecular docking and ADMET prediction showed that two compounds with appropriate pharmacokinetic properties had binding free energies less than −8.0 kcal/mol for the target protein, and the results of molecular dynamics simulations further confirmed that they were stable during the process of inhibition.

* Corresponding author.
** Corresponding author.
E-mail addresses: yangruoqia@163.com (R. Yang), robinyan2002@163.com (B. Yan).

https://doi.org/10.1016/j.heliyon.2022.e10495
Received 19 January 2022; Received in revised form 20 March 2022; Accepted 25 August 2022

1. Introduction

Mitogen-activated protein kinases (MAPKs) belong to the serine-threonine protein kinase family and are an important class of signaling systems that widely exist in mammalian cells to mediate cellular responses. MAPKs can be activated by a range of extracellular stimuli (*e.g.,* neurotransmitters, hormones and growth factors) and then respond accordingly through the cascade amplification of phosphorylation [1, 2]. MAPKs are divided into four subfamilies based on their compositions: extracellular signal-regulated kinase1/2 (ERK1/2), c-Jun N-terminal kinase (JNK), p38 MAPK and ERK5, of which p38 MAPK plays a crucial role in apoptosis, inflammatory response and tumor transformation [3, 4, 5]. Human p38 MAPK is a protein with the molecular weight of 38 kDa, which consists of 360 amino acid residues. So far, four family members have been identified: p38α (encoded by MAPK14), p38β (encoded by MAPK11), p38γ (encoded by MAPK12) and p38δ (encoded by MAPK13) [6]. Studies have demonstrated that these four isoforms have remarkably high homology in amino acid sequence, but their distribution in tissues and regions of action are quite different [7]. P38α MAPK, the first independently identified protein of the four members, is considered as the most important isoform of the p38 MAPK family [8].

P38α MAPK has an N-terminal domain formed by β-pleated sheets and a C-terminal domain formed by α-helix, between which an ATP-binding pocket, i. e. the catalytic site of p38α, is formed [9]. Like other protein kinases, the structure of p38α contains multiple cavity-binding sites, including an adenine-binding region, a ribose-binding region, a phosphate-binding region, and two hydrophobic regions, where hydrophobic pocket I consists of amino acids such as TYR-35, ALA-51, LYS-53 and THR-106, and hydrophobic pocket II consists of amino acids such as VAL-30, LEU-108, MET-109 and GLY-110 [10]. Studies have shown that p38α MAPK is closely associated with the development of cellular inflammation, and activation of p38α MAPK can induce the expression of a series of pro-inflammatory factors including TNF-α and IL-6 [11]. At the same time, p38α MAPK has been confirmed to play a critical role in the development of several cancers, such as colon, breast and lung cancers [12]. On the one hand, p38α MAPK can act as a tumor promoter by enhancing the proliferation of tumor cells, and on the other hand, it can also negatively regulate malignancies by inducing apoptosis [13]. In addition, some researchers have found that inhibitors of p38α MAPK could be applied as synergists with chemotherapeutic agents [14].

p38α MAPK is one of the most attractive anti-cancer targets. In recent years, there have been many researches regarding p38α MAPK inhibitors, and those that have been identified can be divided into direct ATP-competitive inhibition, which inhibits the process of activation through occupying the ATP-binding site of p38α MAPK, and indirect ATP-competitive inhibition, which acts on the allosteric site of p38α MAPK to prevent the binding of ATP [15]. Studies have shown that most p38α MAPK inhibitors are ATP-competitive, typically occupying the hydrophobic pocket I and forming at least one hydrogen bond with the amino acid residue MET-109. There are also a few non-ATP-competitive inhibitors that can alter highly conserved DFG sequences (ASP-168, PHE-169, GLY-170) and thus cause conformational changes [16]. From the perspective of chemical structure, these inhibitors are mainly pyridinyl-imidazole derivatives, as well as some bisamide derivatives and urea derivatives.

A prevalent problem with current p38α MAPK inhibitors is their own toxic side effects. Natural products are widely available and have low toxicity, and therefore may be a valuable source for the discovery of p38α MAPK inhibitors. However, natural products that act on p38α MAPK as a target are rarely reported, so the search for active natural products against p38α MAPK has very promising applications. Recently, the identification of active compounds from natural products has become a hot topic in the field of drug discovery, but traditional methods of experimental screening are time-consuming and laborious, which can hardly meet the needs of modern drug development [17]. With the rapid advances in computer technology and chemical simulation theory, diverse computational approaches have been widely adopted in the field of drug discovery. Virtual screening is one of the representative techniques, which can screen a small number of potentially active compounds from a large number of known compounds, thus significantly reducing the amount of compounds for subsequent experimental validation [18].

Virtual screening is divided into two main categories, one is ligand-based virtual screening (*e.g.,* Pharmacophore Modeling and Machine Learning) and the other is receptor-based virtual screening (*e.g.,* Molecular Docking and Molecular Dynamics Simulation) [19]. It is necessary to point out that different virtual screening techniques have both advantages and limitations. The development of a multi-stage screening system can maximize the advantages of each independent screening method and has already yielded many results in the course of drug discovery for a variety of targets. For example, *Mollica et al.* utilized a hierarchical screening protocol to identify lead compounds targeting human α-glucosidase. They first performed a rapid filtering of the database using computational tools and then the candidate compounds with good prediction results were subjected to synthesis as well as in vitro inhibition assays to confirm the reliability of the established strategy [20]. Similarly, *Poli et al.* searched for novel peptide compounds with μ-opioid receptor (MOR) inhibitory activity from an internally constructed tetrapeptide library by applying a series of virtual screening techniques, such as pharmacophore modeling, molecular docking, and molecular dynamics simulations. The three highest scoring hits were synthesized and biologically validated, eventually they identified a peptide compound with anti-agonist action [21]. Furthermore, *Narendra et al.* combined machine learning methods with different virtual screening techniques such as molecular docking and molecular dynamics simulation to screen compounds from several chemical libraries, and ultimately obtained selective inhibitors targeting aldehyde dehydrogenases 1A1 [22].

In this study, we primarily aimed to find p38α MAPK inhibitors with appropriate pharmacokinetic properties. To achieve this objective, we combined different virtual screening strategies. Firstly, four machine learning classification models (Decision Tree, Random Forest, Support Vector Machine, and Naïve Bayes) were developed to virtually screen the natural products in the ZINC database. Subsequently, the interaction modes between the screened hits and the target protein were analyzed via molecular docking, and then ADMET prediction of the compounds with high binding energies was performed. Finally, we carried out molecular dynamics simulation to investigate the binding stability of the selected compounds.

2. Material and methods

2.1. Data curation

We collected the compounds targeting p38α MAPK from the ChEMBL database (ChEMBL260) as the training set and performed preprocessing to remove redundant data and ensure the reliability of the dataset used for modeling. The specific criteria were as follows: (a) Only those compounds targeting human p38α MAPK were retained; (b) Data with assay type of B (*i.e.,* binding directly to the target) were reserved; (c) Compounds with duplicate or no determined activity values (IC_{50}) were excluded. (d) Compounds in the dataset were classified into inhibitors (tagged as 0) and non-inhibitors (tagged as 1) based on a cut-off value of 500 nM for activity. A total of 2282 active compounds and 1340 inactive compounds were utilized for the construction of machine learning models by normalizing the data as described above. The compounds in the dataset were then divided into a training set and a test set in the ratio of 8:2, with the former being typically used for model training and hyperparameter tuning and the latter being commonly adopted for the evaluation of generated models. We also conducted principal component analysis (PCA) on the curated dataset to assess its chemical space characterization. Moreover, to further verify the generalization ability of the models, we gathered 35 known p38α MAPK inhibitors from the relevant

literature and generated the corresponding non-inhibitors by DUD-E database, both of which together form an external validation set [23].

2.2. Molecular representation and feature engineering

Molecular fingerprinting is an approach to describe the characteristics of compounds by converting the molecular structure into a series of binary codes, which indicate the presence or absence of structural fragments at specific locations in a compound by using the numbers 0 and 1. The types of molecular fingerprints commonly employed today include MACCS fingerprints, PubChem fingerprints and Morgan fingerprints [24]. In this study, we chose MACCS fingerprinting as the molecular representation for machine learning, which is MDL keys based on 166 pre-defined structural features and widely used for fast screening of substructures [25]. In the process of developing machine learning models, feature selection is a necessary step given that an excessive number of features can lead to various problems such as overfitting as well as dimensional disasters. In this study, we extracted the optimal amount of features by using the recursive feature elimination (RFE) algorithm in combination with the learning curve [26].

2.3. Construction and evaluation of machine learning classifiers

In this study, four machine learning models (Decision Tree, Random Forest, Support Vector Machine, and Naïve Bayes) were constructed based on the Scikit-learn library of Python. These models used MACCS fingerprints as descriptors and were subjected to the identical training data. The hyperparameters of the models were tuned by Bayesian Optimization algorithm. Decision Tree (DT) algorithm is capable of classifying data according to a series of rules, where each non-leaf node represents the judgment on a single feature, and each leaf node represents one category label [27]. Random Forest (RF) is an ensemble learning algorithm for prediction, which can solve the problem of overfitting in a single DT to some extent by constructing many independent DTs during training [28]. Support Vector Machine (SVM) is a machine learning algorithm based on the principle of structural risk minimization which maps the input data into a multi-dimensional space through the kernel function and builds a hyperplane with maximum margin to classify the data that are linearly inseparable in the low-dimensional space [29]. Naïve Bayes (NB) is a simple algorithm for classification, which is based on the assumption of conditional independence and can measure the probabilistic relationship between labels and features [30].

The performance of every generated model was compared by computing different statistical indicators. The metrics included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). The specific calculation formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP (true positive) is the amount of p38α MAPK inhibitors predicted correctly, TN (true negative) is the amount of p38α MAPK non-inhibitors predicted correctly, while FN (false negative) represents those p38α MAPK inhibitors that have been wrongly classified as non-inhibitors, FP (false positive) represents those p38α MAPK non-inhibitors that have been wrongly classified as inhibitors.

2.4. Molecular docking

The co-crystallized structure of p38α MAPK was obtained from the Protein Data Bank (https://www.rcsb.org/) with PDB ID-1A9U (2.50 Å resolution, 0.240 R-value free and 0.182 R-value work). All water residues and default ligands were separated from the protein molecule, and then hydrogen atoms were added using AutoDockTools software. Additionally, the energy minimization of the target protein was carried out through ModRefiner web tool (https://zhanglab.ccmb.med.umich.edu/ModRefiner). Prior to the docking study, the small molecule ligands were converted to mol2 format via OpenBabel software and subsequently saved into the pdbqt format, which is necessary for molecular docking.

In this study, we used Autodock Vina to perform molecular docking of the compounds screened by machine learning models to the target protein [31]. The parameters of the grid box were set to X = 1.521, Y = 20.841 and Z = 35.906, and the grid spacing was 57.119 × 58.704× 59.282 Å for X, Y and Z coordinates respectively. At the same time, the exhaustiveness was set to 8 to ensure the accuracy of the prediction results. During the entire process of docking, the small molecule ligands were flexible and the receptor proteins remained rigid. The reliability of the docking protocol was verified by calculating the root mean square deviation (RMSD). Furthermore, the interactions between ligands and proteins were visualized via Pymol software [32].

2.5. ADMET prediction

ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties is an essential criterion for the assessment of drug-like nature. With the advancement of computer technology and cheminformatics, researchers in the field of drug discovery have constructed ADMET prediction models using existing ADMET data for drugs and various algorithms. In this study, we performed an online prediction of ADMET properties with the help of SwissADME web server (http://www.swissadme.ch) and OSIRIS Property Explorer software, and then compared the computed parameters with standard ranges to examine whether these values were within the limits.

2.6. Molecular dynamics simulation

The dynamic properties of biomolecules make it inaccurate at times to predict the binding site from a single static structure. In this study, we conducted a 30 ns molecular dynamics simulation of the screened ligand-protein complexes using GROMACS software [33]. Firstly, we created topology files of ligands and proteins by assigning CHARMM 36 force fields [34]. Afterwards periodic boundary conditions were defined, where the type of grid box was set to orthorhombic, and solventization was performed using the TIP3P water model. Following the ionic balance of the system, the steepest descent algorithm was applied to minimize the energy. Next, the optimized system was subjected to 100 ps NVT equilibrium and NPT equilibrium to stabilize it at a suitable temperature and pressure. Finally, the molecular dynamics simulations of all complexes were carried out with a time interval of 2 fs. The results were analyzed by calculating root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg) and solvent accessible surface area (SASA). In addition, 100 frame trajectory files were extracted from the 20–30 ns stable dynamics trajectories and the binding free energies of the protein-ligand complexes were calculated using MM/PBSA [35].

3. Results and discussion

3.1. Chemical diversity analysis

The sample composition of the modeling datasets greatly affects the accuracy of machine learning algorithms. In this study, we applied MACCS fingerprints as molecular descriptors of compounds in the training set and reduced the feature dimensions of 2282 active and 1340

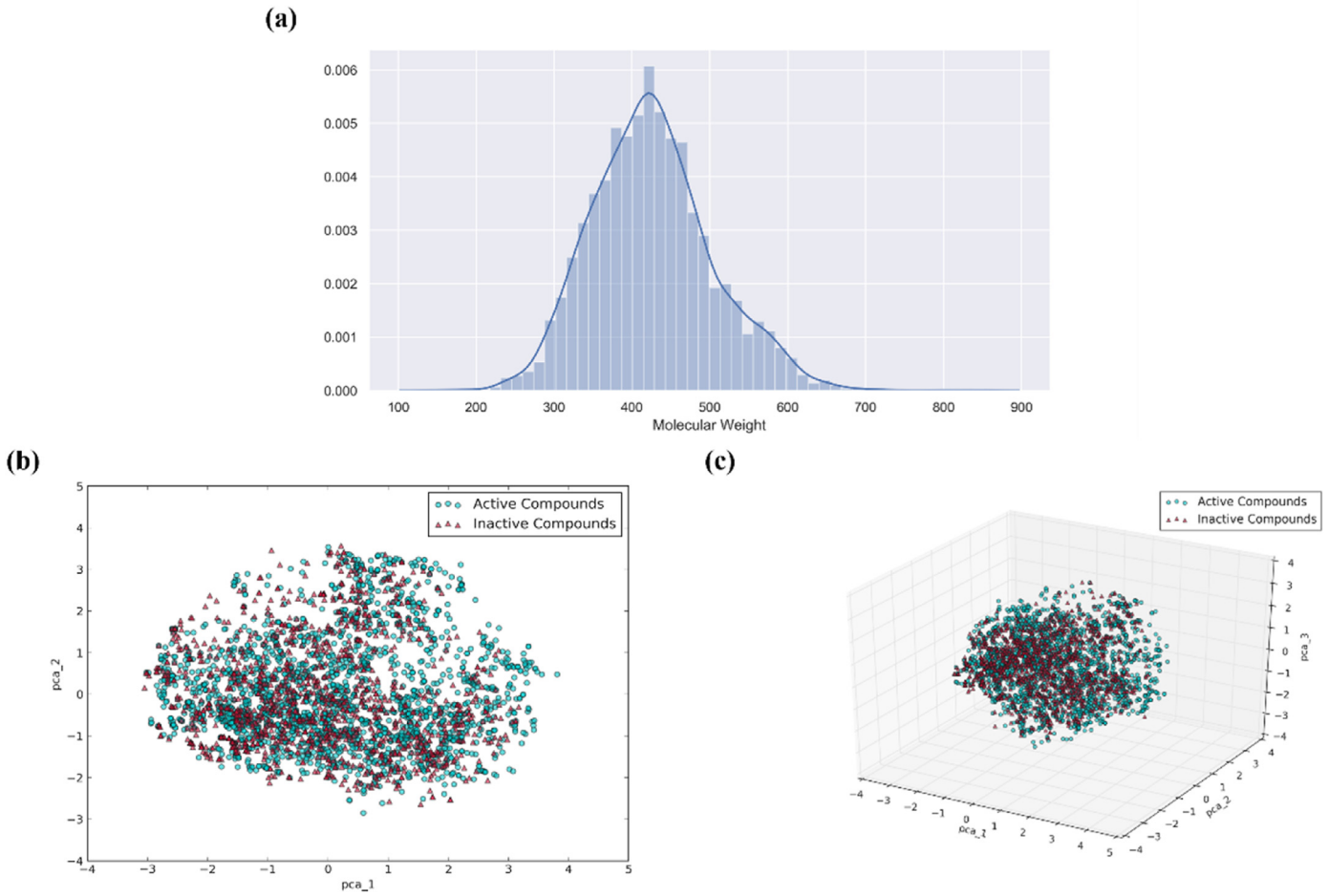


Figure 1. Diversity distribution of the modeling data. (a) The molecular weight histogram. (b) Two dimensional spatial distribution map. (c) Three dimensional spatial distribution map.

inactive molecules to two or three dimensions through PCA. The histogram in Figure 1a showed that the molecular weight of the compounds ranged from 200 to 700 Da. Additionally, it can be seen from the spatial distribution map (Figure 1b, c) that the modeling data had a wide chemical spatial distribution and the model applicability was excellent.

An important issue to be considered in the process of building reliable predictive models is the quality of the datasets. The main advantage of high diversity is that the model based on this data set has great robustness. Our analysis indicated that there exists dissimilarity in the chemical structure for the compounds used in this study.

3.2. Generation and validation of machine learning models

We constructed four prediction models (RF, SVM, DT and NB) for the selected training set and the performances of the developed models were evaluated using 10-fold cross-validation and external independent validation. The results were displayed in Table 1, among 917 compounds in the test set, the quantity of correct classifications (TP + TN) of RF was 780, whereas 794 were for the SVM, 721 for DT and 621 for NB. It is also evident from Table 1 that the RF and SVM classifiers had higher values of accuracy, precision, recall and f1-score for the test set. Furthermore, we plotted the receiver operating characteristic curve (ROC) of different models and calculated the corresponding area under the curve (AUC), which is an important indicator to assess the prediction capability of the generated models, and the closer the AUC value is to 1, the better the classification ability. We can observe from Figure 2 that the AUC values for RF and SVM were 0.932 and 0.931 followed by DT 0.860, and the lowest AUC values for NB were 0.747. These results indicated that the RF and SVM algorithms had greater advantages in this study.

A machine learning model may perform well in the training set, but there was a remarkable difference when validated on an external dataset. To further assess the performance of the models, we made predictions on an independent dataset under the same conditions. The results showed that the RF and SVM models had robust classification capabilities for the compounds in the unknown dataset with an overall accuracy of 78.3% and 79.1%, along with AUC values of 0.834 and 0.850, respectively (Figure 3). These above results indicated that the RF and SVM models performed equally well on the training set and the independent validation dataset.

As a representative of AI technology, machine learning is increasingly deployed in the field of Traditional Chinese medicine, especially after the outbreak of COVID-19, more and more pharmaceutical companies and research institutions are searching for potential active ingredients in natural products by combining machine learning algorithms and virtual screening techniques [36]. In this study, we utilized MACCS fingerprinting to construct four classical machine learning models. Since molecular fingerprints are characterized by high dimensionality and sparsity, which are not conducive to prediction accuracy, we first selected the optimal feature subset through the RFE algorithm. Major machine learning algorithms involve a large number of hyperparameters. In order to obtain the best combination of hyperparameters, we adopted the Bayesian Optimization algorithm to adjust the key parameters of the generated models. Then we found that the RF and SVM models performed better on the test set and external independent validation set by

Table 1. Performance evaluation of four machine learning algorithms using 10-fold cross-validation.

	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-score
RF	393	71	66	387	0.85	0.86	0.85	0.86
SVM	403	61	62	391	0.87	0.87	0.87	0.87
DT	365	99	97	356	0.79	0.79	0.79	0.79
NB	325	139	157	296	0.68	0.68	0.70	0.69

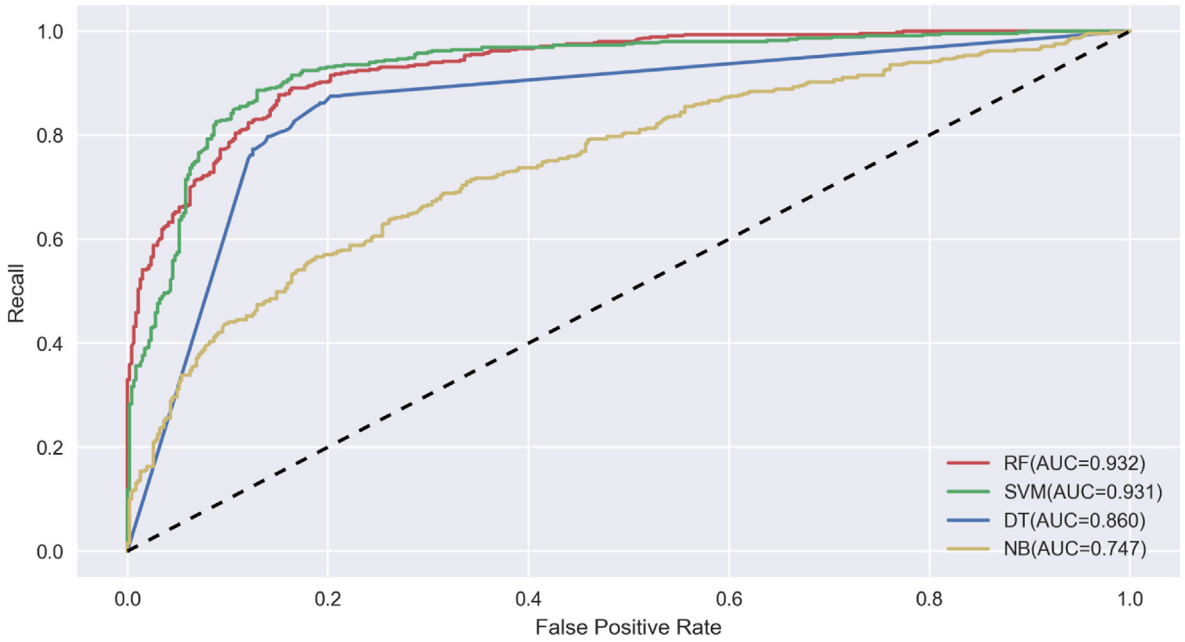


Figure 2. The receiver operator characteristic (ROC) curves for the machine learning models.

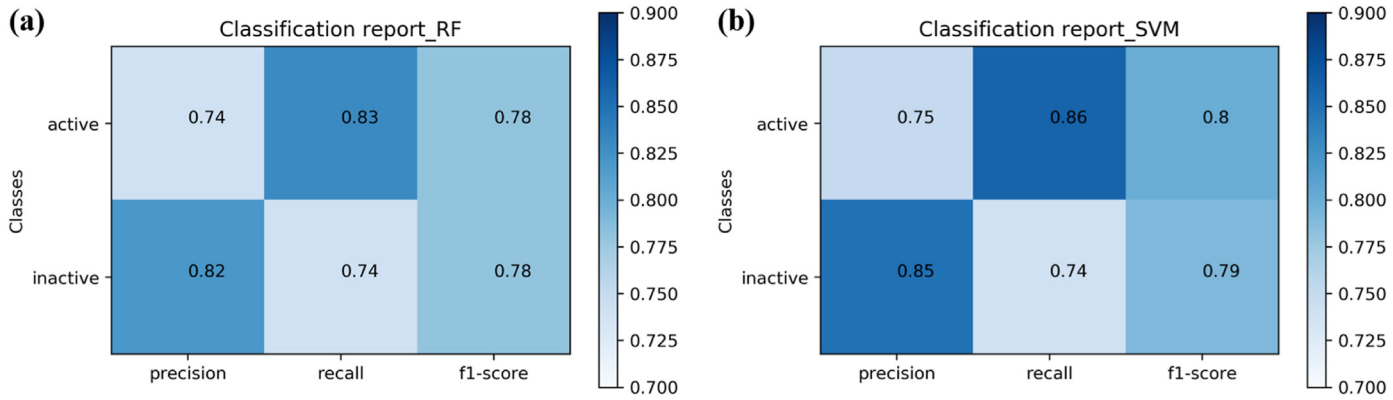


Figure 3. Performance evaluation of RF (a) and SVM (b) models using external independent validation.

comparing the performances between different models, which could be used for virtual screening of natural products in the chemical library.

3.3. Virtual screening of natural products in ZINC database

Since machine learning algorithms use fingerprint space to determine the influences of different chemical groups or molecular skeletons on the biological activity of compounds, the active molecules predicted by the model should have some structural similarities with the compounds in the training data. In this study, we used the more effective RF and SVM models to virtually screen 224,205 natural products in the ZINC database, resulting in 23 p38α MAPK candidate inhibitors with model scores both greater than 0.9, which were used for further molecular docking.

3.4. Confirming the reliability of the docking protocol

Molecular docking is able to describe the pose of the ligand at the binding site of the receptor proteins. To ensure the applicability of the docking process, we used AutoDock Vina to redock the co-crystallized compound and calculated the RMSD of the original ligand in the crystal structure (PDB ID: 1A9U) of the docked complex. As shown in Figure 4, the co-crystallized ligand (CID9543416) reproduced the original docking conformation with the RMSD of 0.47 Å as well as the binding free energy of −9.9 kcal/mol. Meanwhile, the favorable agreement between the two structures implies that our docking protocol is reliable and can be utilized to search the inhibitors of p38α MAPK.

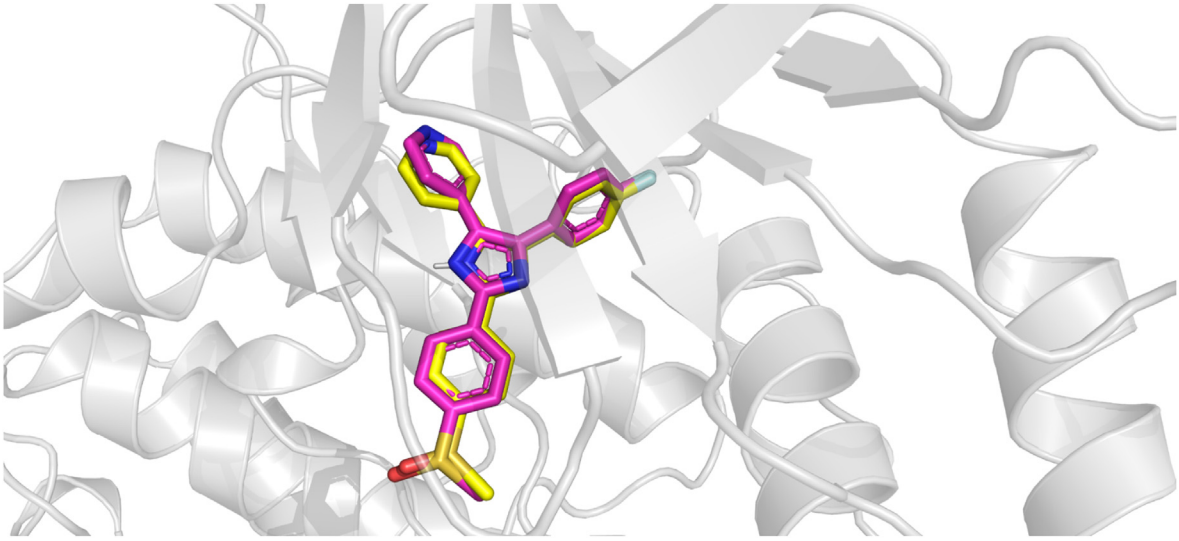
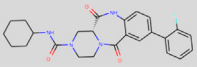
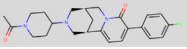
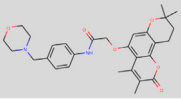
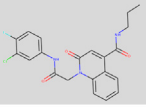
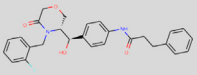
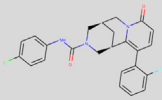
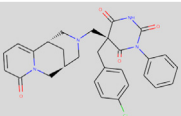
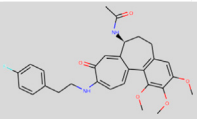


Figure 4. The superimposition of the docked co-crystallized ligand with its X-ray crystal structure (Yellow color indicates docked compound and pink color indicates experimental compound).

3.5. Docking study

We performed docking studies of the compounds screened by the machine learning models with the help of Autodock Vina and ranked them according to the predicted docking binding energy. The results were shown in Table 2, 8 out of 23 compounds have binding free energies ranging from −8.0 kcal/mol to −9.5 kcal/mol Figure 5 displayed the interacting residues of these compounds, specifically, ZINC4236005 had the lowest binding energy of −9.4 kcal/mol and interacted with MET-109 and LYS-53 amino acid residues, followed by ZINC4260400 and ZINC20760122, which showed the binding energy of −8.9 kcal/mol and −8.7 kcal/mol. Both of these compounds formed a hydrogen bond with TYR-35 amino acid residues. ZINC12411147 interacted with LYS-53, TYR-35 and ARG-173 amino acid residues through four hydrogen bonds of 2.3 Å, 3.3 Å, 2.8 Å and 2.8 Å distance. ZINC67896518, ZINC8300300 and ZINC104194468 interacted with LYS-53, ASP-168 and HIS-80 amino acid residues, respectively. ZINC8740044 showed the highest binding energy of −8.0 kcal/mol and formed four hydrogen

Table 2. Binding energy and functional residues of the eight selected compounds.

ZINC ID	Structure	Affinity (kcal/mol)	Interacting residues	Bond length(Å)
ZINC4236005		−9.4	MET-109 LYS-53	3.4 2.4
ZINC4260400		−8.9	TYR-35	2.8
ZINC20760122		−8.7	TYR-35	2.5
ZINC12411147		−8.5	LYS-53 TYR-35 ARG-173	2.3 3.3 2.8, 2.8
ZINC67896518		−8.4	LYS-53	2.5
ZINC8300300		−8.4	ASP-168	2.7
ZINC104194468		−8.3	HIS-80	3.1
ZINC8740044		−8.0	GLY-110 MET-109	2.6 1.9, 2.1, 3.5

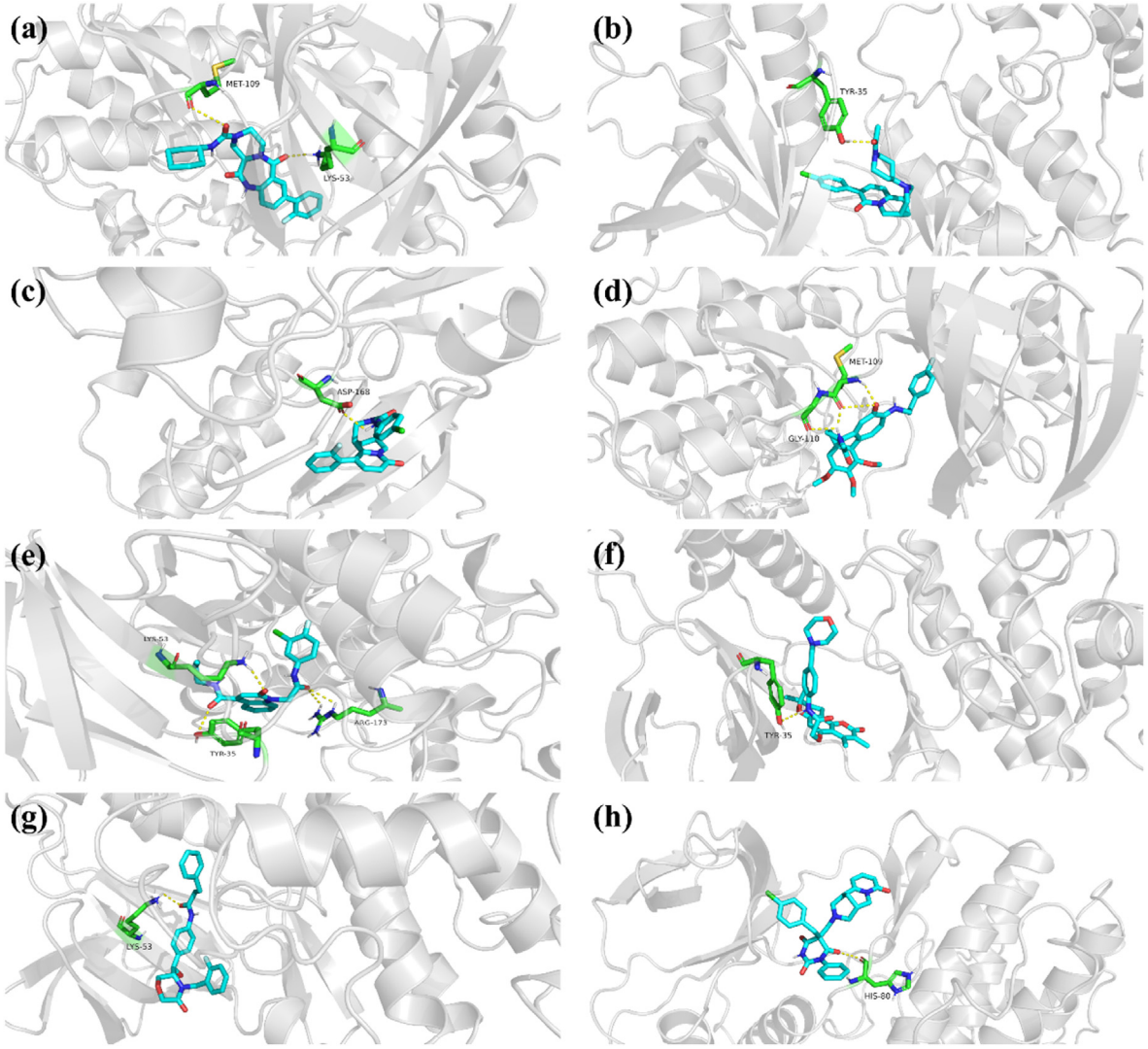


Figure 5. Binding mode of screened compounds (a) ZINC4236005, (b) ZINC4260400, (c) ZINC8300300, (d) ZINC8740044, (e) ZINC12411147, (f) ZINC20760122, (g) ZINC67896518, (h) ZINC104194468 in the active site of target protein (PBD ID: 1A9U), shown as gray ribbon. Active site residues are in green colored sticks. Hydrogen bonds that are formed in between protein and compound are shown as yellow dotted lines.

bonds of 2.6 Å, 1.9 Å, 2.1 Å and 3.5 Å distance with GLY-110 and MET-109 amino acid residues. The results of molecular docking indicated that all eight compounds were bound tightly to the target protein and interacted with the active site residues. Additionally, similar to ZINC4236005, the co-crystallized ligand formed two hydrogen bonds with MET-109 and LYS-53 amino acid residues (Supplementary Figure 1).

Molecular docking is one of the most commonly used virtual screening methods, and the process involves spatial matching and energy matching between ligand and receptor. The results of docking calculations allow predicting the binding mode and affinity between two molecules. In this study, a total of eight compounds displayed binding free energies less than −8.0 kcal/mol, but all of them possessed lower binding affinities to the target protein than the original co-crystallized ligand. According to the results of interaction visualization, we observed that the compound with the lowest binding free energy, ZINC4236005, had the most similar binding mode to the co-crystallized ligand. Both of them occupied hydrophobic pockets I and II, while interacting with the amino acid residues therein through hydrogen bonds. Compounds ZINC4260400, ZINC20760122, ZINC12411147 and ZINC67896518 only occupied hydrophobic pocket I, and no hydrogen bonding interaction was found with residue MET-109. We speculated this might be the primary reason why the binding affinities of these four compounds were lower than that of the co-crystallized ligand. It is worth noting that the compound ZINC12411147 formed two hydrogen bonds with the amino acid residue ARG-173, which was relatively uncommon among inhibitors of p38α MAPK. Compound ZINC8300300 formed one hydrogen bond with the amino acid residue ASP-168, which implied that this compound may be different from other candidate molecules in terms of inhibition pattern. Although compound ZINC8740044 formed multiple hydrogen bonds with the amino acid residue MET-109, we speculated that occupying only hydrophobic pocket II was probably one of the reasons for its lowest binding affinity. Moreover, compound ZINC104194468 was docked at a distinct position from the remaining compounds, hence its interacting amino acid residue HIS-80 has hardly been reported in the literature previously (Supplementary Figure 2).

The discovery of p38α MAPK inhibitors has been very popular for a long time. *Tariq et al.* synthesized a series of N-[3-(substituted-4H-1,2,4-triazol-4-yl)]-benzo[d]thiazol-2-amines and conducted the molecular docking study against p38α MAPK. Binding mode analysis revealed that molecules with lower binding free energy interacted with amino acid residues MET-109, LYS-53, GLY-110 and ALA-34 [37]. *Gangwal et al.* searched for potential p38α MAPK inhibitors from the database via a combined virtual screening approach. The authors found the amino acid residues MET-109, LYS-53, ASP-168, and GLU-71 to be critical for the

formation of the interaction [38]. Notably, the screened compounds in this study also had similar interaction profiles with the above amino acids (except ALA-34 and GLU-71), and these results indicated that our conclusion was in consistent with the findings from previous literature.

3.6. ADMET analysis

Compounds with favorable binding affinity to the target protein are used for further ADMET analysis, which are critical indicators of whether a compound can be considered as a drug. In this study, we evaluated eight compounds for vital physicochemical descriptors and associated pharmacokinetics profiles by SwissADME online server and OSIRIS Property Explorer software. The results showed that two compounds (ZINC4260400 and ZINC8300300) had acceptable ADMET properties (Supplementary Table 1).

Specifically, all 8 compounds followed the Lipinski rule, Veber rule, Egan rule and Muegge rule. They also had stable polarity and promising bioavailability scores. However, ZINC4236005, ZINC67896518, ZINC104194468 and ZINC8740044 compounds didn't follow the Ghose rule due to their molar refractivity being higher than 130 as well as the molecular weights of ZINC104194468 and ZINC8740044 being higher than 480. Many drug candidates have failed in clinical trials due to problems related to absorption. In the present study, all 8 compounds had high gastrointestinal absorption. The ability of a drug to penetrate the blood-brain barrier determines whether it can successfully reach the target tissue and exert the effects. We found that ZINC4260400, ZINC67896518 and ZINC8300300 can cross the blood-brain barrier. The outcome of the toxicity risk estimation displayed that ZINC4260400, ZINC12411147 and ZINC8300300 compound had no risk of mutagenicity, tumorigenicity, irritation and reproductive effects. In addition, we calculated the Drug-Score for all compounds to assess their pharmaceutical potential (Figure 6). The top three compounds were ZINC4260400 (0.75), ZINC12411147 (0.6) and ZINC8300300 (0.54). These results allowed us to search more reliably for natural products with p38α MAPK inhibitory activity.

3.7. Binding stability of the screened compounds to target protein

To investigate the stability of the docked complex, we performed the molecular dynamics simulation via the Gromacs program. Firstly, the changes in the backbone atoms of the receptor protein from the beginning conformation to the end location were analyzed by RMSD. In general, a larger value of RMSD represents a more dramatic alteration during the simulation. In the present study, we could observe that the protein backbones of all complexes maintained low RMSD values throughout the simulation (Figure 7a). It is noteworthy that the trajectory fluctuation of the protein backbone bound to compound ZINC8300300 was relatively prominent, which may be associated with the different inhibition mode of ZINC8300300. We also calculated the RMSD values for each ligand. The results showed that the ligands bound to the target protein all displayed stable RMSD trajectories (Figure 7b). Interestingly, although the RMSD values of compound ZINC8300300 were larger, the trajectory fluctuation was smaller than that of compound ZINC4260400. The effect of amino acid residues on the structural changes of the complexes was subsequently studied by RMSF, where the individual peaks denote the most fluctuating regions of the protein over the simulation process. The results were shown in Figure 7c, the target protein had relatively large variations in three regions, *i.e.,* the index of 30–35, 170–180 and 315–320, which can be considered as flexible. Next, the compactness of the protein structure was characterized using Rg. In Figure 7d, both complexes yielded stable Rg trajectories with average values of 2.23 nm (p38α-ZINC4260400), 2.28 nm (p38α-ZINC8300300) and 2.21 nm (p38α-co-crystallized ligand), respectively. Finally, the hydrophobicity of the protein was assessed through SASA. As it can be seen from Figure 7e, the SASA values of the receptor protein did not exhibit much fluctuation throughout the simulation and the average values for complexes p38α-

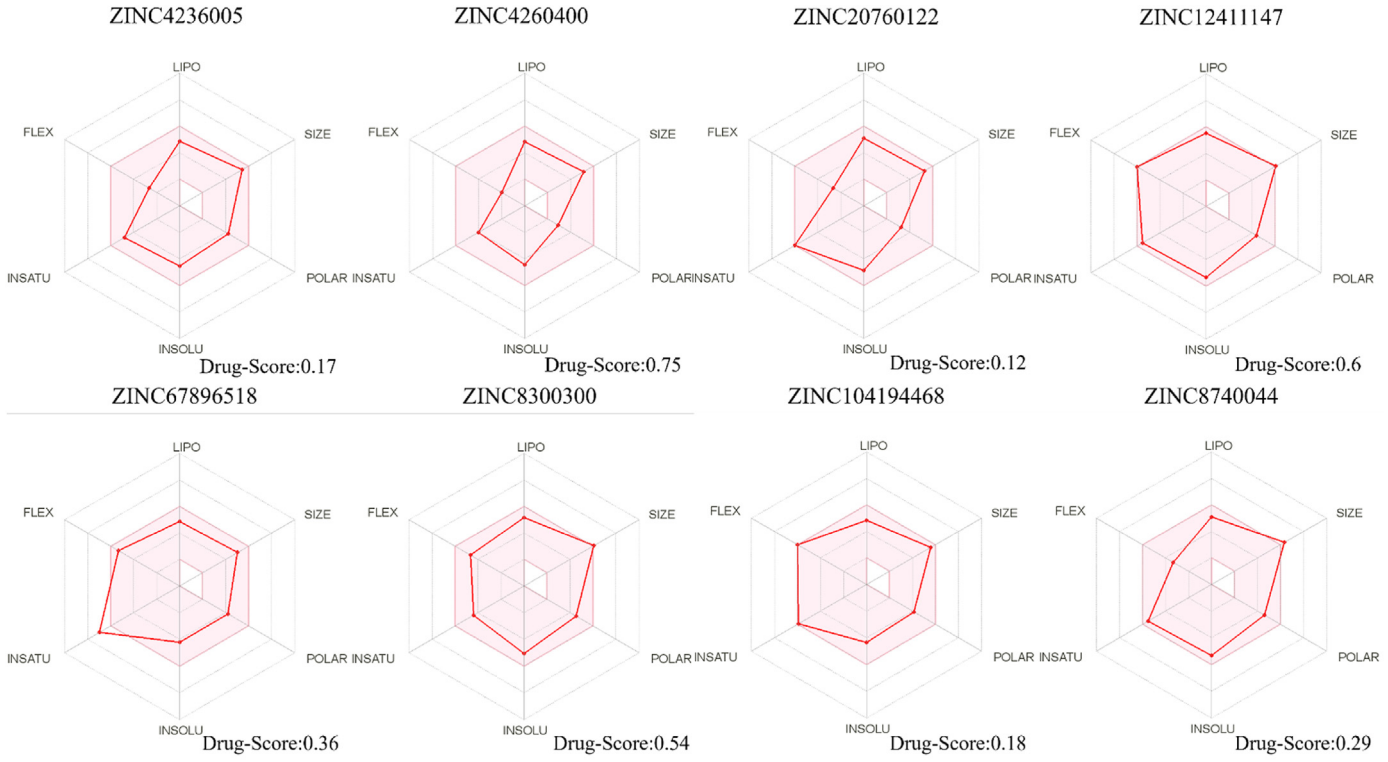


Figure 6. The pharmacokinetic and Drug-Score of screened compounds obtained from SwissADME and OSIRIS Property Explorer (The pink area represents the optimal range for each property).

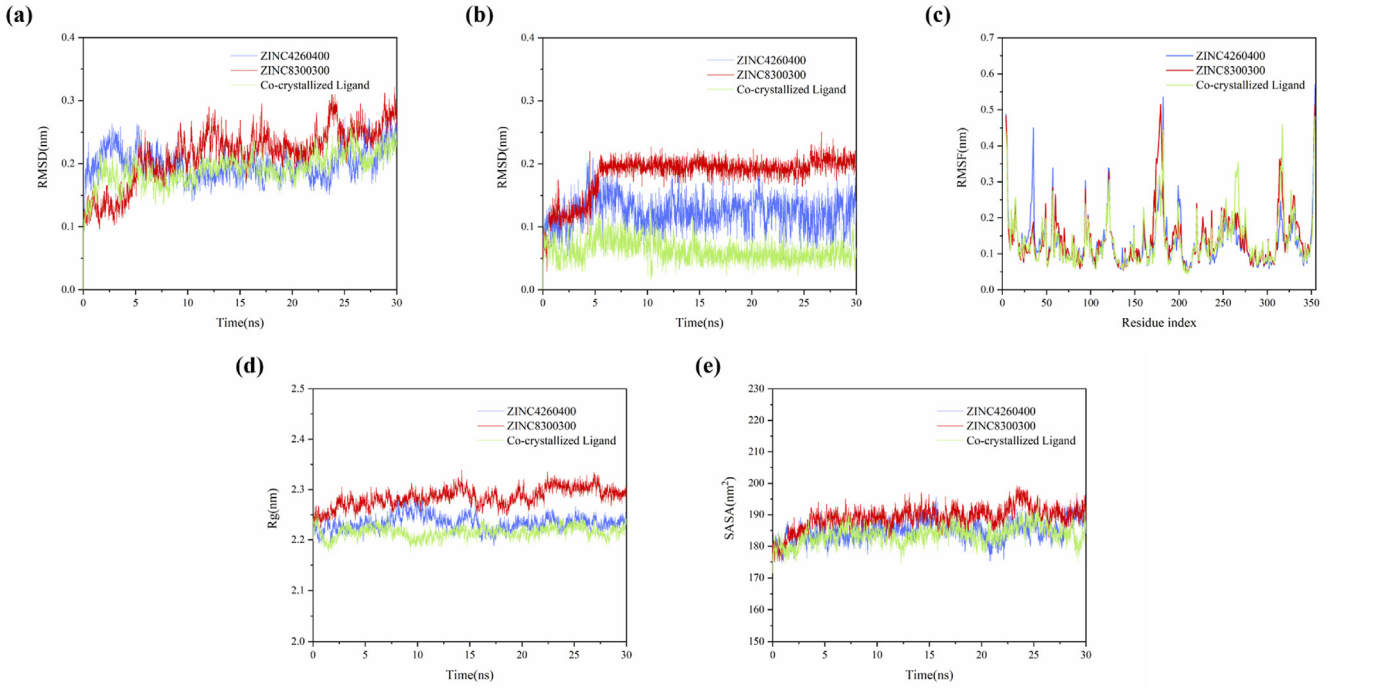


Figure 7. The changes of (a) Protein backbone RMSD, (b) Ligand RMSD, (c) RMSF, (d) Rg and (e) SASA vs. time of the simulation.

Table 3. MM/PBSA binding free energy between p38α MAPK and compounds ZINC4260400, ZINC8300300 and the co-crystallized ligand.

Compound	van der Waal energy (kJ/mol)	Electrostatic energy (kJ/mol)	Polar solvation energy (kJ/mol)	SASA energy (kJ/mol)	Total binding energy (kJ/mol)
ZINC4260400	−206.895 ± 12.21	−36.697 ± 24.26	108.419 ± 18.02	−17.295 ± 0.93	−152.434 ± 19.46
ZINC8300300	−204.483 ± 9.53	−19.452 ± 16.62	111.811 ± 14.66	−15.981 ± 0.69	−128.139 ± 23.34
Co-crystallized ligand	−182.357 ± 11.20	−155.006 ± 31.05	187.454 ± 20.00	−19.574 ± 0.78	−169.414 ± 17.72

ZINC4260400, p38α-ZINC8300300 and p38α-co-crystallized ligand were 184.8 nm^2, 189.1 nm^2 and 183.6 nm^2. Taken together, these results demonstrated that compounds ZINC4260400 and ZINC8300300 could bind steadily to the target protein p38α MAPK.

To further investigate the binding affinities between the ligand molecules and the receptor protein, we selected 20–30 ns trajectory files as the objects and calculated the total binding free energies using MM/PBSA. The results were shown in Table 3. For the co-crystallized ligand, both van der Waals interaction and electrostatic interaction played a key role with −182.357 ± 11.20 kJ/mol and −155.006 ± 31.05 kJ/mol, respectively. However, in compounds ZINC4260400 and ZINC8300300, only van der Waals interaction was dominant. This is the main reason why the binding free energies of the screened compounds were lower than that of the original ligand.

Molecular dynamics simulations mainly rely on Newtonian Mechanics to simulate the dynamic behavior of molecules. In this study, we performed molecular dynamics simulation for 30 ns to verify the binding stability of the screened hits to the target protein. The trajectories of RMSD revealed that both compounds reached the equilibrium state after experiencing different times. The analysis of the RMSF allowed us to observe the flexible and non-flexible regions of the receptor protein. Meanwhile, the results of Rg and SASA further confirmed that these compounds were able to stabilize in the target protein, which provided a powerful support for our screening. Furthermore, the results of MM/PBSA binding free energy were consistent with molecular docking, which also suggested us to increase the electrostatic interactions formed between the candidate compounds and the receptor protein as much as possible in the subsequent structural modification.

4. Conclusion

In recent years, natural products have been extensively applied as conventional medicines for the treatment of various diseases. Many studies have demonstrated that identifying drug candidates against specific targets from natural compounds is an efficient strategy. p38α, one of the important kinases in the MAPK signaling pathway, is usually targeted to inhibit the expression of pro-tumorigenic factors [39]. Despite the great potential of virtual screening techniques for the identification of active compounds, any of these methods alone cannot meet the practical demands of drug discovery. The multi-stage virtual screening can combine the advantages of various methods to improve the screening efficiency and reduce the false positive rate of the results [40].

The aim of this study was to find p38α MAPK inhibitors from natural products, and we established a "machine learning-molecular docking-molecular dynamics simulation" multi-level screening system. Firstly, we constructed multiple classification models based on machine learning algorithms with MACCS molecular fingerprints. Machine learning methods, which develop models by using experimentally verified datasets and make predictions for unknown datasets, are now widely adopted for the discovery of potential drugs and the prediction of biological characteristics [41, 42]. Subsequently, two better performing models (RF and SVM) were selected for virtual screening of natural products in the ZINC database, and the consensus compounds were further preceded by molecular docking. After this, the screened hits with lower binding energies (less than −8.0 kcal/mol) were taken for ADMET analysis. Finally, the stability of receptor-ligand complexes was analyzed through molecular dynamics simulation.

The results indicated that compounds ZINC4260400 and ZINC8300300 showed higher scores in the machine learning models and better binding affinities to the target protein, which could be served as potential p38α MAPK inhibitors. In the future, we will conduct in vitro and in vivo experiments on the two hits to investigate their mechanisms of action, which could provide a basis for the optimization and design of lead compounds.

Declarations

Author contribution statement

Bin Yan: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data.

Ruoqi Yang: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Xuan Zha: Performed the experiments; Analyzed and interpreted the data.

Xingyi Gao: Performed the experiments.

Kangmin Wang: Analyzed and interpreted the data.

Bin Cheng: Contributed reagents, materials, analysis tools or data.

Funding statement

Bin Yan was supported by Sub-project of the National Ministry of Health Major New Drug Creation Science and Technology Major Project [NO. 2014ZX09509001001].

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e10495.

References

[1] J.M. Kyriakis, J. Avruch, Protein Kinase cascades activated by stress and inflammatory cytokines, BioEssays 18 (1996), 567577.
[2] M. Raman, W. Chen, M.H. Cobb, Differential regulation and properties of MAPKs, Oncogene 26 (2007) 3100–3112.
[3] C. Dong, R.J. Davis, R.A. Flavell, MAP kinases in the immune response, Annu. Rev. Immunol. 20 (2002) 55–72.
[4] A. Cuadrado, A.R. Nebreda, Mechanisms and functions of p38 MAPK signalling, Biochem. J. 429 (2010) 403–417.
[5] J. Han, P. Sun, The pathways to tumor suppression via route p38, Trends Biochem. Sci.(Amsterdam. Regular ed.) 32 (2007) 364–371.
[6] T. Zarubin, J. Han, Activation and signaling of the p38 MAP kinase pathway, Cell Res. 15 (2005) 11–18.
[7] A. Cuenda, S. Rousseau, P38 MAP-kinases pathway regulation, function and role in human diseases, Biochim. Biophys. Acta 1773 (2007) 1358–1375.
[8] J. Han, J.D. Lee, L. Bibbs, R.J. Ulevitch, A MAP kinase targeted by endotoxin and hyperosmolarity in mammalian cells, Science 265 (1994) 808–811.
[9] Z. Wang, H.P.C. Harkins, U.R.J. Ulevitch, J. Han, M.H. Cobb, E.J. Goldsmith, The structure of mitogen-activated protein kinase p38 at 2.1-A resolution, P. Natl. Acad. Sci. U. S. A. 94 (1997) 2327–2332.
[10] E.L. Michelotti, K.K. Moffett, D. Nguyen, M.J. Kelly, R. Shetty, X. Chai, K. Northrop, V. Namboodiri, B. Campbell, G.A. Flynn, T. Fujimoto, F.P. Hollinger, M. Bukhtiyarova, E.B. Springman, M. Karpusas, Two classes of p38α map kinase inhibitors having a common diphenylether core but exhibiting divergent binding modes, Bioorg. Med. Chem. Lett. 15 (2005) 5274–5279.
[11] S. Kumar, J. Boehm, J.C. Lee, P38 MAP kinases: key signalling molecules as therapeutic targets for inflammatory diseases, Nat. Rev. Drug Discov. 2 (2003) 717–726.
[12] J. Gupta, A.R. Nebreda, Roles of p38α mitogen-activated protein kinase in mouse models of inflammatory diseases and cancer, FEBS J. 282 (2015) 1841–1857.
[13] A. Igea, A.R. Nebreda, The stress kinase p38α as a target for cancer therapy, Cancer Res. 75 (2015) 3997–4002.
[14] S. Paillas, F. Boissière, F. Bibeau, A. Denouel, C. Mollevi, A. Causse, V. Denis, N. Vezzio-Vié, L. Marzi, C. Cortijo, I. Ait-Arsa, N. Askari, P. Pourquier, P. Martineau, M. Del Rio, C. Gongora, Targeting the p38 mapk pathway inhibits irinotecan resistance in colon adenocarcinoma, Cancer Res. 71 (2011) 1041–1049.
[15] M.M. Madkour, H.S. Anbar, M.I. El-Gamal, Current status and future prospects of p38α/mapk14 kinase and its inhibitors, Eur. J. Med. Chem. 213 (2021), 113216.
[16] S.A. Laufer, D.R.J. Hauser, D.M. Domeyer, K. Kinkel, A.J. Liedtke, Design, synthesis, and biological evaluation of novel tri- and tetrasubstituted imidazoles as highly potent and specific atp-mimetic inhibitors of p38 map kinase: focus on optimized interactions with the enzyme's surface-exposed front region, J. Med. Chem. 51 (2008) 4122–4149.
[17] L. Lin, W. Hsu, C. Lin, Antiviral natural products and herbal medicines, J. Trad. Compl. Med. 4 (2014) 24–35.
[18] T. Hou, X. Xu, Recent development and application of virtual screening in drug discovery: an overview, Curr. Pharmaceut. Des. 10 (2004) 1011–1033.
[19] M. Thafar, A.B. Raies, S. Albaradei, M. Essack, V.B. Bajic, Comparison study of computational prediction tools for Drug-Target binding affinities, Front. Chem. 7 (2019) 782.
[20] A. Mollica, G. Zengin, S. Durdagi, S.R. Ekhteiari, G. Macedonio, A. Stefanucci, M.P. Dimmito, E. Novellino, Combinatorial peptide library screening for discovery of diverse alpha-glucosidase inhibitors using molecular dynamics simulations and binary qsar models, J. Biomol. Struct. Dyn. 37 (2019) 726–740.
[21] G. Poli, M.P. Dimmito, A. Mollica, G. Zengin, S. Benyhe, F. Zador, A. Stefanucci, Discovery of novel μ-opioid receptor inverse agonist from a combinatorial library of tetrapeptides through structure-based virtual screening, Molecules 24 (2019) 3872.
[22] G. Narendra, B. Raju, H. Verma, B. Sapra, O. Silakari, Multiple machine learning models combined with virtual screening and molecular docking to identify selective human ALDH1A1 inhibitors, J. Mol. Graph. Model. 107 (2021), 107950.
[23] M.M. Mysinger, M. Carchia, J.J. Irwin, B.K. Shoichet, Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, J. Med. Chem. 55 (2012) 6582–6594.
[24] F.G. Eli, G.C. R, M. Karina, M.J. Medina-Franco, Database fingerprint (DFP): an approach to represent molecular databases, J. Cheminf. 9 (2017) 9.
[25] D.J. Polton, Installation and operational experiences with MACCS (molecular access system), Online Rev. 6 (1982) 235–242.
[26] B. Gholami, I. Norton, A.R. Tannenbaum, N.Y.R. Agar, Recursive Feature Elimination for Brain Tumor Classification Using Desorption Electrospray Ionization Mass Spectrometry Imaging, United States, 2012, IEEE, United States, 2012, pp. 5258–5261.
[27] L. Rokach, O. Maimon, Top-Down induction of decision trees classifiers—a survey, IEEE Tran. Syst., Man Cybern., Part C (Appl. Rev.) 35 (2005) 476–487.
[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[30] P. Watson, Naïve bayes classification using 2D pharmacophore feature triplet vectors, J. Chem. Inf. Model. 48 (2008) 166–178.
[31] O. Trott, A.J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2009) 455–461.
[32] S. Yuan, H.C.S. Chan, Z. Hu, Using PyMOL as a platform for computational drug design, Wiley Interdiscip. Rev. Comput. Mol. Sci. 7 (2017).
[33] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, Bioinformatics 29 (2013) 845–854.
[34] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A.D. Mackerell Jr., CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields, J. Comput. Chem. 31 (2010) 671–690.
[35] J. Kongsted, U. Ryde, An improved method to predict the entropy term with the mm/pbsa approach, J. Comput. Aided Mol. Des. 23 (2009) 63–71.
[36] M. Maddah, R. Bahramsoltani, N.H. Yekta, R. Rahimi, R. Aliabadi, M. Pourfath, Proposing high-affinity inhibitors from Glycyrrhiza glabra L. Against SARS-CoV-2 infection: virtual screening and computational analysis, New J. Chem. 45 (2021) 15977–15995.
[37] S. Tariq, O. Alam, M. Amir, Synthesis, p38α map kinase inhibition, anti-inflammatory activity, and molecular docking studies of 1,2,4-triazole-based benzothiazole-2-amines, Arch. Pharm. 351 (2018), 1700304.
[38] R.P. Gangwal, N.R. Das, K. Thanki, M.V. Damre, G.V. Dhoke, S.S. Sharma, S. Jain, A.T. Sangamwar, Identification of p38α map kinase inhibitors by pharmacophore based virtual screening, J. Mol. Graph. Model. 49 (2014) 18–24.
[39] J.C. Lee, J.T. Laydon, P.C. Mcdonnell, T.F. Gallagher, S. Kumar, D. Green, D. Mcnulty, M.J. Blumenthal, J.R. Keys, S.W.L. Vatte, J.E. Strickler, M.M. Mclaughlin, I.R. Siemens, S.M. Fisher, G.P. Livi, J.R. White, J.L. Adams, P.R. Young, A protein kinase involved in the regulation of inflammatory cytokine biosynthesis, Nature (1994) 372–739.
[40] H. Kang, Z. Sheng, R. Zhu, Q. Huang, Q. Liu, Z. Cao, Virtual drug screen schema based on multiview similarity integration and ranking aggregation, J. Chem. Inf. Model. 52 (2012) 834–843.
[41] M. Serafim, T. Kronenberger, P.R. Oliveira, A. Poso, K.M. Honorio, B. Mota, V.G. Maltarollo, The application of machine learning techniques to innovative antibacterial discovery and development, Expet Opin. Drug Discov. 15 (2020) 1165–1180.
[42] S. Petra, W.W. Patrick, A.T. Plowright, S. Norman, L. Jennifer, R.A. Goodnow, F. Jasmin, J.M. Jansen, J.S. Duca, T.S. Rush, Z. Matthias, H.J. Edward, K. Elizabeth, K. Matthias, B. Jeff, F. Kimito, L. Chris, S. Gisbert, Rethinking drug design in the artificial intelligence era, Nat. Rev. Drug Discov. 19 (2020) 353–364.