MEDICINAL CHEMISTRY RESEARCH

ORIGINAL RESEARCH ARTICLE

Identification of potential natural product derivatives as CK2 inhibitors based on GA-MLR QSAR modeling, synthesis and biological evaluation

Yanan Xuan[1] · Yue Zhou[2] · Yue Yue[1] · Na Zhang[1] · Guohui Sun[1] · Tengjiao Fan[1,3] · Lijiao Zhao[1] · Rugang Zhong[1]

Abstract

Protein kinase CK2 is a validated target for cancer therapy. Many natural products have shown inhibitory activity against CK2 as potential anti-cancer drug candidates. A compatible quantitative structure-activity relationship (QSAR) model of natural products is necessary to identify the structural determinants related to their biological activities and provides valuable clues for the discovery of natural leads as anticancer drugs. In this study, genetic algorithm (GA) and multiple linear regression (MLR) methods, combined with preferred molecular descriptors, were employed to build QSAR models of CK2 natural product inhibitors. The best model, composed of eight molecular descriptors, yielded $Q^2_{Loo} = 0.7914$ and $R^2 = 0.8220$ for the training set and $Q^2_{ext} = 0.7921$ and $R^2_{ext} = 0.7998$ for the test set, indicating the model's robust reliability and high predictability. As a proof of concept, a true external test set, distinct from the training and test sets, was synthesized and tested in vitro to verify the predictive ability of this model. The predicted pIC_{50} values of 13 compounds showed less than 30% relative error (including 10 compounds with relative errors less than 20%), further validating the predictive performance of this model. And compound M18, M24, and M26 were identified as potential CK2 inhibitors with the predicted pIC_{50} values of 11.29, 8.79, and 12.03 respectively. Furthermore, the underlying structural mechanisms through which key molecular descriptors influenced their inhibitory activities against CK2 were elucidated. All these results provide valuable information for the discovery of CK2 inhibitors.

Graphical Abstract
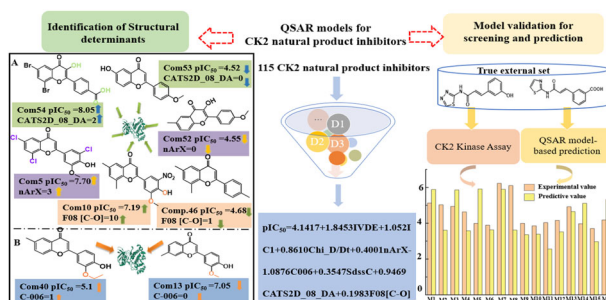
These authors contributed equally: Yanan Xuan, Yue Zhou

✉ Na Zhang
  nanatonglei@bjut.edu.cn

[1] Key Laboratory of Environmental and Viral Oncology, College of Chemistry and Life Science, Beijing University of Technology, Beijing 100124, China

[2] Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100050, China

[3] Department of Medical Technology, Beijing Pharmaceutical University of Staff and Workers (CPC Party School of Beijing Tong Ren Tang (Group) co., Ltd.), Beijing 100079, China

🖄 Springer

Introduction

Natural products refer to endogenous chemical constituents or their metabolites in animals, plants, and microorganisms with structural diversity and broad-spectrum pharmaceutical activities. Historically, natural products have been regarded as valuable sources of drug leads and therapeutic agents [1]. They usually contain aromatic rings and a large number of hydrogen-bond acceptors and donors, which are considered as privileged structures to interact with drug targets [2]. For instance, protein kinase CK2 is involved in the development of cancer as a key regulator of cellular pathways, and many natural compounds have been identified as CK2 inhibitors, therefore representing potential anti-cancer drug candidates [3].

To date, considerable efforts have been devoted to discovering natural products as CK2 inhibitors. A number of polyphenol analogs, such as coumarins, flavones, anthraquinones, and ellagic acid, as well as linear scaffold 2-propenone derivatives, have shown inhibitory activity against CK2. A small library of 7-hydroxycoumarin and trifluoromethyl derivatives of coumarin was synthesized and tested in vitro, demonstrating moderate inhibitory activity (IC_{50}) against CK2 ranging from 0.28 to 39 μM [4, 5]. Meanwhile, attempts have been made to investigate the possibility of flavone derivatives as CK2 inhibitors, including natural flavones [6], synthesized 3-hydroxy-4'-carboxyflavones, and 4'-hydroxyflavone analogs [7, 8]. Among them, FLC26 was the most potent, with an IC_{50} value of 9 nM. Emodin, a representative of anthraquinones, displayed a moderate inhibitory effect on CK2 activity ($IC_{50} = 2$ μM) [9]. Virtual screening was employed to identify ellagic acid [10] as a potent inhibitor of CK2 ($IC_{50} = 20$ nM). Besides the mentioned compounds with polycyclic scaffolds as CK2 inhibitors, linear scaffold 2-propenone derivatives were also described as CK2 hits. Cozza et al. identified curcumin and its degradation product ferulic acid as the most potent CK2 inhibitors ($IC_{50} = 0.84$ μM) [11]. Our group designed and synthesized a series of 2-propenone derivatives using fragment-based drug design, and compounds 8a and 8b exhibited inhibitory activity against CK2 with IC_{50} values of 0.9 μM and 0.6 μM, respectively [12]. Despite the different chemical scaffolds they possess, a general binding mode was found in the CK2α-natural products complexes. Specifically, a hydrophobic deep cleft between the N- and C-terminal lobes of CK2 accommodates the aromatic rings of inhibitors. Polar interactions, such as hydrogen bonds, halogen bonds, or electrostatic interactions formed by amino, hydroxyl, and nitrogen heterocycles with the hinge region (Glu114 and/or Val116), as well as hydroxyl and carboxylate groups with positive areas (Lys68), are responsible for anchoring inhibitor orientation.

Identification of the structural determinants related to natural products' inhibition of CK2 provides valuable clues for the optimization and screening of novel CK2 inhibitors. Computational chemistry and chemical informatics methods are powerful tools for identifying structural features and properties of inhibitors that are strictly related to their biological activities [13]. For instance, based on low-energy and docking conformations, 3D-QSAR models of coumarin derivatives were developed to elucidate the structural mechanisms through which substitutions influence the inhibitory potency of compounds [14]. A reliable QSAR model of indeno[1,2-b]indole derivatives was built to reveal key molecular descriptors responsible for their biological activities and to predict possible hits of virtual screening as active inhibitors [15]. Compound 00082235 was validated by in vitro CK2 activity testing with an IC_{50} value of 2.33 μM.

As a matter of fact, the mentioned QSAR studies were based on a dataset with a certain scaffold, and thus exerted predictive power on the specific scaffold. That is to say, the predictive ability of a QSAR model is largely limited by the training set because only query chemicals similar to the training set can be reliably predicted. In order to enlarge the applicability domain (AD) of QSAR models, it will be significant to pursue a compatible QSAR model with general prediction power on query chemicals without chemical scaffold limitations. The more diverse the training chemicals, the larger the AD [16]. For instance, based on a diverse set of PI3Kγ inhibitors, a robust and highly predictive QSAR model was developed and successfully assessed using another set of compounds outside the training and test sets with various structures [17]. Qualitative classification models of CK2 natural product inhibitors were built to identify privileged substructures related to CK2 inhibition using machine learning algorithms [18].

In this study, a robust and predictive QSAR model of CK2 natural product inhibitors was developed for the discovery of novel CK2 inhibitors. Eight molecular descriptors were identified as key factors related to CK2 inhibition. Furthermore, the prediction ability was also validated by in vitro CK2 activity assay of a true external set that was not involved in modeling. It is expected that this study will convey valuable information to researchers in the design of CK2 inhibitors as anti-cancer drugs.

Table 1 Correlation matrix analysis of eight descriptors of the optimal model

Descriptors	IVDE	IC1	Chi_D/Dt	nArX	C-006	SdssC	CATS2D_08_DA	F08[C–O]
IVDE	1							
IC1	−0.002	1						
Chi_D/Dt	−0.238	0.010	1					
nArX	0.177	0.387	−0.010	1				
C-006	0.117	−0.004	−0.094	−0.106	1			
SdssC	−0.224	0.392	0.091	−0.111	0.090	1		
CATS2D_08_DA	0.208	−0.315	−0.270	−0.003	−0.010	−0.649	1	
F08[C–O]	0.214	−0.431	−0.341	−0.129	0.086	−0.398	0.619	1

Results and analysis

QSAR modeling results

Based on the screening of 613 molecular descriptors, 100 models were generated using a combination of GA and MLR methods. Subsequently, 32 independent models were obtained by filtering these inter-correlated models using QUIK (Q Under Influence of K) rule. Considering the scoring and ranking of these models evaluated by Multi-Criteria Decision Making (MCDM) method, model NO. 1446, involving eight descriptors, was selected as the optimal one. Obviously, the approximate ratio of training set (95) to descriptors (8) is reasonable, meeting the rule of thumb with a ratio exceeding 5. The correlation matrix analysis shown in Table 1 indicated no overfitting correlation between any two descriptors. Furthermore, there was no collinearity among the descriptors in the model, as proven by $Kxx = 0.2751$ (multivariate correlation index) and $\Delta K = 0.0860$ (global correlation among descriptors).

$$pIC_{50}(-LogIC_{50}) = 4.1417 + 1.8453IVDE + 1.052IC1$$
$$+ 0.8610Chi\_D/Dt + 0.4001nArX - 1.0876C - 006$$
$$+ 0.3547SdssC + 0.9469CATS2D\_08\_DA + 0.1983F08[C-O]$$
$$(1)$$

To obtain a robust and predictive model, rigorous validation of the QSAR model was performed using both internal (LOO cross-validation method) and external validation metrics (Y-randomization test). Considering $Q^2_{Loo} > 0.5$ and $R^2 > 0.6$ as indicators of a robust and predictive QSAR model, the statistical parameters $Q^2_{Loo} = 0.7914$ and $R^2 = 0.8220$ for the training set ($Q^2_{ext} = 0.7921$ and $R^2_{ext} = 0.7998$ for the test set) shown in Table 2, as well as the correlation line of the experimental and predicted inhibitory activity values (Fig. 1A), demonstrated the high performance of the optimal model. Meanwhile, the lower values of R^2_{Yscr} and Q^2_{Yscr} of the Y-scrambling test also indicated the reliability of this model, which was not generated by chance. Furthermore, this model was considered

Table 2 Statistical parameters for internal and external validation of the optimal model

Internal validation (training set)		External validation (test set)	
$N_{Training}$	95	N_{Test}	20
Q^2_{LOO}	0.7914	Q^2_{F1}	0.7921
R^2	0.8220	Q^2_{F2}	0.7874
R^2_{adj}	0.8055	Q^2_{F3}	0.7186
$RMSE_{tr}$	0.4228	R^2_{ext}	0.7998
CCC_{tr}	0.9023	$RMSE_{ext}$	0.5315
$RMSE_{cv}$	0.4577	CCC_{ext}	0.8749
CCC_{cv}	0.8857	r^2_m	0.6401
Q^2_{LMO}	0.7649	$\Delta r^2 m$	0.1978
R^2_{Yscr}	0.0847	k	0.9851
Q^2_{Yscr}	−0.1304	$(R^2_{ext}-R^2_0)/R^2_{ext}$	0.0600

acceptable as the statistical parameters $R^2_{ext} = 0.7998$, $k = 0.9851$, and $(R^2_{ext}-R_0^2)/R^2_{ext} = 0.06$ fulfill the requirements proposed by Roy et al. [19].

AD analysis

The applicability domain of QSAR models should be clarified since the developed models can generate trustworthy predictions only if the new compound lies within the applicability domain of the model. The Leverage and standardized approach were employed to define the AD of the optimal model. As shown in Fig. 1B, the inhibitory activity of the whole dataset was predicted well enough that all values fell within the range of less than 3 standard residuals. Five compounds in the training set (39, 40, 97, 108, and 109) were identified as structural outliers with h values higher than the leverage threshold ($h^* = 0.284$) due to the unique chemical groups presented in these compounds, such as the ethyl group in compounds 39 and 40, the sulfonyl group in compound 97, as well as the four rings in compounds 108 and 109. However, small residues (−0.200, 0.253, −0.053, −0.278, −0.109, respectively)
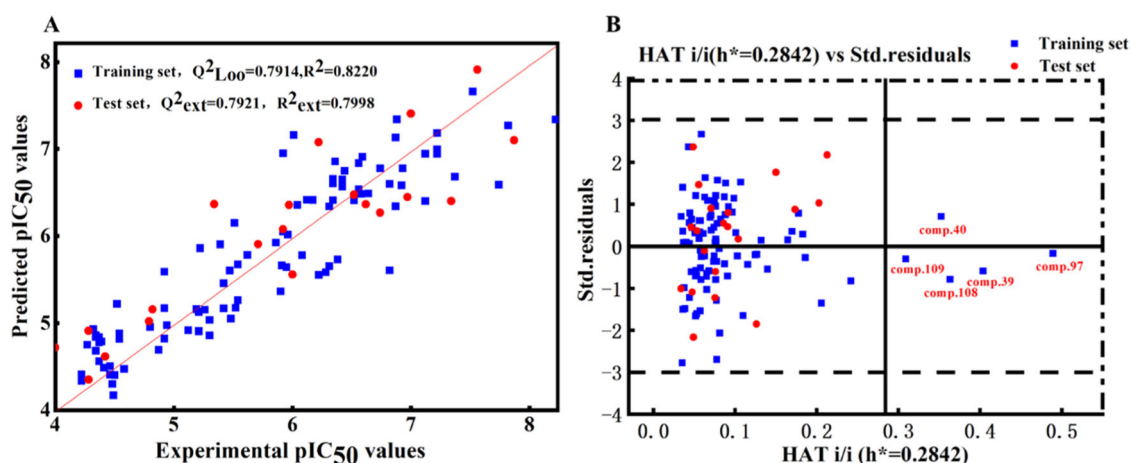
Fig. 1 A Linear fitting graph of experimental versus predicted pIC$_{50}$ values for the optimal QSAR model (B) Williams plot of the optimal model for CK2 natural product inhibitors, where horizontal dotted line means the ±3 standardized residuals for response values, and vertical solid line means the warning leverage value h*
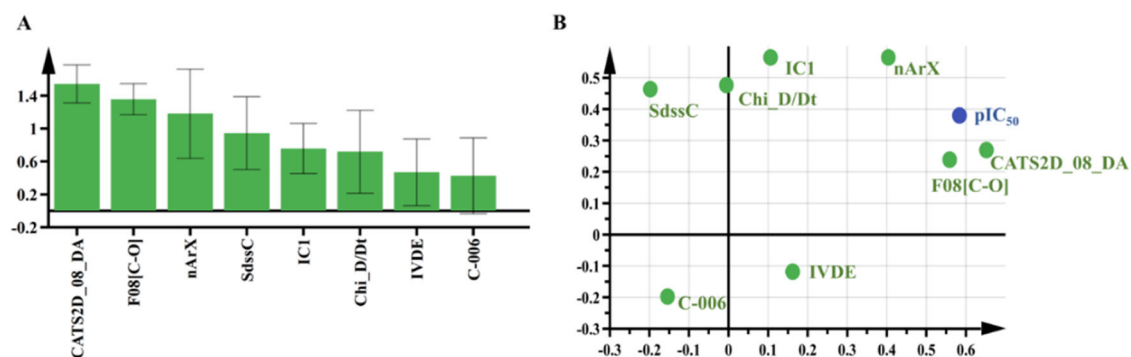

Fig. 2 A Variable importance plot (VIP) and (B) factor loading scatterplot for the eight descriptor variables in the best QSAR model

between their experimental values and predicted ones indicated the high predictive ability of the developed model.

Variable Importance Plot (VIP) and loading plot analysis

Eight molecular descriptors, namely IVDE, IC1, Chi_D/Dt, nArX, C-006, SdssC, CATS2D_08_DA, and F08[C–O], were regarded as determinants of inhibitory activity against CK2. VIP was analyzed to rank the relative importance of the variables (Fig. 2A). Among them, three molecular descriptors—CATS2D_08_DA, F08[C–O], and nArX—scoring higher than 1 (VIP scores of 1.54307, 1.3576, and 1.18082, respectively) were regarded as the most determinant factors of CK2 inhibition, whereas the other descriptors with VIP scores lower than 1 were considered less important. This conclusion could also be made from the loading plot indicated in Fig. 2B, which elucidated the contribution of descriptors to the response pIC$_{50}$. The closer the descriptors to the effect values, the greater the contributions to their inhibitory activity. Obviously,

CATS2D_08_DA, F08[C–O], and nArX may play dominant roles responsible for CK2 inhibition due to the smallest distance from the response pIC$_{50}$, while the other ones far away from the response values or its projection point but near from the origin could be considered as less important descriptors.

Mechanism interpretation of the optimal model

As indicated by Eq. 1, except for C-006, which is negatively correlated with the inhibitory activities of natural products, the other descriptors with positive coefficients would be beneficial for CK2 inhibition. The top three most significant descriptors, CATS2D_08_DA, F08[C–O], and nArX were focused on and discussed in detail (Fig. 3).

CATS2D_08_DA is regarded as the priority favorable factor responsible for CK2 inhibition. This 2D structure-based atom pair descriptor describes the number of potential hydrogen bond donor/acceptor atoms (e.g., oxygen, nitrogen, etc.) at a topological distance of 8 (eight bond distances). For instance, compounds 54, 55, and 56 showed
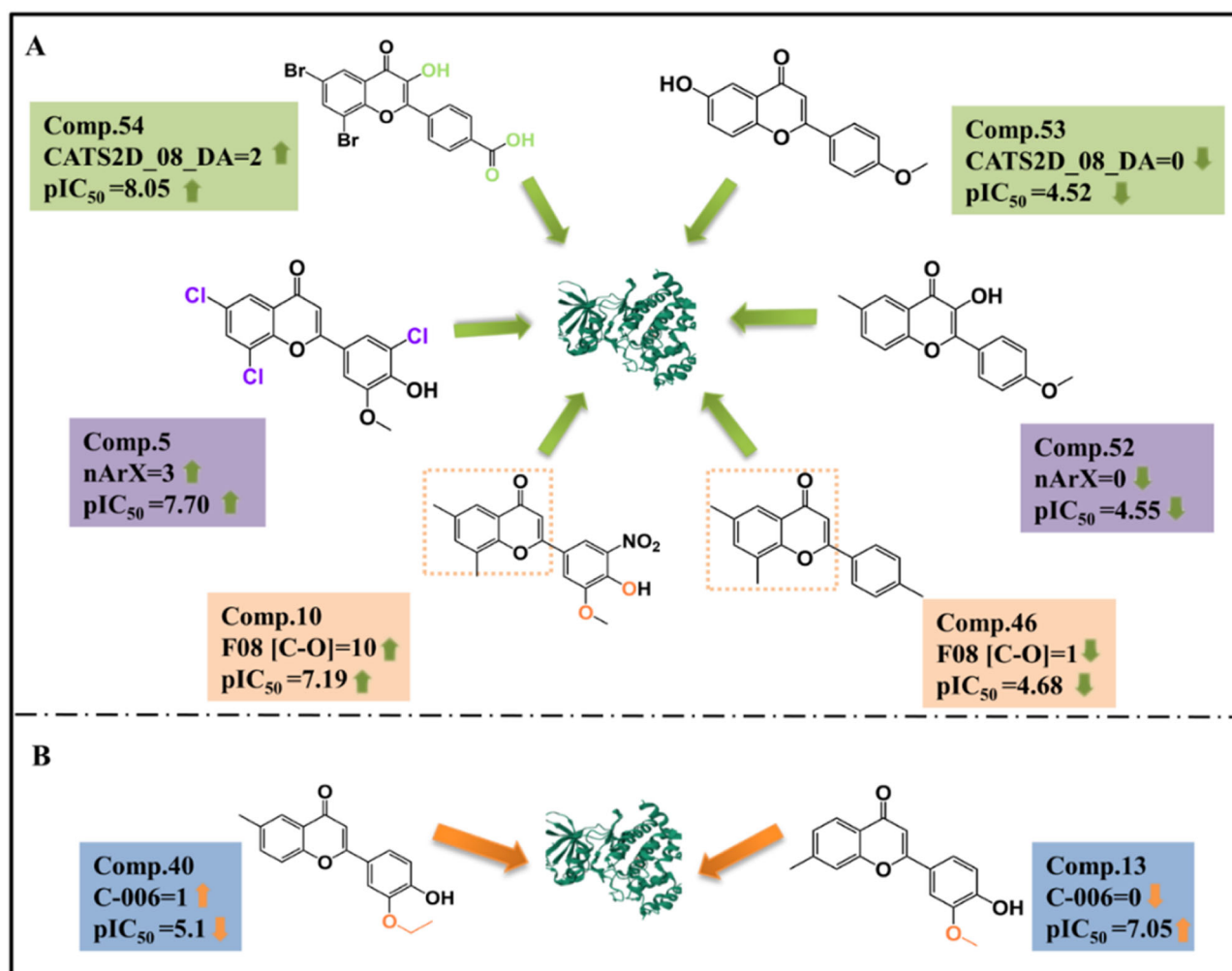
Fig. 3 Mechanistic correlations of (A) positive and (B) negative descriptor variables responsible for the inhibition of CK2

high inhibitory activity against CK2 with CATS2D_08_DA values of 2. This could be explained by the polar interactions involving H-bond donor/acceptor atoms of inhibitors and Glu114 and/or Val116 and/or Lys68 of CK2. Specifically, the hydroxyl at R5 and carboxylate group of compound 54 established H-bond and electrostatic interactions with Lys68. In contrast, the R5 hydrogen atom-substituted compound 53 ($pIC_{50} = 4.52$) generated a CATS2D_08_DA descriptor value of zero, resulting in a significant 10,000-fold reduction in inhibitory activities relative to compound 54 ($pIC_{50} = 8.05$). Interestingly, this descriptor could be used to elucidate the reason for the 5000-fold less potency of R2 non-ionizable substituents tricyclic quinoline (R2 = CONH2, $IC_{50} = 30 \, nM$) and compound 10 (R2 = CN, $IC_{50} > 2500 \, nM$) compared to CX-4945 with (R2 = COOH, $IC_{50} = 0.1$ nM) in vitro [20].

nArX is the number of X atoms (halogen atoms) substituted on the aromatic ring of molecules. This descriptor is positively correlated with inhibitory activity, indicating that

the more halogen atoms on the aromatic ring, the more potent inhibitory activity a compound exhibit. Accelerating studies have confirmed that halogen bonds occurring between halogen atoms and an oxygen atom are the novel driving force for the binding of inhibitors with protein kinase. This was also confirmed by a previous study that the halogen atoms of the benzimidazole analog TBB formed interactions with the hinge region residues Val116 and Glu114 [21], respectively. Additionally, halogen atoms were identified as privileged substructures based on the classification model of CK2 natural product inhibitors [18]. By comparing the structures and activities of compounds 5 ($pIC_{50} = 7.70$) and 52 ($pIC_{50} = 4.55$), three halogen atoms substituted on the flavonoid scaffold generated a 1400-fold increase in inhibitory activity compared to three hydrogen atoms at the corresponding positions.

F08 [C–O] is a two-dimensional frequency fingerprint indicating the frequency of the C–O bond at the corresponding topological distance. A higher frequency of C–O

features located at a topological distance of 8 is beneficial to CK2 inhibition. Since coumarins possess a bicyclic scaffold with an F08 [C–O] value lower than 8, it is reasonable to understand the fact that coumarins exhibited lower inhibitory activity relative to flavonoids with higher F08 [C–O] values. The nitro and hydroxyl groups of compound 10 created an additional center for oxygen and thus generated the F08 [C–O] value of 10, which provided reliable evidence for the lower potency of compound 46 ($pIC_{50} = 4.68$) with an F08 [C–O] value of 1.

The negatively correlated descriptor, C-006, denotes a CH2RX-type fragment in which R indicates an O or S atom connected through C and X. A negative correlation with the descriptor indicates that lower C-006 values may lead to higher biological activity. For example, compounds 97, 39, and 40 with C-006 values of 1 showed moderate inhibitory activities against CK2 with pIC_{50} values of 4.64, 5.3, and 5.1, respectively. Whereas compound 13, without a CH2-group, has a smaller C-006 value but exhibits higher inhibitory activity ($pIC_{50} = 7.05$) than compound 40 ($pIC_{50} = 5.1$).

Other descriptors, including SdssC, IC1, Chi_D/Dt, and IVDE, had less impact on the inhibitory activity of compounds. Regardless of the type of inhibitor (active or inactive), there was no remarkable discrepancy in the values of these descriptors. As indicated in Table S1 (supporting information), the four descriptors of potent compounds 1 and 4 are similar to those of compounds 53, 44, and 46 with lower inhibitory activity, respectively.

Bioassay validation of a true external test set

One important objective of QSAR modeling is to predict the inhibitory activity of new chemical entities outside the training and test sets against CK2. In this study, 26 natural product derivatives (M1-M16 with and M17-M26 without experimental IC_{50} values) were selected as a true external test set to evaluate the predictive ability of the developed model (Table 3 and Fig.S1). Among them, a series of 2-propenone derivatives containing an amine-substituted five-membered heterocycle and a benzoic acid (or phenol) were investigated by combining fragment-hybrid computational design with in vitro assay [12]. Another 10 molecules were identified as theoretical hits based on a hybrid virtual screening of CK2 natural product inhibitors (work in progress).

Enzymatic CK2 inhibition assay indicated that most 2-propenone derivatives exhibited moderate inhibitory effects against CK2α, with the pIC_{50} values ranging from 3.70 to 6.22, which were used to evaluate the predictive ability of the developed model. As shown in Fig. 4, except for compound M5, M8 and M11, the predicted pIC_{50} values of 13 compounds showed less than 30% relative error

(including 10 compounds with relative errors less than 20%), further validating the predictive performance of this model. The less than 10-fold discrepancy between predicted and experimental pIC_{50} values meant that this model was applicable for CK2 inhibitors with linear scaffolds.

Besides the validation of the QSAR model by the inhibitors with known IC_{50} values, this QSAR model was also involved in a hybrid virtual screening of CK2 natural product inhibitors. As shown in Table S2, the top 10 theoretical hits (M17-M26) were considered as potent CK2 inhibitors, with pIC_{50} values ranging from 4.34 to 12.03, which will be further evaluated and tested by the kinase assay. Interestingly, the top three compounds M18, M24, and M26, with predicted pIC_{50} values of 11.29, 8.79, and 12.03 respectively, exhibited higher values for two molecular descriptors, CATS2D_08_DA and F08 [C–O] (Table S2).

Additionally, molecular docking was performed to elucidate the binding modes of two theoretical hits M18 and M19 with CK2. As shown in Fig. 5, both M18 and M19 were sandwiched into the hydrophobic pocket consisting of Leu45, Val53, Val66, Ile95, Phe113, Met163 and Ile174. Polar interactions were also formed between the OH of M18 and the backbone CO of Val116, the carbonyl of M19 with NH2 of Val116, as the OH of M19 with the side chain of Lys68.

Comparison with previous 2D QSAR models of CK2 inhibitors

There has been growing interest over computational methods to construct 2D QSAR models of CK2 inhibitors for predicting the biological activities of new compounds. For instance, 2D-QSAR models of tricyclic quinoline analogs as CK2 inhibitors were developed using MLR and support vector machine methods. The atomic mass and polarizabilities and also number of heteroatoms in molecules were identified as the main independent factors contributing to the CK2 inhibition activity [22]. Ouammou et al performed QSAR studies on phenylaminopyrimidine-(thio) urea derivatives as CK2 inhibitors using MLR approach along with Quantum chemical descriptors, and found two most important descriptors LogP and D.M related to CK2 inhibition [23]. Since the mentioned QSAR studies were based on a dataset with a certain scaffold, and thus have the exclusive predictive ability for query compounds similar to the training set. Here a robust and predictive QSAR model were proposed based on CK2 natural product inhibitors with diverse chemical scaffolds, and showed good predictive ability for the 2-propenone derivatives outside the training and test sets. For natural products, CATS2D_08_DA, F08[C–O], and nArX were considered as three most important molecular descriptors influencing their inhibitory activities against CK2. To get a rigorous validation of the robustness and predictivity of this model, a comprehensive true external

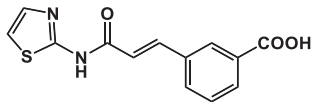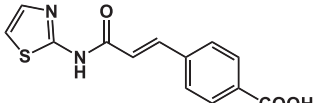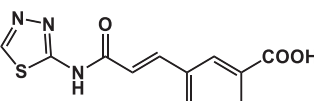**Table 3** The experimental activity (EA pIC$_{50}$), predicted activity (PA pIC$_{50}$) and residuals of the true external compounds
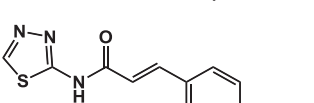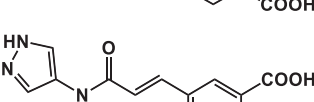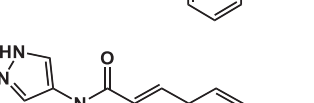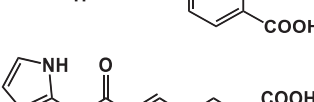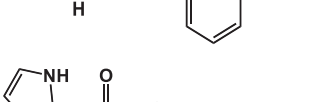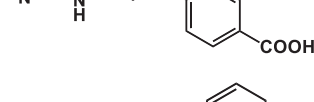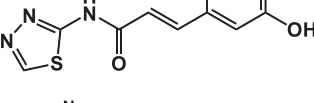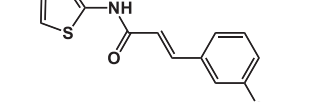
Name	Structures	EA pIC$_{50}$	PA pIC$_{50}$	Residuals
M1		5.13	5.89	−0.76
M2		5.03	3.61	1.42
M3		4.94	5.86	−0.92
M4		4.63	3.58	1.05
M5		3.99	5.91	−1.92
M6		3.89	3.63	0.26
M7		6.22	5.90	0.32
M8		6.10	3.62	2.48
M9		3.99	3.36	0.63
M10		3.85	3.39	0.46
M11		4.03	2.56	1.47

Table 3 (continued)
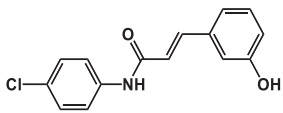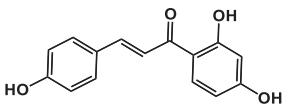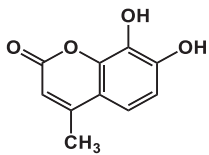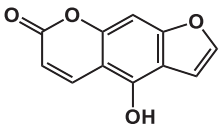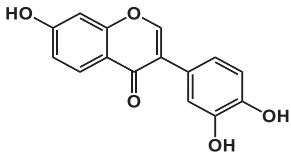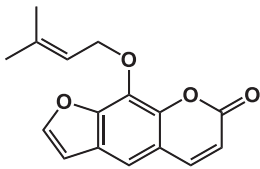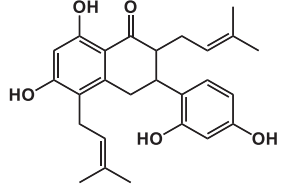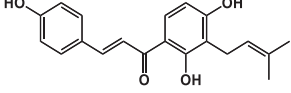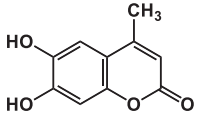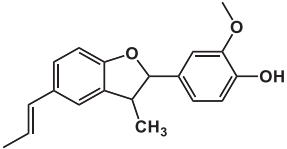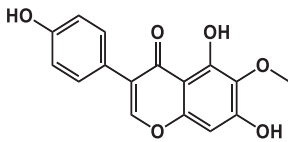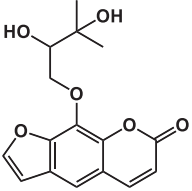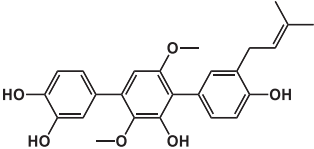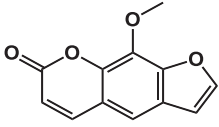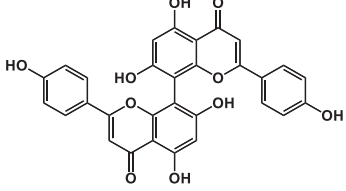
Name	Structures	EA pIC$_{50}$	PA pIC$_{50}$	Residuals
M12		4.16	3.52	0.64
M13		4.92	4.65	0.27
M14		3.97	5.11	−1.14
M15		3.70	2.97	0.73
M16		4.18	5.33	−1.15
M17		−	4.34	−
M18		−	11.29	−
M19		−	6.10	−
M20		−	5.08	−
M21		−	4.89	−

Table 3 (continued)

Name	Structures	EA pIC$_{50}$	PA pIC$_{50}$	Residuals
M22		–	5.81	–
M23		–	6.56	–
M24		–	8.79	–
M25		–	5.42	–
M26		–	12.03	–

test set with more diversified structures and broad-spectrum bioactivities is expected.

Conclusion

Natural products have been regarded as potential therapeutic leads for many diseases. Here, a robust and predictive QSAR model was established based on a diverse set of CK2 natural product inhibitors using the GA-MLR methodology. Besides the statistical parameters of $Q^2_{Loo} = 0.7914$ and $R^2 = 0.8220$ for the training set, and $Q^2_{ext} = 0.7921$ and $R^2_{ext} = 0.7998$ for the test set, the less than one order discrepancy between predicted and experimental pIC$_{50}$ values of the true external test set further verified the powerful predictive ability and high reliability of this developed model. Eight molecular descriptors were analyzed to elucidate the structural mechanisms responsible for the inhibitory activity of compounds. It is expected that this model, as presented in the current study, will be beneficial for the discovery of CK2 inhibitors.

Materials and methods

Data set and molecular descriptors calculation

In viewing of diverse molecular scaffolds and broad-spectrum bioactivities, 115 CK2 natural product inhibitors (Fig. 6) were selected from the published literature [4, 8, 10, 11]. The experimental IC$_{50}$ values of the entire dataset were converted to pIC$_{50}$ values (-Log IC$_{50}$) as the dependent variable distributed from 4.0 to 9.0. Then all these compounds were randomly split into a training set (95) and a test set (20) at a ratio of 5:1 as listed in Table S1 (Supplementary Materials). Following the general workflow shown in Fig. 7, a reliable QSAR model were developed to identify key molecular descriptors related to the inhibitory activities.

In order to get the approximate bioactive conformations of the data set, the co-crystallized compounds 33, 74, 108, 109 and 111–114 complexed with CK2 (Corresponding to PDB ID: 4DGM, 2QC6,1M2P, 2ZJW, 1M2Q, 6HOQ, 6HOR and 6HOT, respectively) were considered as bioactive conformations, and other compounds were constructed
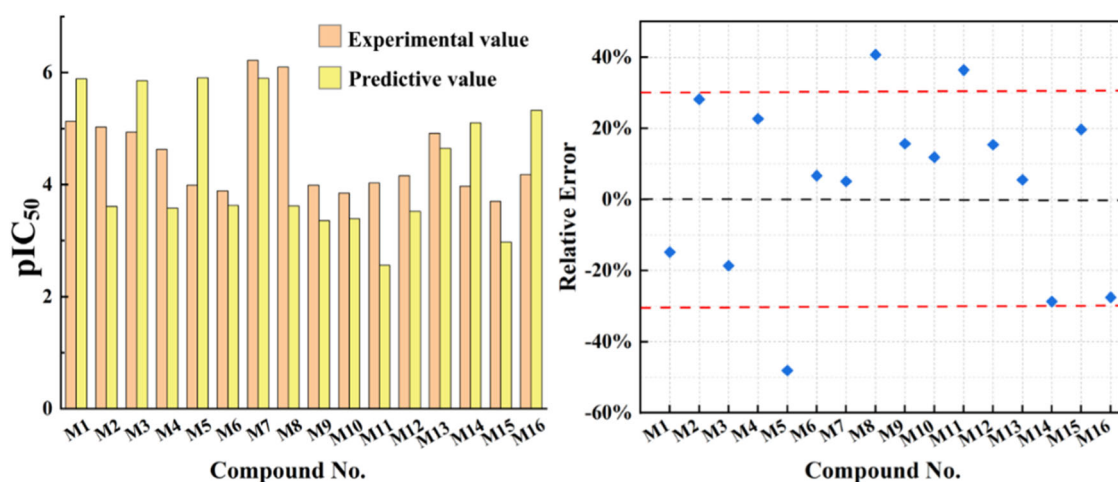
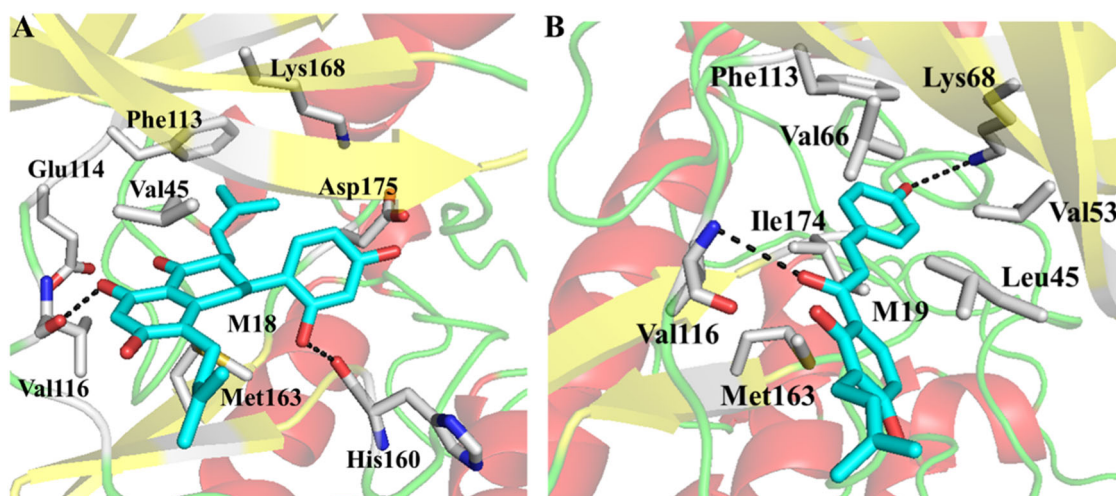Fig. 4 Experimental and Predicted pIC$_{50}$ values of the true external test set

Fig. 5 Binding modes of two theoretical hits, (A) M18, (B) M19 with CK2

with ChemBioDraw Ultra 14.0 (https://www.chemdraw.com.cn/) based on the bioactive conformations of their analogs. Then all compounds were subsequently optimized using 3D module. 2D molecular descriptors including basic descriptors, constitutive indices, ring descriptors, topological indices, connectivity indices, etc. were calculated by DRAGON 7.0 (https://chm.kode-solutions.net/). To reduce the redundancy of constant and intercorrelated descriptors, multicollinearity (the occurrence of high intercorrelation among two or more descriptors) terms need to be removed before multiple linear regression. In this section, constant and near-constant descriptors (more than 80% compounds sharing the same value) and highly inter-correlated descriptors (intercorrelation > 0.95) descriptors were excluded to remove useless information and improve modeling efficiency, and finally, 613 descriptors were retained for developing the regression model.

GA-MLR-OLS-based QSAR models

QSAR modeling was carried out by multiple linear regression (MLR)-ordinary least square (OLS) procedure as implemented in QSARINS 2.2.4 [24, 25]. Genetic algorithm technique was used to select the most relevant descriptors with respect to an objective function. Based on the selected descriptors, MLR analysis was performed on the training set and then, evaluated by test set. Since the ratio of the number of training compounds to the descriptor variables should be greater than 5 (N/M > 5) [26], the maximum number of variables (descriptors) included in the model was no more than 19 (95/5 = 19). Firstly, based on 613 molecular descriptors, a low-dimensional model is generated using the all-subset module to explore all possible combinations between two descriptors to avoid a completely random start of the GA. Subsequently, new combinations with additional

Fig. 6 Molecular skeleton structure of 115 CK2 natural product inhibitors
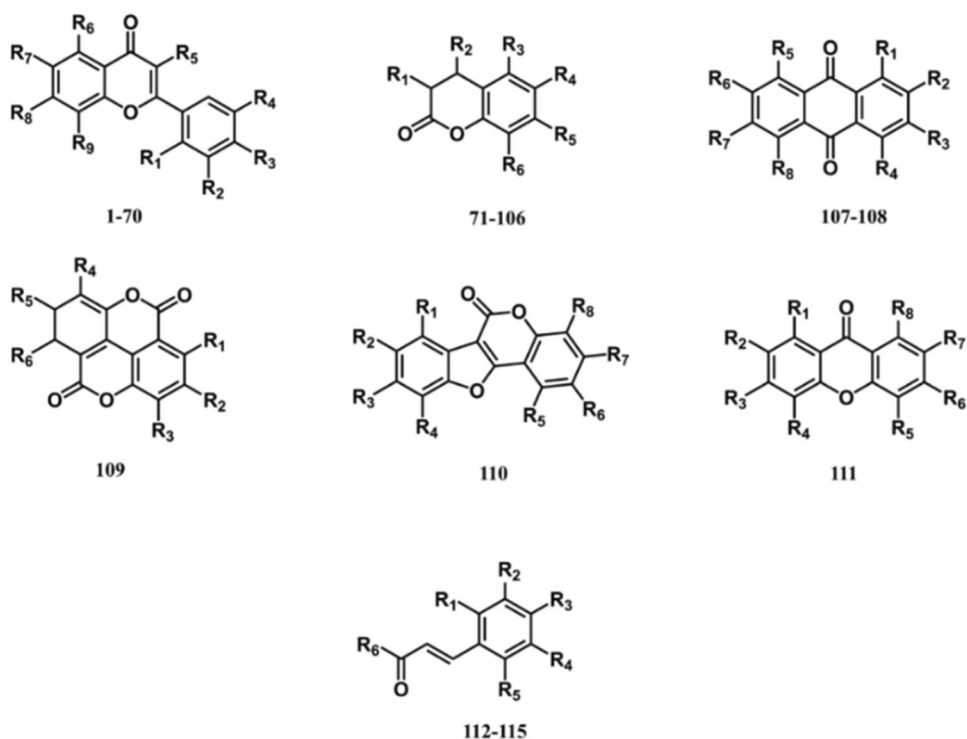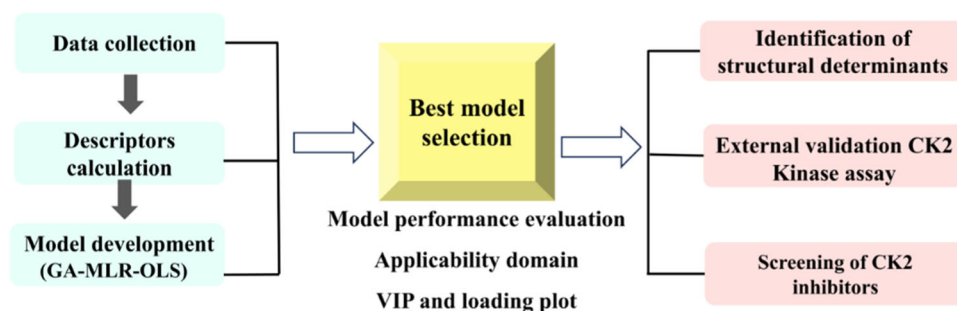

Fig. 7 Schematic illustration and applications of QSAR modeling for CK2 natural product inhibitors

descriptors are explored using the GA module to measure the robustness of the model and generate the best model using the leave-one-out (LOO) cross-validation coefficient (Q^2_{LOO} derived based on MLR) as the fitness function. The next step was to exclude the worse subsets, and then breed the remaining subsets. The modeling parameters include population size, mutation rate and number of genetic generations and set to 200, 20 and 2000, respectively. Only the top 10 models were retained for each set of models with the same size. Furthermore, the QUIK (Q Under Influence of K) rule was employed to exclude the generated models with high predictor collinearity [27].

Model performance evaluation

A LOO cross-validation (internal validation) and test set validation (external validation) were performed to evaluate the performance of models.

As for the LOO-based internal validation of training set, the parameters including cross-validation coefficient Q^2_{LOO}, correlation coefficient R^2 and adjustment coefficient R^2_{adj} were regarded as significant indexes to assess the robustness and stability of a model. Two parameters, $Q^2_{LOO} \geq 0.5$ and $R^2 \geq 0.8$ are considered as the popular criteria to define the robustness and predictive ability of QSAR models [28]. In order to check the stability of QSAR models, Y-scrambling validation was also applied by response scrambling with maximum iterations of 2000 [24, 29]. The model was not considered to be generated randomly when the resulting models obtained with randomized response should have significantly lower Q^2 values than the proposed ones.

The external predictive performance of a model was evaluated by taking Q^2_{ext}, r^2_m, R^2_{ext}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, as well as CCC_{ext}, $RMSE_{ext}$, r^2_m and Δr^2_m into consideration. Here $Q^2_{Fn} > 0.7$; $CCC_{test} > 0.85$; $r^2_m > 0.5$; $\Delta\ r^2_m < 0.2$

were used to a criteria of a good external prediction performance. Moreover, root mean squared error ($RMSE$, including $RMSE_{tr}$ and $RMSE_{test}$) were taken as an additional measure to assess the accuracy of a model.

The model performance was ranked by Multi-Criteria Decision Making (MCDM) technique using scores from 0 to 1, in which 1 means the best performance and 0 the worst [24]. Internal fit (R^2), cross-validation (Q^2), and external validation (Q^2_{ext} and R^2_{ext}) are the parameters evaluated by the MCDM procedure. A QSAR model with the highest value of all statistical parameters and the best MCDM score will be selected as the optimal one.

AD analysis

Only compounds falling into the AD of one QSAR model could be reliably predicted [16]. AD was presented through the Williams plot (leverage value h versus the standardized residual of each compound). The hat value (h) is used to characterize the leverage of a compound in the original variable space, and the leverage value is set to $h* = 3(p + 1)/n$, where p represents the number of descriptors and n represents the number of compounds in the training set. For the effect space, a compound will be recognized as a response outlier if its normalized residuals are greater than 3 [24, 30].

Compounds synthesis and CK2 kinase assay

A true external set was defined as compounds (with or without experimental biological activities) which were not involved in QSAR modeling, and was employed to further evaluate the predictive ability of QSAR models. In this study, 26 compounds including 12 novel CK2 inhibitors synthesized by Qi et al. [12], 4 synthesized compounds and 14 compounds screened from natural product libraries were collected as a true external set.

The synthesis and chemical characterization of compounds M9-M12 is described in Scheme 1. The materials and reagents were purchased from Sigma-Aldrich (Shanghai, China). The intermediate II was obtained by protecting the hydroxyl group of starting material I, which went through substitution with oxalyl chloride to provide Compound III. And then various amines were coupled to afford amide IV. Finally, hydrolysis of the ester group with sodium hydroxide followed by hydrochloric acid treatment were performed to afforded the desired compounds. All synthesized compounds were characterized with various spectroscopic methods. 1HNMR spectra were measured in DMSO-d6 on a Bruker Avance DRX 400 MHz. Mass spectra (MS) were recorded on a Q-TOF maxis (Agilent Technologies 1290 Infinity).

4,3-(3-(2-amino-1,3,4-thiadiazol-5-yl)prop-2-en-1-yl) phenol (M9)

1HNMR(400 MHz, DMSO-d6)δ 8.89 (s, 1H), 8.50 (s, 1H), 7.66 (d, J = 15.8 Hz, 1H), 7.19 (d, J = 7.7 Hz, 2H), 6.80 (dd, J = 17.4, 11.5 Hz, 2H), 6.60 (dd, J = 9.7, 5.2 Hz, 1H).ESI-MS m/z:246.0344 [M-H]-, calcd for C11H9N3O2S :247.0415.

3-(3-(2-aminothiazoly) propenyl) phenol (M10)

1HNMR (400 MHz, DMSO-d6) δ 12.36 (s, 1H), 9.76 (s, 1H), 7.62 (d, J = 15.8 Hz, 1H), 7.51 (d, J = 3.5 Hz, 1H), 7.25 (dd, J = 5.6, 2.1 Hz, 2H), 7.07 – 7.01 (m, 2H), 6.87–6.82 (m, 2H).ESI-MS m/z:245.0386 [M-H]-, calcd for C12H10N2O2S :246.0463.

3-(3-(4-propenamide) propenyl) phenol (M11)

1HNMR (400 MHz, DMSO-d6δ 10.31 (s, 1H), 7.71 (d, J = 7.7 Hz, 2H), 7.47 (d, J = 15.7 Hz, 1H), 7.33 (t, J = 7.9 Hz, 2H), 7.21 (t, J = 8.0 Hz, 1H), 7.06 (t, J = 7.4 Hz, 1H), 6.98 (d, J = 6.4 Hz, 2H), 6.78 (d, J = 15.7 Hz, 2H).ESI-MS m/z:238.0881[M-H]-, calcdfor C15H13NO2: 239.0946.

3-(3-(4-chloroaniline) propenyl) phenol (M12)

1HNMR (400 MHz, DMSO-d6) δ 10.35 (s, 1H), 9.66 (s, 1H), 7.73 (d, J = 8.8 Hz, 2H), 7.50 (d, J = 15.7 Hz, 1H), 7.39 (d, J = 8.8 Hz, 2H), 7.24 (t, J = 7.8 Hz, 1H), 7.07–6.99 (m, 2H), 6.82 (d, J = 8.0 Hz, 1H), 6.73 (d, J = 15.7 Hz, 1H).ESI-MSm/z:272.0486 [M-H]-, calcd for C15H12NO2Cl :273.0557.

ADP-Glo^{TM} Kinase Assay (Promega, Madison, WI, USA) was used to test the inhibitory activity of the true external set against CK2 [31, 32]. The IC$_{50}$ values for test compounds were determined at 5% DMSO with 7 concentrations of each tested inhibitor at the range of 0.002–2000 μM. The kinase reaction was performed in 25 μL mixture containing 10 μL casein kinase 2 solution, 5 μL different concentrations compounds and 10 μL substrate/ATP mixing solution. Incubate for 60 min at room temperature, 25 μL of ADP-Glo^{TM} Reagent was added to terminate the kinase reaction and deplete the remaining ATP. Then 50 μL Kinase Detection Reagent was added to convert ADP to ATP within 30 min, and allow the newly synthesized ATP to produce a luminescence signal using luciferase/luciferin reaction, which was recorded by the luminescence panel of Microplate Reader (Enspire2300-001A, Perkin Elmer, Waltham, MA, USA). The luminescent signal is proportional to the ADP concentration produced and correlated with the kinase activity.
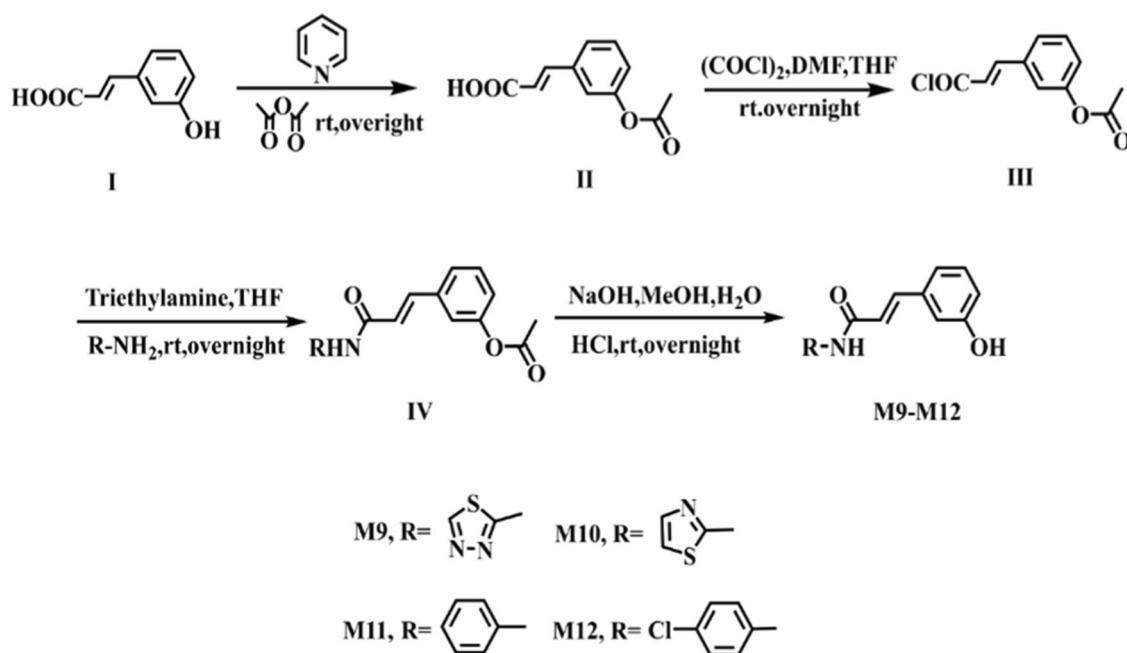
Scheme 1 Synthesis route for the compounds M9-M12

Molecular Docking

In order to predict the binding modes of theoretical hits with CK2, molecular docking was performed by using Auto-Dock Vina v1.2.0 [33]. The active site of the receptor was defined as a three-dimensional grid of $(50 \times 50 \times 50)$ points with a grid spacing of 0.375 Å at the center of mass of the ligand (PDB ID: 4DGN and 6HOQ) [11, 34]. The Lamarckian genetic algorithm (LGA) was employed as the conformational search method to explore the binding modes between CK2 and the inhibitors.

Author contributions YX: conceptualization, data curation, writing-original draft; YZ: data curation, formal analysis, writing-original draft; YY: investigation, validation; NZ: conceptualization, methodology, writing–review & editing; GS: methodology, visualization. TF: methodology; LZ: writing—review and editing; RZ: supervision.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

References

1. Chopra B, Dhingra AK. Natural products: a lead for drug discovery and development. Phytother Res. 2021;35:4660–702. https://doi.org/10.1002/ptr.7099.

2. Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021;20:200–16. https://doi.org/10.1038/s41573-020-00114-z.

3. Chen Y, Wang Y, Wang J, Zhou Z, Cao S, Zhang J. Strategies of targeting CK2 in drug discovery: challenges, opportunities, and emerging prospects. J Med Chem. 2023;66:2257–81. https://doi.org/10.1021/acs.jmedchem.2c01523.

4. Chilin A, Battistutta R, Bortolato A, Cozza G, Zanatta S, Poletto G, et al. Coumarin as attractive casein kinase 2 (CK2) inhibitor scaffold: an integrate approach to elucidate the putative binding motif and explain structure-activity relationships. J Med Chem. 2008;51:752–9. https://doi.org/10.1021/jm070909t.

5. Zhang N, Chen WJ, Zhou Y, Zhao H, Zhong RG. Rational design of coumarin derivatives as CK2 inhibitors by improving the interaction with the hinge region. Mol Inform. 2016;35:15–18. https://doi.org/10.1002/minf.201500091.

6. McCarty MF, Assanga SI, Lujan LL. Flavones and flavonols may have clinical potential as CK2 inhibitors in cancer therapy. Med Hypotheses. 2020;141:109723. https://doi.org/10.1016/j.mehy.2020.109723.

7. Golub AG, Bdzhola VG, Kyshenia YV, Sapelkin VM, Pry-khod'ko AO, Kukharenko OP, et al. Structure-based discovery of novel flavonol inhibitors of human protein kinase CK2. Mol Cell Biochem. 2011;356:107–15. https://doi.org/10.1007/s11010-011-0945-8.

8. Golub AG, Bdzhola VG, Ostrynska OV, Kyshenia IV, Sapelkin VM, Prykhod'ko AO, et al. Discovery and characterization of synthetic 4'-hydroxyflavones-New CK2 inhibitors from flavone family. Bioorg Med Chem. 2013;21:6681–9. https://doi.org/10.1016/j.bmc.2013.08.013.

9. Yim H, Lee YH, Lee CH, Lee SK. Emodin, an anthraquinone derivative isolated from the rhizomes of Rheum palmatum, selectively inhibits the activity of casein kinase II as a competitive inhibitor. Planta Med. 1999;65:9–13. https://doi.org/10.1055/s-1999-13953.

10. Sekiguchi Y, Nakaniwa T, Kinoshita T, Nakanishi I, Kitaura K, Hirasawa A, et al. Structural insight into human CK2 alpha in

complex with the potent inhibitor ellagic acid. Bioorg Med Chem Lett. 2009;19:2920–3. https://doi.org/10.1055/s-1999-13953.

11. Cozza G, Zonta F, Dalle Vedove A, Venerando A, Dall'Acqua S, Battistutta R, et al. Biochemical and cellular mechanism of protein kinase CK2 inhibition by deceptive curcumin. FEBS J. 2020;287:1850–64. https://doi.org/10.1111/febs.15111.

12. Qi X, Zhang N, Zhao L, Hu L, Cortopassi WA, Jacobson MP, et al. Structure-based identification of novel CK2 inhibitors with a linear 2-propenone scaffold as anti-cancer agents. Biochem Biophys Res Commun. 2019;512:208–12. https://doi.org/10.1016/j.bbrc.2019.03.016.

13. Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F. Natural product drug discovery in the artificial intelligence era. Chem Sci. 2022;13:1526–46. https://doi.org/10.1039/d1sc04471k.

14. Zhang N, Zhong R. Docking and 3D-QSAR studies of 7-hydroxycoumarin derivatives as CK2 inhibitors. Eur J Med Chem. 2010;45:292–7. https://doi.org/10.1016/j.ejmech.2009.10.011.

15. Haidar S, Marminon C, Aichele D, Nacereddine A, Zeinyeh W, Bouzina A, et al. QSAR model of indeno[1,2-b]indole derivatives and identification of N-isopentyl-2-methyl-4,9-dioxo-4,9-Dihydronaphtho[2,3-b]furan-3-carboxamide as a potent CK2 inhibitor. Molecules. 2019;25. https://doi.org/10.3390/molecules25010097.

16. Zhong S, Lambeth DR, Igou TK, Chen Y. Enlarging applicability domain of quantitative structure–activity relationship models through uncertainty-based active learning. ACS EST Engg. 2022;2:1211–20. https://doi.org/10.1021/acsestengg.1c00434.

17. Sadeghi F, Afkhami A, Madrakian T, Ghavami R. QSAR analysis on a large and diverse set of potent phosphoinositide 3-kinase gamma (PI3Kγ) inhibitors using MLR and ANN methods. Sci Rep. 2022;12:6090. https://doi.org/10.1038/s41598-022-09843-0.

18. Liu Y, Bi M, Zhang X, Zhang N, Sun G, Zhou Y, et al. Machine learning models for the classification of CK2 natural products inhibitors with molecular fingerprint descriptors. Processes. 2021. https://doi.org/10.3390/pr9112074.

19. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H. Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model. 2012;52:396–408. https://doi.org/10.1021/ci200520g.

20. Pierre F, Chua PC, O'Brien SE, Siddiqui-Jain A, Bourbon P, Haddach M, et al. Discovery and SAR of 5-(3-chlorophenylamino)benzo[c][2,6]naphthyridine-8-carboxylic acid (CX-4945), the first clinical stage inhibitor of protein kinase CK2 for the treatment of cancer. J Med Chem. 2011;54:635–54. https://doi.org/10.1021/jm101251q.

21. Sarno S, Papinutto E, Franchin C, Bain J, Elliott M, Meggio F, et al. ATP site-directed inhibitors of protein kinase CK2: an update. Curr Top Med Chem. 2011;11:1340–51. https://doi.org/10.2174/156802611795589638.

22. Eslam P, Reza A, Mohammad R. Q SAR study of CK2 inhibitors by GA-MLR and GA-SVM methods. Arab J Chem. 2019;12:2141–9. https://doi.org/10.1016/j.arabjc.2014.12.021.

23. Amina G, Abdellah E, Hicham E, Abdelkrim O. QSAR modeling, molecular docking studies and ADMET prediction on a series of phenylaminopyrimidine-(thio) urea derivatives as CK2 inhibitors.

24. Mater Today Proc. 2022;51:1851–62. https://doi.org/10.1016/j.matpr.2020.08.044.

24. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSAR-INS: a new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem. 2013;34:2121–32. https://doi.org/10.1002/jcc.23361.

25. Gramatica P, Cassani S, Chirico N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J Comput Chem. 2014;35:1036–44. https://doi.org/10.1002/jcc.23576.

26. Gramatica P. Principles of QSAR modeling: comments and suggestions from personal experience. IJQSPR. 2020;5. https://doi.org/10.4018/IJQSPR.20200701.oa1.

27. Todeschini R, Consonni V, Maiocchi A. The K correlation index: theory development and its application in chemometrics. Chemom Intell. 1999;46:13–29. https://doi.org/10.1016/S0169-7439(98)00124-5.

28. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20:69–276. https://doi.org/10.1016/s1093-3263(01)00123-1.

29. Hao Y, Sun G, Fan T, Sun X, Liu Y, Zhang N, et al. Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods. Ecotoxicol Environ Saf. 2019;186:109822. https://doi.org/10.1016/j.ecoenv.2019.109822.

30. Roy K, Ambure P, Kar S. How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? ACS Omega. 2018;3:11392–406. https://doi.org/10.1021/acsomega.8b01647.

31. Zegzouti H, Zdanovskaia M, Hsiao K, Goueli SA. ADP-Glo: a bioluminescent and homogeneous ADP monitoring assay for kinases. Assay Drug Dev Technol. 2009;7:560–72. https://doi.org/10.1089/adt.2009.0222.

32. Liu S, Hsieh D, Yang YL, Xu Z, Peto C, Jablons DM, et al. Coumestrol from the National Cancer Institute's natural product library is a novel inhibitor of protein kinase CK2. BMC Pharmacol Toxicol. 2013;14:36. https://doi.org/10.1186/2050-6511-14-36.

33. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Com Chem. 2010;31:455–61. https://doi.org/10.1002/jcc.21334.

34. Lolli G, Cozza G, Mazzorana M, Tibaldi E, Cesaro L, Donella-Deana A, et al. Inhibition of protein kinase CK2 by flavonoids and tyrphostins. A structural insight. Biochemistry. 2021;51:6097–107. https://doi.org/10.1021/bi300531c.