Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

2D-QSAR and docking study of a series of coumarin derivatives as inhibitors of CDK (anticancer activity) with an application of the molecular docking method

Rania Kasmi [a], Elghalia Hadaji [a,*], Oussama Chedadi [a], Abdellah El Aissouq [a], Mohammed Bouachrine [b,c], Abdelkrim Ouammou [a]

[a] LIMOME Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[b] MCNS Laboratory, Faculty of Science, University Moulay Ismail, Meknes, Morocco
[c] EST Khenifra, Sultan Moulay Sliman University, Morocco

ARTICLE INFO

Keywords:
Pharmaceutical chemistry
Theoretical chemistry
DFT
Coumarin
MLR
Validation
Domain of applicability
Docking
Cross validation
Y-randomization
External validation
Qsar

ABSTRACT

Quantitative Structure Activity Relationship (QSAR) analysis techniques are tools largely utilized in many research fields, including drug discovery processes.

In this work electronic descriptors are calculated with the Gaussian 03W software using the DFT method with the BecKe 3-parameters exchange functional and Lee-Yang-Parr correlation functional, with Kohn and Sham orbitals (KS) developed on a Gaussian Basis of type 6-31G (d), in combination with five Lipinski parameters that have been calculated with ChemOffice software, in order to develop a statistically verified 2D-QSAR model able to predict the biological activity of new molecules belonging to the same range of coumarins rather than chemical synthesis and biological evaluations that require more time and resources. Two QSAR models against both MCF-7 and HepG-2 cell lines are obtained using the multiple linear regression method.

The predictive power of these models has been confirmed by internal and external validation. The Leverage method was used to determine the domain of applicability of the 2D-QSAR models developed. The results indicate that the best QSAR model is the one that links the 2D descriptors with the CDK inhibitory activity of the cell line (HepG-2) $R^2 = 0.748$, $R^2cv = 0.618$, MSE = 0.03 for the learning series and $R^2 = 0.73$, MSE = 0.18 for the test series. This model implies that coumarin inhibitory activity is strongly related to dipole moment and the number of hydrogen bond donors. The results obtained suggest the importance of studying structure-activity relationships as a principal axis in drug design. The docking procedure using AutoDOCK Tools was also used to understand the mechanisms of molecular interactions and consequently, to develop new inhibitors.

1. Introduction

There are monitoring points in a cell cycle to supervise them when a cell decides to divide into two identical daughter cells according to fundamental aspects namely: a sufficiently important growth of the cell, the DNA is completely repaired, that this DNA is duplicated (2 complete and identical copies of DNA) and that the mitotic spindle with the microtubules and the alignment of the chromosomes was properly constituted.

The control of different stages of the cell cycle is ensured by molecules (proteins) either alone or in macromolecular complexes (several proteins associated with each other). For this reason, we use two proteins: cyclins and CDK.

Cyclin is a protein that is always produced inside the cell in small quantities, they appear and then suddenly disappear at specific moments of the cycle, periodically. The CDK can, therefore, be activated or deactivated depending on whether or not they are associated with their cyclin.

As soon as CDK is combined with a cyclin, they become enzymatically active and are then able to activate other proteins by phosphorylating them to progress the cell cycle [1].

In brief, cell cycle disorders lead to uncontrolled proliferation, which can lead to cancer.

* Corresponding author.
E-mail address: hadaji_ghalia@hotmail.fr (E. Hadaji).

https://doi.org/10.1016/j.heliyon.2020.e04514
Received 13 February 2020; Received in revised form 14 March 2020; Accepted 16 July 2020

The specific inhibitors of CDK are important targets in drug discovery due to their anti-tumor activity, they induce apoptosis by disrupting the cell cycle.

Morsy and his collaborators [2] tested the anticancer activity of a series of 24 coumarin derivatives in vitro against 2 tumor cell lines, human breast cancer (MCF-7) and hepatocellular carcinoma (HepG-2).

Coumarins are oxygenated heterocycles belonging to the benzopyrone family, whose name according to IUPAC is 2H-1-benzopyran-2-one [3], they are produced by combining a benzene ring with a pyran, having a ketone function in alpha position with respect to oxygen.

Isolated the first time from Coumarounaodorata by Vogel in 1820 [4], today nearly a thousand of coumarins have been described in more than 800 species of plants and microorganisms.

From a structural point of view, they are classified into simple coumarins with substituents on the benzene ring, furanocoumarins, pyranocoumarins, those substituted at positions 3 and or 4 and the latter are dicoumarins and tricoumarins [5].

Coumarins have many biochemical and pharmacological properties. The activity of these molecules depends on the structure and nature of the substituents. The majority of coumarins and their derivatives have been subjected to deep investigations to evaluate their effects on human health. Research has shown that they can be anti-HIV, anti-tumor [6], anti-cancer, anti-microbial [7], anti-inflammatory [8, 9], anti-fungal [10], antioxidant [11] and even vasodilator agents [12].

A 2D-QSAR study was processed to find descriptors that can be correlated to anti-cancer activity expressed in IC_{50}(mol/L) values [the concentration of test compounds required to reduce the cell survival fraction to 50% of the control], they converted to negative logarithms of IC_{50}(pIC_{50}) to obtain the linear relationship with the independent variables.

The principle of QSAR computational methods is to implement a mathematical relationship quantitatively linking molecular descriptors with a macroscopic observable (physicochemical property or biological activity) for a series of similar chemical compounds using statistical data-analytical methods.

The most general mathematical form of QSAR is [13]:

$$\text{Activity} = f\,(X)$$

X: physicochemical and/or structural properties.

The goal of these methods is, therefore to analyze the structural data to detect the determining factors for the property or activity measured.

In the last step, the developed models are subjected to various internal and external validation procedures to test their statistical significance, robustness and predictive power [14]. The current challenge in the QSAR model development process is no longer in developing a statistically robust model to predict activity within the calibration set, but in developing a model that can accurately predict the activity or property of new chemicals [15].

The QSAR study of 24 coumarins was carried out using descriptors from quantum chemistry to quantify the different inter and intra-molecular interactions, and also has the advantage of being directly related to the reactivity properties of molecular systems, the latter is derived from the DFT method which can achieve similar accuracy to other methods in less time and lower cost from a computer point of view, and to describe the compounds that could be orally delivered drugs we used the descriptors that correspond to the Lipinski rule (see Figure 1).

Structural biology is interested in the relationship between the structure of molecules and the activation or inhibition of their biological activity, and this can only be done by predicting the affinity between two molecules, understanding how they function, and defining the residues involved. This is the problem then of molecular docking, which aims to predict the interactions intervening in the formation of molecular complexes, which is considerably easier to implement, cheaper, and faster than the use of in vitro experimental methods [16].

Docking software is therefore practical tools for the design of new ligands likely to interact more favorably with targets of therapeutic interest [17].

2. Materiel and methods

2.1. Experimental data

At this stage, we evaluated the values of the inhibitory and anticancer activities of CDK in vitro against two types of human tumor cell lines from the work of Shaimaa A. Morsy et al [2] who developed the design of a series of 6-methyl-4-substituted coumarin and 4-substituted benzocoumarin as shown in Figure 2. Reported values of IC_{50} (mol/L) were converted to pIC_{50} by taking a negative logarithm ($pIC_{50} = -\log 10\ IC_{50}$) and then used as dependent variables to develop the QSAR model. The IC_{50} values of the compounds 5h, 7d, 7h, 7h, 9h, 9a, 13a and 13d mentioned in Table 1 showed a remarkably high affinity and selectivity towards the MCF-7 and HepG-2 cell lines, in addition to compound 13a which has the greatest cytotoxic activity compared to 5-fluorouracil (the reference molecule).

The structural potential of our derivatives has been confirmed by a QSAR study to find descriptors that can be correlated to activity and subsequently for the successful design of new improved coumarin anticancer structures.

2.2. Molecular descriptors

Obtaining a statistically robust model depends very much on the ability of the descriptors, which are the final result of a logical and mathematical procedure [18], to encode the variation of activity with the structure.

The information coded by the descriptors generally depends on the type of molecular representation and the algorithm defined for its calculation and to predict the correlation between the 24 coumarin derivatives and their antitumor activity against the two cell lines MCF-7 and HepG-2. The chemical structures of the compounds were drawn using ChemDraw Professional 16.0 software [19] the initial optimizations of the geometry were carried out with the Molecular Mechanics (MM) method using the MM2 force field implemented in the Chem3D 16.0 software [19] The values of the 5 Lipinski descriptors [20], for all molecules including molecular weight (MW), lipophilicity (log P), hydrogen receptors (HA), hydrogen bond donors (DH) and the number of NRB rotational bonds, were calculated using the program module "calculation properties" as compiled in Table 2.

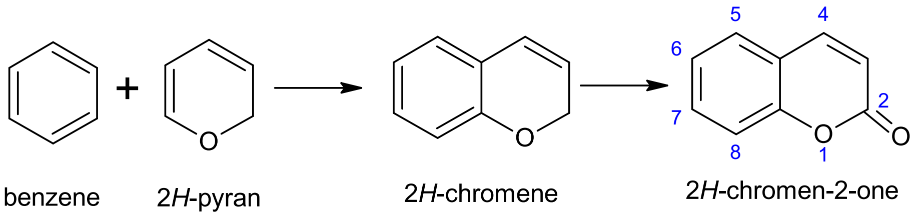

Figure 1. The combination of benzene with a pyran into a coumarin.

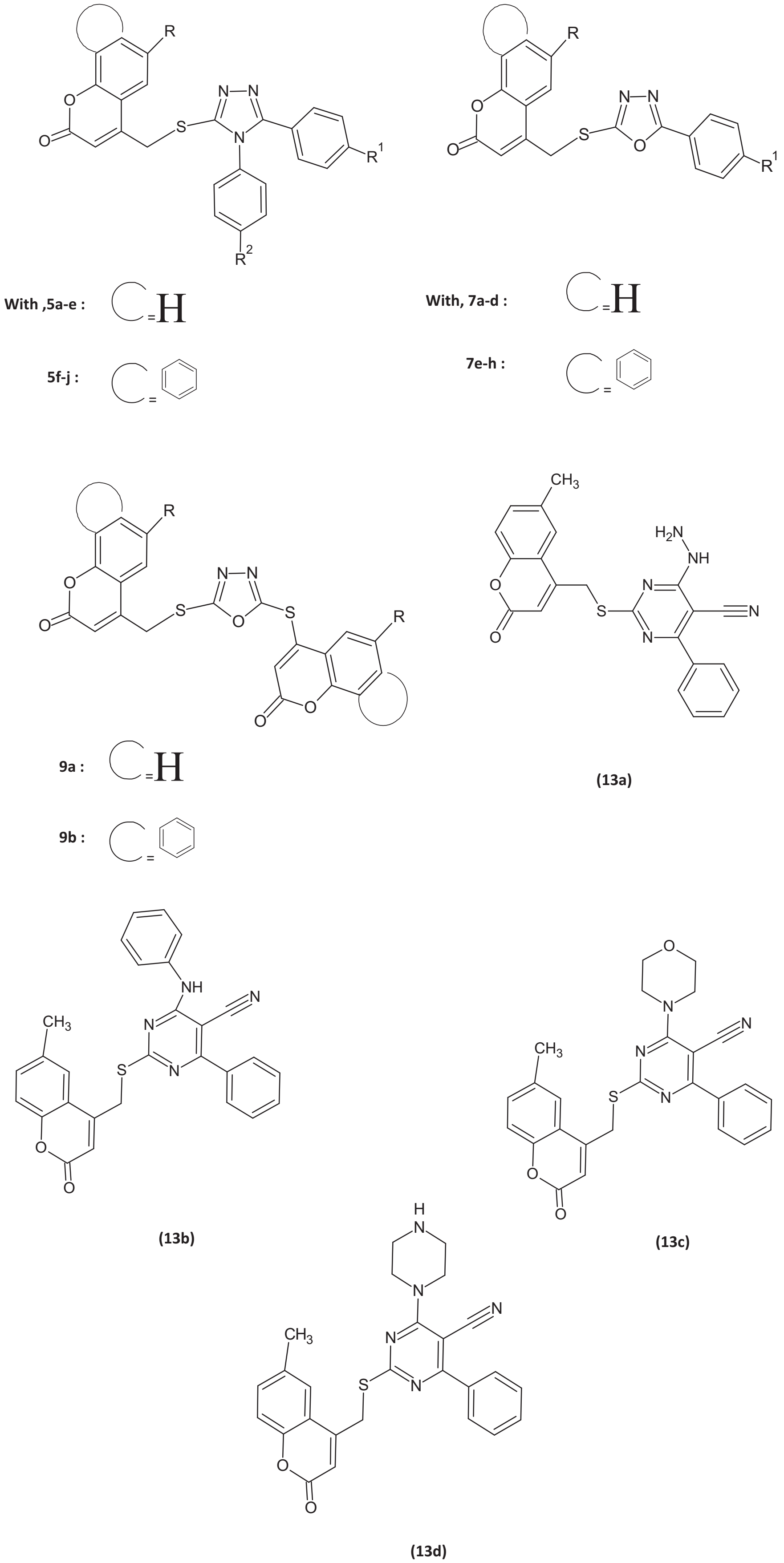

Figure 2. Design of the studied coumarin derivatives.

Table 1. The different substituents associated with coumarin derivatives as well as the experimental values of cytotoxic activity.

Compounds	R	R^1	R^2	$p\mathrm{IC}_{50}$exp(HepG-2)	$p\mathrm{IC}_{50}$exp(MCF-7)
5a	CH_3	H	H	7.114	6.842
5b	CH_3	CH_3	H	6.879	7.301
5c	CH_3	NO_2	H	7.244	6.754
5d	CH_3	H	F	6.848	6.585
5e	CH_3	CH_3	F	7.046	6.896
5f	H	H	H	7.398	6.854
5g	H	CH_3	H	6.693	6.721
5h	H	NO_2	H	7.602	6.807
5i	H	H	F	7.046	7.000
5j	H	CH_3	F	6.807	7.222
7a	CH_3	H	-	6.585	7.770
7b	CH_3	CH_3	-	6.796	7.046
7c	CH_3	NO_2	-	6.770	7.301
7d	CH_3	F	-	7.398	6.921
7e	H	H	-	7.097	7.398
7f	H	CH_3	-	6.745	7.523
7g	H	NO_2	-	6.708	7.097
7h	H	F	-	7.347	6.951
9a	CH_3	-	-	7.770	6.721
9b	H	-	-	6.824	7.699
13a	NHNH2	CN	C6H5	7.886	7.222
13b	C6H6N	CN	C6H5	6.959	7.824
13c	C4H8NO	CN	C6H5	6.721	6.987
13d	C4H9N2	CN	C6H5	7.523	7.699

The 24 structures were transferred to the Gaussian 03W [21] software for optimization with the B3LYP/6–31 G (d) method based on density-functional theory (DFT), to find a geometry for which the energy is minimal and to extract a set of 4 quantum chemistry descriptors, namely the energies of the highest occupied molecular orbit HOMO and the lowest vacant LUMO, the total energy of the molecule (Et) and the dipole moment, the nature of commonly used descriptors (structural, topological, electronic and geometric) and the degree of coding of molecular structural characteristics linked to certain specific physical properties are at the heart of any QSAR study [22].

Table 2. The values of parameters calculated for the 24 molecules of both cell lines MCF-7 and HepG-2.

N°	$p\mathrm{IC}_{50}$ (HepG-2)	$pIC_{50}(MCF7)$	*Et*	*EHOMO*	*ELUMO*	MD	Log P	AH	DH	MW	NRB
1	7.114	6.842	-45622.0823	-6.1666	-1.9475	5.1264	7.27	4	0	425.51	5
2	6.879	7.301	-46691.7168	-6.0444	-1.9301	5.6902	7.71	4	0	439.53	5
3	7.244	6.754	-51184.516	-6.5453	-3.1064	2.161	7.17	5	0	470.5	6
4	6.848	6.585	-48321.721	-6.2882	-1.9986	5.4411	7.44	5	0	443.5	5
5	7.046	6.896	-49391.3555	-6.1600	-1.9826	5.9475	7.87	5	0	457.52	5
6	7.398	6.854	-48732.3731	-6.0047	-1.9712	4.2549	8.10	4	0	461.54	5
7	6.693	6.721	-49802.0068	-5.9799	-1.9540	4.7195	8.54	4	0	475.57	5
8	7.602	6.807	-54294.8076	-6.1472	-3.0974	2.6889	8	5	0	506.54	6
9	7.046	7.000	-51432.0118	-6.0495	-2.0199	4.4166	8.27	5	0	479.53	5
10	6.807	7.222	-52501.6463	-6.0313	-2.0033	4.9825	8.7	5	0	493.56	5
11	6.585	7.770	-39876.3654	-6.5513	-2.1271	5.4651	5.19	4	0	350.39	4
12	6.796	7.046	-40946.008	-6.4876	-2.1034	5.9216	5.63	4	0	364.42	4
13	6.770	7.301	-45438.7237	-6.7154	-3.4250	5.6828	5.09	5	0	395.39	5
14	7.398	6.921	-42576.034	-6.6011	-2.1780	4.6732	5.35	5	0	368.38	4
15	7.097	7.398	-42986.6562	-6.1467	-2.1358	4.4837	6.02	4	0	386.42	4
16	6.745	7.523	-44056.2988	-6.1276	-2.1138	4.9235	6.46	4	0	400.45	4
17	6.708	7.097	-48549.0172	-6.2868	-6.2868	5.7511	5.92	5	0	431.42	5
18	7.347	6.951	-45686.3275	-6.1886	-6.1886	3.9234	6.18	5	0	404.41	4
19	7.770	6.721	-66700.6427	-6.7002	-2.6544	4.6144	5.6	4	0	494.6	6
20	6.824	7.699	-72921.2215	-6.2504	-2.6637	5.2431	7.26	4	0	566.66	6
21	7.886	7.222	-45457.6408	-6.5510	-6.5510	6.6793	3.39	6	2	415.47	6
22	6.959	7.824	-50238.6162	-6.4229	-6.4229	7.0884	5.87	5	1	476.55	7
23	6.721	6.987	-50243.1877	-6.5110	-6.5110	5.9868	4.47	6	0	470.55	6
24	7.523	7.699	-49702.8806	-6.0566	-6.0566	7.4255	4.28	6	1	469.56	6

2.3. Statistical analysis

Chemometric techniques constitute the mathematical basis for the construction of a QSAR model and among the easiest methods to interpret is the descending multiple linear regression analysis contained in XLSTAT 2014 software [20].

The selection of a regression method allows us to specify how the 9 descriptors were entered into the analysis. Using descending selection, we were able to construct two QSAR models from the 24 molecules studied.

In this case, all the variables are introduced into the equation and then eliminated one by one. the variable having the smallest partial correlation with the dependent variable is the variable whose deletion is studied first, if it meets the elimination criteria, it is removed. Once the first variable is eliminated, the elimination of the next variable remaining in the equation and having the lowest partial correlation coefficient is studied [21].

The procedure will be repeated until all the variables stored contribute significantly to the improvement of R^2 [22].

The objective will, therefore, be to develop predictive models for the cytotoxic activity of a series of coumarins against the 2 cell lines MCF-7 and HepG-2 using a reduced number of relevant descriptors and respecting all the protocol of the QSAR methodology.

1. First, the quality of the model is often visualized on a scatter diagram, on which the calculated values of biological activity are displayed, according to the experimental ones. More the points in this graph are close to the adjustment line, more the modeling is better and this can be evaluated by the determination coefficient R^2 [23]which measures the proportion of total variation of Y around the average explained by the regression.
   The Fisher index is also used as an indicator of the degree of statistical significance of the model, i.e., the relevance of the choice of descriptors that compose it.
2. In addition, to verify the stability of the predictive model and to test the influence of each element of the training set on the final model, "leave-one-out" cross-validation techniques are applied.
   Typically, one compound is removed each time, before redefining the model using the n-1 of the remaining compounds as training set, so that the biological activity value for the extracted compound is predicted once for all the compounds. This process is repeated n times for all the compounds of the initial set, thus obtaining a prediction for each object [24]. It is quantified by the R^2cv coefficient [25].
3. Subsequently, the perfect validity of the model is examined by external validation, which evaluates its generalization.
   This validation consists of predicting the activity of a test series that has not been included in any step of the model construction [26], it is characterized by the parameters R^2 (test) R^2cv (test) but they remain insufficient to verify the predictive strength of the QSAR models. Therefore, other parameters must be verified known as the" Tropsha criteria" [27].
4. A QSAR model is not universal, it is developed for a given number of compounds. For this, it is necessary to determine an applicability domain for each model. It is a tool that eliminates molecules from the test set that are located outside the chemical space of the training set, which is required in the validation processes implemented at the OECD level [28].
   Different methods are used to define the domain of applicability of a QSAR model, including that of "leverage" [29].
5. To check the robustness of a QSAR model, the randomization test is often used. It consists of randomly mixing the experimental activities for the learning series (Y) according to the same descriptors (X), new QSAR models are obtained. These latter must have very low R^2 and R^2cv performances. Another metric $^cR^2_p$ is also verified to meet this test, it must not be lower than the threshold value of 0.5, to conclude that the correlations are not fortuitous [30].

3. Results and discussion

3.1. Dataset for analysis

According to research conducted by Shaimaa A. Morsy et al, on the anticancer activity of coumarin derivatives concerning the 2 tumor cell lines MCF-7 and HepG-2, 25% of the synthesized molecules were randomly selected as a test set (6 molecules), while the other compounds (18 molecules) participated in the formation of the 2D-QSAR models.

The quantification of logarithmic values of biological activity pIC_{50} with relevant molecular descriptors was performed by multiple linear regression (MLR) analysis.

3.2. Multiple linear regressions (MLR)

The MLR was used to generate the linear 2D-QSAR models between pIC_{50} values and molecular descriptors. Two molecular descriptors (MD dipole moment and number of hydrogen bond donors DH) were selected to explain the variation of biological activity of coumarin derivatives. The best-selected models of HepG-2 and MCF-7 cell lines are given below:

◆ For the cell line (HepG-2):

$$\boldsymbol{pIC_{50}} = 7.915\text{-}0.196\ \mathbf{MD}+0.598\ \mathbf{DH} \tag{2}$$

Regression statistics:

$\mathbf{N} = 18\ \mathbf{R} = 0.86\ \mathbf{R^2} = 0.748\ \mathbf{R^2_{ajust}} = 0.715\ \mathbf{MSE} = 0.03\ \mathbf{MAE} = 0.13$

$\mathbf{N_{test}} = 6\ \mathbf{R^2_{test}} = 0.73\ \mathbf{MSE_{test}} = 0.18\ \mathbf{MAE_{test}} = 0.38\ \mathbf{F} = \mathbf{22.312}$

◆ For the cell line (MCF-7):

$$\boldsymbol{pIC_{50}} = 7.789–0.159\ \mathbf{MD}+0.520\ \mathbf{DH} \tag{3}$$

Regression statistics:

$\mathbf{N} = 18\ \mathbf{R} = 0.74\ \mathbf{R^2} = 0.545\ \mathbf{R^2_{ajust}} = 0.5\ \mathbf{MSE} = 0.04\ \mathbf{MAE} = 0.15$

$\mathbf{N_{test}} = 6\ \mathbf{R^2_{test}} = 0.52\ \mathbf{MSE_{test}} = 0.20\ \mathbf{MAE_{test}} = 0.4\ \mathbf{F} = \mathbf{8.980}$

For HepG-2, 74.8% of the variability is explained by the dipole moment and the number of hydrogen bond donors, the same descriptors were obtained for MCF-7 but explain only 54.5% of the variability. Generally, more the value of R^2 will be close to 1 (ideal case), more the predicted and observed values are correlated.

According to ANOVA [31] tables S1 and S2 (in the <i>supplementary material</i>), the observed Fisher statistics [32] ($F_{obs1} = 22.312$ and $F_{obs2} = 8.980$) are greater than [F_{crit} (0.05; 2;15) = 3.68], which allows to accept the alternative hypothesis H1 and to confirm that there is at least one coefficient different from zero, i.e. a descriptor correlated with the inhibitory activity explained by pC_{50} values.

From tables S3 and S4 (in the *supplementary material*), the observed student statistic values are higher than those in the distribution table, t (0.025; 15) = 2.131, this allows us to reject the null hypothesis, i.e. the coefficients included in the two different models are significantly different from zero. This judgment is consolidated by the low probability values for the descriptors in Eqs. (1) and (2).

Afterward, we proceeded to the problem of co-linearity and multicollinearity [33] respectively through the examination of the correlation matrix, by calculating the correlation coefficient for all possible pairs of descriptors and the confirmation by the tolerance factor [34] indicated by:

$$\mathrm{TF}(\boldsymbol{xk}) = \frac{1}{\boldsymbol{VIF}(\boldsymbol{xk})} = 1 - \boldsymbol{R}^2_{xk}$$

where VIF (xk) is the inflation factor of the variance for the descriptor xk and R^2xk is the squared correlation coefficient resulting from the regression of the descriptor xk on all other descriptors.

The examination of the correlation matrices (Tables 3 and 4) confirms the absence of collinearity problems between the descriptors of the 2 models, explained by the low values of the correlation coefficients ($R < 0.9$) and since the TF values are all less than 0.5, we can confirm the absence of strong multicollinearity between the descriptors.

The positive sign of DH in regression Eqs. (2) and (3) indicates that $p\mathrm{IC}_{50}$ is directly proportional to this descriptor while the negative coefficients for MD indicate that inhibition is inversely proportional to this descriptor.

A high value of MD dipole moment is expected to participate in hydrogen bonds, as well as dipole-dipole interactions and π-π stacking, it is often used to explain a molecule's activity because it can be directly related to its chemical reactivity.

Since the number of H-bond donors (DH) has a positive sign in Eqs. (2) and (3), we need to increase the number of hydrogen atoms bound to the heteroatoms to boost the activity.

3.3. Internal and external validation

To evaluate the significance of generated models and their precise predictive capacity, internal and external validations were used [35, 36]. The best HepG-2 and MCF-7models revealed a leave one out cross-validation coefficients R^2_{cv} [37] values of 0.618 and 0.509, respectively. The predictability of HepG-2 and MCF-7models were verified by a test set of 6 compounds, which gave a determination correlation coefficient (R^2_{test}) values of 0.73 and 0.52, respectively. The predicted values of the molecular activity of the learning and validation sets of the two models are presented in Table 5.

3.4. Y-randomization

To test the robustness of the obtained models, we then carried out the randomization test [38], which allows us to affirm that the good correlations between the descriptors and the activity are not due to chance. To do this, the observations are randomly disorganized ten times, i. e. the column of the $p\mathrm{IC}_{50}$ response will be changed randomly, but the columns of the descriptors remain unchanged. We, therefore obtain ten models with an average R^2 and R^2cv of 0.24 and 0.15 for HepG-2 as well as 0.2 and 0.102 for MCF-7 respectively. This result indicates that the models obtained are not due to a chance.

3.5. The applicability domain

A QSAR model is not universal, it is developed for a given number of compounds that do not cover all the chemical space and only predictions concerning molecules in this domain can be considered admissible.

The determination of ADs is therefore of great importance as it is explicitly requested in the validation processes put in place at the OECD level [39].

The applicability domain is the region of the chemical space including the compounds of the model learning set.

The analysis of the applicability domain in this work is performed using the "Leverage" method which is based on the variation of the standardized residuals of the dependent variable <i>p</i>IC_{50} with "Leverage" (the distance between the values of the descriptors and their means).

Figures 2 and 3 show that all observations have standardized residues between [-3; 3].

We note the absence of outliers in Figure 3 since the "Leverages" obtained are lower than the threshold value 0.67 ($h^* = 3p/n$). But Figure 4 shows that there is a compound (No. 11) of the training set with an h slightly higher than the critical value ($h > h^* = 0.7$), it is possible that the structure of this molecule influence on the prediction or the descriptors chosen does not give any special consideration to this coumarin derivative.

According to the results obtained by MLR, the QSAR model represented by Eq. (2) shows a major correlation of 2D descriptors with CDK inhibitory activity concerning the cell line (HepG-2). This model is highly predictive and gives very interesting results and structural information that can be guided by other research, on anti-cancer drugs.

If we analyze the experimentally obtained $p\mathrm{IC}_{50}$ (HepG-2) values, we note that in the group of coumarin derivatives (5a-5j), the presence of a nitro substituent as for the compound 5h is favorable to the activity, with a $p\mathrm{IC}_{50}$ value of 7.602, whereas a substitution by a methyl group leads to a reduction of the activity to 6.693. It can be deduced that the greater the effect of the electron donor group, the higher the antitumor activity. For the second group of compounds, including 7a-7h, shows that the presence of a fluorine atom in position 4 of the phenyl ring linked to the oxadiazole fraction generates a very active compound 7d of this range, and the fact of substituting the phenyl ring with a hydrogen atom in this position has decreased the activity to 6.585 as for compound 7a. This reveals that the presence of electronegative groups in this area could present a good anti-tumor activity.

On the other hand, compound 9a was found to be very active ($p\mathrm{IC}_{50} = 7.770$) and this could be attributed to the substitution of thiadiazole with a methyl coumarin group instead of benzocoumarin groups. When introducing the series of compounds 13a-13d, it is to see the effect of the structural modification of the basic skeleton on the antitumor activity, by considering them as starting compounds for future selective modifications of the coumarin molecule, given the very high value of activity ($p\mathrm{IC}_{50} = 7.886$) that provides the structure 13a.

Therefore, based on the skeleton of these molecules and depending on the nature of the chosen substitutions, new series of active coumarins (Table 6) can be designed from the selected MLR model that will reduce the time and cost of synthesis, having higher or similar activity values to the existing one.

3.6. Docking studies

This docking study consists of finding the best position for the ligand (coumarin derivative N.21 of the HepG-2 line) in the receptor-binding site (1KE9), which was obtained from the PDB databank with a resolution of 2Å [40]. Before its use by AutoDock, the complex is separated from its ligand to release the active site after eliminating the water molecules, the prepared files are saved in pdbqt format.

To increase the speed of energy evaluation of the system AutoDock uses a three-dimensional grid broadly encompassing the active site of the 1KE9 protein and allowing free rotation of the ligand in this site. In our case, the center of this box is determined by the coordinates X = -9.477, Y = 50.349, and Z = 11.383 with dimensions 30*40*40 A3. The box is then

Table 3. Correlation matrix of the descriptors of Eq. (2).

	Correlation coefficient			TF
	MD	DH	$p\mathrm{IC}_{50}$	
MD	**1**			0.421
DH	0.459	**1**		0.215
$p\mathrm{IC}_{50}$	-0.310	0.575	**1**	-

The values in bold represent the maximum correlation between the descriptors.

Table 4. Correlation matrix of the descriptors of Eq. (3).

	Correlation coefficient			TF
	MD	DH	pIC_{50}	
MD	**1**			0.427
DH	0.426	**1**		0.234
pIC_{50}	-0.171	0.577	**1**	-

The values in bold represent the maximum correlation between the descriptors.

Table 5. Observed and predicted pIC_{50} values of the training and validation sets of coumarin derivatives of the two MLR models.

Comp.	MLR (HepG-2)			Comp.	MLR (MCF-7)		
	pIC_{50}obs	PredpIC_{50}	Resid		pIC_{50}obs	PredpIC_{50}	Resid
1	7,114	6.911	0.203	1	6.842	6.886	-0.044
2	6,879	6.800	0.079	2	7.301	7.446	-0.145
3	7,244	7.492	-0.248	3	6.754	7.040	-0.285
4	6,848	6.849	-0.001	4	6.585	6.921	-0.336
5	7,046	6.750	0.296	5	6.896	7.007	-0.111
7	6,693	6.991	-0.298	6	6.854	6.957	-0.103
8	7,602	7.389	0.214	7	6.721	6.839	-0.117
9	7,046	7.050	-0.004	8	6.807	6.925	-0.118
11	6,585	6.844	-0.259	9	7.000	6.887	0.113
12	6,796	6.755	0.041	10	7.222	6.845	0.377
13	6,770	6.802	-0.032	11	7.770	7.770	0.000
15	7,097	7.037	0.060	12	7.046	7.077	-0.032
17	6,708	6.788	-0.081	13	7.301	7.166	0.135
18	7,347	7.147	0.200	14	6.921	6.998	-0.077
20	6,824	6.888	-0.064	15	7.398	7.047	0.351
21	7,886	7.803	0.083	16	7.523	7.114	0.409
22	6,959	7.125	-0.166	17	7.097	7.088	0.009
23	6,721	6.742	-0.021	18	6.951	6.975	-0.024
6*	7.398	7.082	0.316	19*	6.721	6.876	-0.155
10*	6.807	6.939	-0.132	20*	7.699	7.362	0.337
14*	7.398	7.000	0.398	21*	7.222	6.849	0.373
16*	6.745	6.951	-0.206	22*	7.824	7.057	0.767
19*	7.770	7.011	0.758	23*	6.987	7.184	-0.197
24*	7.523	7.059	0.464	24*	7.699	7.131	0.568

[*] Indicate the test set compounds.

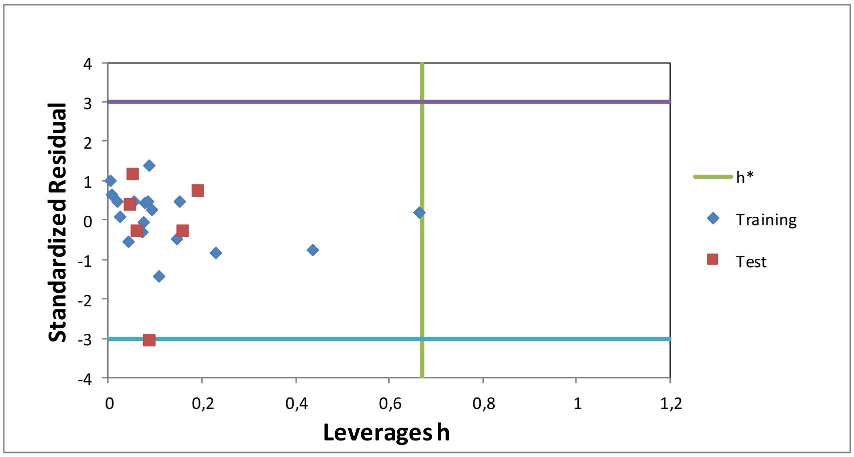

Figure 3. The Williams graph of the model presented by Eq. (02).

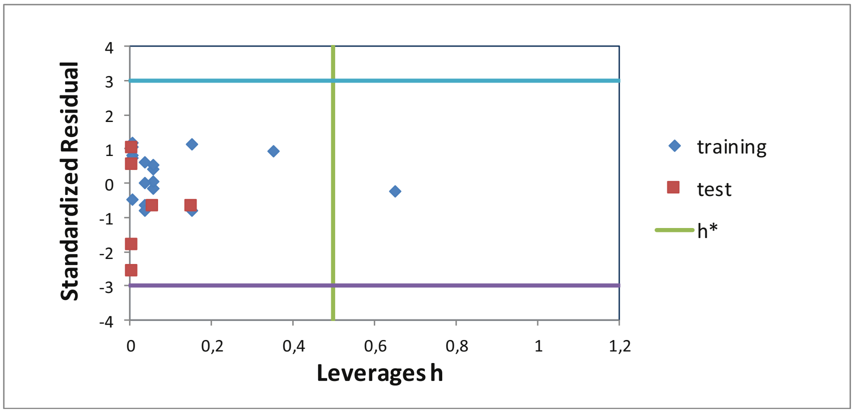

Figure 4. The Williams graph of the model presented by Eq. (03).

Table 6. Chemical structures, molecular descriptors and $p\mathrm{IC}_{50}$ activity with leverage effects (h) of new derivatives.

Novel compounds		MD	DH	$p\mathrm{IC}_{50}$	Leverage
derivatives of the skeleton (5h)	M1 R = H; R_1 = OH; R_2 = H	7.598	1	7.024	0.252
	M2 R = CH_3; R_1 = H; R_2 = OH	3.665	1	7.795	0.187
	M3 R = H; R_1 = H; R_2 = H	4.268	0	7.078	0.083
	M4 R = Br; R_1 = NO_2; R_2 = CN	3.328	0	7.263	0.132
derivatives of the skeleton (13a)	M5 R = OCH_3; R_1 = OCH_3	3.106	1	7.904	0.252
	M6 R = OH; R_1 = OCH_3	9.513	0	6.05	0.476
	M7 R = H; R_1 = OCH_3	6.541	0	6.633	0.132
	M8 R = NO_2; R_1 = NO_2	1.822	0	7.558	0.214
derivatives of the skeleton (9a)	M9 R = $OCOCH_3$; R_1 = CN; R_2 = C6H5	6.515	1	7.236	0.187
	M10 R = COOH; R_1 = CN; R_2 = C6H5	3.504	0	7.228	0.083

centered on the active site and its dimensions are proportional to the size of the ligands studied.

After the generation of the protein and ligand files, the docking can be started, using the Genetic Algorithm (GA) with its default settings, and the results can then be viewed using the Discovery Studio 2016 Client software.

The best docking result is the conformation with the lowest energy -9.43 kcal/mole.

The visual analysis is an essential step to judge the performance of the program, the (Figure 5) shows that the ligand model simulated by AutoDock is correctly placed in the active site of 1KE9 and presents a spatial conformation very close to or even superposable to the one determined experimentally by crystallography that we find in the PDB.

The simulation performed by AutoDock allowed us to obtain a complex formed between the most active coumarin compound and the active site of 1KE9 which is stabilized by two hydrogen bonds between the carbonyl of the inhibitor and the amine function of the LYS A:33 residue with a distance equal to 2.778 Å, the second formed between the NH_2 of the ligand and the amine function of the LEU A:83 residue (d = 3.02 A).The compound is also stabilized by several hydrophobic interactions

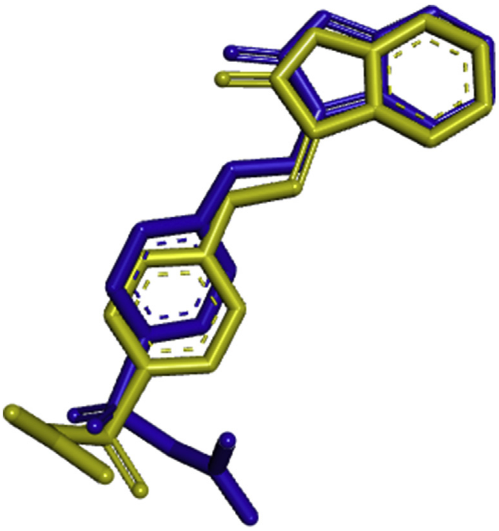

Figure 5. Comparison of the geometry of the ligand obtained by crystallography (colored in blue) with that obtained by AutoDock (colored in green).

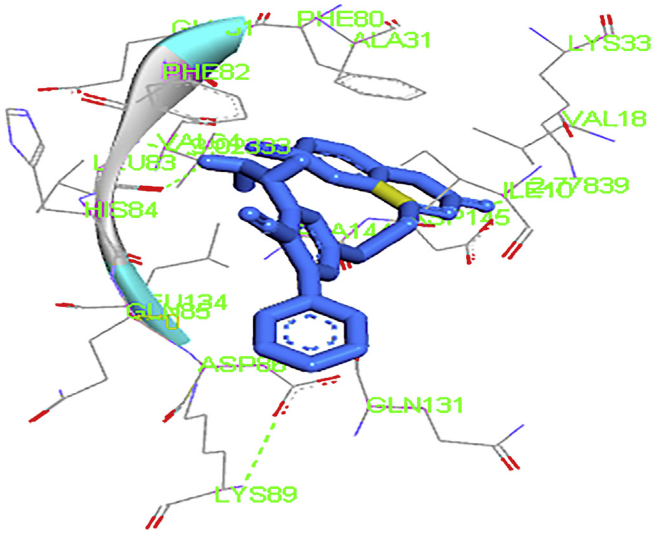

Figure 6. Interaction between the compound and the active site of 1KE9.

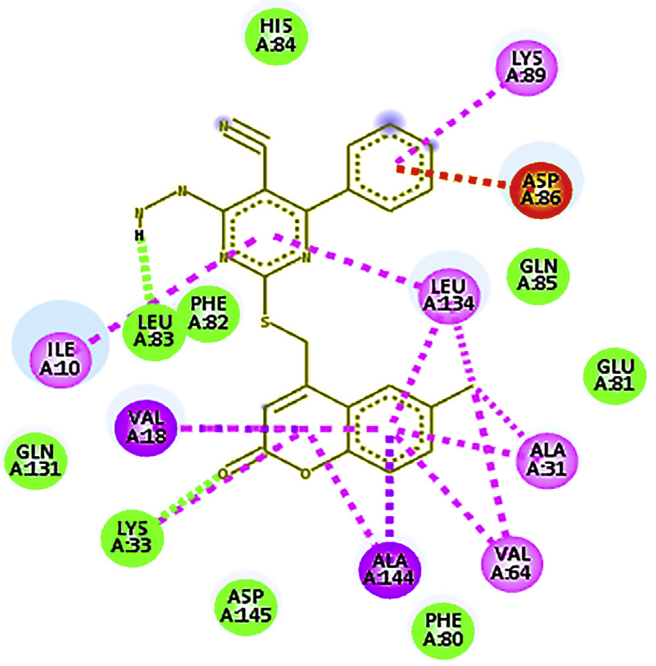

Figure 7. Schematic representation of the interaction model.

with the LYS A:89, LEU A:134, ALA A:31, VAL A:64, ALA A:144, ILE A:10, and VAL A:18 residues (see Figures 6 and 7).

These promising molecular docking results prompted us to predict the most favorable conformation and relative orientation of the M5 molecule proposed in Table 6 and having better activity than those of the original series into the active site of the 1KE9 protein with the same coordinates and dimensions as those used for the most active molecule.

We give in Figure 8 below the illustration of how compound M5 binds to the active site of the protein.

Analysis of the simulation result revealed that the compound M5 interacts with the residues of the active site, mainly through a hydrogen bond established between the sulfanyl of the ligand and the amino residue LEU A:83, as well as through hydrophobic interactions with the residues ALA A:144, VAL A:64, ALA A:31, VAL A:18, ILE A:10, LEU A:134, and PHE A:80.

4. Discussion

The QSAR model linking the inhibitory activity of CDKs concerning the cell line (HepG-2) and two 2D descriptors, provides the best statistical parameters, which led us to treat a molecular docking to predict the probable interactions between the ligand and the amino acids forming the skeleton of the receptor.

The number of H-bond donors (DH) is one of the two parameters which considerably influences activity according to the model equation, which is in agreement with the docking results which show that the most stable conformation of the coumarin derivative (N.21), as well as the predicted compound M5, form a hydrogen bond in common with the same residue LEU A:83, a second one appeared for the derivative (N.21) with LYS A:33 at the active site of 1KE9. We know in advance that the hydrogen bond connects molecules by involving a hydrogen atom, and for this bond to be established, it is necessary to be in the presence of a hydrogen bond donor and an acceptor. The rest of these 2 molecules interact with the active site by hydrophobic interactions.

The results obtained in this work are encouraging and could help in the design of new drugs against human hepatocellular carcinoma, it would, therefore, be interesting to validate and confirm these results experimentally by testing the inhibitory activity of the predicted coumarin derivatives against the CDK protein.

5. Conclusion

In this work, we studied the anticancer activity of CDK in a series of 24 coumarin derivatives, we used electronic descriptors in combination with Lipinski's parameters to relate them to biological activity. The QSAR model for the cell line (HepG-2) was able to show stability and predictive power, confirmed by internal and external validation.

The structural characteristics that can influence the inhibitory activity of compounds have been mentioned, namely the importance of electronegative groups at position 4 of the phenyl ring of the oxadiazole moiety, these functional groups that can form a hydrogen bond with water can affect the hydrophobic behavior and consequently an increase of activity. Which is proven by the strong correlation of the inhibitory activity of the compounds in this series with the dipole moment and the number of hydrogen bond donors. This allows this model to provide rational information for the design of potential new drugs and to complement this 2D-QSAR study with 3D-QSAR analysis as a perspective.

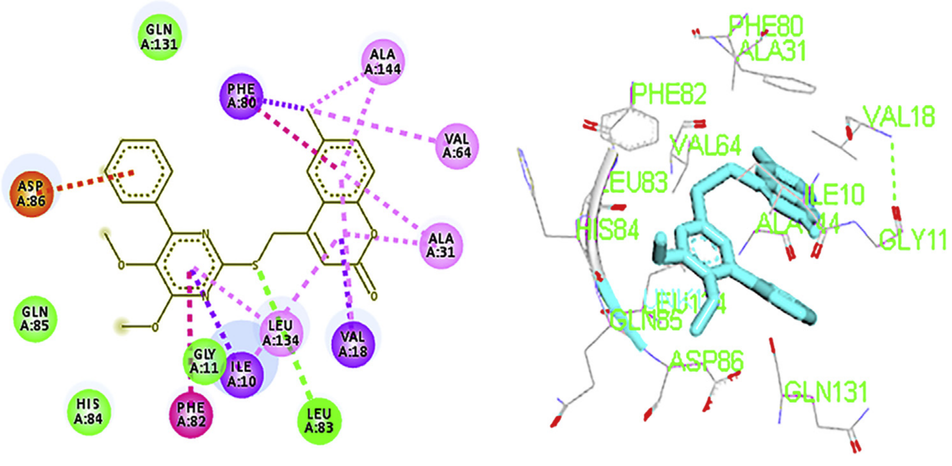

Figure 8. Mode of the interaction of the predicted compound M5 in the binding site of its receptor.

Declarations

Author contribution statement

Rania Kasmi: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

El ghalia Hadaji: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Oussama Chedadi, Abdellah El Aissouq: Performed the experiments.

Mohammed Bouachrine: Contributed reagents, materials, analysis tools or data.

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

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Supplementary material

Table S1. Analysis of variance of the model of Eq. (2).

Source	DOL	Sum of squares	Square average	F	Pr > F
Model	2	1.420	0.710	15.126	0.000
Error	15	0.704	0.047		
Corrected total	17	2.124			

Table S2. Analysis of variance of the model of Eq. (3).

Source	DOL	Sum of squares	Square average	F	Pr > F
Model	3	11.086	3.695	115.877	<0.0001
Error	14	0.446	0.032		
Corrected total	17	11.532			

Table S3. Table of coefficients of the first model.

Ind.Var	Coef	Std.Error	tobs	Pr>$\|t\|$
Intercept	0.792	0.232	3.409	0.004
MD	0.181	0.047	3.859	0.002
DH	-0.634	0.117	-5.423	<0.0001

Table S4. Table of coefficients of the second model.

Ind.Var	Coef	Std.Error	tobs	Pr>$\|t\|$
Intercept	6.287	1.338	4.699	0.000
EHOMO	1.042	0.201	5.191	0.000
MD	0.360	0.048	7.562	<0.0001
DH	-1.591	0.109	-14.544	<0.0001

References

[1] C. Sánchez-Martínez, M.J. Lallena, S.G. Sanfeliciano, A. de Dios, 'Cyclin dependent kinase (CDK) inhibitors as anticancer drugs: recent advances (2015–2019)', Bioorg. Med. Chem. Lett 29 (20) (Oct. 2019) 126637.

[2] S.A. Morsy, A.A. Farahat, M.N.A. Nasr, A.S. Tantawy, Synthesis, molecular modeling and anticancer activity of new coumarin containing compounds, Saudi Pharmaceut. J. 25 (6) (Sep. 2017) 873–883.

[3] A. Garrard, Coumarins, in: Encyclopedia of Toxicology, Elsevier, 2014, pp. 1052–1054.

[4] T. Żołek, D. Maciejewska, Theoretical evaluation of ADMET properties for coumarin derivatives as compounds with therapeutic potential, Eur. J. Pharmaceut. Sci. 109 (Nov. 2017) 486–502.

[5] Y. Miyake, A. Murakami, Y. Sugiyama, M. Isobe, K. Koshimizu, H. Ohigashi, Identification of coumarins from lemon fruit (<i>Citrus limon</i>) as inhibitors of in vitro tumor promotion and superoxide and nitric oxide generation, J. Agric. Food Chem. 47 (8) (Aug. 1999) 3151–3157.

[6] L. Wu, X. Wang, W. Xu, F. Farzaneh, R. Xu, The structure and pharmacological functions of coumarins and their derivatives, Comput. Mater. Continua (CMC) 16 (32) (Nov. 2009) 4236–4260.

[7] K.V. Sashidhara, A. Kumar, M. Kumar, A. Srivastava, A. Puri, Synthesis and antihyperlipidemic activity of novel coumarin bisindole derivatives, Bioorg. Med. Chem. Lett 20 (22) (Nov. 2010) 6504–6507.

[8] M. Curini, F. Epifano, F. Maltese, M.C. Marcotullio, S.P. Gonzales, J.C. Rodriguez, 'Synthesis of collinins, an antiviral coumarin, Aust. J. Chem. 56 (1) (2003) 59.

[9] L.Z. Chen, et al., New arylpyrazoline-coumarins: synthesis and anti-inflammatory activity, Eur. J. Med. Chem. 138 (Sep. 2017) 170–181.

[10] A. El-Agrody, M. Abd El-Latif, N. El-Hady, A. Fakery, A. Bedair, Heteroaromatization with 4-hydroxycoumarin Part II: synthesis of some new pyrano[2,3-d]pyrimidines, [1,2,4]triazolo[1,5-c]pyrimidines and pyrimido[1,6-b]-[1,2,4]triazine derivatives, Molecules 6 (6) (May 2001) 519–527.

[11] J. Yu, L. Wang, R.L. Walzem, E.G. Miller, L.M. Pike, B.S. Patil, Antioxidant activity of citrus limonoids, flavonoids, and coumarins, J. Agric. Food Chem. 53 (6) (Mar. 2005) 2009–2014.

[12] M. Campos-Toimil, F. Orallo, L. Santana, E. Uriarte, Synthesis and vasorelaxant activity of new coumarin and furocoumarin derivatives, Bioorg. Med. Chem. Lett 12 (5) (Mar. 2002) 783–786.

[13] S.M. Free, J.W. Wilson, A mathematical contribution to structure-activity studies, J. Med. Chem. 7 (4) (Jul. 1964) 395–399.

[14] C. Acharya, A. Coop, J.E. Polli, A.D. MacKerell, Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach, CAD 7 (1) (Mar. 2011) 10–22.

[15] S. Yousefinejad, B. Hemmateenejad, Chemometrics tools in QSAR/QSPR studies: a historical perspective, Chemometr. Intell. Lab. Syst. 149 (Dec. 2015) 177–204.

[16] Y. Li, et al., Pharmacophore modeling, molecular docking and molecular dynamics simulations toward identifying lead compounds for Chk1, Comput. Biol. Chem. 76 (Oct. 2018) 53–60.

[17] M. Gupta, R. Sharma, A. Kumar, Docking techniques in pharmacology: how much promising? Comp. Biol. Chem. 76 (Oct. 2018) 210–217.

[18] Danishuddin, A.U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, Drug Discov. Today 21 (8) (Aug. 2016) 1291–1302.

[19] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, 'Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1', Adv. Drug Deliv. Rev. 46 (1–3) (Mar. 2001) 3–26.

[20] The Addinsoft XLSTAT Company | Statistical Software for Excel. https://www.xlstat.com/en/company. (Accessed 26 July 2019).

[21] E. Vittinghoff, C.E. McCulloch, D.V. Glidden, S.C. Shiboski, 5 linear and non-linear regression methods in epidemiology and biostatistics, in: Handbook of Statistics, 27, Elsevier, 2007, pp. 148–186.

[22] Danishuddin, A.U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, Drug Discov. Today 21 (8) (Aug. 2016) 1291–1302.

[23] M. Ghamali, S. Chtita, A. Ousaa, B. Elidrissi, M. Bouachrine, T. Lakhlifi, QSAR analysis of the toxicity of phenols and thiophenols using MLR and ANN, J. Taibah Univ. ScI. 11 (1) (Jan. 2017) 1–10.

[24] A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, and K.-H. Lee, 'Rational Selection of Training and Test Sets for the Development of Validated QSAR Models', p. 13.

[25] K. Roy, On some aspects of validation of predictive quantitative structure-activity relationship models, Expet Opin. Drug Discov. 2 (12) (Dec. 2007) 1567–1577.

[26] J. Shao, Linear model selection by cross-validation, J. Am. Stat. Assoc. 88 (422) (Jun. 1993) 486–494.

[27] T.M. Martin, et al., Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52 (10) (Oct. 2012) 2570–2578.

[28] E. Burello, Review of (Q)SAR models for regulatory assessment of nanomaterials risks, NanoImpact 8 (Oct. 2017) 48–58.

[29] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemomet. Intell, Lab. Syst 145 (Jul 2015) pp. 22–29.

[30] Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Elsevier, 2015.

[31] R.G. Brereton, Introduction to analysis of variance: ANOVA, J. Chemometr. 33 (1) (Jan. 2019), e3018.

[32] N. Khalifa, A. Srour, S. Abd El-Karim, D. Saleh, M. Al-Omar, Synthesis and 2D-QSAR study of active benzofuran-based vasodilators, Molecules 22 (11) (Oct. 2017) 1820.

[33] M. De Bourmont, 'La résolution d'un problème de multicolinéarité au sein des études portant sur les déterminants d'une publication volontaire d'informations : proposition d'un algorithme de décision simplifié basé sur les indicateurs de Belsley, Kuh et Welsch (1980)', in: Comptabilités et innovation, Grenoble, France, May 2012 p. cd-rom, Accessed: Jul. 28, 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00691156.

[34] J.T. Long, S. Neogi, C.M. Caldwell, M.P. DeLange, Variation inflation factor-based regression modeling of anthropometric measures and temporal-spatial performance: modeling approach and implications for clinical utility, Clin. BioMech. 51 (Jan. 2018) 51–57.

[35] A. El Aissouq, H. Toufik, M. Stitou, A. Ouammou, F. Lamchouri, In silico design of novel tetra-substituted pyridinylimidazoles derivatives as c-jun N-terminal kinase-3 inhibitors, using 2D/3D-QSAR studies, molecular docking and ADMET prediction, Int. J. Pept. Res. Therapeut. (Oct. 2019).

[36] A. El Aissouq, H. Toufik, F. Lamchouri, M. Stitou, and A. Ouammou, 'QSAR study of isonicotinamides derivatives as Alzheimer's disease inhibitors using PLS-R and ANN methods', in 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, Dec. 2019, pp. 1–7.

[37] R. Kiralj, M.M.C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, J. Braz. Chem. Soc. 20 (4) (2009) 770–787.

[38] C. Rücker, G. Rücker, M. Meringer, y-Randomization and its Variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (6) (Nov. 2007) 2345–2357.

[39] I.L. Ruiz, M.Á. Gómez-Nieto, Study of the applicability domain of the QSAR classification models by means of the rivality and modelability indexes, Molecules 23 (11) (Oct. 2018) 2756.

[40] H.N. Bramson, et al., Oxindole-based inhibitors of cyclin-dependent kinase 2 (CDK2): design, synthesis, enzymatic activities, and X-ray crystallographic analysis, J. Med. Chem. 44 (25) (Dec. 2001) 4339–4358.