Clinical Complementary Medicine and Pharmacology 2 (2022) 100031

Contents lists available at ScienceDirect

Clinical Complementary Medicine and Pharmacology

journal homepage: www.elsevier.com/locate/ccmp

QSAR Applied to 4-Chloro-3-formylcoumarin Derivatives Targeting Human Thymidine Phosphorylase

Thomas Scior[a,*], Juan Carlos Garcia-Hernandez[a], Hassan H. Abdallah[b], Christian Alexander[c]

[a] Departamento de Farmacia, Facultad de Ciencias Químicas, Benemérita Universidad Autónoma de Puebla, Puebla 72000, Pue., Mexico
[b] Chemistry Department, College of Education, Salahaddin University Erbil, Erbil 44001, Iraq
[c] Division of Cellular Microbiology, Research Center Borstel- Leibniz Lung Center, Borstel 23845, Germany

ARTICLE INFO

Keywords:
Coumarin
Phytopharmacy
R script
Linear model
Multiple linear regression
QSAR

ABSTRACT

Background: Coumarins are secondary metabolites from the phenylpropanoid-type biosynthesis in higher plants. A plethora of potential phytopharmacological activities have been described for derivatives of the coumarin scaffold: hepatoprotective, antineoplastic, antimicrobial, antituberculosis, antiviral, anti-inflammatory anticoagulant, or antithrombotic effects.

Objective: A computer-based quantitative structure – activity relationships (QSAR) study for a series of 4-chloro-3-formylcoumarins was carried out.

Methods: To this end we generated the 3D models of 17 published coumarin structures, calculated their physicochemical properties (descriptors) to correlate them to their experimentally known biological activities measured as inhibition concentrations to block the target enzyme activity. Our proposed approach used free molecular modeling software and applies our scripts written in the programming language R.

Results: The final multiple regression models achieved satisfactory results with a small number of descriptors – all of which were statistically significant and meaningful in the field of pharmacodynamics to develop new 3-formylcoumarins with enhanced activities targeting the human thymidine phosphorylase enzyme.

Conclusion: On theoretical grounds, our *in silico* research contributes in a crucial step in the field of complementary phyto-medicine. This step is located between *in vivo* pharmacological observations of plant extracts on ethnopharmacological, preclinical or controlled clinical levels and the need to identify – at an atomic scale – all those plant ingredients responsible for the biological actions under scrutiny. Our simulations shed light on the modification of phyto-medicine's physicochemical properties to enhance the interaction with their biomolecular target in the patient's body.

1. Introduction

Coumarins are the result of the phenylpropanoid – type biosynthesis of secondary metabolites in higher (developed) plantes. Common dicotyledoneous families that synthesize coumarins are Apiaceae and Rutaceae. Natural coumarins have been studied for a wide range of pharmacological activities: hepatoprotective, antineoplastic (cancer), antimicrobial, antituberculosis, antiviral or anti-inflammatory activities. In recent years, seminal reviews described the coumarin scaffold and laid the groundwork for structure – activity relationships for their hitherto known pharmacological effects (Revankar et al., 2017; Zhu et al., 2018; Singh et al., 2019; Zhang et al., 2019; Annunziata et al., 2020; Prusty et al., 2020; Tafesse et al., 2020; Al-Warhi et al., 2020). Natural coumarins are found especially in seeds, roots, or leaves, and especially in tonka beans from legume trees (Garrard, 2014). Such parts of plants rich in coumarin derivatives have been administered to patients as dried parts of plants to conserve them. Of note, the etymological root of the word "drug" is the old anglo-saxon word "trok", in modern English: "dried". In addition to the interest of complementary medicine in coumarins, they have potential applications in modern pharmacotherapy, too. Commercial importance of Warfarine and Phenprocoumon have achieved as anticoagulant and antithrombotic agents, which are derivatives of antioxidant vitamin K (Quiroga et al., 2017; Quiroga et al., 2018; Scior et al., 2018).

Thymidine phosphorylase (**TP**) is a nucleoside metabolism enzyme which catalyzes the reversible conversion of thymidine to thymine and 2-deoxy-α-D-ribose-1-phosphate. This enzyme is expressed in the nucleus and in the cytoplasm (Fig. 1). TP is the target protein. It has been shown to be overexpressed in several types of cancer in response to stressful cellular conditions. The associated biochemical processes promote tumor angiogenesis.

* Corresponding author.
E-mail address: thomas.scior@correo.buap.mx (T. Scior).

https://doi.org/10.1016/j.ccmp.2022.100031
Received 14 November 2021; Received in revised form 21 March 2022; Accepted 24 March 2022

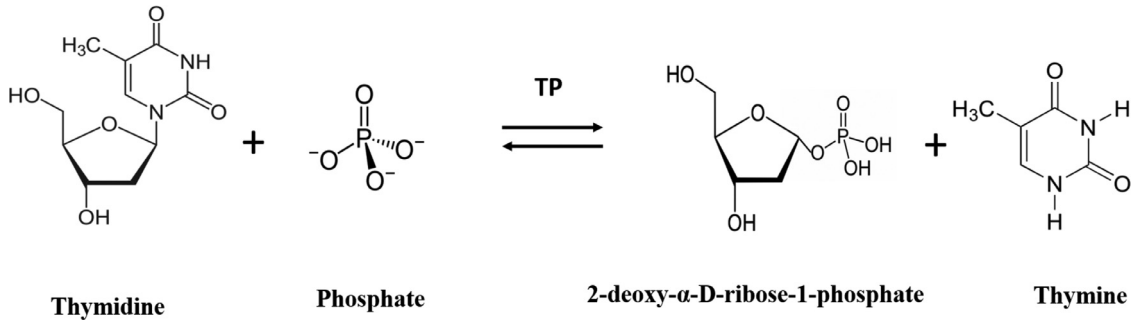


Fig. 1. Reaction catalyzed by thymidine phosphorylase. Thymidine (left most) is recognized by the enzyme and a new ester bond is format between the sugar moiety and the phosphate group, while Thymine is liberated (right most).

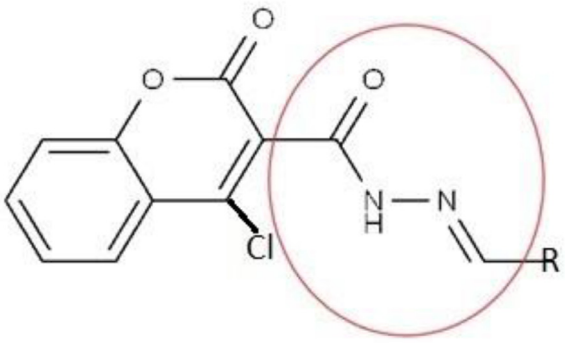


Fig. 2. The general structure (scaffold) of the R-substituted coumarin molecules under scrutiny.

Certain coumarins possess cytostatic (growth inhibitory) properties whereas others exhibit cytotoxic activities (Marshall et al., 1994). It is important to highlight the reports on coumarins with the hydrazide fraction in **position C-3**. They have shown that the combined hydrazide / hydrazone group (-CO–NH–N CH-) has an important role within the molecules that act as antitumor agents (Emami and Dadashpour, 2015). Through various studies, coumarin derivatives substituted at their C-3 carbon position with a hydrazide moiety have been found to have antitumor effects (Huang et al., 2011; Thakur et al., 2015; Luo et al., 2017; Thakur et al., 2021). Here 17 structures were taken from one literature source. They are known thymidine phosphorylase inhibitors. Their measured IC_{50} values were taken from the same publication (Taha et al., 2018). In Fig. 2, the red circle shows the 3-formyl-hydrazone fraction at this position C-3 of the 4-chloro substituted coumarin scaffold. This scaffold constitutes a bicyclic aromatic 1,2-benzopyrone. It comprises a lactone group, i.e. intracellular ester group. The chemical name of the R-substituted scaffold is 4-chloro-2-oxo-2H-chromene-3-carbohydrazone. Its potential cleavage site is contoured by a colored circling line. It also marks the spot where the catalytic reaction takes place when the structures are bond to the active site of the human thymidine phosphorylase. Further structural and mechanistic details can be found in Fig. 1 to Fig. 4 and Table 1 by Taha et al. (Taha et al., 2018). General notes about stability and decay of hydrazone drugs have been published earlier (Scior and Garces-Eisele, 2006). In particular, the presence of pyrrole or thiophene substitutions in position “R” in Fig. 2 is believed to enhance the inhibitory effect. Of note, in isocoumarin, which are also natural products from higher plants, the lactone group has an inverse orientation.

This study seeks to complement the extant literature in the field of complementary medicine and is embedded in our ongoing herbal research. It will connect the acquired knowledge from ethnopharmacology and phytotherapeutical treatment of patients with pharmacological action mechanisms at an atomic scale (Bernard et al., 2002; Blondeau et al., 2010; Do et al., 2015). Modern drug development research is in need of correlating clinical observations with basic research on molecular level – and herbal research work in complementary medicine is no exception.

2. Materials and Methods

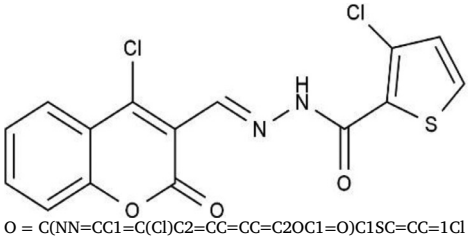
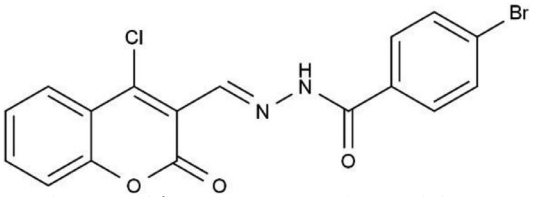
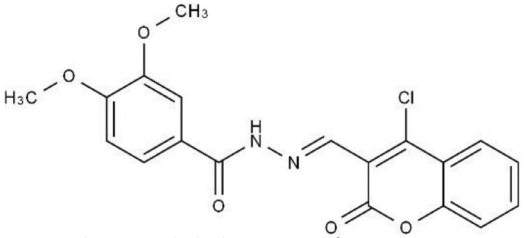
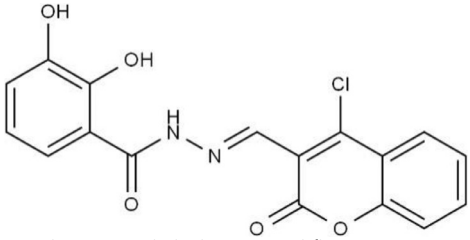
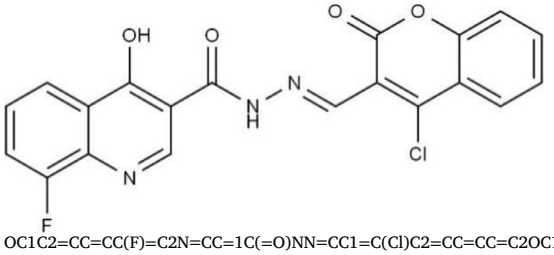
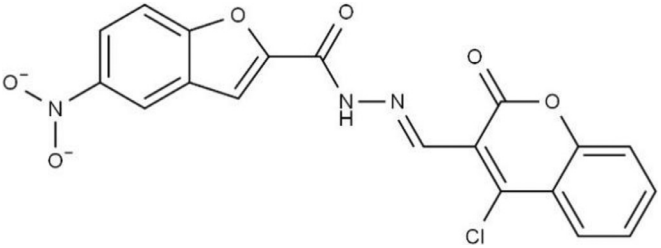
2.1. Drug development assisted by molecular simulations

In the early discovery stages of new drugs from natural sources, a drug profiling procedure may be carried out by which new molecules with therapeutic potential are identified, making possible the combined use of computational, experimental and clinical models (Do et al., 2015). With the advent of computer-aided molecular simulations, it has been possible to recreate several complex natural processes. With the help of this technology a better understanding of physiological processes can be achieved in which a response is sought in order to obtain better results in the development of new treatments. Therefore, the application of the quantitative computational approach proposes numerical equations to change chemical structures. Such structural modifications help strengthen their drug affinity to the biomolecular targets and at the same time enhance their therapeutic effect (Medina-Franco et al., 2015).

2.2. Quantitative structure-activity relationships

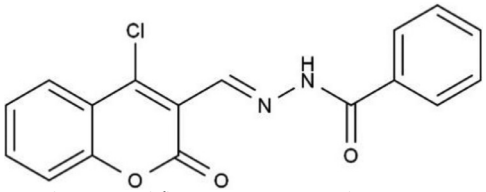
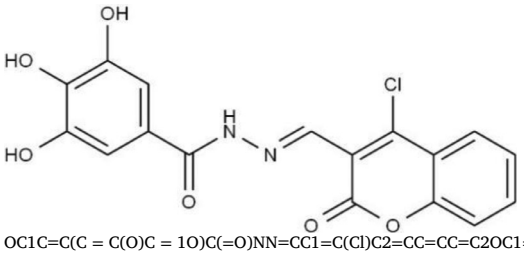
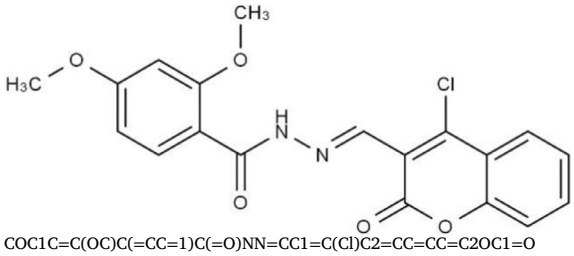
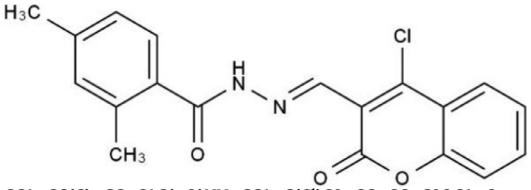
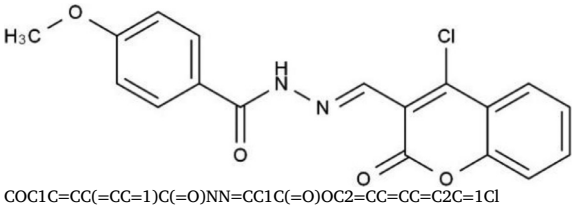
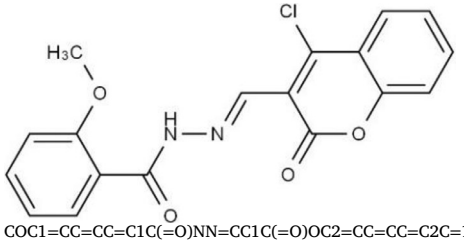
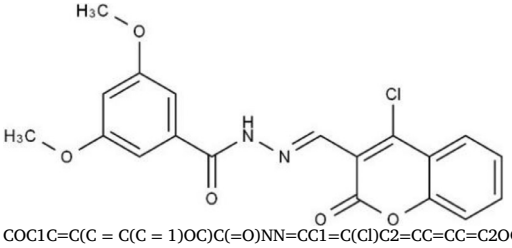
Quantitative structure-activity relationships studies (**QSAR**) generally aim at establishing predictive statistical models of the experimentally observed biological activities. Albeit, they do not only serve as predictive tools to develop new and better drug candidates, but also serve to analyze existing series of analogous drugs (Scior et al., 2009). The results of the present QSAR study contribute for future research to develop new drugs with a 3-formylcoumarin scaffold (Lozano-Aponte and Scior, 2012). In this analysis, an attempt is made to express the biological activity as a linear combination of different descriptors, thus postulating the form of a linear relationship between the activity and the relevant molecular properties. The coefficients in the equations provide contributions to predict the activity. To start any QSAR study a set (or series) of analogous molecules is needed, AKA the structural input data. They are structurally related and share a common pharmacodynamic property, i.e. the same mechanism of action which also implies that they (must) have not only the same molecular target but also a common binidng site in the target structure. On the other hand, biological response data are needed for each molecule, such as $\mathbf{EC_{50}}$ (effective concentration in 50% of the sampled observations), $\mathbf{IC_{50}}$ (inhibition concentration of 50% of the population), $\mathbf{LD_{50}}$ (the dose required for kill half of the total population). Finally, the physicochemical properties also called molecular descriptors are needed. They are calculated using the appropriate software and using the models of the previously designed chemical structures (Lozano-Aponte and Scior, 2012; Roy et al., 2015). Multiple linear regression (**MLR**) is an extension of simple linear regres-

Table 1
All 17 coumarin-derived molecules with their respective molecular structures in SMILES format.

ID	2D drawings with SMILES annotations
1	 O = C(NN=CC1=C(Cl)C2=CC=CC=C2OC1=O)C1SC=CC=1Cl
2	 O = C(NN=CC1=C(Cl)C2=CC=CC=C2OC1=O)C1C=CC(Br)=CC=1
3	 COC1=CC(=CC=C1OC)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
4	 OC1C(=CC=CC=1O)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
5	 OC1C2=CC=CC(F)=C2N=CC=1C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
6	 [O]N([O])C1C=C2C=C(OC2=CC=1)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O

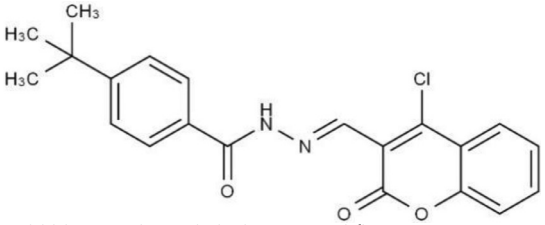
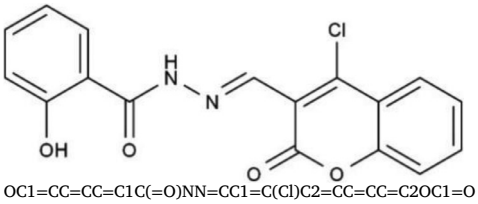
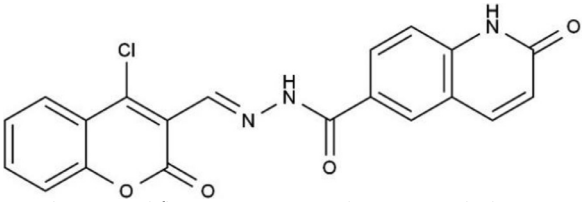
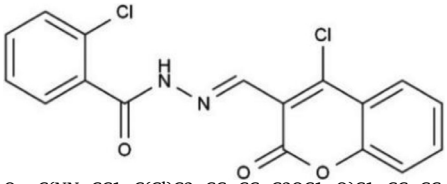
(*continued on next page*)

Table 1 (*continued*)

ID	2D drawings with SMILES annotations
7	 O = C(NN=CC1=C(Cl)C2=CC=CC=C2OC1=O)C1C=CC=CC=1
8	 OC1C=C(C = C(O)C = 1O)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
9	 COC1C=C(OC)C(=CC=1)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
10	 CC1=CC(C)=CC=C1C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
11	 COC1C=CC(=CC=1)C(=O)NN=CC1C(=O)OC2=CC=CC=C2C=1Cl
12	 COC1=CC=CC=C1C(=O)NN=CC1C(=O)OC2=CC=CC=C2C=1Cl
13	 COC1C=C(C = C(C = 1)OC)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O

(*continued on next page*)

Table 1 (*continued*)

ID	2D drawings with SMILES annotations
14	 CC(C)(C)C1C=CC(=CC=1)C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
15	 OC1=CC=CC=C1C(=O)NN=CC1=C(Cl)C2=CC=CC=C2OC1=O
16	 O = C(NN=CC1=C(Cl)C2=CC=CC=C2OC1=O)C1=CC2C=CC(=O)NC=2C=C1
17	 O = C(NN=CC1=C(Cl)C2=CC=CC=C2OC1=O)C1=CC=CC=C1Cl

Abbreviation: ID = identification number for each molecule.

sion (**SLR**) that occupies more than one independent variable (**IDV**) and only one dependent variable (**DV**). In rare cases there is more than one DV denominated at times as multivariate model (Scior et al., 2009). MLR is favored for its simplicity and ease of interpretation since the model assumes a linear relationship between the property of the compound denoted by the letter Y and its vector X of characteristics which are generally calculated as numerical descriptors of ligand structures. Y and X values are traditionally situated on the Y and X axis, respectively. Therefore, with the fitted model, the property of an unknown compound can be predicted (Dehmer et al., 2012). In general terms, the equations of any MLR model are written as a mathematical equation:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_1$$

where:

Y: the DV to predict; X: the IDVs to describe each molecule numerically; α and β: unknown parameters to be estimated by statistical means.

Simplicity of our resulting equations and the use of chemically meaningful descriptors indirectly constitute compelling evidence to justify the postulation of a(n approximately) linear relationship between the activity and the relevant molecular properties.

2.3. Statistical tool and scripts

"R" is a programming language and computing environment for statistics and graphics. Our script used basic R instructions which are integrated in all (newer) R versions (*version 2.x.y* or later). It was downloaded from its original source (*The R Project for Statistical Computing, 2021*). It can be installed on MS Windows, Unix or Linux computers or those running Mac OS. The MLR models and other statistical calculations were carried out by means of our scripts (see Supplementary Materials). To obtain the linear equations of the QSAR models we used our scripts written in R language (*The R Project for Statistical Computing, 2021*). Thanks to its basic graphic packages for linear models our quantitative results were also represented through different types of graphs.

2.4. Molecular modeling

From the Protein Data Bank (Berman et al., 2000), the target protein was retrieved. It constitutes a three-dimensional crystal structure of the protein thymidine phosphorylase from *Escherichia coli*. Said structure is co-crystallized with an appropriate inhibitory ligand. The latter would serve for the modeling and optimization of the molecules to be developed.

With the help of the Vega ZZ program, all 3D structures of the coumarin-derived molecules were generated. Their geometries were optimized (potential energy relaxation) under the Tripos force field (Clark et al., 1989).

Through the use of various specialized computer programs such as *Vega ZZ* (Pedretti et al., 2021), *Swiss ADME* (Daina et al., 2017), *E-Dragon* (Tetko et al., 2005) and *Alva Desc 1.0* (*AlvaDesc: Molecular Descriptors, 2021*), the physicochemical descriptors were obtained, which will serve to construct the multiple regression linear equation. These descriptors will be selected using theoretical and statistical methods to appropriately represent the atoms or chemical groups in the studied ligands.

Once the molecular descriptors had been chosen, our QSAR modeling was carried out, using statistical analysis. The downsides and pitfalls thereof have already been discussed and Published by

Fig. 3. Structural superposition of all 17 compounds with a common 3-formylcoumarin scaffold (on the left side). The highest chemical variation is visible at the right-hand side.

Table 2
Programs and numbers of descriptors obtained from each of them.

Software	Descriptors
Vega ZZ	11
Swiss ADME	23
E-Dragon	21(1666)
AlvaDesc 1.0	21 (5305)

Scior et al. (Scior et al., 2009). In particular, our study avoided the following problems implicated in QSAR modeling: "Pitfall: Linearity Assumption", "Overfitting Test", or "Pitfall: Over- and Under-Determined Equations" (cf. section results with discussion about model 8).

3. Results and Discussion

3.1. Modeling the analog thymidine phosphorylase protein inhibitor molecules

We retrieved several PDB entries for inspection to gain deeper insight and selected the structure with PDB code 4EAD (Timofeev et al., 2013; Timofeev et al., 2014). It contains the X-ray structure of enzyme thymidine phosphorylase, having an inhibitory ligand at its active site. A total of 25 PDB entries (last visit in March 2022) were found less suitable: only three showed a target protein in complex with inhibitory ligands at the active site which resemble less to our series of compounds than 4EAD. Table S1 lists the PDB entries under scrutiny (cf. Supplementary Materials) (Balaev et al., 2016; Norman et al., 2004; Pugmire et al., 1998; Pugmire and Ealick, 1998; Timofeev et al., 2013; Timofeev et al., 2014). The ligand found in PDB code 4EAD constitutes 3′-azido-2′-fluorodeoxyuridine. It was used as a 3D template structure to construct the 17 molecules derived from the 3-formyl coumarin scaffold. Its observed position at the active site of the crystal complex helped guide ligand replacement (cf. Section 3.2).

3.2. Construction of the 17 molecular models

The molecules were built in the *Vega ZZ 3.2.1.0* software version (Pedretti et al., 2021) and saved in SMILE format (Weininger, 1988). The complexed ligand of the target enzyme was replaced by the 17 coumarin molecules (Table S1). In this way, the sensible superposition of the molecules could be obtained (Fig. 3).

3.3. Calculation of molecular descriptors through specialized modeling tools

Table 1 lists the 17 coumarin derivatives with their respective structures in SMILES format. These formats were used to calculate the descriptors in the programs for this purpose.

In the program *Vega ZZ*, 11 descriptors were obtained, while in the *Swiss ADME* program, 23 were obtained. In addition, using the free *E-Dragon* program, 21 were considered, whereas in the *AlvaDesc 1.0* program 21 descriptors were generated. The values taken from *E-Dragon* and *AlvaDesc 1.0* were identical as it turned out that both use the same calculation methods (Table 2).

For the next step in QSAR model generation a Table 3 was prepared to list the selected descriptors to build the preliminary QSAR models.

During the stage of descriptor selection, both statistical tools and various theoretical criteria are used, which is why all the available data were taken into account to build the preliminary QSAR models. Table 4 documents the independent variables taken for each model. The more complex equations were those with a maximum of 4 descriptors and the smallest those with only 2 IDVs.

3.4. Analysis of preliminary QSAR models

The statistical data as well as the graphics were obtained with the R software. Statistical methods explain quantitatively how our independent variables influence the values of the experimental response or dependent variable (IC_{50}). Table 5 reports the values for all coefficients of determination (R^2) for all nine QSAR models. R^2 is a statistical measure reflecting the variation in the response variable explained by the underlying descriptors.

Model number one (model 1, for short) possesses the highest coefficient of determination ("R squared"). It is based on 4 descriptors which are shown in Table 4, however when performing the statistical analyzes of each one, model 8 was considered more appropriate, thanks to a twofold reason: (i) its good prediction of biological activity; (ii) as well as its contribution of fully comprehensive, informative physicochemical descriptors to the inhibitory activity of all 17 molecules under scrutiny. Upon inspection of all nine models, model 8 was seleted as the final model. Hence, the following statistical analysis focused on details for model 8. Its two descriptors were listed in Table 6.

BIC4 describes the nature of bonding for each molecule in a topological way. Bonding is closely related to the ability to form bonds between atoms by exchange of electrons. And this ability is commonly well represented by electronegativity of bonded atoms (Todeschini et al., 2008). **TPSA (Tot)** has been chosen as it shows the correlation with passive molecular transport across membranes. This descriptor allows predicting human intestinal absorption, permeability of cell monolayers, and penetration of the blood-brain barrier. It is obtained from a software that determines it by adding the tabulated surface contributions of the types of polar atoms. Here, polar fragments with nitrogen and oxygen plus "slightly polar" fragments also contain phosphorus and sulfur heteroatoms (Todeschini and Consonni, 2000). This contribution to the inhibitory activity of each molecule can be evaluated with statistical methods as mentioned in the previous section. In this way justify and explain how quantitatively our independent variables influence the values of the experimental response or dependent variable (IC_{50}) in a linear relation conforming the linearity assuption.

3.5. Statistical analysis of the descriptors contained in the final model

The inhibition concentrations of the 17 analogous molecules were evaluated by means of a histogram which displays the range (X axis) and frequency (Y axis) of all IC_{50} values. Fig. 4 shows that more IC_{50} values were found at a concentration range of 0 to 10 μM. Intriguingly, following the claims of the term "rational drug research and development" one would rather expect not to see a "bell-shaped" normal distribution. The latter indicates a higher influence of randomly distributed activities of tested drug candidates. Yet, the complicated and not fully understood R&D leads to results which more or less resemble a normal distribution. This reflects an uncontrolled behavior of hits by mere chance events. Here we learn that it has been more probable to find a lower activity than a true hit, (for further discussions cf. Fig. 3 in Scior et al., 2009).

A correlation matrix (Fig. 5) was generated to inspect the relationship between the inhibition constant IC_{50} which is our DV and the molecular descriptors that are our IDVs.

Table 3
List of selected descriptors.

Descriptor	Short description by key words
IC5	Information Content index (neighborhood symmetry of 5-order)
R4p	R autocorrelation of lag 4/weighted by polarizability
R3s	R autocorrelation of lag 3/weighted by I-state
CATS3D_05_AL	CATS3D Acceptor-Lipophilic BIN 05 (5.000–6.000 Å)
CATS3D_06_AL	CATS3D Acceptor-Lipophilic BIN 06 (6.000–7.000 Å)
CATS3D_08_DL	CATS3D Donor-Lipophilic BIN 08 (8.000–9.000 Å)
SPH	Spherosity
VE1_RG	coefficient sum of the last eigenvector (absolute values) from reciprocal squared geometrical matrix
Mor03i	signal 03/weighted by ionization potential
Mor20s	signal 20/weighted by I-state
F06[N–O]	Frequency of N - O at topological distance 6
TDB10p	3D Topological distance-based descriptors - lag 10 weighted by polarizability
B10[O–Cl]	Presence/absence of O - Cl at topological distance 10
BIC4	Bond Information Content index (neighborhood symmetry of 4-order)
TPSA(Tot)	topological polar surface area using N, O, S, P polar contributions
SpMax3_Bh(s)	largest eigenvalue n. 3 of Burden matrix weighted by I-state

Table 4
Listing of the preliminary equations by multiple linear regression. Estimated regression coefficients AKA beta coefficients.

M	S	MLR – equations
1	4	IC_{50} = 324.5128 - 67.6614x**IC5** + 129.7029x**R4p** - 7.6931x**R3s** + 2.3503x**CATS3D_05_AL**
2	4	IC_{50} = 303.071 - 59.1084x**IC5** + 112.3505x**R4p** - 9.5676x**R3s** + 1.8063x**CATS3D_06_AL**
3	4	IC_{50} = 2242.676 - 78.9514x**IC5** - 1859.98x**SPH** - 3.5349x**VE1_RG** + 5.0594x**CATS3D_05_AL**
4	3	IC_{50} = 344.3994 - 73.9674x**IC5** - 8.8394x**Mor03i** - 5.6325x**F06 [N–O]**
5	3	IC_{50} = 313.1018 - 45.8712x**IC5** - 7.5566x**TDB10p** - 9.1821x**CATS3D_08_DL**
6	3	IC_{50} = 353.5957 - 73.9994x**IC5** - 6.9854x**Mor03i** - 7.0713x**B10 [O–Cl]**
7	2	IC_{50} = 407.029 - 432.727x**BIC4** + 8.2808x**Mor20s**
8	2	IC_{50} = 345.53907 – 333.96x**BIC4** – 0.3131x**TPSA(Tot)**
9	2	IC_{50} = 305.1277 + 285.749x**BIC4** - 5.4597x**SpMax3_Bh(s)**

Descriptors are labeled in **bold face**. Abbreviations: M (model identification number); S (Size, number of descriptors, i.e. independent variables); in the MLR equation formulae: "x" means "multiplied by"; e.g. in model M9, + 285.749xBIC4 means 285.749 is the multiplication factor beta of descriptor value BIC4. Of note, each molecule has its own individual descriptor value and a spread sheet can be provided on request.

Table 5
Listing of QSAR models based on multiple linear regression (MLR) analysis using our R scripts (cf. Supplementary Materials).

MLR -models	Number of descriptors	Value of R^2	Value of Q^2
Model 1	4	0.9203	0.8938
Model 2	4	0.9181	0.8907
Model 3	4	0.8354	0.7805
Model 4	3	0.8275	0.7877
Model 5	3	0.6649	0.5875
Model 6	3	0.8142	0.7713
Model 7	2	0.6794	0.6336
Model 8	2	0.8883	0.8724
Model 9	2	0.8525	0.8314

Abbreviations: R^2 coefficient of determination; Q^2 measures the consistency between the original and cross-validated prediction data applying the MLR equations.

The correlation matrix graph represents the relationship between each of the independent variables with the response or dependent variable. High correlations with values close to −1 or +1 correspond to dark red or blue colors for a relationship with a negative or positive value, respectively. No correlation is given by a value of 0 in either light red or blue colors. The descriptors are related to IC_{50}, but the two IDVs **BIC4** and **TPSA** ought to be really independent variables, so they have no relatedness between them. If both descriptors gave the same or similar molecular information any general rule (here QSAR equations) could not be drawn. They would be redundant, not essential and even a risk to insert unnecessary noise to the data by two different instrumentations (measurement errors): a case of QSAR pitfall by over-fitting (Scior et al., 2009). Here, both descriptors correlate only weakly (0.43). This indicates a very low dependency (redundancy) avoiding the QSAR pitfall of "non-independent variables" (cf. section on "Pitfall: Collinearity" by Scior et al. (Scior et al., 2009)). For practical purpose, both descriptors are considered uncorrelated and henceforth suitable to be inserted in the next step of QSAR model generation.

3.6. Construction of simple linear regression models

In this part, the Simple Linear Regression (**SLR**) models are presented, for the two aforementioned descriptors. The two **IDVs** of **MLR** model 8 are **BIC4** and **TPSA**. In SLR each one is correlated against the biological response variable. Fig. 6–8 show the SLR graph obtained for

Table 6
Descriptors used to form the final model 8 with size 2 from Table 5.

Molecular descriptors	Information	Source
BIC4	Bond Information Content index (neighborhood symmetry of 4-order)	Information indices
TPSA(Tot)	Topological polar surface area using N,O,S,P polar contributions	Molecular properties

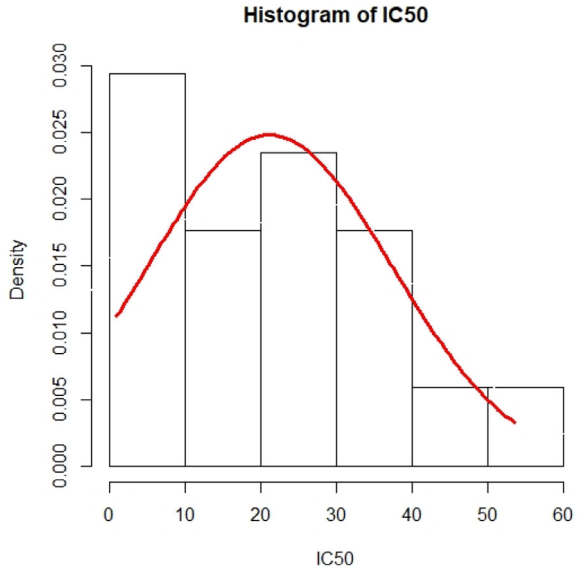


Fig. 4. Histogram by R of the experimentally determined values of the half maximal inhibition concentrations IC_{50}. The Y axis is a measure of occurrence (somehow a "frequency") of activities (probability density) which are represented on the X axis.

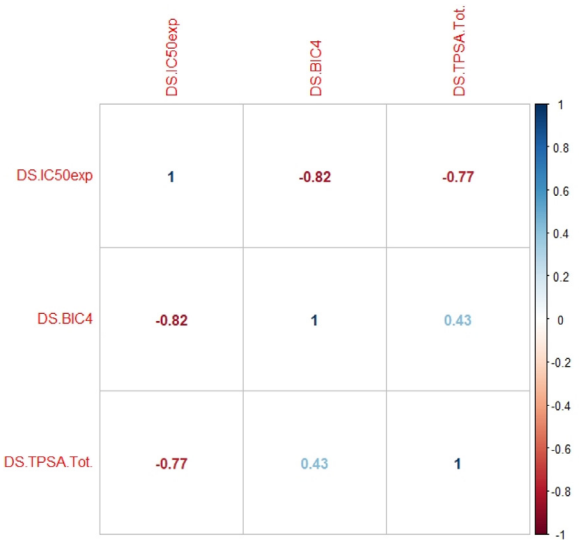


Fig. 5. Color-coded correlation matrix. The colors in dark red (blue) indicate values close to −1 or +1, respectively. Less intense colors symbolize little correlation, that is, close to 0. For abbreviations of the descriptors, cf. Table 3.

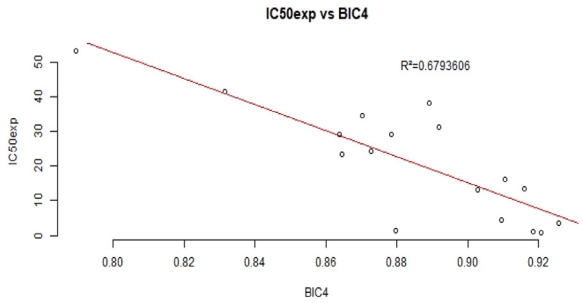


Fig. 6. Simple linear regression by R with one descriptor from model 8 yields a coefficient of determination $R^2 = 0.68$.

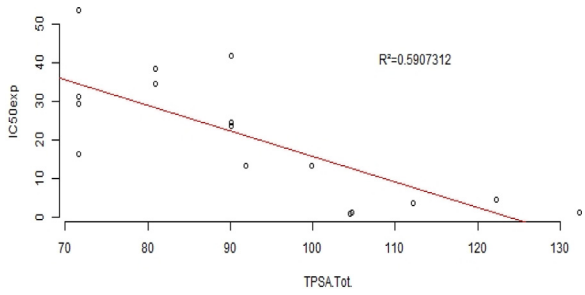


Fig. 7. Simple linear regression by R with a coefficient of determination $R^2 = 0.59$ obtained with the TPSA descriptor.

both equations. In the SLR graph the experimentally determined values of the biological response (**IC_{50} exp**) lie on the Y axis versus (vs) the values of a numeric feature derived from parts of the chemical structure on the X axis. Such a descriptor here constitutes the bonding information content type IV (**BIC4**). The inhibition concentration values decrease when the values of the independent variable increase. This way, when the BIC4 values increase, the inhibitory response activity of the molecules steadily decreases. As a direct consequence, a better inhibition of the TP enzyme will take place and in a similar way the **TPSA (Tot)** descriptor takes on a negative trend, i.e. a steady decrease. Precisely, when the values of the independent variable increase as a direct result, the values of "IC_{50} exp" tend to decrease in a proportional and constant fashion.

The analyses displayed in both figures (Fig. 6 and 7) provide a most valuable hint of how these two variables from final model 8 can influence the inhibitory activity of each of the 17 analogous 3-formyl coumarin molecules. Both variables combined in a Multiple Linear Regression model yield excellent results in form of a very high R^2 value (Table 5).

Table 7
Summary of the results for final model 8 obtained from multiple linear regression (MLR).

Min	1Q	Median	3Q	Max
−6.368	−3.812	−2.177	3.47	13.737
	Estimate (coefficients)	**Std error**	***t*-value**	**pr(>lt l)**
(Intercept)	304.73773	37.25656	8.179	1.06E-06
BIC4	−275.73141	45.13938	−6.108	2.70E-05
TPSA(Tot)	−0.43316	0.08462	−5.119	0.000156
Residual std error	5.736			
	(on 14 DF)	**F-stat**	***P*-value**	
R^2	**R^2-adjusted**			
0.8883	0.8724	55.69	2.165E-07	

Abbreviations: Min/Max = minimum/maximum value, 1Q/3D First/Third Quantile, F-stat = F-statistics, DF = Degrees of Freedom.

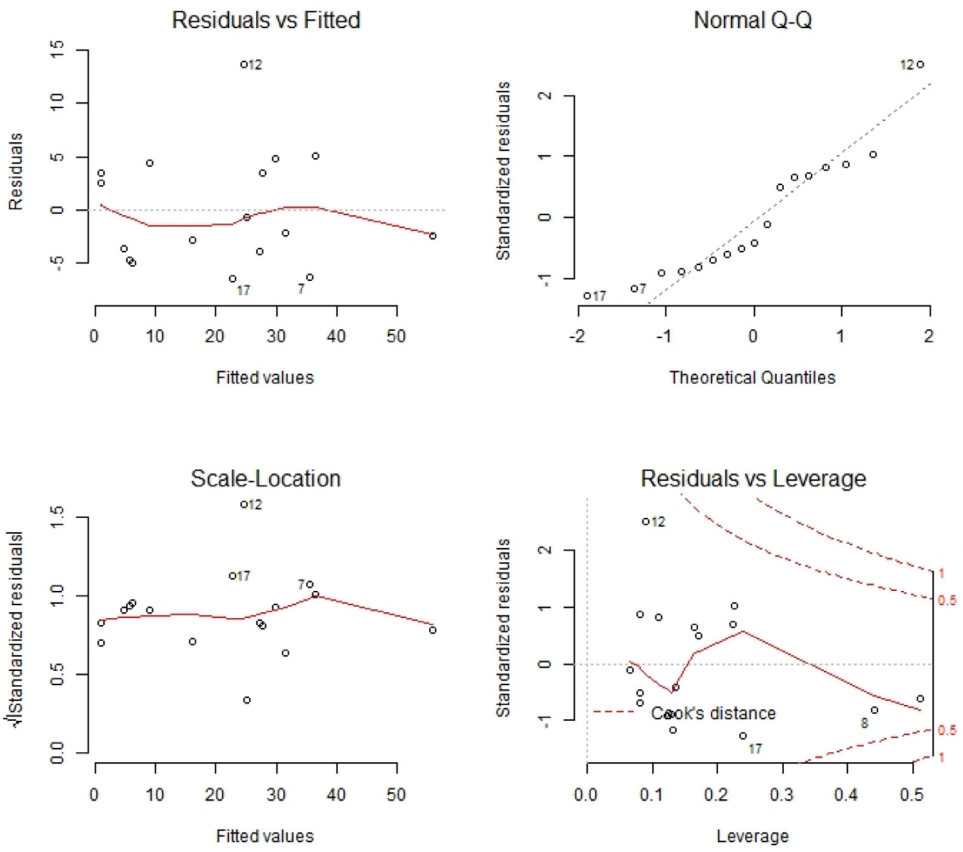


Fig. 8. Statistical analysis of the multiple linear regression model 8. It was obtained by R software with a *P*-value of 2.2×10^{-7} and $R^2 = 0.89$ reflecting a very high statistical significance.

3.7. Statistical analysis of the final model by MLR

We analyzed how the descriptors contribute to the linear equations of our QSAR models. In particular, we examined the question how they perform to predict the biological activity according to their numeric weight, AKA the estimate (of their beta coefficients in Table 7). In terms of a graphical interpretation, the estimates can be seen as the intercept and slope. A more mathematically rigid treatment assumes normalized/standardized values/identical standard deviations of Y and X values. A numerical summary of results for QSAR model 8 is presented in Table 7.

In sight of the well-behaving critical statistics, e.g. the very low p-value, the two values for $R^2 = 0.89$ and R^2-adjusted = 0.87 close to unit 1, it seems not far-fetched to conclude that the final QSAR model was successfully established. It constitutes an acceptable model since for biological test data a larger value range (variability in chemical space and biological response) must be tolerated, and therefore an $R^2 > 0.7$ has been considered as an acceptable threshold in the context of bioassays. In addition, Fig. 8 shows those graphs which validate our multiple linear regression model.

Fig. 8 displays the graph of "Residuals versus fitted values". The graph reflects whether our data have linear or non-linear patterns. In our case we can see that the residuals are well distributed since the red line remains in the center and does not form a parabola. On the other hand, the QQ quantile graph indicates the quartiles or percentiles of data. Here a normal distribution is formed since R takes the data by ordering them in ascending order and compares them with perfectly normally distributed data. Our data approximate a normal distribution, since the points fall close to the straight line and points forming curved tails or very scattered points are not observed.

The "S- Location" graph shows the distribution of the residuals. Their values are well distributed, since the red line is horizontal (not diagonal, much less inclined) while all data points are randomly scattered without moving too far from the line along the X axis. Finally, the graph labeled "Residuals vs Leverage" identifies atypical values with variable influence on the regression results. The variation can be observed by the Cook's distance lines for extreme data values. They lie close to the long lines influencing our linear regression. Removing the corresponding data points would affect the result of our study.

In the graph of "Residuals vs Leverage", data point number 12 is highly influential in our model . On the other hand, omitting this value would improve ("inflate") our coefficient of determination (R^2). Aberrant input data (so called "outliers") omission to optimize R^2 is a sort of data manipulation to obtain better acceptance for publication of otherwise poor results (Scior et al., 2009). This is why here all 17 coumarin compounds were treated without omission (Taha et al., 2018).

3.8. Prediction of biological activity from our equation of the model obtained

Finally, the inhibition constant IC_{50} was calculated, using the proposed **equation** by multiple linear regression. All results are shown in the following Table 8.

The experimental IC_{50} data were compared to the corresponding computed values to obtain pairs of experimental and calculated data points in a XY graph (Fig. 9). In an indirect way, the linearity assumption is also confirmed here. It is needed to justify the choice of linear QSAR models in general and to avoid the aforementioned linearity pitfall. Linearity here can be assumed at least for the limited segment of our input data where IC_{50} have been made available through synthesis and experimental testing. Otherwise MLR would not have provided a determination coefficient as high as almost 0.9 for the final QSAR model 8 with only two (highly informative) descriptors.

Table 8
Results of the IC_{50} from model 8 chosen as the most representative.

ID	Name	IC_{50} experimental	IC_{50} calculated	Residual value (difference)
1	114_L01_Fit16	13.4	8.93182	-4.46818
2	112_L02_Fit16	29.3	31.47711	2.17711
3	110_L03_Fit16	23.5	27.31199	3.81199
4	108_L04_Fit16	3.5	0.92663	-2.57337
5	106_L05_Fit16	1.2	6.10973	4.90973
6	104_L06_Fit16	4.5	1.02980	-3.47020
7	102_L07_Fit16	29.3	35.52186	6.22186
8	100_L08_Fit16	1.3	4.84754	3.54754
9	098_L09_Fit16	24.4	25.03044	0.63044
10	096_L10_Fit16	31.2	27.77017	-3.42983
11	094_L11_Fit16	34.5	29.70530	-4.79470
12	092_L12_FitL16	38.3	24.56265	-13.73735
13	090_L13_FitL16	41.6	36.43819	-5.16181
14	088_L14_FitL16	53.5	55.94658	2.44658
15	086_L15_FitL16	13.2	15.98440	2.78440
16	082_L16_VegaFit	0.9	5.63986	4.73986
17	084_L17_FitL16	16.3	22.66785	6.36785

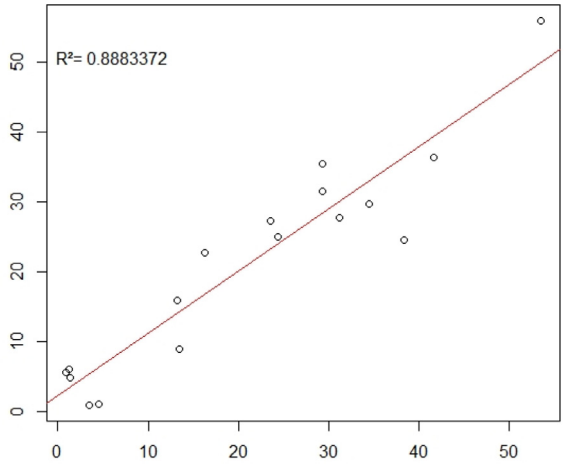


Fig. 9. Correlation diagram of computed versus experimental IC_{50} values. The predicted IC_{50} values from the QSAR model are located on the Y axis, while the X axis takes on the corresponding experimental values. Each pair forms one point in the XY graph. All points reflect a linear tendency. As a reference an ideal regression line was added to the data points. The graph was obtained applying the R software.

4. Conclusion

Computational molecular modeling has ushered a new area to study phytopharmacological data at atomic scale. From the wealth of experimental knowledge about coumarins and their pharmacological effects, we selected a series of 3-formylcoumarins targeting human thymidine phosphorylase (TP). The present QSAR study sheds light on the molecular action mechanism proposing nine linear regression models. They were obtained by running scripts generated under the R statistics program. Precisely, from final model 8 in-depth insight was drawn upon determination of two key molecular descriptors bond index content type 4 (BIC4) and total polar surface area (TPSA). BIC4 and TPSA are definitively the most essential ones to quantitatively explain how the structural diversity relates to experimentally observed inhibitory activity. Upon predicting their biological response power by interacting with a common biomolecular target we sought giving a more general idea of how a quantitative study might improve the understanding of relatedness between plant agents administered to patients in complementary medicine and clinical data. In particular, we demonstrated how QSAR studies from the field of computational medicinal chemistry can be carried out applying the R tool for the development of statistical models with coumarins. Exploiting our findings, a new experimental study for the 3-formylcoumarin-derived molecules can be envisaged. Hence, it is possible to optimize the structures to enhance target affinity. In the context of research and development (R&D), new compounds could be synthesized and tested.

Ethical Approval

Does not apply to theoretical work.

Data Availability

Nil.

Funding

Nil.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Thomas Scior: conceptualization, methodology, R programing, writing original draft; **Juan Carlos Garcia-Hernandez:** model generation, running R scripts; **Hassan H. Abdallah:** MLR statistics; **Christian Alexander:** scientific reading, citations, supervision.

Acknowledgement

We feel much beholden to Dr. Siti Fatimah Zaharah Mustafa from Institute of Marine Biotechnology, Universiti Malaysia Terengganu, Malaysia, for initial participation.

ORCID

Thomas Scior, https://orcid.org/0000-0003-2196-2682.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ccmp.2022.100031.

References

alvaDesc: Molecular Descriptors. (2021). https://www.alvascience.com/alvadesc/.

Al-Warhi, T., Sabt, A., Elkeaeed, E.B., Eldehna, W.M., 2020. Recent advancements of coumarin-based anticancer agents: an up-to-date review. Bioorg. Chem. doi:10.1016/j.bioorg.2020.104163.

Annunziata, F., Pinna, C., Dallavalle, S., Tamborini, L., Pinto, A., 2020. An overview of coumarin as a versatile and readily accessible scaffold whit broad-ranging biological activities. Int. J. Mol. Sci. 21 (13), 4618.

Balaev, V., Lashkov, A., Gabdulkhakov, A., Dontsova, M., Seregina, T., Mironov, A., Betzel, C., Mikhailov, A., 2016. Structural investigation of the thymidine phosphorylase from Salmonella typhimurium in the unliganded state and its complexes with thymidine and uridine. Acta Crystallogr. F Struct. Biol. Commun. 224–233.

Bernard, P., Scior, T., Didier, B., Hibert, M., Berthon, J.Y., 2002. Ethnopharmacology and bioinformatic combination for leads discovery: application to phospholipase A2 inhibitors. Phytochemistry 58 (6), 865–874.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucl. Acid. Res. 28 (1), 235–242.

Blondeau, S., Do, Q.T., Scior, T., Bernard, P., Morin-Allory, L., 2010. Reverse pharmacognosy: another way to harness the generosity of nature. Curr. Pharmaceut. Desi. 16 (15), 1682–1696.

Clark, M., Cramer, R.D., Van Opdenbosch, N., 1989. Validation of the general purpose tripos 5.2 force field. J. Comput. Chem. 10 (8), 982–1012.

Daina, A., Michielin, O., Zoete, V., 2017. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. 7 (1), 42717.

Dehmer, M., Varmuza, K., Bonchev, D. (Eds.), 2012. Statistical Modelling of Molecular Descriptors in QSAR/QSPR. Wiley-Blackwell.

Do, Q., Medina-Franco, J., Scior, T., Bernard, P., 2015. How to valorize biodiversity? Let's go hashing, extracting, filtering, mining, fishing. Planta Med. 81 (06), 436–449.

Emami, S., Dadashpour, S., 2015. Current developments of coumarin-based anti-cancer agents in medicinal chemistry. Eur. J. Med. Chem. 102, 611–630.

Garrard, A., 2014. Coumarins. En A.Garrard, Encyclopedia of Toxicology. Elsevier, pp. 1052–1054 págs.

Huang, X.Y., Shan, Z.J., Zhai, H.L., Su, L., Zhang, X.Y., 2011. Study on the anticancer activity of coumarin derivatives by molecular modeling. Chem. Biol. Drug Des. doi:10.1111/j.1747-0285.2011.01195.x.

Lozano-Aponte, J., Scior, T., 2012. ¿Qué sabe Ud. acerca de...QSAR? What do you know about QSAR? Rev. Mex. Cienc. Farm. 43 (2), 82–84.

Luo, G., Muyaba, M., Lyu, W., Tang, Z., Zhao, R., Xu, Q., You, Q., Xiang, H., 2017. Design, synthesis and biological evaluation of novel 3-substituted 4-anilino-coumarin derivatives as antitumor agents. Bioorg. Med. Chem. Lett. doi:10.1016/j.bmcl.2017.01.013.

Marshall, M.E., Mohler, J.L., Edmonds, K., Williams, B., Butler, K., Ryles, M., Weiss, L., Urban, D., Bueschen, A., Markiewicz, M., Cloud, G., 1994. An updated review of the clinical development of coumarin (1,2-benzopyrone) and 7-hydroxycoumarin. J. Cancer Res. Clin. Oncol. 120 (S1), S39–S42.

Medina-Franco, J.L., Fernández-de Gortari, E., Naveja, J.J., 2015. Avances en el diseño de fármacos asistido por computadora. Educación Química 26 (3), 180–186.

Norman, R.A., Barry, S.T., Bate, M., Breed, J., Colls, J.G., Ernill, R.J., Luke, R.W.A., Minshull, C.A., McAlister, M., S.B., McCall, E.J., McMiken, H., H.J., Paterson, D.S., Timms, D., Tucker, J.A., Pauotit, R.A., 2004. Crystal structure of human thymidine phosphorylase in complex with a small molecule inhibitor. Strcuture 75–84.

Pedretti, A., Mazzolari, A., Gervasoni, S., Fumagalli, L., Vistoli, G., 2021. The VEGA suite of programs: an versatile platform for cheminformatics and drug design projects. Bioinformatics 37 (8), 1174–1175.

Prusty, J.S., Kumar, A., 2020. Coumarins: antifungal effectiveness and future therapeutic scope. Mol. Divers. 24 (4), 1367–1383.

Pugmire, M.J., Cook, W.J., Jasanoff, A., Walter, M.R., Ealick, S.E., 1998. Structural and theoretical studies suggest domain movement produces an active conformation of thymidine phosphorylase. J. Molecul. Biol. 285–299.

Pugmire, M.J., Ealick, S.E., 1998. The crystal structure of pyrimidine nucleoside phosphorylase in a closed conformation. Structure 1467–1479.

Quiroga, I., Melendez, F., Salvador, K., Scior, T., 2018. Identification a new site of metabolism for phenprocoumon by modeling it's CYP2C9 hydroxylation pattern. SAJ Pharm. Pharmacol. 5 (1), 2 1.

Quiroga, I., Scior, T., 2017. Structure-function analysis of the cytochromes P450, responsible for phenprocoumon metabolism. J. Mex. Chem. Soc. 61 (4), 349–360.

Revankar, H.M., Bukhari, S.N., Kumar, G.B., Qin, H.L., 2017. Coumarins scaffolds as COX inhibitors. Bioorg. Chem. 71, 146–159.

Roy, K., Kar, S., Das, R.N., 2015. A Primer on QSAR/QSPR Modeling Fundamental Concepts. Springer International Publishing doi:10.1007/978-3-319-17281-1.

Scior, T., Garces-Eisele, S.J., 2006. Isoniazid is not a lead compound for its pyridyl ring derivatives, isonicotinoyl amides, hydrazides, and hydrazones: a critical review. Curr. Med. Chem. 13 (18), 2205–2219.

Scior, T., Quiroga-Montes, I., Kammerer, B., 2018. Inquiry of literature evidence for induced fit of cytochrome P450 2C9 for Warfarin, phenprocoumon, flurbiprofen and clopidogrel: a critical review. SCIOL Biotechnol. 1, 30–48.

Scior, T., Medina-Franco, J., Do, Q.-.T., Martinez-Mayorga, K., Yunes Rojas, J., Bernard, P., 2009. How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr. Med. Chem. 16 (32), 4297–4313.

Singh, H., Singh, J.V., Bhagat, K., Gulati, H.K., Sanduja, M., Kumar, N., Kinarivala, N., Sharma, S., 2019. Rational approaches, design strategies, structure activity relationship and mechanistic insights for therapeutic coumarin hybrids. Bioorg. Med. Chem. 27 (16), 3477–3510.

Tafesse, T.B., Bule, M.H., Khoobi, M., Faramarzi, M.A., Abdollahi, M., Amini, M., 2020. Coumarin-based Scaffold as alpha-glucosidase Inhibitory Activity: implication for the development of potent antidiabetic agents. Mini Rev. Med. Chem. 20 (2), 134–151.

Taha, M., Adnan Ali Shah, S., Afifi, M., Imran, S., Sultan, S., Rahim, F., Hadiani Ismail, N., Mohammed Khan, K., 2018. Synthesis, molecular docking study and thymidine phosphorylase inhibitory activity of 3-formylcoumarin derivatives. Bioorg. Chem. 78, 17–23. doi:10.1016/j.bioorg.2018.02.028.

Thakur, A., Singla, R., Jaitak, V., 2015. Coumarins as anticancer agents: a review on synthetic strategies, mechanism of action and SAR studies. Eur. J. Med. Chem. doi:10.1016/j.ejmech.2015.07.010.

Thakur, A., Kaur, K., Sharma, P., Singla, R., Singh, S., Jaitak, V., 2021. Synthesis, in vitro, and docking analysis of C-3 substituted coumarin analogues as anticancer agents. Curr. Comput. Aid. Drug Des. doi:10.2174/1573409916666200120114641.

Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y., Prokopenko, V.V., 2005. Virtual computational chemistry laboratory – design and description. J. Comput. Aid. Mol. Des. 19 (6), 453–463.

The R Project for Statistical Computing. (2021). https://www.r-project.org/.

Timofeev, V.I., Abramchik, Y.A., Fateev, I.V., Zhukhlistova, N.E., Murav´eva, T.I., Kuranova, I.P., Esipov, R.S., 2013. Three- Dimensional Structure of Thymidine Phosphorylase from E. coli in complex whit 3´- Azido-2´-Fluoro-2´,3´-Dideoxyuridine. Crystallogr. Rep. 842–853.

Timofeev, V., Abramchik, Y., Zhukhlistova, N., Muravieva, T., Fateev, I., Esipov, R., Kuranova, I., 2014. 3′-Azidothymidine in the active site of *Escherichia coli* thymidine phosphorylase: the peculiarity of the binding on the basis of X-ray study. Acta Crystallograph. Sect. D Biolog. Crystallogr. 1155–1165.

Todeschini, R., & Consonni, V. (2000). Handbook of Molecular Descriptors (R. Mannhold, H. Kubinyi, & Hendrik Timmerman (Eds.)). Wiley.

Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H., & Folkers, G. (Eds.). (2008). Handbook of Molecular Descriptors. Wiley-VCH.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model. 28 (1), 31–36.

Zhang, L., Xu, Z., 2019. Coumarin-containing hybrids and their anticancer activities. J. Med. Chem. doi:10.1016/j.ejmech.2019.111587.

Zhu, J.J., Jiang, J.G., 2018. Pharmacological and nutritional effects of natural coumarins and their structure-activity relationships. Mol. Nutr. Food Res. doi:10.1002/mnfr.201701073.