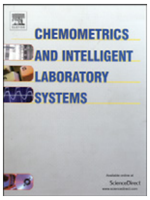
Chemometrics and Intelligent Laboratory Systems 132 (2014) 18–29

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



Predictive QSAR study of chalcone derivatives cytotoxicity activity against HT-29 human colon adenocarcinoma cell lines

Martyna Rybka [a], Andrew G. Mercader [b,*], Eduardo A. Castro [b]

[a] *Warsaw University of Technology, Warsaw, Poland*
[b] *Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina*

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form 1 December 2013
Accepted 14 December 2013
Available online 22 December 2013

Keywords:
Chalcones
Anticancer
Cytotoxicity
HT-29 cell lines
QSAR
Enhanced Replacement Method

ABSTRACT

Chalcones and their derivatives exhibit a wide range of important pharmacological activities; among the most relevant ones is their anticancer activity. For this reason we performed a predictive Quantitative Structure–Activity Relationships (QSAR) analysis of the anticancer activity against HT-29 human colon adenocarcinoma cell lines by chalcones. The best linear model constructed from 136 molecular structures incorporated seven molecular descriptors, selected from more than a thousand geometrical, topological, quantum-mechanical and electronic types of descriptors, showed good predictive ability. The model analysis showed that the descriptors with highest importance on the activity are the number of 10-member rings and a BCUT descriptor weighted by the van der Waals volume.

1. Introduction

Constantly increasing number of cancer cases around the world is one of the main causes of death in both economically developed and developing countries [1]. In view of the unlimited ability of tumor cells to the proliferation, avoiding of apoptosis, degree of the invasion and presence of metastases, cancer is considered to be the biggest challenge of present medicine [2]. In spite of progress in chemotherapy, there are no clinically effective substances acting selectively on tumor cells [3].

One of the pharmacologically relevant group of substances that possess an interesting range of biological activities are chalcones and their derivatives, exhibiting a wide range of pharmacological activities, as anticancer [3,4], anti-inflammatory [5,6], antimalarial [6–8], etc.

Chalcones are 1,3-diphenyl-2-propene-1-one consisting of two aromatic rings linked by a three carbon α, β-unsaturated carbonyl system and are fundamental intermediate compounds in flavonoid and isoflavonoid biosynthesis in plants. Synthetic and naturally occurring chalcones are regarded to be pharmacologically important compounds [9]. Numerous substances of this group have been approved to affect with each level of carcinogenesis and to exhibit activity against cancer cells. This suggests that chalcones may be considered as a group of potential anticancer agents [10]. Changes in their structure have offered a high level of variation that may be useful for the development of new therapeutic agents having improved potency and lower toxicity [11].

A great deal of chalcones derivatives were previously obtained and assessed for their cytotoxic and anticancer activity against series of cancer cell lines. In this paper we have taken into consideration the HT-29 human colon adenocarcinoma cell line. These cells exhibit the ability to form three-dimensional structures called multicellular spheroids showing a great similarity to natural tissue [12].

Within the currently ongoing search for effective anticancer drugs candidates in the present study we have carried out and established a reliable quantitative structure–activity relationship (QSAR) analysis based on 155 chalcones derivatives.

2. Methods

2.1. Data sets

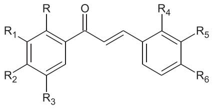
In our QSAR study, a total of 162 chalcone molecules were gathered from the literature [3,13–22], to our knowledge this set of molecules was not employed in a predictive (QSAR) study before. The log IC_{50} value, concentration of the compound (μM) exhibiting 50% inhibition of cell growth [21], for human colorectal cancer cell line, HT-29, was employed as the dependent variable (Table 1). The data-set was divided into a training set of 136 compounds and a test set of 19 compounds (7 compounds were identified as outliers).

All the compounds were evaluated for their cytotoxic activity by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay based on mitochondrial reduction of yellow MTT tetrazolium dye to a highly colored blue formazan product [23].

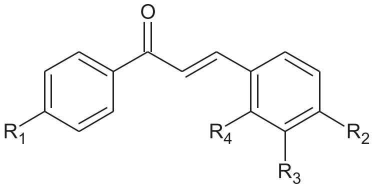
[*] Corresponding author. Tel.:+54 11 6091 3759; fax: +54 11 6091 2100 (int:3759).
E-mail address: amercader@inifta.unlp.edu.ar (A.G. Mercader).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.12.005

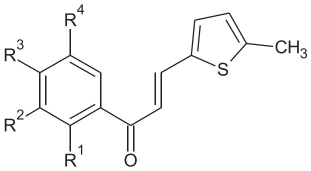
Table 1
Structure of compounds, experimental log IC_{50}, predicted log IC_{50} by Eq. (4), and residuals. Superscript "t" indicates test set substances.



Compound	R	R_1	R_2	R_3	R_4	R_5	R_6	log IC_{50} exp. [13]	log IC_{50} pred.	Res.
1	H	H	NH_2	H	Cl	H	Cl	1.602	1.497	0.105
2	H	H	OCH_3	H	Cl	H	Cl	1.276	1.316	−0.040
3	OH	H	H	H	Cl	H	Cl	1.627	1.355	0.272
4	H	H	NH_2	H	Cl	H	H	1.469	1.665	−0.196
5	H	H	H	H	H	H	OCH_3	1.384	1.720	−0.336
6	H	H	NH_2	H	H	H	OCH_3	1.746	1.759	−0.014
7	H	NH_2	H	H	H	H	OCH_3	1.631	1.754	−0.123
8	H	H	OCH_3	H	H	H	OCH_3	1.691	1.585	0.106
9	OH	H	H	H	H	H	OCH_3	1.775	1.641	0.134
10	H	F	H	F	H	H	OCH_3	2.005	1.595	0.409
11	H	H	NH_2	H	F	H	H	1.680	1.799	−0.119
12	H	H	OCH_3	H	F	H	H	1.276	1.625	−0.349
13	H	H	H	H	H	H	OH	1.391	1.763	−0.372
14	H	H	OCH_3	H	H	H	OH	1.269	1.614	−0.345
15	H	H	H	H	OCH_3	H	OCH_3	2.116	1.595	0.521
16	H	H	F	F	H	OCH_3	H	1.726	1.608	0.118
17	H	H	NH_2	H	OCH_3	OCH_3	H	1.843	1.698	0.145
18	H	NH_2	H	H	OCH_3	OCH_3	H	1.641	1.635	0.006
19	H	H	OCH_3	H	OCH_3	OCH_3	H	1.621	1.563	0.058
20^t	H	H	H	H	CF_3	H	H	1.632	1.672	−0.040
21^t	H	H	OCH_3	H	CF_3	H	H	1.562	1.527	0.035
22^t	H	H	NH_2	H	Br	H	H	1.549	1.498	0.051
23	H	H	OCH_3	H	Br	H	H	1.411	1.320	0.091
24	H	H	OCH_3	H	F	H	F	1.047	1.487	−0.440
25	H	OCH_3	OCH_3	H	Cl	H	Cl	1.272	1.201	0.071
26	H	OCH_3	OCH_3	H	H	H	$N(CH_3)_2$	1.698	1.796	−0.098
27	H	OCH_3	OCH_3	H	Cl	H	H	1.189	1.394	−0.205
28	H	OCH_3	OCH_3	H	H	H	OCH_3	1.737	1.620	0.118
29	H	OCH_3	OCH_3	H	OCH_3	H	OCH_3	1.816	1.384	0.433
30	H	OCH_3	OCH_3	H	OCH_3	OCH_3	H	1.555	1.480	0.075
31	H	OCH_3	OCH_3	H	Br	H	H	1.145	1.210	−0.065



Compound	R_1	R_2	R_3	R_4	log IC_{50} exp. [3]	log IC_{50} pred.	Res.
32	H	H	H	H	1.686	1.899	−0.212
33	CH_3	H	H	H	1.935	1.995	−0.060
34	OCH_3	H	H	H	2.195	1.807	0.388
35	H	H	H	Cl	1.738	1.641	0.097
36	OH	H	H	Cl	1.779	1.550	0.230
37	H	H	CH_3	H	1.932	2.010	−0.079
38	OH	H	CH_3	H	1.706	1.986	−0.280
39	CH_3	H	CH_3	H	1.839	2.058	−0.219
40	OCH_3	H	CH_3	H	1.796	1.917	−0.120
41	CH_3	OCH_3	H	H	2.069	1.847	0.222
42	H	$N(CH_3)_2$	H	H	2.408	1.903	0.505
43	CH_3	$N(CH_3)_2$	H	H	2.419	2.004	0.415
44	H	H	H	OH	1.947	1.822	0.126
45	CH_3	H	H	OH	1.737	1.899	−0.162
46	OCH_3	H	H	OH	1.653	1.665	−0.011
47	H	H	Cl	H	1.507	1.694	−0.188

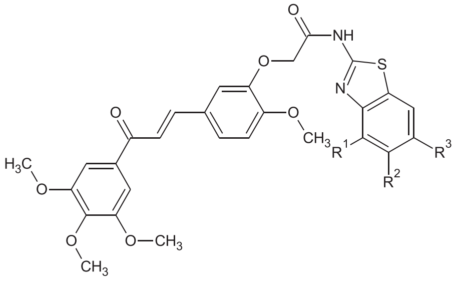


Compound	R_1	R_2	R_3	R_4	log IC_{50} exp. [14]	log IC_{50} pred.	Res.
48	OH	H	H	OC_2H_5	1.638	1.793	−0.156

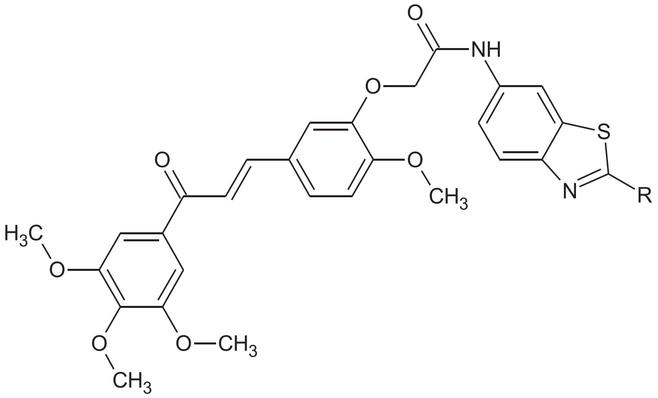
(continued on next page)

Table 1 (*continued*)

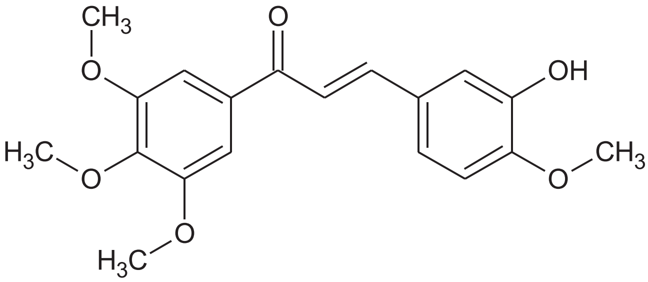
49	OC_2H_5	H	H	OC_2H_5	0.505	–	–
Compound	R_1	R_2	R_3	R_4	log IC_{50} exp. [14]	log IC_{50} pred.	Res.
50	OH	H	H	OC_3H_7	2.099	1.815	0.284
51	OH	H	H	$OCH(CH_3)_2$	1.480	1.749	−0.269
52	OH	H	H	OC_4H_9	1.854	1.753	0.101
53	OH	H	H	$OCH_2CH(CH_3)_2$	1.487	1.911	−0.424
54	OH	H	H	OC_5H_{11}	1.939	1.638	0.301
55	OH	H	H	OCH_3	1.854	1.708	0.146
56	OC_2H_5	H	H	OCH_3	−1.000	–	–
57	OC_3H_7	H	H	OCH_3	1.938	1.788	0.149
58	OC_4H_9	H	H	OCH_3	1.786	1.732	0.054
59	OH	OH	OCH_3	H	1.728	1.599	0.129
60	OH	OH	OH	H	1.555	1.595	−0.040



Compound	R_1	R_2	R_3	log IC_{50} exp. [15]	log IC_{50} pred.	Res.
61	H	H	H	0.185	0.669	−0.484
62	H	H	F	1.076	0.657	0.419
63	H	H	Cl	0.785	0.587	0.198
64	H	H	NO	0.940	0.725	0.214
65	H	H	OCH_3	0.532	0.542	−0.010
66	H	H	OCF_3	0.041	0.349	−0.307
67^t	H	H	CF_3	0.568	0.514	0.054
68	H	H	CH_3	0.322	0.618	−0.296
69	H	H	OC_2H_5	0.380	0.556	−0.176
70	CH_3	H	H	0.785	0.578	0.207
71	H	CH_3	CH_3	0.771	0.681	0.090

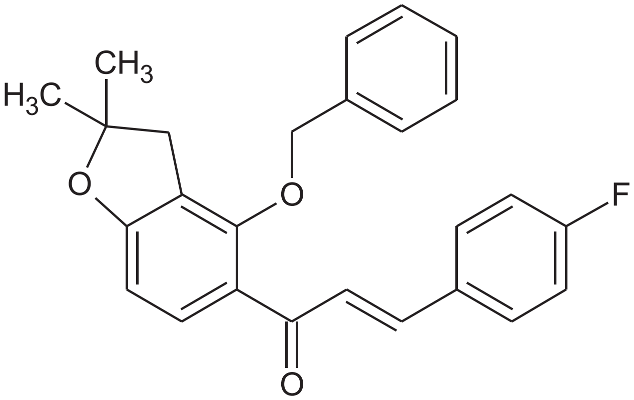


Compound	R	log IC_{50} exp. [15]	log IC_{50} pred.	Res.
72	H	0.924	0.806	0.118
73	CH_3	0.978	1.348	−0.371

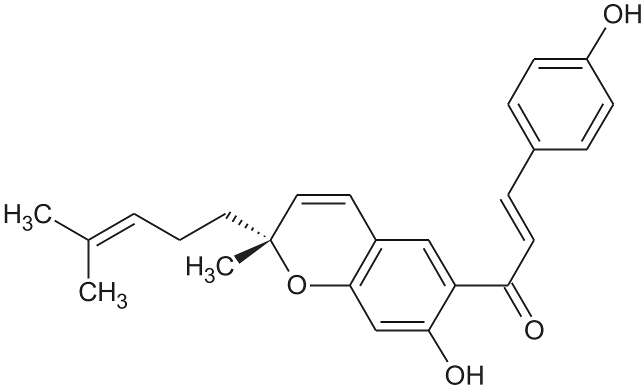


Compound	log IC_{50} exp. [16]	log IC_{50} pred.	Res.
74	0.279	–	–

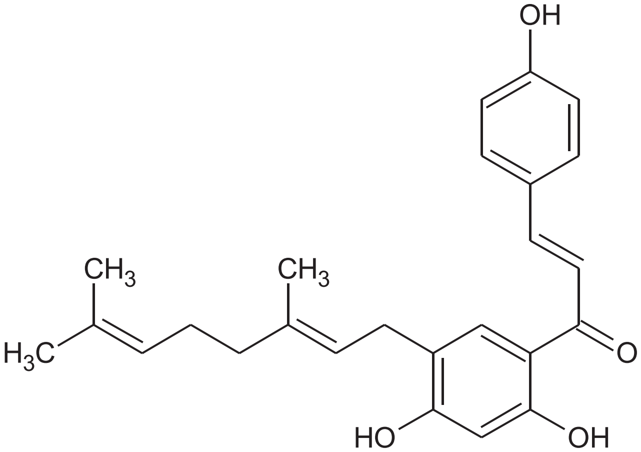
Table 1 (*continued*)



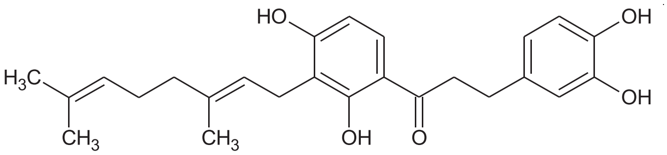
Compound	log IC$_{50}$ exp. [16]	log IC$_{50}$ pred.	Res.
75	1.362	1.130	0.231



Compound	log IC$_{50}$ exp. [17]	log IC$_{50}$ pred.	Res.
76	1.041	0.696	0.346



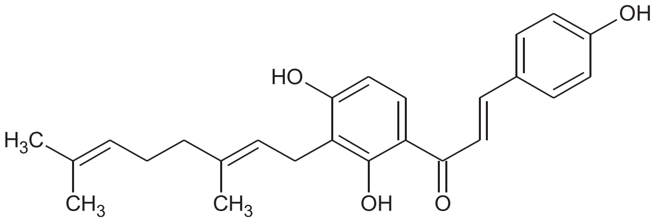
Compound	log IC$_{50}$ exp. [17]	log IC$_{50}$ pred.	Res.
77	1.065	1.092	−0.028



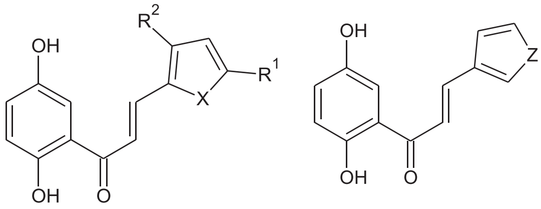
Compound	log IC$_{50}$ exp. [17]	log IC$_{50}$ pred.	Res.
78	1.295	1.419	−0.125

(*continued on next page*)

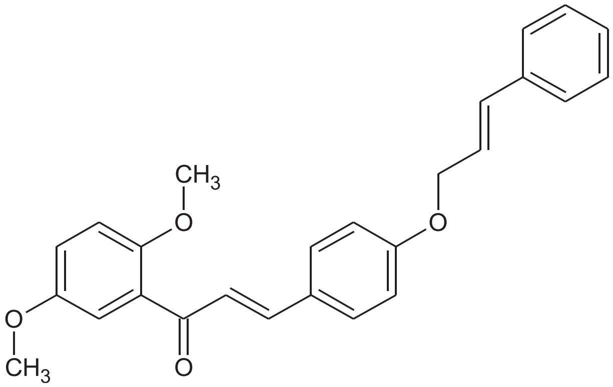
Table 1 (*continued*)



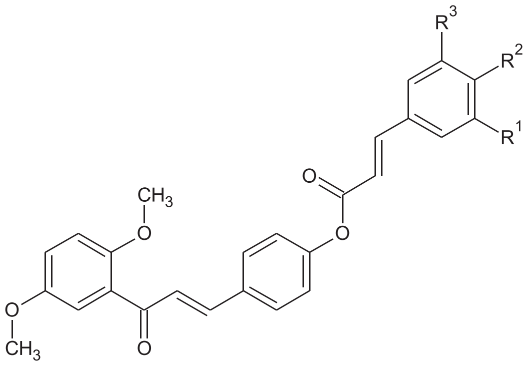
Compound	log IC$_{50}$ exp. [17]	log IC$_{50}$ pred.	Res.
79	1.212	1.250	−0.037



Compound	R$_1$	R$_2$	X	Z	log IC$_{50}$ exp. [18]	log IC$_{50}$ pred.	Res.
80^{t}	H	H	O		1.395	1.384	0.010
81				O	1.328	1.397	−0.069
82^{t}	H	H	S		1.124	1.193	−0.069
83				S	0.724	1.321	−0.596
84	CH_3	H	O		1.634	1.943	−0.309
85	CH_3	H	S		0.255	–	–
86	H	CH_3	S		1.288	1.190	0.098
87	Br	H	O		1.182	1.153	0.029

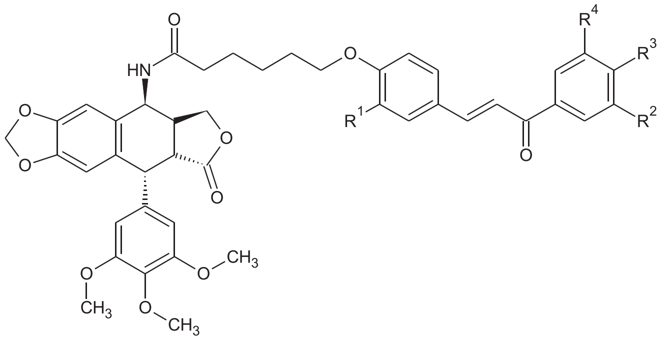


Compound	log IC$_{50}$ exp.[18]	log IC$_{50}$ pred.	Res.
88	1.420	1.346	0.074

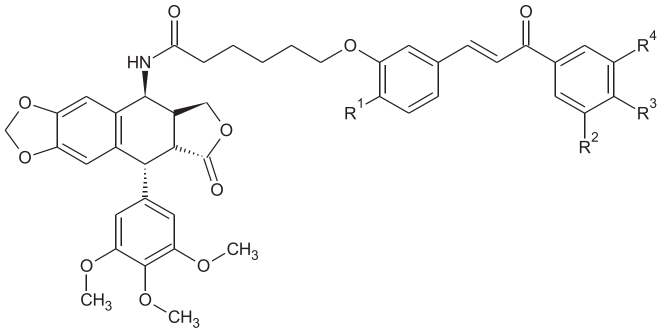


Compound	R$_1$	R$_2$	R$_3$	log IC$_{50}$ exp. [18]	log IC$_{50}$ pred.	Res.
89^{t}	H	CH_3	H	1.215	1.614	−0.399
90	H	OCH_3	H	1.362	1.326	0.036
91	OCH_3	OCH_3	OCH_3	1.377	1.344	0.033

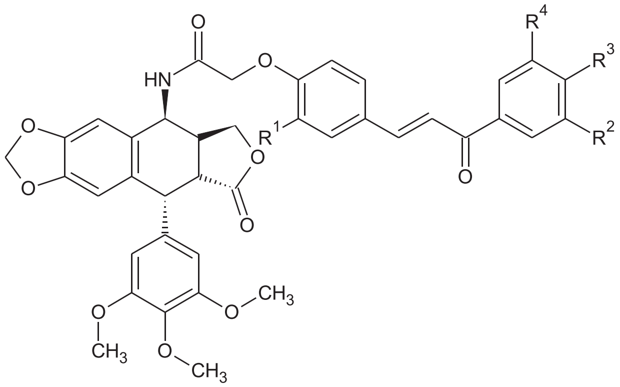
Table 1 (*continued*)



Compound	R_1	R_2	R_3	R_4	log IC$_{50}$ exp. [19]	log IC$_{50}$ pred.	Res.
92	H	OCH_3	OCH_3	OCH_3	1.185	1.085	0.099
93	OCH_3	OCH_3	H	H	1.076	1.093	−0.018
94^{t}	OCH_3	H	OCH_3	H	1.207	1.157	0.050
95	OCH_3	NO	H	H	0.942	1.183	−0.241
96	OCH_3	OCH_3	OCH_3	OCH_3	1.083	1.112	−0.030
97	OCH_3	H	F	H	1.068	1.205	−0.137



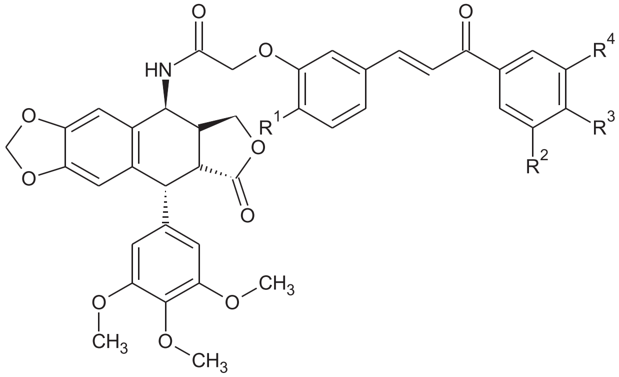
Compound	R_1	R_2	R_3	R_4	log IC$_{50}$ exp. [19]	log IC$_{50}$ pred.	Res.
98	H	OCH_3	OCH_3	OCH_3	1.324	1.138	0.186
99^{t}	OCH_3	OCH_3	OCH_3	OCH_3	0.811	0.926	−0.115



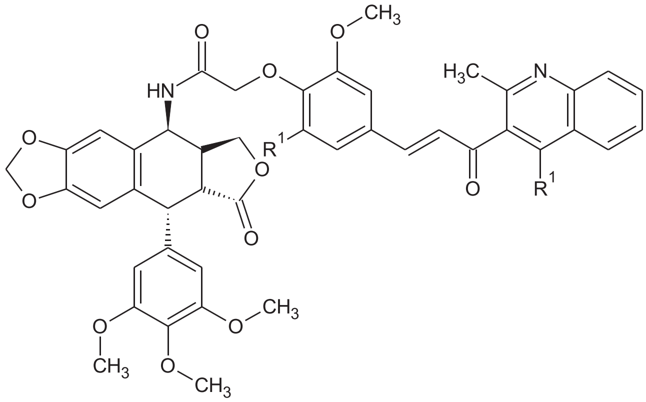
Compound	R_1	R_2	R_3	R_4	log IC$_{50}$ exp. [19]	log IC$_{50}$ pred.	Res.
100	H	OCH_3	OCH_3	OCH_3	1.093	0.969	0.125
101	OCH_3	OCH_3	H	H	0.909	0.785	0.123
102^{t}	OCH_3	H	OCH_3	H	1.228	0.992	0.236
103	OCH_3	NO	H	H	0.925	1.077	−0.152
104	OCH_3	OCH_3	OCH_3	OCH_3	0.980	0.883	0.096
105^{t}	OCH_3	H	F	H	1.335	1.062	0.273

(*continued on next page*)

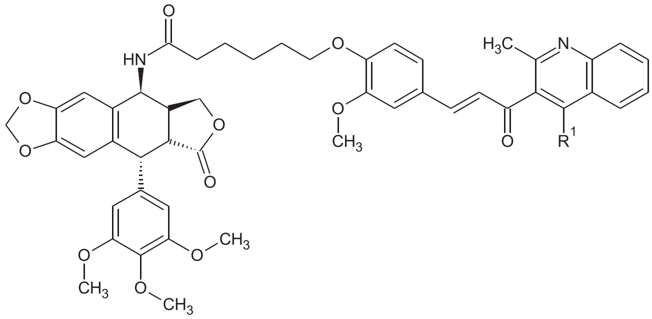
Table 1 (*continued*)



Compound	R_1	R_2	R_3	R_4	log IC_{50} exp. [19]	log IC_{50} pred.	Res.
106	H	OCH_3	OCH_3	OCH_3	1.193	0.938	0.255
107	OCH_3	OCH_3	OCH_3	OCH_3	0.777	0.965	−0.189

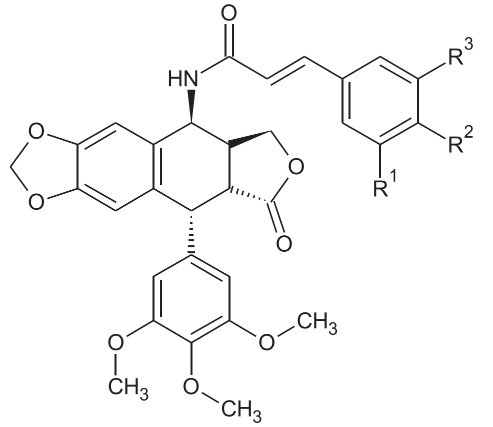


Compound	R_1	log IC_{50} exp. [19]	log IC_{50} pred.	Res.
108	CH_3	1.090	0.922	0.168
109^t	Ph	0.889	0.709	0.181

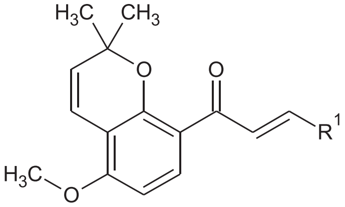


Compound	R_1	log IC_{50} exp. [19]	log IC_{50} pred.	Res.
110	CH_3	1.072	1.046	0.026
111	Ph	0.949	0.949	0.000

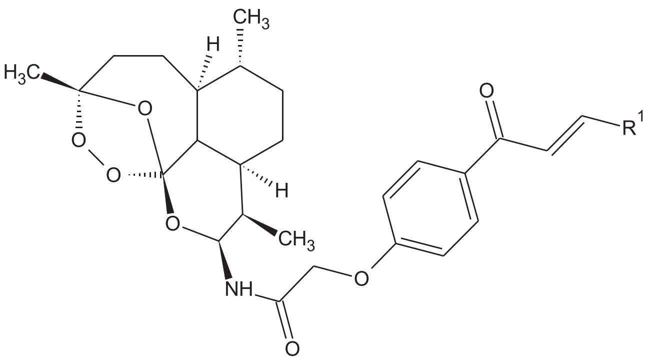
Table 1 (*continued*)



Compound	R_1	R_2	R_3	log IC_{50} exp. [19]	log IC_{50} pred.	Res.
112	H	H	H	0.373	–	–
113^{t}	H	F	H	0.851	1.227	−0.376
114	H	Cl	H	1.057	1.211	−0.154
115	H	Br	H	1.283	1.147	0.136
116^{t}	H	OCH_3	H	1.246	1.179	0.067
117	H	OCF_3	H	−0.432	–	–
118	H	NO	H	1.258	1.259	−0.001
119	Cl	Cl	H	1.322	1.164	0.158
120	OCH_3	OCH_3	OCH_3	0.921	1.212	−0.290



Compound	R_1	log IC_{50} exp. [20]	log IC_{50} pred.	Res.
121^{t}	4-(Dimethylamino)phenyl	0.272	0.858	−0.586
122	Naphthalen-2-yl	0.651	0.906	−0.254
123	4-(Diethylamino)phenyl	−0.260	–	–
124	4-(Pyrrolidin-1-yl)phenyl	0.922	0.763	0.159
125^{t}	4-(1H-Imidazol-1-yl)phenyl	0.820	0.886	−0.066
126	Quinolin-2-yl	0.579	0.453	0.126
127	4-Nitrophenyl	0.572	0.786	−0.215
128	3-Nitrophenyl	0.629	0.835	−0.206
129	3-(Trifluoromethyl)phenyl	0.957	0.687	0.270
130	Pyridin-2-yl	0.326	0.832	−0.506
131	3.4-Dichlorophenyl	0.880	0.625	0.255
132	2-(4-Chloro)phenol	0.696	0.632	0.064

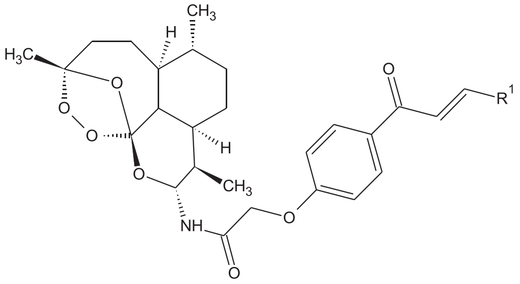


Compound	R_1	log IC_{50} exp. [21]	log IC_{50} pred.	Res.
133	2.4-Difluorophenyl	−0.032	−0.018	−0.013
134	3.4.5-Trimethoxyphenyl	−0.620	−0.242	−0.378
135	4-Trifluoromethoxyphenyl	−0.444	−0.405	−0.039
136	3-Trifluoromethylphenyl	−0.585	−0.322	−0.263
137	4-Fluorophenyl	−0.155	−0.173	0.018
138	4-Chlorophenyl	−0.301	−0.261	−0.040

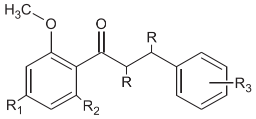
(*continued on next page*)

Table 1 (*continued*)

Compound	R^1	log IC_{50} exp. [21]	log IC_{50} pred.	Res.
139	2.6-duchlorophenyl	−0.854	−0.285	−0.569
140^{t}	3-Methoxyphenyl	−0.143	−0.178	0.035
141	Phenyl	−0.187	−0.137	−0.050
142	2-Furyl	0.362	0.211	0.150
143	2-Thienyl	0.204	0.080	0.124
144	1.3-Benzodioxo-5-yl	0.079	−0.089	0.168



Compound	R_1	log IC_{50} exp. [21]	log IC_{50} pred.	Res.
145	2.4-Difluorophenyl	0.415	−0.022	0.437
146	3.4.5-Trimethoxyphenyl	−0.155	−0.203	0.048
147	4-Trifluoromethoxyphenyl	−0.409	−0.339	−0.070
148	3-Trifluoromethylphenyl	−0.620	−0.315	−0.305
149	4-Fluorophenyl	−0.119	−0.177	0.057
150^{t}	4-Chlorophenyl	−0.174	−0.246	0.073
151	2.6-Duchlorophenyl	−0.301	−0.271	−0.030
152	3-Methoxyphenyl	−0.174	−0.173	−0.001
153	Phenyl	−0.125	−0.131	0.006
154	2-Furyl	0.146	0.152	−0.006
155	2-Thienyl	0.322	0.017	0.306
156	1.3-Benzodioxo-5-yl	0.230	−0.088	0.318



Compound	R	R_1	R_2	R_3	log IC_{50} exp. [22]	log IC_{50} pred.	Res.
157	H	OCH_3	OH	3.4 OCH_2O	2.161	1.688	0.472
158	Double bond	OCH_3	OH	3 OCH_3	1.656	1.535	0.121
159	Double bond	OH	OH	3 OCH_3	1.335	1.600	−0.265
160	Double bond	OH	OH	3.4 OCH_2O	1.316	1.459	−0.143
161	Double bond	$OCH_2CH{=}CH_2$	$OCH_2CH{=}CH_2$	3 OCH_3	1.453	1.389	0.065
162	Double bond	$OCH_2CH{=}CH_2$	$OCH_2CH{=}CH_2$	3.4 OCH_2O	1.382	1.300	0.082

Preliminary tests indicated that molecules **49**, **56**, **74**, **85**, **112**, **117** and **123** presented excessively high errors indicating that they were outliers. The tests were performed comparing the results of approximately 1000 different calibrations, where the outliers presented errors that greatly exceeded the limit of 2.5σ; for these calculations the complete set of molecules was used. Since the prediction from any QSAR model cannot be intrinsically better than the experimental data employed to develop it, and the quality of the input data greatly influences the performance of QSAR models [24], these outliers were preventively left aside and not used in further calculations.

2.2. Molecular descriptors

The structures of the compounds were firstly pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in the Hyperchem 6.03 package [25], and the resulting geometries were further refined by means of the semi-empirical method AM1 (Austin Method 1) using the Polak–Ribière algorithm and a gradient norm limit of 0.01 kcal·Å^{-1}. The molecular descriptors were computed using the software Parameter Client, Virtual Computational Chemistry Laboratory [26,27] which calculates parameters of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centered Fragments [28], electrotopological state molecular indices (ETState) and two sets of fragment descriptors (GSFRAG and GSFRAG-L) [27]. Parameter Client requires MOL2 (Sybyl) input format; in order to translate the structures to this format the software Open Babel was employed [29]. The initial descriptor matrix from Parameter Client contained 1932 descriptors. Nevertheless, some molecules did not permit the calculation of a number of descriptors; after removing these descriptors and linear dependant ones, the resulting total pool consisted of $D = 1433$ descriptors.

2.3. Model search

It is our purpose to search the set **D**, containing D descriptors, for an optimal subset **d**, with $d << D$, and with minimal standard deviation S,

$$S = \sqrt{\frac{1}{(N-d-1)} \sum_{i=1}^{N} {res_i}^2} \qquad (1)$$

by means of the Multivariate Linear Regression (MLR) technique. In this equation N is the number of molecules in the training set, and res_i, the residual for molecule i, is the difference between the experimental property ($\mathbf{p}$) and predicted property ($\mathbf{p_{pred}}$). More precisely, we want to obtain the global minimum of $S(\mathbf{d})$ where $\mathbf{d}$ is a point in a space of size $D!/[d!(D - d)!]$. A full search (FS) of optimal variables is impractical because it requires $D!/[d!(D - d)!]$ linear regressions. Therefore, an alternative method is necessary. We selected the optimum set of descriptors using a new advanced version of the Enhanced Replacement Method (ERM) [30,31] as a search algorithm that produces linear regression QSAR models with results similar to the FS, but with much less computational work. This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1,X_2,\ldots, X_d\}$. The ERM [32] gives models with better statistical parameters than the Forward Stepwise Regression procedure [33], and the more elaborated Genetic Algorithms [34].

Among several other approaches to address this problem, the principle component regression (PCR) and partial least squares (PLS) analyses provide highly predictive QSAR, however they are difficult to understand and interpret for being abstract. A combination of GA and MLR has shown to produce simple, less sophisticated models with better performance on external testing set predictions than PLS [35]. In addition, on an extensive contrast work ERM has shown to further improve the performance of the obtained models when compared to GA [34]. Since ERM provides the same type of models in terms of simplicity compared to GA, ERM was selected for this work.

For the theoretical validation of all models, we chose the well-known Leave-One-Out (loo) and the Leave-More-Out Cross-Validation procedures ($l - n\% - o$) [36], where $n\%$ accounts for the number of molecules removed from the training set. We generated 1,000,000 cases of random data removal for 20 molecules in the case of Leave-More-Out. In our calculations we used the computational environment Matlab 5.0 (MathWorks, Natick, Massachusetts, U.S.A.). The predictive ability of the model was further evaluated by $(r^2 - r^2_0) / r^2$, $(r^2 - r'^2_0) / r^2$, k and k′ [37,38].

The applicability domain (AD) for the QSAR models was explored in order to obtain a reliable prediction for external samples. The AD is a theoretical region in the chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors [39]. The AD can be characterized in various ways such as the leverage approach [40], which allows verifying whether a new chemical can be considered as interpolated and with reduced uncertainty or extrapolated outside the domain. If it is outside the model domain, a warning must be given. The leverage (h) is defined as [40]:

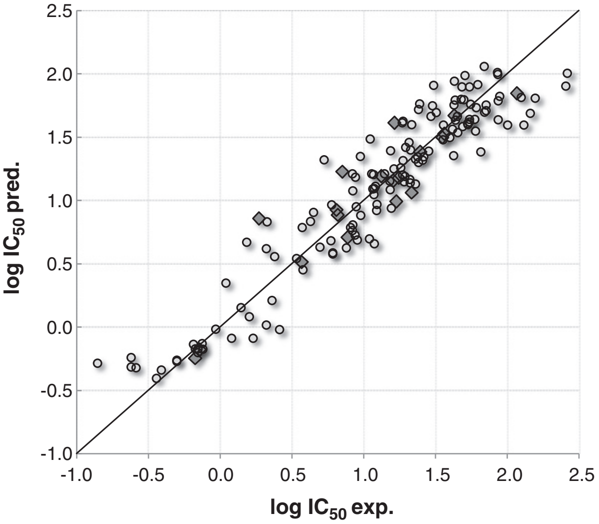


Fig. 1. Predicted (Eq. (4)) vs experimental log IC_{50} for the training (circles) and test (rhombus) sets.

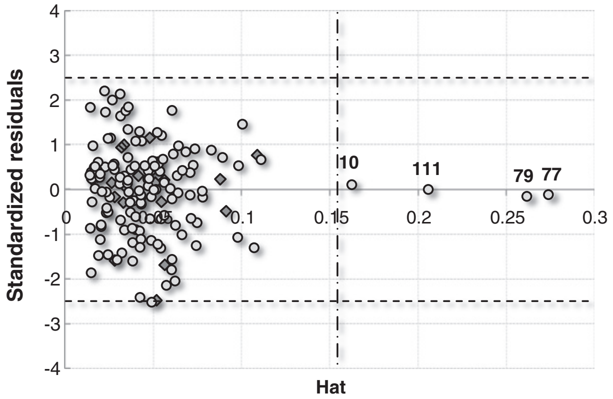


Fig. 2. Williams plot of the Eq. (4) showing the Application Domain. The vertical dashed line indicates the limiting leverage h*.

$$h_i = x_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}x_i^T \quad (i = 1,\ldots,M) \tag{2}$$

where x_i is the 1 × d descriptor row-vector of compound i, M is the number of compounds in the dataset, and $\mathbf{X}$ is the $N \times d$ matrix of the training set (d is the number of model descriptors, and N is the number of training set samples). The leverage is suitable for evaluating the degree of extrapolation, its limit of normal values is set as $h^* = 3(N + 1) / M = 3(\Sigma h_i + 1)/M$, and a leverage greater than h^* for the training set means that the chemical is highly influential in determining the model, while for the test set, it means that the prediction is the result of substantial extrapolation of the model and may not be reliable.

The standardized residual (σ) for molecule i is defined as:

$$\sigma_i = \frac{\text{res}_i}{S_{tr}} \tag{3}$$

where res_i is the residual of molecule i and S_{tr} is the standard deviation of the training set.

In order to visualize the AD of a QSAR model a Williams plot of standardized residuals (σ) vs leverage values (h) can be used to obtain an immediate and simple graphical detection of both the response outliers (Y outliers) and the structurally influential chemicals (X outliers) of a model.

Table 2
Symbols for molecular descriptors involved in the model.

Molecular descriptor	Type	Description
nR10	Constitutional	Number of 10-membered rings
Mor32p	3D-MoRSE	3D-MoRSE — signal 32/ weighted by atomic polarizabilities
SEigm	Topological	Eigenvalue sum from mass weighted distance matrix
BELm1	BCUT	Lowest eigenvalue n. 1 of Burden matrix/ weighted by atomic masses
BEHv1	BCUT	Highest eigenvalue n. 1 of Burden matrix/ weighted by atomic van der Waals volumes
H-051	Constitutional	H attached to alpha-C
SdssC	E-state	Sum of all (═C−) E-state values

Table 3
Correlation matrix for descriptors of Eq. (4) ($N = 136$).

	nR10	*Mor32p*	*SEigm*	*BELm1*	*BEHv1*	*H-051*	*SdssC*
nR10	1	0.3696	0.5333	0.8191	0.3942	0.0712	0.3512
Mor32p		1	0.4048	0.3795	0.2597	0.2015	0.3377
SEigm			1	0.5126	0.7260	0.2795	0.6549
BELm1				1	0.6615	0.1117	0.3299
BEHv1					1	0.3844	0.4739
H-051						1	0.1587
SdssC							1

3. Results and discussion

Using the ERM we searched the total pool of $D = 1433$ descriptors and obtained an optimal model with $d = 7$ parameters (Table 2) linking the molecular structure of the compounds with their activity. The optimal QSAR model according to ERM was:

$$\begin{aligned}\mathrm{logIC}_{50} = {} & -7.4149(\pm 2) - 0.9949(\pm 0.05)nR10 + 1.0716(\pm 0.2)Mor32p \\ & -0.2291(\pm 0.04)SEigm11.8729(\pm 1)BELm1 - 3.6116(\pm 0.8)BEHv1 \\ & + 0.1777(\pm 0.02)H\text{-}051 - 0.278(\pm 0.05)SdssC\end{aligned} \tag{4}$$

$$\begin{aligned}& N = 136, R = 0.9468, S = 0.2360, FIT = 5.993, p{<}10^{-4} \\ & R_{loo} = 0.9405, S_{loo} = 0.2492, R_{l-20\%-o} = 0.9237, S_{l-20\%-o} = 0.2823 \\ & R_{TS} = 0.9250, S_{TS} = 0.2236\end{aligned}$$

here, the standard errors of the regression coefficients are given in parentheses; p is the significance of the model, FIT the Kubinyi function, *loo* and $l - 20\% - o$ stand for the Leave-One-Out and Leave-More-Out Cross Validation techniques respectively and *TS* stands for Test Set.

To demonstrate that Eq. (4) is not the result of happenstance, we resorted to a widely used approach to establish the model robustness: the so-called *y*-randomization [41]. It consists of scrambling the experimental **p** property, so that activities do not correspond to the respective compounds. After analyzing 1,000,000 cases of y-randomization, the smallest *S* value obtained in this way was 0.6256, which is larger than the one coming from the true calibration (0.2360). These results suggest that the model is robust, that its calibration is not a fortuitous correlation, and that we have derived a reliable structure–activity relationship.

The plot of values predicted by Eq. (4) vs. experimental log IC_{50} shown in Fig. 1 suggests that the 136 compounds from the training set and 19 from the test set follow a straight line. The predicted activity given by Eq. (4) for the training and test sets are shown in Table 1. The Williams plot of the standardized residual in terms of the leverages illustrated in Fig. 2 shows that most compounds lie within the AD of Eq. (4) and were calculated correctly. Compounds **10**, **111**, **77** and **79** are X outliers of the training set reinforcing the model [40]; there are no compounds with a standardized residual higher than the limit (2.5σ) that can be considered outliers.

The correlation matrix shown in Table 3 reveals that the descriptors *nR10* and *BeLm1* show a relevant degree of inter-correlation, however the calibration and validation results indicate that they are important for the prediction of the activity.

The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion and/or exclusion of compounds, measured by the statistical parameters $R_{loo} = 0.9405\left(R_{loo}{}^{2} = 0.8845\right)$ and $R_{l-20\%-o} = 0.9237\left(R_{l-20\%-o}{}^{2} = 0.8532\right)$. As general rule $R_{l\ -\ n\%\ -\ o}(\mathrm{Q})$ should be higher than 0.71 ($\mathrm{Q}^2 > 0.5$) to have a validated model [38,42].

The model was further validated by the following conditions [37,38]: $R_{TS}{}^{2} = 0.8556{>}0.6$; $\mathrm{k} = 0.9781$; $\mathrm{k}' = 0.9875$ ($0.85 < \mathrm{k}$ or $\mathrm{k}' < 1.15$); $(\mathrm{r}^2 - \mathrm{r}^2{}_0) / \mathrm{r}^2 = -0.1662 < 0.1$; $(\mathrm{r}^2 - \mathrm{r}'^2{}_0) / \mathrm{r}^2 = -0.1681 < 0.1$.

The molecular descriptors appearing in the linear Eq. (4) combine two- and three-dimensional aspects of the molecular structure. The standardization of their regression coefficients of Eq. (4) allows assigning greater importance to the molecular descriptors that exhibit the largest absolute standardized coefficients [33]. In this case we have,

$$\begin{aligned}& nR10(1.1191){>}BELm1(0.7532){>}SEigm(0.3197){>}H\text{-}051(0.3184) \\ & \quad {>}BEHv1(0.2820){>}Mor32p(0.2209){>}SdssC(0.1909).\end{aligned} \tag{5}$$

By looking at this order we can see that the most significant descriptor is the constitutional descriptor *nR10*, followed by the BCUT descriptor *BELm1*. The Parameter Client [27] provides e-Dragon, ETState, GSFRAG and GSFRAG-L descriptors; however the only descriptor from Eq. (4) that was not calculated using e-Dragon was the ETState descriptor *SdssC*, having the lowest importance in the model. This suggests that e-Dragon descriptors are superior to the rest for calculating the studied activity.

Constitutional descriptors are 0D-descriptors, independent from molecular connectivity and conformations [28]. The descriptor *nR10* is determined by counting the number of 10-member rings present in a molecule, this descriptor has a positive contribution to the cytotoxicity activity since as it increases the log IC_{50} decreases, please refer to Eq. (4). The descriptor *H-051* is determined by counting the number hydrogen linked to an α-carbon, this descriptor has a negative contribution to the activity.

BCUT descriptors are the eigenvalues of a modified connectivity matrix, the Burden matrix (**B**) [43,44]. The matrix is an H depleted molecular graph, defined as follows: diagonal elements are atomic numbers of the elements (Z_i); off diagonal elements (B_{ij}), representing bonded atoms *i* and *j* are equal to $\pi^* \cdot 10^{-1}$ where π^* is the conventional bond order (i.e. 1, 2, 3, 1.5 for single, double, triple and aromatic bonds respectively); off diagonal elements corresponding to terminal bonds are increased by 0.01 and all other matrix elements are set to 0.001. The ordered sequence of the **n** smallest eigenvalues of **B** was proposed as a molecular descriptor based on the assumption that the lowest eigenvalues contain contributions from all the atoms and thus reflects topology of the molecule. The BCUT descriptors are an extension of the Burden eigenvalues and consider 3 classes of matrices, whose diagonal elements account for atomic charge related values, atomic polarizability related values and atomic H bond abilities. A variety of definitions have been used for the off diagonal terms and both 2D and 3D approaches are considered. The highest and lowest eigenvalues of these matrices have been shown to be discriminating descriptors. *BELm1* is the lowest eigenvalue of **B** involving atomic masses as weighting scheme (presenting a negative contribution to the cytotoxicity activity), and *BEHv1* is the highest eigenvalue of **B** involving atomic van der Waals volumes as weighting scheme (showing a positive contribution to the activity).

Topological descriptors are derived from hydrogen-depleted molecular graphs, in which the atoms are represented by vertices and the bonds by edges. The connections between the atoms can be described by various types of topological matrices, which can be mathematically manipulated so as to derive a single number, usually known as graph invariant. *SEigm* is calculated as the eigenvalue sum from the mass weighted distance matrix [28], and this descriptor has a positive contribution to the activity.

The 3D-MoRSE (3D Molecule Representation of Structure based on Electron diffraction) descriptors provide 3D information from the three-dimensional structure of a molecule using a molecular transform derived from an equation used in electron diffraction studies. Several atomic properties can be taken into account, thus giving high flexibility to this representation of a molecule. The simplified form of the transform is:

$$\begin{aligned}& I(s) = \sum\nolimits_{i=2}^{N}\sum\nolimits_{j=1}^{i-1} A_i A_j \frac{\sin sr_{ij}}{sr_{ij}} \\ & \qquad s = 0, \ldots, 31.0 \quad \AA^{-1}\end{aligned} \tag{6}$$

where N is the number of atoms; r_{ij} is the distance between atoms *i* and *j*; A_i can be any atomic property of atom *i* such as atomic number, mass, partial atomic charge, or atomic polarizability; *s* is a reciprocal distance. The value of *s* was considered only at discrete positions within a certain range. Normally 32 equidistant values between 0 and 31 Å^{-1} were chosen. The choice of the range of *s* and the number of values to be considered determined the resolution of the code for representing the 3D structure [45,46]. For the case of *Mor32p*, an atomic polarizability weighting scheme was used and *s* was equal to 31 Å^{-1}, this descriptor has a negative contribution to the cytotoxicity activity.

Finally, *SdssC* belongs to the electrotopological state (E-state) descriptors which are formulated as an intrinsic value plus a perturbation term of a skeletal atom, arising from the electronic interaction within the molecular topological environment of each atom in the molecule [47],and this descriptor presents a positive contribution to the cytotoxicity activity.

4. Conclusion

In this paper we constructed a predictive QSAR model of the inhibitory anticancer HT-29 human colon adenocarcinoma cell line activity of 155 chalcones using seven molecular descriptors that take into account 2D- and 3D-aspects of the molecular structure. The model exhibited very good predictive ability established by the theoretical and test set validations. The analysis of the model suggests that the descriptors with highest importance on the activity depend on the number of 10-member rings and a BCUT descriptor weighted by the van der Waals volume. We expect the proposed model to be a useful tool in the prediction of this anticancer activity, in a fast and costless manner, for any future studies that may require an estimation of this important activity of chalcones, such as determination of candidates for synthesis.

Acknowledgments

The authors thank the National Research Council of Argentina (CONICET) and INIFTA (CONICET, UNLP) for financial support.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA Cancer J. Clin. 61 (2011) 69–90.

[2] D.C. Hiss, G.A. Gabriels, Implications of endoplasmic reticulum stress, the unfolded protein response and apoptosis for molecular cancer therapy. Part I: targeting p53, Mdm2, GADD153/CHOP, GRP78/BiP and heat shock proteins, Expert Opin. Drug Discov. 4 (2009) 799–821.

[3] S. Syam, S.I. Abdelwahab, M.A. Al-Mamary, S. Mohan, Synthesis of chalcones with anticancer activities, Molecules 17 (2012) 6179–6195.

[4] S. Shenvi, K. Kumar, K.S. Hatti, K. Rijesh, L. Diwakar, G.C. Reddy, Synthesis, anticancer and antioxidant activities of 2,4,5-trimethoxy chalcones and analogues from asaronaldehyde: structure–activity relationship, Eur. J. Med. Chem. 62 (2013) 435–442.

[5] Y.H. Kim, J. Kim, H. Park, H.P. Kim, Anti-inflammatory activity of the synthetic chalcone derivatives: inhibition of inducible nitric oxide synthase-catalyzed nitric oxide production from lipopolysaccharide-treated RAW 264.7 cells, Biol. Pharm. Bull. 30 (2007) 1450–1455.

[6] Z.a. Nowakowska, A review of anti-infective and anti-inflammatory chalcones, Eur. J. Med. Chem. 42 (2007) 125–137.

[7] M. Chen, T.G. Theander, S.B. Christensen, L. Hviid, L. Zhai, A. Kharazmi, Licochalcone A, a new antimalarial agent, inhibits in vitro growth of the human malaria parasite *Plasmodium falciparum* and protects mice from *P. yoelii* infection, Antimicrob. Agents Chemother. 38 (1994) 1470–1475.

[8] F. Hayat, E. Moseley, A. Salahuddin, R.L. Van Zyl, A. Azam, Antiprotozoal activity of chloroquinoline based chalcones, Eur. J. Med. Chem. 46 (2011) 1897–1905.

[9] M.R. Ahmed, V.G. Sastry, Synthesis and cytotoxic, anti oxidant activities of new chalcone derivatives, Rasayan J. Chem. 4 (2011) 289–294.

[10] B. Orlikova, D. Tasdemir, Dietary chalcones with chemopreventive and chemotherapeutic potential, Genes Nutr. 6 (2011) 125–147.

[11] M.A. Rahman, Chalcone: a valuable insight into the recent advances and potential pharmacological activities, Chem. Sci. J. (2011) 1–16.

[12] R.-Z. Lin, H.-Y. Chang, Recent advances in three-dimensional multicellular spheroid culture for biomedical research, Biotechnol. J. 3 (2008) 1172–1184.

[13] J. Wu, C. Wang, Y. Cai, J. Peng, D. Liang, Y. Zhao, S. Yang, X. Li, X. Wu, G. Liang, Synthesis and crystal structure of chalcones as well as on cytotoxicity and antibacterial properties, Med. Chem. Res. 21 (2012) 444–452.

[14] J.-H. Cheng, C.-F. Hung, S.-C. Yang, J.-P. Wang, S.-J. Won, C.-N. Lin, Synthesis and cytotoxic, anti-inflammatory, and anti-oxidant activities of 2â€^{2},5â€^{2}-dialkoxylchalcones as cancer chemopreventive agents, Bioorg. Med. Chem. 16 (2008) 7270–7276.

[15] A. Kamal, A. Mallareddy, P. Suresh, T.B. Shaik, V. Lakshma Nayak, C. Kishor, R.V. Shetti, N. Sankara Rao, J.R. Tamboli, S. Ramakrishna, A. Addlagatta, Synthesis of chalcone-amidobenzothiazole conjugates as antimitotic and apoptotic inducing agents, Bioorg. Med. Chem. 20 (2012) 3480–3492.

[16] M. Nagaraju, E. Gnana Deepthi, C. Ashwini, M.V. Vishnuvardhan, V. Lakshma Nayak, R. Chandra, S. Ramakrishna, B.B. Gawali, Synthesis and selective cytotoxic activity of novel hybrid chalcones against prostate cancer cells, Bioorg. Med. Chem. Lett. 22 (2012) 4314–4317.

[17] S.C. Fang, C.L. Hsu, Y.S. Yu, G.C. Yen, Cytotoxic effects of new geranyl chalcone derivatives isolated from the leaves of *Artocarpus communis* in SW 872 human liposarcoma cells, J. Agric. Food Chem. 56 (2008) 8859–8868.

[18] B.-L. Wei, C.-H. Teng, J.-P. Wang, S.-J. Won, C.-N. Lin, Synthetic 2′,5′-dimethoxychalcones as G2/M arrest-mediated apoptosis-inducing agents and inhibitors of nitric oxide production in rat macrophages, Eur. J. Med. Chem. 42 (2007) 660–668.

[19] A. Kamal, A. Mallareddy, P. Suresh, V. Lakshma Nayak, R.V. Shetti, N. Sankara Rao, J.R. Tamboli, T.B. Shaik, M.V. Vishnuvardhan, S. Ramakrishna, Synthesis and anticancer activity of 4β-alkylamidochalcone and 4β-cinnamido linked podophyllotoxins as apoptotic inducing agents, Eur. J. Med. Chem. 47 (2012) 530–545.

[20] G. Wang, W. Wu, F. Peng, D. Cao, Z. Yang, L. Ma, N. Qiu, H. Ye, X. Han, J. Chen, J. Qiu, Y. Sang, X. Liang, Y. Ran, A. Peng, Y. Wei, L. Chen, Design, synthesis, and structure–activity relationship studies of novel millepachine derivatives as potent antiproliferative agents, Eur. J. Med. Chem. 54 (2012) 793–803.

[21] L. Xie, X. Zhai, L. Ren, H. Meng, C. Liu, W. Zhu, Y. Zhao, Design, synthesis and antitumor activity of novel artemisinin derivatives using hybrid approach, Chem. Pharm. Bull. (Tokyo) 59 (2011) 984–990.

[22] J.C. Aponte, M. Verastegui, E. Malaga, M. Zimic, M. Quiliano, A.J. Vaisberg, R.H. Gilman, G.B. Hammond, Synthesis, cytotoxicity, and anti-*Trypanosoma cruzi* activity of new chalcones, J. Med. Chem. 51 (2008) 6230–6234.

[23] T. Mosmann, Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays, J. Immunol. Methods 65 (1983) 55–63.

[24] H. Liu, P. Gramatica, QSAR study of selective ligands for the thyroid hormone receptor beta, Bioorg. Med. Chem. 15 (2007) 5251–5261.

[25] HYPERCHEM, in, 6.03 (Hypercube), http://www.hyper.com.

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

[27] VCCLAB, Virtual Computational Chemistry Laboratory, 2005.

[28] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley VCH, Weinheim, Germany, 2000.

[29] The Open Babel Package, version 2.2.3 http://openbabel.sourceforge.net/2006.

[30] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Advances in the replacement and enhanced replacement method in QSAR and QSPR theories, J. Chem. Inf. Model. 51 (2011) 1575–1581.

[31] A. Lee, A.G. Mercader, P.R. Duchowicz, E.A. Castro, A.B. Pomilio, QSAR study of the DPPH radical scavenging activity of di(hetero)arylamines derivatives of benzo[b] thiophenes, halophenols and caffeic acid analogues, Chemom. Intell. Lab. Syst. 116 (2012) 33–40.

[32] A.G. Mercader, P.R. Duchowicz, F.M. Fernandez, E.A. Castro, Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories, Chemom. Intell. Lab. Syst. 92 (2008) 138–144.

[33] N.R. Draper, H. Smith, Applied regression analysis, John Wiley & Sons, New York, 1981.

[34] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Replacement method and enhanced replacement method *versus* the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories, J. Chem. Inf. Model. 50 (2010) 1542–1548.

[35] A.K. Saxena, P. Prathipati, Comparison of MLR, PLS and GA-MLR in QSAR analysis, SAR QSAR Environ. Res. 14 (2003) 433–445.

[36] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579–586.

[37] V. Ravichandran, S. Shalini, K. Sundram, A.D. Sokkalingam, QSAR study of substituted 1,3,4-oxadiazole naphthyridines as HIV-1 integrase inhibitors, Eur. J. Med. Chem. 45 (2010) 2791–2797.

[38] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, Expert Opin. Drug Discov. 2 (2007) 1567–1577.

[39] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.

[40] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.

[41] S. Wold, L. Eriksson, Statistical validation of QSAR results, in: H.v.d. Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309–318.

[42] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Model. 20 (2002) 269–276.

[43] F.R. Burden, Molecular identification number for substructure searches, J. Chem. Inf. Comput. Sci. 29 (1989) 225–227.

[44] R.B. Frank, A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix, Quant. Struct.-Act. Relat. 16 (1997) 309–314.

[45] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, Chemical information in 3D space, J. Chem. Inf. Comput. Sci. 36 (1996) 1030–1037.

[46] J.H. Schuur, P. Selzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.

[47] L.B. Kier, L.H. Hall, An electrotopological-state index for atoms in molecules, Pharm. Res. 7 (1990) 801–807.