Chemometrics and Intelligent Laboratory Systems 146 (2015) 378–384

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Chalcone derivative cytotoxicity activity against MCF-7 human breast cancer cell QSAR study

Dušan Dimić [a], Andrew G. Mercader [b,*], Eduardo A. Castro [b]

[a] Faculty of Physical Chemistry, University of Belgrade, Studentski trg 12-16, 11581 Belgrade, Serbia
[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 4 May 2015
Received in revised form 17 June 2015
Accepted 19 June 2015
Available online 23 June 2015

Keywords:
Chalcones
Anticancer
Cytotoxicity
MCF-7 breast cancer cells
QSAR
Enhanced Replacement Method

ABSTRACT

Chalcones and their derivatives possess a wide range of significant pharmacological activities; among the most important ones is their anticancer activity. For this reason we performed a Quantitative Structure–Activity Relationships (QSAR) study of their anticancer activity against MCF-7 human breast cancer cell lines. In this work, several descriptor options were tested on the dataset containing 93 molecular structures, using ERM (Enhanced Replacement Method). The best models were found using merely two dimensional descriptors. The two dimensional descriptor pool was further expanded using several nonlinear transformations, which resulted in an optimal five molecular descriptor model that showed very good predictive ability. Thus, ERM was capable of finding a simple to interpret and understand model that nonetheless addresses nonlinearities between the descriptors and the activity. Furthermore, the acquired model is very straightforward to use since it does not require the optimization of chemical structures for the calculation of three dimensional descriptors.

1. Introduction

In a recent report of the American Cancer Society breast cancer was classified as the most frequently diagnosed cancer and the leading cause of cancer death among females [1,2]. Although there are many therapeutic strategies, more than one million new cases of breast cancer are diagnosed every year [3]. Some of the main problems in the treatment of cancer are high cytotoxicity and drug resistance, and no clinically active substances are known to act selectively on tumor cells. Scientists put great effort in finding new therapeutic targets for cancer and to the development of new selective drugs [4].

Chalcones, 1,3-diphenyl-2-propene-1-ones (Fig. 1), present a very interesting group of molecules from pharmaceutical point of view, because of their significant cytotoxic activity against various cancer cells. Besides this important feature chalcones and their analogues display considerable chemoprotective, antiangiogenic, antibacterial, antifugal, antiparasitic, antioxidative, anti-inflammatory activity [4,5] and antimarial activity [6,7]. Chalcones are also precursors for synthesis of flavonoids, which are also known for their anticancer potential [8].

The variations of the structure of chalcones offered a new field of research because of all of the positive effects mentioned above. Many of them were tested against series of cancer cell lines [9]. In order to present the variety of structural differences some of the previously obtained experimental studies on human breast cancer cells (MCF7) are presented. Liu and Go showed that for the large number of chalcones with basic functionalities activity against MCF7 was dependent on the polar volume, hydrogen bonding features, HOMO energies and charge on the β carbon. It was discussed that the change in structure influenced the mechanism of action, so the compounds with basic groups on ring A interfered with cell cycle progression, but compounds with two basics groups had no effect [10]. The same authors synthesized metoxylated chalcones with N-methyl substituents on ring A and showed inhibition of the growth of the MCF7 cell lines, because piperidinyl group adds specificity to the mechanism of activity and also changes physiochemical properties of the whole system [11]. In a study by Ivkovic et al [4] some of the basic modifications of the chalcone structure were introduced, for example electron withdrawing groups (–F, –Cl and $-CF_3$) in the *ortho* and *para* position in the benzyl moiety and their antiproliferative activity was investigated, and structure–activity analysis was performed in order to determine the best molecular descriptors for this class of chalcones.

Shenavi and coworkers obtained interesting class of 2,4,5-trimethoxy chalcones and their analogues from asaronaldehyde. These compounds were tested against the same cell lines and revealed that chalcones with groups that are good electron donors in para position to carbonyl moiety of phenyl ring A had better results. Some of these molecules showed significant nitric oxide scavenging activity [12].

In this paper we examine quantitative structure–activity relationships (QSAR) of chalcones against human breast cancer cells (MCF-7) based on 93 chalcone derivatives taken from the previously mentioned experimental studies, measured under the same conditions. This group

* Corresponding author. Tel.: +54 11 6091 3759; fax: +54 11 6091 2100x3759.
E-mail address: amercader@inifta.unlp.edu.ar (A.G. Mercader).

http://dx.doi.org/10.1016/j.chemolab.2015.06.011
0169-7439/

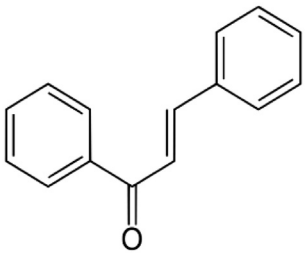

Fig. 1. Chalcone backbone structure.

of chalcones will in principle give us reliable QSAR parameters for the determination of the activity of some newly synthesized chalcone derivative towards MCF-7 cell lines.

2. Material and methods

2.1. Data sets

In our QSAR study, a total of 93 chalcone molecules were gathered from the literature [4,10–12]. To our knowledge this particular set of molecules was not employed in a QSAR study before. The experimental log IC_{50}, concentration of the compounds (μM) exhibiting 50% inhibition of cell growth for human breast cancer (MCF 7), values along with the SMILES structure representation can be found on Table 1. SMILES notation was chosen as a simple way of sharing the dataset with any interested reader, since it allows copying the text string and entering it in many chemical structure representation software that can later be used to calculate descriptor. The data-set was divided into a training set of 63 and a test set of 30 chalcones by applying a *k*-means cluster analysis [13], in order to have representative molecules of the complete dataset in both training and test sets. The basis of the *k*-means cluster analysis is to create *k* clusters or groups of molecules, in such a way that compounds in the same cluster are very similar in terms of a distance metrics and compounds in other clusters are very different; details of the procedure have been presented elsewhere [14].

The cytotoxic activity of all compounds was evaluated 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 [15,16].

2.2. Molecular descriptors

In order to calculate three dimensional descriptors, the structures of the compounds were pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in the Hyperchem 6.03 package [17], and 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 Dragon 5.0 [18] 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 [19]. In addition, descriptors from the software QuBiLs-MIDAS [20], that computes 3D molecular indices by using Quadratic, Bilinear and N-Linear Maps based on N-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings, were included. The settings used in QuBiLs-MIDAS were the following: all algebraic forms, all Matrix Forms, all Groups, and all Properties were selected; only the first Metric and the first Invariant were selected.

The descriptors were separated in three different matrices, a matrix (2D) containing the non-three dimensional descriptors from Dragon 5.0, a matrix (3D) containing the three dimensional descriptors from Dragon and a matrix (Q) with the three dimensional descriptors from QuBiLs-MIDAS. Several combinations of the matrices were tested: 2D; 3D; 2D + 3D; Q; and 2D + Q. In addition, to evaluate nonlinear dependencies the matrix 2D was expanded using the following transformations: ln(x); e^x; x^2; x^3; x^4; x^{-1}; x^{-2}; x^{-3}; and $x^{1/2}$. After the transformations the resulting matrix ($2D_e$) was cleaned removing ∞ and −∞ results; and in addition descriptors with correlations greater than 0.98 were removed.

2.3. Model search

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

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

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 (**p**) and predicted property ($\mathbf{p_{pred}}$). More precisely, the aim is to obtain the global minimum of $S(\mathbf{d})$ where **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. The optimum set of descriptors was selected using a new advanced version of the Enhanced Replacement Method (ERM) [21,22] as a search algorithm that produces linear regression QSAR models with results similar to the FS, nonetheless 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 [23] gives models with better statistical parameters than the Forward Stepwise Regression procedure [24], and the more elaborated Genetic Algorithms [25].

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 [26]. In addition, on an extensive contrast work, ERM has shown to further improve the performance of the obtained models when compared to GA [25]. 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 the models, Leave-One-Out (*loo*) Cross-Validation procedures (*l-n%-o*) [27] was chosen. The computational environment Matlab 5.0 (MathWorks, Natick, Massachusetts, USA) was used for the calculations. The predictive ability of the best model was further evaluated by $(r^2 - r^2_0) / r^2$, $(r^2 - r'^2_0) / r^2$, k and k′ [28,29].

The applicability domain (AD) for the best QSAR model was explored 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, represented in each model by the specific molecular descriptors [30]. The AD can be characterized in various ways such as the leverage approach [31], which allows verifying whether a new chemical can be considered as interpolated and with reduced uncertainty or extrapolated outside the domain. When outside the model domain, a warning must be given. The leverage (*h*) is defined as [31]:

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

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

	SMILES	logIC_{50} exp.	logIC_{50} pred.	Res.	Ref.
1	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)OC)C1CCN(CC1)C	0.4472	0.4970	−0.0499	[10]
2	c1c(c(cc(c1C(=O)/C=C/c1ccccc1)OC)OC)C1CCN(CC1)C	0.5315	0.7299	−0.1984	[10]
3	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)OC)C1CCN(CC1)C	0.3979	0.5499	−0.1520	[10]
4	c1(cc(c(cc1C(=O)/C=C/c1ccncc1)C1CCN(CC1)C)OC)OC	0.8573	0.7103	0.1470	[10]
5	c1(cc(c(cc1C(=O)/C=C/c1cnccc1)C1CCN(CC1)C)OC)OC	0.6335	0.6410	−0.0075	[10]
6	c1(cc(c(cc1C(=O)/C=C/c1ncccc1)C1CCN(CC1)C)OC)OC	0.5911	0.7045	−0.1134	[10]
7	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)N1CCN(CC1)C)OC)OC)C1CCN(CC1)C	0.9294	0.9984	−0.0690	[10]
8	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)C)OC)OC)C1CCN(CC1)C	0.8195	0.7724	0.0471	[10]
9	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)C)OC)OC)C1CCN(CC1)C	0.7324	0.8550	−0.1226	[10]
10	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)OC)OC)OC)C1CCN(CC1)C	0.8451	0.8086	0.0365	[10]
11	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)OC)OC)OC)C1CCN(CC1)C	0.8451	0.7435	0.1016	[10]
12	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)F)OC)OC)C1CCN(CC1)C	0.5798	0.5700	0.0098	[10]
13	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)F)OC)OC)C1CCN(CC1)C	0.5911	0.4702	0.1208	[10]
14	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)C(F)(F)F)OC)OC)C1CCN(CC1)C	0.4472	0.8667	−0.4196	[10]
15	c1(c(c(cc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)OC)C1CCN(CC1)C)O	0.4914	0.5047	−0.0133	[10]
16	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)OC)C1CCN(CC1)C)O	0.8261	0.6914	0.1347	[10]
17	c1(c(c(cc(c1C(=O)/C=C/c1ccccc1)OC)OC)C1CCN(CC1)C)O	0.7559	0.8474	−0.0915	[10]
18	c1(c(c(cc(c1C(=O)/C=C/c1c(cccc1)F)OC)OC)C1CCN(CC1)C)O	0.5315	0.6116	−0.0801	[10]
19	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)F)OC)OC)C1CCN(CC1)C)O	0.8062	0.6183	0.1878	[10]
20	c1(c(c(cc(c1C(=O)/C=C/c1c(cccc1)C)OC)OC)C1CCN(CC1)C)O	1.3483	1.0718	0.2765	[10]
21	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)C)OC)OC)C1CCN(CC1)C)O	0.8195	1.0269	−0.2074	[10]
22	c1(c(c(cc(c1C(=O)/C=C/c1c(cccc1)OC)OC)OC)C1CCN(CC1)C)O	1.0043	0.9852	0.0191	[10]
23	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)OC)OC)OC)C1CCN(CC1)C)O	1.4065	0.9458	0.4607	[10]
24	c1(c(c(cc(c1C(=O)/C=C/c1cc(ccc1)OC)OC)OC)C1CCN(CC1)C)O	1.4942	1.3218	0.1724	[10]
25	c1(c(c(cc(c1C(=O)/C=C/c1c(cc(cc1)OC)OC)OC)OC)C1CCN(CC1)C)O	1.0828	1.0800	0.0028	[10]
26	c1(c(c(cc(c1C(=O)/C=C/c1cc(ccc1)Cl)OC)OC)C1CCN(CC1)C)O	0.4314	0.7491	−0.3178	[10]
27	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)C#N)OC)OC)C1CCN(CC1)C)O	1.1399	1.0945	0.0454	[10]
28	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)C(F)(F)F)OC)OC)C1CCN(CC1)C)O	1.1553	1.0053	0.1500	[10]
29	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)N1CCN(CC1)C)OC)OC)C1CCN(CC1)C)O	1.1584	1.1646	−0.0062	[10]
30	c1(c(c(cc(c1C(=O)/C=C/c1ccc(cc1)N(C)C)OC)OC)C1CCN(CC1)C)O	1.1430	0.9730	0.1701	[10]
31	c1cc(ccc1C(=O)/C=C/c1c(cc(cc1)OC)OC)N1CCNCC1	1.2122	1.2979	−0.0857	[10]
32	c1cc(ccc1C(=O)/C=C/c1c(cccc1)Cl)N1CCNCC1	1.3304	1.3178	0.0126	[10]
33	c1cc(ccc1C(=O)/C=C/c1ccc(cc1)Cl)N1CCNCC1	0.9294	1.1719	−0.2425	[10]
34	c1cc(ccc1C(=O)/C=C/c1ccccc1)N1CCNCC1	1.6990	1.4957	0.2033	[10]
35	c1cc(ccc1C(=O)/C=C/c1c(cccc1)Cl)N1CCCCC1	1.6758	1.3553	0.3205	[10]
36	c1cc(ccc1C(=O)/C=C/c1ccc(cc1)Cl)N1CCCCC1	1.3997	1.1980	0.2016	[10]
37	c1cc(ccc1C(=O)/C=C/c1ccccc1)N1CCCCC1	1.6990	1.5847	0.1142	[10]
38	c1cc(ccc1C(=O)/C=C/c1c(cc(cc1)OC)OC)N1CCCCC1	1.6990	1.3003	0.3987	[10]
39	c1cc(cc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)OC	0.8129	0.9558	−0.1429	[10]
40	c1cc(cc(c1C(=O)/C=C/c1ccncc1)OC)OC	0.8976	1.1305	−0.2328	[10]
41	c1cc(cc(c1C(=O)/C=C/c1cccnc1)OC)OC	1.1335	1.0612	0.0723	[10]
42	c1cc(cc(c1C(=O)/C=C/c1ccccn1)OC)OC	0.8062	1.1305	−0.3243	[10]
43	c1cc(cc(c1C(=O)/C=C/c1ccc(cc1)N1CCN(CC1)C)OC)OC	1.1367	1.2399	−0.1031	[10]
44	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)OC)C1CCN(CC1)c1ccccc1	0.4150	0.4932	−0.0782	[10]
45	c1c(c(cc(c1C(=O)/C=C/c1ccccc1)OC)OC)C1CCN(CC1)c1ccccc1	0.6812	0.7289	−0.0477	[10]
46	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)OC)C1CCN(CC1)c1ccccc1	0.5315	0.5334	−0.0019	[10]
47	c1c(c(cc(c1C(=O)/C=C/c1ccncc1)OC)OC)C1CCN(CC1)c1ccccc1	0.4624	0.7012	−0.2388	[10]
48	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)N1CCCN(CC1)C)OC)OC)C1CCN(CC1)c1ccccc1	0.9445	0.9876	−0.0431	[10]
49	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)OC)N1CCN(CC1)C	0.7924	0.5743	0.2181	[10]
50	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)F)OC)OC)N1CCN(CC1)C	0.8573	0.6346	0.2228	[10]
51	c1c(c(cc(c1C(=O)/C=C/c1ccccc1)OC)OC)N1CCN(CC1)C	0.7782	0.8115	−0.0333	[10]
52	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)N1CCN(CC1)C)OC)OC)N1CCN(CC1)C	1.1399	1.0734	0.0665	[10]
53	c1c(c(cc(c1C(=O)/C=C/c1ccncc1)OC)OC)N1CCN(CC1)C	1.0934	0.7959	0.2975	[10]
54	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)OC)N1CC[C@H](CC1)N1CCCCC1	0.5682	0.7798	−0.2116	[10]
55	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)F)OC)OC)N1CC[C@H](CC1)N1CCCCC1	0.5441	0.8488	−0.3048	[10]
56	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)F)O)OC)OC	1.3503	1.1296	0.2206	[10]
57	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)OC)O)OC)OC	1.0719	1.4141	−0.3422	[10]
58	c1c(c(cc(c1C(=O)/C=C/c1ccccc1)O)OC)OC	1.4065	1.2997	0.1068	[10]
59	c1c(ccc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)C1CCN(CC1)C	0.4624	0.6006	−0.1382	[10]
60	c1c(ccc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)C1CCN(CC1)C	0.5052	0.6216	−0.1165	[10]
61	c1c(ccc(c1C(=O)/C=C/c1ccccc1)OC)C1CCN(CC1)C	0.6233	0.8498	−0.2265	[10]
62	c1(cc(c(c(c1C(=O)/C=C/c1ccccc1Cl)O)C1CCN(CC1)C)OC)OC	0.4914	0.5047	−0.0133	[11]
63	c1(cc(c(c(c1C(=O)/C=C/c1ccc(cc1)Cl)O)C1CCN(CC1)C)OC)OC	0.8261	0.6914	0.1347	[11]
64	c1(cc(c(c(c1C(=O)/C=C/c1ccccc1)O)C1CCN(CC1)C)OC)OC	0.7559	0.8474	−0.0915	[11]
65	c1(cc(c(c(c1C(=O)/C=C/c1c(cccc1)F)O)C1CCN(CC1)C)OC)OC	0.5315	0.6116	−0.0801	[11]
66	c1c(c(cc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)OC)C1CCN(CC1)C	0.4314	0.4970	−0.0657	[11]
67	c1c(c(cc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)OC)C1CCN(CC1)C	0.3979	0.5499	−0.1520	[11]
68	c1c(c(cc(c1C(=O)/C=C/c1ccccc1)OC)OC)C1CCN(CC1)C	0.5315	0.7299	−0.1984	[11]
69	c1c(ccc(c1C(=O)/C=C/c1c(cccc1)Cl)OC)C1CCN(CC1)C	0.4624	0.6006	−0.1382	[11]
70	c1c(ccc(c1C(=O)/C=C/c1ccc(cc1)Cl)OC)C1CCN(CC1)C	0.5052	0.6216	−0.1165	[11]
71	c1c(ccc(c1C(=O)/C=C/c1ccccc1)OC)C1CCN(CC1)C	0.6233	0.8498	−0.2265	[11]
72	c1c(cccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)N(O)O	2.1136	1.8674	0.2462	[12]
73	c1cc(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)Cl	1.5628	1.4262	0.1366	[12]
74	c1cc(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)OC	1.5757	1.6014	−0.0257	[12]
75	c1cc(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)C	1.6249	1.5348	0.0901	[12]

Table 1 (*continued*)

	SMILES	$logIC_{50}$ exp.	$logIC_{50}$ pred.	Res.	Ref.
76	c1(cc(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)O)O	1.5757	1.8240	−0.2484	[12]
77	c1cc(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)F	1.4564	1.3054	0.1510	[12]
78	c1c(c(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)Cl)N(O)O	1.2170	1.7343	−0.5173	[12]
79	c1(cc(cc(c1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)OC)OC)OC	2.0338	1.9244	0.1093	[12]
80	c1cc(ccc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)Br	1.5007	1.3608	0.1398	[12]
81	c1c(cc(cc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)Cc1ccccc1)Cc1ccccc1	1.4296	1.4869	−0.0573	[12]
82	O(C(=O)C)c1ccc(cc1C(=O)/C=C/c1c(cc(c(c1)OC)OC)OC)OC(=O)C	1.7627	1.7601	0.0026	[12]
83	c1cccc(c1C(=O)CCc1c(cccc1)C(F)(F)F)O	1.2279	1.4249	−0.1970	[4]
84	c1cccc(c1C(=O)CCc1c(cccc1)F)O	1.1584	1.3548	−0.1964	[4]
85	c1cccc(c1C(=O)CCc1ccc(cc1)F)O	1.4150	1.1911	0.2239	[4]
86	c1cccc(c1C(=O)CCc1ccccc1C)O	1.5763	1.5378	0.0386	[4]
87	c1cccc(c1C(=O)CCc1ccc(cc1)C)O	1.5302	1.6778	−0.1476	[4]
88	c1cccc(c1C(=O)CCc1c(cccc1)Cl)O	1.4742	1.2864	0.1879	[4]
89	c1cc(c(cc1)OC[C@@H](CNCCC)O)C(=O)CCc1c(cccc1)F	1.2672	1.3142	−0.0470	[4]
90	c1cc(c(cc1)OC[C@@H](CNCCC)O)C(=O)CCc1ccc(cc1)F	1.5185	1.1867	0.3318	[4]
91	c1cc(c(cc1)OC[C@@H](CNCCC)O)C(=O)CCc1c(cccc1)C	1.3692	1.4658	−0.0965	[4]
92	c1cc(c(cc1)OC[C@@H](CNCCC)O)C(=O)CCc1ccc(cc1)C	1.5011	1.5527	−0.0516	[4]
93	c1cc(c(cc1)OC[C@@H](CNCCC)O)C(=O)CCc1c(cccc1)Cl	1.3054	1.2144	0.0909	[4]

where x_i is the $1 \times d$ 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 to evaluate the degree of extrapolation, its limit 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, though for the test set 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:

$$\sigma_i = \frac{res_i}{S_{tr}} \tag{3}$$

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.

3. Results and discussion

By means of ERM the different descriptor matrices were independently searched for the models that best correlate with the activity, the results were summarized in Table 2. The results indicate that:

- The best model in terms of its predicted power reflected in the test set parameters is m_1 obtained using only the two dimensional descriptors (2D).
- The best model obtained from 3D alone (m_2) contains 5 descriptors; and it is of inferior quality in all the parameters compared to m_1. The best 7 descriptor model (m_3) from 3D, added for comparison purposes, shows that it is also of inferior quality than m_1.
- The sum of 3D to 2D descriptors (2D + 3D) gave the same model ($m_4 = m_1$), hence no 3D descriptors were chosen by the descriptor selection algorithm, implying that in this case 3D descriptors are not helpful to model the activity when compared to 2D descriptors.
- The best model obtained from Q alone (m_5) contains 6 descriptors; and it is of inferior quality in all the parameters compared to m_1. The best 7 descriptor model from Q (m_6) shows better calibration parameters but much worse validation parameters compared to m_1, indicating that the model is overfitted. This might be a consequence of the great number of descriptors that Q contains; allowing an excessive fit to the training molecules, probably not by a true structure activity relationship.
- The best model from the addition of 2D to Q (m_7) has 5 descriptors and again is of inferior quality than m_1. The best 7 descriptor model from 2D + Q, shows better calibration but inferior validation parameters, indicating once more an overfitted model.
- The models derived from 3D descriptors present worst calibration but much better validation parameters than those from Q. Again this might be a consequence of the great number of Q descriptors and their problem (in this study) to excessively fit the training set, hence not reflecting a true correlation between the structure and the activity.
- The comparison of the models from Q and those from 2D + Q, reveals that the 2D descriptors help to find better models in terms of the predictive ability with better validation parameters. This further indicates that for this study, the 2D descriptors are better in finding an actual structure activity correlation.

Table 2
Number of descriptors in the matrix (D), number of descriptors in the model (d), Standard deviation (S), correlation coefficient (R), *Kubinyi function* (FIT) and *y*-randomization (loo stands for Leave-One-Out and test indicate external test set results); for the best models found exploring the different descriptors matrices (2D, 3D, 2D + 3D, Q, and 2D + Q).

	D	Model	d	S	R	FIT	S_{loo}	R_{loo}	S_{test}	R_{test}	*y*-rand.
2D	950	m_1	7	0.1595	0.9430	3.9464	0.1864	0.9216	**0.1937**	**0.8661**	0.3476
3D	696	m_2	5	0.2216	0.8824	2.2785	0.2460	0.8536	0.2083	0.8532	0.3632
		m_3	7	0.1717	0.9337	3.3391	0.1987	0.9106	0.2372	0.8337	0.3447
2D + 3D	1734	m_4	7	0.1595	0.9430	3.9464	0.1864	0.9216	**0.1937**	**0.8661**	0.3476
Q	14,178	m_5	6	0.1852	0.9173	3.0036	0.2061	0.8967	0.2355	0.8079	0.3528
		m_6	7	0.1362	0.9570	5.3411	0.1525	0.9459	0.2580	0.7745	0.3436
2D + Q	15,128	m_7	5	0.1730	0.9271	3.9633	0.1911	0.9104	0.2213	0.8290	0.3496
		m_8	7	0.1371	0.9565	5.2763	0.1500	0.9477	0.2276	0.8207	0.3551

Bold numbers indicate the best results.

Table 3
Symbols for molecular descriptors involved in Eq.(4).

Molecular descriptor	Type	Description
F08[C—O]	*2D atom pair frequency*	*Frequency of C—O at topological distance 8*
ATS8m	*2D autocorrelations*	*Broto–Moreau autocorrelation of a topological structure of lag 8 weighted by mass*
F05[O—O]	*2D atom pair frequency*	*Frequency of O—O at topological distance 5*
TI2	*Topological*	*Second Mohar index from Laplace matrix*
GATS4e	*2D autocorrelations*	*Geary autocorrelation of lag 4 weighted by Sanderson electronegativity*

From the above results it seems clear that 2D descriptors are the best option for the database under study. This might be a consequence of the fact that the 3D optimization are done to find the conformer of minimal energy which might differ from the actual *in vitro* conformer. This limitation in the methodology comes from the lack of information about the 3D disposition of the molecules when the experiments were carried away.

The results also seem to indicate that 3D descriptors from Dragon are better than those from QuBiLs-MIDAS; however this comparison might not be entirely fair since the present dataset does not appear as the best option to compare 3D descriptors. On the other hand it seems clear that the great number of descriptors form Q tends to generate models that overfit the training set.

In order to further improve the best obtained model; the 2D matrix was expanded adding several nonlinear transformations. By doing so ERM can be used to find a simple to interpret and understand model, nonetheless that will address to some extend the possible nonlinearity of the correlations between the descriptors and the activity. After the removal of descriptors that had greater dependency than 0.98 compared to any other descriptor, the expanded 2D matrix ($2D_e$) contained a pool of $D = 2115$ descriptors. Subsequently, a search of an optimal set of descriptors from those in the new matrix ($2D_e$) was done using ERM, finding an optimal model with $d = 5$ parameters (Table 3) linking the molecular structure of the compounds with their activity:

$$\begin{aligned}\log IC_{50} = {} & -4.3644(\pm 0.6) + 0.01277\left(\pm 1E^{-3}\right) F08[C\text{-}O]^2 \\ & -0.006185\left(\pm 5E^{-4}\right) ATS8m^4 - 0.01558\left(\pm 3E^{-3}\right) F05[O\text{-}O]^4 \\ & +1.5994\ (\pm 0.2) TI2^{1/2} + 2.5998(\pm 0.5) GATS4e^{1/2} \\ & N = 93,\ R = 0.9060,\ S = 0.1894,\ FIT = 3.379,\ p < 10^{-6} \\ & R_{loo} = 0.8940,\ S_{loo} = 0.2005,\ R_{TS} = 0.9333,\ S_{TS} = 0.1544. \end{aligned} \quad (4)$$

Here, the standard errors of the regression coefficients are given in parentheses; p is the significance of the model, FIT the Kubinyi function, *loo* stand for the Leave-One-Out Cross Validation techniques respectively and *TS* stands for Test Set. The presented model's coefficients were recalculated using all the available data, to include in the correlation all the available structural information.

A second model from $2D_e$ containing $d = 9$ descriptors is also presented, showing that although the number of descriptors is higher than those in m_1 the model still has very good validation parameters:

$$\begin{aligned}\log IC_{50} = {} & 0.2931(\pm 0.1) + 0.001704\left(\pm 2E^{-3}\right) GGI4^4 - 0.002030\left(\pm 3E^{-4}\right) C\text{-}002^4 \\ & -0.02001\left(\pm 7E^{-3}\right) F08[O\text{-}F]^4 - 0.002268\left(\pm 3E^{-4}\right) MATS1e^{-1} \\ & +0.9112\ (\pm 0.1) GATS6m^{1/2} + 0.4028(\pm 0.06) nCrs^{1/2} \\ & -1.1863(\pm 0.08) C\text{-}003^{1/2} + 0.3265\ (\pm 0.05) B07[O\text{-}O]^{1/2} \\ & -0.2633(\pm 0.04) F05[N\text{-}O]^{1/2} \\ & N = 93,\ R = 0.9456,\ S = 0.1489,\ FIT = 4.034,\ p < 10^{-3} \\ & R_{loo} = 0.9357,\ S_{loo} = 0.1617,\ R_{TS} = 0.8864,\ S_{TS} = 0.1890. \end{aligned} \quad (5)$$

Table 4
Correlation matrix for descriptors of Eq. (4) ($N = 93$).

	$F08[C—O]^2$	$ATS8m^4$	$F05[O—O]^4$	$TI2^{1/2}$	$GATS4e^{1/2}$
$F08[C—O]^2$	1	0.3214	**0.6458**	0.1324	0.2017
$ATS8m^4$		1	0.0150	0.2937	0.5106
$F05[O—O]^4$			1	0.0362	0.0114
$TI2^{1/2}$				1	0.2031
$GATS4e^{1/2}$					1

The bold number indicates the highest correlation.

The correlation matrix of the descriptors from Eq. (4), shown in Table 4, reveals that the descriptor do not have a relevant degree of inter-correlation.

The predictive power of the models from Eq. (4) and Eq. (5) is satisfactory as revealed by its stability upon the inclusion and/or exclusion of compounds, measured by the statistical parameters $R_{loo} = 0.8940$ ($R_{loo}^2 = 0.7992$) and $R_{loo} = 0.9357$ ($R_{loo}^2 = 0.8755$) respectively. As a general rule $R_{l-n\%-o}$ (Q) should be higher than 0.71 ($Q^2 > 0.5$) to have a validated model [29,32]. The models were further validated by the following conditions [28,29]: $R_{TS}^2 = 0.8710$ and $R_{TS}^2 = 0.7857 > 0.6$; $k = 0.9819$; $k' = 0.9908$ and $k = 0.9723$; $k' = 0.9870$ ($0.85 < k$ or $k' < 1.15$); $(r^2 - r^2_0) / r^2 = -0.1452$ and $-0.2657 < 0.1$; $(r^2 - r'^2_0) / r^2 = -0.1476$ and $-0.2717 < 0.1$ respectively.

To demonstrate that Eq. (4) and Eq. (5) are not the result of accident, the widely used approach *y*-randomization was used to establish the model robustness [33]. This method consists of scrambling the experimental **p** property, so that activities do not correspond to the respective compounds. After analyzing 10,000,000 cases of *y*-randomization, the smallest *S* values obtained in this way were 0.3626 and 0.3510 respectively, which are larger than the ones coming from the calibration (0.1894 and 0.1490). These results suggest that both models are robust, that the calibration is not fortuitous, and that a reliable structure–activity relationship was found.

To determine the robustness of the models from Table 2 (used for the comparison of the different descriptor matrixes), a similar procedure using 100,000 cases was used. In all cases the y-randomization *S* presented significantly higher values than the calibration *S*.

The plot of values predicted by Eq. (4) *vs.* experimental log IC_{50} shown in Fig. 2 suggests that the 63 compounds from the training set

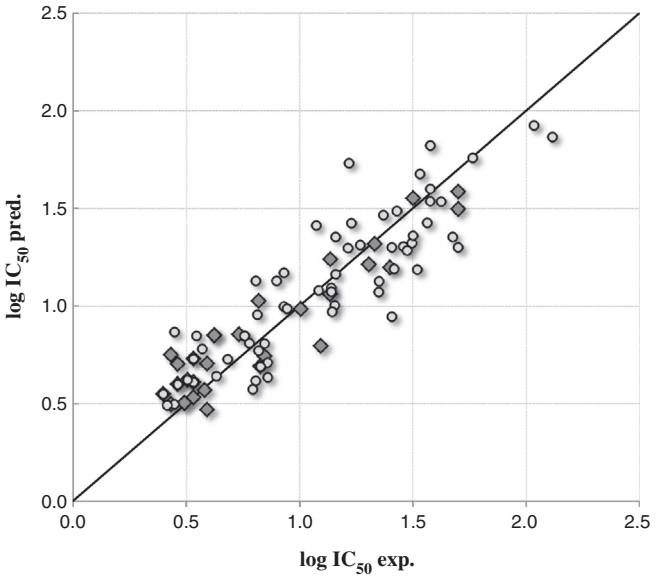

Fig. 2. Predicted (Eq.(4)) *vs* experimental log IC_{50} for the training (circles) and test (rhombus) sets.

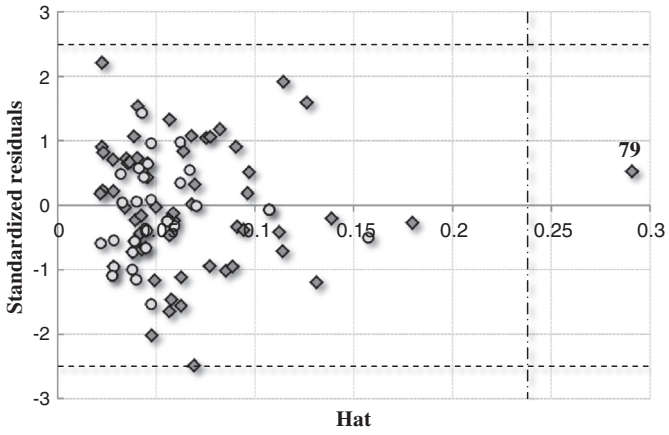

Fig. 3. Williams plot of Eq. (4) showing the Application Domain. The vertical dashed line indicates the limiting leverage h*.

and 30 from the test set follow a straight line. The predicted activity given by Eq. (4) for the training and test sets is shown in Table 1. The Williams plot of the standardized residual in terms of the leverages illustrated in Fig. 3 shows that most compounds lie within the AD of Eq. (4) and were calculated correctly. Compounds **79** and **82** (Hat = 0.9079) are X outliers (Hat > 0.2381) of the training set reinforcing the model [31]; there are no compounds with a standardized residual higher than the limit (2.5σ) that can be considered outliers. The plot of values predicted by Eq. (5) *vs.* experimental log IC_{50} shown in Fig. 4 also suggests that this alternative model follows a straight line.

The standardization of the regression coefficients of Eq. (4) allows assigning greater importance to the molecular descriptors that exhibit the largest absolute standardized coefficients [24]. Taking the descriptor to the power as a new transformed descriptor the standard coefficients are,

$$F08[C-O]^2(0.8232) > ATS8m^4(0.7060) > TI2^{1/2}(0.4318) > F05[O-O]^4(0.3153) > GATS4e^{1/2}(0.3049).$$

By looking at this order we can see that the most significant descriptor is the squared topological descriptor *F08[C—O]*, followed by the 2D autocorrelation to the forth power *ATS8m*.

The review of Eq. (4) also reveals that ERM selected all transformed descriptors.

4. Conclusions

In this work a dataset composed of a series of 93 chalcones with measured activity against human breast cancer cells (MCF-7) was explored revealing that the best models were obtained using only two dimensional descriptors. The comparison between Dragon and QuBiLs-MIDAS for the calculation of three dimensional descriptors showed that better results are found using the first one and the second seems to overfit the training set values. In order to further improve the best obtained models; the two dimensional descriptor matrix was expanded adding several nonlinear transformations. ERM was then used to find a simple to interpret and understand model that nevertheless addressed nonlinearities. The model contained five parameters, all being transformations of descriptors from 2D, and showed better predictive ability than any non-transformed model, established by the theoretical and test set validations. It is expected that the proposed interpretable model may be employed as 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. The fact that no three dimensional descriptors are used additionally simplifies the use of the model since the 3D optimization of structures is not required.

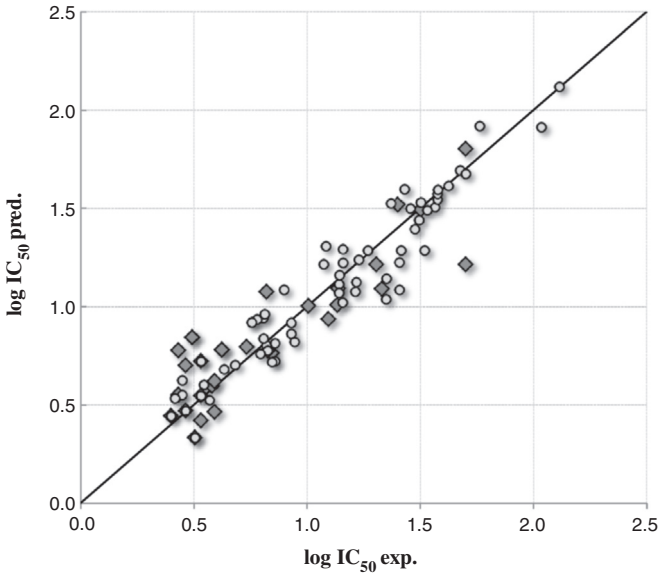

Fig. 4. Predicted (Eq. (5)) *vs* experimental log IC_{50} for the training (circles) and test (rhombus) sets.

Acknowledgments

The authors thank the National Research Council of Argentina (CONICET, PIP11220100100151 project) and INIFTA (CONICET, UNLP) for financial support.

In addition to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities.

References

[1] L. Zhao, X. Zhao, K. Zhao, P. Wei, Y. Fang, F. Zhang, B. Zhang, K. Ogata, A. Mori, T. Wei, The α-tocopherol derivative ESeroS-GS induces cell death and inhibits cell motility of breast cancer cells through the regulation of energy metabolism, Eur. J. Pharmacol. 745 (2014) 98–107.
[2] C. DeSantis, J. Ma, L. Bryan, A. Jemal, Breast cancer statistics, 2013, CA Cancer J. Clin. 64 (2014) 52–62.
[3] T. Rana, A. Chakrabarti, M. Freeman, S. Biswas, Doxorubicin-mediated bone loss in breast cancer bone metastases is driven by an interplay between oxidative stress and induction of TGFβ, PLoS ONE 8 (2013) (e78043).
[4] B.M. Ivković, K. Nikolic, B.B. Ilić, Ž.S. Žižak, R.B. Novaković, O.A. Čudina, S.M. Vladimirov, Phenylpropiophenone derivatives as potential anticancer agents: synthesis, biological evaluation and quantitative structure–activity relationship study, Eur. J. Med. Chem. 63 (2013) 239–255.
[5] D.I. Batovska, I.T. Todorova, Trends in utilization of the pharmacological potential of chalcones, Curr. Clin. Pharmacol. 5 (2010) 1–29.
[6] 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.
[7] Z. Nowakowska, A review of anti-infective and anti-inflammatory chalcones, Eur. J. Med. Chem. 42 (2007) 125–137.
[8] M. López-Lázaro, Flavonoids as anticancer agents: structure–activity relationship study, Curr. Med. Chem.: Anti-Cancer Agents 2 (2002) 691–714.
[9] M. Rybka, A.G. Mercader, E.A. Castro, Predictive QSAR study of chalcone derivatives cytotoxicity activity against HT-29 human colon adenocarcinoma cell lines, Chemom. Intell. Lab. Syst. 132 (2014) 18–29.
[10] X. Liu, M.-L. Go, Antiproliferative activity of chalcones with basic functionalities, Bioorg. Med. Chem. 15 (2007) 7021–7034.
[11] X. Liu, M.-L. Go, Antiproliferative properties of piperidinylchalcones, Bioorg. Med. Chem. 14 (2006) 153–163.
[12] 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.
[13] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience, New York, 2005.
[14] C. Rojas, P.R. Duchowicz, P. Tripaldi, R.P. Diez, QSPR analysis for the retention index of flavors and fragrances on a OV-101 column, Chemom. Intell. Lab. Syst. 140 (2015) 126–132.
[15] T. Mosmann, Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays, J. Immunol. Methods 65 (1983) 55–63.

[16] M.C. Alley, D.A. Scudiero, A. Monks, M.L. Hursey, M.J. Czerwinski, D.L. Fine, B.J. Abbott, J.G. Mayo, R.H. Shoemaker, M.R. Boyd, Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay, Cancer Res. 48 (1988) 589–601.
[17] HYPERCHEM, 6.03 (Hypercube), http://www.hyper.com.
[18] DRAGON, release 5.0 Evaluation Version, http://www.disat.unimib.it/chm.
[19] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley VCH, Weinheim, Germany, 2000.
[20] C.R. García-Jacas, Y. Marrero-Ponce, L. Acevedo-Martínez, S.J. Barigye, J.R. Valdés-Martiní, E. Contreras-Torres, QuBiLS-MIDAS: a parallel free-software for molecular descriptors computation based on multilinear algebraic maps, J. Comput. Chem. 35 (2014) 1395–1409.
[21] 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.
[22] 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.
[23] 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.
[24] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.
[25] 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.
[26] A.K. Saxena, P. Prathipati, Comparison of MLR, PLS and GA-MLR in QSAR analysis, SAR QSAR Environ. Res. 14 (2003) 433–445.
[27] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579–586.
[28] 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.
[29] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, Expert Opin. Drug Discovery 2 (2007) 1567–1577.
[30] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[31] 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.
[32] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Model. 20 (2002) 269–276.
[33] 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.