

The definition of the molecular structure for potential anti-malaria agents by the Monte Carlo method

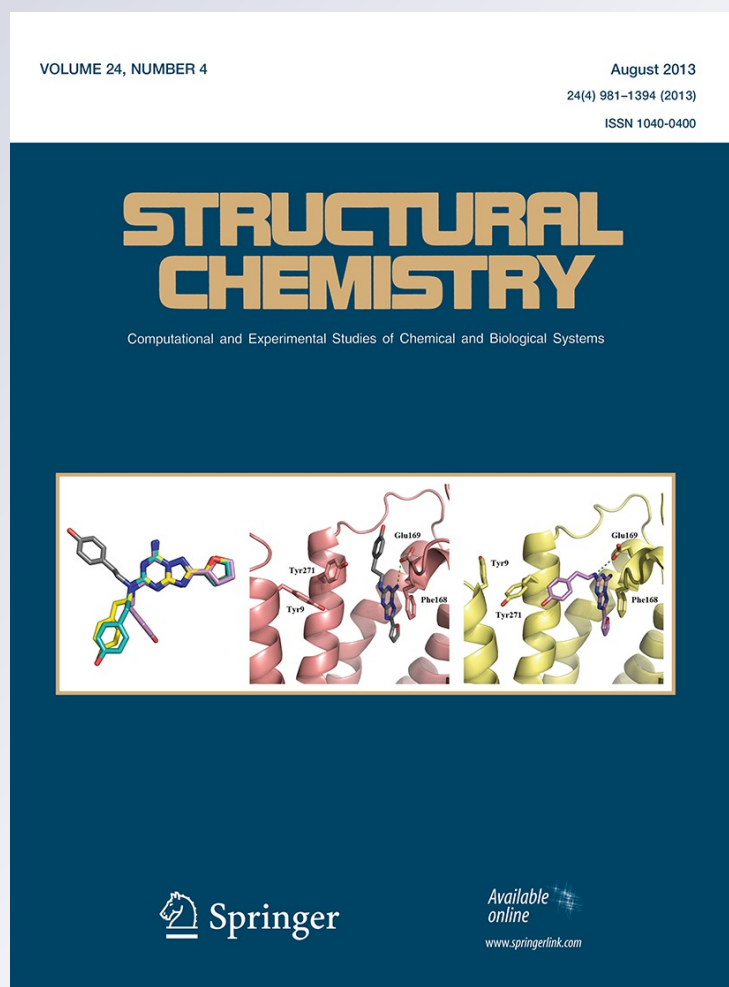
Andrey A. Toropov, Alla P. Toropova, Emilio Benfenati, Giuseppina Gini & Roberto Fanelli

Structural Chemistry

Computational and Experimental Studies of Chemical and Biological Systems

ISSN 1040-0400
Volume 24
Number 4

Struct Chem (2013) 24:1369-1381
DOI 10.1007/s11224-012-0180-2



Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media New York. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

The definition of the molecular structure for potential anti-malaria agents by the Monte Carlo method

Andrey A. Toropov · Alla P. Toropova ·
Emilio Benfenati · Giuseppina Gini ·
Roberto Fanelli

Received: 9 November 2012 / Accepted: 5 December 2012 / Published online: 22 December 2012
© Springer Science+Business Media New York 2012

Abstract A series of 53 endochin analogs (4(1-H)-quinolone derivatives) with anti-malarial activity against the clinically relevant multidrug resistant malarial strain TM-90-C2B has been studied. The CORAL (<http://www.insilico.eu/coral>) software has been used as a tool to build up the quantitative structure–activity relationships (QSAR) for the anti-malaria activity. The QSAR models were calculated with the representation of the molecular structure by simplified molecular input-line entry system and by the molecular graph of atomic orbitals. The method for splitting data into the sub-training set, the calibration set, the test set, and the validation set is suggested. Three various splits were examined. Statistical quality of models for the validation sets (which are not involved in the building up models) is good. Structural indicators (alerts) for increase and decrease of the anti-malaria activity are defined.

Keywords QSAR · Anti-malaria activity · Monte Carlo method · CORAL software

Introduction

Malaria still remains one of the major health problems in the world, consequently the search for new anti-malarial drugs is an important task of medicinal chemistry [1]. Quantitative structure–activity/property relationships (QSARs/QSPRs) can be useful for the solution of this problem [1, 2]. Molecular descriptors which are calculated with molecular graph are basis of the QSPR/QSAR analyses [3, 4] by means of various approaches such as, multiple regression analysis [5], virtual screening [6], artificial neural networks [7].

The CORAL (Correlation And Logic) software is a tool to build up a QSAR model by means of so-called optimal descriptors which are calculated with the Monte Carlo method [8, 9]. The representation of the molecular structure for the software is the simplified molecular input-line entry system (SMILES) [10–12] that can be converted into molecular graphs [9].

The aims of the present study are (i) the estimation of the CORAL models for anti-malaria activity; and (ii) an attempt to define the molecular structure for the potential promising anti-malaria agents.

Method

The data set

A series of 53 endochin analogs (4(1-H)-quinolone derivatives) with anti-malarial activity against the clinically relevant multidrug resistant malarial strain TM-90-C2B (Table 1) has been taken from the literature [1]. The negative decimal logarithm of the effective concentration [$pEC_{50} = \log(10^9/EC_{50} \text{ (nM)})$] has been used as the

Electronic supplementary material The online version of this article (doi:10.1007/s11224-012-0180-2) contains supplementary material, which is available to authorized users.

A. A. Toropov (✉) · A. P. Toropova · E. Benfenati · R. Fanelli
Istituto di Ricerche Farmacologiche Mario Negri,
20156, Via La Masa 19, Milan, Italy
e-mail: andrey.toropov@marionegri.it

G. Gini
Department of Electronics and Information, Politecnico di
Milano, Piazza Leonardo Da Vinci 32, 20133 Milan, Italy

Table 1 The molecular structures of 53 endochin analogs of 4(1-H)-quinolone derivatives

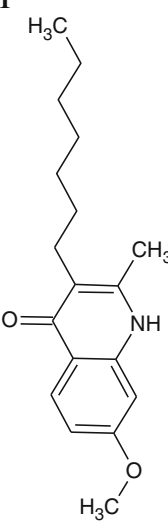
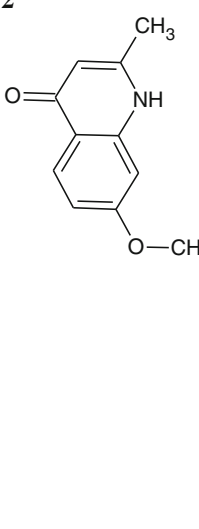
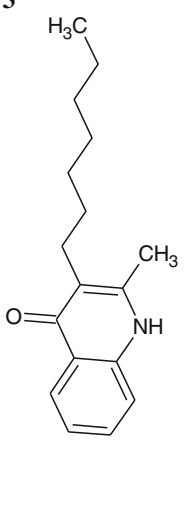
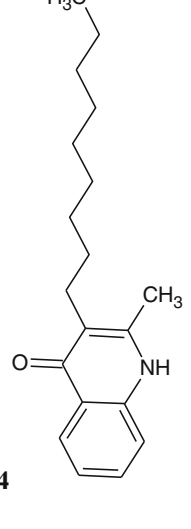
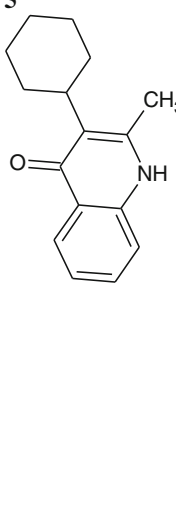
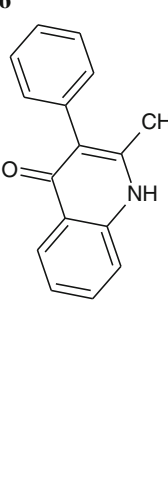
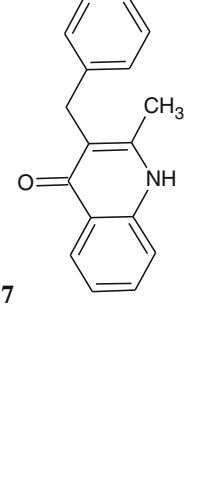
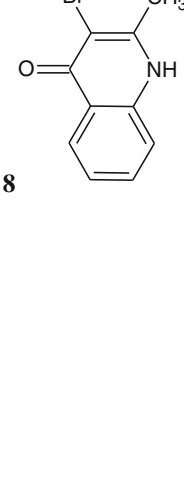
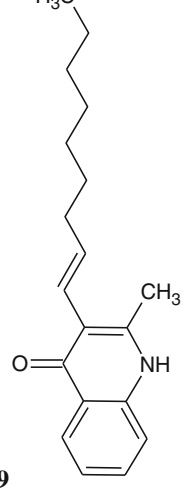
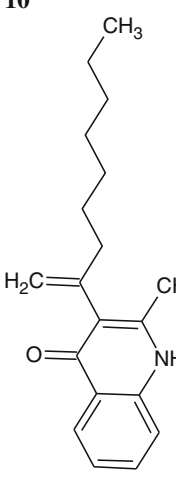
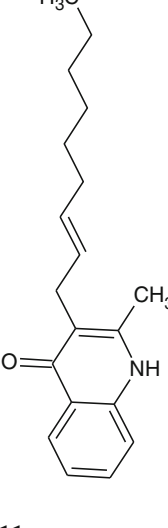
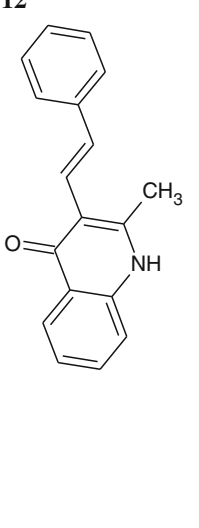
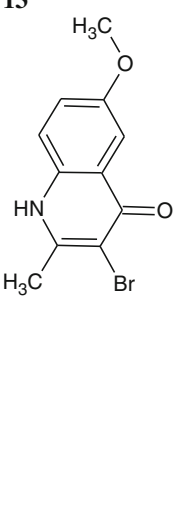
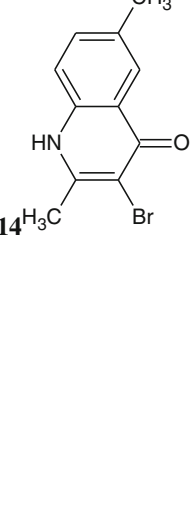
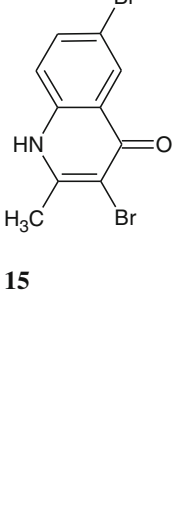
<p>1</p> 	<p>2</p> 	<p>3</p> 	<p>4</p> 	<p>5</p> 
<p>6</p> 	<p>7</p> 	<p>8</p> 	<p>9</p> 	<p>10</p> 
<p>11</p> 	<p>12</p> 	<p>13</p> 	<p>14</p> 	<p>15</p> 

Table 1 continued

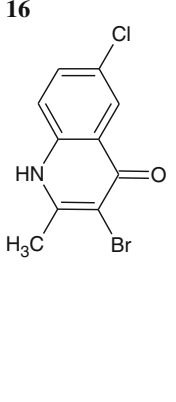
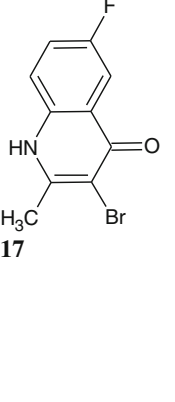
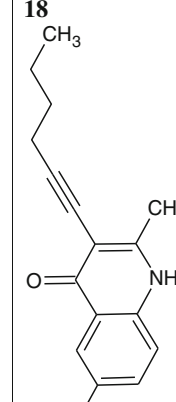
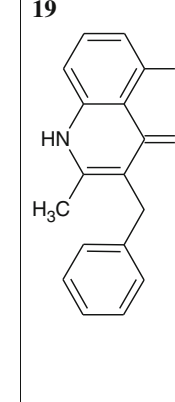
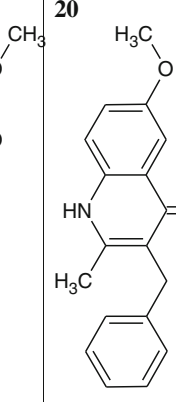
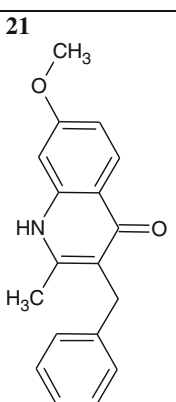
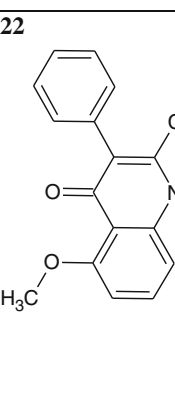
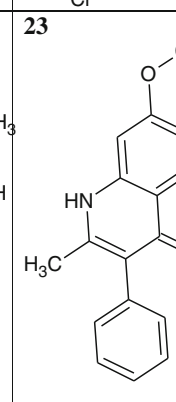
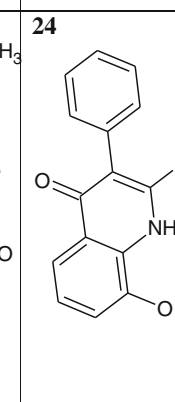
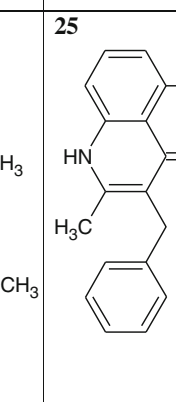
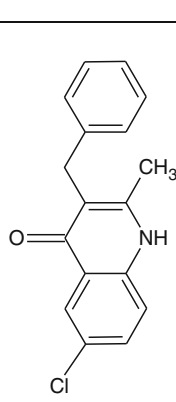
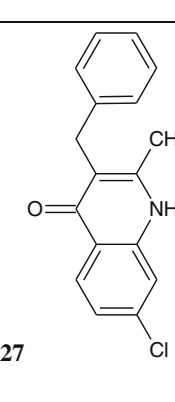
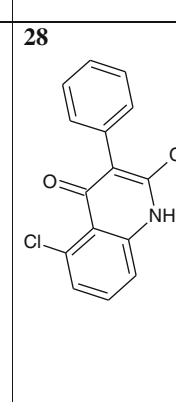
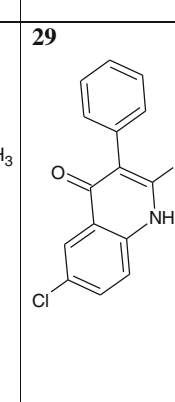
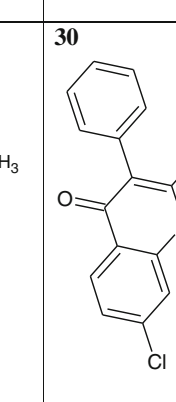
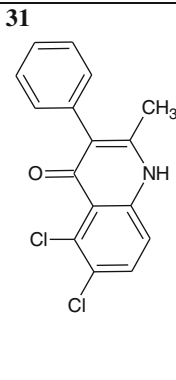
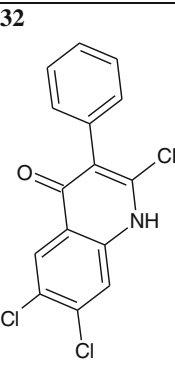
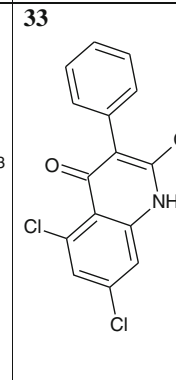
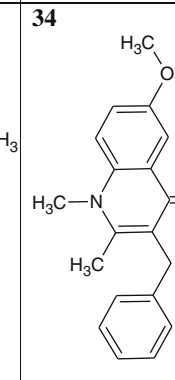
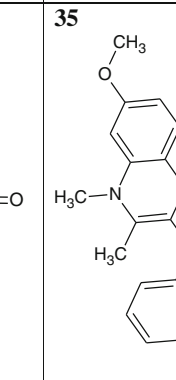
16 	17 	18 	19 	20 
21 	22 	23 	24 	25 
26 	27 	28 	29 	30 
31 	32 	33 	34 	35 

Table 1 continued

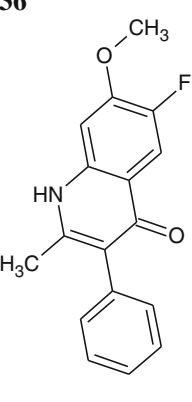
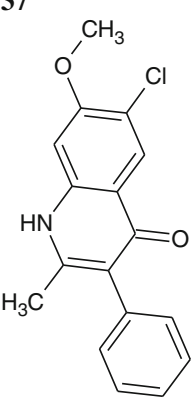
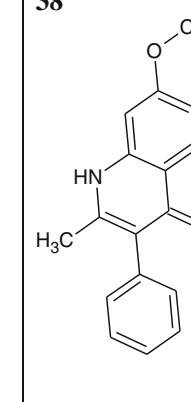
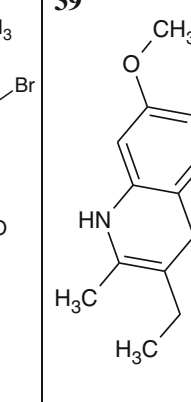
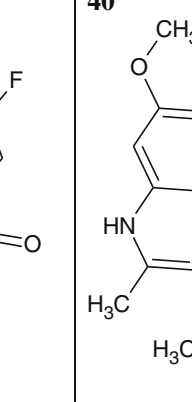
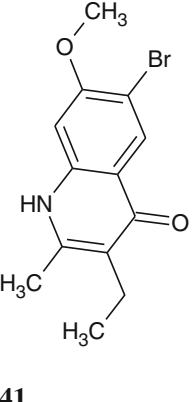
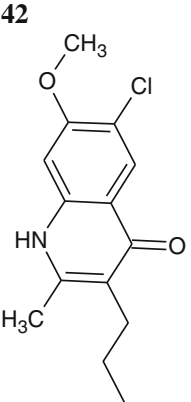
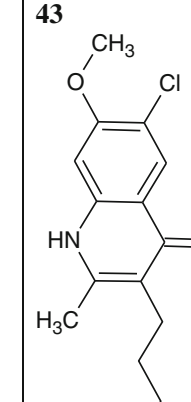
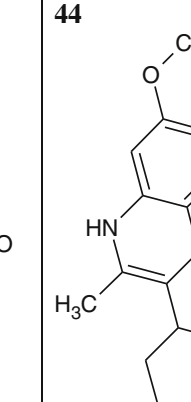
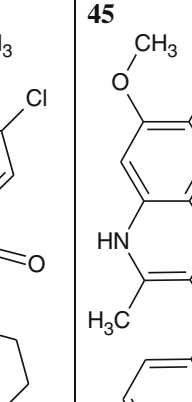
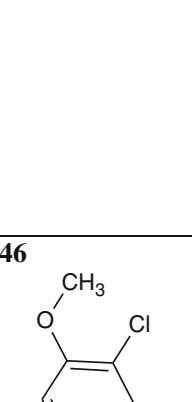
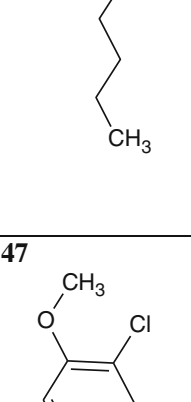
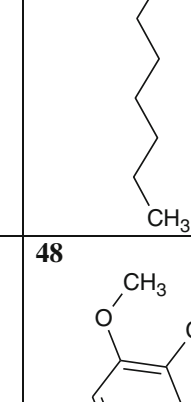
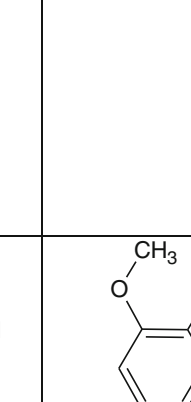
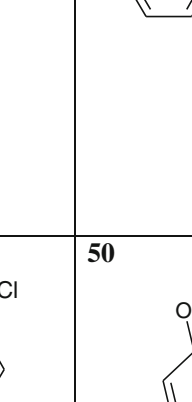
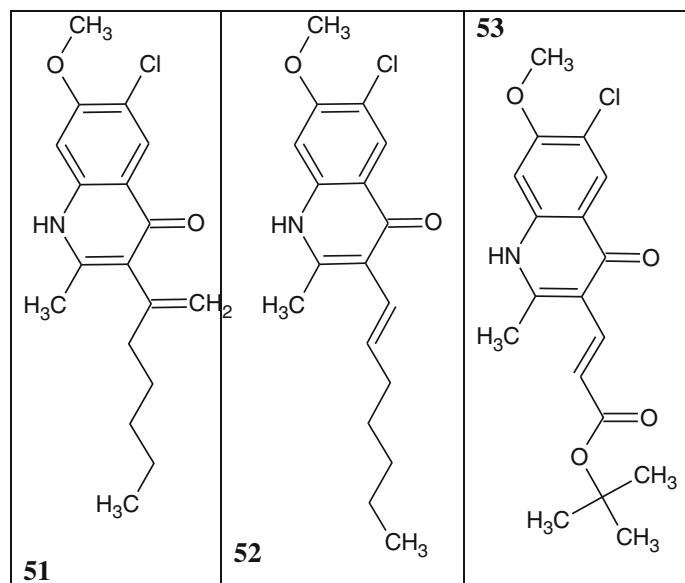
36 	37 	38 	39 	40 
41 	42 	43 	44 	45 
46 	47 	48 	49 	50 

Table 1 continued

endpoint. The QSAR models were built up for three random splits.

Splitting

The proper splitting of the dataset into the corresponding training and test sets plays the important role for the successful development of a QSAR model [13, 14]. In the present study, the splits were selected according to the following principles: (i) the range of the endpoint is approximately the same for each sub-set; (ii) the splits are random; and (iii) the splits are not identical (Table 2). The validation set contains molecular structures which are not involved in the building up of the models. We have checked up level of identity of these splits (Table 2). Table 2 shows that these random splits are different enough.

Optimal descriptors

The molecular structure can be represented by SMILES (Fig. 1) and by molecular graph, in particular, the graph of atomic orbitals (GAO). The SMILES and GAO are different representations of the molecular structure (Fig. 2). We believe that the “hybrid” representation of the molecular structure, i.e., by SMILES together with GAO, can give a model characterized by higher statistical quality than model which is based on the representation of the molecular structure by solely SMILES (or solely GAO).

The hybrid optimal descriptors were used to build up model for the $pEC50$:

$$\text{Hybrid } DCW(T, N_{\text{epoch}}) = \text{SMILES } DCW(T, N_{\text{epoch}}) + \text{GAO } DCW(T, N_{\text{epoch}}) \quad (1)$$

Table 2 Percentage of identity of splits 1–3

	Set	Split 1 (%)	Split 2 (%)	Split 3 (%)
Split 1	Sub-training	100*	18.8	38.9
	Calibration	100	19.4	19.4
	Test	100	0.0	20.0
	Validation	100	0.0	21.1
Split 2	Sub-training		100	28.6
	Calibration		100	33.3
	Test		100	36.4
	Validation		100	40.0
Split 3	Sub-training			100
	Calibration			100
	Test			100
	Validation			100

$$*Identity(\%) = \frac{N_{ij}}{0.5 \times (N_i + N_j)} \times 100$$

where

N_{ij} is the number of substances which are distributed into the same set for both i -th split and j -th split (set = sub-training, calibration, test, validation)

N_i is the number of substances which are distributed into the set for i -th split

N_j is the number of substances which are distributed into the set for j -th split

where

$${}^{\text{SMILES}}DCW(T, N_{\text{epoch}}) = \sum CW(S_k)$$

$${}^{\text{GAO}}DCW(T, N_{\text{epoch}}) = \sum CW(AO_k) + \sum CW(ECO_k)$$

T is threshold for definition of rare (noise) molecular features: $T = 1$ and 2 were examined (if an molecular feature, x , that is extracted from SMILES or GAO takes place less than T times, in the sub-training set, then the correlation weight of the x , $CW(x) = 0$, i.e., the x should be removed from the building up of the model);

N_{epoch} is the number of the epochs of the Monte Carlo optimization: $N_{\text{epoch}} = 150$ has been selected;

S_k is attribute of SMILES: it may be two symbols, which cannot be examined separately (e.g., “Cl”, “Br”, etc.), but in the majority S_k contains one symbol (“C”, “O”, “=”, etc.);

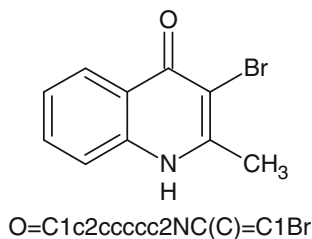
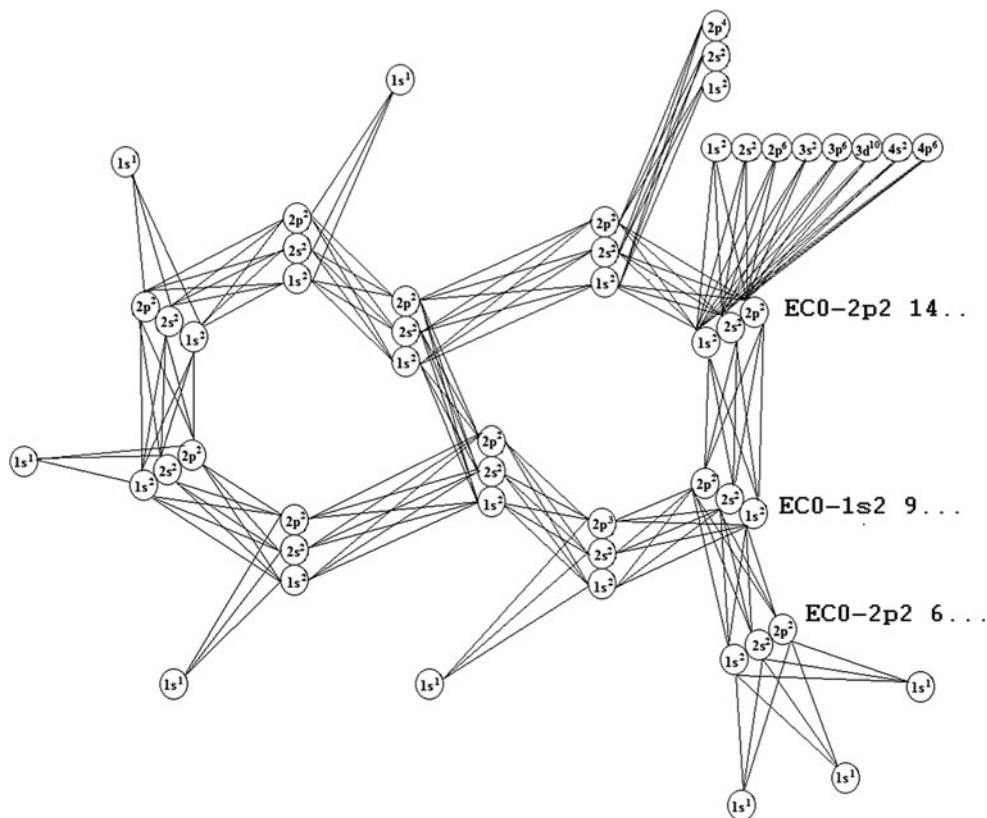


Fig. 1 The molecular structure and SMILES for compound #8

Fig. 2 The architecture of the GAO and several examples of the definitions for Morgan extended connectivity of zero order for compound #8



AO_k is atomic orbitals such as, $1s^2$, $2p^3$, $3d^{10}$, etc. (Fig. 2).

ECO_k is the Morgan extended connectivity of zero order (i.e., the vertex degree) in the GAO (Fig. 2);

$CW(x)$ is the correlation weights for x (some molecular feature which can be extracted from SMILES or from GAO). Table 3 contains an example of the $CW(x)$ calculated with the Monte Carlo method. Table 4 contains an example of the calculation with the correlation weights.

The principle of the building up a QSAR by means of the CORAL software is the calculation of the $CW(x)$ providing the maximum of a target function (TF) [15–17].

$$TF = R + R' - |R - R'| \times \text{Const} \quad (2)$$

where R and R' are correlation coefficients between $pEC50$ and ${}^{\text{Hybrid}}DCW(T, N_{\text{epoch}})$; Const is an empirical constant (we have used Const = 0.1).

In fact the CORAL software should be used according to the following scheme: (i) the phase 1, one should carry out several runs of the Monte Carlo method optimization to define the preferable threshold and N_{epoch} ; which give best statistical quality of the model for the test set; and (ii) the phase 2, one should build up a model using the preferable threshold and N_{epoch} and check up the model with external validation set, which was invisible during the definition of the preferable threshold and N_{epoch} .

Table 3 Correlation weights for calculation of Hybrid $DCW(T, N_{\text{epoch}})$ for the case of split 1 (Eq. 4)

Molecular features (x) extracted from SMILES and GAO	$CW(x)$	N_{TRN}^*	N_{CLB}	N_{TST}
#.....	-5.89900	1	1	0
(.....	-1.22500	20	13	10
/.....	-1.13975	1	0	1
1.....	0.76875	20	13	10
2.....	0.90000	20	13	10
3.....	0.45225	8	9	6
=.....	1.30200	20	13	10
C.....	0.26550	20	13	10
F.....	2.86025	2	0	1
EC0-1s1 3...	0.17075	20	13	10
EC0-1s2 11..	1.71775	5	7	5
EC0-1s2 12..	0.0	0	0	1
EC0-1s2 14..	1.00300	5	1	1
EC0-1s2 3...	0.20300	20	13	10
EC0-1s2 6...	-0.04075	20	13	10
EC0-1s2 7...	0.39900	20	13	10
EC0-1s2 9...	-1.21125	20	13	10
EC0-3d103...	1.52500	5	1	1
EC0-2p2 11..	1.32725	5	7	5
EC0-2p2 12..	0.0	0	0	1
EC0-2p2 14..	1.10400	5	1	1
EC0-2p2 3...	3.54475	20	13	10
EC0-2p2 6...	0.13025	13	5	6
EC0-2p2 7...	0.55925	20	13	10
EC0-2p2 9...	2.28250	20	13	10
EC0-2p3 6...	3.42275	19	13	9
EC0-2p3 9...	1.40300	1	0	1
EC0-2p4 3...	0.67825	20	13	10
EC0-2p4 6...	0.93625	12	6	5
EC0-2p5 3...	2.84375	2	0	1
EC0-2p6 3...	-1.80300	10	8	6
EC0-2s2 11..	1.35400	5	7	5
EC0-2s2 12..	0.0	0	0	1
EC0-2s2 14..	1.25225	5	1	1
EC0-2s2 3...	0.25950	20	13	10
EC0-2s2 6...	-0.02700	20	13	10
EC0-2s2 7...	0.21975	20	13	10
EC0-2s2 9...	-0.89675	20	13	10
EC0-3p5 3...	1.40400	5	7	5
EC0-3p6 3...	1.19600	5	1	1
EC0-3s2 3...	-1.84500	10	8	6
EC0-4p5 3...	1.11550	5	1	1
EC0-4s2 3...	1.05225	5	1	1
Br.....	1.50400	5	1	1
Cl.....	1.10925	5	7	5
N.....	0.66150	20	13	10
O.....	0.12300	20	13	10
\.....	0.0	0	1	0
c.....	-0.73000	20	13	10

* N_{TRN} , N_{CLB} , and N_{TST} are the numbers of the x in the sub-training, calibration, and test set

Table 4 An example of the calculation for ${}^{\text{hybrid}}\text{DCW}(1, 138) = 22.12050$

Molecular features (x)	Correlation Weight, $CW(x)$
Extracted from GAO	
EC0-1s2 3...	0.203
EC0-2s2 3...	0.260
EC0-2p4 3...	0.678
EC0-1s2 9...	-1.211
EC0-2s2 9...	-0.897
EC0-2p2 9...	2.283
EC0-1s2 9...	-1.211
EC0-2s2 9...	-0.897
EC0-2p2 9...	2.283
EC0-1s2 7...	0.399
EC0-2s2 7...	0.220
EC0-2p2 7...	0.559
EC0-1s2 7...	0.399
EC0-2s2 7...	0.220
EC0-2p2 7...	0.559
EC0-1s2 7...	0.399
EC0-2s2 7...	0.220
EC0-2p2 7...	0.559
EC0-1s2 7...	0.399
EC0-2s2 7...	0.220
EC0-2p2 7...	0.559
EC0-1s2 9...	-1.211
EC0-2s2 9...	-0.897
EC0-2p2 9...	2.283
EC0-1s2 6...	-0.041
EC0-2s2 6...	-0.027
EC0-2p3 6...	3.423
EC0-1s2 9...	-1.211
EC0-2s2 9...	-0.897
EC0-2p2 9...	2.283
EC0-1s2 3...	0.203
EC0-2s2 3...	0.260
EC0-2p2 3...	3.545
EC0-1s2 14..	1.003
EC0-2s2 14..	1.252
EC0-2p2 14..	1.104
EC0-1s2 3...	0.203
EC0-2s2 3...	0.260
EC0-2p6 3...	-1.803
EC0-3s2 3...	-1.845
EC0-3p6 3...	1.196
EC0-3d103...	1.525
EC0-4s2 3...	1.052
EC0-4p5 3...	1.115
EC0-1s1 3...	0.171
EC0-1s1 3...	0.171

Table 4 continued

Molecular features (x)	Correlation Weight, $CW(x)$
EC0-1s1 3...	0.171
EC0-1s1 3...	0.171
Extracted from SMILES	
O.....	0.123
=.....	1.302
C.....	0.266
1.....	0.769
C.....	-0.730
2.....	0.900
C.....	-0.730
C.....	-0.730
C.....	-0.730
C.....	-0.730
2.....	0.900
N.....	0.661
C.....	0.266
(.....	-1.225
C.....	0.266
(.....	-1.225
=.....	1.302
C.....	0.266
1.....	0.769
Br.....	1.504

The representations of the molecular structure of this compound by SMILES and by GAO are shown in Figs. 1 and 2, respectively. (ID 8, split 1, Eq. 4)

Results and discussion

Table 5 contains the statistical quality of one-variable models which can be represented by the generalized formula

$$pEC50 = C_0 + C_1 \times {}^{\text{Hybrid}}\text{DCW}(T, N_{\text{epoch}}) \quad (3)$$

One can see (Table 5) that the best statistical quality for all splits takes place if the threshold equals to 1. The average values of the N_{epoch} are 138, 133, and 140 for split 1, split 2, and split 3, respectively. The using of above-mentioned threshold and the number of epochs gives the following models for the $pEC50$:

Split 1

$$pEC50 = 0.6293 (\pm 0.1583) + 0.2784 (\pm 0.0075) * \text{DCW}(1, 138)$$

(4)

Table 5 Statistical characteristics of one-variable models for anti-malaria activity (pEC50) obtained with optimal descriptor calculated with Eq. 1: the threshold $T = 1, 2$ and the number of epochs of the Monte Carlo optimization $N_{epoch} = 150$

Split	Run	T	N_{act}^*	Sub-training set			Calibration set			Test set				
				n	r^2	q^2	s	F	n	r^2	s	n	r^2	s
1	1	1	45	20	0.7156	0.6697	0.561	45	0.7163	0.519	10	0.9729	0.557	
	2	1	45	20	0.7145	0.6676	0.562	45	0.7134	0.525	10	0.9775	0.573	
	3	1	45	20	0.7145	0.6678	0.562	45	0.7159	0.520	10	0.9728	0.540	
	Average				0.7149	0.6684	0.562	45	0.7152	0.521		0.9744	0.557	
	1	2	42	20	0.4438	0.3331	0.785	14	0.4438	0.647	10	0.9557	0.656	
	2	2	42	20	0.4444	0.3336	0.784	14	0.4444	0.651	10	0.9545	0.666	
	3	2	42	20	0.4499	0.3424	0.780	15	0.4499	0.642	10	0.9565	0.643	
	Average				0.4460	0.3364	0.783	14	0.4460	0.647		0.9556	0.655	
	2	1	1	39	12	0.8964	0.8673	0.342	87	0.6964	0.507	12	0.8381	0.547
2	1	1	39	12	0.8973	0.8690	0.340	87	0.6958	0.508	12	0.7736	0.628	
3	1	1	39	12	0.8957	0.8668	0.343	86	0.6971	0.505	12	0.8079	0.584	
Average					0.8965	0.8677	0.342	87	0.6964	0.507		0.8065	0.586	
2	1	2	32	12	0.4761	0.2299	0.769	9	0.6492	0.513	12	0.5113	0.855	
	2	2	32	12	0.4759	0.2314	0.769	9	0.6468	0.514	12	0.5038	0.861	
	3	2	32	12	0.4757	0.2256	0.769	9	0.6467	0.514	12	0.5103	0.855	
	Average				0.4759	0.2290	0.769	9	0.6476	0.514		0.5085	0.857	
	3	1	1	42	16	0.7299	0.6660	0.611	38	0.7320	0.585	10	0.9535	0.351
	2	1	1	42	16	0.7336	0.6700	0.607	39	0.7331	0.581	10	0.9536	0.347
	3	1	1	42	16	0.7433	0.6851	0.596	41	0.7450	0.574	10	0.9562	0.344
	Average					0.7356	0.6737	0.605	39	0.7367	0.580		0.9544	0.347
	1	2	41	16	0.4959	0.3631	0.835	14	0.5804	0.675	10	0.8289	0.587	
2	2	41	16	0.4966	0.3632	0.834	14	0.5802	0.675	10	0.8291	0.588		
3	2	41	16	0.4974	0.3651	0.833	14	0.5790	0.676	10	0.8274	0.587		
Average					0.4966	0.3638	0.834	14	0.5798	0.675		0.8285	0.587	

* N_{act} is the number of molecular features which take place in the sub-training set at least one time, n is the number of substances in the set, r^2 is square correlation coefficient between experimental and calculated values of pEC50, q^2 is the cross-validated squared correlation coefficient, s is standard error of estimation, and F is Fischer F -ratio. Best models are indicated by bold

Table 6 The checking of models calculated with Eqs. 4–6 (for test sets) with Y-randomization

Probe of Y-scrambling	Split 1, Eq. 4 R_r^2	Split 2, Eq. 5 R_r^2	Split 3, Eq. 6 R_r^2
R^2	0.9783	0.7949	0.9553
1	0.4236	0.2104	0.1318
2	0.9698	0.3638	0.0307
3	0.0075	0.5909	0.4882
4	0.4587	0.0550	0.4388
5	0.8696	0.1802	0.4893
6	0.4537	0.1820	0.0400
7	0.1433	0.1681	0.6808
8	0.2080	0.6054	0.2732
9	0.0298	0.2157	0.3258
10	0.0029	0.2309	0.8788
$\overline{R_r^2}$	0.3567	0.2802	0.3777
${}^cR_p^2$	0.7798	0.6396	0.7428

The $\overline{R_r^2}$ is average for ten probes of the Y-scrambling. The ${}^cR_p^2$ (calculated by ${}^cR_p^2 = R\sqrt{R^2 - \overline{R_r^2}}$) should be larger than 0.5 [18]

$n = 20$, $R^2 = 0.7009$, $q^2 = 0.6494$, $s = 0.575$, $F = 42$ (sub-training set)

$n = 13$, $R^2 = 0.7009$, $s = 0.529$ (calibration set)

$n = 10$, $R^2 = 0.9783$, $s = 0.569$ (test set)

$n = 10$, $R^2 = 0.6130$, $s = 0.89$, $R_m^2 = 0.55$ (validation set)

Split 2

$$pEC50 = 0.0003 (\pm 0.1233) + 0.4544 (\pm 0.0097) * DCW(1, 133) \quad (5)$$

$n = 12$, $R^2 = 0.8918$, $q^2 = 0.8618$, $s = 0.350$, $F = 82$ (sub-training set)

$n = 18$, $R^2 = 0.6966$, $s = 0.507$ (calibration set)

$n = 12$, $R^2 = 0.7949$, $s = 0.602$ (test set)

$n = 11$, $R^2 = 0.6252$, $s = 0.641$, $R_m^2 = 0.57$ (validation set)

Split 3

$$pEC50 = -0.0020 (\pm 0.2170) + 0.3613 (\pm 0.0122) * DCW(1, 140) \quad (6)$$

$n = 16$, $R^2 = 0.7365$, $q^2 = 0.6750$, $s = 0.604$, $F = 39$ (sub-training set)

$n = 18$, $R^2 = 0.7367$, $s = 0.577$ (calibration set)

$n = 10$, $R^2 = 0.9553$, $s = 0.343$ (test set)

$n = 9$, $R^2 = 0.6886$, $s = 0.522$, $R_m^2 = 0.66$ (validation set)

in the Eq. 4–6, n is the number of compounds in the set, r is the correlation coefficient, q^2 is the leave-one-out cross-validated correlation coefficient; s is the standard error of estimation, F is the Fischer F -ratio, and R_m^2 is the criterion of the predictability of a model (the model has predictability if $R_m^2 > 0.5$) [1, 13]. The developed QSAR models were further validated using the Y-scrambling (Table 6) to examine their robustness [8, 18].

The statistical characteristics of the model for anti-malaria activity [1] (the same 53 substances) are the following: $n_{\text{train}} = 39$, $R_{\text{train}}^2 = 0.797$, $s_{\text{train}} = 0.517$; $n_{\text{test}} = 14$, $R_{\text{test}}^2 = 0.808$. We believe that the statistical quality of models which are calculated with Eqs. 4–6 comparable with the above-mentioned model [1], but one should consider the analysis for three splits, as more robust data than data for only one split.

The analysis of correlation weights which were obtained in three runs of the Monte Carlo optimization for each split gives the possibility to detect (i) structural features with stable positive values of the correlation weights, which are promoter of the endpoint increase; and (ii) structural features with stable negative values of the correlation weights, which are promoters of the endpoint decrease. By this manner, we have established, that the presence of nitrogen is the stable promoter of the pEC50 increase, vice versa, the presence of branching (this is encoded in SMILES by brackets) is the stable promoter of the pEC50 decrease (Table 7). In other words, models calculated with Eqs 4–6 have the mechanistic interpretation. Table 8 contains an example of the design of structures of 4(1-H)-quinolone derivatives which can be perspective anti-malaria agents.

Thus, the CORAL software can be a tool for QSAR analysis of the anti-malaria activity of 4(1-H)-quinolone derivatives. *Supplementary materials* section contains three splits which were studied (SMILES and numerical data on the pEC50).

Conclusions

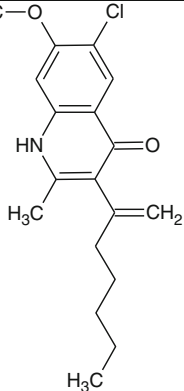
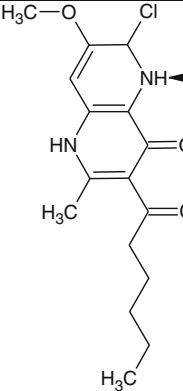
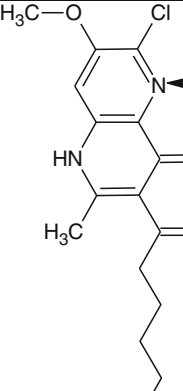
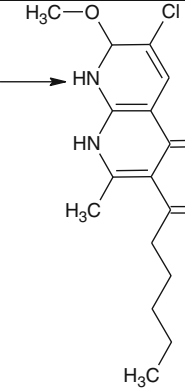
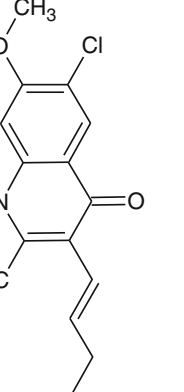
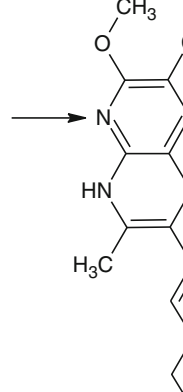
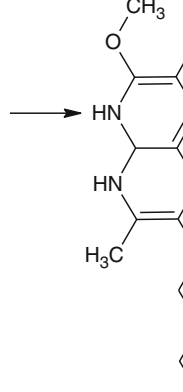
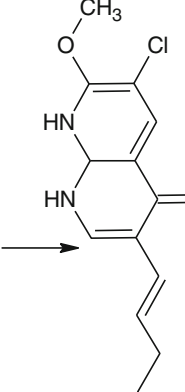
The CORAL software gives models of anti-malaria activity of good statistical quality (Table 5). The statistical quality, however, is considerably dependent on the distribution of available substances into the sub-training set, the calibration set, the test set, and the validation set. The computational experiments gave statistically stable structural alerts (promoters of increase or decrease of pEC50, Table 7). One can use these alerts in the search for new potential anti-malaria agents (Table 8).

Table 7 The analysis of the stability of the correlation weights of molecular features (stable positive or stable negative values for all splits in series of the runs of the Monte Carlo method optimization)

Split	Molecular features, x	Correlation weights, $CW(x)$			N_{TRN}	N_{CLB}	N_{TST}
		Run 1	Run 2	Run 3			
1	1.....	0.92900	0.90425	0.70325	20	13	10
	2.....	1.04800	0.98125	1.23975	20	13	10
	=.....	0.57700	0.85300	0.69600	20	13	10
	C.....	0.37725	0.28425	0.23525	20	13	10
	EC0-1s1 3...	0.37500	0.48425	0.31125	20	13	10
	EC0-1s2 3...	0.00500	0.06050	0.17700	20	13	10
	EC0-1s2 7...	0.29675	0.27700	0.57175	20	13	10
	EC0-2p2 3...	3.90950	4.09675	4.02000	20	13	10
	EC0-2p2 7...	0.46250	0.35725	0.35425	20	13	10
	EC0-2p2 9...	2.16675	2.27375	2.10100	20	13	10
	EC0-2p4 3...	0.97800	0.96025	0.83425	20	13	10
	EC0-2s2 3...	0.07300	0.08525	0.20200	20	13	10
	EC0-2s2 7...	0.35950	0.45625	0.35400	20	13	10
	N.....	0.18250	0.86750	1.11875	20	13	10
	(.....)	-1.21350	-1.22700	-1.22900	20	13	10
	EC0-1s2 9...	-0.89575	-0.89475	-1.04800	20	13	10
	EC0-2s2 9...	-0.94050	-1.24250	-0.84175	20	13	10
C.....	-0.87075	-0.89900	-0.96050	20	13	10	
2	1.....	1.35225	1.37000	1.20300	12	18	12
	2.....	1.33000	1.32925	1.18625	12	18	12
	C.....	1.01475	0.78350	0.68025	12	18	12
	EC0-1s2 3...	0.26650	0.25325	0.25950	12	18	12
	EC0-1s2 7...	0.15825	0.22500	0.43125	12	18	12
	EC0-2p2 3...	2.62925	2.80925	2.82800	12	18	12
	EC0-2p2 7...	0.22375	0.10000	0.39050	12	18	12
	EC0-2p2 9...	1.57625	1.37925	1.53225	12	18	12
	EC0-2p3 6...	2.58125	2.63450	2.49600	12	17	12
	EC0-2s2 6...	0.03225	0.17500	0.16025	12	18	12
	EC0-2s2 7...	0.52900	0.04175	0.31050	12	18	12
	N.....	1.49800	1.27900	1.34675	12	18	12
	(.....)	-1.01650	-0.98350	-1.02275	12	18	12
	=.....	-1.38450	-1.26450	-1.37175	12	18	12
	EC0-1s2 9...	-0.62100	-0.80200	-0.93525	12	18	12
	EC0-2p4 3...	-4.51650	-4.59775	-4.11750	12	18	12
	EC0-2s2 9...	-1.07375	-0.57700	-0.52100	12	18	12
O.....	-0.83650	-0.85125	-0.52100	12	18	12	
3	1.....	1.22100	1.21150	1.30950	16	18	10
	2.....	1.46450	1.62400	1.58425	16	18	10
	EC0-1s2 3...	2.59800	2.61775	2.53125	16	18	10
	EC0-2p2 9...	2.51250	2.29575	2.12300	16	18	10
	EC0-2s2 3...	2.74250	2.65000	2.57725	16	18	10
	N.....	1.45925	1.30225	1.39800	16	18	10
	EC0-2p3 6...	4.12625	3.97825	3.90825	14	18	10
	(.....)	-1.23850	-1.17825	-1.00100	16	18	10
	=.....	-0.84375	-0.91050	-1.04500	16	18	10
	C.....	-0.18425	-0.24475	-0.29775	16	18	10
	EC0-1s2 9...	-1.24075	-1.29075	-1.35300	16	18	10
	EC0-2p2 3...	-0.24675	-0.09175	-0.04600	16	18	10
	EC0-2p4 3...	-1.69600	-1.49075	-1.95325	16	18	10
EC0-2s2 6...	-0.20000	-0.55400	-0.09275	16	18	10	
EC0-2s2 9...	-1.32500	-1.12800	-1.26875	16	18	10	
O.....	-1.73975	-2.04775	-1.59875	16	18	10	

* N_{TRN} , N_{CLB} , and N_{TST} are the numbers of the x in the sub-training, calibration, and test set. The stable promoter of the pEC50 increase for all splits and all runs is the presence of nitrogen; the stable promoter of the pEC50 decrease is the branching. The stable promoters are indicated by bold

Table 8 The design of perspective anti-malaria agents by means of the using of models which are calculated with Eq. 4, Eq. 5, and Eq. 6: compounds #51 and #52 as the basis of the design were used

Experiment	Hypotheses				
 <p>51</p>	 <p>51-1</p>	 <p>51-2</p>	 <p>51-3</p>		
pEC50(expr)=7.830	pEC50(calc)	pEC50(calc)	pEC50(calc)	pEC50(calc)	Eq.
pEC50(calc)=8.354	11.415	9.116	11.053	4	
pEC50(calc)=7.862	8.866	8.548	9.453	5	
pEC50(calc)=7.813	8.413	9.083	8.735	6	
 <p>52</p>	 <p>52-1</p>	 <p>52-2</p>	 <p>52-3</p>		
pEC50(expr)=7.975	pEC50(calc)	pEC50(calc)	pEC50(calc)	pEC50(calc)	Eq.
pEC50(calc)=7.545	8.306	10.606	9.704	4	
pEC50(calc)=7.755	8.440	8.759	8.410	5	
pEC50(calc)=7.708	8.977	8.308	8.032	6	

Perspective modifications of structure are indicated by arrows

Acknowledgments We thank ANTARES (the project number LIFE08-ENV/IT/00435). Also we express our gratitude to Dr. L. Cappellini, Dr. G. Bianchi, and Dr. R. Bagnati for valuable consultations on the computer science, and to J. Baggott for English editing.

References

- Ojha PK, Roy K (2011) Chemom Intell Lab 109:146
- Ibezim E, Duchowicz PR, Ortiz EV, Castro EA (2012) Chemom Intell Lab 110:81

3. Došlić T, Furtula B, Graovac A, Gutman I, Moradi S, Yarahmadi Z (2011) *Match Commun Math Comput Chem* 66:613
4. Furtula B, Gutman I (2011) *J Chemom* 25:87
5. Rasulev BF, Toropov AA, Hamme AT II, Leszczynski J (2008) *QSAR Comb Sci* 27:595
6. Melagraki G, Afantitis A (2011) *Curr Med Chem* 18:2612
7. Afantitis A, Melagraki G, Koutentis PA, Sarimveis H, Kollias G (2011) *Eur J Med Chem* 46:497
8. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *J Comput Chem* 32:2727
9. Toropov AA, Toropova AA, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *Chemom Intell Lab* 109:94
10. Weininger D (1988) *J Chem Inf Comput Sci* 28:31
11. Weininger D, Weininger A, Weininger JL (1989) *J Chem Inf Comput Sci* 29:97
12. Weininger D (1990) *J Chem Inf Comput Sci* 30:237
13. Roy K, Mitra I (2012) *Mini-Rev Med Chem* 12:491
14. Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, Skrzyński M, Worth AP (2011) *Struct Chem* 22:795
15. Toropov AA, Rasulev BF, Leszczynski J (2008) *Bioorg Med Chem* 16:5999
16. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *Cent Eur J Chem* 9:846
17. Toropov AA, Toropova AP, Benfenati E (2010) *Eur J Med Chem* 45:3581
18. Mitra I, Saha A, Roy K (2010) *Mol Simul* 36:1067