Inter*net* Electronic Journal of Molecular Design

October 2002, Volume 1, Number 10, Pages 527–544

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 2

Guest Editor: Jun–ichi Aihara

A QSAR Study on a Set of 105 Flavonoid Derivatives Using Descriptors Derived From 3D Structures

Marjan Vračko[1] and Johann Gasteiger[2]

[1] National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia
[2] Computer Chemie Centrum und Institut fuer Organische Chemie, Universitaet Erlangen – Nuernberg, Naegelsbachstr. 25, D–91052 Erlangen, Germany

Received: July 17, 2002; Revised: September 22, 2002; Accepted: October 3, 2002; Published: October 31, 2002

Citation of the article:
M. Vračko and J. Gasteiger, A QSAR Study on a Set of 105 Flavonoid Derivatives Using Descriptors Derived From 3D Structures, *Internet Electron. J. Mol. Des.* **2002**, *1*, 527–544, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

A QSAR Study on a Set of 105 Flavonoid Derivatives Using Descriptors Derived From 3D Structures[#]

Marjan Vračko[1,*] and Johann Gasteiger[2]

[1] National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia
[2] Computer Chemie Centrum und Institut fuer Organische Chemie, Universitaet Erlangen – Nuernberg, Naegelsbachstr. 25, D–91052 Erlangen, Germany

Received: July 17, 2002; Revised: September 22, 2002; Accepted: October 3, 2002; Published: October 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (10), 527–544

Abstract

A relationship between the 3D structure and biological activity was studied for a set of 105 flavonoid derivatives using a counterpropagation neural network. The 3D structures were determined in two ways, either by the empirical structure generator CORINA or by optimization within the semiempirical AM1 approximation. Furthermore, we compared two types of structure representations, the radial distribution function (RDF) method and the 'spectrum–like' representation. We show how different methods for 3D structure determination and different representations influence the quality of QSAR models. For all methods considered we found comparable models for the relationship between structure and biological activity. The computation times in 3D structure determination are visibly shorter for CORINA as against the AM1 method.

Keywords. 3D structure representation; CORINA; neural networks; flavonoid derivatives.

Abbreviations and notations

AM1, Austin model	PEOE, partial equalization of orbital electronegativity
HF, Hartree–Fock	QSAR, quantitative structure–activity relationships
MPPT2, MPPT3, Moeller–Plesset perturbation theory	QSPR, quantitative structure–property relationships
NMR, nuclear magnetic resonance	RDF, radial density function

1 INTRODUCTION

Drug–receptor interaction is a process of molecular recognition. A drug molecule interacts with the biological target in a specific conformation [1]. It is clear that for understanding this interaction the knowledge of the 3D structure is of crucial importance. The most reliable data on 3D structures are obtained by X–ray diffraction or by multi–dimensional NMR measurements. Unfortunately, these kind of experimental data are rather scarce. Therefore, computational modeling techniques play an important role in drug research.

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; phone: 00386–1–4760315; fax: 00386–1–4259244; E–mail: marjan.vracko@ki.si.

Bearing in mind the 3D drug–receptor picture we can select different strategies in the search for optimum drug. If the 3D geometry of the active site of a receptor is known one can try to tailor the proper drug molecule. Different methods of this kind are often referred to as the receptor–dependent methods [2]. In most cases, the exact structure of the active site is not known, but we know the structures of some highly active molecules. The consideration in such a case follows the idea that 'similar structures show similar activity'. This strategy is often referred to as receptor–independent and represents a basis for quantitative structure–property/activity relationship (QSPR/QSAR) studies. Mathematically, this method has to search for the relationship between descriptors and a property/activity [3].

In a standard QSAR model the descriptors are parameters calculated from a topological [4], a geometric, or a quantum chemical picture [5] of a molecule, but descriptors carry the information on the molecular structure only implicitly. To include the precise 3D structure into QSPR/QSAR studies 3D QSAR seems to be a promising method [1]. Here, molecules are embedded into a 3D box with a grid of points. A molecule is represented by the values of a molecular field (or a similar quantity) in each point [6,7]. Alternatively, a molecular 3D structure can be represented with a 'spectrum–like' representation [8], which is briefly described in section Methods. The next question in QSPR/QSAR modeling is the selection of a proper mathematical tool for the statistical treatment of data. Beside the multivariate linear regression methods several methods for clustering, partitioning, and modeling are available [9].

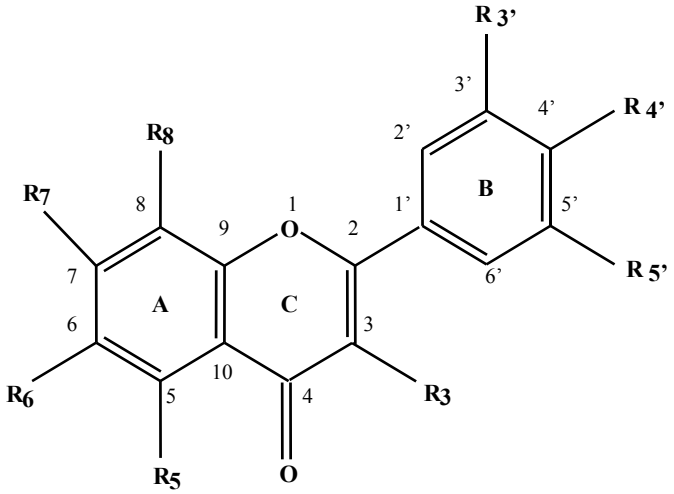
In the present work we treat a set of 105 flavonoid derivatives [10,11] with a counterpropagation neural network [12] using radial distribution functions (RDF) and the 'spectrum–like' representation. A brief description of methods is given in section Methods.

Many flavonoid derivatives play important roles in different biochemical processes and some of them have therapeutic effects (see for example Moon *et al*. [13]). Here, we consider the inhibitory activity of flavonoids against protein–tyrosine kinase $p56^{lck}$. Several QSAR studies were reported on this set of molecules using different descriptors and different methods of modeling. Nikolovska–Coleska *et al*. [14] treated a set of 104 derivatives, which are all included in the set treated in this study, with a standard linear regression technique using classical and quantum chemical descriptors. Novič *et al*. [15] and Oblak *et al*. [16] treated the same data with a counterpropagation neural network and with CODESSA software, respectively. Vračko *et al*. [17] treated the electronic structures of 28 derivatives (a subset of the set treated in this study) with different chemometric methods, such as linear regression, counterpropagation neural networks and principal component analysis. Amić *et al*. [18] reported the QSAR study on the set of 19 flavonoid derivatives using linear regression method. A comprehensive *ab initio* study of 3D structures of some flavonoids is reported by Meyer [19].

2 MATERIALS AND METHODS

In Table 1 we used the set of 105 flavonoid derivatives considered with their biological activities, *i.e.*, the inhibitory activities against protein–tyrosine kinase $p56^{lck}$ [10,11]. The biological activity is given as $\log(1/IC_{50})$, where IC_{50} is the molar concentration of the flavonoids necessary to give half–maximum inhibition. The compounds with activity equal or less than 2.70 are regarded as non–active.

Table 1. 105 flavonoid derivatives with experimental biological activity (Exp)
Ac denotes acetyl, Bn denotes benzyl, R denotes *tert*-butyldimethylsilyl.



No	Substituent	Exp
1	5-OH; 7-OH; 3'-OH; 4'-OH	4.88
2	3-OH; 7-OH; 3'-OH; 4'-OH	4.86
3	5-OH; 7-OH; 4'-OH	4.83
4	5-OH; 4'-OH	4.80
5	6-OH; 3'-OH	4.80
6	5-OH; 7-OH	4.71
7	5-OH; 7-OH; 3'-OH; 4'-OH	4.46
8	7-OH; 3'-OH	4.41
9	6-OH; 3'-OCH3; 4'-OCH3; 5'-OCH3	4.22
10	3-OH; 5-OH; 7-OH; 3'-OCH3; 4'-OH; 5'-OCH3	4.16
11	3-OH; 5-OH; 7-OH; 3'-OH; 5'-OH	4.00
12	6-OH; 4'-OH	3.93
13	7-OH; 8-OH; 3'-OCH3; 4'-OH; 5'-OCH3	3.92
14	6-OH; 4'-OR	3.92
15	6-OH; 3'-OCH3; 4'-OH; 5'-OCH3	3.89
16	7-OH; 4'-OH	3.78
17	7-OH; 8-OH; 3'-OH	3.75
18	3-OH; 5-OH; 7-OH	3.53
19	5-OH; 7-OCH3; 4'-OH	3.55
20	5-OH; 3'-OH	3.50
21	7-OH; 8-OH	3.50
22	7-OH	3.47
23	6-OH; 3'-OCH3; 4'-OR; 5'-OCH3	3.43
24	7-OH; 8-OH; 3'-OCH3; 4'-OCH3; 5'-OCH3	3.40
25	7-OH; 4'-OR	3.01
26	7-OH; 3'-OCH3; 4'-OH; 5'-OCH3	2.90
27	7-OH; 3'-OCH3; 4'-OR; 5'-OCH3	2.82

Table 1. (Continued)

No	Substituent	Exp
28	7-OH; 4'-OBn	2.69
29	3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
30	7-OH; 8-OH; 3'-OCH3; 4'-OR; 5'-OCH3	2.70
31	7-OAc; 8-Ac; 3'-OCH3; 4'-OR; 5'-OCH3	2.70
32	6-OCH3; 3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
33	7-OH; 3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
34	7-OAc; 3'-OCH3; 4'-OH; 5'-OCH3	2.70
35	7-OAc; 3'-OCH3; 4'-OR; 5'-OCH3	2.70
36	7-OCH3; 3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
37	5-OH; 4'-OBn	2.70
38	6-OH; 4'-NH2	5.92
39	5-OH; 7-OH; 4'-NH2	5.13
40	3'-OCH3; 4'-OH; 5'-OCH3	4.57
41	7-OH; 4'-NH2	3.86
42	4'-NH2	3.68
43	3-COOCH3; 4'-OH	3.36
44	4'-OH	3.30
45	3-COOCH3; 4'-NH2	3.09
46	3-COOH; 7-OCH3; 4'-OH	2.99
47	3-COOH; 3'-OCH3; 4'-OH	2.80
48	3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
49	3-COOCH3; 3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
50	3-COOCH3; 3'-OCH3; 5'-OCH3	2.70
51	3-COOCH3; 3'-OCH3; 4'-OCH3	2.70
52	3-COOCH3; 4'-OCH3	2.70
53	3-COOCH3; 4'-Br	2.70
54	3-COOCH3; 4'-OBn	2.70
55	3-COOCH3; 7-OCH3; 4'-OBn	2.70
56	3-COOCH3; 6-OCH3; 4'-OBn	2.70
57	3-COOCH3; 4'-NO2	2.70
58	3-COOCH3; 7-OCH3; 4'-NO2	2.70
59	3-COOCH3; 6-OCH3; 4'-NO2	2.70
60	3-COOCH3; 5-OBn; 7-OBn; 4'-NO2	2.70
61	3-COOH; 3'-OCH3; 4'-OCH3; 5'-OCH3	2.70
62	3-COOH; 3'-OCH3; 5'-OCH3	2.70
63	3-COOH; 3'-OCH3; 4'-OCH3	2.70
64	3-COOH; 4'-OCH3	2.70
65	3-COOH; 4'-Br	2.70
66	3-COOH; 4'-NO2	2.70
67	3-COOH; 7-OCH3; 4'-NO2	2.70
68	3-COOH; 6-OCH3; 4'-NO2	2.70
69	3-COOCH3; 7-OCH3; 4'-OH	2.70
70	3-COOCH3; 6-OCH3; 4'-OH	2.70
71	3-COOCH3; 7-OCH3; 4'-NHAc	2.70
72	3-COOCH3; 6-OCH3; 4'-NHAc	2.70
73	3-COOH; 5-OH; 7-OH; 4'-NO2	2.70
74	4'-NO2	2.70
75	7-OH; 4'-NO2	2.70
76	6-OH; 4'-NO2	2.70
77	5-OH; 7-OH; 4'-NO2	2.70
78	5-NH2; 6-OH; 7-NH2; 4'-NH2	4.74
79	5-NH2; 6-OH; 7-NH2; 3'-NH2	4.34
80	6-OCH3; 8-NH2; 3'-NH2	4.25
81	6-NH2; 4'-NH2	3.99
82	6-NH2; 8-NH2; 4'-NH2	3.97
83	6-OH; 8-NH2; 4'-NH2	3.93
84	8-NH2; 4'-NH2	3.91

Table 1. (Continued)

No	Substituent	Exp
85	6-NH2; 7-OH; 4'-NH2	3.85
86	6-NH2; 3'-NH2	3.70
87	5-OH; 6-NH2; 4'-NH2	3.65
88	5-OH; 8-NH2; 4'-NH2	3.49
89	7-OH; 8-NH2; 4'-NH2	3.48
90	6-OCH3; 8-NH2; 4'-NH2	3.42
91	6-NH2; 7-OH; 3'-NH2	3.30
92	6-NH2; 7-OH; 8-NH2; 4'-NH2	3.12
93	6-NO2; 7-OH; 8-NO2; 4'-NO2	2.81
94	5-CH3; 8-NH2; 4'-NH2	2.79
95	7-OH; 8-NO2; 4'-NO2	2.73
96	6-NO2; 4'-NO2	2.70
97	8-NO2; 4'-NO2	2.70
98	6-NO2; 7-OH; 4'-NO2	2.70
99	5-OH; 8-NO2; 4'-NO2	2.70
100	6-OCH3; 8-NO2; 4'-NO2	2.70
101	5-OCH3; 8-NO2; 4'-NO2	2.70
102	6-NO2; 8-NO2; 4'-NO2	2.70
103	6-OH; 8-NO2; 4'-NO2	2.70
104	5-OH; 6-NO2; 4'-NO2	2.70
105	5-NO2; 6-OH; 7-NO2; 4'-NO2	2.70

3 METHODS

3.1 Molecular Structures

The constitution of a molecule is described by a molecular graph or by a connection table. It contains information on bonds, but it carries no information on metric properties such as bond distances and bond angles. On the other hand, the 3D structure, which is defined by the coordinates of all atoms, carries the complete information on all distances between any pair of atoms. The 3D molecular structures of 105 flavonoid derivatives (see Table 1) were calculated in two different ways. First, we used the program package CORINA [20]. This program determines geometry parameters of a molecule by only taking the connection table of a molecule's constitution as input. As a second method, we selected an optimization of the geometry by the semi–empirical AM1 approximation. The program package MOPAC with the standard input parameters for geometry optimization was used [21].

3.2 Atomic Charges

Beside the 3D structures, the atomic charges were also included into the calculations to account for electronic effects. When CORINA geometries were used, the charges were calculated with the PEOE method [22] contained in the program package PETRA [23]. As input, the PEOE method only needs the constitution of a molecule as expressed by a connection table. When AM1 geometries were used, the charges were calculated within the AM1 approximation.

3.3 Radial Distribution Functions (RDF Descriptors)

The RDF code represents a molecular structure by a radial distribution function. This function can be interpreted as a probability distribution to find an atom in a spherical volume of radius r. Since the details of the RDF representation are described elsewhere [24,25] only the basic equation is given here:

$$g(r) = f\sum_{i}^{N-1}\sum_{j>i}^{N} A_i A_j e^{-B(r-r_{ij})^2} \tag{1}$$

where r_{ij} represents the distance between atoms *i* and *j*, *N* is the number of atoms in a molecule, A_i and A_j are properties associated with the atoms *i* and *j*, respectively. In our study all the parameters *A* were set to one. Furthermore, *B* is the smoothing and *f* is the scaling factor. In our calculations *B* was equal 25 $Å^{-2}$ and *f* was set to one, which were the optimal values according to study reported by Hemmer *et al* [24]. The RDF representation is uniform and invariant under translation and rotation of molecules. In the computational treatment here, a distribution function is given in a discrete form, *i.e.*, it is given as a vector with equidistant values of *r*. The dimension of the vector was set to 64 and the distribution function *g*(*r*) was defined in the interval from 0.0 Å to 12.6 Å.

3.4 The 'Spectrum–Like' Representation

As a second method for structure description, a 'spectrum–like' representation of 3D structures was used. Since the details of the representation have recently been published only some basic ideas will be given here [8]. A representation of a molecule is constructed in three steps. First, from the 3D molecular structure one constructs three projections, on the *xy*, on the *xz*, and on the *yz* plane. In the second step, each projection (figure) is treated separately. A figure is put into a circle of arbitrary radius. A projection beam from the center of a circle produces a pattern of points on the circle where each point represents a particular atom. In the third step each point on the circle is taken as a center for a Lorentzian curve of the form:

$$s_i(\varphi) = \frac{\rho_i}{(\varphi - \varphi_i)^2 + \sigma_i^2} \tag{2}$$

where ρ_i and φ_i are the distance between the origin of the coordinate system and the position of the i–th atom and its polar angle, respectively, σ_i is a free parameter, which can be associated with any atomic property. If we consider only molecular geometries, the σ_i values are set to one, otherwise we selected atomic charges as atomic property. It can be shown that the representation is uniform, unique and reversible. The spectrum related to the figure is a sum of all atomic Lorentzians and it is defined in the interval (0, 2π). By selecting k equidistant points on this interval, each projection is represented by a *k*–dimensional vector. The complete 3D structure is represented with three spectra. In our studies the parameter *k* was set to 60, *i.e.*, a structure was represented with three vectors composed into a vector in 180 dimensional space.

3.5 Counterpropagation Neural Network

The neural network with counterpropagation learning strategy has been described in many articles and textbooks [26,27]. It is a suitable tool for clustering, classification and QSAR modelling, where the neural network models the functional relationship between input and output variables [28,29]. It consists of two layers of neurons (a Kohonen and an output layer), which are arranged in a two–dimensional rectangular matrix. The Kohonen layer (input layer) obtains the input variables that represent the considered objects. During the learning process the output values (in our case biological activities) are given to the output layer, which has the same arrangement of neurons as the Kohonen layer.

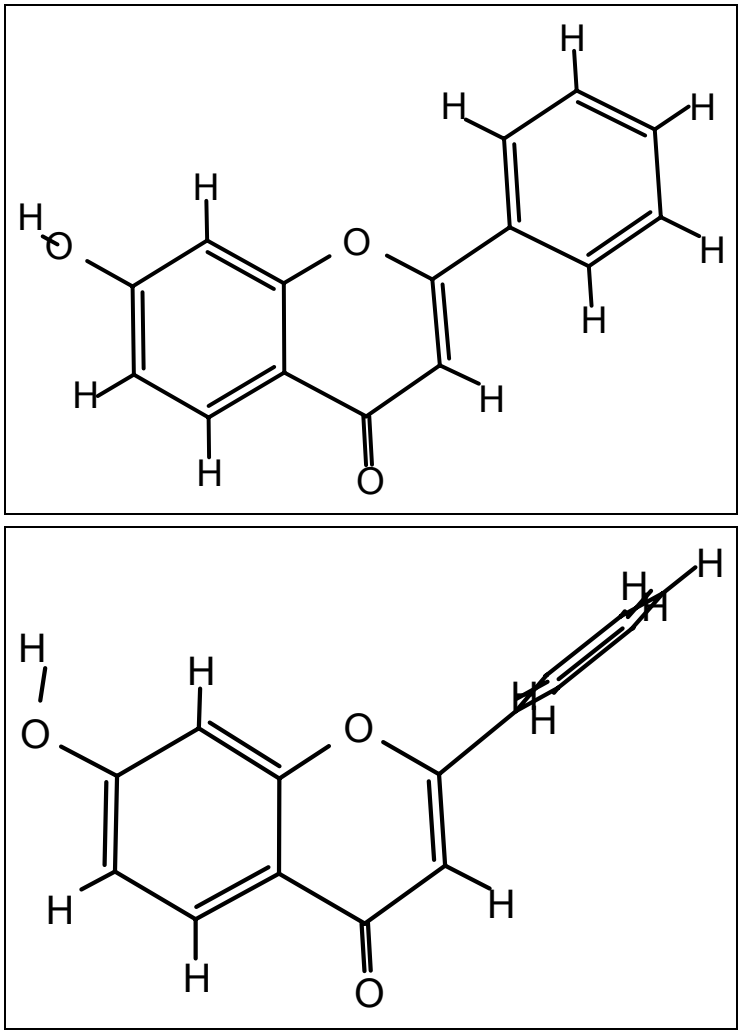


Figure 1. Compound **22**. The 3D structure on the top was calculated by CORINA, the one in the bottom by AM1 optimization procedure.

Learning in the Kohonen layer is done in the same way as in a Kohonen network. This means, a vector of input variables is presented to all neurons. The program selects that neuron that has weights, which are the most close to the input values (winning neuron). The position of the winning neuron is transferred from the Kohonen to the output layer, and the weights in the output layer are

corrected in such a way that they are becoming similar to the given values (biological activity). After the weights are stabilized, the counterpropagation neural network is considered to be trained. In a trained network the objects with similar input vectors are located close to each other. It is expected that they also have similar output values.

In the present study the dimension of network was set to 15×15 neurons. The networks were trained during 400 learning epochs.

4 RESULTS

4.1 3D Structures and Atomic Charges

For some of the compounds considered the geometries as calculated in both approximations, the CORINA model and the one calculated by the AM1 method, are quite different. An example is given in Figure 1. We illustrate the 3D structure of compound No. 22 (see Table 1) as calculated by CORINA program (Figure 1, top) and as derived by the AM1 method (Figure 1, bottom). As we see for this compound CORINA gives a coplanar position for all three rings (A, C, and B) and the hydrogen atom of hydroxy–group, which is attached to position 7, is oriented out of the plane.

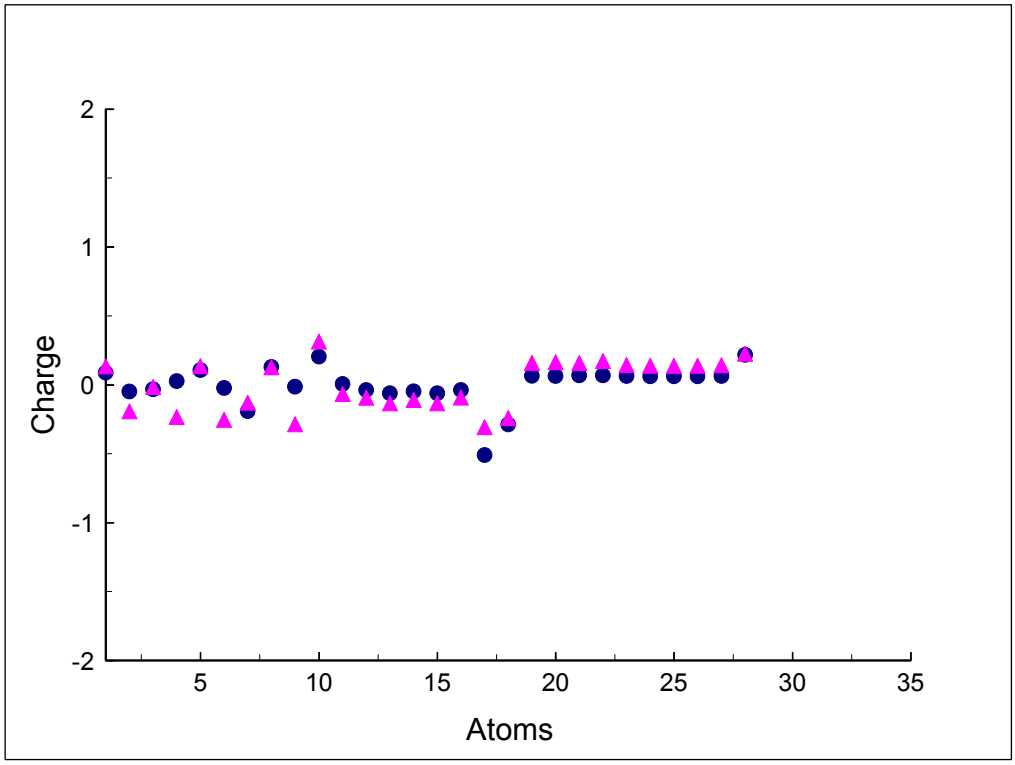


Figure 2. Atomic charges of compound 22 (● represents PEOE charges, ▲ represents AM1 charges).

In the entire set of 105 compounds CORINA produces 71 compounds where the rings A, C, and B are coplanar. In contrast the three rings in the 3D structures obtained with the AM1 approximation are not planar. For example in the compound **22** the ring B is twisted out of the plane while the hydroxy–group is situated in the plane defined by rings A and C. In the entire set of 105 compounds none of the compounds has coplanar rings A, C, and B although in six cases the rings are almost planar. Meyer [19] reported that for this kind of compounds the AM1 optimization provides reasonable results compared to the results obtained with the high level *ab initio* methods, such as HF, MPPT2, and MPPT3 with extended basis sets. He also reported that the energy barrier between coplanar and twisted conformations is very small (~ 0.7 kcal/mol).

The atomic charges were calculated by the PEOE method and in the AM1 approximation. Figure 2 shows the charges in both approaches for the compound **22**. The atomic charges calculated with different methods come out to be quite similar.

4.2 List of Models Considered

In this study six models have been analyzed: (*A*) the 3D geometries were determined with the CORINA geometries, and represented with the RDF method; (*B*) the 3D geometries were determined with the AM1 geometries, and represented with the RDF method; (*C*) the 3D geometries were determined with the CORINA, and represented with the 'spectrum–like' method; (*D*) the 3D geometries were determined with the CORINA, and described with the 'spectrum–like' representation. In addition, the PEOE charges were included into the representation; (*E*) the 3D geometries were optimized within the AM1 approximation, and described with the 'spectrum–like' representation; (*F*) the 3D geometries were optimized within the AM1 approximation, and described with the 'spectrum–like' representation. AM1 charges were included into the representation. The models were tested in their recall ability and their prediction ability.

4.3 The Recall Ability Test

In the first step we analyzed the recognition ability of the models looking for molecules that are recognized as identical by the counterpropagation neural network. Such molecules are situated in the same neuron in the neural networks. The inability to distinguish molecules is not a shortcoming of neural networks. It simply means that the representations of some molecules are too similar to be discriminated by the neural networks. A conflict situation occurs when two (or more) molecules with very different biological activities are located in the same neuron. More exactly, we define a conflict when a non–active compound with a value of activity 2.70 or less and an active compound with activity value larger than 3.00 are located in the same neuron. In such a case, one molecule is an outlier, but the question is which one. To answer this we analyze the neighbors in the network. If the closest neighbors are non–active, the active compound is an outlier and vice versa [29].

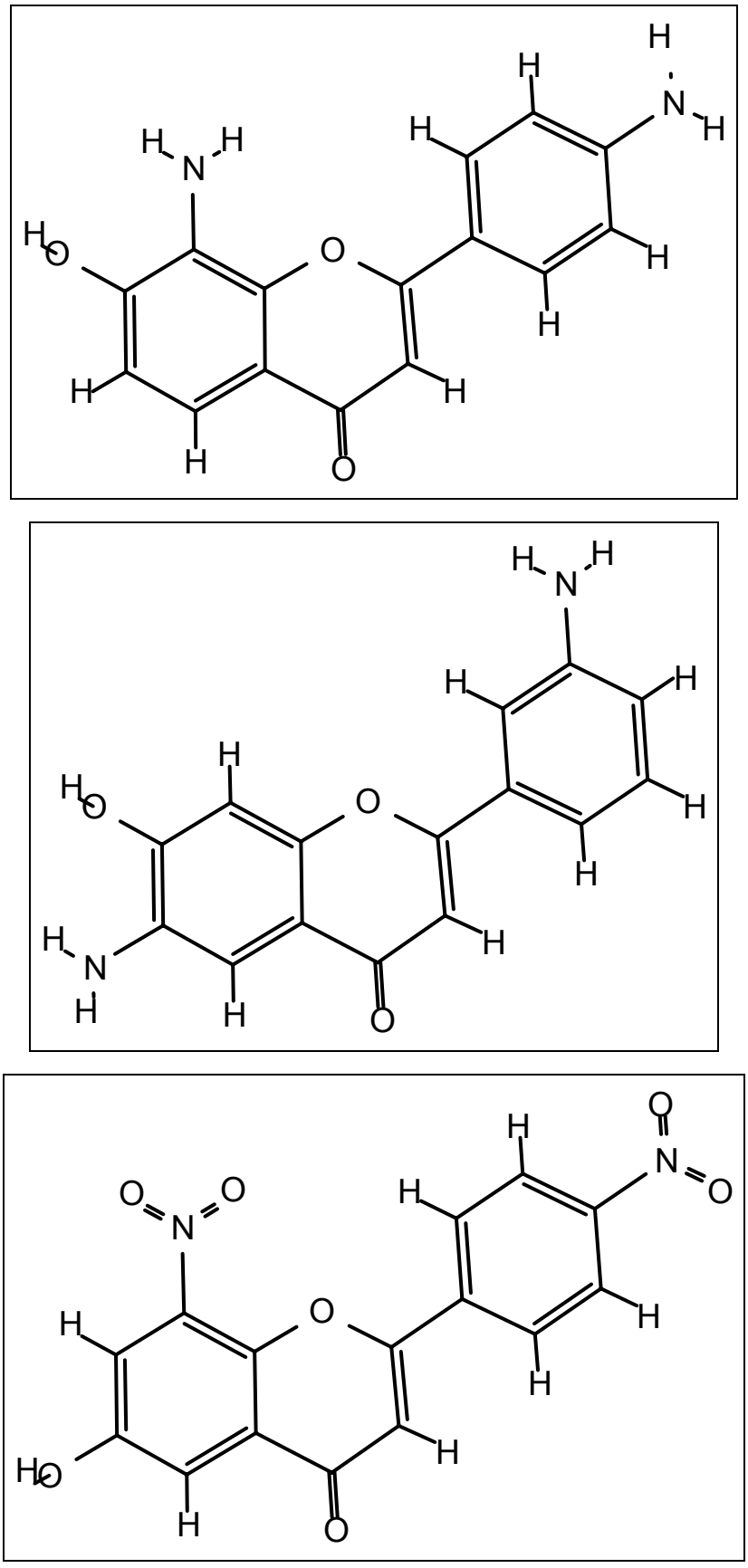


Figure 3. Molecules **89** (top), **91** (in-between) and **103** (bottom), which are recognized as equivalent by neural network. Results are from model C.

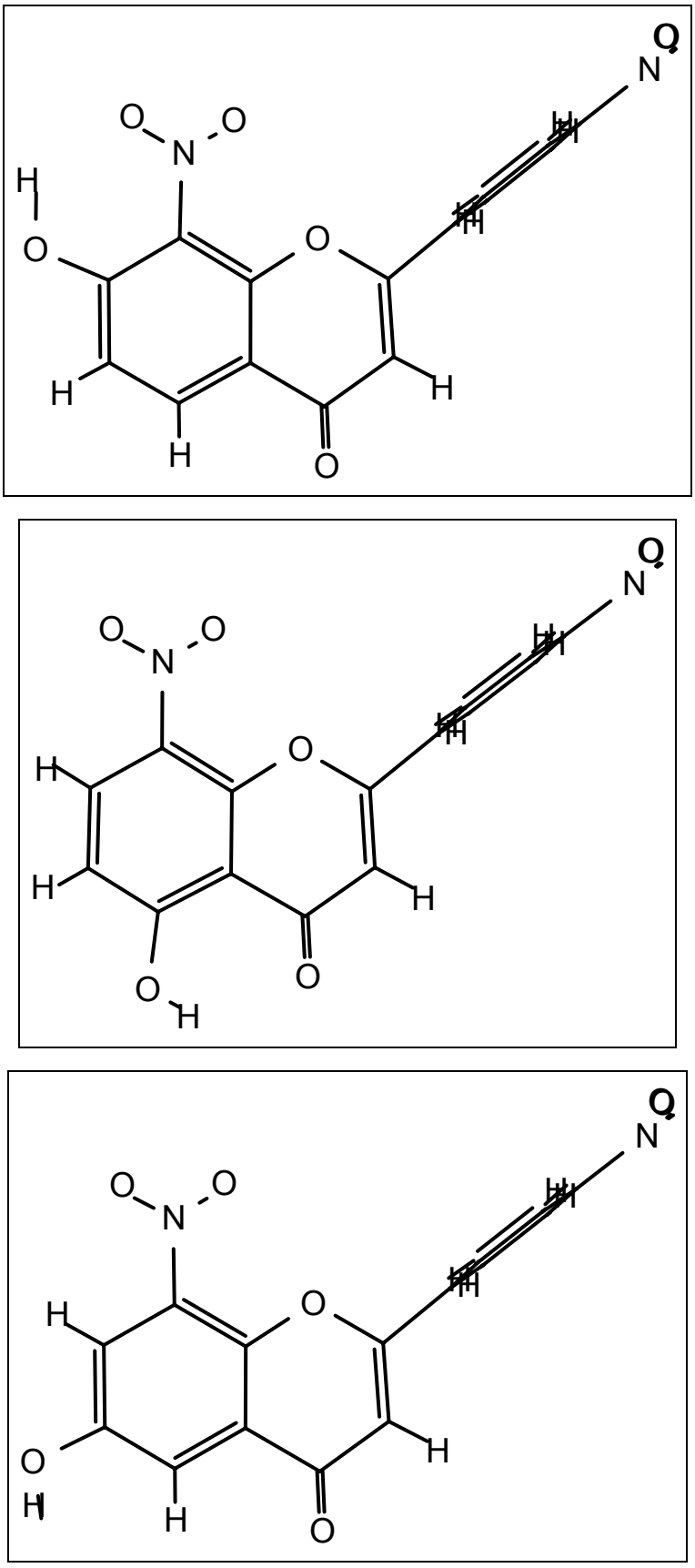


Figure 4. Molecules **95** (top), **99** (in-between) and **103** (bottom), which are recognized as equivalent by neural network. Results are from model E.

Table 2. The groups of molecules which are recognized as equivalent. * denotes outliers (experimental activities are in parentheses). Models A-F are described in text. In model B the compound **38** is an outlier as described in text.

	Model A	Model B	Model C	Model D	Model E	Model F
1	**1** (4.88) **11** (4.00)	**1** (4.88) **2** (4.86)	**1** (4.88) **11** (4.00) **18** (3.53)	**1** (4.88) **11**(4.00)	**1** (4.88) **7** (4.46)	**1** (4.88) **7** (4.46)
	20 (3.50) **21** (3.50) **22** (3.47)	**17** (3.75) **18**(3.53)			**2** (4.86) **17** (3.75) **21** (3.50)	**17** (3.75) **21** (3.50)
						20 (3.50) **22** (3.47)
2	**23** (3.43)* **27** (2.82) **30** (2.70)	**23** (3.43)* **27** (2.82) **30** (2.70)	**23** (3.43) **27** (2.82)	**23** (3.43) **27** (2.82)	**23** (3.43) **27** (2.82)* **30** (2.70)*	**27** (2.82) **30** (2.70)
3	**88** (3.49) **89** (3.48)	**84** (3.91) **88** (3.49)	**83** (3.93) **88** (3.49)	**82** (3.97) **98** (2.70)*	**81** (3.99) **82** (3.97)	**83** (3.93) **88** (3.49)
						84 (3.91) **89** (3.48)
4	**55** (2.70) **56** (2.70)	**54** (2.70) **56** (2.70)		**55** (2.70) **56** (2.70)		
5	**98** (2.70) **104** (2.70)	**96** (2.70) **104** (2.70)				
6			**80** (4.25) **90** (3.42)	**80** (4.25) **90** (3.42)	**68** (2.70)* **80** (4.25) **90** (3.42)	**68** (2.70)* **90** (3.42)
				67 (2.70) **68** (2.70)		
7		**78** (4.74) **79** (4.34)	**78** (4.47) **79** (4.34) **105** (2.70)*	**78** (4.74) **79** (4.34)		**78** (4.74) **79** (4.34)
8			**19** (3.55) **100** (2.70)*	**91** (3.30) **100** (2.70)*	**86** (3.70) **91** (3.30)	**86** (3.70) **91** (3.30)
9	**14** (3.92) **25** (3.01)	**14** (3.92) **25** (3.01)	**14** (3.92) **25** (3.01)	**14** (3.92) **25** (3.01)		
10	**45** (3.09) **57** (2.70)	**43** (3.36) **45** (3.09)	**45** (3.09) **57** (2.70)	**43** (3.36)* **53** (2.70)	**45** (3.09)* **52** (2.70) **57** (2.70)	**52** (2.70) **57** (2.70)
11			**38** (5.92) **74** (2.70)* **76** (2.70)*	**12** (3.93)* **38** (5.92) **76** (2.70)*	**38** (5.92) **76** (2.70)*	
12	**9** (4.22)* **48** (2.70)	**9** (4.22)* **48** (2.70)		**9** (4.22) **24** (3.40)	**9** (4.22) **48** (2.70)*	**9** (4.22)* **48** (2.70)
13	**32** (2.70) **36** (2.70)			**32** (2.70) **36** (2.70)		
14	**6** (4.71) **8** (4.41)	**6** (4.71) **8** (4.41) **20** (3.50)*	**6** (4.71) **16** (3.78)			**6** (4.71) **18** (3.53)
15	**47** (2.80) **65** (2.70)	**47** (2.80) **65** (2.70)		**47** (2.80) **65** (2.70)	**47** (2.80) **65** (2.70)	**47** (2.80) **65** (2.70)
16	**43** (3.36)* **53** (2.70)		**43** (3.36)* **53** (2.70)		**43** (3.36)* **53** (2.70)	**43** (3.36)* **53** (2.70)
17	**28** (2.69) **37** (2.70)				**28** (2.69) **37** (2.70)	
18	**15** (3.89) **40** (4.57)					**26** (2.90)* **40** (4.57)
	13 (3.92) **26** (2.90)*					

Table 2. (Continued)

	Model A	Model B	Model C	Model D	Model E	Model F
19	**4** (4.80) **5** (4.80)	**5** (4.80) **44** (3.30)*	**3** (4.83) **4** (4.80)	**3** (4.83) **4** (4.80)		
20				**92** (3.12) **93** (2.81)	**93** (2.81) **102** (2.70)	**93** (2.81) **102** (2.70)
21			**58** (2.70) **59** (2.70)	**58** (2.70) **69** (2.70)	**58** (2.70) **69** (2.70)	
				59 (2.70) **70** (2.70)		
22			**89** (3.48) **91** (3.30) **103** (2.70)*		**95** (2.73) **99** (2.70) **103** (2.70)	**95** (2.73) **103** (2.70)
23	**85** (3.85) **87** (3.65)	**85** (3.85) **87** (3.65)	**87** (3.65) **98** (2.70)*		**87** (3.65) **104** (2.70)*	
24	**12** (3.93) **44** (3.30)	**12** (3.93) **42** (3.68)		**42** (3.68) **74** (2.70)*	**42** (3.68)* **74** (2.70) **75** (2.70)	
25	**38** (5.92)* **41** (3.86)	**39** (5.13) **41** (3.86)*	**39** (5.13) **75** (2.70)*	**39** (5.13) **41** (3.86)*		
26		**97** (2.70) **99** (2.70)				
27	**49** (2.70) **61** (2.70)					**49** (2.70) **61** (2.70)
28	**24** (3.40)* **33** (2.70)					
29				**31** (2.70) **35** (2.70)		
30					**94** (2.79) **101** (2.70)	

The groups of molecules that are recognized as identical are given in Table 2. With six models 112 groups were found. Table 2 is organized in such a way that we can follow the groups of identical molecules through different models. Groups with at least one common molecule are placed into the same row. For models A, B, C, D, E, and F we found 22, 17, 16, 21, 18, and 18 such groups, respectively. Each group can be analyzed comparing the biological activities of molecules in the group. If the differences in activity within the group are small we conclude that the correlation between structures and activities is good for particular compounds. In 56 (50%) groups the difference in activity lies between 0.0 and 0.5, in 19 (17%) groups the difference lie in the interval 0.5 and 1.0, and in two groups (1%) the difference is between 1.0 and 1.5. 35 (32%) groups show the conflict situation described above. In the group 2 (Table 2) in models A, B the compound 23 is an outlier. The closest neighbors in the network are compounds: **31**, **35**, **36**, and **37**, all with the activity value 2.70. In the model E the closest neighbors are compounds **14** and **25** with activities 3.92 and 3.01, respectively. Contrary to the model A the compounds **27** and **30** can be considered as outliers. Looking to the next neighbors in the map of model E the compounds **14**, **23**, and **25** build a cluster surrounded by non–active compounds (**31**, **35**, **37**, **58**, **71**, **72**).

4.4 Training – Test Set Division

To analyze the prediction ability of models we divided the set into outliers, training and test set. The division of a set into training and test set was performed in three steps. In the first step, we selected the outliers as described above. In the second step, we have chosen a Kohonen network to divide the objects into the training and the test set. Here, the entire Kohonen map is divided into sub–parcels selecting the objects for training set from each sub–parcel equivocally. It is expected that such a training set possesses the information content of the entire set (for details see Simon *et al*. [30]). This division of compounds in training and test set was used for training and testing the model. In the next step, the initial training set was improved by adding a few compounds which were formerly placed in the test set. In this way, the information content of the training set was improved.

Table 3. Number of compounds in training/test sets and outliers for models A-F. Models A-F are described in text.

Model	Training set	Test set	Labels (see Table 1) of outliers
A	66	33	**9**, **23**, **24**, **26**, **38**, **43**
B	63	36	**9**, **20**, **23**, **38**, **41**, **44**
C	62	35	**43**, **74**, **75**, **76**, **98**, **100**, **103**, **105**
D	74	24	**12**, **41**, **43**, **74**, **76**, **98**, **100**
E	62	34	**27**, **30**, **42**, **43**, **45**, **48**, **68**, **76**, **104**
F	76	25	**9**, **26**, **43**, **68**

The number of compounds in the training/test sets and the outliers for all models are given in Table 3. Following the method described above in the model B five compounds were viewed as outliers. In addition, the compound number **38** with the activity 5.92 was also set as an outlier. In this case it was not a conflict in the sense described above, but the compound **38** was situated in a neighborhood in the neural network that was dominated with non–active compounds. The remaining 99 compounds were divided then into the training set with 63 compounds and the test set with 36 compounds. Only in the models F the compound **38** is located in neighborhoods with highly active compounds.

Table 4. Statistical parameters (R, correlation coefficient; b_0 and b_1, parameters of the linear correlation between experimental and predicted values; F, Fisher criterion; RSS, residual sum of squares; MS, mean squares). Models A-F are described in text.

		r	b_0	b_1	F	RSS	MS
A	Training set	0.997	0.058	0.982	10087.814	0.256	0.004
	Test set	0.854	0.180	1.013	83.740	2.411	0.078
B	Training set	0.985	0.154	0.954	2028.155	1.285	0.021
	Test set	0.883	0.383	0.954	120.739	3.923	0.115
C	Training set	0.941	0.479	0.856	466.733	6.081	0.085
	Test set	0.775	1.007	0.693	30.080	4.471	0.203
D	Training set	0.987	0.111	0.967	2856.868	1.119	0.016
	Test set	0.887	0.277	0.952	81.229	2.385	0.108
E	Training set	0.919	0.541	0.838	394.884	7.464	0.102
	Test set	0.904	0.837	1.062	102.864	1.652	0.072
F	Training set	0.965	0.240	0.928	1002.375	3.247	0.044
	Test set	0.907	–0.292	1.183	108.346	1.181	0.056

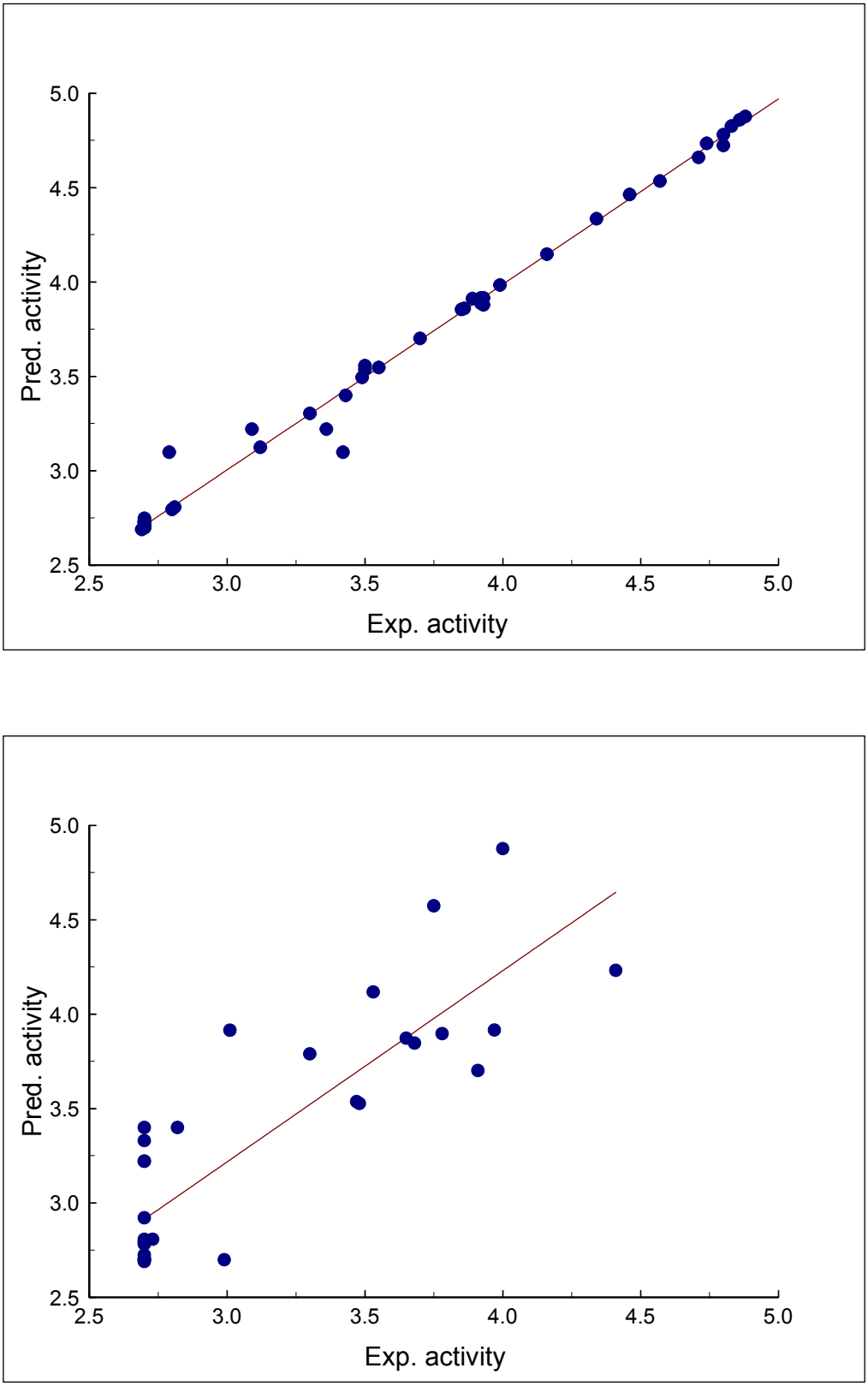


Figure 5. Predicted versus experimental activity for training set (top) and test set (bottom) for model A.

Table 4 shows the statistical parameters for the seven models considered. As the statistical parameters we report the correlation coefficient r, Fisher criterion F, residual sum of squares RSS, mean squares MS, and the parameters b_0 and b_1 of the linear correlation between predicted and experimental biological activities [31]. Here, b_0 and b_1 are the intercept and the slope of the line, respectively. An example of regression lines for training and test set is shown in Figure 5.

5 DISCUSSION AND CONCLUSIONS

The question addressed in this work is how different methods of determination of 3D molecular structures influence the models. The 3D structures were determinated in two alternative ways, with the program CORINA and also by optimization with the AM1 method. Superior models could suggest that corresponding structure determination method and corresponding geometries are 'correct'. This means that the particular geometries probably describe the drug–receptor situation better. From analyzing the 3D geometry parameters we could thus gather information on the unknown receptor geometry. According to the remarks in the Introduction this leads into receptor dependent QSAR. The necessary condition for useful QSAR applications is that the representation of 3D structures is sensitive enough to differences in geometry parameters. This condition is satisfied in both used representations, RDF and ‘spectrum–like’ representation.

The models were tested on the recognition (recall) ability and on the prediction ability. The recall ability can be evaluated considering the groups of molecules that are recognized as identical (Table 2), and considering the statistical parameters for the training sets (Table 4). Generally, none of the models considered is superior to the others. The correlation coefficients for training sets for the CORINA models lie between $0.941 < r < 0.997$, and for the AM1 models between $0.919 < r < 0.985$. The prediction ability can be evaluated considering the statistical parameters of test sets. The correlation coefficients for CORINA models lie between $0.775 < r < 0.887$, and for AM1 models between $0.883 < r < 0.907$.

It was shown that the program CORINA generates geometrical parameters, which give models of comparable quality to the models built with geometry parameters after the quantum chemical optimization procedure. We should emphasize that the program CORINA is significant faster in generating 3D structures. For a molecule considered in this study an average CPU time on SUN IPX working station was 10 minutes using the AM1 procedure in comparison to a tenth of a second needed for the CORINA generator. When considering a large scale data sets with over 100,000 compounds the computation time may become an important factor.

Nikolovska–Coleska *et al*. [14] studied the set of 104 compounds, which are all included in the set reported here. Authors used limited number of classical and quantum chemical descriptors and multiple linear regression for modeling. For the entire set the best correlation coefficient was $r =$

0.750. For different subsets the reported coefficients were between $0.715 < r < 0.820$. Oblak *et al*. [16] reported correlation coefficient for training set of 70 compounds to be $r = 0.8988$. Novič *et al*. [15] considered the same set of 105 compounds using the same modelling technique, but different descriptors (classical and quantum chemical descriptors). The authors found the correlation coefficients for the training set to be larger than 0.9700, and for the test set between $0.8200 < r < 0.9100$. In our approach the correlation coefficients for test sets are comparable to those obtained by linear methods and the correlation coefficients for training sets are even higher.

Acknowledgment

MV thanks the Ministry of education, science and sport of Republic of Slovenia, which support of this work under contract: Program 034 507. The scientific visits have been supported by a Slovenian–Bavarian cooperation project. MV thanks M. C. Hemmer, T. Kleinoeder and C. Schwab from Erlangen for valuable discussions on using the programs RDF, PETRA, and CORINA.

5 REFERENCES

[1] D. Pitea, U. Cosentino, G. Moro, L. Bonati, E. Fraschini, M. Lasagni, and R. Todeschini; in: *Advanced Computer Assisted Techniques in Drug Discovery*, Ed. H. v.d. Waterbeemd, VCH, Weinheim, 1994, p.11.

[2] W. J. Dunn III and A. J. Hopfinger; in: *3D QSAR in Drug Design*. *Volume 5*. *Recent advances*, Eds. H. Kubinyi, G. Folkers, and Y. C. Martin, Kluver ESCOM, Leiden, 1998, p.168.

[3] C. Hansch and A. Leo, *Exploring QSAR*. *Fundamentals and Applications in Chemistry and Biology*, ACS professional reference book, American chemical society, Washington D.C., 1995.

[4] M. Randić, Topological Indices; in: The *Encyclopedia of Computational Chemistry*, Eds. P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H. F. Schaefer III, and P. R. Schreiner, Wiley & Sons, London, 1998, p. 3018.

[5] M. Karelson and V. S. Lobanov, Quantum chemical descriptors in QSAR/QSPR studies, *Chem*. *Rev*. **1996**, *96*, 1027–1043.

[6] R. D. Cramer III, D. E. Patterson, and J. D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape binding of steroids to carrier proteins, *J*. *Am*. *Chem*. *Soc*. **1988**, *110*, 5959–5967.

[7] K. Raghavan, J. K. Buolamwini, M. R. Fesen, Y. Pommier, K. W. Kohn, and J. N. Weinstein, Three–dimensional quantitative structure–activity relationship (QSAR) of HIV integrase inhibitors: a comparative molecular field analysis (CoMFA*)*, *J*. *Med*. *Chem*. **1995**, *38*, 890–897.

[8] J. Zupan and M. Novič, General type of a uniform and reversible representation of chemical structures, *Anal*. *Chim*. *Acta* **1997**, *348*, 409–418. Zupan, J. Zupan, M. Vračko, and M. Novič, New uniform and reversible representation of 3D chemical structures, *Acta Chim*. *Slov*. **2000**, *47*, 19–37.

[9] D. H. Drewry and S. S. Young, Approaches to the design of combinatorial libraries, *Chemom*. *Intell*. *Lab*. *Syst*. **1999**, *48*, 1–20.

[10] M. Cushman, H. Zhu, L. R. Geahlen, and J. Kraker, A synthesis and biochemical evaluation of a series of aminoflavones as potential inhibitors of protein–tyrosine kinases p56, EGFr, p60, *J*. *Med*. *Chem*. **1994**, *37*, 3353–3362.

[11] M. Cushman, D. Nagarathnam, L. D. Burg, and L. R. Geahlen, Synthesis and protein–tyrosine kinase inhibitory activities of flavonoid analogues, *J*. *Med*. *Chem*. **1991**, *34*, 798–806.

[12] J. Zupan and J. Gasteiger, *Neural networks in chemistry and drug design*, WILEY–VCH Verlag GmbH, Weinheim, 1999.

[13] T. Moon, M. H. Chi, D. H. Kim, C. N. Yoon, and Y. S. Choi, Quantitative structure–activity relationship (QSAR) study of flavonoid derivatives of inhibition of cytochrome P450 1A2, *Quant*. *Struct*. *Act*. *Relat*. **2000**, *19*, 257–263.

[14] Ž. Nikolovska–Coleska, L. Suturkova, K. Dorevski, A. Krbavcic, and T. Solmajer, QSAR of flavonoid inhibitors of $p56^{lck}$. Protein tyrosine kinase: a quantum chemical/classical approach, *Quant*. *Struct*.*–Act*. *Relat*. **1998**, *17*, 7–13.

[15] M. Novič, Ž. Nikolovska–Coleska, and T. Šolmajer, Quantitative structure–activity relationship of flavonoid $p56^{lck}$ protein tyrosine kinase inhibitors. A neural network approach, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1997**, *37*, 990–998.

[16] M. Oblak, M. Randic, and T. Solmajer, Quantitative structure–activity relationship of flavonoid analogues. 3. Inhibition of $p56^{lck}$ protein tyrosine kinase, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 994–1001.
[17] M. Vračko, M. Novič, and M. Perdih, Chemometrical treatment of electronic structures of 28 flavonoid derivatives, *Int. J. Quantum Chem.* **2000**, *76*, 733–743.
[18] D. Amić, D. Davidović–Amić, A. Jurić, B. Lučić, and N. Trinajstić, Structure–activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase, *J. Chem. Inf. Comput. Sci.* **1995**, 35, *35*, 1034–1038.
[19] M. Meyer, Ab initio study of flavonoid, *Int. J. Quantum Chem.* **2000**, *76*, 724–732.
[20] http://www2.ccc.uni–erlangen.de/software/corina.
[21] MOPAC 93, Version 2. Copyright © Fujitsu Limited 1993.
[22] J. Gasteiger and M. Marsili, Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges, *Tetrahedron* **1980**, *36*, 3219–3228.
[23] http://www2.ccc.uni–erlangen.de/software/petra.
[24] C. M. Hemmer, V. Steinhauer, and J. Gasteiger, Deriving the 3D structures of organic molecules from their infrared spectra, *Vibrat. Spectroscopy* **1999**, *19*, 151–164.
[25] C. M. Hemmer and J. Gasteiger, Prediction of three–dimensional molecular structures using information from infrared spectra, *Anal. Chim. Acta* **2000**, *420*, 145–154.
[26] R. Hecht–Nielsen, Counter propagation Networks, *Appl. Optics* **1987**, *26*, 4979–4984.
[27] J. Zupan, M. Novič, and J. Gasteiger, Neural networks with counterpropagation learning strategy used for modeling, *Chemom. Intell. Lab. Syst.* **1995**, *27*, 175–187.
[28] M. Vračko, M. Novič, and J. Zupan, Study of structure–toxicity relationship by a counterpropagation neural network, *Anal. Chim. Acta* **1999**, *384*, 319–332.
[29] M. Vračko, A study of structure–carcinogenicity relationship for 86 compounds from NTP data base using topological indices as descriptors, *SAR and QSAR in Environ. Res.* **2000**, *11*, 103–105.
[30] V. Simon, J. Gasteiger, and J. Zupan, A combined application of two different neural network types for the prediction of chemical reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148–9159.
[31] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, and J. Smeyers–Verbeke, *Handbook of chemometrics and qualimetrics: Part A*, Elsevier, Amsterdam, 1997, p.125.