Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 16 (2008) 5733–5742

Bioorganic & Medicinal Chemistry

Prediction of the aqueous solvation free energy of organic compounds by using autocorrelation of molecular electrostatic potential surface properties combined with response surface analysis

Lisa Michielan,[a] Magdalena Bacilieri,[a] Chosei Kaseda[b] and Stefano Moro[a,*]

[a] *Molecular Modeling Section (MMS), Dipartimento di Scienze Farmaceutiche, Università di Padova, via Marzolo 5, I-35131 Padova, Italy*

[b] *Yamatake Corporation, 1-12-2 Kawana, Fujisawa-shi Kanagawa 251-8522, Japan*

Received 5 November 2007; revised 14 March 2008; accepted 25 March 2008
Available online 30 March 2008

Abstract—Several quantitative structure–property relationship (QSPR) approaches have been explored for the prediction of aqueous solubility or aqueous solvation free energies, ΔG_{sol}, as crucial parameter affecting the pharmacokinetic profile and toxicity of chemical compounds. It is mostly accepted that aqueous solvation free energies can be expressed quantitatively in terms of properties of the molecular surface electrostatic potentials of the solutes. In the present study we have introduced *autocorrelation* molecular electrostatic potential (*auto*MEP) vectors in combination with nonlinear response surface analysis (RSA) as alternative 3D-QSPR strategy to evaluate the aqueous solvation free energy of organic compounds. A robust QSPR model ($r_{\mathrm{cv}} = 0.93$) has been obtained by using a collection of 248 organic chemicals. An external test set based on 23 molecules confirmed the good predictivity of the *auto*MEP/RSA model suggesting its further applicability in the in silico prediction of water solubility of large organic compound libraries.

1. Introduction

In last decades a wide interest has been focused on the prediction of aqueous solubility as relevant property affecting the pharmacokinetic profile and toxicity of chemical compounds.[1] Especially in the early phase of drug discovery, molecular solubility represents an important determinant of drug-likeness since it is related, for example, to drug bioavailability.[2] In fact, solvation effects play one of the major contributes influencing the quantity of free drug for biological processes, and increased solubility could correspond to improved therapeutic effectiveness of potential new drugs. From a thermodynamic point of view, the dissolution process was divided into three steps to simplify the examination of the contribution from the different interaction energies. The first step would represent the removal of one molecule from the crystal lattice, the second step the formation of a cavity in the water large enough to accommodate the molecule, and the third step the transfer of the molecule into the water. Steps one and two would be energetically unfavourable, since they involve the breaking of intermolecular bonds in the crystal and the water, respectively; but energy is gained in the third step from favourable interactions between the solute and the water (aqueous solvation free energies, ΔG_{hyd}). For dissolution to occur, the energy gained in the third step has to be numerically larger than the energetic cost of steps one and two.

Microscopically, the solvation effect is due to intermolecular interactions between the solute and the solvent, as well as a change in the intramolecular interactions of the solute and a reorganization of the solvent because of the solute. In general, the calculation of the solvent effects is partitioned into three separate components: electrostatics, short-range effects and cavitation.[3] In

Keywords: Quantitative structure–property relationship (QSPR); Aqueous solvation free energies; Pharmacokinetic profile; Molecular surface electrostatic potentials; Nonlinear response surface analysis (RSA).

* Corresponding author. Tel.: +39 049 8275704; fax: +39 049 8275366; e-mail: stefano.moro@unipd.it

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.03.064

particular, electrostatic forces dominate the interactions among molecules due to their strength and long-range. Consequently, the charge distributions of the solute and solvent play a fundamental role in the solvation process.

Moreover, the design of a tight-binding molecular ligand involves a trade off between an unfavourable electrostatic desolvation penalty incurred when the ligand binds a receptor in aqueous solution and the generally favourable intermolecular interactions made in the bound state.[4] In fact, an accurate treatment of solvation effects is required to determine the most probable binding mode of a receptor–ligand complex as well as for ranking a library of organic compounds according to binding affinity.[5]

The free energy of solvation can be calculated using high level computer simulations of the solute–solvent system of interest. These simulations are computationally demanding and time consuming and are therefore, not suitable for the screening of large compound libraries. They do, however, offer an excellent means for investigating the solute–solvent interactions more closely for a limited number of compounds. However, the methods described here are applicable to any solute–solvent system. The methods available for the calculation of the free energy of solvation were recently reviewed.[6]

The free energy of hydration can also be estimated by a quantitative structure–property relationship (QSPR) approach. One of the first examples of this was provided by Hine and Mookerjee[7] in 1975, but several others have followed since.[8–12] The standard error (RMSE) of the predicted ΔG_{hyd} with these models ranges from about 3–6 kJmol^{-1} which corresponds to approximately 10% of the range of ΔG_{hyd} of the respective training set. Unfortunately, the datasets used for the training of these models are restricted to organic compounds which are not drug-like. It can be expected that the error will be larger for drug molecules, since they are large, often flexible and complex chemical structures in comparison to the small organic molecules used for model development.

Recently, we have reported that *autocorrelation* Molecular Electrostatic Potential (*auto*MEP) vectors in combination to partial least square (PLS) and/or response surface analysis (RSA) techniques can represent an alternative and effective three-dimensional quantitative structure–activity relationship (3D-QSAR) approach.[13–17] In fact, topological and electrostatic complementarities are considered two key concepts in molecular recognition processes. Gasteiger and collaborators investigated the MEP on a molecular surface as particularly useful method for rationalizing the interactions between molecules and the molecular recognition processes.[18–20] The subsequent introduction of the *autocorrelation* vector makes then the MEP information invariant to the spatial rotation and translation of molecules.[21,22]

MEP distribution on a molecular surface plays a critical role in all solvation/desolvation processes. In fact, any implicit solvation model based on the solvent accessible surface area loosely relies on the idea that the major contribution to the free energy of solvation of the solute is determined by the surface of the solute that is accessible to the solvent, and by the screening effect of the solvent. Following this hypothesis, we decided to use *auto*MEP descriptors in combination with response surface analysis (RSA) as alternative 3D-QSPR strategy to evaluate the aqueous solvation free energy of organic compounds, as schematically shown in Figure 1.

2. Computational methodologies

All 3D-QSPR studies were carried out on a linux cluster running under openMosix architecture.[23] Autocorrelation MEP descriptors have been carried out using Adriana suite (version 2.0).[24] Response Surface Analysis (RSA) have been performed using DataFOREST and DataNESIA softwares.[25,26]

2.1. Molecular structure building

3D models of all organic compounds were obtained by using the 3D structure generator Corina, which is an integral part of Adriana suite.[24] It automatically generates 3D atomic coordinates from the constitution of a molecule as expressed by a connection table or linear string. A low energy conformer is generated for each compound, starting from molecules expressed in a valence bond notation, combining monocentric fragments with standard bond lengths and angles and using appropriate dihedral angles. Conformer selection is one of the most crucial step in every 3D-QSAR approaches. Protonation states are selected in agreement with the corresponding pK_a at the physiological pH value (7.4 U).

2.2. Training and test set

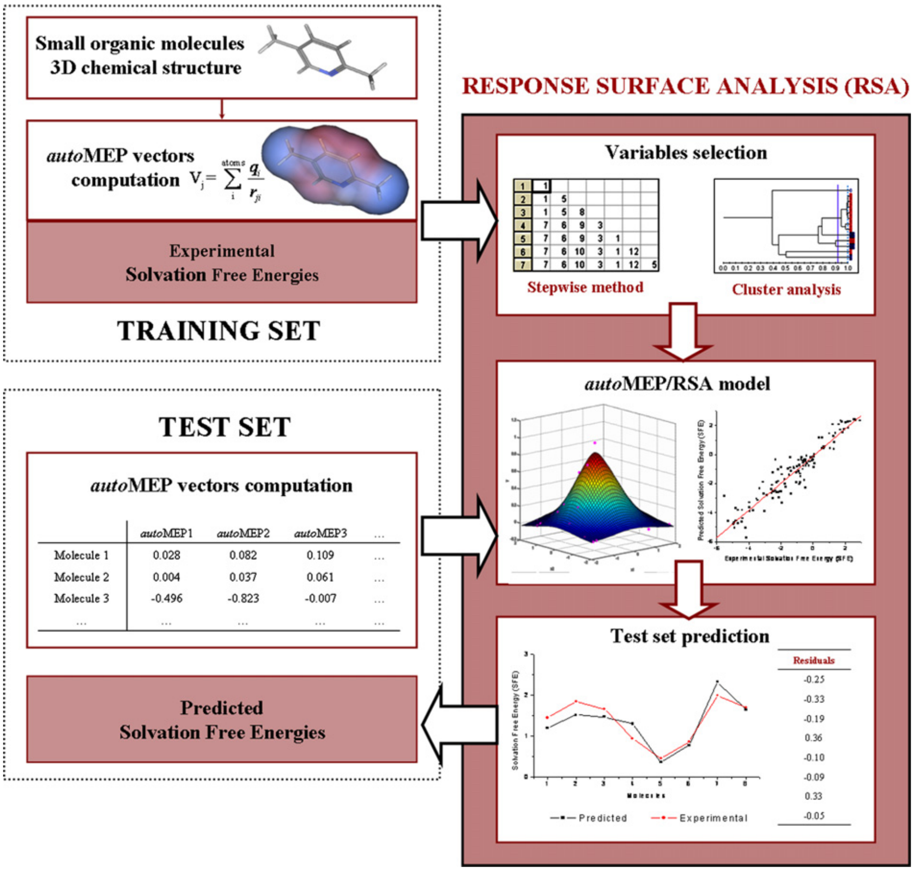
A collection of 271 organic chemicals, for which the experimental values of solvation free energy have been already reported, has been selected to derive our *auto*MEP/RSA model. Two hundred and forty-eight compounds have been considered as training set, while the other 23 derivatives have been used as test set in the validation step,[9] as collected in Tables 1 and 2.

2.3. Molecular electrostatic potential (MEP) calculation

In the present work, MEPs derive from a classical point charge model: the electrostatic potential for each molecule is obtained by moving a unit positive point charge across the molecular surface, and it is calculated at various points j on this surface by the following equation:

$$V_j = \sum_{i}^{\text{atoms}} \frac{q_i}{r_{ji}} \qquad (1)$$

where q_i represents the partial charge of each atom i and r_{ji} are the distance between points j and atom i. Starting from the 3D structure of a molecule and its partial atomic charges, the electrostatic potential is calculated for points on the molecular surface. Partial atomic



	*auto*MEP1	*auto*MEP2	*auto*MEP3	...
Molecule 1	0.028	0.082	0.109	...
Molecule 2	0.004	0.037	0.061	...
Molecule 3	-0.496	-0.823	-0.007	...
...	...	...	...	...

1	1						
2	1	5					
3	1	5	8				
4	7	6	9	3			
5	7	6	9	3	1		
6	7	6	10	3	1	12	
7	7	6	10	3	1	12	5

Residuals
-0.25
-0.33
-0.19
0.36
-0.10
-0.09
0.33
-0.05

Figure 1. Flowchart illustrating 3D-QSPR strategy combining *auto*MEP descriptors computation with response surface analysis (RSA) to predict solvation free energy (ΔG_{hyd}) of small organic molecules.

charges were calculated by the partial equalization of orbital electronegativity (PEOE) method and its extension to conjugated systems implemented by Petra (parameter estimation for the treatment of reactivity applications) module of Adriana suite.[24,27,28] Connolly's solvent accessible surface with a solvent radius of 2.0 Å has been used to project the corresponding MEP. Once the autocorrelation function has been applied, the autocorrelation vector is obtained. Connolly's solvent accessible surfaces and the corresponding MEPs have been calculated by Surface module of Adriana suite.[24]

2.4. Autocorrelation MEP (*auto*MEP) vectors

The idea of using autocorrelation for the transformation of the constitution of a molecule into a fixed length representation was introduced by Moreau and Broto.[21,22] They considered that a certain property p of an atom i can be correlated with the same property p of atom j and the products of p values can be summed over all atom pairs having a certain topological distance d. Each component of the autocorrelation vector is then calculated as follows:

$$A(d) = \sum_{ij} p_i p_j, \qquad (2)$$

where A is the autocorrelation coefficient referring to atom pairs i, j, p_i is the atomic property and d the i, j topological distance.[21,22]

Ligands and proteins interact through molecular surfaces and, therefore, clearly, representations of molecular surfaces have to be sought in the endeavour to understand biological activity. Again, we are under the restriction of having to represent molecular surfaces of different size, and again autocorrelation was employed to achieve this goal, by Gasteiger and collaborators.[18–20] Starting from the topological autocorrelation examples of Moreau and Broto, firstly they generated a set of randomly distributed points on the molecular surface, then, all distances between the surface points were calculated and sorted into preset intervals:

$$A(d_{lower}, d_{upper}) = 1/L \sum p_i p_j (d_{lower} < d_{ij} < d_{upper}) \qquad (3)$$

where the i, j distance d belongs to the d_{lower}, d_{upper} interval and L is the number of distances in the same interval. The application of this concept made possible to compare different molecular properties, as this 3D descriptor represents a compressed expression of the distribution of the property p on the molecular surface. For the calculation of the autocorrelation coefficient we started considering the default values set, changing them consequently in a various way, to see if it was possible to improve the statistical model capability. Finally, parameters values considered for the QSPR model were the following: $d_{lower} = 0$ Å (default is 1 Å); $d_{upper} = 5$ Å (default is 13 Å); $L = 12$ (default value); *point density* = 20 points/Å^2 (default is 10 points/Å^2); *vdW radius reduction factor* = 1.000 (default value).

Table 1. Experimental and predicted solvation free energies (ΔG_{hyd}, in kcal/mol) of the 248 molecules included in the training set

Compound	Molecule name	Experimental SFE (kcal/mol)	Predicted SFE (kcal/mol)
1	Methane	1.98	2.55
2	Ethane	1.81	2.22
3	Propane	2.02	2.08
4	*n*-Butane	2.18	2.26
5	2-Methylpropane	2.32	2.29
6	*n*-Pentane	2.36	2.25
7	2,2-Dimethylpropane	2.69	2.65
8	Cyclopentane	1.22	1.87
9	*n*-Hexane	2.58	2.07
10	3-Methylpentane	2.54	2.57
11	Cyclohexane	1.24	2.08
12	Methylcyclopentane	1.62	1.87
13	2,2-Dimethylbutane	2.63	2.63
14	*n*-Heptane	2.65	2.49
15	2,4-Dimethylpentane	2.92	2.48
16	Methylcyclohexane	1.73	1.92
17	*n*-Octane	2.93	2.43
18	2,2,4-Trimethylpentane	2.89	2.47
19	Ethylene	1.30	0.32
20	Propylene	1.28	1.23
21	1-Butene	1.40	1.51
22	2-Methylpropene	1.30	1.06
23	1-Pentene	1.69	1.80
24	*trans*-2-Pentene	1.35	1.09
25	Cyclopentene	0.57	1.29
26	2-Methyl-2-butene	1.33	0.78
27	3-Methyl-1-butene	1.85	1.58
28	Cyclohexene	0.37	1.20
29	4-Methyl-1-pentene	1.93	1.72
30	*trans*-2-Heptene	1.69	1.61
31	1-Methylcyclohexene	0.68	0.38
32	1-Octene	2.20	1.85
33	1,3-Butadiene	0.57	0.46
34	1,4-Pentadiene	0.95	1.21
35	2-Methyl-1,3-butadiene	0.69	0.44
36	1,5-Hexadiene	1.02	1.39
37	Propyne	−0.48	−1.42
38	Butyne	−0.17	−1.67
39	1-Pentyne	0.01	−0.32
40	1-Hexine	0.29	−0.37
41	1-Heptyne	0.61	0.25
42	1-Octyne	0.72	0.55
43	1-Monyne	1.06	0.37
44	Benzene	−0.90	−0.30
45	Ethylbenzene	−0.62	0.01
46	*o*-Xylene	−0.91	−0.50
47	*m*-Xylene	−0.82	−0.41
48	*p*-Xylene	−0.82	−0.53
49	Propylbenzene	−0.54	−0.10
50	2-Propylbenzene	−0.30	−0.61
51	1,2,4-Trimethylbenzene	−0.87	−0.15
52	Butylbenzene	−0.40	−0.07
53	2-Butylbenzene	−0.46	−0.29
54	*tert*-Amylbenzene	−0.18	−0.30
55	Naphthalene	−2.45	−0.48
56	Acenaphtene	−3.44	−2.07
57	Anthracene	−4.34	−2.97
58	Phenanthrene	−4.12	−3.98
59	Chloromethane	−0.54	−2.39
60	Trichloromethane	−1.04	−2.14
61	Tetrachloromethane	0.10	0.20
62	Bromomethane	−0.80	−1.45
63	Dibromomethane	−1.99	−1.17
64	Tribromomethane	−2.16	−1.91
65	Iodomethane	−0.90	−0.85
66	Chlorofluoromethane	−0.79	−2.47
67	Chlorotrifluoromethane	2.56	−0.56
68	Dichlorodifluoromethane	1.71	−0.11
69	Chloroethane	−0.64	−1.03
70	Bromoethane	−0.70	−0.12
71	Iodoethane	−0.73	−0.22
72	1,1-Dicloroethane	−0.86	−2.12
73	1,2-Dichloroethane	−1.75	−2.34
74	1,2-Dibromoethane	−2.13	−2.99
75	1-Chloro-2-bromoethane	−1.98	−2.88
76	1,1,1-Trichloroethane	−0.25	−0.76
77	1,1,2-Trichloroethane	−1.98	−2.31
78	Pentachloroethane	−1.38	−3.21
79	Hexachloroethane	−1.42	0.38
80	Chloropentafluoroethane	2.90	−0.08
81	1,1,2,2-Tetrachloro-difluoroethane	0.83	1.02
82	1,1,2-Trichloro-trifluoroethane	1.80	1.25
83	1,1-Dichloro-tetrafluoroethane	2.54	0.68
84	1,2-Dichloro-tetrafluoroethane	2.35	1.44
85	1-Chloropropane	−0.36	−0.71
86	2-Chloropropane	−0.25	−0.87
87	1-Bromopropane	−0.57	−0.35
88	2-Bromopropane	−0.48	−0.29
89	1-Iodopropane	−0.62	−0.32
90	2-Iodopropane	−0.47	−0.49
91	1,2-Dichloropropane	−1.27	−2.88
92	1,3-Dichloropropane	−1.92	−3.28
93	1,2-Dibromopropane	−1.96	−2.10
94	1-Chlorobutane	−0.14	−0.24
95	1-Bromobutane	−0.41	−0.35
96	1-Bromo-2-methylpropane	−0.03	−0.33
97	1-Iodobutane	−0.26	−0.29
98	1,1-Dichlorobutane	−0.70	−0.95
99	1-Chloropentano	−0.07	−0.87
100	2-Chloropentane	0.07	0.04
101	3-Chloropentane	0.04	−0.13
102	1-Bromo-3-methylbutane	0.21	−1.30
103	*cis*-1,2-Dichloroethylene	−1.19	−1.37
104	*trans*-1,2-Dichloroethylene	−0.77	−0.82
105	1,2,3-Trichloroethylene	−0.44	−0.82
106	Tetrachloroethylene	0.06	−0.76
107	3-Chloropropene	−0.58	−0.70
108	Chlorobenzene	−1.02	−1.93
109	Bromobenzene	−1.48	−2.15
110	1,2-Dichlorobenzene	−1.38	−1.30
111	1,3-Dichlorobenzene	−0.99	−1.46
112	1,4-Dibromobenzene	−2.33	−2.46
113	*p*-Bromotoluene	−1.41	−1.29
114	1-Bromo-2-ethylbenzene	−1.20	−1.02
115	*o*-Bromocumene	−0.86	−0.80
116	Dimethyl ether	−1.92	−1.88
117	Dimethyl sulfide	−1.56	−1.09
118	1,3-Dioxolane	−4.14	−5.60

Table 1 (*continued*)

Compound	Molecule name	Experimental SFE (kcal/mol)	Predicted SFE (kcal/mol)
119	Diethyl ether	−1.77	−2.00
120	Methylpropyl ether	−1.69	−2.34
121	Methyl isopropyl ether	−2.03	−2.00
122	Tetrahydrofuran	−3.51	−3.20
123	Dioxane	−5.11	−3.67
124	Ethylpropyl ether	−1.84	−1.23
125	Methyl *tert*-butyl ether	−2.24	−2.69
126	2-Methyltetrahydrofuran	−3.34	−3.11
127	Tetrahydropyran	−3.16	−2.32
128	Dipropyl ether	−1.17	−1.10
129	1,2-Dietoxyethane	−3.30	−1.94
130	1,1-Dietoxyethane	−3.32	−3.43
131	Di-*n*-butyl ether	−0.84	−0.38
132	Anisole	−1.05	−3.11
133	Thioanisole	−2.76	−1.21
134	2,2′-Dichlorodietyl sulfide	−3.97	−1.68
135	Methanol	−5.14	−6.09
136	Methan thiol	−1.26	−2.17
137	Ethanol	−4.96	−4.66
138	2,2,2-Trifluoroethanol	−4.35	−4.27
139	1-Propanol	−4.92	−4.24
140	2-Propanol	−4.81	−3.81
141	Allyl alcohol	−5.10	−5.78
142	1,1,1-Trifluoro-2-propanol	−4.21	−6.09
143	2,2,3,3-Tetrafluoropropanol	−4.96	−4.77
144	2,2,3,3,3-Pentafluoropropanol	−4.20	−4.95
145	1-Butanol	−4.78	−4.85
146	2-Butanol	−4.67	−3.80
147	*tert*-Butyl alcohol	−4.57	−2.86
148	2-Methyl-1-propanol	−4.57	−3.34
149	1-Pentanol	−4.55	−4.46
150	2-Pentanol	−4.45	−4.18
151	2-Methyl-1-butanol	−4.48	−4.31
152	2-Methyl-2-butanol	−4.49	−3.42
153	1-Hexanol	−4.42	−4.15
154	Cyclohexanol	−5.02	−4.74
155	2,3-Dimethylbutanol	−3.97	−3.62
156	2-Methyl-3-pentanol	−3.94	−4.33
157	4-Methyl-2-pentanol	−3.79	−4.11
158	2-Methyl-2-pentanol	−3.98	−3.09
159	1-Heptanol	−4.31	−3.94
160	1-Octanol	−4.16	−4.39
161	Phenol	−6.62	−7.13
162	4-Bromophenol	−7.20	−5.30
163	Thiophenol	−2.58	−4.39
164	2-Cresol	−5.94	−6.29
165	4-Nitrophenol	−6.20	−6.26
166	4-*tert*-Butylphenol	−6.00	−4.86
167	Acetaldehyde	−3.55	−3.00
168	Propanal	−3.48	−3.30
169	Butanal	−3.22	−3.56
170	Pentanal	−3.07	−2.86
171	Heptanal	−2.71	−2.24
172	Octanal	−2.32	−2.45
173	Nonanal	−2.10	−3.23
174	*trans*-2-Butenal	−4.28	−4.03
175	*trans*-2-Hexenal	−3.73	−4.26
176	*trans*-2-Octenal	−3.48	−3.44

Table 1 (*continued*)

Compound	Molecule name	Experimental SFE (kcal/mol)	Predicted SFE (kcal/mol)
177	*trans-trans*-2,4-Hexadienal	−4.70	−3.21
178	Benzaldehyde	−4.08	−3.43
179	Acetone	−3.85	−2.22
180	2-Pentanone	−3.56	−2.33
181	2-Heptanone	−3.11	−2.75
182	2-Octanone	−2.92	−2.22
183	2-Nonanone	−2.51	−1.56
184	2-Undecanone	−2.18	−1.16
185	Acetophenone	−4.64	−4.32
186	Ethylformate	−2.68	−2.39
187	Methylacetate	−3.36	−4.03
188	Propylformate	−2.51	−2.71
189	Isopropylformate	−2.04	−2.28
190	Ethylacetate	−3.12	−3.08
191	Methylpropionate	−3.01	−3.53
192	Isobutylformate	−2.25	−2.44
193	Propylacetate	−2.89	−2.86
194	Isopropylacetate	−2.68	−2.95
195	Methylbutyrate	−2.87	−2.93
196	Isoamylformate	−2.16	−2.70
197	Butylacetate	−2.58	−2.45
198	Isobutylacetate	−2.39	−2.48
199	Propylpropionate	−2.49	−2.38
200	Isopropylpropionate	−2.25	−2.61
201	Ethylbutyrate	−2.53	−2.47
202	Methylpentanoate	−2.57	−2.66
203	Amylacetate	−2.49	−2.53
204	Propylbutyrate	−2.31	−2.51
205	Ethylpentanoate	−2.56	−2.09
206	Methylhexanoate	−2.51	−2.87
207	Hexylacetate	−2.29	−3.13
208	Amylpropionate	−2.02	−2.38
209	Methyloctanoate	−2.07	−3.73
210	Ethylheptanoate	−2.33	−2.10
211	Methylbenzoate	−4.34	−4.76
212	Ethylamine	−4.67	−3.76
213	Butylamine	−4.43	−3.81
214	Pentylamine	−4.14	−4.15
215	Hexylamine	−4.09	−2.43
216	Dimethylamine	−4.34	−3.83
217	Diethylamine	−4.12	−2.39
218	Pyrrolidine	−5.54	−3.71
219	Piperidine	−5.17	−3.74
220	Dipropylamine	−3.70	−1.16
221	Hexamethyleneimine	−4.97	−3.42
222	Trimethylamine	−3.27	−4.15
223	Triethylamine	−3.07	−3.61
224	*n*-Methylpirrolidine	−4.02	−5.29
225	*n*-Methylpiperidine	−3.94	−5.04
226	Propionitrile	−3.90	−4.34
227	Butyronitrile	−3.69	−5.07
228	Nitroethane	−3.76	−4.51
229	2-Nitropropane	−3.18	−3.58
230	Nitrobenzene	−4.17	−4.54
231	2-Nitrotoluene	−3.63	−3.67
232	3-Nitrotoluene	−3.50	−3.87
233	Pyridine	−4.75	−4.26
234	2-Methylpyridine	−4.68	−4.58
235	3-Methylpyridine	−4.84	−4.70
236	4-Methylpyridine	−4.99	−5.01
237	2-Ethylpyridine	−4.38	−4.38
238	4-Ethylpyridine	−4.78	−2.45

(*continued on next page*)

Table 1 (*continued*)

Compound	Molecule name	Experimental SFE (kcal/mol)	Predicted SFE (kcal/mol)
239	2,3-Dimethylpyridine	−4.88	−5.02
240	2,4-Dimethylpyridine	−4.92	−4.97
241	2,5-Dimethylpyridine	−4.77	−4.55
242	2,6-Dimethylpyridine	−4.66	−3.66
243	3,4-Dimethylpyridine	−5.28	−4.85
244	3,5-Dymethylpyridine	−4.90	−4.76
245	2-Methylpirazine	−5.58	−4.45
246	2-Ethylpyrazine	−5.53	−4.23
247	2-Ethyl-3-methoxypirazine	−4.45	−4.19
248	2-Isobutyl-3methoxypirazine	−3.73	−4.56

Consequently, we derived 12 autocorrelation vectors per molecule, computed at the 12 (L value) distances in the interval from 0 to 5 Å with a step width of 0.4 Å. We considered that the step width of 0.4 Å, derived from the partition in 12 intervals of the global distance 0–5 Å, was sufficient to describe in an accurate way the distribution of the MEP property. This transformation produces a unique fingerprint of each molecule under consideration. Autocorrelation vector has been calculated by Surface module of Adriana suite.[24]

2.5. Response surface analysis (RSA)

Response surface analysis (RSA) is a collection of mathematical and statistical techniques useful for analyzing the effects of several independent variables.[29] In most RSA problems, the relationship between the response and the independent variables is unknown. Thus, the first step in RSA is to approximate the mathematical function (*f*). Usually, this process employs a low-order polynomial in some region of the independent variables. If the response is well-modelled by a linear function of the independent variables, then the approximating function (*f*) is a first-order model. If there is curvature in the system or in the region of the optimum, then a polynomial of higher degree of (*f*) must be used to approximate the response, which is analyzed to locate the optimum, that is, the set of independent variables such that the partial derivatives of the model response with respect to the individual independent variables are equal to zero. The eventual objective of RSA is to determine the optimum operating conditions for the system, or to determine a region which satisfies the operating specifications. Almost all RSA problems utilize one or both of these approximating polynomials.

In this work, our RSA is based on a multivariate *thin plate spline* algorithm derived by the Green's theorem:

$$y = \sum_{i-1}^{n} \alpha_i g(d_i) + \sum_{j=1}^{p} c_j x_j, \tag{4}$$

where α_i and c_j are the weight coefficients, p the number of independent variables x, n the number of data points and d_i the Green's function applied to the Euclidean distances between data i and any coordinate in x- axis.[30] According to this algorithm, the response surface function is the result of an elastic beam displacement in the x_n space, where the elastic beam has to bend to reach the data points in the y space. Input values are therefore regarded as points of force actions, while output values as displaced values.[30]

In the RSA technique, the selection of the most informative independent variables (in our specific case, *auto*-MEP descriptors) is strongly recommended to reduce the dimensionality of the final model and improving model predictivity. Consequently, linear stepwise regression and a nonlinear cluster analysis have been applied to select the most statistically relevant independent *auto*-MEP descriptors to use in our RSA model. DataFOREST and DataNESIA software have been used, respectively, for descriptors selection and RSA model calculation.[25,26]

3. Results and discussion

The solvation process is defined as 'the process in which a particle of the solute is transferred from a fixed position in the gas phase into a fixed position in solution at constant temperature'.[31] The key parameter to describe the effects of the solvent is the free energy of solvation, ΔG_{hyd} and is defined as the reversible work spent in the transfer of the solute under the aforementioned conditions at equal number densities in the gas phase and in solution. As previously said in the introduction, the electrostatic forces dominate the interactions among molecules and consequently the solvation process is strongly dependent from the charge distributions of solute and solvent.

As anticipated, *auto*MEP descriptors encode into autocorrelation vectors the three-dimensional spatial distribution and the intensity of the electrostatic potential projected on the molecular surface. Indeed, *auto*MEPs can be considered a sort of electrostatic fingerprint of each chemical structure. We decided to use *auto*MEP descriptors in combination with a response surface analysis technique (*auto*MEP/RSA) to predict the solvation free energy of a set of 248 organic chemicals.

As requested from the RSA technique, stepwise regression analysis together with the cluster analysis on the original 12 *auto*MEP descriptors led us to select five of them as final combination to utilize as independent variables into RSA model: *auto*MEP 1, 7, 8, 10 and 12.

The application of the thin spline plate algorithm in the calibration step has provided a very high correlation coefficient (r = 0.99) value, confirming the good choice of the independent variables selection, as summarized in Figure 2 and Table 3.

Leave-one-out (LOO) cross-validation technique has been used to validate the *auto*MEP/RSA model confirming the robustness in prediction of our statistical model

Table 2. Experimental and predicted solvation free energies (ΔG_{hyd}, in kcal/mol) of the test set of 23 molecules applying our *auto*MEP/RSA model

Compound	Molecule name	Experimental ΔG_{hyd} (kcal/mol)	Predicted *auto*MEP/RSA ΔG_{hyd} (kcal/mol)	Predicted HLOGS ΔG_{hyd} (kcal/mol)	Predicted ALOGS ΔG_{hyd} (kcal/mol)	Predicted ATOGEN ΔG_{hyd} (kcal/mol)	Residuals[a] *auto*MEP/RSA ΔG_{hyd} (kcal/mol)	Residuals[a] HLOGS ΔG_{hyd} (kcal/mol)	Residuals[a] ALOGS ΔG_{hyd} (kcal/mol)	Residuals[a] ATOGEN ΔG_{hyd} (kcal/mol)
1	2-Methylpentane	2.56	2.53	3.09	2.23	1.28	−0.03	0.53	−0.33	−1.28
2	*cis*-1,2-Dimethylcyclohexane	1.60	2.38	0.76	1.36	1.71	0.78	−0.84	−0.24	0.11
3	1-Hexene	1.73	1.96	1.96	1.62	1.19	0.23	0.23	−0.11	−0.54
4	2,3-Dimethyl-1,3-butadiene	0.40	0.33	0.08	1.18	1.11	−0.07	−0.32	0.78	0.71
5	Toluene	−0.77	0.28	−0.59	−0.71	−1.23	1.05	0.18	0.06	−0.46
6	*tert*-Butylbenzene	−0.44	−0.3	−0.26	−0.01	−0.59	0.14	0.18	0.43	−0.15
7	Dichloromethane	−1.42	−1.82	−1.26	−1.51	−1.67	−0.40	0.16	−0.09	−0.25
8	1,3-Dibromopropane	−1.99	−1.6	−0.94	−2.12	−1.18	0.39	1.05	−0.13	0.81
9	Chloroethylene	0.50	−0.76	−1.19	0.11	−0.40	−1.26	−1.69	−0.39	−0.90
10	1,4-Dichlorobenzene	−1.02	−1.75	−0.90	−1.22	−3.33	−0.73	0.12	−0.20	−2.31
11	Diethyl sulfide	−1.45	−0.8	−2.02	−2.29	−3.34	0.65	−0.57	−0.84	−1.89
12	Diisopropyl ether	−0.54	−0.96	−0.14	−2.00	−2.75	−0.42	0.4	−1.46	−2.21
13	Ethane thiol	−4.08	−2.98	−1.65	−1.61	–	1.10	2.43	2.47	–
14	3-Hexanol	−3.73	−3.8	−3.33	−3.31	−5.30	−0.07	0.40	0.42	−1.57
15	Hexanal	−2.85	−2.54	−2.89	−2.96	−2.92	0.31	−0.04	−0.11	−0.07
16	2-Butanone	−3.76	−2.63	−3.35	−3.54	−3.35	1.13	0.41	0.22	0.41
17	Methylformate	−2.82	−4.34	−2.86	−2.57	−5.40	−1.52	−0.04	0.25	−2.58
18	Ethylpropionate	−2.83	−2.75	−2.75	−2.90	−4.76	0.08	0.08	−0.07	−1.93
19	Isoamylacetate	−2.24	−2.54	−2.66	−2.38	−4.33	−0.30	−0.42	−0.14	−2.09
20	Propylamine	−4.56	−3.88	−3.82	−4.55	−5.41	0.68	0.74	0.01	−0.85
21	Dibutylamine	−3.38	−1.46	−4.91	−3.35	−5.13	1.92	−1.53	0.03	−1.75
22	1-Nitropropane	−3.38	−4.18	−1.61	−3.26	−3.88	−0.80	1.77	0.12	−0.50
23	2-Isobutylpyrazine	−5.11	−3.84	−7.25	−7.04	−7.10	1.27	−2.14	−1.93	−1.99

Previously published HLOGS and ALOGS models, and ATOGEN model predictions are also reported. Residuals in kcal/mol are also shown[9,12].

[a] Predicted ΔG_{hyd} (kcal/mol) – experimental ΔG_{hyd} (kcal/mol).

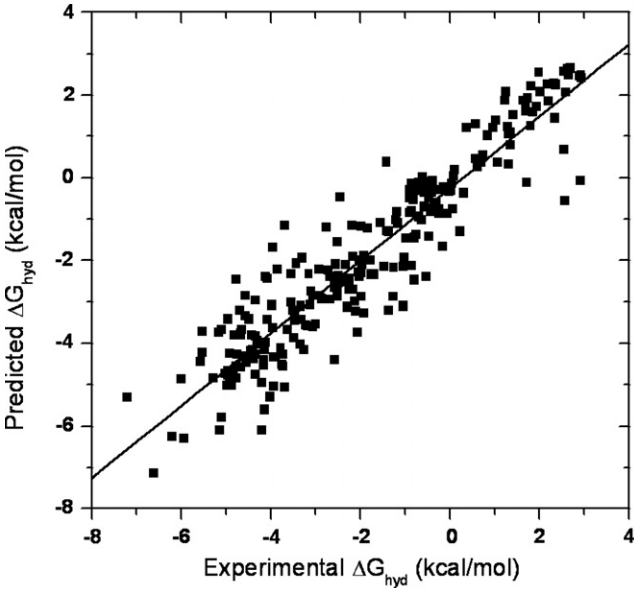


Figure 2. *auto*MEP/RSA model experimental versus predicted values of solvation free energy (ΔG_{hyd} in kcal/mol) after the cross-validation step.

(r_{cv} = 0.93). Significantly, *auto*MEP/RSA model is comparable or better, in terms of performance, to all previously reported QSPR models.[7–12]

Interestingly, *auto*MEPs 1 and 7 seem to play a major role in describing the complexity of the final response surface. A representation of solvation free energy as a function of *auto*MEP 1 and *auto*MEP 7 is shown in Figure 3.

As shown in Table 1, the predictivity of *auto*MEP/RSA model does not present any particular dependence from the chemical scaffold or from the subsistent decoration of the analysed organic compounds. The residuals of 48 derivatives of the training set overcome 1 kcal/mol, and it happens especially when chlorine and fluorine atoms are present (see molecules **59**, **60**, **66–68**, **72**, **78–80**, **83**, **91–92**, **134** and **142** in Table 1), or for some aliphatic and aromatic alcohols (see molecules **147–148**, **152**, **162** and **166** in Table 1) or aromatic amines (see molecules **218–219**, **224–225**, **238**, **245** and **246** in Table 1). In most cases the solvation free energy of halogen derivatives is overestimated, while alcohols and aromatic amines are generally underestimated, if compared to the respective experimental values. Conformational equilibria, influence of both short-range and cavitation effects, and the accuracy of the calculated MEPs might explain the weakness of the above-mentioned predictions. However, it is interesting to note that as all other 3D-QSPR approaches, also *auto*MEP/RSA model are able to discriminate among stereoisomers, improving the limits of some models that have utilized, for example, atomic constants as molecular descriptors.[12] As an example, the prediction of solvation free energies for *cis* and *trans* isomers of 1,2-dichloroethylene is correctly assigned (molecules **103** and **104**, Table 1). Moreover, 1-bromopropane and 2-bromopropane are also correctly recognized (molecules **87** and **88**, Table 1).

Table 3. Summary of *auto*MEP/RSA model statistics

	*auto*MEP/RSA model
Number of molecules	248
X variables	5
r	0.99
r_{cv}[a]	0.93
Slope	0.87
Offset	−0.25
q[b]	0.92
RMR[c]	0.069
RSS[d]	1.19

[a] Cross-validated *r* after leave-one-out procedure: $r_{cv} = [\mathrm{SXY}/(\mathrm{SXX})^{1/2}(\mathrm{SYY})^{1/2}]$, $\mathrm{SXY} = \sum(X - X_{mean})(Y - Y_{mean})$, $\mathrm{SXX} = \sum(X - X_{mean})^2$ and $\mathrm{SYY} = \sum(Y - Y_{mean})^2$ with $X = Y_{experimental}$ and $Y = Y_{predicted}$.

[b] *r* of the internal test set.

[c] Root mean square of residuals: RMR.

[d] Residual sum of squares: RSS.

A test set of 23 molecules with a different chemical structure and solvation free energy values has been selected to further validate our *auto*MEP/RSA model. This collection represents almost the 10% of the number of compounds selected as training set. The experimental versus predicted solvation free energies values are collected in Table 2. Again, a very good correlation coefficient calculated on the test set (q = 0.92) is an additional evidence about the good predictivity of the *auto*MEP/RSA model.

The predicted solvation free energies result very close to the experimental values, as shown in Figure 4. Indeed, less accurate estimation is again reported for halogen derivatives and amines (see e.g., molecules **9**, **21** and **23** in Fig. 4 and Table 2).

We have then compared the predictions of *auto*MEP/RSA model with those already reported for HLOGS and ALOGS models by Viswanadhanand and collaborators (see Predicted ALOGS and Predicted HLOGS in Table 2) and ATOGEN model by Kang (see Predicted ATOGEN in Table 2).[9,12] Analyzing residuals profiles of predictions, it is interesting to underline that *auto*MEP/RSA model produces results (and statistical values) comparable or better respect to the above-mentioned predictors/models results. The highest residuals correspond to molecules having a chemical structure in common with the worst predicted compounds in the training set, such as chloroethylene, dibutylamine and 2-isobutylpyrazine (molecules **9**, **21** and **23** in Table 2), for which *auto*MEP/RSA deviations from the corresponding experimental values are more than 1 kcal/mol.

Overall, we can consider the combination of autocorrelation MEP vectors in combination with a response surface analysis an alternative tool to evaluate the aqueous solvation free energy of organic compounds.

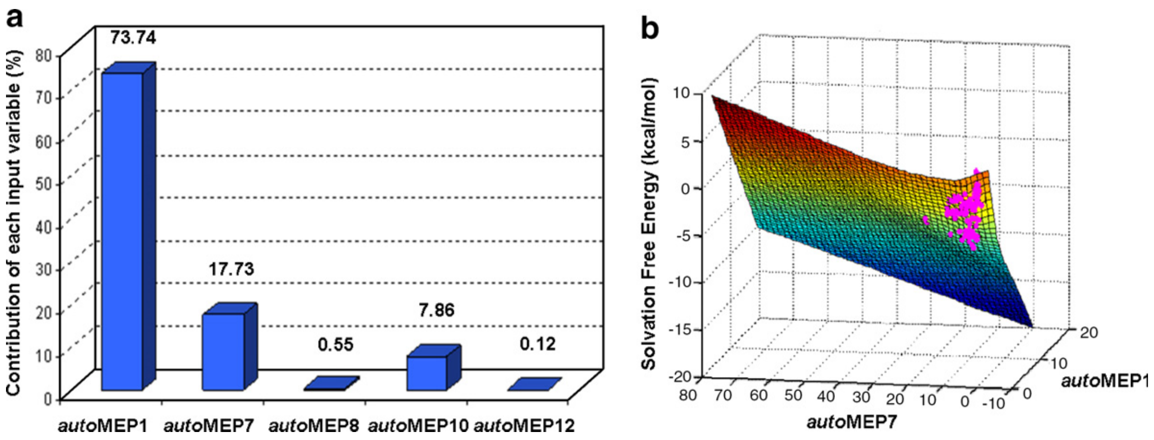


Figure 3. (a) Contribution of each *auto*MEP vectors in defining the final *auto*MEP/RSA model; (b) Contribution of *auto*MEP 1 and *auto*MEP 7 to the response surface.

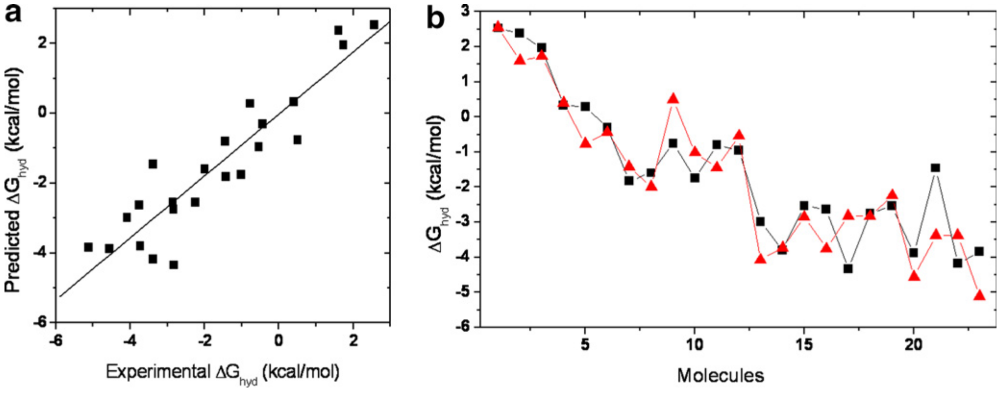


Figure 4. (a) *auto*MEP/RSA experimental versus predicted values of solvation free energy (ΔG_{hyd} in kcal/mol) of test set; (b) Comparison between experimental (▲) and predicted (■) ΔG_{hyd} data of test set.

4. Conclusion

The solvent environment of molecules plays a very important role in their structure and function. In biological systems it is well known that water has profound effects in the functions of proteins. Simulations assist us in microscopic studies of chemical and biological phenomena. It is important then to include solvation effects accurately and efficiently in molecular property prediction and simulations. In particular, the solvation energy is partitioned into long-range and short-range contributions. The long-range contributions are due to polar interactions between the solvent and the solute and the short-range are due to van der Waals and entropic effects.

In this work, we present an alternative 3D-QSPR approach combining *auto*MEP molecular descriptors with a response surface analysis (RSA) technique to evaluate the aqueous solvation free energy of organic compounds. Considering our results, *auto*MEP vectors can be considered an interesting electrostatic fingerprint able to describe both pharmacodynamic and pharmacokinetic processes. In particular, we would like to extend our studies verifying the applicability of *auto*MEP/RSA approach in other ligand-based approaches to in silico pharmacology.

Acknowledgments

The molecular modelling work coordinated by S.M. was carried out with financial support from the University of Padova, Italy, and the Italian Ministry for University and Research (MIUR), Rome, Italy. We thank Molecular Network GmbH (Erlangen, Germany) for the assistance in using Adriana suite. S.M. is also very grateful to Chemical Computing Group for the scientific and technical partnership.

References and notes

1. Johnson, S. R.; Zheng, W. *AAPS J.* **2006**, *8*, 27–40.
2. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
3. Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.
4. Wang, H.; Ben-Naim, A. A. *J. Med. Chem.* **1996**, *39*, 1531–1539.
5. Zou, X.; Sun, Y.; Kuntz, I. D. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.
6. Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187–4226.
7. Hine, J.; Mookerjee, P. K. *J. Org. Chem.* **1975**, *40*, 292–298.
8. Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.

9. Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405–412.
10. Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E., Jr. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794–1805.
11. Mansson, R. A.; Frey, J. G.; Essex, J. W.; Welsh, A. H. *J. Chem. Inf. Model.* **2005**, *45*, 1791–1803.
12. Kang, H.; Choi, H.; Park, H. *J. Chem. Inf. Model* **2007**, *47*, 509–514.
13. Moro, S.; Bacilieri, M.; Ferrari, C.; Spalluto, G. *Curr. Drug Discov. Technol.* **2005**, *2*, 13–21.
14. Moro, S.; Bacilieri, M.; Cacciari, B.; Spalluto, G. *J. Med. Chem.* **2005**, *48*, 5698–5704.
15. Moro, S.; Bacilieri, M.; Cacciari, B.; Bolcato, C.; Cusan, C.; Pastorin, G.; Klotz, K. N.; Spalluto, G. *Bioorg. Med. Chem.* **2006**, *14*, 4923–4932.
16. Bacilieri, M.; Kaseda, C.; Spalluto, G.; Moro, S. *Lett. Drug Des. Discov.* **2007**, *4*, 122–127.
17. Bacilieri, M.; Varano, F.; Deflorian, F.; Marini, M.; Catarzi, D.; Colotta, V.; Filacchioni, G.; Galli, A.; Costagli, C.; Kaseda, C.; Moro, S. *J. Chem. Inf. Model.* **2007**, *47*, 1913–1922.
18. Gasteiger, J.; Li, X.; Rudolph, C.; Sadovski, J.; Zupan, J. *J. Am. Chem. Soc.* **1994**, *116*, 4608–4620.
19. Wagener, M.; Sadovski, J.; Gasteiger, J. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7778.
20. Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1213.
21. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.
22. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757–764.
23. OpenMosix, version 2.4.26; Moshe Bar, Tel Aviv University, Israel, 2004.
24. Adriana, version 2.0; Molecular Networks GmbH: Erlangen, Germany, 2004, 2005.
25. DataFOREST, version 9; Yamatake Corporation: Fujisawa-shi Kanagawa, Japan, 2007.
26. DataNESIA, version 3.2; Yamatake Corporation: Fujisawa-shi Kanagawa, Japan, 2007.
27. Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219–3228.
28. Gasteiger, J.; Saller, H. *Angew. Chem.* **1985**, *97*, 699–701.
29. Myers, R.; Montgomery, D. C. In *Response Methodology Surface*; Wiley, John, Ed.; Wiley Interscience: New York, USA, 1995.
30. Kaseda, C. *Response Surface Methodology using a spline algorithm*; G.d.a.c.S., Ed.; Nashboro Press: Fujisawa-shi Kanagawa, Japan, 2004.
31. Lorentz, H. A. *Theory of Electrons*; Dover: NY, 1952.