TARTU ÜLIKOOL
FÜÜSIKA-KEEMIA TEADUSKOND
KEEMILISE FÜÜSIKA INSTITUUT

Deniss Savtšenko

Toksilisuse QSAR modelleerimine kvantkeemiliste orbitaalsete deskriptorite abil

Magistritöö

Juhendaja: Professor Mati Karelson
Vanemteadur Tarmo Tamm

Tartu 2007

Contents

1. Introduction

Drug design **[1]** is an iterative and sometimes very expensive process. Trial-and-error testing of chemical substances on animals, and matching the apparent effects to treatment becomes even more complicated when there is no detailed understanding of the biochemical processes responsible for the activity. Rational drug design begins with the knowledge of specific chemical responses in the body or target organism, and tailoring combinations of these to fit a treatment profile. Examining structural similarities and differences for active and inactive molecules helps to create a correct hypothesis. For selecting working hypotheses from those not working, a *Quantitative Structure Activity Relationship* (*QSAR*) approach **[2]** could be used.

In order to obtain good *QSAR* models one should have reliable data objects: molecular descriptors and experimental properties. In the present Thesis, we wish to examine the applicability of molecular orbital related descriptors for the prediction of the toxicity of chemical compounds. In a series of cases, the *QSAR* modeled toxicity has been related to the electrophilicity of compounds, and accordingly quantum-chemically calculated „electrophilic“ descriptors. In most cases, those are the energies of the lowest unoccupied molecular orbital, or *LUMO*. However, the unfavorable delocalization of the *LUMO*, steric hindrance etc. may forbid its participation in the given toxic action (chemical reaction). In such cases, it should be concretized, which unoccupied molecular orbital, or *UMO*, should be used for the calculation of related molecular descriptors.

In this Thesis, we will study the applicability of molecular orbital descriptors for the prediction of the toxicity of aromatic compounds. The main goal was to investigate, if an appropriate orbital selection can make *QSAR* predictions more exact and robust.

2. Literature overview

2.1. QSAR and drug design

QSAR dates back to the 19th century. It has been established that already in 1863, A.F.A. Cros at the University of Strasbourg observed that the toxicity of alcohols to mammals increased as the water solubility of the alcohols decreased **[3]**. Little additional development of *QSAR* occurred until the work of Louis Hammett (1894-1987) **[4]**, who correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity. As an example, Hammett showed that substituents have an ordered and quantitative effect on the dissociation of other organic acids and bases. Electron-withdrawal by the nitro group increases dissociation, with the effect being less for the meta than for the para substituent.

Models based on the simplest relationships that utilize only electronic properties as the descriptors of structures has difficulties if applied to biological systems: other structural descriptors are also needed. Researchers have attempted for many years to develop drugs based on *QSAR*. As access to high-speed computers and graphics workstations became commonplace, this field has evolved into what is often termed rational drug design or computer-assisted drug design **[5]**.

Most generally the mathematical expression for *QSAR* models has the following form:

Property of a compound = function(physico-chemical properties and/or molecular structural properties)

But it is not always true that all similar molecules should behave similarly. Therefore, the main problem is how to resolve a small difference on the molecular level for getting a better predicted parameter on the macromolecular level. *QSAR* attempts to create consistent relationships between molecular differences and compound property value variations.

2.2. QSAR approaches

Multilinear regression

The multilinear regression (*MLR*) approach is historically the oldest that has been used in the *QSAR* area. *MLR* results are easy to interpret but not always handy to use, because most of the biological activities and physico-chemical properties possess nonlinear relationships between them by their nature. In order to get more complex dependencies, more complex descriptors or derivative descriptors could be used.

The multilinear regression (*MLR*) method is the simplest one used in the *QSAR* analysis. It asumes that given a dataset in the form (x_i, y_i) for i = 1, ... N, a straight line is desired to be fit with this dataset. Therefore, the goal is to find the slope and y-intercept that makes a line fit the given data "best". The common approach used for linear regression is to minimize the sum of the squares of the differences between the data and a line; that is, to find the values of a and b that minimize the sum:

$$R^2 = \sum_{i=1}^{N} (ax_i + b - y_i)^2$$

Minimizing this sum is called *least squares minimization*.

Several statistical characteristics provide information about the quality of the model: R^2 - squared correlation coefficient, R^2_{cv} - squared cross-validated correlation coefficient, F – Fisher criterion value, s^2 - squared standard error.

Usually a large pool of molecular descriptors is used for *QSAR* modelling. Selecting the best model among lots of others is no trivial task. Various approaches have been suggested for this purpose **[6].** In the current work, the *Best Multilinear Regression* (*BMLR*) algorithm was used **[7]**, which combines *MLR* and the following descriptor selection procedure:

1. search and select all orthogonal (to a given threshold) descriptor pairs in a given descriptor space.
2. build two-parameter regression equations using descriptor subspace created in step 1.
3. create three-parameter regressions by adding additional noncollinear descriptor using the best two-parameter equations from step 2.
4. create higher rank regression equations recursively like in step 3.

Firstly, the number of descriptors in final equation should not exceed certain limits because it could lead to a overfitted model. Secondly, the initial descriptor space should be constructed using chemical intuition. As an example extended *BMLR* approach could be used for fixing descriptors that a researcher wants to see in final models, if he wants to investigate concrete parameter dependence.

Artificial neural networks

Artificial neural networks (*ANN*) **[8]** were originally designed as a model for the activity of the human brain. Recently, computational neural networks have been employed as nonlinear models for *QSAR***[9,10]**.

The word *network* in the term "*artificial neural network*" arises because the function $f(x)$ is defined as a composition of other functions $g_i(x)$,which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables.

There are many different kinds of neural networks:

- single- and multilayer perceptron networks
- Kohonen self-organizing networks
- Recurrent networks
- Stochastic neural networks
- Modular neural networks

2.3. Molecular descriptors

Molecular descriptors **[6]** are mathematical values that describe the structure or shape of molecules. These can be calculated using several available *QSAR* programs (*CODESSA PRO* **[11]**, *ADRIANA.Code***[12]**, *DRAGON***[13]** , *etc*) and then used for creating structure-activity or structure-property relationships.

All molecular descriptors can be organized to more general groups:

- Constitutional descriptors - depend on types of atoms and bonds forming the molecule
- Topological descriptors – depend on the graph representation of the molecule
- Geometrical descriptors – depend on molecular structure and conformation
- Electrostatic descriptors – depend on the molecular charge distribution and charge density
- Quantum chemical descriptors – depend on the quantum chemical representation of the molecule (electronic and nuclear interactions, orbital energies, location, etc).

The latter includes molecular orbital related descriptors – *LUMO/HOMO* energies, *Fukui indices* **[14]**, which in turn is represented by 3 main types of descriptors:

Fukui atomic nucleophilic reactivity index:

$$N_A = \sum_{i \in A} c_{iHOMO}^2 / (1 - E_{HOMO})$$

Fukui atomic electrophilic reactivity index:

$$E_A = \sum_{j \in A} c_{jLUMO}^2 / (E_{LUMO} + 10)$$

Fukui atomic one-electron reactivity index:

$$R_A = \sum_{i \in A} \sum_{j \in A} c_{iHOMO} * c_{jLUMO} / (E_{LUMO} - E_{HOMO})$$

E_{HOMO} – highest occupied molecular orbital energy

E_{LUMO} – lowest unoccupied molecular orbital energy

c_{iHOMO} – highest occupied molecular orbital MO coefficients

c_{jLUMO} – lowest unoccupied molecular orbital MO coefficients

The difference of E_{HOMO} and E_{LUMO}, termed the *band gap*, can sometimes serve as a measure of the excitability of the molecule: the smaller the energy, the more easily it will be excited. Such properties like compound electrophilic or nucleophilic capability can be easily described with molecular Fukui indices.

2.4. Frontier orbital theory

The frontier orbital theory **[15]** developed by K. Fukui and others postulates that only two molecular orbitals are essential for determining a wide range of chemical reactions. These two *frontier orbitals* are frequently abbreviated to *HOMO, the highest occupied molecular orbital*, and *LUMO, the lowest unoccupied molecular orbital*.

The chemical concepts such as acid and base, oxidation and reduction appeared a long time ago and are connected to molecular reactivity. Their analogs such as electrophilicity, nucleophilicity, and electron donors and acceptors are directly connected to molecular orbital structure and electron density.

Electron density theory explains electrophile vs. nucleophile behavior in the simplest manner: that an electrophilic reagent will attack locations of large electron density in a molecule while a nucleophilic reaction will occur at the site of small electron density. Apparently this does not work in all cases. The frontier orbital theory developed by K. Fukui **[16]** is based on moleacular orbitals stating that only two of them are essential in a wide range of chemical reactions.

Apart from electrostatic interactions, the overlap between orbitals may favor the reaction between an electron donor and an electron acceptor. A high (in energy) lying occupied orbital in the donor may overlap with a low lying empty one in the acceptor, leading to a net stabilization. The strength of the interaction is determined by:

- the energy difference between the two orbitals involved (the smaller the better)
- the amount of overlap between the orbitals (the larger the better)

These effects can be, however, diminished by:

- steric effects which hinder overlap of frontier orbitals
- unfavorable symmetry of participating frontier orbitals
- inaccessible orbital localization centers by chemical substitution

The *HOMO vs. LUMO* interaction has proven useful for the interpretation of the sign of reaction constants and the scale of a substituent constant in Hammett-type equations. The relative easiness of occurrence of cyclic addition reactions and interesting phenomena like regioselection and periselection have also been interpreted with considerable success simply by the knowledge of the *HOMO* and *LUMO* energies, their localization and nodal

structure.

According to the molecular orbital theory **[17]** there is a set of molecular orbitals of discrete energies for every molecule. In contrast to the orbitals described by valence bond theory, which are usually localized between two specific atoms, molecular orbitals can extend over the entire molecule. Extensions to frontier orbitals theory may sound like: essential are those close to the frontier orbitals that have the best localization on the hypothetic active centers of the molecule. It means that if some biomolecule reacts specifically with only a specific molecular fragment, the orbital participating in this process should have maximum and specific localization in the respective molecular space.

3. QSAR modeling of the toxic action of phenols

Chemical compounds may exhibit toxicity via various mechanisms of toxic action **[18]**. The mechanisms frequently involved in aquatic toxicity include (but are not limited to) narcosis, respiratory uncoupling, electrophilic reactivity, and central nervous system seizure. Compounds with the narcosis mechanism exhibit baseline toxicity or toxicity associated with hydrophobicity, and compounds with other mechanisms have toxicity higher than the baseline toxicity. A correct identification of toxicity mechanism helps the understanding of the toxic effects of a xenobiotic compound of a living organism.

Phenols are widely used both in industry and as consumer products. Widely used derivatives include 2,4-dichlorophenol and 2,4,5-trichlorophenol, which are precursors for the herbicides 2,4-dichlorophenoxy acetic acid and 2,4,5-trichlorophenoxy acetic acid, respectively, and chlorophenols, which are themselves used as bactericides, fungicides and herbicides **[19]**. Cresols form an important group of disinfectants, and some naturally occuring phenols such as thymol (2-isopropyl-5-methylphenol) and carvacrol (5-isopropyl-2-methylphenol) are also known for their antiseptic action. Environmentally important nitrophenols include 3-trifluoromethyl-4-nitrophenol (TFM), which has been used as larval lampricide to control the sea lamprey (Petromyzon marinus) in the Great Lakes for more than 30 years.

The toxicity of phenols involves a number of different mechanisms and modes of action (*MOA*) **[20]**. The ability to act as oxidative uncouplers is associated to pK_a values (negative logarithms of ionization constant) in the range 3.8 to 8.5. The energy of the lowest unoccupied molecular orbital, E_{LUMO}, has been demonstrated to discriminate various *MOAs* **[21].** This parameter may reflect both the tendency of phenols to attack electron-rich sites of endogenous macromolecules directly, and their ability to undergo metabolic activation following 1-electron reduction. Nitrobenzene derivatives often show acute toxicities above the baseline toxicity toward aquatic organisms, which has typically been explained by the electrophilic nature of these chemicals.

Existing methods for classifying phenols according to *MOAs* can be grouped into two types of approaches – a qualitative approach, based on simple structural characteristics (such as the presence of a certain substituent), or a quantitative approach based on statistical analyses of physico-chemical properties **[22]**. The first approach is simple and relatively

successful for phenols with only few substituents. However, it becomes inefficient in case substituents associated with different *MOAs* are present in a molecule. Classification based on physico-chemical properties also has some disadvantages. These include the availability and use of the descriptors, the difficulty of mechanistic interpretation with some types of descriptors, and the fact that the property profile of the initial compound may differ significantly from the metabolically activated toxicant.

4. Molecular orbital selection approach in toxicity modeling

4.1. Molecular orbital selection

In order to facilitate orbital specific descriptor calculations the respective interface of the *QSAR* program was developed. The in-house *QSARModel* program enables the user to select the atoms and orbitals of interest and calculate descriptors specifically for those. Examples of *Mopac* calculated frontier molecular visualization for p-nitrophenol are given on *Figures 1* and *2*:

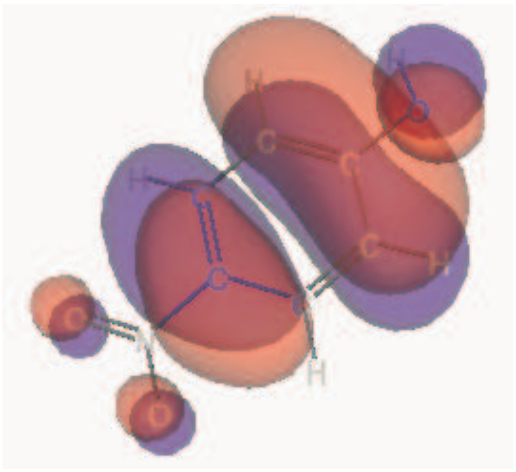


Figure 1: p-nitrophenol HOMO 3D isosurface view from QSARModel

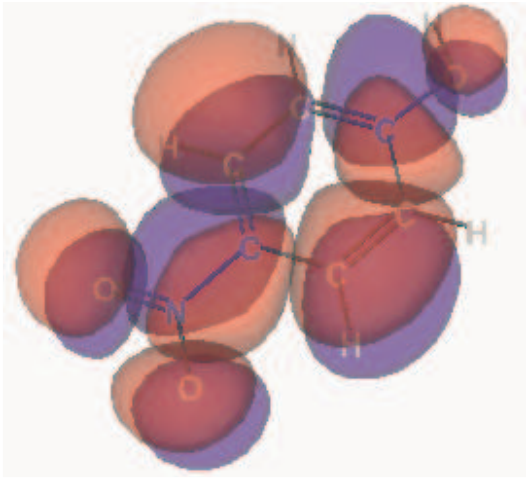


Figure 2: p-nitrophenol LUMO 3D isosurface view from QSARModel

The graphical user interface also supports comfortable data analysis with the help of data tables, data charts, multilinear regression analyzers, etc. Data manipulation is maximally dynamic. *Figure 3* (see in the supplementary data) shows one of the possibilities what can be done with data: chosen calculation menu for molecule in *„New Table“* table.

This program was used in current work for unoccupied molecular orbital selection. Suitable semi-empirical *AM1* **[23]** unoccupied molecular orbitals were selected by visual criteria. The main selection rule for this process was the *UMO's* preferable localization on the most electrophilic carbon atoms that visually had the most intensive positive charge coloring (see *Figures 4,5,6* in the supplementary data). In most cases $LUMO_{+1}$ was preferred, but some molecules (especially alkoxyphenols) showed acceptable localization of the default *LUMO*.

4.2. Initial datasets and their pre-analysis

The toxicity of phenols has been widely studied by different methods. In *QSAR* modeling, the toxicity has been frequently correlated with *LUMO* energy, especially for electrophilic phenol derivatives **[21,24,25,26]**. In this work, we have examined a possibility to use a wider selection of orbitals. First, a data set on the toxicities of phenol derivatives was prepared and correlations made with default frontier orbital (*LUMO*) dependent descriptor values. However, specific unoccupied molecular orbitals, or *UMOs*, were selected thereafter by using exact rules, and new correlations were build up using the new descriptor values corresponding to these orbitals.

Mopac 6.0 and force-field *AM1* **[23]** were used for all quantum-chemical calculations . Molecular orbitals were visualized and descriptors were calculated using *AM1* output files.

Pre-analysis of molecular structure vs. molecular orbital showed that benzenes and their derivatives generally have two lowest degenerate and complementary unoccupied molecular orbitals. In case of non-substituted benzenes these two orbitals are highly symmetric with close energy values. Substituents change the picture making one of them more localized on aromatic ring atoms, and the other more localized on bonds. Strong electron withdrawal groups in most cases change this picture very significantly. For example a nitro group makes the *LUMO* to spread over the entire molecule with preferred localization on N-C bond and oxygen atoms. The more substituents benzene has the more brindled picture would be observed.

For testing our approach three different datasets from different sources were used.

Dataset 1

Based on 40 substituted benzenes **[25]:** nitrobenzenes, nitroanilines, halogenated benzenes. Article authors used Mopac 6.0 - AM1 for energy-minimization and the LUMO energy calculation. The final model obtained (R^2 =0.793, SE = 0.316, F = 71.07, n = 40):

$$\log 1/EC_{50} = 0.272 * \log K_{OW} - 0.659 * E_{LUMO} + 2.54$$

We used the same dataset but with the following remarks:

- 4-aminoethylphenolate was not taken into account, as article does not contain precise description of its structure (if authors used the deprotonated form of this compound or not)

- different values for the LUMO energies were obtained

36 compounds from the 39 were used as the training set, 3 were left for the testing set (picked randomly):

- 2,4 – dichlorophenol (nr. 5)
- 3-bromoaniline (nr. 18)
- m-dinitrobenzene (nr. 32)

Table 1. The AM1 calculated E_{LUMO}, E_{UMO} and E_{LUMO}, log K_{OW}, log $1/EC_{50}$ values from **[25]** *for dataset nr. 1*

N	Name	E_{LUMO} (eV, calc.)	E_{UMO} (eV, calc.)	E_{LUMO} (eV,[25])	log K_{OW} [25]	log $1/EC_{50}$ (mol/l,[25])
1	2,3-dichloroaniline	-0.0100	-0.0100	-0.198	2.86	3.98
2	2,4,6-tribromoaniline	-0.300	-0.300	-0.342	4.03	4.37
3	2,4,6-trichlorophenol	-0.502	-0.502	-0.821	3.69	3.81
4	2,4-dichloroaniline	0.0158	0.0917	-0.179	2.91	3.74
5	2,4-dichlorophenol	-0.245	-0.105	-0.427	2.92	3.62
6	2,4-dinitroaniline	-1.474	-1.474	-1.662	1.84	4.68
7	2,4-dinitrophenol	-1.887	-1.561	-1.808	1.54	4.16
8	2,4-dinitrotoluene	-1.841	-1.450	-2.096	2.04	4.52
9	2,5-dichloroaniline	-0.0665	0.172	-0.263	2.92	3.82
10	2,5-dichloronitrobenzene	-1.284	-0.886	-1.343	2.9	4.31
11	2,6-dinitrotoluene	-1.749	-1.452	-2.000	2.02	4.06
12	2-chloroaniline	0.285	0.414	0.192	1.76	2.89
13	2-methylaniline	0.601	0.691	0.594	1.4	2.34
14	2-nitroaniline	-0.795	0.0647	-0.942	1.85	3.33
15	2-nitroanisole	-0.931	-0.220	-1.077	1.8	3.44
16	2-nitrophenol	-1.184	-0.274	-1.153	1.89	3.51
17	3,4-dichloronitrobenzene	-1.524	-0.824	-1.518	3.29	4.52

18	3-bromoaniline	0.165	0.367	0.190	2.1	2.8
19	3-bromonitrobenzene	-1.306	-0.703	-1.352	2.64	4.32
20	3-chloroaniline	0.263	0.391	0.160	1.88	2.79
21	3-nitroaniline	-0.950	-0.153	-1.031	1.37	3.48
22	3-nitroanisole	-1.072	-0.258	-1.216	2.16	3.71
23	3-nitrophenol	-1.160	-0.340	-1.281	2.0	3.75
24	4-bromonitrobenzene	-1.413	-0.562	-1.461	2.55	3.88
25	4-methylaniline	0.615	0.675	0.173	1.39	3.19
26	4-nitroaniline	-0.785	-0.0908	-0.875	1.39	3.4
27	4-nitroanisole	-0.983	-0.316	-1.146	2.03	3.65
28	4-nitrophenol	-1.065	-0.416	-1.205	2.04	3.57
29	4-nitrotoluene	-1.045	-0.235	-1.263	2.34	3.74
30	aniline	0.639	0.639	0.755	1.03	2.56
31	m-chloronitrobenzene	-1.285	-0.637	-1.359	2.49	3.95
32	m-dinitrobenzene	-1.911	-1.911	-2.087	1.52	4.85
33	nitrobenzene	-1.068	-0.312	-1.218	1.89	3.26
34	o-chloronitrobenzene	-1.070	-0.596	-1.226	2.26	3.94
35	o-dinitrobenzene	-1.841	-1.841	-2.157	1.69	5.04
36	p-chlornitrobenzene	-1.344	-0.547	-1.420	2.35	4.01
37	p-dinitrobenzene	-2.208	-2.208	-2.365	1.46	4.96
38	pentachlorophenol	-0.977	-0.977	-1.431	5.12	4.63
39	phenol	0.398	0.398	0.395	1.46	2.46

Dataset 2

Based on 19 nitrobenzenes and their derivatives [26]. This article contains a brief analysis of the dataset. Authors used the same scheme for quantum-chemical calculations in good agreement (with 1/1000 precision) with our obtained values for descriptors. Their model including the LUMO energy was based on 18 nitrobenzenes (picric acid was excluded, the only compound from the dataset that is deprotonated to more than 99.9% at the pH of the biotest medium) abd has the following properties ($R^2 = 0.89$, $R^2_{cv} = 0.88$, SE = 0.42, $F_{2,15} = 61$, n = 18):

$$\log EC_{50} = -0.61(\pm 0.11) * \log K_{OW} + 1.595(\pm 0.299) * E_{LUMO} - 1.19(\pm 0.36)$$

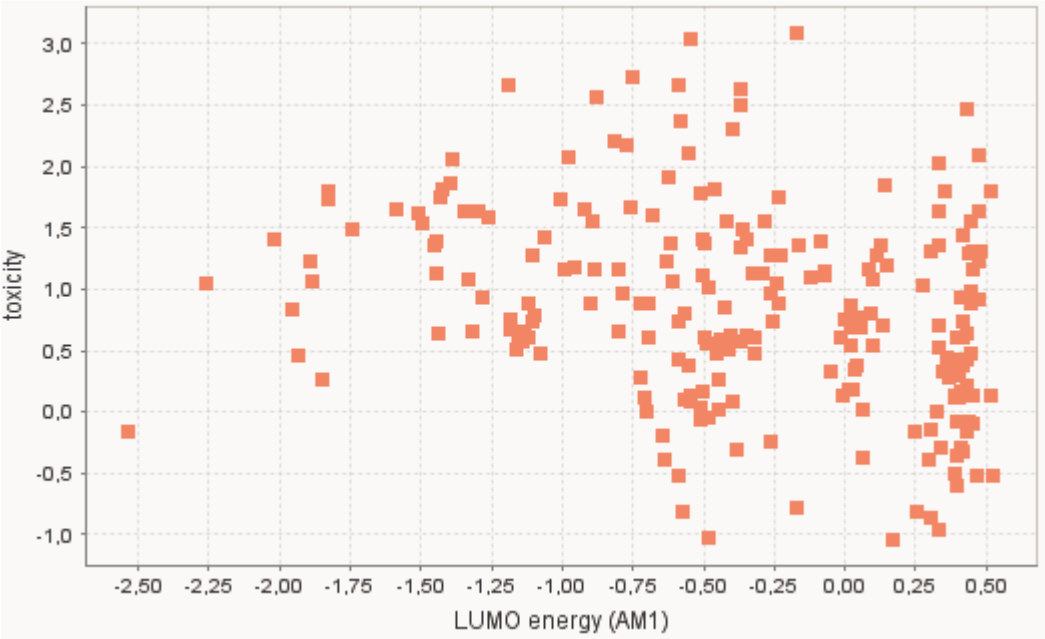
*Table 2. The AM1 calculated E_{LUMO}, E_{UMO} and log K_{OW},log EC_{50} values from **[26]** for the dataset nr. 2*

N	Name	E_{LUMO} (eV, calc.)	E_{UMO} (eV, calc.)	log K_{OW} **[26]**	log EC_{50} (mol/l,**[26]**)
1	2,3-diaminonitrobenzene	-0.842	-0.0546	1.27	-3.62
2	2,4,5-trichloronitrobenzene	-1.692	-1.049	3.48	-5.71
3	2,4,6-trinitrophenol (picric acid)	-2.534	-1.259	0.89	-2.96
4	2-amine-5-chloronitrobenzene	-1.003	-0.252	2.72	-4.26
5	3,4-dichloronitrobenzene	-1.524	-0.824	3.12	-5.78
6	3-nitroaniline	-0.950	-0.153	1.37	-3.56
7	4-amine-3-methylnitrobenzene	-0.757	-0.103	1.83	-4.04
8	4-chloro-1,3-dinitrobenzene	-2.061	-1.565	2.17	-5.53
9	4-chloro-3-methylnitrobenzene	-1.283	-0.517	2.9	-4.89
10	4-nitroaniline	-0.785	-0.0908	1.39	-3.48
11	4-nitrotoluene	-1.045	-0.235	2.37	-3.9
12	TFM	-1.586	-1.036	2.77	-4.78
13	dintiramine	-1.805	-1.805	4.3	-7.34
14	m-chloronitrobenzene	-1.285	-0.637	2.46	-5.11
15	m-dinitrobenzene	-1.911	-1.911	1.49	-5.6
16	nitrobenzene	-1.068	-0.312	1.85	-3.58
17	o-chloronitrobenzene	-1.070	-0.596	2.24	-3.82
18	p-chloronitrobenzene	-1.344	-0.547	2.39	-4.66
19	trifluralin	-1.994	-1.994	5.34	-7.14

Dataset 3

Based on 221 phenols and their derivatives [21]. The authors used E_{LUMO} mostly for the classification of compounds into different categories. According to this article only 23 molecules from the whole dataset show strong electrophilic capability (all have one nitro- or nitroso- group as substitient). We prepeared a similar set of 229 phenols and used it for the QSAR model building. The picture in Graph 1 shows how poor is the correlation between the toxicity data and the AM1 lowest unoccupied molecular orbital energies for the whole dataset. It means that a single mechanism of toxic action connected to compound electrophilicity is not applicable for the whole dataset. The charge distribution and orbital locations of all of these compounds was visually examined for the existence of electrophilic behavior, resulting in the selection of more or less suitable compounds for the training set. Overall, 60 molecules with high positive charge density on aromatic carbons and adjacent hydrogens were selected. Some demonstrative examples for the selection are given on *Figures 4,5* and *6* (see in supplementary data). In *Figure 4*, 3,5-dimethyl phenol has no electrophilic aromatic carbons (no “positive” blue coloring on aromatic carbons), as shown in the colored charge map of this molecule. 4-methylphenol (*Figure 5*) to some extent electrophilic aromatic carbons, that was acceptable for current *QSAR* modeling. Finally, 2-chlorophenol (*Figure 6*) has good visible “positive” regions, especially at the meta-position to the hydroxyl group.

Similarly, 10 molecules having appropriate visual appearance for electrophilic aromatic carbons were selected for the test set. The electrophilic phenols having 3 or more additional substituents were also excluded from the calculations (for example 2,3,4,6-tetrachlorophenol), due to the absence of accessible aromatic carbon for direct reaction on it. However, compound 60, pentafluorophenol, was included because of the much smaller *van der Waals’* radius of fluorine as compared to the chloro or methyl groups. Phenols having more than one *+I* substituent (for example 2-chloro-4,5-dimethylphenol), show much smaller positive charge density on aromatic carbons and thus they were also excluded. Preliminary calculations also showed that phenols containing amino groups significantly mixed up the final picture. These aminophenols are classified as precursors of soft electrophiles because of their ability to undergo oxidation to nitrophenols.



Graph 1: Toxicity vs. LUMO energy values map for initial data set of 229 phenol derivatives

Table 3. The AM1 calculated E_{LUMO}, E_{UMO} and log 1/EC_{50} values for dataset nr. 3

N	Name	E_{LUMO} (eV, calc.)	E_{UMO} (eV, calc.)	log 1/EC_{50} (mmol/l,**[27,28,29,30]**)
1	2,3-dichlorophenol	-0.262	-0.262	1.276
2	2,4-dibromophenol	-0.348	-0.167	1.4
3	2,4-dichloro-6-nitrophenol	-1.431	-0.921	1.75
4	2,4-dichlorophenol	-0.245	-0.105	1.04
5	2,4-difluorophenol	-0.318	-0.318	0.604
6	2,5-dichlorophenol	-0.325	-0.0121	1.125
7	2,6-dichloro-4-fluorophenol	-0.568	-0.392	0.804
8	2,6-dichlorophenol	-0.258	-0.258	0.73
9	2,6-difluorophenol	-0.321	-0.159	0.471
10	2,6-dimethoxyphenol	0.396	0.396	-0.598
11	2-(chloromethyl)-4-nitrophenol	-1.186	-0.704	0.75
12	2-bromo-4-methylphenol	-0.0132	0.226	0.6
13	2-bromophenol (test set)	-0.0494	0.235	0.33
14	2-chloro-4-hydroxybenzaldehyde(test set)	-0.696	-0.193	0.89
15	2-chloro-5-methylphenol	0.0205	0.0205	0.54
16	2-chlorophenol	0.0295	0.0295	0.183
17	2-ethoxyphenol	0.398	0.398	-0.358

18	2-fluoro-4-nitrophenol (test set)	-1.333	-0.770	1.073
19	2-fluorophenol	0.0133	0.180	0.19
20	2-hydroxybenzaldehyde	-0.591	0.217	0.42
21	2-hydroxybenzamide	-0.264	0.259	-0.242
22	2-hydroxybenzonitrile	-0.509	0.0435	0.034
23	2-hydroxybenzyl alcohol	0.329	0.583	-0.954
24	2-methoxyphenol	0.391	0.391	-0.51
25	2-methyl-3-nitrophenol	-1.100	-0.293	0.779
26	2-methylphenol	0.409	0.531	-0.29
27	2-nitrophenol	-1.184	-0.274	0.67
28	2-nitroresorcinol (test set)	-1.320	-0.241	0.66
29	3,4-dichlorophenol	-0.236	-0.236	1.745
30	3,5-dimethoxyphenol	0.453	0.453	-0.09
31	3-acetamidophenol	0.244	0.489	-0.16
32	3-bromophenol (test set)	-0.0742	-0.0742	1.145
33	3-chloro-4-fluorophenol	-0.292	-0.112	1.131
34	3-chlorophenol	0.0187	0.232	0.87
35	3-ethoxy-4-methoxyphenol	0.337	0.337	-0.299
36	3-ethylphenol	0.400	0.400	0.29
37	3-fluoro-4-nitrophenol	-1.285	-0.721	0.935
38	3-fluorophenol	0.0417	0.171	0.381
39	3-hydroxybenzaldehyde	-0.547	0.188	0.08
40	3-hydroxybenzonitrile	-0.514	0.0253	-0.064
41	3-methoxyphenol (test set)	0.413	0.413	-0.33
42	3-methyl-2-nitrophenol (test set)	-1.121	-0.238	0.61
43	3-methyl-4-nitrophenol	-1.007	-0.367	1.729
44	3-nitrophenol	-1.160	-0.340	0.51
45	4-acetamidophenol	0.253	0.431	-0.82
46	4-bromo-2,6-dichlorophenol	-0.514	-0.453	1.78
47	4-bromo-6-chloro-2-methylphenol	-0.225	-0.151	1.28
48	4-bromophenol	0.0201	0.121	0.68
49	4-chloro-2-methylphenol	0.135	0.135	0.7
50	4-chloro-2-nitrophenol	-1.388	-0.594	2.06
51	4-chloro-3-methylphenol	0.0930	0.0930	0.8

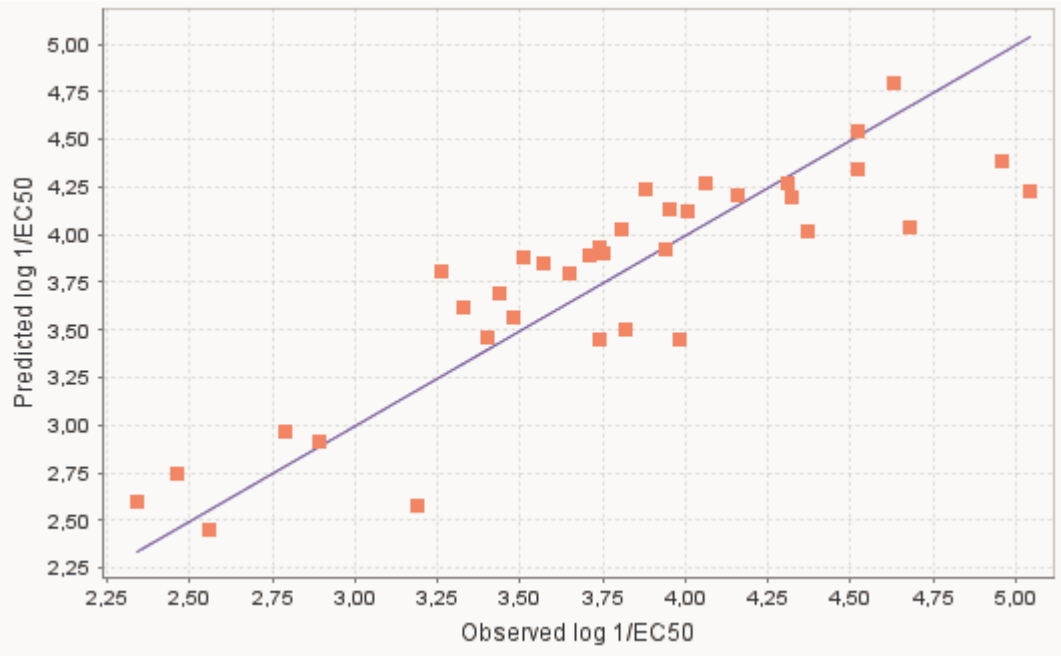
52	4-chloro-3-nitrophenol (test set)	-1.104	-0.630	1.27
53	4-chlorophenol	0.0946	0.143	0.54
54	4-ethoxyphenol	0.328	0.328	0.01
55	4-fluoro-2-nitrophenol (test set)	-1.447	-0.625	1.384
56	4-fluorophenol	0.0592	0.0592	0.02
57	4-hydroxybenzaldehyde	-0.450	0.101	0.266
58	4-hydroxybenzonitrile	-0.413	-0.0676	0.516
59	4-hydroxybenzotrifluoride	-0.348	-0.141	0.62
60	4-hydroxybenzyl cyanide	0.0631	0.153	-0.38
61	4-methoxyphenol	0.304	0.304	-0.14
62	4-methyl-3-nitrophenol	-1.107	-0.305	0.74
63	4-methylphenol (test set)	0.429	0.487	-0.16
64	4-nitro-3-(trifluoromethyl)phenol	-1.588	-1.015	1.65
65	4-nitrophenol	-1.065	-0.416	1.43
66	4-nitrosophenol	-0.799	0.0301	0.654
67	5-fluoro-2-nitrophenol	-1.448	-0.556	1.125
68	5-methyl-2-nitrophenol	-1.152	-0.192	0.59
69	N,2-dihydroxybenzamide	-0.556	0.123	0.379
70	Pentafluorophenol	-1.296	-1.133	1.63

4.3. Results and discussion

Dataset 1

Two 2-parameter models (see *Graph 2 vs. Graph 3*) were constructed with 36 molecules using values from *Table 1*. Like the model from the original paper **[25]** it uses the lowest molecular orbital E_{LUMO} energy and *log* K_{OW} values (R^2 = 0.760, R^2_{cv} = 0.709, F = 52.2, s^2 = 0.111):

Model 1: $\log 1/EC_{50} = 0.324 * \log K_{OW} - 0.630 * E_{LUMO} + 2.52$

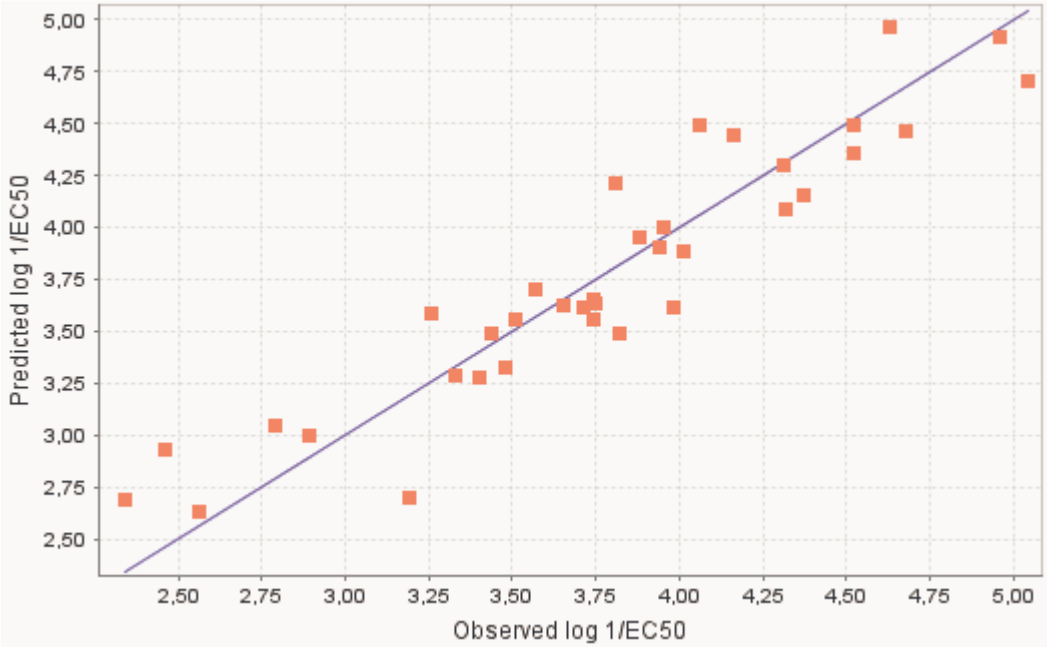


Graph 2: The plotting of predicted vs. experimental toxicity according to Model 1

Using selected UMO energies instead of LUMOs significant improvement is observed (R^2 = 0.869, R^2_{cv} = 0.832, F = 109.7, s^2 = 0.061):

Model 2: $\log 1/EC_{50} = 0.270 * \log K_{OW} - 0.764 * E_{UMO} + 2.84$

The coefficients of E_{LUMO} or E_{UMO} in *Model 1* and *Model 2* equations, respectively, are more than 2 times larger than the coefficient of *log* K_{OW} although their values have similar magnitude. Therefore, it may be concluded that the electrophilic factor is the main controlling influence of the toxicity of nitrobenzenes, especially for compounds with low *log* K_{OW} value.



Graph 3: The plotting of predicted vs. experimental toxicity according to Model 2

Table 4. The deviations of the toxicity values predicted by Model 1 and Model 2

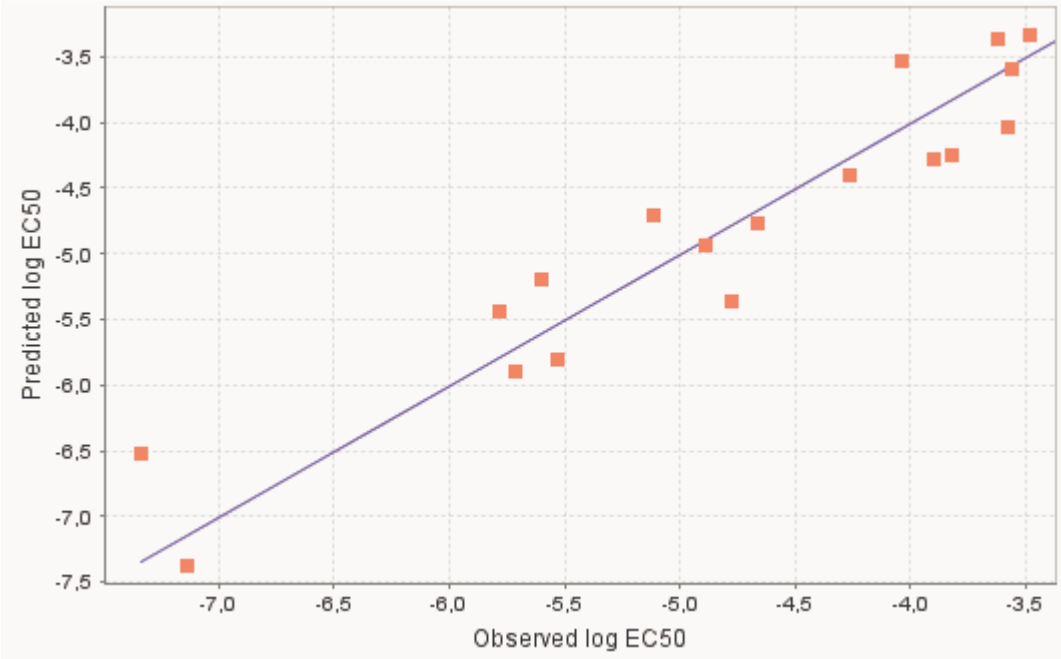
Name	*log 1/EC*$_{50}$ (mol/l,**[25]**)	fitted with *Model 1*	residuals	fitted with *Model 2*	residuals
2,4-dichlorophenol	3.62	3.62	0.0	3.71	0.09
3-bromoaniline	2.80	3.10	0.30	3.12	0.32
m-dinitrobenzene	4.85	4.22	-0.63	4.71	-0.14
standard deviation, s^2			0.15		0.04

Table 4 shows that in overall *Model 2* is more precise for predicting the toxicity values for the test set compounds, although *Model 1* was more exact in case of 2,4-dichlorophenol.

Dataset 2

The 2-parameter model (see also *Graph 4*) that uses the lowest molecular orbital E_{LUMO} energy and *log* K_{OW} values was created (R^2 = 0.894, R^2_{cv} = 0.836, F = 63.4, s^2 = 0.168):

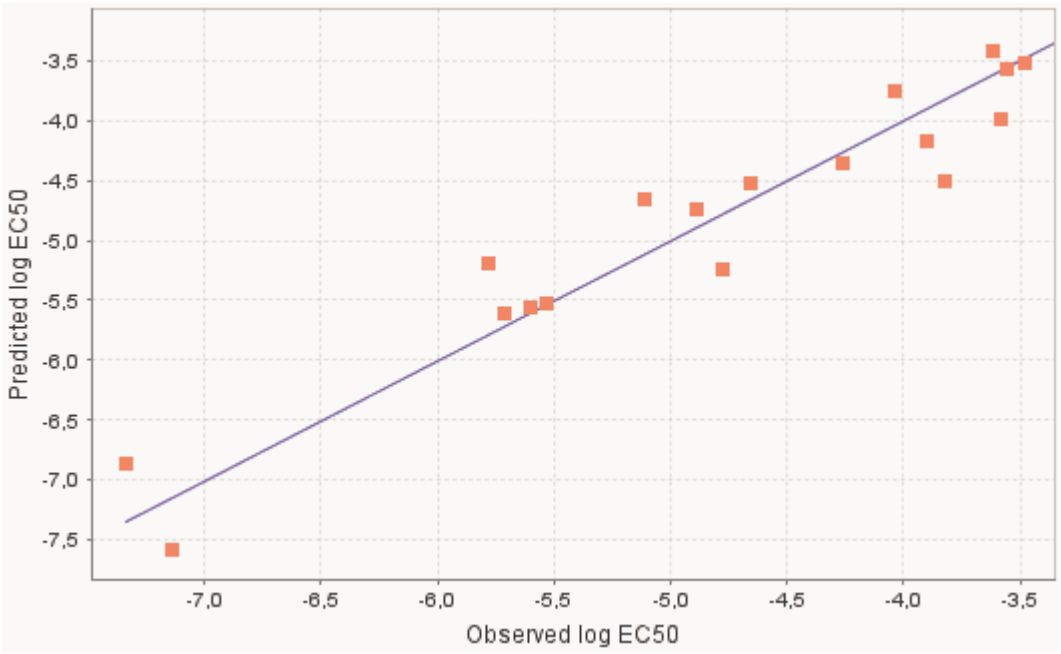
Model 3: $\log EC_{50} = 1.61 * E_{LUMO} - 0.530 * \log K_{OW} - 1.34$



Graph 4: The plotting of predicted vs. experimental toxicity according to Model 3

Model 3 has very little differencies with the original one **[26]**. The model (see *Graph* 5) using selected orbitals (*UMO* analog) (R^2 = 0.912, R^2_{cv} =0.873, F = 77.8, s^2 = 0.140) showes slightly better R^2 value, but not as significant as in the first case.

Model 4: $\log EC_{50} = 1.10 * E_{UMO} - 0.501 * \log K_{OW} - 2.72$

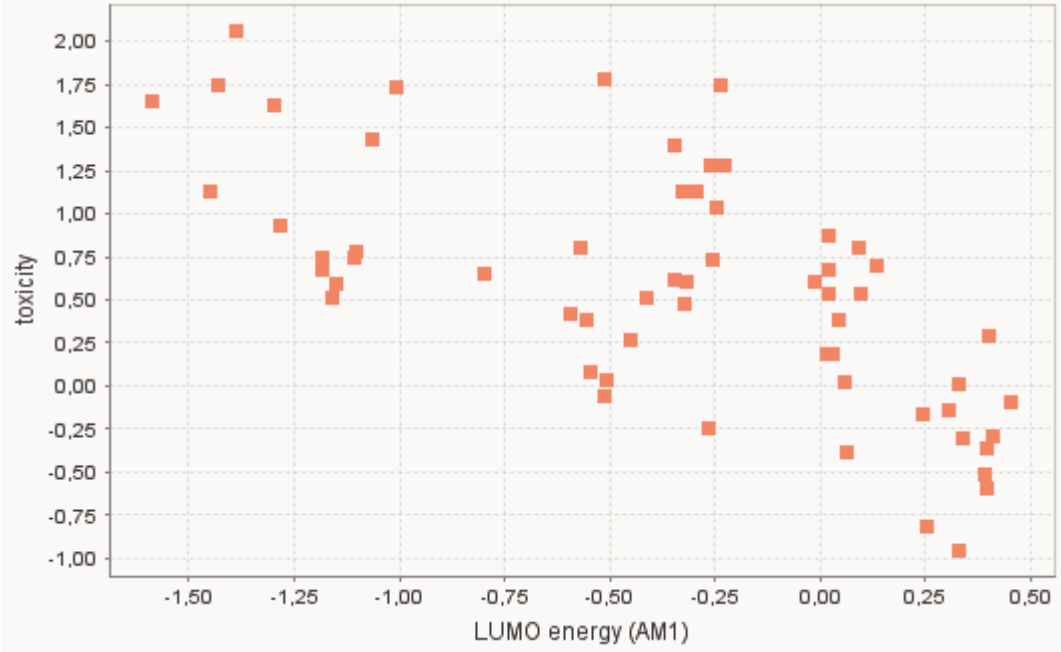


Graph 5: The plotting of predicted vs. experimental toxicity according to Model 4

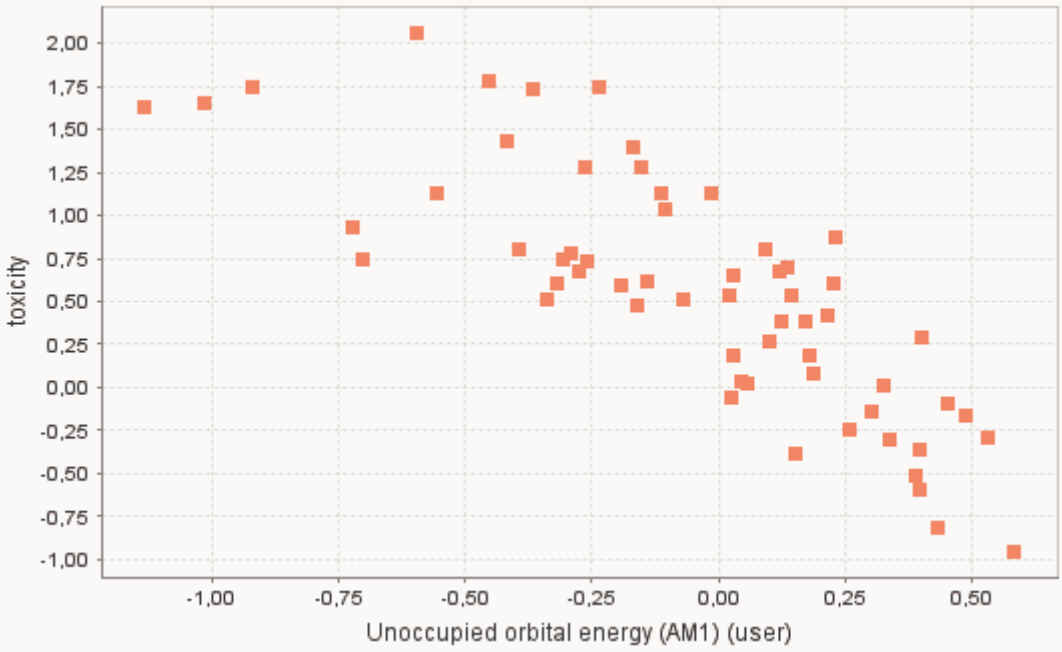
All created models (*Model 1*, *Model 2*, *Model 3*, *Model 4*) are in accordance with the explanation. The signs of coefficients show that toxicity increases with the decrease of EC_{50}, which in turn (according to equations) decreases with increasing *log* K_{OW} and descreasing (more negative) E_{LUMO} (E_{UMO})

Dataset 3

With given 60 molecules 1-parameter and 2-parameter models were constructed. It is obvious that the orbital selection procedure significantly improves the correlation between the toxicity values and the unoccupied molecular orbital energies (see *Graph 6* vs *Graph 7*):



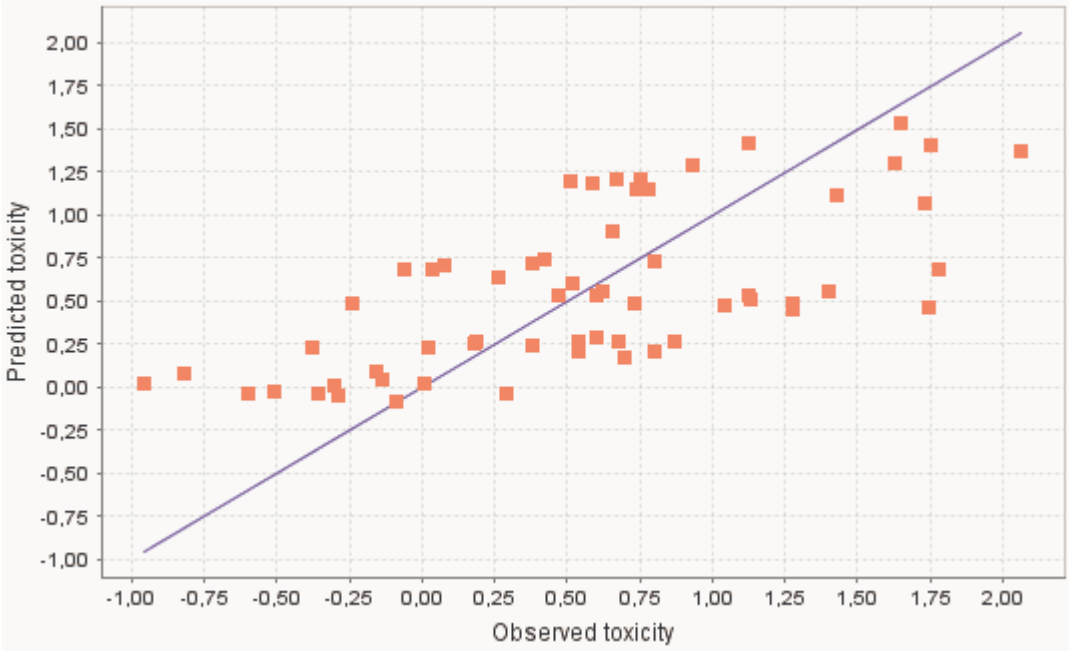
Graph 7: Toxicity vs. LUMO energy values map for training dataset 3



Graph 6: Toxicity vs. UMO energy values map for training dataset 3

The respective 1-parameter *QSAR* models (see also *Graph 8 vs. Graph 9*) are as follows. Model 5 (R^2 = 0.435, R^2_{cv} =0.401, F = 44.6, s^2 = 0.278) formula (training data is shown on *Graph 8*):

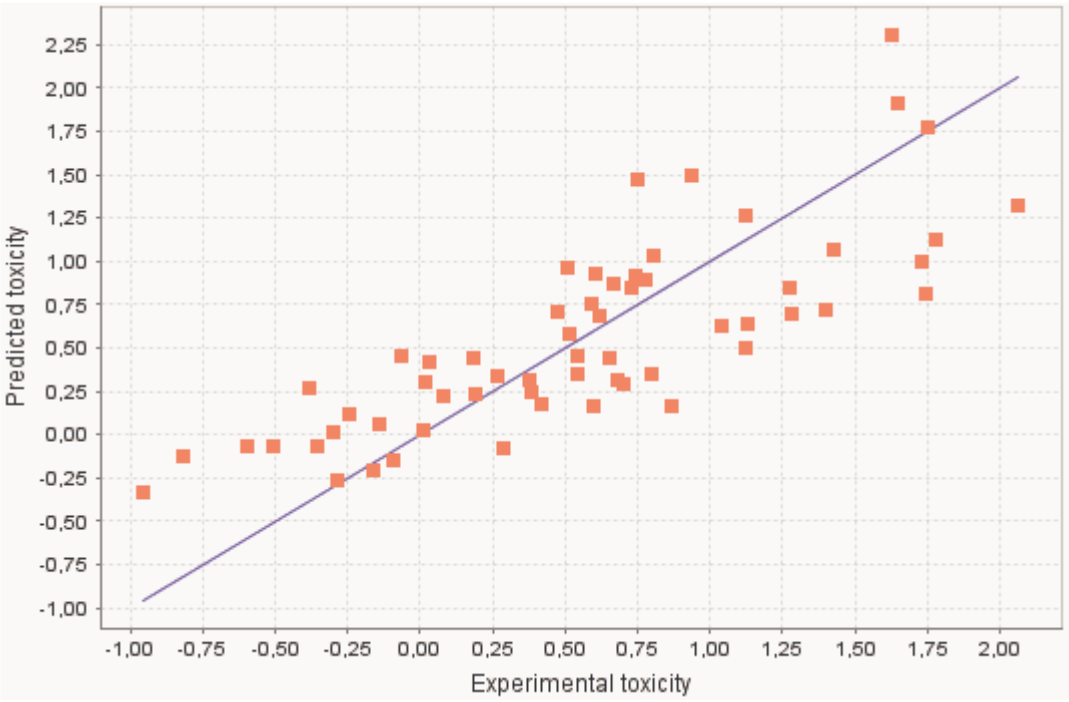
$$\log 1/EC_{50} = -0.787 * E_{LUMO} + 0.278$$



Graph 8: The plotting of predicted vs. experimental toxicity according to Model 5

Model 6 (R^2 = 0.640, R^2_{cv} =0.612, F = 103.1, s^2 = 0.177) formula (training data is shown on *Graph 9*):

$$\log 1/EC_{50} = -1.44 * E_{UMO} + 0.486$$



Graph 9: The plotting of predicted vs. experimental toxicity according to Model 6

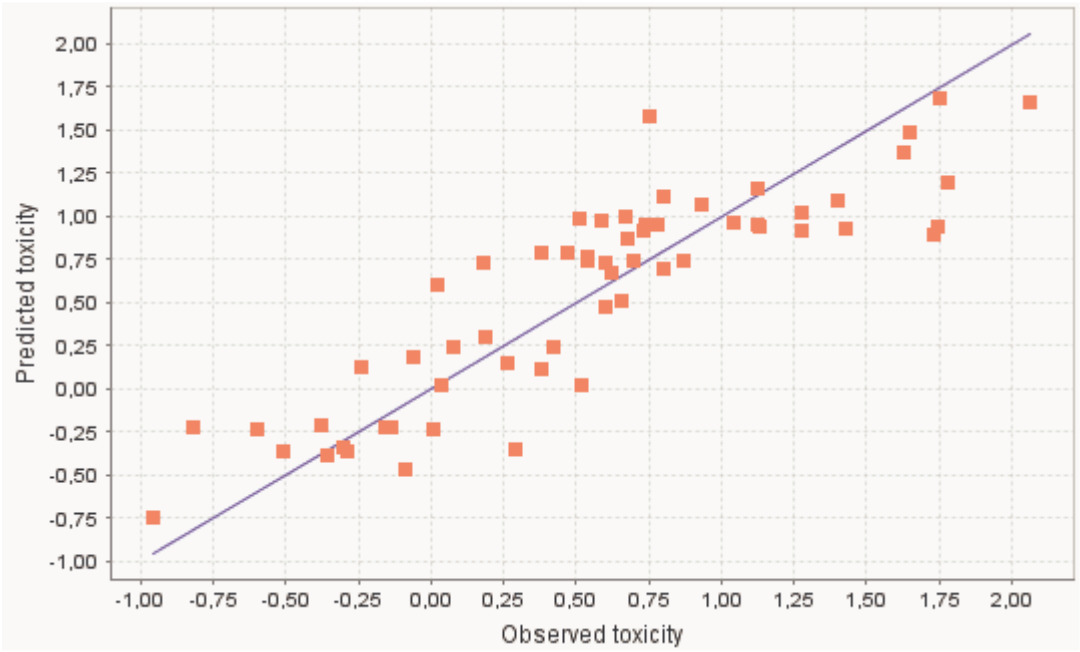
The prediction of toxicities for the test set of 10 compounds also improves significantly when user selected *UMO* energies were used instead of *LUMO* energies (cf. *Table 5*).

Table 5. The deviations of the toxicity values predicted by Model 5 and Model 6

Name	log *1/EC*$_{50}$ (mmol/l)	fitted with *Model 5*	residuals	fitted with Model 6	residuals
2-bromophenol	0.33	0.32	-0.01	0.15	-0.18
2-chloro-4-hydroxybenzaldehyde	0.89	0.83	-0.06	0.76	-0.13
2-fluoro-4-nitrophenol	1.07	1.33	0.26	1.59	0.52
2-nitroresorcinol	0.66	1.32	0.66	0.83	0.17
3-bromophenol	1.15	0.34	-0.81	0.59	-0.56
3-methoxyphenol	-0.33	-0.05	0.28	-0.11	0.22
3-methyl-2-nitrophenol	0.61	1.16	0.55	0.83	0.22
4-chloro-3-nitrophenol	1.27	1.15	-0.12	1.39	0.12
4-fluoro-2-nitrophenol	1.38	1.42	0.03	1.39	0.01
4-methylphenol	-0.16	-0.06	0.10	-0.21	-0.05
standard deviation, s^2			0.15		0.08

One of the best 2-parameter models depending on E_{LUMO} contains *Max atomic orbital electronic population* (*AM1*) **[6]** descriptor as a second predictor.

Model 7 (R^2 = 0.757, R^2_{cv} = 0.732, F = 89.0, s^2 = 0.121) has the following general



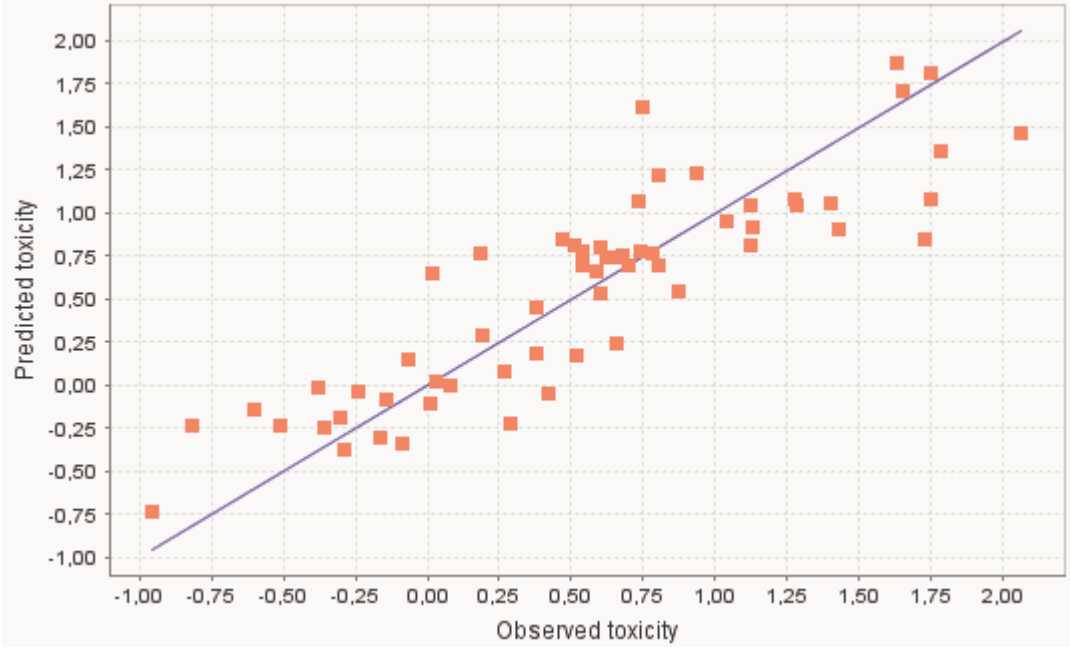
Graph 10: The plotting of predicted vs. experimental toxicity according to Model 7

formula (training data is shown on *Graph 10*):

$$\log 1/EC_{50} = -0.622 * E_{LUMO} + 12.5 * Max.AO\ electr.\ pop. - 23.9$$

Similar 2-parameter model that uses E_{UMO} values has a little better precision. Model 8 (R^2 = 0.764, R^2_{cv}=0.739, F = 92.0, s^2 = 0.118) has the following general formula (training data is shown on *Graph 11*):

$$\log 1/EC_{50} = -1.07 * E_{UMO} + 8.68 * Max.AO\ electr.\ pop. - 16.4$$



Graph 11: The plotting of predicted vs. experimental toxicity according to Model 8

Even in this "directed" model use of E_{UMO} shows better results. This is also seen from model testing (cf. *Table 6*). *Graph 9* and data in *Table 6* show that unoccupied orbital energies have a clear correlation with toxicity values, but it is also true that additional parameters are needed to create reliable models featuring the additional interactions related to the toxicity and possibly reducing the *AM1* calculation errors for this kind of energies. 3- or more parameter containing models for this dataset show less improvement in R^2 between E_{LUMO} and E_{UMO} using models. Best 4-parameter model has R^2 = 0.813 for E_{LUMO} and R^2 = 0.826 for E_{UMO}. In these multiparameter models orbital energies are less important, so they are not shown in current work.

Table 6. The deviations of the toxicity values predicted by Model 7 and Model 8

Name	*log 1/EC_{50}* (mmol/l)	fitted with *Model 7*	residuals	fitted with *Model 8*	residuals
2-bromophenol	0.33	0.75	0.42	0.52	0.19
2-chloro-4-hydroxybenzaldehyde	0.89	1.19	0.3	1.01	0.12
2-fluoro-4-nitrophenol	1.07	1.13	0.06	1.31	0.23
2-nitroresorcinol	0.66	1.07	0.41	0.7	0.04
3-bromophenol	1.15	0.68	-0.47	0.79	-0.36
3-methoxyphenol	-0.33	-0.4	-0.07	-0.27	0.06
3-methyl-2-nitrophenol	0.61	0.96	0.35	0.71	0.09
4-chloro-3-nitrophenol	1.27	1.44	0.17	1.47	0.2
4-fluoro-2-nitrophenol	1.38	1.37	-0.01	1.26	-0.12
4-methylphenol	-0.16	-0.37	-0.21	-0.32	-0.16
standard deviation, s^2			0.08		0.03

5. Conclusions

The goal to improve the accuracy and predicting power of *QSAR* models involving frontier orbital energies by selecting orbitals of suitable symmetry and localization was achieved. Orbital dependent descriptor calculation with human interference and their subsequent use for model building showed better results (statistically better models with more predicting power) than using the same orbital descriptors with default values for all three datasets.

However, the semi-empirical *AM1* method that was used in current work in order to have some comparison with originally proposed models, is not very precise in describing molecular electronic structure, including unoccupied orbital energies and compounds containing nitrogen atoms. Therefore, the next step could be the use of descriptors obtained from *ab initio* Hartree-Fock (*HF*) or post-HF calculations.

Default Fukui indexes, namely the atomic electrophilic reactivity index, did not show acceptable correlations with toxicity data for using them in *QSAR* model building. Some specific localized indexes like electrophilic index for aromatic centers, or selected atom electrophilic index could be implemented in future work.

6. Kokkuvõte

Eesmärk parandada QSAR mudelite täpsust ja ennustusvõimet, kasutades selleks orbitaalide valimist sümmeetria ja lokalisatsiooni alusel õnnestus täita. Inimese poolt juhitud orbitaalsõltuv deskriptoriarvutus ja selle edasine kasutamine mudelite ehitamiseks näitab paremaid tulemusi (statistiliselt paremad mudelid suurema ennustusvõimega) võrreldes samade orbitaalsete molekulaardeskriptorite vaikimisi väärtuste kasutamisega kõigi kolme andmekomplekti jaoks.

Käesolevas töös kasutati sama pool-empiirilist AM1 meetodit nagu originaalallikates, et oleks võimalus tulemusi artiklitest leitud originaalsete mudelitega võrrelda. See meetod ei ole väga täpne molekulaarse elektroonilise struktuuri kirjeldamiseks, eriti LUMO energiate ja lämmastiku sisaldavad ühendite korral. Seega võiks järgmiseks loogiliseks sammuks olla selliste deskriptorite kasutamine, mis on saadud *ab initio* Hartree-Fock (HF) or post-HF arvutuste põhjal.

Klassikaline Fukui indeks - *„atomic electrophilic reactivity index”* ei näidanud loodetud statistiliselt olulisi korrelatsioone toksilisuse andmetega, kasutamaks teda QSAR mudelite ehitamiseks. Edasise arengu võimalusena võiks defineerida spetsiaalsed lokaliseeritud indeksid, nagu aromaatse tsentri elektrofiilsuse indeks, või valitud aatomi elektrofiilsuse indeks.

References

1. http://en.wikipedia.org/wiki/Drug_design
2. http://en.wikipedia.org/wiki/QSAR
3. Borman, S. New QSAR techniques eyed for environmental assessments, *Chem. Eng. News*, **1990**, 68, 20-23.
4. http://books.nap.edu/html/biomems/lhammett.html
5. Stuper, A.; Brugger, W.; Jurs. Computer assisted studies of chemical structure and biological function. P. Wiley: New York, **1979**.
6. Karelson, M.. Molecular Descriptors in QSAR/QSPR. J. Wiley & Sons, New York, **2000**.
7. Katritzky, A.R., Lan Mu, Lobanov, V.S., Karelson, M.. Correlation of Boiling Points with Molecular Structure. I A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.*, **1996**, 100, 10400-10407.
8. http://en.wikipedia.org/wiki/Artificial_neural_network
9. Sild, S; Karelson, M. A General QSPR Treatment for Dielectric Constants of Organic Compounds. *J. Chem. Inf. Comput. Sci.*, **2002**, 42, 360-367.
10. Zupan, J.; Gasteiger, J. Neural networks for Chemists: an introduction. VCH-Verlag: Weinheim, **1993**, 213-228.
11. www.codessa-pro.com/
12. http://www.mol-net.de/software/adrianacode/index.html
13. http://www.talete.mi.it/dragon_net.htm
14. Franke, R. Theoretical Drug Design Methods. Elsevier, Amsterdam, **1984.**
15. Fukui, K.; Yonezawa, T.; Shingu, H. Perspective on "A molecular orbital theory of reactivity in aromatic hydrocarbons". *J. Chem. Phys.*, **1952**, 20, 722.
16. Fukui, K. The role of frontier orbitals in chemical reactions (Nobel Lecture). Angewandte Chemie International Edition in English, **1982**, 21, 801-809.
17. Leach, A.R. Molecular modelling. Longman, **1996**.
18. http://en.wikipedia.org/wiki/Toxic
19. Hattula, M. L.; Wasenius, V.-M.; Reunanen, H.; Arstila, A.U. Acute toxicity of some chlorinated phenols, catechols and cresols to trout. *Bull. Environ. Contam. Toxicol.*, **1981**, 26, 295-298.
20. Ren, S. Phenol mechanism of toxic action classification and prediction: a decision tree

approach.*Toxicol. Lett.*, **2003**, 144, 313-323.

21. Aptula, A. O.; Netzeva, T. I.; Valkova, I.V.; Cronin, M. T. D.; Schultz, T. W.; Kühne R.; Schüürmann G. Multivariate discrimination between modes of toxic action of phenols. *QSAR*, **2002** , 21, 12 – 22.
22. Livingstone D.J. Data analysis for chemists: application to QSAR and chemical product design. Oxford University Press, Oxford **1995.**
23. Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. AM1: a new general purpose quantum mechanical model. *J.Am.Chem.Soc*., **1977**, 99, 4899-4907.
24. Xiaodong Wang; Chunsheng Yin; Liansheng Wang. Structure-activity relationships and response-surface analysis of nitroaromatics toxicity to the yeast (Saccharomyces cerevisiae). *Chemosphere*, **2002**, 46, 1045-1051.
25. Guang-Hua Lu; Xing Yuan; Yuan-Hui Zhao. QSAR study on the toxicity of substituted benzenes to the algae (Scenedesmus obliquus). *Chemosphere*, **2001,** 44, 437-440.
26. Schmitt, H.; Altenburger R.; Jastorff, B.; Schüürmann, G. Quantitative structure-activity analysis of the algae toxicity of nitroaromatic compounds. *Chem. Res. Toxicol.*, **2000**, 13, 441-450.
27. Bearden, A.P.; Schultz, T.W. Comparison of Tetrahymena and Pimephales toxicity based on mechanism of action. *SAR QSAR Environ., Res.* **1998**, 9, 127-153.
28. Cronin, M.T.D.; Aptula, A.O.; Duffy, J.C.; Netzeva, T.I.; Rowe, P.H.; Valkova, I.V.; Schultz, T.W. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. *Chemosphere*, **2002**, 49, 1201-1221.
29. Schultz, T.W. Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. *Chem. Res. Toxicol.*, **1999**, 12, 1262-1267.
30. Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. The Use of Diversity versus Representativity in the Training and Validation of Quantitative Structure-Activity Relationships. *SAR and QSAR Environ. Res.*, **2003**, 14, 59-81.

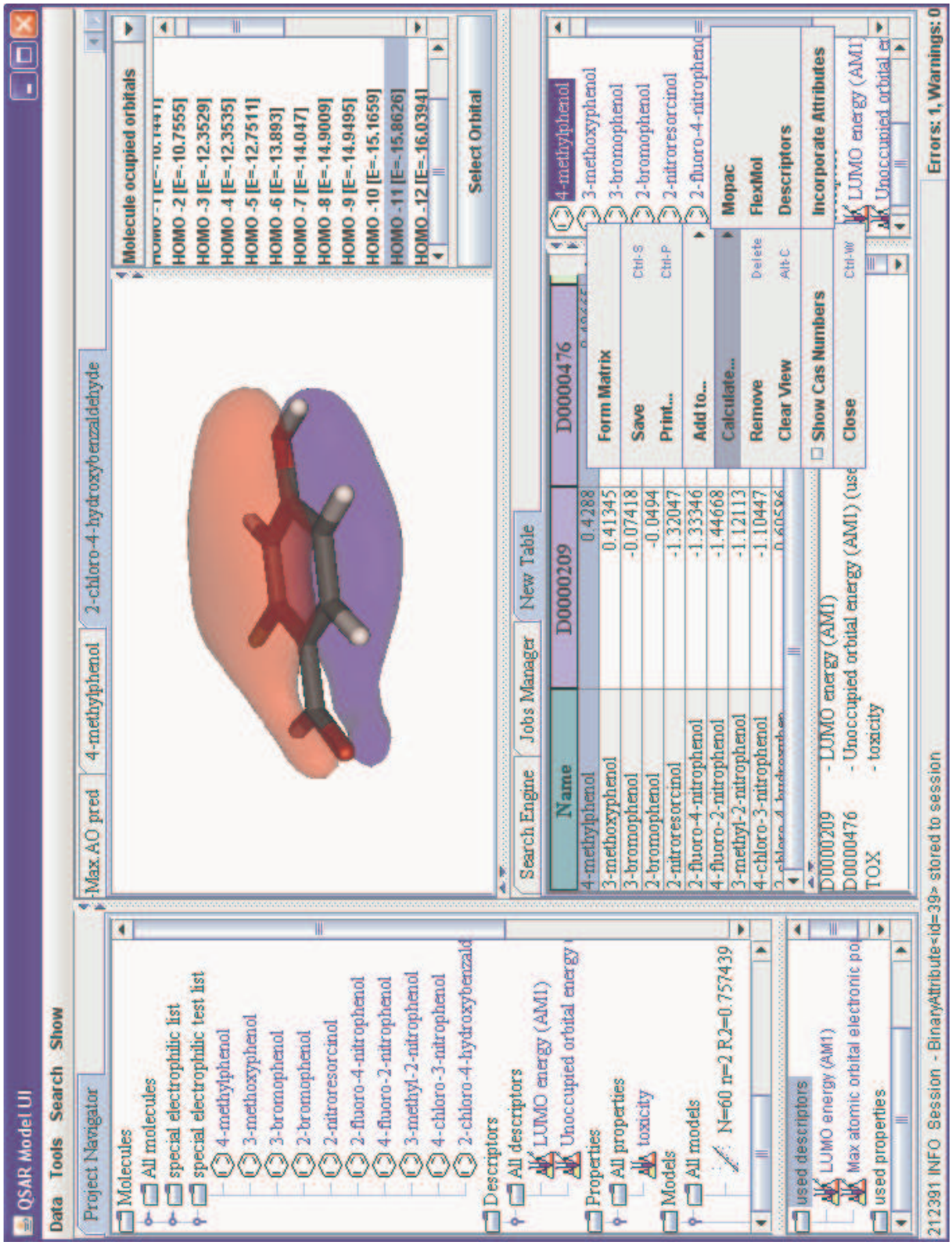


Figure 3: Data manipulation is maximally dynamic

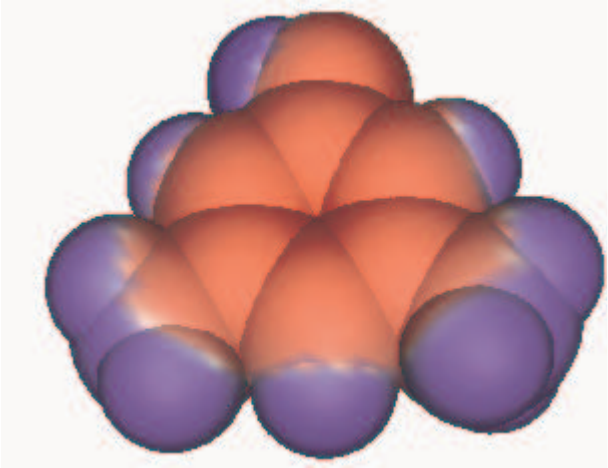


Figure 4: Color map of 3,5-dimethylphenol

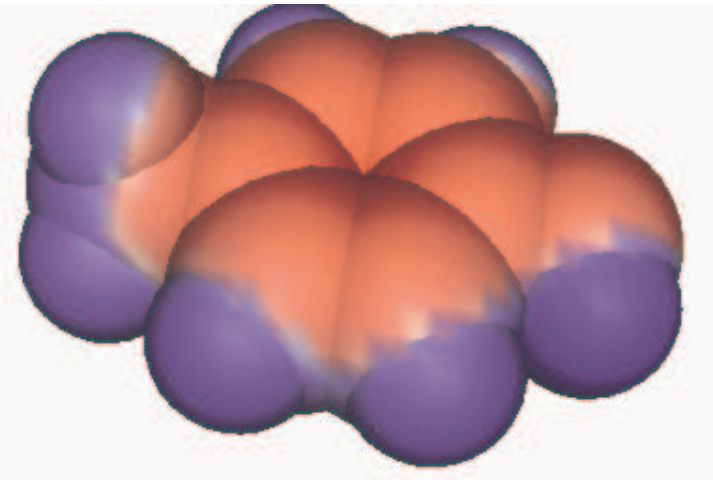


Figure 5: Color map of 4-methylphenol

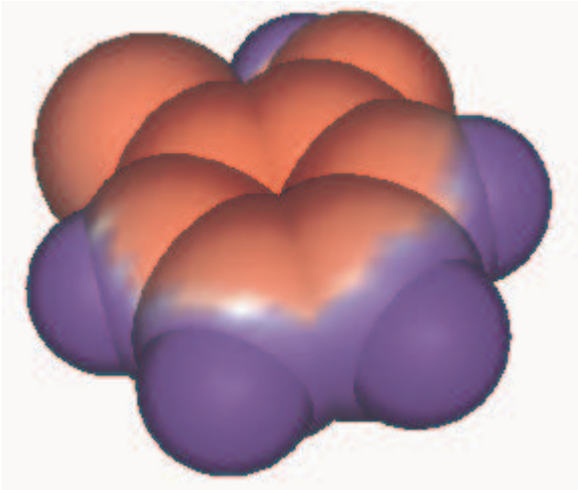


Figure 6: Color map of 2-chlorophenol