QSPR/QSAR: state-of-art, weirdness, the future

Andrey A. Toropov, Alla P. Toropova*

Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri, Via Mario Negri 2, 20156 Milano, Italy

Abstracts
Ability of quantitative structure – property / activity relationships (QSPRs/QSARs) to serve for epistemological processes in natural sciences is discussed. Some weirdness of QSPR/QSAR state-of-art are listed. There are some contradictions in the research results in this area. Sometimes, these should be classified as paradoxes or weirdness. These points often are ignored. Here these are listed and briefly commented. In addition, hypotheses on the future evolution of the QSPR/QSAR theory and practice are suggested.

Keywords: QSAR evolution; Multi-target QSAR; Monte Carlo method; Fuzzy sets

*) Corresponding author
Alla P. Toropova, PhD
Laboratory of Environmental Chemistry and Toxicology
Department of Environmental Health Science (Dipartimento: Ambiente e Salute)
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Via Mario Negri 2, 20156 Milano, Italy
Tel: +39-02-39014595
Fax: +39-02-39014735
e-mail: alla.toropova@marionegri.it
1. Introduction

Everything starts by contradiction or even conflict. Evolution of sciences does not provide evolution for intellectual comfort. Any learning is a hard process. Development of sciences does not improves the learning process from point of view of students at all. Intellectual distance between researchers of different scientific fields gradually becomes insurmountable abyss. However, basis for optimistic interpretation of the above facts also exists. Internet provides comfort of information service. One can get more or less satisfactory consultation on majority of questions related to everyday life and even questions related to education and sciences.

The quantitative structure – property activity relationships (QSPRs/QSARs) relatively new field of natural sciences. There are a large group of aims associated with QSPRs/QSARs technique; main of these are probably the follows (i) prediction of physicochemical behavior of various substances in industry and their further ecologic impacts; (ii) biochemical behavior of various substances in ecological and medicinal aspects; (iii) selection of substances, which can be perspective candidates to defined roles.

Results of traditional experiments were depended to properties of substances, masses, radiation, head capacity, electronic, physicochemical and biochemical conditions as well as many others conditions and circumstances. Computational experiments related to QSPR/QSAR concerned to “information conditions” (available datasets) and “statistical conditions” (diversity of substances in datasets), as well as preference of the user.

Wiener has carried out the pioneer works in the field of correlation “molecular structure – macro-effect of a substance” in 1940s [1-4]. This was the start of QSPR/QSAR history, in other words, this is the first stage of evolution of QSPR/QSAR theory and practice.

The main task of the QSPR/QSAR at this period was to establish correlation between an endpoint and descriptor for a set of substances. Criteria of quality of those models were (i) the total number of compounds in the available set; (ii) correlation coefficient; (iii) standard error of estimation; and (iv) the Fischer F-ratio [1-5]. In this period the family of topological indices [6-17], indices based on the mathematical theory of information [18-28], various 3D descriptors [29-32], and descriptors of quantum mechanics [33-38] were the basis of the QSPR/QSAR theory and practice.

However, absence of the reliable statistical checking up of these models had led to intensive criticism of the QSPR/QSAR research. This criticism is continuing up to now [39-42].

A set of principles were proposed for evaluating the validity of QSAR models at conference held in Setubal, Portugal in 2002. According to the Setubal principles, QSARs should:

(1) Be associated with a defined endpoint of regulatory importance;
(2) Take the form of an unambiguous algorithm;
(3) Have a clear domain of applicability;
(4) Be associated with appropriate measures of goodness of robustness, and predictivity.
(5) Have a mechanistic interpretation.

In further these principles were renamed in OECD principles (Organisation for Economic Co-operation and Development) http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

The OECD principles open the second stage of QSAR history: “not only to establish a correlation, but to checking up predictive potential of the correlation”.

2. QSPR/QSAR: State of art

There are improvement of the QSPR/QSAR technique during last decade. However, some “unpleasant peculiarities” remain still. The list of "main unpleasant peculiarities" of QSPR/QSAR analysis is follows: (i) possibility of "chance correlations" [43-46]; (ii) possibility of overtraining [47]; (iii) possibility of weak reproducibility of statistical quality of an approach suggested [48,49].

A person who would like to apply a model hardly will be pleasant to necessity to get a group of descriptor via hard to understand software and with further necessity to carry out calculation with other hard to understand software that provides the multiple linear regression analysis or the artificial neuron networks or something else. Attempts to solve problems related to the above "unpleasant peculiarities" of QSPR/QSAR are performing. However, these attempts provide three weirdness points.

2.1. The first weirdness of QSPR/QSAR

The distribution of available data for QSPR/QSAR analyses into the training and validation sets can be done by various manner [50, 51]. The distribution have key influence for the statistical quality of QSPR/QSAR models [52, 53]. Here, one can see first weirdness in the modern QSPR/QSAR researches: majority of models are based solely one distribution available data into the training and validation sets.

According many authors, it is a matter of course, the rational split into training and validation set gives better statistical results for the validation sets than models based on random splits [54]. However, the experiment confirms that there are splits successful for one approach, which are unsuccessful for other approach [55-59]. For example, three different splits (Table 1) into training and validation sets of 87 anticancer inhibitors [60] give models with different predictive abilities (Table 2).
Table 1
Distribution of 87 anticancer inhibitors [60] into training and validation sets

<table>
<thead>
<tr>
<th>Split</th>
<th>Training set</th>
<th>Validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>1, 4, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 25, 27, 29, 30, 31, 33, 34, 36, 37, 38, 40, 41, 42, 45, 46, 48, 51, 52, 54, 55, 56, 57, 59, 60, 61, 63, 64, 65, 67, 69, 70, 73, 74, 75, 77, 90, 94, 98, 99, 109, 112, 116, 117, 118, 120, 121, 122, 123, 124, 126, 130, 136;</td>
<td>2, 3, 5, 10, 11, 22, 26, 32, 35, 39, 43, 47, 68, 71, 92, 103, 125, 143,</td>
</tr>
<tr>
<td>#2</td>
<td>1, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 23, 25, 27, 31, 33, 34, 36, 40, 41, 42, 45, 46, 48, 51, 54, 55, 56, 57, 59, 61, 63, 65, 67, 69, 73, 74, 75, 77, 98, 109, 112, 116, 117, 121, 123, 124, 130, 136, 5, 10, 11, 22, 26, 32, 39, 43, 47, 68, 71, 92, 103, 125, 143;</td>
<td>4, 17, 20, 21, 29, 30, 37, 38, 52, 60, 64, 70, 90, 94, 99, 118, 120, 122, 126, 2, 3, 35,</td>
</tr>
<tr>
<td>#3</td>
<td>1, 4, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 21, 23, 25, 27, 29, 31, 36, 37, 38, 40, 41, 42, 45, 48, 51, 54, 55, 56, 57, 59, 60, 64, 65, 69, 70, 73, 74, 75, 77, 94, 98, 99, 109, 116, 118, 121, 124, 130, 136, 2, 5, 10, 11, 26, 32, 35, 39, 43, 47, 68, 71, 125, 143;</td>
<td>20, 30, 33, 34, 46, 52, 55, 61, 63, 67, 90, 112, 117, 120, 122, 123, 126, 3, 22, 92, 103,</td>
</tr>
</tbody>
</table>

Table 2
The predictive potential of different approaches observed for different splits

<table>
<thead>
<tr>
<th>Method</th>
<th>Split</th>
<th>Number of compounds in validation set</th>
<th>Determination coefficient for validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-QSAR [60]</td>
<td>#1</td>
<td>18</td>
<td>0.77</td>
</tr>
<tr>
<td>Method 1</td>
<td>#1</td>
<td>18</td>
<td>0.43</td>
</tr>
<tr>
<td>Method 2</td>
<td>#1</td>
<td>18</td>
<td>0.53</td>
</tr>
<tr>
<td>Method 1</td>
<td>#2</td>
<td>22</td>
<td>0.84</td>
</tr>
<tr>
<td>Method 1</td>
<td>#3</td>
<td>21</td>
<td>0.81</td>
</tr>
<tr>
<td>Method 2</td>
<td>#2</td>
<td>22</td>
<td>0.82</td>
</tr>
<tr>
<td>Method 2</td>
<td>#3</td>
<td>21</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Method 1 is one-variable model calculated with the Monte Carlo technique [61-64] for hybrid optimal descriptors, which are calculated with simplified molecular input-line entry system (SMILES) [65], together with molecular graph [66-71]:

\[
DCW(1,10) = \sum CW(EC0_k) + \sum CW(EC1_k) + \sum CW(S_j) + \sum CW(SS_j)
\]  

(1)

The EC0_k is vertex degree in hydrogen-suppressed graph (HSG); EC1_k is Morgan extended connectivity [72-73] of first order; S_j are SMILES atoms i.e. one symbol (e.g. ‘C’, ‘N’, ‘O’, etc.) or a
group of symbols which cannot be examined separately (e.g. ‘Cl’, ‘Br’, ‘@@’, %12, etc.); the $S_j$ are connected pairs of the SMILES atoms.

Method 2 is one-variable model calculated with the Monte Carlo technique for hybrid optimal descriptors:

$$DCW_{(1,10)} = CW(C5) + CW(C6) + \sum CW(S_j) + \sum CW(SS_j)$$

(2)

The $C5$ and $C6$ are codes of molecular rings extracted from the adjacency matrix of HSG [74]. The $CW(EC0_k)$, $CW(EC1_k)$, $CW(S_j)$, $CW(SS_j)$, $CW(C5)$, and $CW(C6)$ are correlation weights of the above listed SMILES attributes and invariants of HSG calculated with the Monte Carlo method (http://www.insilico.eu/coral). The numerical data on the correlation weights are calculated with the Monte Carlo method.

The described experiment confirms successful and unsuccessful splits exist. Excellent split (Split 1) for 3D-QSAR approach is poor for 2D approaches, i.e. models calculated with Eq. 1 or Eq. 2. However (Table 2), split 2 is excellent (at least successful) for method 1, whereas the split 3 is excellent (at least successful) for method 2.

2.2. The second weirdness of QSPR/QSAR

The number of statistical characteristics aimed to measure the predictive potential of a model gradually increase (Table 3), despite apparent attractiveness of small number of criteria of the predictive potential for practical applications.

Table 3

<table>
<thead>
<tr>
<th>Criterion of the predictive potential</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$</td>
<td>[75]</td>
</tr>
<tr>
<td>$CCC = \frac{2 \sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$</td>
<td>[76]</td>
</tr>
<tr>
<td>$R_0^2 = 1 - \frac{\sum(\hat{y}_i - \hat{y}_i^{opt})^2}{\sum(\hat{y}_i - \bar{y})^2}$</td>
<td>[77]</td>
</tr>
<tr>
<td>$R'_0^2 = 1 - \frac{\sum(y_i - \hat{y}_i^{opt})^2}{\sum(y_i - \bar{y})^2}$</td>
<td></td>
</tr>
</tbody>
</table>
\[ k = \frac{y_i \hat{y}_i}{\bar{y}_i^2} \]
\[ k' = \frac{y_i \hat{y}_i}{\bar{y}_i^2} \]

\[
Q^2 = 1 - \frac{\sum (y_k - \hat{y}_k)^2}{\sum (y_k - \bar{y}_k)^2} \]
\[
Q^2_{F_1} = 1 - \left[ \frac{\sum_{i=1}^{N_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{EXT}} (y_i - \bar{y}_{EXT})^2} \right] / N_{EXT} \]
\[
Q^2_{F_2} = 1 - \left[ \frac{\sum_{i=1}^{N_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{EXT}} (y_i - \bar{y}_{EXT})^2} \right] / N_{EXT} \]
\[
Q^2_{F_3} = 1 - \left[ \frac{\sum_{i=1}^{N_{TR}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{TR}} (y_i - \bar{y}_{TR})^2} \right] / N_{TR} \]

\[
r_m^2 = r^2 \left( 1 - \sqrt{|r^2 - r_0^2|} \right) \]

\[
IIC_{CLB} = r_{CLB} \frac{\min (-\text{MAE}_{CLB}, +\text{MAE}_{CLB})}{\max (-\text{MAE}_{CLB}, +\text{MAE}_{CLB})} \]
\[
-\text{MAE}_{CLB} = \frac{1}{N} \sum_{k=1}^{N} |\Delta_k|, \Delta_k < 0; \text{ ~N is the number of } \Delta_k < 0 \]
\[
+\text{MAE}_{CLB} = \frac{1}{N} \sum_{k=1}^{N} |\Delta_k|, \Delta_k > 0; \text{ ~N is the number of } \Delta_k \geq 0 \]
\[
\Delta_k = \text{observed}_k - \text{calculated}_k \]

On one hand, the diversity of different criteria of predictive potential is a tool to improve quality of QSAR/QSPR models. On the other hand, this situation causes sometimes the uncertainty in choice of the best model.

### 2.3. The third weirdness of QSAR/QSPR

Naturally, the contribution of the molecular structure is key importance to an endpoint. However, any biological activity is a mathematical function of many different conditions and circumstances. In other words, toxicity or pharmaceutical effect are caused by not only molecular structure, but also physicochemical conditions (e.g. temperature, humidity) and circumstances (noise/silence, illumination/darkness). Hardly, somebody disagree with the above postulate, but majority of QSAR/QSPR have built up without taking into account something besides of molecular structure.

It is to be noted, however, in some cases the molecular structure is not informative to build up predictive model of endpoints [81-95].

### 3. Discussion
There are problems. There are solutions. Hierarchy of problems in the field of the modelling of various endpoints is not established. One group of researchers believe that validation of a model is key importance. Other group believes that main result is the statistical quality of a model. Third group concentrates on mechanistic interpretation. It is curiously, but non-standard tasks and solutions also exist and sometimes these are very important. Examples are below.

3.1. Multi-target QSAR models

Limitation of almost all QSAR models is that they predict the biological activity for only one endpoint. In other words, traditional QSAR give model for biological activity of drugs against only one parasite species [96] one species of virus [97] one type cancer [98]. So-called, multi-target QSAR as a tool to build up models for several endpoints is suggested [96-98].

Apparently, this conception has attractive advantages; nonetheless, traditional approaches serve to solve the task of building up multi-target QSARs, e.g. using multiple regression [99], partial least squares (PLS) [100], artificial neural networks (ANN) [101-103], and random forest [104]. It is to be noted, that interest to researches dedicated to multi-target QSAR in drug discovery gradually increases during past decade, whereas interest to general QSAR in drug discovery is approximately constant. Figure 1 confirms this situation.

Figure 1.
Comparison of frequencies of use QSAR and mt-QSAR in drug discovery researches.
3.2. Similarity of endpoints
As noted in previous section, the simultaneous examination of two endpoints is an attractive way in the QSPR/QSAR analysis. In addition to multi-target QSAR, the similarity of endpoints may be a heuristic tool of control of the biochemical knowledge [105-107]. Similarity/dissimilarity of endpoints can be expressed via correlation weights of molecular features extracted from SMILES [105]. In principle, spectrum of physicochemical conditions with clear impact to biochemical endpoints (toxicity, therapeutic potential) able to provide hints to establish similarity (dissimilarity) for two endpoints relevant to drug discovery, toxicity, risk assessment, and other.

3.3. The simplicity or the efficiency: which is better?
QSAR should be assessed as surrogate of real experiment. QSAR aimed to measure an endpoint value. However, to expect adequate prediction physicochemical and biochemical behaviour for arbitrary substance by means of QSPR/QSAR is naively. Despite of the above sad thesis, QSPR/QSAR has become an integral part of modern science as a tool to detect "fuzzy tendencies" in behaviour of groups of substances. This fact logically echoes the theory of fuzzy sets [108]. This is not surprisingly; fuzzy set theory has success for solving of some problems of QSPR/QSAR analysis [109-111].

One can extract two components in the total big variety of QSAR studies: (i) "extensive" studies; and (ii) "intensive" studies. The aim of "extensive" studies is integration of results of applying of current approaches to solve practical tasks. The aim of "intensive" studies is attempts to develop new conceptions of the QSPR/QSAR analysis. Naturally, small part of results of the "intensive" studies gradually becomes a tool of robust "extensive" studies. Nowadays, multi-target QSAR is a part of “intensive” studies [96-104]. Development of criteria of predictive potential of models (Table 3) also is a part of the “intensive” studies. Maybe, search for similarity of endpoints [105-107], also, will become part of "intensive" QSPR/QSAR researches.

4. Conclusions
Evolution of the field of QSPR/QSAR has two components: intensive and extensive. The intensive component responsible to development quality and epistemology potential of various QSPR/QSAR approaches. The multi-target QSAR is perspective field of evolution of the QSAR theory and practices. Other perspective components of the "intensive" evolution of the QSPR/QSAR are (i) applying of fuzzy sets theory; and (ii) development of statistical methods to detect similarity of biochemical endpoints.
Acknowledgement
The authors are grateful for the contribution of the project LIFE-VERMEER contract (LIFE16 ENV/ES/000167) for financial support.

Disclosure statement
No potential conflict of interest was reported by the authors.
References


53. Ghaemian, P., Shayanfar, A. Quantitative structure activity relationship (QSAR) of methylated polyphenol derivatives as permeability glycoprotein (P-gp) inhibitors: A


