Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

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, hypothesises 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

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

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;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

(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 QSPR/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).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

Table 1

Distribution of 87 anticancer inhibitors [60] into training and validation sets

Split #1	Training set = 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; Validation set = 2, 3, 5, 10, 11, 22, 26, 32, 35, 39, 43, 47, 68, 71, 92, 103, 125, 143,
Split #2	Training set = 1, 6, 7, 8, 9, 12, 13, 14, 15, 16, 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; Validation set = 4, 17, 20, 21, 29, 30, 37, 38, 52, 60, 64, 70, 90, 94, 99, 118, 120, 122, 126, 2, 3, 35,
Split #3	Training set = 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, 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; Validation set = 20, 30, 33, 34, 46, 52, 55, 61, 63, 67, 90, 112, 117, 120, 122, 123, 126, 3, 22, 92, 103,

Table 2

The predictive potential of different approaches observed for different splits

Method	**Split**	**Number of compounds in validation set**	**Determination coefficient for validation set**
3D-QSAR [60]	#1	18	0.77
Method 1	#1	18	0.43
Method 2	#1	18	0.53
Method 1	#2	22	0.84
Method 1	#3	21	0.81
Method 2	#2	22	0.82
Method 2	#3	21	0.85

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]:

$$DCW(1,10) = \sum CW(EC0_k) + \sum CW(EC1_k) + \sum CW(S_j) + \sum CW(SS_j) \qquad (1)$$

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

group of symbols which cannot be examined separately (e.g. 'Cl', 'Br', '@@', %12, etc.); the SS_j are connected pairs of the SMILES atoms.

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

$$DCW(1,10) = CW(C5) + CW(C6) + \sum CW(S_j) + \sum CW(SS_j) \quad (2)$$

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

Statistical criteria of the predictive potential for QSPR/QSAR models

Criterion of the predictive potential	**Reference**
$R = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2}}$	[75]
$CCC = \frac{2\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2 + \sum (y-\bar{y})^2 + n(\bar{x}-\bar{y})^2}$	[76]
$R_0^2 = 1 - \frac{\sum(\tilde{y}_i - y_i^{ro})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2}$ $R'^2_0 = 1 - \frac{\sum(y_i - \tilde{y}_i^{ro})^2}{\sum(y_i - \bar{y}_i)^2}$	[77]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

$k = \frac{y_i \tilde{y}_i}{\tilde{y}_i^2}$ $k' = \frac{y_i \tilde{y}_i}{y_i^2}$	
$Q^2 = 1 - \frac{\sum(y_k - \acute{y}_k)^2}{\sum(y_k - \bar{y}_k)^2}$ $Q^2_{F1} = 1 - \frac{\left[\sum_{i=1}^{N_{EXT}}(\acute{y}_i - y_i)^2\right]/N_{EXT}}{\left[\sum_{i=1}^{N_{EXT}}(y_i - \overline{y}_{TR})^2\right]/N_{EXT}}$ $Q^2_{F2} = 1 - \frac{\left[\sum_{i=1}^{N_{EXT}}(\acute{y}_i - y_i)^2\right]/N_{EXT}}{\left[\sum_{i=1}^{N_{EXT}}(y_i - \overline{y}_{EXT})^2\right]/N_{EXT}}$ $Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{N_{EXT}}(\acute{y}_i - y_i)^2\right]/N_{EXT}}{\left[\sum_{i=1}^{N_{TR}}(y_i - \overline{y}_{TR})^2\right]/N_{TR}}$	[78]
$r_m^2 = r^2\left(1 - \sqrt{\lvert r^2 - r_0^2 \rvert}\right)$	[79]
$IIC_{CLB} = r_{CLB}\frac{\min\left({}^{-}MAE_{CLB}, {}^{+}MAE_{CLB}\right)}{max\left({}^{-}MAE_{CLB}, {}^{+}MAE_{CLB}\right)}$ ${}^{-}MAE_{CLB} = \frac{1}{{}^{-}N}\sum_{k=1}^{{}^{-}N}\lvert\Delta_k\rvert, \Delta_k 0; {}^{-}N \text{ is the number of } \Delta_k < 0$ ${}^{+}MAE_{CLB} = \frac{1}{{}^{+}N}\sum_{k=1}^{{}^{+}N}\lvert\Delta_k\rvert, \Delta_k 0; {}^{+}N \text{ is the number of } \Delta_k \geq 0$ $\Delta_k = observed_k - calculated_k$	[80]

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

2.3. The third weirdness of QSPR/QSAR

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 QSPR/QSAR 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

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.

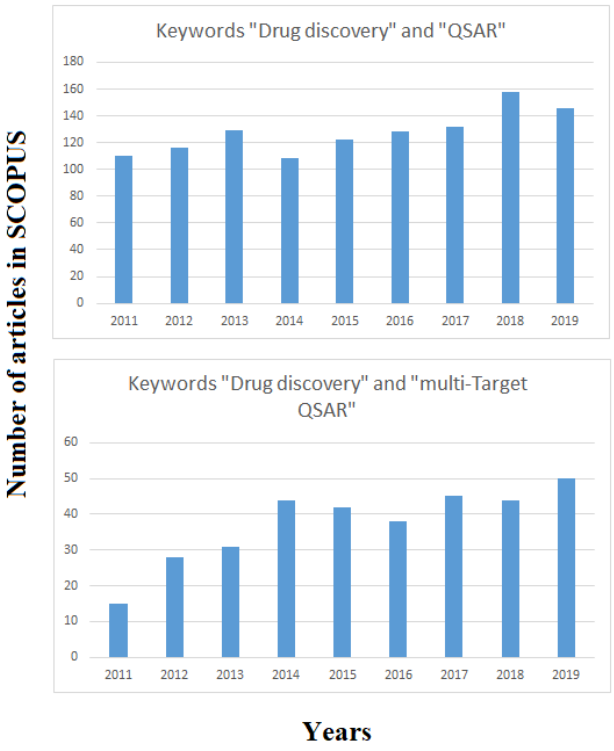

Figure 1.

Comparison of frequencies of use QSAR and mt-QSAR in drug discovery researches.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

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 vie 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 assessment 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

References

1. Wiener, H. Structural Determination of Paraffin Boiling Point. *J. Am. Chem. Soc*. **1947**, *69*, 17-20.
2. Wiener, H. Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons. *J. Am. Chem. Soc*. **1947**, *69(11)*, 2636-2638.
3. Wiener, H. Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure. Surface Tension, Specific Dispersion, and Critical Solution Temperature in Aniline. *J. Phys. Chem*. **1948**, *52(6)*, 1082-1089.
4. Wiener, H. Vapor Pressure-Temperature Relationships Among the Branched Paraffin Hydrocarbons. *J. Phys. Chem*. **1948**, *52(2)*, 425-430.
5. Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn*. **1971**, *44*, 2332-2339.
6. Katritzky, A. R.; Gordeeva, E. V. Traditional Topological Indexes vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 835.
7. Balaban, T.; Mills, D.; Ivanciuc, O.; Basak, S. C. Reverse Wiener Indices. *Croat. Chem. Acta*. **2000**, *73*, 923.
8. Ivanciuc, O.; Ivanciuc, T.; Klein, D. J.; Seitz, W. A.; Balaban, A. T. Wiener Index Extension by Counting Even/Odd Graph Distances. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 536-549.
9. Torrens, F. Valence Topological Charge-Transfer Indices for Dipole Moments. *Molecules* **2003**, *8*, 169-185.
10. Rios Santamarina, I.; Garc a Domenech, R. Cortijo, J.; Santamaria, P.; Morcillo, E. J.; GÆlvez, J. Natural Compounds with Bronchodilator Activity Selected by Molecular Topology. *Internet Electron. J. Mol. Des*. **2002**, *1*, 70-79. http://www.biochempress.com.
11. Balaban, A. Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR, Similarity/Dissimilarity and Drug Design. *SAR QSAR Environ. Res*. **1998**, *8*, 1-21.
12. Estrada, E. On the Topological Sub-Structural Molecular Design (TOSS-MODE) in QSPR/QSAR and Drug Design Research. *SAR QSAR Environ. Res*. **2000**, *11*, 55-73.
13. Randić, M. in Encyclopedia of Computational Chemistry, Schleyer, P. V. R. (Ed.) John Wiley & Sons: New York, **1998**, *5*, 3018-3032.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

14. Randić, M. Generalized Molecular Descriptors. J. Math. Chem. 1991, 7, 155-168.
15. Estrada, E., Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem*. **2001**, *8*, 1699-1714.
16. Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 891-898.
17. Patel, H.; Cronin, M. T. D. A Novel Index for the Description of Molecular Linearity. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1228-1236.
18. Ivanciuc, O.; Ivanciuc, T.; Cabrol Bass, D.; Balaban, A. T. Evaluation in Quantitative StructureProperty Relationship Models of Structural Descriptors Derived from Information Theory Operators. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 631-643.
19. Randić. M. Characterization of Molecular Branching. *J. Am. Chem. Soc*. **1975**, *69*, 6609-6615.
20. Plavšić, D.; Nikolić, S.; Trinajstić, N.; Mihalić, Z. On the Harary Index for the Characterization of Chemical Graphs. *J. Math. Chem*. **1993**, *12*, 235-250.
21. Estrada, E. Spectral Moment of Edge Adjacency Matrix in Molecular Graphs.1. Definition and Application to the Prediction of Physical Properties of Alkanes. *J. Chem. Inf. Comp. Sci*. **1996**, *36*, 846-849.
22. Estrada, E. Spectral Moment of Edge Adjacency Matrix in Molecular Graphs. 2. Molecules Containing Heteroatom and QSAR Applications. *J. Chem. Inf. Comp. Sci*. **1997**, *37*, 320-328.
23. Estrada, E. Spectral Moment of Edge Adjacency Matrix in Molecular Graphs 3. Molecules Containing Cycles. *J. Chem. Inf. Comp. Sci*. **1998**, *38*, 123-127.
24. Mihalic, Z.; Trinajstić, N. A Graph-Theoretical Approach to Structure-Property Relationships. *J. Chem. Educ*. **1992**, *69*, 701-712.
25. Diudea, M. V. (Eds.) QSPR/QSAR Studies by Molecular Descriptors; Nova Science, Huntington: New York, **2001**.
26. Marino, D. J. G.; Peruzzo, P. J.; Castro, E. A.; Toropov, A. A. QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants. *Internet Electron. J. Mol. Des*. **2002**, *1*, 115-133, http://www.biochempress.com.
27. Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1412-1422.
28. Broach, J. R.; Thorner, J. High-Throughput Screening for Drug Discovery. *Nature* **1996**, *384*, 14-16.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

29. Raichurkar, A.V.; Shah, U.A.; Kulkarni, V.M. 3D-QSAR of novel Phosphodiesterase-4 inhibitors by genetic function approximation. *Med. Chem*. **2011**, *7(6)*, 543-552.
30. Gaurav, A.; Yadav, M.R.; Giridhar, R.; Gautam, V.; Singh, R. 3D-QSAR studies of 4-quinolone derivatives as high-affinity ligands at the benzodiazepine site of brain GABAA receptors. *Med. Chem. Res*. **2011**, *20(2)*, 192-199.
31. Sinha, N.; Sen, S. Predicting hERG activities of compounds from their 3D structures: Development and evaluation of a global descriptors based QSAR model. *Eur. J. Med. Chem*. **2011**, *46(2)*, 618-630.
32. Singh, H.P.; Chaturvedi, A.P.; Sharma, C.S. 3D-QSAR and insilico study: Modeling parameters for designing new selective 12-LO enzymes inhibitors. *Int. J. Pharm. Tech. Res*. **2011**, *3(1)*, 231-236.
33. Clare, B.W., Supuran, C.T. Carbonic anhydrase inhibitors. Part 61. Quantum chemical QSAR of a group of benzenedisulfonamides. *Eur. J. Med. Chem*. **1999**, *34(6)*, 463-474.
34. Hameed, A.J., Ibrahim, M., ElHaes, H. Computational notes on structural, electronic and QSAR properties of [C60]fulleropyrrolidine-1-carbodithioic acid 2; 3 and 4-substituted-benzyl esters. *J. Mol. Struct. THEOCHEM* **2007**, *809(1-3)*, 131-136.
35. Tekiner-Gulbas, B., Temiz-Arpaci, O., Oksuzoglu, E., Eroglu, H., Yildiz, I., Diril, N., Aki-Sener, E., Yalcin, I. QSAR of genotoxic active benzazoles. *SAR QSAR Environ. Res*. **2007**, *18(3-4)*, 251-263.
36. Panda, P., Samanta, S., Alam, Sk.M., Basu, S., Jha, T. QSAR for analogs of 1,5-N,N′-disubstituted-2-(substituted benzenesulphonyl) glutamamides as antitumor agents. *Internet Elect. J. Mol. Des*. **2007**, *6(9)*, 280-301.
37. Eroglu, E., Türkmen, H. A DFT-based quantum theoretic QSAR study of aromatic and heterocyclic sulfonamides as carbonic anhydrase inhibitors against isozyme, CA-II. *J. Mol. Graph. Model*. **2007**, *26(4)*, 701-708.
38. Pasha, F.A., Cho, S.J., Beg, Y., Tripathi, Y.B. Quantum chemical QSAR study of flavones and their radical-scavenging activity. *Med. Chem. Res*. **2007**, *16(7-9)*, 408-417.
39. Doweyko, A.M. Is QSAR relevant to drug discovery? *IDrugs* **2008**, *11(12)*, 894-899.
40. Doweyko, A.M. QSAR: Dead or alive? *J. Comput. Aid. Mol. Des*. **2008**, *22(2)*, 81-89.
41. Maggiora, G.M. On outliers and activity cliffs - Why QSAR often disappoints? *J. Chem. Inf. Model*. **2006**, *46(4)*, 1535.
42. Johnson, S.R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model*. **2008**, *48(1)*, 25-26.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

43. Topliss, J.G., Costello, R.J. Chance Correlations in Structure-Activity Studies Using ultiple Regression Analysis. *J. Med. Chem*. **1972**, *15(10)*, 1066-1068.

44. Clark, M., Cramer, R.D. III. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Rel*. **1993**, *12(2)*, 137-145.

45. Baumann, K. Chance correlation in variable subset regression: Influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR Comb. Sci*. **2005**, *24(9)*, 1033-1046.

46. Toropov, A.A., Toropova, A.P. QSAR as a random event: criteria of predictive potential for a chance model. *Struct. Chem*. **2019**, *30(5)*, 1677-1683.

47. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model*. **2008**, *48(9)*, 1733-1746.

48. Roy, P.P.; Kovarich, S.; Gramatica, P. QSAR model reproducibility and applicability: A case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles. *J. Comput. Chem*. **2011**, *32(11)*, 2386-2396.

49. Patel, M.; Chilton, M.L.; Sartini, A.; Gibson, L.; Barber, C.; Covey-Crump, L.; Przybylak, K.R.; Cronin, M.T.D.; Madden, J.C. Assessment and Reproducibility of Quantitative Structure-Activity Relationship Models by the Nonexpert. *J. Chem. Inf. Model*. **2018**, *58(3)*, 673-682.

50. Roy, P.P., Leonard, J.T., Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. *Chemometr. Intell. Lab. Syst*. **2008**, *90(1)*, 31-42.

51. Roy, K., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., Kabir, H. Introduction of rm2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. *Chemometr. Intell. Lab. Syst*. **2012**, *118*, 200-210.

52. Masand, V.H., Mahajan, D.T., Nazeruddin, G.M., Hadda, T.B., Rastija, V., Alfeefy, A.M. Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model. *Med. Chem. Res*. **2015**, *24(3)*, 1241-1264.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

comparison of different training and test set selection methods. *Lett. Drug Des. Discov*. **2017**, *14(9)*, 999-1007.

54. Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H., Tropsha, A. Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model*. **2012**, *52(10)*, 2570-2578.
55. Toropov, A.A.; Toropova, A.P.; Puzyn, T.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere* **2013**, *92(1)*, 31-37.
56. Veselinović, J.B.; Nikolić, G.M.; Trutić, N.V.; Živković, J.V.; Veselinović, A.M. Monte Carlo QSAR models for predicting organophosphate inhibition of acetylcholinesterase. *SAR QSAR Environ. Res*. **2015**, *26(6)*, 449-460.
57. Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. QSAR model as a random event: A case of rat toxicity. *Bioorg. Med. Chem*. **2015**, *23(6)*, 1223-1230.
58. Toropova, A.P.; Toropov, A.A.; Veselinović, J.B.; Veselinović, A.M. QSAR as a random event: a case of NOAEL. *Environ. Sci. Pollut. Res*. **2015**, *22(11)*, 8264-8271.
59. Toropova, M.A.; Raska, I.; Jr.; Toropova, A.P.; Raskova, M. CORAL software: Analysis of impacts of pharmaceutical agents upon metabolism via the optimal descriptors. *Curr. Drug Metab*. **2017**, *18(6)*, 500-510.
60. Alam, S.; Khan, F. 3D-QSAR, Docking, ADME/Tox studies on Flavone analogs reveal anticancer activity through Tankyrase inhibition. *Sci. Rep*. **2019**, *9(1)*, 5414.
61. Jain, S.; Amin, S.A.; Adhikari, N.; Jha, T.; Gayen, S. Good and bad molecular fingerprints for human rhinovirus 3C protease inhibition: identification, validation, and application in designing of new inhibitors through Monte Carlo-based QSAR study. *J. Biomol. Struct. Dyn*. **2020**, *38(1)*, 66-77.
62. Kumar, P.; Kumar, A.; Sindhu, J. In silico design of diacylglycerol acyltransferase-1 (DGAT1) inhibitors based on SMILES descriptors using Monte-Carlo method. *SAR QSAR Environ. Res*. **2019**, *30(8)*, 525-541.
63. Ahmadi, S.; Mardinia, F.; Azimi, N.; Qomi, M.; Balali, E. Prediction of chalcone derivative cytotoxicity activity against MCF-7 human breast cancer cell by Monte Carlo method. *J. Mol. Struct*. **2019**, *1181*, 305-311.
64. Bhargava, S.; Patel, T.; Gaikwad, R.; Patil, U.K.; Gayen, S. Identification of structural requirements and prediction of inhibitory activity of natural flavonoids against Zika virus

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

through molecular docking and Monte Carlo based QSAR Simulation. *Nat. Prod. Res*. **2019**, *33(6)*, 851-857.

65. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28(1)*, 31-36.
66. Kumar, A.; Chauhan, S. Monte Carlo method based QSAR modelling of natural lipase inhibitors using hybrid optimal descriptors. *SAR QSAR Environ. Res*. **2017**, *28(3)*, 179-197.
67. Bhargava, S.; Adhikari, N.; Amin, S.A.; Das, K.; Gayen, S.; Jha, T. Hydroxyethylamine derivatives as HIV-1 protease inhibitors: a predictive QSAR modelling study based on Monte Carlo optimization. *SAR QSAR Environ. Res*. **2017**, *28(12)*, 973-990.
68. Kumar, P.; Kumar, A. Monte Carlo Method Based QSAR Studies of Mer Kinase Inhibitors in Compliance with OECD Principles. *Drug Res*. **2018**, *68(4)*, 189-195.
69. Gaikwad, R.; Ghorai, S.; Amin, S.A.; Adhikari, N.; Patel, T.; Das, K.; Jha, T.; Gayen, S. Monte Carlo based modelling approach for designing and predicting cytotoxicity of 2-phenylindole derivatives against breast cancer cell line MCF7. *Toxicol. in Vitro* **2018**, *52*, 23-32.
70. Kumar, P.; Kumar, A.; Sindhu, J.; Lal, S. QSAR Models for Nitrogen Containing Monophosphonate and Bisphosphonate Derivatives as Human Farnesyl Pyrophosphate Synthase Inhibitors Based on Monte Carlo Method. *Drug Res*. **2019**, *69(3)*, 159-167.
71. Manisha; Chauhan, S.; Kumar, P.; Kumar, A. Development of prediction model for fructose- 1,6- bisphosphatase inhibitors using the Monte Carlo method. *SAR QSAR Environ. Res*. **2019**, *30(3)*, 145-159.
72. Toropov, A.A., Toropova, A.P., Ismailov, T.T., Voropaeva, N.L., Ruban, I.N. Extended molecular connectivity: Prediction of boiling points of alkanes. *J. Struct. Chem.* **1997**, *38(6)*, 965-969.
73. Toropov, A.A., Toropova, A.P. QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struct.: THEOCHEM*, **2002**, *578(1-3)*, 129-134.
74. Toropov, A.A., Toropova, A.P., Marzo, M., Benfenati, E. Use of the index of ideality of correlation to improve aquatic solubility model. *J. Mol. Graph. Model.* **2020,** *96*, 107525.
75. Pearson K. Notes on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond*. **1895**, *58*, 240–242.
76. Lin L.I.-K. Assay validation using the concordance correlation coefficient. *Biometrics* **1992**, *48(2)*, 599-604.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

77. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20(4)*, 269-276.

78. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation (2009) *J. Chem. Inf. Model.* **2009**, *49(7)*, 1669-1678.

79. Mitra, I.; Roy, P.P.; Kar, S.; Ojha, P.K.; Roy, K. On further application of r_m^2 as a metric for validation of QSAR models. *J. Chemometr.* **2010**, *24(1)*, 22-33.

80. Toropov A.A.; Toropova A.P. The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models? *Muta.t Res. Genet. Toxicol. Environ. Mutagen.* **2017,** *819*, 31-37

81. Toropov, A.A.; Toropova, A.P. Quasi-SMILES and nano-QFAR: United model for mutagenicity of fullerene and MWCNT under different conditions. *Chemosphere* **2015**, *139*, 18-22.

82. Toropova, A.P.; Toropov, A.A.; Benfenati, E. A quasi-QSPR modelling for the photocatalytic decolourization rate constants and cellular viability (CV%) of nanoparticles by CORAL. *SAR QSAR Environ. Res.* **2015**, *26(1)*, 29-40.

83. Toropova, A.P.; Toropov, A.A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicol. Environ. Saf.* **2015**, *112*, 39-45.

84. Toropova, A.P.; Toropov, A.A. Mutagenicity: QSAR -quasi-QSAR -nano-QSAR. *Mini-Rev. Med. Chem.* **2015,** *15(8)*, 608-621.

85. Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Korenstein, R.; Leszczynska, D.; Leszczynski, J. Optimal nano-descriptors as translators of eclectic data into prediction of the cell membrane damage by means of nano metal-oxides. *Environ. Sci. Pollut. Res.* **2015**, *22(1)*, 745-757.

86. Toropov, A.A.; Achary, P.G.R.; Toropova, A.P. Quasi-SMILES and nano-QFPR: The predictive model for zeta potentials of metal oxide nanoparticles. *Chem. Phys. Lett.* **2016**, *660*, 107-110.

87. Toropov, A.A.; Toropova, A.P.; Begum, S.; Achary, P.G.R. Towards predicting the solubility of CO2 and N2 in different polymers using a quasi-SMILES based QSPR approach. *SAR QSAR Environ. Res.* **2016**, *27(4)*, 293-301.

88. Toropova, A.P.; Toropov, A.A.; Manganelli, S.; Leone, C.; Baderna, D.; Benfenati, E.; Fanelli, R. Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials. *NanoImpact* **2016**, *1*, 60-64.

89. Toropova, A.P.; Toropov, A.A.; Veselinović, A.M.; Veselinović, J.B.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. Nano-QSAR: Model of mutagenicity of fullerene as a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

mathematical function of different conditions. *Ecotoxicol. Environ. Saf.* **2016**, *124*, 32-36.

90. Toropova, A.P.; Toropov, A.A. Nano-QSAR in cell biology: Model of cell viability as a mathematical function of available eclectic data. *J. Theor. Biol.* **2017**, *416*, 113-118.
91. Toropova, A.P.; Toropov, A.A.; Leszczynska, D.; Leszczynski, J. CORAL and Nano-QFAR: Quantitative feature – Activity relationships (QFAR) for bioavailability of nanoparticles (ZnO, CuO, Co_3O_4, and TiO_2). *Ecotoxicol. Environ. Saf.* **2017**, *139*, 404-407.
92. Trinh, T.X.; Choi, J.-S.; Jeon, H.; Byun, H.-G.; Yoon, T.-H.; Kim, J. Quasi-SMILES-Based Nano-Quantitative Structure-Activity Relationship Model to Predict the Cytotoxicity of Multiwalled Carbon Nanotubes to Human Lung Cells. *Chem. Res. Toxicol.* **2018,** *31(3)*, 183-190.
93. Toropov, A.A.; Sizochenko, N.; Toropova, A.P.; Leszczynski, J. Towards the development of global nano-quantitative structure–property relationship models: Zeta potentials of metal oxide nanoparticles. *Nanomaterials* **2018**, *8(4)*, 243.
94. Choi, J.-S.; Trinh, T.X.; Yoon, T.-H.; Kim, J.; Byun, H.-G. Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. *Chemosphere*, **2019**, *217*, 243-249.
95. Ahmadi, S. Mathematical modeling of cytotoxicity of metal oxide nanoparticles using the index of ideality correlation criteria. *Chemosphere* **2020**, *242*, 125192.
96. Prado-Prado, F.J.; de la Vega, O.M.; Uriarte, E.; Ubeira, F.M.; Chou, K.-C.; González-Díaz, H. Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg. Med. Chem.* **2009**, *17(2)*, 569-575.
97. Prado-Prado, F.J.; García-Mera, X.; González-Díaz, H. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. *Bioorg. Med. Chem.* **2010**, *18(6)*, 2225-2231
98. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N.D.S. Rational drug design for anti-cancer chemotherapy: Multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg. Med. Chem.* **2012**, *20(15)*, 4848-4855.
99. Liu, Q.; Che, D.; Huang, Q.; Cao, Z.; Zhu, R. Multi-target QSAR study in the analysis and design of HIV-1 inhibitors. *Chin. J. Chem.* **2010**, *28(9)*, 1587-1592.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2020 doi:10.20944/preprints202001.0325.v1

100. Nikolic, K., Filipic, S., Agbaba, D. Multi-target QSAR and docking study of steroids binding to corticosteroid-binding globulin and sex hormone-binding globulin. *Curr. Comput. Aided. Drug Des.* **2012,** *8(4)*, 296-306.

101. Garcia-Domenech, R., Zanni, R., Galvez-Llompart, M., Galvez, J. Predicting antiprotozoal activity of benzyl phenyl ether diamine derivatives through QSAR multi-target and molecular topology. *Mol. Divers.* **2015**, *19 (2)*, 357-366.

102. Abdolmaleki, A., Ghasemi, J.B., Ghasemi, F. Computer aided drug design for multi-target drug design: SAR /QSAR, molecular docking and pharmacophore methods. *Curr. Drug Targets*, **2017**, 18 (5), pp. 556-575.

103. Speck-Planche, A., Scotti, M.T. BET bromodomain inhibitors: fragment-based in silico design using multi-target QSAR models. *Mol. Divers.* **2019,** *23(3)*, 555-572.

104. Halder, A.K., Cordeiro, M.N.D.S. Development of multi-target chemometric models for the inhibition of class I PI3K enzyme isoforms: A case study using QSAR-Co tool. *Int. J. Mol. Sci.* **2019**, *20(17)*, 4191.

105. Toropov, A.A., Toropova, A.P., Benfenati, E., Salmona, M. Mutagenicity, anticancer activity and blood brain barrier: similarity and dissimilarity of molecular alerts. *Toxicol. Mech. Methods*, **2018**, *28(5)*, 321-327.

106. Toropova, A.P., Toropov, A.A., Begum, S., Achary, P.G.R. Blood brain barrier and Alzheimer's disease: Similarity and dissimilarity of molecular alerts. *Curr. Neuropharmacol.* **2018**, *16(6)*, 769-785.

107. Toropova, A.P., Toropov, A.A., Veselinović, A.M., Veselinović, J.B., Leszczynska, D., Leszczynski, J. Quasi-SMILES as a Novel Tool for Prediction of Nanomaterials' Endpoints. In book: Speck-Planche, A. (Eds.) Multi-Scale Approaches in Drug Discovery: From Empirical Knowledge to In silico Experiments and Back. **2017**, 191-221.

108. Zadeh, L.A. Fuzzy sets. *Inform. Control* **1965**, *8(3)*, 338-353.

109. Kumar, M., Thurow, K., Stoll, N., Stoll, R. Robust fuzzy mappings for QSAR studies *Eur. J. Med. Chem.* **2007,** *42(5)*, 675-685.

110. Pérez-Garrido, A., Girón-Rodríguez, F., Bueno-Crespo, A., Soto, J., Pérez-Sánchez, H., Helguera, A.M. Fuzzy clustering as rational partition method for QSAR. *Chemometr. Intell. Lab. Syst.* **2017,** *166*, 1-6.

111. Abdolmaleki, A., Ghasemi, J.B. Inhibition activity prediction for a dataset of candidates' drug by combining fuzzy logic with MLR/ANN QSAR models. *Chem. Biol. Drug Des.* **2019**, *93(6)*, 1139-1157.