Molecular Design and QSARs/QSPRs with Molecular Descriptors Family

Sorana D. BOLBOACĂ[1], Lorentz JÄNTSCHI[2*] and Mircea V. DIUDEA[3]

[1] "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, 400023 Cluj, Romania. E-mail: sbolboaca@umfcluj.ro

[2] University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Mănăştur, 400372 Cluj, Romania. E-mail: lori@academicdirect.org

[3] Babeş-Bolyai University, Cluj-Napoca, 11 Arany Janos, 400028 Cluj, Romania. E-mail: diudea@chem.ubbcluj.ro

* Corresponding author: Phone: +4-0264-401775; Fax: +4-0264-401768.

Abstract:

The aim of the present paper is to present the methodology of the molecular descriptors family (MDF) as an integrative tool in molecular modeling and its abilities as a multivariate QSAR/QSPR modeling tool. An algorithm for extracting useful information from the topological and geometrical representation of chemical compounds was developed and integrated to calculate MDF members. The MDF methodology was implemented and software is available online at the AcademicDirect gateway – Chemistry – SARs – MDS_SARs (http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/). This integrative tool was developed in order to maximize performance, functionality, efficiency and portability. The MDF methodology is able to provide reliable and valid multiple linear regression models. Furthermore, in many cases, the MDF models were better than the results in the literature in terms of correlation coefficients (statistically significant Steiger's Z test at a significance level of 5%) and/or in terms of values of information criteria and Kubinyi function. The MDF methodology developed and implemented as a platform for investigating and characterizing quantitative relationships between the chemical structure and the activity/property of active compounds was used on more than 50 study cases. In almost all cases, the methodology allowed obtaining of QSAR/QSPR models improved in explanatory power of structure - activity / property relationship. The algorithms applied in the computation of geometric and topological descriptors (useful in modeling physical-chemical or biological properties of molecules) and those used in searching for reliable and valid multiple linear regression models certain enrich the pool of low-cost low-time drug design tools.

Keywords:

Information criteria, model generation, model validation, molecular descriptors family (MDF), quantitative structure-activity relationship (QSAR), quantitative structure-property relationship (QSPR)

Introduction

Crum-Brown and Fraser conjectured, in 1868, that the biological activity of a compound is a function of its structure and chemical composition [1]. A few years later, Richet showed that the cytotoxicity of a set of organic compounds was inversely related to their water solubility [2]. Since then, many researchers have investigated the link, between the chemical structure and its activity or property [3-6].

Hammett introduced free energy relationship (a type of QSAR/QSPR for chemical kinetics) in the 19th century [7]. The concept of quantitative structure-activity/property relationships (QSAR/QSPR), a mathematical tool able to quantitatively describe the link between chemical structure and biological activity/property for a given set of compounds was introduced in 1937 [8].

Therefore, since the 1960's, the QSAR/QSPR paradigm found its usefulness in agro-chemistry, pharmaceutical chemistry, toxicology and other related research fields [9]. To date, the scientific literature contains many research reports on various methodologies of QSARS/QSPRS such as use of NMR chemical shifts [10], reliability and uncertainty assessment of regressors [11], integration of high-throughput screening [12], design via alignment of molecules [13]. In this respect, two monographs [14, 15] comprehend and can be use as guidelines for details of different approaches.

Quantitative structure-activity/property relationships, mathematical approaches able to identify and characterize the link between chemical structure and activity/property [9], are applied when the activity/property is a quantitative (linear models) or qualitative (non-linear models [16]) variable. The structural information is collected by various molecular descriptors (2D descriptors [17-22]).

The following methods were introduced in order to include structural 3D information in QSARs/QSPRs: CoMFA - comparative molecular field analysis and its variants CoMSIA (MSIA - molecular similarity indices Analysis) [23]; ▪ WHIM - weighted holistic invariant molecular (and its variant MS-WHIM - molecular surface WHIM) [24];▪ MTD - minimal topological distance (and its variant MSD, S - steric) [25]; ▪ FPIF - fragmental property index family [26]; ▪ MDF - molecular descriptors family [27]; ▪ MDFV - molecular descriptors family on vertices [28-30]; ▪ TOPS-MODE - topological sub-structural molecular design [31-35], ▪ other approaches [36-39].

The selection of descriptors [40] is as important in QSAR/QSPR analysis as the statistical method (regression method [41, 42], factor analysis [43], discriminant analysis [44], principal component analysis [45], cluster analysis [46] applied using genetic algorithms (GAs) [47] and/or neural network [48]) applied to identify the structure-activity/property relationship. Furthermore, internal and external validation methods are used to characterize QSAR/QSPR models. Cross-validation [49], randomization [50] and the assessment of QSAR/QSPR equations [51-53] represent the most frequently used methods.

The MDF approach has already been applied in more than fifty activities/properties on various classes of active compounds. The aim of the present review is to present the methodology of the molecular descriptors family (MDF) as an integrative tool in molecular modeling and its abilities as a multivariate QSAR/QSPR modeling tool.

Material and Method

Methodology of the molecular descriptors family

The MDF methodology is an original approach used to translate the complex topological and geometrical information obtained from the structure of chemical compounds into the so-called “molecular descriptors family” [27].

The input data are the topological and geometrical information provided by the HyperChem (v. 8.0/2007) software. The molecular structures were optimized by two molecular mechanics procedures (AMBER - assisted model building with energy refinement [54] and Polak-Ribiere optimization algorithm [55]). The energy calculations were performed at the semi-empirical AM1 [56] level of theory.

The MDF methodology [27, 57] works with molecular fragments and the input/output of the descriptors characteristics below:

- Two distance operators (D_M): geometric distance (*g*), topological distance (*t*). The distance operator implements both the topological and the geometric feature of the molecule.
- Seven atomic properties (A_P): relative atomic mass (*M*), atomic partial charge, Extended Hückel energy (*Q*), cardinality (*C*), atomic electronegativity (*E*), group electronegativity (*G*), number of hydrogen atoms adjacent to the investigated atom (*H*).
- Twenty-four interaction descriptors (P_D): $D=d$, $d=1/d$, $O=p_1$, $o=1/p_1$, $P=p_1p_2$, $p=1/p_1p_2$, $Q=\sqrt{}p_1p_2$, $q=1/\sqrt{(p_1p_2)}$, $J=p_1d$, $j=1/p_1d$, $K=p_1p_2d$, $k=1/p_1p_2d$, $L=d\sqrt{(p_1p_2)}$, $l=1/d\sqrt{(p_1p_2)}$, $V=p_1/d$, $E=p_1/d_2$, $W=p_1^2/d$, $w=p_1p_2/d$, $F=p_1^2/d^2$, $f=p_1p_2/d^2$, $S=p_1^2/d^3$, $s=p_1p_2/d^3$, $T=p_1^2/d^4$, $t=p_1p_2/d^4$; where d = distance operator and p = atomic property. This implements a series of descriptors of some physical entities (e.g., force, field, energy, potential), as they occur in magnetism, electrostatics, gravity or quantum mechanics.
- Six overlapping interactions I_M: models for sporadic and remote interactions (*R*, *r*), models for frequent and remote interactions (*M*, *m*), models for frequent and closed interactions (*D*, *d*). The overlapping interaction models provide either scalar or vectorial description of the interactions at fragment level.
- Four algorithms of molecular fragmentation applied to atomic pairs (F_C): fragmentation based on paths (Cluj [27]) (*P*) or on distances (Szeged [58]) (*D*); fragmentation in maximal fragments (*M*) or in minimal fragments (*m*). Note

that some parts of a molecule are more active than others and explain most of the activity/property of a molecule, as Hammet observed in the first SAR study [7].

- Nineteen algebraic operations based on the four fragmentation methods (S_M) below: ▪ *group of values*: minimum value (*m*), maximum value (*M*), lowest absolute value (*n*), highest absolute value (*N*); ▪ *arithmetic group*: sum (*S*), arithmetic mean of the number of fragment properties (*A*), arithmetic mean of the number of fragments (*a*), arithmetic mean of the number of atoms (*B*), arithmetic mean of the number of bonds (*b*); ▪ *geometric group*: multiplication (*P*), geometric mean of the number of fragment properties (*G*), geometric mean of the number of fragments (*g*), geometric mean of the number of atoms (*F*), geometric mean of the number of bonds (*f*); ▪ *harmonic group*: harmonic sum (*s*), harmonic mean of the number of fragments properties (*H*), harmonic mean of the number of fragments (*h*), harmonic mean of the number of atoms (*I*), harmonic mean of the number of bonds (*i*).
- Six linearization operators (L_O): identity (*I*), inverse (*i*), absolute value (*A*), inverse of absolute value (*a*), logarithm (*L*), logarithm of absolute value (*l*).
- The above-described functions are found in the descriptors name (in descending order). The number of members in the molecular descriptors family (MDF) is obtained by multiplying all the above operators:

$$\text{MDF} = D_M \times A_P \times P_D \times I_M \times F_C \times S_M \times L_O = 2 \times 6 \times 6 \times 24 \times 4 \times 19 \times 6 = 787968 \text{ (descriptors)}$$

The number of MDF members depends on the set of studied molecules. In the pool of descriptors, there are neither descriptors without a physical meaning (e.g. the logarithm of a negative number) nor descriptors with infinite value (e.g. those resulting from the division by zero). Finally, the degenerated values (e.g., a descriptor has the same value for two different molecules in the set) are removed from the descriptors pool by a bias procedure.

Using $d_G(i,j)$ for distances in the graph *G* between vertices *i* and *j*, we may recall the following definitions:

- Cluj distance fragments are sets of vertices obeying the relation [14, 59]:

  $$[CJ_{i,j,p}(G) = \{v \mid v \in \text{V(G)}, d_{G\setminus p}(i,v) < d_{G\setminus p}(j,v)\}, \text{ for any } p \in D(G)$$

  where $V(G)$ is the set of vertices in *G*, $D(G)$ is the set of all the shortest paths in *G* while $d_{G|p}(i,j)$ refers to the distance between vertices *i* and *j* measured in the graph *G* from which the path *p* was subtracted. The entries in the Cluj matrix are considered the maximum of all such fragments: $\mathbf{[UCJ]}_{i,j} = \max_p|CJ_{i,j,p}|$. The half sum of elements in the unsymmetrical Cluj matrix [**UCJ**] provides the hyper-Cluj index CJ_p while the half sum of [**UCJ**]·[**A**] provides the CJ_e index (where **[A]** is the adjacency matrix).
- The Szeged index was proposed by Gutman [60] in an attempt to expand the Wiener index calculation in cycle-containing graphs:

  $$[SZ(G)]_{i,j} = |SZ_{i,j}| \cdot |SZ_{j,i}|$$
  $$SZ_{i,j} = \{v \in V(G), d_G(i,v) < d_G(j,v)\}$$

  The half sum of elements in the Szeged matrix [SZ] provides the hyper-Szeged index SZ_p while the half sum of [SZ]·[A] provides the Szeged index *SZ*.
- A minimal subgraph of *G* can be defined in order to always contain at least one vertex, *i*:

  $$MinF(G)_{i,j} = (\{\text{i}\}, \varnothing)$$
- A maximal (connected) subgraph of *G* [61], containing the vertex *i* but not the vertex *j*, abbreviated as $MaxF(G)_{i,j}$ can be defined both by vertex and edge:

  $$V(MaxF(G)_{i,j}) = \{s \in VTemp(G)_{i,j} \mid D(VTemp(G)_{i,j})_{s,i} < \infty\}$$
  $$E(MaxF(G)_{i,j}) = \{(s,t) \in E(G) \mid s,t \in V(MaxF(G)_{i,j})\}$$
- Temporary graphs ($VTemp(G)_{i,j}$, $ETemp(G)_{i,j}$) are disconnected graphs, whose sets are defined as:

  $$VTemp(G)_{i,j} = \{s \in V(G) \mid s \neq j\},\ ETemp(G)_{i,j} = \{(u,v) \in E(G) \mid u,v \neq j\}$$

The online resource http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/j_mdf_demo.php could be used to calculate any MDF descriptors by providing the structure as *.hin file upon request.

- calculate the MDF descriptors and the modality of calculus may be followed there.

MDF software

The PHP (Pre Hypertext Processor) - MySQL database - FreeBSD server triad was used to implement the MDF methodology.

Three databases (one temporary `MDFSARtmp` for the set in work, one permanent `MDFSARs` for the final data and a dedicated one, `MDFSAR15aa`, for the set of amino-acids) were stored on a FreeBSD server using MySQL database.

Four separate programs were designed to assist the users for applying MDF methodology based on the schedule of the operations: preparation of the database structure (*0_mdf_prepare.php*), generating of the descriptors (*1_mdf_generate.php*), adaptation of the descriptors to the observed measures via a list of six alternatives (*2_mdf_linearize.php*), reducing the descriptors pool size based on explanatory power (*3_mdf_bias.php*) and hierarchizing of descriptors based on explaining amount (*4_mdf_order.php*).

As long as MDF methodology was exclusively built on a search space encoded in respect of all rigors of a genetic code sequence, the creation of a genetic algorithm to search the performing models was the natural way forward in development of fast and efficient searching application [62-64]. The genetic algorithm (GA) [65] for QSAR/QSPR modeling was implemented to act as follows (details could be found on [66-68]:

- Inheritance and mutation. The six operators described above (the linearization operator was excluded) represent the solution domain of the 131328 size (2 × 6 × 6 × 24 × 4 × 19). An offspring is obtained from a parent (inheritance) through a transformation (mutation). The number of obtained offspring's is six time higher than the number of parents. A fitness function was defined in order to obtain offspring with real and distinct values. Nearly half of offspring died during this step due to mutation. About 300000 offspring with a seven letters genetic representation were valid at the end of this step.

- Selection. A bias procedure (selection) was applied to the valid offspring obtained in the previous step. A determination coefficient with distinct first nine digits was defined as the first fitness function in this step. About 100000 members were obtained by applying the first fitness function. The second fitness function was applied to the valid members in order to identify the descriptors with the highest performances in terms of correlation with the measured activity/property.
- Crossover. Pairs of valid offspring (MDF members) identified after selection were crossed over in order to obtain models with two MDF members. Two fitness functions were defined and applied during this step: the highest value of the determination coefficient, its adjusted value, and the highest value of the cross-validation leave-one-out score.

Two different approaches were implemented in order to identify simple and multiple linear regressions.

The first approach refers to a series of programs implemented as FreePascal client-server programs able to identify, withdraw degenerated and correlated MDF members, and search for simple and multiple linear regressions [63, 69]. Our programs used the MySQL dynamic libraries to connect to the databases and implement five search strategies: four systematic searches (two descriptors, three descriptors, two pairs of descriptors, more than two pairs of descriptors) and one heuristic (random search in *i* variables).

The second approach implemented the search strategy using a genetic algorithm [68] also implemented as a FreePascal program. The parameters and ranges of genetic algorithm are specific to the investigated set of molecules and details for some sets could be found in [70, 71]. The genetic algorithm evolved as follows [64, 67]:

```
1. Read the configuration file (connection to database)
2. Read the genetic code of MDF
3. Allocates the memory space for the genetic code
4. Creates the translation functions between *storage address* and *genetic code*
5. Connects to the database and selects the experimental data (also provides the sample size)
6. Allocates the memory space for experimental data and for MDF values; reads the experimental values from the database
7. Displays the sample size of the MDF population (based on genetic code calculation) and the number of molecules in the set
8. Read and select the values of descriptors for all molecules in the set
9. Read the cultivar configuration file
10. Creates the output files
11. Iterate
11.1. Initiate the values of global minimum and global maximum
11.2. Create the first generation of evolution
11.3. Iterate
    1st step:
        ▪ Selects pairs of MDF members using the selection operator.
    2nd step:
        ▪ Calculates selection and survival scores for the selected descriptors
        ▪ Calculates the objected function for the set of molecules
        ▪ Identifies the group of descriptors from the sample which meets the objective in its generation and automatically includes the descriptors belonging to this group in the next generation
    3rd step:
        ▪ Identifies (with a low probability and using a discrete uniform probability function) the segment to be mutated and applies the mutation to the selected descriptors
    4th step:
        ▪ Identifies (using a discrete uniform probability function) the segment to be crossed over and produces offspring through cross-over
    5th step:
        ▪ Identifies (with a low probability and using a discrete uniform probability function) the segment to be mutated in offspring and applies the mutation to the offspring descriptors
    6th step:
        ▪ Uses the survival operator to replace some parent descriptors with offspring descriptors.
12. Repeat until an imposed condition of the objective function is met (e.g. a value of the determination coefficient) or the required number of iterations occurs.
```

The performances of the MDF methodology have been tested on various sets of chemical (bioactive) compounds. The following criteria were used to test the model [51, 72, 73]:

- Goodness-to-fit of the model (highest and significant value of correlation coefficient and lowest difference related to the adjusted value (min(r^2-r_{adj}^2)) and leave-one-out score (min(r^2-r_{loo}^2))).

- Statistically significant regression model (at a significance level of 5%) with the smallest possible number of descriptors.
- Absence of co-linearity between descriptors (correlation coefficient not statistically significant when applying all correlation methods (Pearson (r), semi-Quantitative (r_{sQ}), Spearman (ρ), Kendall's (τ_a, τ_b, τ_c) and Gamma (Γ)) [74].
- Internal valid models (cross-validation leave-one-out analysis) [75].
- Information criteria used to compare performances of two different models (the lowest values indicate the best model [76, 77]): corrected Akaike information criteria (AIC_c) / AIC based on the determination coefficient (AIC_{R2}) / McQuarrie and Tsai corrected AIC (AIC_u) [78, 79], Schwarz (or Bayesian) Information Criterion (BIC) [80], Amemiya Prediction Criterion (APC) [81], Hannan-Quinn Criterion (HQC) [82].
- Kubinyi function (FIT) [83, 84] (the highest value indicates the best model).

Whenever possible, a correlation-correlated analysis was also conducted in order to compare the QSAR/QSPR model with previously reported models using the Steiger's test at a significance level of 5% [53].

The representations of the main steps involved on the MDF-based QSAR/QSPR modeling approach are presented in Figure 1.

[Figure 1 comes about here]

The estimation and prediction abilities of the MDF were also evaluated and the results are presented in the Results and Discussion section in order to support the usefulness of the MDF modeling approach.

Results and Discussion

The MDF methodology was successfully implemented as a tool for QSAR/QSPR modeling. The software is available online at the AcademicDirect gateway – Chemistry – SARs – MDS_SARs (http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/). The main data-interfaces and their characteristics are presented in Figure 2. The MDF computer-based system did not include programs to draw molecules or to convert them into different formats. Various stand-alone programs were developed to prepare molecules for modeling (conversion of *.sdf files into *.hin files, conversion of *.mol files as *.hin files, geometry optimization by HyperChem [85], etc.).

[Figure 2 comes about here]

Although the MDF integrative system was created as an Intranet system, it also has some features for Internet users (e.g. MDF members can be calculated as both Intranet and Internet networks). The value of MDF members for one *.hin file molecule can be calculated by using the `BorQ SARs by Sets`. External users are able to analyze the internal validation of already investigated sets of compounds in both leave-one-out (`LOO Analysis (LOO: Leave-One-Out)`) and leave-many-out (`TvT Experiment (TvT: Training vs. Test)`) analyses. The system also allows the prediction of activity/property of new compounds from the studied sets of compounds, based on a previously found MDF QSAR/QSPR model (`SARs (SAR: Structure-Activity Relationships)`). The ` Investigator` tool was created for the management of current jobs (last set in the `MDFSARtmp` temporary database).

The values of calculated MDF descriptors proved reliable since identical values are obtained if the computations are done more than once. The time needed to calculate the MDF descriptors is linearly related to the complexity of molecules in the investigated set. The time needed to search for multiple linear regressions using complete search approach is most time-consuming step (over 90% of the total time needed to analyze a set of active compounds) but was significantly reduced by introduction of GA search approach [68, 86-89]. The developed system proved able to function properly on any type of chemical structures if the molecules are of small or medium size (100 atoms). Unfortunately, we were not able to figure out to date how to surpass the problem of exponential increase of execution time related to molecule complexity general problem for all programs that analyze complex chemical structure, but we work on this problem. Moreover, the most time-expensive parts of the implemented system were translated into executable programs able to run under Windows and using computer resources at maximum.

The value of the MDF methodology and of the MDF integrative system can be analyzed in terms of ability to identify accurate and valid MDF QSAR/QSPR models. More than fifty various activities / properties on different sets of compounds were already investigated. A summary of MLR MDF QSAR/QSPR models is presented Table 1.

[Table 1 comes about here]

The number of adapted to measured properties/activities MDF members ranged in investigated datasets from 62712 to 111477, with an average of 82702±13580 members (95% confidence interval [78622; 86781]). A number of one hundred and thirty two descriptors proved able to explain the structure-property/activity and were used in the best models. Only two MDF descriptors were identified in more than one model: inPRlQg (the MLR of 408461 set & 408464 set – the same set of compounds but different investigated property) and lAmrfEt (the MLR of MR10 set & the MLR of DevMTOp00 set – different sets and different property/activity). The most frequent MDF descriptors (identified according to the letters in their name) used in the MDF QSAR/QSPR models are listed in Table 2.

[Table 2 comes about here]

In most cases, the studied property/activity correlated with MDF descriptors of both geometric and topological nature (~53%) while 31% of cases proved to be only of geometric nature. The top three atomic properties related to the studied property/activity were: atomic partial charge, total molecular energy (51% of descriptors in the models), number of hydrogen atoms adjacent to the focused atom (17% of descriptors in the models), and relative atomic mass (11% of descriptors in the models). Only 13 out of 24 interaction descriptors occurred in the best MDF QSAR/QSPR models (the most frequently identified was $K=p_1p_2d$ – 17% of descriptors). The most useful algorithm for molecular fragmentation proved to be the fragmentation in maximal fragments (identified in 39% of descriptors used in the best models) closely followed by fragmentation based on distances (identified in 34% of cases). Almost 53% of possible global overlapping of fragment interactions could be identified in the studied MDF QSAR/QSPR models along with 50% of possible linearization operators.

The analysis of the MDF QSAR/QSPR in terms of goodness-of-fit revealed that all models had significantly higher correlation coefficients provided by statistically significant regression models at a significance level of 5% ($p \leq 0.00035$, see Table 2). The smallest difference, either for the difference between the determination coefficient and its adjusted value or for the difference between the determination coefficient and the r^2_{loo} score, proved to be below $1.20\cdot10^{-6}$.

In 98% of the models, both differences were below 0.05; 67% and 47% were below 0.01 and 40%, and 13% were below 0.0005. Therefore, the MDF methodology proved able to provide valid and reliable MDF QSAR/QSPR models. Furthermore, the MDF models demonstrated statistically significant higher goodness-of-fit compared to previous models reported in the literature (confirmed by the correlated correlation analysis: Steiger's test [53], significance level of 5%) (19654 set [115], 22583 set [116], 23159 set [117], 26449 set [118], IChr10 set [27], MR10 set [119], PCB_rrf set [120], PCB_lkow set [121], RRC_lkow set [122], Tax385 & Tox395 [123], 41521 set [124], Triazines [125], 52344 set [126], 23151 set [127], 408462 set [128], 408464 set [129], 52730 set [130], cqdmdfv set [28]). The metaheuristic search was as good as the complete search and had the advantage of smaller costs in terms of resources and time [66, 68, 70].

The values of information criteria were used to compare the MLR models obtained by using different approaches on the same dataset of compounds. The lowest values of information criteria and the highest FIT function obtained for MDF QSAR/QSPR models compared to previously reported models showed the performances of the MDF approach [28, 29, 30].

The MDF methodology opens a new low-cost pathway in understanding the link between the chemical structure of compounds and their property/activity, the investigation of already known compounds as well as in the discovery of new compounds.

Conclusions

The molecular descriptors family methodology, developed and implemented as a platform for investigating and characterizing quantitative relationships between the chemical structure and the activity/property of active compounds, proved successful.

The MDF methodology showed estimation and prediction abilities through valid and reliable MLR QSAR/QSPR models. Its implementation provides a risk-free environment for molecular modeling and could be used as a valid and reliable tool in the investigation of structure-activity/property relationships. Daily computer and Internet skills, besides strong knowledge of QSAR/QSPR models, are needed for using the implemented MDF system.

List of abbreviations

2D descriptors = descriptors derived from a two-dimensional graph representation of a molecule
3D = three dimensional
CoMFA = **Co**mparative **M**olecular **F**ield **A**nalysis
CoMSIA = **Co**mparative **M**olecular **S**imilarity **I**ndices **A**nalysis
FPIF = **F**ragmental **P**roperty **I**ndex **F**amily
FreeBSD = Free Berkeley Systems Distribution
GA = genetic algorithm
MDF = **M**olecular **D**escriptors **F**amily
MDFV = **M**olecular **D**escriptors **F**amily on **V**ertices
MLR = multiple linear regression
MSD = **M**inimal **S**teric **D**istance
MS-WHIM = **M**olecular **S**urface **W**eighted **H**olistic **I**nvariant **M**olecular
MTD = **M**inimal **T**opological **D**istance
MySQL = My Structured Query Language
PHP = Pre Hypertext Processor
QSARs = **Q**uantitative **S**tructure-**A**ctivity **R**elationships
QSPRs = **Q**uantitative **S**tructure-**P**roperty **R**elationships
SAR = **S**tructure-**A**ctivity **R**elationship
SPR = **S**tructure-**P**roperty **R**elationship

TOPS-MODE = **TOP**ological **S**ub-structural **MO**lecular **DE**sign
WHIM = **W**eighted **H**olistic **I**nvariant **M**olecular

Acknowledgements

The study was supported by POSDRU/89/1.5/S/62371 through a fellowship for L. Jäntschi.
This paper is dedicated to Professor Milan Randic on his 80th birthday.

References

[1] Crum-Brown, A.; Fraser, T.R. On the Connection between Chemical Constitution and Physiological Action. Part I. On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *Phil. Trans. R. Soc. London*, **1868**, *25*, 151-203.

[2] Richet, C.R. Sur rapport entre la toxicité et les propriétés physiques des corps. *C. R. Seances Soc. Biol. Fi.*, **1893**, *9*, 775-776.

[3] Meyer, H. Zur Theorie der Alkoholnarkose Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung?. *N.-S. Arch. Pharmacol.*, **1899**, *42*, 109-118.

[4] Gao, Q.; Yang, L.; Zhu, Y. Pharmacophore based drug design approach as a practical process in drug discovery. *Curr. Comput.-Aided Drug Des.*, **2010**, *6*, 37-49.

[5] Novič, M.; Vračko, M. QSAR models for reproductive toxicity and endocrine disruption activity, *Molecules*, **2010**, *15(3)*, 1987-1999.

[6] Valerio, L.G.; Yang, C.; Arvidson, K.B.; Kruhlak, N.L. A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, **2010**, *6*, 505-518.

[7] Hammett, L.P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.*, **1935**, *17*, 125-136.

[8] Hammett, L.P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.*, **1937**, *59*, 96-103.

[9] Hansch, C. Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.*, **1969**, *2*, 232-239.

[10] Verma, R.P.; Hansch, C. Use of 13C NMR chemical shift as QSAR/QSPR descriptor. *Chem. Rev.*, **2011**, *111*, 2865-2899.

[11] Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health. Persp.*, **2003**, *111*, 1361-1375.

[12] Bajorath, J. Integration of virtual and high-throughput screening. Nat. Rev. Drug. Discov., **2002**, *1*, 882-894.

[13] Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput. Aid. Mol. Des.*, **2000**, *14*, 215-232.

[14] Diudea, M.; Gutman, I.; Jäntschi, L. *Molecular Topology*. 2nd ed.; Nova Science, Huntington: New York, **2002**.

[15] Diudea, M.V. *QSPR/QSAR Studies by Molecular Descriptors*. Nova Science, Huntington: New York, **2001**.

[16] Michielan, L.; Moro, S. Pharmaceutical perspectives of nonlinear QSAR strategies. *J. Chem. Inf. Model.,* **2010**, *50*, 961-978.

[17] Helguera, A.M.; Combes, R.D.; González, M.P.; Cordeiro, M.N. Applications of 2D descriptors in drug design: a DRAGON tale. *Curr. Top. Med. Chem.*, **2008**, *8*, 1628-1655.

[18] Kier, L.B. My journey through structure: The structure of my journey. *Internet Electron. J. Mol. Des.*, **2006**, *5*, 181-191.

[19] Bonchev, D. My life-long journey in mathematical chemistry. *Internet Electron. J. Mol. Des.*, **2005**, *4*, 434-490.

[20] Trinajstić, N. A life in science. *Internet Electron. J. Mol. Des.*, **2003**, *2*, 413-434.

[21] Gute, B.D.; Basak, S.C.; Mills, D.; Hawkins, D.M. Tailored similarity spaces for the prediction of physicochemical properties. *Internet Electron. J. Mol. Des.*, **2002**, *1*, 374-387.

[22] Roy, K.; Saha, A. Comparative QSPR studies with molecular connectivity, molecular negentropy and TAU indices. Part 2. Lipid-water partition coefficient of diverse functional acyclic compounds. *Internet Electron. J. Mol. Des.*, **2003**, *2*, 288-305.

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

[24] Todeschini, R.; Lasagni, M.; Marengo, E. New molecular descriptors for 2D and 3D structures. Theory. *J. Chemom*, **1994**, *8*, 263-272.

[25] Simon, Z.; Chiriac, A.; Holban, S.; Ciubotariu, D.; Mihalaş, G.I. *Minimum Steric Difference. The MTD Method for QSAR Studies*. Research Studies Pr, **1984**.

[26] Jäntschi, L.; Katona, G.; Diudea, M.V. Modeling Molecular Properties by Cluj Indices. *MATCH - Commun. Math. Comput. Chem.*, **2000**, *41*, 151-188.

[27] Jäntschi, L. MDF - A New QSAR/QSPR Molecular Descriptors Family. *Leonardo J. Sci.*, **2004**, *3*, 68-85.

[28] Bolboacă, S.D.; Jäntschi, L. Comparison of QSAR Performances on Carboquinone Derivatives. *THESCIENTIFICWORLDJO*, **2009**, *9*, 1148-1166.

[29] Bolboacă, S.D.; Marta, M.M.; Jäntschi, L. Binding affinity of triphenyl acrylonitriles to estrogen receptors: quantitative structure-activity relationships. *Folia Med.*, **2010**, *52*, 37-45.

[30] Bolboacă, S.D.; Marta, M.M.; Stoenoiu, C.E.; Jäntschi, L. Molecular Descriptors Family on Vertex Cutting: Relationships between Acelazolamide Structures and their Inhibitory Activity. *Appl. Med. Inform.*, **2009**, *25*, 65-74.

[31] Estrada, E. How the parts organize in the whole? A top-down view of molecular descriptors and properties for QSAR and drug design. *Mini Rev. Med. Chem.*, **2008**, *8*, 213-221.

[32] Verma, J.; Khedkar, V.M.; Coutinho, E.C. 3D-QSAR in drug design--a review. *Curr. Top. Med. Chem.*, **2010**, *10*, 95-115.

[33] Andrade, C.H.; Pasqualoto, K.F.; Ferreira, E.I.; Hopfinger, A.J. 4D-QSAR: perspectives in drug design. *Molecules*, **2010**, *15*, 3281-3294.

[34] Polanski, J. Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. *Curr. Med. Chem.*, **2009**, *16*, 3243-3257.

[35] Vedani, A.; Dobler, M.; Lill, M.A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.*, **2005**, *48*, 3700-3703.

[36] Costescu, A.; Diudea, M.V. QSTR study on aquatic toxicity against Poecilia reticulata and Tetrahymena pyriformis using topological indices. *Internet Electron. J. Mol. Des.*, **2006**, *5*, 116-134.

[37] Miličević, A.; Nikolić, S.; Plavšić, D.; Trinajstić, N. On the Hosoya Z index of general graphs. *Internet Electron. J. Mol. Des.*, **2003**, *2*, 160-178.

[38] Skvortsova, M.I.; Fedyaev, K.S.; Palyulin, V.A.; Zefirov, N.S. Molecular design of chemical compounds with prescribed properties from QSAR models containing the Hosoya index. *Internet Electron. J. Mol. Des.*, **2003**, *2*, 70-85.

[39] Roy, K.; Ghosh, G. Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. *Internet Electron. J. Mol. Des.*, **2003**, *2*, 599-620.

[40] González, M.P.; Terán, C.; Saíz-Urra, L.; Teijeira, M. Variable selection methods in QSAR: an overview. *Curr. Top. Med. Chem.*, **2008**, *8*, 1606-1627.

[41] Yap, C.W.; Li, H.; Ji, Z.L.; Chen, Y.Z. Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini Rev. Med. Chem.*, **2007**, *7*, 1097-1107.

[42] Hasegawa, K.; Funatsu, K. Non-linear modeling and chemical interpretation with aid of support vector machine and regression. *Curr. Comput. Aided Drug Des.*, **2010**, *6*, 24-36.

[43] Thurstone, L.L. Multiple Factor Analysis. *Psychol. Rev.*, **1931**, *38*, 406-427.

[44] Wang, G.; Li, Y.; Liu, X.; Wang, Y. Understanding the aquatic toxicity of pesticide: Structure-activity relationship and molecular descriptors to distinguish the ratings of toxicity. *QSAR Comb. Sci.*, **2009**, *28*, 1418-1431.

[45] Linusson, A.; Elofsson, M.; Andersson, I.E.; Dahlgren, M.K. Statistical molecular design of balanced compound libraries for QSAR modeling. *Curr .Med. Chem.*, **2010**, *17*, 2001-2016.

[46] Liu, W.; Johnson, D.E. Clustering and its application in multi-target prediction. *Curr. Opin. Drug Discov. Devel.*, **2009**, *12*, 98-107.

[47] Ghosh, P.; Bagchi, M.C. QSAR modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. *Curr. Med. Chem.*, **2009,** *16*, 4032-4048.

[48] Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. Neural networks in building QSAR models. *Methods Mol. Biol.*, **2008**, *458*, 137-158.

[49] Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs, *Environ. Health Perspec.*, **2003**, *111*, 1361-1375.

[50] Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.*, **2007**, *47*, 2345-2357.

[51] Bolboacă, S.D.; Jäntschi, L. Modelling the property of compounds from structure: statistical methods for models validation. *Environ. Chem. Lett.*, **2008**, *6*, 175-181.

[52] Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.*, **2007**, *13*, 3494-3504.

[53] Steiger, J.H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, **1980**, *87*, 245-251.

[54] Weiner, S.J.; Kollman, P.A.; Case, D.A.; Singh, U.C.; Chio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **1984**, *106*, 765-784.

[55] Jug, K.; Nanda, D.N. SINDO1 II. Application to ground states of molecules containing carbon, nitrogen and oxygen atoms. *Theor. Chim. Acta*, **1980**, *57*, 107-130.

[56] Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.*, **1985**, *107*, 3902-3909.

[57] Jäntschi, L.; Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. *Leonardo El. J. Pract. Technol.*, **2005**, *4*, 76-98.

[58] Khadikar, P.V.; Deshpande, N.V.; Kale, P.P.; Dobrynin, A.; Gutman, I.; Domotor, G.J. The Szeged index and an analogy with the Wiener index. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 547-550.

[59] Diudea M. V., Cluj Matrix Invariants. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 300-305.

[60] Gutman, I. A Formula for the Wiener Number of Trees and Its Extension to Graphs Containing Cycles, Graph Theory Notes New York, **1994**, *27*, 9-15.

[61] Jäntschi, L.; Bolboacă, S.D.; Furdui, C.M. Characteristic and Counting Polynomials: Modelling Nonane Isomers Properties. *Mol. Simul.*, **2009**, *35*, 220-227.

[62] Jäntschi, L.; Bolboacă, S.D., Diudea, M.V. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization. *Int. J. Mol. Sci.*, **2007**, *8*, 1125-1157.

[63] Jäntschi, L. Delphi Client - Server Implementation of Multiple Linear Regression Findings: a QSAR/QSPR Application. *Appl. Med. Inform.*, **2004**, *15*, 48-55.

[64] Jäntschi, L. A genetic algorithm for structure-activity relationships: software implementation. http://arxiv.org/abs/0906.4846 (accessed June 7, 2011).

[65] Chambers, D.L. *The Practical Handbook of Genetic Algorithms: Applications*. 2nd ed.; Chapman and Hall/CRC: Florida, **2001**.

[66] Jäntschi, L.; Bolboacă, S.D.; Sestraş, R.E. Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls. *J. Mol. Model.*, **2010**, *16*, 377-386.

[67] Jäntschi, L. Genetic algorithms and their applications. University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, Cluj-Napoca, 2010.

[68] Jäntschi, L.; Bolboacă, S.D.; Sestraş, R.E. A Study of Genetic Algorithm Evolution on the Lipophilicity of Polychlorinated Biphenyls. *Chem. Biodivers.*, **2010**, *7*, 1612-1872.

[69] Jäntschi, L.; Bolboacă, S.D. Structure-Activity Relationships on the Molecular Descriptors Family Project at the End. *Leonardo El. J. Pract. Technol.*, **2007**, *11*, 163-180.

[70] Jäntschi, L.; Bolboacă, S.D.; Sestraş, R.E. Recording Evolution Supervised by a Genetic Algorithm for Quantitative Structure-Activity Relationship Optimization. *Appl. Med. Inform.*, **2010**, *26*, 89-100.

[71] Jäntschi, L.; Bolboacă, S.D., Bălan, M.; Sestraş, R.E.; Diudea, M.V. Results of Evolution Supervised by Genetic Algorithms. *Not. Sci. Biol.*, **2010**, *2*, 12-15.

[72] Bolboacă, S.D.; Pică, E.M.; Cimpoiu, C.V.; Jäntschi, L. Statistical Assessment of Solvent Mixture Models Used for Separation of Biological Active Compounds. *Molecules*, **2008**, *13*, 1617-1639.

[73] Hawkins, D.M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1-12.

[74] Bolboacă, S., Jäntschi, L. Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds. *Leonardo J. Sci.*, **2006**, *9*, 179-200.

[75] Hurvich, C.M.; Tsai, C. Regression and Time Series Models of Finite but Unknown Order. *Biometrika*, **1989**, *76*, 297-307.

[76] Burnham, K.P.; Anderson, D.R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.*, **2004**, *33*, 261-304

[77] Bozdogan, H. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*, **1987**, *52*, 345-370.

[78] Hurvich, C.M.; Tsai, C.-L. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, **1991**, *78*, 499-509.

[79] McQuarrie, A.D.R.; Tsai, C.-L. *Regression and time series model selection*. World Scientific Publishing Co.: **1998**.

[80] Schwarz, G. Estimating the dimension of a Model. *Ann. Stat.*, **1978**, *6*, 461-464.

[81] Amemiya, T. Qualitative response models: A survey. *J. Econ. Lit.*, **1981**, *19*, 1483-1536.

[82] Hannan, E.J.; Quinn, B.G. The determination of the Order of an Autoregression. *J. Roy. Statistical Society B*, **1979**, *41*, 190-195.

[83] Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Rel.*, **1994**, *3*, 393-401.

[84] Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Rel.*, **1994**, *13*, 285-294.

[85] Bolboacă, S.D.; Jäntschi, L. Computer Assisted Geometry Optimization for in silico Modeling. **2010**.

[86] Jäntschi, L.; Bolboacă, S.D.; Bălan, M.; Sestraş, R.E. Tendency of Evolution Supervised by Genetic Algorithms. . *Bull. UASVM Hort.*, **2010**, *67*, 80-85.

[87] Jäntschi, L.; Bolboacă, S.D.; Diudea, M.V.; Sestraş, R.E. Average Trends over Millennia of Evolution Supervised by Genetic Algorithms. 1. Analysis of Genotypes. . *Bull. UASVM Hort.*, **2010**, *67*, 72-79.

[88] Jäntschi, L.; Bolboacă, S.D.; Diudea, M.V.; Sestraş, R.E. Average Trends over Millennia of Evolution Supervised by Genetic Algorithms. 2. Analysis of Phenotypes. . *Bull. UASVM Agric.*, **2010**, *67*, 161-168.

[89] Jäntschi, L.; Bolboacă, S.D.; Diudea, M.V.; Sestraş, R.E. Average Trends over Millennia of Evolution Supervised by Genetic Algorithms. 3. Analysis of Phenotypes Associations. *Bull. UASVM Agric.*, **2010**, *67*, 169-174.

[90] Zhou, Y.-X.; Xu, L.; Wu, Y.-P.; Liu, B.-L. A QSAR study of the antiallergic activities of substituted benzamides and their structures. *Chemometr. Intell. Lab.*, **1999**, *45*, 95-100.

[91] Castro, E.A.; Torrens, F.; Toropov, A.A.; Nesterov, I.V.; Nabiev, O.M. QSAR Modeling ANTI-HIV-1 Activities by Optimization of Correlation Weights of Local Graph Invariants. *Mol. Simul.*, **2004**, *30*, 691-696.

[92] Toporov, A.A.; Toporova, A.P. QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struc.-THEOCHEM*, **2002**, *578*, 129-134.

[93] Baker, J.R.; Mihelcic, J.R.; Sabljic, A. Reliable QSAR for estimating KOC for persistent organic pollutants: correlation with molecular connectivity indices. *Chemosphere*, **2001**, *45*, 213-221.

[94] Wei, D.; Zhang, A.; Wu, C.; Han, S.; Wang, L., Progressive study and robustness test of QSAR model based on quantum chemical parameters for predicting BCF of selected polychlorinated organic compounds (PCOCs). *Chemosphere*, **2001**, *44*, 1421-1428.

[95] Ungwitayatorn, J.; Pickert, M.; Frahm, A.W. Quantitative structure-activity relationship (QSAR) study of polyhydroxyxanthones. *Pharm. Acta Helv.*, **1997**, *72*, 23-29.

[96] Abraham, M.H.; Kumarsingh, R.; Cometto-Muniz, J.E.; Cain, W.S. A Quantitative Structure-Activity Relationship (QSAR) for a Draize Eye Irritation Database. *Toxicol. in Vitro*, **1998**, *12*, 201-207.

[97] Morita, H.; Gonda, A.; Wei, L.; Takeya, K.; Itokawa, H. 3D QSAR Analysis of Taxoids from Taxus Cuspidata Var. Nana by Comparative Molecular Field Approach. *Bioorg. Med. Chem. Lett.*, **1997**, *7*, 2387-2392.

[98] Toropov, A.; Toropova, A.; Ismailov, T.; Bonchev, D. 3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes. *J. Mol. Struc.-THEOCHEM*, **1998**, *424*, 237-247.

[99] Zakarya, D.; Larfaoui, E.M.; Boulaamail, A.; Tollabi, M.; Lakhlifi, T. QSARs for a series of inhibitory anilides. *Chemosphere*, **1998**, *36*, 2809-2818.

[100] Brasquet, C.; Le Cloirec, P. QSAR for Organics Adsorption Onto Activated Carbon In Water: What About The Use Of Neural Networks?. *Wat. Res.*, **1999**, *33*, 3603-3608.

[101] Supuran, C.T.; Clare, B.W. Carbonic anhydrase inhibitors - Part 57: Quantum chemical QSAR of a group of 1,3,4-thiadiazole- and 1,3,4-thiadiazoline disulfonamides with carbonic anhydrase inhibitory properties. *Eur. J. Med. Chem.*, **1999**, *34*, 41-50.

[102] Hasegawa, K.; Arakawa, M.; Funatsu, K. 3D-QSAR study of insecticidal neonicotinoid compounds based on 3-way partial least squares model. *Chemometr. Intell. Lab.*, **1999**, *47*, 33-40.

[103] Shertzer, G.H.; Tabor, M.W.; Hogan, I.T.D.; Brown, J.S.; Sainsbury, M. Molecular modeling parameters predict antioxidant efficacy of 3-indolyl compounds. *Arch. Toxicol.*, **1996**, *70*, 830-834.

[104] Ade, T.; Zaucke, F.; Krug, H.F. The structure of organometals determines cytotoxicity and alteration of calcium homeostasis in HL-60 cells. *Anal. Bioanal. Chem.*, **1996**, *354*, 609-614.

[105] Hodisan, T.; Culea, M.; Cimpoiu, C.; Cot, A. Separation, identification and quantitative determination of free amino acids from plant extracts. *J. Pharm. Biomed. Anal.*, **1998**, *18*, 319-323.

[106] Kawakami, J.; Hoshi, K.; Ishiyama, A.; Miyagishima, S.; Sato, K. Application of a Self-Organizing Map to Quantitative Structure-Activity Relationship Analysis of Carboquinone and Benzodiazepine. *Chem. Pharm. Bull.*, **2004**, *52*, 751-755.

[107] *Development of Marine Sediment Toxicity for Ordnance Compounds and Toxicity Identification Evaluation Studies at Select Naval Facilities*. NFESC Contract Report CR 01-002-ENV. **2000**.

[108] Selassie, C.D.; Li, R.L.; Poe, M.; Hansch, C. On the optimization of hydrophobic and hydrophilic substituent interactions of 2,4-diamino-5-(substituted-benzyl)pyrimidines with dihydrofolate reductase. *J. Med. Chem.*, **1991**, *34*, 46-54.

[109] Jäntschi, L.; Mureşan, S.; Diudea, M., Modeling Molecular Refraction and Chromatographic Retention by Szeged Indices. *Stud. Univ. Babes-Bolyai Chem.*, **2000**, *XLV*, 313-318.

[110] Opriş, D.; Diudea, M.V. Peptide property modeling by Cluj indices. *SAR QSAR Environ. Res.*, **2001**, *12*, 159-179.

[111] Eisler, R.; Belisle, A.A. *Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review*, **1996**, p 75.

[112] Ivanciuc, O. Artificial neural networks applications. Part 4. Quantitative structure-activity relationships for the estimation of the relative toxicity of phenols for Tetrahymena. *Rev. Roum. Chim.*, **1998**, *43*, 255-260.

[113] Smith, C.J.; Hansch, C.; Morton, M.J. QSAR treatment of multiple toxicities: the mutagenicity and cytotoxicity of quinolines. *Mutat. Res.*, **1997**, *379*, 167-175.

[114] Diudea, M.; Jäntschi, L.; Pejov, L. Topological Substituent Descriptors. *Leonardo El. J. Pract. Technol.*, **2002**, *1*, 1-18.

[115] Jäntschi, L.; Bolboacă, S.D. Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction. *Clujul Med.*, **2007**, *LXXX*, 125-132.

[116] Bolboacă, S.; Ţigan, Ş.; Jäntschi, L. Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives. In *European Federation for Medical Informatics Special Topic Conference & ROMEDINF*, Reichert, A.; Mihalas, G.; Stoicu-Tivadar, L.; Schultz, S.; Engelbrecht, R., Eds. ISO Press: Timisoara, **2006**; pp 222-226.

[117] Jäntschi, L. Delphi Client - Server Implementation of Multiple Linear Regression Findings: a QSAR/QSPR Application. *Appl. Med. Inform.*, **2004**, *15*, 48-55.

[118] Bolboacă, S.; Jäntschi, L. Molecular Descriptors Family on Structure Activity Relationships 3. Antituberculotic Activity of some Polyhydroxyxanthones. *Leonardo J. Sci*, **2005**, *7*, 58-64.

[119] Jäntschi, L.; Bolboacă, S. Molecular Descriptors Family on Structure Activity Relationships 4. Molar Refraction of Cyclic Organophosphorus Compounds. *Leonardo El. J. Pract. Technol.*, **2005**, *7*, 55-102.

[120] Jäntschi, L. QSPR on Estimating of Polychlorinated Biphenyls Relative Response Factor using Molecular Descriptors Family. *Leonardo El. J. Pract. Technol.*, **2004**, *5*, 67-84.

[121] Jäntschi, L.; Bolboacă, S. Molecular Descriptors Family on Structure Activity Relationships 6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls. *Leonardo El. J. Pract. Technol.*, **2006**, *8*, 71-86.

[122] Jäntschi, L.; Bolboacă, S.D. Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information. *Int. J. Quantum Chem.*, **2007**, *107*, 1736-1744.

[123] Jäntschi, L.; Bolboacă, S. Molecular Descriptors Family on QSAR Modeling of Quinoline-based Compounds Biological Activities. In *The 10th Electronic Computational Chemistry Conference*, Online, **2005**.

[124] Bolboacă, S.; Jäntschi, L. Molecular Descriptors Family on Structure Activity Relationships 2. Insecticidal Activity of Neonicotinoid Compounds. *Leonardo J. Sci.*, **2005**, *6*, 78-85.

[125] Bolboacă, S.; Jäntschi, L. Molecular Descriptors Family on Structure-Activity Relationships: Modeling Herbicidal Activity of Substituted Triazines Class. *Bull. UASVM Agric.*, **2006**, *62*, 35-40.

[126] Bolboacă, S.; Filip, C.; Ţigan, Ş.; Jäntschi, L. Antioxidant Efficacy of 3-Indolyl Derivates by Complex Information Integration. *Clujul Med.*, **2006**, *LXXIX*, 204-209.

[127] Jäntschi, L.; Bolboaca, S. Molecular Descriptors Family on Structure Activity Relationships 5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates. *Leonardo J. Sci.*, **2006**, *8*, 77-88.

[128] Jäntschi, L.; Ungureşan, M.L.; Bolboacă, SD. Integration of Complex Structural Information in Modeling of Inhibition Activity on Carbonic Anhydrase II of Substituted Disulfonamides. *Appl. Med. Inform.*, **2005**, *17*, 12-21.

[129] Jäntschi, L.; Bolboacă, S. Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information. *Electron. J. Biomed.*, **2006**, *2*, 22-33.

[130] Bolboacă, S.D.; Jäntschi, L. Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information. *Ther. Pharmacol. Clin. Toxicol.*, **2006**, *10*, 110-114.

Figures/illustrations

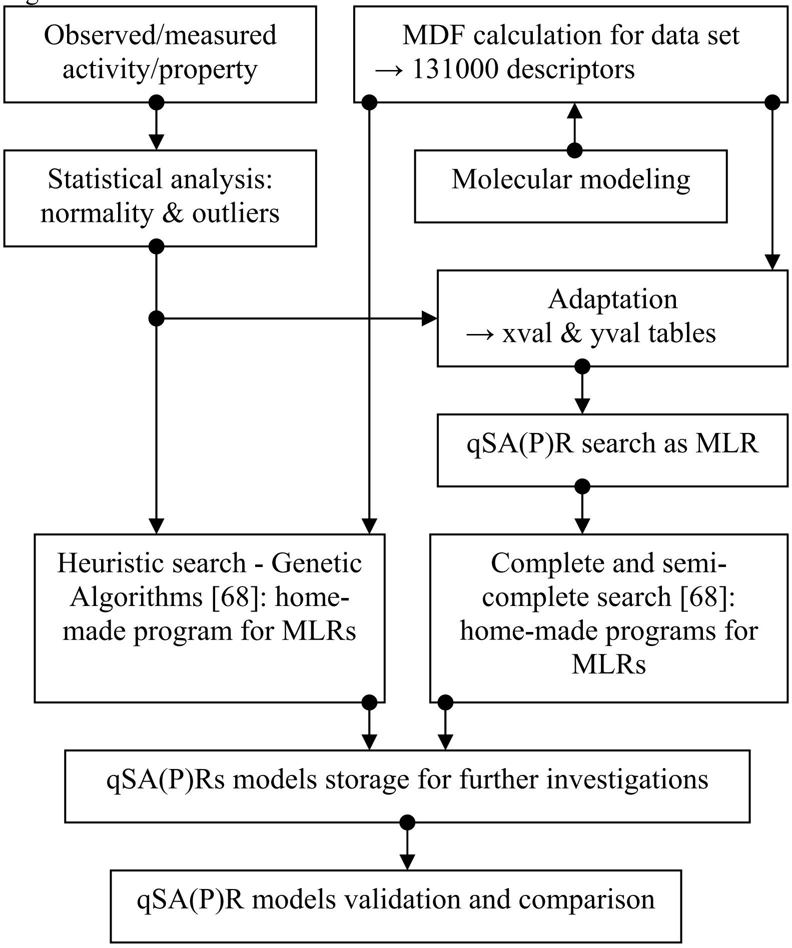
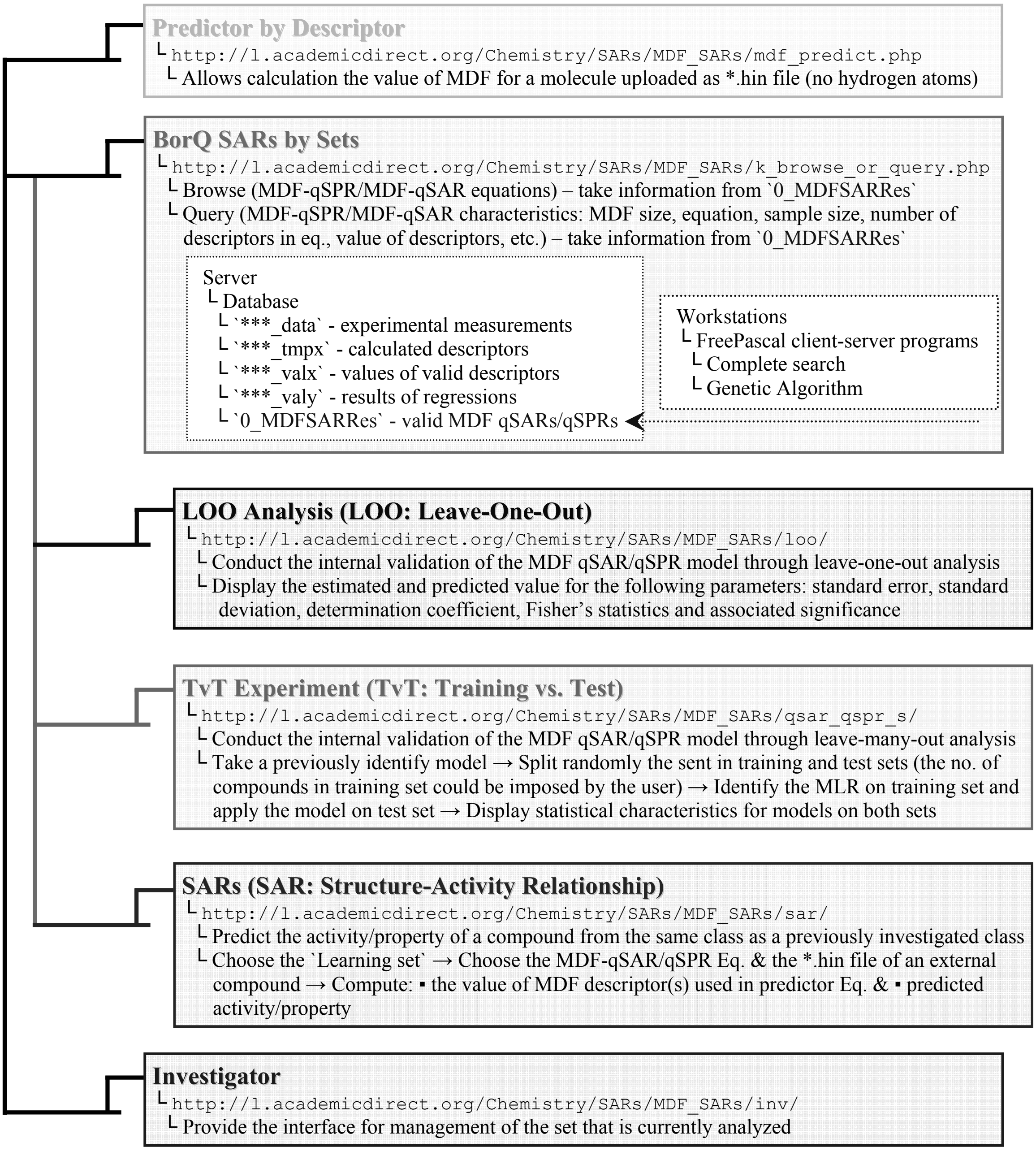

Observed/measured activity/property
MDF calculation for data set → 131000 descriptors
Molecular modeling
Statistical analysis: normality & outliers
Adaptation → xval & yval tables
qSA(P)R search as MLR
Heuristic search - Genetic Algorithms [68]: home-made program for MLRs
Complete and semi-complete search [68]: home-made programs for MLRs
qSA(P)Rs models storage for further investigations
qSA(P)R models validation and comparison

Figure 1. Process flow of the main steps involved on the MDF-based QSAR/QSPR modeling.



Predictor by Descriptor
└ http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/mdf_predict.php
└ Allows calculation the value of MDF for a molecule uploaded as *.hin file (no hydrogen atoms)

BorQ SARs by Sets
└ http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/k_browse_or_query.php
└ Browse (MDF-qSPR/MDF-qSAR equations) – take information from `0_MDFSARRes`
└ Query (MDF-qSPR/MDF-qSAR characteristics: MDF size, equation, sample size, number of descriptors in eq., value of descriptors, etc.) – take information from `0_MDFSARRes`

Server
└ Database
└ `***_data` - experimental measurements
└ `***_tmpx` - calculated descriptors
└ `***_valx` - values of valid descriptors
└ `***_valy` - results of regressions
└ `0_MDFSARRes` - valid MDF qSARs/qSPRs

Workstations
└ FreePascal client-server programs
└ Complete search
└ Genetic Algorithm

LOO Analysis (LOO: Leave-One-Out)
└ http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/loo/
└ Conduct the internal validation of the MDF qSAR/qSPR model through leave-one-out analysis
└ Display the estimated and predicted value for the following parameters: standard error, standard deviation, determination coefficient, Fisher's statistics and associated significance

TvT Experiment (TvT: Training vs. Test)
└ http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/qsar_qspr_s/
└ Conduct the internal validation of the MDF qSAR/qSPR model through leave-many-out analysis
└ Take a previously identify model → Split randomly the sent in training and test sets (the no. of compounds in training set could be imposed by the user) → Identify the MLR on training set and apply the model on test set → Display statistical characteristics for models on both sets

SARs (SAR: Structure-Activity Relationship)
└ http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/sar/
└ Predict the activity/property of a compound from the same class as a previously investigated class
└ Choose the `Learning set` → Choose the MDF-qSAR/qSPR Eq. & the *.hin file of an external compound → Compute: ▪ the value of MDF descriptor(s) used in predictor Eq. & ▪ predicted activity/property

Investigator
└ http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/inv/
└ Provide the interface for management of the set that is currently analyzed

Figure 2. The main data-interfaces of MDF tools.

Tables and captions

Table **1**. Most important characteristics of MDF QSAR/QSPR models.

Set [ref]	**n**	**k**	$\mathbf{r^2}$	$\mathbf{r^2_{adj}}$	**SErr**	**F (p)**	$\mathbf{r^2_{loo}}$	**AICc**	**AICR^2**	**AICu**	**BIC**	**APC**	**HQC**	**FIT**
19654 [90]	23	4	0.9979	0.9974	0.06	$2089\ (9.78\cdot10^{-24})$	0.9840	-122	1	-4	-115	0.004	-125	164
22583 [91]	57	5	0.9175	0.9094	0.45	$113\ (2.14\cdot10^{-26})$	0.8997	-83	5	0	-66	0.226	-80	6
23110 [92]	69	3	0.9011	0.8965	0.24	$197\ (1.38\cdot10^{-32})$	0.8904	-190	1	-2	-177	0.063	-187	7
23159 [63]	8	2	0.9505	0.7755	0.15	$58\ (3.54\cdot10^{-4})$	0.8989	-21	1	0	-23	0.037	-28	5
23167 [94]	31	3	0.9394	0.9327	0.15	$140\ (1.51\cdot10^{-16})$	0.9240	-114	2	-2	-105	0.024	-114	9
26449 [95]	10	2	0.9974	0.9966	0.03	$1330\ (9.25\cdot10^{-10})$	0.9948	-63	-2	-4	-63	0.001	-68	120
31572 [96]	24	4	0.9583	0.9495	0.14	$109\ (7.85\cdot10^{-13})$	0.9353	-87	4	-2	-79	0.024	-89	8
3300 [97]	34	5	0.9758	0.9715	0.21	$226\ (1.00\cdot10^{-21})$	0.9654	-99	5	-2	-86	0.050	-99	16
33504 [98]	73	2	0.9982	0.9981	1.75	$1.9\cdot10^{4}\ (9.38\cdot10^{-97})$	0.9980	85	-5	2	94	3.171	87	465
34121 [99]	76	4	0.7147	0.6986	0.62	$44\ (1.21\cdot10^{-18})$	0.6811	-68	4	0	-52	0.405	-64	2
36638 [100]	16	3	0.9950	0.9938	0.03	$799\ (4.47\cdot10^{-14})$	0.9812	-108	0	-5	-104	0.001	-112	69
408461 [101]	40	4	0.9175	0.9081	0.16	$97\ (1.84\cdot10^{-18})$	0.8911	-139	4	-2	-127	0.030	-138	6
408462 [101]	40	4	0.9037	0.8927	0.17	$82\ (2.74\cdot10^{-17})$	0.8804	-135	4	-2	-123	0.033	-134	5
408464 [101]	40	4	0.9202	0.9111	0.16	$101\ (1.04\cdot10^{-18})$	0.9034	-140	4	-2	-128	0.029	-139	6
41521 [102]	8	2	0.9987	0.9981	0.05	$1889\ (6.35\cdot10^{-8})$	0.9981	-38	-3	-2	-40	0.004	-46	178
52344 [103]	8	2	0.9998	0.9997	0.01	$1.3\cdot10^{4}\ (5.49\cdot10^{-10})$	0.9994	-65	-5	-5	-67	0.000	-72	1191
52730 [104]	10	2	0.9976	0.9970	0.06	$1473\ (6.49\cdot10^{-10})$	0.9959	-51	-2	-3	-51	0.004	-56	133
a_acids [105]	12	2	0.9871	0.9842	0.27	$344\ (3.15\cdot10^{-9})$	0.9777	-26	-1	0	-24	0.090	-30	29
cqdmdfv [106]	37	5	0.9368	0.9266	0.17	$92\ (1.22\cdot10^{-17})$	0.9015	-122	6	-2	-109	0.034	-121	6
DevMTOp00 [107]	8	2	1.0000	1.0000	12.53	$9.1\cdot10^{5}\ (1.26\cdot10^{-14})$	1.0000	49	-9	9	47	216	41	85370
DevMTOp03 [107]	8	2	0.9992	0.9988	0.04	$2946\ (2.09\cdot10^{-8})$	0.9977	-44	-3	-3	-46	0.002	-52	277
DevMTOp04 [107]	8	2	0.9998	0.9997	0.02	$1.1\cdot10^{4}\ (7.21\cdot10^{-10})$	0.9985	-57	-4	-4	-59	0.000	-64	1067
DevMTOp05 [107]	8	2	0.9992	0.9989	0.03	$3134\ (1.79\cdot10^{-8})$	0.9985	-46	-3	-3	-48	0.002	-54	295
DevMTOp07 [107]	8	2	0.9996	0.9994	0.03	$6314\ (3.12\cdot10^{-9})$	0.9991	-50	-4	-4	-52	0.001	-57	594
DevMTOp12 [107]	8	2	0.9991	0.9987	0.04	$2644\ (2.74\cdot10^{-8})$	0.9980	-44	-3	-3	-46	0.002	-51	249
DevMTOp14 [107]	8	2	0.9996	0.9994	0.03	$5941\ (3.63\cdot10^{-9})$	0.9988	-49	-4	-3	-51	0.001	-56	559
DevMTOp16 [107]	8	2	0.9989	0.9985	0.03	$2374\ (3.59\cdot10^{-8})$	0.9975	-48	-3	-3	-50	0.001	-56	223
DevMTOp17 [107]	8	2	0.9995	0.9994	0.02	$5437\ (4.53\cdot10^{-9})$	0.9986	-58	-4	-5	-60	0.000	-66	512
DevMTOp20 [107]	8	2	0.9992	0.9988	0.04	$3013\ (1.98\cdot10^{-8})$	0.9983	-45	-3	-3	-47	0.002	-53	284
DevMTOp21 [107]	8	2	0.9990	0.9986	0.04	$2539\ (3.03\cdot10^{-8})$	0.9980	-44	-3	-3	-46	0.002	-52	239
DevMTOp23 [107]	8	2	0.9999	0.9998	0.01	$2.2\cdot10^{4}\ (1.39\cdot10^{-10})$	0.9997	-60	-5	-5	-62	0.000	-67	2063
DHFR [108]	67	4	0.9058	0.8997	0.19	$149\ (4.60\cdot10^{-31})$	0.8932	-215	3	-2	-200	0.040	-212	6
Dipeptides [14]	58	4	0.9036	0.8963	0.32	$124\ (3.02\cdot10^{-26})$	0.8831	-125	4	-1	-111	0.113	-123	6
IChr10 [109]	10	2	0.9992	0.9990	0.10	$4369\ (1.45\cdot10^{-11})$	0.9985	-40	-3	-2	-39	0.013	-45	394
JCCS2001 [110]	47	4	0.9403	0.9346	0.16	$165\ (4.06\cdot10^{-25})$	0.9238	-165	3	-2	-152	0.029	-163	9
MR10 [109]	10	2	1.0000	0.9999	0.07	$8.3\cdot10^{4}\ (4.87\cdot10^{-16})$	0.9999	-46	-6	-3	-45	0.007	-51	7490
PCB_lkow [111]	206	4	0.9168	0.9151	0.24	$554\ (2.75\cdot10^{-107})$	0.9093	-579	2	-2	-558	0.060	-573	10
PCB_rrf [111]	209	4	0.7367	0.7316	0.18	$143\ (5.77\cdot10^{-58})$	0.7169	-705	3	-2	-684	0.034	-699	2
PCB_rrt [111]	209	2	0.9972	0.9972	0.01	$3.7\cdot10^{4}\ (1.13\cdot10^{-263})$	0.9971	-1939	-5	-8	-1926	0.000	-1935	335
RRC_lbr [112]	30	4	0.9737	0.9695	0.12	$231\ (2.35\cdot10^{-19})$	0.9650	-118	3	-3	-108	0.018	-119	16
RRC_lkow [112]	30	4	0.9781	0.9745	0.17	$279\ (2.44\cdot10^{-20})$	0.9680	-98	3	-2	-88	0.035	-98	19
RRC_pka [112]	30	4	0.9638	0.9580	0.19	$166\ (1.27\cdot10^{-17})$	0.9474	-92	3	-2	-82	0.044	-92	12
Ta395 [113]	15	2	0.9766	0.9727	0.17	$250\ (1.65\cdot10^{-10})$	0.9614	-48	0	-2	-45	0.035	-51	19
Tox395 [113]	14	2	0.9568	0.9490	0.18	$122\ (3.11\cdot10^{-8})$	0.9343	-43	0	-2	-41	0.038	-46	10
Triazines [114]	30	4	0.9885	0.9867	0.08	$537\ (7.63\cdot10^{-24})$	0.9850	-144	2	-4	-134	0.008	-144	38

Abbreviations:
Set = abbreviation of the set; [ref] = reference for the experimental data source;
n = number of compounds in the set; k = number of independent variables in qSAR/qSPR equation;
r^2 = determination coefficient; r^2_{adj} = adjusted determination coefficient;
SErr = standard error of estimate; F = Fisher's statistics;
p = significance of Fisher's statistics; r^2_{loo} = determination coefficient in leave-one-out analysis;
AIC_c = corrected Akaike information criteria; AIC_{R2} = AIC based on the determination coefficient
AIC_u = McQuarrie and Tsai corrected AIC; BIC = Schwarz (or Bayesian) Information Criterion
APC = Amemiya Prediction Criterion; HQC = Hannan-Quinn Criterion; FIT = Kubinyi function

Table **2**. The characteristics of the MDF models according to descriptors.

Set (v[a])	**7th letter (freq)**	**6th letter (freq)**	**4th letter (freq)**	**1st letter (freq)**
19654 (4)	g(3)-t(1)	C(1)-E(1)-M(1)-Q(1)	D(1)-r(3)	i(3)-l(1)
22583 (5)	g(2)-t(3)	E(1)-H(1)-M(1)-Q(2)	D(2)-m(2)-r(1)	A(1)-i(2)-l(2)
23110 (3)	g(1)-t(2)	H(1)-Q(2)	m(1)-r(2)	A(1)-i(2)
23159 (2)	t(2)	E(1)-H(1)	m(1)-r(1)	i(1)-l(1)
23167 (3)	g(2)-t(1)	Q(3)	r(3)	i(2)-l(1)
26449 (2)	t(2)	G(1)-Q(1)	D(1)-r(1)	i(1)-l(1)
31572 (4)	g(1)-t(3)	H(1)-Q(3)	r(4)	i(3)-l(1)
3300 (5)	g(3)-t(2)	C(1)-H(1)-Q(3)	D(1)-m(1)-r(3)	i(3)-l(2)
33504 (2)	t(2)	G(1)-H(1)	r(2)	i(1)-l(1)
34121 (4)	g(4)	Q(4)	D(3)-r(1)	i(3)-l(1)
36638 (3)	g(2)-t(1)	H(1)-Q(2)	D(1)-m(2)	i(2)-l(1)
408461 (4)	g(3)-t(1)	M(1)-Q(3)	m(1)-r(3)	i(3)-l(1)
408462 (4)	g(4)	C(1)-G(1)-Q(2)	D(2)-r(2)	i(3)-l(1)
408464 (4)	g(3)-t(1)	Q(4)	D(2)-m(1)-r(1)	i(4)
41521 (2)	g(2)	H(1)-Q(1)	m(1)-r(1)	i(1)-l(1)
52344 (2)	g(1)-t(1)	G(1)-H(1)	m(1)-r(1)	i(1)-l(1)
52730 (2)	g(2)	M(1)-Q(1)	m(1)-r(1)	i(1)-l(1)
a_acids (2)	g(2)	G(1)-H(1)	D(1)-r(1)	i(2)
cqdmdfv (5)	g(5)	G(1)-H(2)-M(1)-Q(1)	D(1)-m(4)	i(5)
DevMTOp00 (2)	g(1)-t(1)	E(1)-Q(1)	m(1)-r(1)	i(1)-l(1)
DevMTOp03 (2)	g(2)	M(1)-Q(1)	r(2)	i(2)
DevMTOp04 (2)	g(2)	H(1)-Q(1)	D(1)-r(1)	i(1)-l(1)
DevMTOp05 (2)	g(2)	Q(2)	D(1)-m(1)	A(1)-i(1)
DevMTOp07 (2)	t(2)	C(1)-Q(1)	m(1)-r(1)	A(1)-i(1)
DevMTOp12 (2)	g(1)-t(1)	G(1)-M(1)	m(1)-r(1)	l(2)
DevMTOp14 (2)	g(1)-t(1)	E(1)-Q(1)	m(1)-r(1)	i(2)
DevMTOp16 (2)	g(1)-t(1)	H(1)-Q(1)	m(2)	A(2)-i(2)
DevMTOp17 (2)	g(2)	C(1)-Q(1)	D(1)-r(1)	i(2)
DevMTOp20 (2)	g(1)-t(1)	G(1)-M(1)	m(1)-r(1)	i(1)-l(1)
DevMTOp21 (2)	t(2)	E(2)	r(2)	i(2)
DevMTOp23 (2)	g(1)-t(1)	E(1)-G(1)	m(1)-r(1)	l(2)
DHFR (4)	g(2)-t(2)	H(2)-Q(2)	D(1)-m(1)-r(2)	i(2)-l(2)
Dipeptides (4)	g(1)-t(3)	E(2)-H(1)-M(1)	D(1)-m(2)-r(1)	i(4)
IChr10 (2)	g(1)-t(1)	M(1)-Q(1)	D(1)-m(1)	i(1)-l(1)
JCCS2001 (4)	g(4)	M(2)-Q(2)	D(2)-r(2)	i(3)-l(1)
MR10 (2)	t(2)	E(1)-M(1)	m(1)-r(1)	l(2)
PCB_lkow (4)	g(3)-t(1)	E(1)-G(1)-Q(2)	D(1)-m(2)-r(1)	A(2)-i(2)
PCB_rrf (4)	g(3)-t(1)	H(2)-Q(2)	D(1)-m(1)-r(2)	i(4)
PCB_rrt (2)	g(1)-t(1)	H(2)	m(1)-r(1)	i(1)-l(1)
RRC_lbr (4)	g(3)-t(1)	Q(4)	D(1)-m(2)-r(1)	A(1)-i(1)-l(2)
RRC_lkow (4)	g(4)	G(1)-Q(3)	D(2)-m(1)-r(1)	i(2)-l(2)
RRC_pka (4)	g(4)	H(1)-Q(3)	m(3)-r(1)	A(2)-i(1)-l(1)
Ta395 (2)	t(2)	M(1)-Q(1)	m(1)-r(1)	i(1)-l(1)
Tox395 (2)	g(2)	Q(2)	r(2)	A(1)-l(1)
Triazines (4)	g(2)-t(2)	H(1)-Q(1)	m(3)-r(1)	i(3)-l(1)

Abbreviations:

[a] v = number of descriptors in the model; freq = absolute frequency;

7th letter - interaction via: *g* = geometric distance, *t* = topological distance.

6th letter - dominant atomic property: *M* = relative atomic mass,
Q = atomic partial charge, semi-empirical Extended Hückel model, single point approach, *C* = cardinality, *E* = atomic electronegativity, *G* = group electronegativity, *H* = number of hydrogen atoms adjacent to the investigated atom.

4th letter - overlapping interaction: *M*, *m* = frequent and distant interactions, *D*, *d* = frequent and closed interactions, *R*, *r* = sporadic and distant interactions.

1st letter - structure on activity scale: *A* = absolute value, *i* = inverse of identity, *l* = logarithm of absolute value.