View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Archivio della Ricerca - Università di Pisa

Modeling of the acute toxicity of benzene derivatives by complementary QSAR methods

Carlo Bertinetto, Celia Duce, Roberto Solaro, Maria Rosaria Tiné*,
Department of Chemistry and Industrial Chemistry, University of Pisa, Pisa, Italy
E-mail: mrt@dcci.unipi.it

Alessio Micheli,
Department of Computer Science, University of Pisa, Pisa, Italy

Károly Héberger, Ante Miličević,
The Institute of Medical Research and Occupational Health, Zagreb, Croatia

Sonja Nikolić*
The Rugjer Bošković Institute, Bijenička 54, HR-10000 Zagreb, Croatia
E-mail: sonja@irb.hr

(Received May 15, 2012)

Abstract

A data set containing acute toxicity values (96-h LC_{50}) of 69 substituted benzenes for fathead minnow (*Pimephales promelas*) was investigated with two Quantitative Structure-Activity Relationship (QSAR) models, either using or not using molecular descriptors, respectively. Recursive Neural Networks (RNN) derive a QSAR by direct treatment of the molecular structure, described through an appropriate graphical tool (variable-size labeled rooted ordered trees) by defining suitable representation rules. The input trees are encoded by an adaptive process able to learn, by tuning its free parameters, from a given set of structure-activity training examples. Owing to the use of a flexible encoding approach, the model is target invariant and does not need *a priori* definition of molecular descriptors. The results obtained in this study were analyzed together with those of a model based on molecular descriptors, i.e. a Multiple Linear Regression (MLR) model using CROatian MultiRegression selection of descriptors (CROMRsel). The comparison revealed interesting similarities that could lead to the development of a combined approach, exploiting the complementary characteristics of the two approaches.

1. Introduction

Evaluation of the risk posed by the multitude of chemicals produced every year by industry and agriculture is a complicated task. Proper evaluation of the risk is of increasing importance. It is not practically and economically feasible to conduct toxicity tests on all substances released into the environment. Therefore, experimental measurements need to be integrated with theoretical predictive methods that can fill gaps in the data and identify those compounds that are most promising for empirical assessment. Many predictive methods correlate the toxicity to other, simpler, physico-chemical property and biological activity [1]. However, several classes of property/activity data are unavailable for many substances. It is estimated that roughly half of chemicals do not have any experimental data at al. [2]. Therefore, increasing preference is being given to methods that correlate the investigated property/activity (or *target* property/activity) with representation(s) of the molecular structure alone.

Quantitative Structure–Activity Relationships (QSARs) are widely recognized as scientifically credible tools for the prediction of acute toxicity [2-4]. The basic aim of QSAR is to find a function F that relates the appropriate representation of the molecular structure to the target activity. In more detail, F can be decomposed into an encoding or feature representation function f and a mapping function g. The choice of functions f and g is what discriminates among the different approaches, with most differences arising from the f function. Standard QSAR approaches employ structural molecular descriptors or calculated molecular activity to encode the molecules (f function), while the output value is computed through either linear or nonlinear regression models (g function).

In order to make adequate assessment of the quality of a QSAR model and to obtain useful predictions, it is of fundamental importance to use accurate empirical data [5]. The MED-Duluth Database [6] provides toxicity data for more than 750 assays on over 600 compounds towards the freshwater fish fathead minnow (*Pimephales promelas*). The measured activity is the Lethal Concentration for 50% of the tested sample after 96 hours of exposure (96h-LC_{50}). These data are considered highly reliable by regulatory authorities such as the USA Environmental Protection Agency, both because of their experimental accuracy and their significance in the evaluation of acute and chronic toxicity in vertebrate animals and in aquatic environments [7]. Among the classes of compounds included in the MED-Duluth database, benzene derivatives received much attention in previous QSAR studies because of their widespread use in the chemical and pharmaceutical industry [8]. In 1984 Hall *et al.* [9]

derived group contributions to the LC_{50} of a homogeneous set of 66 substituted benzenes, 26 of which were experimentally determined by the authors themselves. They found a decreasing contribution to toxicity in the order Cl > Br > NO_2 > CH_3 > OCH_3 > NH_2 > OH. The additivity model obtained from those contributions fitted the whole data set with a square correlation coefficient, R^2, of 0.904 and a standard error of estimate, S, of 0.25. Approximately the same dataset (69 compounds) was later investigated by Basak et al. [10] who built MLR models using the CROMRsel procedure for descriptor selection. The descriptors included in their best model were P_9 (path of length nine), ${}^2\chi^v$, ${}^4\chi^v$ (valence path connectivity indices of order two and four, respectively), ${}^6\chi^v{}_{Pc}$ (valence path-cluster connectivity index of order six), E_{lumo} (energy of the lowest unoccupied molecular orbital), μ (dipole moment) and ${}^{3D}W_H$ (3-D Wiener number for the hydrogen-filled structures computed using geometric distance matrices). This model showed R^2 = 0.884, S = 0.26 for the fitting of the whole data set and $R_{CV}{}^2$ = 0.856, S_{CV} = 0.29 for the leave-one-out cross-validation. Toropov and Toropova [11] calculated descriptors for the 69 benzenoids in the presence of correlation weights in the molecular graph and different values of third-order Morgan extended connectivity. Their model yielded R^2 = 0.898, S = 0.25 for the training set of 44 compounds and R^2 = 0.918, S = 0.23 for the test set of 25 compounds. Perez-Gonzalez *et al.* [8] used TOPological Sub-structural MOlecular DEsign (TOPS-MODE), based on the calculation of the spectral moments of the bond matrix. The obtained model was function of $\mu_5{}^{dip}$ (fifth spectral moment weights with dipole moment), $\mu^*\mu_1{}^{hyb}$ (square of first spectral moment weights with hydrophobicity) and $\mu_1{}^{dist}$ (first spectral moment weights with atomic distance). It showed R^2 = 0.888, S = 0.25 for the training set of 50 molecules and R^2 = 0.908, S = 0.28 for the test set of 19 molecules.

In the present work we have re-investigated this data set of 69 benzenoids with the Recursive Neural Network (RNN) model, developed in the last years by the Department of Computer Science and the Department of Chemistry & Industrial Chemistry of the University of Pisa for QSPR/QSAR analysis [12-15]. This approach differs radically from standard methods: it automatically learns the f and g functions and it treats a variable-size structured representation of molecules instead of numerical descriptors directly. Its main advantages are generality and adaptability, as it can be applied with little or no modification to different classes of compounds and target properties. In particular, it is not necessary to calculate and select a new set of descriptors each time a new property or compound type is investigated [16]. This characteristic qualifies the method as *target invariant*. Its previous applications successfully predicted the boiling points of linear and branched alkanes [12, 13], the

pharmacological activity of series of substituted benzodiazepines [12-14] and 8-azaadenine derivates [15], the free energy of solvation of mono- and poly-functional organic compounds [16, 17], the glass transition temperature of (meth)acrylic polymers and copolymers [18-24] and the melting point of pyridinium bromides [18, 25]. This method is particularly suitable for tasks in which no background knowledge is available *a priori* because the molecular representation retains all structural information whereas the RNN automatically learns the correlation functions. On the other hand, the input data are of much higher complexity than in standard approaches (usually vectors of less than 10 descriptors); therefore a greater number of training examples are needed to build an accurate relationship. Moreover, due to the recursive and non-linear nature of these relationships, their physical interpretation is more difficult than a direct observation of known physicochemical descriptors.

The results obtained in this study were analyzed together with those of Multiple Linear Regression (MLR) models obtained through the **CRO**atian **M**ulti**R**egression **sel**ection of descriptors (CROMRsel) method developed at the Ruđer Bošković Institute of Zagreb [26]. In addition to making a performance comparison, the purpose of this research was to find a starting point from which to develop a combined approach, exploiting the complementary characteristics of the two methods.

2. Methods

The RNN model is explained in detail elsewhere [12-16]. We here briefly summarize its main characteristics. The RNNs are an extension of standard neural networks able to directly deal with labeled hierarchical structured representation of molecules, in particular in the form of rooted trees, a subclass of DPAGs (Directed Positional Acyclic Graphs). Trees have variable size and give a richer and more flexible vehicle of information than the flat vectors of descriptors employed in traditional QSAR approaches. Moreover, RNNs can adaptively encode the input structures by learning from the given structure-activity training examples. To this end, the RNN recursively encodes each structure through a bottom-up approach that dynamically mimics its morphology. For each vertex of the input structure, the model computes a numerical code by using information of both the vertex label and, recursively, the code of the sub-graphs descending from the current vertex. The process returns a code for the whole molecular structure, as depicted in Fig. 1. This code is then mapped to the output activity value. The learning algorithm allows the model to tune the free parameters of the neural network functions on the basis of the training examples and by this

process the RNN models find a direct and adaptive relationship between molecular structures and target properties/activities.

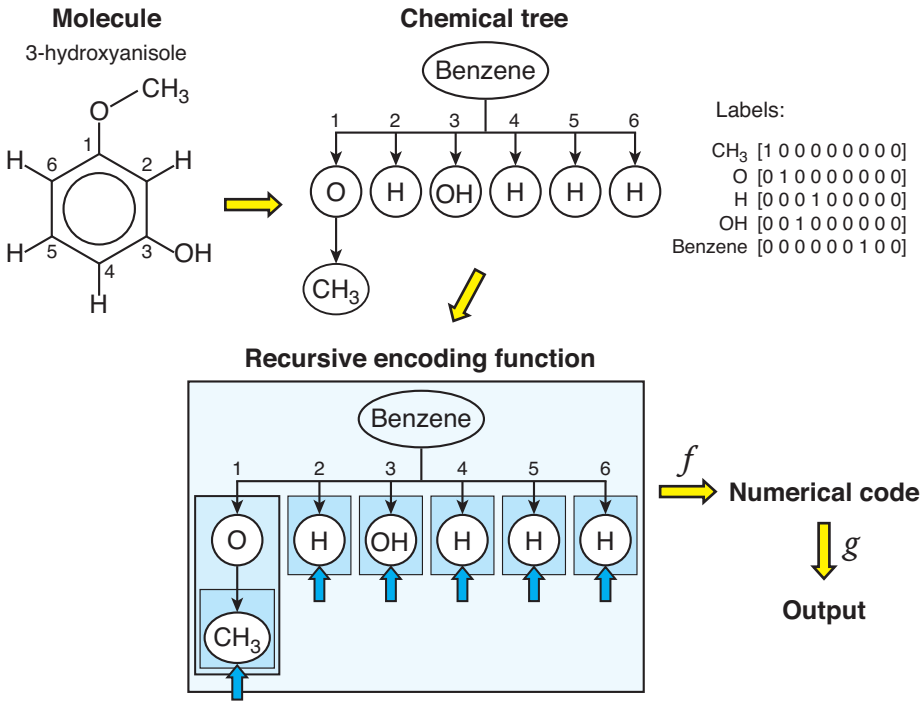

Figure 1: Tree representation and encoding of 3-hydroxyanisole. The first child of *Benzene* is the oxygen atom linked to a methyl group, because it has higher priority than *OH* and *H*. The ordering is, according to the drawing of the molecule, clockwise in order to assign a lower position (3) to the next group with highest priority (*OH*). *f* and *g* indicate the encoding and mapping functions, respectively.

Every chemical compound is represented as a labeled rooted ordered tree by a 2-D graph that could easily be obtained from its structural formula. The molecule is fragmented into defined atomic groups: each group corresponds to a vertex of the tree and each bond between them corresponds to an edge, see Fig. 1. An appropriate set of rules is defined in order to have a unique correspondence between each molecule and its chemical tree. Each vertex is assigned a label, which is a tuple of variables categorically distinguishing the symbol of the atomic group. Despite being conventionally defined, a label can convey chemical information through orthogonality or similarity to other labels. In the current study the following groups were used: *Benzene*, NO_2, NH_2, *O*, *OH*, *Cl*, *Br*, CH_3 and *H*. They were rated according to a priority scale [16], which corresponds to the order in which they are written, that was used to determine the tree root and the total order on each vertex's subtree. In this work the tree root was always placed on the *Benzene*, as shown in Fig. 1. *Benzene* has 6

children, ordered according to their position on the ring. The first child is the one with highest priority and the direction of the ordering (clockwise or counter-clockwise) is the one that assigns the lowest position to the next fragment with highest priority. This kind of structure-based representation is general and able to represent any sort of chemical compound [16,21]. Its flexibility also allows for choosing the most suitable representation for each data set and predictive task at hand [18].

In the MLR approach, the regression equation expressing the QSAR model is in the form of a linear combination of descriptors, with their coefficients determined by the least-squares method. The MLR models were developed using CROMRsel [26] for a more efficient stepwise (*one-by-one*) model selection. The best models are selected in the orthogonal basis in order to have a simpler and faster procedure, which also allows for taking into account the many higher-order and cross-product terms; this point is well illustrated in ref. 26. The algorithm on which the computer program is based can be presented as follows [27].

1. Initial data: the set of N descriptors and target activity values.
2. Two descriptors are selected in unbiased fashion from the initial set of descriptors. They are orthogonalized, then the correlation coefficient between the activity *A* and each individual orthogonalized descriptor is computed. From this, individual computations of the total correlation coefficient, which is equal to the correlation coefficient between the experimental activity values *A* and its computed values *A*', are obtained. The same procedure is repeated for the selected *n* descriptors, where $n \leq N$ and $n < m$ (*m* = the number of molecules). The value of *n* is fixed at the beginning of the computation. The *n* value is the size of multiregression model, i.e., the number of descriptors we want to have in the model.
3. For every *I*-tuple of descriptors ($I = 2, 3, ..., n$), the combination with the highest *R* is singled out, and this combination necessarily possesses the smallest value of *S* among all possibilities generated by use of an *I*-tuple of the same class (there are exactly $\binom{N}{I}$ of them). In such a way, the *n* best *I*-tuples are obtained.
4. Among the *n* best *I*-tuples ($I = 2, 3, ..., n$), the one that gives the smallest *S*, which is at the same time the best total solution, is selected. This solves the problem of selecting the optimum number of descriptors and detecting the optimum *I*-tuple of descriptors to produce the best estimation of the activity *A*.
5. Finally, statistical parameters, for training and prediction on an independent test set are calculated for the selected best model.

3. Results and discussion

The data set was taken from Hall [9] and consists of the 96h- LC_{50} (concentration that kills 50% of the tested sample after 96 hours of exposure) of 69 substituted benzenes towards the freshwater fish fathead minnow (*Pimephales promelas*). The toxicity is expressed as -log(LC_{50}), with the concentration measured in mol/L; the values range from 3.04 to 6.37. These data represent a relatively small and homogeneous subset of the MED-Duluth Database and therefore provide a mainly local predictive problem. The total data set of 69 compounds was split into a training set and an external test set of 51 and 18 molecules, respectively. A few of the included molecular structures are depicted in Fig. 2.

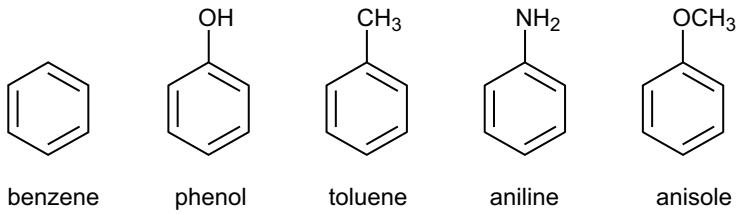

Figure 2: Structures of benzene and the simplest benzenoids used in the current study.

The random initialization of the RNN connection weights can lead to different outcomes because of the use of a stochastic gradient–based technique to solve a least-mean square problem. In order to have a significant result, in each of our experiments sixteen trials were carried out for the RNN simulation and the results were averaged over the different trials. Learning was terminated when the maximum error for each compound of the training set was below a preset threshold value, which was set at 0.6 units of -log(LC_{50}) ([LC_{50}] = mol/L). This value was determined on the basis of the standard deviation in the experimental measurements performed by Hall, which is reported as 0.15 [9]. The chosen training threshold corresponds to four times the standard deviation value and, according to Gaussian statistics, encompasses 94.5% of the total variance. The average statistics on the results are reported in Table 1, whereas Tables 2 and 3 list the detailed outcomes for each compound of the training and test set, respectively.

A scatter plot of the experimental -log(LC_{50}) *vs*. calculated values obtained by the RNN method is provided in Fig. 3.

Moreover, we carried out a prediction on the same data set by using Multiple Linear Regression (MLR) analysis. In particular, we used CROMRsel method [10, 26] for the selection of the descriptors that yield the best model. The descriptors were calculated with

DRAGON 5.0 software and filtered by eliminating constants, near constants and highly intercorrelated descriptors, i.e. with a correlation coefficient greater than 0.95.

	Training set 51 molecules		Test set 18 molecules	
	RNN	MLR	RNN	MLR
MAR[a]	0.19	0.21	0.22	0.21
Max[b]	0.51	0.65	0.45	0.58
S[c]	0.24	0.27	0.25	0.27
R^2 [d]	0.910	0.886	0.821	0.806

[a] Mean Absolute Residual (log units). [b] Maximum absolute residual (log units). [c] Standard error of estimate/prediction (log units). [d] Square correlation coefficient.

Table 1: Experimental results for the RNN and MLR methods.

The resulting pool of 169 descriptors included constitutional, topological, geometrical and charge descriptors, walk/path and functional group counts, connectivity, information, edge adjacency and topological charge indices, atom-centered fragments and structure-calculated molecular properties. CROMRsel [26] was applied to this 169-descriptors set to build and select models, allowing a maximum of 4 descriptors per model. The one that gave the best performance, with respect to the correlation coefficient, on the training set is:

$$-\log(LC_{50}) = 2.75(\pm0.16) + 0.31(\pm0.04)\mathbf{nCL} + 2.14(\pm0.37)\mathbf{X5v} + 0.32(\pm0.04)\mathbf{MPC09} + 4.27(\pm 0.50)\mathbf{DP18} \quad \mathbf{(1)}$$

where **nCL** is the number of chlorine atoms, **X5v** is the valence connectivity index ${}^5\chi^v$, **DP18** is the molecular profile no. 18 and **MPC09** is the molecular path count of order 9 [28-31].

The average results yielded by MLR for training and test sets are reported in Table 1, while the detailed outcomes are shown in Table 2 and Table 3. Comparison of experimental data *vs.* values calculated by equation (1) is plotted in Fig. 4.

The RNN and MLR methods gave very similar results on this data set, and comparable to those obtained in the reported literature. The overall statistical parameters in Table 1 show approximately the same values for both methods, though indicating slightly better performance by RNN. It was not obvious *a priori* that RNN should give better performance on this small and homogeneous data set, because RNN, as explained in the introduction, is best suited for general, non-local problems.

Molecule name	Experimental -log(LC_{50})	Calculated -log(LC_{50}) **RNN**	Calculated -log(LC_{50}) **MLR**
Benzene	3.40	3.33	3.23
Bromobenzene	3.89	3.56	3.63
Chlorobenzene	3.77	3.86	3.74
Phenol	3.51	3.45	3.27
Toluene	3.32	3.56	3.40
1,2-dichlorobenzene	4.40	4.37	4.23
1,3-dichlorobenzene	4.30	4.19	4.19
2-chlorophenol	4.02	3.96	3.77
3-chlorotoluene	3.84	3.86	3.86
4-chlorotoluene	4.33	4.10	4.28
1,3-dihydroxybenzene	3.04	3.41	3.29
3-hydroxyanisole	3.21	3.32	3.51
3-methylphenol	3.29	3.45	3.41
4-methylphenol	3.58	3.69	3.57
4-nitrophenol	3.36	3.72	3.56
1,4-dimethoxybenzene	3.07	3.54	3.70
1,4-dimethylbenzene	4.21	3.80	3.89
2-nitrotoluene	3.57	3.67	3.57
3-nitrotoluene	3.63	3.53	3.68
1,2-dinitrobenzene	5.45	4.94	4.95
1,4-dinitrobenzene	5.22	4.91	5.07
2-methyl-3-nitroaniline	3.48	3.64	3.66
2-methyl-4-nitroaniline	3.24	3.71	3.85
3-methyl-6-nitroaniline	3.80	3.88	3.84
4-methyl-2-nitroaniline	3.79	3.98	3.89
4-hydroxy-3-nitroaniline	3.65	3.72	3.63
4-methyl-3-nitroaniline	3.77	3.84	3.85
1,2,4-trichlorobenzene	5.00	5.02	5.08
1,3,5-trichlorobenzene	4.74	4.92	4.60
3,4-dichlorotoluene	4.74	4.62	4.68
2,4-dichlorotoluene	4.54	4.43	4.68
4-chloro-3-methylphenol	4.27	4.09	4.03
2,4-dimethylphenol	3.86	3.82	3.67
2,6-dimethylphenol	3.75	3.90	3.54
2,3-dinitrotoluene	5.01	4.93	5.17
2,4-dinitrotoluene	3.75	4.22	4.40
2,5-dinitrotoluene	5.15	5.05	5.20
2,6-dinitrotoluene	3.99	4.11	4.12
1,3,5-trinitrobenzene	5.29	4.97	4.90
2-methyl-3,6-dinitroaniline	5.34	5.02	5.18
5-methyl-2,4-dinitroaniline	4.92	4.49	4.42
4-methyl-2,6-dinitroaniline	4.21	4.43	4.59
5-methyl-2,6-dinitroaniline	4.18	4.31	4.39
4-methyl-3,5-dinitroaniline	4.46	4.46	4.41
2,4,6-tribromophenol	4.70	4.67	4.36
1,2,3,4-tetrachlorobenzene	5.43	5.35	5.46
2,4,6-trichlorophenol	4.33	4.77	4.75
2-methyl-4,6-dinitrophenol	5.00	4.52	4.42
2,3,6-trinitrotoluene	6.37	6.31	6.71
2,3,4,5-tetrachlorophenol	5.72	5.67	5.52
2,3,4,5,6-pentachlorophenol	6.06	5.84	6.18

Table 2: Detailed outputs of RNN and MLR experiments for the training set.

Molecule name	Experimental -log(LC_{50})	Calculated -log(LC_{50})	Calculated -log(LC_{50})
		RNN	**MLR**
1,4-dichlorobenzene	4.62	4.50	4.68
2-methylphenol	3.77	3.57	3.43
1,2-dimethylbenzene	3.48	3.69	3.55
4-nitrotoluene	3.76	3.77	3.83
1,3-dinitrobenzene	4.38	3.99	4.08
2-methyl-5-nitroaniline	3.35	3.75	3.93
2-methyl-6-nitroaniline	3.80	3.63	3.72
1,2,3-trichlorobenzene	4.89	4.70	4.65
2,4-dichlorophenol	4.30	4.62	4.36
3,4-dimethylphenol	3.90	3.68	3.68
2,4-dinitrophenol	4.04	4.17	4.12
1,2,4-trimethylbenzene	4.21	3.93	3.97
3,4-dinitrotoluene	5.08	5.17	5.42
3,5-dinitrotoluene	3.91	4.34	4.39
2-methyl-3,5-dinitroaniline	4.12	4.57	4.49
3-methyl-2,4-dinitroaniline	4.26	4.28	4.17
1,2,4,5-tetrachlorobenzene	5.85	5.74	5.85
2,4,6-trinitrotoluene	4.88	5.10	5.08

Table 3: Detailed outputs of RNN and MLR experiments for the test set.

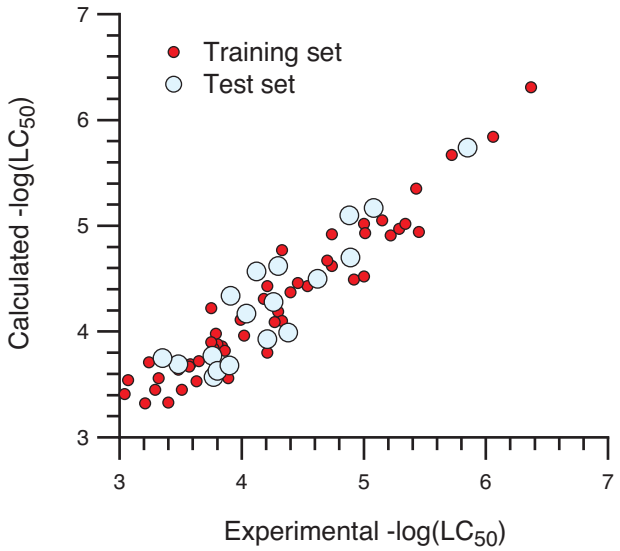

Figure 3: Plot of the experimental *vs*. calculated -log(LC_{50}) obtained with the RNN method.

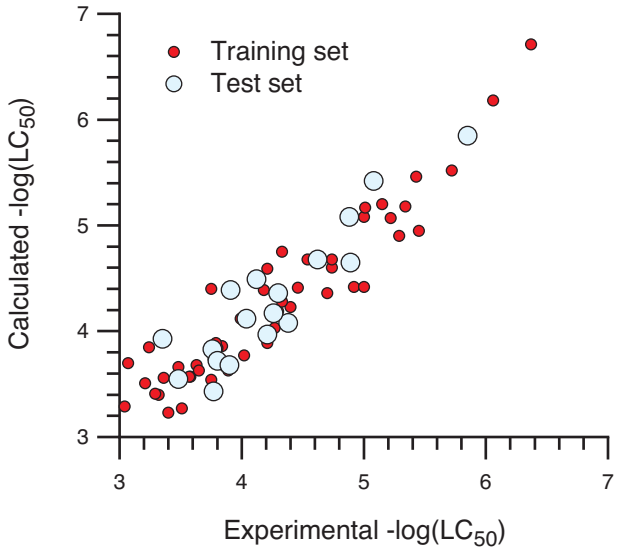

Figure 4: Plot of the experimental *vs*. calculated -log(LC_{50}) obtained with the MLR method.

Observation of the detailed outcomes in Tables 2 and 3 reveals a great similarity between the two methods, even at the level of the outcomes of individual molecules. Indeed, the average absolute difference between the output for each molecule by RNN and MLR is 0.13 log units for the training set and 0.11 for the test set. These values are considerably smaller than the mean average residuals reported in Table 1. Only 12 molecules in the training test and 2 in the test set show residuals of opposite sign between RNN and MLR. These observations could suggest that the encoding procedures of the two methods, although using very different procedures (encoding by learning from examples versus selection of predefined features) arrive at approximately the same interpretation of the molecular structure. Although this hypothesis needs further investigation, it could provide a key to better understand the internal representation of RNN methodology. As mentioned in the introduction, the physical interpretation of the results is relatively easy in descriptor-based methods but problematic in structure-based NN approaches. On the other hand, RNN does not use descriptors and therefore, by construction, does not require fixing *a priori* their number, type and selection [12, 16]: the encoding of molecules into numerical data for the QSAR modeling is generated by learning the direct map between molecular structures and target property values in the data set.

3.1 Validation

An additional validation tool became available only recently [32, 33]. The sum of absolute values for ranking differences (SRDs) between ‘reference’ and actual rankings (experimental, RNN- and MLR-predicted values) will show which calculation method is

better, and whether they are superior to the measured values. If the SRD values are smaller then the model is better. The average of all three methods has been accepted as 'reference' ranking. The ordering by sum of ranking differences is compared with simulated random numbers. A set of 250 vectors was generated with uniform discrete distribution between 0 and 1 for 51 objects (training set) and for 18 objects (test set).

Results for both training and test sets are reported in Table 4. The last 5 entries in the table refer to simulated random numbers, where XX1 is the first vigintile (5%), Q1 is the first quartile (25%), Med is the median, Q3 is the last quartile (75%) and XX19 is the last vigintile (95%). The results are shown in Fig. 5 for the 51 compounds of the training set. The SRD values are scaled between 0 and 100 and plotted on the X-axis. The Gaussian fit (Mean = 66.67, StD = 6.03) of the discrete distribution of SRD values for random numbers is reported in Fig. 5 as well.

	Ranking SRDs							
	RNN	Exp	MLR	XX1	Q1	Med	Q3	XX19
Training set	86	102	122	742	806	866	920	990
Test set	18	18	26	79	95	108	120	134

Table 4: Non-scaled sum of ranking differences for training, 51 molecules, and test sets, 18 molecules.

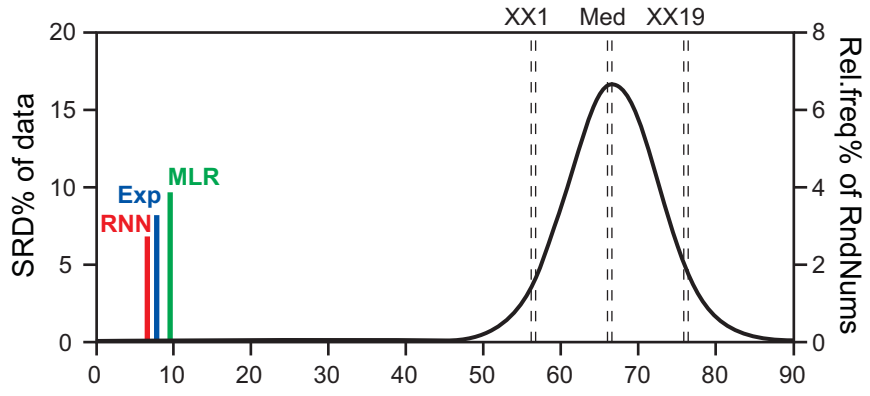

Figure 5: Scaled SRD values (between 0 and 100) using average as reference (X and left Y axes) for the training set of 51 compounds. The continuous line is the Gaussian fit (Mean = 66.67, StD = 6.03) of the discrete distribution of SRD values for random numbers. XX1 is the 5% percentile, Med is the median, XX19 is the 95% percentile for the discrete distribution (right Y axis).

The RNN provides the best representation for the training set, the experimental values are somewhat worse and the multilinear approximation is even worse. The experimental SRDs are far away from the generated (random) values, which is reassuring: the probability that the real SRDs derive from a random sequence is negligible.

The smaller SRD values evaluated for the test set (Table 4) reflect the smaller number of compounds involved as compared to the training set. Rescaled SRD values for the test set can be seen in Fig. 6.

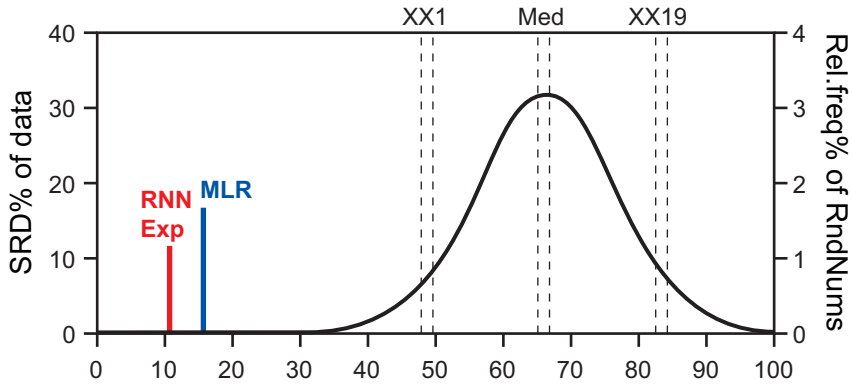

Figure 6: Scaled SRD values (between 0 and 100) using average as reference (X and left Y axes) for the test set of 18, compounds. The continuous line is the Gaussian fit (Mean = 66.73, StD = 10.39) of the discrete distribution of SRD values for random numbers. XX1 is the 5% percentile, Med is the median, XX19 is the 95% percentile for the discrete distribution (right Y-axis).

The real and random SRD are closer for the test set than for the training set, although the probability of this being a random sequence is still negligible. The empirical random distribution is well approximated by the normal distribution. By chance, for the test set the calculated values are better on this criterion than the experimental ones.

4. Conclusions

We investigated a data set containing LC_{50} values for 69 benzene derivatives to obtain structure-activity relationships by applying two distinct methods: Recursive Neural Networks and Multiple Linear Regression. These methods are radically different in procedure and capabilities and many of their aspects are complementary. The RNN uses a more general chemical representation, in the form of labeled trees, which does not need any background knowledge of the specific problem. On the other hand, MLR provides simpler and physically understandable relationships between the property and selected molecular descriptors. Both

methods provided good results as compared to other studies available in the literature. This is particularly satisfactory for the RNN method, although this method is in general best suited for non-local tasks, where the mechanism of action(s) may not depend on few known features as in the present task. Moreover, it is also satisfactory from a statistical learning point of view that RNN, which is potentially more complex than the MLR approach using only 4 descriptors, could achieve a comparable predictive performance even on a small data set.

Despite the differences between the methods, their outcomes were very similar, in terms of average parameters as well as of individual outputs. This observation suggests that they arrive at more or less the same interpretation of the molecular structure, although by following very different strategies. In particular, RNN, by encoding the structures through *ex novo* calculation of "adaptive topological descriptors", seems to reach a situation analogous to that obtained by MLR through selection of the most significant structural features. This hypothesis could be exploited in future work for the development of an approach that combines the flexibility of RNN with the more direct physical understanding of MLR.

5. Acknowledgments

The authors CD, AM, RS and MRT acknowledge the financial support of the University of Pisa. The financial support by Regione Toscana (Prot. n. AOOGRT/102715/Q.20.70.20 of 21/04/2011) is also gratefully acknowledged. The authors SN and KH are indebted to the Croatian - Hungarian TÉT project no Cro16/2006. This work was supported by grants nos. 098-1770495-2919 (SN) and 022-1770495-2901 (SN and AM) awarded by the Ministry of Science, Education, and Sport of the Republic of Croatia.

6. References

1. S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, D. Opit, Use of Statistical and Neural Net Approaches in Predicting Toxicity of Chemicals, *J. Chem. Inf. Comput. Sci.* **40** (2000) 885–890.
2. C. M. Auer, J. V. Nabholz, K. P. Baetcke, Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure–activity relationships (SAR) under TSCA Section 5, *Environ. Health. Perspect.* **87** (1990) 183–197
3. J. D. Walker, Applications of QSAR in toxicology: a US Government Perspective, *J. Mol. Struct. (Theochem)* **622** (2003) 167–184

4. Regulation (EC) No 1907/2006 of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC, *Official Journal of the European Union* **L 396/1** (2006) 1-849.
5. S. P. Bradbury, Quantitative structure–activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research, *Toxicol. Lett.* **79** (1995) 229–237.
6. C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister, R. A. Drummond, Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*), *Environ. Toxicol. Chem.* **16** (1997) 948–967
7. http://wwwepagov/ncct/dsstox/sdf_epafhmhtml.
8. M. Pérez González, A. Morales Helguera, M. Angel Cabrera, Quantitative structure–activity relationship to predict toxicological properties of benzene derivative compounds, *Bioorg. Med. Chem.* **13** (2005) 1775–1781.
9. L. H. Hall, L. B. Kier, G. Phipps, Structure–activity relationship studies on the toxicities of benzene derivatives: I an additivity model, *Environ. Toxicol. Chem.* **3** (1984) 355–365.
10. S. C. Basak, B. D. Gute, B. Lučić, S. Nikolić, N. Trinajstić, A comparative QSAR study of benzamidines complement–inhibitory activity and benzene derivatives acute toxicity, *Comput. Chem.* **24** (2000) 181–191.
11. A. A. Toropov, A. P. Toropova, QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity, *J. Mol. Struct. (Theochem)* **578** (2002) 129–134.
12. A. Micheli, A. Sperduti, A. Starita, A. M. Bianucci, A novel approach to QSPR/QSAR based on neural networks for structures, *Study Fuzz. Soft. Comput.* **120** (2003) 265–296.
13. A. M. Bianucci, A. Micheli, A. Sperduti, A. Starita, Application of cascade correlation networks for structures to chemistry, *Appl. Int. J.* **12** (2000) 117–146.
14. A. Micheli, A. Sperduti, A. Starita, A. M. Bianucci, Analysis of the internal representations developed by neural networks for structures applied to quantitative

structure–activity relationship studies of benzodiazepines, J Chem Inf Comput Sci 41 (2001) 202–218.

15. A. Micheli, A. Sperduti, A. Starita, A. M. Bianucci, Design of new biologically active molecules by recursive neural networks, Proc. Int. Joint Conference on Neural Networks 4 (2001) 2732–2737.
16. L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tiné, Predicting physical chemical properties of compounds from molecular structures by recursive neural networks, *J. Chem. Inf. Model.* **46** (2006) 2030–2042.
17. L. Bernazzani, C. Duce, A. Micheli, V. Mollica, M. R. Tiné, Quantitative Structure–Property Relationship (QSPR) Prediction of Solvation Gibbs Energy of Bifunctional Compounds by Recursive Neural Networks, *J. Chem. Eng. Data* **55** (2010) 5425–5428.
18. C. Bertinetto, C. Duce, A. Micheli, R. Solaro, A. Starita, M. R. Tiné, Evaluation of hierarchical structured representations for QSPR studies of small molecules and polymers by recursive neural networks, .*J. Mol. Graphics Model.* **27** (2009) 797–802.
19. C. Duce, A. Micheli, R. Solaro, A. Starita, M. R Tiné, Recursive neural networks prediction of glass transition temperature from monomer structure. An application to acrylic and methacrylic polymers, *J. Math. Chem.* **46** (2009) 729–755.
20. C. Duce, A. Micheli, A. Starita, M. R Tiné, R. Solaro, Prediction of polymer properties from their structure by recursive neural networks, *Macromol. Rapid Commun.* **27** (2006) 712–716.
21. C. Bertinetto, C. Duce, A. Micheli, R. Solaro, A. Starita, M. R. Tiné, Prediction of the glass transition temperature of (meth)acrylic polymers containing phenyl groups by recursive neural network, *Polymer* **48** (2007) 7121–7129.
22. C. Bertinetto, C. Duce, A. Micheli, R. Solaro, A. Starita, M. R. Tiné, Modelling Structure–Property Relationship for Copolymers by Structured Representation of Repeating Units, in: G. Maroulis, T. E. Simos TE (Eds.), *Advances in Computational Science*, Lectures presented at the International Conference on Computational Methods in Sciences and Engineering 2008 (ICCMSE 2008), *AIP Conf. Proc.* **1148** (2008) 33–36.
23. A. Micheli, C. Bertinetto, C. Duce, R. Solaro, M. R. Tiné, Recursive Neural Networks for Cheminformatics QSPR for Polymeric Compounds (towards Biomaterial Design), in: F. Masulli, A. Micheli, A. Sperduti (Eds.), *Knowledge–Based Intelligent*

Engineering Systems – Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, Vol 196, pp 37–52.

24. C. G. Bertinetto, C. Duce, A. Micheli, R. Solaro, M. R. Tiné, QSPR analysis of copolymers by recursive neural networks Prediction of the glass transition temperature of (meth)acrylic random copolymers, *Mol. Inf.* **29** (2010) 635–643.
25. R. Bini, C. Chiappe, C. Duce, A. Micheli, R. Solaro, A. Starita, M. R. Tiné, Ionic liquids prediction of their melting points by a recursive neural network model, *Green Chem.* **10** (2008) 306–309.
26. B. Lučić, N. Trinajstić, Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR modeling, *J. Chem. Inf. Comput. Sci.* **39** (1999) 121–132.
27. B. Lučić, S. Nikolić, N. Trinajstić, D. Juretić, The Structure–Property Models Can Be Improved Using the Orthogonalized Descriptors, *J. Chem. Inf. Comput. Sci.* **35** (1995) 532–538.
28. R. Todeschini, V. Consonni, Handbook of molecular descriptors. Wiley–VCH, Weinheim, 2000.
29. M. Karelson, Molecular descriptors in QSAR/QSPR. Wiley–Interscience, New York, 2000.
30. A. T. Balaban, From chemical topology to 3D molecular geometry. Plenum Press, New York, 1997.
31. H. Kubinyi, G. Folkers, Y. C. Martin, 3D QSAR in drug design, Kluver/ESCOM Leiden, Vol 1–3, 1996-1998.
32. K. Héberger, Sum of Ranking Differences A Novel Measure for Fair Comparison of Methods or Models, *Trends Anal. Chem.* **29** (2010) 101–109.
33. K. Héberger, K. Kollár–Hunek, Sum of ranking differences for method discrimination and its validation comparison of ranks with random numbers, *J. Chemometr.* **25** (2011) 151–158.