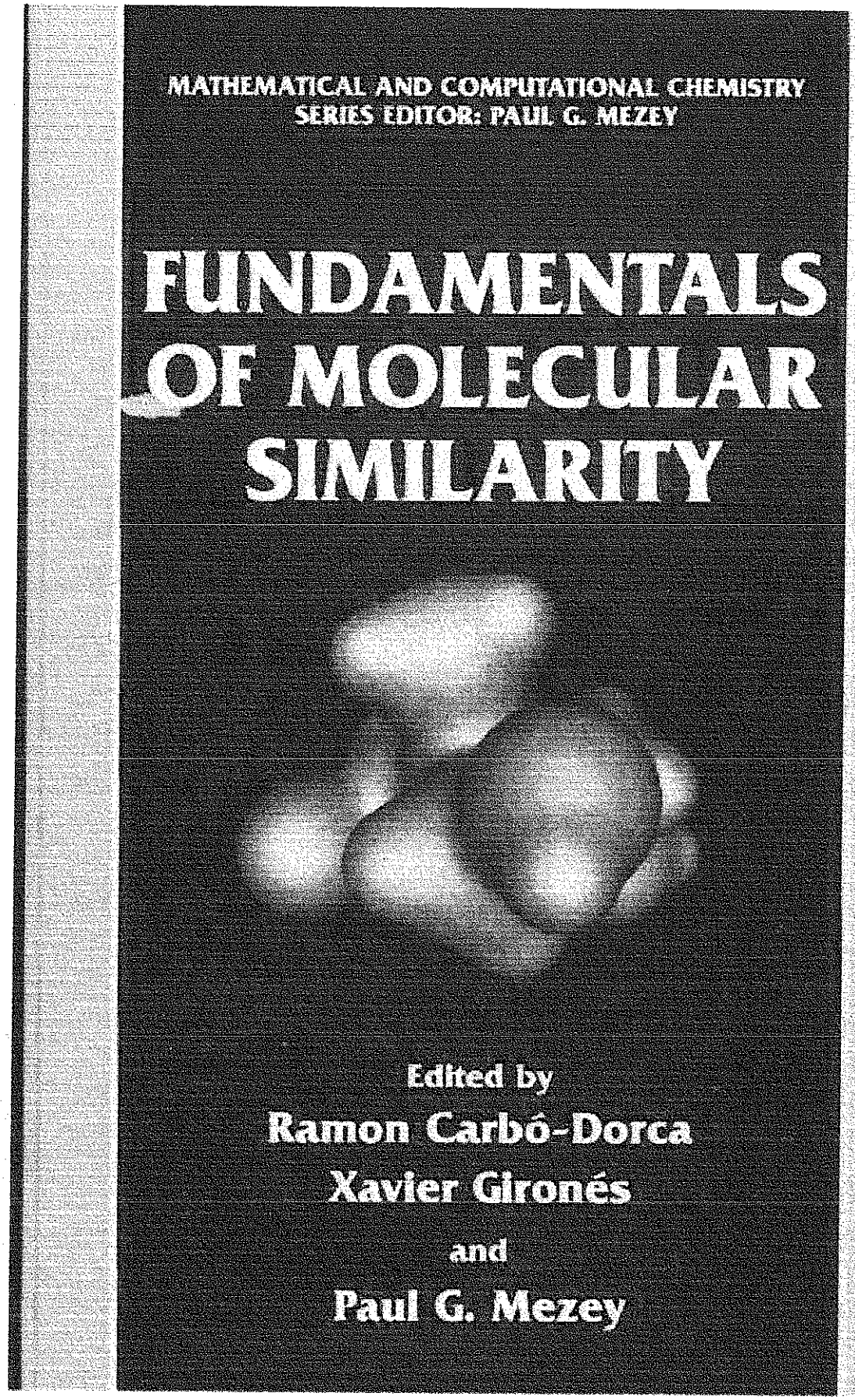

MATHEMATICAL AND COMPUTATIONAL CHEMISTRY
SERIES EDITOR: PAUL G. MEZEY
FUNDAMENTALS OF MOLECULAR SIMILARITY
Edited by
Ramon Carbó-Dorca
Xavier Gironés
and
Paul G. Mezey

Chapter 1

Prediction of boiling points of organic compounds from molecular descriptors by using backpropagation neural network

G. Espinosa, A. Arenas, and Francesc Giralt
Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química (ETSEQ), Departament d'Enginyeria Informàtica i Matemática (ETSE), Universitat Rovira i Virgili, Tarragona, Catalunya, Spain

Key words: Neural networks, QSPR, Boiling point

Abstract: Artificial neural networks (ANN) with a backpropagation learning algorithm are used to determine the relationship between the molecular structures of organic compounds and their boiling points using molecular descriptors. Two sets of descriptors, formed by up to four molecular connectivity and four valence connectivity indices, are considered to predict the boiling points of an heterogeneous set of 1116 organic compounds. The optimal number of descriptors or dimension of the input layer needed to attain the best predictions has been determined. In addition, different architectures have been considered either by directly expanding the number of units in the hidden layer of the standard backpropagation architecture or by the use of cascade correlation. The minimum dimension of the most representative training set is determined with a specific algorithm reported in the literature for image classification problems. For the two best backpropagation architectures 6-12-1 and 8-12-1, a number of 509 and 536 compounds were respectively required to capture the significant relationships between the structure and the boiling point of the complete set. The corresponding mean absolute errors in testing are 11.6 K and 19.7 K. For a subset of 242 alkanes and alcohol's, with 200 used for training and 42 for testing, the error decrease to 4.3 K for the 6-12-1 architecture. This error is lower than the 5.5 K obtained by multilineal regression analysis of the same data.

1. INTRODUCTION

The design and optimisation of industrial process require the knowledge of thermophysical properties. Available data banks can provide this information. However in specific cases, such as those related to drug activity or enviromental impact assessment, data are scarce and difficult or expensive to obtain experimentally. To overcome this lack of ready information, several thermodynamic models and correlations have been developed for a wide range of conditions. Among these models, the methods based on quantitative structure property relationships (QSPR) are promising. The basic concept of QSPR is to relate the structure of a compound with the property of interest. The compound's structure is expressed in terms of molecular descriptors that characterise a given molecular feature. Molecular descriptors, such as the connectivity indices and the corresponding valence connectivity indices, that encode features such as size, branching, unsaturation, heteroatom content and cyclicity [1,2] are useful. For example, the first order connectivity index was used in 1982 to correlated the solubility of hydrocarbons in water [3]. The connectivity indices are based on local molecular properties and are bond-additive quantities so that in bonds of different kinds make different contribution to the overall molecular descriptors. The key step is to build the structure property relationship. This involves two major activities: 1 The representation of compounds using molecular descriptors and multivariate statistical methods or artificial neural networks [4,5]. And 2 the mapping of the descriptors to built a relationship with the properties of interest. Among the physical properties correlated by QSPR are boiling points, [1,6,7], melting points, [7], solubilities, [3], partition coefficients, [8]. The success of regression analysis in QSPR model building depends upon the degree of linearity between the physical property of interest and the descriptors selected. As the number of descriptors increases the capability of regression analysis decreases due to the redundancy of information incorporated by the different descriptors. Some techniques, such as principal component analysis and partial least square regression, have been used to minimise this problem. Nevertheless, these techniques require the a priori assumption of the form of the model. To solve this issue, multilineal regression analyses (MLR) is commonly used as an alternative. Recently, artificial neural networks have become an option to build QSPR models. The purpose of the current study is to apply QSPR and neural networks to better correlate the boiling points of organic compounds.

2. BOILING POINT

The boiling point of organic compound is useful for identifying substances and for estimating other physical properties [9]. There are different methods to predict boiling points. For example, group contribution methods are widely used for this purpose [10]. These contribution methods are limited to the class of compounds for which the groups have been established. The QSPR approach employs descriptors derived solely from molecular structure [4,11]. One of the pioneering works to predict the boiling points of paraffins was carried out by Wiener [2]. Other topological indices, such as the connectivity indices [1,4] and the Randic indices [2], have been successfully applied to correlate the boiling points of alkanes and alcohols.

In the present study, 1116 organic compounds were considered. The boiling points were taken from the Design for Physical Property Data (DIPPR) database. Two subsets that contained the same data as those used by Kier and Hall [1] and Hall and Story [12] were chosen to validate the results. The complete set is structurally heterogeneous, includes saturated and unsaturated hydrocarbons, aromatic, and halogenated compounds, with groups cyano, amino, ester, ether, carbonyl, hydroxyl, and carboxyl. The structures and connectivity indices of these compounds were obtained using the Molecular Modelling Pro software. Four molecular connectivity indices and four valence molecular connectivity indices for each compound were considered.

3. ARTIFICIAL NEURAL NETWORKS

The next step is to establish the relationship between molecular descriptors and the boiling points by using artificial neural networks. A standard neural architecture consists of many simple interconnected processors (units). The weight of each connection or synapse stores the information learned from examples [13]. The successful application of neural networks depends on three factors. First the design is critical because the network will overfit data if too few hidden units are used. Second, the size of the training set must be correct to avoid over or under training. Finally, it is important to select an appropriate training set, because it has to represent the entire dataset.

Two supervised neural algorithms, where input patterns are associated with known output patterns, were used. The first one was backpropagation architecture [14,15]. Its implementation involves a forward pass through the layers of units (nodes) to estimate the error, followed by a backward pass

that modifies the weights (synapses or connections) to decrease the error. Networks with one input and one output layer, and with one or two hidden layers with different nodes in each layer, were examined. Six or eight nodes in the input layer were considered so that input vectors with six or eight connectivity indices could be presented to the network. The output node was the boiling point. To train the network, the weights of the synapses between the nodes of each layer and those of the next layer were optimised with a steepest descent method by propagating the error back throughout the layers.

One problem with backpropagation is to find the appropriate topology. To overcome this constrain an auto constructive algorithm was also implemented. The cascade correlation method was selected because it is one of the most relevant constructive algorithms [16,17]. The hidden units are added to this network one at a time without changing the connection weights after they have been added. It supports a variety of learning algorithms, but a backpropagation scheme was used for consistency. An initial, minimal network with only input and output units was trained. Training continued until a given criteria, such as the maximum number of epochs or a patience indicator, were met. If the network did not fulfilled the error criteria during the initial phase, a new hidden unit was added to maximise the correlation. This hidden unit should account for missing features.

4. RESULTS AND DISCUSSION

Three sets of compounds were used to evaluate the present model and to compare its performance with previous proposals reported in the literature. The first set of compounds include 242 alcohols, with up to ten carbon atoms, and all the alkane isomers with five to ten carbon atoms. The range of boiling points is 282.65K-504.15K. A number of 42 compounds were selected for the testing phase. The results obtained with several backpropagation architectures indicated that expansion of the input space yielded better results than a contraction in all cases. This shows that extra dimensionality can represent better additional features of the training set, and that these extra features make a favourable contribution to the performance of the network. The best configuration for this set of alkane isomers and alcohols is a 8-12-1, i.e., the combination of eight connectivity indices as input (four molecular connectivity indices plus four valence connectivity indices), one hidden layer with twelve units, and the boiling point as output. The absolute mean error between the predicted and experimental boiling points was 4.3K. Only two compounds (heptane and nonanol) yielded residuals greater than 10K The standard deviation of the predictions was 3.3K and the average relative error 2.9%. For the same set of connectivity

indices, a multilineal regression analysis yielded predictions with a mean absolute error of 10.3K, a standard deviation of 5.5K and a relative mean error of 7.7%. It is not possible to make a direct comparison with the previous work of Hall and Kier [1] for a similar set of compounds, because they used different molecular descriptors. Still it should be noted that those authors reported a mean absolute error of 5.9K. This corresponds to a 4.1% relative error which is higher than for the model presented here. The predicted and measured boiling points are plotted in Fig. 1. The data in Fig.1 were correlated with a coefficient r^2=0.983, which validates the model established by the network. These results show that the non linear relationship between molecular structure and boiling point is well extracted by the neural network, increasing the extrapolation capabilities of this model to other different but similar sets of compounds.

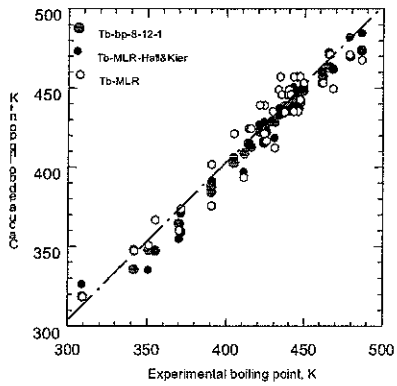


Figure 1. Boiling points for the 42 alcohols and alkanes used for testing with an 8-12-1 backpropagation architecture (bp). Comparison with Hall and Kier, [1], and with MLR with connectivity indices.

The second set is formed by 220 heterogeneous organic compounds with three to nineteen carbon atoms, including saturated and unsaturated hydrocarbons, and the groups ester, ether, carbonyl, hydroxyl, and carboxyl. Their boiling point range was 225.51K-608K. Training was carried out with 30 randomly selected compounds. In this case the expansion of the input space also yielded better results than contraction. The best configuration was 8-12-1. The standard deviation between predictions and measurements was

11.3K, with an absolute mean error of 21.8K and an average relative error of 5.38%. Hall and Story [12] reported a mean absolute error of 4.57K, for a 19-5-1 architecture (nineteen electrotopological indices as input) for the same group of compounds, which correspond to a 1.12% relative error. This better performance is due to the type of indices used by these authors, which allows the complete characterisation of all functional groups involved. Fig. 2 depicts the boiling points predicted versus the experimental data, for the set of heterogeneous organic compounds.

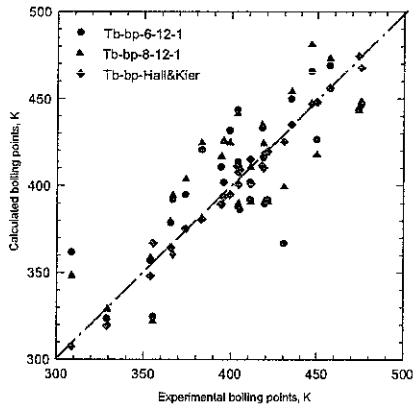


Figure 2. Boiling points for the 30 organic compounds used for testing with an 8-12-1 and 6-12-1 backpropagation architectures (bp). Comparison with the neural architecture 19-5-1 of Hall and Story, [12] (electrotopological indices)

The data in Fig. 2 were correlated with a coefficient r2=0.8. This low correlation coefficient indicates that increasing the diversity of compounds without increasing the information about the functional groups involved decreases the capability of neural of networks based on processing units. It should be noted that a neural network requires a minimum number of data for training and that the training set has to represent the majority of characteristics of the whole set of compounds considered. Also, connectivity indices alone do not provide enough information to characterise a heterogeneous set. To solve the first issue, the two previous sets were unified and incremented to 1116 heterogeneous organic compounds, with boiling points spanning the range 111.7K-711.5K. About 60% of the compounds were used for training and the rest were used to evaluate the model. The best

configuration was again 8-12-1. The mean absolute error was 28 K and the relative mean error 7%. In all trials carried out to select the best architecture the compounds that could be consider as outliers (compounds with high residuals) were the polyfluorine compounds, substituted aromatics, and pyridines. The results are summarised in Fig. 3.

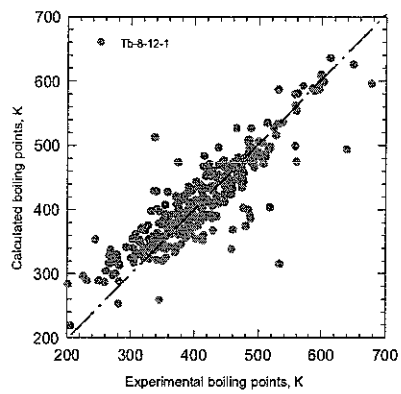


Figure 3. Boiling points for the 416 organic compounds used for testing with an 8-12-1 backpropagation architecture (bp).

The performance of the architectures tested did not improve by randomly selecting different compounds for training the network, within 60% of the total. To overcome this deficiency, the size and the content of training set was optimised using a specific algorithm reported in the literature for image classification problems [18]. This means the minimum number of patterns with the maximum of information was finally determined. The best configurations for this set were 8-12-1 and 6-12-1, eight or six inputs respectively, twelve units in the hidden layer and the boiling point of the output layer. The standard deviations of the predicting set were respectively 12.8K and 11.3K, absolute mean errors were 19.7K and 11.6K and the average relative errors were 4.62% and 4.33%. The predictions obtained with the 8-12-1 and 6-12-1 architectures and the best training set are shown in Figs. 4 and 5. The comparison of these results with Figure 2 show the importance of the definition of the training set to build a good model.

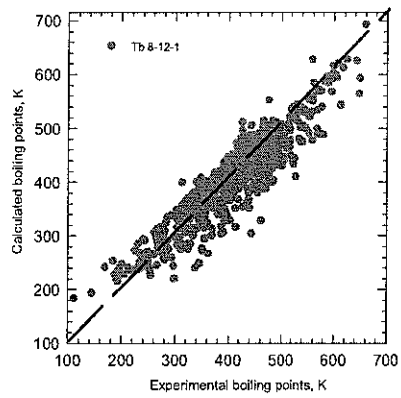


Figure 4. Boiling points for the 509 organic compounds used for testing with an 6-12-1 backpropagation architecture (bp) and the minimum training set.

To overcome the second issue raised above about the insufficient molecular information of topological indices, Espinosa et al., [21] considered the dipole moment and the kappa index as additional descriptors to model de boiling points of a homogeneous set of aliphatic hydrocarbons. The inclusion of these two indices in the input vector does not improve the correlation of the boiling points of the current heterogeneous sets with backpropagation algorithm. Thus, the use of cognitive systems such as FuzzyARTMAP, [22] should be considered, together with additional three-dimensional molecular information. Finally, it is worth noting that contrary to previous reports, cascade correlations does not improve the performance of backpropagation, [16,17]. For the compounds in Fig 6, the cascade correlation algorithm yields a mean absolute error of 33K, compared to the 12.8K and 11.3K obtained with the 6-12-1 (Fig. 4) and 8-12-1 (Fig. 5) backpropagation architectures, respectively.

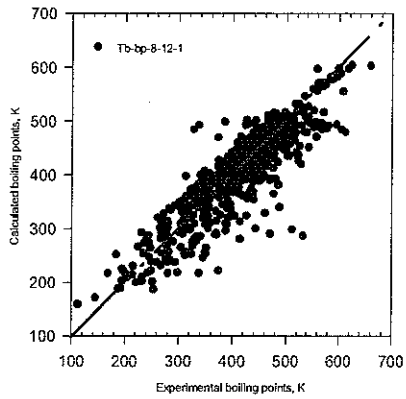


Figure 5. Boiling points for the 536 organic compounds used for testing with an 8-12-1 backpropagation architecture (bp) and the minimum training set.

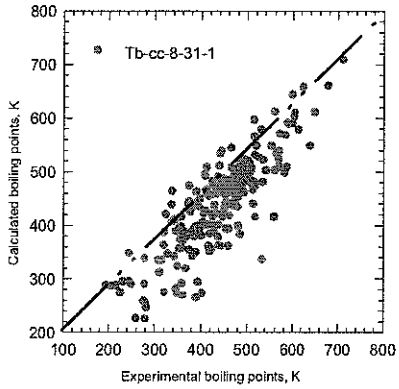


Figure 6. Boiling points for the 416 organic compounds used for testing with an 8-33-1 cascade correlation architectures (cc).

5. CONCLUSION

The present model, which combines neural networks with QSPR, performs better than previous correlations for similar input information. The best backpropagation configurations to p redict the boiling points of 1116 organic compounds were 8-12-1 and 6-12-1. This implies using eight or six connectivity indices as input nodes, twelve middle nodes, and a single node for boiling point. The cascade correlation constructive algorithm didn't yield better r esults than backpropagation. The determination o f the minimum training set reduces the absolute mean error and improves predictions.

REFERENCES

1. L. Hall and L. Kier,*J. Chem. Inf. Compt. Sci.* **35**, pp 1039(1995).
2. M. Randic and N. Trinajstic,*J. Mol Struct.* **284**, 209 (1993).
3. M. Medir and F. Giralt,*AICHE Journal* **28**, 341 (1982).
4. Katritzky, M. Karelson and V. Lobanov,*Pure Appl. Chem.* **69**, 245 (1997).
5. P. Jurs, 214 th ACS National Meeting, 1997.
6. Katritzky, Lan Mu, and V. Lobanov,*J. Phys. Chem.* **100**, 10400 (1996).
7. Katritzky,Lan Mu, and M. Karelson,*J. Chem. Inf. Compt. Sci.* **38**, 293 (1998).
8. Patil,. *J. Hazard. Mater.* **19**, 35 (1994)
9. R. Reid, J. Prausnitz and B.Poling,*The Properties of Gases and liquids*, 4th ed., McGraw-Hill, New York, 1987.
10. Joback and R. Reid,*Chem. Eng. Commun.* **57**, 233 (1987).
11. Bünz, B. Braun, and R. Janowsky *Ind. Eng. Chem. Res.* **37**, 3043 (1998).
12. Hall and C. Story *J. Chem. Inf. Compt. Sci.* **36**, 1004 (1996).
13. Hertz and K. Palmer,*Introduction to the Theory of Neural Computation*, Addison Wesley, The Advanced Book Program, pp. 115 (1991).
14. D. Rumelhart, G. Hinton, and R.Williams,*Nature* **323**, 533 (1986).
15. S. Falhman,*An Empirical Study of Learning Speed in Backpropagation Networks*, Technical Reports CMU-CS-88-162 (1988).
16. S. Fahlman, and C. Lebiere,*The Cascade Correlation Learning Architecture, Advances in Neural Information Processing System* II, pp. 524 (1990).
17. Squieres and J. Savlik,*Experimental Analysis of Aspects of the Cascade Correlation Learning Architectures*, Machine Learning Research Group Working, paper 91-1 (1991).
18. F.Tamburini and R. Davoli,*An Algorithm Method to Build Good Training Sets for Neural Networks Classifiers*, Technical Report UBLCS-94-18, (1994).
19. G. Carpenter and S. Grossberg,*Computer Vision, Graphics, and Image Processing.* **37**, 54 (1987).
20. G. Carpenter and S. Grossberg,*Computer*, **21**, 77 (1988).
21. Espinosa G., Yaffe D., Cohen, Y., Arenas, A., and Giralt F*J, Chem. Inf. Compt. Sci.* **40**, 859 (2000).
22. Giralt, F., Arenas, A., Ferre-Gine, J., Rallo, R. and Kopp, G*Physics of Fluids*, **12**, 1826 (2000).

605211462

Slow transition from star configuration to homogenous configuration

December 4, 2002

Within the formalism introduced for representing the network as a set of discrete points in N-dimensional space, where the position of a network is defined by its betweenness vector b_i, we can measure a euclidean distance between any two points. In order to see which regions of the space we are visiting in the optimisation process, we want to measure the *shortest* distance d between a real network and the its closest point on the bisector.

$$d^2 = \sum_{i=1}^{N} (b_i - \langle b \rangle)^2$$

where $\langle b \rangle = \frac{1}{N} \sum_{i=1}^{N} b_i$.

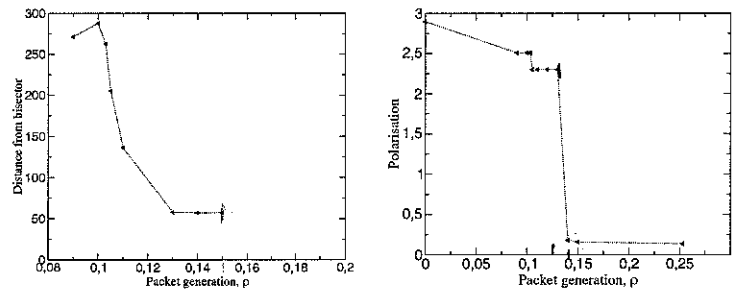
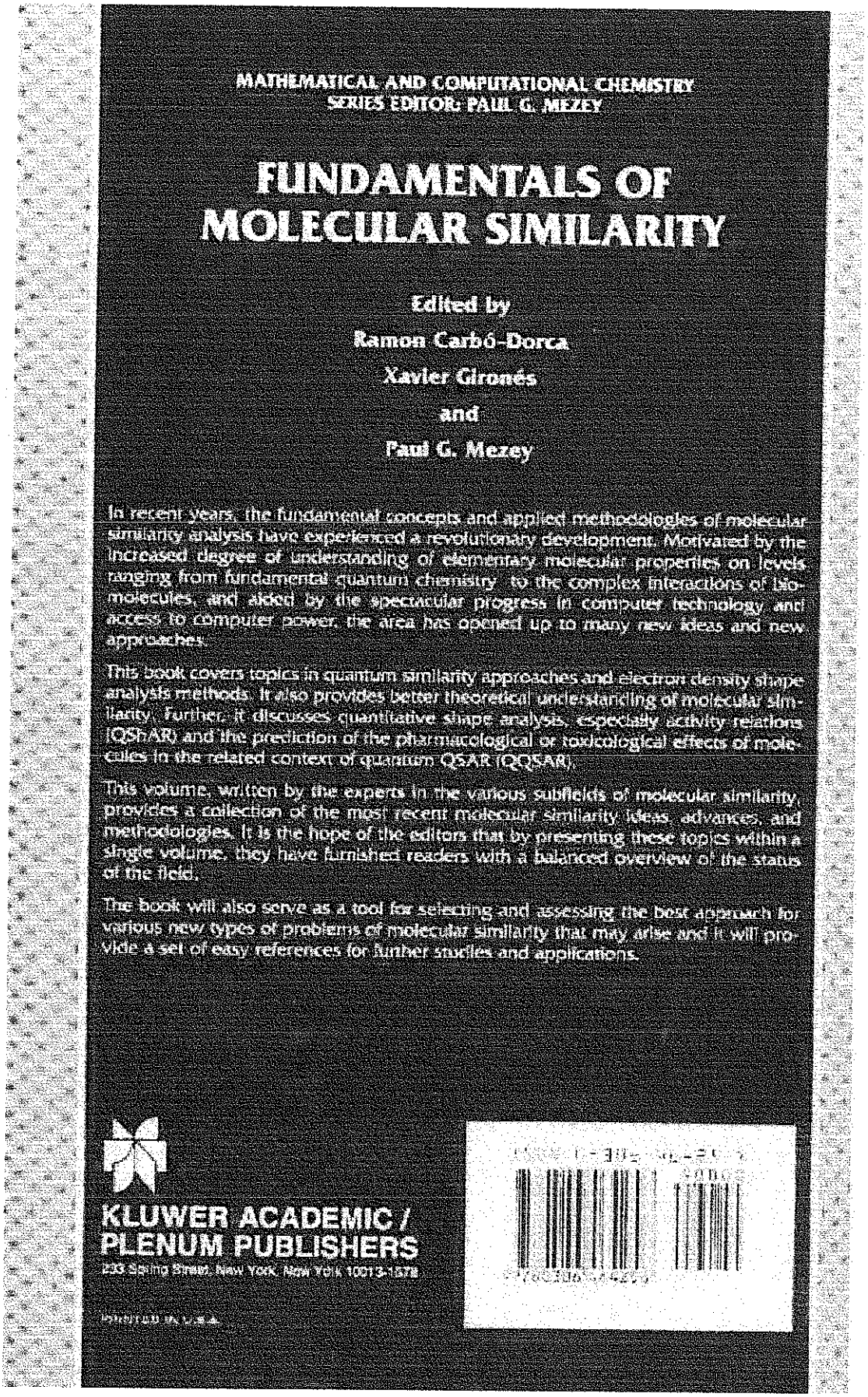


Figure 1: (a) Shortest distance from the bisector to the network in N dimensional space for different values of ρ. The network has 32 nodes and 128 links. Most importantly, the distance does not fall as rapidly as expected. There exist optimal configurations which are neither stars, nor homogeneous networks (for example $rho = 0.105$ and $\rho = 0.110$ which correspond to the networks shown in figure 2 (d) and (e)).(b) Polarisation of the network with 32 nodes and 128 links. The transition from starlike configurations to homogeneous ones is very sharp in this representation.


MATHEMATICAL AND COMPUTATIONAL CHEMISTRY
SERIES EDITOR: PAUL G. MEZEY
FUNDAMENTALS OF MOLECULAR SIMILARITY
Edited by
Ramon Carbó-Dorca
Xavier Gironés
and
Paul G. Mezey
In recent years, the fundamental concepts and applied methodologies of molecular similarity analysis have experienced a revolutionary development. Motivated by the increased degree of understanding of elementary molecular properties on levels ranging from fundamental quantum chemistry to the complex interactions of bio-molecules, and aided by the spectacular progress in computer technology and access to computer power, the area has opened up to many new ideas and new approaches.
This book covers topics in quantum similarity approaches and electron density shape analysis methods. It also provides better theoretical understanding of molecular similarity. Further, it discusses quantitative shape analysis, especially activity relations (QSHAR) and the prediction of the pharmacological or toxicological effects of molecules in the related context of quantum QSAR (QQSAR).
This volume, written by the experts in the various subfields of molecular similarity, provides a collection of the most recent molecular similarity ideas, advances, and methodologies. It is the hope of the editors that by presenting these topics within a single volume, they have furnished readers with a balanced overview of the status of the field.
The book will also serve as a tool for selecting and assessing the best approach for various new types of problems of molecular similarity that may arise and it will provide a set of easy references for further studies and applications.
KLUWER ACADEMIC / PLENUM PUBLISHERS
233 Spring Street, New York, New York 10013-1578
PRINTED IN U.S.A.