Estimation of Infinite Dilution Activity Coefficients of Organic Compounds in Water with Neural Classifiers

Francesc Giralt and G. Espinosa
Departament d'Enginyeria Química, ETSEQ, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain

A. Arenas and J. Ferre-Gine
Departament d'Enginyeria Informática i Matemàtiques, ETSE, Universitat Rovira i Virgili, Tarragona, Catalunya, Spain

L. Amat, X. Gironés, and R. Carbó-Dorca
Institut de Química Computacional, Universitat de Girona, Girona, Catalunya, Spain

Y. Cohen
Dept. of Chemical Engineering, University of California, Los Angeles, CA 90095

DOI 10.1002/aic.10116
Published online in Wiley InterScience (www.interscience.wiley.com).

A new approach is presented for the development of quantitative structure–property relations (QSPR) based on the extraction of relevant molecular features with self-organizing maps and the use of a modified fuzzy-ARTMAP classifier for variable prediction. The present methodology is demonstrated for the development of a QSPR for the aqueous-phase infinite dilution activity coefficient γ^{∞}, based on a data set of 325 diverse organic compounds. The QSPR was developed using a set of 11 molecular descriptors (four connectivities ${}^{v}\chi^{1-4}$, Coulomb self-similarity measure, electron–nuclear attraction, dipole moment, sum of atomic numbers, number of filled levels, average polarizability, and nuclear–nuclear repulsion). The final set of molecular descriptors was selected from an initial pool of 23 topological and quantum chemical descriptors, including six molecular quantum similarity measures, by means of a topological analysis of self-organization of the data set. Additional interpolated information to enhance the training of the neural system was obtained from the self-organization analysis. The resulting fuzzy-ARTMAP–based QSPRs performed with errors that were on the average seven times smaller compared to previous published models. The use of only four molecular quantum similarity measures proved to be sufficient for building a $\ln\gamma^{\infty}$ fuzzy-ARTMAP–based QSPR with reasonable accuracy. © 2004 American Institute of Chemical Engineers *AIChE J,* 50: 1315–1343, 2004

Keywords: self-organizing maps, fuzzy-ARTMAP neural classifier, QSPR, infinite dilution activity

Introduction

The distribution of organic chemicals in the environment is affected by their physicochemical and thermodynamic properties, among which air–water partitioning (or Henry's law constant) and aqueous solubility are of particular importance (Mackay et al., 1992; Yalkowsky, 1999; Yalkowsky and He, 2003). Knowledge of the above parameters is of fundamental interest and also of importance in various industrial processes (Fredenslund et al., 1975; Mackay et al., 1992) and in groundwater remediation by air stripping. The Henry's law constant of sparingly water soluble organics is directly proportional to the

Correspondence concerning this article should be addressed to F. Giralt at fgiralt@etseq.urv.es.

infinite dilution activity coefficient γ^{∞}, which in turn is essentially inversely proportional to the aqueous solubility. The infinite dilution activity coefficient is a fundamental thermodynamic parameter important for the estimation of aqueous solubility and Henry's law constant (Fredenslund et al., 1975; Mackay et al., 1992).

Various theoretical models and group contribution methods for predicting activity coefficients in dilute solutions have been proposed in the literature (Fredenslund et al., 1975; Lazaridis and Paulaitis, 1993; Mackay and Shiu, 1977; Mackay et al., 1992; Medir and Giralt, 1982; Mitchell and Jurs 1998; Sherman et al., 1996; Tochigi et al., 1990). These predictions have been successful when dealing with athermal systems and, to a lesser extent, with polar systems. For example, it has been shown that the combinatorial term in UNIFAC is underestimated, whereas the residual is overestimated (Voutsas and Tassios, 1996) when dealing with aqueous mixtures and with increasing polarity of the organic solute. The modification of the combinatorial and residual terms in the original UNIFAC model has significantly improved predictions for athermal systems but such modifications have been less satisfactory for polar systems (Voutsas and Tassios, 1997). Alternative approaches that are specifically designed for aqueous solutions, such as the linear solvation energy relationship (LSER; Sherman et al., 1996), might be more accurate (average absolute deviation of 0.294 $\ln\gamma^{\infty}$ units for 336 organics in aqueous systems).

Improvements of group contribution methods by direct calculations of interaction energies are possible by means of quantum chemical or variational methods; however, such methods are typically computationally demanding and restricted to relatively small molecules (Sandler, 2002). A brief review on the application of computational quantum chemistry methods, either to obtain the interaction energy surface for a pair of molecules or to improve current group contribution methods by less intensive calculations, can be found elsewhere (Sandler, 2002).

Over the last three decades various quantitative structure–property relationships (QSPRs) have been proposed to estimate γ^{∞} for organics in water. The premise of QSPRs is that there is a unique relationship between molecular chemical descriptors and a target physicochemical property. Given a selected set of molecular descriptors, one searches for optimized correlations between the descriptors and the desired chemical specific property. For example, Mackay and Shiu (1977) correlated the aqueous phase $\ln\gamma^{\infty}$ for hydrocarbons with the number of carbon atoms. Medir and Giralt (1982) correlated $\ln \gamma^{\infty}$ for aliphatic and aromatic hydrocarbons using molecular descriptors that included the first-order molecular connectivity index, surface area, dipole moment, number of carbon atoms, total electronic energy, and acentric factor. Mitchell and Jurs (1998) applied several correlation techniques and perceptron neural networks to estimate $\ln\gamma^{\infty}$. They used three topological descriptors, four charged partial surface area (CPSA) indices, two hydrogen bonding descriptors, heat of formation, and two theoretical linear solvation energy relationship (TLSER) indices to describe the basicity of hydrogen bonding. Also noted is a related study by Yalkowsky and Valvani (1979), in which the aqueous solubility of organic compounds was correlated with the molecular surface area, and a series of studies by Yalkowsky and coworkers (Peterson and Yalkowsky, 2001; Ran and Yalkowsky, 2001; Ran et al., 2001; Yang et al., 2002) that contributed significantly to this area of aqueous solubility estimation with models based on group contribution or fragment structural information.

Backpropagation neural networks have recently emerged as an alternative for the development of QSPRs and quantitative structure–activity relationships (QSARs) to predict physicochemical properties and biological activities, respectively (Bünz et al., 1998; Chow et al., 1995; Egolf and Jurs, 1993; Espinosa et al., 2000, 2001a,b; Gakh et al., 1994; Hall and Story, 1996; Mitchell and Jurs, 1998; Simamoea et al., 1993; Stanton and Jurs, 1990; Stanton et al., 1991; Viswanadhan et al., 2001; Yaffe et al., 2001, 2003). This alternative modeling strategy for QSPR development yields significantly higher prediction accuracy compared to that of traditional regression-based correlations. For example, the 12–6–1 neural network model proposed by Mitchell and Jurs (1998), to estimate $\ln\gamma^{\infty}$ of organics in water, performed with an average root-mean-square error of 0.376 ln units, 0.406 ln units, and 0.434 for the training (271 compounds), validation (25 compounds), and test (25 compounds) sets, respectively. Also, 92 γ^{∞} values of 19 halocarbons in water and 18 organic compounds in five hydrofluoroparaffins solvents over a temperature range of 291–333 K were predicted by Rani and Dutt (2002) with a feedforward network trained with 351 data points, with an average absolute deviation of 11.8% on the basis of γ^{∞}, compared with 94.3% obtained by multilinear regression. The interpretation of results obtained with these feedforward neural architectures is not straightforward, given that the structure–property or structure–activity relationships are embedded within the weights distributed within the network. More recently, neural network–based QSPRs and QSARs (Espinosa et al., 2002, 2003; Gasteiger et al., 1994a,b) have been developed based on Kohonen (self-organizing or feature maps) and ARTMAP neural network architectures. This latter approach proved to be particularly useful in the recognition of coherent structures embedded in turbulent flows (Ferre-Gine et al., 1996) and in the development of industrial virtual sensors (Rallo et al., 2002a,b) attributed to the ability of these algorithms to classify patterns in complex data sets, even in the presence of other correlated information and noise.

The current study presents a comprehensive approach to developing neural network–based QSPR for the aqueous infinite dilution activity coefficient of organics based on a predictive fuzzy-ARTMAP architecture and the use of self-organizing maps (SOMs), also known as a Kohonen neural network (Kohonen, 1982, 1990), for extracting molecular features relevant to the target property. The initial set of descriptors was selected to contain topological and quantum molecular information to capture both two- and three-dimensional (3-D) (Carbó-Dorca and Besalú, 1998; Cramer et al., 1988) information (such as conformational, stereochemical, electronic, and binding information). The present set of descriptors also included molecular quantum similarity measures (MQSM; Carbó-Dorca and Besalú, 1998). The selection of the most suitable set of descriptors from the initial set was accomplished with a SOM analysis, which also served to identify chemical classes and their characteristics with respect to the molecular information included in the set of descriptors. In addition, in the current work we demonstrate that the integration of SOM with fuzzy ARTMAP (Carpenter et al., 1987, 1991, 1992; Giralt et al.,

2000) improves not only the accuracy and predictive capabilities of the QSPR (Espinosa et al., 2002, 2003) but enables one to explore the relative contribution of any given descriptor (or group of descriptors), with respect to both chemical classification and estimation of the target property.

Neural Network Architectures

Kohonen self-organizing maps: cluster analysis and selection of descriptors

The Kohonen neural network is a self-organizing map suitable for classification analysis (Erwing et al., 1992; Kaski and Lagus, 1996; Kohonen, 1982, 1990; Vesanto, 1999). In the present work SOM analysis was used to: (1) select the most suitable set of descriptors by measuring the dissimilarity of the different maps that are formed when clustering into the nodes of the map the complete set of compounds according to their molecular characteristics, as specified by different sets of molecular descriptors and the target $\ln\gamma^{\infty}$ variable; and (2) use the vectors characterizing each neuron or node in the trained map, that is, the prototype vectors of clustered compounds into the nodes during training, in addition to the compounds themselves, to train the fuzzy-ARTMAP–based QSPR.

A self-organizing map automatically adapts itself such that similar inputs are associated with topologically close units (or neurons) in a two-dimensional (2D) grid. During the training process, N-dimensional input data are self-organized in a discretized 2D plane formed usually by a grid of $K \times K$ units. The number of units is specified based on the sought population distribution of input data among the neurons, in such a way that close input vectors of descriptors characterizing compounds in this N-dimensional space are mapped into close-neighborhood neurons while minimizing the number of empty or overcrowded classes. The SOM type of neural network is especially useful for capturing underlying relationships within the input data. Briefly, the main steps in the generation of a SOM are as follows:

(1) An input vector x_i of dimension N is presented to the network. Each cycle of presentation including all input vectors x_i is called an *epoch*.

(2) The Euclidian distance between this input vector and all nodes (neurons) in the network lattice is calculated, as follows

$$\delta_j = \sum_{i=0}^{N-1} [x_i(t) - w_{ij}(t)]^2 \qquad (1)$$

where $x_i(t)$ is the ith component of the N-dimensional input vector and $w_{ij}(t)$ is the connection strength (weight) between the input vector component i and the mapping array node j at position t, in the sequence of data presentation to the network. Initially these weights are assigned random values.

(3) Node j^* with the minimum distance δ_j defined by Eq. 1 is selected as the winner neuron or best matching unit (BMU).

(4) The weights of node j^* and those of its neighbor nodes, identified by the neighborhood $N_{j^*}(t)$, are updated

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)[x_i(t) - w_{ij}(t)] \qquad (2)$$

for $j \in N_{j^*}(t)$ and $1 \le i \le N$.

The function $\eta(t)$, which decreases monotonically over the environment of the winner neuron, defines the region of influence that the input vector has on the SOM. This function is defined by the neighborhood function η_0 and the learning rate $\alpha(t)$ according to

$$\eta(t) = \eta_0(\|r_c - r\|, t)\alpha(t) \qquad (3)$$

where r is the location of the units or neurons on the grid of the map. The simplest neighborhood function is the bubble function, which is constant over the whole neighborhood of the winner neuron (node) and zero elsewhere (Kaski and Kohonen, 1994; Vesanto, 1999). However, a more convenient function is the Gaussian neighborhood function, defined by

$$\eta_0 = \exp\left(\frac{-\|r_c - r\|^2}{2\sigma^2(t)}\right) \qquad (4)$$

where the neighborhood radius $\sigma(t)$ self-adapts after each epoch. The type of neighborhood function and the number of neurons determine the sensitivity and the granularity of the map, respectively. Finally, it is noted that the learning rate $\alpha(t)$ in Eq. 3 is a decreasing function of t over the range [0, 1] and it is usually defined by the power series

$$\alpha(t) = \alpha_0\left(\frac{\alpha_T}{\alpha_0}\right)^{t/T} \qquad (5)$$

where α_0 and α_T are the initial and final learning rates, respectively, and T is the size of the training set cycles (i.e., the number of epochs selected for training).

The trained SOM can be used to visualize different features of the data (Kaski and Kohonen, 1994; Vesanto, 1999). The graph representations of the clustered set of data into the nodes of the map facilitate a clearer identification of the underlying relationships among data. This is accomplished by visualizing either the matrix of distances between nodes (**U**-matrix) or the contribution of different input information into this organization [component planes (C-planes)]. A component plane is the distribution over the map of the values of one of the elements (weights w_{ij}) of the vectors (prototypes) characterizing each neuron or node. Visualization and analysis of the SOM were accomplished using the Matlab SOM toolbox, as reported elsewhere (Kaski and Kohonen, 1994; Vesanto, 1999).

Identification of prototype classes of compounds and selection of relevant descriptors required setting the optimal size of the SOM. The optimal SOM size should accommodate, during training, the compounds characterized by the N molecular descriptors of the pool (that is, initial set) plus the target variable $\ln\gamma^{\infty}$ into the $K \times K$ grid units, with about 80% of the nodes occupied by compounds to ensure both continuity of clusters within the nodes of the map and a sufficient population density per node. The above ensures the generation of compact clusters in the nodes [that is, small average distance (n) between the members of each cluster], and that all nodes, whether or not occupied by clustered compounds, are trained according to Eqs. 2 and 3. Subsequently, a curvilinear component analysis and/or visual inspection of the component C-planes corresponding to all descriptors was carried out to identify similarities in clustering topology (that is, how each descriptor clustered the compounds in the nodes in relation to the target variable $\ln\gamma^{\infty}$). Thus, descriptors favoring a similar topology or distribution of compounds among the nodes (clusters) in the map could be identified as similar and grouped into a common class of descriptors with respect to the target variable. Descriptors were ordered by picking from each class of similar descriptors those with the highest correlation or absolute covariance, with the restriction that each absolute covariance with the target variable was higher than the average value for the pool of descriptors (that is, the complete set of descriptors). The ordering process continued by sorting indices according to the value of the absolute covariance. Following this procedure (Espinosa et al., 2002), nonredundant information of the different classes of descriptors, formed by grouping their respective *C*-planes, as well as their correlation with respect to the target variable, were accounted for in the construction of the most suitable set of descriptors.

The most representative set (that is, most suitable set) of descriptors was defined as the set with the smallest number of descriptors that provided the highest representation of the compound data set as determined by the SOM analysis. The selection was carried out by first selecting the most descriptive index from each of the *N* self-organizing maps, and then successively adding the remaining indices in the order of decreasing covariance with the target variable. To determine when all relevant information had been considered in the succession of *N* SOMs, changes in topology caused by the progressive incorporation of the molecular information were quantified. This was accomplished by measuring the dissimilarity between any of two maps *L* and *M*. Dissimilarity was defined as the averaged difference in the SOM representation of the sample vectors used for training, as follows

$$D(L, M) = E\left[\frac{d_L(x) - d_M(x)}{d_L(x) + d_M(x)}\right] \tag{6}$$

in which E is the average expectation, the subscripts L and M designate the two different maps, and $d(x)$ is the distance from x to the second BMU, denoted by $m_{c'(x)}$, beginning at the first BMU or winner neuron, denoted by $m_{c(x)}$. Of all possible paths between $m_{c(x)}$ and $m_{c'(x)}$ the shortest continuous path

$$d(x) = \|x - m_c(x)\| + \min_i \sum_{k=0}^{K_{c'(x)}-1} \|m_{I_i(k)} - m_{I_j(k+1)}\| \tag{7}$$

between neighbor units was selected. This distance, which is similar to that first proposed by Kaski and Lagus (1996), reflects the continuity of the SOM. Also, it indicates the relative capacity of any map to represent the data set when trained with compounds characterized by a given number of descriptors with respect to any other map trained using more or fewer descriptors.

The smallest average dissimilarity value calculated by applying Eq. 6 to the SOMs, corresponding to all combinations of the ordered pool of descriptors, indicates the maximum coherence and compactness of the information represented by that particular map. Thus, the process of including indices to form the most suitable set of molecular descriptors can conclude when the average dissimilarity measure of the corresponding map with the rest of maps stabilizes. The above dissimilarity analysis provides a systematic methodology of determining similarity among maps even when the dimension or number of indices of the input vectors may be very different. The indices of SOM with minimal average dissimilarity provide a good representation of all clusters of compounds formed in the nodes and constitute the most suitable set of molecular descriptors for QSPR modeling (Espinosa et al., 2002).

Fuzzy ART and Fuzzy ARTMAP

The selection of a neural architecture with predictive capabilities has been the subject of numerous studies in relation to time series analyses (Cybenko, 1989; Fessant et al. 1995; Giralt et al., 2000; Hornik et al. 1989), data mining (Agrawal et al., 1993; Bishop, 1995; Fayyad, 1996; Hertz et al., 1991), or pattern recognition (Agrawal et al., 1993; Ferre-Gine et al., 1996; Gutfreund and Mézard, 1988; Hecht-Nielsen, 1995; Hertz et al., 1991). The most commonly used architecture in the above fields and in other engineering applications has been the multilayer perceptron (Bishop, 1995; Hertz et al., 1991) with the learning mechanism of backpropagation. This approach is simple to use, has a sound mathematical foundation, and yields excellent results for most engineering applications. However, it is not suitable when pattern recognition or feature extraction capabilities are desired because relationships between variables in such networks are embedded within the weights in a distributed form (Bishop, 1995; Hecht-Nielsen, 1995; Hertz et al., 1991). In difficult problems involving pattern recognition, such as those found in the development of QSPRs for data sets of heterogeneous compound classes, it is advantageous to use neural network classifiers, as shown in a number of recent studies (Espinosa et al., 2000, 2001b, 2002, 2003; Yaffe et al., 2001, 2003) on QSPR development.

One of the most powerful classifiers is ARTMAP, which is based on adaptive resonance theory (ART) and has been shown to be capable of learning the dynamics of large-scale structures in a turbulent wake flow (Giralt et al., 2000). The application of fuzzy ARTMAP networks for QSPR development (Espinosa et al., 2002) has several advantages because of their capability to classify and analyze noisy and incomplete data sets with a reduced number of required model parameters and avoidance of local minima trapping (Carpenter et al., 1987, 1991, 1992). This architecture is also sufficiently transparent to allow continuous checking of the goodness of the classification during the training process, as well as how the relationships between the inputs and the outputs are established.

Adaptive resonance theory (ART) initially emerged from research on human cognitive information processing (Carpenter et al., 1987, 1991). Fuzzy ARTMAP (Carpenter et al., 1992) is one of the algorithms of the ART family that overcomes the *stability–plasticity dilemma* by creating as many new classes as needed to incorporate new information presented to the network in a stable manner while preserving the old knowledge contained in previously created classes. It associates prototypes of input patterns with their target outputs. The key feature is a control parameter that measures the similarity between the prototype patterns, stored in different network categories or classes, and any current input pattern. If the control parameter (vigilance) is not satisfied within a given accuracy, a new class or category is created during

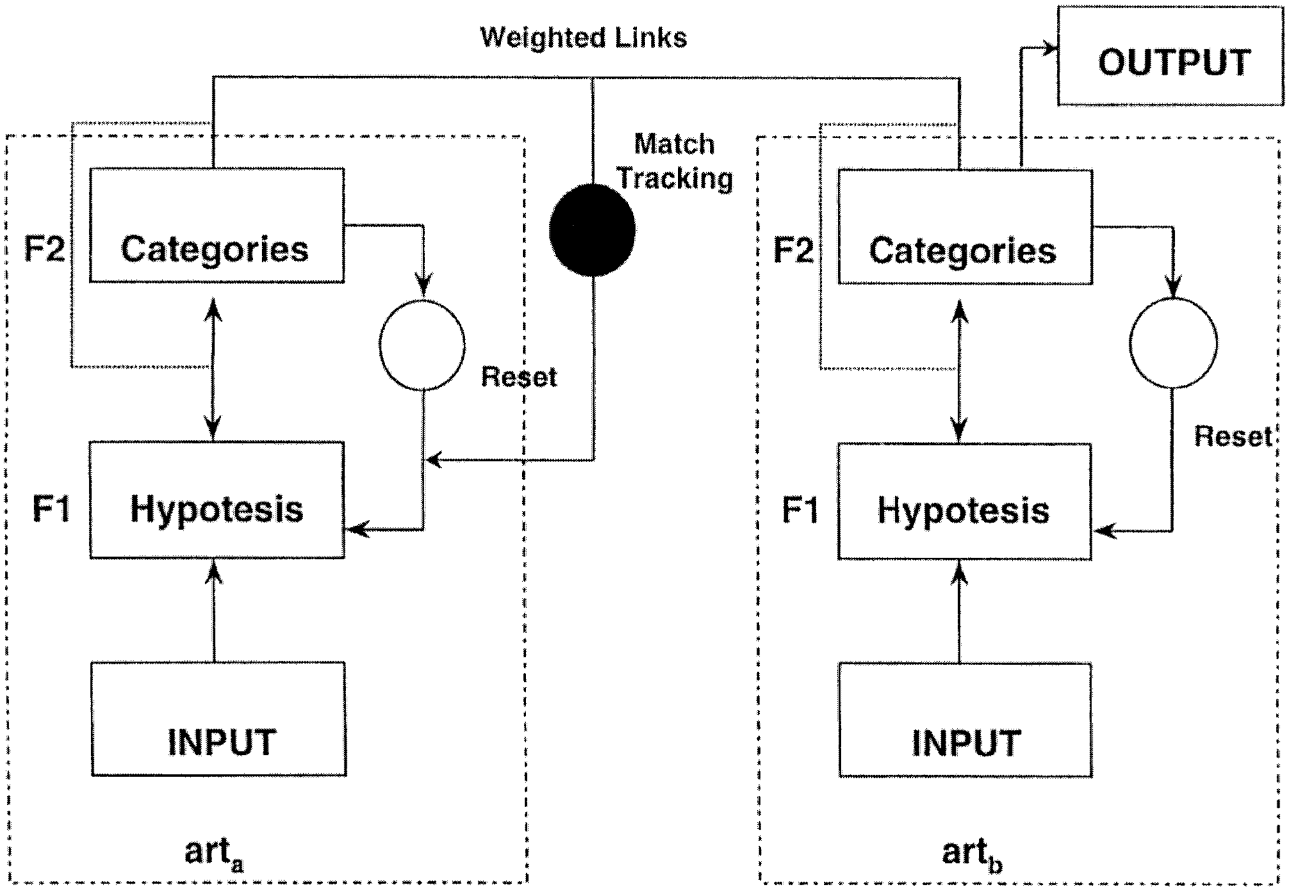

Figure 1. Fuzzy ARTMAP architecture.

learning. As a result, the number of categories (or prototype vectors) grows until a network structure has been built that is able to model the output based on the input data.

The architecture of fuzzy ARTMAP consists of two fuzzy ART modules, ART$_a$ and ART$_b$, interconnected by a map field, F_{ab}, as shown in Figure 1. The ART$_a$ (ART$_b$) module has two layers of nodes: F_{1a} (F_{1b}) is the input layer and F_{2a} (F_{2b}) is a dynamic layer where each node (class or category) encodes a prototype of a cluster of the input patterns. The number of such nodes can be increased when necessary. During supervised learning, an input pattern vector (molecular descriptors) is fed to the ART$_a$ module and the output vector (target variable) to ART$_b$. These are independently classified in each module. A map field (F_{ab}) adaptively associates prototype nodes in ART$_a$ with their respective target classes in ART$_b$. The node with the highest value of the activation function is selected as the winner, and all other nodes are suppressed in accordance with the winner-take-all rule (Carpenter et al., 1987, 1991, 1992).

The search cycle ends when either the current prototype is able to satisfy the vigilance parameter criterion (accuracy in the classification) or a new node is recruited in F_{2a} with the input pattern coded as its prototype pattern. During testing only the category F_{2b} layer is activated so that for any input presented to F_{1a} an output can be effected from F_{2b}, according to the predictive fuzzy ARTMAP system proposed by Giralt et al. (2000). The network reaches a resonant state when a category prototype vector matches the current input vector sufficiently such that the orienting subsystem will not send a reset signal to the F_2 layer. The network learns only in its resonant state, where it is capable of developing stable classification of arbitrary sequences of input patterns by self-organization. The voting strategy of training the system several times, using different orderings of the input set, not only improves predictions but can also be used to assign confidence estimates to competing predictions when dealing with small, noisy, or incomplete training sets. Further information on this issue and on the ART family of neural networks can be found elsewhere (Carpenter et al., 1987, 1991, 1992).

Data Set and Molecular Descriptors

The present aqueous infinite dilution activity coefficient γ^∞ data set consisted of 325 organic compounds, originally compiled by Sherman et al. (1996) and later also used by Mitchell and Jurs (1998), to correlate the aqueous phase $\ln\gamma^\infty$ with molecular descriptors. This heterogeneous data set includes hydrocarbons, alcohols, ethers, aldehydes, ketones, acids, halogenated hydrocarbons, amines, amides, nitriles, and compounds containing sulfur. The complete data set, with the corresponding $\ln\gamma^\infty$ values, is included in the supporting information (see Table 5 below). A subset of 280 compounds was selected for training (*tr*) and 45 compounds were used for testing (*te*) the fuzzy-ARTMAP–based QSPR models. Both data sets were selected by a fuzzy ART (Carpenter et al., 1987, 1991, 1992) neural system following the procedure described by Espinosa et al. (2001b), to ensure that the training set represented the complete data set. The 10 × 10 SOM trained to identify the most suitable molecular information, in relation to the target ln γ^∞ property, contained the classified information of the input data among the prototype vectors representing the nodes with clustered compounds (that is, the occupied nodes of the map). The map also contained interpolated information in the prototype vectors of empty nodes because they were also trained as neighbors of occupied (winning) nodes by Eqs. 2 and 3. Thus, the current work also explores the benefit of using the information of the 100 prototypes of the map for training the current fuzzy-ARTMAP–based QSAR model, in addition to the 280 compounds of the training set. In this latter case, however, the 10 × 10 SOM was built from training information only (280 compounds) to avoid contaminating these prototypes with test information.

An initial set of 23 molecular descriptors was first developed following the criteria and procedure described below. An optimal (that is, most suitable) subset consisting of the minimum number of descriptors that also provided the relevant information necessary to develop the present QSPR was then obtained. In this systematic selection process (Espinosa et al., 2002) all indices were classified according to the topology of their *C*-planes in the SOM (i.e., according to their similar capability to cluster compounds in the 10 × 10 nodes of the map). Thus, both the most representative indices from these *C*-plane classes and those with the highest correlation with the target variable can be selected.

The initial pool (that is, initial set of descriptors) was formed with the topological and quantum chemical descriptors considered in the studies of Medir and Giralt (1982) and Yaffe et al. (2001) to respectively predict lnγ^∞ of hydrocarbons in water and aqueous solubilities of diverse set of organic compounds. Topological indices provide information about the adjacency of the atoms in the molecular structure. Among those more frequently used in QSPRs are the Wiener index (Wiener, 1947), the connectivity indices used by Randic (Randic, 1975; Randic and Trinajstic, 1993), and the connectivity indices defined by Basak and Maguson (1988). For the present ln γ^∞ QSPR models the selected indices included the valence connectivity indices of order zero to four (${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi^v$, ${}^4\chi^v$) (Kier and Hall, 1976, 1985, 1999), the kappa index of second order (Kier and Hall, 1976), the sum of atomic numbers, and the Hansen indices (Hansen, 1979) of hydrogen, polarity, and dispersion. The above molecular indices were generated from the 2D molecular structures of the data set compounds using Molecular Modeling Pro (1998) 3.01 software.

The quantum chemical descriptors included the average molecular polarizability, dipole moment, number of filled molecular orbital levels, electron–nuclear attraction energy, nuclear–nuclear repulsion energy, exchange energy, and resonance energy. The above initial set of molecular descriptors (Yaffe et al., 2001) describes the interactions among the atoms in a molecule at the quantum level. These quantum indices were calculated by semiempirical Parametric Method 3 (PM3), with molecular structures optimized in 3-D. Additional descriptors that quantify 3-D similarity between molecules were generated by means of molecular quantum similarity measures (Amat and Carbó-Dorca, 1997; Carbó-Dorca and Besalú, 1998) determined by atomic shell approximation (ASA).

Three-dimensional similarity measures can be calculated from the scalar product of atomic density functions, as described elsewhere (Amat and Carbó-Dorca, 1997; Carbó-Dorca and Besalú, 1998). Briefly, quantum similarity matrices are formed from the scalar product (or projection) of the quantum atomic density functions of two molecules, using the metrics given by different quantum operators. The MQSM between two molecules A and B is given by the following integral

$$Z_{AB}(\Omega_\alpha) = \int\int \rho_A(\vec{r}_1)\Omega_\alpha(\vec{x}_1, \vec{r}_2)\rho_B(\vec{r}_2)d\vec{r}_1 d\vec{r}_2 \qquad (8)$$

where $\{\rho_A(r_1), \rho_B(r_2)\}$ are the density functions of each molecule and $\Omega_\alpha(r_1, r_2)$ is a positive definite operator. The particular operators considered in the present work constitute the overlap operator to measure similarity of molecular shape, and the coulomb operator to evaluate electrostatic similarities. Detailed information on the calculation of these quantum chemical descriptors and examples of previous applications in QSPR/QSAR modeling can be found elsewhere (Amat and Carbó-Dorca, 1997; Amat et al., 1998, 1999; Karelson et al., 1996; McWeeny, 1989).

To select the most relevant information contained in the large 325 × 325 overlap and coulomb quantum similarity matrices for the 325 compounds of the data set, it is common to use the diagonal elements of these matrices [that is, the quantum self-similarity measures (Carbó-Dorca and Besalú, 1998)]. Another alternative is to apply a dimensional reduction based on multidimensional scaling (Amat et al., 1998, 1999). This simplification, however, implies the use of principal-component analysis (PCA). In such an approach, it is difficult to discriminate the influence of the different descriptors and their relationship with the original similarity measures. It is noted that it is also possible to reduce the dimension of the MQS matrices, without losing relevant information, by projecting their respective elements onto SOMs, as explained in the previous section. Any one of these 2-D maps retains the topology (relationships) of the original matrix elements. Therefore, the resulting classification allows the selection of the more relevant elements representing each of the formed classes. This last alternative was adopted in this study.

The selection of the more relevant elements in any of the two overlap (*Ove*) and coulomb (*Cou*) 325 × 325 MQS matrices involved training two SOMs formed by square grids of 5 × 5 neurons or units by using the rows of the *Ove* matrix or of the *Cou* matrix as input patterns. After training, each class was identified by its prototype vector. In addition to the elements of the diagonals of both matrices (that is, the molecular self-similarities), herein identified as *Ove* and *Cou*, the following cross-similarities between compounds were selected: (1) those that were prototypes in both maps, if any, because this was an indication that they represented molecules within the set of γ^∞ data that behave similarly in front of the two different overlap and coulomb projections; (2) the prototypes of the classes with the highest population density of compounds because they represented the largest conglomerates of common behaviors. Following the above selection procedure, 1-4-cyclohexadiene (C_6H_8) was identified as the only prototype in both maps that could represent the largest group of chemicals with similar behavior. The MQSM of this compound can be accurately calculated because it contains no heavy atoms in its structure and the ASA approach is not biased. The corresponding cross-similarity measures or projections, using the overlap operator or the coulomb operator of all compounds with the 1-4-cyclohexadiene, have been respectively identified as molecular descriptors Ove_{C6H8} and Cou_{C6H8}.

Two additional prototypes of the more densely populated classes were also selected: *N*-methyl-2-pyrrolidone (C_5H_9NO) from the *Ove* matrix (Ove_{C5H9NO}) and 1-chloropropane (C_3H_7Cl) from the *Cou* matrix (Cou_{C3H7Cl}). The main assumption of the current approach is that descriptors from the same class contribute similar type of information to the QSPRs. Thus, the indices *Ove*, *Cou*, Ove_{C6H8}, Cou_{C6H8}, Ove_{C5H9NO}, and Cou_{C3H7Cl} were included in the pool of descriptors to incorporate the more relevant information contained in the 325 × 325 MQS matrices. It should be noted that the natural logarithms of the above six MQS indices were used in all calculations to reduce the differences in ranges, and that values reported in tables and figures are expressed in this format.

Following the calculation of descriptors and analysis as specified above, the initial pool of molecular information was formed by 23 descriptors: five valence connectivity indices of order zero to four, kappa index of second order, sum of atomic numbers, three Hansen indices, average polarizability, dipole moment, number of filled levels, electron–nuclear energy, nuclear–nuclear repulsion energy, exchange energy, resonance energy, and the six MQS matrix defined above.

Studies on QSPRs often involve the generation of large sets of molecular descriptors. Although neural networks can deal with a large set of input parameters, it is prudent to seek the smallest possible descriptor subset that would retain a reasonable accuracy of the QSPR. In the present study, for example, one may argue that because the MQS matrix is calculated, for any given chemical data set, using the metrics given by all the relevant quantum operators, it should contain all relevant structural information. Therefore, it was hypothesized that quantum similarity indices alone would be sufficient to establish reasonably accurate QSPR models. To test the above hypothesis, the same procedure for selecting the most suitable set of descriptors, from the initial pool of molecular descriptors, was applied to the initial set formed only by the six MQS matrices extracted from the quantum similarity matrix (that is, to *Cou*, Cou_{C6H8}, Cou_{C3H7Cl}, *Ove*, Ove_{C6H8}, and Ove_{C5H9NO}).

Results and Discussion

Classification of chemicals

The optimal SOM grid size, which was found to be 10 × 10, classified the 325 compounds characterized by a 24-dimensional vector (the above 23 molecular descriptors plus the target variable $\ln\gamma^{\infty}$) into 80 nodes (neurons) of the map during training, with an adequate population density, as illustrated by the component plane for $\ln\gamma^{\infty}$ depicted in Figure 2. This figure identifies 13 chemical families that cluster into the 80 nodes with clustered compounds (occupied nodes) according to both their generic family label and molecular similarity. Gray levels indicate the clustering compactness in each node, as measured by the average distance (*n*) between the clustered compounds. Table 1 lists the main characteristics of the 80 occupied nodes, along with an indication on how compounds cluster into them, according the molecular information provided by the 23 descriptors and the corresponding $\ln\gamma^{\infty}$ values.

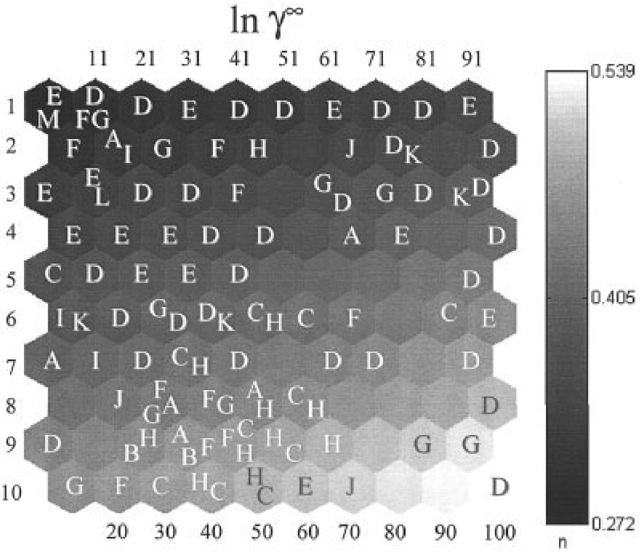

Figure 2. Distribution of 13 families of organic compounds over the component plane for $\ln\gamma^{\infty}$.

(A) Monoaromatic hydrocarbons (20 compounds); (B) polyaromatic hydrocarbons (5 compounds); (C) aliphatic hydrocarbons (36 compounds); (D) hydrocarbons with oxygen substituents (123 compounds); (E) halogenated aliphatic hydrocarbons (66 compounds); (F) aromatic hydrocarbons with nitrogen and/or oxygen substituents (11 compounds); (G) hydrocarbons with sulfur substituents and/or oxygen and/or nitrogen (19 compounds); (H) cyclic hydrocarbons (15 compounds); (I) aromatic hydrocarbons with oxygen substituents (8 compounds); (J) halogenated aromatic hydrocarbons (9 compounds); (K) heterocyclic hydrocarbons (10 compounds); (L) aliphatic hydrocarbons with halogen and oxygen substituents (2 compounds); (M) aliphatic hydrocarbons with nitro and halogen groups (1 compound). The gray levels indicate the clustering intensity of the $\ln\gamma^{\infty}$ data set into the SOM nodes measured by the average distance (*n*) between members of each cluster.

A close inspection of Figure 2 and Table 1 reveals several important features about the data set, molecular information embedded in the molecular descriptors, and classification process, as listed below:

(1) The SOM captures the distinctiveness of the 13 chemical families (A–M) as characterized by the 23 molecular descriptors and the $\ln\gamma^{\infty}$ values, with a dominant presence in the nodes of clusters formed by the two most populated chemical families D (hydrocarbons with oxygen substituents) and E (halogenated aliphatic hydrocarbons). The neighborhood between clusters is also consistent with the formal segregation of the data set into 13 chemical families A–M.

(2) Nearly 70% of the 80 occupied nodes (Figure 2) contain compounds of the same chemical family. The remaining 30% of the occupied nodes are formed by similar compounds from other neighbor ones. For example, clusters (nodes) 40, 46, 49, 58, and 59 are formed by chemicals from families C (aliphatic hydrocarbons) and H (cyclic hydrocarbons).

(3) Nodes consistently cluster chemicals with similar structures. However, there are significant differences between the degree of membership by which nodes cluster chemicals and of the span of the corresponding $\ln\gamma^{\infty}$ values. This is clearly observed in Table 1, where the square of the cluster average

Table 1. Information on Self-Organization of Chemicals in Terms of the 13 Chemical Families of the Data Set That Cluster into the 80 Occupied Nodes in the 10 × 10 SOM, as Illustrated in Figure 2 for the Component Plane Corresponding to the Target Property ln γ^{∞}*

Identification of occupied SOM nodes	Chemical family (see caption of Fig. 2)	Compound with the minimum atomic number	Compound with the maximum atomic number	Number of compounds per node	n^2 for U-matrix	$[\ln \gamma^{\infty}]_{min}$	$[\ln \gamma^{\infty}]_{max}$
1	E	C2H3Cl1	C3H7I1	12	0.7828	5.53	11.86
1	M	C1N1O2Cl3	C1N1O2Cl3	1	0.7828	8.64	8.64
2	F	C5H5N1	C6H7N1	2	0.6633	2.99	3.89
3	E	C4H9Cl1	C5H11Br1	2	0.5584	8.62	11.10
4	E	C1F4	C1F2Cl2	2	0.4682	10.02	12.63
5	C	C5H8	C5H8	1	0.3927	7.79	7.79
6	I	C5H10O1	C5H10O1	1	0.3317	4.36	4.36
6	K	C4H8O2	C4H8O2	1	0.3317	1.69	1.69
7	A	C6H6	C9H12	2	0.2855	7.82	11.67
9	D	C2H4O2	C9H18O1	6	0.2369	2.74	9.70
10	G	C2H5N1O2	C3H7N1O2	2	0.2346	4.48	5.70
11	D	C3H6O2	C3H6O2	1	0.7460	3.86	3.86
11	F	C7H5N1	C7H7N1O2	2	0.7460	7.46	8.87
11	G	C6H11N1	C6H11N1	1	0.7460	7.24	7.24
12	A	C7H8	C10H14	7	0.6265	9.13	11.82
12	I	C7H8O1	C8H10O1	3	0.6265	5.62	9.66
13	E	C1F3Cl1	C1F3Cl1	1	0.5216	11.07	11.07
13	L	C2H3O1F3	C2H3O1F3	1	0.5216	2.16	2.16
14	E	C1H2F2	C2H2Br4	5	0.4314	5.72	10.29
15	D	C5H10O2	C17H36O1	6	0.3559	4.84	21.30
16	D	C6H14O1	C6H12O2	2	0.2949	5.14	6.40
17	I	C7H8O1	C7H6O2	4	0.2487	5.51	8.20
18	J	C6H4Cl2	C6Cl6	2	0.2171	11.53	21.94
20	F	C5H9N1O1	C5H9N1O1	1	0.1978	-0.99	-0.99
21	D	C6H12O2	C6H12O2	1	0.7129	6.70	6.70
22	G	C4H11N1	C5H9N1	2	0.5934	1.69	6.00

*The average distance of compounds within each cluster of the nodes in the U-matrix is given by n^2.

Table 1. Information on Self-Organization of Chemicals in Terms of the 13 Chemical Families of the Data Set That Cluster into the 80 Occupied Nodes in the 10 × 10 SOM, as Illustrated in Figure 2 for the Component Plane Corresponding to the Target Property ln $\gamma^{\infty *}$ (Continued)

Identification of occupied SOM nodes	Chemical family (see caption of Fig. 2)	Compound with the minimum atomic number	Compound with the maximum atomic number	Number of compounds per node	n^2 for U-matrix	$[\ln \gamma^{\infty}]_{\min}$	$[\ln \gamma^{\infty}]_{\max}$
23	D	C 4 H 10 O 1	C 5 H 12 O 1	2	0.4885	4.23	5.71
24	E	C 1 H 1 F 3	C 4 F 8	8	0.3983	5.98	12.31
25	E	C 3 H 5 Cl 1	C 3 H 6 Br 2	6	0.3227	6.97	8.97
26	D	C 12 H 26 O 2	C 8 H 18 O 1	8	0.2618	5.18	18.77
26	G	C 4 H 10 S 1	C 4 H 10 S 1	1	0.2618	9.03	9.03
27	D	C 6 H 14 O 1	C 9 H 18 O 1	6	0.2155	5.63	9.99
27	L	C 4 H 9 O 1 Cl 1	C 4 H 9 O 1 Cl 1	1	0.2155	7.54	7.54
28	A	C 8 H 8	C 9 H 12	5	0.1839	9.80	11.67
28	F	C 3 H 7 N 1 O 2	C 6 H 5 N 1 O 2	2	0.1839	8.17	9.37
28	G	C 3 H 7 N 1 O 2	C 3 H 7 N 1 O 2	1	0.1839	5.69	5.69
29	H	C 6 H 12	C 6 H 12	1	0.1670	11.60	11.60
29	B	C 9 H 10	C 9 H 10	1	0.1670	11.01	11.01
30	C	C 6 H 12	C 6 H 14	2	0.1646	11.39	12.76
31	E	C 3 H 7 Br 1	C 1 H 2 Br 2	2	0.6834	6.75	7.64
32	F	C 4 H 7 N 1	C 4 H 7 N 1	1	0.5638	4.76	4.76
33	D	C 6 H 14 O 1	C 8 H 18 O 1	2	0.4590	6.44	10.76
34	D	C 5 H 12 O 1	C 6 H 14 O 1	5	0.3688	4.73	7.75
35	E	C 1 H 3 F 1	C 2 Cl 6	7	0.2932	5.98	14.31
36	D	C 5 H 12 O 1	C 18 H 38 O 1	21	0.2323	4.85	23.34
36	K	C 6 H 10 O 1	C 6 H 6 O 2	2	0.2323	3.99	4.85
37	C	C 5 H 10	C 7 H 12	5	0.1860	9.45	11.46
37	H	C 6 H 10	C 7 H 12	2	0.1860	10.25	11.30
38	F	C 6 H 7 N 1	C 7 H 15 N 1	2	0.1544	4.84	4.99
38	G	C 6 H 15 N 1	C 6 H 15 N 1	2	0.1544	4.87	4.90
39	A	C 10 H 14	C 10 H 14	2	0.1374	12.44	14.58
39	B	C 11 H 10	C 12 H 12	4	0.1374	12.55	13.90
39	F	C 7 H 7 N 1 O 3	C 7 H 7 N 1 O 3	1	0.1374	8.52	8.52
40	C	C 5 H 10	C 8 H 16	7	0.1351	8.04	14.65
40	H	C 8 H 16	C 8 H 16	1	0.1351	13.85	13.85
41	D	C 5 H 10 O 1	C 6 H 12 O 1	2	0.6575	4.67	6.68
42	H	C 6 H 8	C 7 H 12	2	0.5380	8.60	11.54
43	F	C 7 H 9 N 1	C 7 H 9 N 1	1	0.4331	5.91	5.91
44	D	C 5 H 10 O 2	C 9 H 18 O 2	2	0.3429	5.28	9.87

Table 1. Information on Self-Organization of Chemicals in Terms of the 13 Chemical Families of the Data Set That Cluster into the 80 Occupied Nodes in the 10 × 10 SOM, as Illustrated in Figure 2 for the Component Plane Corresponding to the Target Property ln γ^{∞}* (Continued)

Identification of occupied SOM nodes	Chemical family (see caption of Fig. 2)	Compound with the minimum atomic number	Compound with the maximum atomic number	Number of compounds per node	n^2 for U-matrix	[ln γ^{∞}]$_{min}$	[ln γ^{∞}]$_{max}$
45	D	C 5 H 12 O 1	C 6 H 14 O 1	2	0.2673	4.88	5.08
46	C	C 5 H 10	C 5 H 12	4	0.2064	9.86	11.70
46	H	C 5 H 10	C 5 H 10	1	0.2064	10.12	10.12
47	D	C 9 H 20 O 1	C 11 H 24 O 2	6	0.1601	6.41	12.38
48	A	C 9 H 12	C 10 H 14	3	0.1285	11.16	12.96
48	H	C 7 H 14	C 7 H 14	1	0.1285	12.11	12.11
49	C	C 7 H 14	C 7 H 16	6	0.1116	12.80	14.51
49	F	C 8 H 17 N 1	C 8 H 17 N 1	1	0.1116	6.85	6.85
49	H	C 7 H 14	C 7 H 14	1	0.1116	12.80	12.80
50	C	C 6 H 14	C 8 H 18	5	0.1093	12.35	16.02
50	H	C 7 H 8	C 8 H 12	2	0.1093	8.98	11.70
51	D	C 2 H 6 O 1	C 8 H 16 O 2	7	0.6353	1.32	9.43
56	C	C 4 H 10	C 4 H 10	1	0.1842	10.87	10.87
58	C	C 5 H 8	C 5 H 8	1	0.1063	8.68	8.68
58	H	C 5 H 8	C 6 H 12	2	0.1063	8.86	11.29
59	C	C 6 H 12	C 7 H 12	2	0.0893	9.23	11.70
59	H	C 5 H 8	C 5 H 8	1	0.0893	13.58	13.58
60	E	C 4 H 9 Cl 1	C 5 H 11 Cl 1	4	0.0870	8.94	10.38
61	E	C 2 H 1 Cl 3	C 2 H 1 Cl 3	1	0.6167	9.08	9.08
62	J	C 6 H 5 F 1	C 6 H 5 Br 1	4	0.4972	8.48	11.10
63	D	C 2 H 4 O 2	C 7 H 16 O 1	4	0.3923	-0.08	8.09
63	G	C 2 H 6 O 1 S 1	C 2 H 6 O 1 S 1	1	0.3923	-2.41	-2.41
64	A	C 8 H 10	C 9 H 12	2	0.3021	10.41	11.53
66	F	C 2 H 3 N 1	C 4 H 9 N 1 O 1	2	0.1656	0.04	2.41
67	D	C 1 H 4 O 1	C 4 H 10 O 1	7	0.1193	0.33	3.92
69	H	C 5 H 8	C 5 H 8	1	0.0708	8.82	8.82
70	J	C 7 H 7 Cl 1	C 6 H 5 I 1	3	0.0684	10.37	10.90
71	D	C 5 H 10 O 1	C 7 H 14 O 1	2	0.6017	5.39	7.24
72	D	C 7 H 14 O 2	C 7 H 14 O 2	1	0.4822	8.00	8.00
72	K	C 8 H 14 O 2	C 8 H 14 O 2	1	0.4822	7.91	7.91
73	G	C 1 H 3 N 1 O 2	C 3 H 7 N 1 O 2	2	0.3773	-0.19	3.45
74	E	C 2 H 2 Cl 2	C 1 F 1 Cl 3	4	0.2871	7.14	13.79
77	D	C 12 H 26 O 1	C 12 H 26 O 1	1	0.1044	15.31	15.31

Table 1. Information on Self-Organization of Chemicals in Terms of the 13 Chemical Families of the Data Set That Cluster into the 80 Occupied Nodes in the 10 × 10 SOM, as Illustrated in Figure 2 for the Component Plane Corresponding to the Target Property ln γ^{∞}* (Continued)

Identification of occupied SOM nodes	Chemical family (see caption of Fig. 2)	Compound with the minimum atomic number	Compound with the maximum atomic number	Number of compounds per node	n^2 for U-matrix	$[\ln \gamma^{\infty}]_{min}$	$[\ln \gamma^{\infty}]_{max}$
81	D	C 4 H 8 O 1	C 4 H 8 O 1	1	0.5904	3.88	3.88
83	D	C 7 H 14 O 2	C 7 H 14 O 2	1	0.3660	8.29	8.29
86	C	C 4 H 10	C 4 H 10	1	0.1393	11.10	11.10
89	F	C 6 H 7 N 1	C 6 H 7 N 1	1	0.0445	3.74	3.74
89	G	C 3 H 5 N 1	C 6 H 15 N 1	3	0.0445	3.56	4.77
91	E	C 2 H 4 Cl 2	C 3 H 6 Cl 2	10	0.5828	6.46	10.49
92	D	C 1 H 2 O 1	C 6 H 12 O 1	8	0.4632	1.03	6.02
93	D	C 1 H 2 O 2	C 6 H 12 O 1	3	0.3584	-0.33	5.06
93	K	C 2 H 4 O 1	C 2 H 4 O 1	1	0.3584	1.83	1.83
94	D	C 5 H 12 O 1	C 6 H 12 O 2	5	0.2682	5.29	6.70
95	D	C 2 H 4 O 1	C 7 H 14 O 2	5	0.1926	1.37	9.02
96	E	C 2 H 3 Cl 3	C 2 F 5 Cl 1	2	0.1317	7.31	11.90
97	D	C 4 H 10 O 1	C 7 H 14 O 2	5	0.0854	4.57	8.03
98	D	C 3 H 6 O 1	C 3 H 6 O 1	1	0.0538	2.56	2.56
99	G	C 1 H 4 S 1	C 4 H 10 S 1	4	0.0368	4.72	7.38
100	D	C 5 H 12 O 1	C 5 H 12 O 1	1	0.0345	4.91	4.91

distance for each node in the **U**-matrix, (n^2), and the corresponding minimum and maximum values of lnγ^{∞} are listed for the 80 occupied nodes. The span of lnγ^{∞} is very noticeable and significant in the more populated clusters (number of compounds ≥ 5) (that is, nodes 1, 9, 12, 14, 15, 24–28, 35–40, 46, 47, 49–51, 63, and 67). It should be noted that lnγ^{∞} is only one component in the 24-dimensional vectors (23 descriptors plus lnγ^{∞}) characterizing the nodes in the map. It is also well known that it is difficult to account for heteroatoms, particularly halogens, within any given chemical structure with molecular descriptors (Basak and Maguson, 1988; Carbó-Dorca and Besalú, 1998). The dispersion (n^2) of compounds around nodes in Table 1 clearly illustrates this difficulty. The dispersion of compounds is smallest ($0.1093 \leq n^2 \leq 0.2064$) in the more populated nodes 37, 40, 46, and 50, which cluster only families of hydrocarbons with no substituents containing heteroatoms (families A, B, C, H, and K). This dispersion increases ($0.0854 \leq n^2 \leq 0.6353$) in the populated nodes 9, 12, 15, 26, 28, 36, 39, 47, 49, 51, 63, 67, 92, 94, 95, and 97 that cluster hydrocarbons with oxygen, nitrogen, and/or sulfur substituents (families D, F, G, and I). The largest dispersion ($0.2155 \leq n^2 \leq 0.7828$) is encountered in nodes 1, 14, 24, 25, 27, 35, and 91 with hydrocarbon families with halogen substituents (E, J, L, and M).

Finally, it should be noted that the component plane for lnγ^{∞} in Figure 2 is compact, given that three or more nodes with clustered compounds (occupied nodes) usually surround the empty ones, implying that the latter were also updated by means of Eqs. 2 and 3 during training. Thus, vectors of the empty nodes are bound to be a good source of additional (interpolated) information for training the QSAR models with the aim at increasing generalization during testing or predictive operation mode, even at the expense of introducing some noise in the training set. In the current study, the 100 node vectors of a 10 × 10 SOM, built only from information of the training set, were also used as additional information for training the fuzzy-ARTMAP–based QSPR model. In such a way, the interpolated information obtained from this SOM may enhance the classification capabilities of any new information presented to the predictive fuzzy ARTMAP neural system during testing.

Selection of most suitable descriptors

The component planes of the 10 × 10 SOM, built from molecular and target lnγ^{∞} information of the complete set of 325 compounds, were used to select the most suitable set of molecular descriptors from the initial pool of 23 topological and quantum chemical descriptors. The selection was carried out according to the descriptors' contribution to the classification of the 325 compounds in relation to lnγ^{∞} and following the methodology of index selection proposed by Espinosa et al. (2002). Accordingly, the 23 component *C*-planes shown in Figure 3 are grouped into six classes according to the similarity in the contribution of descriptors to the topological organization of the map. The six classes, identified with the Roman numerals I–VI, were formed by curvilinear component analysis. A visual inspection of Figure 3 shows that the classification into six classes of similar descriptors is consistent with the similarity shown by their respective *C*-planes. The first three classes include the topological (connectivities, kappa index, and the sum of atomic numbers) and quantum information (average polarizability, number of filled levels, and nuclear–nuclear repulsion). Because each of these classes contains descriptors with 2-D molecular size–related information they should help to explain how the increase in chain length or hydrophobicity is related with $\ln\gamma^{\infty}$. The dipole moment is appropriately grouped in class IV with the Hansen indices, whereas the remaining atomic energies and all MQS matrices are classified into classes V and VI, respectively. The classification of *C*-planes shown in Figure 3 is also consistent with the covariances between descriptors and target variable listed in Table 2. The highest covariance with $\ln\gamma^{\infty}$ corresponds to $^1\chi^v$ (class I), whereas the first descriptor with a *C*-plane organization is significantly different from cluster I and with a still high covariance, is μ (class IV). It is interesting to note that the use of these two descriptors proved to be very effective in the linear correlations of $\ln\gamma^{\infty}$ proposed by Medir and Giralt (1982) for hydrocarbons.

The ordering of indices, selected descriptors, and the basis for the descriptors' selection process are summarized in Table 3. The first indices selected from the initial pool to form the most suitable set were those with highest absolute covariance with $\ln\gamma^{\infty}$ in each of the six classes of Figure 3 (that is, $^1\chi^v$, *N*, $^4\chi^v$, μ, *ENA*, and *Cou*). In this case, all six indices were selected because their covariances are higher than the average value for the whole pool of 23 descriptors, as may be deduced from the information provided in Table 2. As can be verified from Table 3, the successive addition of descriptors in the process of selecting the most suitable set causes a decrease in the value of the dissimilarity function, defined by Eq. 6, from 0.703 for $^1\chi^v$ to 0.229, when the six indices mentioned above were incorporated. This significant decrease highlights the importance of adding indices according to their topological impact on the studied QSPR. Further addition of $^3\chi^v$, *NFL*, *AP*, $^2\chi^v$, and *NNR*, chosen according to their absolute covariance with $\ln\gamma^{\infty}$, lowered the dissimilarity to the minimum value of 0.190. It is important to recognize that molecular information beyond the point of minimum dissimilarity (Table 3) can be redundant, it may introduce noise if added to the QSPR, and it could also induce errors attributed to conflicting information with the previous and more relevant indices of the most suitable set. Values of the 11 descriptors, included in the most suitable set (constructed from the initial pool of descriptors), for the compounds in the overall data set, are provided in the supporting information (see Table 5 below). The final set of the most suitable 11 descriptors identified in the present work includes nine of the 11 descriptors used by Yaffe et al. (2001) to predict the aqueous solubility of organic compounds. These authors used a nonlinear selection method, based on a dynamic backpropagation neural network genetic algorithm, to identify 11 suitable descriptors from an initial set of 30 descriptors. In the present set, the RE and EE descriptors used by Yaffe et al. (2001) were substituted here by *Cou* and *N*. Also, it is noted that the current incorporation of the sum of atomic numbers reinforces the well-known dependency of $\ln\gamma^{\infty}$ with chain length (Mackay and Shiu, 1977; Medir and Giralt, 1982).

To assess QSPR performance when solely using significant and coherent quantum chemical information, the above selection procedure was also applied to the six MQS descriptors (Table 3), which constitute class VI shown in Figure 3. Two classes were identified, one formed by the three coulomb descriptors and the other including all overlap information. The most suitable set, formed by *Cou*, Cou_{C6H8}, Cou_{C3H7Cl}, and *Ove*, was selected based on the minimum average dissimilarity value. It is interesting to note that the *Ove* descriptor is needed despite its low covariance with $\ln\gamma^{\infty}$ (see Table 2), which is below the average for the complete MQS set, given that it is the only source of quantum information for molecular shape available in this set of most suitable MQS matrices.

QSPR models

Fuzzy-ARTMAP–based QSPR for $\ln\gamma^{\infty}$ were developed using either the most suitable set of 11 descriptors from the initial set listed in Table 3 [$^1\chi^v$, *N*, $^4\chi^v$, μ, *ENA*, *Cou*, $^3\chi^v$, *NFL*, *AP*, $^2\chi^v$, and *NNR*] or solely using the most significant MQS matrix [*Cou*, Cou_{C6H8}, Cou_{C3H7Cl}, and *Ove*]. In each case, training was carried out by using either the training set of 280 compounds or this subset complemented with the interpolated information obtained from a 10 × 10 SOM, trained only with the molecular descriptors of these 280 compounds and the target variable $\ln\gamma^{\infty}$, as discussed previously.

The fuzzy-ARTMAP algorithm is a neural classifier and only the presentation of new information during training will trigger the creation of new classes in ART_b according to the required accuracy for the target variable $\ln\gamma^{\infty}$, and correspondingly the creation of an equal or larger number of classes for the training compounds in ART_a. This ensures the generation of many (compounds in ART_a) to one ($\ln\gamma^{\infty}$ value in ART_b) relationships in the network. The accuracy in ART_b is set by a fixed vigilance parameter ρ_b, whereas that of ART_a increases dynamically from zero according to the needs of classification and of the *if–then* relationships. Thus, overtraining in fuzzy ARTMAP will manifest itself in the creation of an unnecessarily large architecture (Georgiopoulos et al., 2001) and as degradation in generalization capabilities of the predictive model (that is, poor performance on samples or data not present in the training set).

In the current study the generation of a reasonable number of classes has been monitored and reasonable generalization ensured by choosing a training set (Espinosa et al., 2001b) of 280 compounds that was representative of the complete set of 325 compounds. The alternative of applying cross-validation procedures (Koufakou et al., 2001) during the training stage did not yield significantly different results since the predictive capabilities of the current model are very demanding, that is, the number of chemical families and the range of $\ln\gamma^{\infty}$ values are both very large. It should be noted that the current approach would allow the construction of a nearly zero biased network (close to zero training error) that will also have good generalization capabilities and yield a finite small variance for the test set. This is possible in the framework of the bias/variance dilemma (Geman et al., 1992) since any test data presented to the network that are not well represented by the current training set (that is, test data not included in the current test set of 45 compounds) would be labeled as "unable to classify" in ART_a within the classification error tolerance and considered as a new candidate for training. Very large or infinite errors could be assumed for these hypothetical, unclassifiable test data so that the product of the error of the training set times the error of a test set, formed by the classifiable and unclassifiable compounds, would still be finite. In addition, the use of SOM prototypes, for enhanced learning in a fuzzy-ARTMAP–based

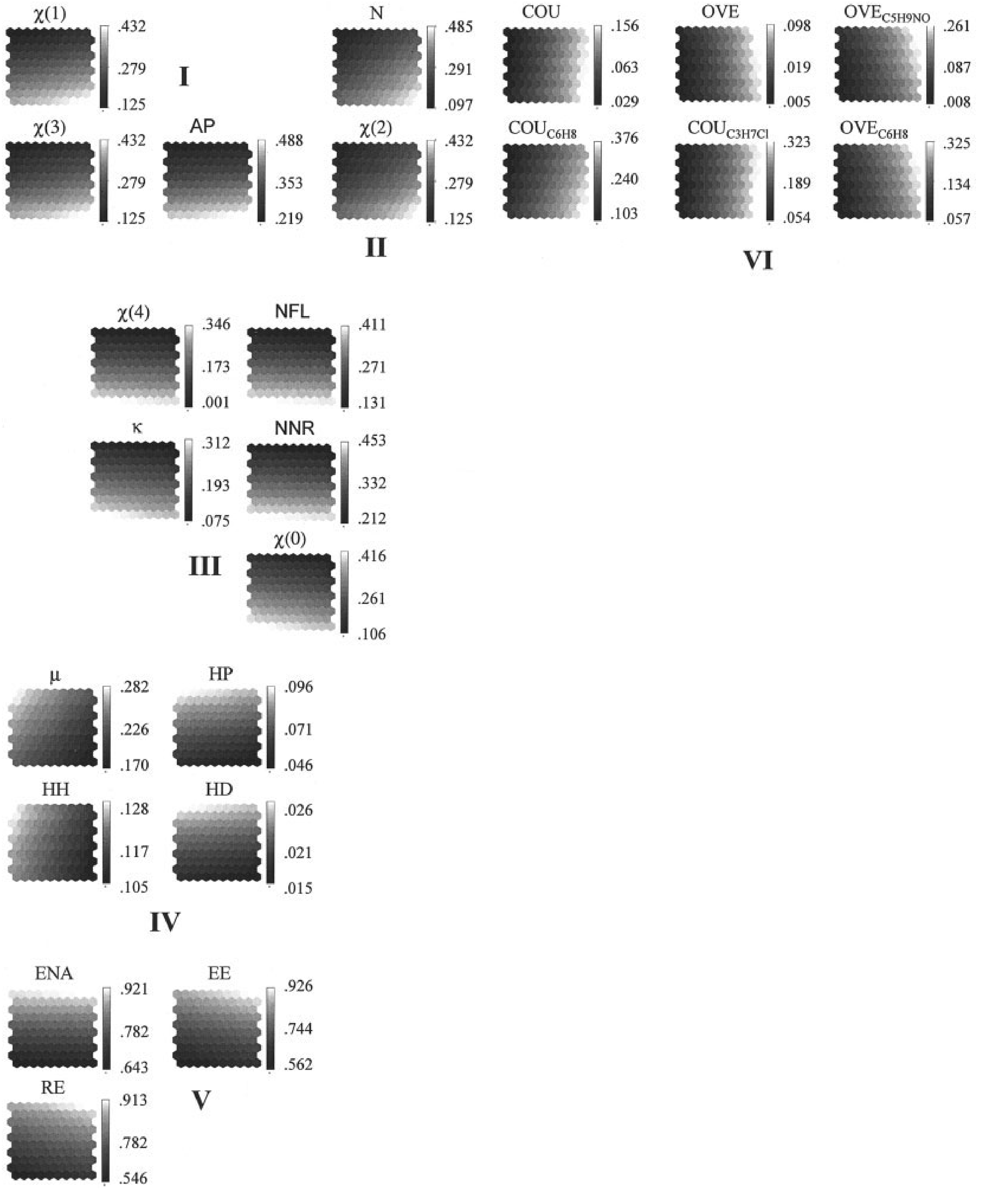

Figure 3. Topological classification of the component planes for the 23 molecular descriptors into six classes. The gray levels indicate distances between compounds within clusters in the nodes of the SOM.

Table 2. Covariance Matrix for the 23 Molecular Descriptors, Grouped According to the Classification of Figure 3, and Experimental Infinity Dilution Activity Coefficient, ln γ^{∞}

ID		$^1\chi^v$	$^3\chi^v$	AP	N	$^2\chi^v$	$^4\chi^v$	NFL	NNR	κ	$^0\chi^v$	μ	HP	HH	HD	ENA	EE	RE	Cou	Cou C_6H_8	Cou $C_3H_7Cl_1$	Ove	Ove C_6H_8	Ove $C_5H_9N_1O_1$	lnγ^{∞}
I	$^1\chi^v$	1.000	.846	.793	.843	.766	.823	.872	.834	.764	.850	-.149	-.254	-.033	-.108	-.845	-.821	-.809	.551	.535	.398	.130	.044	-.017	.632
I	$^3\chi^v$	.846	1.000	.744	.729	.645	.756	.773	.745	.523	.675	-.214	-.215	-.041	-.073	-.749	-.691	-.688	.443	.480	.307	.051	.006	-.101	.584
I	AP	.793	.744	1.000	.678	.607	.756	.771	.726	.539	.650	-.133	-.227	-.071	-.119	-.723	-.752	-.756	.405	.481	.265	-.023	-.003	-.151	.541
II	N	.843	.729	.678	1.000	.868	.568	.726	.691	.519	.509	-.104	-.133	.005	-.051	-.700	-.477	-.485	.841	.817	.719	.499	.404	.351	.583
II	$^2\chi^v$	.766	.645	.607	.868	1.000	.486	.600	.576	.326	.536	-.195	-.191	-.040	-.083	-.580	-.457	-.443	.692	.713	.580	.362	.308	.228	.527
III	$^4\chi^v$	.823	.756	.756	.568	.486	1.000	.794	.781	.623	.740	-.143	-.229	-.067	-.058	-.783	-.815	-.822	.237	.276	.084	-.157	-.189	-.283	.553
III	NFL	.872	.773	.771	.726	.600	.794	1.000	.974	.754	.831	-.035	-.181	.040	-.051	-.980	-.876	-.902	.343	.382	.169	-.140	-.217	-.260	.543
III	NNR	.834	.745	.726	.691	.576	.781	.974	1.000	.713	.806	-.020	-.152	.065	-.051	-.980	-.850	-.879	.300	.334	.127	-.161	-.244	-.277	.496
III	κ	.764	.523	.539	.519	.326	.623	.754	.713	1.000	.787	.048	-.161	.058	-.064	-.725	-.768	-.761	.117	.081	.034	-.135	-.279	-.236	.437
III	$^0\chi^v$	.850	.675	.650	.509	.536	.740	.831	.806	.787	1.000	-.086	-.255	.058	-.119	-.818	-.949	-.929	.134	.133	-.015	-.289	-.362	-.374	.435
IV	μ	-.149	-.214	-.133	-.104	-.195	-.143	-.035	-.020	.048	-.086	1.000	.603	.351	.038	.016	.063	.025	-.101	-.140	-.094	-.037	-.071	.101	-.535
IV	HP	-.254	-.215	-.227	-.133	-.191	-.229	-.181	-.152	-.161	-.255	.603	1.000	.637	.064	.153	.244	.220	-.001	-.075	-.017	.033	.023	.068	-.494
IV	HH	-.033	-.041	-.071	.005	-.040	-.067	.040	.065	.058	.058	.351	.637	1.000	.014	-.069	-.026	-.044	-.043	-.078	-.058	-.050	-.087	.016	-.491
IV	HD	-.108	-.073	-.119	-.051	-.083	-.058	-.051	-.051	-.064	-.119	.038	.064	.014	1.000	.050	.107	.093	-.010	-.022	-.019	-.003	-.011	-.011	-.002
V	ENA	-.845	-.749	-.723	-.700	-.580	-.783	-.980	-.980	-.725	-.818	.016	.153	-.069	.050	1.000	.856	.884	-.309	-.340	-.136	.152	.237	.267	-.508
V	EE	-.821	-.691	-.752	-.477	-.457	-.815	-.876	-.850	-.768	-.949	.063	.244	-.026	.107	.856	1.000	.993	-.072	-.118	.090	.390	.416	.476	-.451
V	RE	-.809	-.688	-.756	-.485	-.443	-.822	-.902	-.879	-.761	-.929	.025	.220	-.044	.093	.884	.993	1.000	-.073	-.129	.093	.402	.426	.484	-.439
VI	Cou	.551	.443	.405	.841	.692	.237	.343	.300	.177	.134	-.101	-.001	-.043	-.010	-.309	-.072	-.073	1.000	.954	.974	.849	.788	.745	.425
VI	Cou C_6H_8	.535	.480	.481	.817	.713	.276	.382	.334	.081	.133	-.140	-.075	-.078	-.022	-.340	-.118	-.129	.954	1.000	.919	.733	.749	.638	.419
VI	Cou $C_3H_7Cl_1$	.398	.307	.265	.719	.580	.084	.169	.127	.034	-.015	-.094	-.017	-.058	-.019	-.136	.090	.093	.974	.919	1.000	.922	.869	.848	.334
VI	Ove	.130	.051	-.023	.499	.362	-.157	-.140	-.161	-.135	-.289	-.037	.033	-.050	-.003	.152	.390	.402	.849	.733	.922	1.000	.923	.954	.170
VI	Ove C_6H_8	.044	.006	-.003	.404	.308	-.189	-.217	-.244	-.279	-.362	-.071	.023	-.087	-.011	.237	.416	.426	.788	.749	.869	.923	1.000	.900	.121
VI	Ove $C_5H_9N_1O_1$	-.017	-.101	-.151	.351	.228	-.283	-.260	-.277	-.236	-.374	.101	.068	.016	-.011	.267	.476	.484	.745	.638	.848	.954	.900	1.000	.-037
	ln γ^{∞}	.632	.584	.541	.583	.527	.553	.543	.496	.437	.435	-.535	-.494	-.491	-.002	-.508	-.451	-.439	.425	.419	.334	.170	.121	-.037	1.000

Table 3. Initial Set of Descriptors, Covariances, and Cumulative Dissimilarity Measures Ordered According to the SOM Variable Selection Procedure

Descriptor	Abbreviation	Class	Covariance with Target Variable	Cumulative Dissimilarity
Valence connectivity index of first order	$^1\chi^v$	I	0.632	0.703
Sum of atomic numbers	*N*	II	0.583	0.553
Valence connectivity index of fourth order	$^4\chi^v$	III	0.553	0.366
Dipole moment	μ	IV	−0.535	0.276
Electron–nuclear attraction	*ENA*	V	−0.508	0.243
Coulomb self-similarity	*Cou*	VI	0.425	0.229
Valence connectivity index of the third order	$^3\chi^v$	I	0.584	0.212
Number of filled levels	*NFL*	II	0.543	0.204
Average polarizability	*AP*	I	0.541	0.195
Valence connectivity index of second order	$^2\chi^v$	II	0.527	0.191
Nuclear–nuclear repulsion energy	*NNR*	III	0.496	0.190
Hansen polarizability	*HP*	IV	−0.494	0.192
Hansen hydrogen bonding	*HH*	IV	−0.491	0.196
Exchange energy	*EE*	V	−0.451	0.198
Resonance energy	*RE*	V	−0.439	0.201
Kappa second-order index	$^2\kappa$	III	0.437	0.211
Valence connectivity index of zero order	$^0\chi^v$	III	0.435	0.218
Molecular quantum cross-similarity of Coulomb with respect to C_6H_8	Cou_{C6H8}	VI	0.419	0.234
Molecular quantum cross-similarity of Coulomb with respect to $C_3H_7Cl_1$	$Cou_{C3H7Cl1}$	VI	0.334	0.253
Overlap self-similarity	*Ove*	VI	0.170	0.257
Molecular quantum cross-similarity of Overlap with respect to C_6H_8	Ove_{C6H8}	VI	0.121	0.270
Molecular quantum cross-similarity of Overlap with respect to $C_5H_9N_1O_1$	$Ove_{C5H9N1O1}$	VI	−0.037	0.377
Hansen dispersivity	*HD*	IV	−0.002	0.381

model, does not imply weighted fitting of input data because at most only a few extra classes in the ART modules will be created during training. More information and discussions on statistics and generalization in relation to neural networks can be found elsewhere (Cheng and Titterington, 1994).

The performance of the fuzzy-ARTMAP–based QSPR with the most suitable set of 11 descriptors is summarized in Figure 4. This fuzzy-ARTMAP–based QSPR model performed with an average absolute error and standard deviation of 0.09 (1.21%) and 0.29 (3.65%) $\ln\gamma^\infty$ units, respectively, for the complete set of 325 compounds. The remarkable generalization capability of this model is illustrated by the prediction of $\ln\gamma^\infty$ for the test set of 45 chemicals, as shown in Figure 4a, with an average absolute error and standard deviation of 0.52 (6.64%) and 0.51 (6.23%) $\ln\gamma^\infty$ units, respectively. The performance of the above QSPR for the training set of 280 compounds [Figure 4(a)] was with a very low average absolute error and standard deviation of 0.02 (0.36%) and 0.02 (0.60%) $\ln\gamma^\infty$ units, respectively. This high level of performance for the training set is expected, given that fuzzy ARTMAP is a neural classifier. In fact, errors for the training set could ultimately be reduced to zero by increasing precision to the point where all fuzzy ARTMAP classes are occupied each with a single compound. However, this would hinder or prevent predictive generalization during testing, which is the main goal of QSPR models. Thus, a vigilance parameter of 0.999 was used in the current study as a compromise between adequate precision for training and reasonably accurate predictive generalization for the testing phase, as illustrated in Figure 4a.

The performance of the fuzzy ARTMAP QSPR with the most suitable set of descriptors was further increased upon training using both the 100 prototypes (obtained from the SOM analysis as described before) and the training set of 280 compounds. In this case of enhanced training, the vigilance parameter of the ARTMAP algorithm was relaxed from 0.999 to 0.995 to promote generalization in the testing, despite the likely increase in training errors that this strategy could cause. The resulting QSPR performance for the test set was with average absolute error and standard deviation that decreased to 0.40 (5.35%) and 0.48 (5.85%) $\ln\gamma^\infty$ units, respectively, relative to the performance obtained without the prototypes [Figure 4b]. We note that, as a consequence of this strategy of enhancing generalization during testing, the average absolute errors and standard deviations when training with the prototypes increased to 0.05 (1.07%) and 0.04 (6.42%) $\ln\gamma^\infty$ units. Comparison of Figure 4(a) and (b) indicates that the inclusion of prototypes in the training phase improves the classification of 12 chemicals out of the 45 test set chemicals. This is so because 11 of these chemicals were assigned during testing to fuzzy ARTMAP classes formed by prototypes of occupied SOM nodes and one was assigned to a class formed by a prototype of an empty but trained node, as indicated in Figure 4(b).

In addition to the above QSPRs, the four most suitable MQS measures alone (*Cou*, Cou_{C6H8}, Cou_{C3H7Cl}, and *Ove*) were used to construct a QSPR using the same training set of 280 compounds. The performance of the resulting QSPR, for the test set of 45 compounds, was with an average absolute error and standard deviation of 0.92 (11.2%) and 1.09 (11.5%) $\ln\gamma^\infty$ units, respectively. Although the above error and standard deviation are approximately twice those obtained for the most suitable set of 11 indices (Figure 4), the performance is remarkable considering that only four MQS matrices were used. It is also interesting to note that the inclusion of the 100 SOM prototypes in the training phase did not improve the QSPR's

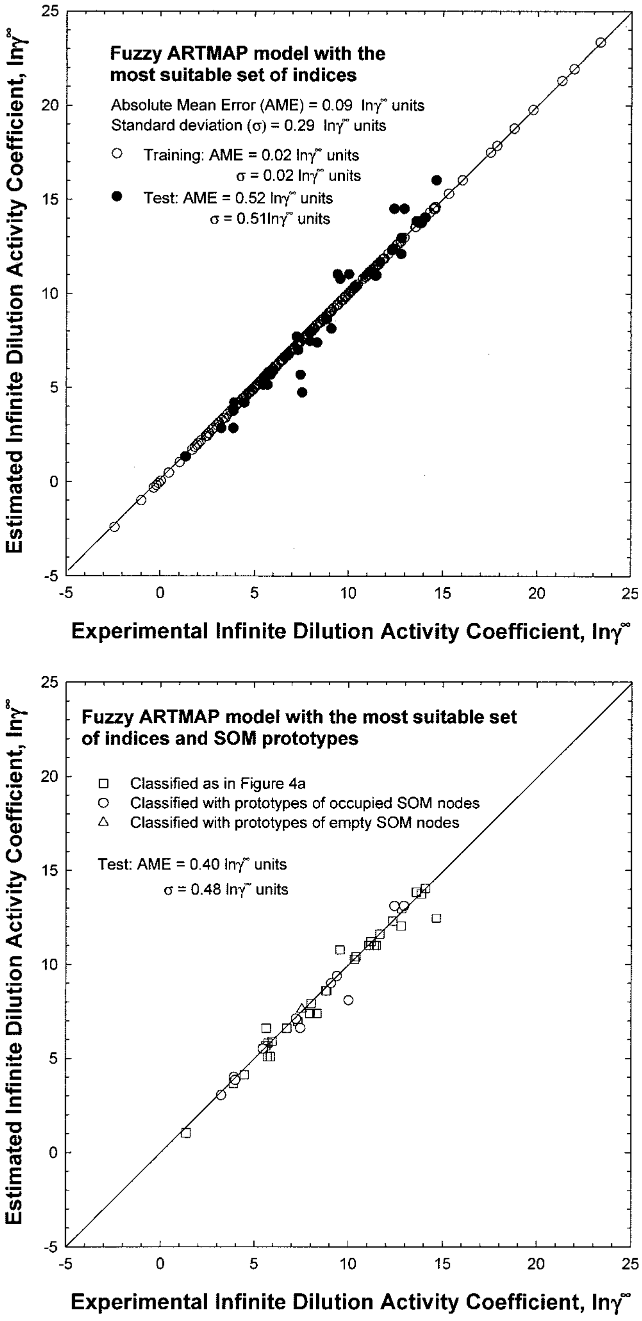

Figure 4. Comparison between experimental $\ln\gamma^{\infty}$ values and those predicted with the two current fuzzy ARTMAP models developed using the most suitable set of eleven descriptors and trained with (a) 280 compounds or (b) 280 compounds complemented with 100 prototypes of clustered compound in the SOM nodes.

Note that (b) depicts results only for the test data set of 45 compounds.

performance in this case, given that errors associated with the description halogen groups with only MQS matrices also propagate over the map and thus negate the increase in interpolation capabilities. It is noted that the errors associated with hydrocarbon families containing halogen substituents tend to be the largest in Table 4.

To illustrate the potential of using MQSM as the fundamental source of molecular information in QSPR building, an additional model was developed with the set of four similarity measures mentioned above (*Cou*, Cou_{C6H8}, Cou_{C3H7Cl}, and *Ove*), the sum of atomic numbers, and the sum of atomics numbers of heteroatoms, to better account for size effects and for the presence of such atoms in the target molecules. The performance of the above QSPR, for the test set of 45 compounds, was with an average absolute error and standard deviation of 0.57 (7.06%) and 0.57 (7.06%) $\ln\gamma^{\infty}$ units, respectively, a level of performance similar to that obtained with the most suitable set of 11 indices [Figure 4a].

It is instructive to explore the performance of the present QSPRs for specific families of organic compounds, as illustrated in Table 4. Clearly, the MQS-based model performs very well when the training data set is large and the range of molecular sizes (that is, number of carbon atoms) is narrow. This is an indication of the need for more and/or improved shape information than provided by the Overlap operator. It is noted that the performance results given in Table 4, in terms of the average absolute errors and standard deviations, are for the complete data set (training plus test data); thus differences between the different models are not as evident as with comparisons based on the test set (such as Figures 4 and 5). This choice, however, facilitates comparison with previous studies, as discussed in the next section.

Comparison with previous studies

The current fuzzy ARTMAP models were compared with the best performing neural network–based QSPR reported by Mitchell and Jurs (1998) and with the regression-based QSPR of Medir and Giralt (1982) for specific chemical families. The model of Mitchell and Jurs (1998) is a multilayer perceptron, with an input–hidden-output layer architecture of 12–6–1 units, respectively, trained with a quasi-Newton BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm. These authors used 12 indices that included three topological descriptors, four charged partial surface area indices, two hydrogen bonding measures, the heat of formation, and two theoretical linear solvation energy relationships. To consistently compare previous and current models for the different chemical families present in the data set of 325 compounds (Mitchell and Jurs, 1998; Sherman et al., 1996), and to avoid the impediments caused by the different training and test sets used, the results are presented in Table 4 in terms of chemical families without distinguishing training and test compounds. Also, the correlation reported by Medir and Giralt (1982) was recalculated for each of the chemical families A, B, and C to adapt it to the more complete set of current data (Mitchell and Jurs, 1998; Sherman et al., 1996).

The performance of the present fuzzy-ARTMAP–based QSPR model was better than the neural network model of Mitchell and Jurs (1998) for each of the 13 homogeneous families present in the data set of 325 compounds (Table 4). The current two models, developed with the most suitable set of descriptors and trained by either the training set of 280 compounds or by these compounds with the 100 prototypes from the SOM, performed with absolute mean errors and

Table 4. Comparison of Performances of Current and Previous QSPR Models in Terms of Absolute Mean Error (Standard Deviation) in the Prediction of $\ln\gamma^\infty$ for the 13 Families of Organic Compounds Included in the Data Set

ID	Family	Number of Compounds	Number of Carbon Atoms	Fuzzy ARTMAP Best Set	Fuzzy ARTMAP MQSM	Fuzzy ARTMAP Best Sets and SOM Prototypes	12–6–1 NN(*) (Mitchell and Jurs[10])	Medir and Giralt[9]
A	Monoaraomatic hydrocarbons	21	C6–C10	0.22 (0.57)	0.53 (1.34)	0.07 (0.15)	0.54 (1.13)	0.56 (1.09)(1)
B	Polyaromatic hydrocarbons	5	C9–C12	0.07 (0.10)	0.16 (0.38)	0.06 (0.10)	0.22 (0.28)	0.30 (0.15)(2)
C	Aliphatic hydrocarbons	35	C4–C8	0.10 (0.25)	0.32 (0.75)	0.15 (0.38)	0.45 (0.47)	0.46 (0.32)(3)
D	Hydrocarbons with oxygen substituents	124	C1–C18	0.06 (0.17)	0.09 (0.31)	0.09 (0.18)	0.70 (0.78)	—
E	Halogenated aliphatic hydrocarbons	66	C1–C6	0.12 (0.34)	0.12 (0.43)	0.12 (0.27)	0.82 (2.25)	—
F	Aromatic hydrocarbons with nitrogen and/or oxygen substituents	16	C5–C8	0.05 (0.12)	0.06 (0.20)	0.06 (0.06)	1.12 (2.11)	—
G	Hydrocarbons with sulfur substituents and/or oxygen and/or nitrogen	19	C1–C4	0.09 (0.14)	0.26 (0.61)	0.10 (0.16)	0.18 (0.12)	—
H	Cyclic hydrocarbons	15	C5–C9	0.07 (0.17)	0.10 (0.34)	0.10 (0.19)	0.51 (0.52)	—
I	Aromatic hydrocarbons with oxygen substituents	8	C7–C8	0.04 (0.03)	0.03 (0.03)	0.03 (0.04)	0.25 (0.23)	—
J	Halogenated aromatic hydrocarbons	9	C6–C7	0.01 (0.02)	0.06 (0.12)	0.07 (0.05)	0.27 (0.24)	—
K	Heterocyclic hydrocarbons	5	C4–C8	0.02 (0.03)	0.00 (0.00)	0.03 (0.04)	0.36 (0.25)	—
L	Aliphatic hydrocarbons with halogen and oxygen substituents	2	C2–C4	1.41 (1.99)	1.00 (1.42)	0.10 (0.05)	0.15 (0.11)	—
M	Aliphatic hydrocarbons with nitro and halogen groups	1	C1	0.01 (0.00)	0.01 (0.00)	0.06 (0.00)	0.08 (0.00)	—

*The performance of this model by families was not reported in the original reference. The values included in this table have been currently calculated with the original 12–6–1 feedforward architecture.

(1) $\ln\gamma^\infty = 2.99 + 2.58\ {}^1\chi^v$.

(2) $\ln\gamma^\infty = -0.054 + 3.1\ {}^1\chi^v$.

(3) $\ln\gamma^\infty = 5.461 + 2.739\ {}^1\chi^v + 5.884\mu$.

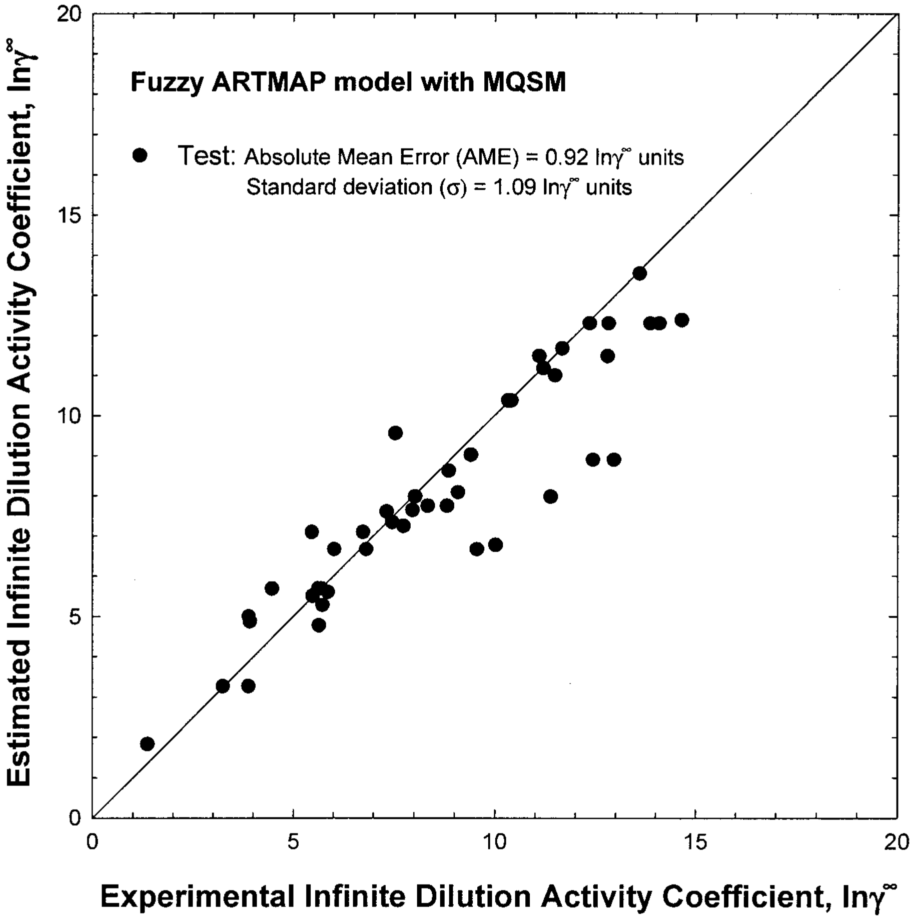

Figure 5. Comparison between experimental $\ln\gamma^\infty$ values and those predicted with the current fuzzy ARTMAP model developed using only the most suitable set of four MQSM and trained with 280 compounds.

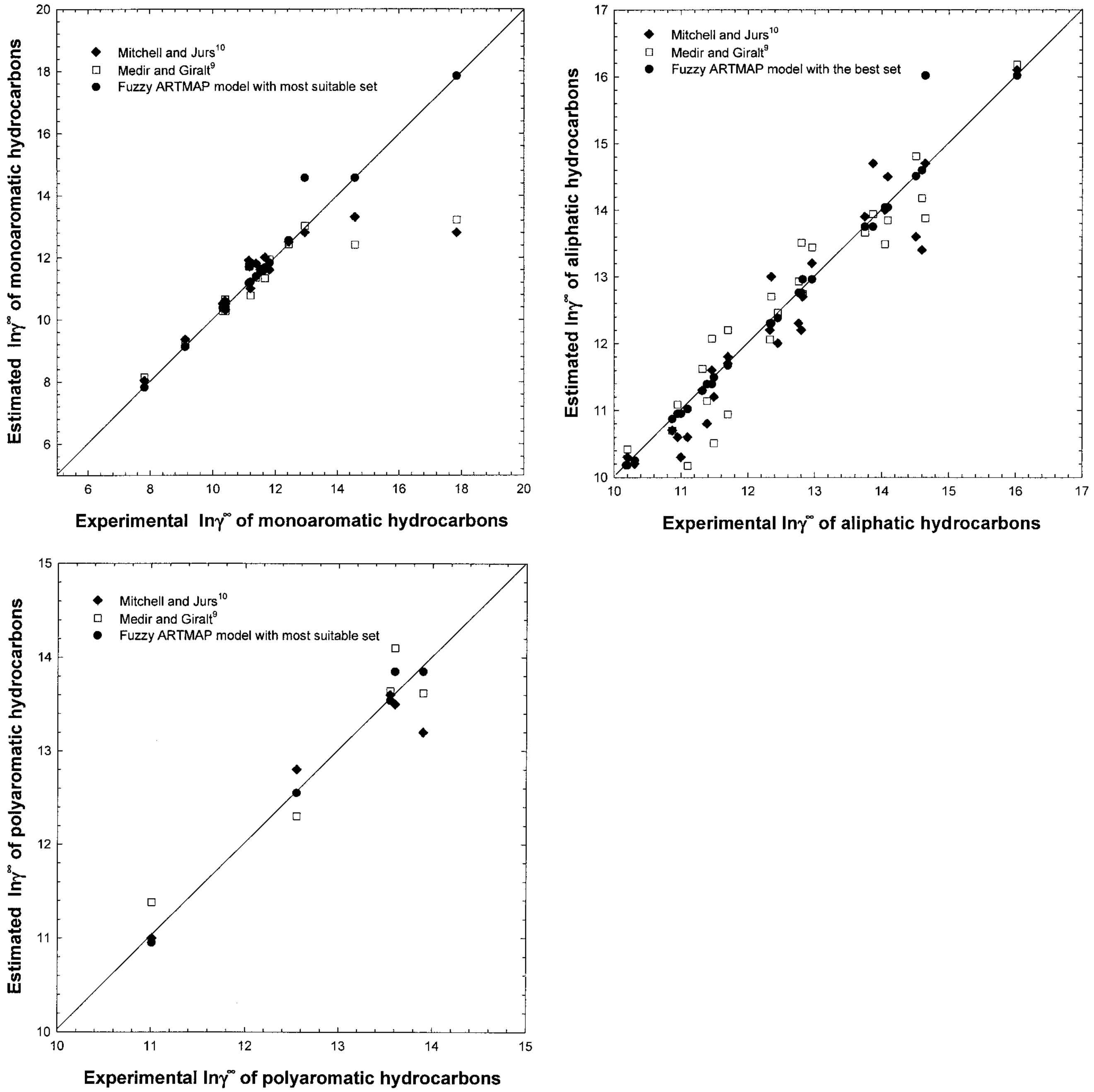

Figure 6. Comparison of $\ln\gamma^{\infty}$ between the current fuzzy ARTMAP model (developed using the most suitable set of eleven descriptors and trained with 280 compounds) and previous QSAR models for (a) monoaromatic hydrocarbons, (b) polyaromatic hydrocarbons, and (c) aliphatic hydrocarbons.

standard deviations that are on the average seven times smaller than errors obtained with the model of Mitchell and Jurs (1998). Improved QSPR performance is also noted relative to the linear regression models of Medir and Giralt (1982). It is interesting to note that, for the two families of aromatic hydrocarbons A and B, this linear model with only the first-order connectivity as descriptor yields comparable predictions to those obtained with the model of Mitchell and Jurs (1998) calculated for the whole data set. In the case of aliphatic hydrocarbons (family C) the dipole moment was also included in the linear correlation to maintain performance.

It is instructive to compare the present fuzzy ARTMAP model, trained with the most suitable set of descriptors and 280 compounds, to the Mitchell and Jurs (1998) and Medir and Giralt (1982) QSPRs for specific chemical families, as depicted in Figure 6(a)–(c) for monoaromatic, polyaromatic, and aliphatic hydrocarbons. In the case of monoaromatic hydrocarbons (benzene and toluene derivatives), deviations of previous models are observed mainly at the highest $\ln\gamma^{\infty}$ values [Figure 6(a)], whereas the polyaromatics deviations [Figure 6(b)] are more evenly distributed over the range studied. Both previous models deviate over the whole range of $\ln\gamma^{\infty}$ values covered for

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds*

No.	Name	ID	Formula	N	$^1\chi^V$	$^2\chi^V$	$^3\chi^V$	$^4\chi^V$	ln(MQS Coulomb)	ENA	NNR	μ	NFL	AP	ln γ^∞
1	tetrachloromethane	tr	C 1 Cl 4	7.40E+01	2.26E+00	3.85E+00	0.00E+00	0.00E+00	7.85E+00	-4.37E+03	2.22E+03	1.00E-02	1.60E+01	4.28E+01	9.64
2	trichlorofluoromethane	te	C 1 F 1 Cl 3	6.60E+01	1.89E+00	2.57E+00	0.00E+00	0.00E+00	7.65E+00	-4.45E+03	2.27E+03	5.60E-01	1.60E+01	3.54E+01	8.86
3	dichlorodifluoromethane	te	C 1 F 2 Cl 2	5.80E+01	1.51E+00	1.57E+00	0.00E+00	0.00E+00	7.42E+00	-4.56E+03	2.33E+03	8.00E-01	1.60E+01	2.69E+01	10.02
4	chlorotrifluoromethane	tr	C 1 F 3 Cl 1	5.00E+01	1.13E+00	8.56E-01	0.00E+00	0.00E+00	7.13E+00	-4.73E+03	2.33E+03	1.04E+00	1.60E+01	1.70E+01	11.07
5	tetrafluoromethane	tr	C 1 F 4	4.20E+01	7.56E-01	4.29E-01	0.00E+00	0.00E+00	6.76E+00	-4.91E+03	2.53E+03	0.00E+00	1.60E+01	8.02E+00	12.63
6	tribromomethane	tr	C 1 H 1 Br 3	1.12E+02	3.40E+00	6.66E+00	0.00E+00	0.00E+00	9.14E+00	-2.81E+03	1.43E+03	9.50E-01	1.30E+01	4.00E+01	8.13
7	trichloromethane	tr	C 1 H 1 Cl 3	5.80E+01	1.96E+00	2.22E+00	0.00E+00	0.00E+00	7.50E+00	-2.87E+03	1.47E+03	1.02E+00	1.30E+01	3.22E+01	6.81
8	dichlorofluoromethane	tr	C 1 H 1 F 1 Cl 2	5.00E+01	1.53E+00	1.23E+00	0.00E+00	0.00E+00	7.23E+00	-2.94E+03	1.51E+03	1.25E+00	1.30E+01	2.51E+01	5.72
9	chlorodifluoromethane	te	C 1 H 1 F 2 Cl 1	4.20E+01	1.09E+00	5.77E-01	0.00E+00	0.00E+00	6.89E+00	-3.06E+03	1.58E+03	1.52E+00	1.30E+01	1.71E+01	7.46
10	trifluoromethane	tr	C 1 H 1 F 3	3.40E+01	6.55E-01	2.47E-01	0.00E+00	0.00E+00	6.43E+00	-3.21E+03	1.58E+03	1.89E+00	1.30E+01	8.27E+00	8.37
11	triiodomethane	tr	C 1 H 1 I 3	1.66E+02	4.33E+00	1.08E+01	0.00E+00	0.00E+00	9.93E+00	-2.68E+03	1.36E+03	6.10E-01	1.30E+01	6.41E+01	12.30
12	dibromomethane	tr	C 1 H 2 Br 2	7.80E+01	2.77E+00	2.72E+00	0.00E+00	0.00E+00	8.67E+00	-1.65E+03	8.48E+02	1.45E+00	1.00E+01	2.80E+01	6.75
13	dichloromethane	tr	C 1 H 2 Cl 2	4.20E+01	1.60E+00	9.07E-01	0.00E+00	0.00E+00	7.03E+00	-1.71E+03	8.81E+02	1.36E+00	1.00E+01	2.24E+01	5.53
14	difluoromethane	tr	C 1 H 2 F 2	2.60E+01	5.35E-01	1.01E-01	0.00E+00	0.00E+00	6.00E+00	-1.88E+03	9.83E+02	1.81E+00	1.00E+01	8.19E+00	6.49
15	diidomethane	tr	C 1 H 2 I 2	1.14E+02	3.54E+00	4.42E+00	0.00E+00	0.00E+00	9.46E+00	-1.58E+03	7.83E+02	1.20E+00	1.00E+01	3.85E+01	9.39
16	formaldehyde	tr	C 1 H 2 O 1	1.60E+01	2.89E-01	0.00E+00	0.00E+00	0.00E+00	5.17E+00	-7.24E+02	3.92E+02	2.16E+00	6.00E+00	9.92E+00	1.03
17	formic acid	tr	C 1 H 2 O 2	2.40E+01	4.94E-01	1.05E-01	0.00E+00	0.00E+00	5.80E+00	-1.70E+03	9.05E+02	3.94E+00	9.00E+00	1.37E+01	-0.33
18	bromomethane	tr	C 1 H 3 Br 1	4.40E+01	1.96E+00	0.00E+00	0.00E+00	0.00E+00	7.91E+00	-8.35E+02	4.38E+02	1.55E+00	7.00E+00	1.71E+01	5.98
19	chloromethane	tr	C 1 H 3 Cl 1	2.60E+01	1.13E+00	0.00E+00	0.00E+00	0.00E+00	6.31E+00	-8.76E+02	4.59E+02	1.38E+00	7.00E+00	1.40E+01	6.17
20	fluoromethane	tr	C 1 H 3 F 1	1.80E+01	3.78E-01	0.00E+00	0.00E+00	0.00E+00	5.38E+00	-9.33E+02	4.95E+02	1.44E+00	7.00E+00	7.92E+00	6.67
21	nitromethane	tr	C 1 H 3 N 1 O 2	3.20E+01	7.57E-01	4.24E-01	0.00E+00	0.00E+00	6.20E+00	-3.06E+03	1.61E+03	3.99E+00	1.20E+01	2.00E+01	3.45
22	methanol	tr	C 1 H 4 O 1	1.80E+01	4.47E-01	0.00E+00	0.00E+00	0.00E+00	5.27E+00	-1.07E+03	5.73E+02	1.49E+00	7.00E+00	1.09E+01	0.46
23	methyl mercaptane	tr	C 1 H 4 S 1	2.60E+01	1.34E+00	0.00E+00	0.00E+00	0.00E+00	6.23E+00	-7.40E+02	3.93E+02	2.23E+00	6.00E+00	1.86E+01	4.72
24	nitrotrichloromethane	tr	C 1 N 1 O 2 Cl 3	8.00E+01	2.25E+00	2.87E+00	5.93E-01	0.00E+00	7.86E+00	-7.99E+03	3.94E+03	2.80E+00	2.10E+01	4.66E+01	8.64
25	tetrachloroethene	tr	C 2 Cl 4	8.00E+01	2.51E+00	2.41E+00	1.28E+00	0.00E+00	7.91E+00	-5.45E+03	2.77E+03	0.00E+00	1.80E+01	5.28E+01	10.49
26	hexachloroethane	tr	C 2 Cl 6	1.14E+02	3.65E+00	5.54E+00	2.88E+00	0.00E+00	8.43E+00	-1.06E+04	5.34E+03	1.00E-02	2.50E+01	6.48E+01	14.31
27	1-1-2-trichlorotrifluoroethane	tr	C 2 F 3 Cl 3	9.00E+01	2.52E+00	2.70E+00	1.25E+00	0.00E+00	8.01E+00	-1.09E+04	5.94E+03	1.56E+00	2.50E+01	4.13E+01	11.02
28	tetrafluoroethene	tr	C 2 F 4	4.80E+01	1.01E+00	5.21E-01	1.43E-01	0.00E+00	6.88E+00	-5.80E+03	2.98E+03	0.00E+00	1.80E+01	1.96E+01	10.47
29	1-2-dichlorotetrafluoroethane	tr	C 2 F 4 Cl 2	8.20E+01	2.14E+00	2.23E+00	7.49E-01	0.00E+00	7.84E+00	-1.11E+04	5.67E+03	1.33E+00	2.50E+01	3.32E+01	11.15
30	chloropentafluoroethane	tr	C 2 F 5 Cl 1	7.40E+01	1.76E+00	1.47E+00	5.35E-01	0.00E+00	7.63E+00	-1.13E+04	5.75E+03	1.10E+00	2.50E+01	2.53E+01	11.90
31	hexafluoroethane	tr	C 2 F 6	6.60E+01	1.38E+00	9.96E-01	3.21E-01	0.00E+00	7.38E+00	-1.15E+04	5.66E+03	6.00E-03	2.50E+01	1.67E+01	13.79
32	trichloroethene	tr	C 2 H 1 Cl 3	6.40E+01	2.07E+00	1.62E+00	7.40E-01	0.00E+00	7.58E+00	-3.85E+03	1.96E+03	4.90E-01	1.50E+01	4.21E+01	9.08
33	pentachloroethane	tr	C 2 H 1 Cl 5	9.90E+01	3.29E+00	4.30E+00	2.22E+00	0.00E+00	8.20E+00	-8.24E+03	4.10E+03	8.46E-01	2.20E+01	5.57E+01	10.08
34	1-1-2-2-tetrabromoethane	tr	C 2 H 2 Br 4	1.54E+02	4.86E+00	7.06E+00	5.13E+00	0.00E+00	9.50E+00	-6.02E+03	3.05E+03	1.01E+00	1.90E+01	6.00E+01	10.29
35	trans-1-2-dichloroethene	tr	C 2 H 2 Cl 2	4.80E+01	1.64E+00	7.55E-01	4.27E-01	0.00E+00	7.13E+00	-2.47E+03	1.27E+03	0.00E+00	1.20E+01	3.27E+01	7.14

*$^{(1-4)}\chi^v$ = valence connectivity index; N = sum of atomic numbers; *NFL* = number of filled levels; μ = dipole moment; *AP* = average polarizability (PM3); *ENA* = electron–nuclear attraction; *NNR* = nuclear–nuclear repulsion; *Cou* (Coulomb MQS).

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

36	cis-1-2-dichloroethene	tr	C 2 H 2 Cl 2	4.80E+01	1.64E+00	7.55E-01	4.27E-01	0.00E+00	7.15E+00	-2.47E+03	1.27E+03	0.00E+00	1.20E+01	3.27E+01	6.77
37	1-1-2-2-tetrachloroethane	tr	C 2 H 2 Cl 4	8.20E+01	2.95E+00	2.99E+00	1.71E+00	0.00E+00	7.93E+00	-1.08E+04	5.47E+03	0.00E+00	2.50E+01	5.12E+01	8.15
38	1-1-1-2-tetrachloroethane	te	C 2 H 2 Cl 4	8.20E+01	2.85E+00	3.52E+00	1.36E+00	0.00E+00	7.94E+00	-6.25E+03	3.17E+03	1.21E+00	1.90E+01	4.74E+01	9.09
39	chloroethene	tr	C 2 H 3 Cl 1	3.20E+01	1.06E+00	4.62E-01	0.00E+00	0.00E+00	6.50E+00	-1.51E+03	7.90E+02	9.10E-01	9.00E+00	2.30E+01	7.16
40	1-1-2-trichloroethane	tr	C 2 H 3 Cl 3	6.60E+01	2.52E+00	2.13E+00	1.05E+00	0.00E+00	7.60E+00	-4.45E+03	2.27E+03	1.08E+00	1.60E+01	3.85E+01	7.31
41	1-1-1-trichloroethane	tr	C 2 H 3 Cl 3	6.60E+01	2.20E+00	3.62E+00	0.00E+00	0.00E+00	7.62E+00	-4.58E+03	2.34E+03	1.38E+00	1.60E+01	4.01E+01	8.68
42	acetonitrile	tr	C 2 H 3 N 1	2.20E+01	7.24E-01	2.24E-01	0.00E+00	0.00E+00	5.49E+00	-1.29E+03	6.89E+02	3.21E+00	8.00E+00	1.86E+01	2.41
43	2-2-2-trifluoroethanol	tr	C 2 H 3 O 1 F 3	5.00E+01	1.24E+00	7.73E-01	1.79E-01	0.00E+00	6.91E+00	-6.90E+03	3.54E+03	3.48E+00	1.90E+01	1.89E+01	2.16
44	1-2-dibromoethane	tr	C 2 H 4 Br 2	8.60E+01	3.27E+00	1.96E+00	1.92E+00	0.00E+00	8.69E+00	-2.89E+03	1.43E+03	2.00E-03	1.30E+01	4.02E+01	7.84
45	1-bromo-2-chloroethane	tr	C 2 H 4 Cl 1 Br 1	6.80E+01	2.69E+00	1.55E+00	1.11E+00	0.00E+00	8.18E+00	-2.94E+03	1.51E+03	2.10E-01	1.30E+01	3.48E+01	7.06
46	1-2-dichloroethane	tr	C 2 H 4 Cl 2	5.00E+01	2.10E+00	1.13E+00	6.41E-01	0.00E+00	7.16E+00	-2.99E+03	1.54E+03	1.00E-02	1.30E+01	3.04E+01	6.46
47	1-1-dichloroethane	tr	C 2 H 4 Cl 2	5.00E+01	1.88E+00	2.05E+00	0.00E+00	0.00E+00	7.18E+00	-3.09E+03	1.59E+03	1.62E+00	1.30E+01	3.08E+01	6.98
48	acetaldehyde	te	C 2 H 4 O 1	2.40E+01	8.13E-01	2.36E-01	0.00E+00	0.00E+00	5.67E+00	-1.69E+03	8.93E+02	2.46E+00	9.00E+00	1.73E+01	1.37
49	ethylene oxide (oxirane)	tr	C 2 H 4 O 1	2.40E+01	1.08E+00	6.12E-01	0.00E+00	2.20E-01	5.70E+00	-1.81E+03	8.89E+02	1.78E+00	9.00E+00	1.72E+01	1.83
50	acetic acid	tr	C 2 H 4 O 2	3.20E+01	9.28E-01	5.19E-01	0.00E+00	0.00E+00	6.15E+00	-3.01E+03	1.57E+03	1.83E+00	1.20E+01	2.05E+01	-0.08
51	methyl formate	tr	C 2 H 4 O 2	3.20E+01	8.80E-01	3.32E-01	9.62E-02	0.00E+00	6.14E+00	-3.00E+03	1.57E+03	3.86E+00	1.20E+01	2.20E+01	2.74
52	bromoethane	tr	C 2 H 5 Br 1	5.20E+01	2.09E+00	1.39E+00	0.00E+00	0.00E+00	7.98E+00	-1.90E+03	9.87E+02	1.85E+00	1.00E+01	2.60E+01	6.52
53	chloroethane	tr	C 2 H 5 Cl 1	3.40E+01	1.51E+00	8.01E-01	0.00E+00	0.00E+00	6.54E+00	-1.94E+03	1.01E+03	1.55E+00	1.00E+01	2.22E+01	5.98
54	iodoethane	tr	C 2 H 5 I 1	7.00E+01	2.47E+00	1.77E+00	0.00E+00	0.00E+00	8.75E+00	-1.86E+03	9.67E+02	1.83E+00	1.00E+01	2.96E+01	7.69
55	nitroethane	tr	C 2 H 5 N 1 O 2	4.00E+01	1.35E+00	6.10E-01	2.47E-01	0.00E+00	6.47E+00	-4.59E+03	2.25E+03	4.12E+00	1.50E+01	2.74E+01	4.48
56	ethanol	tr	C 2 H 6 O 1	2.60E+01	1.02E+00	3.16E-01	0.00E+00	0.00E+00	5.75E+00	-2.16E+03	1.13E+03	1.45E+00	1.00E+01	1.82E+01	1.32
57	dimethylsulfoxide	tr	C 2 H 6 O 1 S 1	4.20E+01	2.09E+00	1.58E+00	0.00E+00	0.00E+00	6.77E+00	-3.19E+03	1.57E+03	4.49E+00	1.30E+01	3.55E+01	-2.41
58	dimethyl sulfide	tr	C 2 H 6 S 1	3.40E+01	2.44E+00	1.22E+00	0.00E+00	0.00E+00	6.48E+00	-1.89E+03	9.90E+02	1.96E+00	1.00E+01	2.68E+01	5.18
59	ethanethiol	te	C 2 H 6 S 1	3.40E+01	1.65E+00	9.46E-01	0.00E+00	0.00E+00	6.48E+00	-1.73E+03	8.51E+02	2.33E+00	9.00E+00	2.46E+01	5.46
60	1-3-dichloropropene	te	C 3 H 4 Cl 2	5.60E+01	2.20E+00	1.08E+00	5.34E-01	3.02E-01	7.25E+00	-3.84E+03	1.90E+03	1.30E+00	1.50E+01	3.99E+01	7.24
61	3-bromo-1-propene	tr	C 3 H 5 Br 1	5.80E+01	2.20E+00	1.09E+00	5.66E-01	0.00E+00	8.03E+00	-2.65E+03	1.37E+03	1.69E+00	1.20E+01	3.42E+01	7.47
62	3-chloro-1-propene	tr	C 3 H 5 Cl 1	4.00E+01	1.62E+00	7.51E-01	3.27E-01	0.00E+00	6.70E+00	-2.70E+03	1.40E+03	1.43E+00	1.20E+01	3.05E+01	6.97
63	1-2-3-trichloropropane	tr	C 3 H 5 Cl 3	7.40E+01	3.07E+00	2.14E+00	1.70E+00	3.70E-01	7.70E+00	-6.21E+03	3.17E+03	2.29E+00	1.90E+01	4.41E+01	8.37
64	propanenitrile	tr	C 3 H 5 N 1	3.00E+01	1.28E+00	5.12E-01	1.58E-01	0.00E+00	5.91E+00	-2.42E+03	1.20E+03	3.25E+00	1.10E+01	2.58E+01	3.56
65	1-3-dibromopropane	tr	C 3 H 6 Br 2	9.40E+01	3.77E+00	2.31E+00	1.39E+00	1.36E+00	8.72E+00	-4.36E+03	2.16E+03	1.82E+00	1.60E+01	4.43E+01	8.81
66	1-2-dibromopropane	tr	C 3 H 6 Br 2	9.40E+01	3.50E+00	3.14E+00	2.37E+00	0.00E+00	8.74E+00	-4.54E+03	2.25E+03	5.15E-01	1.60E+01	4.76E+01	8.97
67	1-3-dichloropropane	tr	C 3 H 6 Cl 2	5.80E+01	2.60E+00	1.49E+00	8.01E-01	4.53E-01	7.28E+00	-4.46E+03	2.29E+03	1.50E+00	1.60E+01	3.70E+01	7.74
68	1-2-dichloropropane	tr	C 3 H 6 Cl 2	5.80E+01	2.44E+00	1.99E+00	9.86E-01	0.00E+00	7.31E+00	-4.66E+03	2.39E+03	3.70E-01	1.60E+01	3.78E+01	7.75
69	acetone	tr	C 3 H 6 O 1	3.20E+01	1.20E+00	9.08E-01	0.00E+00	0.00E+00	6.05E+00	-2.98E+03	1.47E+03	2.78E+00	1.20E+01	2.45E+01	1.95
70	propionaldehyde	tr	C 3 H 6 O 1	3.20E+01	1.35E+00	5.75E-01	1.67E-01	0.00E+00	6.04E+00	-2.91E+03	1.52E+03	2.51E+00	1.20E+01	2.46E+01	2.56
71	methyl acetate	tr	C 3 H 6 O 2	4.00E+01	1.32E+00	6.96E-01	2.87E-01	0.00E+00	6.44E+00	-4.59E+03	2.38E+03	1.83E+00	1.50E+01	2.87E+01	3.12
72	ethyl formate	tr	C 3 H 6 O 2	4.00E+01	1.47E+00	5.52E-01	2.35E-01	6.80E-02	6.42E+00	-4.46E+03	2.31E+03	3.91E+00	1.50E+01	2.97E+01	3.86
73	2-bromopropane	tr	C 3 H 7 Br 1	6.00E+01	2.29E+00	2.84E+00	0.00E+00	0.00E+00	8.05E+00	-3.28E+03	1.69E+03	2.04E+00	1.30E+01	3.40E+01	7.64
74	1-bromopropane	te	C 3 H 7 Br 1	6.00E+01	2.59E+00	1.48E+00	9.81E-01	0.00E+00	8.04E+00	-3.18E+03	1.64E+03	1.81E+00	1.30E+01	3.29E+01	7.96
75	2-chloropropane	tr	C 3 H 7 Cl 1	4.20E+01	1.81E+00	1.88E+00	0.00E+00	0.00E+00	6.75E+00	-3.32E+03	1.71E+03	1.66E+00	1.30E+01	2.98E+01	7.30
76	1-chloropropane	tr	C 3 H 7 Cl 1	4.20E+01	2.01E+00	1.07E+00	5.66E-01	0.00E+00	6.74E+00	-3.23E+03	1.67E+03	1.55E+00	1.30E+01	2.93E+01	7.47

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

77	2-iodopropane	tr	C 3 H 7 I 1	7.80E+01	2.60E+00	3.46E+00	0.00E+00	0.00E+00	8.79E+00	-3.22E+03	1.67E+03	2.08E+00	1.30E+01	3.82E+01	8.82
78	1-iodopropane	tr	C 3 H 7 I 1	7.80E+01	2.97E+00	1.75E+00	1.25E+00	0.00E+00	8.78E+00	-3.13E+03	1.62E+03	1.77E+00	1.30E+01	3.66E+01	9.05
79	N-N-dimethylformamide	tr	C 3 H 7 N 1 O 2	4.80E+01	1.55E+00	1.18E+00	3.83E-01	0.00E+00	6.40E+00	-4.53E+03	2.35E+03	3.46E+00	1.50E+01	3.33E+01	-0.19
80	2-nitropropane	tr	C 3 H 7 N 1 O 2	4.80E+01	1.74E+00	1.32E+00	4.03E-01	0.00E+00	6.71E+00	-6.44E+03	3.34E+03	4.16E+00	1.80E+01	3.42E+01	5.69
81	1-nitropropane	te	C 3 H 7 N 1 O 2	4.80E+01	1.85E+00	1.03E+00	3.79E-01	1.75E-01	6.69E+00	-6.22E+03	3.23E+03	4.21E+00	1.80E+01	3.44E+01	5.70
82	2-propanol	tr	C 3 H 8 O 1	3.40E+01	1.41E+00	1.09E+00	0.00E+00	0.00E+00	6.12E+00	-3.56E+03	1.76E+03	1.52E+00	1.30E+01	2.50E+01	2.03
83	1-propanol	tr	C 3 H 8 O 1	3.40E+01	1.52E+00	7.24E-01	2.24E-01	0.00E+00	6.11E+00	-3.46E+03	1.80E+03	1.43E+00	1.30E+01	2.52E+01	2.60
84	octafluorocyclobutane	tr	C 4 F 8	9.60E+01	2.51E+00	2.30E+00	1.58E+00	6.64E-01	7.91E+00	-2.24E+04	1.11E+04	1.20E-02	3.60E+01	3.35E+01	12.31
85	n-butane	tr	C 4 H 10	3.40E+01	1.91E+00	1.00E+00	5.00E-01	0.00E+00	6.02E+00	-3.47E+03	1.80E+03	0.00E+00	1.30E+01	2.90E+01	10.87
86	2-methyl-propane	tr	C 4 H 10	3.40E+01	1.73E+00	1.73E+00	0.00E+00	0.00E+00	6.03E+00	-3.56E+03	1.77E+03	6.00E-03	1.30E+01	2.86E+01	11.10
87	tert-butanol	tr	C 4 H 10 O 1	4.20E+01	1.72E+00	2.17E+00	0.00E+00	0.00E+00	6.43E+00	-5.28E+03	2.72E+03	1.54E+00	1.60E+01	3.16E+01	2.48
88	2-butanol	tr	C 4 H 10 O 1	4.20E+01	1.95E+00	1.26E+00	5.91E-01	0.00E+00	6.41E+00	-5.14E+03	2.65E+03	1.50E+00	1.60E+01	3.21E+01	3.27
89	2-methyl-1-propanol	tr	C 4 H 10 O 1	4.20E+01	1.88E+00	1.58E+00	3.65E-01	0.00E+00	6.41E+00	-5.14E+03	2.55E+03	1.39E+00	1.60E+01	3.20E+01	3.89
90	1-butanol	te	C 4 H 10 O 1	4.20E+01	2.02E+00	1.08E+00	5.12E-01	1.58E-01	6.39E+00	-4.94E+03	2.45E+03	1.42E+00	1.60E+01	3.23E+01	3.92
91	diethylether	tr	C 4 H 10 O 1	4.20E+01	1.99E+00	7.81E-01	4.08E-01	2.04E-01	6.40E+00	-5.05E+03	2.60E+03	1.15E+00	1.60E+01	3.36E+01	4.23
92	methyl propyl ether	tr	C 4 H 10 O 1	4.20E+01	1.90E+00	9.93E-01	4.08E-01	2.04E-01	6.40E+00	-3.53E+03	1.83E+03	1.21E+00	1.30E+01	2.61E+01	4.88
93	diethyl sulfide	tr	C 4 H 10 S 1	5.00E+01	3.14E+00	2.34E+00	1.22E+00	6.11E-01	6.88E+00	-4.56E+03	2.35E+03	1.94E+00	1.60E+01	4.31E+01	7.38
94	1-butanethiol	tr	C 4 H 10 S 1	5.00E+01	2.65E+00	1.52E+00	8.27E-01	4.73E-01	6.86E+00	-4.40E+03	2.27E+03	2.38E+00	1.50E+01	3.91E+01	9.03
95	diethyl amine	tr	C 4 H 11 N 1	4.20E+01	2.12E+00	9.57E-01	5.00E-01	2.50E-01	6.36E+00	-4.96E+03	2.56E+03	1.18E+00	1.60E+01	3.57E+01	1.69
96	thiophene	tr	C 4 H 4 S 1	4.40E+01	2.41E+00	1.61E+00	1.05E+00	6.79E-01	6.81E+00	-3.32E+03	1.72E+03	6.70E-01	1.30E+01	4.22E+01	7.35
97	isobutyronitrile	tr	C 4 H 7 N 1	3.80E+01	1.67E+00	1.28E+00	2.58E-01	0.00E+00	6.25E+00	-3.87E+03	1.92E+03	3.28E+00	1.40E+01	3.27E+01	4.76
98	butyronitrile	tr	C 4 H 7 N 1	3.80E+01	1.78E+00	9.08E-01	3.62E-01	1.12E-01	6.23E+00	-3.76E+03	1.86E+03	3.30E+00	1.40E+01	3.31E+01	4.77
99	tetrahydrofuran	tr	C 4 H 8 O 1	4.00E+01	2.08E+00	1.32E+00	8.27E-01	5.10E-01	6.40E+00	-4.81E+03	2.39E+03	1.67E+00	1.50E+01	3.04E+01	2.83
100	2-butanone	te	C 4 H 8 O 1	4.00E+01	1.76E+00	1.06E+00	4.98E-01	0.00E+00	6.35E+00	-4.47E+03	2.31E+03	2.70E+00	1.50E+01	3.17E+01	3.24
101	butyraldehyde	te	C 4 H 8 O 1	4.00E+01	1.85E+00	9.55E-01	4.07E-01	1.18E-01	6.33E+00	-4.32E+03	2.24E+03	2.54E+00	1.50E+01	3.17E+01	3.88
102	1-4-dioxane	tr	C 4 H 8 O 2	4.80E+01	2.15E+00	1.22E+00	7.44E-01	4.40E-01	6.72E+00	-6.78E+03	3.36E+03	6.00E-03	1.80E+01	3.52E+01	1.69
103	ethyl acetate	tr	C 4 H 8 O 2	4.80E+01	1.90E+00	9.25E-01	3.48E-01	2.03E-01	6.66E+00	-6.26E+03	3.22E+03	1.91E+00	1.80E+01	3.64E+01	4.18
104	methyl propanoate	te	C 4 H 8 O 2	4.80E+01	1.88E+00	9.30E-01	5.16E-01	1.44E-01	6.67E+00	-6.29E+03	3.24E+03	1.76E+00	1.80E+01	3.60E+01	4.47
105	propyl formate	tr	C 4 H 8 O 2	4.80E+01	1.97E+00	9.67E-01	3.90E-01	1.66E-01	6.64E+00	-6.06E+03	3.13E+03	3.90E+00	1.80E+01	3.70E+01	5.13
106	ethyl ethanoate	tr	C 4 H 8 O 2	4.80E+01	1.90E+00	9.25E-01	3.48E-01	2.03E-01	6.87E+00	-8.97E+03	4.46E+03	1.93E+00	2.20E+01	4.38E+01	5.60
107	2-bromobutane	tr	C 4 H 9 Br 1	6.80E+01	2.82E+00	2.75E+00	1.21E+00	0.00E+00	8.11E+00	-4.83E+03	2.40E+03	2.05E+00	1.60E+01	4.10E+01	9.03
108	1-bromobutane	te	C 4 H 9 Br 1	6.80E+01	3.09E+00	1.83E+00	1.05E+00	6.93E-01	8.10E+00	-3.65E+01	-2.94E+03	1.82E+00	1.60E+01	4.02E+01	9.41
109	2-chlorobutane	tr	C 4 H 9 Cl 1	5.00E+01	2.35E+00	1.93E+00	8.70E-01	0.00E+00	6.93E+00	-4.88E+03	2.51E+03	1.68E+00	1.60E+01	3.70E+01	8.54
110	1-chloro-2-methyl-propane	tr	C 4 H 9 Cl 1	5.00E+01	2.36E+00	1.86E+00	9.25E-01	0.00E+00	6.93E+00	-4.94E+03	2.55E+03	1.42E+00	1.60E+01	3.23E+01	8.62
111	1-chlorobutane	tr	C 4 H 9 Cl 1	5.00E+01	2.51E+00	1.42E+00	7.54E-01	4.00E-01	6.91E+00	-4.70E+03	2.42E+03	1.56E+00	1.60E+01	3.65E+01	8.94
112	N-n-dimethyl acetamide	tr	C 4 H 9 N 1 O 1	4.80E+01	1.82E+00	1.41E+00	6.30E-01	0.00E+00	6.65E+00	-6.35E+03	3.15E+03	3.57E+00	1.80E+01	4.02E+01	0.04
113	hypochlorous acid tert-butyl ester	te	C 4 H 9 O 1 Cl 1	5.80E+01	2.17E+00	2.34E+00	6.93E-01	0.00E+00	7.15E+00	-7.00E+03	3.48E+03	1.59E+00	1.90E+01	4.04E+01	7.54
114	2-methyl-2-butene	tr	C 5 H 10	4.00E+01	1.87E+00	1.37E+00	5.77E-01	0.00E+00	6.28E+00	-4.46E+03	2.30E+03	2.20E-01	1.50E+01	3.90E+01	9.79
115	2-pentene	tr	C 5 H 10	4.00E+01	2.03E+00	9.77E-01	4.71E-01	2.36E-01	6.26E+00	-4.30E+03	2.14E+03	4.10E-02	1.50E+01	3.84E+01	9.86
116	cyclopentane	tr	C 5 H 10	4.00E+01	2.50E+00	1.77E+00	1.25E+00	8.84E-01	6.32E+00	-4.74E+03	2.44E+03	1.00E-02	1.50E+01	3.39E+01	10.12
117	1-pentene	tr	C 5 H 10	4.00E+01	2.02E+00	1.08E+00	4.93E-01	2.04E-01	6.26E+00	-4.36E+03	2.25E+03	2.10E-01	1.50E+01	3.76E+01	10.18

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

118	3-methyl-1-butene	tr	C 5 H 10	4.00E+01	1.90E+00	1.48E+00	4.71E-01	0.00E+00	6.28E+00	-4.49E+03	2.32E+03	1.60E-01	1.50E+01	3.71E+01	10.31
119	tetrahydropyran	tr	C 5 H 10 O 1	4.80E+01	2.58E+00	1.67E+00	1.08E+00	6.87E-01	6.66E+00	-6.68E+03	3.32E+03	1.38E+00	1.80E+01	3.86E+01	4.36
120	3-methyl-2-butanone	tr	C 5 H 10 O 1	4.80E+01	2.15E+00	1.77E+00	8.13E-01	0.00E+00	6.62E+00	-6.30E+03	3.24E+03	2.60E+00	1.80E+01	3.83E+01	4.43
121	2-pentanone	tr	C 5 H 10 O 1	4.80E+01	2.26E+00	1.45E+00	6.02E-01	3.52E-01	6.60E+00	-6.08E+03	3.13E+03	2.68E+00	1.80E+01	3.87E+01	4.54
122	3-pentanone	tr	C 5 H 10 O 1	4.80E+01	2.33E+00	1.25E+00	7.89E-01	2.50E-01	6.60E+00	-6.15E+03	3.17E+03	2.61E+00	1.80E+01	3.89E+01	4.67
123	pentanal	tr	C 5 H 10 O 1	4.80E+01	2.35E+00	1.31E+00	6.76E-01	2.87E-01	6.57E+00	-5.90E+03	2.92E+03	2.57E+00	1.80E+01	3.88E+01	5.39
124	pentanoic acid	tr	C 5 H 10 O 2	5.60E+01	2.49E+00	1.50E+00	7.44E-01	3.28E-01	6.85E+00	-7.88E+03	4.05E+03	1.80E+00	2.10E+01	4.18E+01	4.84
125	isopropyl acetate	tr	C 5 H 10 O 2	5.60E+01	2.30E+00	1.66E+00	4.02E-01	3.32E-01	6.88E+00	-8.32E+03	4.13E+03	1.82E+00	2.10E+01	4.33E+01	5.28
126	n-propyl acetate	te	C 5 H 10 O 2	5.60E+01	2.40E+00	1.34E+00	5.09E-01	2.46E-01	6.86E+00	-8.03E+03	4.13E+03	1.91E+00	2.10E+01	4.36E+01	5.49
127	ethyl propanoate	tr	C 5 H 10 O 2	5.60E+01	2.46E+00	1.16E+00	5.94E-01	2.63E-01	6.87E+00	-8.11E+03	4.02E+03	1.82E+00	2.10E+01	4.38E+01	5.55
128	methyl isobutyrate	te	C 5 H 10 O 2	5.60E+01	2.38E+00	1.33E+00	6.82E-01	3.06E-01	6.88E+00	-8.06E+03	4.00E+03	1.76E+00	2.10E+01	4.32E+01	5.73
129	methyl butyrate	tr	C 5 H 10 O 2	5.60E+01	2.38E+00	1.33E+00	6.82E-01	3.06E-01	6.86E+00	-8.06E+03	4.14E+03	1.74E+00	2.10E+01	4.31E+01	5.80
130	1-bromopentane	tr	C 5 H 11 Br 1	7.60E+01	3.59E+00	2.19E+00	1.30E+00	7.40E-01	8.15E+00	-6.27E+03	3.22E+03	1.84E+00	1.90E+01	4.74E+01	11.10
131	2-chloro-2-methyl-butane	tr	C 5 H 11 Cl 1	5.80E+01	2.63E+00	3.09E+00	1.11E+00	0.00E+00	7.12E+00	-6.89E+03	3.53E+03	1.66E+00	1.90E+01	4.36E+01	7.48
132	1-chloropentane	tr	C 5 H 11 Cl 1	5.80E+01	3.01E+00	1.77E+00	1.00E+00	5.33E-01	7.06E+00	-6.32E+03	3.24E+03	1.57E+00	1.90E+01	4.36E+01	10.38
133	2-methylbutane	tr	C 5 H 12	4.20E+01	2.27E+00	1.80E+00	8.16E-01	0.00E+00	6.35E+00	-5.12E+03	2.55E+03	1.00E-02	1.60E+01	3.57E+01	11.32
134	pentane	tr	C 5 H 12	4.20E+01	2.41E+00	1.35E+00	7.07E-01	3.54E-01	6.32E+00	-4.94E+03	2.55E+03	0.00E+00	1.60E+01	3.60E+01	11.46
135	2-2-dimethyl-propane	tr	C 5 H 12	4.20E+01	2.00E+00	3.00E+00	0.00E+00	0.00E+00	6.37E+00	-5.26E+03	2.70E+03	0.00E+00	1.60E+01	3.51E+01	11.70
136	2-pentanol	tr	C 5 H 12 O 1	5.00E+01	2.45E+00	1.64E+00	7.06E-01	4.18E-01	6.65E+00	-6.85E+03	3.52E+03	1.53E+00	1.90E+01	3.93E+01	4.57
137	terbutyl methyl ether	tr	C 5 H 12 O 1	5.00E+01	2.11E+00	2.32E+00	6.12E-01	0.00E+00	6.69E+00	-7.23E+03	3.70E+03	1.28E+00	1.90E+01	3.96E+01	4.73
138	2-2-dimethyl-1-propanol	tr	C 5 H 12 O 1	5.00E+01	2.17E+00	2.72E+00	4.74E-01	0.00E+00	6.68E+00	-7.14E+03	3.55E+03	1.35E+00	1.90E+01	3.85E+01	4.91
139	2-methyl-1-butanol	tr	C 5 H 12 O 1	5.00E+01	2.42E+00	1.70E+00	1.01E+00	1.29E-01	6.66E+00	-6.91E+03	3.55E+03	1.35E+00	1.90E+01	3.91E+01	5.08
140	1-pentanol	tr	C 5 H 12 O 1	5.00E+01	2.52E+00	1.43E+00	7.62E-01	3.62E-01	6.62E+00	-6.57E+03	3.38E+03	1.41E+00	1.90E+01	3.93E+01	5.29
141	ethyl isopropyl ether	tr	C 5 H 12 O 1	5.00E+01	2.39E+00	1.50E+00	5.00E-01	3.33E-01	6.66E+00	-6.94E+03	3.45E+03	1.19E+00	1.90E+01	4.04E+01	5.30
142	3-methyl-1-butanol	tr	C 5 H 12 O 1	5.00E+01	2.38E+00	1.91E+00	7.06E-01	2.58E-01	6.65E+00	-6.82E+03	3.39E+03	1.43E+00	1.90E+01	3.90E+01	5.34
143	ethyl propyl ether	tr	C 5 H 12 O 1	5.00E+01	2.49E+00	1.20E+00	5.53E-01	2.89E-01	6.63E+00	-6.69E+03	3.33E+03	1.14E+00	1.90E+01	4.07E+01	5.55
144	methyl sec-butyl ether	tr	C 5 H 12 O 1	5.00E+01	2.34E+00	1.45E+00	9.77E-01	1.67E-01	6.66E+00	-6.99E+03	3.48E+03	1.25E+00	1.90E+01	4.01E+01	5.71
145	isobutyl methyl ether	tr	C 5 H 12 O 1	5.00E+01	2.26E+00	1.85E+00	5.00E-01	3.33E-01	6.66E+00	-6.91E+03	3.54E+03	1.15E+00	1.90E+01	4.00E+01	6.09
146	methyl butyl ether	tr	C 5 H 12 O 1	5.00E+01	2.40E+00	1.35E+00	7.02E-01	2.89E-01	6.63E+00	-6.66E+03	3.42E+03	1.18E+00	1.90E+01	4.03E+01	6.30
147	pyridine	tr	C 5 H 5 N 1	4.20E+01	1.85E+00	1.02E+00	5.66E-01	3.13E-01	6.47E+00	-4.60E+03	2.38E+03	1.93E+00	1.50E+01	4.36E+01	2.99
148	1-pentyne	tr	C 5 H 8	3.80E+01	1.85E+00	9.54E-01	3.94E-01	1.44E-01	6.20E+00	-3.78E+03	1.96E+03	3.60E-01	1.40E+01	3.56E+01	7.79
149	2-methyl-1-3-butadiene	tr	C 5 H 8	3.80E+01	1.55E+00	1.05E+00	3.48E-01	0.00E+00	6.22E+00	-3.90E+03	2.02E+03	1.90E-01	1.40E+01	3.94E+01	8.68
150	1-4-pentadiene	te	C 5 H 8	3.80E+01	1.63E+00	8.13E-01	3.33E-01	1.18E-01	6.20E+00	-3.79E+03	1.97E+03	8.00E-02	1.40E+01	3.95E+01	8.82
151	cyclopentene	tr	C 5 H 8	3.80E+01	2.15E+00	1.40E+00	9.08E-01	5.89E-01	6.26E+00	-4.13E+03	2.14E+03	1.50E-01	1.40E+01	3.59E+01	8.86
152	cyclopentanone	tr	C 5 H 8 O 1	4.60E+01	2.41E+00	1.75E+00	1.16E+00	7.69E-01	6.59E+00	-5.79E+03	2.99E+03	2.71E+00	1.70E+01	3.70E+01	3.37
153	pentanenitrile	tr	C 5 H 9 N 1	4.60E+01	2.28E+00	1.26E+00	6.42E-01	2.56E-01	6.49E+00	-5.28E+03	2.62E+03	3.33E+00	1.70E+01	4.02E+01	6.00
154	N-methyl-2-pyrrolidone	tr	C 5 H 9 N 1 O 1	5.40E+01	2.54E+00	1.92E+00	1.31E+00	7.83E-01	6.84E+00	-7.80E+03	3.87E+03	3.30E+00	2.00E+01	4.53E+01	-0.99
155	hexachlorobenzene	tr	C 6 Cl 6	1.38E+02	4.90E+00	4.15E+00	4.00E+00	2.00E+00	8.60E+00	-1.70E+04	8.58E+03	0.00E+00	3.30E+01	1.09E+02	21.94
156	1-hexyne	tr	C 6 H 10	4.60E+01	2.35E+00	1.31E+00	6.75E-01	2.79E-01	6.47E+00	-5.30E+03	2.73E+03	3.70E-01	1.70E+01	4.28E+01	9.45
157	1-5-hexadiene	tr	C 6 H 10	4.60E+01	2.13E+00	1.15E+00	5.75E-01	2.36E-01	6.54E+00	-5.31E+03	2.74E+03	1.50E-01	1.70E+01	4.64E+01	10.20
158	cyclohexene	tr	C 6 H 10	4.60E+01	2.65E+00	1.76E+00	1.16E+00	7.60E-01	6.60E+00	-5.81E+03	2.99E+03	1.70E-01	1.70E+01	4.36E+01	10.25

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

159	cyclohexanone	tr	C 6 H 10 O 1	5.40E+01	2.91E+00	2.10E+00	1.41E+00	9.46E-01	6.82E+00	-7.80E+03	4.00E+03	2.81E+00	2.00E+01	4.48E+01	3.99
160	1-ethenyl ethyl acetate	tr	C 6 H 10 O 2	6.20E+01	2.46E+00	1.55E+00	6.84E-01	3.10E-01	7.02E+00	-9.56E+03	4.75E+03	1.77E+00	2.30E+01	5.18E+01	6.16
161	hexanenitrile	tr	C 6 H 11 N 1	5.40E+01	2.78E+00	1.62E+00	8.92E-01	4.54E-01	6.71E+00	-6.65E+03	3.30E+03	3.61E+00	1.90E+01	6.19E+01	7.24
162	2-3-dimethyl-1-butene	tr	C 6 H 12	4.80E+01	2.30E+00	2.00E+00	9.86E-01	0.00E+00	6.57E+00	-6.30E+03	3.14E+03	3.38E-01	1.80E+01	4.42E+01	9.23
163	2-methyl-1-pentene	tr	C 6 H 12	4.80E+01	2.41E+00	1.71E+00	6.77E-01	4.27E-01	6.54E+00	-6.09E+03	3.13E+03	3.70E-01	1.80E+01	4.46E+01	11.00
164	cyclohexane	tr	C 6 H 12	4.80E+01	3.00E+00	2.12E+00	1.50E+00	1.06E+00	6.47E+00	-6.58E+03	3.38E+03	0.00E+00	1.80E+01	4.19E+01	11.29
165	1-hexene	te	C 6 H 12	4.80E+01	2.52E+00	1.43E+00	7.62E-01	3.48E-01	6.52E+00	-5.93E+03	3.05E+03	2.10E-01	1.80E+01	4.47E+01	11.39
166	4-methyl-1-pentene	te	C 6 H 12	4.80E+01	2.38E+00	1.92E+00	6.38E-01	3.33E-01	6.54E+00	-6.13E+03	3.15E+03	2.50E-01	1.80E+01	4.43E+01	11.49
167	methylcyclopentane	tr	C 6 H 12	4.80E+01	2.89E+00	2.39E+00	1.64E+00	1.13E+00	6.59E+00	-6.52E+03	3.25E+03	3.90E-02	1.80E+01	4.07E+01	11.60
168	4-hexen-1-ol	tr	C 6 H 12 O 1	5.60E+01	2.64E+00	1.43E+00	7.65E-01	3.76E-01	6.78E+00	-7.56E+03	3.76E+03	1.39E+00	2.10E+01	4.91E+01	4.95
169	cyclohexanol	tr	C 6 H 12 O 1	5.60E+01	3.07E+00	2.29E+00	1.57E+00	1.08E+00	6.86E+00	-8.59E+03	4.40E+03	1.50E+00	2.10E+01	4.51E+01	5.06
170	2-methyl-4-penten-3-ol	tr	C 6 H 12 O 1	5.60E+01	2.49E+00	1.97E+00	9.25E-01	2.72E-01	6.84E+00	-8.28E+03	4.12E+03	1.46E+00	2.10E+01	4.72E+01	5.18
171	1-hexen-3-ol	tr	C 6 H 12 O 1	5.60E+01	2.62E+00	1.59E+00	8.57E-01	4.14E-01	6.82E+00	-8.00E+03	3.97E+03	1.68E+00	2.10E+01	4.78E+01	5.38
172	3-methyl-2-pentanone	tr	C 6 H 12 O 1	5.60E+01	2.69E+00	1.92E+00	1.31E+00	2.87E-01	6.83E+00	-8.21E+03	4.08E+03	2.66E+00	2.10E+01	4.55E+01	5.56
173	2-hexanol	**te**	C 6 H 12 O 1	5.00E+01	2.95E+00	1.99E+00	9.75E-01	5.00E-01	6.85E+00	-8.65E+03	4.44E+03	1.52E+00	2.20E+01	4.63E+01	5.64
174	3-3-dimethyl-2-butanone	tr	C 6 H 12 O 1	5.60E+01	2.45E+00	2.81E+00	1.06E+00	0.00E+00	6.85E+00	-8.48E+03	4.34E+03	2.71E+00	2.10E+01	4.51E+01	5.66
175	4-methyl-2-pentanone	tr	C 6 H 12 O 1	5.60E+01	2.62E+00	2.30E+00	6.96E-01	5.75E-01	6.82E+00	-8.12E+03	4.17E+03	2.62E+00	2.10E+01	4.53E+01	5.68
176	2-hexanone	tr	C 6 H 12 O 1	5.60E+01	2.76E+00	1.81E+00	8.82E-01	4.26E-01	6.80E+00	-7.83E+03	4.02E+03	2.66E+00	2.10E+01	4.58E+01	5.87
177	2-methyl-3-pentanone	tr	C 6 H 12 O 1	5.60E+01	2.71E+00	1.97E+00	9.92E-01	4.08E-01	6.83E+00	-8.20E+03	4.21E+03	2.61E+00	2.10E+01	4.58E+01	5.89
178	3-hexanone	tr	C 6 H 12 O 1	5.60E+01	2.83E+00	1.64E+00	9.23E-01	4.56E-01	6.81E+00	-7.92E+03	4.06E+03	2.59E+00	2.10E+01	4.59E+01	6.02
179	1-hexanol	tr	C 6 H 12 O 1	5.00E+01	3.02E+00	1.78E+00	1.01E+00	5.39E-01	6.82E+00	-8.32E+03	4.14E+03	1.41E+00	2.20E+01	4.64E+01	6.68
180	hexanal	tr	C 6 H 12 O 1	5.60E+01	2.85E+00	1.66E+00	9.26E-01	4.78E-01	6.78E+00	-7.61E+03	3.91E+03	2.58E+00	2.10E+01	4.59E+01	6.70
181	hexanoic acid	tr	C 6 H 12 O 2	6.40E+01	2.99E+00	1.85E+00	9.94E-01	5.26E-01	7.01E+00	-9.73E+03	5.00E+03	1.80E+00	2.40E+01	4.89E+01	6.40
182	ethyl butyrate	tr	C 6 H 12 O 2	6.40E+01	2.96E+00	1.56E+00	7.59E-01	4.37E-01	7.03E+00	-1.00E+04	5.13E+03	1.81E+00	2.40E+01	5.09E+01	6.59
183	n-butyl-acetate	tr	C 6 H 12 O 2	6.40E+01	2.90E+00	1.69E+00	8.03E-01	3.60E-01	7.03E+00	-9.89E+03	5.07E+03	1.92E+00	2.40E+01	5.08E+01	6.70
184	isobutyl acetate	te	C 6 H 12 O 2	6.40E+01	2.76E+00	2.20E+00	6.20E-01	2.84E-01	7.05E+00	-1.02E+04	5.24E+03	1.94E+00	2.40E+01	5.04E+01	6.74
185	n-propyl propanoate	tr	C 6 H 12 O 2	6.40E+01	2.96E+00	1.57E+00	7.56E-01	3.18E-01	7.03E+00	-1.00E+04	4.97E+03	1.82E+00	2.40E+01	5.09E+01	6.99
186	methyl pentanoate	tr	C 6 H 12 O 2	6.40E+01	2.88E+00	1.68E+00	9.62E-01	4.23E-01	7.03E+00	-9.94E+03	4.93E+03	1.74E+00	2.40E+01	5.03E+01	7.14
187	1-chlorohexane	tr	C 6 H 13 Cl 1	6.60E+01	3.51E+00	2.13E+00	1.25E+00	7.10E-01	7.20E+00	-8.07E+03	4.13E+03	1.58E+00	2.20E+01	5.07E+01	11.86
188	2-2-dimethylbutane	tr	C 6 H 14	5.00E+01	2.56E+00	2.91E+00	1.06E+00	0.00E+00	6.63E+00	-7.13E+03	3.66E+03	7.00E-02	1.90E+01	4.22E+01	12.33
189	2-3-dimethyl-butane	te	C 6 H 14	5.00E+01	2.64E+00	2.49E+00	1.33E+00	0.00E+00	6.62E+00	-7.04E+03	3.61E+03	0.00E+00	1.90E+01	4.26E+01	12.35
190	2-methyl-pentane	tr	C 6 H 14	5.00E+01	2.77E+00	2.18E+00	8.66E-01	5.77E-01	6.60E+00	-6.81E+03	3.49E+03	2.00E-02	1.90E+01	4.27E+01	12.76
191	3-methylpentane	te	C 6 H 14	5.00E+01	2.81E+00	1.92E+00	1.39E+00	2.89E-01	6.60E+00	-6.90E+03	3.54E+03	7.00E-02	1.90E+01	4.27E+01	12.82
192	hexane	tr	C 6 H 14	5.00E+01	2.91E+00	1.71E+00	9.57E-01	5.00E-01	6.57E+00	-6.56E+03	3.37E+03	0.00E+00	1.90E+01	4.30E+01	12.96
193	3-methyl-3-pentanol	tr	C 6 H 14 O 1	5.80E+01	2.84E+00	2.20E+00	1.52E+00	2.50E-01	6.89E+00	-9.22E+03	4.72E+03	1.56E+00	2.20E+01	4.59E+01	4.85
194	2-3-dimethyl-2-butanol	tr	C 6 H 14 O 1	5.80E+01	2.67E+00	2.81E+00	1.41E+00	0.00E+00	6.91E+00	-9.34E+03	4.65E+03	1.50E+00	2.20E+01	4.56E+01	4.88
195	2-methyl-2-pentanol	tr	C 6 H 14 O 1	5.80E+01	2.78E+00	2.56E+00	8.62E-01	6.12E-01	6.88E+00	-9.07E+03	4.64E+03	1.55E+00	2.20E+01	4.60E+01	5.14
196	3-3-dimethyl-2-butanol	tr	C 6 H 14 O 1	5.80E+01	2.62E+00	3.04E+00	1.25E+00	0.00E+00	6.90E+00	-9.34E+03	4.65E+03	1.45E+00	2.20E+01	4.55E+01	5.43
197	2-methyl-3-pentanol	tr	C 6 H 14 O 1	5.80E+01	2.86E+00	2.22E+00	1.19E+00	4.71E-01	6.88E+00	-9.06E+03	4.51E+03	1.59E+00	2.20E+01	4.61E+01	5.63
198	3-methyl-2-pentanol	tr	C 6 H 14 O 1	5.80E+01	2.86E+00	2.13E+00	1.47E+00	3.41E-01	6.88E+00	-9.05E+03	4.50E+03	1.46E+00	2.20E+01	4.62E+01	5.66
199	3-hexanol	tr	C 6 H 14 O 1	5.80E+01	2.99E+00	1.85E+00	1.09E+00	5.37E-01	6.86E+00	-8.76E+03	4.36E+03	1.63E+00	2.20E+01	4.63E+01	5.85

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

200	4-methyl-2-pentanol	re	C 6 H 14 O 1	5.80E+01	2.81E+00	2.48E+00	8.13E-01	6.82E-01	6.87E+00	-8.94E+03	4.44E+03	1.50E+00	2.20E+01	4.62E+01	5.86
201	2-methyl-sec butyl methyl ether	tr	C 6 H 14 O 1	5.80E+01	2.67E+00	2.32E+00	1.40E+00	1.44E-01	6.90E+00	-9.30E+03	4.63E+03	1.29E+00	2.20E+01	4.69E+01	6.11
202	di-isopropylether	tr	C 6 H 14 O 1	5.80E+01	2.78E+00	2.23E+00	5.44E-01	5.44E-01	6.88E+00	-9.09E+03	4.65E+03	1.31E+00	2.20E+01	4.74E+01	6.44
203	2-2-dimethyl-1-butanol	tr	C 6 H 14 O 1	5.80E+01	2.73E+00	2.68E+00	1.39E+00	1.12E-01	6.90E+00	-9.21E+03	4.58E+03	1.30E+00	2.20E+01	4.57E+01	6.61
204	isopropyl propyl ether	tr	C 6 H 14 O 1	5.80E+01	2.89E+00	1.92E+00	6.55E-01	3.54E-01	6.86E+00	-8.76E+03	4.36E+03	1.17E+00	2.20E+01	4.76E+01	7.09
205	di-propylether	tr	C 6 H 14 O 1	5.80E+01	2.99E+00	1.61E+00	6.97E-01	3.91E-01	6.83E+00	-8.47E+03	4.21E+03	1.12E+00	2.20E+01	4.78E+01	7.75
206	tri-ethylamine	tr	C 6 H 15 N 1	5.80E+01	3.07E+00	1.62E+00	1.34E+00	6.71E-01	6.85E+00	-8.98E+03	4.47E+03	1.08E+00	2.20E+01	5.05E+01	4.21
207	butyl ethyl amine	tr	C 6 H 15 N 1	5.80E+01	3.12E+00	1.71E+00	9.57E-01	4.79E-01	6.80E+00	-8.34E+03	4.15E+03	1.15E+00	2.20E+01	4.99E+01	4.87
208	dipropyl amine	tr	C 6 H 15 N 1	5.80E+01	3.12E+00	1.75E+00	8.54E-01	4.79E-01	6.80E+00	-8.35E+03	4.28E+03	1.16E+00	2.20E+01	4.98E+01	4.90
209	1-3-dichlorobenzene	te	C 6 H 4 Cl 2	7.40E+01	2.95E+00	2.31E+00	1.26E+00	8.95E-01	7.58E+00	-7.72E+03	3.95E+03	8.80E-01	2.10E+01	6.55E+01	11.10
210	1-2-dichlorobenzene	tr	C 6 H 4 Cl 2	7.40E+01	2.96E+00	2.23E+00	1.58E+00	7.10E-01	7.60E+00	-7.85E+03	4.01E+03	1.35E+00	2.10E+01	6.45E+01	11.13
211	1-4-dichlorobenzene	tr	C 6 H 4 Cl 2	7.40E+01	2.95E+00	2.31E+00	1.31E+00	6.81E-01	7.57E+00	-7.70E+03	3.94E+03	0.00E+00	2.10E+01	6.60E+01	11.53
212	bromobenzene	tr	C 6 H 5 Br 1	7.60E+01	2.89E+00	2.21E+00	1.26E+00	7.19E-01	8.19E+00	-5.98E+03	2.97E+03	1.18E+00	1.80E+01	5.76E+01	10.02
213	chlorobenzene	tr	C 6 H 5 Cl 1	5.80E+01	2.48E+00	1.73E+00	9.85E-01	5.60E-01	7.14E+00	-6.06E+03	3.11E+03	9.30E-01	1.80E+01	5.51E+01	9.55
214	fluorobenzene	tr	C 6 H 5 F 1	5.00E+01	2.10E+00	1.30E+00	7.33E-01	4.15E-01	6.76E+00	-6.15E+03	3.05E+03	1.60E+00	1.80E+01	4.69E+01	8.48
215	iodobenzene	tr	C 6 H 5 I 1	9.40E+01	3.16E+00	2.52E+00	1.44E+00	8.23E-01	8.86E+00	-5.92E+03	3.04E+03	7.90E-01	1.80E+01	6.32E+01	10.90
216	nitrobenzene	tr	C 6 H 5 N 1 O 2	6.40E+01	2.46E+00	1.56E+00	9.45E-01	5.37E-01	7.13E+00	-9.67E+03	4.98E+03	5.24E+00	2.30E+01	5.99E+01	8.17
217	benzene	tr	C 6 H 6	4.20E+01	2.00E+00	1.15E+00	6.67E-01	3.85E-01	6.44E+00	-4.58E+03	2.37E+03	0.00E+00	1.50E+01	4.56E+01	7.82
218	5-methyl-furfural(5-methylfuraldehyde)	tr	C 6 H 6 O 2	5.80E+01	2.34E+00	1.57E+00	8.77E-01	5.17E-01	6.96E+00	-8.13E+03	4.03E+03	3.10E+00	2.10E+01	5.58E+01	4.85
219	4-methyl pyridine	tr	C 6 H 7 N 1	5.00E+01	2.26E+00	1.52E+00	8.47E-01	4.26E-01	6.70E+00	-6.29E+03	3.24E+03	2.27E+00	1.80E+01	5.20E+01	3.74
220	3-methyl pyridine	te	C 6 H 7 N 1	5.00E+01	2.26E+00	1.53E+00	8.09E-01	4.48E-01	6.70E+00	-6.29E+03	3.24E+03	2.07E+00	1.80E+01	5.22E+01	3.89
221	aniline	tr	C 6 H 7 N 1	5.00E+01	2.20E+00	1.41E+00	8.00E-01	4.53E-01	6.70E+00	-6.29E+03	3.24E+03	1.30E+00	1.80E+01	5.59E+01	4.99
222	1-4-cyclohexadiene	tr	C 6 H 8	4.40E+01	2.30E+00	1.41E+00	8.78E-01	5.42E-01	6.49E+00	-5.18E+03	2.67E+03	0.00E+00	1.60E+01	4.47E+01	8.60
223	1-heptyne	tr	C 7 H 12	5.40E+01	2.85E+00	1.66E+00	9.25E-01	4.77E-01	6.69E+00	-6.96E+03	3.58E+03	3.70E-01	2.00E+01	4.99E+01	10.95
224	cycloheptene	tr	C 7 H 12	5.40E+01	3.15E+00	2.11E+00	1.41E+00	9.37E-01	6.83E+00	-7.83E+03	4.01E+03	2.10E-01	2.00E+01	4.97E+01	11.30
225	1-methyl-cyclohexene	tr	C 7 H 12	5.40E+01	3.05E+00	2.30E+00	1.52E+00	9.99E-01	6.77E+00	-7.72E+03	3.96E+03	1.60E-01	2.00E+01	5.16E+01	11.54
226	1-6-heptadiene	tr	C 7 H 12	5.40E+01	2.63E+00	1.51E+00	8.16E-01	4.07E-01	6.69E+00	-6.97E+03	3.58E+03	8.00E-02	2.00E+01	5.37E+01	11.70
227	cycloheptane	tr	C 7 H 14	5.60E+01	3.50E+00	2.47E+00	1.75E+00	1.24E+00	6.82E+00	-8.60E+03	4.40E+03	2.00E-02	2.10E+01	4.86E+01	12.11
228	methylcyclohexane	te	C 7 H 14	5.60E+01	3.39E+00	2.74E+00	1.89E+00	1.31E+00	6.83E+00	-8.65E+03	4.42E+03	2.00E-02	2.10E+01	4.86E+01	12.80
229	2-heptene	tr	C 7 H 14	5.60E+01	3.03E+00	1.71E+00	9.60E-01	4.89E-01	6.73E+00	-7.58E+03	3.89E+03	4.00E-02	2.10E+01	5.36E+01	12.80
230	2-4-dimethyl-3-pentanone	tr	C 7 H 14 O 1	6.40E+01	3.09E+00	2.71E+00	1.14E+00	6.67E-01	7.03E+00	-1.05E+04	5.21E+03	2.45E+00	2.40E+01	5.24E+01	7.01
231	2-heptanone	tr	C 7 H 14 O 1	6.40E+01	3.26E+00	2.16E+00	1.13E+00	6.24E-01	6.97E+00	-9.67E+03	4.96E+03	2.66E+00	2.40E+01	5.28E+01	7.24
232	5-methyl-2-hexanone	te	C 7 H 14 O 1	6.40E+01	3.12E+00	2.63E+00	1.07E+00	4.92E-01	7.06E+00	-9.99E+03	5.12E+03	2.67E+00	2.40E+01	5.25E+01	7.33
233	4-heptanone	tr	C 7 H 14 O 1	6.40E+01	3.33E+00	2.04E+00	1.06E+00	6.83E-01	6.98E+00	-9.80E+03	5.02E+03	2.58E+00	2.40E+01	5.30E+01	7.41
234	heptanal	te	C 7 H 14 O 1	6.40E+01	3.35E+00	2.02E+00	1.18E+00	6.54E-01	6.96E+00	-9.42E+03	4.83E+03	2.59E+00	2.40E+01	5.30E+01	8.34
235	ethyl pentanoate	tr	C 7 H 14 O 2	7.20E+01	3.46E+00	1.91E+00	1.04E+00	5.54E-01	7.18E+00	-1.20E+04	5.97E+03	1.80E+00	2.70E+01	5.80E+01	7.97
236	isopentyl acetate	tr	C 7 H 14 O 2	7.20E+01	3.26E+00	2.52E+00	1.00E+00	4.38E-01	7.20E+00	-1.22E+04	6.27E+03	1.88E+00	2.70E+01	5.77E+01	8.00
237	isopropyl butyrate	te	C 7 H 14 O 2	7.20E+01	3.36E+00	2.29E+00	8.21E-01	5.26E-01	7.21E+00	-1.24E+04	6.16E+03	1.71E+00	2.70E+01	5.78E+01	8.03
238	n-pentyl-acetate	tr	C 7 H 14 O 2	7.20E+01	3.40E+00	2.05E+00	1.05E+00	5.68E-01	7.17E+00	-1.19E+04	6.07E+03	1.92E+00	2.70E+01	5.79E+01	8.08
239	methyl hexanoate	tr	C 7 H 14 O 2	7.20E+01	3.38E+00	2.03E+00	1.21E+00	6.21E-01	7.18E+00	-1.19E+04	6.09E+03	1.74E+00	2.70E+01	5.74E+01	8.29

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

240	1-ethyl piperidine	tr	C 7 H 15 N 1	6.40E+01	3.66E+00	2.49E+00	1.90E+00	1.28E+00	7.04E+00	-1.08E+04	5.38E+03	1.17E+00	2.40E+01	5.56E+01	4.84
241	3-3-dimethylpentane	tr	C 7 H 16	5.80E+01	3.12E+00	2.87E+00	1.91E+00	2.50E-01	6.85E+00	-9.16E+03	4.68E+03	6.00E-02	2.20E+01	4.94E+01	13.75
242	2-3-dimethyl-pentane	te	C 7 H 16	5.80E+01	3.18E+00	2.63E+00	1.78E+00	4.71E-01	6.83E+00	-9.01E+03	4.61E+03	4.00E-02	2.20E+01	4.97E+01	13.87
243	2-2-dimethyl-pentane	tr	C 7 H 16	5.80E+01	3.06E+00	3.31E+00	1.00E+00	7.50E-01	6.84E+00	-9.02E+03	4.61E+03	6.00E-02	2.20E+01	4.94E+01	14.05
244	2-4-dimethyl-pentane	te	C 7 H 16	5.80E+01	3.13E+00	3.02E+00	9.43E-01	9.43E-01	6.83E+00	-8.90E+03	4.55E+03	3.00E-02	2.20E+01	4.97E+01	14.09
245	n-heptane	tr	C 7 H 16	5.80E+01	3.41E+00	2.06E+00	1.21E+00	6.77E-01	6.78E+00	-8.31E+03	4.26E+03	0.00E+00	2.20E+01	5.01E+01	14.51
246	2-methyl-hexane	tr	C 7 H 16	5.80E+01	3.27E+00	2.54E+00	1.14E+00	6.12E-01	6.80E+00	-8.62E+03	4.42E+03	4.00E-02	2.20E+01	4.99E+01	14.60
247	3-ethyl-3-pentanol	tr	C 7 H 16 O 1	6.60E+01	3.41E+00	2.29E+00	1.97E+00	7.50E-01	7.08E+00	-1.14E+04	5.68E+03	1.48E+00	2.50E+01	5.32E+01	5.94
248	2-3-dimethyl-3-pentanol	te	C 7 H 16 O 1	6.60E+01	3.23E+00	2.87E+00	1.96E+00	4.08E-01	7.09E+00	-1.16E+04	5.78E+03	1.55E+00	2.50E+01	5.28E+01	5.96
249	2-3-dimethyl-2-pentanol	te	C 7 H 16 O 1	6.60E+01	3.20E+00	2.96E+00	1.82E+00	5.00E-01	7.08E+00	-1.15E+04	5.74E+03	1.44E+00	2.50E+01	5.29E+01	6.02
250	2-4-dimethyl-2-pentanol	tr	C 7 H 16 O 1	6.60E+01	3.14E+00	3.41E+00	9.08E-01	9.99E-01	7.07E+00	-1.14E+04	5.67E+03	1.56E+00	2.50E+01	5.30E+01	6.17
251	3-methyl-3-hexanol	tr	C 7 H 16 O 1	6.60E+01	3.34E+00	2.60E+00	1.55E+00	7.15E-01	7.06E+00	-1.13E+04	5.61E+03	1.62E+00	2.50E+01	5.30E+01	6.29
252	2-methyl-2-hexanol	tr	C 7 H 16 O 1	6.60E+01	3.28E+00	2.92E+00	1.14E+00	6.09E-01	7.05E+00	-1.10E+04	5.48E+03	1.48E+00	2.50E+01	5.29E+01	6.49
253	2-2-dimethyl-3-pentanol	tr	C 7 H 16 O 1	6.60E+01	3.16E+00	3.29E+00	1.39E+00	6.12E-01	7.08E+00	-1.15E+04	5.74E+03	1.53E+00	2.50E+01	5.26E+01	6.66
254	2-4-dimethyl-3-pentanol	tr	C 7 H 16 O 1	6.60E+01	3.23E+00	2.98E+00	1.37E+00	7.70E-01	7.07E+00	-1.12E+04	5.58E+03	1.54E+00	2.50E+01	5.31E+01	6.82
255	1-heptanol	tr	C 7 H 16 O 1	6.60E+01	3.52E+00	2.14E+00	1.26E+00	7.15E-01	6.99E+00	-1.02E+04	5.22E+03	1.41E+00	2.50E+01	5.35E+01	8.09
256	benzonitrile	tr	C 7 H 5 N 1	5.40E+01	2.38E+00	1.48E+00	9.03E-01	5.13E-01	6.81E+00	-6.65E+03	3.43E+03	3.61E+00	1.90E+01	6.19E+01	7.46
257	benzoic acid	tr	C 7 H 6 O 2	6.40E+01	2.59E+00	1.67E+00	1.02E+00	5.81E-01	7.10E+00	-9.54E+03	4.90E+03	2.25E+00	2.30E+01	6.14E+01	7.60
258	benzylchloride	tr	C 7 H 7 Cl 1	6.60E+01	3.06E+00	1.89E+00	1.31E+00	7.45E-01	7.28E+00	-7.90E+03	3.93E+03	1.33E+00	2.10E+01	6.15E+01	10.37
259	3-nitrotoluene	tr	C 7 H 7 N 1 O 2	7.20E+01	2.87E+00	2.07E+00	1.19E+00	7.24E-01	7.28E+00	-1.18E+04	5.85E+03	5.46E+00	2.60E+01	6.86E+01	8.87
260	2-nitrotoluene	tr	C 7 H 7 N 1 O 2	7.20E+01	2.88E+00	2.01E+00	1.28E+00	7.52E-01	7.29E+00	-1.21E+04	6.22E+03	5.00E+00	2.60E+01	6.87E+01	9.37
261	2-methoxynitrobenzene (2-nitroanisole)	tr	C 7 H 7 N 1 O 3	8.00E+01	2.99E+00	1.90E+00	1.26E+00	7.63E-01	7.46E+00	-1.47E+04	7.28E+03	6.24E+00	2.90E+01	7.47E+01	8.52
262	1-6heptadiyne	tr	C 7 H 8	5.00E+01	2.28E+00	1.26E+00	6.42E-01	2.93E-01	6.59E+00	-7.05E+03	3.62E+03	4.10E-01	1.90E+01	5.38E+01	8.04
263	1-3-5-cycloheptatriene	tr	C 7 H 8	5.00E+01	2.48E+00	1.48E+00	8.78E-01	5.21E-01	6.67E+00	-6.26E+03	3.11E+03	3.79E-01	1.80E+01	5.60E+01	8.98
264	toluene	tr	C 7 H 8	5.00E+01	2.41E+00	1.65E+00	9.40E-01	5.34E-01	6.68E+00	-6.27E+03	3.23E+03	2.60E-01	1.80E+01	4.56E+01	9.13
265	o-cresol (2-hydroxytoluene)	tr	C 7 H 8 O 1	5.80E+01	2.55E+00	1.79E+00	1.12E+00	5.63E-01	6.93E+00	-8.30E+03	4.27E+03	1.40E+00	2.10E+01	5.90E+01	5.51
266	m-cresol (3-hydroxytoluene)	te	C 7 H 8 O 1	5.80E+01	2.54E+00	1.84E+00	1.00E+00	6.28E-01	6.93E+00	-8.19E+03	4.21E+03	9.60E-01	2.10E+01	5.92E+01	5.62
267	p-cresol (4-hydroxytoluene)	tr	C 7 H 8 O 1	5.80E+01	2.54E+00	1.84E+00	1.03E+00	5.45E-01	6.92E+00	-8.16E+03	4.05E+03	1.18E+00	2.10E+01	5.96E+01	5.74
268	methoxybenzene(anisole)	tr	C 7 H 8 O 1	5.80E+01	2.52E+00	1.52E+00	9.79E-01	5.57E-01	6.93E+00	-8.25E+03	4.10E+03	1.09E+00	1.80E+01	5.97E+01	8.20
269	2-aminotoluene	tr	C 7 H 9 N 1	5.80E+01	2.62E+00	1.86E+00	1.19E+00	5.87E-01	6.91E+00	-8.27E+03	4.12E+03	1.34E+00	2.10E+01	6.39E+01	5.91
270	o-xylene	te	C 8 H 10	5.80E+01	2.83E+00	2.08E+00	1.43E+00	6.63E-01	6.89E+00	-8.25E+03	4.10E+03	4.44E-01	2.10E+01	6.25E+01	10.33
271	ethyl benzene	tr	C 8 H 10	5.80E+01	2.97E+00	1.84E+00	1.25E+00	7.14E-01	6.88E+00	-8.13E+03	4.17E+03	3.30E-01	2.10E+01	6.15E+01	10.40
272	m-xylene	tr	C 8 H 10	5.80E+01	2.82E+00	2.16E+00	1.17E+00	8.07E-01	6.88E+00	-8.14E+03	4.18E+03	2.60E-01	2.10E+01	6.26E+01	10.41
273	p-xylene	te	C 8 H 10	5.80E+01	2.82E+00	2.15E+00	1.22E+00	6.37E-01	6.88E+00	-8.12E+03	4.16E+03	5.00E-02	2.10E+01	6.30E+01	10.41
274	ethoxybenzene	tr	C 8 H 10 O 1	6.60E+01	3.11E+00	1.75E+00	1.05E+00	6.83E-01	7.09E+00	-1.02E+04	5.08E+03	1.12E+00	2.40E+01	6.78E+01	9.66
275	4-ethenylcyclohexene	tr	C 8 H 12	6.00E+01	3.21E+00	2.28E+00	1.64E+00	1.03E+00	6.93E+00	-9.08E+03	4.65E+03	1.40E-01	2.20E+01	5.93E+01	11.70
276	1-octyne	tr	C 8 H 14	6.20E+01	3.35E+00	2.01E+00	1.17E+00	6.54E-01	6.88E+00	-8.74E+03	4.48E+03	3.70E-01	2.30E+01	5.70E+01	12.45
277	cyclohexyl acetate	tr	C 8 H 14 O 2	7.80E+01	3.96E+00	2.87E+00	1.88E+00	1.42E+00	7.36E+00	-1.46E+04	7.27E+03	1.80E+00	2.90E+01	6.36E+01	7.91
278	cyclooctane	tr	C 8 H 16	6.40E+01	4.00E+00	2.83E+00	2.00E+00	1.41E+00	7.02E+00	-1.08E+04	5.50E+03	0.00E+00	2.40E+01	5.51E+01	13.58
279	cis-1-2-dimethylcyclohexane	tr	C 8 H 16	6.40E+01	3.80E+00	3.24E+00	2.54E+00	1.50E+00	7.04E+00	-1.09E+04	5.59E+03	1.00E-02	2.40E+01	5.53E+01	13.85

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

280	1-octene	te	C 8 H 16	6.40E+01	3.52E+00	2.14E+00	1.26E+00	7.15E-01	6.92E+00	-9.45E+03	4.84E+03	2.10E-01	2.40E+01	5.89E+01	14.65
281	octanal	tr	C 8 H 16 O 1	7.20E+01	3.85E+00	2.37E+00	1.43E+00	8.31E-01	7.11E+00	-1.14E+04	5.82E+03	2.59E+00	2.70E+01	6.01E+01	9.02
282	n-hexyl-acetate	tr	C 8 H 16 O 2	8.00E+01	3.90E+00	2.40E+00	1.30E+00	7.45E-01	7.31E+00	-1.39E+04	7.10E+03	1.92E+00	3.00E+01	6.50E+01	9.43
283	1-propyl piperidine	tr	C 8 H 17 N 1	7.20E+01	4.16E+00	2.90E+00	1.99E+00	1.50E+00	7.19E+00	-1.29E+04	6.45E+03	1.15E+00	2.70E+01	6.27E+01	6.85
284	octane	tr	C 8 H 18	6.60E+01	3.91E+00	2.41E+00	1.46E+00	8.54E-01	6.96E+00	-1.02E+04	5.20E+03	0.00E+00	2.50E+01	5.71E+01	16.02
285	2-2-3-trimethyl-3-pentanol	tr	C 8 H 18 O 1	7.40E+01	3.53E+00	3.88E+00	2.30E+00	5.30E-01	7.27E+00	-1.44E+04	7.16E+03	1.54E+00	2.80E+01	5.95E+01	6.95
286	1-octanol	te	C 8 H 18 O 1	7.40E+01	4.02E+00	2.49E+00	1.51E+00	8.92E-01	7.14E+00	-1.21E+04	6.21E+03	1.40E+00	2.80E+01	6.05E+01	9.56
287	di-butyl-ether	tr	C 8 H 18 O 1	7.40E+01	3.99E+00	2.32E+00	1.28E+00	5.95E-01	7.16E+00	-1.23E+04	6.30E+03	1.12E+00	2.80E+01	6.21E+01	10.76
288	styrene	tr	C 8 H 8	5.60E+01	2.61E+00	1.61E+00	1.04E+00	5.89E-01	6.84E+00	-7.42E+03	3.81E+03	2.00E-02	2.00E+01	6.63E+01	9.80
289	acetophenone	tr	C 8 H 8 O 1	6.40E+01	2.86E+00	1.92E+00	1.18E+00	6.73E-01	7.00E+00	-9.49E+03	4.87E+03	2.79E+00	2.30E+01	6.54E+01	6.88
290	indan	tr	C 9 H 10	6.40E+01	3.53E+00	2.62E+00	2.01E+00	1.51E+00	7.06E+00	-9.80E+03	5.02E+03	4.50E-01	2.30E+01	6.79E+01	11.01
291	m-methyl styrene	tr	C 9 H 10	6.40E+01	3.02E+00	2.12E+00	1.28E+00	7.97E-01	7.03E+00	-9.41E+03	4.68E+03	2.73E-01	2.30E+01	7.47E+01	11.21
292	p-methyl styrene	te	C 9 H 10	6.40E+01	3.02E+00	2.11E+00	1.31E+00	7.03E-01	7.02E+00	-9.36E+03	4.66E+03	2.94E-01	2.30E+01	7.59E+01	11.21
293	p-ethyl-toluene	tr	C 9 H 12	6.60E+01	3.38E+00	2.34E+00	1.53E+00	8.24E-01	7.06E+00	-1.01E+04	5.19E+03	4.00E-02	2.40E+01	7.03E+01	11.16
294	o-ethyl-toluene	tr	C 9 H 12	6.60E+01	3.39E+00	2.28E+00	1.64E+00	1.01E+00	7.07E+00	-1.04E+04	5.15E+03	5.24E-01	2.40E+01	7.02E+01	11.18
295	1-2-3-trimethyl-benzene	tr	C 9 H 12	6.60E+01	3.24E+00	2.52E+00	1.88E+00	8.98E-01	7.08E+00	-1.04E+04	5.33E+03	5.20E-01	2.40E+01	7.07E+01	11.39
296	isopropylbenzene	tr	C 9 H 12	6.60E+01	3.35E+00	2.57E+00	1.47E+00	8.38E-01	7.07E+00	-1.03E+04	5.14E+03	2.14E-01	2.40E+01	6.79E+01	11.53
297	1-2-4-trimethyl-benzene	tr	C 9 H 12	6.60E+01	3.24E+00	2.59E+00	1.66E+00	8.91E-01	7.07E+00	-1.03E+04	5.26E+03	2.30E-01	2.40E+01	7.12E+01	11.67
298	1-3-5-trimethylbenzene	te	C 9 H 12	6.60E+01	3.23E+00	2.67E+00	1.37E+00	1.20E+00	7.06E+00	-1.02E+04	5.22E+03	0.00E+00	2.40E+01	7.10E+01	11.67
299	propyl benzene	tr	C 9 H 12	6.60E+01	3.47E+00	2.24E+00	1.38E+00	9.33E-01	7.05E+00	-1.00E+04	5.00E+03	2.14E-01	2.40E+01	6.85E+01	11.82
300	2-6-dimethyl-4-heptanone	tr	C 9 H 18 O 1	8.00E+01	4.04E+00	3.73E+00	1.27E+00	9.94E-01	7.32E+00	-1.46E+04	7.47E+03	2.48E+00	3.00E+01	6.62E+01	9.08
301	2-nonanone	tr	C 9 H 18 O 1	8.00E+01	4.26E+00	2.87E+00	1.63E+00	9.77E-01	7.27E+00	-1.37E+04	6.98E+03	2.62E+00	3.00E+01	6.69E+01	9.70
302	5-nonanone	tr	C 9 H 18 O 1	8.00E+01	4.33E+00	2.75E+00	1.62E+00	8.73E-01	7.28E+00	-1.39E+04	6.90E+03	2.53E+00	3.00E+01	6.72E+01	9.99
303	nonanal	tr	C 9 H 18 O 1	8.00E+01	4.35E+00	2.72E+00	1.68E+00	1.01E+00	7.25E+00	-1.34E+04	6.84E+03	2.59E+00	3.00E+01	6.72E+01	11.23
304	butyl pentanoate	tr	C 9 H 18 O 2	8.80E+01	4.46E+00	2.68E+00	1.50E+00	7.24E-01	7.44E+00	-1.63E+04	8.10E+03	1.80E+00	3.30E+01	7.25E+01	9.87
305	1-nonanol	tr	C 9 H 20 O 1	8.20E+01	4.52E+00	2.84E+00	1.76E+00	1.07E+00	7.28E+00	-1.42E+04	7.25E+03	1.41E+00	3.10E+01	6.76E+01	11.03
306	1-3-nonanediol	tr	C 9 H 20 O 2	9.00E+01	4.60E+00	3.02E+00	1.85E+00	1.10E+00	7.46E+00	-1.71E+04	8.49E+03	1.94E+00	3.40E+01	7.10E+01	6.41
307	tert-butyl-benzene	te	C 10 H 14	7.40E+01	3.66E+00	3.62E+00	1.64E+00	9.38E-01	7.25E+00	-1.29E+04	6.56E+03	3.00E-01	2.70E+01	7.49E+01	12.44
308	sec-butyl-benzene	te	C 10 H 14	7.40E+01	3.89E+00	2.72E+00	1.98E+00	1.02E+00	7.23E+00	-1.25E+04	6.41E+03	2.70E-01	2.70E+01	7.55E+01	12.96
309	1-2-4-5-tetramethylbenzene	tr	C 10 H 14	7.40E+01	3.65E+00	3.02E+00	2.11E+00	1.10E+00	7.24E+00	-1.23E+04	6.27E+03	0.00E+00	2.70E+01	5.16E+01	14.58
310	butyl benzene	tr	C 10 H 14	7.40E+01	3.97E+00	2.59E+00	1.66E+00	1.03E+00	7.20E+00	-1.20E+04	1.10E+02	3.40E-01	2.70E+01	7.61E+01	17.85
311	1-decanol	tr	C 10 H 22 O 1	9.00E+01	5.02E+00	3.20E+00	2.01E+00	1.25E+00	7.40E+00	-1.63E+04	8.33E+03	1.41E+00	3.40E+01	7.47E+01	12.38
312	2-4-dimethyl-2-4-octanediol	tr	C 10 H 22 O 2	9.80E+01	4.71E+00	4.59E+00	1.88E+00	1.58E+00	7.65E+00	-2.16E+04	1.08E+04	1.83E+00	3.70E+01	7.75E+01	6.66
313	2-propyl-1-3-heptanediol	tr	C 10 H 22 O 2	9.80E+01	5.05E+00	3.44E+00	2.29E+00	1.43E+00	7.51E+00	-1.83E+04	9.12E+03	1.79E+00	3.40E+01	7.09E+01	6.85
314	1-methyl naphthalene	tr	C 11 H 10	7.60E+01	3.82E+00	2.80E+00	2.01E+00	1.39E+00	7.31E+00	-1.26E+03	6.52E+03	2.70E-01	2.70E+01	9.23E+01	12.55
315	2-4-dimethyl-2-4-nonanediol	tr	C 11 H 24 O 2	1.06E+02	5.21E+00	4.94E+00	2.13E+00	1.78E+00	7.75E+00	-2.41E+04	1.20E+04	1.84E+00	4.00E+01	8.46E+01	7.82
316	1-4-dimethyl naphthalene	tr	C 12 H 12	8.40E+01	4.24E+00	3.24E+00	2.36E+00	1.60E+00	7.45E+00	-1.53E+04	7.79E+03	1.00E-02	3.00E+01	1.01E+02	13.55
317	1-ethyl naphthalene	te	C 12 H 12	8.40E+01	4.38E+00	2.99E+00	2.26E+00	1.60E+00	7.45E+00	-1.51E+04	7.73E+03	3.10E-01	3.00E+01	9.97E+01	13.60
318	1-3-dimethyl naphthalene	tr	C 12 H 12	8.40E+01	4.23E+00	3.30E+00	2.21E+00	1.69E+00	7.45E+00	-1.51E+04	7.71E+03	3.60E-01	3.00E+01	1.02E+02	13.90
319	1-dodecanol	tr	C 12 H 26 O 1	1.06E+02	6.02E+00	3.91E+00	2.51E+00	1.60E+00	7.62E+00	-2.08E+04	1.03E+04	1.41E+00	4.00E+01	8.88E+01	15.31
320	2-butyl-1-3-butanediol	tr	C 12 H 26 O 2	1.14E+02	6.05E+00	4.15E+00	2.81E+00	1.71E+00	7.83E+00	-2.60E+04	1.29E+04	1.61E+00	4.30E+01	9.23E+01	9.84

Table 5. Molecular Descriptors and Experimental Aqueous Infinity Dilution Activity Coefficients for the Data Set of 325 Organic Compounds* (Continued)

321	1-tetradecanol	tr	C 14 H 30 O 1	1.22E+02	7.02E+00	4.61E+00	3.01E+00	1.95E+00	7.80E+00	-2.55E+04	1.27E+04	1.40E+00	4.60E+01	1.03E+02	17.50
322	1-pentadecanol	tr	C 15 H 32 O 1	1.30E+02	7.52E+00	4.97E+00	3.26E+00	2.13E+00	7.88E+00	-2.79E+04	1.39E+04	1.40E+00	4.90E+01	1.10E+02	18.77
323	1-hexadecanolcomo	tr	C 16 H 34 O 1	1.38E+02	8.02E+00	5.32E+00	3.51E+00	2.31E+00	7.96E+00	-3.04E+04	1.52E+04	1.40E+00	5.20E+01	1.17E+02	19.77
324	1-heptadecanol	tr	C 17 H 36 O 1	1.46E+02	8.52E+00	5.67E+00	3.76E+00	2.48E+00	8.03E+00	-3.29E+04	1.64E+04	1.40E+00	5.50E+01	5.51E-01	21.30
325	1-octadecanol	tr	C 18 H 38 O 1	1.54E+02	9.02E+00	6.03E+00	4.01E+00	2.66E+00	8.10E+00	-3.55E+04	1.77E+04	1.40E+00	5.80E+01	1.31E+02	23.34

aliphatic hydrocarbons, with greater scatter in the Mitchell and Jurs (1998) model. The current "best" model performed reasonably well over the entire data range covered except for the outlier 1-octene.

Conclusions

The integration of self-organizing maps (SOMs) with a fuzzy ARTMAP neural system was applied to develop QSPRs for the aqueous infinite dilution activity coefficient of organic compounds based in a heterogeneous data set of 325 organic compounds. The present study demonstrated that SOMs can be effectively used to classify organic chemicals according to their structural information (that is, in terms of molecular descriptors). A SOM-based analysis was shown to be effective for selecting the most suitable set of descriptors from an initial set, and for generating complementary interpolated input information for training. The QSPRs developed for $\ln\gamma^{\infty}$ performed with remarkable predictive generalization capabilities.

The fuzzy-ARTMAP–based QSPR developed with 11 descriptors, based on the data set of 325 compounds, performed with average absolute errors of 0.02 (0.36%) and 0.52 (6.64%) $\ln\gamma^{\infty}$ units, for the training and test sets, respectively. This performance was superior to that of other QSPRs reported in the literature. When the prototypes were added to the training set, the average absolute error slightly increased to 0.05 (1.07%) $\ln\gamma^{\infty}$ units and for the training set and decreased to 0.40 (5.36%) $\ln\gamma^{\infty}$ units for the test set. The performance of the $\ln\gamma^{\infty}$ QSPR, based only on four molecular quantum similarity measures, also selected by means of SOMs from a limited pool of six similarity measures, was better than that of previous QSPR models, with average absolute errors of 0.02 (0.38%) and 0.92 (11.2%) $\ln\gamma^{\infty}$ units for the training and test sets, respectively. The present results suggest that it should be possible to develop accurate QSPRs using the information contained in the quantum similarity matrices. Such an approach, however, will require improvements in the calculation of the quantum atomic density functions of molecules with heteroatoms using the metrics given by different quantum operators.

Although the present study focused on the aqueous infinite dilution activity coefficient as a case study, the present approach of using SOM analysis for features extraction, in combination with the modified fuzzy-ARTMAP classifier for variable prediction, could be an effective tool in various chemical engineering applications for the identification of critical (or significant) variables or parameters, pattern recognition, and establishing parameter–property relations.

Available supporting information

We provide Table 5 as supplemental, supporting data, listing the best set of molecular descriptors and the experimental infinity dilution activity coefficients for the 325 organic compounds considered.

Acknowledgments

The authors are grateful for the financial support received from the "Dirección General de Investigación Científica y Técnica," Projects PPQ2000-1339 and PPQ2001-1519, and from the CIRIT "Programa de Grups de Recerca Consolidats de la Generalitat de Catalunya," Projects 1998SGR-00102 and 2000SGR-00103.

Literature Cited

Agrawal, T., T. Imielinsky, and A. Swami, "Database Mining: A Performance Perspective," *IEEE Trans. Knowl. Data Eng.*, **5**, 6 (1993).

Amat, L., and R. Carbó-Dorca, "Quantum Similarity Measures under Atomic Shell Approximation: First Order Density Fitting Using Elementary Jacobi Rotations," *J. Comput. Chem.*, **18**, 2023 (1997).

Amat, L., R. Carbó-Dorca, and R. Ponec, "Simple Linear QSAR Models Based on Quantum Similarity Measures," *J. Med. Chem.*, **42**, 5169 (1999).

Amat, L., D. Robert, E. Besalú, and R. Carbó-Dorca, "Molecular Quantum Similarity Measures Tuned 3D QSAR: An Antitumoral Family Validation Study," *J. Chem. Inf. Comput. Sci.*, **38**, 624 (1998).

Basak, S., and V. Maguson, "Determining Structural Similarity of Chemicals Using Graph Theoretic Indices," *Discrete Appl. Math.*, **19**, 17 (1988).

Bishop, C. M., *Neural Networks for Pattern Recognition*, Oxford Univ. Press, Oxford, UK (1995).

Bünz, P., B. Braun, and R. Janowsky, "Application of Quantitative Structure-Performance Relationship and Neural Network Models for the Prediction of Physical Properties from Molecular Structure," *Ind. Eng. Chem. Res.*, **37**, 3043 (1998).

Carbó-Dorca, R., and E. Besalú, "A General Survey of Molecular Quantum Similarity," *Theor. Chem.*, **45**, 11 (1998).

Carpenter, G., and S. Grossberg, "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine." *Comput. Vis. Graphics Image Process.*, **37**, 54 (1987).

Carpenter, G., S. Grossberg, N. Marcuzon, J. Reynolds, and D. Rosen, "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps," *IEEE Trans. Neural Networks*, **3**, 698 (1992).

Carpenter, G., S. Grossberg, N. Marcuzon, and D. B. Rosen, "Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System," *Neural Networks*, **4**, 759 (1991).

Cheng, B., and D. M. Titterington, "Neural Networks: A Review from a Statistical Perspective," *Stat. Sci.*, **9**(1), 2 (1994).

Chow, H., H. Chen, T. Ng, P. Myrdal, and S. H. Yalkowsky, "Using Backpropagation Networks for the Estimation of Aqueous Activity Coefficients of Aromatic Organic Compounds," *J. Chem. Inf. Comput. Sci.*, **35**, 723 (1995).

Cramer, R., D. Patterson, and J. Bunce, "Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carried Proteins," *J. Am. Chem. Soc.*, **110**(18), 5959 (1988).

Cybenko, G., "Approximation by Superposition of Sigmoidal Functions," *Math. Control Signal Syst.*, **2**, 303 (1989).

Egolf, L., and P. Jurs, "Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques," *J. Chem. Inf. Comput. Sci.*, **33**, 616 (1993).

Erwing, E., K. Obermayer, and K. Schulten, "Self-Organizing Maps: Stationary States, Metastability and Convergence Rate," *Biol. Cybern.*, **67**(1), 35 (1992).

Espinosa, G., A. Arenas, and F. Giralt, "Prediction of Boiling Points of Organic Compounds from Molecular Descriptors by Using Backpropagation Neural Networks," in *Fundamentals of Molecular Similarity*, R. Carbó-Dorca, ed., Kluwer Academic, Dordrecht, 1 (2001a).

Espinosa, G., A. Arenas, and F. Giralt, "Integrated SOM-Fuzzy ARTMAP Neural System for the Evaluation of Toxicity," *J. Chem. Inf. Comput. Sci.*, **42**(2), 343 (2002).

Espinosa, G., A. Arenas, and F. Giralt, "QSAR for TD50 of Aromatic Compounds by Using an Integrated SOM-Fuzzy ARTMAP Based Neural System with Topological and Quantum Molecular Similarity Descriptors," in *Fundamentals of Molecular Similarity* (V Girona Seminar on Molecular Similarity, Girona, Spain), Kluwer Academic, Dordrecht, The Netherlands (2003).

Espinosa, G., D. Yaffe, A. Arenas, Y. Cohen, and F. Giralt, "A Fuzzy ARTMAP Based on Quantitative Structure-Property Relationships (QSPRs) for Predicting Physical Properties of Organic Compounds," *Ind. Eng. Chem. Res.*, **40**(12), 2757 (2001b).

Espinosa, G., D. Yaffe, Y. Cohen, A. Arenas, and F. Giralt, "Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons," *J. Chem. Inf. Comput. Sci.*, **40**, 859 (2000).

Fayyad, U. M., "Data Mining and Knowledge Discovery: Making Sense Out of Data," *IEEE Expert*, 20 (1996).

Ferre-Gine, J., R. Rallo, A. Arenas, and F. Giralt, "Identification of Coherent Structures in Turbulent Shear Flows with a Fuzzy ARTMAP Neural Network," *Int. J. Neural Syst.*, **7**(5), 559 (1996).

Fessant, F., S. Bengio, and D. Collobert, "Use of Modular Architectures for Time Series Prediction," *Neural Proc. Lett.* (1995).

Fredenslund, A., R. Jone, and J. Prausnitz, "Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures," *AIChE J.*, **21**(6), 1086 (1975).

Gakh, A., E. Gakh, B. Sumpter, and D. Noid, "Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds," *J. Chem. Inf. Comput. Sci.*, **34**, 832 (1994).

Gasteiger, J., X. Li, C. Rudolph, J. Sadowski, and J. Zupan, "Representation of Molecular Electrostatic Potentials by Topological Feature Maps," *J. Am. Chem. Soc.*, **116**, 4608 (1994a).

Gasteiger, J., X. Li, and A. Uschold, "The Beauty of Molecular Surfaces as Revealed by Self-Organizing Neural Networks," *J. Mol. Graphics*, **12**, 90 (1994b).

Geman, S., E. Bienenstock, and R. Doursal, "Neural Networks and the Bias/Variance Dilemma," *Neural Comput.*, **4**, 1 (1992).

Georgiopoulos, M., A. Koufakou, G. C. Anagnostopoulos, and T. Kasparis, "Overtraining in Fuzzy ARTMAP: Myth or Reality?" *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks* (IJCNN '01), **2**, 1186 (2001).

Giralt, F., A. Arenas, J. Ferre-Gine, R. Rallo, and G. A. Kopp, "The Simulation and Interpretation of Turbulence with a Cognitive Neural System," *Physics of Fluids*, **12**(7), 1826 (2000).

Gutfreund, H., and M. Mézard, "Processing of Temporal Sequences in Neural Networks," *Phys. Rev. Lett.*, **61**, 235 (1988).

Hall, L., and C. Story, "Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR Atom Type Electrotopological State Indices Using Artificial Neural Networks," *J. Chem. Inf. Comput. Sci.*, **36**, 1004 (1996).

Hansen, C. M., "The Three Dimensional Solubility Parameter—Key to Paint Component Affinities. II. Dyes, Emulsifiers, Mutual Solubility and Compatibility, and Pigments," *J. Paint Technol.*, **39**, 509 (1979).

Hecht-Nielsen, R., "Replicator Neural Networks for Universal Optimal Source Coding," *Science*, **269**, 186 (1995).

Hertz, J., A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Addison–Wesley, Reading, MA (1991).

Hornik, K., M. Stinchcombe, and H. White, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, **2**, 359 (1989).

Karelson, M., V. Lobanov, and A. Katritzky, "Quantum-Chemical Descriptors in QSAR/QSPR Studies," *Chem. Rev.*, **96**, 1027 (1996).

Kaski, S., and T. Kohonen, "Winner-Take-All Networks for Physiological Models of Competitive Learning," *Neural Networks*, **7**, 973 (1994).

Kaski, S., and K. Lagus, "Comparing Self-Organizing Maps," *Proc. of ICANN'96*, 809 (1996).

Kier, L., and L. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York (1976).

Kier, L., and L. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Wiley, New York (1985).

Kier, L. B., and L. H. Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, Orlando, FL (1999).

Kohonen, T., "Self-Organizing Formation of Topologically Correct Feature Maps," *Biol. Cybern.*, **43**, 59 (1982).

Kohonen, T., "The Self-Organizing Map," *Proc. IEEE*, **78**, 1464 (1990).

Koufakou, A., M. Georgiopoulos, G. Anagnostopoulos, and T. Kasparis, "Cross-Validation in Fuzzy ARTMAP for Large Databases," *Neural Networks*, **14**, 1279 (2001).

Lazaridis, T., and M. Paulaitis, "Activity Coefficients in Dilute Aqueous Solutions from Free Energy Simulations," *AIChE J.*, **39**(6), 1051 (1993).

Mackay, D., W. Y. Shiu, and K. Ch. Ma, Illustrated Handbook of Physical Chemical Properties and Environmental Fate for Organic Chemicals, Vol. 1, Lewis Publishers, Chelsea, MI (1992)

Mackay, D., and Y. Shiu, "Aqueous Solubility of Polynuclear Aromatic Hydrocarbons," *J. Chem. Eng. Data.*, **22**, 399 (1977).

McWeeny, R., *Methods of Molecular Quantum Mechanics*, Academic Press, New York (1989).

Medir, M., and F. Giralt, "Correlation of Activity Coefficients of Hydrocarbons in Water at Infinite Dilution with Molecular Parameters," *AIChE J.*, **28**(2), 341 (1982).

Mitchell, B., and P. Jurs, "Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure," *J. Chem. Inf. Comput. Sci.*, **38**, 200 (1998).

Molecular Modeling Pro™, Revision 3.1, ChemSM™ Inc. (1998).

Peterson, D. L., and S. H. Yalkowsky, "Comparison of Two Methods for Predicting Aqueous Solubility," *J. Chem. Inf. Comput. Sci.*, **41**, 1531 (2001).

Rallo, R., A. Arenas, J. Ferre-Gine, and F. Giralt, "A Neural Virtual Sensor for the Inferential Prediction of Product Quality from Process Variables," *Comput. Chem. Eng.*, **26**(12), 1735 (2002a).

Rallo, R., J. Ferre-Gine, A. Arenas, and F. Giralt, "Forecasting Product Quality in Industrial Processes with Virtual Sensors," *Proceedings of the Topical Conference on Sensor Technology*, C. C. Liu, H. B. Martin, and J. C. Angus, eds., AIChE Annual Meeting, AIChE Pub. No. P-172, 127 (2002b).

Ran, Y., N. Jain, and S. H. Yalkowsky, "Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE)," *J. Chem. Inf. Comput. Sci.*, **41**, 1208 (2001).

Ran, Y., and S. H. Yalkowsky, "Prediction of Drug Solubility by the General Solubility Equation (GSE)," *J. Chem. Inf. Comput. Sci.*, **41**, 354 (2001).

Randic, M., "On Characterization of Molecular Branching," *J. Am. Chem. Soc.*, **97**, 6609 (1975).

Randic, M., and N. Trinajstic, "Comparative Structure–Property Studies: The Connectivity Basis," *J. Mol. Struct.*, **284**, 209 (1993).

Rani, Y. K., and N. V. K. Dutt, "Estimation of Activity Coefficients at Infinite Dilution of Halocarbons in Water and Organic Compounds in Hydrofluoroparaffins Using Neural Networks," *Chem. Eng. Commun.*, **189**(3), 372 (2002).

Sandler, S. I., "Quantum Mechanics: The New Engineering Thermodynamics," *Proceedings of the AspenWorld 2002*, Washington, DC, Oct.–Nov. 2002 (https://www.aspenworld2002.com/).

Sherman, S. R., D. B. Trampe, D. M. Bush, M. Schiller, C. A. Eckert, A. J. Dallas, J. Li, and P. W. Carr, "Compilation and Correlation of Limiting Activity Coefficients of Nonelectrolytes in Water," *Ind. Eng. Chem. Res.*, **35**, 1044 (1996).

Simamora, P., A. Miller, and S. Yalkowsky, "Melting Point and Normal Boiling Point Correlations: Applications to Rigid Aromatic Compounds," *J. Chem. Inf. Comput. Sci.*, **33**, 437 (1993).

Stanton, D., L. Egolf, P. Jurs, and M. Hicks, "Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans and Thiophenes," *J. Chem. Inf. Comput. Sci.*, **31**, 301 (1991).

Stanton, D., L. Egolf, P. Jurs, and M. Hicks, "Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles," *J. Chem. Inf. Comput. Sci.*, **32**, 306 (1992).

Stanton, D., and P. Jurs, "Development and Use of Charged Partial Surface Area Structural Descriptors for Quantitative Structural-Property Relationship Studies," *Anal. Chem.*, **62**, 2323 (1990).

Tochigi, K., D. Tiegs, J. Gmehling, and K. Kojima, "Determination of New ASOG Parameters," *J. Chem. Eng. Jpn.*, **23**, 453 (1990).

Vesanto, J., "SOM-Based Data Visualization Methods," *Intell. Data Anal.*, 6, 111 (1999).

Viswanadhan, W., G. Mueller, S. Basak, and J. Weinstein, "Comparison of a Neural Net-Based QSAR Algorithm (PCANN) with Hologram and Multiple Linear Regression-Based QSAR Approaches: Application to 1-4-Dihydropyridine-Based Calcium Channel Antogonist," *J. Chem. Inf. Comput. Sci.*, **41**, 505 (2001).

Voutsas, E. C., and D. P. Tassios, "Prediction of Infinite-Dilution Activity Coefficients in Binary Mixtures with UNIFAC. A Critical Evaluation," *Ind. Eng. Chem. Res.*, **35**, 1438 (1996).

Voutsas, E. C., and D. P. Tassios, "Analysis of the UNIFAC-Type Group Contribution Models at the Highly Dilute region. 1. Limitations of the Combinatorial and Residual Expressions," *Ind. Eng. Chem. Res.*, **36**, 4965 (1997).

Wiener, H., "Structural Determination of Paraffin Boiling Points," *J. Am. Chem. Soc.*, **69**, 17 (1947).

Yaffe, D., Y. Cohen, G. Espinosa, A. Arenas, and F. Giralt, "A Fuzzy ARTMAP Based Quantitative Structure-Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds," *J. Chem. Inf. Comput. Sci.*, **41**(5), 1150 (2001).

Yaffe, D., Y. Cohen, G. Espinosa, A. Arenas, and F. Giralt, "A Fuzzy ARTMAP Based Quantitative Structure-Property Relationships (QSPRs) for the Henry's Law Constant of Organic Compounds," *J. Chem. Inf. Comput. Sci.*, **43**(1), 85 (2003).

Yalkwosky, S., and S. Valvani, "Solubilities and Partitioning. 2. Relationship between Aqueous Solubilities, Partition Coefficients, and Molecular Surface Areas of Rigid Aromatic Hydrocarbons," *J. Chem. Eng. Data*, **24**(2), 127 (1979).

Yalkowsky, S. H., *Solubility and Solubilization in Aqueous Media*, American Chemical Society, Oxford, UK (1999).

Yalkowsky, S. H., and Y. He, *Handbook of Aqueous Solubility Data*, CRC Press, Boca Raton, FL (2003).

Yang, G., Y. Ran, and S. H. Yalkowsky, "Prediction of the Aqueous Solubility: Comparison of the General Solubility Equation and the Method Using an Amended Solvation Energy Relationship," *J. Pharm. Sci.*, **91**, 517 (2002).

Manuscript received Jan. 2, 2002, revision received July 7, 2003, and final revision received Sept. 13, 2003.