

IMPROVED QUANTITATIVE STRUCTURE PROPERTY RELATIONSHIP MODELS FOR INFINITE-DILUTION ACTIVITY COEFFICIENTS OF AQUEOUS SYSTEMS

Brian J. Neely
Srini S. Godavarthy
Robert L. Robinson Jr.
Khaled A.M. Gasem

School of Chemical Engineering
Oklahoma State University
423 Engineering North
Stillwater, OK 74078

Proceedings of the Sixth International Petroleum Environmental Conference,
Albuquerque, NM, November 11, 2004

Phase equilibrium data are essential for the proper design and operation of most chemical processes. When experimental data are unavailable, thermodynamic models, such as group contribution methods, are used to predict phase equilibrium. The accuracy of these models in predicting infinite-dilution activity coefficients (γ^∞) of aqueous systems is questionable. Moreover, model development is hampered by a lack of (a) γ^∞ data at temperatures above 300 K, and (b) γ^∞ data for water in hydrocarbon systems. Using quantitative structure-property relationships (QSPR), mathematical models are developed relating the structure of a diverse set of organic molecules to γ^∞ values for hydrocarbons in water and water in hydrocarbons.

The database used for this study contains over 1400 data points at temperatures ranging from 283.2 to 373.2 K. The data include both direct and indirect measurements for a variety of hydrocarbons, which include alkanes, alkenes, aromatics, halogenates, alcohols, phenols, aldehydes, ketones, acids, esters, ethers, amines, amides, nitriles, nitro compounds, and sulfur compounds. QSPR models were developed using linear as well as non-linear modeling tools, and results indicate these models are satisfactory in correlating single temperature aqueous solubility data, but fail when correlating multiple temperature data.

A suitable theoretical backbone, which could account for the effect of temperature on solubility, was required. Bader and Gasem (1993), previously at OSU, had developed an equation of state (EOS) to correlate γ^∞ of aqueous systems, however, the parameters used in this EOS could not be generalized satisfactorily. Structure based generalizations were developed for these parameters using existing QSPR tools, and preliminary results indicate this combined approach of an EOS to account for temperature effects and structure based parameter generalizations provide accurate estimates for γ^∞ .

INTRODUCTION

Water, which is the most common industrial solvent, plays an important role in many different areas including separation processes, distillation units, chromatographic systems, waste treatment, and environmental concerns [1-7]. With growing application of biotechnologies, there also exists an increased need for phase equilibria of aqueous systems in those processes [8].

Due to the unique molecular structure of water and the attendant physical characteristics, including hydrogen bonding, systems containing hydrocarbons and water often exhibit strong nonideality when compared to systems comprised only of hydrocarbons. The activity coefficient, γ , is a parameter that quantifies the amount of nonideality present in a system. When a component of a hydrocarbon-water binary system is sufficiently dilute, the infinite-dilution activity coefficient, γ^∞ , is reflective of only intermolecular solute-solvent interactions without the additional complication of solute-solute interactions. Insight into the chemical and physical forces present in an aqueous system is provided by these coefficients.

The solubility of any solute in a given solvent may be described in terms of the activity coefficients (γ_i) at a given temperature and pressure. For a given temperature and pressure, the mole fraction of a solute (x_i) can be expressed as follows, when the hydrocarbon is at low concentration:

$$x_i = \frac{(p - p_j^\circ)}{p_i^\circ \gamma_i^\infty - p_j^\circ} \quad (1)$$

where p is system pressure, p° is the pure vapor pressure and γ_i^∞ is the infinite-dilution activity coefficient. The subscripts i and j indicate the solute and the solvent, respectively. In deriving this relation, we assume low-pressure operations, where ideal-gas behavior applies to the vapor phase.

While several experimental methods exist for the investigation of infinite-dilution activity coefficients, these methods often suffer serious limitations [9-11] and are time consuming. Models for the prediction or calculation of infinite-dilution activity coefficients would be useful and are represented by examples from theoretical regular solution theory models [12-18], theoretical equation of state models [19], pure component models [20-22], group contribution models [23-25], empirical models [26-31], the LSER model [32, 33], and computational chemistry models [34-39]. These models generally do not provide satisfactory predictions, and early QSPR studies were limited by the involvement of only single temperature data of one component of the aqueous systems.

The molecular structure of a chemical determines the chemical and physical properties of a particular chemical compound. Continuing investigation has centered on elucidation of the relationship between physical properties and molecular structure. As

computational capability has improved, research revolved around developing free energy relationships by molecular mutation using Monte Carlo (MC) simulators [40]. Although this approach remains attractive, Monte Carlo is being replaced in many applications by QSPR models. The QSPR approach often provides predictions for chemical and physical properties of as-yet-unmeasured or unknown compounds based on structure information. High quality predictions are obtained using these descriptors since structure-property mapping is at an atomic level rather than at a functional group level. QSPR models will be influential in enabling advances in chemical design, where a key challenge is the development of tools permitting the rapid creation of unique molecules. Over the last ten years, QSPR have played an increasingly important role in drug screening and discovery [41] and application is appearing in areas outside the pharmaceutical industry. While standard methodologies for chemical design result in a discovery phase of research and development from two to three years, QSPR methodologies are estimated to result in a reduction of this phase to three to six months.

The objectives of this work are to (a) develop a quantitative structure property relationship (QSPR) for prediction of γ_i^∞ values of hydrocarbon-water systems, (b) evaluate the efficacy of QSPR models using multiple linear regression analyses and back propagation neural networks, (c) evaluate the ability of the model to predict aqueous and hydrocarbon solubility at multiple temperatures.

DATABASE

The database, which is culled from 96 journal literature sources dating from 1927 to 1995, consists of 1400 infinite-dilution activity coefficients (IDAC) at temperatures ranging from 283.15 K to 373.15 K from a diverse set of structural classifications [42]. Data available consist of both hydrocarbon in water and water in hydrocarbon IDAC's, and the hydrocarbon in water data is further sorted with reference to experimental origin.

The two origin classifications are direct measurements and indirect measurements. Examples of direct measurements are gas-liquid chromatography method (GLC), headspace GLC method, gas-stripping method, liquid-liquid chromatography method, differential ebulliometry method, and differential static method. Included under the general title of GLC methods are stationary phase GLC, non-steady-state GLC, and relative GLC. The indirect measurements include extrapolations from vapor-liquid equilibrium data, and calculations from other thermodynamic data such as liquid-liquid equilibrium data and gas-liquid partition coefficient data.

Where provided by the source material, the database also contains error estimates. These error estimates were then used in the case studies. Table 1 and 2 provide a numerical analysis of the database and a list of the different hydrocarbon structures found in the database.

The database was used to develop six case studies for this study. The first three case studies, DIRECT, INDIRECT, and WATER, consisted of all available data, including error points, for each of the three sections of the database; direct, indirect, and

water in hydrocarbon, respectively. The fourth case study, INDEX1, used all data, but did not use error points due to a software limitation. The fifth and sixth case studies, INDEX2 and WATERIND respectively, involved only matched hydrocarbon in water and water in hydrocarbon data. For example a measurement of hexane in water, whether from the direct or indirect set, must have a corresponding measurement of water in hexane for inclusion in both INDEX2 and WATERIND. A summary of the case studies is available in Table 3. Regardless of the type of measurement, namely the water in hydrocarbon data, the hydrocarbon structure was used exclusively for the study with the exception of WATERIND. In this particular case study, the structure of water was used to represent the water in hydrocarbon data. After the initial step in QSPR model development, the DIRECT and INDIRECT case studies were combined to form a case study, HC, comprised of all hydrocarbon in water data.

The molecular structures found in the database were prepared in the following manner:

1. Molecular structures were drawn and optimised using the MMX molecular mechanics force field module available in ChemDraw Ultra [43].
2. 2D structures were generated using ChemDraw Ultra.
3. Chem3D Pro [44] was employed to generate 3D molecular structures from exported 2D structures.
4. These structures were initially optimized using the MOPAC [45] module available in Chem3DUltra.
5. The pre-optimized structures were submitted to the AMPAC 6.0 [46] program for further geometry refinement and for the calculation of molecular orbital parameters. The AM1 parameterizations were used to calculate the quantum-chemical molecular descriptors.
6. Output from AMPAC was used in CODESSA [47] to calculate various molecular descriptors.

In addition to a small number of constructed descriptors, over 1400 descriptors from such categories as constitutional, topographical, geometric, electrostatic, quantum chemical, and thermodynamic [48] were generated for each molecular structure and are briefly described as follows:

1. Constitutional Descriptors: These simple descriptors reflect only the molecular composition of the compound without using the geometric or electronic structure of the molecule e.g., number of atoms, number of bonds, number of rings, and molecular weight.
2. Topological Descriptors: These descriptors provide the atomic connectivity in the molecule, which include molecular connectivity indices, substructure counts, molecular weights, weighted paths, molecular distance edge descriptors, kappa indices, electro topological state indices, and many other graph invariants [49, 50].
3. Geometric Descriptors: These descriptors are calculated to encode the 3D aspects of the structures and include such descriptors as moments of

inertia, solvent-accessible surface area, length-to-breadth ratios, shadow areas, and gravitational index [51, 52].

4. **Electrostatic Descriptors:** These descriptors are calculated to encode aspects of the structures that are electron related, which include partial atomic charges, HOMO energies, LUMO energies, and dipole moment.
5. **Quantum Chemical Descriptors:** These descriptors represent quantum-chemically calculated charge distribution in the molecules and, therefore, describe the polar interactions between molecules and their chemical reactivity. The descriptors also provide the value of the partial charge on the atoms in the molecule (e.g., dH_{\min} represents the minimum partial charge on a hydrogen atom). Additionally, these descriptors relate to the strength of intramolecular interactions and characterize the stability of the molecules, their conformational flexibility, and other valency-related properties, such as the maximum bond order (P_{AB}) for a given pair of atomic species A and B in the molecule [53].
6. **Thermodynamic Descriptors:** These descriptors are calculated on the basis of the total partition function (Q) of the molecule and its electronic, translational, rotational, and vibrational components.
7. **Constructed Descriptors:** The descriptors generated by CODESSA do not provide the best modeling approach because functional group descriptors are entirely neglected. However, it has been shown that functional groups play an important role in estimating properties [54, 55]. Forty functional group descriptors were constructed for each molecule. This data set then was analyzed to develop a single descriptor, which was representative of the whole data set. The concept of group contributions is based on the premise that each functional group in the molecule provides either a positive or negative increment to the molecular properties. Specifically, addition of functional groups is likely to alter the properties by increasing the polarizability and possibly the dipole moment of the molecule; thus, these functional groups redistribute electrons, increase or decrease internal strains, and also change the molecular symmetry and rotational entropy [54-56]. This approach initially identifies the molecules that are best represented by the model, and then finds the descriptors that provide the most reduction in the squared error for the outliers.

QSPR MODEL DEVELOPMENT

Development of a QSPR model for each case study consists of a strategy to reduce the number of descriptors and three different analyses.

The Type I analysis employs CODESSA to generate a linear model by means of multiple linear regression. During this first analysis, the descriptor set is reduced by elimination of non-orthogonal descriptors. The result of an analysis specifying 25 parameters was employed to determine outliers in the data. If there was a deviation greater than 2σ , a datum was determined to be an outlier and was eliminated from the case studies. An example of this is shown for INDEX1 in Figure 1. The analysis was

repeated to generate a final set of approximately 200 descriptors, which are the most significant. CODESSA was then used with the corrected data and the final descriptor set at specifications of 14, 12, 10, 8, 6, and 4 descriptors in order to generate R^2 plots for the determination of the optimum combination of R^2 value and number of descriptors. Results are presented in Table 4.

Prior to commencement of the Type II analyses, each case study was randomly divided into a training set, prediction set, and cross validation set composed of 70, 20 and 10%, respectively, of the total number of data in each case study. The different sets will be employed to test the viability of the *a priori* predictive capability of the models. Type II analysis involves the addition of linear and nonlinear descriptors, descriptor reduction using a genetic algorithm, and linear analysis with CODESSA. The added descriptors included melting point, boiling point, octanol-water partition, functional group parameter based on molecular structure, and various mathematical manipulations of such descriptors as the molecular weight, gravitational index, and molecular volume.

Using the descriptor set from the Type I analysis and the additional descriptors, a genetic algorithm in NeuralPower [57] is employed to reduce the descriptor set to 50 descriptors in a stepwise fashion where the set is reduced by approximately 25% each time over the course of five iterations of the genetic algorithm. With the new descriptor set CODESSA was then used at various specifications of descriptors in order to generate R^2 plots for the determination of the optimum combination of R^2 value and number of descriptors. The result for the Type II analysis is shown in Table 5 and a plot for INDEX1 is provided in Figure 1.

The twenty most significant descriptors from the Type II analyses are used as a descriptor set for the Type III analyses, which are non linear models using neural networks. A back propagation neural network was used in NeuralPower. The initial weights for the network, type of transfer function, and network architecture were determined through trial and error.

Once a transfer function and architecture were selected, ten replicate analyses using randomized initial weights were performed. During these analyses, the root mean square error (RMSE) of the training set and cross validation set was compared as training cycles accumulated. An increase in the RMSE of the cross validation set is indicative of over training or a loss of general predictive capability of the neural network. A contour plot can then be constructed using the cross validation RMSE, which is utilized in determining the region of least RMSE. The identified region will be the replicate analysis used for the Type III model. When a replicate in a contour plot contains an extended “valley” of relatively unchanging RMSE values, such as shown by replicate number five in Figure 3, training is halted at a point in which the lowest RMSE value is obtained while using the fewest possible number of training cycles, which reduces the computational burden. Using Figure 3, two possible selections are replicate five and ten at 6000 and 2000 training cycles, respectively. Calculation of the %AAD for the training, prediction, and cross validation sets of replicate five results in 5.7, 13.5, and 13.0, respectively and for replicate ten results in 6.0, 12.8, and 11.8, respectively. As

shown by the improvement in the training set %AAD, these numbers illustrate that improvement in the training set correlation with additional training comes at the expense of a decrease in the predictive capability shown by the increase in %AAD for the prediction and cross validation sets. In this case for INDEX1, replicate number ten with 2000 training cycles would be selected. After selection of a particular replicate and number of training cycles, the results are obtained for the Type III models. The result for the Type III analysis is presented in Table 5 and a plot for INDEX1 is provided in Figure 4.

RESULTS AND DISCUSSION

A summary of the QSPR modeling results for INDEX1 is presented in Table 5 (for other case study results see [58]). In general the non-linear model performs better than the linear model in predicting the infinite-dilution activity coefficient. The back-propagation network was used for generating the non-linear model. The resulting descriptor set obtained for the Type III analysis provides insight into the relationship between structural molecular features and physical properties of an organic molecule. Among the final descriptor set were two constructed descriptors, a functional group parameter and a mathematical manipulation of the gravitational index. Hybrid models, which involve the use of descriptors obtained from linear methods to develop non-linear models, are increasingly being employed due to the decrease in the amount of computational time required when using only non-linear methods.

The new Type III model developed for predicting γ^∞ of hydrocarbon-water systems provided satisfactory prediction of γ^∞ data (6.0 %AAD and R^2 of 0.988). The descriptors currently given in the literature and used in software packages do not adequately describe the molecular structure relationship with γ^∞ , but the addition of constructed descriptors improved the model predictions. However, predictions at extended temperatures are still poor. Since the majority of the data in the database are collected at ambient temperature, the lack of extended temperature data results in network training skewed to ambient temperature data.

A possible solution is the provision of a theoretical backbone to the model, which accounts for temperature dependence in the data. Preliminary study has been given to the development of an improved QSPR model, which is based on the Bader-Gasem equation of state [19]. Due to the need of inclusion of extended temperature data, application of this model has been limited to a small subset of the main database, but initial results of 3.5 %AAD show marked improvement and are encouraging. Details of the other case studies and further work concerning theoretical backbones for temperature dependency are available from the Thermodynamics Group of the School of Chemical Engineering at Oklahoma State University [58].

References

- [1] G. Malmary, A. Vezier, A. Robert, J. Mourgues, T. Conte, and J. Molinier, "Recovery of tartaric and malic acids from dilute aqueous effluents by solvent extraction technique," *Journal of Chemical Technology and Biotechnology*, vol. 60, pp. 67-71, 1994.
- [2] L. D. Skrylev, A. G. Nevinskii, and A. N. Purich, "Flotation extraction concentration of dilute aqueous solutions of g-hexachlorocyclohexane," *Zhurnal Prikladnoi Khimii (Sankt-Peterburg, Russian Federation)*, vol. 57, pp. 2026-30, 1984.
- [3] C. L. Yaws, H. C. Yang, J. R. Hopper, and K. C. Hansen, "Organic chemicals: water solubility data," *Chemical Engineering (New York, NY, United States)*, vol. 97, pp. 115-16, 118, 1990.
- [4] C. L. Yaws, H. C. Yang, J. R. Hopper, and K. C. Hansen, "Hydrocarbons: water solubility data," *Chemical Engineering (New York, NY, United States)*, vol. 97, pp. 177-8, 180, 182, 1990.
- [5] T. Lazaridis and M. E. Paulaitis, "Activity coefficients in dilute aqueous solutions from free energy simulations," *AIChE Journal*, vol. 39, pp. 1051-60, 1993.
- [6] J. Li, A. J. Dallas, D. I. Eikens, P. W. Carr, D. L. Bergmann, M. J. Hait, and C. A. Eckert, "Measurement of large infinite dilution activity coefficients of nonelectrolytes in water by inert gas stripping and gas chromatography," *Analytical Chemistry*, vol. 65, pp. 3212-18, 1993.
- [7] D. L. Bergmann and C. A. Eckert, "Measurement of limiting activity coefficients for aqueous systems by differential ebulliometry," *Fluid Phase Equilibria*, vol. 63, pp. 141-50, 1991.
- [8] P. A. Belter, E. L. Cussler, and W. Hue, *Bioseparations*. New York: John Wiley & Sons, Inc., 1988.
- [9] M. S. H. Bader, "Vapor-Liquid Equilibrium Properties of Aqueous and Supercritical fluids at Infinite Dilution." Stillwater, OK: Oklahoma State University, 1993.
- [10] M. S. H. Bader and K. A. M. Gasem, "Determination of infinite dilution activity coefficients for organic-aqueous systems using a dilute vapor-liquid equilibrium method," *Chemical Engineering Communications*, vol. 140, pp. 41-72, 1996.
- [11] D. Tiegs, J. Gmehling, A. Medina, M. Soares, J. Bastos, P. Alessi, and I. Kikic, "Activity Coefficients at Infinite Dilution," in *Chemistry Data Series*, vol. IX, Part 1, D. Behrens and R. Eckermann, Eds. Frankfurt, Germany: Schoon & Wetzels GmbH, 1986.
- [12] G. Scatchard, "Equilibria in nonelectrolyte solutions in relation to the vapor pressures and densities of the components," *Chemical Reviews (Washington, DC, United States)*, vol. 8, pp. 321-33, 1931.
- [13] J. H. Hildebrand and S. E. Wood, "Derivation of equations for regular solutions," *Journal of Chemical Physics*, vol. 1, pp. 817-22, 1933.
- [14] R. F. Weimer and J. M. Prausnitz, "Screen extraction solvents this way," *Hydrocarbon Processing and Petroleum Refiner*, vol. 44, pp. 237-42, 1965.

- [15] C. M. Hansen, "Three-dimensional solubility parameter-key to paint component affinities. II. Dyes, emulsifiers, mutual solubility and compatibility, and pigments," *Journal of Paint Technology*, vol. 39, pp. 505-10, 1967.
- [16] J. G. Helpinstill and M. Van Winkle, "Prediction of infinite-dilution activity-coefficients for polar-polar binary systems," *Industrial & Engineering Chemistry Process Design and Development*, vol. 7, pp. 213-20, 1968.
- [17] B. Karger, L. R. Snyder, and C. Eon, "An expanded solubility parameter treatment for classification and use of chromatographic solvents and adsorbents. Parameters for dispersion, dipole and hydrogen bonding interactions," *Journal of Chromatography*, vol. 125, pp. 71-88, 1976.
- [18] R. Tijssen, H. A. H. Billiet, and P. J. Schoenmakers, "Use of the solubility parameter for predicting selectivity and retention in chromatography," *Journal of Chromatography*, vol. 122, pp. 185-203, 1976.
- [19] M. S. H. Bader and K. A. M. Gasem, "Modeling infinite dilution activity coefficients of organic-aqueous systems using a modified regular solution equation and cubic equations-of-state," *Canadian Journal of Chemical Engineering*, vol. 76, pp. 94-103, 1998.
- [20] E. R. Thomas and C. A. Eckert, "Prediction of limiting activity coefficients by a modified separation of cohesive energy density model and UNIFAC," *Industrial & Engineering Chemistry Process Design and Development*, vol. 23, pp. 194-209, 1984.
- [21] W. J. Howell, A. M. Karachewski, K. M. Stephenson, C. A. Eckert, J. H. Park, P. W. Carr, and S. C. Rutan, "An improved MOSCED equation for the prediction and application of infinite dilution activity coefficients," *Fluid Phase Equilibria*, vol. 52, pp. 151-60, 1989.
- [22] M. J. Hait, C. L. Liotta, C. A. Eckert, D. L. Bergmann, A. M. Karachewski, A. J. Dallas, D. I. Eikens, J. J. Li, and P. W. Carr, "Space predictor for infinite dilution activity coefficients," *Industrial & Engineering Chemistry Research*, vol. 32, pp. 2905-14, 1993.
- [23] E. L. Derr and C. H. Deal, "Analytical Solutions of Groups Correlation of Activity Coefficients Through Structural Group Parameters," *Institute of Chemical Engineering Symposium Series*, vol. 32, pp. 40, 1969.
- [24] A. Fredenslund, R. L. Jones, and J. M. Prausnitz, "Group-contribution estimation of activity coefficients in nonideal liquid mixtures," *AIChE Journal*, vol. 21, pp. 1086-99, 1975.
- [25] J. Gmehling, J. Li, and M. Schiller, "A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties," *Industrial & Engineering Chemistry Research*, vol. 32, pp. 178-93, 1993.
- [26] G. J. Pierotti, C. H. Deal, and E. L. Derr, "Activity coefficients and molecular structure," *Journal of Industrial and Engineering Chemistry (Washington, D. C.)*, vol. 51, pp. 95-102, 1959.
- [27] C. Tsionopoulos and M. Prausnitz, "Activity Coefficients of Aromatic Solutes in Dilute Aqueous Solutions," *Industrial & Engineering Chemistry Fundamentals*, vol. 10, pp. 593-600, 1971.

- [28] M. Medir and F. Giralt, "Correlation of activity coefficients of hydrocarbons in water at infinite dilution with molecular parameters," *AIChE Journal*, vol. 28, pp. 341-3, 1982.
- [29] N. V. K. Dutt and D. H. L. Prasad, "Estimation of infinite dilution activity coefficients of hydrocarbons in water from molar refraction," *Fluid Phase Equilibria*, vol. 45, pp. 1-5, 1989.
- [30] S. H. Yalkowsky and S. C. Valvani, "Solubilities and partitioning. 2. Relationships between aqueous solubilities, partition coefficients, and molecular surface areas of rigid aromatic hydrocarbons," *Journal of Chemical and Engineering Data*, vol. 24, pp. 127-9, 1979.
- [31] D. Mackay and W. Y. Shiu, "Aqueous solubility of polynuclear aromatic hydrocarbons," *Journal of Chemical and Engineering Data*, vol. 22, pp. 399-402, 1977.
- [32] R. W. Taft, J. L. M. Abboud, M. J. Kamlet, and M. H. Abraham, "Linear solvation energy relations," *Journal of Solution Chemistry*, vol. 14, pp. 153-86, 1985.
- [33] S. R. Sherman, 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," *Industrial & Engineering Chemistry Research*, vol. 35, pp. 1044-58, 1996.
- [34] K. S. Shing, "Infinite-dilution activity coefficients from computer simulation," *Chemical Physics Letters*, vol. 119, pp. 149-51, 1985.
- [35] I. G. Economou, "Monte Carlo Simulation of Phase Equilibria of Aqueous Systems," *Fluid Phase Equilibria*, vol. 183-184, pp. 259-269, 2001.
- [36] T. M. Nelson and P. C. Jurs, "Prediction of Aqueous Solubility of Organic Compounds," *Journal of Chemical Information and Computer Sciences*, vol. 34, pp. 601-9, 1994.
- [37] B. E. Mitchell and P. C. Jurs, "Prediction of infinite dilution activity coefficients of organic compounds in aqueous solution from molecular structure," *Journal of Chemical Information and Computer Sciences*, vol. 38, pp. 200-209, 1998.
- [38] J. He and C. Zhong, "A QSPR study of infinite dilution activity coefficients of organic compounds in aqueous solutions," *Fluid Phase Equilibria*, vol. 205, pp. 303-316, 2003.
- [39] P. D. T. Huibers and A. R. Katritzky, "Correlation of the Aqueous solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure," *Journal of Chemical Information and Computer Science*, vol. 38, pp. 283-292, 1998.
- [40] M. G. Martin, J. I. Siepmann, and M. R. Schure, "Exploring multi-component phase equilibria by Monte Carlo simulations: Towards a description of gas-liquid chromatography," *Unified Chromatography, ACS Symposium Series*, vol. 748, 2000.
- [41] H. Kubinyi, "QSAR and 3D QSAR in Drug Design. 2. Applications and Problems," *Drug Discov. Today*, vol. 2, pp. 538-546, 1997.
- [42] K. Kojima, S. Zhang, and T. Hiaki, "Measuring Methods of Infinite-Dilution Activity Coefficients and a Database for Systems Including Water," *Fluid Phase Equilibria*, vol. 131, pp. 145-179, 1997.

- [43] "Chem3D Ultra," 6.0 ed. Cambridge, MA: CambridgeSoft.com, 2000.
- [44] "Chem3D Pro," 6.0 ed. Cambridge, MA: CambridgeSoft.com, 2000.
- [45] J. P. P. Stewart, "MOPAC Program Package," 1990.
- [46] "AMPAC," 6.7 ed. Shawnee Mission, KS: Semichem, 2000.
- [47] A. R. Katritzky, M. Karelson, V. S. Lobanov, R. Dennington, and T. Keith, "CODESSA," 2.63 ed. Shawnee, KS: Semichem, Inc., 1999.
- [48] A. R. Katritzky, V. S. Lobanov, and M. Karelson, "CODESSA User's Manual." University of Florida, Gainesville, 1994.
- [49] L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*. New York: Research Studies Press, 1986.
- [50] A. T. Balaban, "From Chemical Topology to 3D Geometry," *J. Chem. Inf. Comput. Sci.*, vol. 37, pp. 645-650, 1997.
- [51] M. Karelson, *Molecular Descriptors in QSAR/QSPR*. New York: John Wiley & Sons, 2000.
- [52] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2000.
- [53] M. Karelson, V. S. Lobanov, and A. R. Katritzky, "Quantum-Chemical Descriptors in QSAR/QSPR Studies," *Chem. Rev. (Washington, D.C.)*, vol. 96, pp. 1027-1043, 1996.
- [54] K. G. Joback and R. C. Reid, "Estimation of Pure-Component Properties from Group-Contributions," *Chem. Eng. Communications*, vol. 57, pp. 233, 1987.
- [55] L. Constantinou and R. Gani, "New Group Contribution Method for Estimating Properties of Pure Compounds," *AIChE Journal*, vol. 40, pp. 1697, 1994.
- [56] M. Karelson, U. Maran, Y. Wang, and A. R. Katritzky, "QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of Codessa Applications. Collect.," *Czech. Chem. Commun.*, vol. 64, pp. 1551-1571, 1999.
- [57] "NeuralPower," 2.5 ed: CPC-X Software, 2003.
- [58] B. J. Neely, "Improved Quantitative Structure Property Relationship Models for Infinite-Dilution Activity Coefficients of Aqueous Systems," in *School of Chemical Engineering*. Stillwater: Oklahoma State University, 2004.

Table 1: Numerical Analysis of the Database

Type	Data	Error Data
Hydrocarbon in Water		
Direct Measurement	776	438
Indirect Measurement	388	0
Water in Hydrocarbon	236	66

Table 2: Database Hydrocarbon Structures

Hydrocarbon in Water		Water in Hydrocarbon
Direct	Indirect	
Alkanes	Aliphatic Alkanes	Aliphatic Alkanes
Alkenes	Cyclic Alkanes	Aromatic Hydrocarbons
Aromatic Hydrocarbons	Aliphatic Alkenes	Halogenated Hydrocarbons
Halogenated Hydrocarbons	Cyclic Alkenes	Alcohols
Alcohols	Alkynes	Ketones
Phenol and Derivatives	Monocyclic Aromatics	Acids
Aldehydes	Polycyclic Aromatics	Aldehydes
Ketones	Halogenated Hydrocarbons	Ethers
Acids	Alcohols	Esters
Esters	Phenol Derivatives	Compounds with Nitrogen
Ethers	Ketones	
Amines and Amides	Acids	
Nitriles	Esters	
Nitro Compounds	Ethers	
Compounds with Sulfur	Aldehydes	
	Amines and Amides	
	Nitro Compounds	
	Compounds with Sulfur	

Table 3: Summary of Case Studies

Case Study	Number of Values	Case Study	Number of Values
DIRECT		WATER	
Data Values	776	Data Values	236
Error Values	438	Error Values	66
Total	1214	Total	302
INDIRECT		INDEX2	
Data Values	388	DIRECT	410
Total	388	INDIRECT	30
INDEX1		WATERIND	
DIRECT	776	DIRECT	410
INDIRECT	388	INDIRECT	30
WATER	236	WATER	154
Total	1400	Total	594

Table 4: Summary of Type I Results

Results	Case Study					
	DIRECT	INDIRECT	WATER	INDEX1	INDEX2	WATERIND
R ² with all Data at 25 Parameters	0.9802	0.9607	0.9800	0.9358	0.9048	0.9064
Numbers of Outliers	64	21	17	72	40	46
% of Outliers	5.3	5.4	5.6	5.1	6.7	7.7
R ² with Corrected Data at 25 Parameters	0.9879	0.9767	0.9880	0.9686	0.9555	0.9507
Number of Descriptors in Final Set	35	40	33	39	46	47
R ² at 14 Parameters	0.9785	0.9617	0.9753	0.9485	0.9505	0.9364
R ² at 12 Parameters	0.9757	0.9533	0.9703	0.9370	0.9481	0.9323
R ² at 10 Parameters	0.9708	0.9465	0.9543	0.9336	0.9419	0.9230
R ² at 8 Parameters	0.9619	0.9243	0.9359	0.9197	0.9323	0.9057
R ² at 6 Parameters	0.9434	0.8997	0.8962	0.8979	0.9100	0.8724
R ² at 4 Parameters	0.8880	0.8622	0.8271	0.7871	0.8216	0.8048

Table 5: Summary of QSPR Model Results

Results	INDEX1		
	Type I	Type II	Type III
Numbers of Descriptors	10	3	20
R ²	0.937	0.956	0.988
%AAD	--	10.0	6.0

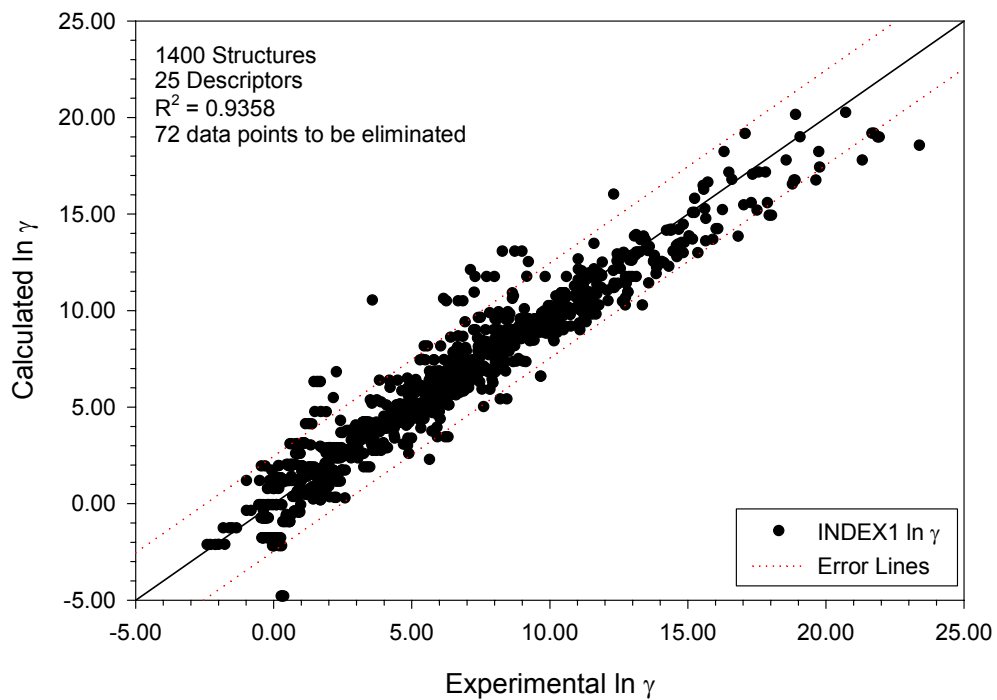


Figure 1: Infinite-Dilution Activity Coefficients of INDEX1 Case Study (Type I)

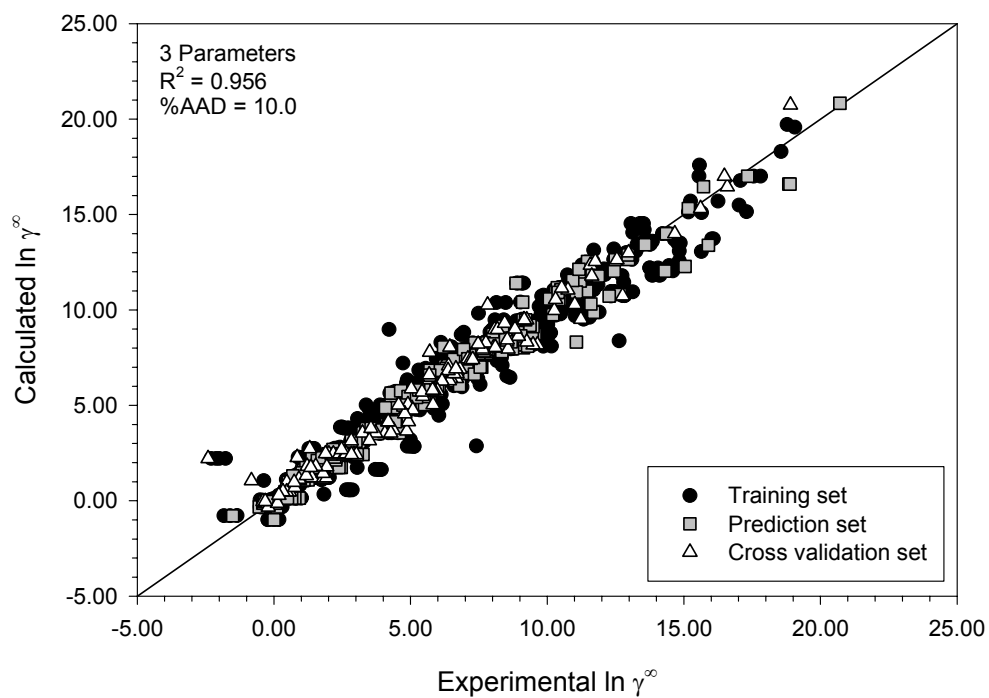


Figure 2: Infinite-Dilution Activity Coefficients of INDEX1 Case Study (Type II)

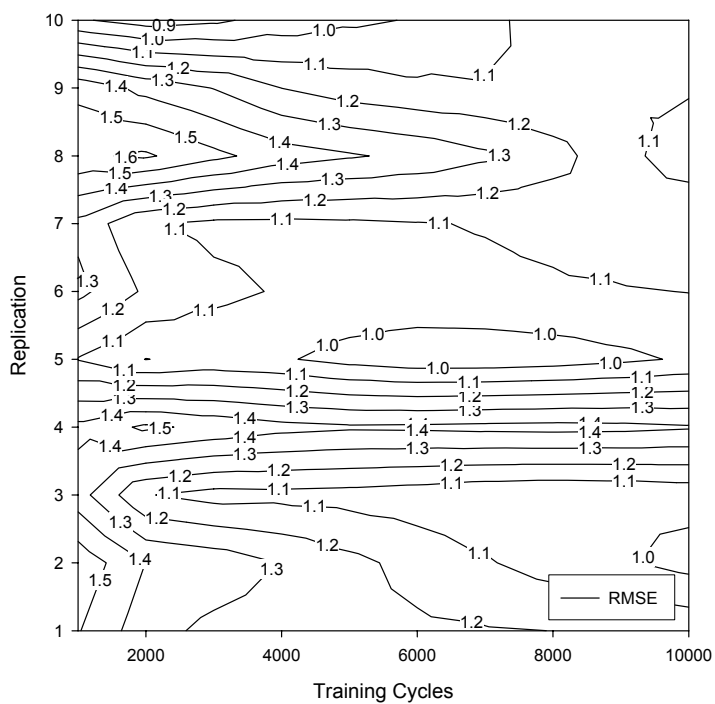


Figure 3: Contour plot for INDEX1 (Type III)

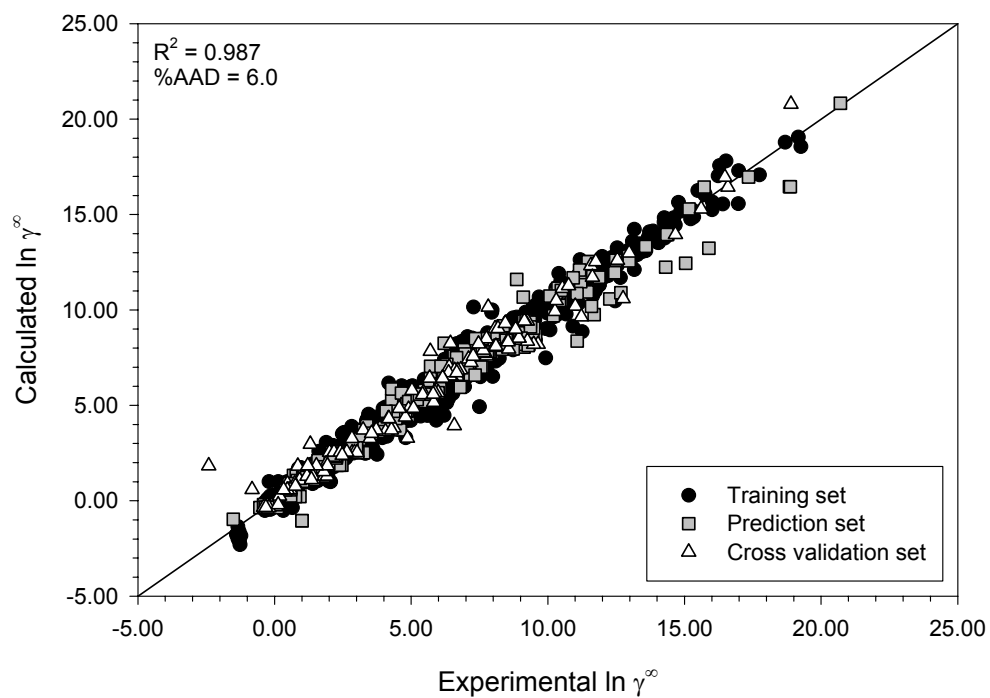


Figure 4: Infinite-Dilution Activity Coefficients of INDEX1 Case Study (Type III)