2006

The Modeling of Solubility

Frank Christopher Campanell
Wright State University

The Modeling of Solubility

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

By

Frank Campanell
B.S., Wright State University, 2006

2006
Wright State University

WRIGHT STATE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

Sept 26, 2006

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY FRANK C. CAMPANELL ENTITLED "THE MODELING OF SOLUBILITY" BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

_____

Paul G. Seybold, Ph.D.
Thesis Director

_____

Kenneth Turnbull, Ph.D.
Department Chair

Committee on
Final Examination

_____

Paul G. Seybold, Ph.D.

_____

Rubin Battino, Ph.D.

_____

David A. Dolson, Ph.D.

_____

Joseph F. Thomas, Jr., Ph.D.
Dean, School of Graduate Studies

ABSTRACT

Campanell, Frank C. M.S., Department of Chemistry, Wright State University, 2006. The Modeling of Solubility.

In this work the solubilities of gases in liquids and liquids in liquids were modeled using both physical properties and topological descriptors of the solutes. Quantitative structure-activity relationship (QSAR) methods were employed to create single-linear regression (SLR) and multiple-linear regression (MLR) models of the solubilities. Factor analysis was employed to determine the number of significant factors present in the solubilities. The solubilities of monoalcohols in water, halogenated alkanes in water, gases in water, gases in alkanes, and gases in alcohols were examined and modeled.

Table of Contents

1. Review Solubility, QSAR Descriptors, and the Modeling of Solubility……..…..1

    1.1 Overview…….……..1

    1.2 The Three-Step Solvation Process……2

    1.3 Molecular Modeling of Solubility…….…...3

    1.4 Features used in the Modeling of Solubility…….……9

        1.4.1 Experimental Properties……..……...10

        1.4.2 Calculated Properties……..……...11

        1.4.3 Topological Descriptors……..……...13

    1.5 References……..…….17

2. The Aqueous Solubilities of Monoalcohohols……..……...20

    2.1. Abstract……..……...20

    2.2. Introduction……..…….21

    2.3. Methods……..……..22

    2.4. Results and Discussion……..……...23

        2.4.1. Using Calculated Physical Properties ……..…….23

        2.4.2. Using Topological Descriptors ……..……...27

        2.4.3. Calculated Physical Properties and Topological Descriptors together..…31

    2.5 Appendix..…….33

    2.6 References..……...40

3. Modeling the Aqueous Solubilities of Halogenated Alkanes..……42

    3.1. Abstract..……...42

3.2. Introduction..…….43

3.3. Methods..……..44

3.4. Results and Discussion..……...45

3.4.1. Chloroalkanes..……..45

3.4.1.1. Calculated Physical Properties...……46

3.4.1.2. Calculated Topological Descriptors..…….47

3.4.1.3. Best Overall Model..……..48

3.4.2. Bromoalkanes..……..50

3.4.2.1. Calculated Physical Properties.....……..51

3.4.2.2. Calculated Topological Descriptors..…….52

3.4.2.3. Best Overall Model..……..53

3.4.3. Iodoalkanes..……..53

3.4.4. All Halogenated Alkanes..…….54

3.4.4.1. Calculated Physical Properties..……56

3.4.4.2. Calculated Topological Descriptors and Best Overall Model..…….57

3.5. Conclusions..……...59

3.5. Appendices…..…….60

3.5.1. Appendix A…..……..60

3.5.2. Appendix B…..……..64

3.5.3. Appendix C..……..66

3.5.4. Appendix D..……..67

3.6. References..……...69

4. The Solubilities of Gases in Water..……70

4.1. Abstract..……...70

4.2. Introduction...……71

4.3. Methods…..……..72

4.4. Results and Discussion..……...72

4.4.1. All Gases in Water…..……...73

4.4.2. Taking Subsets of Gases..……..74

4.4.3. Groups of Gases that are not well modeled ..……....75

4.5. Appendix..……77

4.6. References..……...81

5. The Solubilities of Gases in Alkanes..…….82

5.1. Abstract..……...82

5.2. Introduction..…….83

5.3. Methods..……..83

5.4. Results..…….84

5.5. Discussion..……...88

5.6. Appendix..…….90

5.7. References..……...92

6. The Solubilities of Gases in Alcohols..……93

6.1. Abstract..……...93

6.2. Introduction..…….94

6.3. Methods……94

6.4. Results……..…….95

6.5. Discussion……...……..…..100

6.6. Appendix…….103

6.7. References……...105

7. Factor Analysis of Gas Solubilities……106

7.1. Abstract……...106

7.2. Introduction……107

7.3. Methods……..108

7.4. Results and Discussion……...109

7.4.1. Results for the Noble Gases……109

7.4.2. Results for all 13 Gases……113

7.5. Appendix……118

7.6. References……...120

List of Figures

1. 2-Methylbutane…….13

2. 2-methyl-1-propanol…….13

3. Solubility modeled by Volume…….33

4. Solubility modeled by Surface Area…….33

5. Solubility modeled by Polarizability……...…….34

6. Alcohol #30……..……34

7. Molecule #62……35

8. Solubility modeled by Volume, with the top two outliers removed……..…….35

9. Solubility modeled by Volume and Polarizability…….36

10. Solubility modeled by ABSQ……...……..36

11. Solubility modeled by X_1……37

12. Solubility modeled by ABSQ and Xo…….37

13. Solubility modeled by ABSQ with the top two outliers removed…..……38

14. Solubility modeled by Vol and Sur…….60

15. Solubility modeled by k1 and Q……..…60

16. Solubility modeled by 4sC, sCl, and k1……..61

17. Solubility modeld by 4sC, sCl, and Pol, with the top outlier removed…….…..61

18. Solubility modeled by Vol with Tetrabromomethane removed…….……63

19. Solubility modeled by x1 with tetrabromomethane removed……...……..63

20. Solubility modeled by x1, xvc3, and Q……...…64

21. Solubility modeled by Volume……...…66

22. Solubility modeled by 4sC, sF, sCl, and ka1 with top outlier removed…….67

23. Solubility modeled by 4sC, sF, sCl, ka1 and sBr with top three outliers removed…67

24. Solubility modeled by DM…….……77

25. Solubility modeled by BP and Pol…….77

26. Solubility modeledby BP and DM……..78

27. Noble Gas Solubility modeled by BP…….78

28. Solubility of Simple Gases modeled by BP and Pol……...……79

29. Hexane Solubility modeled by Boiling Point…….90

30. Hexane Solubility modeled by Polarizability…….90

31. The Solubility of Methanol modeled by BP……...…..103

32. The Solubility of Methanol modeled by Pol…….103

33. Factor plot for analysis 1: He, Ne, Ar, and Kr……..111

34. Factor Plot for Analysis 1: all Noble Gases but Rn……..……117

35. Factor Plot for Analysis 2: All the Noble Gases……...……117

List of Tables

1. Important Molecular Descriptors for this Work……..23

2. Table of the Correlation matrix of calculated Physical Properties and Solubility for alcohols in water…….……...24

3. Summary of R^2 for the Best 1-term, 2-term, and 3-term Regressions using Topological Descriptors…….…….28

4. Correlation Matrix for the Calculated Descriptors and Solubility and the Solubility…….…28

5. All Monoalcohols and their Calculated Physical Properties…….……39

6. Descriptors Used and Their Abbreviations……...……44

7. Correlation Matrix for the Chloroalkanes…….45

8. Regression with similar results to Eq. 3.6…….49

9. Correlation Matrix for the Bromoalkanes…….50

10. Summary of the Best Models for the Iodoalkanes……..53

11. Factor Analysis for All Halogenated Alkanes……...….55

12. Calculated Physical Property Models for all Halogenated Alkanes……...…57

13. Aqueous Solubilities, Calculated Physical Descriptor values, and Topological Descriptor values for the Chloroalkanes……..62

14. Aqueous Solubilities, Calculated Physical Descriptor Values, and Topological Descriptor Values for Bromoalkanes……...65

15. Table of Solubilities and Volume Values for Iodoalkanes…….66

16. All Halogenated Alkanes……68

17. Physical Properties of the Gases Examined……72

18. Correlation Matrix of BP, Pol, and DM……..73

19. Other Groupings Attempted……76

20. All Relevant Values for Gases in Water…….....80

21. Physical Properties Used in this Study (Gases in Alkanes)……...….84

22. Correlation Matrix of Physical Property and Factor Analysis……...…….84

23. Physical Properties of the Gases, and ln Mole Fraction (ln $X_{2)}$ experimental solubilities……...…...91

24. Physical Properties of the Gases, and ln X_2 experimental solubilities……..…104

25. Correlation Matrix of ln(mole fraction) Solubilities of Noble Gases……...109

26. Solvents used in the Second Factor Analysis……...…113

27. A Representative Correlation Matrix……..…….114

28. Solubilities of the Noble gases, expressed as ln(molefraction) in different Solvents at 298……118

Chapter 1

Review of Solubility, QSAR Descriptors, and the Modeling of Solubility

1.1. Overview

The phenomenon of solubility has been studied since the inception of the science of chemistry as we know it today. The solubility of a substance is defined as the concentration of dissolved solute in equilibrium with undissolved solute at a specified temperature and pressure[1]. Hence solubility occurs in any two-component system in which equilibrium can be reached. This definition holds for a variety of phases, such as gases in liquids, liquids in liquids, solids in liquids, gases in solids, and solids in solids. The study of solubility is relevant for a wide variety of processes, including oxygen transport in the blood, drug interactions, environmental pollution, and industrial processes. [1]

[1] References for each chapter are gathered at the end of that chapter.

1.2. The Three-Step Solvation Process

Solubility is traditionally thought of as a three-step energetic process.[2] The first step involves the use of energy to create a cavity in the solvent. The cavity is defined as a void of excluded volume in the bulk solvent prepared in order to accommodate the solute molecule.[3] This energy of cavity formation involves the separation and reorganization of their bulk solvent in preparation for the solute. The energy required should therefore be proportional to the volume or surface area of the solute. The second step involves separating solute molecules from the bulk solute aggregates. This second step can be ignored for the solubility of gases in liquids, since in a gas solute-solute interactions are minimal. In other cases the energy required for solute separation will be related to the bonding forces involved in the solute-solute interactions. The third step releases energy when the solute is inserted into the solvent cavity and interacts via attractive forces with the solvent. The energy released is normally due mainly to dispersion forces and any hydrogen bonding forces active between the solute and the solvent.

An elementary examination of thermodynamics is useful as a reference for the energy changes involved in solvation. The ideal solution provides a useful limit for considering the process of solvation; in this

$$\Delta V_{m,mix} = 0 \text{ and } \Delta H_{m,mix} = 0 \qquad (1.1)$$

This indicates that there are no volume changes or heat effects in the mixing process. This leads to[1]:

$$-R \ln X_2 = \Delta H_{vap}(1/T_b - 1/T) \qquad (1.2)$$

where R is the gas constant, ΔH_{vap} is the enthalpy change on vaporization of the pure solute at its normal boiling point T_b, T is the temperature of the measurement, and X_2 is the mole fraction solubility.

From statistical thermodynamics the entropy on forming an ideal solution is:[1]

$$\Delta S = k \ln [W(\text{mixed}) / W(\text{unmixed})] \quad (1.3)$$

where k is Boltzmann's constant, and W is the number of different microscopic ways in which the configuration of the system can be achieved. For an ideal liquid the interactive forces between all the molecules of the mixture are identical and the solute and solvent molecules completely mix. This allows the molar entropy of mixing for the ideal solution is:

$$\Delta_{mix}S_m = -X_1 R \ln X_1 - X_2 R \ln X_2 \quad (1.4)$$

where X_i is the mole fraction of liquid *i* in the mixture, and $\Delta_{mix}S_m$ is always positive.

In a real liquid mixture equation 1.3 still holds true, but it is now impossible to calculate W due to the complexity of the system. To find a workable approach consider the criterion for spontaneous change for a real mixture:

$$dG_{T,p} \leq 0 \quad (1.5)$$

$$\mu_i = \mu_i^* + RT \ln X_i Y_i \quad (1.6)$$

The term μ_i is the chemical potential of *i* in the liquid mixture, μ_i^* is the chemical potential of pure liquid *i* at the temperature and pressure of the mixture, and Y_i is the activity coefficient defined so that $Y_i \rightarrow 1$ as $X_i \rightarrow 1$.

$$\Delta_{mix}G_m = \Sigma X_i (\mu_i - \mu_i^*) \quad (1.7)$$

$$\Delta_{mix}G_m = \Sigma X_i RT \ln X_i + \Sigma X_i RT \ln Y_i \quad (1.8)$$

The last term is called the excess Gibbs energy of mixing.

1.3. Molecular Modeling of Solubility

The general aim of molecular modeling is to find a relationship between some property P and features of the molecules considered, P = f(S). In the present case P =

solubility. Solubility is an experimentally determined quantity. In some cases solubility can be accurately modeled by analyzing the factors affecting the solvation process. Ideally such a model should accurately calculate the solubility, have explanatory value, and be based on reasonable structural features.

The goal of modeling solubility is to produce an accurate and easily interpreted relationship between solute properties and solubility. Such a model might utilize experimental physical properties, calculated physical properties, or structural descriptors of the solute or solvent. For the present work only features of the solute are used. The use of calculated properties or topological descriptors lead to an approach called "QSAR" or "QSPR" modeling. QSAR stands for "quantitative structure-activity relationship", and QSPR is a similar term that stands for "quantitative structure-property relationship". The term QSAR is more common due to its use in the pharmaceutical industry. Hansch and Fujita[4] are credited with coining the term QSAR. Their early work involved octanol/water partition coefficients and was based on the Hammet approach. This approach uses easily calculated properties or descriptors to correlate solubility with the solute's structural features. It is important to note that the features employed in making QSARs are not usually derived from quantum mechanical calculations, and hence are quicker to calculate and easily applied to the types and ranges of molecules of interest.

The first modern attempt at a structure-property relationship came in the 1940's from a pre-med student named Harry Wiener at The City College, NY.[5,6] The "Wiener Index" used mathematical theory to successfully estimate the boiling points and other properties of a set of alkanes by considering only their topology. This work was largely ignored until the 1970's when Randić[7] developed a "branching index" technique from mathematical graph theory. Both of these studies used only the two-dimensional

connectivity of the atoms of the molecules. More recently many studies have been done using this general idea and applying it to solubility. Abraham et. al.[8] have developed a number of linear solvation energy relationships studies (LSER) of solutes in water, alcohols, and alkanes.

QSAR relationships contain a wealth of information about the molecules for which they are generated. Descriptors for shape, branching, charges, and volume are all easily estimated. QSAR studies can be used for a variety of purposes, such as giving insight into the factors that control chemical phenomena, predicting physical or chemical properties, and checking experimental data.

The models used in the present work take the form of multiple linear regressions (MLR) or single linear regressions (SLR). These types of models are easily generated and interpreted. In MLR or SLR a dependent variable y (here solubility) is modeled in terms of independent variable(s) (X_1, X_2,….X_o) to yield a model of the form:[9]

$$y = A_1x_1 + A_2x_2 + \ldots + A_nx_n + C \qquad (1.9)$$

where A_1, A_2, … A_n are coefficients of the descriptors, x_1, the x_2, …., x_n are the structural descriptors, and C is a constant. The success of such an equation in an application can be ascertained from the statistics generated by the model. Many of these statistics assume a *null hypothesis*, which states that any observed relationships are merely random occurrences.[10] The null hypothesis must be rejected on statistical grounds for the statistical correlation to have validity. The statistics used in this work will mainly consist of the coefficient of determination (R^2), the standard error (s), the t-statistic or t-test (t), the robustness of fit statistic or leave-one-out statistic (q^2), and the Fisher statistic (F).

The coefficient of determination R^2 is the square of the Pearson correlation coefficient (r). Mathematically r is a measure of the strength of the linear relationship between the dependant variable y and the independent variable(s).[4]

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} \tag{1.10}$$

where

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Here n is the sample size. The properties of r include: $0 \le r \le 1$, r and the slope of the least squares line have the same sign, and a low value of r implies that little or no linear relationship exists between y and x, whereas when r is close to 1, or to –1, a strong the relationship exists between y and x. The coefficient of determination is the square of this value (r), hence R^2 is always a positive value, with 0 indicating absolutely no correlation, and 1 indicating a correlation that has exhausted all the variation in the dependant variable.[12] R^2 represents the fraction of the variance in the data explained by the model.

R^2 = explained variance in the data / total variance in the data. (1.11)

For example, an R^2 value of 0.92 implies that 92% of the variance in the dependent variable is explained by the independent variable(s) x_i.[13] The R^2 value will always increase as more independent variables are added, ending when $R^2 = 1$ and the number of independent variables equals the total number of data points. It is possible to take an

adjusted coefficient of determination that takes into account the number of independent variables:[14]

$$\text{adjusted } R^2 = R^2*((n-1)/v)) \tag{1.12}$$

Where n is the number of response values and v = n – number of fitted independent variables. The correlation coefficient is the most telling statistic for the quality of a model, but R^2 by itself gives an incomplete picture. The parsimony of the model, or the number of independent variables used to model the property, is not covered well by R^2. Unger and Hansch[15] have suggested the *principle of parsimony* be used to judge if a model is acceptable. Their criterion for an acceptable model is that one should have at least six data points per independent variable.

The standard error (s) is the square root of the residual sum-of squares. This is a measure of the difference between the predicted and actual values of the dependent variable. There are two forms of s; the overall standard error, and the standard error associated with each single independent variable in a MLR. The overall standard error follows the equation:[9]

$$s = \sqrt{\frac{\sum_{k=1}^{N}(Y_k - \overline{Y}_k)^2}{N - p - 1}} \tag{1.13}$$

N is the sample size, $\overline{Y}_k$ is the mean of the dependant variables, Y_k is a given dependant variable, and p is the number of independent variables. This statistic gives a good criterion for finding outliers. An outlier is often defined as any data point whose calculated value falls more than two or three standard deviations from the observed value, depending on what type of work one is doing. The s statistic is also useful for

interpreting how good the model is at high R^2 values, since s is a much more sensitive measure in this region. The independent variable standard error is calculated much the same way[16]:

$$s = \sqrt{\frac{A_{kk} * \sum_{k=1}^{N} (Yk - \overline{Y}k)^2}{N - p - 1}} \tag{1.14}$$

where A is a matrix of and A_{kk} is the K^{th} diagonal element of matrix A, and the other terms are as defined in Eq. 1.13. This individual standard error is useful in determining if that particular independent variable adds any statistical significance to the model. The standard error can range from 0 to ∞, with a lower s value being more desirable.

The t-test is an offshoot of the latter standard error. A t-test value exists for each individual independent variable. The t-test value is defined as:

$$t = \text{coefficient value for } x_i \text{ / uncertainty in the coefficient} \tag{1.15}$$

It is obvious that higher absolute values for the t-test signify a more statistically significant variable. It is generally accepted that any t-test value above 2 has some statistical significance. For the present work, any t-test value below 5 will be considered statistically unsound, and all t-test values will be taken as absolute values. Each t-test value is directly associated with two other statistics, the Pearson p-value and the Fisher Statistic (q.v).

The robustness of fit, or leave-one-out statistic, (q^2), provides a way of cross-validating the model. Each experimental point in the data set is removed in turn and a new regression equation is obtained. The new regression equation is then used to create a

calculated data point for the removed experimental datum. This newly generated set of calculated data points is then used to form q^2. The formula for q^2 is:[17]

$$q^2 = 1 - \frac{\sum (y_i - \bar{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{1.16}$$

This method of statistical analyses is useful to ensure that the model is not heavily dependant on any single data point. The q^2 value can fall between R^2 and 0, with values close to R^2 value indicating more robust fit.

The Fisher statistic (F) gauges the likelihood that a given correlation between the dependent and independent variable(s) occurs by chance. The F statistic follows the equation:

$$F = \frac{\sum_{k=1}^{N} (y_k - \bar{y})^2 (N - p - 1)}{\sum_{K=1}^{N} (y_o - y_k)^2 (p)} \tag{1.17}$$

where y_o is the observed value of the dependent variable, y_k is the predicted value of the dependent variable, N is the sample size, and p is the number of independent variables. The F statistic can range from 0 to ∞, with a higher value being more desirable. Like the s value, the F value can be directly related to the Pearson p-value.

Together these five statistics provide a suitably comprehensive picture of the statistical significance of the model.

1.4. Features used in the Modeling of Solubility

First it is important to define some of the terms that will be used in this work. A *gas* is technically defined as a substance that is above its critical temperature at a given pressure. Fogg and Gerrard[18] have used the looser working definition that a gas has a

normal boiling point below 298.15 K. For the purposes of this work when a gas is referred to, it will meet the Fogg and Gerrard criterion. Similarly, the definition of a liquid will be a substance with a boiling above 298.15 K. There are three basic categories of properties that can be used to model solubility: experimentally determined properties, calculated molecular properties, and topological descriptors. Ideally a model should not mix descriptors or properties from different categories. This may sometimes be ignored, however, if the properties or descriptors form an acceptable model that still intuitively relates to theory.

1.4.1. Experimental Properties

Experimentally determined properties can sometimes be used to accurately model solubility. The inherent difficulty is that they must be experimentally determined, whereas one reason for modeling is to determine property values without performing experiments.

The experimentally determined properties used in this work, with their general applications to the study of solubility, include the following:

Boiling Point (BP)[19] – Defined as the temperature at which the vapor pressure of a liquid becomes equal to the applied pressure at any pressure. The *normal boiling point* is used in the present work, and can apply to a vapor pressure of either one bar or one atmosphere. Because historically most measurements were done in atmospheres, the latter definition is used in this work. The boiling point of a solute is often closely related its molecular size and mass, since dispersion forces tend to increase with size and mass. In certain cases other types of forces, such as hydrogen bonding, may be important.

Melting Point (MP)[20] – Defined as the temperature at which the solid and liquid phases of a substance are in equilibrium at a given pressure. This can be especially useful for solid solutes. The *normal melting point* is taken at 1 ATM for this work.

Critical Volume (V_c)[21] – Defined as the molar volume at the critical pressure P_c and critical temperature T_c. $V_c^{2/3}$ provides an estimate of the surface area of a molecule (in this cause the solute). Estimates of the volume and surface area of a molecule are useful since they may be related to the energy need to create the cavity in the solvent, and, also, to the interaction energy with the solvent.

Polarizability (Pol)[22] – Defined as the ease of distortion of the electronic cloud of a molecule by an electric field. It is the ratio of the induced dipole (μ_{ind}) moment to the field (E) that generated it.

$$\alpha = \mu_{ind} / E \qquad (1.18)$$

The dispersion forces generated by a molecule are in general proportional to the molecule's polarizability. The polarizability is usually proportional to the number of electrons in a molecule. The polarizability data used in this work comes from the CRC.[23] Pol's units are $Å^3$.

Dipole Moment (DM) – A dipole consists of two charges q and –q separated by distance r. The dipole moment is a vector product and has a vector value, but its absolute value is normally used. The dipole moment can be a useful indicator of either hydrogen bonding acceptor ability or hydrogen bond donor ability. DM has units of Debyes.

1.4.2. Calculated Properties

Calculated expressions of physical properties are often used in lieu of the actual physical properties. These calculated properties have the advantage of avoiding experimental work, but their accuracy needs examinations. Calculated physical properties may have a range for which they produce acceptable values. For example, the

additive method of polarizability presented below produces statistically well-behaved data for any molecule with at least 2 non-hydrogen atoms. Any smaller molecules, such as H_2, yield inaccurate calculated polarizabilities using this formula.

Radius of Gyration (RG)[24] - A parameter characterizing the size of a molecule of any shape. The molecule is assumed to be a rigid structure.

$$RG = \left(\frac{\sum_i m_i r_i^{\,2}}{\sum_i m_i} \right)^{1/2} \tag{1.19}$$

where atoms of a mass m_i are located a distance r_i from the center of mass. (There is another definition that uses the moment of inertia instead of the total mass of the atoms in the denominator). This is an often-used parameter in the study of polymers, and it should be noted that the method of calculating RG in this work is different from that used in the case of polymers. The radius of gyration is indicative of the size of a solute molecule, and commonly has units of Å.

Inherent Properties of the Molecule - There are many properties of a molecule that can be used to create an accurate MLR and that can be extracted with basic chemical knowledge. These include; the formula weight (fw), the total number of electrons in the molecule (N_E), the number of lone pairs of electrons (N_{LP}), the total number of valence electrons in the molecule (N_{VE}), the total number of electrons in a molecule without including electrons in a bond (N_{VB}), and N_{VB} plus the electrons involved in bonds to hydrogen (N_{VH}). There are other inherent properties of use in the study of solubility, and they will be discussed as needed. There are no units, as the data simply enumerates the occurrence of the property.

Calculated Polarizability (Polc) - There are several ways to obtain a *calculated* polarizability. The semiempirical AM1, PM3, and MNDO quantum chemical methods

all include methods for calculating polarizabilities. The additive method used in this work is taken from the work of Bosque and Sales[25]. This additive method is useful because it is said to be a better approximation for polarizabilities than most other additive methods. The Bosque and Sales equation is:

$$\alpha/\ \text{Å}^3 = 0.32 + 1.51\ N_C + 0.57 N_O + 0.17\ N_H + 2.99\ N_S + 3.29 N_{Br} + 1.03\ N_N + 0.22\ N_F + 2.16\ N_{Cl} + 2.48\ N_P \quad (1.20)$$

Here N_C is the number of carbon atoms, N_O the number of oxygen atoms, etc. For example, for methane:

(CH_4): $\alpha = 0.32 + 1.51*1 + 0.17*4 = 2.51\ \text{Å}^3$

QsarIS[26] generates another form of calculated polarizability sometimes used in this work.

Parachor[27] - This is a simple additive property that has been used historically for either individual atoms, or group contributions to estimate the surface tension of a liquid.

Surface Area, Volume, and Dipole Moment - The program, used in the studies to be reported, has methods to estimate each of these properties. The use of calculated properties will be noted.

1.4.3. Topological Descriptors

Often the general composition of a molecule, in terms of the number and types of atoms and how they are bonded to each other, can be used to create descriptors that have the ability to model solubility. Topological descriptors use the topology of a molecule to create a unitless descriptor. QsarIS[26] was used to create the majority of the topological descriptors used in this work. A calculational method is employed to generate these values. First a hydrogen-suppressed graph of the molecule created. Examples are shown in Figures 1.1 and 1.2.

Figure 1.1.

2-Methylbutane

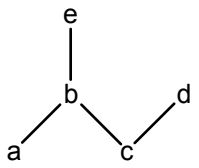
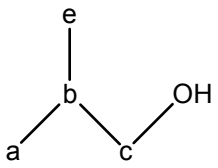


Figure 1.2.

2-methyl-1-propanol



Molecular Connectivity Indicies[26] - The molecular connectivity (chi) indices provide quantitive characterization of the skeletal variation in a molecule. The molecular graph is evaluated as a collection of fragments (subgraphs) of different sizes and complexity. A chi index is a summation over a count of a given type of subgraph weighted by a function of the "weighted" delta values (δ = the number of non-hydrogen atoms connected to the atom) of the atoms that comprise each subgraph. In simple connectivity, each non-hydrogen atom has the same basic delta value. In valence connectivity, each type of atom is given its own delta value. The following equations define the simple and valence chi indices along with the simple and valence subgraph terms:

$$^{m}x_{t} = \sum_{k} {}^{m}C_{k} \text{ (for simple chi index)} \qquad (1.21)$$

$$^{m}C_{k} = \Pi_{k}(\delta_{i})^{1/2} \text{ (simple subgraph term)}^{2} \qquad (1.22)$$

$$^{m}x_{t}^{v} = \sum_{k} {}_{m}C_{k}^{v} \text{ (for valence chi index)} \qquad (1.23)$$

The resulting index is a summation over all subgraph terms in the molecular graph. Each subgraph term is computed for one subgraph of (m) connected edges between 2 or more vertices, which are defined above. For the zero order chi index (chi zero) or (chi valence

[2] e.g. $^{1}\chi = \sum \frac{1}{\sqrt{\delta_i \delta_j}}$

zero) the subgraph has one vertex and zero edges so the subgraph is an individual atom. Simple connectivity treats all non-hydrogen atoms the same, and valence connectivity treats all heteroatoms with different constants. For example, the first order connectivity ($^1\chi$) for Figures 1.1 and 1.2 is calculated as

$$^1\chi = \frac{1}{\sqrt{a*b}} + \frac{1}{\sqrt{c*b}} + \frac{1}{\sqrt{d*c}} + \frac{1}{\sqrt{e*b}} = \frac{1}{\sqrt{1*3}} + \frac{1}{\sqrt{2*3}} + \frac{1}{\sqrt{1*2}} + \frac{1}{\sqrt{1*3}} = 2.27 \quad (1.24)$$

Here carbon *a* is connected to one other non-hydrogen atom, and is hence given a value of 1. Carbon *b* is connected to three non-hydrogen atoms, and is hence given a value of 3. The first order valence connectivity ($^1\chi^v$)is the same as the simple first order for 2-methylbutane, but for 2-methyl-1-propanol it is calculated by

$$^1\chi = \frac{1}{\sqrt{a*b}} + \frac{1}{\sqrt{c*b}} + \frac{1}{\sqrt{d*c}} + \frac{1}{\sqrt{e*b}} = \frac{1}{\sqrt{1*3}} + \frac{1}{\sqrt{2*3}} + \frac{1}{\sqrt{4*2}} + \frac{1}{\sqrt{1*3}} = 1.92 \quad (1.25)$$

The value for δ (*d)* is 4 as assigned to alcohol oxygen.

Kier Shape Indicies - The Kappa Shape Indices are a family of graph-based structure descriptors designed with the specific objective of encoding relative shape characteristics into a manifold of values that are computed for each molecule being described. These kappa values form the basis of a method of molecular structure quantification in which multiple attributes of molecular shape are encoded into three indices. Kappa values are derived from counts of one-bond (paths of length 1), two-bond (paths of length 2) and three-bond fragments (paths of length 3). With each count being made relative to corresponding fragment counts from a pair of reference structures with the same number of graph vertices (atoms or hydride groups) as the molecule being described. These reference structures define, for each graph with a given number of vertices, the maximum and minimum values for each encoded shape characteristic. Each

of the three indices encodes separate characteristic shape attributes of the molecular graph. For traditional kappa indices:

$$m_k = C({}^mP_{\max}\ {}^mP_{\min})/({}^mP_i)^2 \qquad (1.26)$$

Where P_{max} is the number of paths of order m in an unbranched reference molecule with the same number of atoms as the molecule being described, P_{min} is the number of paths of order m in a reference molecule with an extreme structure feature and the same number of atoms as the molecule being described, P_i is the number of paths of order m in the actual molecule "i" being described, m is the order of the path, and C is a constant (2 for path orders 1 and 2, and 4 for path order three).

ABSQ - The sum of the absolute values of the charges on the atoms of the molecule. Gasteiger[28] charges are used.

MaxHp, MaxNeg, MaxQp - These are the largest positive charge on a hydrogen atom, the largest negative charge over the atoms in a molecule, and the largest positive charge over the atoms in a molecule, respectively.

1.5. References

[1] Battino, R.; Letcher, T. M., An Introduction to the Understanding of Solubility. *J. Chem. Ed.,* **2001**, *78*, 103-111.

[2] Kamlet, M. J.; Doherty, R. M.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W., Solubility A New Look. *ChemTech,* **1986**, 566-576.

[3] Basilevk, M. V.; Geigoriev, F. V.; Leontyev, I.V.; Sulimov,V.B., Excluded Volume Effect for Large and Small Solutes in Water. *J. Phys. Chem.,* **2005**, *109*, 6939-6946.

[4] Hansch, C.; Fujita, T., A Method for the Correlation of Biological Activity and Chemical Structures. *J. Amer. Chem. Soc.,* **1964**, *86*, 1616-1626.

[5] Wiener, H., Influence of Interatomic Forces on Paraffin Properties. *J. Chem. Phys*., **1947**, *15*, 766-782

[6] Wiener, H., Structural determination of Paraffin Boiling Points. *J. Am. Chem. Soc.*, **1947**, *69*, 2636-2638.

[7] Randić, M., Characterization of Molecular Branching. *J. Am. Chem. Soc.*, **1975**, *97*, 6609-6615.

[8] Abraham, M. H., Linear Solvation Energy Relationships *J. Phys. Chem.,* **1987**, *91*.

[9] Schroeder, L. D.; Sjoquist, D. J.; Stephan, P.E., *Understanding Regression Analysis, An Introductory Guide.* **1987**, Sage, London.

[10] Borowski, E. J.; Bowewin, J.M., *Harper Collins Dictionary of Mathematics*. HarperCollins: New York, **1991**.

[11] Quang, D., Bui Hong Institute of Technology Statistics http://www.netnam.vn/unescocourse/statistics/11_6.htm

[12] Lewis, B., *Applied Regression, An Introduction.* Sage Publicantion: London **1986**.

[13] Achen, C.H., *Interpreting and Using Regression.* Sage Puclication: London **1986**.

[14] Stats @ MTSU http://mtsu32.mtsu.edu:11308/dictionary/r

[15] Unger, S.; Hansch, C., On Model Building in Structure-Activity Relationships. *J. Med. Chem*., **1972**, *16*, 745-749.

[16] QsarIS Version 1.2 SciVision, Burlinqton, MA **2000**.

[17] Golbraikh, A.; Tropsha, A., Beware of q^2!, *J. Mol. Graphics and Modeling*, **2002**, *20*, 269-276.

[18] Fogg, P. G. T.; Gerrard, W., *Solubility of Gases in Liquids*. **1991** John Whiley & Sons Ltd: Baffins Lane, Chichester, England.

[19] The Chemical Institute of Canada
http://home.nas.net/~dbc/cic_hamilton/dictionary/b.html

[20] Web Elements Chemistry definitions
http://www.webelements.com/webelements/properties/text/definitions/melting-point.html

[21] Atkins, P., *The Elements of Physical Chemistry 2nd edition.*, Freeman and Company: New York, **1997**.

[22] IUPAC Compendium of Chemical Terminology
http://www.iupac.org/publications/compendium/P.html

[23] Lide, D. R. (ed.), *CRC handbook of Chemistry and Physics 2002-2003.* CRC Press **2003**.

[24] IUPAC Compendium of Chemical Terminology
http://www.iupac.org/goldbook/R05121.pdf

[25] Bosque, R.; Sales, J., Polarizabilities of Solvents from the Chemical Composition. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 1154-1163.

[26] QsarIS ver. 1.2 copyright **2001** SciVision 2000 Wheeler Road, Burlington, MA. 01803

[27] Poling, J. M.; Prausnitz; J. M; O'Connell, J. P., *The Properties of Gases and Liquids Fifth Edition*. McGraw-Hill Publishing: Washington D.C., **2000**.

[28] Gasteiger, J.; Marsili, M., Iterative Partial Equalization of Orbital Electronnegativity-A Rapid Access to Atomic Charges, *Tetrahedron*, **1980**, *56*, p. 3219-3228

Chapter 2

The Aqueous Solubilities of Monoalcohols.

2.1. Abstract. The experimental solubilities of 66 alcohols in water were modeled using calculated values for the alcohol physical properties as well as sets of 1, 2, and 3 topological descriptors. The physical properties used were volume, surface area, and polarizability. The R^2 values for all three physical property single-term regressions were higher than 0.960. The R^2 value for the best single topological descriptor (ABSQ) was 0.9847, that for the best two-term topological descriptors regression was R^2= 0.9872, and that for the best set of 3-term topological descriptors regression was R^2 = 0.9879.

2.2. Introduction

The solubility of monoalcohols in water is a fairly well-understood phenomenon. The ideal and regular solution approaches use thermodynamic relations, as shown in chapter one of this work, and do not apply to polar solutes. Modifications for polar solutes include: (1) a correction factor for the entropy of mixing to account for the size differences (the Flory-Higgins correction term), and (2) a correction term for the effect of hydrogen bonding. Hermann[1] has shown that the molecular surface area of the solute and a corresponding surface tension or "interfacial tension" term can describe the solubilities of hydrocarbons in water. Amidon, Yalkowsky, and Leung[2] extended this approach to describe the solubilities of aliphatic alcohols in water. Surface areas were estimated using van der Waals radii to account for the size of each individual atom as a sphere. This method was used for both the solutes and the solvent. The effect of the hydroxyl group was included by adding another term, in which the presence or absence of a hydroxyl group was accounted for by using an indicator term (1 if a hydroxyl group was present, 0 if not present). The two-term regression model made from this simple method had an $R^2 = 0.978$. Yalkowsky and Benerjee[3] have reviewed the usefulness of molecular volume in the determination of aqueous solubilities in their book *Aqueous Solubility Methods of Estimation for Organic Compounds*.

Randić's[4] molecular connectivity index has also been used to describe the aqueous solubilities of alcohols. In 1978 Cammarata,[5] created a topological method for modeling the solubilities of aliphatic alcohols based partly on Randic's connectivity index. Cammarata used a set of 51 aliphatic alcohols in water, and was able to obtain $R^2 =$

0.976. In 1980 Kier[6] modeled the solubilities of the same set of 51 aliphatic alcohols in water, using the connectivity index itself, obtaining an $R^2 = 0.956$.

The general trends seen so far suggest that the energy given back by solute-solvent interaction should normally be due mostly to dispersion forces, and hence the size of the solute should be important. Hydrogen bonding might also be important, but in the present case most of the alcohol molecules are large and hydrogen bonds only contribute a small portion of this energy.

2.3. Methods

The alcohol solubilities were obtained from the work of Yalkowsky and Valvani,[7] and are expressed as ln S, where S is the molar solubility. Cyclohexanol, the only cyclic alcohol measured, was excluded from the original list of 67 compounds for this work. The appendix lists the alcohols ordered by compound number. The alcohols are identified as follows: the number after the X is the total number of carbons in the longest chain, and the number before the X is the carbon to which the OH group is attached. A number preceding an M signifies a methyl group on that number carbon. E signifies an ethyl group. For example, 35MM4X7 is 3,5-dimethyl-4-heptanol.

The software used to generate all of the calculated alcohol physical properties, the topological descriptors, and the regression analyses was QsarIS.[8] A list of over 50 different topological descriptors produced by the program was screened using a genetic algorithm. The calculated topological descriptors are all without units. All factor analysis work was done using the StatMost program.[9]

Polarizabilities were generated using the simple additive method employed by QsarIS. Volumes were obtained from QsarIS using a grid method. The most important topological descriptors employed are defined below. It is important to note that the

connectivity indicies χ_0^v and χ_0 are linearly related for the present set of monoalcohols, and therefore only χ_0 has been retained. The same is true for κ2 and κa2, with only κ2 being retained**.** This effect occurs because all the solute molecules are composed of only carbon and hydrogen, with one oxygen atom each.

Table 2.1. Important Molecular Descriptors for this Work.

ABSQ	The sum of absolute values of the charges on each atom of a molecule, in electrons. (Gasteiger charges used)[10]
Dipole	Calculated molecular dipole moment in Debyes.
χ_o	Simple zeroth order connectivity index as defined in Kier and Hall[11]
κ2	Kier Kappa 2 shape index as defined by Kier and Hall[12]
SSCH3*	The Sum of E-State values for all methyl groups

*Note that these descriptors are formed from electrotopological indices.[13]

2.4. Results and Discussion

2.4.1. Using Calculated Physical Properties

Table 2 is a correlation matrix for the calculated physical properties and the solubilities of the monoalcohols. A negative value indicates that the two compared properties are negatively correlated. All correlations are based on a sample of 66 alcohols. It is clear that all the properties are highly correlated. The polarizability, surface area, and volume correlated positively as expected.

Table 2.2. Table of the Correlation matrix of calculated Physical Properties and Solubility

	Solubility	Polarizability	Surface Area	Volume
Solubility	1			
Polarizability	-0.9800	1		
Surface Area	-0.9802	0.9764	1	
Volume	-0.9848	0.9876	0.9974	1

It is also of interest to note that principal component analysis (PCA) of the three calculated physical descriptors and the solubility shows that a single underlying factor explains over 98% of the variance in the set. Figure 2 shows the percent loading results.

Table 2.3. Factor Percentage Loading

	Factor 1	Factor 2	Factor 3
ln X_2	98.43%	0.27%	1.31%
Polarizability	98.37%	0.99%	0.64%
Surface	98.88%	1.10%	0.00%
Volume	99.66%	0.22%	0.08%

The factor analysis indicates that the remaining variance in the solubility data after factor one is not well described by any of the calculated physical properties. The solubility data have 1.31% of their variance described by factor three. None of the properties considered has a significant correlation with factor three. Volume creates the best model because it most accurately describes factor one.

The following single-variable linear regressions were obtained in this study for the 66 compounds. Using the volume (V) as a single parameter:

$$\ln S = 7.68\ (\pm 0.27) - 0.0838(\pm 0.0018)^* V \qquad (2.1)$$

$$n = 66 \quad R^2 = 0.9698 \quad s = 0.669 \quad F = 2058 \quad Q^2 = 0.968$$

Using the surface (SA) area as a single parameter:

$$\ln S = 7.80(\pm 0.31) - 0.0610(\pm 0.0015)*SA \qquad (2.2)$$

$$n = 66 \quad R^2 = 0.9609 \quad s = 0.762 \quad F = 1572 \quad Q^2 = 0.9582$$

Using the polarizability (POLcalc) as a single parameter:

$$\ln S = 8.13(\pm 0.32) - 1.830(\pm 0.047)*POLcalc \qquad (2.3)$$

$$n = 66 \quad R^2 = 0.9604 \quad s = 0.766 \quad F = 1552 \quad Q^2 = 0.9577$$

Figures 2.A1, 2.A2, and 2.A3 in the appendix show plots of the experimental vs calculated results using the volume, surface area, and polarizability, respectively.

All three physical properties model the solubilities extremely well. The polarizability might be expected to be the best of the three descriptors, since it accounts for the size, composition, and energy of interaction of the molecule. In fact, it is marginally the poorest of the three descriptors (by an R^2 value difference of less than 0.01 from the best fit with volume), perhaps due to the relatively crude method of estimation. The polarizability is "clumped" for several groups of alcohols. This clumping is due to the additive method of calculation and the alcohols having the same chemical formula. The surface area and the volume take into account the amount of branching in the alcohol, and thus produce values more individually tailored to each alcohol molecule. All three regressions have R^2 values within 0.01 of each other, making all three equally able to model the solubility. As the factor analysis shows, one property can estimate all three steps in the solvation process. The volume or surface area is proportional to the amount of energy used in creation of the cavity. The energy required for separation of the alcohol solute in step two is proportional to the interaction forces present between the solute alcohols. The energy given back in the third step by solute-solvent interaction is due to dispersion forces and hydrogen bonding. Since there are no "small" alcohols (with

5 carbons or fewer) in the data set hydrogen bond acceptor and donor sites should make only relatively small contributions. The importance of hydrogen bonding should be minimal, as is clearly the case.

The top two outliers for the surface area and volume regression are the same, molecules #30 (22MM1X5) and #62 (2X11), respectively. Both of these outliers deviate by more than 3 s. The top two outliers for the polarizability regression are #64 (1X14) and #62(2X11). Both of these outliers are outside 3 s. Figures 2.A4 and 2.A5 in the appendix show 3-D representations of molecules #30, and #62, respectively. #30 has an unusual amount of branching that might make the –OH group much more sequestered than in most of the other alcohols, thus lowering the effectiveness of its hydrogen bonding to water. Molecules #62 and #64 are not unusual in any noticeable way, indicating possible flaws in either in the model or in the original experimental data. The top two outliers for each physical property were removed from the data set and the regressions rerun. Using the volume as the single parameter:

$$\ln S = 7.71(\pm 0.24) - 0.0840(\pm 0.0017)^* V \qquad (2.4)$$

$$n = 64 \quad R^2 = 0.980 \quad s = 0.557 \quad F = 2971 \quad Q^2 = 0.978$$

Using the surface area as a single parameter:

$$\ln S = 8.01(\pm 0.27) - 0.062(\pm 0.0013)^*SA \qquad (2.5)$$

$$n = 64 \quad R^2 = 0.972 \quad s = 0.657 \quad F = 2114 \quad Q^2 = 0.970$$

Using the polarizability as a single parameter:

$$\ln S = 8.08(\pm 0.30) - 1.823(\pm 0.045)^*POLcalc \qquad (2.6)$$

$$n = 64 \quad R^2 = 0.964 \quad s = 0.697 \quad F = 1679 \quad Q^2 = 0.962$$

The improvement of the model by removing the top two outliers is most significant for polarizability. The overall improvement of all three models is slight, indicating that the

original models are adequate. Figure 2.A6 in the appendix is the plot of the volume equation 2.4.

The best physical two-property model uses volume and polarizability. This supports the factor analysis findings, since volume was the best representation of factor 1, and polarizability was the best representation of factor 3. Figure 2.A7 in the appendix is the corresponding plot. Using the volume and the polarizability:

$$\ln S = 7.89(\pm 0.27) - 0.0585(\pm 0.011)*\text{volume} - 0.562(\pm 0.25)*\text{polarizability} \qquad (2.7)$$

$$n = 66 \quad R^2 = 0.970 \quad s = 0.649 \quad F = 1097 \quad Q^2 = 0.969$$

The improvement of the two-term model is statistically insignificant over the best single-term model. R^2 improves by less then 0.001, s does not improve by more then 0.02, and the t-test value for the polarizability is very low. The top two outliers are again #62 and #64 respectively. This shows there is no real improvement in using two physical properties. A three-term regression yielded similar results, with no real statistical improvement. The single physical property of volume is adequate to model the solubilities of this set of compounds.

2.4.2 Using the Topological Descriptors

The topological descriptors generated by QsarIS create better models than the calculated physical properties. QsarIS employs a genetic algorithm to parse the 50+ topological parameters it generates. The best regression models for all 66 alcohols are presented with their R^2 values in Table 3. Only the best 1, 2, and 3 term regressions are shown, with some regression with the repeating descriptors (κa2 and χ_0^v) removed.

Table 2.4 Summary of R^2 for the Best 1-term, 2-term, and 3-term Regressions using Topological Descriptors

R^2	Variables		
0.9847	ABSQ		
0.9801	X_1		
0.9109	κ2		
0.9872	Xo	ABSQ	
0.9872	ABSQ	Qs	
0.9872	SsCH3	ABSQ	
0.9872	κa2	ABSQ	
0.9879	Xo	ABSQ	Dipole
0.9879	ABSQ	Dipole	Qv
0.9878	κ2	ABSQ	Dipole
0.9878	κ2	ABSQ	Dipole

Clearly the best regression does not exceed R^2 =0.9879. It is also apparent that adding a third term does not significantly improve the regression.

Table 2.5 is a correlation matrix of the pertinent topological descriptors.

Table 2.5 Correlation Matrix for the Calculated Descriptors and Solubility and the Solubility

	ln S	X_1	k2	ABSQ	Dipole	Pol	Sur	Vol
ln S	1	-0.993	-0.751	-0.9949	0.1131	0.176	-0.9835	-0.988
X_1		1	0.413	0.9979	-0.101	0.994	0.989	0.996
k2			1	0.261	-0.007	0.659	0.644	0.953
ABSQ				1	-0.1029	0.993	0.986	0.993
Dipole					1	-0.121	-0.055	-0.069
Pol						1	0.976	0.987
Sur							1	0.997
Vol								1

Factor analysis of the solubility data and the descriptors can be found in Table 2.6. As was the case with the calculated physical properties, one factor accounts for

almost all (98.9%) of the variance in the solubility data. DP clearly has no significant correlation with any of the other parameters. Factors 4 and 5 are not well matched by any of the topological descriptors or calculated physical properties.

Table 2.6 Factor Percentage Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Ln(S)Obs	98.96%	0.00%	0.06%	0.76%	0.20%
x1	94.93%	0.15%	4.82%	0.00%	0.07%
k2	90.88%	0.51%	8.43%	0.01%	0.16%
ABSQ	99.60%	0.00%	0.08%	0.19%	0.01%
Dipole	1.18%	98.68%	0.12%	0.00%	0.00%
Polarizability	98.33%	0.03%	1.57%	0.00%	0.05%
Surface	98.62%	0.31%	0.35%	0.62%	0.07%
Volume	99.46%	0.16%	0.00%	0.28%	0.05%

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
% accounted for	85.24%	12.48%	1.93%	0.23%	0.08%
total %	85.24%	97.73%	99.66%	99.89%	99.97%

The best single topological descriptor is ABSQ.

$$\ln S = 14.50(\pm 0.29) - 13.01(\pm 0.20)*\text{ABSQ} \qquad (2.8)$$

$n = 66 \quad R^2 = 0.9847 \quad s = 0.477 \quad F = 4109 \quad Q^2 = 0.9834$

ABSQ is the sum of absolute values of the (Gasteiger) charges on each atom of the molecule, in electrons. The ability of ABSQ to accurately model the aqueous solubility may indicate that although dispersion forces are dominant in the solubility process, the charges on atoms provide some small interactions. Figure 2.A8 in the appendix gives a plot of equation 2.8.

As noted earlier, Cammarata[5] used a technique derived from Randic's[4] connectivity index to model the aqueous solubilities of 51 aliphatic alcohols. The simple first order connectivity for the larger group of 66 aliphatic alcohols used here generates a

better model than Cammarata's, most likely due to the addition of the15 aqueous solubility data points.

$$\ln S = 6.91\ (\pm 0.20) - 2.739\ (\pm 0.049)*\chi_1 \qquad (2.9)$$

$$n = 66 \quad R^2 = 0.9802 \quad s = 0.542 \quad F = 3177 \quad Q^2 = 0.9787$$

Figure 2.A9 in the appendix is a plot of equation 2.9.

The best 2-term topological descriptor regression uses χ_o and ABSQ:

$$\ln S = 15.43(\pm 0.37) - 16.19(\pm .0.91)*ABSQ + 0.52(\pm 0.15)*\chi_o \qquad (2.10)$$

$$n = 66 \quad R^2 = 0.9872 \quad s = 0.438 \quad F = 2438 \quad Q^2 = 0.9857$$

Note that with a t-test of just 3.6, χ_o is not very significant. It is reasonable that χ_o adds some statistical significance to the ABSQ model. χ_o is informative of the general size, (correlating highly with both surface area and volume) of the solute alcohol. The amount of branching is relevant to whether the hydroxyl site is "hidden" in the solute. Figure 2.A10 in the appendix is a plot of equation 2.10.

Adding a third variable did not significantly improve the regression, with the third term having a very low t-test value (<1). Clearly any variance in the solubility data described by the dipole descriptor is covered by ABSQ and Xo.

The top two outliers in the models given by equations 2.8 and 2.9 were #30 and #62, the same compounds as for the volume regressions. The two outliers were removed and the regressions rerun. The best single topological descriptor was again ABSQ:

$$\ln S = 14.63(\pm 0.24) - 13.13(\pm 0.17)*ABSQ \qquad (2.12)$$

$$n = 64 \quad R^2 = 0.9899 \quad s = 0.391 \quad F = 6058 \quad Q^2 = 0.9891$$

The first-order connectivity regression has one significant outlier, #62. This supports the idea that #62 is an incorrectly taken datum point.

$$\ln S = 7.01(\pm 0.18) - 2.771(\pm 0.043)^* \chi_1 \quad (2.13)$$

$$n = 65 \quad R^2 = 0.9851 \quad s = 0.472 \quad F = 4156 \quad Q^2 = 0.984$$

The best 2-term topological descriptor regression uses Xo and ABSQ.

$$\ln S = 15.54(\pm 0.29) - 16.25(\pm .0.71)^*ABSQ + 0.51(\pm 0.11)^* \chi_o \quad (2.14)$$

$$n = 64 \quad R^2 = 0.9924 \quad s = 0.341 \quad F = 3979 \quad Q^2 = 0.9915$$

Figure 2.A11 in the appendix is a plot of equation. 2.12. There is considerable statistical improvement with the removal of the outliers. The s value decreased roughly 25% for both models. The F value increased dramatically. The best 3-term topological descriptor regression contains ABSQ, χ_o, and Dipole.

$$\ln S = 15.16(\pm 0.35) - 16.44(\pm .0.70)^*ABSQ + 0.55(\pm 0.11)^*Xo - 0.28(\pm 0.16)^*Dipole$$

$$n = 64 \quad R^2 = 0.9929 \quad s = 0.335 \quad F = 2752 \quad Q^2 = 0.9913 \quad (2.15)$$

Removal of the two outliers did not make dipole moment or the addition of a third term to the model any more relevant.

2.4.3. Calculated Physical Properties and the Topological Descriptors together

It is difficult to improve upon the two-term topological models. Mixing the two types of parameters is undesirable, as noted in Chapter 1. The best mixed two-term model contains polarizability and ABSQ. This regression is slightly weaker statistically than the two-term topological descriptor model of Eq. 2.9. If the top two outliers are removed (#30 and #62) and the regression rerun, the best statistical model in this chapter is found. The model is:

$$\ln S = 17.63(\pm 0.29) - 19.30(\pm 0.71)^*ABSQ + 0.88(\pm 0.11)^*polarizability \quad (2.16)$$

$$n = 64 \quad R^2 = 0.9948 \quad s = 0.284 \quad F = 5809 \quad Q^2 = 0.9941$$

Once again, ABSQ has the best ability to explain the “Factor 1” variance in the data, and polarizability seems best attuned to addressing the leftover variance.

2.5. Appendix

Figure 2.A1. - Solubility modeled by Volume

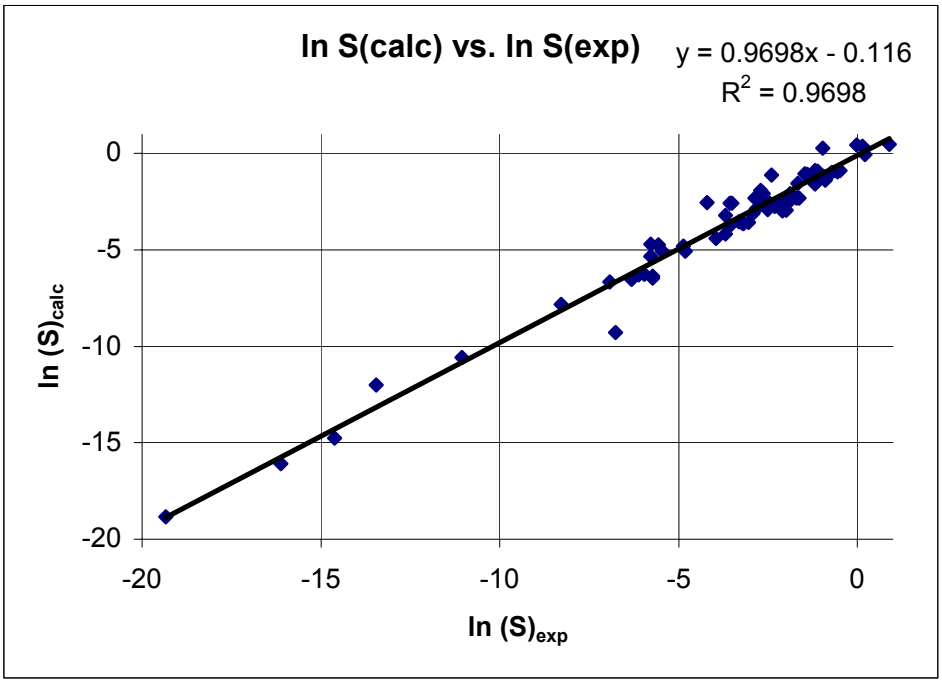


Figure 2.A2. - Solubility modeled by Surface Area

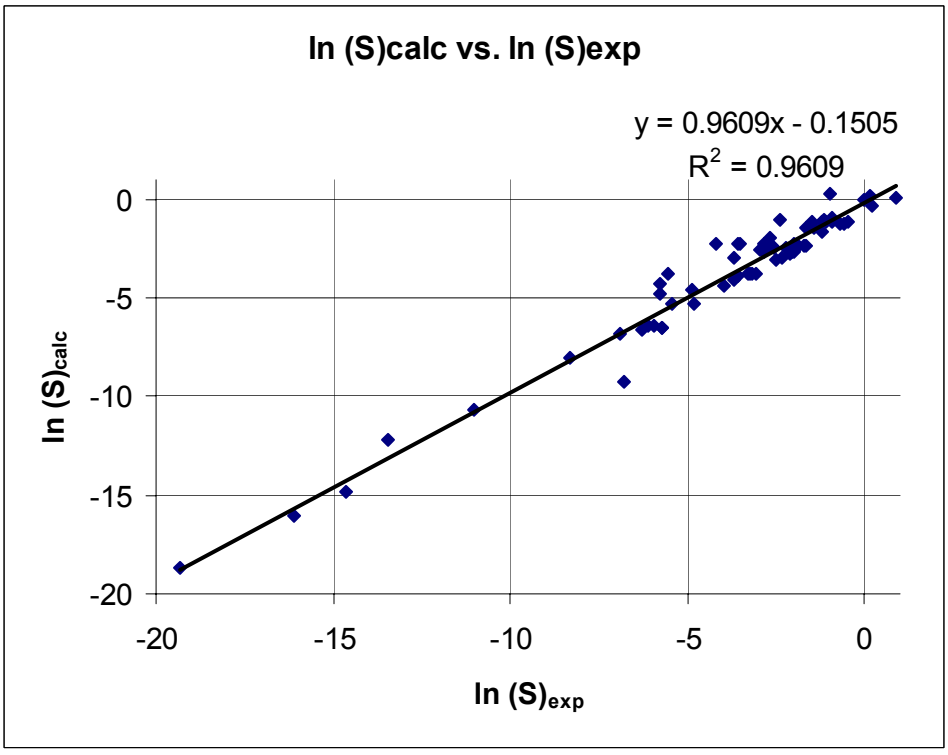


Figure 2.A3. - Solubility modeled by Polarizability

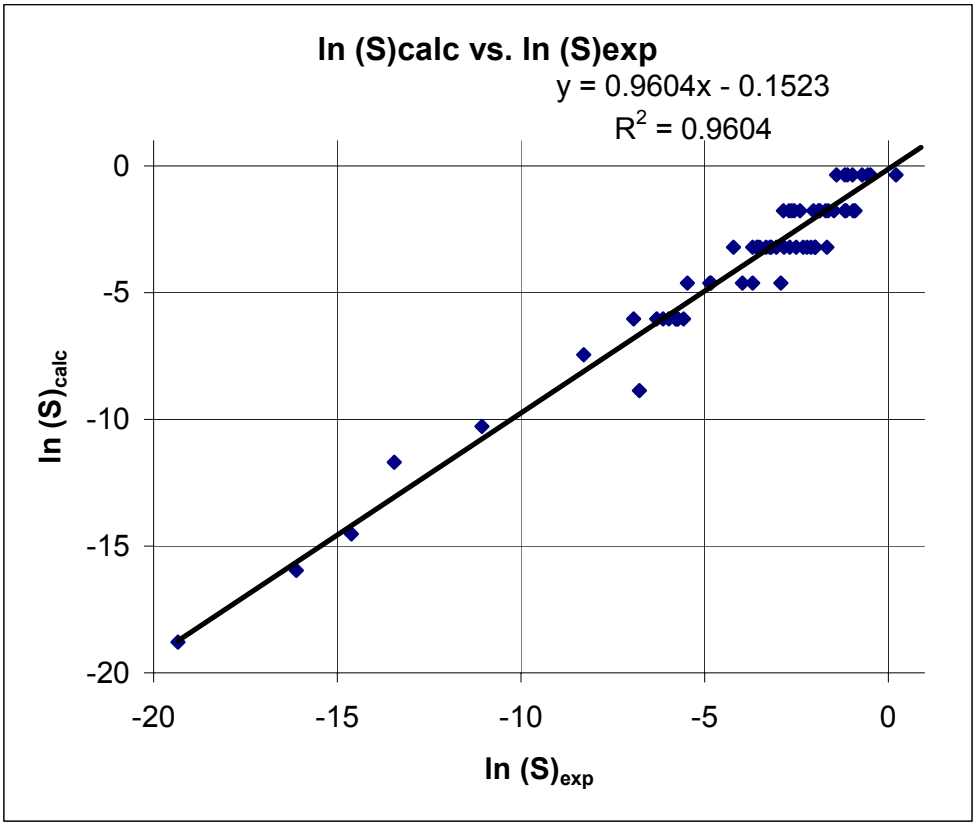


Figure 2.A4. - Alcohol #30

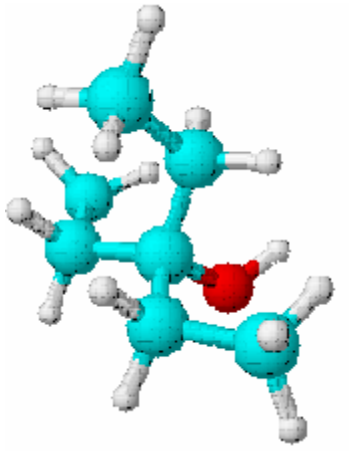


Figure 2.A5. - Molecule #62

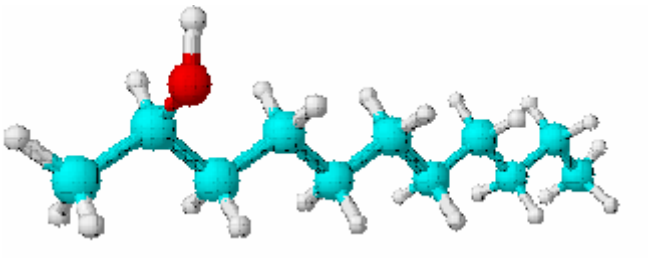


Figure 2.A6. – Solubility modeled by Volume, with the top two outliers removed

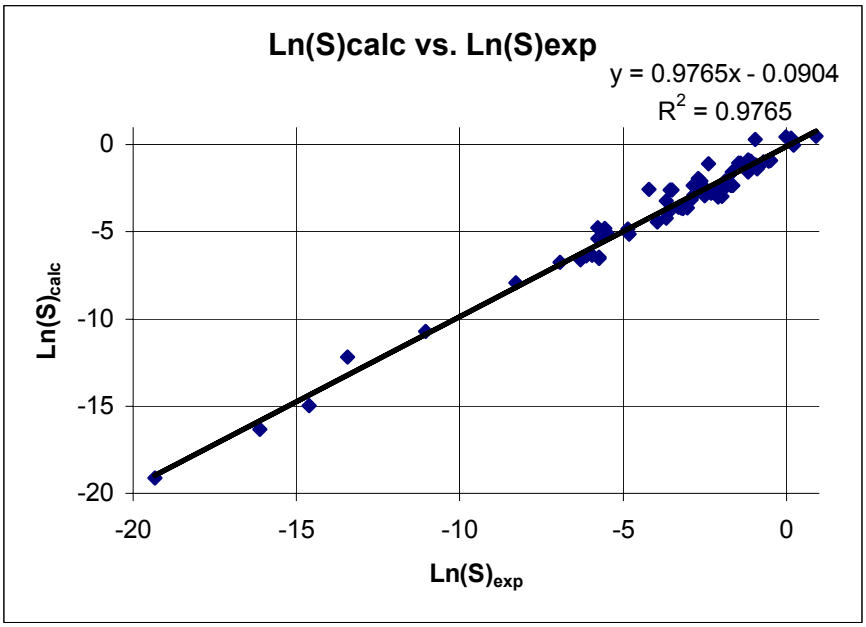


Figure 2.A7. – Solubility modeled by Volume and Polarizability

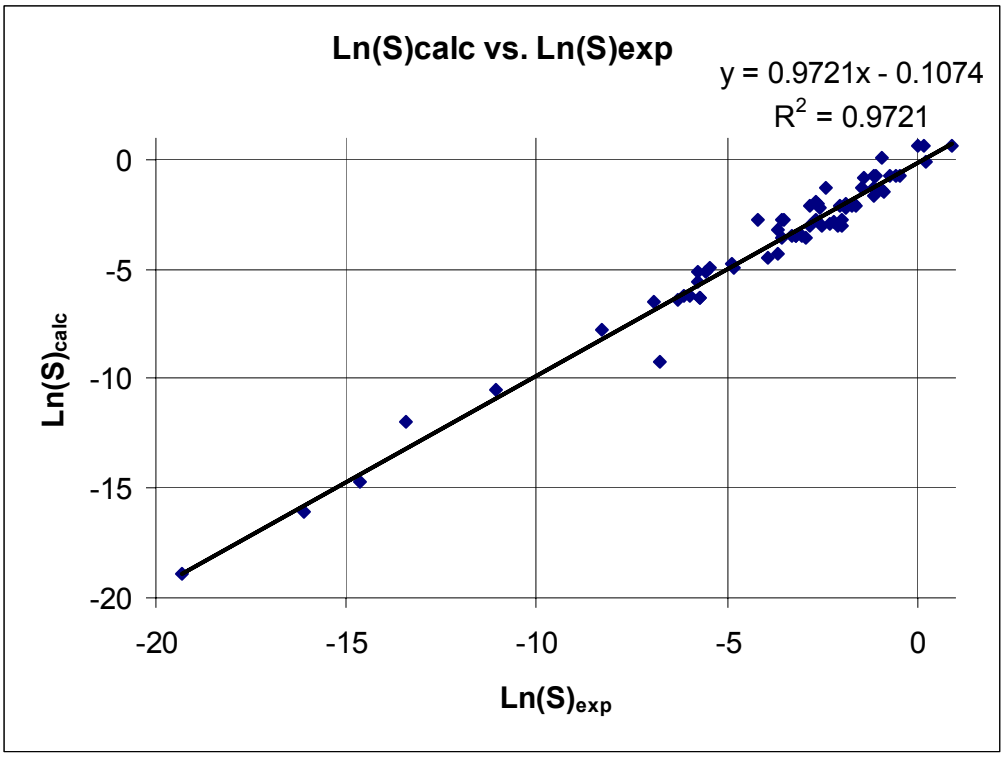


Figure 2.A8. - Solubility modeled by ABSQ

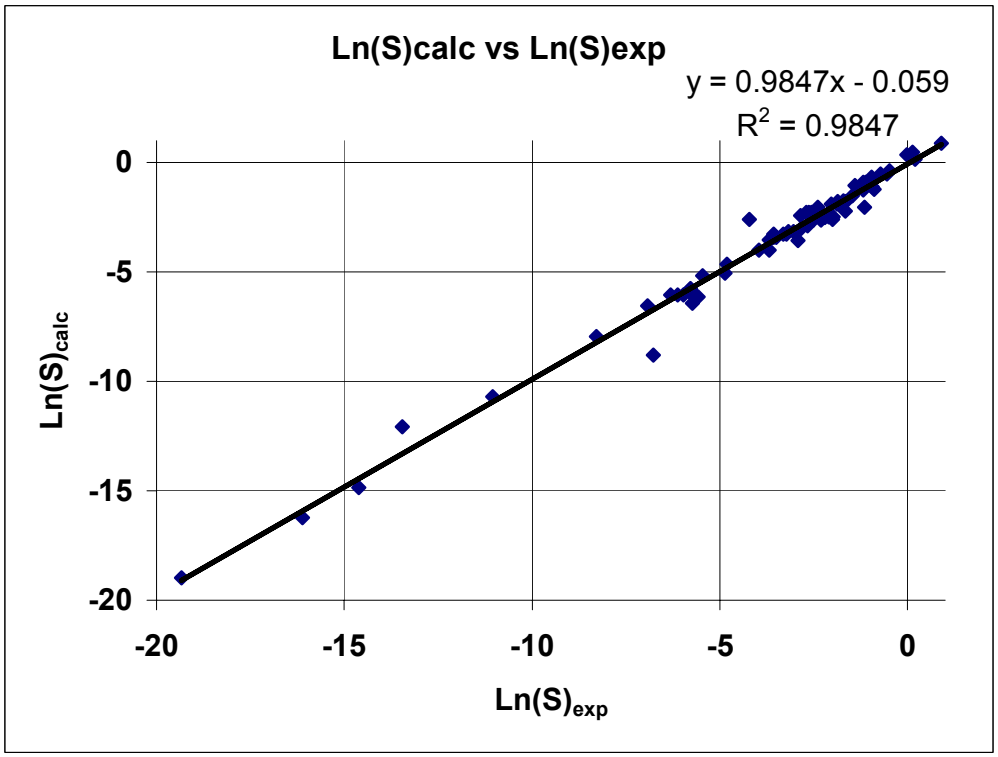


Figure 2.A9. - Solubility modeled by X_1

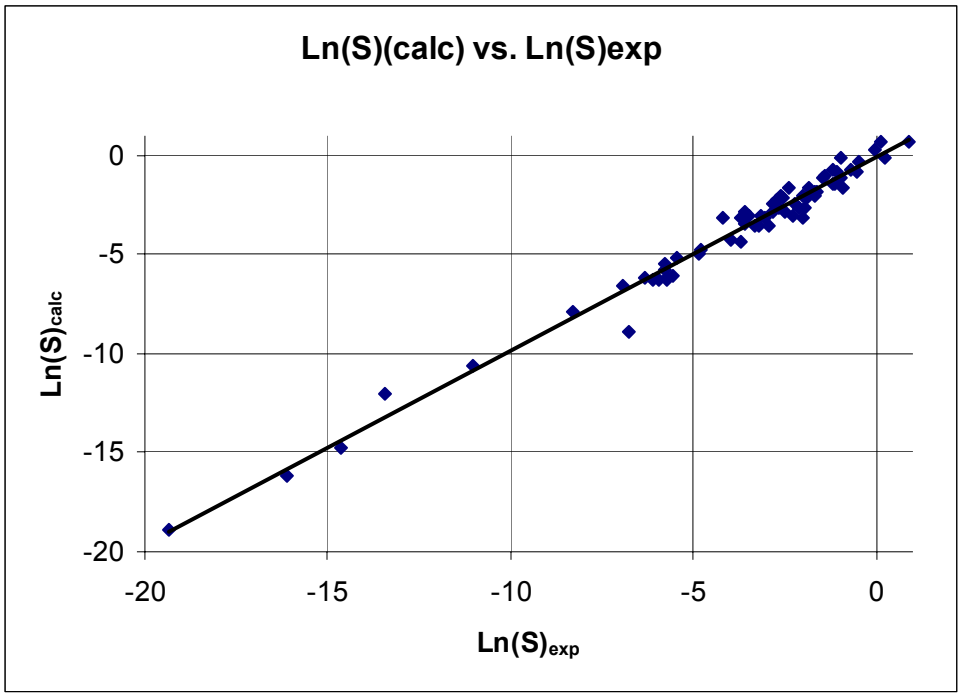


Figure 2.A10. - Solubility modeled by ABSQ and Xo

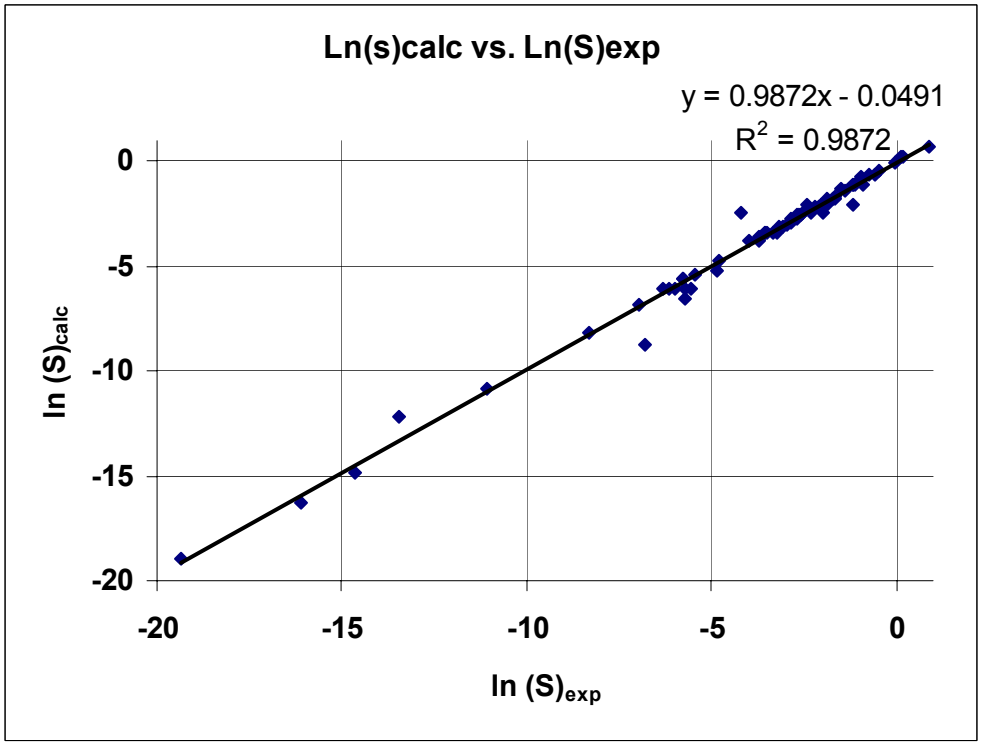


Figure 2.A11. - Solubility modeled by ABSQ with the top two outliers removed

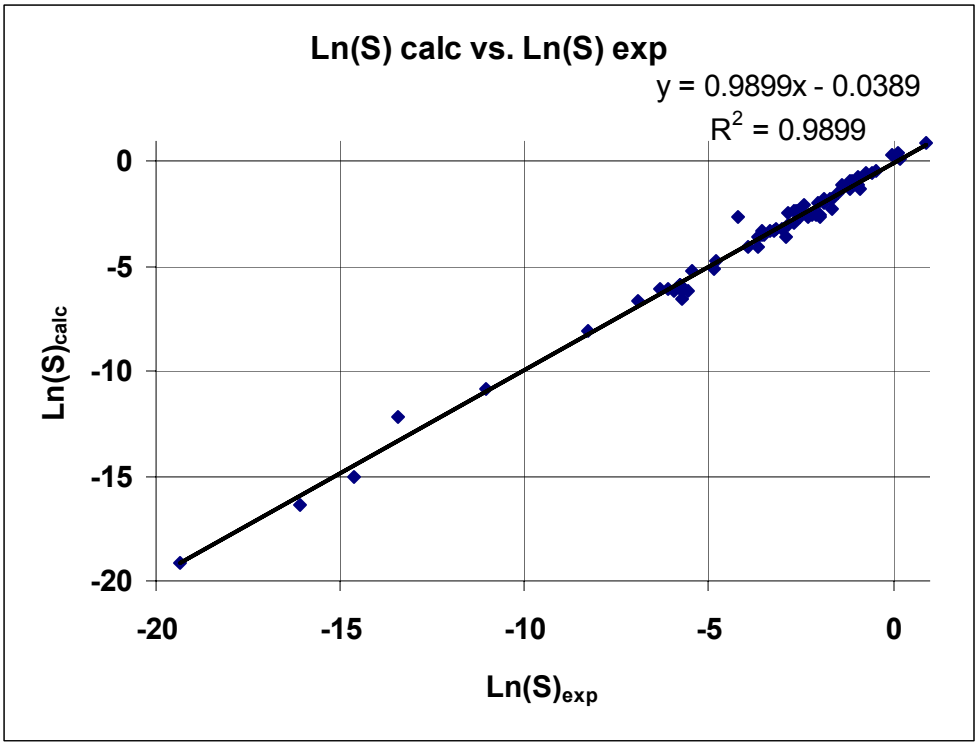


Table 2.A1. – All Monoalcohols and their Calculated Physical Properties

Molecules	Name	ln S	Polarizability	Surface	Volume	Molecules	Name	ln S	Polarizability	Surface	Volume
1	1X4	-0.023	3.87	127.652	86.5772	35	4X7	-3.224	6.192	190.178	134.69
2	2M1X3	0.138	3.87	125.377	87.5113	36	5M2X6	-3.178	6.192	189.452	135.21
3	2X4	0.898	3.87	126.993	86.1798	37	2M3X6	-3.039	6.192	189.334	134.509
4	1X5	-1.405	4.644	151.129	104.595	38	22MM3X5	-2.671	6.192	166.833	121.485
5	2M1X4	-1.105	4.644	146.189	103.001	39	24MM3X5	-2.832	6.192	165.588	126.099
6	3M1X4	-1.174	4.644	146.554	102.292	40	2M2X6	-2.51	6.192	178	126.426
7	22MM1X3	-0.967	4.644	123.373	88.4597	41	3M3X6	-2.303	6.192	175.823	124.659
8	2X5	-0.714	4.644	148.034	103.497	42	23MM2X5	-2.095	6.192	172.501	127.303
9	3X5	-0.553	4.644	147.554	103.132	43	23MM3X5	-1.98	6.192	172.098	126.935
10	3M2X4	-0.484	4.644	145.869	102.543	44	3E3X5	-2.003	6.192	165.318	122.208
11	2M2X4	0.207	4.644	133.455	92.5527	45	233MMM2X4	-1.658	6.192	150.835	110.388
12	1X6	-2.855	5.418	171.736	119.411	46	1X8	-5.457	6.966	215.636	152.262
13	2M1X5	-2.556	5.418	168.325	120.034	47	2E1X6	-4.858	6.966	202.74	149.042
14	4M1X5	-2.625	5.418	166.412	116.525	48	2X8	-4.812	6.966	215.515	152.181
15	22MM1X4	-2.395	5.418	144.394	104.993	49	2M2X7	-3.96	6.966	200.316	144.195
16	33MM1X4	-1.151	5.418	145.116	105.079	50	3M3X7	-3.684	6.966	194.508	141.392
17	2E1X4	-2.694	5.418	160.068	114.861	51	223MMM3X5	-2.924	6.966	170.274	128.968
18	2X6	-2.026	5.418	171.01	119.386	52	1X9	-6.931	7.74	240.49	171.121
19	3X6	-1.888	5.418	167.257	116.568	53	7M1X8	-5.733	7.74	234.883	167.795
20	3M2X5	-1.704	5.418	165.928	119.352	54	22EE1X5	-5.572	7.74	189.333	148.219
21	4M2X5	-1.865	5.418	165.772	120.573	55	2X9	-6.309	7.74	236.029	169.53
22	2M3X5	-1.635	5.418	166.286	119.317	56	3X9	-6.125	7.74	233.406	166.754
23	33MM2X4	-1.474	5.418	146.488	104.502	57	4X9	-5.964	7.74	232.663	166.548
25	2M2X5	-1.174	5.418	155.487	110.633	58	5X9	-5.733	7.74	235.245	168.798
26	3M3X5	-0.898	5.418	147.052	108.242	59	26MM4X7	-5.779	7.74	205.816	155.342
27	23MM2X4	-0.944	5.418	143.357	105.812	60	35MM4X7	-5.779	7.74	198.962	147.965
28	24MM2X5	-2.21	6.192	168	123.661	61	1X10	-8.289	8.514	259.567	185.082
29	1X7	-4.214	6.192	165.318	122.208	62	2X11	-6.77	9.288	280.175	202.566
30	22MM1X5	-3.5	6.192	165.655	122.597	63	1X12	-11.05	10.062	303.409	217.932
31	24MM1X5	-3.684	6.192	176.204	130.038	64	1X14	-13.45	10.836	327.336	234.857
32	44MM1X5	-3.569	6.192	164.194	122.686	65	1X15	-14.62	12.384	370.955	267.846
33	2X7	-3.569	6.192	192.071	136.462	66	1X16	-16.12	13.158	391.052	283.667
34	3X7	-3.316	6.192	189.851	133.991	67	1X18	-19.34	14.706	434.776	316.614

2.6. References

[1] Hermann R. B., Theory of Hydrophobic Bonding II, *J. Phys. Chem.*, **1972**, *76*, 2574-2582.

[2] Amidon, G. L.; Yalkowsky, S. H.; Leung, S., Solubility of Nonelectrolytes in Polar Solvents II: Solubility of Aliphatic Alcohols in Water, *J. Pharm. Sci,*. **1974**, *63*, 1858-1866.

[3] Yalkowsky, S. H.; Benerjee, S., *Aqueous Solubility Methods of Estimation for Organic Compounds*, Dekker, N.Y. **1992** p. 75-98.

[4] Randić, M., On Characterization of Molecular Branching *J. Am. Chem. Soc.*, **1975,** *97*, 6609-6615.

[5] Cammarata, A., Molecular Topology and Aqueous Solubility of Aliphatic Alcohols, *J. Pharm. Sci*., **1978**, *68* 839-842.

[6] Kier, L. B., *Physical Chemical Properties of Drugs*, Dekker, N.Y. **1980**, p. 277-320.

[7] Valvani, S. C; Yalkowsky, S.H., Solubility and Partitioning I: Solubility of Nonelectrolytes in Water, *J. Pharm. Sci* **1980**, *69*, 912-922.

[8] QsarIS ver. 1.2 **2001,** SciVision Inc., 200 Wheeler Road, Burlington, MA. 01803.

[9] StatMost 3.0 for Windows ©**1995,** DataMost Corporation Los Angeles, California.

[10] Gasteiger, J.; Marsili, M., Iterative Partial Equalization of Orbital Electronnegativity-A Rapid Access to Atomic Charges, *Tetrahedron*, **1980**, *56*, p. 3219-3228.

[11] Kier, L. B.; Hall, L.H., *Molecular Connectivity in Structure-Activity Analysis,* Research Studies Press LTD, Letchworth England **1986.**

[12] Kier, L. B,. Inclusion of Symmetry as a Shape Attribute in Kappa Index Analysis, *Quant. Struct.-Act. Relat.* **1987**, *6*, 8-12.

[13]Kier, L. B.; Hall, L.H., *Molecular Structure Descriptions: The Electrotopological State*, Academic Press: New York, **1995**, p.65ff.

Chapter 3

Modeling the Aqueous Solubilities of Halogenated Alkanes

3.1. Abstract The solubilities of 72 halogenated alkanes in water were modeled using topological descriptors, calculated physical properties, and combinations of the two to provide the best one-, two-, three-, and four-term regressions. The halogenated alkanes were first broken down into subgroups, and later modeled as a whole. The chloroalkanes were modeled best by topological descriptors, with an R^2 of 0.9827. The bromoalkanes were modeled best by topological descriptors, with the best overall model having R^2 of 0.9916. The iodoalkanes were modeled best by a calculated physical property (volume), with $R^2 = 0.9908$. The best overall models for all 72 halogenated alkanes utilized only calculated physical properties, and had an R^2 of 0.9684.

3.2. Introduction

The aqueous solubilities of halogenated alkanes are important for the areas of industrial production and environmental pollution. Modeling of aqueous solubilities has been carried out using a variety of methods. These methods all use properties of the solutes, e.g., physical properties, topological descriptors, and calculated physical properties, to describe the solubility.

Yalkowsky, Orr, and Valvani[1] analyzed the aqueous solubilities of halobenzenes and polycyclic aromatic hydrocarbons in 1979 using melting points, molecular surface areas, and octanol-water partition coefficients.

Dunnivant, Elzerman, Jurs, and Hasan used a similar approach[2] in 1992. Melting points, molecular surface areas, and various topological descriptors were used to predict the molar solubilites of polychlorinated biphenyls.

In 1998 Katritzky and Huibers[3] used solute topological descriptors to model the aqueous solubilities of a diverse set of molecules. The set included halogenated hydrocarbons, hydrocarbons, and PCBs. Using a three-term regression Katritzky and Huibers obtained an overall $R^2 = 0.959$, and an $R^2 = 0.962$ for the halogenated hydrocarbons as a group. The three terms were molecular volume, bonding information content order, and atomic charge weighted over partial negative surface area.

Other properties of the solutes have been used to predict the aqueous solubilites for a large number of compounds. Ruelle and Kesselring[4] predicted aqueous solubilities of a variety of halocompounds and PCBs using thermodynamic properties of the solutes. This work obtained an overall $R^2 = 0.950$.

3.3. Methods

Experimental solubility data for 72 haloalkanes were obtained from the work of Ruelle and Kesselring[4], and Katritzky and Huibers[3]. The solubilities were modeled as the natural logarithm of the molar solubility S. The fluoroalkanes were not examined as a group due to a lack of data.

The software used to generate the calculated physical properties, the topological descriptors, and all regression analyses, was QsarIS.[5] A set of over 50 different topological descriptors, and 30 calculated physical properties was screened using a genetic algorithm. QsarIS gives no units for the calculated physical properties and topological descriptors it generates. Their corresponding abbreviations were retained from the program.

All factor analysis work was done using the StatMost[6] software program. Only principal components analysis (PCA) was used.

Table 3.1. Descriptors Used and Their Abbreviations

Parameter	Description
x1	Simple first order connectivity index, as defined in chapter 1
xvc3	Valence third order connectivity index, as defined in chapter 1
Vol	Calculated Volume
Sur	Calculated Surface Area
Q	The magnitude of the principal quadrupole moment[i]
4sC	E-state descriptor for carbon atom with 4 single bonds[7]
sCl	E-state descriptor for Cl with one single bond[7]
sF	E-state descriptor for Fl with one single bond[7]
k1	1st order Kier Kappa Alpha shape index, defined in chapter 1[7]
Q	The magnitude of the principal quadrupole moment.
Pol	Calculated Polarizability as defined in Chapter 1

3.4. Results and Discussion

3.4.1. Chloroalkanes

The aqueous solubilities of 29 chloroalkanes were modeled. The principle of parsimony suggests that no more then four terms should be used to model the solubilities. Lists of the solutes, their calculated physical properties, and topological descriptors can be found in appendix 3A. Table 3.2. is the correlation matrix, and Table 3.3. is the factor analysis of the pertinent physical properties and topological descriptors for all the compounds.

Table 3.2. Correlation Matrix for the Chloroalkanes (29 solutes)

	ln S	4sC	sCl	k1	Pol	Sur	Vol	Q
ln S	1	0.512	-0.239	-0.922	-0.591	-0.785	-0.853	0.272
4sC		1	-0.799	-0.409	-0.789	0.049	-0.087	0.232
sCl			1	0.346	0.911	-0.123	0.014	0.242
k1				1	0.702	0.853	0.919	-0.102
Pol					1	0.284	0.416	0.164
Sur						1	0.987	0.009
Vol							1	0.011
Q								1

The solubility data do not correlate highly with any of the physical properties or topological descriptors except Vol and k1. This is to be expected, as the Vol is used successfully as a parameter for modeling in the examined literature.[2,3] The Pol was correlated highly to both of the e-states indicies 4sC and sCl. Since the e-states take into account the electronic character of the atoms, this result is expected. Clearly Q stands about, having little correlation with any of the other parameters.

Table 3.3. Factor Analysis for the Chloroalkanes (29 solutes)

Percent Factor Weight

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
% Factor	54.38%	28.31%	15.20%	1.56%	0.36%	0.18%
total %	54.38%	82.68%	97.88%	99.44%	99.80%	99.98%

Percent Factor Loadings

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
ln S	86.79%	3.93%	5.71%	2.59%	0.50%
4sC	34.19%	49.34%	11.99%	3.83%	0.48%
sCL	24.92%	70.51%	2.84%	1.54%	0.03%
k1	94.79%	3.28%	0.36%	0.22%	0.74%
Pol	65.45%	31.00%	1.98%	1.53%	0.02%
Sur	57.29%	39.80%	1.87%	0.14%	0.85%
Vol	71.44%	26.64%	1.52%	0.07%	0.24%
Q	0.16%	1.95%	95.36%	2.54%	0.01%

The factor analysis indicates that Factor one provides most of the variance in the solubility data. Factor two and Factor three also play a significant role for solubility. Factors six and seven were removed from the percent factor loadings table because they have minimal contributions (less then 0.1%) to the variance in the solubility data. k1 and Vol have factor loadings similar to that of ln S. This may be why Vol and k1 are the most significant (as evidence by the t-test) parameters in the physical property and topological descriptor regression, respectively.

3.4.1.1. Chloroalkanes: Calculated Physical Properties (29 solutes)

The best single-term regression is Vol:

$$\ln S = 3.36(\pm 0.99) - 0.0892(\pm 0.010)*\text{Vol} \qquad (3.1)$$

$n = 29 \quad R^2 = 0.7277 \quad s = 1.04 \quad F = 72.2 \quad Q^2 = 0.692$

There is one major outlier outside of 2s, hexachloroethane. There is a second minor outlier outside of 1s, 1,2,3-trichloropropane. Hexachloroethane may be an incorrect data point, or it may be such a large outlier because it has a unique lack of ability (relative to the rest of the data set) to hydrogen bond. The low R^2 can be expected, as the volume gives no indication of hydrogen bonding ability. The best two-term regression is:

$$\ln S = 1.96(\pm 0.81) - 0.313(\pm 0.049)*\text{Vol} + 0.168(\pm 0.036)*\text{Sur} \qquad (3.2)$$

$$n = 29 \quad R^2 = 0.8505 \quad s = 0.78 \quad F = 73.9 \quad Q^2 = 0.758$$

The top outlier is slightly more than 2s, and 1,1,2,2-tetrachloroethane. Here the Sur is added and the model improves. The second step in solubility, where energy is used to create a hole is made in the solvent, should be related to the size and volume of the solute. The third step, in which energy is given back by solute-solvent interaction, should be related to the surface area of the solute. No term in this model accounts for hydrogen bonding activity. It is of interest that neither Pol, nor the specific specPol (polarizability/volume) performs better than Sur. This could be due to the crude method of estimation for polarizability, or could suggest that polarizability has little correlation with hydrogen bonding ability. Figure 3.A1 in the appendix is a plot of this model.

The best three-term model includes Vol, Sur, and Pol. This model is not included in this work, as there is only a very slight statistical improvement to the model. (R^2 = 0.86)

3.4.1.2. Chloroalkanes: Topological Descriptors (29 solutes)

The topological descriptors create better models for the chloroalkanes than the calculated physical properties. The best one-term regression utilizes k1 and has an $R^2 = 0.86$. This model is omitted because the best two-term model is more statistically significant. The best two-term descriptor regression follows:

$$\ln S = 2.24(\pm 0.48) - 1.392(\pm 0.085)*k1 + 0.63(\pm 0.14)*Q \quad (3.3)$$

$$n = 29 \quad R^2 = 0.9177 \quad s = 0.585 \quad F = 145 \quad Q^2 = 0.890$$

Figure A3 in the appendix is a plot of this model. This model is better than the two-term physical properties model Eq 3.2. k1 has factor loadings closer to ln S than Vol, indicating that it should be a better first parameter. Vol and k1 are highly correlated, indicating that k1 is in some way related to volume. Q, or the principle quadrupole moment, could be giving an indication of possible ability to hydrogen bond. There is one outlier more than 3s, hexachloroethane. This molecule seems problematic for the all the chloroalkane models.

The best three-term regression is:

$$\ln S = 1.01(\pm 0.31) + 1.35(\pm 0.15)*4sC + 0.227(\pm 0.018)*sCl - 1.336(\pm 0.061)*k1$$

$$n = 29 \quad R^2 = 0.9672 \quad s = 0.377 \quad F = 246 \quad Q^2 = 0.947 \quad (3.4)$$

Figure 3.A3. in the appendix is a plot of this model. Two e-state indices and the Kier shape index provide the best regression. The e-state indices contain information about the topology and electronic states of atoms. The chlorine and carbon indicies are used in this regression. An e-state index combines both the electronic character and the topological environment for each type atom.[8] In regression (3.4) the e-state index for chlorine atoms and that for carbon atoms with 4 single bonds are used. The e-state index for the chlorine may also indicative of the ability of the chlorine atom to foster hydrogen bonding (part of the e-state considers electronegativity and interatomic interactions). There is only one outlier more than 2σ, 2,3-dichlorobutane. This datum was removed and the regression rerun to yield the following.

$$\ln S = 1.01(\pm 0.31) + 1.41(\pm 0.13)*4sC + 0.227(\pm 0.015)*sCl - 1.315(\pm 0.052)*k1$$

$$n = 28 \quad R^2 = 0.9767 \quad s = 0.321 \quad F = 335 \quad Q^2 = 0.959 \quad (3.5)$$

The best four-term regression is not more statistically significant than the best three-term model, and so is not shown.

3.4.1.3. Chloroalkanes: Best overall Model

The best overall model combines the e-state for saturated carbons and the chlorines, with the polarizability of the solutes.

$$\ln S = 2.90(\pm 0.41) + 1.30(\pm 0.14)*4sC + 0.574(\pm 0.025)*sCl - 1.847(\pm 0.080)*Pol$$

$$n = 29 \quad R^2 = 0.9703 \quad s = 0.358 \quad F = 269 \quad Q^2 = 0.947 \qquad (3.6)$$

It is of interest that this is the best three-term regression, because it does not contain either k1 or Vol. There are several similar regressions that give R^2 values close to this value, as is shown in Table 3.3. Table 3.3. also indicates that adding another term to the regression adds little statistical significance to the model. The polarizability should relate to the dispersion forces in the solute, and sCl may relate to hydrogen bonding ability with the solvent. All three terms are significant. The largest outlier is again 2,3-dichlorobutane. This datum was removed and the regression rerun to yield:

$$\ln S = 2.77(\pm 0.31) + 1.38(\pm 0.11)*4sC + 0.572(\pm 0.020)*sCl - 1.818(\pm 0.052)*Pol$$

$$n = 28 \quad R^2 = 0.9827 \quad s = 0.277 \quad F = 455 \quad Q^2 = 0.963 \qquad (3.7)$$

Figure 3.A5. in the appendix is a plot of Eq. 3.7.

Table 3.4. – Regression with similar results to Eq. 3.6.

R^2	Parameters			
0.9703	4sC	sCl	Pol	
0.9727	Vol	4sC	SsCl	Pol
0.9657	4sC	sCl	k1	
0.9706	Vol	4sC	sCl	k1
0.9703	4sC	sCl	k1	Pol
0.9633	Vol	4sC	sCl	

3.4.2. Bromoalkanes (18 solutes)

The aqueous solubility of 18 bromoalkanes was modeled. The principle of parsimony dictates that no more than three terms should be used to model the solubility. A list of the solutes and their properties or topological descriptors can be found in appendix. Table 3.5. is a correlation matrix for the ln S, the calculated physical properties, and the topological descriptors.

Table 3.5. Correlation Matrix for the Bromoalkanes

	ln S	x1	xvc3	Sur	Vol	Q	Pol
ln S	1	-0.842	-0.109	-0.959	-0.961	0.12	-0.548
x1		1	0.194	0.858	0.921	0.3456	0.843
xvc3			1	-0.119	0.00	-0.114	-0.595
Sur				1	0.988	0.041	0.494
Vol					1	0.105	0.621
Q						1	0.418
Pol							1

The solubility is highly correlated with x1, Sur, and Vol. This is intuitively verified, as these three parameters should be related to dispersion forces. xcv3 does not correlate highly with any other parameter due to the data point tetrabromomethane.

Table 3.6. Factor Analysis of the Bromoalkanes

Percent Factor Weight

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 6
Factor %	61.56%	20.63%	16.63%	0.90%	0.09%
total %	61.56%	82.18%	98.81%	99.71%	99.80%

Percent Factor Loadings

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 6
ln S	86.10%	7.37%	5.28%	0.88%	0.31%
xv1	95.77%	1.32%	1.47%	0.65%	0.05%
xvc3	3.56%	60.50%	34.59%	1.31%	0.04%
Sur	85.62%	13.96%	0.02%	0.21%	0.07%
Vol	94.36%	5.32%	0.01%	0.02%	0.06%
Q	4.14%	19.54%	75.01%	1.31%	0.00%
Pol	61.36%	36.40%	0.01%	1.95%	0.07%

The factor analysis indicates that one factor provides almost all of the variance in the solubility data. Factor two is the variance in the data due mostly to tetrabromomethane, as shown by the % factor loadings for xvc3. This is verified by the regressions for the topological descriptors. A single property regression provides an adequate model when tetrabromomethane is removed from the data. Factors three and four were removed from the % factor loadings table because they contribute less than 0.1% to the variance in the solubilities.

3.4.2.1. Bromoalkanes: Calculated Physical Properties (18 solutes)

The same trends seen in the models utilizing the physical properties of the chloroalkanes are evident in the bromoalkanes. The best single-term regression uses Vol.

$$\ln S = 3.56(\pm 0.69) - 0.0887(\pm 0.0064)*\text{Vol} \quad (3.8)$$

$n = 18 \quad R^2 = 0.9231 \quad s = 0.755 \quad F = 192.1 \quad Q^2 = 0.906$

The largest outlier is tetrabromomethane. This data point was removed and the regression rerun to yield:

$$\ln S = 3.65(\pm 0.60) - 0.0887(\pm 0.0056)*\text{Vol} \quad (3.9)$$

$n = 17 \quad R^2 = 0.9433 \quad s = 0.662 \quad F = 250.0 \quad Q^2 = 0.926$

Figure 3.1B is a plot of this model. Tetrabromomethane is unique in the bromoalkane data set, as it has no spot for possible hydrogen bonding. This effect was seen with hexachloroethane in the chloroalkanes. There are no outliers (> 1s) in this model. Tetrabromomethane could be an incorrect data point, but that is unlikely given the topological descriptor regression (q.v.).

The best two-term calculated physical property regression is Vol and Sur, which is the same as the chloroalkanes.

$$\ln S = 3.68(\pm 0.81) - 0.0520(\pm 0.049)*\text{Vol} + 0.0273(\pm 0.036)*\text{Sur} \quad (3.10)$$

$n = 18 \quad R^2 = 0.9271 \quad s = 0.760 \quad F = 95.5 \quad Q^2 = 0.854$

The second term provides for no statistical improvement. The only large outlier (more than 1s) is again tetrabromomethane.

3.4.2.2. Bromoalkanes: Topological Descriptors (18 solutes)

The quality of the topological regression is slightly higher then the physical property regressions. This is due to x1 being a better parameter than Vol for creating a model. In addition, xvc1 is able to "account" for tetrabromomethane.

$$\ln S = 1.1(\pm 0.53) - 2.87(\pm 0.21)*x1 \quad (3.11)$$

$n = 18 \quad R^2 = 0.9215 \quad s = 0.763 \quad F = 188 \quad Q^2 = 0.901$

The same regression with tetrabromomethane removed is:

$$\ln S = 1.48(\pm 0.25) - 2.95(\pm 0.099)*x1 \quad (3.12)$$

$n = 17 \quad R^2 = 0.9834 \quad s = 0.357 \quad F = 894 \quad Q^2 = 0.980$

Figure 3.B2. in the appendix is a plot of Eq. 3.12.

The best two-term regression is

$$\ln S = 1.6(\pm 0.22) - 2.99(\pm 0.087)*x1 - 0.19(\pm 0.021)*xvc3 \quad (3.13)$$

$n = 18 \quad R^2 = 0.9876 \quad s = 0.314 \quad F = 596 \quad Q^2 = 0.980$

There are no outliers more than 1s in this model, and tetrabromomethane has the smallest residual in the data set. Here a second descriptor is added to account for the variance in the data added by tetrabromomethane. The xvc3 value for tetrabromomethane is an entire order of magnitude different from the xvc3 values for the other solutes.

3.4.2.3. Bromoalkanes: Best overall Models (18 solutes)

There is no better two-term model found than that of Eq. 3.13 when the calculated physical properties and topological descriptors are mixed. The best overall three-term regression is:

$$\ln S = 1.21(\pm 0.26) - 3.155(\pm 0.098)*x1 - 0.241(\pm 0.027)*xvc3 + 0.122(\pm 0.047)*Pol$$

$$n = 18 \quad R^2 = 0.9916 \quad s = 0.267 \quad F = 550 \quad Q^2 = 0.979 \qquad (3.14)$$

This model provides a term for the volume or size of the molecule (x1), a term to account for tetrabromomethane (xvc3), and the Pol. The result is a highly accurate model. Figure 3.B3. in the appendix is a plot of Eq. 3.14.

3.4.3. Iodoalkanes (8 solutes)

There are 8 iodoalkanes, which are modeled adequately by one calculated property, volume. The regression is

$$\ln S = 3.13(\pm 0.37) - 0.091(\pm 0.0036)*Vol \qquad (3.15)$$

$$n = 8 \quad R^2 = 0.9908 \quad s = 0.284 \quad F = 645.3 \quad Q^2 = 0.973$$

The calculated volume is the best single-term regression, and is sufficient to model the solubilities of the 8 iodoalkanes. The best single-term topological descriptor is k1 (R^2 = 0.827). This regression is omitted due to the ability of Eq. 3.15 to model the solubility. The single term regression for the iodoalkanes is more statistically significant than either of the single term regressions for the bromoalkanes. This may be due to a smaller data set. It is may also be indicative of hydrogen bonding being less important in the iodoalkanes than in bromoalkanes. Iodine is the largest halogen considered, so volume may be more important in this set. Appendix C contains the plot and data for the iodoalkanes.

The best two-term regression is Pol and Sur. Table 3.7. shows a summary of the best two-term and single-term models for the iodoalkanes.

Table 3.7. – Summary of the Best Models for the Iodoalkanes

R^2.	Parameters	
0.9995	Pol	Surface
0.9991	x1	Pol
0.9991	Vol	k1
0.999	Vol	Pol
0.9987	Vol	x1
0.9908	Vol	
0.8266	k1	

3.4.4. For all Halogenated Alkanes (71 solutes)

The aqueous solubility of 71 halogenated alkanes is modeled in this section. This is the entire set of halogenated alkanes reviewed before, plus 16 alkanes with mixed halogen groups on them. Some of the new solutes include fluorine atom(s). Table 3.8. is the correlation matrix for all the halogenated alkanes reviewed, their calculated physical properties, and topological descriptors. Table 3.9. is the factor analysis.

Table 3.8. Correlation Matrix for All Halogenated Alkanes

	ln S	Sur	4sC	sF	sCl	ka1	Pol	Vol
ln S	1	-0.854	0.188	-0.022	-0.034	-0.8274	-0.593	-0.877
Sur		1	0.155	-0.223	-0.115	0.696	0.478	0.987
4sC			1	-0.926	-0.246	-0.472	0.113	0.161
sF				1	0.037	0.325	-0.314	-0.249
sCl					1	0.368	0.253	-0.099
ka1						1	0.517	0.715
Pol							1	0.597
Vol								1

Table 3.9. Factor Analysis for All Halogenated Alkanes

Percent Factor Weight

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
% Factor	38.58%	29.30%	15.68%	11.82%	3.06%	1.25%	0.25%
total %	38.58%	67.88%	83.56%	95.38%	98.44%	99.68%	99.93%

Percent Factor Loadings

Parameter	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
ln S	89.61%	1.44%	1.64%	0.06%	6.90%	0.30%	0.05%
4sC	0.20%	96.06%	0.62%	1.22%	0.67%	1.11%	0.12%
sF	1.22%	87.76%	8.38%	0.51%	1.30%	0.27%	0.56%
sCl	1.31%	12.10%	78.41%	8.00%	0.03%	0.01%	0.14%
ka1	73.38%	22.46%	0.84%	1.10%	1.11%	0.42%	0.70%
Pol	50.37%	2.81%	18.40%	27.98%	0.41%	0.01%	0.03%
Sur	84.86%	3.85%	5.77%	4.46%	0.44%	0.39%	0.19%
Vol	91.01%	4.43%	3.17%	0.69%	0.54%	0.10%	0.00%

For the first time in this chapter the correlation table indicates that no property correlates with ln S better than 0.88. The variability in the ln S data is still largely dependant on Factor 1. It is of interest that Factor 3 is slightly more important to ln S than Factor 2. 4sC and sCl load highly on Factor 3, which may explain why the topological descriptors are better than the calculated physical properties when used to model the entire set of halogenated alkanes. As with the smaller groups of halogenated alkanes,Vol and the kappa shape index both load highly on Factor 1, and will be used to create the basis for the calculated physical properties model and the topological descriptor model, respectively.

3.4.4.1. Halogenated Alkanes: Calculated Physical Properties (71 solutes)

The best single-term regression is again Vol:

$$\ln S = 2.89(\pm 0.54) - 0.0850(\pm 0.0056)*\text{Vol} \quad (3.16)$$

$n = 71 \quad R^2 = 0.7698 \quad s = 1.11 \quad F = 230.8 \quad Q^2 = 0.759$

There are two outliers beyond 3s, hexachloroethane and chloropentafluoroethane. Both of these ethanes are completely surrounded by halogens, making them unable to hydrogen bond. Hexachloroethane was the most problematic of the chloroalkanes data set, and continues to be problematic here. It could be an incorrect data point, but since chloropentafluoroethane is also a large outlier, it seems likely that that these two solutes simply do not model well. It is of interest that this regression is more statistically significant than the Vol model for just the chloroalkanes. Note that chlorine is more electronegative than bromine or iodine. This might lead to hydrogen bonding being more important in the solvation process for chloroalkanes. The volume, being unable to account for hydrogen bonding, creates better models for halogenated alkanes where hydrogen bonding plays less of a role.

Adding either Sur or Pol or both to the model adds little, as shown by Table 3.10. For all the regressions given in Table 3.10. the top two outliers are again hexachloroethane and chloropentafluoroethane, in that order.

Table 3.10. – Calculated Physical Property Models for all Halogenated Alkanes

R^2	Physical Properties		
0.7698	Vol		
0.7773	Vol	Pol	
0.7753	Vol	Sur	
0.7739	Pol	Sur	
0.7773	Vol	Pol	Sur
0.3515	Pol		

3.4.4.2. Halogenated Alkanes: Topological Descriptors (71 solutes)

The best topological descriptor models follow the same general trends as the best topological descriptor models for the chloroalkanes. The best three-term regression has an R^2 of 0.86, and so is not given here. The best four-term regression is

$$\ln S = 1.69(\pm 0.38) + 0.68(\pm 0.13)*4sC + 0.177(\pm 0.024)*sF + 0.154(\pm 0.0163)*sCl - 1.321(\pm 0.066)*ka1$$

$$n = 71 \quad R^2 = 0.8958 \quad s = 0.760 \quad F = 141.2 \quad Q^2 = 0.869 \qquad (3.17)$$

The kappa shape index is the most heavily weighted on factor one, and correlates highly with volume. This suggests that ka1 is the descriptor that is most closely related to the size of the solute. The other three descriptors are the e-states for the carbons, fluorines, and chlorines. There is one significant outlier, 1,1,2,2-tetrabromoethane. Since the e-state index for bromine was not included, it is logical that the top outlier contains several bromines. This datum was removed and the regression rerun:

$$\ln S = 1.65(\pm 0.37) + 0.89(\pm 0.15)*4sC + 0.219(\pm 0.028)*sF + 0.165(\pm 0.0161)*sCl - 1.319(\pm 0.064)*ka1$$

$$n = 70 \quad R^2 = 0.9022 \quad s = 0.727 \quad F = 150.0 \quad Q^2 = 0.886 \qquad (3.18)$$

There is a small improvement of regression quality when this outlier is removed. Figure 3.D1. in the appendix is a plot of this model.

The best five-term regression simply adds the e-state index for the bromine atoms. Doing so increases the overall statistical significance of the model while raising the t-test values for the other four descriptors. This validates the need for a fifth descriptor in a model made from such a large set of data.

$$\ln S = 1.53(\pm 0.27) + 0.782(\pm 0.092)*4sC + 0.208(\pm 0.017)*sF + 0.208(\pm 0.013)*sCl + 0.213(\pm 0.026)*sBr - 1.417(\pm 0.048)*ka1$$

$$n = 71 \quad R^2 = 0.9488 \quad s = 0.537 \quad F = 240.9 \quad Q^2 = 0.923 \qquad (3.19)$$

There are three outliers beyond 3s, tetrabromomethane, hexachloroethane, and 1,1-difluroethane. Two of these outliers where problems before, most likely due to the unique lack of any hydrogen bonding sites on two of them. 1,1-difluoroethane, which is a newly added solute in this section, is problematic for many of the regressions, and is mostly likely an incorrect data point. If the outliers are removed the improved regression equation is

$$\ln S = 1.36(\pm 0.21) + 0.890(\pm 0.089)*4sC + 0.231(\pm 0.017)*sF + 0.228(\pm 0.010)*sCl + 0.254(\pm 0.022)*sBr - 1.409(\pm 0.036)*ka1$$

$$n = 68 \quad R^2 = 0.9684 \quad s = 0.403 \quad F = 380.1 \quad Q^2 = 0.963 \qquad (3.20)$$

Figure 3.D2. in the appendix is a plot of this model. Eq 3.20 is the best overall topological model. Adding new topological descriptors adds little statistical significance.

Mixing the calculated physical properties and topological descriptors does not provide for any better models than the topological descriptors alone.

3.4.5. Overall Conclusions

The models presented for the bromoalkanes and iodoalkanes are well able to model the solubility. The chloroalkanes are a bit more problematic, but are still able to be modeled accurately. A successful five-parameter model for 71 haloalkanes has been presented.

3.5. Appendix

3.5.1. Appendix A –Chloroalkanes Results

Figure 3.A1. – Solubility modeled by Vol and Sur

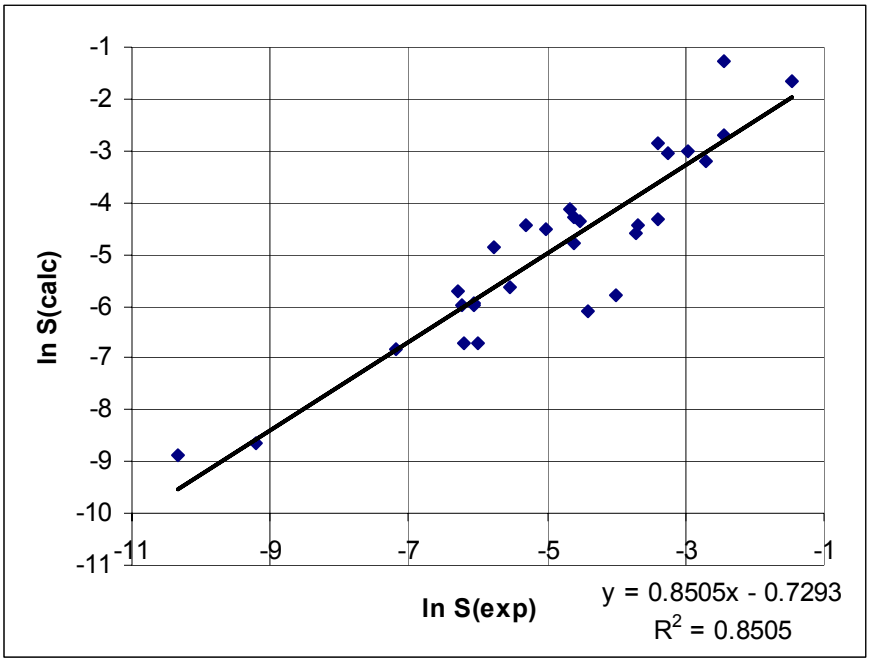


Figure 3.A2. – Solubility modeled by k1 and Q

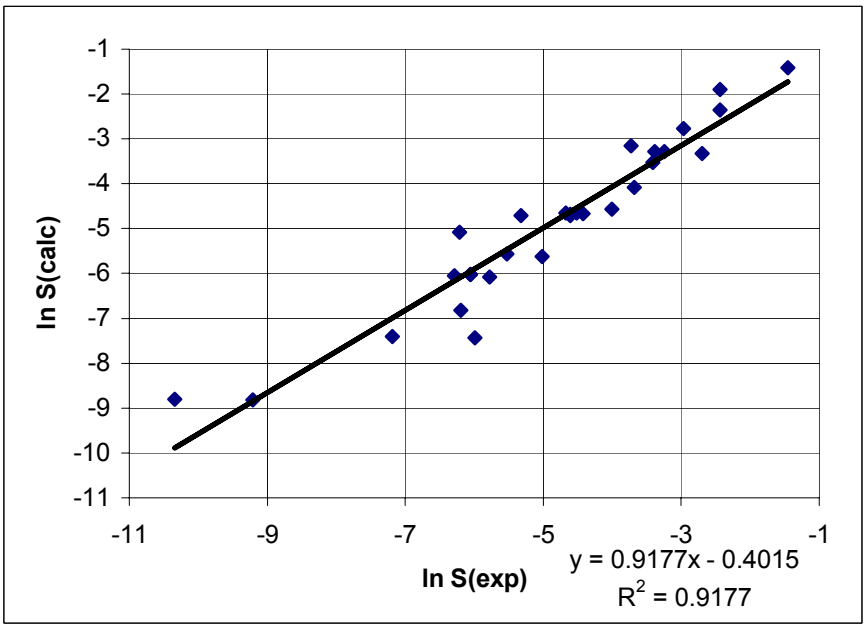


Figure 3.A3. – Solubility modeled by 4sC, sCl, and k1

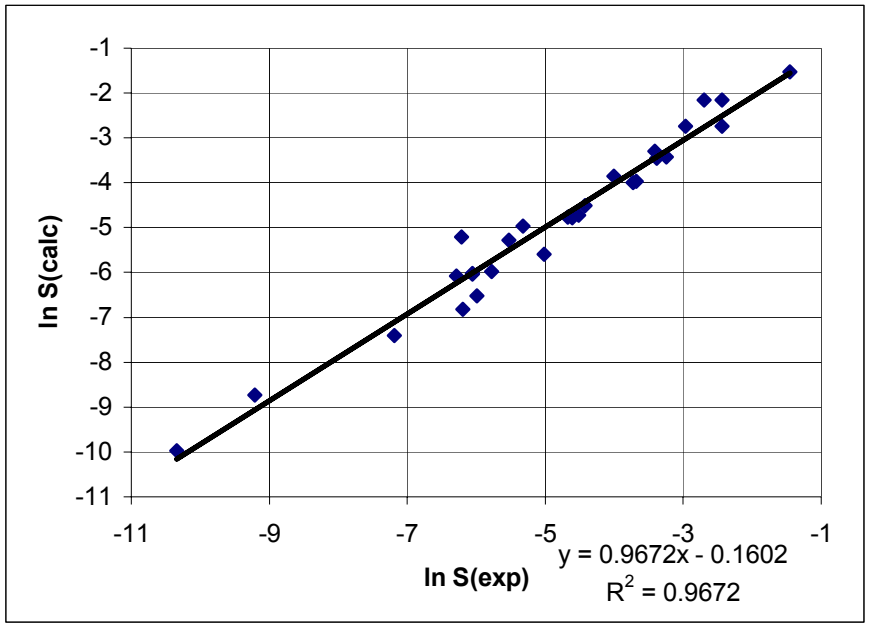


Figure 3.A4. – Solubility modeld by 4sC, sCl, and Pol, with the top outlier removed

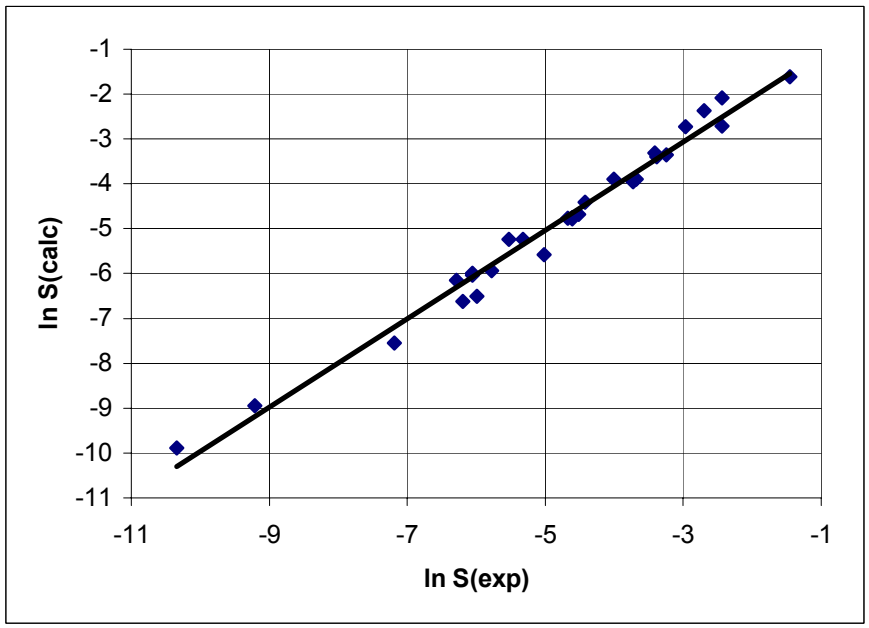


Table 3.A1. – Aqueous Solubilities, Calculated Physical Descriptor values, and Topological Descriptor values for the Chloroalkanes

#	ln S	4sC	sCl	k1	Pol	Sur	Vol	Q
1	-2.44	0	4.99846	3	4.25	90.7915	59.05	0.0623
2	-3.38	0	5.18596	4	5.024	113.068	76.05	0.0615
3	-3.25	0	5.27469	4	5.024	107.785	73.89	0.0749
4	-4.67	0	5.30165	5	5.798	135.436	92.18	0.1030
5	-4.61	0	5.33642	5	5.798	132.008	92.46	0.0656
6	-4.51	0	5.46219	5	5.798	128.552	89.24	0.1090
7	-6.29	0	5.38029	6	6.572	157.165	108.85	0.0950
8	-6.06	0	5.57789	6	6.572	151.043	106.37	0.1257
9	-6.06	0	5.64969	6	6.572	151.234	106.49	0.1300
10	-5.78	0.0138889	5.71759	6	6.572	134.916	94.17	0.0471
11	-7.18	0	5.43727	7	7.346	180.469	124.99	0.1436
12	-9.21	0	5.48046	8	8.12	201.411	141.99	0.1283
13	-1.45	0	9.52778	3	5.404	87.0042	58.29	0.8198
14	-2.97	0	10.0802	4	6.178	100.879	70.05	0.8751
15	-2.44	0	10.108	4	6.178	112.148	75.14	1.5474
16	-3.68	0	10.5347	5	6.952	128.679	89.51	1.0050
17	-3.73	0	10.4344	5	6.952	129.972	90.66	2.4871
18	-5.53	0	10.6866	6	7.726	151.142	105.44	0.8759
19	-6.22	0	10.9614	6	7.726	144.785	103.14	1.6434
20	-6.19	-0.262346	11.358	7	8.5	149.71	108.08	1.0742
21	-2.69	0	14.4167	4	7.332	100.663	70.58	0.0000
22	-4.61	-1.08333	15.1829	5	8.106	107.513	77.65	0.0000
23	-3.41	0	15.2639	5	8.106	128.201	88.82	1.9031
24	-4.42	0	15.8573	6	8.88	145.438	103.74	2.2984
25	-5.32	-1.61111	19.3056	5	9.26	105.531	77.08	0.0000
26	-4.01	0	20.4568	6	10.034	132.276	95.69	2.4521
27	-5.02	-1.27623	20.4313	6	10.034	129.002	89.92	0.7755
28	-5.99	-1.55247	25.652	7	11.962	135.659	100.53	0.1078
29	-10.34	-3.69907	30.8657	8	13.89	139.966	109.78	0.1350

3.5.2. Appendix B: Bromoalkanes Results

Figure 3.B1. - Solubility modeled by Vol with Tetrabromomethane Removed

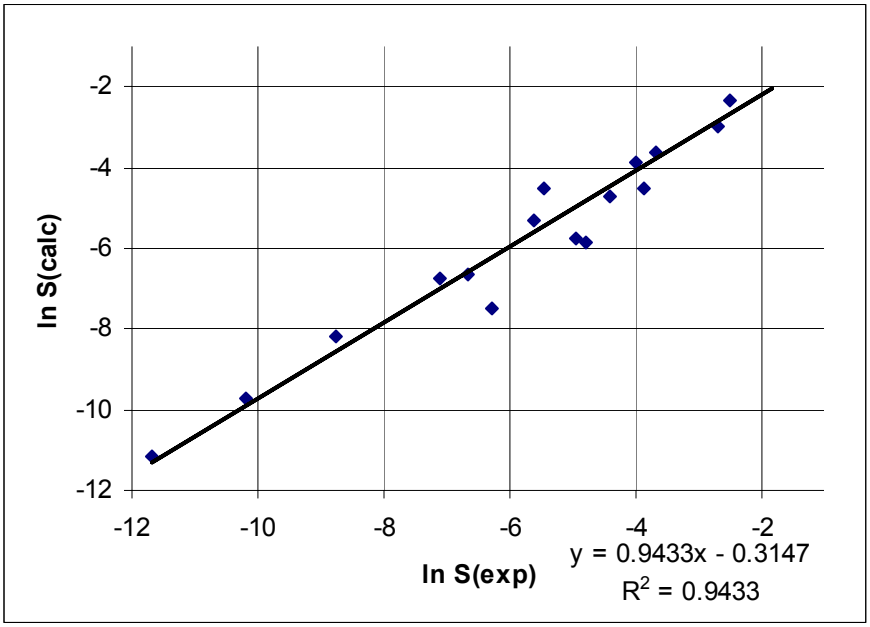


Figure 3.B2. - Solubility modeled by x1 with tetrabromomethane removed

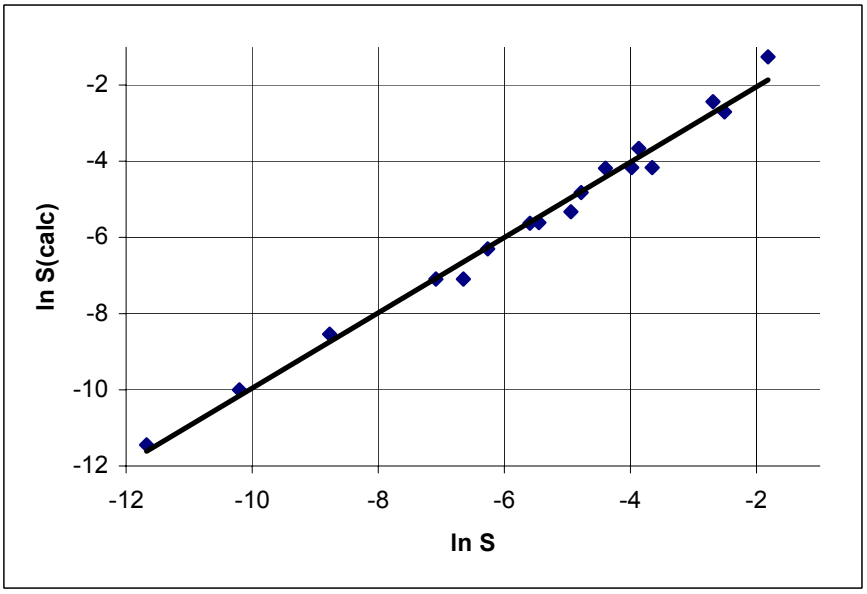


Figure 3.B3. - Solubility modeled by x1, xvc3, and Q

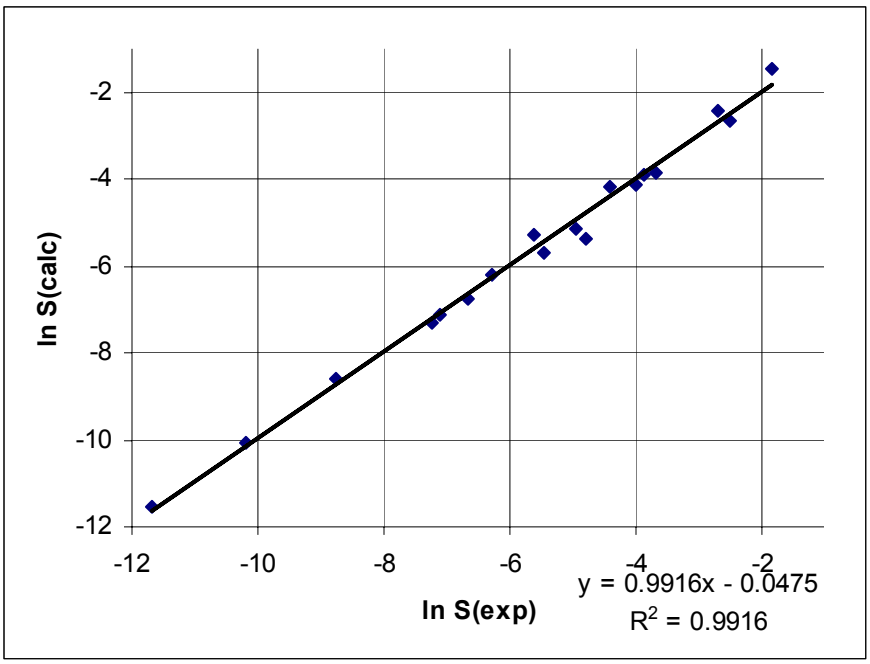


Table 3.B1. – Aqueous Solubilities, Calculated Physical Descriptor Values, and Topological Descriptor Values for Bromoalkanes

#	Solute	ln S	x1	xvc3	Polarizability	Surface	Volume
30	bromomethane	-6.26	2.64	2.57	12.83	159.56	125.62
31	dibromomethane	-3.87	1.91	0.00	7.57	130.10	92.27
32	tribromomethane	-4.95	2.27	0.80	8.35	144.76	105.93
33	tetrebromomethane	-4.79	2.41	0.00	8.35	147.03	107.14
34	bromoethane	-5.60	2.27	0.41	6.50	140.36	100.95
35	1,2-dibromoethane	-6.65	2.77	0.41	7.27	162.75	116.09
36	1,1,2,2-tetrabromoethane	-5.46	2.41	0.00	5.80	135.44	92.18
37	1-bromopropane	-10.20	3.91	0.00	8.82	210.73	150.53
38	2-bromopropane	-8.77	3.41	0.00	8.04	189.62	133.70
39	1,2-dibromopropane	-11.67	4.41	0.00	9.59	234.54	166.80
40	1,3-dibromopropane	-7.09	2.91	0.00	7.27	166.00	117.44
41	1-bromobutane	-3.98	1.91	0.00	5.72	122.24	84.67
42	1-bromopentane	-3.66	1.73	1.13	5.72	115.10	81.79
43	1-bomo-2-methylbutane	-2.51	1.41	0.00	4.95	99.80	67.48
44	1-bromo-3-methylbutane	-1.82	1.00	0.00	4.17	77.47	52.09
45	1-bromohexane	-2.69	1.41	0.00	6.80	104.51	74.88
46	1-bromoheptane	-7.23	2.00	15.15	12.05	130.20	104.15
47	1-bromooctane	-4.40	1.73	4.37	9.43	124.20	94.28

3.5.3. Appendix C – Iodoalkanes Results

Solubility modeled by Volume

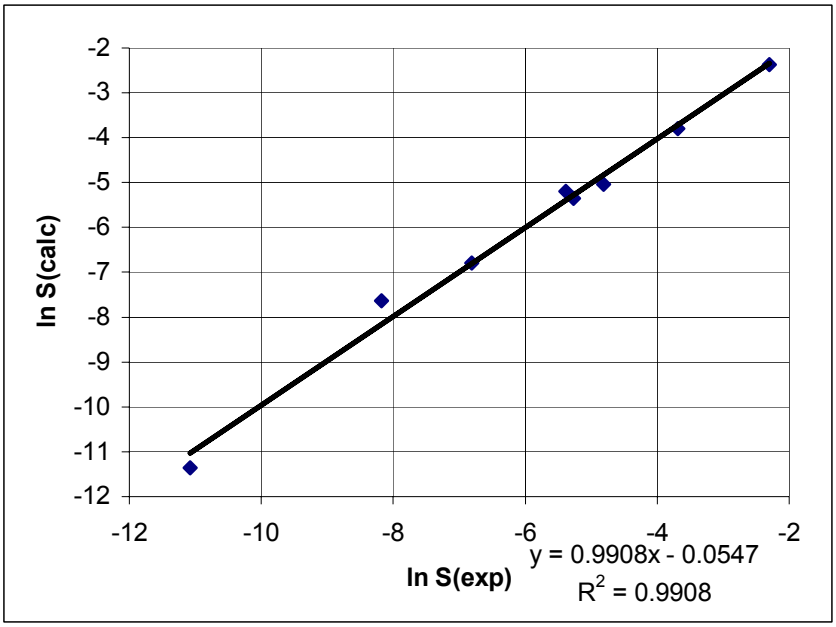


Table of Solubilities and Volume Values for Iodoalkanes

#	Molecules	ln S	Volume
48	iodomethane	-2.3	60.1852
49	diiodomethane	-5.39	91.0554
50	triiodomethane	-8.17	117.738
51	iodoethane	-3.68	75.7207
52	1-iodopropane	-5.27	92.8167
53	2-iodopropane	-4.81	89.3479
54	1-iodobutane	-6.82	108.496
55	1-iodoheptane	-11.08	158.453

3.5.4. Appendix D

Figure 3.D1. – Solubility modeled by 4sC, sF, sCl, and ka1 with top outlier removed

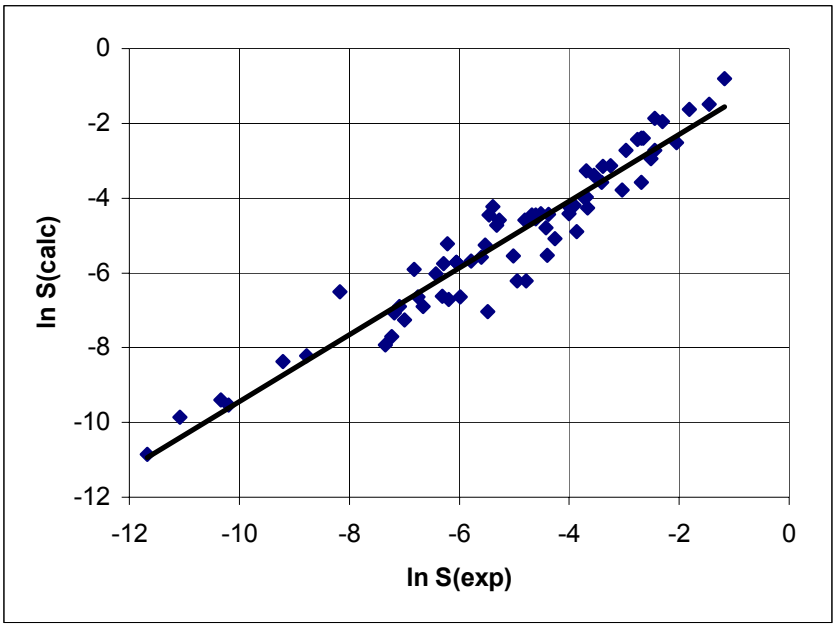


Figure 3.D2. – Solubility modeled by 4sC, sF, sCl, ka1 and sBr with top three outliers removed

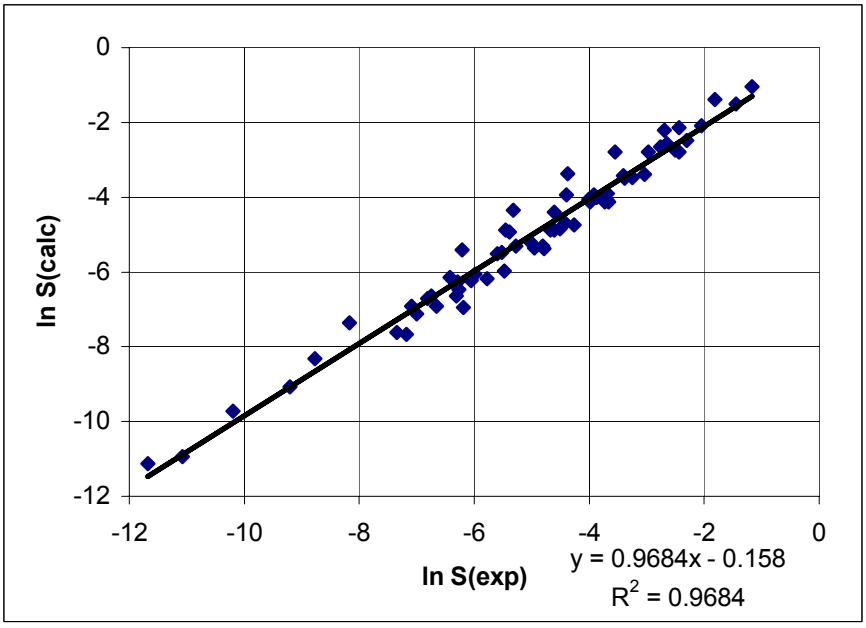


Table 3.D1. All Halogenated Alkanes

#	Solute	ln S	#	Solute	ln S
				Bromoalkanes	
	Chloroalkanes		30	bromomethane	-1.82
1	dichloromethane	-1.45	31	dibromomethane	-2.69
2	trichloromethane	-2.69	32	tribromomethane	-4.4
3	tetrachloromethane	-3.35	33	tetrebromomethane	-7.23
4	chloroethane	-2.44	34	bromoethane	-2.51
5	1,1-dichloroethane	-2.97	35	1,2-dibromoethane	-3.87
6	1,2-dichloroethane	-2.44	36	1,1,2,2-tetrabromoethane	-6.26
7	1,1,1-thrichloroethane	-4.61	37	1-bromopropane	-3.98
8	1,1,2-trichloroethane	-3.41	38	2-bromopropane	-3.66
9	1,1,2,2-tetrachloroethane	-4.01	39	1,2-dibromopropane	-4.95
10	1,1,1,2-tetrachloroethane	-5.02	40	1,3-dibromopropane	-4.79
11	pentachloroethane	-5.99	41	1-bromobutane	-5.46
12	hexachloroethane	-10.34	42	1-bromopentane	-7.09
13	1-chloropropane	-3.38	43	1-bomo-2-methylbutane	-5.6
14	2-chloropropane	-3.25	44	1-bromo-3-methylbutane	-6.65
15	1,2-dichloropropane	-3.68	45	1-bromohexane	-8.77
16	1,3-dichloropropane	-3.73	46	1-bromoheptane	-10.2
17	1,2,3-trichloropropane	-4.42	47	1-bromooctane	-11.67
18	1-chlorobutane	-4.67		Iodoalkanes	
19	2-chlorobutane	-4.51	48	iodomethane	-2.3
20	1,1-dichlorobutane	-5.53	49	diiodomethane	-5.39
21	2,3-dichlorobutane	-6.22	50	triiodomethane	-8.17
22	1-chloro-2-methylpropane	-4.61	51	iodoethane	-3.68
23	1-chloropentane	-6.29	52	1-iodopropane	-5.27
24	2-chloropentane	-6.06	53	2-iodopropane	-4.81
25	3-chloropentane	-6.06	54	1-iodobutane	-6.82
26	2-chloro-2-methylbutane	-5.78	55	1-iodoheptane	-11.08
27	2,3-dichloro-2-methylbutane	-6.19		Mixed Halogen Alkanes	
28	1-chlorohexane	-7.18	56	bromochloromethane	-2.05
29	1-chloroheptane	-9.21	57	bromodichloromethane	-3.55
			58	chlorodibromomethane	-4.37
			59	chloropentafluoroethane	-6.42
			60	1,1-difluoroethane	-1.31
			61	1,1,2,2-tetrachlorodifluoroethane	-7.35
			62	1,1,2-trichlorotrifluoroethane	-7.00
			63	1,1-dichlorotetrafluoroethane	-6.74
			64	1-chloro-1,1-difluoroethane	-2.76
			65	1-chloro-2-fluoroethane	-1.17
			66	2-bromo-2-chloro-1,1,1-trifluoroethane	-3.91
			67	1-chloro-1,1,1-trifluoroethane	-2.65
			68	1,2-dichlorotetrafluoroethane	-6.31
			69	1-bromo-2chloroethane	-3.04
			70	1-bromo-3-chloropropane	-4.26
			71	1,2-dibromo-3-chloropropane	-5.48

3.6. References

[1] Yalkowsky, S. H.; Orr, R.J.; Calcani, S. C., Solubility and Partitoning 3. The Solubility of Halobenzenes in Water. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 351-353.

[2] Dunnicant, F. M.; Elzerman, A. W.; Just, P. C.; Hasan, M. N., Quantitative Structure-Property Relationships for Aqueous Solubilities and Henry's Law Constants of Polychlorinated Biphenyls. *Environ. Sci. Technol.* **1992**, *79*, 2239-2246.

[3] Huibers, P. D. T.; Katritzky, A. R., Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 183-292.

[4] Ruelle, P; Kesselring, U. W., Aqueous Solubility Predictions of Environmentally Important Chemicals *Chemosphere*, **1997,** *34*, 275-298.

[5] QsarIS ver. 1.2 **2001,** SciVision Inc., 200 Wheeler Road, Burlington, MA. 01803.

[6] StatMost 3.0 for Windows ©**1995,** DataMost Corporation Los Angeles, California.

[7] Kier, L. B.; Hall, L.H., *Molecular Structure Descriptions: The Electrotopological State*, Academic Press: New York, **1995**, p.65ff.

[8] QsarIS Molecular Descriptor Help Menu, QsarIS ver. 1.2 **2001,** SciVision Inc., 200 Wheeler Road, Burlington, MA. 01803.

Chapter 4
Solubilities of Gases in Water

4.1. Abstract

The solubilities of 45 gases in water were modeled using one- or two-term regressions. The models use only experimentally determined physical properties of the solutes. 20 different physical properties are reviewed for their ability to model solubility. No single property can effectively model the solubility ($R^2 > 0.5$). The boiling point and polarizability together provide the best model, with $R^2 = 0.86$.

4.2. Introduction

The modeling of gases in water has traditionally been done using properties of the solute. Abraham et. al.[1] has done extensive work in this field. In 1994 Abraham created a model for 408 gases and vapor in water using a five-term regression with an R^2 = 0.9976. The terms include a hydrogen bond accepting term, a hydrogen bond donating term, a volume term, polarizability, and the excess molar refraction. Almost all of the solutes measured in this study are considered to be in the vapor phase. Modeling the solubility becomes much more tricky when only using solutes in the gas phase at 298 K.

The solubilities of gases in solution can in principle be broken down into two energetic steps: first, formation of a cavity in the interior of the liquid, and second, placing the solute into the cavity and its interaction with the surrounding solvent. The energy needed to form the cavity is directly proportional to the size of the solute molecule. In practice the boiling point and polarizability are both closely related to the "size" of the solute molecule.

In most solvents the energy given back by solute-solvent interaction is due mostly to dispersion forces. Water is a unique solvent. The extensive hydrogen bonding in water makes it difficult to model the solubilities of gases using a small number of physical properties. The solutes reviewed are all gases at STP. This creates additional problems in modeling, since the gas molecules examined can be very different from one another. The great variety of gases included and the very large range of gas solubilities in water present special challenges.

4.3. Methods

For this paper a gas is defined as a substance that has a normal boiling point at 298.15 ˚K. The gas solubility data were obtained from the *IUPAC Solubility Series*[2]. All solubility data were as ln X_2, where X_2 is the mole fraction solubility at 1 atm partial pressure of gas. The boiling points (BPs) are in kelvins at one atm, and come from Fogg and Gerrard.[3] The polarizability (Pol) data and the dipole moment (DM) data were taken from the *CRC 2001*.[4] The polarizabilities have units of Å^3, and the DMs have units of Deybes. The programs QsarIS[5] and StatMost[6] were used to create the regressions.

4.4. Results

The solubilities of the gases in water were modeled using QsarIS. The goal was to model the solubilities of all possible gases in water, or subsets of the gases in water, using one- two- or three-term regressions. QsarIS has the employs a genetic algorithm to quickly identify which combinations of a large number of properties can successfully model a given dependant variable. Table 1 lists the physical properties of the gases screened. Most of the experimental properties were not useful in modeling the solubility of the 45 gases in water. All of the gas physical properties (except Van Der Waals area, Van Der Waals volume, and Radius of Gyration) are experimentally determined and at 298.15 K where appropriate.

Table 4.1. - Physical Properties of the Gases Examined

Molecular Mass	Critical Pressure (P_C)
Boiling Point (BP)	Critical Volume (V_C)
Polarizability (Pol)	(Critical Volume)$^{2/3}$ ($V_C^{2/3}$)
Dipole Moment (DM)	Critical Compressibility Factor
Enthalpy of Fusion	Critical Density
Solubility Parameter	Acentric Factor
Enthalpy of Vaporization	VanDer Waals Area
Heat Capacity (C_P)	VanDer Waals Volume
Absolute Entropy	Gibbs Energy of Formation
Refractive Index	Viscosity
Critical Temperature	Radius of Gyration

Only three physical properties of the gases had any significant ability to model the aqueous solubility as part of a 1- or 2-term regression, and this paper will deal with BP, Pol, and DM. Table 2 shows a correlation matrix for them.

Table 4.2. - Correlation Matrix of BP, Pol, and DM[a]

	BP	Pol	DM
BP	1	0.84 (35)	0.45 (45)
Pol		1	0.085 (35)

[a] The number of data points is in parenthesis

It is clear that BP and Pol are well correlated, but dipole moment is not strongly correlated with either.

4.4.1. Taking all Possible Gases in Water

No single property regression adequately models the solubilities of the full data set. The best single-property regression uses DM. It is important to note that only 18 of the 45 gases in the regression have nonzero DM s. BP and Pol do not function well as single-property descriptors. Figure 4.A1 in the appendix is the graph for the DM regression. The regression equations are:

$$\ln X_2 = 2.85(\pm 0.39)*DM - 9.96(\pm 0.32) \qquad (4.1)$$

$$N = 45 \quad R^2 = 0.496 \quad s = 1.75 \quad F = 42.4 \quad Q^2 = 0.439$$

$$\ln X_2 = .018(\pm 0.0047)*BP - 12.39(\pm 0.96) \qquad (4.2)$$

$$N = 45 \quad R^2 = 0.281 \quad s = 2.1 \quad F = 17.3 \quad Q^2 = 0.225$$

$$\ln X_2 = 0.03(\pm 0.2)*\text{Pol} - 9.16(\pm 0.94) \quad (4.3)$$

$$N = 35 \quad R^2 = 0.00 \quad s = 2.56 \quad F = 0.02 \quad Q^2 = 0.00$$

Since DM is 0 for most of the gases, it is only sensible to use it as a second descriptor.

The two-property models improve significantly over the single property models. The best model presented in this work uses BP and Pol. It is important to note that a single-property regression of BP using only the 35 common solubility data points with Pol has an $R^2 = 0.28$. Figure 4.A2 in the appendix gives the BP - Pol regression plot. Figure 4.A3 in the appendix shows the BP - DM regression graph. The regression equations for the best two-property models are:

$$\ln X_2 = -1.36(\pm 0.12)*\text{Pol} + 0.055(\pm 0.004)*\text{BP} - 13.20(\pm 0.45) \quad (4.4)$$

$$N = 35 \quad R^2 = 0.858 \quad s = 0.94 \quad F = 96.7 \quad Q^2 = 0.82$$

$$\ln X_2 = 2.27(\pm 0.44)*\text{DM} + 0.012(\pm 0.0029)*\text{BP} - 11.87(\pm 0.72) \quad (4.5)$$

$$N = 41 \quad R^2 = 0.634 \quad s = 1.56 \quad F = 31.2 \quad Q^2 = 0.57$$

Obviously, there is a synergistic effect between BP and Pol.[7]

4.4.2. Taking subsets of the Gases in Water

Since the models of solubility in water do not work well for the full data set, smaller subsets were modeled. The goal was to find a 1- or 2-property model that exceeds the ability of Pol and BP to model the solubility of the full data set

Noble Gases in Water

The noble gases in water modeled well using only a single property. BP works the best ($R^2 = 0.995$). Pol as the single-property has an R^2 of 0.972. BP and Pol together

does not provide a significantly better model. Figure 4.A4 in the appendix contains the BP regression graph. The regression equation for BP is:

$$\ln X_2 = 0.0158(\pm 0.0006)*BP - 12.01(\pm 0.07) \quad (4.6)$$

$$N = 6 \quad R^2 = 0.995 \quad s = 0.10 \quad F = 750 \quad Q^2 = 0.988$$

Gases composed of three atoms or less

This subset of "simpler" gases models well with either one- or two-property regressions. This group excludes the noble gases, as they have already been shown to model well. Figure 4.A5 in the appendix contains the BP and Pol regressions. The regression equations are:

$$\ln X_2 = -1.22(\pm 0.23)*Pol + 0.049(\pm 0.0042)*BP - 12.67(\pm 0.29) \quad (4.7)$$

$$N = 13 \quad R^2 = 0.964 \quad s = 0.51 \quad F = 133.0 \quad Q^2 = 0.926$$

$$\ln X_2 = 0.0268(\pm 0.0035)*BP - 13.04(\pm 0.53) \quad (4.8)$$

$$N = 13 \quad R^2 = 0.859 \quad s = 0.95 \quad F = 67.7 \quad Q^2 = 0.763$$

4.4.3. Groups of Gases that are not well-modeled

Equally important in this work are groups that do not model well. There are many groups of gases which do not model well. Listed below is a table of groups which one would expect to model well, but do not. Several are listed in this section, with the best 2-property regression R^2 value obtain from any combination of the 3 physical properties. These groups are presented in Table 4.3.

Table 4.3. - Other Groupings Attempted.

Group	N	Best Model	R^2
Alkanes	4	BP	0.1
Alkanes,Alkenes,Alkynes	11	DM + Pol	0.14
Gases with DM	18	BP + DM	0.64
Gases without DM	20	BP + Pol	0.65

4.5. Appendix

Figure 4.A1.

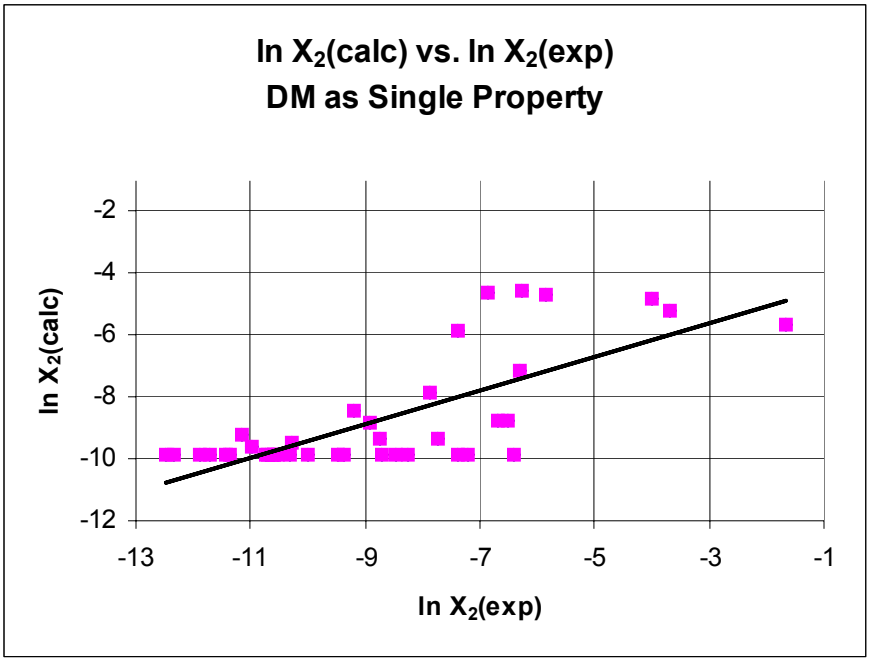


Figure 4.A2.

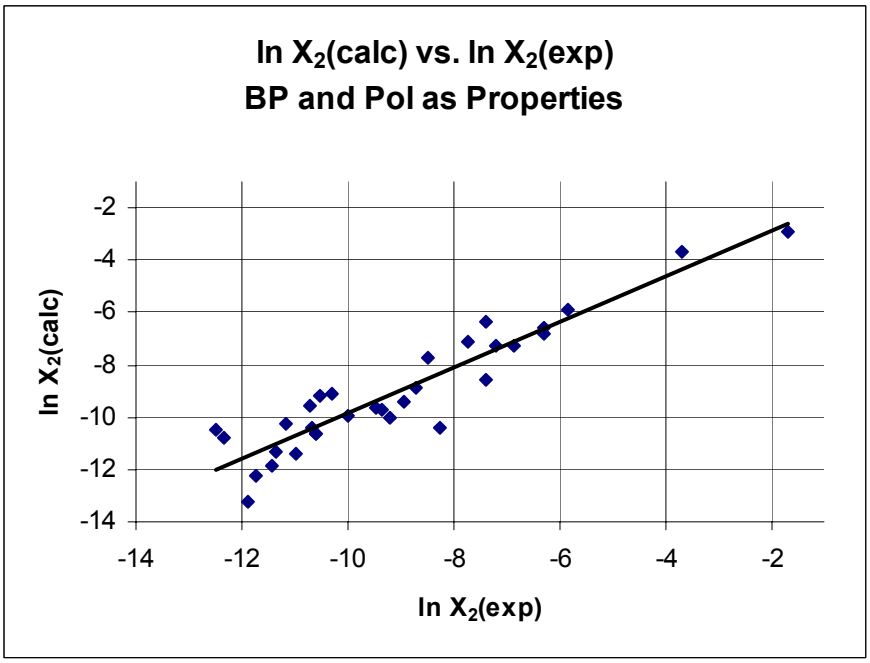


Figure 4.A3.

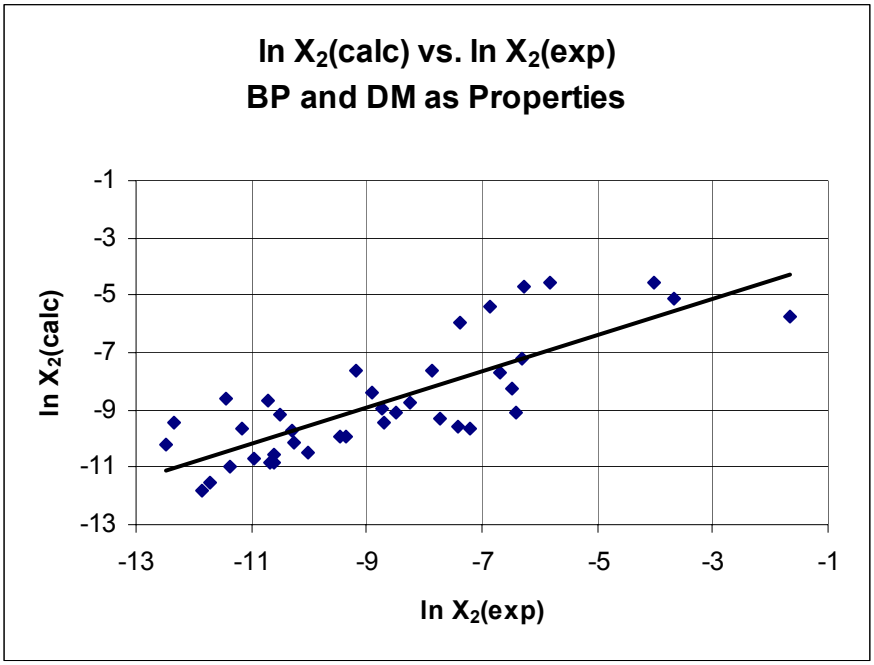


Figure 4.A4.

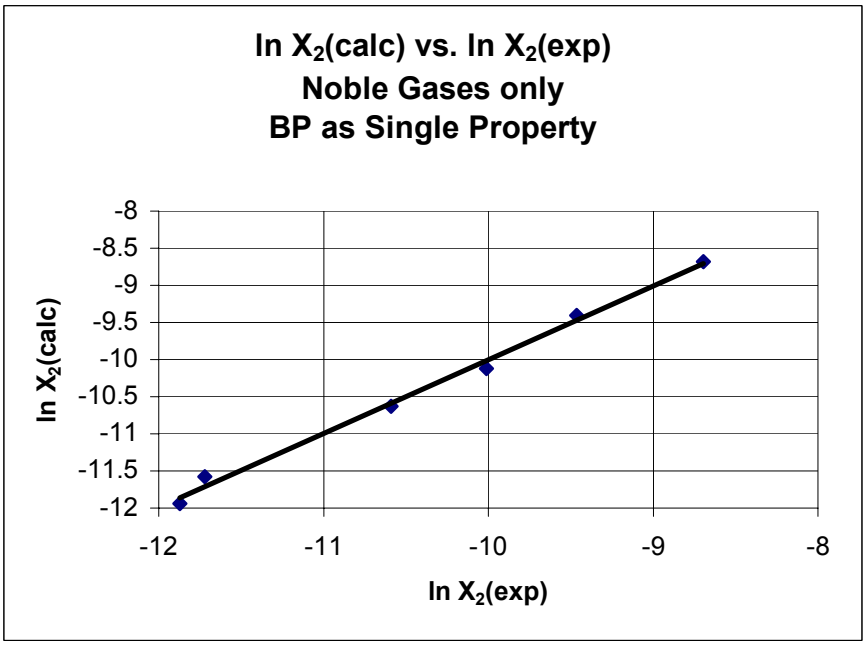


Figure 4.A5.

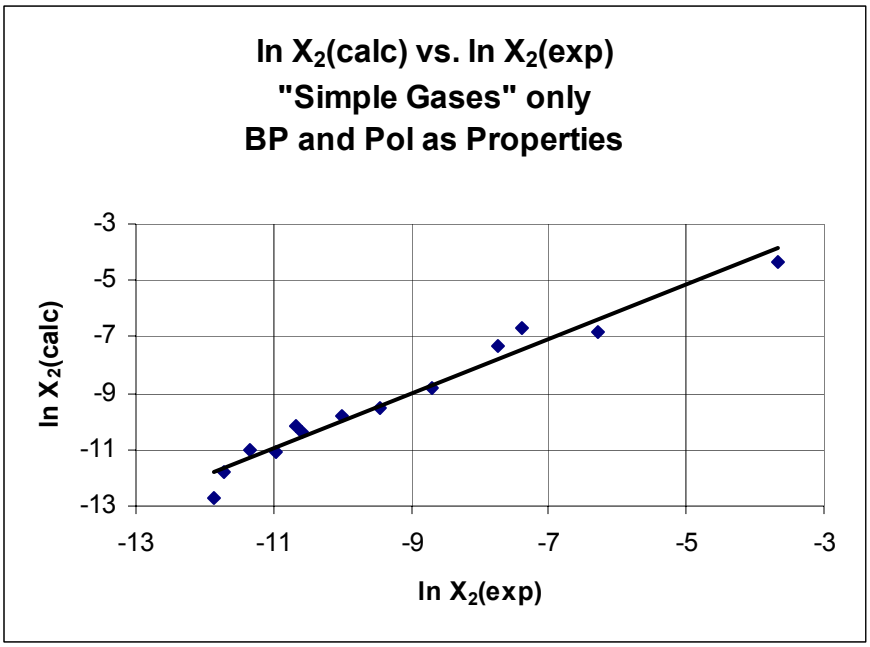

ln X2(calc) vs. ln X2(exp)
"Simple Gases" only
BP and Pol as Properties
ln X2(calc)
ln X2(exp)
-3
-5
-7
-9
-11
-13
-13
-11
-9
-7
-5
-3

Table 4.A1. - All Relevant Values

Gases	ln X_2	BP/(T/K)	Pol/ $Å^3$	DM/ Deybe
NH3	-1.6734	239.80	2.26	1.47
SO2	-3.6773	263.15	3.72	1.63
ClO2	-4.0041	283.05		1.78
CH3Br	-5.8331	276.61	5.78	1.81
CH3Cl	-6.2775	248.93	5.35	1.87
H2S	-6.2920	212.88	3.78	0.97
Cl2	-6.4082	239.04		0.00
CH2F2	-6.4546	221.5500		1.96
H2Se	-6.4950	231.15		0.40
CH2=CHC=-CH	-6.6983	278.30		0.40
CH3F	-6.8476	194.80	3.54	1.85
C2H2	-7.1982	189.15	3.33	0.00
CHClF2	-7.3798	232.35	5.91	1.42
CO2	-7.4002	194.67	2.91	0.00
N2O	-7.7326	184.20	3.03	0.18
COS	-7.8662	222.95		0.71
CH2=CHCH=CH2	-8.2544	268.74	8.64	0.00
CHF3	-8.3556	191.0500	3.54	1.85
c-C3H6	-8.4809	240.29	5.66	0.00
CFCl3	-8.6798	296.9200	9.47	0.45
Rn	-8.6945	211.15	5.30	0.00
AsH3	-8.7316	210.67		0.20
C3H6	-8.9207	225.45	6.26	0.37
CH2=C(CH3)2	-9.1818	266.25	8.29	0.50
C2H4	-9.3634	169.45	4.25	0.00
Xe	-9.4638	165.15	4.04	0.00
CCl2F2	-9.9650	243.3600	7.93	0.51
Kr	-10.0102	119.90	2.48	0.00
NO	-10.2668	121.45		0.15
C2H6	-10.3055	184.55	4.43	0.00
CF2=CF2	-10.4646	197.15		0.00
C3H8	-10.5182	231.08	6.29	0.00
Ar	-10.5903	87.29	1.64	0.00
CH4	-10.5938	111.54	2.59	0.00
O2	-10.6809	90.18	1.58	0.00
C4H10	-10.7258	272.65	8.20	0.00
CO	-10.9683	81.65	1.95	0.11
NF3	-11.1490	144.12	3.62	0.24
H2	-11.1672	20.37	0.80	0.00
N2	-11.3534	77.34	1.74	0.00
C(CH3)4-neopen	-11.4387	282.65	10.20	0.00
Ne	-11.7196	27.25	0.39	0.00
He	-11.8720	4.23	0.20	0.00
C3F6	-12.1482	244.15		0.00
SF6	-12.3353	209.25	6.54	0.00
CF4	-12.4755	145.15	3.84	0.00

4.6. References

[1] Abraham, Michael H.; Andonian-Haftvan, Jenik; Whiting, Gary S.; Leo, Albert; Taft, Robert S., Hydrogen Bonding. Part 34. The Factors that Influence the Solubility of Gases and Vapors in Water at 298 K, and a New Method for its Determination. *Jol. Chem. Soc*. Journal of the Chemical Society, **1994**, *8*, 1777-1788.

[2] IUPAC Solubility Data Series gas solubility volumes

[3] Fogg, P. G. T.; Gerrard, W., *Solubility of gases in Liquids* John Wiley & Sons Ltd.: Barrins Lane, Chichester West Sussex, England, **1991.**

[4] Lide, D.R. (ed.) *CRC handbook of Chemistry and Physics 2002-2003* CRC Press **2003**

[5] QsarIS ver. 1.2 **2001,** SciVision Inc., 200 Wheeler Road, Burlington, MA. 01803.

[6] StatMost 3.0 for Windows ©**1995,** DataMost Corporation Los Angeles, California.

[7] Peterangelo, Stephen C.; Seybold, Paul G., Synergistic interactions among QSAR descriptors., *Int. J. Quantum. Chem.*, **2003**, *96*, 1-9.

Chapter 5
Solubilities of Gases in Alkanes

5.1. Abstract The solubilities of gases in eight alkane solvents were modeled using single experimental physical properties of the gases. The properties used were (1) normal boiling point, (2) polarizability, (3) critical volume, and (4) (critical volume)$^{2/3}$. Boiling point provided the best model in every case. Polarizability provided the second best model. The boiling point models had an average $R^2 = 0.961$.

5.2. Introduction

The solubilities of gases in liquids can be broken down into two energetic steps: first, an energy-demanding formation of a cavity for the solute in the liquid, and second, placement of the solute in the cavity and its interaction with the surrounding solvent. The energy needed to form the cavity should be related to the size of the solute molecule. The solute-solvent interaction energy is expected to be related to the solute's surface area and characteristics. The critical volume (V_C) is one measure of solute volume, and $V_C^{2/3}$ is often taken as a measure proportional to the surface area of the solute molecule. The solute polarizability (Pol) is related to the numbers and kinds of atoms in the solute gas.

The energy given back by solute-solvent interaction is usually due mostly to dispersion forces. Both the gas polarizability and its normal boiling point are related to its internal cohesive forces and thus may be related to the gas's interaction energy with the solvent.

5.3. Methods

All experimental solubility data used in this paper were obtained from the *IUPAC Data Solubility Series*[1], and are in the form of ln X_2, where X_2 is the mole fraction solubility. The software used to perform the regressions was StatMost.[2] Table 5.1 lists the physical properties used, their units, and their sources. All the physical properties for the gases used in this study were experimentally determined values. A table of the specific values can be found in the appendix to this chapter.

Table 5.1. - Physical Properties Used in this Study

Name	Abbreviation	Units	Source
Normal Boiling Point	BP	K at 1 bar	Fogg and Gerrard[3]
Polarizability	Pol	$Å^3$	CRC[4]
Critical Volume$^{2/3}$	$V_C^{2/3}$	$(cm^3/mol)^{2/3}$	CRC[4]
Critical Volume	V_C	cm^3/mol	CRC[4]

5.4. Results

Table 2 gives a correlation matrix of the four physical descriptors used. All correlations have 22 points of comparison. Table 5.2 shows the principal component factor analysis (PCA) results for the physical properties.

Table 5.2. - Correlation Matrix of Physical Property and Factor Analysis

	BP	V_C	$Vc^{2/3}$	Pol
BP	1	0.826	0.8495	0.9172
V_C		1	0.9965	0.9720
$V_C^{2/3}$			1	0.9789
Pol				1

Percent Factor Weight

	Factor 1	Factor 2	Factor 3	Factor 4
Factor %	94.33%	5.35%	0.26%	0.06%
Total %	94.33%	99.68%	99.94%	100.00%

Percent Factor loadings

	Factor 1	Factor 2	Factor 3	Factor 4
BP	85.17%	14.75%	0.08%	0.00%
Pol	99.22%	0.00%	0.77%	0.00%
$Vc^{2/3}$	97.20%	2.54%	0.13%	0.12%
Vc	95.71%	4.10%	0.07%	0.12%

The factor analysis indicates that BP has slightly weaker correlations with the other experimental properties. The other properties all correlate above 95%. The small variance in the data accounts for why BP is the best property to model with. In particular, only BP has any sizeable amount of its variance due to Factor 2.

The results are given in the form of a table for each solvent alkane. The regressions have been given in order of descending R^2 values. Each regression has the general form of

$$\ln X_2 = -A + B * \text{physical property}$$

Key to tables:

A	Intercept in regression ± uncertainty of intercept
t(A)	t-Test value for A
B	Coefficient of the physical property constant ± uncertainty
t(B)	t-test value for B
s	Standard Error of Regression
N	number of solubility data points used in regression
F	F test value

Table 5.3. n-Pentane Single Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	16	0.967	8.38 ± 0.21	39.9	0.0259 ± 0.0013	20.1	405.1	0.401	0.961
Pol	16	0.915	7.47 ± 0.28	26.8	0.841 ± 0.0686	12.3	150.2	0.641	0.888
$Vc^{2/3}$	16	0.807	10.01 ± 0.74	13.6	0.229 ± .030	7.7	58.8	0.963	0.765
V_C	16	0.775	8.02 ± 0.53	15	0.029 ± 0.0040	7.12	51.7	1.01	0.717

The general trend present for all the gases in alkanes is evident in this set of regressions. BP always produces the best single physical property model. There are no major outliers.

Table 5.4. n-Hexane Single Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	20	0.960	8.37 ± 0.20	41.4	0.0257 ± 0.0012	20.9	435.9	0.427	0.952
Pol	20	0.928	7.61 ± 0.23	32.6	0.881 ± 0.058	15.2	230.4	0.577	0.905
$Vc^{2/3}$	20	0.816	10.38 ± 0.67	15.5	0.245 ± .027	9	80.1	0.919	0.774
V_C	20	0.775	8.24 ± 0.51	16.2	0.031 ± 0.0039	7.9	62	1.02	0.718

A graph for BP and Pol can be found in the appendix.

Table 5.5. n-Heptane Single Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	18	0.966	8.41 ± 0.20	42.9	0.02620 ± 0.0012	21.4	458.2	0.413	0.957
Pol	18	0.929	7.65 ± 0.24	31.3	0.889 ± .062	14.5	209.4	0.599	0.906
$Vc^{2/3}$	18	0.827	10.46 ± 0.68	15.3	0.245 ± .027	8.8	76.9	0.932	0.789
V_C	18	0.791	8.33 ± 0.52	16.1	0.31 ± 0.0039	7.8	60.6	1.027	0.739

Table 5.6. n-Octane Single Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	18	0.966	8.43 ± 0.20	42.7	0.0265 ± 0.0012	21.4	458.4	0.417	0.957
Pol	18	0.931	7.67 ± 0.24	31.6	0.899 ± 0.061	14.7	216.3	0.595	0.909
$Vc^{2/3}$	18	0.831	10.49 ± 0.68	15.4	0.248 ± .028	8.9	78.8	0.932	0.793
V_C	18	0.796	8.36 ± 0.52	16.2	0.0311 ± 0.0039	7.9	62.3	1.02	0.745

Table 5.7. n-Nonane Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	18	0.962	8.45 ± 0.21	40.1	0.0267 ± 0.0013	20.2	409.8	0.444	0.952
Pol	18	0.931	7.69 ± 0.25	31.4	0.907 ± 0.062	14.7	217	0.601	0.909
$Vc^{2/3}$	18	0.832	10.54 ± 0.69	15.3	0.251 ± .028	8.9	79.3	0.939	0.764
V_C	18	0.796	11.22 ± 0.72	15.6	0.276 ± 0.032	8.7	62.9	1.03	0.745

Table 5.8. n-Decane Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	18	0.966	8.56 ± 0.29	42	0.0270 ± 0.0013	21.4	458.8	0.425	0.957
Pol	18	0.931	7.86 ± 0.38	31.1	0.915 ± 0.062	14.7	217	0.608	0.908
$VC^{2/3}$	18	0.830	11.0 ± 1.0	15.2	0.252 ± .029	8.9	78.9	0.95	0.793
V_C	18	0.795	8.40 ± 0.53	15.9	0.032 ± 0.0040	7.9	62	1.05	0.743

Table 5.9. n-Dodecane Single Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	17	0.951	8.44 ± 0.24	35.9	0.02649 ± 0.0016	17.1	291.6	0.485	0.936
Pol	17	0.925	7.85 ± 0.26	30.5	0.978 ± .072	13.6	185.7	0.599	0.886
$Vc^{2/3}$	17	0.804	10.99 ± 0.80	13.7	0.273 ± .035	7.8	61.4	0.971	0.751
V_C	17	0.765	8.71 ± 0.60	14.6	0.035 ± 0.0050	7	48.8	1.06	0.682

Table 5.10. n-Hexadecane Single Property Regression Details

Property	N	R^2	A	t(A)	B	t(B)	F	s	Q^2
BP	15	0.945	7.91 ± 0.25	33.1	0.0262 ± 0.0018	14.9	221.2	0.526	0.922
Pol	15	0.933	7.86 ± 0.38	31.9	0.968 ± 0.072	13.4	180.5	0.579	0.884
$Vc^{2/3}$	15	0.854	11.22 ± 0.72	15.6	0.27 ± .032	8.6	76.2	0.853	0.802
V_C	15	0.818	11.22 ± 0.72	15.6	0.276 ± 0.032	8.7	58.6	0.952	0.731

5.5. Discussion

The process of solvation is usually thought of in three steps. The normal first step, in which a solute molecule is separated from its bulk condition, can be ignored here. The solutes are all gases, and so the solute-solute attractive forces are correspondingly already weak. Step two, where a hole is created in the solvent, should require less energy than in the previous chapters, since the alkanes exhibit none of the hydrogen bonding found in water. Step three, in where the solute and solvent interact, should be dependant entirely on dispersion forces. Each of the four physical properties can be related to one or both of the two remaining energetic steps in the solvation of a gas in a liquid. The normal boiling point is indicative of the energy needed to change the state of matter of the gas molecules, and is a loose measure of the intermolecular forces. This energy is generally related to the total size of the gas molecule (dispersion forces) and possible other bonding forces due to charges on the individual gas atoms of the gases. The polarizability is essentially the distortability of the electron clouds in the gas molecule. The polarizability is proportional to the number and types of constituent atoms in the gas molecule, and should be related to the energy released through interaction with the solvent. $V_C^{2/3}$ is a rough indicator of the surface area of the molecule. This property might therefore be related to the cavity-forming step in solvation.

The correlation matrix given in the results shows a fair amount of correlation (> 0.83) for all 4 physical properties. This is to be expected, as all four properties are somewhat related to the size of the gas molecule. $V_C^{2/3}$, and Vc correlate very highly (> 0.99). Since gas molecules are small and fairly rigid, the surface area and volume should be highly correlated. Boiling point and polarizability are highly correlated (0.91). Since

both are related to dispersion forces, this is logical. It is interesting to note that $V_C^{2/3}$ work better than V_C.

It is evident that a single experimentally determined physical property can effectively model the solubilities of the present set of gases in alkane solvents. In all of the cases the gases' boiling points were the best property for modeling, followed closely by polarizability. The boiling point works very well as a single property regression: it has $R^2 > 0.94$ in every case, and an average R^2 of 0.961 for all eight solvents. This makes sense, as the boiling point is highly related to both energetic steps in the solvation process. The polarizability also works well as a single property regressor ($R^2 > 0.92$), with an average R^2 value of 0.928 for the eight solvents.

$V_C^{2/3}$ should be related to the second step in solvation, and loosely related to the third step in solvation. The average R^2 value for $V_C^{2/3}$ was 0.825, and for V_C the average $R^2 = 0.789$.

A two-term regression provides no increase of statistical significance to the models. The best two-term regression in each solvent always employed boiling point and polarizability, and the Pol term never has a t-test better than 3. The adjusted R^2 value of the model increases slighlty.

In general, the quality of the regressions decreased as the size of the solvent alkane increased. This is most likely to due to shape-dependant solvent-solvent interactions that complicate the two-step energetic process.

5.6. Appendix

Figure 5.A1. Hexane Solubility modeled by Boiling Point

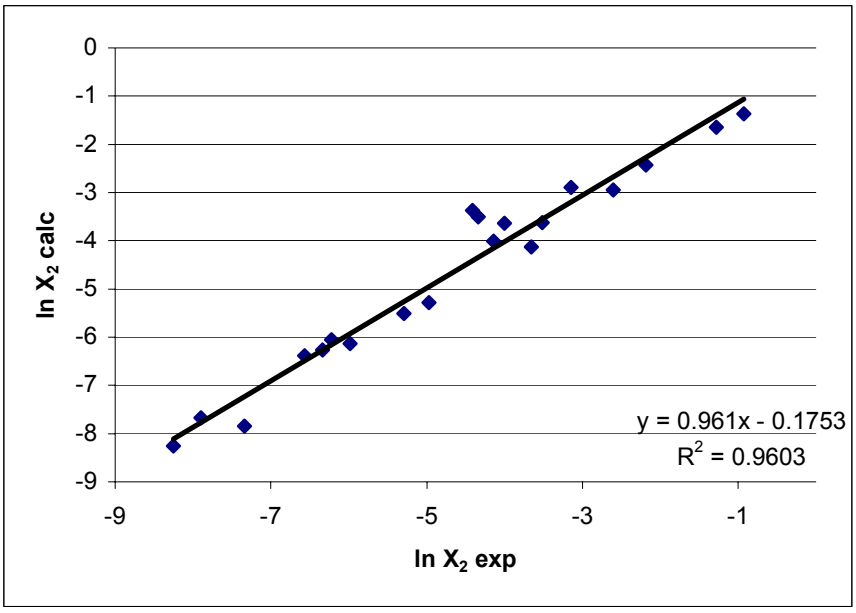


Figure 5.A2. Hexane Solubility modeled by Polarizability

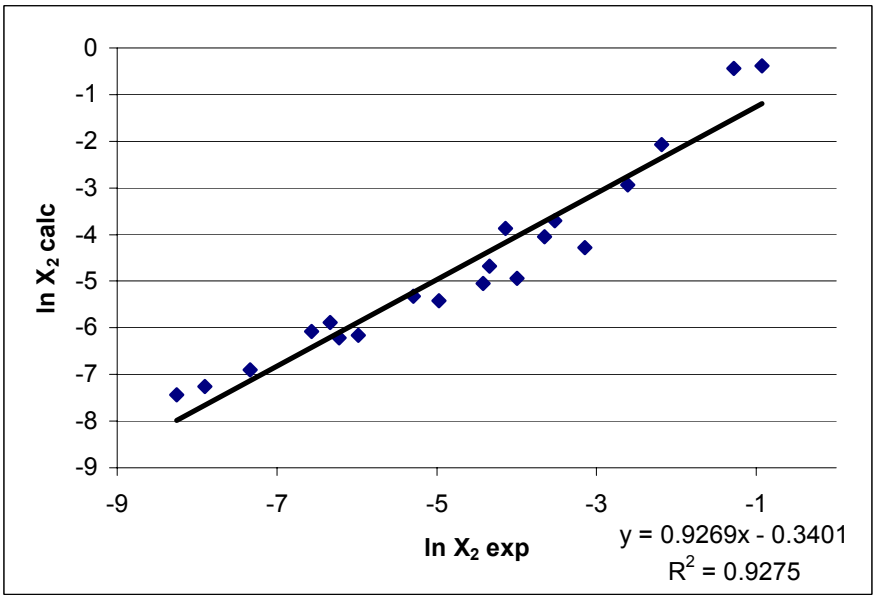


Table 5.A1. - Physical Properties of the Gases, and ln Mole Fraction (ln $X_{2)}$ experimental solubilities.

gas	BP/K	Pol/ Å^3	RG/ Å^3	VC$^{2/3}$/cm^2	VC/cm^3
O2	90.18	1.58	0.680	17.500	73.208
N2	77.34	1.74	0.547	20.100	90.114
CO	81.65	1.95	0.558	20.500	92.818
CO2	194.67	2.91	1.040	20.700	94.179
H2	20.37	0.80	0.371	16.000	64.000
CH4	111.54	2.59	1.118	21.400	98.997
C2H6	184.55	4.43	1.826	28.000	148.162
C2H4	169.45	4.25	1.548	25.500	128.769
N20	184.2	3.03	0.954	21.200	97.612
C2H2	189.15	3.33	0.744	23.400	113.194
Ar	87.29	1.64	1.076	17.700	74.466
Kr	119.9	2.48	1.138	20.300	91.463
Xe	165.15	4.04	1.296	24.100	118.311
Rn	211.15	5.30		27.000	140.296
He	4.23	0.20	0.808	14.900	57.515
Ne	27.25	0.39	0.839	12.000	41.569
C3H8	231.08	6.29	2.431	34.500	202.642
C4H10	272.65	8.20	2.886	40.200	254.882
Isobutane	261.45	8.14	2.98	41.000	262.528
H2S	212.88	3.78	0.638	21.30	98.30
SF6	209.25	6.54		34.000	198.252
CF4	145.15	3.84	2.269	27.000	140.296

5.7. References

[1] IUPAC *Solubility Data Series gas solubility volumes*

[2] QsarIS ver. 1.2 copyright **2001** SciVision 2000 Wheeler Road Burlington, MA. 01803.

[3] Fogg, P. G. T.; Gerrard, W., *Solubility of gases in Liquids* John WIley & Sons Ltd.: Barrins Lane, Chichester West Sussex, England **1991.**

[4] Lide, D.R. (ed.) *CRC handbook of Chemistry and Physics 2002-2003* CRC Press **2003.**

Chapter 6

Solubilities of Gases in Alcohols

6.1. ABSTRACT. The ln X_2 solubilities of 7-16 gases in twelve different monoalcohol solvents were modeled using single physical properties of the gases. The physical properties, in order of ability to model were (1) normal boiling point (2) polarizability, (3) critical volume$^{2/3}$, and (4) critical volume. The boiling point had an average R^2 = 0.959 for the solvents studied

6.2. Introduction

The solubilities of gases in liquids can be broken down into two energetic steps: first, an energy-demanding formation of a cavity for the solute in the liquid, and second, placement of the solute in the cavity and its interaction with the surrounding solvent. The energy needed to form the cavity should be related to the size of the solute molecule. The solute-solvent interaction energy is expected to be related to the solute's surface area and characteristics. The critical volume (V_C) is one measure of solute volume, and $V_C^{2/3}$ is often taken as a measure proportional to the surface area of the solute molecule. The solute polarizability (Pol) is related to the numbers and kinds of atoms in the solute gas.

The energy given back by solute-solvent interaction is normally due mostly to dispersion forces. Hydrogen bonding can also sometimes be important, but in the present study most of the alcohol molecules are large and form only relatively weak hydrogen bonds. The longer the chain of carbons in the alcohol, the less important the hydrogen bonding becomes. The gas polarizability and boiling point are related to cohesive forces and might be related to the interaction energy of the gas with the solvent.

6.3. Methods

All experimental solubility data used in this paper were obtained from the *IUPAC Data Solubility Series*[1], and are in the form of ln X_2, where X_2 is the mole fraction solubility. The software used to perform the regressions was StatMost.[2] Table 5.1 lists the physical properties used, their units, and their sources. All the physical properties for the gases used in this study were experimentally determined. A table of the specific values can be found in the appendix for this chapter.

Table 6.1. - Physical Properties Used in this Study

Name	Abbreviation	Units	Source
Normal Boiling Point	BP	K at 1 bar	Fogg and Gerrard[3]
Polarizability	Pol	$Å^3$	CRC[4]
Critical Volume$^{2/3}$	$V_C^{2/3}$	$(cm^3/mol)^{2/3}$	CRC[4]
Critical Volume	V_C	cm^3/mol	CRC[4]

6.4. Results

Table 2 gives a correlation matrix of the 4 physical descriptors used. All correlations have 22 points of comparison. Table 6.2 gives shows the factor analysis (PCA) of the physical properties.

Table 6.2. - Correlation Matrix of Physical Property and Factor Analysis

	BP	V_C	$Vc^{2/3}$	Pol
BP	1	0.826	0.8495	0.9172
V_C		1	0.9965	0.9720
$V_C^{2/3}$			1	0.9789
Pol				1

Percent Factor Weight

	Factor 1	Factor 2	Factor 3	Factor 4
Factor %	94.33%	5.35%	0.26%	0.06%
Total %	94.33%	99.68%	99.94%	100.00%

Percent Factor loadings

	Factor 1	Factor 2	Factor 3	Factor 4
BP	85.17%	14.75%	0.08%	0.00%
Pol	99.22%	0.00%	0.77%	0.00%
$Vc^{2/3}$	97.20%	2.54%	0.13%	0.12%
Vc	95.71%	4.10%	0.07%	0.12%

The factor analysis indicates that BP has slightly weaker correlations with the other experimental properties. The other properties all correlate above 95%. The small variance in the data accounts for why BP is the best property to model with. In particular, only BP has any sizeable amount of its variance due to Factor 2.

The results are given in the form of a table for each alcohol solvent. The regressions have been given in order of descending R^2 values. Each regression has the general form of

$$\ln X_2 = -A\ (\pm A(S)) + B\ (\pm B(S))*\text{physical property}$$

For example, the first set of statistics for the BP in Methanol study using the gases in Table 4 is

$$\ln X_2 = -9.81\ (\pm\ 0.17) + 0.0247\ (\pm\ 0.0011)*BP$$

T test values 58.0 23.3

Regression quality N = 16 R^2= 0.975 S = 0.349 F = 545.08

Key for the table:

A	Intercept in regression
A(S)	Uncertainty of intercept
t(A)	t-Test value for A
B	Coefficient of the physical property constant
B(S)	Uncertainty of the coefficient of the physical property constant
t(B)	t-test value for B
s	Standard Error of Regression
N	number of solubility data points used in regression
F	F test value

Table 6.3. Methanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	16	0.975	9.81	0.17	58.0	0.0247	0.0011	23.3	545.1	0.349	0.966
Pol	16	0.793	8.92	0.42	21.1	0.78	0.11	7.31	53.5	1.006	0.726
$V_c^{2/3}$	16	0.642	11.2	1.0	11.2	0.211	0.042	5.01	25.2	1.32	0.549
V_c	16	0.597	9.38	0.74	12.7	0.026	0.0057	4.6	20.7	1.41	0.484

Table 6.4. Ethanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
B.P.	16	0.989	9.58	0.11	83.7	0.0251	0.0007	35.4	1255.9	0.237	0.985
Pol	16	0.909	8.85	0.29	30.8	0.817	0.069	11.8	139.3	0.683	0.909
$V_c^{2/3}$	16	0.828	11.58	0.71	16.4	0.231	0.028	8.2	67.3	0.939	0.783
V_c	16	0.792	9.61	0.54	17.7	0.029	0.004	7.3	53.4	1.03	0.733

Table 6.5. 1-propanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	10	0.959	9.3	0.23	40.5	0.024	0.0017	13.8	191.5	0.307	0.927
Pol	10	0.828	8.29	0.42	18.8	0.78	0.11	6.2	38.6	0.637	0.749
$V_c^{2/3}$	10	0.594	12.8	1.9	6.8	0.31	0.09	3.4	11.7	0.979	0.441
V_c	10	0.579	10.59	1.29	8.2	0.043	0.013	3.3	11.1	0.996	0.401

Table 6.6. 2-propanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	10	0.997	9.66	0.088	109.2	0.0271	0.0005	56.1	3149.8	0.109	0.996
Pol	10	0.931	8.38	0.37	22.9	0.767	0.074	10.4	107.5	0.567	0.898
$V_c^{2/3}$	10	0.834	10.86	0.95	11.4	0.215	0.034	6.3	40.1	0.879	0.762
V_c	10	0.816	8.84	0.69	12.7	0.0261	0.0044	6.0	356	0.924	0.735

Table 6.7. 1-butanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
B	14	0.9821	9.33	0.176	53.0	0.0262	0.001	25.7	656.8	0.285	0.972
Pol	14	0.8522	8.26	0.41	19.8	0.795	0.095	8.3	69.2	0.819	0.809
$V_c^{2/3}$	14	0.7102	10.4	1.01	10.4	0.210	0.039	5.4	29.4	1.15	0.638
V_c	14	0.648	8.53	0.71	12.1	0.0257	0.005	5.2	26.6	1.18	0.608

Table 6.8. 1-pentanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
B.P.	12	0.937	9.27	0.29	30.8	0.0263	0.0022	12.2	149.14	0.426	0.895
Polar	12	0.891	9.01	0.38	23.7	1.14	0.13	9.0	81.3	0.562	0.861
$V_c^{2/3}$	12	0.695	13.61	1.64	8.3	0.361	0.076	4.8	22.83	0.938	0.575
V_c	12	0.689	10.95	1.11	9.8	0.051	0.011	4.7	21.9	0.952	0.552

Table 6.9. 1-hexanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	9	0.956	9.12	0.256	35.7	0.256	0.0021	12.9	153.47	0.328	0.914
Pol	9	0.88	8.97	0.42	21.4	1.11	0.15	7.2	51.41	0.545	0.826
$V_c^{2/3}$	9	0.698	12.94	1.68	7.7	0.32	0.079	4.0	16.24	0.864	0.586
V_c	9	0.692	10.63	1.14	9.3	0.045	0.011	4.0	15.75	0.873	0.562

Table 6.10. 1-cyclohexanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	8	0.942	9.84	0.28	35.0	0.0253	0.0025	9.9	98.41	0.479	0.885
Pol	8	0.958	9.92	0.244	40.6	1.25	0.11	11.7	136.61	0.408	0.902
$V_c^{2/3}$	8	0.821	14.07	1.27	11.1	0.34	0.065	5.2	27.55	0.841	0.701
V_c	8	0.809	11.85	0.89	13.3	0.051	0.011	5.1	25.55	0.867	0.627

Table 6.11. 1-heptanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	11	0.961	8.89	0.24	37.1	0.0247	0.0017	14.9	222.1	0.343	0.931
Pol	11	0.86	8.38	0.41	20.3	0.954	0.13	7.5	55.6	0.649	0.786
$V_c^{2/3}$	11	0.699	11.3	1.2	8.9	0.257	0.057	4.6	21.0	0.954	0.585
V_c	11	0.681	9.26	0.87	10.7	0.0339	0.0077	4.4	19.2	0.982	0.522

Table 6.12. 1-octanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	15	0.972	9.11	0.19	48.8	0.0271	0.0013	21.2	446.3	0.382	0.959
Pol	15	0.924	8.47	0.269	31.5	0.958	0.076	12.6	158.6	0.625	0.891
$V_c^{2/3}$	15	0.848	11.74	0.74	15.9	0.272	0.032	8.5	72.7	0.885	0.811
V_c	15	0.819	9.51	0.55	17.3	0.0349	0.0046	7.7	58.7	0.967	0.758

Table 6.13. 1-decanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	11	0.951	8.79	0.19	44.6	0.0241	0.0018	13.2	173.31	0.354	
Pol	11	0.932	9.24	0.269	34.4	1.51	0.14	11.1	122.75	0.417	
$V_c^{2/3}$	11	0.654	13.98	1.81	7.7	0.401	0.097	4.1	17.06	0.937	
V_c	11	0.664	11.79	1.26	9.4	0.0651	0.015	4.3	17.79	0.924	

Table 6.14. 1-undecanol Single Property Regression Details

Descriptor	N	R^2	A	A(S)	t(A)	B	B(S)	t(B)	F	s	Q^2
BP	7	0.952	8.73	0.267	29.5	0.0243	0.0024	10.0	100.3	0.367	0.907
Pol	7	0.887	9.89	0.635	15.6	1.8	0.287	6.3	39.5	0.564	0.741
$V_c^{2/3}$	7	0.467	16.16	4.81	3.4	0.51	0.24	2.1	4.4	1.23	0.179
V_c	7	0.471	13.0	3.3	3.9	0.079	0.037	2.1	4.5	1.22	0.185

6.5. Discussion

The process of solvation is usually thought of in three steps. The normal first step, in which a solute molecule is separated from its bulk, can be ignored here. The solutes are all gases, and so the solute-solute attractive forces are correspondingly already

weak. Step two, where a hole is made in the solvent, should require slightly more energy than in the previous chapter, since the alcohols exhibit some hydrogen bonding. Step three, in where the solute and solvent interact, should be dependant on dispersion forces and hydrogen bonding.. The hydrogen bonding effect on energy should be particularly noticeable in methanol. The results indicate that BP produces an excellent model for methanol ($R^2 = 0.975$). The other physical properties produce models that are uncharacteristically poor. This suggests that only BP can adequately account for the energy of hydrogen bonding.

Each of the four physical properties can be related to one or both of the two energetic steps in the solvation of a gas in a liquid. The normal boiling point is indicative of the energy needed seperate the gas molecules, and is a loose measure of their intermolecular forces. This energy is related to the total size of the gas molecule (dispersion forces) and any other bonding forces that may be present. The polarizability is essentially the distortability of the electron cloud in the gas molecule. It is proportional to the number and types of constituent atoms in the gas molecule, and should be related to the energy released by interaction with the solvent. $V_C^{2/3}$ is a rough indicator of the surface area of the molecule. This property should therefore be highly proportional to the second cavity-forming step in solvation.

The correlation matrix given in the results shows a fair amount of correlation (> 0.83) for all 4 physical properties. This is to be expected, as all four properties are somewhat related to the size of the gas molecule. $V_C^{2/3}$ and Vc correlate very highly, as expected (> 0.99). Since gas molecules are small and fairly rigid, the surface area and volume should be highly correlated. Boiling point and polarizability are highly correlated (0.91). Since both are related to dispersion forces, this is logical. It is

interesting to note that only $V_C^{2/3}$ work better than V_C. This may suggest that the surface area of the molecule is more important than its volume in these models.

In all the cases but one the gas boiling point was the best property for modeling, followed by polarizability. Boiling point works very well as a single property regression: it has $R^2 > 0.9$ in every case, and an average R^2 of 0.959 for all twelve solvents. The polarizability also works well as a single property regression ($R^2 > 0.80$) with an average R^2 value of 0.883 for all twelve solvents.

Radius of gyration and critical volume work are related to the first energetic step of solvation, but have no real correlation with the second step. This seems to be the case, as both do not work as well as boiling point or polarizability. The average R^2 value for $V_C^{2/3}$ was 0.702, and for radius of gyration the average was $R^2 = 0.668$.

$V_C^{2/3}$ was used to change V_C into an estimation of the surface area of the solute gas. It is noteworthy that $V_C^{2/3}$ worked slightly better than V_C by itself.

Two-term regression provided no increase of statistical significance to the models. The best two-term regression in each solvent was always boiling point and polarizability. For any of the models listed here the addition of a second term (Pol) is unnecessary, and the Pol term never had a t-test better than 4. The adjusted R^2 value of the models increased slightly. The two-term regressions were in general of slightly higher quality than the corresponding two-term regressions for the solubilities of gases in alkanes (previous chapter).

6.6. Appendix

Figure 6.A1. – The Solubility of methanol modeled by BP

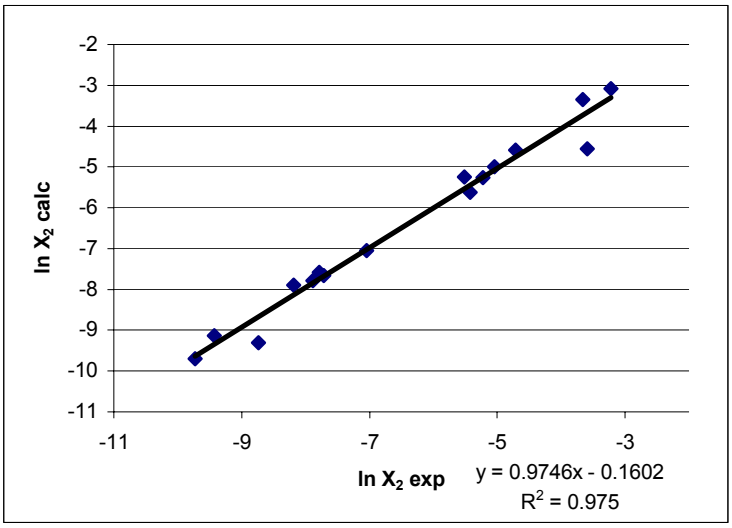


Figure 6.A2. - The Solubility of methanol modeled by Pol

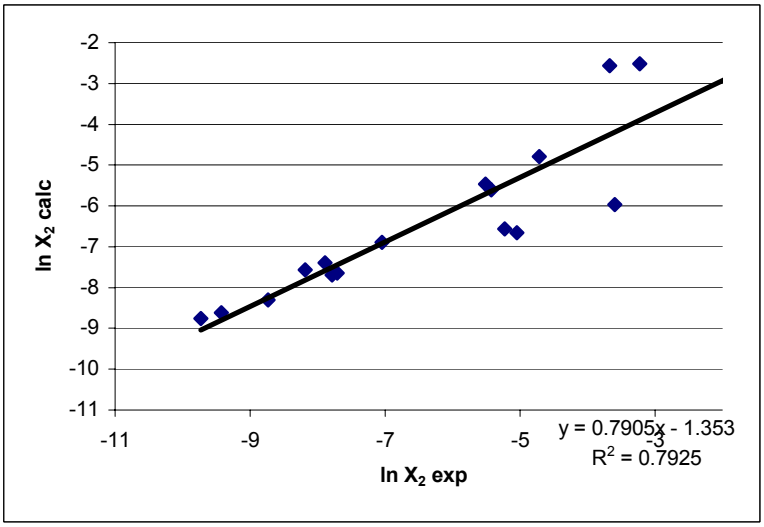


Table 6.A1. - Physical Properties of the Gases, and ln X_2 experimental solubilities.

gas	BP	Pol	RG	$CV^{2/3}$	CV	methanol	ethanol	1-propanol	2-propanol	1-butanol	1-pentanol
O2	90.18	1.58	0.680	17.500	73.208	-7.786	-7.463	-7.595	-7.154	-7.170	-7.370
N2	77.34	1.74	0.547	20.100	90.114	-8.184	-7.932	-7.802	-7.683	-7.684	-7.569
CO	81.65	1.95	0.558	20.500	92.818	-7.886	-7.731	-7.507	-7.412	-7.379	-7.451
CO2	194.67	2.91	1.040	20.700	94.179	-5.047	-5.044	-4.851		-4.730	-4.820
H2	20.37	0.80	0.371	16.000	64.000	-8.734	-8.488	-8.373		-8.240	-8.225
CH4	111.54	2.59	1.118	21.400	98.997	-7.049	-6.661	-6.463	-6.543	-6.282	-6.171
C2H6	184.55	4.43	1.826	28.000	148.162	-5.509	-5.015	-4.739		-4.528	-4.390
C2H4	169.45	4.25	1.548	25.500	128.769	-5.428	-5.093	-4.912	-5.001	-4.799	-4.601
N20	184.2	3.03	0.954	21.200	97.612	-5.221	-4.934	-4.835	-4.835	-4.657	-4.558
C2H2	189.15	3.33	0.744	23.400	113.194						
Ar	87.29	1.64	1.076	17.700	74.466	-7.713	-7.378	-7.161		-7.003	-6.915
Kr	119.9	2.48	1.138	20.300	91.463						-5.751
Xe	165.15	4.04	1.296	24.100	118.311						
Rn	211.15	5.30		27.000	140.296	-4.711	-4.320		-3.932		-3.053
He	4.23	0.20	0.808	14.900	57.515	-9.730	-9.445				
Ne	27.25	0.39	0.839	12.000	41.569	-9.425	-9.124				
C3H8	231.08	6.29	2.431	34.500	202.642		-3.790		-3.523	-3.305	
C4H10	272.65	8.20	2.886	40.200	254.882	-3.224	-2.507		-2.138	-1.959	
Isobutane	261.45	8.14	2.98	41.000	262.528	-3.665	-2.978		-2.602	-2.420	
H2S	212.88	3.78	0.638	21.30	98.30	-3.590				-3.458	

Table 1 continued

gas	1-hexanol	cyclohexanol	1-heptanol	1-octanol	1-undecanol
O2	-7.256		-6.831	-6.784	-6.878
N2	-7.446	-8.232	-7.402	-7.402	-7.071
CO	-7.363		-7.272	-7.071	-7.195
CO2			-4.538	-4.669	-4.153
H2	-8.108	-8.692	-7.862	-7.844	-7.924
CH4	-6.084		-5.952	-5.892	-5.613
C2H6	-4.305	-4.791	-4.148	-4.075	
C2H4	-4.538		-4.448	-4.351	
N20	-4.459	-5.738	-4.370	-4.286	-4.075
C2H2					
Ar	-6.777	-7.543	-6.685	-6.677	
Kr		-6.389		-5.580	
Xe					
Rn					
He		-9.959		-9.022	
Ne		-9.542		-8.684	
C3H8			-2.908	-2.810	
C4H10				-1.252	
Isobutane					
H2S					

6.7. References

[1] IUPAC *Solubility Data Series gas solubility volumes*

[2] QsarIS, ver. 1.2 copyright **2001,** SciVision Inc., 2000 Wheeler Road, Burlington, MA. 01803.

[3] Fogg, P. G. T.; Gerrard, W., *Solubility of gases in Liquids,* John WIley & Sons Ltd.: Barrins Lane, Chichester West Sussex, England **1991.**

[4] Lide, D.R. (ed.) *CRC handbook of Chemistry and Physics 2002-2003,* CRC Press, **2003.**

Chapter 7

Factor Analysis of Gas Solubilities

7.1. Abstract Factor analysis was performed using the SAS software program to determine the number of important underlying factors influencing the solubilities of gases in different solvents. A set of thirteen gases was studied, including five noble gases, two simple alkanes, ethylene, and six permanent gases. For the noble gases two underlying factors were found to be present. For the set of thirteen gases, two underlying factors were also found to account for most of the variance. The more important of the two factors changes from gas to gas for the group of all thirteen gases.

7.2. Introduction

Factor analysis has been used to correlate the solubilities of gases in liquids with the thermodynamic properties of the gases. In 1976 Van der Veen and Ligny[1] reviewed the solubilities of a set of 20 gases in 39 solvents using factor analysis. The standard partial molar entropy of solution and mole fraction solubility were used and correlated in the factor analysis. Each solvent was modeled individually by regression analysis using the results of the factor analysis. The models took the general form:

$$y_g = \sum_{j=1}^{n} G_j * S_j \qquad (7.1)$$

Where y_g is the experimental datum (solubility), and G_j and S_j are adjustable parameters that depend on the gas and solvent, respectively.

In 1988 Gargas, Seybold, and Anderson[2] preformed principal component factor analysis on the partition coefficients of halogenated methanes, ethanes, and ethylenes in various tissues. It was determined that there were 2 dominant dimensions, which could be related to the solubilities of the compounds in saline and oily environments.

In 2005 Sharghi et al.[3] used factor analysis to identify and quantify variations in the energies of solvation. Various properties of the solutes and solvents were used for factor analysis and to create linear energy relationships with the experimental energies of solvation.

Also in 2005 Katritzky et al.[4] used factor analysis in combination with multiple linear regression (MLR) to classify various types of solvents.

The purpose of this work is to use factor analysis on the mole fraction solubility of gases in solvents to determine the number of underlying factors needed to explain the

variance in the solubility data. The number of factors and their importance or "weight" can then be considered when trying to model the solubility of the gases.

Factor analysis utilizes a correlation matrix to determine patterns in the data that indicate the number of important influences. The correlation matrix for the set of relevant influences is first formed and then diagonalized to yield eigenvalues for each factor generated. The higher the eigenvalue, the more that factor accounts for the variance in the data. The sum total of the eigenvalues equals the number of components (in this case gases) examined.

There are several ways to analyze the data in factor analysis. The analysis types are referred to as transformations. In an orthogonal transformation the factors extracted from the data are orthogonal, i.e., independent. This is sometimes called principal components analysis (PCA). In an oblique transformation one seeks to find possible factors regardless of overlap. The latter will sometimes produce more useful patterns, but a consequence of correlated factors is that there is no unambiguous measure of the importance of each factor in explaining the variance. Only orthogonal transformations were used in the present study.

7.3. Methods

All gas solubility data were taken from the IUPAC *Solubility Data Series for gas solubility.*[5] The gases were broken down into related groups to maximize the number of shared solubilities for each analysis. Appendix A lists the natural log of the mole fraction ($\ln X_2$) data and some simple statistics regarding the data. All calculations were performed using the SAS[6] software package. Information on the procedures used in SAS to produce the results can be found in *SAS/STAT User's Guide.*[7]

7.4. Results

7.4.1. Noble Gases

A correlation matrix was first generated for all six noble gases. The matrix was first reviewed for compromising holes in the data set and for basic trends that should be apparent in the factor analysis. Table 1 shows the correlation matrix. Note that, with the exception of helium, the solubilities of the noble gases are highly correlated.

Table 7.1. - Correlation Matrix of ln(mole fraction) Solubilities of Noble Gases.

(correlation coefficient and number of common values)

	He	Ne	Ar	Kr	Xe	Rn
He	1	0.73344	0.61624	0.6834	0.87172	0.87129
	42	39	42	36	23	10
Ne	0.73344	1	0.98179	0.95998	0.87798	0.89408
	39	39	39	36	23	10
Ar	0.61624	0.98179	1	0.99449	0.92528	0.95243
	42	39	44	38	24	12
Kr	0.6834	0.95998	0.99449	1	0.94057	0.97276
	36	36	38	39	24	11
Xe	0.87172	0.87798	0.92528	0.94057	1	0.98675
	23	23	24	24	25	8
Rn	0.87129	0.89408	0.95243	0.97276	0.98675	1
	10	10	12	11	8	14

Analysis 1

The factor analysis results for this analysis show results for the four lowest molecular mass noble gases (He, Ne, Ar, and Kr). The gases are shown below.

Analysis 1: Factor analysis for He, Ne, Ar, and Kr.

Factor	Eigenvalue	Proportion	% loading	total %
1	3.5117	3.0677	87.79%	87.79%
2	0.4439	0.4019	11.10%	98.89%
3	0.042	0.0398	1.05%	99.94%
4	0.0024	0.0034	0.06%	100.00%

gas	Factor 1	% loading	Factor 2	% loading	Factor 3	% loading	Factor 4	% loading
He	0.7978	63.65%	0.6027	36.32%	0.00292	0.00%	0.000259	0.0000%
Ne	0.9718	94.44%	-0.1724	2.97%	0.16005	2.56%	0.000267	0.0000%
Ar	0.986	97.22%	-0.1575	2.48%	-0.03884	0.15%	0.001489	0.0002%
Kr	0.9789	95.82%	-0.1615	2.61%	-0.12216	1.49%	0.000749	0.0001%

The gases in this analysis share 35 solvents. Here SAS was instructed to look for the 4 most important unrelated factors. The proportion column indicates how important each factor is in the overall effect for all four gases. It is clear that there are two important factors here, since factor three accounts for only 1.05% of the variance in the data. The cumulative column indicates the percent of the variance in the data that is explained by the factors in that row or above; here factors 1 and 2 together explain almost 99% of the variance in the solubility. The individual breakdown of each factor's importance to each gas comes next. This follows the formula

$$(\text{factor 1})^2+(\text{factor 2})^2+(\text{factor 3})^2+(\text{factor 4})^2= 1.00 \qquad (7.2)$$

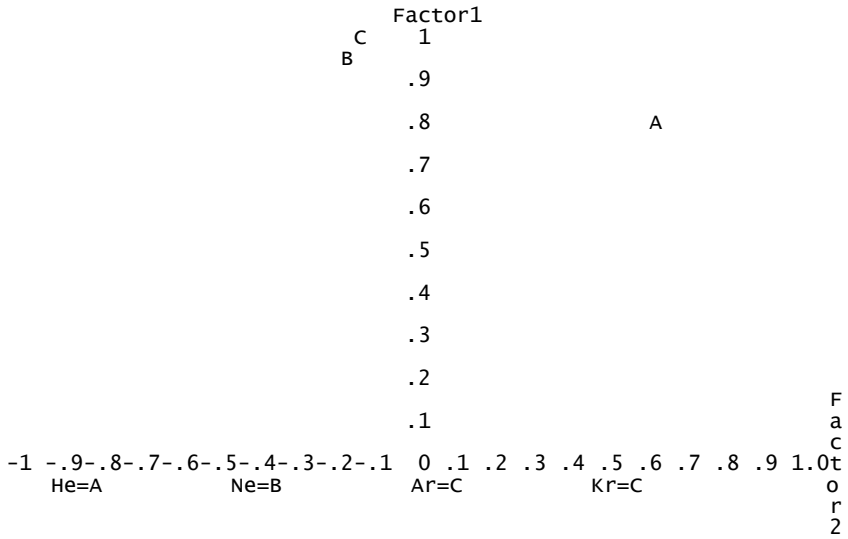
For example, for He

$$(0.79789)^2+(0.60279)^2+(0.00292)^2+(0.00269)^2= 1.00$$

Ne, Ar, and Kr are all >94% dependent on factor 1, whereas helium is significantly dependant on factor 2. The factor pattern may also be used to infer how important each factor is to each specific noble gas. A negative factor loading indicates a

negative dependence on that factor. Figure 7.2 shows a plot of how each gas depends on the first two factors.

Figure 7.2. - Factor plot for analysis 1: He, Ne, Ar, and Kr



Analysis 2

Analysis 2 adds Xe to the previous group of He, Ne, Ar, and Kr. There were 23 common solubility points. Solvents included in the second factor analysis, and left out of anaylsis 2 that are in analysis 1, are listed in Table 7.2. Fewer solvents were used in this factor analysis because solubility data for Xe in some solvents was not available.

As seen in the results below, once again there are only two important factors, with the third factor accounting for only 0.52% of the variance. Factor 4 was left out of this analysis, and all following analyses, due to its lack of importance. The importance of factor two to Helium plummets, possibly due to the missing shared solubility values. The eight solvents excluded in this case may therefore be responsible for making helium so dependent on factor 2 in the earlier analysis. It is also clear that factor 1 is still by far the predominant factor, accounting for >88% of the variance for Xenon and > 96% for the other considered gases. Figure 7.5. is a plot of the two important factors.

Analysis 2: Factor analysis for He, Ne, Ar, Kr, and Xe.

Factor	Eigenvalue	Proportion	% loading	total %
1	4.8019	4.6359	96.04%	96.04%
2	0.1659	0.1401	3.32%	99.36%
3	0.0258	0.0204	0.52%	99.87%

gas	Factor 1	% loading	Factor 2	% loading	Factor 3	% loading
He	0.9821	96.45%	-0.1688	2.85%	0.0664	0.44%
Ne	0.9852	97.06%	-0.1557	2.42%	0.0481	0.23%
Ar	0.997	99.40%	-0.029	0.08%	-0.0668	0.45%
Kr	0.9937	98.74%	0.0348	0.12%	-0.1042	1.09%
Xe	0.9408	88.51%	0.3334	11.12%	0.0611	0.37%

Table 7.3. - Solvents used in the Second Factor Analysis

2,2,4-trimethylpentane	n-heptane	n-nonane
benzene	n-hexadecane	n-octane
carbon disulfide	n-hexane	n-pentane
chlorobenzene	methylcyclohexane	n-tetradecane
cyclohexane	methylhydrazine	n-pentadecane
n-decane	nitrobenzene	toluene
n-dodecane	nitromethane	n-tridecane
n-undecane	Water	

Analysis 3

Analysis three includes all the noble gases. There are only 6 shared solvents in this example. It is important to note that any conclusions obtained from this example are tentative because of the scarcity of data points. The six solvents used are water, hexane, cyclohexane, carbon disulfide, benzene, and toluene.

Analysis 3: Factor analysis for He, Ne Ar, Kr, Xe and Rn.

Factor	Eigenvalue	Proportion	% loading	total %
1	5.8439	5.6886	97.40%	97.40%
2	0.1553	0.1546	2.59%	99.99%
3	0.0006	0.0006	0.01%	100.00%

gas	Factor 1	% loading	Factor 2	% loading	Factor 3	% loading
He	0.9757	95.20%	0.2187	4.78%	0.0124	0.02%
Ne	0.9853	97.08%	0.1705	2.91%	0.0007	0.00%
Ar	0.9992	99.84%	0.0373	0.14%	-0.0146	0.02%
Kr	0.9988	99.76%	-0.0465	0.22%	-0.0122	0.01%
Xe	0.9881	97.63%	-0.1532	2.35%	0.0047	0.00%
Rn	0.9739	94.85%	-0.2266	5.13%	0.0095	0.01%

Here again there are only two important factors, with factor one being responsible for almost all the variation in the data. Factor 2 plays a small role in helium and radon. Figure 7.6. shows a plot of the two important factors.

7.4.2. Results for all 13 gases

20 different groupings of the 13 gases were analyzed by orthogonal transformations. In every case the % factor loading of the third factor was never above 0.15, indicating just two important factors. Not all of the 20 groups are shown, as the trends can be shown in a couple of example analyses.

A correlation matrix was first generated for all the gases involved. The matrix was first reviewed for compromising holes in the data set and for basic trends that that should be confirmed in the factor analysis. Table 7.4. is a representative correlation matrix used in the factor analysis. A full correlation matrix can be found in the appendix.

Table 7.4. – A Representative Correlation Matrix[a]

	Ar	O_2	CO	CO_2	CH_4	N_2O
Ar	1	0.71108	0.54159	0.48486	0.80391	0.66055
#	50	39	31	36	37	25
O_2	0.71108	1	0.6432	0.46405	0.92111	0.14843
#	39	65	42	49	50	34
CO	0.54159	0.6432	1	0.15536	0.57081	0.12729
#	31	42	45	39	39	33
CO_2	0.48486	0.46405	0.15536	1	0.27197	0.21137
#	36	49	39	54	44	34
CH_4	0.80391	0.92111	0.57081	0.27197	1	0.09648
#	37	50	39	44	54	33
N_2O	0.66055	0.14843	0.12729	0.21137	0.09648	1
#	25	34	33	34	33	37

a - a # indicate the number of compared data points

Analysis 4

The gases were first analyzed by their chemical groups. The noble gases are already reviewed, so the next group is composed of CH_4, C_2H_6, and C_2H_4. Data was available for 19 common solvents for these gases.

Factor	Eigenvalue	Proportion	% loading	total %
1	2.6999	2.4468	90.00%	90.00%
2	0.2531	0.2062	8.44%	98.44%
3	0.0469	0.0365	1.56%	100.00%

gas	Factor 1	% loading	Factor 2	% loading	Factor 3	% loading
CH_4	0.94109	88.57%	-0.31886	10.17%	0.1126	1.27%
C_2H_6	0.92027	84.69%	0.38524	14.84%	0.06851	0.47%
C_2H_4	0.98357	96.74%	-0.05536	0.31%	-0.17184	2.95%

In this case each gas is more than 88% dependant on one factor. Factor three did not contribute in a meaningful way.

Each of the similar groups of gases reviewed exhibited similar behavior. There were always two relevant factors, with the first factor explaining 80% or more of the variance.

The next step is to cross the groups of gases. This allows conclusions to be drawn about the data set a whole while maximizing the amount of data analyzed. This increases the accuracy of any conclusions drawn. In analysis 5 two noble gases and are crossed with the last group. For Analysis 5 there are 16 common solvents.

Analysis 5

Factor	Eigenvalue	Proportion	% loading	total %
1	3.3727	2.8461	84.32%	84.32%
2	0.5265	0.4397	13.16%	97.48%
3	0.0868	0.073	2.17%	99.66%

gas	Factor 1	% loading	Factor 2	% loading	Factor 3	% loading
Ar	0.94421	89.15%	-0.31634	10.01%	0.03226	0.10%
He	0.93892	88.16%	-0.32302	10.43%	0.08878	0.79%
C_2H_6	0.82636	68.29%	0.5451	29.71%	0.14123	1.99%
C_2H_4	0.9575	91.68%	0.15826	2.50%	-0.24075	5.80%

The correlations found here can be applied to the gases reviewed in the first two examples. In analysis 1 Ar and Kr are 99.2% similar in how the first two factors are weighted in importance. Thus, the results for Ar in Analysis 5 give a good indication of how Kr would act As expected, there are only two relevant factors. Each gas is still primarily dependent on one factor. Factor two plays a slightly more important role in this example than the previous two. As the groupings of gases become more unrelated the relative importance of the first two factors became closer. In no case is there a noteworthy third factor. Analysis 6 has 19 common solvents.

Analysis 6

Factor	Eigenvalue	Proportion	% loading	total %
1	3.6851	3.5063	0.9213	92.13%
2	0.1788	0.0859	0.0447	96.60%
3	0.0928	0.04943	0.0232	98.92%

gas	Factor 1	% loading	Factor 2	% loading	Factor 3	% loading
Ar	0.9563	91.45%	-0.25711	6.61%	-0.07273	0.53%
N_2O	0.9773	95.51%	-0.01628	0.03%	-0.15345	2.35%
CO	0.96532	93.18%	-0.05138	0.26%	0.25171	6.34%
C_2H_4	0.93999	88.36%	0.33127	10.97%	-0.02495	0.06%

Even with four radically different gases no important third factor is present. The trend noted in example 3 is actually downplayed in this example, with only 4.63% of the variance explained by factor 2.

Conclusion

20 different groupings of the 13 gases were analyzed by orthogonal transformations. In every case the eigenvalue of the third factor was never above 0.15, indicating 2 important factors.

Figure 7.5. - Factor Plot for Analysis 1: all Noble Gases but Rn

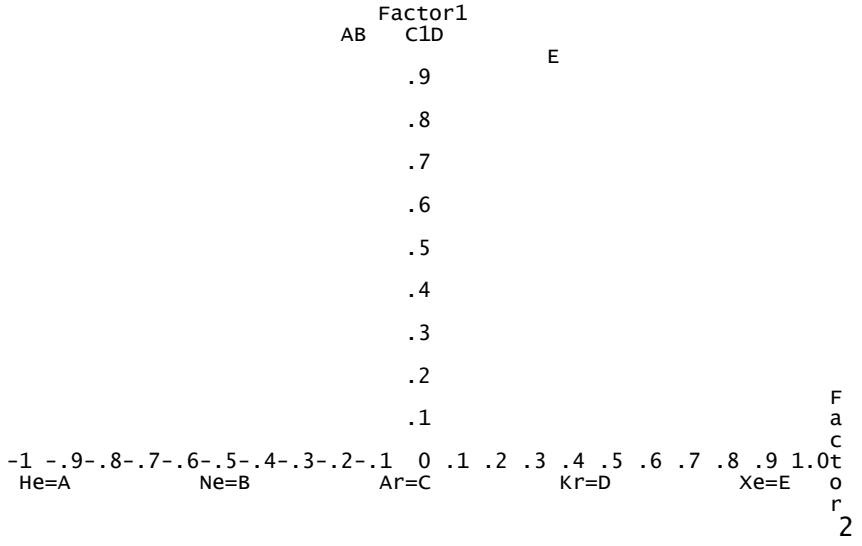
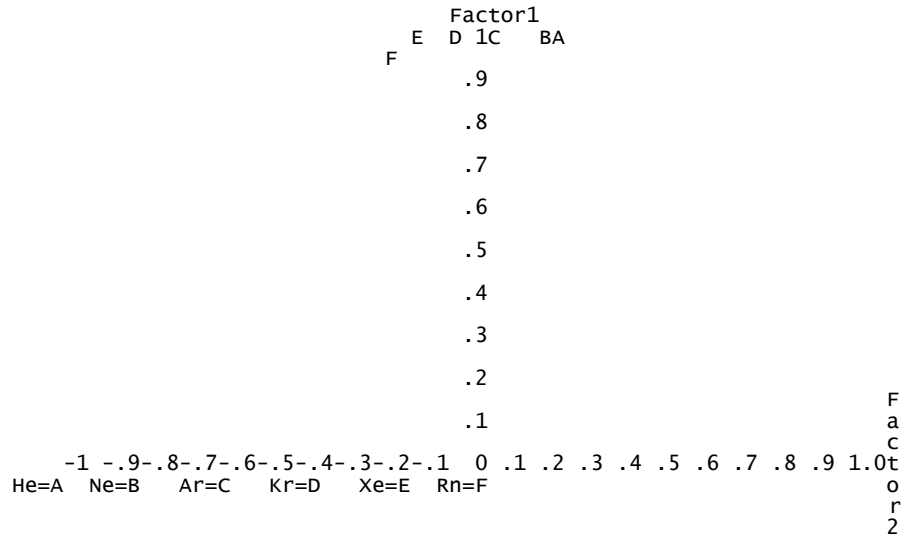


Figure 7.6. - Factor Plot for Analysis 2: All the Noble Gases



7.5. Appendix

Table 7.A1 Solubilities of the Noble gases, expressed as ln(molefraction) in different Solvents at 298

Solubility/Gas	He	Ne	Ar	Kr	Xe	Rn
water	-1.187E+01	-1.172E+01	-1.059E+01	-1.001E+01	-9.447E+00	-8.697E+00
D2O	-1.167E+01	-1.157E+01	-1.052E+01	.	.	.
pentane	-8.255E+00	-7.799E+00	-5.864E+00	-4.847E+00	-3.601E+00	.
hexane	-8.255E+00	-7.899E+00	-5.983E+00	-4.969E+00	-3.654E+00	-2.608E+00
heptane	-8.298E+00	-7.943E+00	-5.991E+00	-4.945E+00	-3.649E+00	.
octane	-8.350E+00	-7.924E+00	-6.020E+00	-4.952E+00	-3.674E+00	.
2,2,4-Trimethylpentane	-8.085E+00	-7.689E+00	-5.836E+00	-4.830E+00	-3.605E+00	.
nonane	-8.352E+00	-7.978E+00	-5.999E+00	-4.945E+00	-3.688E+00	.
decane	-8.340E+00	-7.963E+00	-5.999E+00	-4.935E+00	-3.700E+00	.
undecane	-8.568E+00	-7.902E+00	-5.968E+00	-4.948E+00	-3.703E+00	.
dodecane	-8.413E+00	-8.069E+00	-5.964E+00	-4.884E+00	-3.483E+00	.
tridecane	-8.568E+00	-7.929E+00	-6.004E+00	-4.968E+00	-3.721E+00	.
tetradecane	-8.386E+00	-8.035E+00	-5.956E+00	-4.852E+00	-3.725E+00	.
pentadecane	-8.623E+00	-7.958E+00	-6.004E+00	-4.956E+00	-3.737E+00	.
hexadecane	-8.623E+00	-8.047E+00	-6.004E+00	-4.968E+00	-3.743E+00	.
cyclohexane	-9.011E+00	-8.598E+00	-6.509E+00	-5.360E+00	-3.868E+00	-2.578E+00
methylcyclohexane	-8.728E+00	-8.377E+00	-6.293E+00	-5.160E+00	-3.785E+00	.
cis-1,2dimethylcyclohexa	-8.874E+00	-8.417E+00	-6.329E+00	-5.213E+00	.	.
trans-1,2dimethylcyclohe	-8.623E+00	-8.236E+00	-6.218E+00	-5.121E+00	.	.
cyclooctane	-7.059E+00	-8.894E+00	-6.670E+00	-5.672E+00	.	.
benzene	-9.486E+00	-9.070E+00	-7.033E+00	-5.900E+00	-4.457E+00	-3.214E+00
toluene	-9.220E+00	-8.849E+00	-6.816E+00	-5.699E+00	-4.227E+00	-3.006E+00
o-Xylene	-4.680E+00	-8.909E+00	-6.828E+00	-5.690E+00	.	.
m-Xylene	-9.181E+00	-8.731E+00	-6.737E+00	-5.620E+00	.	.
p-Xylene	-9.143E+00	-8.786E+00	-6.689E+00	-5.578E+00	.	.
methanol	-9.730E+00	-9.425E+00	-7.713E+00	.	.	-4.711E+00
ethanol	-9.445E+00	-9.124E+00	-7.378E+00	.	.	-4.320E+00
isobutanol	-9.191E+00	-8.785E+00	-6.957E+00	-5.972E+00	.	-3.471E+00
1-pentanol	.	.	-6.915E+00	-5.751E+00	.	-3.053E+00
1-octanol	-9.022E+00	-8.684E+00	-6.677E+00	-5.580E+00	.	.
1-decanol	-8.786E+00	-8.528E+00	-6.506E+00	-5.461E+00	.	.
1,2,3-propanetriol	.	.	.	-8.517E+00	.	-7.346E+00
cyclohexanol	-9.959E+00	-9.542E+00	-7.543E+00	-6.389E+00	.	.
acetone	-9.124E+00	-8.759E+00	-7.006E+00	-5.745E+00	.	-4.075E+00
acetic acid	.	.	-7.785E+00	-6.803E+00	-5.960E+00	-4.547E+00
hexafluorobenzene	-8.454E+00	-7.969E+00	-6.032E+00	-5.133E+00	.	.
chlorobenzene	-9.577E+00	-9.231E+00	-7.057E+00	-5.903E+00	-4.398E+00	.
carbon disulfide	-1.015E+01	-9.738E+00	-7.628E+00	-6.345E+00	-4.566E+00	-3.041E+00
sulfinylbismethane	-1.047E+01	-1.021E+01	-8.779E+00	-7.715E+00	-6.377E+00	.
nitromethane	-1.016E+01	-9.827E+00	-8.063E+00	-7.084E+00	-3.907E+00	.
nitrobenzene	-1.026E+01	-1.004E+01	-7.717E+00	-6.578E+00	-5.143E+00	.
aniline	.	.	.	.	-6.230E+00	-4.313E+00
methylhydrazine	-1.059E+01	.	-8.628E+00	.	.	.
1,1-dimethylhydrazine	-9.279E+00	.	-7.280E+00	.	.	.
1,2-dimethylhydrazine	-1.268E+01	.	-5.770E+00	.	.	.

Table 7.A2. Correlation Matrix of Solubility for all 13 Gases

	He	Ne	Ar	Kr	H2	N2	O2	CH4	C2H6	C2H4	CO	CO2	N2O
He	1.0000	-0.028	-0.0202	-0.0425	0.7836	0.8284	0.0104	-0.0607	0.4617	0.7860	0.4509	0.0453	0.9121
#	41	39	41	36	23	31	34	30	16	17	27	31	21
Ne	-0.028	1.0000	0.9712	0.9484	0.8421	0.8463	0.6840	0.7613	0.3720	0.7896	0.4716	0.7434	0.9305
#	39	39	39	36	23	30	34	30	16	17	27	30	20
Ar	-0.020	0.971	1.0000	0.9936	0.8750	0.8709	0.7111	0.8039	0.4336	0.8041	0.5416	0.4849	0.6606
#	41	39	50	38	28	38	39	37	22	23	31	36	25
Kr	-0.042	0.948	0.9936	1.0000	0.8591	0.8277	0.5722	0.7807	0.3532	0.7941	0.4485	0.5754	0.9193
#	36	36	38	39	22	30	34	30	15	16	26	30	19
H2	0.783	0.842	0.8750	0.8591	1.0000	0.9681	0.9725	0.9787	0.3794	0.9212	0.5824	0.5173	0.0937
#	23	23	28	22	45	44	43	41	28	28	35	38	32
N2	0.828	0.846	0.8709	0.8277	0.9681	1.0000	0.9411	0.9431	0.1642	0.8805	0.6415	0.6020	0.1067
#	31	30	38	30	44	59	53	48	31	30	43	44	34
O2	0.010	0.684	0.7111	0.5722	0.9725	0.9411	1.0000	0.9211	0.2507	0.9173	0.6432	0.4641	0.1484
#	34	34	39	34	43	53	65	50	28	30	42	49	34
CH4	-0.060	0.761	0.8039	0.7807	0.9787	0.9431	0.9211	1.0000	0.5670	0.9086	0.5708	0.2720	0.0965
#	30	30	37	30	41	48	50	54	28	29	39	44	33
C2H6	0.461	0.372	0.4336	0.3532	0.3794	0.1642	0.2507	0.5670	1.0000	0.8836	0.2889	0.0688	0.1249
#	16	16	22	15	28	31	28	28	32	27	27	27	25
C2H4	0.786	0.789	0.8041	0.7941	0.9212	0.8805	0.9173	0.9086	0.8836	1.0000	0.4890	0.5723	0.2232
#	17	17	23	16	28	30	30	29	27	33	27	27	24
CO	0.450	0.4716	0.5416	0.4485	0.5824	0.6415	0.6432	0.5708	0.2889	0.4890	1.0000	0.1554	0.1273
#	27	27	31	26	35	43	42	39	27	27	45	39	33
CO2	0.045	0.743	0.4849	0.5754	0.5173	0.6020	0.4641	0.2720	0.0688	0.5723	0.1554	1.0000	0.2114
#	31	30	36	30	38	44	49	44	27	27	39	54	34
N2O	0.912	0.930	0.6606	0.9193	0.0937	0.1067	0.1484	0.0965	0.1249	0.2232	0.1273	0.2114	1.0000
#	21	20	25	19	32	34	34	33	25	24	33	34	37

7.6. References

[1] Ligny, C. L.; Van deer Veen, N.G., Correlation and Prediction of Solubility and Entropy of Solution of Gases in Liquids by Means of Factor Analysis, *Ind.Eng.Chem.*, **1976**, *15*, 336-341.

[2] Gargas, M. L.; Seybold, P. G.; Anderson, M. E., Modeling the Tissue Solubilities and Metabolic Rate Constant of Halogenated Methanes, Ethanes, and Ethylenes, *Tox.Lett.*, **1988**, *43*, 235-256.

[3] Sharghi, H., Hemmateenejad, B., Ghavi, R.; Shamsipur, M., Solvatochromic Linear Solvation Energy Relationships for Solubility in Various Solvents by Target Factor Analysis, *J. Chi. Chem. Soc*., **2005**, *52*, 11-19.

[4]Katritzky, A.R.; Fara, D.C.; Kuanar, M.; Hur, E.; Karelson, M., The Classification of Solvents by Combining Classical QSPR Methodology with Principle Component Analysis, *J. Phys. Chem. A.*, **2005**, *109*, 10323-10341.

[5]IUPAC *Solubility Data Series gas solubility volumes*

[6] SAS for windows, version 8., © 2003, SAS Institute Inc, Cary, NC, USA.

[7] Cody, Smith *Applied Statistics and the programming Language Fourth Edition*; Prentice Hall, **1997.**