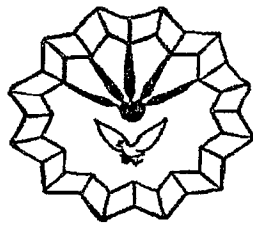




٧٣٧٦٦

۸۷/۱۱/۲۶۳۴
۸۷/۱۱/۷



Razi University
Faculty of Science
Department of Chemistry

PhD Thesis

**Quantitative Structure–Property Relationships (QSPRs) Studies
of Aromatic Acids, Refrigerants, Cationic Surfactants, Diverse
Drugs and Macromolecules Using Chemometrics Methods**

Supervisor:

Prof. Jahanbakhsh Ghasemi

۸۷/۱۱/۲۶۳۴

۸۷/۱۱/۷

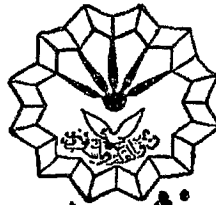
By:

Saadi Saaidpour

July 2008

۸۷/۱۱/۷

۸۷/۱۱۰۲۹۳۴



دانشگاه رازی

دانشکده علوم

گروه شیمی

پایان نامه جهت اخذ درجه دکتری تخصصی

رشته ی شیمی گرایش تجزیه

مطالعات روابط کمی ساختار ملکولی ترکیبات آلی اسیدهای آروماتیک،
خنک کننده ها، دترجنت های کاتیونی، ترکیبات شبه دارویی و دارویی،
وماکرومولکولها با خواص آنها بوسیله روشهای مختلف کمومتریکس

استاد راهنما:

پروفسور جهانبخش قاسمی

نگارش:

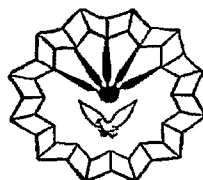
سعدی سعیدپور

۹۹۷۷

تیر ماه ۱۳۸۷

۱۳۸۷ / ۱۷ / ۱۱

ص



Razi University

Faculty of Science
Department of Chemistry

PhD Thesis

Quantitative Structure–Property Relationships (QSPRs) Studies of Aromatic Acids, Refrigerants, Cationic Surfactants, Diverse Drugs and Macromolecules Using Chemometrics Methods

By:

Saadi Saaidpour

Evaluated and approved by thesis committee: as *EXCELLENT*

J. Ghaseemi.....Supervisor: Jahanbakhsh Ghaseemi, Prof. of Analytical Chemistry (Chairman)

A. Afkhami.....External Examiner: Abbas Afkhami, Prof. of Analytical Chemistry

T. Madrakian.....External Examiner: Tayyebeh Madrakian, Assoc. Prof. of Analytical Chemistry

M. B. Gholivand.....Internal Examiner: Mohammad Bagher Gholivand, Prof. of Analytical Chemistry

M. Irandoust.....Internal Examiner: Mohsen irandoust, Assist. Prof. of Analytical Chemistry

F. Jalali.....Internal Examiner: Fahimeh Jalali, Assist. Prof. of Analytical Chemistry

July 2008

کلیه حقوق مادی مترتب بر نتایج مطالعات، ابتکارات و
نوآوری های ناشی از تحقیق موضوع این پایان نامه
متعلق به دانشگاه رازی است.

DEDICATION

Dedicated to:

My Parents

My Brothers

My Sister

My Wife Mina

and

My Children

Soroush and Samanah

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisor, Prof. Jahanbakhsh Ghasemi for his patience, guidance and encouragement all through the project. I admire him for his attitude to explore and learn new aspects of science and for his enthusiasm in teaching and mentoring. Prof. Ghasemi helped me identify my strengths and also guided me through difficult situations. This dissertation would not have been possible without his guidance and persistent help.

I would like to thank the thesis committee, Prof. M.B. Gholivand, Dr. F. Jalali, Dr. M. Irandoust, Prof. A. Afkhami, and Dr. T. Madrakian for their direction, assistance and guidance.

I want to give special thanks to the Dr. M. Irandoust head of chemistry department for his contribution.

I would like to thank my parents and my family members who have been a constant source of support and encouragement. It is their confidence in me that has allowed me to achieve my goals.

Finally, I would like to express my gratitude to my wife, who as a friend helped me through my project with constant support, love and encouragement. Finally, I thank anyone who taught me even one word.

ABSTRACT

A quantitative structure-activity/property relationship (QSAR/QSPR) has become an important branch of modern chemistry in past decades. A fundamental goal of QSAR/QSPR studies is to predict complex physical, chemical, biological, and technological properties of chemicals from simpler descriptors, preferably those calculated solely from molecular structure. This thesis focuses upon the methodology of QSPR, and the results of seven studies that implement that methodology. The goal of these works is to create predictive models that will link the molecular structures of sets of organic compounds to their physicochemical properties and/or biological activities.

The first study (chapter2) involves a very simple, strong, descriptive and interpretable model, based on the quantitative structure-property relationship (QSPR), is developed using multiple linear regression approach and quantum chemical descriptors derived from AM1-based calculations (MOPAC7.0) for determination of the acidity constants of some aromatic acids derivatives. A multiple linear regression (MLR) model with 74 molecules as training set has been developed for the prediction of the acidity constants of some aromatic acids using quantum chemical descriptors. The pK_a values of aromatic acids generally decreased with increasing positive partial charges of acidic hydrogen atom. This model was applied for the prediction of the pK_a of some aromatic acids (33 test acids), which were not used in the modeling procedure. The average relative error ($\overline{RE}\%$) of prediction is lower than 1% and square correlation coefficient (R^2) of 0.9882 is obtained.

In the second study (chapter3) of QSPR model for the estimation of boiling points of organic compounds containing halogens, oxygen, or sulfur without hydrogen bonding were established with the Molecular Modeling Pro Plus (MMPP) software. A QSPR study was performed to develop models that relate the structures of 90 refrigerants compounds to their boiling point temperatures. The optimal QSPR model was developed based on a 4-4-1 artificial neural network (ANN) architecture using molecular descriptors calculated from molecular structure alone. The root mean square errors (RMSE) in normal boiling points predictions were 4.46°C for the training set, 3.86°C for the validation set and 4.99 °C for the prediction set.

In the third study (chapter 4) relationships between the molecular structure and the critical micelle concentration (CMC) of cationic surfactants were investigated using a quantitative structure-micellization relationship (QSMR) approach. The CMC of a set of

29 tetra-alkyl ammonium and 15 alkylpyridinium salts was related to molecular structure descriptors using an ordinary least squares regression (OLS) method. Among different models obtained, three equations were selected as the best and their specifications are given. The results obtained for the simultaneous modeling of tetra-alkyl ammonium and alkylpyridinium salts indicate that geometric characteristics such as the hydrophobic chain length (L_C), hydrophobic volume (V_H), area of hydrophilic portion (A_{HP}) and radius of the hydrated counter ions (R_{HCl}) play a major role in micelle formation. Root mean square error of prediction (RMSEP) and average relative error ($\overline{RE}\%$) of prediction set for simultaneous model were about 0.0938 and 2.1124%, respectively.

The fourth study (chapter 5) involves a QSPR study was performed to develop models those relate the structures of 150 drug organic compounds to their *n*-octanol–water partition coefficients ($\log P_{o/w}$). A genetic algorithm was also applied as a variable selection tools in QSPR analysis. The models were constructed based on 110 training compounds, and predictive ability was tested on 40 compounds reserved for that purpose. Modeling of logarithm of $\log P_{o/w}$ of these compounds as a function of the theoretically derived descriptors was established MLR and ANN. The neural network employed here is a connected back-propagation model with a 4-4-1 architecture. Four descriptors for these compounds molecular volume (MV) (Geometrical), hydrophilic-lipophilic balance (HLB) (Constitutional), hydrogen bond forming ability (HB) (Electronic) and polar surface area (PSA) (Electrostatic) are taken as inputs for the models. The root mean square error of prediction (RMSEP) and square correlation coefficient (R^2) for MLR and ANN models were 0.2158, 0.9864 and 0.1838, 0.9876 for the prediction set $\log P_{o/w}$, respectively.

The fifth study (chapter 6) involves quantitative structure-retention relationship (QSRR) analysis is a useful technique capable of relating chromatographic retention time to the chemical structure of a solute. A QSRR study has been carried out on the reversed-phase high-performance liquid chromatography (RP-HPLC) retention times ($\log t_R$'s) of 62 diverse drugs (painkillers) by using molecular descriptors. MLR is utilized to construct the linear QSRR model. The applied MLR is based on a variety of theoretical molecular descriptors selected by the stepwise variable subset selection procedure. Stepwise regression was employed to develop a regression equation based on 50 training compounds, and predictive ability was tested on 12 compounds reserved for that purpose. The geometry of all drugs were optimized by the semi-empirical method AM1 and used to calculate different molecular descriptors. The regression equation included three parameters that consisted of *n*-octanol–water partition coefficient ($\log P$), molecular surface

area (SM) and hydrophilic-lipophilic balance (HLB) of the drug molecules, all of which could be related to retention time property. The results indicate that a strong correlation exists between the $\log t_R$ and mentioned descriptors for drug compounds.

In the sixth study (chapter 7) a QSPR study was performed to develop a model that relates the structures of 150 drug organic compounds to their aqueous solubility ($\log S_w$). Molecular descriptors derived solely from 3D structure were used to represent molecular structures. A subset of the calculated descriptors selected using stepwise regression that used in the QSPR model development. Stepwise regression was employed to develop a regression equation based on 110 training compounds, and predictive ability was tested on 40 compounds reserved for that purpose. The final regression equation included three parameters that consisted of octanol/water partition coefficient ($\log P$), molecular volume (MV) and hydrogen bond forming ability (HB), of the drug molecules, all of which could be related to solubility property. The prediction results are in good agreement with the experimental values. The root mean square error of prediction (RMSEP) and square correlation coefficient (R^2) of prediction of $\log S_w$ were 0.0959 and 0.9954, respectively.

The seventh study (chapter 8) involves QSPR models for the stability constants of 58 complexes of 1,4,7,10,13-pentaoxacyclopentadecane ethers (15C5) were established with the CODESSA program. Experimental stability constants ($\log K$) for a diverse set of 58 complexes of 15C5 structures are correlated with computed structural descriptors using CODESSA. Stability constants for complexes of 15C5 ethers with potassium cation (K^+) have been determined at 25 °C in methanol solution. The best multilinear regression method (BMLR) encoded in CODESSA software was used to select significant descriptors for building multilinear QSPR model and the predictive power of model is estimated with the leave-one-out (LOO) cross-validation method. The proposed model can be used for the prediction of the stability constants of 15C5 complexes. The best QSPR model with five descriptors has $R^2 = 0.9452$, $s^2 = 0.0110$, and $F = 67.0312$.

TABLE OF CONTENTS

Content	Page
TABLE OF CONTENTS-----	A
LIST OF TABLES-----	G
LIST OF FIGURES -----	I
LIST OF ABBREVIATIONS-----	L
PREFACE-----	O
Chapter 1	1
Introduction	1
1.1. Introduction to QSAR and chemometrics	2
1.2. QSPR/QSAR methodology	4
1.3. Statistical and QSAR/QSPR software's	6
1.4. Data set selection	8
1.5. Data entry	9
1.6. Molecular modeling.....	10
1.7. Molecular descriptors calculation.....	11
1.7.1. Constitutional descriptors	11
1.7.2. Topological descriptors	12
1.7.3. Geometrical descriptors.....	13
1.7.4. Electrostatic descriptors.....	14
1.7.5. Quantum chemical descriptors.....	15
1.7.6. Thermodynamic descriptors	15
1.7.7. Solvation descriptors	16
1.7.8. Constructed descriptors	16
1.8. Feature selection	16
1.8.1. Objective feature selection.....	17
1.8.2. Subjective feature selection	18
1.8.2.1. Stepwise regression method	19

1.8.2.2. Genetic algorithm (GA)	19
1.9. Model construction	23
1.9.1. Data pretreatment	23
1.9.2. Ordinary least squares (OLS) regression.....	24
1.9.3. Multiple linear regressions (MLR).....	25
1.9.3.1. Multicollinearity	26
1.9.4. Principal component regression (PCR)	27
1.9.5. Partial least squares regression (PLS)	29
1.9.6. Neural networks	31
1.9.6.1. Introduction	31
1.9.6.2. Neural network construction	32
1.9.6.3. Artificial neural networks in QSPR modeling.....	36
1.10. Model validation	39
1.11. Applications of QSPR analysis.....	40
1. References	42
Chapter 2	50
QSPR Study for Estimation of Acidity Constants of Some Aromatic Acids Derivatives Using Multiple Linear Regression (MLR) Analysis	50
2.1. Introduction to prediction of acidity constants.....	51
2.2. Materials and methods	52
2.2.1. Computer hardware and software	52
2.2.2. Data set	52
2.2.3. Molecular modeling and structural descriptors.....	56
2.2.4. MLR modeling.....	56
2.3. Results and discussion	57
2.3.1. MLR analysis	58
2.3.2. Statistical parameters.....	63
2.4. Conclusions	64
2. References	66
Chapter 3	68

Artificial Neural Network Based Quantitative Structural Property Relationship for Predicting Boiling Points of Refrigerants	68
3.1. Introduction to predicting of boiling points	69
3.2. Methodology.....	71
3.2.1. Data set	72
3.2.2 Structure entry and Optimization	75
3.2.3. Descriptor generation	75
3.2.4. Genetic algorithm for descriptor selection	75
3.2.5. Artificial neural network modeling.....	77
3.3. Results and discussion	79
3.3.1. ANN analysis	80
3.3.2. Interpretation of the selected descriptors.....	84
3.3.2.1. Vapor pressure ($\ln P_{\text{vap}}$) at 25°C	84
3.3.2.2. First order valence connectivity (${}^1\chi_v$).....	85
3.3.2.3. Second order kier Shape index (${}^2\kappa$).....	86
3.3.2.4. Enthalpy of vaporization at boiling point (ΔH_{vap})	87
3.4. Conclusions	87
3. References	89
Chapter 4.....	91
Quantitative Structure - Micellization Relationship Study of Cationic Surfactants Using Ordinary Least Squares Regression	91
4.1. Introduction to critical micelle concentration of Surfactants	92
4.2. Materials and experimental methods	94
4.2.1. Data set	94
4.2.2. Computer hardware and software	95
4.2.3. Molecular modeling and calculation of descriptors	95
4.2.4. Selection of descriptors	95
4.2.5. Ordinary least squares regression modeling	96
4.3. Results and discussion	97
4.3.1. OLS analysis	97
4.3.2. Interpretation of descriptors.....	103

4. References	107
Chapter 5	110
Quantitative Structure–Property Relationship Study of N-Octanol-Water Partition Coefficients of Some of Diverse Drugs Using Artificial Neural Networks (ANN) And Multiple Linear Regressions (MLR)	110
5.1. Introduction to n-octanol/water partition coefficient.....	111
5.2. Data and methods.....	112
5.2.1. Data set	112
5.2.2. Computer hardware and software	114
5.2.3. Molecular modeling and theoretical molecular descriptors.....	114
5.2.4. Genetic algorithm for descriptor selection	114
5.2.5. Multiple linear regression modeling.....	116
5.2.6. Artificial neural network.....	117
5.3. Results and discussion	118
5.3.1. MLR and ANN analysis	118
5.3.2. Interpretation of descriptors.....	123
5.3.2.1. Molecular volume term	123
5.3.2.2. Hydrophilic-lipophilic balance term	123
5.3.2.3. Hydrogen-bonding term	124
5.3.2.4. Polar surface area term.....	125
5.3.3. Statistical parameters	125
5.4. Conclusions	127
5. References	128
Chapter 6	130
QSRR Prediction of the Chromatographic Retention Behavior of Painkiller Drugs	130
6.1. Introduction to retention time in chromatography.....	131
6.2. Materials and methods	134
6.2.1. Data set	134
6.2.2. Computer hardware and software	136

6.2.3. Molecular modeling and theoretical molecular descriptors.....	136
6.2.4. Stepwise regression for descriptor selection.....	136
6.2.5. Multiple linear regressions (MLR).....	137
6.3. Results and discussion.....	138
6.3.1. MLR analysis	138
6.3.2. The effect of the selected descriptors on the retention time	142
6.3.2.1. N-octanol- water partition coefficient (logP)	142
6.3.2.2. Molecular surface area (SM)	144
6.3.2.3. Hydrophilic-lipophilic balance (HLB).....	144
6.3.3. Statistical parameters.....	145
6.4. Conclusions	146
6. References	148
Chapter 7	151
QSPR Prediction of Aqueous Solubility of Drug-Like Organic Compounds	151
7.1. Introduction to aqueous solubility	152
7.2. Data and methods.....	153
7.2.1. Data set	153
7.2.2. Computer hardware and software	153
7.2.3. Molecular modeling and theoretical molecular descriptors.....	154
7.2.4. Stepwise regression for descriptor selection.....	154
7.2.5. Multiple linear regression modeling.....	155
7.3. Results and discussion	159
7.3.1. MLR analysis	159
7.3.2. Interpretation of descriptors.....	163
7.3.2.1. n-Octanol-water partition coefficient.....	163
7.3.2.2. Molecular volume	164
7.3.2.3. Hydrogen bond forming ability	164
7.3.3. Statistical parameters.....	165
7.4. Conclusions	166
7. References	168

Chapter 8	170
QSPR Modeling of Stability Constants of Diverse 15-Crown-5 Ethers Complexes Using Best Multiple Linear Regressions	170
8.1. Introduction to properties of crown ethers	171
8.2. Materials and methods	173
8.2.1. Data set	173
8.2.2. Molecular modeling and descriptor calculation.....	176
8.2.3. Multilinear regression modeling	177
8.3. Results and discussions	179
8.3.1. BMLR analysis	181
8.3.2. Interpretation of the selected descriptors.....	186
8.4. Conclusions	188
8. References	190

LIST OF TABLES

Table	Page
Chapter1	
Table1.1. Software's in the QSPR/QSAR area, where MM – Molecular Modeling, DC – Descriptors Calculator, SA – Statistical Analysis, and MV – Model Validation.....	8
Chapter2	
Table2.1. Experimental values of pKa for various aromatic acids in water at 25 °C for training (a) and prediction (b) sets.	54
Table2.2. Molecular structure descriptors employed for the proposed QSPR model.	56
Table2.3. Standardized coefficients and other statistical parameters for MLR model.	60
Table2.4. Comparison between experimental and calculated pKa values for external prediction set.....	61
Table2.5. Statistical parameters obtained by applying the MLR method to the test set. ...	64
Chapter3	
Table3.1. Experimental and calculated boiling points of refrigerants by ANN modeling..	73
Table3.2. Correlation matrix of descriptors and boiling points.....	79
Table3.3. Statistical parameters for ANN modeling.....	84
Chapter4	
Table4.1. Experimental values of logCMC cationic surfactants for train set (a-I, a-II) and prediction set (b-I, b-II).....	94
Table4.2. Correlation matrix for the dependence of logCMC with the descriptors.	97
Table4.3. Comparison between experimental and calculated logCMC values for external prediction set.....	101
Chapter5	
Table5.1. Experimental values of logP _{OW} for drug organic compounds at 25 °C for training set.	113
Table5.2. Parameters of the genetic algorithm	116

Table5.3. Molecular descriptors, experimental $\log P_{o/w}$, predicted $\log P_{o/w}$ and residuals values for external prediction set by MLR and ANN methods.	121
Table5.4. Statistical parameters obtained by applying the MLR and ANN models to the test set.	127

Chapter6

Table6.1. Apparatus and analysis conditions for RP-HPLC retention time data.	134
Table6.2. Experimental retention times and molecular descriptors values for 62 drug compounds.	135
Table6.3. Experimental $\log t_R$, predicted $\log t_R$, residuals and percent relative errors values for train and external test sets.	141
Table6.4. Statistical parameters obtained by applying the MLR method to the test set. ...	146

Chapter7

Table7.1. Experimental values of $\log S_w$ for drug-like organic compounds in water at 25 °C for training (a) and prediction (b) sets.	157
Table7.2. Experimental $\log S_w$, molecular descriptors, predicted $\log S_w$, residuals and percent relative errors values for external prediction set.	161

Chapter8

Table8.1. Chemical Structures of fifty-eight 15-crown-5 ethers.	174
Table8.2. The multilinear QSPR model obtained for 46 crown ethers for $\log K$	179
Table8.3. The descriptors values for 58 compounds.	180
Table8.4. Experimental and predicted stability constants of 15-crown-5 ethers complexes.	183

LIST OF FIGURES

Figure	Page
Chapter1	
Figure1.1. QSPR/QSAR model generation flowchart.	5
Figure1.2. 2-D Representation of 2-pentyl-1,4,7,10,13-pentaoxacyclopentadecane.	9
Figure1.3. 3-D Representation of 2-pentyl-1,4,7,10,13-pentaoxacyclopentadecane.	11
Figure1.4. A flow chart describing the working of a genetic algorithm.....	20
Figure1.5. A schematic diagram of the single point crossover operation. The grids on the left represent the parents and the grids on the right represent the children formed after crossover. The portion of the chromosomes to the left of the split point are swapped.	22
Figure1.6. A schematic diagram of the mutation operation.....	22
Figure1.7. Auto-scaling as usually performed prior to QSAR analysis. First column mean-centering followed by column-wise scaling with the inverse of the corresponding standard deviation.	24
Figure1.8. A typical QSAR data set: X is of the dimensions $I \times J$ where $J > I$ with a single response variable y ($I \times 1$)......	25
Figure1.9. A graphical representation of the first two Principal Components. In PCR the component scores $T = [t_1 t_2]$ are the regressors and the following holds true: $t_1^T t_2 = 0$ and $p_1^T p_2 = 0$	28
Figure1.10. A graphical representation of the first two PLS components. In Wolds PLS algorithm, the following holds true: $t_1^T t_2 = 0$, $w_1^T w_2 = 0$ and $p_1^T p_2 \neq 0$	30
Figure1.11. Schematic of a Neural Network.....	33
Figure1.12. Abbreviated Notation for a Neural Network	33
Figure1.13. Common activation functions. Soft nonlinearity: (a) Sigmoid and (b) tanh; Hard non-linearity: (c) Signum and (d) Step.....	34
Figure1.14. A more detailed view of a single hidden layer neuron. The x_i 's represent the input value of the neurons in the preceding layer and w_i 's correspond to the weights for the connections between this neuron and those in the preceding layer. b represents the bias term for this neuron.	38
Figure1.15. Typical three-layer, feed-forward, fully-connected artificial neural network used in QSPR.....	38

Chapter2

- Figure2.1.** The experimental pKa values aromatic acids correlate well with the (a) partial positive charge on the acidic hydrogen and (b) changing of bond length O-H..... 58
- Figure2.2.** (a) Plots of predicted pKa and residuals pKa estimated by MLR modeling versus experimental pKa for test molecules in prediction set. (b) Plots of experimental and predicted pKa values versus sample number of prediction set. 62

Chapter3

- Figure3.1.** The typical architecture of the ANN..... 79
- Figure3.2.** Scatter plot of the calculated versus experimental T_b values for training, validation and testing sets of 90 refrigerants compounds. 82
- Figure3.3.** Scatter plot of the experimental T_b values versus residuals. 82

Chapter4

- Figure4.1.** Three-dimensional structure of (a) $C_{14}H_{29}N^+(Et)_3Br^-$ and (b) $C_{14}H_{29}PY^+Br^-$. ..93
- Figure4.2.** The experimental logCMC values cationic surfactants correlate well with the (a) hydrophobic chain length [L_C] and (b) hydrophobic volume [V_H]. 100
- Figure4.3.** Plots of predicted logCMC and residuals logCMC versus experimental logCMC using model (I) for tetra-alkyl ammonium surfactants. 102
- Figure4.4.** Plots of predicted logCMC and residuals logCMC versus experimental logCMC using model (II) for alkyipyridinium surfactants..... 102
- Figure4.5.** Plots of predicted logCMC and residuals logCMC versus experimental logCMC using simultaneous model for test set of cationic surfactants. 103

Chapter5

- Figure5.1.** Network architecture for studying the $\log P_{o/w}$ of drug. 118
- Figure5.2.** Plot of standardized coefficients versus descriptors in MLR model. 120
- Figure5.3.** Predicted $\log P_{o/w}$ and residuals $\log P_{o/w}$ estimated by MLR modeling vs. experimental $\log P_{o/w}$ 122
- Figure5.4.** Predicted $\log P_{o/w}$ and residuals $\log P_{o/w}$ estimated by ANN modeling vs. experimental 122